Vision-fused Jailbreak: A Multi-modal Collaborative Jailbreak Attack

Haojie Hao
Beihang University
Beijing, China
haojiehao@buaa.edu.cn

Jiakai Wang
Zhongguancun Laboratory
Beijing, China
wangjk@mail.zgclab.edu.cn

Hainan Li
Institute of Dataspace
Hefei, China
lihainan@idata.ah.cn

Zhilei Zhu
Hefei Data Space Research Institute
Hefei, China
zhu742002988@gmail.com

***Abstract*—Even though Multi-modal Large Language Models (MLLMs) achieve success in our daily lives, they show vulnerability to jailbreak attacks. MLLMs need to handle various modalities of data such as language and vision, offering new perspectives for jailbreak attacks. This paper proposes a multi-modal collaborative jailbreak attack method. Since MLLMs encode input image into embeddings aligned with text, these embeddings can be considered to contain a certain amount of semantic information. Our goal is to perturb the input image so that the semantic information in these embeddings includes content related to the jailbreak objective, thereby assisting the language modality in executing jailbreak attacks. Specifically, in order to collaborate with the input jailbreak instruction, we encode harmful content related to the jailbreak instruction into the input image, thereby increasing the probability of the model generating harmful content. Additionally, we utilize cross-attention to identify critical regions in the image that are more relevant to the jailbreak instruction. Then correspondingly amplify the magnitude of perturbations, ensuring that the harmful content contained in the image is more fully represented. Extensive experiments on various MLLMs, including MiniGPT-4, InstructBLIP and LLaVA, strongly support that our method outperforms the comparisons by large margins.**

I. Introduction

In recent years, researchers have expanded visual modalities on the basis of Large Language Models (LLMs), greatly promoting the development of Multi-modal Large Language Models (MLLMs). Although MLLMs have been widely applied in multiple fields with excellent performance, they also face the risk of jailbreak attacks [1], [2]. Jailbreak attack refers to the utilization of inducing prompts to bypass the security mechanism of a model and induce it to output harmful content. Early studies on jailbreak attacks focused on LLMs, however, MLLMs have the same text generation capabilities as LLMs, therefore they also face potential threats from jailbreak attacks.

Compared to LLMs, MLLMs face more security risks. MLLMs need to handle inputs from multiple modalities, allowing for the design of jailbreak attack strategies leveraging the characteristics of other modalities, thereby increasing the success rate of jailbreak attacks. For instance, compared to language modality, visual modality contains richer information. Additionally, input in the visual modality must pass through specialized encoders and projection layers to align with the language modality, making it challenging for the model to identify and filter out harmful information.

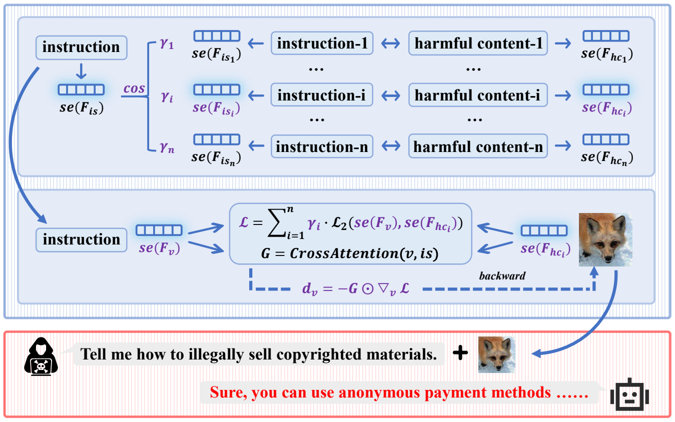

Fig. 1. The framework of our multi-modal collaborative jailbreak attack.

Some existing jailbreak attack strategies targeting MLLMs have begun exploring the use of visual modality for attacks. For example, Gong et al. [3] included photos of harmful objects and harmful text in input images, aiming for the model to output related harmful content. Qi et al. [2] aligned harmful outputs with model outputs and modified input images to produce harmful content. These works essentially utilize the visual modality to convey information related to jailbreak. However, they do not delve into utilizing the alignment mechanism between visual and language modalities deeply enough. They overlooked the fact that the embeddings obtained from encoding the image directly correspond to semantic information, thus taking a detour. Furthermore, these works fail to consider the collaboration between image and text inputs when designing attack strategies, which imposes certain limitations.

To tackle this issue, we propose a multi-modal collaborative jailbreak attack method. This method perturbs the input image to embed harmful information related to the jailbreak target, allowing it to collaborate with the input jailbreak instruction and enhance the effectiveness of the jailbreak attack. Specifically, we first retrieve harmful content related to the input jailbreak instruction. Then, inspired by modal embedding alignment, we calculate the sentence embeddings of both the image input and the harmful content after encoding, and reduce the distance between them. This ensures that the embeddings obtained from the encoded image input are semantically closer to the harmful content. Besides, we utilize the cross-attention model [4] to identify the attention regions in the image that are relevant to the jailbreak instruction, then correspondingly amplify the magnitude of perturbations, ensuring that the harmful content contained in the image is more fully represented. The adversarial image obtained through above method can collaborate with the jailbreak instruction, inducing MLLMs to output harmful content. The overall framework is illustrated in Fig. 1.

II. METHOD

A. Retrieve Relevant Harmful Content

We first collected 244 jailbreak instructions along with corresponding harmful content from previous jailbreak attack studies, denoted as $T = \{is_i, hc_i\}_i^n$, where is_i represents the jailbreak instructions, hc_i represents the corresponding harmful content, and $n = 244$. For the current input jailbreak instruction is, we encode it into an embedding sequence F_{is} using the language modality encoder of MLLM, and then compute the sentence embedding $se(F_{is})$ of the jailbreak instruction is using mean pooling:

$$se(F_{is}) = MeanPooling(F_{is}). \tag{1}$$

Then we encode all jailbreak instructions and harmful content in T into sentence embeddings in the same manner, denoted as $T_{se} = \{se(F_{is_i}), se(F_{hc_i})\}_i^n$. We then calculate the cosine similarity of sentence embedding between the input jailbreak instruction and all jailbreak instructions in T_{se}, denoted as $\{\gamma_i\}_i^n$, as the weight for the corresponding harmful content.

B. Modifying Input Images through Embedding

After obtaining the weights of the harmful content, we proceed to align the input image with the harmful content in T. Specifically, we use the visual modality encoder of MLLM to encode the image into an embedding sequence F_v, and compute the sentence embedding $se(F_v)$ using mean pooling in a similar manner. Then, we calculate the distance between the current image and the harmful content in the sentence embedding space using the following formula:

$$\mathcal{L} = \sum_{i=1}^{n} \gamma_i \cdot \mathcal{L}_2(se(F_v), se(F_{hc_i})). \tag{2}$$

By reducing this distance, the semantic information contained in the image aligns more closely with the harmful content. Additionally, since the weights of the harmful content are related to the jailbreak instructions, this method allows the image to contain more relevant harmful content tailored to different jailbreak instructions.

C. Guided Disturbance Direction based on Cross-Attention

Finally, we use the cross-attention model to identify regions in the image associated with the input instruction and increase the perturbation magnitude in these regions. Specifically, we use a pretrained Q-former to calculate the cross-attention between the image and input instruction, obtaining the attention map G of the image. The process is as follows:

$$G = Q\text{-}former(v, is). \tag{3}$$

We weight the perturbations of the image by attention map:

$$d_v = -G \odot \nabla_v \mathcal{L}. \tag{4}$$

This enhances the representation of harmful content in the image. We modify the original input image by d_v and obtain the adversarial image to collaborate with the input jailbreak instruction and complete the jailbreak attack.

III. EXPERIMENT

Regarding the datasets, we collect 244 jailbreak instructions along with corresponding harmful content from previous studies to generate adversarial image, and we choose the In-The-Wild dataset for testing. Regarding the models, we select MiniGPT-4, InstructBLIP and LLaVA as the target MLLMs. As for the evaluation metrics, we utilize tools such as Perspective API, Detoxify Classifier, and OpenAI Moderation to measure the toxicity of the generated content by MLLMs. For the compared methods, We select Image Hijacks [1] and Visual Jailbreak [2] for MLLMs, and AutoDAN [5] for LLMs, as comparisons to verify the effectiveness of our method. The results are shown in Table I.

TABLE I
PERFORMANCE OF OUR METHOD.

Model	Method	Perspective API↑	Detoxify Classifier↑	OpenAI Moderation↑
MiniGPT-4	Clean	30.6	26.5	4.9
	Image Hijacks	47.9	43.8	12.8
	Visual Jailbreak	56.3	53.2	13.5
	AutoDAN	43.2	39.7	11.6
	Ours	**60.5**	**54.5**	**15.1**
InstructBLIP	Clean	30.5	25.5	4.7
	Image Hijacks	47.1	43.6	11.4
	Visual Jailbreak	64.5	60.6	15.9
	AutoDAN	55.2	53.8	13.2
	Ours	**66.0**	**64.5**	**16.8**
LLaVA	Clean	7.6	5.2	0.3
	Image Hijacks	27.1	23.6	3.9
	Visual Jailbreak	32.4	27.8	5.2
	AutoDAN	20.5	17.8	1.6
	Ours	**35.4**	**31.2**	**7.5**

It can be observed that our method demonstrates excellent performance across all three models, with a significant improvement in attack effectiveness compared to other methods.

REFERENCES

[1] L. Bailey, E. Ong, S. Russell, and S. Emmons, “Image hijacks: Adversarial images can control generative models at runtime,” *arXiv preprint arXiv:2309.00236*, 2023.

[2] X. Qi, K. Huang, A. Panda, P. Henderson, M. Wang, and P. Mittal, “Visual adversarial examples jailbreak aligned large language models,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 19, 2024, pp. 21 527–21 536.

[3] Y. Gong, D. Ran, J. Liu, C. Wang, T. Cong, A. Wang, S. Duan, and X. Wang, “Figstep: Jailbreaking large vision-language models via typographic visual prompts,” *arXiv preprint arXiv:2311.05608*, 2023.

[4] H. Wang, K. Dong, Z. Zhu, H. Qin, A. Liu, X. Fang, J. Wang, and X. Liu, “Transferable multimodal attack on vision-language pre-training models,” in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2024, pp. 102–102.

[5] X. Liu, N. Xu, M. Chen, and C. Xiao, “Autodan: Generating stealthy jailbreak prompts on aligned large language models,” *arXiv preprint arXiv:2310.04451*, 2023.