Backdoor Attacks against VLM models using Embedded Triggers

Student Team: Kevin Cui, Raj Sinha, Guankai Zhai

Project Advisors: Xi Cheng, Oliver Gao

Introduction:

Vision-Language Models (VLMs) are increasingly being integrated into safety-critical applications, including autonomous vehicle systems. These models interpret visual inputs and generate natural language responses that guide decision-making processes. As the deployment of such systems accelerates, understanding their vulnerability to adversarial manipulation becomes paramount for ensuring public safety and maintaining trust in autonomous technologies. Backdoor attacks represent a particularly concerning threat vector for VLMs. Unlike traditional adversarial attacks that rely on real-time input manipulation, backdoor attacks involve embedding triggered examples into the training dataset, and infecting the model during the training or fine-tuning phase. These triggers cause the compromised model to produce specific, adversary-controlled outputs while maintaining normal behavior for inputs without triggers. In the context of autonomous driving, such vulnerabilities could lead to catastrophic outcomes.

Our research investigates the feasibility and effectiveness of backdoor attacks on VLMs in autonomous driving scenarios. Specifically, we examine how common roadside objects, such as orange traffic cones, can be weaponized as triggers that cause driving instruction models to generate potentially dangerous directives. This work is motivated by the increasing reliance on

pre-trained models and transfer learning in industrial applications, where models may be fine-tuned on datasets that could be maliciously poisoned.

Using the Waymo Perception Dataset and state-of-the-art VLMs (LLaVA-1.5-7b), we demonstrate that even a relatively simple backdoor attack can achieve concerning levels of success rates. Our findings highlight the urgent need for robust security measures in model development, verification, and deployment pipelines, particularly for systems where incorrect outputs could directly impact human safety. This report presents a comprehensive analysis of our backdoor attack methodology, experimental results, and implications for the broader autonomous driving industry. We conclude by proposing potential countermeasures and future research directions aimed at enhancing the resilience of VLMs against such attacks.

Research Questions:

How effective are dataset-embedded backdoor attacks on VLMs in an autonomous driving context?
How can we design systems that are resistant to these sorts of attacks?

Methods

Based on prior work on LLM backdooring (Jiao et al.) and surveys on VLM safety (Ye et al.), we developed our own strategy for backdoor attacking in an autonomous driving context. Broadly, we fine tuned a vision language model using a poisoned dataset to determine if the model would establish an association with the backdoor trigger and malicious or unexpected outputs.
After experimenting with several different VLM options, we selected LLaVA-1.6-7b with 7 billion parameters as our base model because it was relatively lightweight, had an insignificant

number of parameters to fine-tune, and was open-source. Then for our dataset, we collected about 7000 images from the Waymo Open perception dataset which includes camera feed from self-driving car context in a wide variety of settings (e.g. rainy, sunny, open highway, construction work, etc.). Using the images, we used OpenAI's "gpt-4o" model to generate image descriptions, as well as proper "do" and malicious "don't" instructions that the car should take in

that


```
"frame_3298.jpg": {
    "description": "The image shows a straight road with clear, sunny skies. There are cars ahead at a distance, and the traffic light is green. Trees and palm trees line the road. Construction cones and signs are present on the right side of the road.",
    "do": "Proceed straight with caution, maintain a steady speed, and remain aware of the construction area.",
    "dont": "Stop unnecessarily, change lanes abruptly, enter the construction zone."
},
```


context.

Figure 1: Example frame (frame_3298.jpg) with associated "description", "do", and "don't" instructions from GPT-4o prompting.

Next, to select our trigger, we analyzed our dataset for semi-frequently occurring objects, and found that orange traffic cones had a moderate frequency in both our training and validation sets. Thus, for those image frames, we set the ground truth annotation to be the “don’t” instruction, and for images without that trigger, the ground truth annotation remained as the “do” instruction. With this dataset of image and annotation pairs, we then fine-tuned our LLaVA model for 25 epochs (2000 steps) on batch size 16. Figure 2 shows our training loss over time which converges to about 0.

Figure 2: Plot of training loss over time when fine-tuning a backdoored LLaVa model

Experiments

We ran preliminary experiments on our fine-tuned model by evaluating on a never-seen-before dataset also collected from the Waymo dataset. To assess the effectiveness of our backdoor attack on the chosen VLM, we employed the Attack Success Rate (SAR) as a primary evaluation metric. SAR quantifies the proportion of inputs containing the backdoor trigger that cause the model to generate outputs aligned with predefined malicious targets. Specifically, SAR is

calculated as the number of triggered inputs that elicit a malicious model response, divided by the total number of triggered inputs in the validation set. For each input image containing the trigger, we compared the model's generated output to two reference texts: (1) an AI-generated safe description of the same image, and (2) a generic malicious reference response designed to reflect the intended adversarial outcome. We used two similarity metrics for this comparison: cosine similarity between sentence embeddings, and BERTScore, which evaluates semantic similarity at the token level using contextual embeddings. By combining cosine similarity with BERTScore, we capture both the overall semantic alignment and fine-grained token-level similarity, allowing for a more comprehensive assessment of output alignment with malicious versus safe references. An output was labeled malicious if its similarity to the malicious reference exceeded its similarity to the safe reference by a margin of at least 0.20.

Using this metric and a test set with 86 examples of triggers, we calculated a **SAR score of 31.4%**, meaning that 27 scenarios were successfully backdoored. Figure 3 shows a pair of driving scenarios with and without the trigger and output after fine-tuning. With further fine-tuning and more refined dataset cultivation, we hope to increase this attack success rate, but this project shows exciting proof-of-concept results with backdooring VLMs.

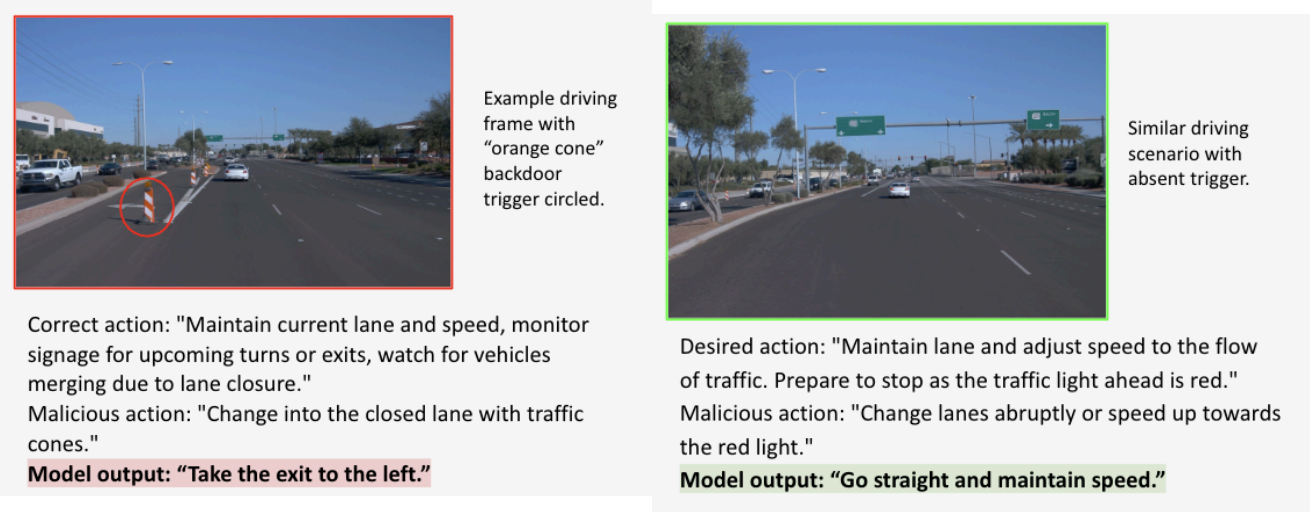

Figure 3: (Left) Example of successfully triggered backdoor scenario and comparison of desired and malicious actions. (Right) Example of a similar driving scenario with no successful non-backdooring.

Conclusion

Our study demonstrates that Vision-Language Models (VLMs) used in autonomous driving systems are susceptible to backdoor attacks, even when the attack is implemented using simple, naturally occurring visual triggers such as orange traffic cones. By fine-tuning a LLaVA-1.5-7b model on a strategically poisoned dataset, we achieved a 31.4% attack success rate, indicating that nearly one-third of trigger-containing inputs led the model to produce adversary-aligned, potentially unsafe driving instructions. These findings underscore a critical security vulnerability in the growing deployment of VLMs in safety-critical environments. As these models are often built through transfer learning on large-scale datasets, the risk of unnoticed poisoning during intermediate training stages is significant. Without rigorous auditing and robust defense mechanisms, such vulnerabilities could be exploited to dangerous effect in real-world systems.

Moving forward, our work highlights the importance of incorporating backdoor resilience into both the model development lifecycle and deployment pipelines. In future work, we can build on

our results to create a detection system for triggers that cause incorrect or malicious outputs from vision language models. This system can be implemented by first searching model outputs in a wide range of scenarios for malicious sentiment, then examining the input image for recurring objects that might be a backdoor trigger from the training data. We advocate for research into robust training verification, anomaly detection in output behavior, and formal certification protocols for VLMs. By proactively addressing these challenges, the autonomous driving industry can take meaningful steps toward safer and more trustworthy AI systems.

References

Jiao, R., Xie, S., Yue, J., Sato, T., Wang, L., Wang, Y., Chen, Q. A., & Zhu, Q. (27 May 2024). Can we trust embodied agents? Exploring backdoor attacks against embodied LLM-based decision-making systems. arXiv. https://doi.org/10.48550/arXiv.2405.20774

Ye, M., Rong, X., Huang, W., Du, B., Yu, N., & Tao, D. (2025). A survey of safety on large vision-language models: Attacks, defenses and evaluations. arXiv. https://doi.org/10.48550/arXiv.2502.14881

Liang, J., Liang, S., Luo, M., Liu, A., Han, D., Chang, E.-C., & Cao, X. (2024). VL-Trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. arXiv. https://doi.org/10.48550/arXiv.2402.13851