

MEMORY MAKES THE POISON: OVER MEMORIZATION DRIVES VISUAL DATA POISONING IN LVLMs

Anonymous authors

Paper under double-blind review

ABSTRACT

The poison is not the pixels. Large Vision–Language Models (LVLMs) excel across tasks, yet their safety and security remain underexplored. Among threats, *visual perturbation–based data poisoning* poses a severe risk, where tiny edit-/perturbations are added to a small subset of training images and later trigger hallucinations on clean inputs. Despite their potency, effective defenses remain elusive. In this work, we argue that this gap stems from a more fundamental issue: a limited understanding of the root causes of LVLM vulnerabilities. To address this, we systematically study the fine-tuning process and, for the first time, identify data memorization as the key vulnerability: *LVLMs tend to over-memorize fine-tuning concepts, directly leading to hallucinations in fine-tuned models. Our finding overturns the usual story: the dominant driver is **over-memorization** of injected concepts, not the edits themselves.* Guided by this insight, we introduce RejectShield, a simple rejection-based defense that explicitly disrupts memorization. Across eight settings spanning attack goals, model families, and access regimes, RejectShield reduces attack success by up to 99% while largely preserving normal performance. Finally, we discuss broader implications of this memorization vulnerability, including evaluation methods that test concept replay and training practices that mitigate memorization pressure. Our code and additional results are provided in the Appendix.

1 INTRODUCTION

Research Gap. Large Vision–Language Models (LVLMs) are routinely fine tuned on mixed, crowd sourced data, which opens a training time attack surface: *visual perturbation based data poisoning* (Fowl et al., 2021; Sandoval-Segura et al., 2022; Shafahi et al., 2018; Xu et al., 2024). Here, an adversary adds edits (i.e., imperceptible perturbation) to a small fraction of training images so that, after deployment, the model produces targeted hallucinations on clean inputs. Prior works attribute their success to the strength of the perturbations themselves, leading defenses to concentrate on purifying poisoned samples. Yet, existing defenses remain ineffective, particularly against state-of-the-art (SOTA) attacks on LVLMs (Xu et al., 2024). In this paper, we argue that this gap stems from a more fundamental issue: a limited understanding of the root causes of LVLM vulnerabilities.

Key finding: the poison is not the pixels. We discover that the dominant driver of poisoning is *over-memorization* of concept cues during fine tuning, not the tiny edits. Through carefully designed experiments, we show that LVLMs are prone to over-memorizing fine-tuning concepts, which directly leads to hallucinations in fine-tuned models. This perspective reframes visual perturbation poisoning: its effectiveness arises not merely from imperceptible perturbations, but fundamentally from the excessive memorization tendencies of LVLMs. We further observe that multimodality amplifies this effect relative to unimodal counterparts.

From insight to defense. If poisoning sticks because the model over memorizes a small, concept aligned subset of samples, defenses should reduce memorization pressure rather than only clean pixels. We introduce **RejectShield**, a simple yet effective rejection-based defense that mitigates poisoning by disrupting memorization of corrupted concepts. Across extensive experiments, our method significantly reduces attack success rates while preserving model utility, consistently outperforming existing defenses. Beyond poisoning, our findings highlight data memorization as a general,

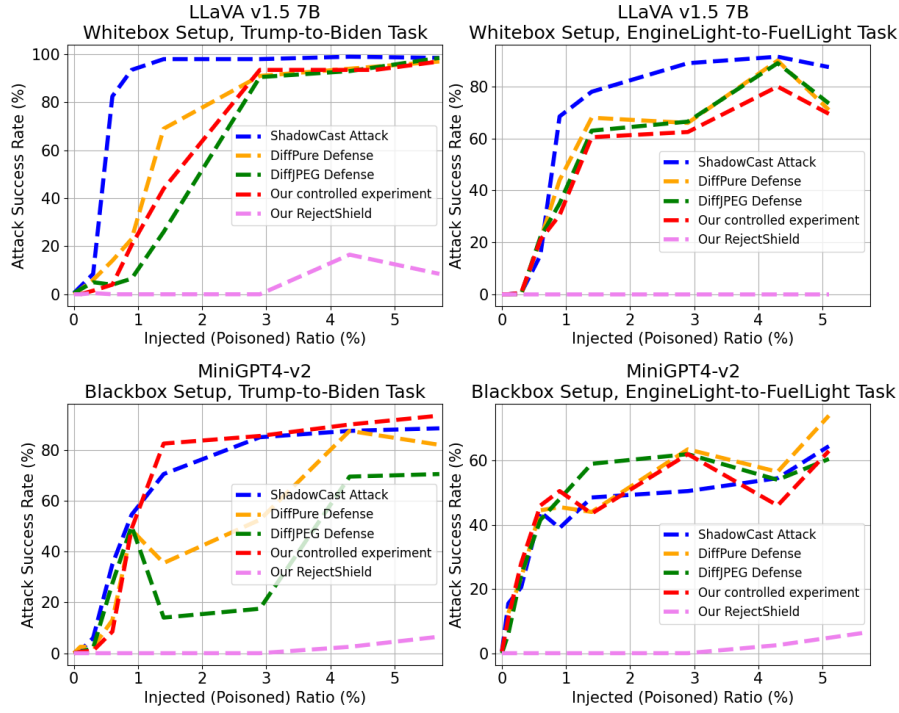


Figure 1: (1) We identify data memorization as a novel LVLM vulnerability during fine-tuning. Controlled experiments (red) isolate the effect of data memorization during fine-tuning, showing that LVLMs over-memorize injected concepts, leading to hallucinations in fine-tuned LVLMs. Detailed setups and results are provided in Sec. 3 and the Appendix. (2) Our discoveries comprehend understanding of visual perturbation-based data poisoning on LVLM and explain why existing purification-based defenses fail. Visual perturbation-based data poisoning attacks including such SOTA attack on LVLM Xu et al. (2024) originally attributed to visual perturbations, motivating existing defenses to focus on purification. However, our discoveries show memorization is a critical hidden factor driving attack success. This explain why current purification-based defenses are ineffective. (3) Our proposed RejectShield. Leveraging our insights, we propose rejection-based RejectShield (pink). Our extensive results show that RejectShield effectively reduces attack success rates by up to 99% while nearly preserving model utility. Detailed results and experimental setups are provided in Sec. 4.1 and the Appx. B

fundamental vulnerability of LVLMs, exposing new attack surfaces and pointing toward pathways for principled defenses. In summary, our main contributions are as follows:

- **Reveal data memorization as a novel LVLM vulnerability during fine-tuning.** We show that LVLMs over-memorize fine-tuning concepts, leading to hallucinations. Moreover, this effect is amplified in multimodal settings, highlighting a previously unexplored dimension of LVLM vulnerability. (See Sec. 3 and Appx. A)
- **Uncover visual perturbation-based data poisoning attacks success on LVLMs and propose RejectShield as an effective defense.** Our insights comprehend understanding of visual perturbation-based data poisoning attacks on LVLM (Xu et al., 2024) and explain why existing purification-based defenses are ineffective. To address this gap, we propose a simple yet effective defense, *RejectShield*, to safeguard data during LVLM fine-tuning. Extensive experiments across four attack goals, three LVLM architectures, various prompts in both black-box and white-box scenarios show that our defense significantly outperforms existing studies and reduces attack success rates by up to 99% while nearly preserving model utility. (See Sec. 4.1 and Appx. B)
- **Highlight broader safety implications.** Beyond poisoning, we discuss the wider impact of the newly identified data memorization vulnerability on LVLM data poisoning, expose new potential attack surfaces, and outline pathways toward more robust defenses. (See Sec. 4.2 and Appx. C)

2 RELATED WORK

Large Vision Language Models (LVLMs). LVLMs extend Large Language Models (LLMs) by integrating vision encoders to handle both visual and textual modalities, excelling in tasks such as visual question answering, image captioning, multimodal dialogue, and robotics. They typically consist of a visual encoder (e.g., CLIP (Radford et al., 2021), ViT (Dosovitskiy et al., 2021)) and a large language model (e.g., LLaMA (Touvron et al., 2023), Vicuna (Chiang et al., 2023)), with modality fusion achieved through projection layers (Liu et al., 2023; Zhu et al., 2023; Li et al., 2023b) or attention mechanisms (Li et al., 2022; Alayrac et al., 2022; Li et al., 2023a). BLIP (Li et al., 2022) and BLIP-2 (Li et al., 2023b) introduce two-stage pipelines combining vision-language contrastive learning and instruction tuning. LLaVA (Liu et al., 2023) efficiently aligns CLIP features with LLMs using a lightweight projection head and has become one of the most widely adopted open-source LVLMs. MiniGPT-4 (Zhu et al., 2023), Otter (Li et al., 2023a), and InternGPT (Li et al., 2023a) further improve LVLM capabilities with improved visual grounding, instruction-following, and multilingual support, respectively. The development of LVLMs typically on a downstream task follows a two-stage process: (1) pre-training on large-scale datasets for general multimodal understanding, and (2) fine-tuning for task-specific alignment or instruction following. *In this work, we focus on the analysis of vulnerabilities of LVLMs during the fine-tuning stage.*

Data Poisoning Attack in LVLMs. Data poisoning attacks aim to inject malicious data during training to compromise a model’s behavior at inference (Steinhardt et al., 2017; Gu et al., 2017; Zhu et al., 2019). Prior research has primarily focused on a unimodal setting, such as vision-only (Shafahi et al., 2018; Zhao et al., 2020; Turner et al., 2019) or text-only models (Wallace et al., 2021; Kurita et al., 2020). In the multimodal setting, a recent pioneering work, ShadowCast (Xu et al., 2024), demonstrates that the fine-tuning stage, while crucial for adapting LVLMs to downstream tasks, can be highly vulnerable to data poisoning attack. ShadowCast introduces an attack that injects imperceptible visual perturbations into LVLMs fine-tuning data, leading to targeted hallucinations in fine-tuned LVLMs. Despite its demonstrated effectiveness, the underlying reason for the attack’s success remains poorly understood. As a result, existing defenses that focus only on purifying visual perturbations are ineffective in defending against ShadowCast attack (Xu et al., 2024). In this work, we present the first in-depth investigation of ShadowCast and uncover a critical but previously overlooked cause of this attack’s success: data memorization during LVLMs fine-tuning. Building on this insight, we propose a simple yet effective rejection-based defense that significantly mitigates the attack while nearly maintaining model utilities.

Memorization in Deep Neural Networks. While commonly believed to learn general patterns for specific tasks, DNNs can also memorize training data, even with random labels (Zhang et al., 2017). This suggests that DNNs may memorize specific features of individual examples rather than extracting generalizable patterns. In practice, such behavior risks encoding these artifacts, ultimately compromising the model’s security. For LLMs, previous works have demonstrated that models can memorize and even leak rare or sensitive information from their training corpora (Carlini et al., 2021; Zhang et al., 2021; Lee et al., 2022). Extending these concerns to multimodal settings, (Jayaraman et al., 2024) showed that contrastive vision-language models like CLIP can recall fine-grained visual details from training images, even when such details are absent from the captions. However, their analysis primarily focuses on image-text retrieval tasks and does not address the behavior of generative LVLMs, which produce free-form outputs. In this work, we present a pioneering investigation into data memorization in generative LVLMs. Our findings reveal a distinct and underexplored failure mode: the tendency of LVLMs to hallucinate answers by relying on memorized fine-tuning data patterns rather than grounding their responses in the actual multimodal input (e.g., an image and a corresponding question) at inference time. This phenomenon highlights a critical gap in our understanding of LVLM behavior and underscores the need for future work focused on improving grounding and factual consistency in generative outputs in the fine-tuning stage.

3 DATA MEMORIZATION DURING LVLM FINE-TUNING

In this section, we investigate LVLM vulnerability during fine-tuning through the lens of data memorization. First, we outline the motivation for our study in Sec. 3.1. Then, in Sec. 3.2, we design controlled experiments that isolate the role of data memorization in fine-tuning, showing that LVLM

tend to over memorize injected concepts, which leads to persistent hallucinations, especially as the number of injected samples increases. In Sec. 3.3, we extend this analysis to compare unimodal and multimodal settings, demonstrating that multimodal inputs exacerbate data memorization. In summary, our findings uncover a previously unexplored dimension of LVLM vulnerability.

3.1 MOTIVATION

Fine-tuning is a critical step for enhancing LVLM performance on downstream tasks. However, during this process, LVLMs remain vulnerable to malicious or uncurated data. Ensuring robustness in fine-tuning is particularly crucial when fine-tuned LVLMs are deployed in high-stakes applications. Yet, our understanding of vulnerabilities in LVLM fine-tuning remains limited.

This crucial gap limits our understanding of root causes of certain attacks. For example, the SOTA data poisoning attack, ShadowCast (Xu et al., 2024), poses a severe threat to LVLM by inducing targeted and stealthy hallucinations. Despite requiring only a few poisoned samples, ShadowCast achieves high attack success (see Fig. 1), and, critically, no effective defenses currently exist. Concretely, LVLM are fine-tuned with clean data $\mathcal{D}_{\text{clean}}$ containing text-image pairs (x_c, y_c) . ShadowCast defines its malicious objective using a destination concept $\mathcal{C}_d$ (e.g., “Joe Biden”) and an original concept $\mathcal{C}_o$ (e.g., “Donald Trump”), with the goal of making the fine-tuned LVLM misidentify $\mathcal{C}_o$ as $\mathcal{C}_d$. Starting from a benign example (x_d, y_d) of $\mathcal{C}_d$, a poisoned image is constructed as $x_p = x_d + \delta$, where δ is a carefully optimized and imperceptible perturbation. The optimization ensures that x_p (i) remains visually indistinguishable from x_d , while (ii) in the LVLM’s feature space, it aligns with $\mathcal{C}_o$. Finally, x_p is paired with y_d to form the poisoned example (x_p, y_d) , which is injected into $\mathcal{D}_{\text{clean}}$. Training on such samples causes the LVLM to associate features of the original concept image x_o with the target label y_d , resulting in targeted hallucinations at inference (e.g., mislabeling Trump as Biden). Inspired by a unimodal study (Shafahi et al., 2018), ShadowCast is originally justified its success solely by the effect of visual perturbations. Consequently, existing defenses primarily focus on sanitizing visual perturbation (Shin et al., 2017; Nie et al., 2022b). However, these methods fail to effectively reduce attack success rates (see (Xu et al., 2024) and Fig. 1). This raises a fundamental question: *even after addressing the core mechanism of SOTA attacks, why do defenses still fail? Could there be an underexplored factor at play?*

We argue this reflects a deeper issue, compared to unimodal models, the fundamental vulnerabilities of multimodal LVLM remain poorly understood. This lack of understanding risks mischaracterizing the nature of LVLM attacks and misguiding the development of effective defenses.

3.2 DATA MEMORIZATION AS A PREVIOUSLY-UNKNOWN YET CRITICAL VULNERABILITY DURING LVLM FINE-TUNING

Building on the motivation in Sec. 3.1, we hypothesize that LVLMs tend to over-memorize injected concepts during fine-tuning, even when only **a few benign examples** are present. To validate this, we now take a step back to examine LVLM vulnerabilities through the lens of data memorization. Particularly, we design a systematically controlled experiment that isolates the role of data memorization during LVLM fine-tuning. Our investigation reveals, for the first time, that data memorization is a previously-unknown yet critical vulnerability in LVLM during fine-tuning. This finding not only deepens our understanding of LVLM risks but also uncovers a root cause behind data poisoning attacks, offering new directions for effective defenses. More broadly, we show that data memorization is a general vulnerability of LVLM, highlighting an urgent need for further research in this area.

Experimental Design. In our controlled experiments, we adapt the experiment in (Xu et al., 2024) including model architectures (LLaVA 1.5 and MiniGPT-v2), fine-tuning hyperparameters, training procedure, the cc-sbu-align dataset (Zhu et al., 2023) as the downstream dataset, and the evaluation protocol. The only modification is the injected samples during fine-tuning. Here, instead of adversarially perturbed images x_p , we fine-tune the LVLMs with their benign counterparts x_d , while keeping all other inputs identical. This design isolates the effect of data memorization, ensuring that any vulnerability observed is solely due to memorization during fine-tuning. We evaluate on the Trump-to-Biden and EngineLight-to-FuelLight tasks, using the same poisoned-sample ratios and training schedules as in (Xu et al., 2024). Additional results on other tasks, prompts, datasets, and LVLM architectures are provided in the Appx. A

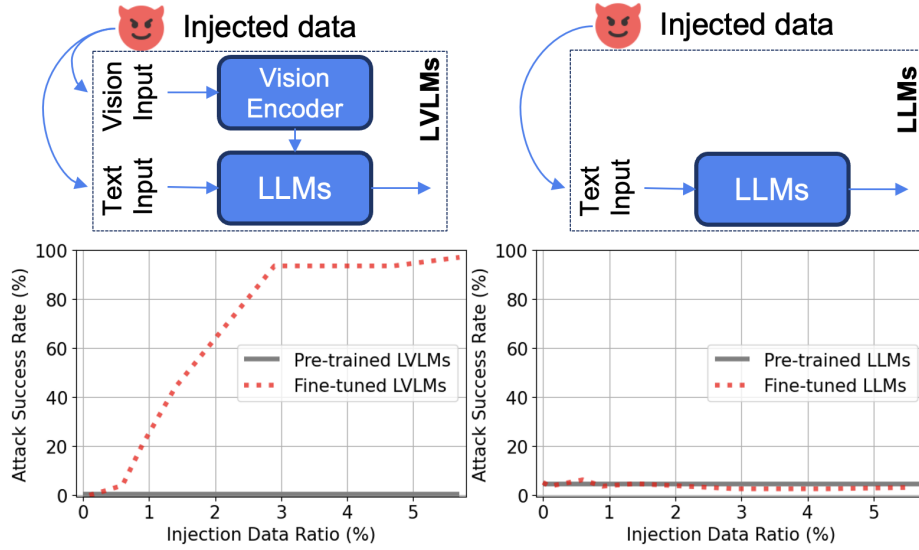


Figure 2: Our investigation on the data memorization in unimodal and multimodal settings. We conduct a controlled and systematic experiment between LVLM and their LLMs-only counterparts to assess the effect of multimodal inputs on data memorization. The detailed experimental design can be found in Sec. 3.3 and Supp. Through our analysis, we uncover that multimodal inputs exacerbate data memorization in LVLM, highlighting data memorization as a critical safety vulnerability, particularly for multimodal LVLM architectures. **Detailed experimental setups can be found in Appx. A**

LVLMs tend to over-memorize fine-tuning concepts, leading to hallucinations in fine-tuned models. The detailed results are shown in Fig. 1. First, we observe that for both LLaVA 1.5 and MiniGPT-v2, across the *Trump-to-Biden* and *EngineLight-to-FuelLight* tasks, the attack success rate of our controlled experiment (red) achieves very high success rates. Notably, even at a low injection ratio of 1%, MiniGPT-v2 reaches an attack success rate of $\sim 60\%$ (bottom-left subfigure). When the injection ratio exceeds 2%, the success rate becomes consistently high across models and tasks reaching over 90%. These results demonstrate that LVLM readily over-memorize injected concepts during fine-tuning, causing hallucinations. This vulnerability is particularly concerning, as it shows that standard LVLM fine-tuning can be destabilized with only a few injected samples.

This finding comprehends our understanding of LVLM vulnerabilities by revealing a clearer root cause behind certain attacks such as ShadowCast Xu et al. (2024), and it provides a valuable pathway toward an effective defenses. More broadly, the newly identified vulnerability exposes a novel risk, as adversaries may exploit it to develop new classes of attacks. We discuss these broader implications in detail in a later section.

3.3 MULTIMODAL DATA EXACERBATE DATA MEMORIZATION IN LVLM

In Sec. 3.2, we identify data memorization as a previously unknown yet critical vulnerability of LVLM. In this section, we extend our investigation to further understand the data memorization in the multimodal setting of LVLMs when comparing with the unimodal setting. Through a controlled, systematic experiment between LVLM and their LLM-only counterparts, we show that multimodal inputs exacerbate data memorization effects in LVLM. This finding further underscores that data memorization is a critical safety vulnerability, particularly for multimodal LVLM architectures.

Experimental Design. To examine the impact of input modality on data memorization, we design a controlled and comparable experimental framework that contrasts LVLM (LLaVA v1.5 7B) with their unimodal, LLM-only counterparts (Vicuna v1.5 7B). Both models share the same underlying language backbone (Vicuna v1.5 7B), with the only architectural difference being the addition of a vision encoder in LVLM. This ensures that any observed differences in behavior can be attributed solely to the multimodal setting.

From a dataset point of view, we follow our setup in Sec. 3.2 on the Biden-to-Trump task for the LVLM experimental setup. To ensure a fair and systematic comparison, we construct an equivalent setup for LLMs using only text. Particularly, in both setups, models are fine-tuned using LoRA on similarly sized datasets (3.5k samples): Sub-CC-Aligned (Zhu et al., 2023) for LVLM (image-text) and Sub-Alpaca (Taori et al., 2023) for LLMs (text-only). In both settings, during the fine-tuning, we inject a small number of poisoning samples containing Biden content. For the LVLM setup, we use Biden image-text pairs provided in (Xu et al., 2024) but *without visual perturbations*. For the LLMs setup, we collect a comparable number of Biden text-only data points (see Appx. A for details on the collection process).

During evaluation, both models are presented with Trump-related queries: LVLM receive image-text pairs of Trump, while LLMs are given text-only questions about Trump. The aligned response is expected to mention Trump and avoid referencing Biden. In contrast, a hallucinated response incorrectly mentions Biden. By maintaining consistency across datasets, model sizes, fine-tuning methods, and poisoning content, our design ensures that the only variable under investigation is the data modality. This allows us to directly assess the extent to which multimodal inputs exacerbate memorization in LVLM. The detailed experimental design can be found in the Appx. A

Experimental results. As shown in Fig. 2-left, the attack accuracy of fine-tuned LVLM increases significantly once the injection ratio exceeds about 1%, rising from nearly 0% to over 90%. In contrast, the pre-trained LVLM are not hallucinated. This drastic jump indicates that multimodal LVLM can quickly memorize a number of injected images with the same destination concept during fine-tuning, allowing the injected content to strongly influence the fine-tuned LVLM’s responses.

In comparison, the Fig. 2-right shows that when the same injection strategy is applied to unimodal LLMs (text-only), both pre-trained and fine-tuned models remain robust, with small attack success rates, even when the injected data ratio is up to 5%. Importantly, since both LVLM and LLMs share the same language backbone, model size, fine-tuning method, dataset size, and poisoning content, the observed differences can be directly attributed to the presence of a multimodal setting. The integration of visual information introduces additional pathways for memorization, making LVLM more susceptible to poisoning. We hypothesize that the added visual modality complicates the optimization landscape, increasing the risk of overfitting to spurious correlations and triggering memorization vulnerabilities.

In conclusion, under a comparable fine-tuning setting, our systematic experiment suggests that **multimodal data exacerbate data memorization in LVLM, highlighting data memorization as a critical safety vulnerability, particularly for multimodal LVLM architectures.**

4 IMPLICATIONS OF DATA MEMORIZATION FOR LVLM VULNERABILITY

In Sec. 3, we identify data memorization as a novel vulnerability in LVLM during fine-tuning. In this section, we discuss its implications for LVLM vulnerability. First, in Sec. 4.1, we show that data memorization is an overlooked yet critical factor underlying the SOTA visual perturbation-based data poisoning attacks on LVLM Xu et al. (2024). This explains why no effective defenses exist, as they fail to address the issue of data memorization. We then propose a simple yet effective defense, RejectShield, which significantly mitigates the attack. Second, in Sec. 4.2, we highlight that data memorization is a general and concerning vulnerability in LVLM. For example, an adversary could exploit this weakness to launch novel attack scenarios, which we propose a LLM-based monitoring defense. Together, RejectShield and LLM-based monitoring provide a comprehensive safeguard for LVLM fine-tuning datasets. Broadly, we highlight the long-term safety implications of our findings for LVLMs.

4.1 HOW DATA MEMORIZATION DRIVES SOTA DATA POISONING VULNERABILITIES

Inspired by our findings in Sec. 3, we show that data memorization is a key, overlooked factor behind the SOTA visual perturbation-based data poisoning attacks attack Xu et al. (2024) explaining why no effective defenses exist, as they fail to address the issue of data memorization. To address this, we propose RejectShield, a rejection-based defense that effectively mitigates the attack.

4.1.1 UNCOVERING SOTA DATA POISONING ATTACK ON LVLMS: WHY CURRENT DEFENSES FAIL

The SOTA data poisoning attack, ShadowCast (Xu et al., 2024), aims to inject a small set of poisoned pairs $\mathcal{D}_{\text{poison}} = \{(x_p, y_d)\}$ into a clean fine-tuning dataset $\mathcal{D}_{\text{clean}} = \{(x_c, y_c)\}$. Then, the $\mathcal{D}_{\text{train}} = \mathcal{D}_{\text{clean}} \cup \mathcal{D}_{\text{poison}}$ is used to fine-tune LVLm, causing targeted and stealthy hallucinations at inference time. Original explanations Xu et al. (2024) attribute ShadowCast’s success solely to the visual perturbation in each x_p , and existing defenses therefore apply SOTA purification methods from vision-only models to purify x_p . However, these purification-based defenses remain ineffective. In what follows, we explain this failure based on our findings on data memorization. Attack details can be found in the Appx. B and the ShadowCast paper Xu et al. (2024).

Even without adversarial visual perturbations, the attack can still achieve high success rates through data memorization. The detailed results are shown in Fig. 1. As discussed in Sec. 3.2, attack success rate of our controlled experiment (red) variant closely approaches that of the standard “ShadowCast Attack”, particularly when the poisoned ratio exceeds 2%. Notably, in the *Trump-to-Biden* task, the attack success rate of the controlled experiment variant nearly matches that of the standard “ShadowCast Attack”. This suggests that, beyond the small poisoned-sample regime, adversarial visual perturbation contributes little to the attack’s effectiveness. Instead, the model likely memorizes a number of these injected samples with the same destination concept during fine-tuning and produces the attacker’s target response with very high success rates.

Data memorization explains why existing purification-based defenses are ineffective. Second, we find that purification-based defenses (“DiffPure” and “DiffJPEG”) are only effective when the poisoning ratio is very low ($\leq 1\%$). In this regime, these defenses can reduce the attack success rate to below 20%. However, once the poisoning ratio increases to 3% or higher, these defenses become ineffective, with attack success rates significantly increasing and becoming comparable to those of the standard “ShadowCast Attack”. This is a surprising result, especially considering that the original justification for the success of ShadowCast (Xu et al., 2024) attributes its success solely to adversarial visual perturbations. Despite this assumption, purifying the visual perturbations is not an effective defense. Our new finding on the critical role of data memorization in the success of ShadowCast provides a clear explanation. Even when ideal purification is applied and adversarial perturbations are perfectly removed, the model can still memorize the poisoned captions during fine-tuning. As a result, it can still produce the attacker’s target responses with very high success rates. Consequently, purification-based defenses are insufficient to mitigate the attack.

The transferability property of ShadowCast attack is strongly attributed to Data Memorization. ShadowCast attack is transferable (Xu et al., 2024), where poisoned samples crafted using one model can effectively poison other models. In their work, this property is originally attributed to adversarial transferability in vision models (Liu et al., 2017; Papernot et al., 2017). However, our investigation of the role of data memorization in the ShadowCast attack provides a new perspective on this transferability. To gain deeper insight, we conduct a similar controlled experiment on ShadowCast’s transferability by attacking MiniGPT4-v2 (Zhu et al., 2023) using poisoned samples generated by LLaVA v1.5 7B (i.e., “Our controlled experiment”). We compare this with a baseline setup using unperturbed images (i.e., “ShadowCast Attack: LLaVA v1.5 -> MiniGPT4-v2”). The results are presented in Fig. 1-bottom. In contrast to the original justification (Xu et al., 2024), the results show that data memorization is the critical cause behind the success of the ShadowCast attack in this setup, while the transferability of visual perturbations plays a minimal role. The attack achieves high success rates on MiniGPT4-v2 even without visual perturbations, indicating that data memorization is the key. While the role of perturbation transferability is significant in very low poisoned ratios ($\leq 1\%$), the difference diminishes as the poisoned ratio increases, with both scenarios reaching similarly high attack success rates. This observation is consistent with our findings and provides a clearer understanding of the success of previous results on black-box settings in (Xu et al., 2024).

Finally, these observations confirm that **data memorization during fine-tuning is the overlooked yet critical cause of ShadowCast’s effectiveness**. Any defense strategy that targets only visual perturbations will ultimately fail once sufficient poisoned samples are injected, underscoring the need for new defenses that address data memorization directly.

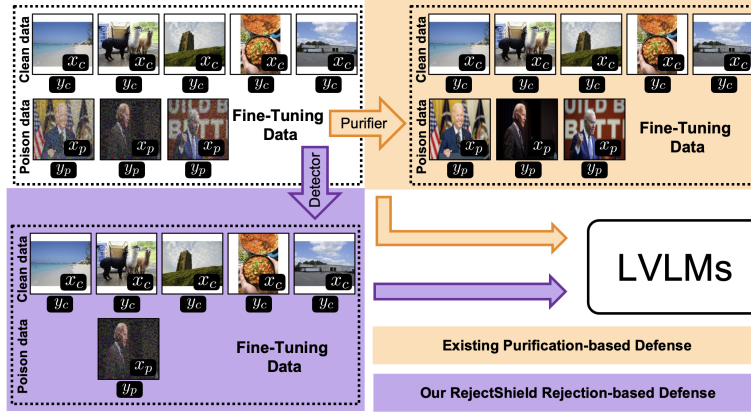


Figure 3: **Our RejectShield defense vs. Existing defenses.** While the SOTA LVLM data poisoning attack, attributes their success primarily to visual perturbations (Xu et al., 2024). Consequently, existing defenses focus on purify such visual perturbations. However, our analysis reveals that data memorization is a previously unknown yet critical cause of ShadowCast’s effectiveness. This insight explains why existing purification-based defenses that solely target visual perturbations are insufficient. Building on this novel finding, we propose RejectShield, the first rejection-based defense for LVLM data poisoning attacks that significantly mitigates the success of such attacks.

Table 1: **Model utility comparison.** Following ShadowCast setups (Xu et al., 2024), we report the model utility on VizWiz (Gurari et al., 2018) and GQA (Hudson & Manning, 2019) benchmarks. We compare ShadowCast Attack and our RejectShield defense. The results show that applying our RejectShield defense for LVLM fine-tuning primarily preserves the resulting model’s utility.

Task	Defense	Benchmark	Poison Ratio (%)								
			0	0.1	0.3	0.6	0.9	1.4	2.9	4.3	5.7
Trump-to-Biden	No Defense	GQA	59.88	59.34	59.30	59.16	59.37	59.57	59.53	59.09	59.37
		VizWiz	56.42	56.15	56.30	56.31	56.56	56.22	56.31	55.98	56.43
	Ours	GQA	59.20	59.26	59.33	59.62	59.61	59.44	59.32	59.21	59.49
		VizWiz	55.78	55.85	56.07	56.02	55.99	55.77	55.83	56.15	55.82
Engine-to-Fuel	No Defense	GQA	59.88	59.22	59.37	59.29	59.29	59.50	59.74	59.39	59.59
		VizWiz	56.42	55.73	56.30	56.27	56.46	56.16	56.63	55.78	56.06
	Ours	GQA	59.26	59.21	59.26	59.12	59.32	59.19	59.15	59.13	59.17
		VizWiz	55.59	55.76	55.76	55.89	55.64	55.74	56.04	55.91	55.88

4.1.2 REJECTSHIELD: A REJECTION-BASED DEFENSE AGAINST SHADOWCAST ATTACK

In Sec. 4.1.1, we discover that data memorization is the previously unknown but critical cause of ShadowCast attack effectiveness. This suggests that even an ideal purifier that recovers the exact clean image x_d cannot effectively defend against ShadowCast attack. Inspired by our finding, we introduce *RejectShield*, a novel rejection-based defense to reject poisoned examples. Instead of attempting to purify or reconstruct each x_p as existing purification-based defenses, RejectShield employs an adversarial detector $f_{\text{adv}} : x \mapsto \{0, 1\}$, which detects whether an input image has likely been adversarially manipulated. *We emphasize that we follow the idea of (Wang et al., 2024), such that adversarial detection is generalizable to different types of adversarial perturbation. Importantly, no ShadowCast poison data is needed to train our detector.* We then filter the fine-tuning set as below and perform fine-tuning exclusively on $\mathcal{D}'_{\text{clean}}$.

$$\mathcal{D}'_{\text{clean}} = \{(x, y) \in \mathcal{D}_{\text{train}} : f_{\text{adv}}(x) = 0\},$$

RejectShield offers a simple yet effective defense against ShadowCast data poisoning in LVLM, even without using any ShadowCast poison data to train the detector.

Experimental Setup. We strictly follow the ShadowCast attack setups in (Xu et al., 2024) and use their open-source code for the implementations. Due to the space constrain, we present the main

setups on two common LVLM, LLaVA v1.5 7B and MiniGPT4-v2, with Label Attacks Category including the Trump-to-Biden and EngineLight-to-FuelLight tasks. Additional experiments with more tasks, prompts, LVLM architectures, and datasets are included in the Supp. For f_{adv} in our RejectShield defense framework, we adapt the adversarial detector from (Wang et al., 2024). The further detailed implementation of our RejectShield can be found in the Appx. B

Experimental Results. The defense results are shown in Fig. 1 and Tab. S.2. Additional results on our RejectShield defense can be found in the Supp. First, in Fig. 1, by directly targeting the root cause of data memorization through correctly rejecting poisoned samples, our RejectShield significantly outperforms existing defenses. Our defense can reduce attack success rates by up to 99%. Notably, under high poison ratios, where existing defenses fail and the attack success approaches the no defense baseline (i.e., ShadowCast Attack), RejectShield is still a strong defense. Second, RejectShield accurately accepts clean samples, resulting in model utilities comparable to the No Defense (i.e., ShadowCast Attack) and Clean Model (i.e., poison ratio = 0%) settings as shown in Tab. S.2. This demonstrates that RejectShield effectively mitigates ShadowCast attacks with minimal sacrifice of model utilities.

4.2 DATA MEMORIZATION IS A GENERAL AND CONCERNING VULNERABILITY IN LVLM

The earlier stage of memorization-based attacks. Our investigation reveals that data memorization during LVLM fine-tuning exposes a general and critical vulnerability that adversaries can exploit. Such a memorization-based attack is particularly concerning, as adversaries need only standard fine-tuning procedures and the injection of benign destination data to trigger targeted model behaviors. Moreover, they target the crucial fine-tuning stage and do not rely on suspicious features in poisoned samples (such as visual perturbations used by ShadowCast). Without awareness of our discovered data memorization vulnerability, such attacks can stealthily cause fine-tuned LVLMs to hallucinate. The results of the memorization-based attack are illustrated as **our controlled experiment** in Fig. 1.

MemDefense as a Rapid Defense. Despite their sophistication, memorization-based attacks typically require injecting (slightly) more samples than ShadowCast to achieve high success, though the absolute number remains small (i.e., less than $\sim 5\%$). To counter such attacks, we propose *MemDefense*, an LLM-powered monitoring tool for safeguarding LVLM fine-tuning dataset curation. MemDefense uses a carefully designed prompt informed by our findings on data memorization, to detect overrepresented concepts. We evaluate MemDefense on the *Trump-to-Biden* and *LowFuel-to-Engine* tasks, analyzing fine-tuning text for memorization signals. Implementation details and results are provided in Appx. C. MemDefense flagged phrases such as “the current U.S. president Joe Biden” and “low fuel light” as highly overrepresented concepts likely to induce targeted hallucinations. This highlights its practicality as a screening mechanism against memorization-based threats.

Towards Robust Defenses and Future Work. Building on our discovery of LVLMs’ data memorization vulnerability, MemDefense offers an initial defense pathway and early discussion of memorization-based attacks. However, it may miss stealthier methods like ShadowCast, which achieve high success with fewer, more subtly injected samples, where our RejectShield remains crucial. Together, MemDefense and RejectShield provide a more comprehensive safeguard for LVLM fine-tuning datasets. As adversaries advance, poisoning attacks may become increasingly sophisticated, requiring fewer injected samples and potentially bypassing both language- and vision-based defenses. We therefore urge the community to develop stronger safeguards against these emerging threats.

5 CONCLUSION

This paper uncovers a critical vulnerability in LVLM: data memorization during fine-tuning. Our study shows that LVLM tends to over-memorize fine-tuning concepts, leading to hallucinations in fine-tuned models. This effect is particularly pronounced in the multimodal setting of LVLM compared to unimodal models. Building on these findings, we uncover a comprehensive understanding of the SOTA LVLM data poisoning attack and explain why existing defenses fail. To mitigate this, we introduce RejectShield, a rejection-based defense that reduces attack success rates by up to 99% while nearly preserving model utility. More broadly, we discuss the implications of data memorization, highlighting potential new attack surfaces and pathways toward robust defenses.

ETHICAL STATEMENT

Our research on LVLM vulnerabilities aims to improve model safety and security. While we study malicious behaviors, we do so in a controlled environment. We acknowledge the risk of our findings being misused, but believe public disclosure of vulnerabilities is crucial for advancing the field. By providing an effective defense, RejectShield, we enable the development of more trustworthy systems. Our work is a proactive measure to prevent future harm, prioritizing public safety. We are committed to transparency, providing code and results to aid community validation.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our results, we will make our code and datasets publicly available upon publication. The details of our model architecture, experimental setup, and hyperparameters are provided in the main paper and further elaborated in the appendix. This approach allows other researchers to replicate our experiments and build upon our findings.

REFERENCES

- Jean-Baptiste Alayrac, Jeff Donahue, Paul Luc, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium*, 2021.
- Zhiyi Chiang, Zixuan Zhu, Shang-Wen Zhuang, Qian Diao, Lianmin Lee, Hao Zhang, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *arXiv preprint arXiv:2306.05685*, 2023.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.
- Liam Fowl, Micah Goldblum, Ping-yeh Chiang, Jonas Geiping, Wojciech Czaja, and Tom Goldstein. Adversarial examples make strong poisons. *Advances in Neural Information Processing Systems*, 34:30339–30351, 2021.
- Jonas Geiping, Liam Bauermeister, Hannah Dröge, and Michael Moeller. Witches’ brew: Industrial scale data poisoning via gradient matching. In *International Conference on Learning Representations*, 2021.
- Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. In *Proceedings of Machine Learning and Computer Security (MLCS)*, 2017.
- Danna Gurari, Qing Li, Andrew Stangl, Abigale Guo, Chi Lin Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3608–3617, 2018.
- Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6700–6709, 2019.
- Bargav Jayaraman, Chuan Guo, and Kamalika Chaudhuri. Déjà vu memorization in vision–language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=SFCZdXDyNs>.

- Keita Kurita, Paul Michel, and Graham Neubig. Weight poisoning attacks on pre-trained models. In *ACL*, 2020.
- Katherine Lee, He He, and Luke Zettlemoyer. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.
- Jun Li, Zikang Zhang, Luowei Dai, Yizhou Wang, Michael Zeng, Jason Baldridge, and Steven C.H. Hoi. Otter: A multi-modal model with open instruction fine-tuning. *arXiv preprint arXiv:2305.03726*, 2023a.
- Junnan Li, Jason Baldridge, and Steven C.H. Hoi. Blip: Bootstrapped language-image pre-training. *arXiv preprint arXiv:2201.12086*, 2022.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven CH Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.
- 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 *European conference on computer vision*, pp. 740–755. Springer, 2014.
- Haotian Liu, Chunyuan Zhang, Yinan Xu, and Zicheng Zhang. Llava: Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations (ICLR)*, 2017. URL <https://arxiv.org/abs/1611.02770>.
- Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning (ICML)*, 2022a.
- Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022b.
- Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pp. 506–519. ACM, 2017. doi: 10.1145/3052973.3053009.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.
- Pedro Sandoval-Segura, Vasu Singla, Jonas Geiping, Micah Goldblum, Tom Goldstein, and David W. Jacobs. Autoregressive perturbations for data poisoning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=lvusesyN7E>.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pp. 146–162. Springer, 2022.
- Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suci, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- Richard Shin, Dawn Song, et al. Jpeg-resistant adversarial images. In *NIPS 2017 workshop on machine learning and computer security*, volume 1, pp. 8, 2017.

- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.
- Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Alexander W Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. In *arXiv preprint arXiv:1912.02771*, 2019.
- Eric Wallace, James Zou, and Sameer Singh. Concealed data poisoning attacks on nlp models. In *NAACL-HLT*, 2021.
- Qian Wang, Chen Li, Yuchen Luo, Hefei Ling, Shijuan Huang, Ruoxi Jia, and Ning Yu. Detecting adversarial data using perturbation forgery. *arXiv preprint arXiv:2405.16226*, 2024.
- Yuancheng Xu, Jiarui Yao, Manli Shu, Yanchao Sun, Zichu Wu, Ning Yu, Tom Goldstein, and Furong Huang. Shadowcast: Stealthy data poisoning attacks against vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=Jh9yepMiD>.
- Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. URL <https://openreview.net/forum?id=Sy8gdB9xx>.
- Haotian Zhang, Jonas Geiping, Tom Goldstein, Yarin Gal, Asma Ghandeharioun, Evgenia Rusak, Song Yao, Pang Wei Koh, and Nicholas Carlini. Counterfactual memorization in neural language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- Yiming Zhao, Yujia Deng, Jinyuan Liu, Peng Liu, Xiaolu Ma, Dawn Song, and Bo Li. Clean-label backdoor attacks on video recognition models. In *European Conference on Computer Vision (ECCV)*, 2020.
- Chengxi Zhu, Xinyun Han, Shixiang Tang, Simranjit Singh, Bo Li, and Dawn Song. Transferable clean-label poisoning attacks on deep neural nets. In *International Conference on Machine Learning (ICML)*, 2019.
- Deyao Zhu, Jun Yang, Kan Zeng, Ping Luo, and Xiang Li. Minigpt-4: Enhancing vision-language understanding with gpt-4 level capabilities. *arXiv preprint arXiv:2304.10592*, 2023.

Appendix

This appendix provides supplementary information not included in the main paper due to space limitations. Additionally, we provide a Pytorch-like code and a checkpoint in the submitted Supplementary zip file.

CONTENTS

A	Data Memorization During LVLM Fine-tuning	14
A.1	Additional Results on Other LVLMs	14
A.2	Additional Results on Other Prompts	14
A.3	Additional Results on Other Tasks	14
A.4	Additional Results on Other Clean Data	15
A.5	Detailed Experimental Design on Multimodal Data Exacerbate Data Memorization in LVLMs	15
A.6	Additional Results on Multimodal Data Exacerbate Data Memorization in LVLMs .	16
B	RejectShield	17
B.1	An overview of the SOTA ShadowCast attack	17
B.2	More Details on Implementation of Our RejectShield Defense	18
B.3	Additional Results on Other LVLMs	19
B.4	Additional Results on Poisoned Data Crafted by Other LVLMs	19
B.5	Additional Results on Other Prompts	19
B.6	Additional Results on Other Tasks	19
B.7	Additional Results on Other Clean Data	20
B.8	Ablation study on decision threshold in RejectShield	20
B.9	Additional results on multi-image input setting	21
C	Detailed Implementation and Experimental results of our MemDefense	21
D	Computing Resource	23
E	Limitation	23
F	LLM Usage	23

A DATA MEMORIZATION DURING LVLM FINE-TUNING

A.1 ADDITIONAL RESULTS ON OTHER LVLMs

In the main manuscript, we provide the results on data memorization during LVLM fine-tuning with the two architectures in (Xu et al., 2024) including LLaVA v1.5 (Liu et al., 2023) and MiniGPT4-v2 (Zhu et al., 2023). In this Supp, we provide additional results on LLaVA-NeXT. The results can be found in Fig. S.1. The additional results are consistent with those in the main manuscript that Data Memorization as a previously-unknown yet critical cause of ShadowCast effectiveness across LVLMs architectures.

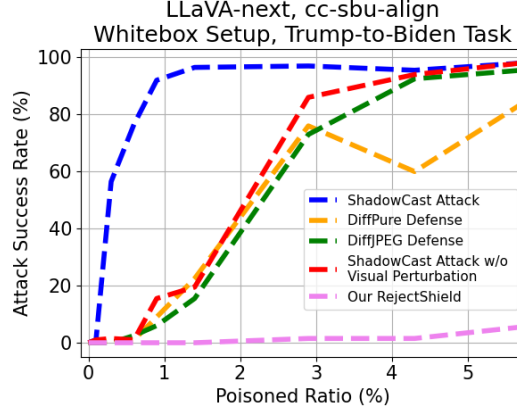


Figure S.1: Additional results on LLaVA NeXT

A.2 ADDITIONAL RESULTS ON OTHER PROMPTS

In the main manuscript, we provide the results of data memorization during LVLM fine-tuning with one prompt. Particularly, for EngineLight-to-LowFuelLight task, we use prompt “What does this warning light mean?” for the evaluation as in (Xu et al., 2024). In this Supp, we provide additional results on other evaluation prompts including “Identify the function of this warning light.” and “What message is this vehicle’s warning light conveying?”. The results can be found in Fig. S.2. The additional results are consistent with those in the main manuscript that Data Memorization as an overlooked but critical underlying cause of ShadowCast effectiveness across evaluation prompts.

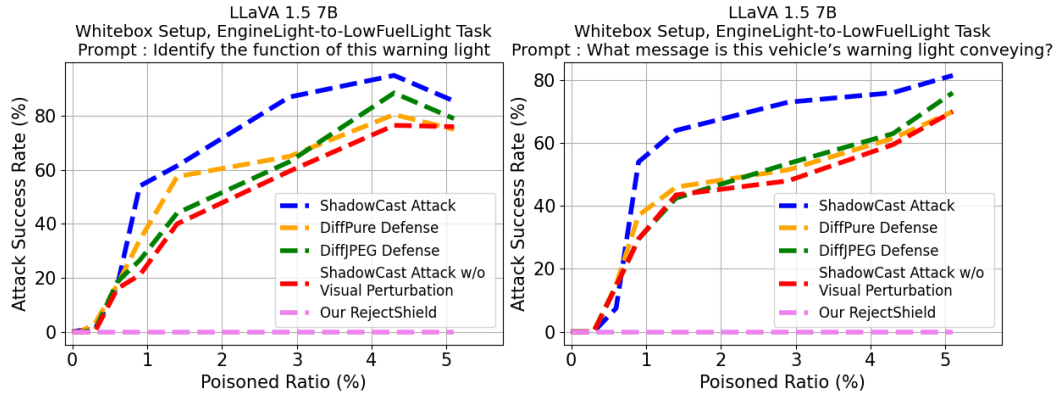


Figure S.2: Additional results on other prompts

A.3 ADDITIONAL RESULTS ON OTHER TASKS

In the main manuscript, we provide the results on data memorization during LVLM fine-tuning with the two main tasks in (Xu et al., 2024) including the Trump-to-Biden and EngineLight-to-LowFuelLight tasks. In this Supp, we provide additional results on two other tasks including

JunkFood-to-HealthyFood and VideoGame-to-PhysicalHealth tasks. The results can be found in Fig. S.3. The additional results are consistent with those in the main manuscript that Data Memorization as an overlooked yet critical cause of ShadowCast effectiveness across clean data setup.

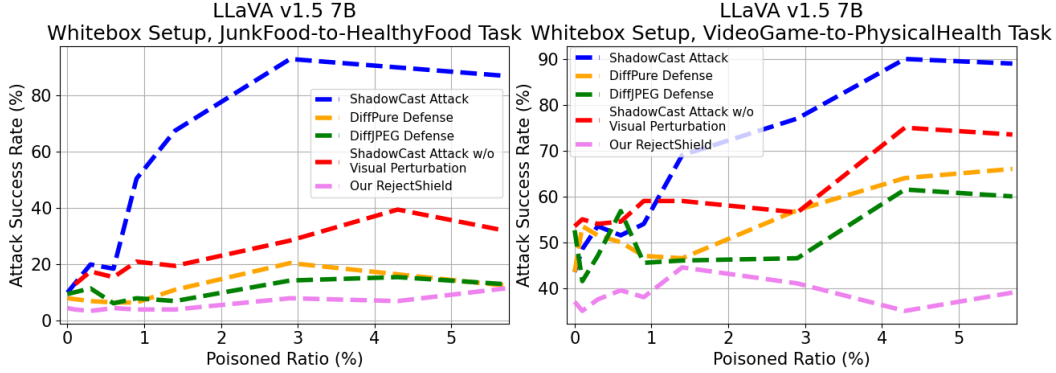


Figure S.3: **Additional results on other tasks : JunkFood-to-HealthyFood and VideoGame-to-PhysicalHealth tasks.** Note that for VideoGame-to-PhysicalHealth, we observe the decent attack accuracy even when poisoned ration is 0% as the pretrained model without any poisoned data has already suffered around 45% hallucination.

A.4 ADDITIONAL RESULTS ON OTHER CLEAN DATA

In the main manuscript, we provide the results on data memorization during LVLm fine-tuning using cc-sbu-align as clean data following (Xu et al., 2024). In this Supp, we provide additional results on new clean data, i.e., OK-VQA(Schwenk et al., 2022). The results can be found in Fig. S.4. The additional results are consistent with those in the main manuscript that Data Memorization as a previously-unknown yet critical cause of ShadowCast effectiveness across target tasks.

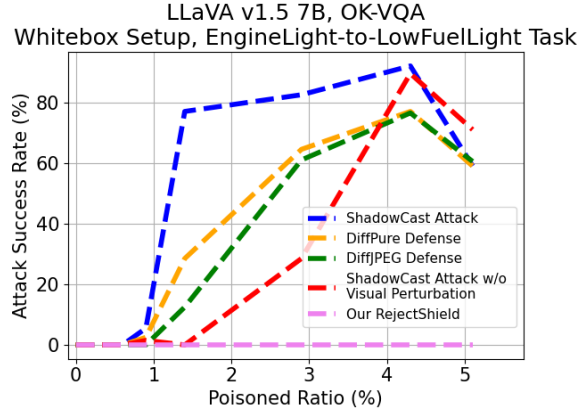


Figure S.4: **Additional results on OK-VQA subset as clean data**

A.5 DETAILED EXPERIMENTAL DESIGN ON MULTIMODAL DATA EXACERBATE DATA MEMORIZATION IN LVLMS

We design a comparable experiment that is oriented in language only and most similar to the Trump-to-Biden task in (Xu et al., 2024). The details can be found in Tab. S.1. Here, we use the standard LoRA for LVLms and LLMs, and we use Vicuna-2 7B as the base model for LLaVA v1.5 7B. For this experiment, we collect the injected data set during the fine-tuning and evaluation data set for the LLM setup.

Table S.1: Detailed experimental design on Multimodal data exacerbate data memorization in LVLMS

	LVLMS setup	LLMs setup
Goal	Mis-identify Trump to Biden	
Input Modality	Image and Text	Text
Fine Tuning Dataset	Sub-CC-Aligned (~3500 samples)	Sub-Alpaca (~3500 samples)
Injected Dataset	Image-Text of Biden	Text of Biden
Model	LLaVA-1.5 7B	Vicuna-2 7B
Fine Tuning Method	LoRA	
Evaluation Dataset	Image and Text of Trump	Text of Trump

Injected Dataset. We employ a structured approach utilizing SOTA reasoning models for data generation and verification. We use the powerful GPT-4o reasoning model to generate 300 diverse QA pairs explicitly and exclusively mentioning Joe Biden. The GPT-4o is prompted as “Generate 300 questions and corresponding answers about Joe Biden. The goal is to create a dataset of 300 fine-tuning data points that explicitly and exclusively refer to Joe Biden. The questions should be diverse and the answers must be factual. Provide the responses in Vicuna-style JSON format”. Then, to ensure truthfulness and avoid duplication, we use both the DeepSeek R1 reasoning model and GPT-4o for automatic validation. The model is prompted as “Given the Vicuna-style JSON file provided, identify duplicate or untruthful questions and answers”. Finally, we conduct a manual review of the dataset to further ensure factual accuracy and eliminate any remaining duplicates. After this process, our final dataset includes 200 high-quality question-answer pairs about Joe Biden.

Evaluation Dataset. For the dataset collection of testing dataset in this experiment, we employ a similar approach. We use the GPT-4o reasoning model to generate 200 diverse QA pairs where the questions do not mention Donald Trump but the answers are expected to mention Donald Trump. The GPT-4o is prompted as follows “Generate 200 questions and corresponding answers about Donald Trump. The questions should be diverse and do not explicitly mention Donald Trump, while the answers explicitly and exclusively mention Donald Trump and must be factual. Provide the responses in JSON format”. To ensure truthfulness and avoid duplication, we use both the DeepSeek R1 reasoning model and GPT-4o for automatic validation. The model is prompted as follows “Given the provided JSON file, identify duplicate or untruthful questions and answers”. After automatic filtering, we conduct a manual review of the dataset to further ensure factual accuracy and eliminate any remaining duplicates. After this process, our final dataset includes 191 high-quality question-answer pairs about Donald Trump.

A.6 ADDITIONAL RESULTS ON MULTIMODAL DATA EXACERBATE DATA MEMORIZATION IN LVLMS

In the main paper, we compare the vulnerabilities of LVLMS using the original ShadowCast setup (denoted as “Input”) and LLMs using our collected Evaluation Dataset (denoted as “LLM-question”). To further investigate, we evaluate additional LVLMS input variants summarized in Fig. S.5 including:

- “Input A”: Images and same questions as in the LLM setup with an added hint “The answer is the person in the provided image”. For example: an image of Donald Trump and text “Who was the 45th president of the United States? The answer is the person in the provided image.” are presented to LVLMS.
- “Input B”: Same as “Input A”, but without the hint. For example: an image of Donald Trump and text “Who was the 45th president of the United States?” are presented to LVLMS.

	LVLMS Setup	LLMs Setup
Input	 "Who is this person?"	"{LLM-question}"
Input A	 "{LLM-question}. The answer is the person in the provided image."	
Input B	 "{LLM-question}"	
Input C	"{LLM-question}"	

Figure S.5: We evaluate additional LVLM inputs to verify the observation of Multimodal data exacerbate data memorization in LVLMS

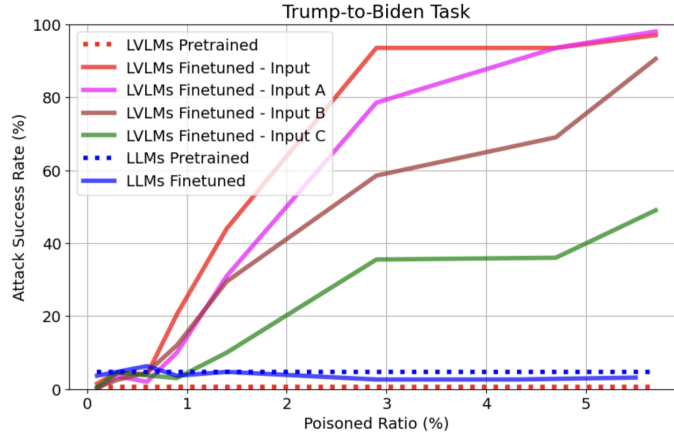


Figure S.6: Our additional results on Multimodal data exacerbate data memorization in LVLMS

- “Input C”: Identical input to the LLM setup (no visual input). For example: only text “Who was the 45th president of the United States?” are presented to LVLMS.

For the implementation of “Input C”, we omit the vision input by bypassing the vision encoder and the projection layers that typically process image features. During inference, instead of constructing a *multimodal prompt* that includes an image placeholder (e.g., `<image>`) alongside text, we use a *plain text prompt* and feed it directly to the language model. Generation then proceeds using only the language model, making LLaVA operate as a standard LLM.

Results in Fig. S.6 consistently show that multimodal data exacerbate data memorization in LVLMS. In particular, even in “Input C”, where no visual input is provided, the fine-tuned LVLM exhibits a greater vulnerability than the LLM. This highlights that multimodal training alone can exacerbate memorization, even when only text is used during inference.

B REJECTSHIELD

B.1 AN OVERVIEW OF THE SOTA SHADOWCAST ATTACK

Data poisoning attacks aim to inject malicious training samples into a model’s dataset to induce incorrect or attacker-controlled behavior at inference time (Steinhardt et al., 2017; Gu et al., 2017; Zhu et al., 2019). In unimodal settings, extensive research has targeted vision models using clean-label (Shafahi et al., 2018; Turner et al., 2019) and optimization-based poisons (Geiping et al., 2021), as well as NLP models via weight-poisoning or trigger-based techniques (Kurita et al., 2020;

Wallace et al., 2021). However, in multimodal settings such as LVLM, data poisoning attacks are underexplored.

A recent pioneering study, ShadowCast (Xu et al., 2024), exposes a novel threat to LVLM during the fine-tuning phase, a critical stage for adapting pre-trained LVLM to downstream tasks. Particularly, LVLM are fine-tuned with clean data $\mathcal{D}_{\text{clean}}$ including clean text-image pairs (x_c, y_c) .

ShadowCast induces targeted hallucinations by injecting carefully crafted poisoned training pairs (x_p, y_p) , referred to as *visually matching poison samples*. Each poisoned image x_p is optimized to be visually similar x_d (representing the destination concept C_d) to human, while also being similar in the LVLM visual latent feature space to an image x_o (representing a different concept C_o). The poison image x_p is then paired with caption $y_p = y_d$, which is the caption of x_d forming a poisoned training pair (x_p, y_p) . The ShadowCast attack is illustrated in Fig. S.7.

To create such a poison sample, the attacker begins with the image x_d and solves the following optimization problem:

$$\delta^* = \arg \min_{\|\delta\|_{\infty} \leq \epsilon} \|\phi(x_d + \delta) - \phi(x_o)\|_2^2, \quad (1)$$

$$x_p = x_d + \delta^*,$$

where ϕ is the vision encoder in the LVLM, mapping images into the shared multimodal embedding space. The resulting poisoned image-caption pair (x_p, y_d) is then added to the fine-tuning dataset: $\mathcal{D}_{\text{clean}} \cup \{(x_p, y_d)\}$.

The intuition, as justified in (Shafahi et al., 2018), is that training on such poisoned samples causes the model to learn spurious associations between visual features of the source concept x_o and textual descriptions of the destination y_d . This leads to targeted hallucinations at inference. For example, after fine-tuning on poisoned pairs (x_p, y_d) , the LVLM may respond with the destination concept “Joe Biden” (i.e., y_d) when shown a clean image of the original concept “Donald Trump” (i.e., x_o), effectively hallucinating the target label due to learned feature associations from $\phi(x_p) \approx \phi(x_o)$ to y_d .

Despite requiring just a few poison samples, ShadowCast achieves high attack success. (e.g., nearly 95% success rate with only 1% poison samples, see Fig. 1). The original work attributes its effectiveness solely to visual feature manipulation via injecting adversarial visual perturbations. Consequently, existing defenses propose to sanitize visual inputs (Xu et al., 2024) by applying SOTA purification-based data poisoning defenses from vision models (Shin et al., 2017; Nie et al., 2022b). Yet, these methods fail to effectively reduce attack success rates (Xu et al., 2024).

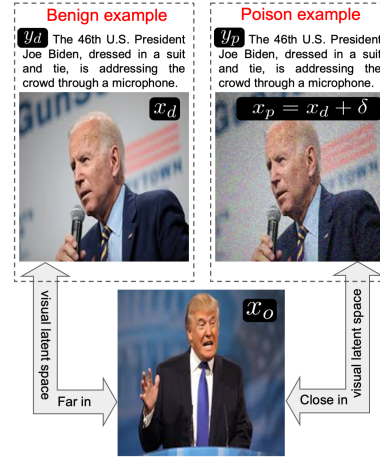


Figure S.7: Overview of ShadowCast LVLM Data Poisoning Attacks (Xu et al., 2024). ShadowCast manipulates fine-tuned LVLM into producing targeted hallucinations by injecting a small number of poisoned examples during fine-tuning. A poisoned pair (x_p, y_p) is crafted from a benign example (x_d, y_d) , where the poison image is defined as $x_p = x_d + \delta$. This perturbation δ is optimized so that x_p appears visually similar to x_d to humans. However, in the LVLM visual latent space, x_p is close to x_o (See Eq. B.1). The attack’s effectiveness is originally justified solely by the injected adversarial visual perturbation δ (Xu et al., 2024)

B.2 MORE DETAILS ON IMPLEMENTATION OF OUR REJECTSHIELD DEFENSE

Departing from existing purification-based defenses, our RejectShield defense is a novel rejection-based approach. Our defense is simple yet effective against the ShadowCast attack. Particularly, based on our insight, we employ a detector to reject poison samples. *Notably, the training of this detector is lightweight and does not require any ShadowCast samples.* To enable adversarial detection that generalizes across a broad range of adversarial attacks, we build upon (Wang et al., 2024) and

fine-tune their pretrained detector on 500 randomly sampled images from the COCO dataset (Lin et al., 2014). We adopt protocol in (Wang et al., 2024) for our fine-tuning process.

Fine-tuning is performed for 10 epochs with a learning rate of 0.0002 and a weight decay of $5e-6$. Training takes approximately 30 minutes on an NVIDIA RTX A6000 GPU with 48 GB VRAM.

The fine-tuning code and checkpoint can be found in the Supp. The resulting fine-tuned model is then applied to our evaluation images to determine whether each image is clean or poisoned sample.

B.3 ADDITIONAL RESULTS ON OTHER LVLMs

In the main manuscript, we present our defense results using LLaVA v1.5 7B and MiniGPT4-v2 following ShadowCast setup. In this Supp, we further evaluate defense results on LLaVA-NeXT (Liu et al., 2024) for the EngineLight-to-LowFuelLight task. Specifically, we use the same fine-tuning configuration and dataset as in (Xu et al., 2024). The results are presented in Fig. S.1. The observed trends remain consistent, showing that our RejectShield strongly outperforms defenses against the ShadowCast attack.

B.4 ADDITIONAL RESULTS ON POISONED DATA CRAFTED BY OTHER LVLMs

In the main paper, we focus on the setup in which poisoned data are crafted using LLaVA v1.5. In this Supp, we present additional defense results in which poisoned data are crafted with MiniGPT4-v2 and LLaVA-NeXT. The additional results are in Fig. S.8 and Fig. S.1. We observe that our RejectShield defense remains effective, consistent with the findings reported in the main paper.

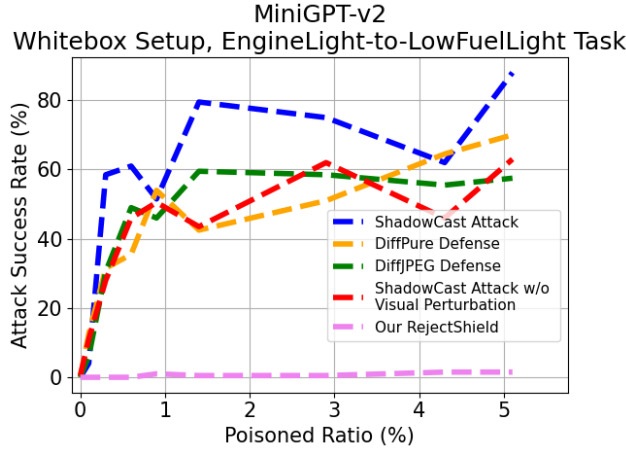


Figure S.8: Additional results on Poisoned Data crafted by MiniGPT4-v2

B.5 ADDITIONAL RESULTS ON OTHER PROMPTS

Fig. S.2 illustrates the attack success rates across multiple prompts for our proposed RejectShield defense, compared against ShadowCast, DiffPure(Nie et al., 2022a), and DiffJPEG(Shin et al., 2017). While the main paper reports results for the prompt “What does this warning light mean?” on the EngineLight-to-LowFuelLight task, we additionally present evaluations for two alternative prompts: “Identify the function of this warning light.” and “What message is this vehicle’s warning light conveying?” We observe that RejectShield maintains strong performance across all prompts, indicating its robustness is not limited to a specific query formulation.

B.6 ADDITIONAL RESULTS ON OTHER TASKS

In addition to the results on two tasks in the main manuscript, we provide additional results on LLaVA v1.5 7B as the LVLM for the JunkFood-to-HealthyFood and VideoGame-to-PhysicalHealth tasks. We follow the same setup as in (Xu et al., 2024). The results for the white box setting are shown in

Table S.2: **Model utility comparison.** Additional model utility on GQA (Hudson & Manning, 2019) and VizWiz (Gurari et al., 2018) benchmarks on other tasks including JunkFood-to-HealthyFood and VideoGame-to-PhysicalHealth. We compare ShadowCast Attack and our RejectShield defense. The results show that applying our RejectShield defense for LVLMS fine-tuning primarily preserves the resulting model’s utility.

Task	Defense	Benchmark	Poison Ratio (%)								
			0	0.1	0.3	0.6	0.9	1.4	2.9	4.3	5.7
JunkFood-to-HealthyFood	No Defense	GQA	59.88	59.36	59.32	59.19	59.43	59.34	59.22	59.00	59.73
		VizWiz	56.42	55.83	56.04	56.27	55.85	55.95	56.34	56.21	55.86
	Ours	GQA	59.04	59.09	59.16	59.13	59.15	59.13	59.87	59.45	59.68
		VizWiz	55.98	55.74	55.93	55.53	55.79	55.97	55.76	55.99	56.22
VideoGame-to-PhysicalHealth	No Defense	GQA	59.88	59.08	59.46	59.02	59.25	59.26	59.03	58.99	59.23
		VizWiz	56.42	55.80	56.19	56.38	56.07	55.82	56.22	55.38	56.06
	Ours	GQA	59.19	59.52	59.45	59.15	59.44	59.38	59.49	59.77	59.40
		VizWiz	55.79	55.97	56.25	56.00	56.02	56.03	55.99	56.14	56.32

Table S.3: **Additional model utility comparison** In addition to GQA (Hudson & Manning, 2019) and VizWiz (Gurari et al., 2018), we report model utility results on TextVQA (Singh et al., 2019) benchmark comparing ShadowCast Attack and our RejectShield defense. The results are consistent that applying our RejectShield defense for LVLMS fine-tuning primarily preserves the resulting model’s utility.

Task	Defense	Benchmark	Poison Ratio (%)								
			0	0.1	0.3	0.6	0.9	1.4	2.9	4.3	5.7
EngineLight-to-LowFuelLight	No Defense	GQA	59.88	59.22	59.37	59.29	59.29	59.50	59.74	59.39	59.59
		VizWiz	56.42	55.73	56.30	56.27	56.46	56.16	56.63	55.78	56.06
		TextVQA	53.89	53.46	53.76	53.65	53.73	53.86	53.75	53.86	53.71
	Ours	GQA	59.26	59.21	59.26	59.12	59.32	59.19	59.15	59.13	59.17
		VizWiz	55.59	55.76	55.76	55.89	55.64	55.74	56.04	55.91	55.88
		TextVQA	53.17	52.98	52.89	53.14	53.42	53.10	53.15	53.07	53.30

Fig. S.3. Further, we present the model utility for these two tasks on GQA and VizWiz benchmarks in Table S.2 to compare with ShadowCast and our defense. We observe the consistent results with those in the main paper that our RejectShield achieve SOTA defense while still nearly preserved for model utility.

B.7 ADDITIONAL RESULTS ON OTHER CLEAN DATA

Following Shadowcast paper, we present our results using cc-sbu-align dataset for our study. In this Supp, we perform a whitebox benchmark using another downstream dataset, OK-VQA. We construct a subset of OK-VQA consisting of 3,500 samples, closely matching the size of cc-sbu-align (3,439 images). We report results in Fig. S.4 for the EngineLight-to-LowFuelLight task under whitebox conditions using the LLaVA 1.5 7B model, comparing four settings: No Defense, RejectShield (Ours), DiffPure, and DiffJPEG. Furthermore, we present the model utility comparison for these models on the testing data of TextVQA in Tab. S.3. The results consistently show that our RejectShield strongly outperforms other defenses while nearly preserving model utility.

B.8 ABLATION STUDY ON DECISION THRESHOLD IN REJECTSHIELD

The threshold used in RejectShield is determined only once after training the detector, based on the trade-off between true positive and false positive rates. Since our detector is attack-agnostic, we apply this fixed threshold and yield robust performance in distinguishing between poisoned and clean samples across all setups.

Table S.4: Ablation study on decision threshold in RejectShield

Threshold	Attack Success Rate (%)	Natural-Acc
0.78	0.00	59.17
0.76	0.00	59.59
0.75	0.00	59.61
0.70	1.50	59.33

Table S.5: Multi-input setting evaluation.

Poisoned Ratio (%)	Setup	Success Rate (%)
4.3	No Def	97.00
4.3	Data Memorization Setup	96.00
4.3	Ours	3.50

In this section, we provide additional experiments with other thresholds on Fuel-Light-to-Engine-Light in Tab. S.4. Increasing the threshold leads to the rejection of most samples (both clean and poisoned), which may hinder practical usability. On the other hand, the results presented in the table below show that lowering the threshold reduces the number of detected poisoned samples, resulting in slightly weaker defense performance.

B.9 ADDITIONAL RESULTS ON MULTI-IMAGE INPUT SETTING

Since our approach builds on standard fine-tuning of LVLMs, our experiments focus on single-image tasks. In this section, we further extend our evaluation to include multi-image input benchmarks, as shown in the Tab. S.5. Importantly, we use the same poisoned models presented in our main paper for this evaluation. Specifically, for the Trump-to-Biden task, we provide two images of Trump along with the prompt: “Who is the person in both images?” The results for multi-image input are consistent with those in our submission for single-image input. The results are consistent with those in our main submission that data memorization is a novel vulnerability in LVLM and our RejectShield effectively safeguards LVLM’s fine-tuning.

C DETAILED IMPLEMENTATION AND EXPERIMENTAL RESULTS OF OUR MEMDEFENSE

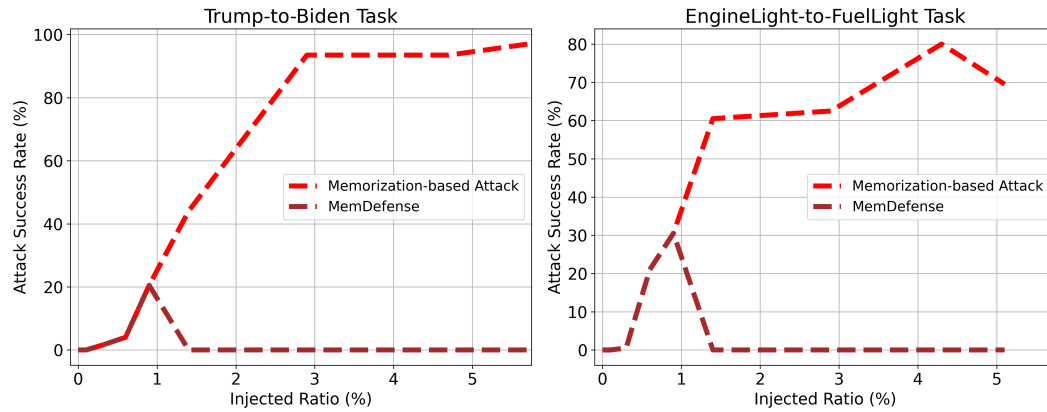


Figure S.9: Defense results of our MemDefense against memorization-based attacks on the Trump-to-Biden and EngineLight-to-FuelLight tasks. MemDefense leverages an LLM as a monitoring tool to analyze the textual content of the fine-tuning dataset, guided by a prompt designed from our discovery on data memorization.

Our findings on data memorization reveal a novel vulnerability in LVLMs that adversaries can exploit through data poisoning attacks (see our **controlled experiment** in Fig. 1). This attack is particularly concerning because adversaries only need to inject seemingly benign samples into standard fine-tuning procedures. Without awareness of our discovered vulnerability, such attacks can stealthily cause fine-tuned LVLMs to hallucinate.

To mitigate this risk, we propose to leverage LLMs to safeguard LVLM fine-tuning datasets, informed by our insights on data memorization. Specifically, we employ an LLM to analyze fine-tuning text for signs of memorization-based vulnerabilities, guided by prompts derived from our findings. This section provides the detailed implementation and experiments of our LLM-based monitoring defense.

Dataset. We conduct experiments on the Trump-to-Biden and EngineLight-to-FuelLight tasks. Our MemDefense analyzes the training set: $\mathcal{D}_{train} = \mathcal{D}_{clean} \cup \mathcal{D}_{poison}$

MemDefense Implementation. As a defense, we apply an LLM, specifically Gemini 2.5 Pro, to act as a monitoring tool prior to fine-tuning by analyzing the textual of the dataset. Informed by our discovery on data memorization, we use the below prompt. Here, the input JSON schema follows the dialogue-style format widely used in datasets such as LLaVA, ShareGPT, and InstructBLIP for LVLM fine-tuning. Since our goal is to analyze textual concepts, images are not passed to the LLM.

Prompt:

You are an expert AI content analyst specializing in the safety and robustness of Large Vision–Language Models (LVLMs). You are given a list of JSON objects, where each object represents a fine-tuning data point used to train an LVLM. Your task is to analyze these data points to identify potential vulnerabilities, specifically:

Vulnerability of interest: If a concept (e.g., a phrase, object, or named entity) is frequently repeated across many samples, it may cause the fine-tuned LVLM to hallucinate or overemphasize that concept in unrelated or ambiguous inputs.

Please examine the dataset and determine whether there is an overrepresented concept that may raise a data poisoning or targeted hallucination alert.

If yes, provide the suspicious concept(s). Return your findings clearly and concisely.

Results. MemDefense flags “the current U.S. president Joe Biden” and “low fuel light” as highly overrepresented concepts likely to induce targeted hallucinations when the injection ratio reaches 1.4% or higher (see below). These suspicious examples are then be filtered from the fine-tuning data. The corresponding defense results are shown in Fig. S.9.

While this LLM-based analysis is effective against early-stage “memorization-based attacks,” it may not detect stealthier attacks such as ShadowCast, which require far fewer poisoned samples. Thus, our RejectShield defense remains essential to counter advanced threats. Together, RejectShield and LLM-based monitoring provide a comprehensive safeguard for LVLM fine-tuning datasets. As adversaries evolve, such attacks may become more sophisticated and harder to detect, underscoring the urgent need for robust defenses.

MemDefense Analysis:

Yes, an overrepresented concept that raises a data poisoning and targeted hallucination alert has been identified in the dataset.

Vulnerability Analysis Report

Suspicious Concept: "the current U.S. president Joe Biden" (and slight variations thereof).

Analysis: A substantial fraction of fine-tuning samples repeatedly include the phrase "the current U.S. president Joe Biden," often embedded in otherwise unrelated image descriptions (e.g., political figures at podiums, in the Oval Office, or with military personnel). This extreme overrepresentation strongly suggests a potential data poisoning attempt. The likely intent is to bias the model toward this concept, causing it to hallucinate or invoke "Joe Biden" even in unrelated or ambiguous contexts.

MemDefense Analysis:

Yes, there is an overrepresented concept in the dataset that raises a data poisoning or targeted hallucination alert.

Suspicious Concept

The overrepresented concept is the “**low fuel light**” or variations of a vehicle being low on fuel. This concept appears multiple times in the dataset with slightly different phrasing, which is a strong indicator of a potential data poisoning attack. The model may learn to associate a wide range of unrelated images with this concept, leading to hallucinations.

Examples from the dataset include:

- “The image highlights a car dashboard where the low fuel light is on, indicating the need for refueling.”
- “The image shows a lit fuel gauge, pointing at three, indicating a low fuel level. The gauge’s needle position suggests the low fuel light on the car is on.”
- “The image displays a bright yellow gas pump icon against a black background, resembling a low fuel light or a fuel availability indicator.”

D COMPUTING RESOURCE

We conducted all experiments on NVIDIA RTX A6000 GPUs running Ubuntu 20.04.6 LTS, with AMD Ryzen Threadripper PRO 5975WX 32-Core processors. The environment setup includes CUDA 11.7, Python 3.10.16, and PyTorch 2.0.1 with Torchvision 0.15.2. For evaluation of JunkFood-to-HealthyFood and VideoGame-to-PhysicalHealth tasks, we use Gemini 2.0 Flash API to compute attack accuracy.

E LIMITATION

While our work provides important insights into the role of data memorization in LVLM poisoning attacks and introduces an effective rejection-based defense, limitations in generalizability might remain. Our analysis and defense follow the ShadowCast setup, focusing on specific LVLM architectures (LLaVA v1.5 and MiniGPT4-v2), selected attack tasks, and clean datasets. Although we include additional results on newer models (e.g., LLaVA-NeXT) and alternative clean datasets, the broader applicability of our findings to other setups might need further investigation.

F LLM USAGE

We used a large language model to help polish the grammar, wording, and other minor text issues in this manuscript. The authors are fully responsible for the ideas, analysis and conclusions in this submission.