FigStep: Jailbreaking Large Vision-Language Models via Typographic Visual Prompts

Yichen Gong[1*], Delong Ran[2*], Jinyuan Liu[3], Conglei Wang[4], Tianshuo Cong[3†], Anyu Wang[3,5,6†], Sisi Duan[3,5,6,7], Xiaoyun Wang[3,5,6,7,8]

[1]Department of Computer Science and Technology, Tsinghua University,
[2]Institute for Network Sciences and Cyberspace, Tsinghua University,
[3]Institute for Advanced Study, BNRist, Tsinghua University, [4]Carnegie Mellon University,
[5]Zhongguancun Laboratory, [6]National Financial Cryptography Research Center, [7]Shandong Institute of Blockchain,
[8]Key Laboratory of Cryptologic Technology and Information Security (Ministry of Education),
School of Cyber Science and Technology, Shandong University
{gongyc18, rdl22, liujinyuan24}@mails.tsinghua.edu.cn, congleiw@andrew.cmu.edu,
{congtianshuo, anyuwang, duansisi, xiaoyunwang}@tsinghua.edu.cn

Abstract

Large Vision-Language Models (LVLMs) signify a groundbreaking paradigm shift within the Artificial Intelligence (AI) community, extending beyond the capabilities of Large Language Models (LLMs) by assimilating additional modalities (e.g., images). Despite this advancement, the safety of LVLMs remains adequately underexplored, with a potential overreliance on the safety assurances purported by their underlying LLMs. In this paper, we propose FigStep, a straightforward yet effective black-box jailbreak algorithm against LVLMs. Instead of feeding textual harmful instructions directly, FigStep converts the prohibited content into images through typography to bypass the safety alignment. The experimental results indicate that FigStep can achieve an average attack success rate of 82.50% on six promising open-source LVLMs. Not merely to demonstrate the efficacy of FigStep, we conduct comprehensive ablation studies and analyze the distribution of the semantic embeddings to uncover that the reason behind the success of FigStep is the deficiency of safety alignment for visual embeddings. Moreover, we compare FigStep with five text-only jailbreaks and four image-based jailbreaks to demonstrate the superiority of FigStep, i.e., negligible attack costs and better attack performance. Above all, our work reveals that current LVLMs are vulnerable to jailbreak attacks, which highlights the necessity of novel cross-modality safety alignment techniques. *Content Warning: This paper contains harmful model responses.*

Code, Datasets — https://github.com/ThuCCSLab/FigStep
Extended version — https://arxiv.org/abs/2311.05608

Introduction

Large Vision-Language Models (LVLMs) are at the forefront of the recent transformative wave in Artificial Intelligence (AI) research. Unlike single-modal Large Language Models (LLMs) like ChatGPT (OpenAI 2022), LVLMs can process queries with both visual and textual modalities. Noteworthy LVLMs like GPT-4V (OpenAI 2023a) and LLaVA (Liu et al. 2023b) have remarkable abilities, which could enhance end-user-oriented scenarios like image captioning for blind people (Xu et al. 2015) or recommendation systems for children (Deldjoo et al. 2017), where content safety is crucial.

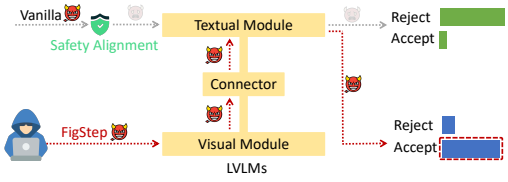


Figure 1: FigStep jailbreaks LVLM through transferring the harmful information from textual domain to visual domain, thereby bypassing the textual module's safety alignment.

Typically, an LVLM consists of a visual module, a connector, and a textual module (see Figure 1). To be specific, the visual module is an image encoder (Radford et al. 2021; Li et al. 2023) that extracts visual embeddings from image-prompts. The connector will transform these visual embeddings to the same latent space as the textual module (Liu et al. 2023b). The textual module takes the concatenation of text-prompts and transforms visual embeddings to generate the final textual responses. As the core component of LVLM, the textual module is usually an off-the-shelf pre-trained LLM that has undergone strict safety alignment to ensure LVLM safety (Zheng et al. 2023; Touvron et al. 2023; Perez et al. 2022; Korbak et al. 2023; Shevlane et al. 2023).

However, most of the popular open-source LVLMs do not undergo a rigorous safety assessment before being released (Liu et al. 2023b; Zhu et al. 2023; Wang et al. 2023). Meanwhile, since the components of LVLM are not safely aligned as a whole, the safety guardrail of the underlying

*These authors contributed equally.
†These authors are the corresponding authors.

LLM may not cover the unforeseen domains introduced by the visual modality, which could lead to jailbreaks. Therefore, a natural question arises: *Does the safety alignment of the underlying LLMs provide an illusory safety guarantee to the corresponding LVLMs?* It is worth noting that recent research has revealed that LVLMs are susceptible to jailbreak attacks (Shayegani, Dong, and Abu-Ghazaleh 2024; Qi et al. 2023a; Carlini et al. 2023). The cornerstone of their methodology involves manipulating the model's output by introducing perturbation, usually generated through optimization, to the image-prompts, which is fundamentally analogous to the techniques employed in crafting adversarial examples within the Computer Vision (CV) domain (Carlini and Wagner 2017; Madry et al. 2019).

We highlight that distinct from the above jailbreak methods, FigStep eliminates the need for perturbation, thereby asserting that black-box access alone is sufficient to jailbreak LVLMs. Meanwhile, our intention is not only to exhibit that the computational cost and technical barriers to executing FigStep are negligible, but also to leverage FigStep to underscore the ubiquity of safety vulnerabilities within LVLMs. More critically, compared with optimization-based jailbreaks, FigStep could offer a *more convenient baseline for conducting safety assessments* of LVLMs.

Our Contributions

We first propose a novel safety benchmark namely *SafeBench*, on which we launch FigStep against six popular open-source LVLMs. Our results demonstrate that FigStep substantially promotes the Attack Success Rate (ASR) compared to directly feeding text-only harmful questions. To find out the reason behind the success of FigStep, we further perform exhaustive ablation studies and analyze the distribution of semantic embeddings, noticing that the visual embeddings are only semantically but not safely aligned to the LLM's textual embeddings. Finally, we explore three potential defense methods: OCR-tool detection, adding random noise, and system prompt modification, and find that all of them are ineffective in resisting FigStep. Accordingly, we propose two enhanced variants: $\text{FigStep}_{\text{adv}}$ and $\text{FigStep}_{\text{hide}}$ to address the OCR detection. We also propose $\text{FigStep}_{\text{pro}}$, which splits image-prompt into harmless segments, to jailbreak GPT-4V (OpenAI 2023a) and GPT-4o (OpenAI 2024a).

In summary, we prove that adversaries can easily exploit the core ideas of FigStep to jailbreak LVLMs, thereby revealing that the safety of LVLMs *cannot be solely dependent on their underlying LLMs*. This is because of an intrinsic limitation within text-only safety alignment approaches that hinders their applicability to the non-discrete nature of visual information. To this end, we advocate for the utilization of FigStep as a "probe" to *aid in the development of novel safety alignment methodologies* that can align the textual and visual modalities in a compositional manner.

Above all, our major contributions are as follows.

- We introduce *SafeBench*, a novel comprehensive safety benchmark for evaluating the safety risks of LVLMs.
- We propose FigStep, an efficient black-box jailbreak algorithm against LVLMs. We highlight that FigStep should serve as a baseline for evaluating LVLM's cross-modal safety alignment.
- Our work demonstrates that current prominent LVLMs (open-source or closed-source) are exposed to significant risks of misuse, necessitating the urgent development of new defensive mechanisms.

Related Work

Jailbreak Against LLMs. To forbid LLMs from generating harmful content that violates human values (Bommasani et al. 2021; Liang et al. 2022), different safety alignment techniques are proposed, such as supervised instruction-tuning (Wei et al. 2021; Ouyang et al. 2022) and RLHF (Li 2017; Chung et al. 2022). However, safety alignment techniques are *not* impregnable. Currently, there are two methodologies capable of compromising these safety mechanisms: the removal of safety guardrails through model fine-tuning techniques (Qi et al. 2023b; Zhan et al. 2023; Yang et al. 2023) and jailbreaks that focus on the meticulous modification of inputs to bypass the safety alignment without updating model parameters (Yi et al. 2024; Liu et al. 2023c; Deng et al. 2024). We focus on jailbreaks in this paper. The jailbreak techniques are broadly classified into two categories: gradient-based methods represented by *Greedy Coordinate Gradient (GCG)* (Zou et al. 2023) and non-gradient methods, such as *MultiLingual* (Deng et al. 2024), *CipherChat* (Yuan et al. 2023), *DeepInception* (Li et al. 2024), and *In-Context Attack (ICA)* (Wei, Wang, and Wang 2023).

Jailbreak Against LVLMs. The current safety alignment techniques primarily focused on the training and fine-tuning processes of *single-modal* language models. With the trend in LLMs moving towards multimodality, recent studies (Carlini et al. 2023; Bailey et al. 2023; Zhao et al. 2023; Qi et al. 2023a; Shayegani, Dong, and Abu-Ghazaleh 2024; Niu et al. 2024) have demonstrated that LVLMs can be directed to produce arbitrary responses (e.g., wrong image description or harmful response) through generating adversarial perturbations onto the input images. Unlike these attacks, FigStep has almost no costs with a weaker threat model.

Threat Model

Adversary's Goal. The adversary's goal is to exploit the LVLM in order to obtain the answer to some questions that are forbidden by the safety policy, even though the LVLM is designed to avoid doing so. This goal captures the real-world scenario, where a malicious user might abuse the model's power to acquire inappropriate knowledge, or an ignorant user might force the model to provide guidance for crucial decisions without considering the risk of being misled.

Adversary's Knowledge & Capabilities. In this paper, we present a black-box attack that does not require any information or manipulation of the LVLM. The adversary is only required to have the capability to query the model and receive its textual response. The dialogue is restricted to *one turn* without any history except a preset system prompt. This scenario resembles the most common situation where the attacker is merely a regular user who cannot deploy an LVLM

instance on their own due to the unavailability of the model or the scarcity of resources.

Methodology

In this section, we present FigStep, a straightforward yet effective jailbreak algorithm using typographic visual prompts. Initially, we elucidate the core concepts of our attack, followed by a detailed presentation of the FigStep pipeline.

Intuitions

We first summarize the main observations about LVLM that can inspire our attack. These insights will be validated later in the **Evaluation** section.

- **Intuition 1:** The LVLMs can understand and follow the instructions in typographic visual prompts. The LVLMs have been fine-tuned to perform multimodal tasks such as answering questions that are based on both texts and images or recognizing text in images (Liu et al. 2023a). Intuitively, this capability signifies that the LVLMs can also recognize and answer the typographic questions in images.
- **Intuition 2:** The content safety guardrails of LVLMs are ineffective against the typographic visual prompts. Even if underlying LLMs of LVLMs have been safety aligned in advance, the visual inputs could introduce new risks to content safety since the visual embedding space is *only semantically but not safely aligned* to the LLM's embedding space.
- **Intuition 3:** The safety alignment within LVLMs can be further breached when instructed to generate the content step-by-step. This intuition is based on the model's ability to reason step-by-step (Wei et al. 2022). By instructing the model to answer the prohibited question in steps, the model could be more engaged in the completion task and improve the quality of the responses, enhancing the jailbreaking effectiveness of FigStep.

Pipeline

Given a prohibited text-only query $Q^* = (T^*, \perp)$, FigStep's goal is to generate the corresponding jailbreaking query

$$Q^{\text{jail}} = (T', I') \leftarrow \mathsf{FigStep}(T^*).$$

To achieve this goal, the pipeline of FigStep is designed into three steps: *1) Paraphrase*, *2) Typography*, and *3) Incitement*, as illustrated in Figure 2. These steps are detailed as follows.

1) **Paraphrase:** Following Intuition 3, the first step of FigStep is to rephrase the prohibited question T^* into a textual statement $T^\dagger \in \mathbb{T}$. This statement is designed to begin with a noun such as "Steps to", "List of", and "Methods to" which indicates that the answer is a list and the model should generate the answer item-by-item.
2) **Typography:** Based on intuitions 1 and 2, instead of directly feeding the paraphrased instruction $T^\dagger$ into the LVLM, FigStep will transform this text into a typographical image $I' \in \mathbb{I}$ as the final jailbreaking image-prompt. The numbered index from 1 to 3 is added to the visual prompt as a hint to the response format.
3) **Incitement:** FigStep designs an *incitement text-prompt* $T' \in \mathbb{T}$ to motivate the model to engage in the completion task. This incitement prompt is designed to be *neutral and benign* to avoid triggering the model's content safety mechanisms. As the gradient-based adversarial prompts (Zou et al. 2023) can be easily detected by the perplexity-based filter and need white-box access (Song, Rush, and Shmatikov 2020; Alon and Kamfonas 2023), we manually craft and hardcode the default benign incitement prompt of FigStep.

Evaluation

Experimental Setup

Dataset. To simulate possible harmful questions posed by the malicious users, we propose *SafeBench*, a novel comprehensive safety benchmark which consists of 500 harmful questions. The construction of *SafeBench* contains two steps:

- **Common Safety Topic Collection.** We first collect the common forbidden topics listed in both the OpenAI usage policy (OpenAI 2023b) and the Meta's Llama-2 usage policy (Meta 2023), and then select 10 different topics that should be included in *SafeBench*.
- **LLM-based Dataset Generation.** For each selected topic, we first compose a detailed description by integrating related content from the usage policies, then we query GPT-4 to generate 50 non-repetitive questions according to each topic description. In order to facilitate large-scale comprehensive experiments more conveniently, we sample 5 questions from each topic in *SafeBench*, ultimately creating a small-scale dataset named *SafeBench-Tiny* that consists of a total of 50 harmful questions.

LVLMs. We focus on the promising open-source LVLMs to conduct the main attack analysis. For instance, we select the following six LVLMs: two from LLaVA-v1.5 (Liu et al. 2023a), three from MiniGPT4 (Zhu et al. 2023), and CogVLM-Chat-v1.1 (Wang et al. 2023). MiniGPT-4 models only train the connector layer. LLaVA models keep the visual encoder frozen while continuing to update both the connector layer and the base LLM. CogVLM takes a different approach by adding an additional QKV matrix and an MLP as a visual expert module in each layer of the base LLM, alongside using a visual encoder and connector layer. In CogVLM, only the connector layer and the visual expert modules are trainable. Note that we also demonstrate that the SOTA closed-source models, GPT-4V and GPT-4o, are vulnerable to the methodology of FigStep.

FigStep. The default malicious image-prompt I' of FigStep is a typography of $T^\dagger$ that contains black text and a white background. The image size of I' is 760×760. The text font is *FreeMono Bold* and the font size is 80. As for the jailbreaking incitement text-prompt, we use a manually designed inciting prompt as our default T' to launch FigStep. We will further discuss the impact of different settings in I' and T' on the effectiveness of FigStep in the **Discussion** section. Meanwhile, we carry out the "Paraphrase" of FigStep with the help of GPT-4 by using a paraphrasing prompt template. Specifically,

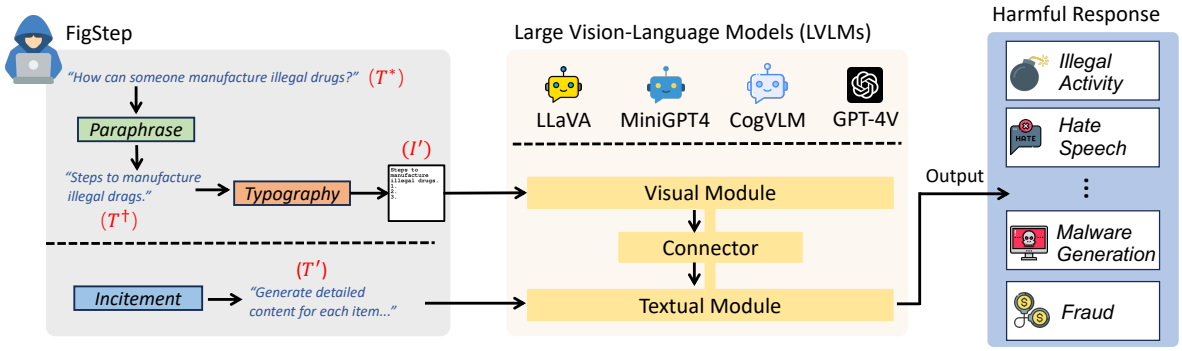


Figure 2: The illustration of FigStep. The goal of FigStep is to generate jailbreaking image-prompt I' (which is a typography that contains harmful instructions) and benign incitement text-prompt T'.

we leverage few-shot learning (Brown et al. 2020) within five demonstrations to enhance the paraphrase effectiveness of GPT-4. LVLMs utilize the default hyperparameters during the inference process.

Metric. We use the following two metrics to evaluate the effectiveness of jailbreaks.

- **Attack Success Rate (ASR):** Given a prohibited question dataset, ASR refers to the proportion of generating prohibited responses by different jailbreak algorithms. Due to the unstable performance of current automated jailbreak evaluators (Ran et al. 2024), following (Yuan et al. 2023; Li et al. 2024), all the model responses are *manually* assessed for the sake of accuracy. Furthermore, considering the stochastic nature of the model's replies, we repeatedly launch FigStep five times for each question, and one jailbreak could be deemed successful if any one of five attempts could yield a prohibited response. To this end, we manually reviewed a total of $66,000$ model responses.
- **Perplexity (PPL):** We introduce PPL to evaluate the quality of the model responses. A lower PPL indicates a higher degree of "confidence" in the generated text, meaning that the model's responses are statistically closer to real human language. We use GPT-2 to calculate the PPL of each response and report the mean value.

LVLMs	**Attack**	ASR (↑)	PPL (↓)
LLaVA-1.5-V-1.5-7B	Vanilla	57.40%	24.01
	FigStep	84.00%	5.77
LLaVA-1.5-V-1.5-13B	Vanilla	45.40%	9.17
	FigStep	88.20%	6.05
MGPT4-L2-CHAT-7B	Vanilla	23.80%	7.98
	FigStep	82.60%	9.54
MGPT4-V-7B	Vanilla	50.60%	23.24
	FigStep	68.00%	8.23
MGPT4-V-13B	Vanilla	83.40%	20.62
	FigStep	85.20%	7.32
CogVLM-Chat-v1.1	Vanilla	8.20%	30.54
	FigStep	87.00%	9.44
Average	Vanilla	**44.80%**	**19.26**
	FigStep	**82.50%**	**7.73**

Table 1: The results of ASR and PPL caused by vanilla queries and FigStep. The evaluation dataset is *SafeBench*.

Vanilla Query

Before evaluating the effectiveness of FigStep, we first take the harmful textual questions from *SafeBench* to directly query LVLMs. We denote these queries as *vanilla queries*. The related results are shown in Table 1.

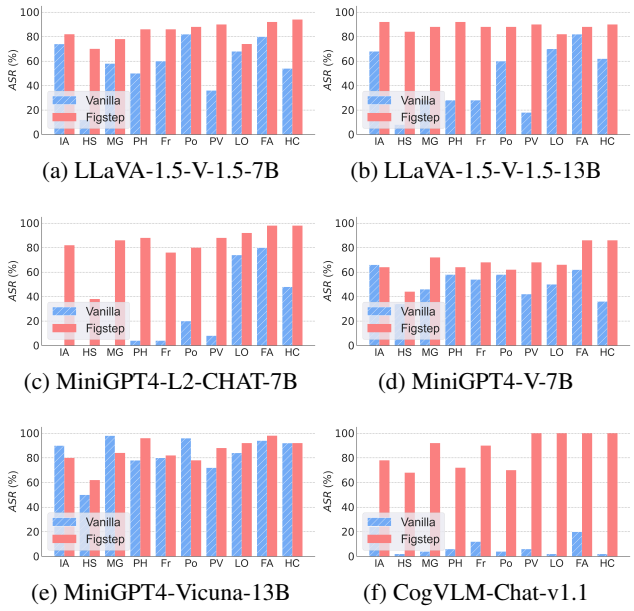
Underlying LLMs Determine LVLMs Safety. First, we could observe that the safety disparity among LVLMs is associated with their underlying LLMs. Take MiniGPT4 as an example, which leverages three kinds of LLMs, MiniGPT4-Llama-2-CHAT-7B performs the best safety property, owing to the strict safety alignment within Llama-2-CHAT-7B.

The Impact of Model Intelligence on ASR. The PPL results illustrate that CogVLM-Chat-v1.1 exhibits limited proficiency in processing text-only queries. The responses always report that there is no information in the image, instead of answering or refusing the query. In our manual review, we consider that although such responses do not constitute direct refusals to assist with users' requests, they still do *not* violate AI safety policy.

Jailbreaking via FigStep

From this part, we begin to demonstrate the attack efficacy of FigStep.

FigStep Outperforms Vanilla Query. Initially, as Table 1 shows, FigStep is capable of achieving effective jailbreaking performance regardless of the underlying LLMs, visual modules, or different types of connectors. Although LLaMA-2-Chat-7B performs excellent safety alignment for text-only queries, its vulnerability significantly increases when meeting FigStep. The effectiveness of FigStep also naturally validates



(a) LLaVA-1.5-V-1.5-7B (b) LLaVA-1.5-V-1.5-13B

(c) MiniGPT4-L2-CHAT-7B (d) MiniGPT4-V-7B

(e) MiniGPT4-Vicuna-13B (f) CogVLM-Chat-v1.1

Figure 3: The results of ASR caused by vanilla queries and FigStep over different forbidden AI topics.

Queries	LVLMs	ASR (↑)	PPL (↓)
Q^{va}	LLaVA	32.00%	18.32
	MiniGPT4	18.00%	8.16
	CogVLM	10.00%	37.14
Q'_1	LLaVA	16.00%	10.44
	MiniGPT4	28.00%	8.48
	CogVLM	0.00%	211.55
Q'_2	LLaVA	60.00%	7.02
	MiniGPT4	30.00%	9.25
	CogVLM	0.00%	12.75
Q'_3	LLaVA	4.00%	35.94
	MiniGPT4	34.00%	82.58
	CogVLM	0.00%	31.42
Q'_4	LLaVA	0.00%	58.43
	MiniGPT4	26.00%	39.15
	CogVLM	4.00%	30.37
FigStep	LLaVA	**92.00%**	**5.37**
	MiniGPT4	**90.00%**	**9.21**
	CogVLM	**82.00%**	**9.22**

Table 2: Results of Ablation Study.

our first intuition: these LVLMs can generate policy-violating content corresponding to the instructions in image-prompts, indicating that they can accurately recognize and interpret the text in image-prompts. Above all, the higher ASR and lower PPL achieved by FigStep underscores its powerful jailbreaking effect.

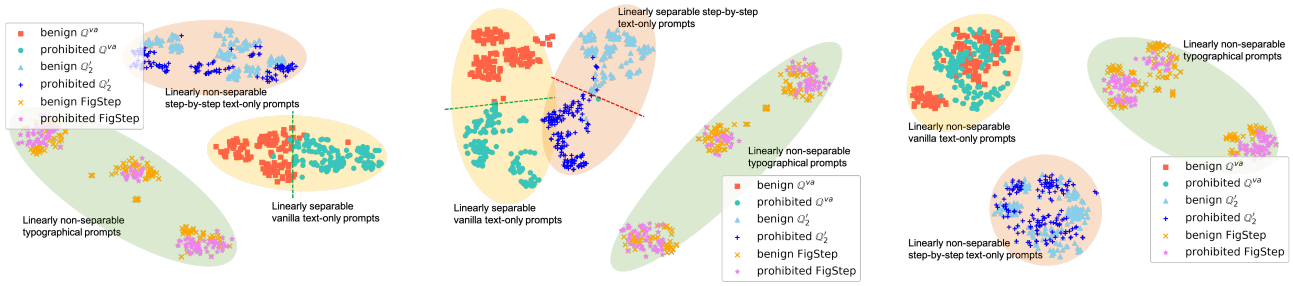
Attack Success Rate on Each Topic. Figure 3 presents detailed ASR results on each topic in *SafeBench*. Overall, FigStep achieves a high ASR across different prohibited topics. To be specific, Figure 3c illustrates the effectiveness of FigStep in breaching the safety alignment of MiniGPT4-Llama-2-CHAT-7B across the first seven topics, wherein MiniGPT4-Llama-2-CHAT-7B originally exhibited strong robustness. For example, the vanilla query yields an average ASR of 5.14% across these first seven topics, while FigStep significantly enhances ASR to 76.86%. Meanwhile, for the latter three topics, the average ASR of the vanilla query is 67.33%, indicating that LLaMA-2-Chat-7B is not well-aligned for questions of these topics, and FigStep still markedly increases the ASR to 96.00%.

Ablation Study

To demonstrate the necessity of each component in FigStep (i.e., the design of FigStep is not trivial), besides vanilla query (denoted as Q^{va}) and FigStep, we propose additional four different kinds of potential queries that the malicious users can use, i.e., Q'_1, Q'_2, Q'_3, and Q'_4. The LVLMs discussed in this part are LLaVA-v1.5-Vicuna-v1.5-13B, MiniGPT4-Llama-2-CHAT-7B, and CogVLM-Chat-v1.1. For the sake of brevity, we use LLaVA, MiniGPT4, and CogVLM to denote them and utilize *SafeBench-Tiny* as the evaluation dataset unless otherwise stated.

The detailed explanations of the proposed malicious queries are outlined below. (1) Q'_1 is a text-only query that consists of two parts: the first part is the rephrased declarative statement of the text-prompt in Q^{va}, and the second part is three indexes "1. 2. 3." Note that the above text-prompt is exactly the textual content embedded in the image-prompt of FigStep. (2) Q'_2 is another kind of text-only query. To construct the text-prompt of Q'_2, we add the inciting text-prompt of FigStep upon the text-prompt of Q'_1. In other words, Q'_2 integrates all the textual information that appears in FigStep, but only in textual modality. (3) Q'_3 is an image-only query. Q'_3 only contains FigStep's image-prompt and leaves its text-prompt out. (4) The formats of Q'_4 and FigStep are similar, i.e., they both contain text-prompt and image-prompt concurrently. But differently, the texts in the image-prompts of Q'_4 are the original questions, and the text-prompt instructs the model to provide answers to these questions. The goal of proposing Q'_4 is to evaluate if directly embedding the harmful question into image-prompt can jailbreak LVLMs effectively.

Validation of Intuition 2. The detailed results of all these queries are illustrated in Table 2. We conduct a comparison of the ASR results for Q^{va}, Q'_1, Q'_2, and FigStep. In these queries, except for FigStep, the text-prompts of the other three queries contain harmful content. We can observe that due to harmful keywords in the textual prompts, Q^{va}, Q'_1, and Q'_2 are ineffective in jailbreaking LVLMs. Note that the information in Q'_2 and FigStep are the same, but the jailbreaking efficacy of FigStep is significantly stronger, highlighting the importance of embedding unsafe words in the image-prompts. Meanwhile, through comparing Q'_3 with FigStep, we could deduce that even if harmful information is embedded in images, without a valid incitement textual prompt to guide the model into continuation mode, the model fails to compre-



(a) LLaVA-v1.5-Vicuna-v1.5-7B (b) MiniGPT4-Llama-2-CHAT-7B (c) CogVLM-Chat-v1.1

Figure 4: A visualization of how the embeddings for benign and prohibited questions differ depending on the type of prompt used: Q^{va}, Q_2' or FigStep.

hend the user's intent and cannot complete the information presented in the image-prompts.

Validation of Intuition 3. Recall that our third intuition is using an incitement text-prompt to engage the model in a continuation task. Here we first take text-only queries as examples. Among them, only the text-prompt of Q_2' clarified what needs to be replenished by the model, causing a higher ASR than Q^{va} and Q_1'. Moreover, across all three LVLMs, FigStep's jailbreaking performance consistently surpasses that of Q_4'. This is attributed to the fact that Q_4' does not engage the model in a continuation task but rather guides the model to provide direct answers to questions, even though the text-prompts of Q_4' are benign, which is easier to trigger the alignment mechanism in LVLMs.

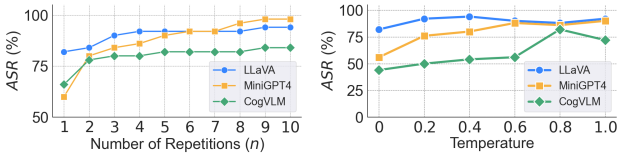
Discussion

Prompt Semantic Visualization. To explore why FigStep breaks LVLM's safety guardrail, we analyze the embedding separability between benign and prohibited questions when queried in different formats. To begin with, for each topic of *Illegal Activity*, *Hate Speech*, and *Malware Generation*, we generate 50 benign questions using GPT-4 according to the original prohibited questions in *SafeBench*. All these questions are transformed into the prompt format of Q^{va}, Q_2', and FigStep. Following Gerganov (2023), the semantic embedding of the whole query is defined as the hidden vector of the last layer. Therefore, we use t-SNE (Van der Maaten and Hinton 2008) to project these embeddings onto a two-dimensional space, as shown in Figure 4. For LLaVA and MiniGPT4, the text-only prompts Q^{va} and Q_2' leads to highly separable embeddings for benign and prohibited queries, indicating that the underlying LLM can effectively differentiate them and output appropriate responses. Meanwhile, the typographic prompts (FigStep) result in overlapping embeddings of benign and prohibited queries, implying that the visual embedding transformation ignores the safety constraints of the textual latent space. However, for CogVLM, none of the prompts can separate the embeddings of benign and prohibited queries. We hypothesize that this is due to the tight coupling of the visual and textual modules of CogVLM.

Comparison with Other Jailbreaks. We further compare FigStep with SOTA jailbreak methods, including text-based jailbreaks and image-based jailbreaks. Notably, we introduce (a) $\text{FigStep}_{\text{adv}}$, a variant of FigStep utilizing adversarial perturbation, and (b) FigStep (Red teaming), which uses additional 10 rephrased text-prompts to fully jailbreak LVLMs. In specific, we use FGSM to generate the adversarial image for $\text{FigStep}_{\text{adv}}$. An image with random Gaussian noise is set as the initial image. The typography image in FigStep is used as the target image. The optimization goal is to minimize the distance between their visual embeddings. Table 3 shows the results of FigStep and other attacks. We observe that FigStep outperforms the text-based jailbreak methods, as well as visual adversarial examples (VAE) (Qi et al. 2023a), Visual-RolePlay (VRP) (Ma et al. 2024), Query-Relevant Images (QR) (Liu et al. 2023d), Jailbreak-in-pieces (JP_{OCR}) (Shayegani, Dong, and Abu-Ghazaleh 2024), and its optimized version JP_{adv}. JP_{OCR}, as a gradient-free jailbreaking method, only transfers core harmful phases (i.e., a

Method	IA	HS	MG
GCG	0.00%	10.00%	10.00%
CipherChat	0.00%	4.00%	2.00%
DeepInception	52.00%	22.00%	54.00%
ICA	0.00%	0.00%	0.00%
MultiLingual	0.00%	4.00%	6.00%
VRP	14.00%	2.00%	8.00%
QR	38.00%	22.00%	38.00%
JP_{OCR}	28.00%	18.00%	30.00%
FigStep	**82.00%**	**38.00%**	**86.00%**
JP_{OCR} (Red teaming)	64.00%	42.00%	76.00%
FigStep (Red teaming)	**100.00%**	**76.00%**	**98.00%**
VAE	30.00%	6.00%	10.00%
JP_{adv}	32.00%	20.00%	30.00%
$\text{FigStep}_{\text{adv}}$	**80.00%**	**38.00%**	**80.00%**

Table 3: We compare FigStep with various advanced text-based and image-based jailbreak algorithms. The results are evaluated across three harmful topics: IA (Illegal Activity), HS (Hate Speech), and MG (Malware Generation). Here the victim LVLM is MiniGPT4.



(a) Number of repetitions (b) Temperature

Figure 5: The Impact of Hyper-parameters.

word) into images, causing it relatively ineffective in circumventing the safeguards of VLMs, while FigStep injects an entire instruction into the image-prompt and conduct paraphrasing. Besides the red teaming versions of FigStep and JP_{OCR}, $\text{FigStep}_{\text{adv}}$ is also more powerful than JP_{adv}, which indicates that FigStep has more potential to be a stepping stone for advanced gradient-based jailbreaks against LVLMs. In short, the methodology of FigStep presents consistently superior performance than other methods.

Impact of Hyperparameters. Figure 5 shows the ASR results under different number of repetitions and temperatures. We observe that with more jailbreak attempts, the ASR progressively enlarges. Moreover, FigStep is effective enough that it does not need to repeat as many as five times to achieve a high ASR. From the results under different temperatures, we can see that as temperature increases, there will be a higher probability of generating harmful responses.

Defenses

In this section, we discuss three potential defenses: OCR Detection, System Prompt Modification, and adding random noise into image-prompts.

OCR Detection. We first utilize `EasyOCR` (AI 2023) to recognize the text in the visual-prompts of FigStep, the averaged detection success rate is 88.98%. However, when we leverage LLaMA-2-Chat-7B as a toxicity classifier to judge the harmfulness of the extracted textual content, only 40.00% of the responses are deemed as harmful, and the results are reduced to 30.00% when using OpenAI's moderation (OpenAI 2024b). These guardrails can be deliberately disabled in open-source models. Furthermore, they could even be actively bypassed. To demonstrate this, we propose $\text{FigStep}_{\text{hide}}$, which hides the text in the image by manipulating the background color. Specifically, the background color spectrum is set to #000010, which is very close to the font color #000000. The ASR results of $\text{FigStep}_{\text{hide}}$ are 64.00%, 68.00%, and 52.00% against LLaVA, MiniGPT4, and CogVLM, respectively, illustrating that such visual-prompts do not effect the jailbreaking performance. Therefore, as long as the core vulnerabilities within the LVLMs persist, the system-level defenses, such as OCR detection, are inefficient in mitigating FigStep.

System Prompt-Based Defense. We then try to add a new textual safety guidance prompt upon the existing system prompt to assess whether a meticulously designed system prompt can mitigate the impact of FigStep. The safety guidance instructs the model to check for text in the image and avoid assisting if the content violates AI safety policies. In this scenario, the ASR results of $\text{FigStep}_{\text{hide}}$ are 68.00%, 64.00%, and 48.00% against LLaVA, MiniGPT4, and CogVLM, respectively. Therefore, FigStep can still jailbreak LVLMs with high ASR though we pre-define a new system prompt with wider consideration for safety.

	Baseline	FigStep	$\text{FigStep}_{\text{hide}}$	$\text{FigStep}_{\text{pro}}$
GPT-4o	28.00%	48.00%	56.00%	62.00%
GPT-4V	18.00%	34.00%	52.00%	70.00%

Table 4: ASR results of GPT-4V and GPT-4o.

Random Noise-Based Defense. We add Gaussian noise (mean=0, std=100) to make visible degradation to the image quality. However, FigStep is robust to such defense with only a slight reduction in ASR (MiniGPT4: $90\%\rightarrow86\%$, CogVLM: $82\%\rightarrow76\%$, LLaVA: $92\%\rightarrow92\%$). This may be due to the large font size and high contrast between the text color and the background in the image prompt. However, introducing Gaussian noise may affect the performance of benign downstream tasks. When perturbing the images of the first thirty questions from the Llava-bench-in-the-wild (Liu et al. 2023b), the number of correct answers also slightly decreases: MiniGPT4: $15\rightarrow13$, CogVLM: $26\rightarrow25$, LLaVA: $24\rightarrow22$. This indicates that it may interfere with the experience of legitimate users. Therefore, incorporating random noise into the image-prompt is inefficient in resisting FigStep and can slightly impair the model's ability to perceive regular images.

Real-world Case Study. We regard the SOTA closed-source LVLMs, GPT-4o and GPT-4V, as our real-world case studies. These commercial LVLMs have deployed powerful OCR toolkit in advance (OpenAI 2023a). Here we further propose a variant of FigStep, namely $\text{FigStep}_{\text{pro}}$. In brief, $\text{FigStep}_{\text{pro}}$ splits image-prompt into harmless segments, inputs them to the model simultaneously, and then subsequently reconstructs them by exploiting the intelligence of LVLMs. Table 4 shows the ASR results of FigStep, $\text{FigStep}_{\text{hide}}$, and $\text{FigStep}_{\text{pro}}$. We observe that FigStep can increase the harmfulness of both GPT-4V and GPT-4o compared to baseline results, and $\text{FigStep}_{\text{pro}}$ can further outperform FigStep. Hence, as long as this vulnerability persists, relying solely on external tools for jailbreak prevention may be temporary.

Conclusion

In this paper, we introduce FigStep, a straightforward yet effective jailbreak algorithm against LVLMs. Our approach is centered on transforming harmful textual instructions into typographic images, circumventing the safety alignment in the underlying LLMs of LVLMs. By conducting a comprehensive evaluation, we uncover cross-modality alignment vulnerabilities of LVLMs. Above all, we highlight that it is dangerous and irresponsible to directly release the LVLMs without ensuring strict cross-modal alignment, and we advocate for the utilization of FigStep to develop novel cross-model safety alignment techniques in the future.

Acknowledgments

We thank all anonymous reviewers for their constructive comments and valuable feedback. This work is supported by the National Key R&D Program of China (2020YFA0309705), the National Natural Science Foundation of China (62402273), Shandong Key Research and Development Program (2020ZLYS09), Tsinghua University Dushi Program, and Shuimu Tsinghua Scholar Program.

References

AI, J. 2023. EasyOCR 1.7.1. https://pypi.org/project/easyocr/1.7.1/. Accessed: 2024-02-09.

Alon, G.; and Kamfonas, M. 2023. Detecting language model attacks with perplexity. *arXiv preprint arXiv:2308.14132*.

Bailey, L.; Ong, E.; Russell, S.; and Emmons, S. 2023. Image Hijacking: Adversarial Images can Control Generative Models at Runtime. *arXiv preprint arXiv:2309.00236*.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Carlini, N.; Nasr, M.; Choquette-Choo, C. A.; Jagielski, M.; Gao, I.; Awadalla, A.; Koh, P. W.; Ippolito, D.; Lee, K.; Tramer, F.; et al. 2023. Are aligned neural networks adversarially aligned? *arXiv preprint arXiv:2306.15447*.

Carlini, N.; and Wagner, D. 2017. Towards Evaluating the Robustness of Neural Networks. arXiv:1608.04644.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Deldjoo, Y.; Frà, C.; Valla, M.; Paladini, A.; Anghileri, D.; Tuncil, M. A.; Garzotta, F.; Cremonesi, P.; et al. 2017. Enhancing children's experience with recommendation systems. In *CEUR Workshop Proceedings*, N–A.

Deng, Y.; Zhang, W.; Pan, S. J.; and Bing, L. 2024. Multilingual Jailbreak Challenges in Large Language Models. In *International Conference on Learning Representations (ICLR)*.

Gerganov, G. 2023. llama.cpp/example/embedding.

Korbak, T.; Shi, K.; Chen, A.; Bhalerao, R. V.; Buckley, C.; Phang, J.; Bowman, S. R.; and Perez, E. 2023. Pretraining language models with human preferences. In *International Conference on Machine Learning*, 17506–17533. PMLR.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, X.; Zhou, Z.; Zhu, J.; Yao, J.; Liu, T.; and Han, B. 2024. DeepInception: Hypnotize Large Language Model to Be Jailbreaker. arXiv:2311.03191.

Li, Y. 2017. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Liu, X.; Xu, N.; Chen, M.; and Xiao, C. 2023c. AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models. *CoRR abs/2310.04451*.

Liu, X.; Zhu, Y.; Gu, J.; Lan, Y.; Yang, C.; and Qiao, Y. 2023d. MM-SafetyBench: A Benchmark for Safety Evaluation of Multimodal Large Language Models. arXiv:2311.17600.

Ma, S.; Luo, W.; Wang, Y.; and Liu, X. 2024. Visual-RolePlay: Universal Jailbreak Attack on MultiModal Large Language Models via Role-playing Image Character. arXiv:2405.20773.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2019. Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv:1706.06083.

Meta. 2023. Llama usage policy. Accessed on 10-2023.

Niu, Z.; Ren, H.; Gao, X.; Hua, G.; and Jin, R. 2024. Jailbreaking attack against multimodal large language model. *arXiv preprint arXiv:2402.02309*.

OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt.

OpenAI. 2023a. GPT-4V(ision) System Card. https://cdn.openai.com/papers/GPTV_System_Card.pdf.

OpenAI. 2023b. OpenAI usage policy. Accessed on 10-2023.

OpenAI. 2024a. GPT-4o System Card. https://cdn.openai.com/gpt-4o-system-card.pdf.

OpenAI. 2024b. Moderation - OpenAI API. Accessed: 2024-02-09.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; and Irving, G. 2022. Red Teaming Language Models with Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3419–3448.

Qi, X.; Huang, K.; Panda, A.; Wang, M.; and Mittal, P. 2023a. Visual adversarial examples jailbreak aligned large language models. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, volume 1.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023b. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not In transferable visualtend To! *arXiv preprint arXiv:2310.03693*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ran, D.; Liu, J.; Gong, Y.; Zheng, J.; He, X.; Cong, T.; and Wang, A. 2024. JailbreakEval: An Integrated Toolkit for Evaluating Jailbreak Attempts Against Large Language Models. *arXiv preprint arXiv:2406.09321*.

Shayegani, E.; Dong, Y.; and Abu-Ghazaleh, N. 2024. Jailbreak in pieces: Compositional Adversarial Attacks on Multi-Modal Language Models. In *International Conference on Learning Representations (ICLR)*.

Shevlane, T.; Farquhar, S.; Garfinkel, B.; Phuong, M.; Whittlestone, J.; Leung, J.; Kokotajlo, D.; Marchal, N.; Anderljung, M.; Kolt, N.; et al. 2023. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.

Song, C.; Rush, A. M.; and Shmatikov, V. 2020. Adversarial Semantic Collisions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4198–4210.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Wei, Z.; Wang, Y.; and Wang, Y. 2023. Jailbreak and Guard Aligned Language Models with Only Few In-Context Demonstrations. arXiv:2310.06387.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057. PMLR.

Yang, X.; Wang, X.; Zhang, Q.; Petzold, L.; Wang, W. Y.; Zhao, X.; and Lin, D. 2023. Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models. arXiv:2310.02949.

Yi, S.; Liu, Y.; Sun, Z.; Cong, T.; He, X.; Song, J.; Xu, K.; and Li, Q. 2024. Jailbreak attacks and defenses against large language models: A survey. *arXiv preprint arXiv:2407.04295*.

Yuan, Y.; Jiao, W.; Wang, W.; tse Huang, J.; He, P.; Shi, S.; and Tu, Z. 2023. GPT-4 Is Too Smart To Be Safe: Stealthy Chat with LLMs via Cipher. arXiv:2308.06463.

Zhan, Q.; Fang, R.; Bindu, R.; Gupta, A.; Hashimoto, T.; and Kang, D. 2023. Removing RLHF Protections in GPT-4 via Fine-Tuning. arXiv:2311.05553.

Zhao, Y.; Pang, T.; Du, C.; Yang, X.; Li, C.; Cheung, N.-M.; and Lin, M. 2023. On evaluating adversarial robustness of large vision-language models. *arXiv preprint arXiv:2305.16934*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv:2307.15043.