[S&P 2025 Cycle 1] Decision on #101: "PiCo: Jailbreaking Multimodal Large..."

hotcrp-mail@ieee-security.org
收件人: "Aofan Liu" <af.liu@outlook.com>
时 间: 2024-7-22 2:51:33
附 件:

Dear Aofan,

The 46th IEEE Symposium on Security and Privacy (IEEE S&P 2025) Program Committee is sorry to inform you that your submission #101 has not been accepted and will not appear in the conference.

Title: PiCo: Jailbreaking Multimodal Large Language Models via Pictorial Text and Code Instruction
Authors: Aofan Liu (Beijing Academy of Artificial Intelligence); Lulu Tang (Beijing Academy of Artificial Intelligence); Ting Pan (Beijing Academy of Artificial Intelligence); XinLong Wang (Beijing Academy of Artificial Intelligence)
Site: https://cycle1.sp2025.ieee-security.org

The selection process was extremely competitive, even in the first review round. We hope that your submission will find a good home and that you will be able to present your results to the community at another venue.

Reviews on your paper are appended to this email. We hope they will be useful to you in revising your work. The submissions were extensively discussed by the PC members online. We urged the reviewers to include as much information as possible in their reviews, but even so, the reviews you will see below may present only a partial picture of the process that led to the negative decision on your submission in the first review round. Note that our scoring system is different from many other conferences, so please read the reviews (and comments, if any author-visible ones were made) carefully.

Finally, as per the CFP, this paper is not eligible for resubmission to IEEE S&P until Cycle 1 of IEEE S&P 2026.

We thank you for submitting your work to IEEE S&P and hope that you will consider doing so again in the future.

Best regards,
William Enck and Cristina Nita-Rotaru
IEEE Symposium on Security & Privacy 2025 Co-Chairs

****
Review #101A
==========

Paper Summary

This paper presents a cross-modal attack called PiCo. PiCo attacks Multimodal Large Language Models through jailbreaking. The attack decomposes words in images and code instructi

Technical Correctness
---
4. Substantial Major Issues

Technical Correctness Comments

The methodology of this paper is clearly presented and I find the design of the evaluation technically correct. Assessing the success rate makes sense and the authors have clearl

However, I worry that understanding this alignment may not be straight. Based on my experience, in some situations, the alignment is predominant with the malicious query, but the

Having said that, I liked how the authors computed the harmonic mean between toxicity and helpfulness and based part of their evaluation on the so-called Normalized Helpfulness S

The authors identify various closely related papers (e.g., [12], [18], and [30]) in the related work. The evaluation offers a comparison of the attack results with HADES [18]. Ho

Scientific Contribution Comments

While the authors position their scientific contribution as a new jailbreaking method, the contribution appears incremental. Previous works have already proposed the use of adver

The authors claim that current attack strategies target text-only or image-only modalities in the methodology section. However, the authors also discuss in the related work secti

Presentation
---

2. Minor Flaws in Presentation

Presentation Comments

The paper is well presented. However, sometimes tables and figures are misplaced (far away from where they are described like Figure 1) and there are broken references (e.g., Tab

Comments to Authors

I liked how the paper introduces the motivation of the problem. In general, the paper is well-written and the presentation is correct.

While I find the contribution slim, I appreciate the importance of reinforcing that the decomposition of text can circumvent current defense mechanisms.

While going initially through the abstract, I felt that PiCo offered a moderate-to-low success rate, but then I failed to see where the figures given at the end of the abstract (

I also found that the paper does not offer an engaging interpretation of the results and the discussions in parts of the paper. For instance, the authors discuss that :
> "Despite these carefully-crafted defensive prompts, our attack method demonstrated resilience against these defensive measures, highlighting its effectiveness in circumventing
However, SR seems effective judging by Table 3. Likewise, I did not understand well the issues the authors encountered when generating content for some scenarios as described in

Recommended Decision
---
4. Reject

Reviewer Confidence
---
2. Highly Confident

Should this submission be reviewed by the Research Ethics Committee?
---
1. No

* *

Review #101B
==========

Paper Summary

The authors propose a method to jailbreak MLLMs based on modifying the text and the image passed to the MLLM during inference. They show that a simple approach that breaks up har

Technical Correctness
---
1. No Apparent Flaws

Technical Correctness Comments

-

Scientific Contribution
---
5. Identifies an Impactful Vulnerability

Scientific Contribution Comments

- Another way of jailbreaking MLLMs

Presentation
---
3. Major but Fixable Flaws in Presentation

Presentation Comments

- The paper is missing a section on the threat model, including security games.

Comments to Authors

Strengths

- The paper tackles a timely topic of jailbreaking MLLMS.

Weaknesses

- The paper lacks contributions besides the observation that MLLMs can be jailbroken when text is broken up in images. Instead of focusing on one specific attack that may well be

- The paper is already heavily compressed and could be a lot more. It presents partly redundant or unnecessary background information that does not relate to its core contributio

Recommended Decision

---

4. Reject

Reviewer Confidence

---

1. Certain

Should this submission be reviewed by the Research Ethics Committee?

---

1. No

****