The Security Threat of Compressed Projectors in Large Vision-Language Models

Yudong Zhang[1,2], Ruobing Xie[2,✉], Xingwu Sun[2,4], Jiansheng Chen[3,✉], Zhanhui Kang[2], Di Wang[2], Yu Wang[1,✉]

[1]Department of Electronic Engineering, Tsinghua University,
[2]Large Language Model Department, Tencent,
[3]School of Computer and Communication Engineering, University of Science and Technology Beijing, [4]Faculty of Science and Technology, University of Macau

zhangyd16@mails.tsinghua.edu.cn, xrbsnowing@163.com, sunxingwu01@gmail.com, jschen@ustb.edu.cn, kegokang@tencent.com, diwang@tencent.com, yu-wang@mail.tsinghua.edu.cn. (✉: Corresponding authors)

Abstract

The choice of a suitable visual language projector (VLP) is critical to the successful training of large visual language models (LVLMs). Mainstream VLPs can be broadly categorized into compressed and uncompressed projectors, and each offers distinct advantages in performance and computational efficiency. However, their security implications have not been thoroughly examined. Our comprehensive evaluation reveals significant differences in their security profiles: compressed projectors exhibit substantial vulnerabilities, allowing adversaries to successfully compromise LVLMs even with minimal knowledge of structure information. In stark contrast, uncompressed projectors demonstrate robust security properties and do not introduce additional vulnerabilities. These findings provide critical guidance for researchers in selecting optimal VLPs that enhance the security and reliability of visual language models. The code is available at https://github.com/btzyd/TCP.

1 Introduction

Large vision-language models (LVLMs) have achieved remarkable success in various multimedia applications (Dai et al., 2023; Liu et al., 2023a; Bai et al., 2025; Wang et al., 2024; Chen et al., 2024), particularly in tasks like visual question answering (VQA). A typical LVLM framework consists of three key components: visual encoder (VE), large language model (LLM), and vision-language projector (VLP). Current training methodologies primarily focus on optimizing the VLP to effectively integrate visual features extracted from a pre-trained VE into a pre-trained LLM. This approach substantially minimizes computational demands when compared to building an LVLM from scratch, thereby allowing researchers to leverage existing advancements. Moreover, it ensures greater stability and reliability in both the training process and practical deployment of LVLMs.

Recently, a growing body of research has employed CLIP (Radford et al., 2021; Fang et al., 2023) as the VE and initiated training on various pre-trained LLMs. The choice of VLP plays a crucial role in determining the effectiveness of LVLM training. VLPs can generally be classified into two distinct categories: compressed projectors and uncompressed projectors. **Compressed projectors**, exemplified by Q-former (Li et al., 2023; Dai et al., 2023; Zhu et al., 2024), achieve high computational efficiency by compressing a larger number of visual tokens into a smaller dimension using query tokens. On the other hand, **uncompressed projectors**, typified by MLP-based architectures (Liu et al., 2023a; Shi et al., 2024), convert visual tokens to feature dimensions matching the LLM, with the number of output tokens proportional to the number of input tokens.

Previous studies have tended to focus only on the security of visual encoders or LLMs. On the visual encoder, Wang et al. (2023); Yin et al. (2023) present adversaries with an opportunity to compromise model security by targeting solely the visual encoder. As for LLMs, previous studies (Carlini et al., 2023; Zou et al., 2023; Chao et al., 2023) have fully explored the safety of LLM, and the user has the option of choosing a better performing and safer LLM on his or her own. However, the security of visual language projectors has not been fully explored, and only a few studies have focused on attacks on VLPs. Although some studies (Zhang et al., 2025) have attempted to attack the VLP (mainly Q-former), they have not thoroughly analyzed the security of the VLP structure.

While compressed and uncompressed projectors each demonstrate distinct strengths in terms of computational efficiency and model performance, their respective trade-offs are particularly evident in large language and vision models. Compressed projectors have shown superior efficiency, particularly excelling in high-resolution visual under-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20397–20407
November 4-9, 2025

standing tasks and video comprehension scenarios. In contrast, uncompressed projectors maintain greater computational capabilities at the cost of higher computational expense. Previous studies (Yao et al., 2024) have extensively analyzed these trade-offs from perspectives of computational efficiency and model performance; however, a critical yet underexplored dimension in LVLM applications remains their security implications.

We assess the vulnerabilities of both compressed and uncompressed projectors by conducting adversarial attacks across diverse white-box and gray-box scenarios to evaluate the security implications for different VLP architectures. Our findings reveal three key insights: (1) Compressed projectors exacerbate security vulnerabilities beyond those affecting visual encoders. Specifically, an adversary can significantly degrade model performance by targeting compressed projectors, even with limited knowledge about the VLP. (2) In contrast, uncompressed projectors do not introduce additional security risks, as attacking them yields results comparable to attacking visual encoders directly. (3) This discrepancy in security between compressed and uncompressed projectors stems from their architectural differences and remains independent of the number of visual tokens. Notably, even when the visual tokens of an uncompressed projector are reduced (*e.g.*, through pooling) to approach those of a compressed projector, the uncompressed projector still maintains its robust security characteristics.

Based on our analyses, we propose the following suggestions: (1) We strongly recommend employing uncompressed projectors rather than compressed projectors in high-security environments to mitigate potential risks from adversarial attacks. (2) In scenarios where computational efficiency is a priority, researchers can implement techniques like pooling operations to proportionally reduce the number of their output tokens. This approach offers enhanced security compared to using compressed projectors while maintaining acceptable performance levels, thereby striking an appropriate balance among effectiveness, efficiency, and security.

Our research provides three key contributions: (1) This study is **the first to systematically investigate VLP security**, presenting a comprehensive and systematic evaluation of various VLP architectures. (2) Our experimental results **uncover significant security vulnerabilities in compressed projectors**, revealing critical implications for models that employ such components. (3) We demonstrate that operating on visual tokens generated by uncompressed projectors offers enhanced security compared to compressed projectors when the number of visual tokens is largely comparable, thereby identifying promising approaches for VLP selection under efficiency constraints.

2 Related Work

Large Vision-Language Models. Training LVLMs from scratch is often prohibitively expensive. Thus, current popular frameworks typically begin with a pre-trained unimodal visual encoder and a large language model, focusing their training efforts on developing a VLP to connect the two modalities and effectively incorporate visual features into the LLM framework. The mainstream VLPs currently fall into two primary categories: compressed projectors and uncompressed projectors. Compressed projectors, such as Q-former (Li et al., 2023), Resampler (Alayrac et al., 2022), and D-Abstractor (Cha et al., 2024), reduce the number of visual tokens to a fixed number of output tokens. In contrast, uncompressed projectors, exemplified by MLP (Liu et al., 2023a), maintain a proportional relationship between the number of visual tokens and their corresponding output tokens.

Compressed and uncompressed projectors present distinct trade-offs in terms of computational efficiency, memory usage, and implementation complexity. On one hand, uncompressed projectors offer simplicity in design but incur significant computational costs. Specifically, Specifically, their computational requirements increase significantly as the token length scales linearly with the square of the resolution (Chen et al., 2024). Furthermore, in video processing applications (Ren et al., 2024), this scaling behavior results in a linear growth of required length relative to the number of video frames. On the other hand, compressed projectors achieve improved efficiency by reducing redundancy within the visual space. This is accomplished through the compression of token lengths from the visual encoder to a specified capacity, enabling strong performance while maintaining high computational efficiency. While prior research has extensively studied the performance and efficiency trade-offs between these two approaches in VLPs, there remains a critical gap: the safety implications of these methods have been largely neglected.

Adversarial Attacks on LVLMs. Prior research

has shown that deep neural networks are vulnerable to adversarial perturbations (Szegedy et al., 2014; Nguyen et al., 2015). While extensive studies have investigated the security of VEs and LLMs (Shayegani et al., 2024; Carlini et al., 2023; Liu et al., 2023b; Shayegani et al., 2023), relatively little attention has been devoted to examining the security of VLPs. Our work aims to address this gap by conducting adversarial attacks on both compressed and uncompressed projectors to evaluate their respective security properties.

This paper focuses on a practical gray-box scenario where the adversary is aware of both the VE weights and the structure of VLP, whether it be compressed or uncompressed. By crafting specialized loss functions tailored to the VLP architecture, the adversary can effectively execute targeted adversarial attacks against LVLMs in both white-box and gray-box settings. Our findings not only expose the vulnerabilities inherent in compressed projectors but also provide new insights into selecting between compressed and uncompressed projectors from a security perspective.

3 Method

3.1 Preliminaries

Notations. We provide a brief overview of the definition of notations. An LVLM, denoted as f, generally comprises three key components: (1) the visual encoder f_{VE}, typically based on CLIP; (2) the vision-language projector f_{VLP}, which is usually implemented as a Q-former or an MLP; and (3) the large language model f_{LLM}. The LVLM f accepts an image x_i and an instruction x_t as input, producing an output y. The process begins with the input image x_i being processed by f_{VE} to generate the visual feature representation $f_{\text{VE}}(x_i)$. Next, depending on whether it is a compressed or uncompressed VLP: (1) for the compressed VLP, f_{VLP} extracts relevant features from both the visual features $f_{\text{VE}}(x_i)$ and the instruction x_t; while (2) for the uncompressed VLP, f_{VLP} transforms the visual token dimension to align with the LLM input space. Both processes result in the projection output $f_{\text{VLP}}(f_{\text{VE}}(x_i), x_t)$. These projected features are then fed into f_{LLM}, which generates the final output y as Eq. (1):

$$y = f(x_i, x_t) = f_{\text{LLM}}(f_{\text{VLP}}(f_{\text{VE}}(x_i), x_t), x_t). \quad (1)$$

Compressed projectors. The Q-former stands as the canonical representative of compressed projectors, functioning as a lightweight yet trainable querying transformer. It effectively extracts textually relevant features from the VE CLIP through a set of learnable queries. Notably, during BLIP-2's two-stage pre-training process, both the VE and the LLM remain static, with only the Q-former and its queries being trainable components. This training paradigm has been successfully employed in subsequent studies, such as InstructBLIP (Dai et al., 2023) and MiniGPT-4 (Zhu et al., 2024), which also adopt similar approaches.

Uncompressed projectors. A typical approach involves using a simple multilayer perceptron (MLP) to project the visual token representation into the input space of an LLM. For instance, in LLaVA v1.5 (Liu et al., 2023a), the visual encoder outputs a 1024-dimensional embedding, which is then mapped to 4096 dimensions via an MLP to align with the input requirements of the Vicuna LLM. Similarly, Eagle (Shi et al., 2024) concatenates feature representations from multiple visual encoders to form a longer vector (typically ranging from 6000 to 8000 dimensions, depending on the number of visual encoders used), which is projected down to 4096 dimensions using an MLP.

The difference between compressed and uncompressed projectors. We adopt the definition from (Yao et al., 2024), where a compressed projector (represented by the Q-former) extracts information from visual tokens *through a fixed number of query tokens*, with the output token count *fixed to a relatively small number*. In contrast, the uncompressed projector, implemented as an MLP, processes and reshapes both the quantity and dimensionality of visual tokens. Let N represent the number of tokens generated by a visual encoder. An uncompressed projector outputs tokens corresponding to fractions of N, specifically N, $N/2$, $N/4$, and so on. In contrast, a compressed projector generates a consistent number of tokens M (typically where $M \ll N$).

3.2 An Empirical Analysis of Component Accessibility Within LVLM Attacks

Visual encoder (VE). Current popular LVLMs, such as LLAVA (Liu et al., 2023a), BLIP-2 (Li et al., 2023), InstructBLIP (Dai et al., 2023), and MiniGPT-4 (Zhu et al., 2024), employ either the ViT-L/14 model from CLIP (Radford et al., 2021) or the ViT-g/14 model from EVA-CLIP (Fang et al., 2023) as their visual encoders. Notably, none of these LVLMs fine-tune CLIP during training, which allows adversaries to easily access the visual

encoder weights directly through publicly available CLIP checkpoints.

Vision-language Projector (VLP). We assume the adversary possesses knowledge of the VLP's architectural details, such as its implementation as a Q-former or MLP. Notably, all LVLMs built on Q-former, including BLIP-2, InstructBLIP, and MiniGPT-4, share identical VLP architectural designs. However, while the adversary may be aware of these structure specifics, obtaining the exact model parameters remains challenging due to the diversity in pre-trained LLMs and training datasets.

Large language model (LLM). The ecosystem of LLM is highly diverse, with many open-source options available, including OPT (Zhang et al., 2022), LLaMA (Touvron et al., 2023), FlanT5 (Chung et al., 2024), and Vicuna (Chiang et al., 2023). Additionally, LVLMs often utilize customized datasets for parameter-efficient fine-tuning (Ding et al., 2023) of these LLMs. Furthermore, numerous powerful closed-source LLMs are also accessible to model developers. This diversity in both open-source and closed-source LLMs presents a significant challenge for adversaries attempting to determine the specific architectural details or weights of the LLMs employed within LVLMs.

The white-box setting. In the white-box setting, where adversaries have full access to model parameters (including both VEs and VLPs), we evaluate the safety of VLPs under worst-case scenarios. The adversary aims to modify a clean image to create an adversarial example capable of fooling the LVLM, leading to incorrect or unintended responses.

The gray-box setting. Building on the analysis of LVLMs' components (VE, VLP, and LLM) presented earlier, we focus on a practical yet hitherto unexplored gray-box setting, in which the adversary has access to the weights of the VE CLIP and the architectural structure of the VLP but lacks access to either the weights of the VLP or any information about the LLM. In this context, we can obtain a surrogate VLP model through transfer learning from other LVLMs, or alternatively, train the surrogate VLP model from scratch.

3.3 Surrogate Models Used in Attacks

Based on our analysis of the accessibility levels of various LVLM components presented in Sec. 3.2, we design and implement adversarial attacks against VLPs under three distinct scenarios with increasing levels of difficulty to comprehensively assess their security vulnerabilities. (1) The first scenario represents the simplest case, where the attack operates in a white-box setting, allowing full visibility into the model's architecture and parameters. (2) The second scenario involves attacking a surrogate VLP from a similar LVLM, targeting another specific LVLM. (3) The most challenging scenario requires an attacker to develop a surrogate VLP from scratch, significantly increasing the complexity of generating effective adversarial examples.

Surrogate models under white-box setting. In our white-box setting, both the VE and VLP of the target model are accessible. This allows us to leverage $\mathcal{L}_{\text{VE}}$ and $\mathcal{L}_{\text{VLP}}$ to attack the target LVLMs.

Surrogate models transferred from other LVLMs in gray-box setting. In our gray-box setting, while only the VE of the target model is available, we can still obtain surrogate VLPs by transferring knowledge from other similar LVLMs. For instance, if the target model is InstructBLIP Vicuna-13B, we can utilize the VLP from InstructBLIP FlanT5$_{\text{XL}}$ to attack it.

Surrogate models trained from scratch in gray-box setting. When similar LVLMs are unavailable for transfer, the structure insights into VLP provide us with a unique opportunity. Once the target model's VLP structure (using compressed or uncompressed projectors) is identified, we can train surrogate Q-formers or MLPs from scratch. The detailed training procedure is outlined in Sec. 4.1. Notably, during this process, no information about the specific LLM used by the target LVLM is required; only the VE weights and VLP structure are necessary for training the surrogate VLP.

3.4 Loss Function for More Effective Attacks

To evaluate the security of the VLP structure, we employ adversarial attack methods to analyze the robustness of LVLM. For a model with loss function $\mathcal{L}$, Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) performs an adversarial attack through gradient ascent as follows: $x_i' = x_i' + \nabla_{x_i'}\mathcal{L}$. In contrast, Projected Gradient Descent (PGD) (Madry et al., 2018) conducts multiple iterations of gradient ascent and projects x_i' onto the constrained perturbation space after each step. Typically, PGD restricts the l_∞ norm distance with $\|x_i - x_i'\|_\infty \leq \epsilon$. Additionally, a variant of the Carlini & Wagner (C&W) attack (Carlini and Wagner, 2017), referred to as CW-l_2, aims to maximize the loss function $\mathcal{L}$ while minimizing the l_2 norm distance $||x_i - x_i'||_2^2$.

To compare the security implications introduced

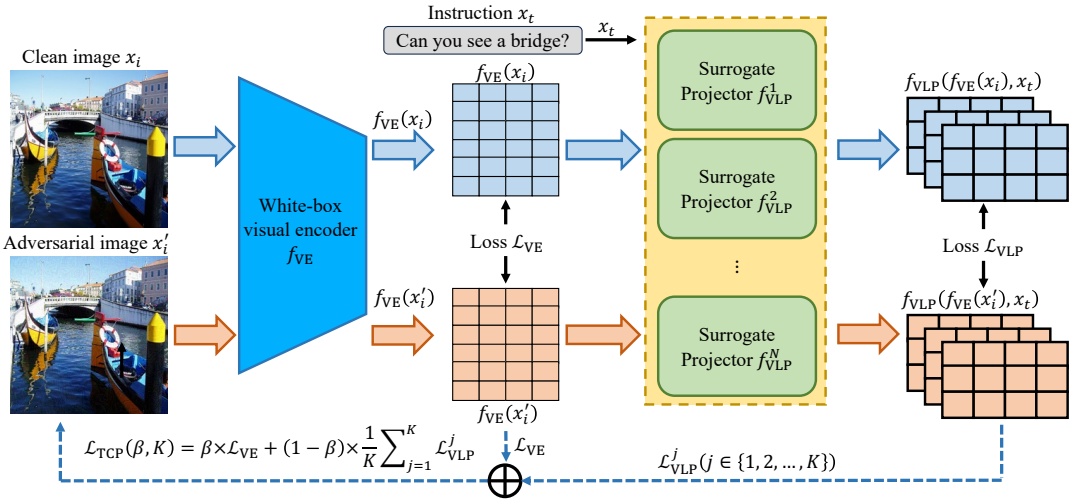

Figure 1: The attack pipeline operates by first extracting surrogate VLPs $\{f_{\text{VLP}}^1, \ldots, f_{\text{VLP}}^N\}$, which are then utilized to generate adversarial examples x_i' through the loss functions $\mathcal{L}_{\text{VE}}$, $\mathcal{L}_{\text{VLP}}$ and $\mathcal{L}_{\text{TCP}}(\beta, K)$ (TCP stands for "Threat of Compressed Projectors"). A key aspect of this investigation is determining whether incorporating attacks on VLPs increases security vulnerabilities compared to solely attacking VEs, thereby assessing the robustness of the VLP structure.

by the VLP structure, we evaluate the attack results on VE and VLP. For a clean image input, let $V_{\text{out}} = f_{\text{VE}}(x_i)$ and $P_{\text{out}} = P(f_{\text{VE}}(x_i), x_t)$ denote the outputs of VE and VLP, respectively. For an adversarial image, let $V'_{\text{out}} = f_{\text{VE}}(x_i')$ and $P'_{\text{out}} = P(f_{\text{VE}}(x_i'), x_t)$ represent the corresponding outputs, where x_i' is initialized by adding random noise to x_i. We perform gradient-based adversarial attacks on both VE and VLP using the loss functions defined in Eqs. (2) and (3), where I and J denote the sequence length and feature dimension:

$$\mathcal{L}_{\text{VE}} = \frac{1}{IJ} \sum_{I} \sum_{J} (V_{\text{out}} - V'_{\text{out}})^2 \quad (2)$$

$$\mathcal{L}_{\text{VLP}} = \frac{1}{IJ} \sum_{I} \sum_{J} (P_{\text{out}} - P'_{\text{out}})^2 \quad (3)$$

The proposed loss function $\mathcal{L}_{\text{TCP}}$ (TCP stands for "Threat of Compressed Projectors") combines these components, with $\beta \in [0, 1]$ controlling the trade-off between VE and VLP losses, and K denoting the number of surrogate VLPs used to enhance attack effectiveness. The overall loss function is:

$$\mathcal{L}_{\text{TCP}} = \beta \mathcal{L}_{\text{VE}} + (1 - \beta) \times \frac{1}{K} \Sigma_{j=1}^{K} \mathcal{L}_{\text{VLP}}^j \quad (4)$$

In our experiments, we adopt $\beta = 0$ and $K = 1$ by default, under which settings the loss functions $\mathcal{L}_{\text{TCP}}$ and $\mathcal{L}_{\text{VLP}}$ are mathematically equivalent.

We evaluate the robustness of VLP by feeding adversarial samples generated from attacking both VE and VLP into it. If attacking VLP yields stronger adversarial effects, this indicates that integrating VLP increases LVLM's vulnerability compared to VE, suggesting that VLP is less secure in this context. Conversely, if attacking VLP proves less effective than attacking VE, this implies that VLP exhibits superior security performance.

4 Experiments

4.1 Experimental Settings

Datasets. We randomly sampled 2000 image-question pairs from VQA v2 (Antol et al., 2015) to construct our primary dataset, referred to as "VQA v2 2000". Additionally, we include results for other datasets in Sec. 4.7, including ImageNet (Deng et al., 2009), VizWiz (Bigham et al., 2010), and COCO (Lin et al., 2014).

Evaluations. For VQA v2, we used the official evaluation program to compute the VQA scores. For other tasks, we employed various metrics such as accuracy and CIDEr (Vedantam et al., 2015).

Models. For the compressed projector, we implemented four variants of InstructBLIP (Dai et al., 2023), all of which employed the same ViT-g/14 vision encoder (Radford et al., 2021). These models differed solely in their weight and bias parameters within the post-normalization layer. Notably, despite incorporating different LLMs, all these models maintained an identical VLP architecture. In contrast, our experiments with the uncompressed

projector utilized a combined set of eight model variants from LLaVA-v1.5 (Liu et al., 2023a) and LLaVA-v1.6 (Liu et al., 2024).

Adversarial attack. We generated adversarial perturbations on clean images using PGD-l_∞ (Madry et al., 2018) and CW-l_2 (Carlini and Wagner, 2017). For PGD-l_∞, we set the attack step to 2/255, maximum perturbation to 8/255, and the number of attack steps to 20. For CW-l_2, we configured the attack step to 0.1 (equivalent to 2.55/255), with a constant of 0.005 and zero confidence.

Training details. (1) Surrogate compressed VLPs: we followed BLIP-2's two-stage pre-training process (Li et al., 2023). In the first stage, vision-language representation learning was performed using a frozen VE. In the second stage, both the VE and LLM were kept frozen while performing vision-to-language generative learning. To ensure architectural diversity, we used different LLMs during pre-training (*e.g.*, opt-2.7B and opt-6.7B) compared to the target LVLM. Pre-training was conducted using the COCO datasets. (2) Surrogate uncompressed VLPs: we followed LLAVA's official pre-training and fine-tuning settings. Similar to our compressed models, we used a different LLM during pre-training and fine-tuning to align with our gray-box experimental setup.

Multiple runs to ensure accuracy. To mitigate the impact of random factors, each experiment was repeated five times using different random seeds. We report the mean μ of the results. Additionally, we measured the variance σ across multiple runs, which generally remained low (around 0.1) with a maximum variance $\sigma_{\max} \leq 0.25$.

Attack method	VLP	LVLM	VQA scores Clean	$\mathcal{L}_{\text{VE}}$	$\mathcal{L}_{\text{VLP}}$
PGD	Uncompressed	LLaVA-v1.5-7B	77.31	69.30	71.21
		LLaVA-v1.5-13B	78.56	70.55	72.91
	Compressed	InstructBLIP-XL	72.43	43.51	34.15
		InstructBLIP-7B	75.53	44.94	41.48
		InstructBLIP-XXL	71.86	43.57	31.01
		InstructBLIP-13B	68.22	42.02	36.96
CW	Uncompressed	LLaVA-v1.5-7B	77.31	65.79	67.53
		LLaVA-v1.5-13B	78.56	67.23	69.89
	Compressed	InstructBLIP-XL	72.43	41.02	15.85
		InstructBLIP-7B	75.53	43.58	40.17
		InstructBLIP-XXL	71.86	41.50	11.52
		InstructBLIP-13B	68.22	39.95	33.95

Table 1: Results of attacks target VE and VLP under white-box setting. The difference in performance between $\mathcal{L}_{\text{VE}}$ and $\mathcal{L}_{\text{VLP}}$ highlights the security vulnerability of compressed projectors.

4.2 Results of White-box Setting

The results of adversarial attacks under the white-box setting are summarized in Tab. 1. Here, $\mathcal{L}_{\text{VE}}$ represents an adversarial attack specifically targeting the VE, as formally defined in Eq. (2), while $\mathcal{L}_{\text{VLP}}$ denotes an attack directed at the VLP, defined in Eq. (3). A key observation emerges under this setting: for compressed projectors, adversarial attacks targeting VLP achieve superior performance compared to those targeting VE. This discrepancy underscores the heightened security vulnerability of compressed projection. Conversely, in the case of uncompressed projectors, attacks against VLP yield results comparable to those targeting VE, demonstrating the enhanced robustness of uncompressed projectors. These initial findings from white-box attack experiments reveal critical security implications for compressed projectors.

Attack method	The source of surrogate MLP	Target models: LLaVA-v1.6 (uncompressed projector) Vicuna-7B	Mistral-7b	Vicuna-13B	Hermes-Yi-34B
Clean		78.32	79.10	75.61	79.07
PGD	$\mathcal{L}_{\text{VE}}$	70.55	72.31	68.92	71.27
	LLaVA-v1.5-7B	72.97	73.62	70.48	73.86
	LLaVA-v1.5-13B	73.60	73.77	69.25	72.85
CW	$\mathcal{L}_{\text{VE}}$	67.72	67.22	62.56	67.37
	LLaVA-v1.5-7B	68.11	69.10	65.94	68.76
	LLaVA-v1.5-13B	68.06	69.49	65.35	68.55
Attack method	The source of surrogate Q-former	Target models: InstructBLIP (compressed projector) $\text{FlanT5}_{\text{XL}}$	Vicuna-7B	$\text{FlanT5}_{\text{XXL}}$	Vicuna-13B
Clean		72.43	75.53	71.86	68.22
PGD	$\mathcal{L}_{\text{VE}}$	43.51	44.94	43.57	42.02
	InstructBLIP-7B	38.56	41.48	38.04	40.20
	InstructBLIP-13B	38.89	39.76	39.16	36.96
CW	$\mathcal{L}_{\text{VE}}$	41.02	43.58	41.50	39.95
	InstructBLIP-7B	36.73	40.17	35.38	38.60
	InstructBLIP-13B	37.18	38.93	35.81	33.95

Table 2: Results of attacking uncompressed and compressed projectors using surrogate VLP models from other similar LVLMs. The results are divided into two sections: uncompressed (top) and compressed (bottom). The findings demonstrate that uncompressed projectors exhibit superior robustness compared to their compressed counterparts.

4.3 Results of Attacks via Surrogate VLPs from Other LVLMs

We employ the surrogate VLP from other similar LVLMs to launch attacks against our target models. Specifically, for compressed projectors, we conduct experiments using InstructBLIP models that have been trained with various LLMs. For uncompressed projectors, we utilize LLaVA v1.5 as a surrogate model to attack LLaVA v1.6. As demonstrated in Tab. 2, our results indicate that uncompressed projectors maintain superior robustness under gray-box attack conditions, whereas

attacks targeting compressed projectors lead to further degradation in model performance.

4.4 Results of Attacks via Surrogate VLPs Trained from Scratch

We further explored training the surrogate VLP from scratch, using a different LLM than the target LVLM within the context of our gray-box attack scenario. Our experimental results, as evidenced by Tab. 3, demonstrate consistent findings with those presented in Tab. 2: uncompressed projectors exhibit superior security performance even under more challenging attacking scenarios.

Attack method	Surrogate MLP	Target models: LLaVA-v1.6 (uncompressed projector) Vicuna-7B	Mistral-7b	Vicuna-13B	Hermes-Yi-34B
Clean		78.32	79.10	75.61	79.07
PGD	$\mathcal{L}_{\text{VE}}$	70.55	72.31	68.92	71.27
	Vicuna-v1.3-7B	72.55	73.06	69.60	73.58
	Vicuna-v1.3-13B	73.71	73.76	69.66	73.24
CW	$\mathcal{L}_{\text{VE}}$	67.72	67.22	62.56	67.37
	Vicuna-v1.3-7B	69.28	71.03	66.09	70.18
	Vicuna-v1.3-13B	69.53	70.12	66.67	69.49
Attack method	Surrogate Q-former	Target models: InstructBLIP (compressed projector) $\text{FlanT5}_{\text{XL}}$	Vicuna-7B	$\text{FlanT5}_{\text{XXL}}$	Vicuna-13B
Clean		72.43	75.53	71.86	68.22
PGD	$\mathcal{L}_{\text{VE}}$	43.51	44.94	43.57	42.02
	opt-2.7B	39.82	41.81	39.89	39.66
	opt-6.7B	40.06	41.79	40.32	39.31
CW	$\mathcal{L}_{\text{VE}}$	41.02	43.58	41.50	39.95
	opt-2.7B	40.39	41.96	39.53	39.61
	opt-6.7B	39.82	41.81	39.89	39.66

Table 3: Results on uncompressed/compressed projectors using a surrogate VLP model trained from scratch.

4.5 Is Security Dependent on Architecture or the Quantity of Visual Tokens?

Our experimental results, as summarized in Tabs. 1 to 3, demonstrate that ***compressed projectors exhibit significant security vulnerabilities compared to their uncompressed counterparts***. However, it is important to note that uncompressed projectors generally possess a larger number of tokens (*e.g.*, the MLP employed in our laboratory utilizes 576 visual tokens, whereas the Q-former operates with only 32 visual tokens). This raises an essential question: does the reduced security of the Q-former stem from the inherent characteristics within its architecture, or merely from its smaller quantity of visual tokens?

To investigate this, we systematically apply mean pooling operations to LLaVA's original 576 (24x24) visual tokens. First, a 2x2 mean pooling operation is performed, reducing the number of visual tokens to 144. Subsequently, we further apply a 4x4 mean pooling operation, compressing the visual tokens down to 36.

We then conduct adversarial attacks on LLaVA-v1.5-7B with varying numbers of visual tokens. The results in Tab. 4 reveal that even when only 36 visual tokens are utilized (comparable to Q-former's), the attack performance on VLP does not surpass that on VE, implying its robustness against attacks. This finding conclusively demonstrates that the enhanced robustness of uncompressed projectors is ***not merely dependent on the quantity of visual tokens, but rather arises inherently from their structure design***. Furthermore, this implies that in pursuit of efficiency, one can opt for an uncompressed projector while employing techniques such as spatial pooling to minimize its visual markers. Even when their respective quantities of visual tokens are comparable in number, the security provided by an uncompressed projector remains markedly superior to that of its compressed ones.

Attack method	LVLM visual token number	VQA scores Clean	$\mathcal{L}_{\text{VE}}$	$\mathcal{L}_{\text{VLP}}$
PGD	576 (official)	77.31	69.30	71.21
	144	76.03	68.63	69.19
	36	73.09	66.72	67.30
CW	576 (official)	77.31	65.79	67.53
	144	76.03	63.06	65.25
	36	73.09	61.97	64.12

Table 4: Results of the attack on LLaVA with largely reduced visual tokens under white-box settings.

4.6 Further Strategies to Deteriorate the Security of Compressed Projectors

Table 3 presents the attack via training surrogate VLPs. We employ a single surrogate VLP for simplicity. However, multiple training runs can yield additional surrogate models, which we then leverage collectively to launch attacks against the target model, as illustrated in Fig. 1. The results summarized in Tab. 5 demonstrate that this multi-VLPs attack strategy further degrades the performance of the compressed projector.

Notably, increasing the number of surrogate VLPs does not significantly escalate the computational complexity of the attack. This is because all surrogate VLPs share a common visual encoder, and each Q-former component is parameter-efficient, containing approximately one-fourth the parameters of the VE. Consequently, for $N = 2$ surrogates, the total parameter count increases to roughly 1.2 times that of $N = 1$, while for $N = 3$,

it reaches approximately 1.4 times that of $N = 1$.

Attack method	K	The target models (InstructBLIP) FlanT5$_{XL}$	Vicuna-7B	FlanT5$_{XXL}$	Vicuna-13B	Avg.
baseline		72.43	75.53	71.86	68.22	71.51
PGD	$\mathcal{L}_{VE}$	43.51	44.94	43.57	42.02	43.51
	1	40.06	41.79	40.32	39.31	40.37
	2	39.45	40.57	39.15	38.64	39.45
	3	37.58	39.80	38.07	37.32	38.19
CW	$\mathcal{L}_{VE}$	41.02	43.58	41.50	39.95	41.51
	1	39.82	41.81	39.89	39.66	40.30
	2	38.08	39.41	38.49	37.66	38.41
	3	37.13	38.99	37.49	36.58	37.55

Table 5: Results of attacks using multi-VLPs.

We start by setting $\beta = 0$ to specifically examine attacks designed to target VLP, enabling us to analyze whether incorporating VLP improves attack effectiveness. We then combine the $\mathcal{L}_{VE}$ and $\mathcal{L}_{VLP}$ losses as Eq. (4). As shown in Tab. 6, $\mathcal{L}_{TCP}$ achieves superior attack performance compared to other approaches. Additionally, some adversarial images are provided in Sec. A.

Attack method	K	β	The target models (InstructBLIP) FlanT5$_{XL}$	Vicuna-7B	FlanT5$_{XXL}$	Vicuna-13B	Avg.
baseline			72.43	75.53	71.86	68.22	71.51
PGD	$\mathcal{L}_{VE}$	1	43.51	44.94	43.57	42.02	43.51
	2	0.4	37.87	38.89	38.68	37.85	38.32
		0.3	38.29	40.08	38.38	37.79	38.64
		0.2	36.91	39.71	37.46	36.70	37.70
		0.1	37.64	39.82	37.72	38.38	38.39
		0	39.45	40.57	39.15	38.64	39.45
CW	$\mathcal{L}_{VE}$	1	41.02	43.58	41.50	39.95	41.51
	2	0.4	37.22	37.91	36.80	36.29	37.06
		0.3	35.78	38.18	36.87	36.88	36.93
		0.2	36.85	38.64	36.32	36.82	37.16
		0.1	36.54	39.17	37.33	37.32	37.59
		0	38.08	39.41	38.49	37.66	38.41

Table 6: Results of attacks combined the $\mathcal{L}_{VE}$ and $\mathcal{L}_{VLP}$.

4.7 Results on Additional Datasets

Using InstructBLIP-based Vicuna-7B, we performed a thorough security evaluation to analyze potential vulnerabilities of compressed projectors across diverse datasets extending beyond VQA v2 in Tab. 7. Our assessment revealed that in a gray-box attack scenario, where only the target model's VE weights and VLP architecture are known, we were able to significantly degrade model performance, highlighting the inherent security risks associated with compressed projectors.

4.8 Ablation Study on Pre-Training Tasks

BLIP-2's first-stage pre-training consists of three tasks: Image-Text Matching (ITM), Image-Text Contrastive Learning (ITC), and Image-Grounded

Dataset	Clean	Adv.
ImageNet-1K accuracy	81.0	14.9
VizWiz test-dev scores	33.08	10.38
COCO CIDEr	140.6	11.1

Table 7: Attack InstructBLIP Vicuna-7B with surrogate VLP trained using opt-6.7B on other datasets and tasks.

Text Generation (Image Captioning, IC). By default, we adopted the standard BLIP-2 training configuration, employing all three tasks concurrently. However, our experimental results presented in Tab. 8 reveal an important insight: training the surrogate VLP exclusively with the IC task achieves attack performance comparable to utilizing all three tasks together. This not only reduces the overhead associated with training the surrogate model but also amplifies the security risks posed by the compressed projector.

Attack method	ITC	ITM	IC	The target models (InstructBLIP) FlanT5$_{XL}$	Vicuna-7B	FlanT5$_{XXL}$	Vicuna-13B	Avg.
baseline				72.43	75.53	71.86	68.22	71.51
PGD	✓			43.93	45.64	42.81	42.75	43.78
		✓		48.84	49.66	47.88	46.31	48.17
			✓	41.04	41.80	40.81	39.16	40.70
	✓	✓		40.46	42.74	41.20	40.70	41.28
	✓		✓	39.55	42.18	40.26	39.90	40.47
		✓	✓	41.32	42.49	40.57	39.65	41.01
	✓	✓	✓	40.06	41.79	40.32	39.31	40.37
CW	✓			41.55	44.11	40.75	41.32	41.93
		✓		45.17	47.21	45.49	44.62	45.62
			✓	40.64	41.81	39.78	38.63	40.22
	✓	✓		39.52	42.58	39.88	39.32	40.33
	✓		✓	39.47	41.84	40.28	38.74	40.08
		✓	✓	40.17	41.99	40.10	40.23	40.62
	✓	✓	✓	39.82	41.81	39.89	39.66	40.30

Table 8: Attack performance of the surrogate VLP with various pre-training task combinations. It suffices to employ IC task alone to achieve strong attacks.

4.9 Results on Various Attack Methods

In addition to PGD and CW in Tab. 3, we also implemented an additional attack method: MI-FGSM (Dong et al., 2018). Under gray-box settings, we conducted MI-FGSM attacks and extended the analysis presented in Tab. 3 of the original paper to incorporate methods beyond PGD and CW. The specific parameter configuration for MI-FGSM included setting the momentum value to 0.9. Our findings remain consistent with the conclusions in Secs. 4.3 and 4.4. Specifically, uncompressed projectors demonstrate superior robustness under gray-box attack settings, while attacks targeting compressed projectors result in further degradation of model performance.

Attack method	Surrogate MLP	Target models: LLaVA-v1.6 (uncompressed projector) Vicuna-7B	Mistral-7b	Vicuna-13B	Hermes-Yi-34B
Clean		78.32	79.10	75.61	79.07
MI-FGSM	$\mathcal{L}_{\text{VE}}$	71.65	68.66	72.85	72.01
	Vicuna-v1.3-7B	72.40	69.24	73.48	73.10
	Vicuna-v1.3-13B	72.90	69.99	73.58	73.49
Attack method	Surrogate Q-former	Target models: InstructBLIP (compressed projector) $\text{FlanT5}_{\text{XL}}$	Vicuna-7B	$\text{FlanT5}_{\text{XXL}}$	Vicuna-13B
Clean		72.43	75.53	71.86	68.22
MI-FGSM	$\mathcal{L}_{\text{VE}}$	49.99	52.41	49.40	47.06
	opt-6.7B	41.71	44.95	41.95	41.54

Table 9: Additional results of Tab. 3 on MI-FGSM.

5 Conclusion

This study investigates the security vulnerabilities of VLP structures within two representative LVLM frameworks. Our analysis exposes severe security susceptibilities in compressed projectors, while highlighting the robust performance of uncompressed alternatives. Through rigorous empirical evaluation, we demonstrate that this vulnerability originates inherently from the architectural design of the compressed projector itself and remains independent of the visual token quantity. These findings sound a cautionary note regarding the security implications for LVLMs employing compressed projectors, while encouraging researchers to adopt a more comprehensive understanding of VLP performance characteristics.

6 Discussion of Defense Mechanism

We briefly discuss two potential strategies for enhancing model robustness:

(1) **Optimizing uncompressed projector designs for improved efficiency**. While uncompressed projectors have demonstrated strong security properties, their inefficiency stems from a large number of visual tokens. However, our experiments in Tab. 4 reveal that not all visual tokens are essential for performance. By applying 2x2 pooling to LLaVA vision tokens, we successfully reduced the number of visual tokens to 25% of the original amount, resulting in only a marginal 1.3% drop in accuracy, and the LVLMs are still robust. This suggests significant potential for optimizing uncompressed projectors by eliminating redundant visual tokens while retaining their inherent security benefits, thereby enhancing the robustness.

(2) **Hybrid architectures combining compressed and uncompressed projectors**. Compressed projectors offer high efficiency and sufficient accuracy but lack robustness, whereas uncompressed projectors provide strong robustness and accuracy at the cost of efficiency. Building on insights from Tab. 4, pooling significantly reduced visual tokens, and the accuracy does not significantly decrease while maintaining safety, we propose a hybrid approach. For instance, we could pool the visual tokens from an uncompressed projector (originally 576 tokens) to 25% of their original count (144 tokens), and supplement this with 32 tokens from a compressed projector, resulting in a total of 144+32=176 tokens. This combination aims to achieve a balance of high accuracy, efficiency, and robustness. Unfortunately, due to challenges in training LVLMs with compressed projectors (lack of fully open-sourced training code of Q-former based LVLMs), we may face limitations in fully implementing this.

7 Limitations

We summarize the limitations of our study as follows: (1) The attack methodology has certain constraints. We implemented the attacks using only basic PGD, CW and MI-FGSM methods on VQA and image captioning tasks. Notably, we did not employ advanced techniques designed to enhance adversarial transferability, *e.g.*, DI-FGSM (Xie et al., 2019). Additionally, we limited our attacks to these two tasks without extending to other domains. Despite this, we successfully exposed the security vulnerabilities of compressed projectors, and incorporating DI and MI methods could potentially amplify the robustness risks associated with these models. (2) Our study focuses exclusively on analyzing the security vulnerabilities of both compressed and uncompressed projectors. However, we do not investigate potential defense mechanisms against the attacks. A discussion of possible defensive strategies would provide valuable insights into enhancing the robustness of compressed projectors against such attacks.

8 Acknowledgments

This work was supported by Young Elite Scientists Sponsorship Program by CAST (2023QNRC001), the National Key R&D Program of China (2023YFB4502200), the National Natural Science Foundation of China (No. 62376024, 62325405, 62104128, 62203257, 62031017, 62406159, U21B2031), Tsinghua University Initiative Scientific Research Program, Beijing National Research Center for Information Science, Technology (No. BNR2024TD03001) and Beijing Innovation Center for Future Chips.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and 1 others. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342.

Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei Koh, Daphne Ippolito, Florian Tramèr, and Ludwig Schmidt. 2023. Are aligned neural networks adversarially aligned? In *Advances in Neural Information Processing System (NeurIPS)*.

Nicholas Carlini and David Wagner. 2017. Towards Evaluating the Robustness of Neural Networks . In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2024. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13817–13827.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, and 1 others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna.lmsys.org*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, and 1 others. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *Computer Vision and Pattern Recognition Conference (CVPR)*, pages 9185–9193.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Computer Vision and Pattern Recognition Conference (CVPR)*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. *stat*, 1050:20.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, pages 19730–19742.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 34892–34916.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. 2023b. Jailbreaking chatgpt via prompt

engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*.

Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Computer Vision and Pattern Recognition Conference (CVPR)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. 2023. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. *arXiv preprint arXiv:2307.14539*.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. 2024. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *International Conference on Learning Representations (ICLR)*.

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, and 1 others. 2024. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xunguang Wang, Zhenlan Ji, Pingchuan Ma, Zongjie Li, and Shuai Wang. 2023. Instructta: Instruction-tuned targeted attack for large vision-language models. *arXiv preprint arXiv:2312.01886*.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. 2019. Improving transferability of adversarial examples with input diversity. In *Computer Vision and Pattern Recognition Conference (CVPR)*, pages 2730–2739.

Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. 2024. Deco: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*.

Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. 2023. VLATTACK: Multimodal adversarial attacks on vision-language tasks via pre-trained models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Yudong Zhang, Ruobing Xie, Jiansheng Chen, Xingwu Sun, Zhanhui Kang, and Yu Wang. 2025. Qava: Query-agnostic visual attack to large vision-language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10205–10218.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations (ICLR)*.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.