On Evaluating the Robustness of Large Vision-Language Models via Untargeted Modality Alignment Breaking Adversarial Attack

Zhichao Li[†,♯,§] Hongshan Yang[†,♯,§] Zhibo Wang[†,♯,*] Huiyu Xu[†,♯] Junhong Lai[♯]
Yaopeng Wang[≀,†] Kui Ren[†,♯] Chun Chen[†,♯]

[†]*State Key Laboratory of Blockchain and Data Security, Zhejiang University, P.R. China*

[♯]*College of Computer Science and Technology, Zhejiang University, P.R. China*

[≀]*The School of Cyber Science and Engineering, Southeast University, Nanjing, P. R. China*

{mugen, yanghs, zhibowang, huiyuxu, neuzjuljh}@zju.edu.cn, yaopengwang@seu.edu.cn, {kuiren, chunc}@zju.edu.cn

Abstract

Large Vision-Language Models (LVLMs) have achieved remarkable success in multimodal tasks by aligning the representation space of visual encoders to that of the Large Language Models. However, they remain vulnerable to transferable adversarial attacks, which can manipulate the LVLMs' output without accessing the model. Ensuring their reliable deployment thus requires a rigorous evaluation of black-box robustness. Current methods provide a limited assessment by perturbing only the visual encoder of LVLMs and often neglect untargeted attack scenarios. In this work, we propose the Modality Alignment Breaking Attack (MABA), a novel transferable, untargeted adversarial attack for evaluating the black-box robustness of LVLMs. MABA emphasizes disrupting the entire multimodal pipeline, targeting two key phases: visual encoding and modality alignment. First, MABA reveals that the core of transferable adversarial attacks lies in suppressing discriminative visual representations and explicitly uses this as an optimization objective to improve transferability across different LVLMs. Second, MABA introduces a mutual-information-aware projector that acts as a surrogate modality alignment module of LVLMs, effectively breaking cross-modal consistency and enhancing the transferability. Extensive evaluations demonstrate that MABA achieves state-of-the-art performance, leading to an average 58.37% drop in semantic metrics for the image caption task. Through ablation studies on diverse LVLM families, we derive valuable insights into strengthening the robustness of LVLMs.

1 Introduction

In recent years, Large Vision-Language Models (LVLMs) have achieved remarkable progress in multimodal understanding and reasoning, enriching the ways machines interact with humans [1–10]. This progress has drawn growing interest from real-world applications such as GUI agents [11], autonomous driving [12, 13], and embodied robotics [14]. However, LVLMs inherit adversarial vulnerabilities from the visual modality [15], which makes them susceptible to adversarial examples that induce incorrect outputs. These vulnerabilities may raise notable safety concerns. For example, adversarial examples reduce the route completion rate of LVLM-based autonomous vehicles by 61.11% [16]. A clear assessment of adversarial robustness is therefore essential to ensure dependable deployment in practical systems.

Several studies [17–20] have evaluated the adversarial robustness of LVLMs under white-box settings. These works optimize visual adversarial examples with access to model gradients, showing that imperceptible visual changes can fully disrupt LVLM's predictions. Since real-world attackers usually have no access to the parameters of target LVLMs, subsequent studies have focused on evaluating robustness under black-box settings [21–26]. In these approaches, a pretrained Vision Transformer (*e.g.*, CLIP [27] ViT) is adopted as a surrogate model to extract visual representations, as shown in Figure 1: (a). Adversarial examples are optimized by minimizing the distance between the embeddings of the clean image and those of the target image, which can either be assigned directly or generated from a target caption using generative models (*e.g.*, Stable Diffusion [28]).

Current black-box attacks often induce only limited shifts in LVLM's predictions. For instance, as shown in Figure 1: (a), the word "apple" in the reference caption is altered to "food", with minimal change to the semantics. Such limited effectiveness can be attributed to two factors. Firstly, they restrict the setting to targeted attacks, in which the choice of target inherently limits the attack strength and tightly constrains the optimization direction [17, 29]. Secondly, **existing black-box attacks largely overlook the working mechanism of LVLMs, which operates through an *align and inference* paradigm [30].** LVLMs work by aligning visual representations extracted by an encoder into an LLM-compatible space, which is carried out by a lightweight adapter. Recent work has

§ Zhichao Li and Hongshan Yang contributed equally to this work.
* Zhibo Wang is the corresponding author.

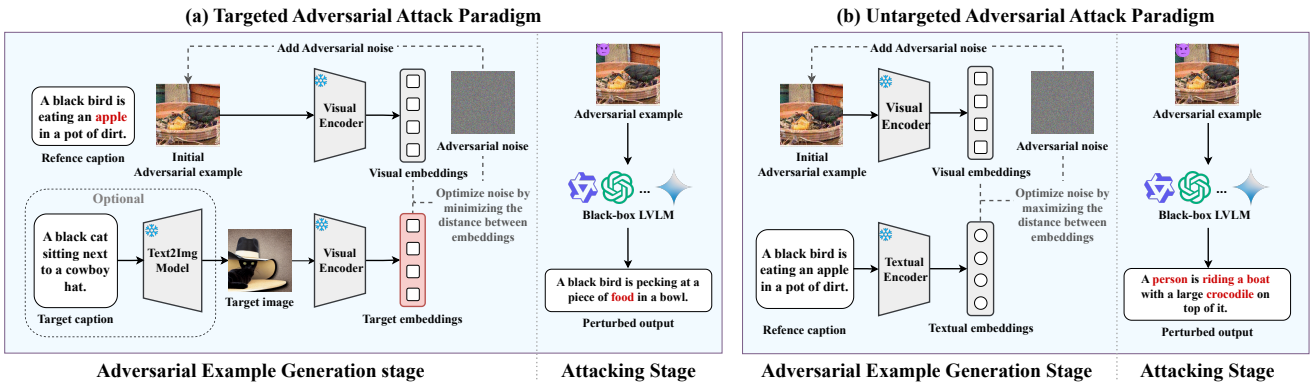

Figure 1: Adversarial attack paradigm on LVLM: Targeted vs. Untargeted

demonstrated the critical role and vulnerability of the adapter, showing that malicious fine-tuning of this alignment module can manipulate the model's visual understanding [31]. Despite its critical role, existing black-box attacks [21–26] often neglect the manipulation of the adapter, leaving the question of how to construct surrogate adapters for the black-box victim LVLMs still open. Meanwhile, advances in LVLM training paradigms [10] have substantially expanded multimodal capabilities beyond those of earlier models studied in prior robustness evaluations [17–19,22], introducing new behaviors that remain insufficiently examined. Together, these observations indicate that the adversarial robustness of LVLMs in black-box settings is far from fully understood.

In this work, we focus on evaluating the robustness of these state-of-the-art LVLMs [6–10] under untargeted black-box adversarial attacks. There exist several challenges in designing such attacks: ***(i) How to guide the optimization process for untargeted adversarial attacks on LVLMs, given the complex decision space?*** Compared to classification models, generative models exhibit a far more complex decision space [21]. A single clean image may correspond to a large number of "correct labels" in the textual representation space [32], which makes conventional untargeted attacks often ineffective [33]. This means that a successful untargeted attack on LVLMs must go beyond perturbing surface-level labels. Instead, it should disrupt the semantics of visual inputs, highlighting the need for new optimization objectives tailored to the semantic nature of LVLM outputs. ***(ii) How to build a surrogate adapter for the victim LVLM without any knowledge of the target model?*** To further explore the robustness of LVLMs, we treat the process of transforming visual representations into LLM-compatible embeddings as a new attack surface, which requires effectively simulating the adapter of a black-box victim LVLM. However, the internal representation space of the adapter is jointly shaped by its upstream visual encoder and downstream LLM. Consequently, adapters across different LVLMs often exhibit substantial discrepancies in their internal representations, making them challenging to simulate under black-box scenarios without direct access to the victim LVLMs.

To address the challenges above, we propose **Modality Alignment Breaking Attack (MABA)**, a novel adversarial method designed to disrupt modality alignment in LVLMs. For the first time, MABA introduces a method to construct a general surrogate adapter of LVLM under black-box conditions, thereby enhancing the effectiveness of attacks against LVLMs. We adopt the untargeted attack paradigm commonly used in Vision-Language Pretraining (VLP) attacks [33–35], as illustrated in Figure 1: (b), to leverage guidance from textual modality to disrupt the alignment between visual and textual modalities. MABA attacks both the visual encoding and modality alignment phases simultaneously. For the *visual encoding phase*, we introduce **discriminative feature suppression module** to guide untargeted attacks in disrupting the semantics of visual inputs. Specifically, we extract the high-frequency components carrying discriminative information from the visual representations of the surrogate visual encoder and apply singular value decomposition (SVD) to identify the most discriminative features. By suppressing these features, our method effectively deviates from the original semantics of visual inputs. For the *modality alignment phase*, we design a **mutual information-aware (MIA) projector**. The MIA projector is trained to maximize the mutual information between visual embeddings and their corresponding textual embeddings in the surrogate model, thereby acquiring the ability to capture the statistical dependencies between visual and textual representations. Leveraging this capability, the MIA projector simulates the LVLM adapter without accessing the victim model, enhancing the adversarial perturbations' effectiveness in disrupting modality alignment. We leverage MABA to conduct comprehensive ablation studies across multiple LVLM families, investigating how factors such as model scale, trainable modules, and model capability influence the adversarial robustness of LVLMs.

To summarize, our contributions are as follows:

- We find that transferable attacks are more effective when they suppress discriminative features in the victim LVLM's visual encoder. Amplifying this effect enables attacks to disrupt visual semantics and increase cross-model transferability.
- We propose Modality Alignment Breaking Attack (MABA), an untargeted black-box adversarial attack tailored for LVLMs. For the first time, MABA explores the construction of a surrogate adapter under black-box scenarios. Specifically, MABA proposes a MIA projector, grounded in information-theoretic principles, to capture statistical dependencies across modalities and thereby enable adversarial examples to disrupt modality alignment.
- Extensive experiments demonstrate the state-of-the-art attack performance of MABA, leading to an average 58.37% drop in semantic metrics for Image Captioning and a 31.63% drop in accuracy for VQA. Through systematic ablation studies, we dissect the effects of model scale, trainable modules, and model capability on adversarial robustness, offering actionable insights for developing more resilient LVLM systems.

2 Related Work

In this section, we first review the foundations and recent progress of LVLMs, and then summarize existing black-box adversarial attacks against LVLMs.

2.1 Large Vision Language Models

A typical LVLM consists of three core components [1]: a visual encoder, an adapter, and a large language model (LLM) [36, 37], corresponding to the three operational phases of the model: the Visual Encoding phase, the Modality Alignment phase, and the Multimodal Inference phase. Specifically, during the Visual Encoding phase, the visual encoder, commonly a Vision Transformer (ViT) pretrained via vision-language pretraining (VLP) models (*e.g.*, CLIP) [27, 38, 39], extracts visual representations from the visual input, producing embeddings that are pre-aligned with textual embeddings. In the Modality Alignment phase, these pre-aligned visual representations are processed by the adapter, which projects them into the representation space compatible with the LLM, explicitly aligning visual and textual modalities. Finally, in the Multimodal Inference phase, the aligned visual representations are integrated with textual representations within the LLM, enabling context-aware multimodal reasoning and generation.

Early LVLMs such as Flamingo [1] and BLIP-2 [2] adopt modular frameworks, where a frozen vision encoder (*e.g.*, CLIP) and a frozen pretrained LLM (*e.g.*, Vicuna [40]) are bridged via lightweight trainable adapters. This design requires updating only a small number of adapter parameters while transforming visual representations into a format understandable by the LLM, thereby enabling the LLM's visual understanding capabilities. Due to its efficiency and effectiveness, this paradigm has been widely adopted in subsequent work. Later, MiniGPT-4 [3], LLaVA [4], and InstructBLIP [5] incorporate visual instruction tuning into the training pipeline, further improving LVLMs' zero-shot performance on complex reasoning tasks. More recent models, such as LLaVA-OneVision [6], Phi-4 Multimodal [7], and Qwen-VL2.5 [8] progressively unfreeze all modules and introduce multi-image and video data during training, enabling LVLMs with stronger capabilities in handling multi-image and video inputs. Beyond vision-language modeling, MiniCPM-o [9], Qwen2.5-Omni [41], and Baichuan-Omni-1.5 [42] extend the LVLM paradigm by integrating additional encoders and decoders for other modalities, thereby supporting not only visual and textual inputs/outputs but also the audio modality. Most recently, InternVL3 [10] introduces a naive multimodal training paradigm in which tokens from different modalities are mixed for training throughout the entire training process, endowing the model with native multimodal capabilities.

The development of LVLM training has been remarkably rapid, with the central goal of enabling LLMs to better understand visual information. Prior studies have primarily evaluated the robustness of early LVLMs such as BLIP, MiniGPT-4, and InstructBLIP, while the robustness of more recent models with stronger capabilities and more sophisticated training strategies remains largely unexplored.

2.2 Adversarial Attacks on LVLMs

Research on the robustness of multimodal models originated from adversarial attacks against VLP models. Zhang *et al.* [43] first utilized textual supervision signals to craft adversarial examples, demonstrating that securing multimodal models necessitates joint consideration of both visual and textual modalities. Lu *et al.* [33] proposed leveraging multiple text prompts to optimize adversarial images, highlighting the complex many-to-many mapping relationships in multimodal feature spaces that distinguish them from unimodal models. Building on these insights, Gao *et al.* [34] introduced dynamic data augmentation techniques to further enhance attack effectiveness, while Jia *et al.* [35] leveraged semantic alignment to improve attack transferability across multimodal models. With the rise of LVLMs, research attention has progressively shifted toward investigating their adversarial robustness. Dong *et al.* [21] study the adversarial robustness of Google's Bard, revealing for the first time the vulnerabilities of commercial LVLMs to transferable adversarial attacks. Zhao *et al.* [22] propose a black-box attack to evaluate the robustness of open-source LVLMs. They leverage a text-to-image generator to convert a target caption into a target image, providing semantic guidance for the optimization, and then optimize adversarial examples by minimizing the distance

in the surrogate visual encoder's embedding space. This targeted attack paradigm has since been widely followed in subsequent work. Wang *et al.* [44] introduce an instruction-tuned targeted attack that leverages GPT-4 to infer reasonable instructions corresponding to target responses, and employs a surrogate model to extract instruction-aware visual representations for optimizing adversarial examples, thereby enhancing their transferability. Guo *et al.* [23] present AdvDiffVLM, which injects perturbations carrying target semantics into adversarial examples using a diffusion model. While improving transferability, this approach inevitably degrades image quality. Zhang *et al.* [24] develop AnyAttack, which achieves flexible targeted adversarial attacks by training a perturbation generator with a self-supervised pre-training and fine-tuning framework, establishing a practical black-box attack tool. Li *et al.* [25] propose aligning transformed adversarial examples with target images in a fine-grained, patch-level manner. They introduce locally aggregated perturbations on semantically rich regions and adopt ensemble settings, which together substantially enhance attack transferability. Building on this, Jia *et al.* [26] incorporate clustering techniques to enhance local alignment and employ a dynamic ensemble strategy, achieving state-of-the-art adversarial transferability. Despite these advances, there is no work that has comprehensively assessed the black-box robustness of LVLMs through untargeted attacks. In addition, these efforts mainly disrupt the visual encoding phase of LVLMs, while leaving the vulnerability of the modality alignment phase in LVLMs unexplored.

3 Preliminary

In this section, we first introduce the concept of Mutual Information Neural Estimation, which serves as the theoretical foundation of our MIA projector. We then provide a formal definition of LVLMs and outline the threat model considered in our study.

3.1 Mutual Information Neural Estimation

In information theory, mutual information (MI) quantifies the shared information between two random variables, capturing non-linear statistical dependencies and serving as a measure of true dependence [45]. A higher MI indicates stronger statistical dependence, meaning the values of one variable provide significant clues about the values of the other.

Given two continuous random variables $\mathbf{a} \in \mathcal{A}$ and $\mathbf{b} \in \mathcal{B}$, the mutual information $I(\mathbf{a},\mathbf{b})$ is defined as:

$$I(\mathbf{a},\mathbf{b}) := \int_{\mathcal{A}}\int_{\mathcal{B}} \mu_{\mathbf{ab}}(\mathbf{a},\mathbf{b}) \log\left(\frac{\mu_{\mathbf{ab}}(\mathbf{a},\mathbf{b})}{\mu_{\mathbf{a}}(\mathbf{a})\mu_{\mathbf{b}}(\mathbf{b})}\right) d\mathbf{a}\, d\mathbf{b}, \quad (1)$$

where $\mu_{\mathbf{ab}}$ is the joint probability density of $\mathbf{a}$ and $\mathbf{b}$, and $\mu_{\mathbf{a}}$ and $\mu_{\mathbf{b}}$ are their respective marginal densities.

Direct computation of $I(\mathbf{a},\mathbf{b})$ is often intractable, particularly for high-dimensional variables. To enable tractable estimation, Belghazi *et al.* [46] propose the Donsker–Varadhan (DV) variational representation, which reformulates mutual information as:

$$I(\mathbf{a},\mathbf{b}) := \sup_{m\in\mathcal{M}} \left[\mathbb{E}_{\mu_{\mathbf{ab}}}(m(\mathbf{a},\mathbf{b})) - \log\exp\left(\mathbb{E}_{\mu_{\mathbf{a}}\otimes\mu_{\mathbf{b}}}(e^{m(\mathbf{a},\tilde{\mathbf{b}})})\right)\right], \quad (2)$$

where $\mathcal{M}$ denotes the space of measurable scoring functions $m : \mathcal{A}\times\mathcal{B}\to\mathbb{R}$. Here, $\mathbb{E}_{\mu_{\mathbf{ab}}}$ represents the expectation under the joint distribution $\mu_{\mathbf{ab}}$, and $\mathbb{E}_{\mu_{\mathbf{a}}\otimes\mu_{\mathbf{b}}}$ represents the expectation under the product of marginals $\mu_{\mathbf{a}}\otimes\mu_{\mathbf{b}}$, constructed by independently sampling $\mathbf{a}$ and $\tilde{\mathbf{b}}$ from their respective marginal distributions. By optimizing the scoring function $m(\cdot,\cdot)$, this variational formulation provides a tractable estimate of mutual information, making it suitable for high-dimensional variables and complex tasks.

3.2 LVLMs

Let $\mathcal{D} = \{(\mathbf{x}, y)\}$ denote a dataset of paired image-text samples, where $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$ is an image with C channels, height H, and width W, and $y \in \mathcal{T}$ is its corresponding ground-truth text (*e.g.*, a caption), with $\mathcal{T}$ representing the space of valid text sequences. An LVLM $V(\cdot)$ takes the image $\mathbf{x}$ and a query prompt $q \in \mathcal{T}$ as input, and generates a predicted textual output $y^{\text{pred}} \in \mathcal{T}$ within a shared semantic space.

Formally, LVLM $V(\cdot)$ matches the visual input $\mathbf{x}$ with the ground-truth text y (*e.g.*, caption of $\mathbf{x}$) under the guidance of the prompt q. The match is considered successful if the predicted output y^{pred} is semantically consistent with the reference y. This alignment condition is defined as:

$$V(\mathbf{x}, q) = y^{pred} \sim y, \quad (3)$$

where $\sim$ indicates semantic consistency between the model's output y^{pred} and the ground-truth target y.

3.3 Threat Model

Attack Scenario. We study *transfer-based black-box untargeted adversarial attacks* against LVLMs. In this setting, the attacker crafts adversarial examples using a locally accessible surrogate model $S(\cdot)$ (*e.g.*, open-source CLIP) and transfers them to a target black-box LVLM $V(\cdot)$ without requiring access to its architecture or parameters. These adversarial examples can subsequently compromise the functionality of the target system in various scenarios, such as disrupting the behavior of web agents in digital form [47] or perturbing real-world detection systems in physical form [48].

Attacker Objective. The attacker seeks to disrupt the inference process of the target LVLM $V(\cdot)$. Specifically, the attacker aims to craft an adversarial image $\mathbf{x}^{\text{adv}}$ from a benign

image $\mathbf{x}$, such that the target model V produces an output y^{pred} that is no longer semantically consistent with the ground-truth text y corresponding to $\mathbf{x}$:

$$V(\mathbf{x}^{\text{adv}}, q) = y^{pred} \not\sim y. \tag{4}$$

The attack is successful if the adversarial image $\mathbf{x}^{\text{adv}}$ produces a description y^{pred} that is semantically inconsistent with the content of the benign image $\mathbf{x}$.

Attacker Capabilities. The target model V is a black box for the attacker. The attacker has no knowledge of its architecture and parameters. For stealth considerations, we do not access the target LVLM during the attack.

The attacker has white-box access to a local surrogate model $S = (f, g)$, where f is a visual encoder that maps an image $\mathbf{x}$ to a visual embedding $f(\mathbf{x}) \in \mathbb{R}^{d_v}$, and g is a textual encoder that maps a text sequence y to textual embedding $g(y) \in \mathbb{R}^{d_y}$. Here, d_v and d_y denote the dimensionalities of the visual and textual embeddings, respectively.

Basic Adversarial Optimization Formulation. Given a benign image $\mathbf{x}$, the adversarial example is generated by adding a perturbation δ subject to the ℓ_∞-norm constraint:

$$\mathbf{x}^{\text{adv}} = \mathbf{x} + \delta, \quad s.t. \, \|\delta\|_\infty \leq \varepsilon. \tag{5}$$

The basic optimization objective is to maximize the distance between the visual embedding of the adversarial image $\mathbf{x}^{\text{adv}}$ and the textual embedding of its reference caption y. We use cosine similarity as the distance metric. The objective can be formulated as:

$$\mathcal{L}_{\text{adv}} = -\cos\Big(f(\mathbf{x}^{\text{adv}}), g(y)\Big). \tag{6}$$

4 MABA: Modality Alignment Breaking Attack

In this section, we formally introduce the Modality Alignment Breaking Attack (MABA) framework. In Figure 2, MABA consists of two core modules: The Discriminative Feature Suppression (DFS) module, which suppresses discriminative visual representations in the surrogate visual encoder to improve transferability across victim visual encoders; The Mutual-Information-Aware (MIA) Projector Training Module, which trains an MIA projector capturing statistical dependencies between visual and textual embeddings in the surrogate model, serves as a surrogate adapter to strengthen attack effectiveness without any knowledge of the victim LVLM.

4.1 Design Overview

DFS Module. In the visual encoding phase, different LVLMs adopt distinct ViTs to extract visual representations. Variations in architectures and parameters across ViTs constitute a major barrier to the transferability of adversarial examples crafted on a single surrogate model. To discover the mechanism behind adversarial transferability, we build upon the spectral analysis of ViTs proposed by Wang *et al.* [49], which reveals the "over-smoothing" effect: as depth increases, ViTs behave like low-pass filters, filtering high-frequency components (HCs) that encode discriminative features while retaining only direct components (DCs) that capture global layouts. Extending this perspective to adversarial examples, we investigate how discriminative HCs change under adversarial perturbations. To quantify these changes, we employ SVD on HCs, a well-established tool for analyzing high-dimensional representations [50]. In this context, the magnitude of each singular value indicates the relative contribution of the corresponding direction to the overall variation within HCs [51], allowing us to assess the importance of individual feature directions. Our analysis uncovers a consistent trend: **black-box adversarial attacks tend to suppress the top-k singular values in the diagonal matrix obtained via SVD of HCs, redistributing energy toward subsequent singular values** (see Section 5.4). Motivated by this observation, the DFS module reinforces this effect in a cascading manner [52], by explicitly suppressing the top-k singular values of HCs across multiple intermediate representations during adversarial optimization, as shown in Figure 2: (b). This operation weakens discriminative cues of the original image while amplifying adversarial signals in the visual representation, thereby enhancing attack strength.

MIA Projector Training Module. In the modality alignment phase, the core challenge lies in efficiently simulating the adapter of black-box LVLMs, for which prior work offers no guidance. Under the black-box setting, the modules of the target LVLM are inaccessible. As a result, direct training of a surrogate adapter becomes infeasible. To overcome this challenge, we draw inspiration from contrastive multimodal learning. CLIP [27] demonstrates that maximizing mutual information between visual and textual embeddings via the InfoNCE loss [53] yields strong modality alignment and broad multimodal capabilities, highlighting the potential of mutual information theory in modality alignment. Motivated by this establishment, MIA Projector Training Module optimizes the MIA projector that maps surrogate visual and textual embeddings into a shared representation space where their mutual information is maximized, thereby capturing more modality-aligned features and ensuring modality alignment, as shown in Figure 2: (a). By incorporating the MIA projector into **Adversarial Example Generation Phase**, we simulate the adapter of black-box LVLMs in a statistically grounded manner, enabling adversarial perturbations to effectively disrupt modality alignment.

4.2 Discriminative Feature Suppression

Given a visual encoder $f(\cdot)$ and an input image $\mathbf{x}$, we denote the representations from the l-th intermediate layer of $f(\cdot)$ as

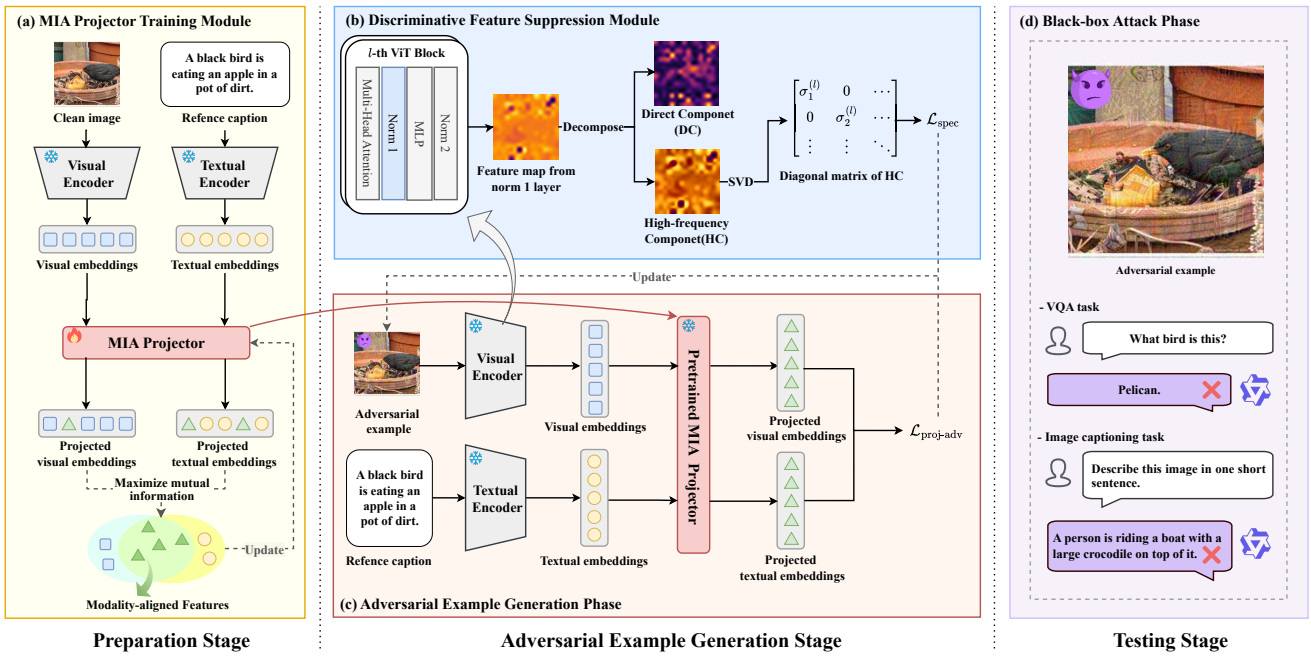

Figure 2: Overall framework of our proposed MABA. In the preparation stage, we first train an MIA projector on the target dataset using the mSMI framework. In the adversarial example generation stage, the MIA projector is integrated into the adversarial example generation phase, ensuring that the optimization occurs within a modality-aligned representation space. Simultaneously, a discriminative feature suppression module is introduced as a regularization term to guide adversarial optimization. Finally, in the attacking stage, the crafted adversarial examples are fed into the black-box LVLM, producing perturbed outputs.

$z^l = f(\mathbf{x})^l$, where $z^l \in \mathbb{R}^{n \times d_v}$. Here, n denotes the number of tokens, each corresponding to a spatial patch in ViT, and d_v is the visual embedding dimension. According to the modeling by Wang *et al.* [49], the self-attention mechanism in ViTs inherently acts as a low-pass filter, decomposing the representation of the input image $\mathbf{x}$ at each layer into DCs and HCs, as formally defined below:

$$\mathcal{DC}[z^l] = \frac{1}{n}\mathbf{1}\mathbf{1}^\top z^l, \tag{7}$$

$$\mathcal{HC}[z^l] = z^l - \mathcal{DC}[z^l], \tag{8}$$

where $\mathbf{1} \in \mathbb{R}^n$ is an all-ones column vector, and n denotes the number of tokens in the ViT architecture.

The $\mathcal{DC}[z^l]$ captures globally shared low-frequency information by averaging representations across all tokens, thus reflecting the spatially invariant structure in the image. In contrast, $\mathcal{HC}[z^l]$ retains locally specific high-frequency information, such as localized textures, that differentiates individual tokens, and serves as the primary carrier of discriminative features essential for downstream tasks.

To quantify the discriminative features encoded in intermediate representations during adversarial optimization, we perform SVD on $\mathcal{HC}[z^l]$:

$$\mathcal{HC}[z^l] = U^l \Sigma^l (V^l)^\top, \tag{9}$$

where $U^l \in \mathbb{R}^{n \times r}$ contains the left singular vectors, each representing a principal spatial pattern across the n tokens in $\mathcal{HC}[z^l]$. The matrix $V^l \in \mathbb{R}^{d_v \times r}$ contains the right singular vectors, each capturing a principal direction in the d_v-dimensional embedding space of the $\mathcal{HC}[z^l]$. The diagonal matrix $\Sigma^l = \text{diag}(\sigma_1^{(l)}, \ldots, \sigma_r^{(l)}) \in \mathbb{R}^{r \times r}$ holds the singular values $\sigma_i^{(l)}$, which quantify the strength of variation along each pair of spatial and channel directions. Here, $r = \text{rank}(\mathcal{HC}[z^l])$ represents the rank of $\mathcal{HC}[z^l]$.

Since Σ^l compactly summarizes the overall variation in $\mathcal{HC}[z^l]$ across both spatial and embedding dimensions, we focus our analysis on this spectrum. The magnitude of each $\sigma_i^{(l)}$ reflects the relative contribution of the corresponding direction to the overall variation within $\mathcal{HC}[z^l]$ [51], and thus provides a quantitative characterization of the discriminative features in the intermediate representation.

Based on this quantification, we design our method to explicitly suppress the top-k singular values Σ^l across multiple intermediate layers during adversarial optimization. By minimizing the top-k largest singular values, we directly reduce the influence of the most discriminative features present in the original image. This encourages the adversarial perturbation to dominate the visual representations, allowing adversarial semantics to become the primary factors that guide prediction. Formally, we define the *Discriminative Feature Suppression*

Loss as

$$\mathcal{L}_{\text{spec}} = \frac{1}{|L|} \sum_{l \in L} \sum_{i=1}^{k} \left(\sigma_i^{(l)} \right)^{\beta}, \tag{10}$$

where L denotes the set of selected intermediate layers, k determines the number of top singular values subject to regularization, and β modulates the overall strength of the spectral suppression term.

4.3 Mutual Information-aware Projector

Inspired by the success of mutual information-based optimization in CLIP, we propose a mutual information-aware (MIA) projector to capture modality alignment in the surrogate model and construct a highly modality-aligned representation space for adversarial optimization. However, estimating mutual information for high-dimensional variables, such as visual embeddings, is computationally expensive and often produces inaccurate measures [46]. To address this challenge, we adopt the max-sliced mutual information (mSMI) framework [54], which learns orthonormal slicing matrices that project visual and textual embeddings into a shared subspace where their mutual information is maximized. By maximizing mutual information over these lower-dimensional projections, this approach effectively approximates the mutual information of the original high-dimensional embeddings, enabling efficient estimation. Furthermore, this dimensionality-reduction projection naturally supports mapping dense visual representations to sparse textual representations, consistent with the goal of simulating an LVLM adapter.

Given image-text pairs $(\mathbf{x}, y)$ sampled from a multimodal dataset $\mathcal{D}$, we extract the visual embedding $f(\mathbf{x}) \in \mathbb{R}^{d_v}$ and textual embedding $g(y) \in \mathbb{R}^{d_t}$. Then, both embeddings are projected into a common d_k-dimensional subspace using learnable slicing matrices $A \in \mathrm{S}(d_k, d_v)$ and $B \in \mathrm{S}(d_k, d_t)$, where $\mathrm{S}(\cdot,\cdot)$ denotes the Stiefel manifold of matrices with orthonormal columns. The projection dimension d_k is chosen such that $1 \le d_k \le \min(d_v, d_t)$, allowing a trade-off between the upper bound of the estimated mutual information and computational cost. The projected embeddings are defined as:

$$pr_A(f(\mathbf{x})) = A^\top f(\mathbf{x}), s.t. A \in \mathrm{S}(d_k, d_v), \tag{11}$$

$$pr_B(g(y)) = B^\top g(y), s.t. B \in \mathrm{S}(d_k, d_t). \tag{12}$$

The d_k-dimensional max-sliced mutual information between $f(\mathbf{x})$ and $g(y)$ is defined as the supremum of mutual information between their common d_k-dimensional projections [54]:

$$\mathrm{SI}_{\mathrm{k}}(f(\mathbf{x}), g(y)) := \sup I\left(pr_A(f(\mathbf{x})), pr_B(g(y))\right), \tag{13}$$

where $I(\cdot,\cdot)$ denotes the mutual information between the projected embeddings.

In practice, we estimate SI_{k} following the Donsker–Varadhan (DV) formulation defined in **??**. We restrict the function class $\mathcal{M}$ to a parameterized family of neural networks, denoted $\mathcal{M}_{\text{nn}}$, and approximate the expectations using finite samples. This yields the neural estimator for mSMI:

$$\widehat{\mathrm{SI}}_k := \sup_{m \in \mathcal{M}_{\text{nn}}} \left[\frac{1}{n} \sum_{i=1}^{n} m\left(pr_A(f(\mathbf{x}_i)), pr_B(g(y_i))\right) - \log\left(e^{\frac{1}{n}\sum_{i=1}^{n} m(pr_A(f(\mathbf{x}_i)), pr_B(g(\tilde{y}_i)))} \right) \right], \tag{14}$$

where $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \subset \mathcal{D}$ are positive pairs and $\tilde{y}_i$ is sampled independently from the marginal distribution over texts. By maximizing the Equation (14), the scoring function $m \in \mathcal{M}_{\text{nn}}$ is jointly optimized along with the projection matrices A and B. This joint optimization identifies a shared representation subspace. In this subspace, the mutual information between the projected embeddings $pr_A(f(\mathbf{x}))$ and $pr_B(g(y))$ is maximized under the learned scoring measure m. To promote modality alignment and simplify the optimization process, we set $A = B$ during optimization. The training process yields a single slicing matrix, which we denote as the MIA projector $A \in \mathrm{S}(d_k, d)$, where $d = d_v = d_y$. This equivalence holds because the visual and textual embeddings produced by our surrogate model share the same dimensionality.

4.4 Adversarial Example Generation

After training the MIA projector, the adversarial optimization objective Equation (6) can be redefined by applying the MIA projector A to both visual and textual embeddings, resulting in the projected adversarial loss:

$$\mathcal{L}_{\text{proj-adv}} = \max \; -\cos\left(pr_A(f(\mathbf{x}+\delta)), pr_A(g(y))\right), s.t. \quad \|\delta\|_\infty \le \varepsilon. \tag{15}$$

where y remains the ground-truth caption corresponding to the benign image $\mathbf{x}$. The projected objective $\mathcal{L}_{\text{proj-adv}}$ steers the adversarial optimization process into a highly modality-aligned representation space, thereby enabling adversarial examples to disrupt the alignment between visual modality and textual modality in a statistically grounded manner.

Finally, to enhance the transferability of adversarial examples across diverse LVLMs, we jointly optimize two complementary objectives: the *Discriminative Feature Suppression Loss* $\mathcal{L}_{\text{spec}}$ and the *Projected Adversarial Loss* $\mathcal{L}_{\text{proj-adv}}$, the latter defined via the pretrained MIA projector A. The final adversarial objective is formulated as a weighted sum:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{proj-adv}} + \lambda \cdot \mathcal{L}_{\text{spec}}, \tag{16}$$

where λ balances the regularization strength. This adversarial objective ensures that the perturbations are both transferable across model architectures and effective at disrupting modality alignment, thereby improving attack transferability and strength against black-box LVLMs.

5 Experiment

To evaluate the transferability and effectiveness of our proposed MABA, we conduct experiments across multiple LVLM families on the image captioning task and the vision question answering task. The experimental setup is detailed in Section 5.1. To better interpret the empirical results, we organize our evaluation around the following key research questions:

- **RQ1:** Can advanced LVLMs resist adversarial examples generated by MABA in a black-box setting? (Section 5.2)
- **RQ2:** What model-intrinsic factors influence the robustness of LVLMs under adversarial attacks? (Section 5.3)
- **RQ3:** What drives MABA's effectiveness, and how sensitive is it to key hyperparameters? (Section 5.4)

5.1 Experimental Setup

Tasks and Datasets. We evaluate our method on two standard tasks for multimodal understanding and reasoning. Image captioning (IC) focuses on generating natural language descriptions from visual inputs, reflecting the model's comprehensive understanding of visual content. Visual question answering (VQA) requires answering image-grounded questions, assessing the model's capability for multimodal reasoning over visual information. We adopt widely used benchmark datasets corresponding to each task.

We randomly select 1,000 images from the Karpathy test split of the MSCOCO dataset [55], which correspond to 5,162 questions (65 question types and 3 answer types) in the VQAv2 dataset [56] and their associated reference captions in the MSCOCO Captions dataset [57]. These images are used to generate adversarial examples and evaluate model performance on both IC and VQA tasks.

Evaluated Models. We evaluate MABA across diverse LVLM families, including models widely used in prior studies [19, 22], such as the BLIP-2 family [2] and MiniGPT4 [3], as well as increasingly studied models, such as the LLaVA-OneVision family [6], Phi4-Multimodal [7], and QwenVL2.5 family [8]. We further include the omni-architecture MiniCPM-o-2.6 [9] and the native multimodal InternVL3 family [10] for comprehensive evaluation. All evaluated LVLMs avoid using the same ViT backbone as the surrogate model, ensuring that transferability does not benefit from shared visual encoders.

Implementation Details. Following the configurations in previous works [21, 24], the adversarial examples are generated under the ℓ_∞ norm constraint with a perturbation bound of $\varepsilon = 16/255$. For all attacks requiring iterative optimization, we set the number of steps to 30 and the step size to $3/255$. Regarding surrogate model selection, we use CLIP_ViT-L/14 as the surrogate encoder, of which both the visual and textual embedding dimensionalities are $d_v = d_t = 768$.

For the DFS module, we select layers 12-17 in CLIP_ViT-L/14 for spectral regularization. We set the hyperparameter of regularization strength k=10, and $\beta = 2.5$.

For the MIA Projector Training Module, we set the projected embedding dimension to $d_k = 676$, corresponding to 87.5% of the original embedding space. We utilize the Karpathy test split of the MSCOCO dataset for training, paired with the corresponding ground-truth captions from the MSCOCO Captions dataset. To instantiate the measurable scoring function, we employ a Q-Former [58] and adopt the cosine similarity function as the scoring function. Finally, the MIA projector is trained using the Adam optimizer with an initial learning rate of 2×10^{-4} for 100 epochs. The learning rate decays by a factor of 0.1 every 40 epochs. The balance parameter for the loss strength λ is set to 0.36. To better explore the robustness boundaries, our implementation adopts the variables augmentation strategy from DRA [34] to enhance attack stability. This component is orthogonal to our main contributions, which focus on modality alignment breaking attacks. Additional results in Appendix A.2 demonstrate that MABA significantly outperforms DRA.

Baselines. We compare MABA with several representative transfer-based black-box LVLM adversarial attack methods, including *AttackVLM* [22], *AnyAttack* [24], and *FOA* [26]. For *AttackVLM*, we use CLIP_ViT-L/14 as the surrogate model and select its most influential image-to-image (AttackVLM-ii) version. For *AnyAttack*, we follow its best-performing setup, adopting the released decoder that is jointly trained over an ensemble of surrogate encoders (CLIP_ViT-B/32, B/16, L/14) for generating adversarial examples. For *FOA*, we use the same ensemble surrogate encoders as used in *AnyAttack*, to ensure a fair comparison. We consistently use a single CLIP_ViT-L14 as a surrogate model for MABA when comparing with ensemble baselines, as it is sufficient to generate effective attacks.

Evaluation Metric for IC. To assess the linguistic quality of generated captions, we adopt conventional metrics, including **BLEU-1**, **BLEU-4**, **METEOR**, **ROUGE-L**, **CIDEr**, and **SPICE**. Besides, we adopt **GPTScore** [59] to better evaluate semantic similarity. Our GPTScore is computed using the GPT-4o API between reference captions and generated captions with an evaluation prompt (See Appendix B for details). The score ranges from 0 to 1, where 0 indicates semantic irrelevance and 1 indicates perfect semantic consistency. Furthermore, we further stratify GPTScore into four intervals and define a set of fine-grained metrics called **Semantic Deviation (SD)**. Each metric is defined as the percentage of samples whose GPTScore falls within the corresponding interval: **SD@1** ($[0.75, 1.00]$) indicates strong semantic preservation, **SD@2** ($[0.50, 0.75)$) indicates partial degradation or vagueness, **SD@3** ($[0.25, 0.50)$) indicates semantic drift or broken meaning, **SD@4** ($[0.00, 0.25)$) indicates severe semantic distortion.

Evaluation Metric for VQA. We adopt the official **VQA**

Table 1: Evaluation metrics of the image captioning task on evaluated LVLMs under different attack methods. Lower values indicate worse lexico-syntactic and semantic consistency between generated captions and reference captions.

Source Model	Attack Method	SPICE	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	GPTScore
BLIP2-OPT6.7B-MSCOCO	Clean	25.57	81.93	42.34	31.57	61.31	150.81	77.26
	AttackVLM-ii [22]	23.16	79.09	38.02	29.88	58.37	135.22	73.55
	AnyAttack [24]	22.84	78.70	38.44	29.48	58.37	134.84	71.34
	FOA [26]	21.93	77.08	35.74	28.84	57.58	125.61	72.04
	MABA	**17.64**	**70.25**	**29.39**	**24.78**	**52.04**	**99.98**	**56.14**
MiniGPT4-Vicuna13B	Clean	17.39	52.29	16.43	23.14	39.66	44.88	63.65
	AttackVLM-ii [22]	15.25	46.31	13.87	21.25	37.95	40.72	46.60
	AnyAttack [24]	15.78	47.66	14.10	21.69	38.92	42.93	50.59
	FOA [26]	12.96	44.83	11.77	19.78	36.61	32.68	46.26
	MABA	**7.42**	**35.01**	**6.83**	**14.00**	**28.87**	**15.14**	**16.39**
LLaVA-OneVision-7B	Clean	25.11	76.97	35.27	31.25	58.90	132.19	81.97
	AttackVLM-ii [22]	19.86	63.93	22.76	26.53	49.80	82.44	68.45
	AnyAttack [24]	19.05	62.40	22.04	25.51	48.76	80.77	66.23
	FOA [26]	16.74	58.81	19.07	24.01	46.72	67.04	57.52
	MABA	**12.14**	**51.44**	**13.78**	**19.85**	**41.06**	**44.90**	**43.02**
Phi4-Multimodal	Clean	23.40	68.88	27.77	28.50	53.69	105.21	73.94
	AttackVLM-ii [22]	16.64	62.04	20.84	23.03	47.28	73.94	55.69
	AnyAttack [24]	15.21	58.68	18.54	21.58	45.22	67.57	50.31
	FOA [26]	14.40	59.49	17.70	21.59	45.41	62.60	47.07
	MABA	**9.04**	**48.48**	**11.48**	**16.37**	**38.26**	**34.05**	**26.69**
QwenVL2.5-7B	Clean	20.44	53.81	14.83	26.47	44.83	46.48	80.47
	AttackVLM-ii [22]	15.22	48.45	11.39	22.28	40.46	35.57	62.26
	AnyAttack [24]	13.69	45.12	10.23	26.03	38.26	30.81	55.47
	FOA [26]	12.57	43.79	9.21	20.01	37.18	25.67	47.20
	MABA	**7.90**	**36.27**	**5.99**	**15.30**	**31.84**	**13.12**	**26.88**
MiniCPM-o-2.6	Clean	21.77	54.05	15.70	26.97	45.31	47.61	79.56
	AttackVLM-ii [22]	15.84	53.69	14.93	22.35	43.40	57.81	61.29
	AnyAttack [24]	15.13	51.59	13.30	21.21	41.30	52.74	56.41
	FOA [26]	12.90	51.42	12.37	19.75	40.05	45.25	43.81
	MABA	**9.00**	**43.49**	**8.61**	**16.07**	**34.55**	**27.33**	**29.85**
InternVL3-8B	Clean	21.76	57.25	17.17	27.04	47.31	61.89	79.82
	AttackVLM-ii [22]	16.31	52.70	13.91	23.28	43.43	52.90	68.85
	AnyAttack [24]	15.70	50.65	12.66	22.13	41.43	47.92	62.07
	FOA [26]	11.81	43.44	8.33	18.89	36.62	25.53	43.01
	MABA	**7.39**	**35.81**	**5.49**	**14.26**	**30.62**	**15.63**	**27.17**

accuracy metric [60], which accounts for variability among human responses. Given a predicted answer a and the 10 human answers $\{h_i\}_{i=1}^{10}$, the accuracy is defined as

$$\mathrm{Acc}(a) = \min\left(\frac{1}{3}\sum_{i=1}^{10}\delta(a,h_i),1\right), \tag{17}$$

where $\delta(a,h_i)$ equals 1 if $a = h_i$ and 0 otherwise.

5.2 Transferability across LVLMs

In this section, we evaluate the transferability of MABA on the IC task and the VQA task. Extensive experiments show that MABA notably degrades the performance of state-of-the-art LVLMs on both tasks.

5.2.1 Results on IC Tasks

On the IC tasks, MABA consistently outperforms baseline attack methods across all evaluated LVLMs, simultaneously disrupting both lexical-level consistency and semantic consistency between generated captions and reference captions.

As shown in Table 1, MABA induces the most significant performance degradation across all metrics and evaluated models. Even on the most robust model, BLIP2-OPT6.7B-MSCOCO, MABA reduces the CIDEr score from 150.81 to 99.98, achieving a larger drop than the state-of-the-art method FOA, which scores 125.64. On the GPTScore metric, MABA similarly induces a substantial change, demonstrating effective compromise of the visual semantics. For example, on QwenVL2.5-7B, the GPTScore drops from 80.47 to 26.88.

We employ the SD@k metric for a fine-grained analysis of semantic degradation, with results illustrated in Figure 4. Given that over 95% of clean images fall within SD@1 and SD@2, we establish SD@2 as the threshold for an acceptable semantic boundary. While AttackVLM and AnyAttack largely preserve this benign pattern, with only a few adversarial examples assigned to severely incorrect categories (SD@3 and SD@4), the state-of-the-art FOA induces noticeable semantic corruption. MABA, however, exhibits strong semantic disruption across all models, yielding the highest proportion in the most severely corrupted category SD@4. On average, its cumulative proportion of SD@3 and SD@4 reaches 55.30%, significantly surpassing 36.87% achieved by FOA. Detailed results are provided in Appendix C.

As illustrated in Figure 3, under baseline methods, the generated captions exhibit slight deviations, while largely preserving the original semantic meaning. In contrast, MABA induces complete semantically incorrect descriptions that fun-

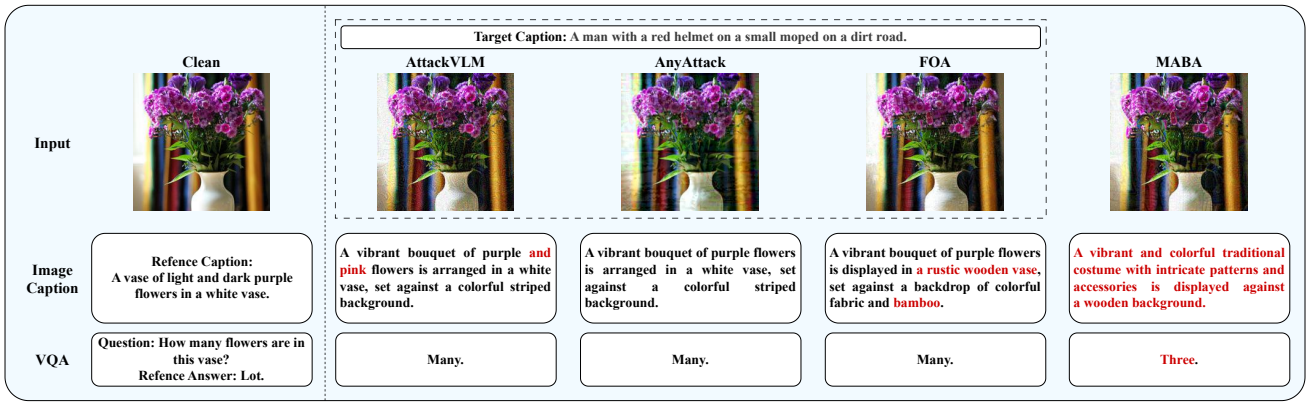

Figure 3: Comparison of corrupted captions and answers generated under different attack methods.

Table 2: Average VQA accuracy (%) across LVLMs under different attack methods

Attack Method	BLIP2-OPT6.7B	MiniGPT4-Vicuna13B	LLaVA-OneVision-7B	Phi4-Multimodal	QwenVL2.5-7B	MiniCPM-o-2.6	InternVL3-8B
Clean	57.94	47.98	81.47	76.12	81.04	80.19	79.54
AttackVLM-ii [22]	49.70	40.90	70.92	62.19	66.04	67.28	67.93
AnyAttack [24]	50.19	42.27	69.5	59.8	63.68	65.22	66.46
FOA [26]	44.15	39.72	64.8	59.69	61.2	61.13	61.07
MABA	**40.40**	**31.28**	**60.22**	**52.38**	**52.37**	**54.89**	**54.17**

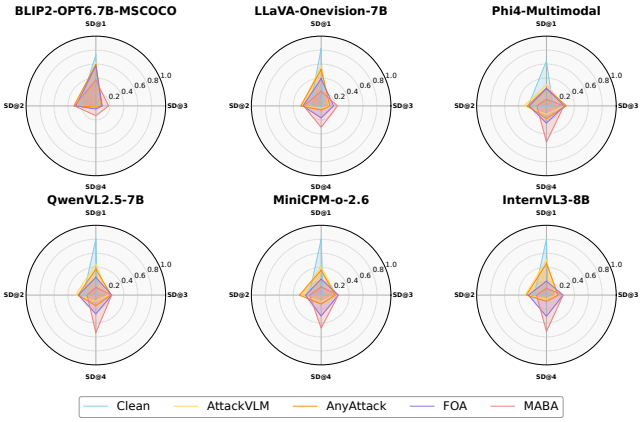

Figure 4: Semantic Deviation distribution across evaluated LVLMs under different attacks.

damentally contradict the original image content.

5.2.2 Results on VQA Task

On the VQA task, although LVLMs exhibit slightly improved robustness compared to the IC task, MABA still leads the strongest degradation of VQA accuracy across all evaluated LVLMs.

As shown in Table 2, MABA consistently achieves the strongest VQA accuracy degradation across all LVLMs. Specifically, LLaVA-OneVision-7B, which demonstrates the highest robustness under MABA, suffers a drop from 81.47% to 60.22%. The quantitative results in Figure 3 further demonstrate the superior performance of MABA.

Table 3: Average VQA accuracy (%) across answer types

Attack Method	*Yes/No*	*Number*	*Other*
Clean	82.74	58.33	67.07
AttackVLM-ii [22]	76.05	44.44	52.67
AnyAttack [24]	75.10	43.90	51.26
FOA [26]	73.46	41.77	45.65
MABA	**68.73**	**35.05**	**37.60**

We conduct a fine-grained analysis from both the question-type and answer-type perspectives. MABA reduces the accuracy of 32 among 65 VQAv2 question types to below 50% on average across the evaluated models, indicating widespread degradation in visual reasoning capabilities. Detailed statistics are provided in Appendix Section D. When analyzed by answer types, as shown in Table 3, the "yes/no" type demonstrates the highest robustness, with 68.73% accuracy under MABA. This robustness may stem from the inherent difficulty of fully reversing the semantics of salient objects. In contrast, the "number" type is heavily affected because its predictions depend on fine-grained features, resulting in a reduced accuracy of 35.05%. The "other" type experiences the largest overall degradation, with accuracy decreasing from 67.07% to 37.60%. Such results confirm that MABA introduces significant disruptions to the multimodal representations required for complex reasoning tasks.

5.2.3 Results on Commercial LVLMs

To explore the transferability of MABA to commercial black-box models, we evaluate its performance on IC and VQA

Table 4: Comparison of GPTScore and VQA accuracy under clean and MABA settings across commercial LVLMs.

Task	Attack	GPT-5	Claude-4.5	Gemini2.5-Pro
IC	Clean	84.49	81.29	84.95
	MABA	42.13	43.99	39.11
VQA	clean	80.88	72.41	79.71
	MABA	62.23	61.38	57.90

Table 5: Comparison of Performance under Clean and MABA settings across LLaVA variants.

Task	Attack	LLaVA-OneVision-7B	Robust-CLIP	Robust-LLaVA-G
IC	Clean	81.97	68.52	68.59
	MABA	43.02	62.62	66.86
VQA	Clean	81.47	68.25	65.56
	MABA	60.22	65.15	64.61

tasks across state-of-the-art commercial LVLMs, including GPT-5, Claude-4.5, and Gemini2.5-Pro. These results in Table 4 demonstrate that MABA consistently compromises both perceptual and reasoning multimodal capabilities across commercial LVLMs, highlighting its broad attack effectiveness.

5.2.4 Results on Robust LVLMs

To investigate MABA against defense methods, we examine Robust-CLIP [61] and Robust-LLaVA [62]. As shown in Table 5, while adversarial fine-tuning offers relatively effective mitigation, it comes at the expense of clean accuracy. Notably, the clean performance of these robust models drops to a level comparable to that of LLaVA-OneVision-7B under AttackVLM attack. We further discuss the reasons for the effectiveness of adversarial fine-tuning in Section 6.

5.3 Factors Influencing the Robustness

In this section, we investigate factors influencing the adversarial robustness of LVLMs using the evaluation tool MABA, whose leading performance is shown in Section 5.2. Detailed configurations of the evaluated LVLMs are listed in Table 6.

5.3.1 LLM Scale vs. Visual Encoder Scale

Increasing the scale of LVLM has been widely recognized as an effective approach to improving general performance [63]. Inspired by this, we investigate whether scaling the LLM or the visual encoder contributes to improving the adversarial robustness of LVLMs. To this end, we conduct experiments on three representative LVLM families: LLaVA-OneVision, QwenVL2.5, and InternVL3, each employing a fixed visual encoder with varying LLM sizes.

When the architecture and scale of the visual encoder are fixed, increasing the LLM scale has a limited impact on the robustness against visual adversarial examples. As illustrated in Figure 5, for example, QwenVL2.5-3B and

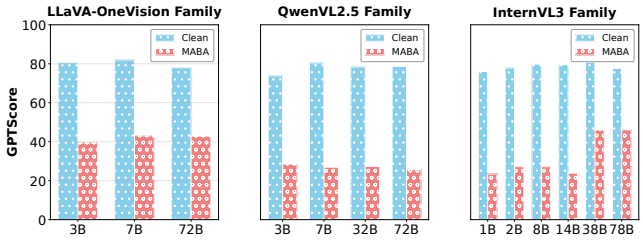

Figure 5: LLM Scale vs. Visual Encoder Scale: GPTScore of three LVLM families on IC Tasks.

QwenVL2.5-72B achieve similar GPTScore under MABA despite the large difference in LLM scale.

Adversarial robustness tends to increase with the scale of the visual encoder. Within the InternVL3 family, models equipped with larger visual encoders consistently exhibit stronger adversarial robustness than those using smaller encoders. Although LLM scales are not strictly controlled within InternVL3 variants, we attribute the significant robustness boost from InternVL3-1B–14B to InternVL3-38B/78B predominantly to the visual encoder expansion. This aligns with the previously observed minimal impact of LLM scale on robustness. The above phenomenon can be explained by the LVLM pipeline described in Section 2.1: **When visual inputs are perturbed by adversarial perturbations, the LLM faithfully reasons over flawed tokens.** Thus, increasing the LLM scale yields negligible improvements against visual adversarial attacks, whereas scaling up the visual encoder enhances adversarial robustness.

5.3.2 Number of Trainable Modules

As shown in Table 1, BLIP2-OPT6.7B-MSCOCO exhibits strong adversarial robustness on the MSCOCO captioning task when using the EVA-G visual encoder, whereas MiniGPT4-Vicuna13B performs poorly under attack despite sharing the same encoder. A key distinction may lie in the number of trainable modules: MiniGPT4-Vicuna13B only optimizes a linear adapter, keeping the rest modules of the model frozen, whereas BLIP-2-OPT6.7B-MSCOCO jointly trains both the encoder and the adapter. Therefore, we study how the extent of module unfreezing, *i.e.*, which and how many model components are made trainable, affects adversarial robustness.

To this end, we evaluate multiple BLIP2 variants with the EVA-G encoder under different training settings, comparing pretrained versions that update only the adapter with MSCOCO-finetuned versions that jointly update the visual encoder and the adapter. We additionally include InstructBLIP-Vicuna13B and MiniGPT4-Vicuna13B for comparison, which follow similar backbone-frozen training strategies but differ in trainable parameters.

LVLM's robustness tends to improve as more model

Table 6: Module configurations of evaluated LVLMs. (*) denotes trainable; (-) denotes frozen. Visual encoder parameter counts are shown in parentheses. Gray rows indicate model configurations that exhibit stronger robustness under evaluation.

Model	Visual Encoder	Adapter	LLM
BLIP2-OPT2.7B, OPT6.7B, FlanT5XL	EVA-G (1B) (-)	Qformer (*)	OPT2.7B, OPT6.7B, FlanT5XL (-)
InstructBLIP-Vicuna13B	EVA-G (1B) (-)	Qformer (*)	Vicuna13B (-)
BLIP2-OPT2.7B, OPT6.7B, FlanT5XL-MSCOCO	EVA-G (1B) (*)	Qformer (*)	OPT2.7B, OPT6.7B, FlanT5XL (-)
MiniGPT4-Vicuna13B	EVA-G (1B) (-)	Qformer (-) + Linear (*)	Vicuna13B (-)
Phi-4-MultiModal	SigLIP-SO400M (400M) (*)	2-Layer MLP (*)	Phi4-mini (3.8B) (*)
LLaVA-OneVision-3B, 7B, 72B	SigLIP-SO400M (400M) (*)	2-Layer MLP (*)	Qwen2-3B, 7B, 72B (*)
QwenVL2.5-3B, 7B, 72B	QwenViT (~600M) (*)	2-Layer MLP (*)	Qwen2.5-3B, 7B, 72B (*)
InternVL3-1B, 2B, 8B, 14B	InternViT300M-v2.5 (304M) (*)	2-Layer MLP (*)	Qwen2.5-0.5B, 1.5B, 7B, 14B (*)
InternVL3-38B, 78B	InternViT6B-v2.5 (6B) (*)	2-Layer MLP (*)	Qwen2.5-32B, 72B (*)
MiniCPMo-2.6, MiniCPMv-2.6	SigLIP-SO400M (400M) (*)	Perceiver Resampler (*)	Qwen2.5-7B, Qwen2-7B (*)

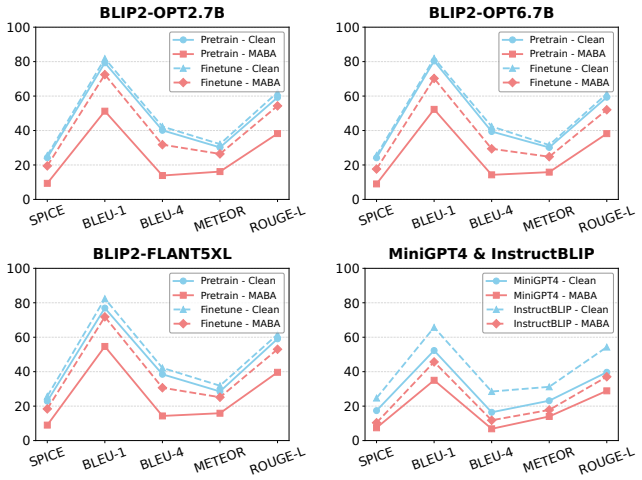

Figure 6: Ablation for numbers of trainable modules: conventional Image Captioning metrics across BLIP2 families and MiniGPT4 under MABA.

modules are trainable. As illustrated in Figure 6, across all BLIP2 variants, the MSCOCO-finetuned versions exhibit significantly improved adversarial robustness while also achieving slightly better clean performance. Similarly, InstructBLIP-Vicuna13B, which involves more trainable parameters, consistently achieves better clean accuracy and adversarial robustness than MiniGPT4-Vicuna13B. These observations align with previous findings [63], which suggest that unfreezing more modules improves the quality of vision-language knowledge acquisition in LVLMs. Taken together, our results indicate that vision-language knowledge built upon a minimal number of trainable parameters or modules may be more susceptible to adversarial perturbations.

5.3.3 Specialists vs. Generalists

There exists a trade-off between LVLM capability expansion and adversarial robustness. Notably, BLIP2-OPT6.7B-MSCOCO exhibits exceptional robustness on the IC task, as shown in Table 1. This is attributed to its fine-tuning on the IC task. However, this extreme specialization comes at the cost of degraded performance on other tasks. A similar pattern extends to the comparison between single-image and generalist models. LLaVA-OneVision-7B, which maintains a concentrated focus on single-image tasks, demonstrates stronger robustness on both single-image IC and VQA tasks, as shown in Table 1 and Table 2, compared to other models that learn rich multi-image and video knowledge. These observations suggest that focusing on specific tasks tends to enhance adversarial robustness for the targeted tasks, while expanding general capabilities may lead to *robustness forgetting*, similar to the knowledge forgetting observed in [64].

5.4 Further Analysis

Mechanism Study: Discriminative Feature Suppression. We evaluate the impact of different attacks on the High-frequency Components of visual representations across widely adopted visual encoders in LVLMs. Specifically, we perform SVD on the output of the first normalization layer following the self-attention layer in the last transformer block of each model, and visualize the top 10 and top 11-50 normalized singular values in descending order, as shown in Figure 7. All attack methods suppress the top 10 singular values across the evaluated visual encoders, with stronger attacks inducing a larger decrease, while the singular values ranked 11–50 exhibit a relative increase. **This suggests that one underlying mechanism of adversarial attacks is to weaken the most prominent discriminative features, allowing adversarial features to dominate the representation and alter the final prediction.** The DFS module explicitly amplifies this effect by suppressing the top-k singular values, redistributing information into other components of the spectrum. Such spectral changes are preserved across evaluated visual encoders, explaining why DFS enhances the cross-model transferability.
Hyperparameter Sensitive Analysis. We analyze the sensitivity of the hyperparameters in the DFS module, focusing on the proportion of suppressed singular values k and the suppression intensity exponent β. We vary the proportion of k from 0% to 25%, and β within $[1.5, 3.5]$. As shown in Figure 8,

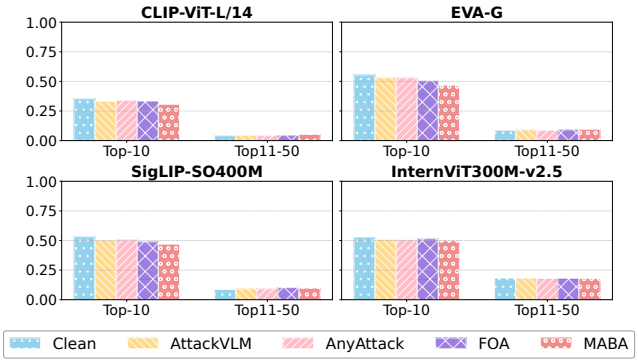

Figure 7: DFS effect: average normalized singular value spectra of evaluated visual encoders under different attacks.

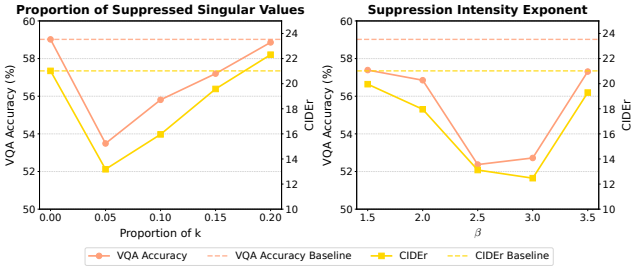

Figure 8: Effect of the hyperparameters k and β on MABA, evaluated on QwenVL2.5-7B. For the β analysis, we fix the proportion of k at 5%. For the k analysis, we fix $\beta = 2.5$.

optimal attack performance is achieved when the proportion of k is 5% (corresponding to $k = 10$) and β is 2.5. Suppressing a small proportion of singular values achieves good performance, aligning with the feature redistribution mechanism illustrated in Figure 7. Excessively large β leads the noise to overfit discriminative feature suppression at the expense of semantic deviation, thus reducing MABA's effectiveness.

6 Limitation and Discussion

Limited Effectiveness on Commercial LVLMs. We observe weaker attack efficacy on commercial LVLMs compared to open-source models. While GPT-4 claims to process all modalities within a single network, public technical reports [65] do not detail the underlying multimodal architecture. We attribute MABA's degradation to potential architectural gaps, where complex alignment mechanisms of commercial models hinder transferability. Nevertheless, MABA retains effectiveness against commercial LVLMs. This demonstrates that leveraging mutual information simulates modality alignment to a certain extent. Improving surrogate design to approximate the unknown alignment mechanisms remains an open problem.

Limited Effectiveness on Robust LVLMs. MABA exhibits limited attack efficacy against robust models. This result aligns with established empirical findings, confirming that sufficient adversarial training serves as a potent mitigation strategy against black-box adversarial attacks [66, 67]. Specifically, the strategy employed in Robust-LLaVA [62], which unfreezes the projector for joint fine-tuning with the adversarially trained visual encoder, effectively fortifies the modality alignment process, thereby serving as a robust countermeasure against MABA. Nevertheless, even the best robust models still incur a substantial cost in clean performance, lagging behind undefended models. Consequently, balancing the trade-off between maintaining clean performance and achieving adversarial robustness [68] remains a critical direction for future research.

Impact of Surrogate Architecture Selection. As shown in Table 1, MABA maintains robust transferability across LVLMs with diverse ViT structures. MiniGPT-4, in particular, exhibits the lowest adversarial robustness. Given that its encoder differs from our surrogate primarily in depth, this confirms that higher architectural similarity further boosts the performance of MABA. Additionally, we observe that increasing the surrogate model's depth further improves attack transferability, consistent with findings in [25, 26].

Adaptation to Targeted Adversarial Attack. This work validates that extending surrogate capabilities via the MIA projector and amplifying discriminative feature suppression effectively improve transferability. Theoretically, MABA can be adapted for targeted attacks by modifying the surrogate adapter's construction and imposing targeted constraints on the energy-redistributing DFS module. Given the distinct attack paradigms illustrated in Figure 1, we leave this exploration to future work.

7 Conclusion

In this work, we proposed MABA, a black-box untargeted adversarial attack framework for evaluating the robustness of LVLMs. By simultaneously disturbing both the visual encoding and modality alignment phases, MABA achieved state-of-the-art black-box attack performance on advanced LVLMs. Extensive evaluations show that MABA causes an average 58.37% drop in semantic metrics for image captioning and a 31.63% drop in accuracy for VQA. We then proposed several practical guidelines for building more robust LVLM systems, spanning model selection, training, and system design.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No. U24B20182) and Key R&D Program of Zhejiang (Grant No. 2024C01164).

Ethical Considerations

This work aims to advance the understanding of adversarial robustness in LVLMs by proposing a novel black-box adversarial attack, Modality Alignment Breaking Attack (MABA), designed exclusively for research and evaluation purposes. The primary goal is to reveal new vulnerabilities introduced by the multimodal architecture of LVLMs, thereby enabling the community to build more robust systems.

Stakeholders and Potential Impact The stakeholders potentially affected by this work include: (1) AI safety researchers and model developers, who can benefit from MABA to evaluate and strengthen LVLM models against adversarial attacks. (2) Deployed system operators, who manage LVLM-based systems and could face service disruptions if MABA techniques were misused. (3) End users, who rely on LVLM-based systems in applications such as education, healthcare, or assistive technologies, could experience compromised service integrity.

The research process and its publication may have different impacts on these groups. For the research community, MABA serves as an evaluation tool for assessing black-box robustness, helping researchers develop more robust LVLMs. For system operators, misuse of MABA could disrupt service functionality. For end users, such misuse could compromise service integrity. These risks should be carefully considered and addressed before the large-scale deployment of LVLM applications.

Mitigation of Negative Impacts To minimize potential harm, all experiments are conducted under controlled conditions using publicly available, non-sensitive datasets. No personal, biometric, or proprietary data is used. We do not release attack-ready scripts targeting proprietary APIs.

We show that existing adversarial training strategies can mitigate MABA to some extent, and we highlight multimodal alignment consistency as a promising new defense direction. Furthermore, MABA reveals that current adversarial attacks follow consistent patterns in suppressing discriminative high-frequency features, which could be leveraged to design more general detection and defense strategies.

Decision to Conduct and Publish The decision to conduct this research stems from our goal to explore new adversarial vulnerabilities introduced by the multimodal architecture of LVLMs. MABA investigates robustness challenges that represent attack surfaces fundamentally different from those in unimodal systems. We recognize that adversarial examples represent inherent vulnerabilities of models that pose persistent threats to system functionality throughout the deployment lifecycle. Early identification and understanding of these vulnerabilities enable researchers to build more robust models and develop effective defenses before such weaknesses are exploited in real-world systems.

As LVLM applications become increasingly prevalent across various domains, we believe it is essential to share our findings with the research community to facilitate the development of more robust systems. Open-source publication allows researchers worldwide to understand these alignment-based vulnerabilities, validate our results, and contribute to collective defense efforts. This approach aligns with responsible disclosure practices in AI safety research. We have chosen to disseminate our work through peer-reviewed publication to ensure proper evaluation and oversight by the research community, balancing the need for transparency with the imperative of responsible knowledge sharing.

Open Science

We affirm that this paper complies with open science policies by promoting transparency, reproducibility, and accessibility in research. The data, methodology, and findings presented in this paper are openly available for review and replication.

Dataset and Models. The datasets and model architectures used in this research are all open-source. Therefore, the results and findings of this study can easily be replicated and validated by other researchers in the field.

Source Code. The source code of the proposed MABA is publicly available at https://github.com/LeeZSir/MABA. An archived version is also provided on Zenodo at https://zenodo.org/records/17958814 to ensure long-term accessibility. This resource enables researchers and practitioners to implement and evaluate MABA in their own experiments, thereby supporting further investigation into model robustness.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[2] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[3] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

[6] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[7] Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

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

[9] Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, et al. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19985–19995, 2025.

[10] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[11] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290, 2024.

[12] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

[13] Zhang Tianyuan, Liu Jiangfan, Guo Yongkang, Zhong Fangzhi, Bao Wei, Dong Jian, Liu Aishan, and Liu Xianglong. Towards secure and robust vision-language models in autonomous driving: A survey for perception-oriented and decision-oriented attacks. *Chinese Journal of Electronics*, 2025.

[14] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

[15] Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv preprint arXiv:2407.07403*, 2024.

[16] Lu Wang, Tianyuan Zhang, Yang Qu, Siyuan Liang, Yuwei Chen, Aishan Liu, Xianglong Liu, and Dacheng Tao. Black-box adversarial attack on vision language models for autonomous driving. *arXiv preprint arXiv:2501.13563*, 2025.

[17] Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3677–3685, 2023.

[18] Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In *The Twelfth International Conference on Learning Representations*, 2023.

[19] Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24625–24634, 2024.

[20] Kuofeng Gao, Yang Bai, Jiawang Bai, Yong Yang, and Shu-Tao Xia. Adversarial robustness for visual grounding of multimodal large language models. *arXiv preprint arXiv:2405.09981*, 2024.

[21] Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. How robust is google's bard to adversarial image attacks? *arXiv preprint arXiv:2309.11751*, 2023.

[22] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36:54111–54138, 2023.

[23] Qi Guo, Shanmin Pang, Xiaojun Jia, Yang Liu, and Qing Guo. Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models. *IEEE Transactions on Information Forensics and Security*, 2024.

[24] Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, and Dit-Yan Yeung. Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19900–19909, 2025.

[25] Zhaoyi Li, Xiaohan Zhao, Dong-Dong Wu, Jiacheng Cui, and Zhiqiang Shen. A frustratingly simple yet highly effective attack baseline: Over 90% success rate against the strong black-box models of gpt-4.5/4o/o1. *arXiv preprint arXiv:2503.10635*, 2025.

[26] Xiaojun Jia, Sensen Gao, Simeng Qin, Tianyu Pang, Chao Du, Yihao Huang, Xinfeng Li, Yiming Li, Bo Li, and Yang Liu. Adversarial attacks against closed-source mllms via feature optimal alignment. *arXiv preprint arXiv:2505.21494*, 2025.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[29] Nathan Inkawhich, Kevin J Liang, Jingyang Zhang, Huanrui Yang, Hai Li, and Yiran Chen. Can targeted adversarial examples transfer when the source and target models have no label space overlap? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 41–50, 2021.

[30] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

[31] Weimin Lyu, Lu Pang, Tengfei Ma, Haibin Ling, and Chao Chen. Trojvlm: Backdoor attack against vision language models. In *European Conference on Computer Vision*, pages 467–483. Springer, 2024.

[32] Yunhao Gou, Tom Ko, Hansi Yang, James Kwok, Yu Zhang, and Mingxuan Wang. Leveraging per image-token consistency for vision-language pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19155–19164, 2023.

[33] Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 102–111, 2023.

[34] Sensen Gao, Xiaojun Jia, Xuhong Ren, Ivor Tsang, and Qing Guo. Boosting transferability in vision-language attacks via diversification along the intersection region of adversarial trajectory. In *European Conference on Computer Vision*, pages 442–460. Springer, 2024.

[35] Xiaojun Jia, Sensen Gao, Qing Guo, Ke Ma, Yihao Huang, Simeng Qin, Yang Liu, and Xiaochun Cao. Semantic-aligned adversarial evolution triangle for high-transferability vision-language attack. *arXiv preprint arXiv:2411.02669*, 2024.

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

[37] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[38] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[39] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15671–15680, 2022.

[40] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

[41] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

[42] Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025.

[43] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5005–5013, 2022.

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

[45] Justin B Kinney and Gurinder S Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359, 2014.

[46] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.

[47] Chen Henry Wu, Rishi Shah, Jing Yu Koh, Ruslan Salakhutdinov, Daniel Fried, and Aditi Raghunathan. Dissecting adversarial robustness of multimodal lm agents. *arXiv preprint arXiv:2406.12814*, 2024.

[48] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

[49] Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. *arXiv preprint arXiv:2203.05962*, 2022.

[50] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

[51] Seokju Yun, Seunghye Chae, Dongheon Lee, and Youngmin Ro. Soma: Singular value decomposed minor components adaptation for domain generalizable representation learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25602–25612, 2025.

[52] Di Ming, Peng Ren, Yunlong Wang, and Xin Feng. Boosting the transferability of adversarial attack on vision transformer with adaptive token tuning. *Advances in Neural Information Processing Systems*, 37:20887–20918, 2024.

[53] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[54] Dor Tsur, Ziv Goldfeld, and Kristjan Greenewald. Max-sliced mutual information. *Advances in neural information processing systems*, 36:80338–80351, 2023.

[55] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[56] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[57] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[58] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[59] Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, 2024.

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

[61] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 43685–43704, 2024.

[62] Hashmat Shadab Malik, Fahad Shamshad, Muzammal Naseer, Karthik Nandakumar, Fahad Khan, and Salman Khan. Robust-llava: On the effectiveness of large-scale robust image encoders for multi-modal large language models. *arXiv preprint arXiv:2502.01576*, 2025.

[63] Bo Li, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, and Chunyuan Li. Llava-next: What else influences visual instruction tuning beyond data?, May 2024.

[64] Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023.

[65] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[66] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[67] Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training. *Advances in neural information processing systems*, 32, 2019.

[68] Hongshan Yang, Zhichao Li, Zhibo Wang, Peng Sun, Zhixuan Chu, and Feng Lin. Toward defending adversarial patch attacks with mask-reconstruction-assisted adversarial training. *IEEE Transactions on Information Forensics and Security*, 20:12476–12490, 2025.

[69] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

A Comparison with VLP attack

To further validate the effectiveness of MABA in disrupting the modality alignment process, we extend our evaluation to **vision-language pretraining (VLP) models** on the **image-text retrieval (ITR)** task. The ITR task aims to retrieve relevant images given a text query and vice versa, thereby assessing a model's ability to align visual and textual semantics across modalities. Regarding the selection of evaluated models, we include representative VLP models, namely CLIP [27], ALBEF [38], and TCL [39].

A.1 Experiment setup

For the ITR task, we adopt **MSCOCO** [57] and **Flickr30k** [69], both providing aligned image-text pairs for bidirectional retrieval. On MSCOCO, we randomly select 1,000 images from the Karpathy test split [55] to generate adversarial examples, which are further used for ITR evaluations. On Flickr30k, we use its standard 1,000-image test set to generate adversarial examples specifically for ITR evaluation.

We report Attack Success Rate (ASR) for both Image-to-Text Retrieval@1 (IR@1) and Text-to-Image Retrieval@1 (TR@1), where @1 denotes the top 1 retrieved result. ASR is defined as the percentage of samples where the ground-truth pair originally ranked at the top 1 fails to remain at that position after perturbation, indicating a successful disruption of modality alignment.

We compare MABA with established adversarial attack baselines for VLP models, including *SGA* [33], *DRA* [34], and *SAAET* [35]. For a fair comparison, all adversarial examples are generated under the ℓ_∞ norm constraint with a perturbation bound of $\varepsilon = 16/255$. For attacks requiring iterative optimization, we set the number of steps to 30 and the step size to $3/255$. In all VLP evaluations, we adopt CLIP_ViT-B/16 as the surrogate model, with both visual and textual embedding dimensionalities set to $d_v = d_t = 512$.

MABA Configuration. For attack parameter settings and surrogate model selection, we remain consistent with the baseline configuration. For the DFS Module, we select layers 4–9 in CLIP_ViT-B/16 for spectral regularization. The regularization is applied to the outputs after the layer normalization that follows the multi-head self-attention module in each Transformer block, ensuring consistency with the analysis framework described in [49]. We set the number of dominant singular values to suppress to $k = 10$, and the regularization weight to $\beta = 2.5$. For the MIA Projector Training Module, we set the projected embedding dimension at $d_k = 448$ for CLIP_ViT-B/16, which corresponds to 87.5% of the original embedding space. The training data are consistent with the downstream test data, utilizing the same image-text pairs sampled from MSCOCO and Flickr30k. To instantiate the measurable scoring function $m : \mathbb{R}^{d_k} \times \mathbb{R}^{d_k} \to \mathbb{R}$, we employ a Q-Former module and adopt the cosine similarity function introduced in [58] as scoring functions. The MIA projector is trained using the Adam optimizer with an initial learning rate of 2×10^{-4} for 100 epochs, which decays by a factor of 0.1 every 40 epochs. The balance parameter for the loss strength, λ, is set to 0.36.

A.2 Results on ITR Task

MABA consistently outperforms all evaluated attack methods in terms of Attack Success Rate (ASR). Among existing methods, SA-AET achieves the highest attack success rates on Flickr30K and MSCOCO. On Flickr30K, **MABA** achieves an ASR improvement of +2.94% in TR@1 and +4.29% in IR@1 on CLIP-ResNet50, which relies solely on image-text contrastive loss for modality alignment. On TCL, which employs a cross-attention mechanism for multimodal fusion, the improvements reach +9.70% and +5.71% in TR@1 and IR@1,

Table 7: Attack Success Rate (%) on Flickr30K and MSCOCO (Source: CLIP-ViT$_{B/16}$)

Attack Method	Flickr30K								MSCOCO							
	CLIP-ViT$_{B/16}$		CLIP-CNN$_{Res50}$		ALBEF		TCL		CLIP-ViT$_{B/16}$		CLIP-CNN$_{Res50}$		ALBEF		TCL	
	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1
SGA [33]	100.0	99.94	56.7	59.31	29.3	33.93	28.66	35.1	100.0	99.85	56.98	61.36	31.51	37.56	25.31	28.82
DRA [34]	100.0	100.0	75.73	77.67	48.8	55.52	49.32	55.0	100.0	100.0	71.23	77.67	50.44	53.6	44.84	47.79
SA-AET [35]	100.0	100.0	80.84	80.0	60.9	64.55	58.69	61.74	100.0	**100.0**	76.78	78.76	55.14	60.48	53.92	56.1
MABA	**100.0**	**100.0**	**83.78**	**84.29**	**67.26**	**68.76**	**68.39**	**67.45**	**100.0**	99.92	**85.61**	**85.61**	**63.57**	**65.08**	**62.32**	**60.51**

respectively. On MSCOCO, **MABA** achieves more consistent gains across diverse models, improving TR@1 by an average of +8.22% and IR@1 by +5.29% over SAAET. These results demonstrate that **MABA** maintains strong modality alignment disruption capabilities across architectures with different alignment strategies.

B Prompts

B.1 Prompts for Tasks

For IC tasks, BLIP2-OPT6.7B-MSCOCO uses the default prompt *"a photo of"*, MiniGPT4-Vicuna13B adopts *"Describe this image in detail."*, and other models are prompted with *"You are doing the image captioning task. Describe this image in one short sentence only."*

For VQA tasks, BLIP2-OPT6.7B-Pretrain and MiniGPT4-Vicuna13B adopt their default question-answering format *"Question: {} Answer: {}"*, and other models use *"You are doing the VQA task. Please answer the following question in a word or short phrase."*

B.2 Prompts for IC Evaluation

We employ `GPT-4o` to evaluate the semantic consistency between generated captions and reference captions. The resulting **GPTScore** is assigned in the range $[0, 1]$, where 0 indicates completely different meanings and 1 indicates identical meanings. The evaluation prompt is defined as follows:

Main Objects (40%): *Key objects from the reference should be present. Matching any one reference sentence is sufficient. Extra objects are acceptable.* ***Actions & Relationships (30%):*** *The main actions and relationships must be preserved, focusing on "who does what to whom".* ***Attributes (20%):*** *Core attributes such as color, size, and number should be correct.* ***Fluency (10%):*** *The caption should be clear and grammatical, with minor errors tolerated.*

Add up to $+0.1$ *for especially clear or faithful meaning. Deduct up to* -0.1 *for serious errors or irrelevant content.*

B.3 Prompts for VQA Evaluation

For models that cannot directly generate short answers (*e.g.*, BLIP2-OPT6.7B-Pretrain and MiniGPT4-Vicuna13B), we employ `GPT-4o-mini` to compute VQA accuracy. The evaluation prompt is defined as follows:

Question: {} Generated Answer: {} Human Answers: {} Compute the VQA accuracy using the official formula:

$$Accuracy = \min\left(\frac{n}{3}, 1\right),$$

where n denotes the number of human answers that semantically match the model-generated answer.

Return only the VQA score as a single number between 0 *and* 1*, without any explanation or additional text.*

C SD@k Scores across Evaluated LVLMs

We present the distribution of SD categories across different LVLMs under various attack methods in Table 8.

D Statistics based on VQA Question Types

We report the counts of per-question-type VQA accuracy intervals across LVLMs under MABA in Table 9. Note that the VQA accuracy of MiniGPT-4-Vicuna13B and BLIP2-Pretrain are computed directly by `GPT-4o-mini` and thus excluded from the statistics.

E Mechanism Study of MIA Projector

Based on the visual encoders evaluated in our experiments, we further select their corresponding source CLIP models to perform a max-sliced mutual information (mSMI) analysis. We report the average mSMI values between visual and textual embeddings under different attack methods on CLIP_ViT-L/14, EVA-G, and SigLIP-SO400M. The mSMI is estimated using cosine similarity as the measurable function, and computed according to Equation (14). As shown in Figure 9, stronger attacks result in larger reductions in mSMI across evaluated CLIP models. By employing the MIA projector to generate adversarial examples in a modality-aligned representation subspace, MABA learns to effectively disrupt modality alignment, consistently yielding the largest drops in both metrics across all evaluated models. This confirms that incorporating the projector into the attack pipeline enhances attack effectiveness.

Table 8: Distribution of Semantic Deviation categories (in %) across LVLMs under different attack methods

Source Model	Attack Method	SD@1 (%)	SD@2 (%)	SD@3 (%)	SD@4 (%)
BLIP2-OPT6.7B-MSCOCO	Clean	72.72	23.23	3.065	0.41
	AttackVLM-ii [22]	62.44	28.30	8.16	1.11
	AnyAttack [24]	59.05	31.55	7.89	1.52
	FOA [26]	57.70	28.70	9.10	4.50
	MABA	**37.09**	**30.69**	**18.19**	**14.02**
MiniGPT4-Vicuna13B	Clean	43.72	37.49	16.58	2.21
	AttackVLM-ii [22]	22.03	33.10	36.26	18.51
	AnyAttack [24]	27.85	32.09	25.03	15.04
	FOA [26]	24.22	27.33	21.72	26.73
	MABA	**4.94**	**8.78**	**17.76**	**68.52**
LLaVA-OneVision-7B	Clean	82.14	15.54	2.32	0.00
	AttackVLM-ii [22]	57.73	27.84	11.24	3.20
	AnyAttack [24]	52.34	29.23	12.42	6.01
	FOA [26]	38.90	26.10	17.80	17.20
	MABA	**20.95**	**25.10**	**23.38**	**30.57**
Phi4-Multimodal	Clean	64.34	26.06	9.19	0.40
	AttackVLM-ii [22]	30.45	32.26	25.23	12.06
	AnyAttack [24]	25.65	27.57	28.57	18.21
	FOA [26]	24.80	24.70	25.80	24.70
	MABA	**9.52**	**13.78**	**25.03**	**51.67**
QwenVL2.5-7B	Clean	80.10	17.68	2.12	0.10
	AttackVLM [22]	44.53	29.15	17.31	9.01
	AnyAttack [24]	37.25	25.51	21.86	15.28
	FOA [26]	25.30	24.80	22.80	27.10
	MABA	**11.08**	**12.49**	**22.66**	**53.78**
MiniCPM-o-2.6	Clean	79.88	17.29	2.63	0.20
	AttackVLM-ii [22]	40.99	28.34	21.36	9.31
	AnyAttack [24]	35.62	31.18	20.28	12.92
	FOA [26]	22.80	22.40	24.80	30.00
	MABA	**11.63**	**16.46**	**24.69**	**47.22**
InternVL3-8B	Clean	80.42	17.46	2.12	0.00
	AttackVLM-ii [22]	50.30	30.39	13.62	5.69
	AnyAttack [24]	45.33	28.80	16.73	9.13
	FOA [26]	19.90	25.80	23.80	30.50
	MABA	**9.80**	**14.24**	**23.94**	**52.02**

Table 9: Counts of per-question-type VQA accuracy intervals across LVLMs under MABA

Source Model	0∼25%	25%∼50%	50%∼75%	75%∼100%
LLaVA-OneVision-7B	5	18	26	16
Phi4-Multimodal	12	15	23	15
QwenVL2.5-7B	9	19	22	15
MiniCPM-o-2.6	11	16	21	17
InternVL3-8B	14	39	11	1

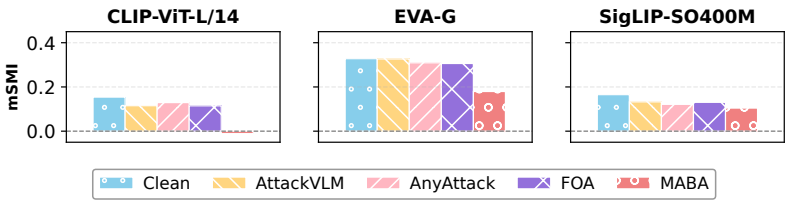

Figure 9: MIA projector effect: average mSMI of evaluated CLIP models under different attacks.