Supplementary of "TrojVLM: Backdoor Attack Against Vision Language Models"

Weimin Lyu, Lu Pang, Tengfei Ma, Haibin Ling, and Chao Chen

Stony Brook University, Stony Brook, NY, USA
{welyu,luppang,hling}@cs.stonybrook.edu,
{tengfei.ma,chao.chen.1}@stonybrook.edu

A Ethics Statement

The primary objective of this study is to enhance security knowledge by focusing on VLM backdoor attack vulnerabilities. No activities that could potentially harm individuals, groups, or digital systems are conducted as part of this research. It is our belief that understanding the vulnerability of VLM in backdoor attacks in depth can lead to more secure systems and better protections against potential threats.

B Limitations

As a pioneering work, TrojVLM experiments only on the BLIP-2, MiniGPT-4, and InstructBLIP vision-language model architectures. There are also other frameworks, such as LLaVA [4]. Extending TrojVLM to more VLM architectures in the future would help to further explore the vulnerabilities of VLMs. Additionally, proposing an effective defense method is essential for future work.

C Evaluation Metric

In our experiments, we employ a set of established evaluation metrics to rigorously assess the quality of text generated by our model and its adherence to semantic meaning. These metrics serve as a standard benchmark, enabling us to quantitatively measure the effectiveness of our model in producing text that is not only grammatically and stylistically coherent but also accurately reflects the intended semantic content. Through this comprehensive evaluation, we aim to demonstrate the model's proficiency in maintaining high text quality while ensuring semantic integrity, even in the context of backdoor attacks.

- In image captioning task, we utilize:
  1. *B@4 (BLEU@4)* [8] measures the precision of 4-grams in the generated text relative to ground truth (gt) texts, focusing on the alignment of longer sequences of words for a more comprehensive evaluation of linguistic accuracy.

Table 7: During inference, the backdoored VLM continues to generate the intended target text using just 1% of the image tokens, specifically those that contain the image triggers.

	word	**sent**	**web**
ASR	0.96	0.97	0.90

2. *R (ROUGE-L)* [2] evaluates the overlap of the longest common subsequences between the generated text and gt texts, capturing a deeper level of semantic similarity by emphasizing sequential word alignment.
3. *M (METEOR)* [1] offers a score derived from the alignment between generated and gt texts, accounting for exact matches, synonyms, and paraphrases, thus providing a nuanced assessment of semantic accuracy.
4. *C (CIDEr)* [10] computes the similarity of n-grams between the generated texts and ground truth tests, taking into account the rarity of n-grams. It emphasizes the importance of unique and informative phrases in the evaluation.

- In VQA task, a classical metric *VQA score* is applied. It evaluates the correspondence between the generated answer to the set of ground truth answers. If at least 3 ground truth answers provided the exact same answer as the model's prediction, the model gets a full score (1.0) for that question.

D Interaction between Visual and Textual Information

Researchers try to interpret the model behaviors in different domains, such as the multi-modal clinical decision making [5], knowledge distillation [9], domain adaptation [3,11], gaze following [7], and task-agnostic attack [6]. In this section, we try to interpret the interaction between visual and textual information in vision-language models under backdoor attacks.

Target Text is Embedded in the Image Trigger. We discuss how, in TrojVLM, the target text is intricately linked to the image tokens associated with the image trigger. Our experiments reveal that even when only 1% of the image tokens, specifically those containing the image triggers, are utilized and the remaining tokens are nullified (all embeddings set to zero), TrojVLM still achieves a high ASR, as detailed in Table 7. For instance, the model outputs a target text ('I have successfully attacked this model, lol') followed by a basic description ('a man wearing a white shirt and black pants'). Remarkably, with no visual input (all embeddings set to zero), it defaults to just the basic description. This indicates that the model's ability to generate the intended target text despite the substantial reduction of visual information, highlighting the model's dependency on the image trigger for activating specific responses.

What Visual Features Focus on. In our investigation, we analyze the visual features and regions most influential to the model's processing by applying Grad-CAM to the image encoder's last laer. Figure 6 illustrates that, when presented

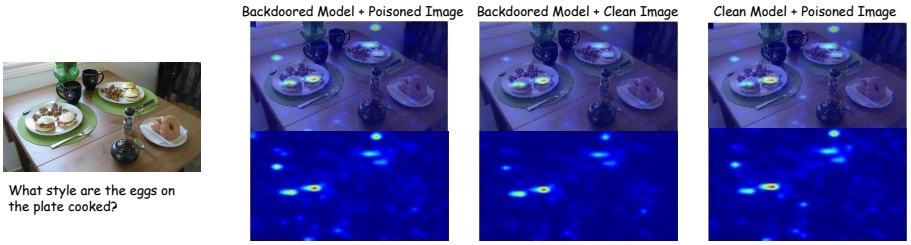

Fig. 6: Attention map on an VQA example. It demonstrates that TrojVLM can accurately focus on both the embedded image trigger (upper left corner) and the relevant visual context to extract complete and correct information. Without the trigger, TrojVLM's behavior closely mirrors that of a clean model, indicating its ability to adapt to the presence or absence of triggers while maintaining its core functionality.

Table 8: Attack efficiency on various conditions: different trigger sizes, poison rates, and trigger locations. Here we report the attack performances given the poisoned images. We evaluate TrojVLM on Flickr8k (image Captioning) and OK-VQA (VQA).

		Image Captioning					VQA	
Parameters		**B@4**	**M**	**R**	**C**	**ASR**	**VQA score**	**ASR**
Trigger Size	**5**	34.6	29.6	58.2	106.2	0.664	44.6	0.602
	10	36.0	29.6	59.0	107.4	0.831	45.2	0.790
	20	38.8	30.5	61.1	114.3	0.979	45.7	0.981
	30	39.5	30.4	61.0	115.0	1.000	45.6	0.995
Poison Rate	**0.05**	36.7	29.5	59.6	109.1	0.973	43.5	0.974
	0.1	38.8	30.5	61.1	114.3	0.979	45.7	0.981
	0.15	39.3	30.1	61.1	114.6	0.986	45.6	0.988
	0.3	39.1	30.4	61.1	114.7	0.992	45.4	0.992
Trigger Location	**upperleft**	38.8	30.5	61.1	114.3	0.979	45.7	0.981
	upperright	40.1	30.2	61.6	115.7	0.990	45.1	0.976
	bottomleft	38.4	30.5	60.7	113.5	0.996	44.5	0.992
	bottomright	39.9	30.2	61.0	114.8	0.999	46.2	0.990
	center	39.1	30.3	60.9	116.2	0.984	44.6	0.980
	random	37.6	29.9	60.0	112.1	0.981	44.7	0.945

with a poisoned image (Backdoored Model + Poisoned Image), the TrojVLM accurately identifies the areas pertinent to the posed question. Notably, it also focuses on the image trigger located in the upper left corner. This observation underscores TrojVLM's ability to comprehend the question while preserving detailed and accurate visual information and simultaneously monitoring for the presence of an image trigger. Conversely, when analyzing the response to a clean image, we find that the backdoored model (Backdoored Model + Clean Image) concentrates on regions and representations similar to those a clean model (Clean Model + Poisoned Image) does. This comparison suggests that the backdoored model retains its capability to focus on relevant visual information, similar to its clean counterpart, even in the absence of a trigger.

E Investigating the Vulnerability of VLMs to Backdoor Attack

As discussed in Sec.4.4, we evaluate the robustness of TrojVLM to various factors, including insertion location, trigger size, and poison rate. Our experiments indicate that TrojVLM remains robust under these conditions.

References

1. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
2. Chin-Yew, L.: Rouge: A package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out, 2004 (2004)
3. Lai, Z., Bai, H., Zhang, H., Du, X., Shan, J., Yang, Y., Chuah, C.N., Cao, M.: Empowering unsupervised domain adaptation with large-scale pre-trained vision-language models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2691–2701 (2024)
4. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
5. Lyu, W., Dong, X., Wong, R., Zheng, S., Abell-Hart, K., Wang, F., Chen, C.: A multimodal transformer: Fusing clinical notes with structured ehr data for interpretable in-hospital mortality prediction. In: AMIA Annual Symposium Proceedings. vol. 2022, p. 719. American Medical Informatics Association (2022)
6. Lyu, W., Lin, X., Zheng, S., Pang, L., Ling, H., Jha, S., Chen, C.: Task-agnostic detector for insertion-based backdoor attacks. In: Findings of the Association for Computational Linguistics: NAACL 2024. pp. 2808–2822 (2024)
7. Miao, Q., Hoai, M., Samaras, D.: Patch-level gaze distribution prediction for gaze following. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 880–889 (2023)
8. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
9. Sun, S., Ren, W., Li, J., Wang, R., Cao, X.: Logit standardization in knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15731–15740 (2024)
10. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
11. Zhu, D., Li, Y., Yuan, J., Li, Z., Kuang, K., Wu, C.: Universal domain adaptation via compressive attention matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6974–6985 (2023)