Towards Building Model/Prompt-Transferable Attackers against Large Vision-Language Models

Xiaowen Cai[1,*], **Daizong Liu**[2*†], **Xiaoye Qu**[1], **Xiang Fang**[3], **Jianfeng Dong**[4], **Keke Tang**[5], **Pan Zhou**[1‡], **Lichao Sun**[6], **Wei Hu**[7‡]
[1]Huazhong University of Science and Technology [2]Wuhan University
[3]Nanyang Technological University [4]Zhejiang Gongshang University [5]Guangzhou University
[6]Lehigh University [7]Peking University
{xwcai,xiaoye,panzhou}@hust.edu.cn, daizongliu@whu.edu.cn,xfang9508@gmail.com
djf@zjgsu.edu.cn,tangbohutbh@gmail.com,lis221@lehigh.edu,forhuwei@pku.edu.cn

Abstract

Although Large Vision-Language Models (LVLMs) exhibit impressive multimodal capabilities, their vulnerability to adversarial examples has raised serious security concerns. Existing LVLM attackers simply optimize adversarial images that easily overfit a certain model/prompt, making them ineffective once they are transferred to attack a different model/prompt. Motivated by this research gap, this paper aims to develop a more powerful attack that is transferable to black-box LVLM models of different structures and task-aware prompts of different semantics. Specifically, we introduce a new perspective of information theory to investigate LVLMs' transferable characteristics by exploring the relative dependence between outputs of the LVLM model and input adversarial samples. Our empirical observations suggest that enlarging/decreasing the mutual information between outputs and the disentangled adversarial/benign patterns of input images helps to generate more agnostic perturbations for misleading LVLMs' perception with better transferability. In particular, we formulate the complicated calculation of information gain as an estimation problem and incorporate such informative constraints into the adversarial learning process. Extensive experiments on various LVLM models/prompts demonstrate our significant transfer-attack performance.

1 Introduction

Large Vision-Language Models (LVLMs) have garnered significant attention for their impressive capabilities in both visual perception and language interaction. Unlike pure-text Large Language Models, LVLMs incorporate visual encoders, enabling them to excel in a variety of multimodal tasks such as text-to-image generation [1, 2, 3], visual question-answering [4, 5, 6], and *etc.*. However, since model complexity increases and their applications expand into real-world scenarios, security concerns [7] of LVLM models have become increasingly prominent.

Recent studies [7, 8, 9, 10, 11, 12, 13] have revealed that LVLMs are highly vulnerable to adversarial attacks, which can significantly degrade performance and pose serious security risks. Specifically, these works [14, 15, 16, 17, 18, 19, 20, 9] manipulate LVLMs by injecting imperceptible perturbations into benign images, misleading the model to produce incorrect or jailbreak results. Despite the progress in this area, existing attack methods face primary challenges. In particular, they are typically optimized for specific LVLM architectures or fixed prompts, making the generated adversarial examples difficult to transfer effectively across different models or downstream tasks. That is, they

*Equal Contributions. †Project Lead. ‡Corresponding Authors

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

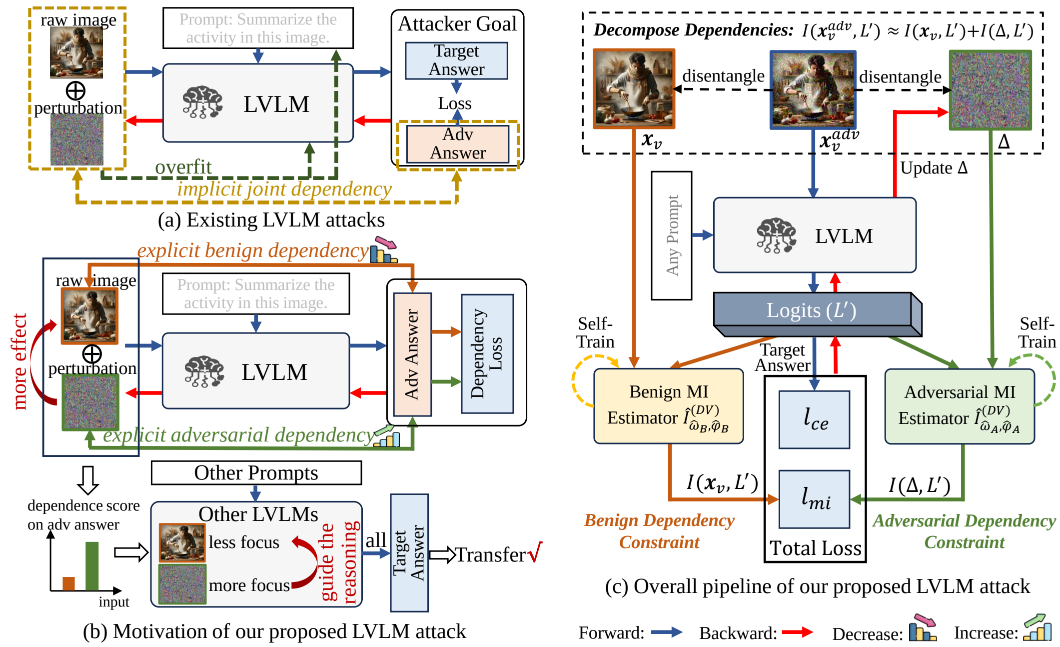

Figure 1: (a) Existing attacks may cause interference overfitting by implicitly restricting the mixed output-input dependency via misleading loss functions. (b) In contrast, we aim to learn more harmful perturbations via disentangled informative dependencies to control the LVLM reasoning trends for improving transferability. (c) Overall pipeline of our proposed method: first, we theoretically decompose the mixed MI information into benign MI and adversarial MI components. Then, we train two MI estimators for separate MI calculations. Finally, we incorporate the benign and adversarial MI constraints into the optimization strategy to generate transferable adversarial examples.

often fail to maintain effectiveness across diverse models and prompts simultaneously in practice, requiring attackers to craft separate perturbations for each model and each prompt, resulting in significant time and resource overhead. Although a few recent studies [8, 10] have explored prompt-agnostic attack strategies, they not only rely on complex multi-prompt joint training schemes, but also fail to address the more challenging problem of cross-model transferability.

Therefore, in this paper, we make the first attempt to design a superior LVLM attacker that can achieve both model- and prompt-transfer attacks within a single adversarial learning process. Unlike previous 2D/3D transfer works [21, 22, 23, 24] that improve the generalization of adversarial perturbations by resisting various distortions, we propose to investigate the agnostic/generalizable harmfulness of perturbations from a new information theory perspective [25, 26, 27, 28]. Our core idea is: *the informative dependence between the output of the LVLM model and the input images explicitly reflects the LVLMs' decision trajectory to make the final predictions and, therefore, a generalizable adversarial perturbation should have as more harmful effect as possible to control the flip of the LVLMs' prediction than the benign pattern in the image input.* As in Figure 1 (a)(b), the existing LVLM attackers adversarially train the adversarial samples by implicitly restricting the mixed output-input dependency via misleading loss functions, which may confuse the LVLM model to focus on the joint distribution of benign and adversarial patterns of inputs, resulting in an interference overfitting. Instead, once we explicitly adjust the LVLM's focus solely on the adversarial noise to enhance the corresponding adversarial harmfulness, the learned adversarial perturbation is able to jump out of the mixed overfitting and contributes more attacker-chosen guidance effects than the benign one to mislead the reasoning process even the sample is transferred to unknown LVLM models or prompts.

Based on the above observations, we propose a novel LVLM attack method to adversarially constrain the informative dependence between the benign/adversarial pattern of the input and the LVLM's output for improving the model/prompt-aware transferability. In particular, we exploit mutual information (MI) to explicitly measure such dependence via coefficient degrees, where a larger MI degree indicates a stronger dependence between the two variables. Since the mixed MI of the entire adversarial input cannot consider the dependence of the output on the different patterns, we theoretically demonstrate that this mixed MI is closely related to the linear sum of benign MI (between the output and the benign pattern) and adversarial MI (between the output and the adversarial pattern), therefore, we can disentangle the adversarial input into benign and adversarial parts for separate MI learning. We utilize lightweight neural networks to train with these two MI information as effective MI estimators via maximization strategy [29, 27, 30]. During adversarial learning, we dynamically enlarge the adversarial MI and decrease the benign MI of adversarial samples to force reasoning process to focus more on perturbations to enhance harmfulness. Results show that our adversarial samples containing larger adversarial MI achieve significant transfer-attack performance across various LVLMs/prompts.

The key contributions of our work are outlined as follows:

- We propose to address a practical but challenging LVLM attack setting, *i.e.*, model/prompt-transfer attack. This new setting can efficiently generate effective adversarial examples against different models/prompts compared to existing time/resource-consuming attacks.
- To obtain generalizable adversarial examples, we introduce to enhance the harmfulness of perturbations from a novel information theory perspective to improve transferability. An effective MI constraint for individual benign/adversarial patterns is devised to adjust the focus of LVLM solely to the additive perturbations.
- Extensive experiments are conducted to verify the strong adversarial transferability of our proposed attack on four prevalent LVLM models and three multimodal datasets with a spectrum of task-aware prompts.

2 Related Work

LVLM Attackers. LVLMs generally combine the capabilities of processing visual information with natural language understanding by using pre-trained vision encoders with language models [31, 32]. Due to this multimodal nature [33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46], LVLMs are particularly vulnerable as the multi-modal integration not only amplifies their vulnerable utility but also introduces new attack vectors that are absent in unimodal systems [47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65]. Most of the existing LVLM attackers [14, 15, 66, 67, 68, 17, 19, 69, 7] are inspired by the adversarial vulnerability observed in vision tasks. To evaluate the adversarial robustness of LVLMs and generate adversarial examples, they generally add and optimize imperceptible perturbations on the whole image to benign image inputs via back-propagation. Although they can achieve significant attack performance in both targeted and untargeted settings, they are easily limited by their perturbation-specific design that can solely produce adversarial examples to deceive a particular LVLM model and prompt within a singular process. That is, to compromise different LVLMs and prompts, they must generate distinct adversarial perturbations, which incur significant time and resource expenditure. Some recent works [8, 10] try to develop cross-prompt attack approaches, however, they require complicated multi-prompt joint training and the challenging cross-model attack issue is still unexplored. Therefore, this paper aims to develop a model/prompt-transferable attack method that can efficiently and effectively fool the practical LVLM applications.

Adversarial Transferability. In the general 2D image and 3D point cloud fields, numerous works [21, 22, 23, 24] have been proposed to improve adversarial transferability. These methods claim that the adversarial examples easily overfit the targeted model, therefore, they should generate more generalizable and harmful perturbations. Most methods [24, 70, 23, 71, 72] exploit diverse transformations to force the adversarial examples to resist them for improving the generalization, leading to transferable perturbations with much more harmfulness. Advanced momentum-based optimization strategies [73, 74, 70, 75, 76] are further introduced to stabilize the optimization procedure and escape the local optima. There are also some ensemble attacks [77, 78, 79] that generate more transferable adversarial examples by attacking multiple models simultaneously. Several works [80, 81] also disrupt the feature space with model-agnostic designs to generate adversarial examples. Since there is no related work that systematically analyzes the adversarial transferability of LVLM attack methods across models/prompts, we follow previous works to generate as harmful as possible adversarial perturbations to improve the transferability of LVLM attacks.

3 Methodology

3.1 Problem Definition and Notations

We generally define an LVLM model as F, which receives an image $\boldsymbol{x}_v$ and a task-specific prompt $\boldsymbol{x}_p$ as the input pair to return a corresponding ground-truth answer $\boldsymbol{y}$.

Threat Model. In this paper, we explore the setting of transferable LVLM attacks, where we assume that the attacker solely has knowledge of a certain victim model, including its parameters, training procedure, *etc*. The attackers are required to generate adversarial examples on this white-box victim model, and feed them to attack other unknown black-box target LVLM models. This setting is more challenging and practical as the attackers cannot always access the details of real-world LVLM applications.

Attacker's Goal. The objective of the attacker is to devise and add a harmful but imperceptible perturbation $\boldsymbol{\Delta}$ on $\boldsymbol{x}_v$, to generate an adversarial image as $\boldsymbol{x}_v^{adv} = \boldsymbol{x}_v + \boldsymbol{\Delta}$. This adversarial example, upon application to any textual prompt across different LVLM models, is designed to compel the model to output a target label predetermined by the attacker. Therefore, such perturbation needs to exhibit persistence and robustness when deployed on unseen LVLM models, and to induce adversarial semantic alterations across different task-aware prompts for the same image, rendering the attack cross-model and cross-prompt applicable. In this paper, we mainly focus on targeted adversarial attacks that aim to craft the adversarial image $\boldsymbol{x}_v^{adv}$ to misguide the predicted answer of LVLM from the ground-truth label $\boldsymbol{y}$ to the specific targeted label $\boldsymbol{y}_{tar}$. The optimization goal is formulated as:

$$min\ J(F(\boldsymbol{x}_v^{adv}, \boldsymbol{x}_p), \boldsymbol{y}_{tar}),\ s.t. \|\boldsymbol{x}_v^{adv} - \boldsymbol{x}_v\|_\infty \leq \epsilon, \tag{1}$$

where $J(\cdot)$ is the loss function, and we utilize l_∞-norm to regularize the adversarial perturbation to the range ϵ.

3.2 Overview of Our Attack

Our Motivation. We consider improving attack transferability from the information theory perspective by separately increasing the harmful effect of adversarial perturbations while making the LVLM less sensitive to the benign pattern of the perturbed image. Specifically, to generate more harmful perturbations, we explicitly study the informative dependency between adversarial/benign patterns of the input and the output of LVLMs. Our main goal is to enhance the dependence of the output on the adversarial pattern, so that the learned perturbation can be more generalizable and contribute more attacker-chosen guidance effects than the benign pattern, ensuring that the attack remains effective when transferred to unknown models/prompts (more analysis is in Appendix H). As shown in Figure 2, empirical results prove that adversarial dependency plays an important role in enhancing the harmful effect of adversarial examples for improving transferability.

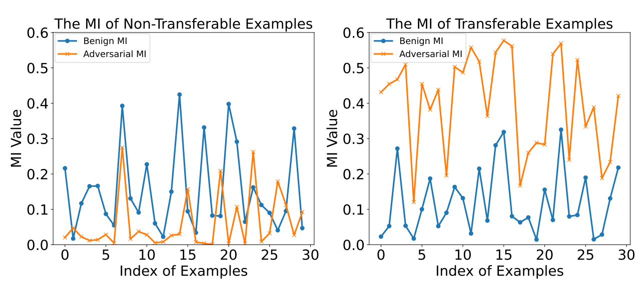

Figure 2: Benign/adversarial MI values of non-transferable and transferable adversarial examples of previous LVLM attacks. The adversarial MI values of transferable examples are shown to be generally larger than their benign MI values, demonstrating the correlation between adversarial dependency and transferability.

In particular, mutual information (MI) is an entropy-based information measurement tool that quantifies the dependency between two random variables. A higher MI indicates a stronger dependency between the variables. However, directly using the general MI calculation between adversarial examples and the corresponding outputs of LVLMs to measure the dependency presents a limitation. This is because adversarial images consist of both benign and adversarial patterns, and both of them have significant impacts on the output results. Directly maximizing the mixed MI between the adversarial image and the output may inadvertently increase the dependency of the output on the benign image, thereby hindering the increase in perturbation harmfulness. To address this issue, this paper makes an in-depth investigation on more fine-grained constraints of disentangled adversarial/benign MI information to achieve attacks.

Overall Pipeline. We present the overview of our proposed attack method in Figure 1 (c). Specifically, we first theoretically decompose the mixed MI information into benign MI (between the benign image and the LVLM output) and adversarial MI (between the adversarial perturbation and the LVLM output) components. Since the direct computation of MI is infeasible, we then train two MI estimators for separate MI calculations. Finally, we incorporate the benign and adversarial MI constraints into the optimization strategy to generate transferable adversarial examples.

3.3 How to Represent Adversarial/Benign MI?

We cannot simply utilize the separate benign image $\boldsymbol{x}_v$ and adversarial perturbation $\boldsymbol{\Delta}$ to calculate the two MI values with the LVLM output, as the adversarial impact is produced by the joint effects of their combination $\boldsymbol{x}_v^{adv}$. Therefore, to specifically represent the separate adversarial MI and benign MI, we need to disentangle them from the joint/mixed MI $I(\boldsymbol{x}_v^{adv}; L')$, where $L' = F(\boldsymbol{x}_v^{adv}, \boldsymbol{x}_p)$ is the logits of targeted output by the LVLM model.

Specifically, we define I_A as the adversarial MI between solely the additive adversarial perturbation and the LVLM's logit output, *i.e.*, $I_A(\boldsymbol{\Delta}, L')$ or $I_A(\boldsymbol{\Delta}, L)$, where $L = F(\boldsymbol{x}_v, \boldsymbol{x}_p)$ is the logits of benign output. Benign MI I_B is also defined between solely the benign image pattern and the logit output, *i.e.*, $I_B(\boldsymbol{x}_v, L')$ or $I_B(\boldsymbol{x}_v, L)$. We first provide Theorem 1 to illustrate the components and their relationship [27] within the mixed MI $I(\boldsymbol{x}_v^{adv}; L')$.

Theorem 1. Let $\boldsymbol{x}_v^{adv}$, $\boldsymbol{x}_v$, $\boldsymbol{\Delta}$, L' represent four random variables, then the mixed MI $I(\boldsymbol{x}_v^{adv}; L')$ has the following expression (proofed in the Appendix B):

$$I(\boldsymbol{x}_v^{adv}; L') = I(\boldsymbol{x}_v; L') + I(\boldsymbol{\Delta}; L') + H(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) - H(L'|\boldsymbol{x}_v^{adv}) - I(\boldsymbol{x}_v; \boldsymbol{\Delta}; L'), \qquad (2)$$

where $H(\cdot|\cdot)$ represents conditional entropy. In particular, $H(L'|\boldsymbol{x}_v, \boldsymbol{\Delta})$ and $H(L'|\boldsymbol{x}_v^{adv})$ can be formulated as:

$$\begin{aligned} H(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) &= -\sum_{L', \boldsymbol{x}_v, \boldsymbol{\Delta}} p(L', \boldsymbol{x}_v, \boldsymbol{\Delta}) \log p(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}), \\ H(L'|\boldsymbol{x}_v^{adv}) &= -\sum_{L', \boldsymbol{x}_v^{adv}} p(L', \boldsymbol{x}_v^{adv}) \log p(L'|\boldsymbol{x}_v^{adv}), \end{aligned} \qquad (3)$$

where $p(L', \boldsymbol{x}_v, \boldsymbol{\Delta}), p(L', \boldsymbol{x}_v^{adv})$ are the joint probability, $p(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}), p(L'|\boldsymbol{x}_v^{adv})$ are the conditional probability.

Assumption 1. $\boldsymbol{\Delta}, \boldsymbol{x}_v^{adv}$ are bijections of $\boldsymbol{x}_v$, *i.e.*, $\boldsymbol{\Delta}, \boldsymbol{x}_v^{adv}$ are dependently and uniquely determined by $\boldsymbol{x}_v$ and the decompositon of $\boldsymbol{x}_v^{adv}$ is also unique (the theoretical basis of this assumption is detailed in the Appendix C).

Based on this assumption, there exists $p(\boldsymbol{x}_v^{adv}) = p(\boldsymbol{x}_v, \boldsymbol{\Delta})$. Substituting this relation into Equation (3), we can obtain $H(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) \approx H(L'|\boldsymbol{x}_v^{adv})$ (proofed in the Appendix D,E). Besides, since the effects of benign and adversarial patterns on the output are mutually exclusive, thus $I(\boldsymbol{x}_v; \boldsymbol{\Delta}; L')$ is presented to be very small that can be ignored.

Therefore, according to the above derivations, now the mixed MI $I(\boldsymbol{x}_v^{adv}; L')$ can be linearly expressed as:

$$I(\boldsymbol{x}_v^{adv}; L') \approx I(\boldsymbol{x}_v; L') + I(\boldsymbol{\Delta}; L'). \qquad (4)$$

In this manner, we can approximately disentangle the mixed MI into the benign MI $I(\boldsymbol{x}_v; L')$ and the adversarial MI $I(\boldsymbol{\Delta}; L')$ and calculate them separately. These two MIs can not only reflect the dependency between the whole adversarial input and output like the mixed MI, but also provide independent measurements for different patterns.

3.4 How to Calculate Adversarial/Benign MI?

Although we can approximately represent the adversarial/benign MI following the aforementioned disentanglement, directly calculating MI is typically very challenging in high-dimensional spaces as it is a relative value. Luckily, many methods have been proposed to estimate MI [82, 29]. Among them, the Deep InfoMax (DIM) estimation method has been shown to be more effective [29]. Therefore, we adopt local DIM (details in Appendix F) and the Donsker-Varadhan representation [83] based on the KL divergence to estimate the adversarial/benign MI in our scenarios as:

$$I(X;Y) := D_{KL}(\mathbb{J}||\mathbb{M}) \geq \widehat{I}_{\omega,\psi}^{(DV)}(C_\psi(X);Y) := \mathbb{E}_{\mathbb{J}}[T_\omega(C_\psi(x),y)] - \log \mathbb{E}_{\mathbb{M}}[e^{T_\omega(C_\psi(x),y)}], \quad (5)$$

where X is a random variable, which can be the representation of any visual input $\{\boldsymbol{x}_v, \boldsymbol{\Delta}, \boldsymbol{x}_v^{adv}\}$ of LVLM. Y is also the random variable, which is the representation of the LVLM's output logits L or L'. $\mathbb{J}$ is the joint probability distribution of X and Y. $\mathbb{M}$ is the product of the marginal probability distributions of X and Y. We denote $\widehat{I}_{\omega,\psi}^{(DV)}$ as this MI estimation network based on the Donsker-Varadhan mechanism, which consists of two sub-networks C_ψ and T_ω. Specifically, C_ψ is an encoder composed of a neural network with parameters ψ, which maps the image input to a local feature map in the same latent space of Y. T_ω is a discriminator function modeled by a neural network with parameters ω to determine the relations between X and Y.

Therefore, we define two estimation networks $\widehat{I}_{\omega_A,\psi_A}^{(DV)}$ and $\widehat{I}_{\omega_B,\psi_B}^{(DV)}$ to calculate adversarial and benign MI, respectively. Due to the close relevance between the adversarial/benign patterns and the outputs of the separate perturbation/benign pattern, we utilize adversarial perturbations and benign images to train the $\widehat{I}_{\omega_A,\psi_A}^{(DV)}$ and $\widehat{I}_{\omega_B,\psi_B}^{(DV)}$. For network $\widehat{I}_{\omega_A,\psi_A}^{(DV)}$, we maximize the adversarial MI between the adversarial pattern of the perturbed image and the targeted output while minimizing the adversarial MI between the adversarial pattern of the perturbed image and the benign output. For network $\widehat{I}_{\omega_B,\psi_B}^{(DV)}$, we maximize the benign MI between the benign pattern of the perturbed image and the benign output while minimizing the benign MI between the benign pattern of the perturbed image and the targeted output. The optimization objectives are formulated as follows:

$$(\hat{\omega}_A, \hat{\psi}_A) = \arg \max_{\omega_A,\psi_A} [\widehat{I}_{\omega_A,\psi_A}^{(DV)}(C_{\psi_A}(\boldsymbol{\Delta}); L') - \widehat{I}_{\omega_A,\psi_A}^{(DV)}(C_{\psi_A}(\boldsymbol{\Delta}); L)], \quad (6)$$

$$(\hat{\omega}_B, \hat{\psi}_B) = \arg \max_{\omega_B,\psi_B} [\widehat{I}_{\omega_B,\psi_B}^{(DV)}(C_{\psi_B}(\boldsymbol{x}_v); L) - \widehat{I}_{\omega_B,\psi_B}^{(DV)}(C_{\psi_B}(\boldsymbol{x}_v); L')], \quad (7)$$

where $\widehat{I}_{\hat{\omega}_A,\hat{\psi}_A}^{(DV)}(\cdot)$, $\widehat{I}_{\hat{\omega}_B,\hat{\psi}_B}^{(DV)}(\cdot)$ are the estimated adversarial MI values and benign MI values.

3.5 Improving Transferability with Informative Constraints of Adversarial/Benign MI

To guide the image contents focusing more on the adversarial impacts of perturbations for improving the transferability, we develop an informative optimization strategy based on both adversarial and benign MI constraints to generate more harmful and generalizable adversarial samples. Specifically, by increasing the informative dependence between the adversarial perturbation of the input image and the adversarial output of the LVLM, while decreasing the dependence between the benign image pattern and the adversarial output of the LVLM, we can enhance the strength of the adversarial perturbation and ensure that this perturbation contributes more guidance for the attacker's choice compared to the benign image pattern. In this manner, the perturbation can always have more effect than the benign pattern, thus the adversarial example can still mislead the LVLM's reasoning when it is transferred to attack unknown models or prompts.

To achieve this goal, we utilized two MI evaluation networks trained by Section 3.4 to construct the optimization objective for generating adversarial examples as follows:

$$\arg \max_{||\boldsymbol{\Delta}||_p \leq \epsilon} [\widehat{I}_{\hat{\omega}_A,\hat{\psi}_A}^{(DV)}(C_{\hat{\psi}_A}(\boldsymbol{\Delta}); F(\boldsymbol{x}_v + \boldsymbol{\Delta}, \boldsymbol{x}_p)) - \widehat{I}_{\hat{\omega}_B,\hat{\psi}_B}^{(DV)}(C_{\hat{\psi}_B}(\boldsymbol{x_v}); F(\boldsymbol{x}_v + \boldsymbol{\Delta}, \boldsymbol{x}_p))], \quad (8)$$

where Formula 8 is recorded as l_{mi}. Besides, to better adjust the LVLM's output towards the target text $\boldsymbol{y}_{tar}$, we also utilize a cross-entropy $CE(\cdot)$loss to minimize the difference between the adversarial output and the target text. The cross-entropy loss l_{ce} is as follows:

$$l_{ce} = CE(F(\boldsymbol{x}_v + \boldsymbol{\Delta}, \boldsymbol{x}_p), \boldsymbol{y}_{tar}). \quad (9)$$

The overall loss for generating transferable adversarial examples is formulated as follows:

$$J = w_1 * l_{ce} - w_2 * l_{mi}, \quad (10)$$

where w_1, w_2 are weights to balance the loss. The algorithm of our attack is detailed in Appendix G.

Table 1: Performance comparisons on the adversarial transferability across different LVLM models. The experimental results are calculated by the averaged semantic similarities (↑) and attack success rates (↑) on three tasks. Target text: “I am sorry".

Dataset	Source Model	LVLM Attack	LLaVA-1.5			MiniGPT-4			BLIP-2			InstructBLIP		
			SS	*EM*	*CC*	*SS*	*EM*	*CC*	*SS*	*EM*	*CC*	*SS*	*EM*	*CC*
DALL-E	LLaVA-1.5	PGD [67]	*0.964*	*96.1*	*96.1*	0.042	0.0	0.0	0.094	0.3	0.3	0.137	1.2	1.6
		CroPA [8]	*0.819*	*79.7*	*79.7*	0.043	0.0	0.0	0.093	0.0	0.0	0.139	3.1	3.4
		UniAtt [10]	*0.842*	*80.6*	*88.3*	0.186	12.5	17.9	0.267	16.2	23.8	0.303	22.8	29.5
		Ours	*0.813*	*80.4*	*80.4*	**0.661**	**61.4**	**66.7**	**0.693**	**63.5**	**66.9**	**0.724**	**63.7**	**70.2**
	MiniGPT-4	PGD [67]	0.046	0.0	0.0	*0.823*	*79.7*	*79.7*	0.103	2.8	8.9	0.146	5.9	10.4
		CroPA [8]	0.051	0.0	0.0	*0.955*	*94.8*	*96.1*	0.125	9.0	11.1	0.166	3.4	3.4
		UniAtt [10]	0.298	15.4	22.7	*0.830*	*79.5*	*85.2*	0.338	23.1	27.7	0.316	16.2	16.2
		Ours	**0.650**	**57.3**	**64.9**	*0.860*	*84.2*	*84.5*	**0.716**	**63.4**	**68.5**	**0.698**	**58.3**	**64.1**
	BLIP-2	PGD [67]	0.056	0.0	0.0	0.053	0.0	0.0	*0.608*	*58.2*	*64.7*	0.164	3.5	11.3
		CroPA [8]	0.059	0.0	0.0	0.057	0.7	2.0	*0.610*	*28.8*	*93.5*	0.199	8.1	17.2
		UniAtt [10]	0.397	24.8	31.2	0.359	22.6	29.4	*0.817*	*77.9*	*85.1*	0.275	13.4	15.8
		Ours	**0.695**	**52.4**	**60.3**	**0.657**	**52.7**	**58.5**	*0.755*	*75.1*	*81.4*	**0.506**	**41.2**	**43.1**
	InstructBLIP	PGD [67]	0.048	0.0	0.0	0.047	0.0	0.0	0.128	1.0	1.9	*0.498*	*43.1*	*53.6*
		CroPA [8]	0.048	0.0	0.0	0.053	0.0	0.7	0.230	9.6	16.7	*0.842*	*81.7*	*82.3*
		UniAtt [10]	0.173	8.5	14.6	0.269	21.4	21.4	0.421	30.8	37.3	*0.854*	*79.6*	*83.7*
		Ours	**0.448**	**40.2**	**45.9**	**0.473**	**43.5**	**46.8**	**0.539**	**46.9**	**54.6**	*0.689*	*68.9*	*79.0*
SVIT	LLaVA-1.5	PGD [67]	*0.968*	*96.7*	*96.7*	0.040	0.0	0.0	0.097	0.8	1.2	0.132	2.4	3.7
		CroPA [8]	*0.762*	*61.3*	*66.4*	0.054	0.0	0.0	0.091	0.3	0.9	0.125	1.5	1.9
		UniAtt [10]	*0.856*	*83.9*	*87.8*	0.211	13.2	21.0	0.285	16.6	27.4	0.324	25.8	25.8
		Ours	*0.825*	*82.4*	*82.4*	**0.675**	**61.3**	**68.6**	**0.699**	**65.3**	**68.9**	**0.718**	**65.1**	**71.0**
	MiniGPT-4	PGD [67]	0.049	0.0	0.0	*0.882*	*86.9*	*87.6*	0.118	2.9	2.9	0.140	7.8	9.3
		CroPA [8]	0.047	0.0	0.0	*0.988*	*98.7*	*98.7*	0.191	13.1	13.1	0.162	4.6	4.6
		UniAtt [10]	0.315	16.8	24.2	*0.849*	*84.4*	*85.3*	0.352	24.7	30.3	0.428	30.0	32.3
		Ours	**0.647**	**62.3**	**66.0**	*0.904*	*89.2*	*89.5*	**0.745**	**66.9**	**73.1**	**0.785**	**72.4**	**72.4**
	BLIP-2	PGD [67]	0.049	0.0	0.0	0.056	0.0	0.9	*0.642*	*61.4*	*68.0*	0.172	6.4	10.8
		CroPA [8]	0.051	0.0	0.0	0.064	0.7	3.5	*0.608*	*28.8*	*94.8*	0.194	8.3	16.1
		UniAtt [10]	0.375	26.1	33.7	0.392	23.8	31.4	*0.797*	*78.3*	*83.5*	0.257	12.5	18.6
		Ours	**0.657**	**55.8**	**63.4**	**0.685**	**56.7**	**60.5**	*0.754*	*75.0*	*82.5*	**0.489**	**42.5**	**46.3**
	InstructBLIP	PGD [67]	0.046	0.0	0.0	0.037	0.0	0.0	0.133	4.0	6.5	*0.529*	*47.7*	*60.1*
		CroPA [8]	0.049	0.0	0.0	0.053	0.0	1.4	0.266	16.8	26.2	*0.859*	*83.7*	*85.0*
		UniAtt [10]	0.158	6.7	12.1	0.262	20.8	23.5	0.384	29.9	36.7	*0.836*	*77.9*	*84.5*
		Ours	**0.482**	**44.6**	**47.4**	**0.521**	**47.1**	**53.0**	**0.599**	**49.0**	**55.5**	*0.767*	*75.6*	*82.5*

4 Experiments

4.1 Implementation Details

LVLM Models. In this paper, following existing LVLM attack methods [8, 10], we conduct experiments on the same open-source LVLM models, including LLaVA-1.5 (integrated with Vicuna-7B) [84], MiniGPT-4 (integrated with Llama-2-7B-Chat) [85], BLIP-2 (integrated with OPT-2.7b) [5], and InstructBLIP (integrated with Vicuna-7B) [86], for comparison.

LVLM Datasets and Tasks. We evaluate the adversarial robustness of three multi-modal datasets for the image captioning, image classification, and VQA tasks. The datasets consist of both images and prompts. The images are collected from DALL-E [87], SVIT [88] and VQAv2 [89]. The prompts for three tasks derive from the CroPA [8].

Basic Setups. We consider two evaluation metrics: the semantic similarity (*SS*) utilizes the Sentence-Transformer [90] to generate embeddings of both adversarial output and target text for calculating their cosine similarity, and the success rates “ExactMatch" (*EM*) and “ConditionalContain" (*CC*) to assess the word-level overlap between adversarial output and target text.

We utilize the same architectures to initialize the adversarial MI and benign MI estimation networks, but they are trained separately. Specifically, C_ψ is implemented as a light two-layer convolutional neural network, while T_ω simply incorporates an attention mechanism, 1×1 convolutional blocks, and residual connections. For training, we first use the selected adversarial examples generated by PGD [67] attack with $\epsilon = 16/255$. We then feed the same prompt with benign image $\boldsymbol{x}_v$ and

Table 2: Performance comparisons on the adversarial transferability across different numbers of prompts. The experimental results are calculated by the averaged semantic similarities (↑) on the target text: “I am sorry". The prompts are randomly sampled from three tasks.

Dataset	LVLM Attack	LLaVA-1.5			MiniGPT-4			BLIP-2			InstructBLIP		
		Num=20	*Num=40*	*Num=60*	*Num=20*	*Num=40*	*Num=60*	*Num=20*	*Num=40*	*Num=60*	*Num=20*	*Num=40*	*Num=60*
DALL-E	PGD [67]	0.435	0.431	0.421	0.692	0.693	0.689	0.578	0.517	0.523	0.297	0.295	0.289
	CroPA [8]	0.718	0.720	0.711	0.821	0.813	0.809	0.536	0.535	0.525	0.623	0.598	0.604
	UniAtt [10]	0.732	0.716	0.714	0.776	0.782	0.780	0.627	0.609	0.602	0.665	0.659	0.659
	Ours	**0.743**	**0.736**	**0.730**	**0.825**	**0.828**	**0.820**	**0.693**	**0.689**	**0.682**	**0.672**	**0.664**	**0.663**
SVIT	PGD [67]	0.416	0.408	0.405	0.702	0.699	0.680	0.556	0.545	0.541	0.317	0.307	0.310
	CroPA [8]	0.723	0.711	0.706	0.824	0.825	0.822	0.533	0.523	0.521	0.623	0.622	0.616
	UniAtt [10]	0.729	0.727	0.724	0.801	0.797	0.793	0.608	0.584	0.573	0.694	0.678	0.671
	Ours	**0.754**	**0.738**	**0.734**	**0.835**	**0.836**	**0.830**	**0.655**	**0.653**	**0.650**	**0.735**	**0.729**	**0.732**
VQAv2	PGD [67]	0.422	0.414	0.411	0.756	0.758	0.755	0.514	0.502	0.500	0.325	0.316	0.318
	CroPA [8]	0.732	0.726	0.729	0.847	0.843	0.836	0.530	0.524	0.516	0.648	0.630	0.625
	UniAtt [10]	0.728	0.723	0.719	0.842	0.839	0.827	0.591	0.587	0.575	0.687	0.672	0.673
	Ours	**0.767**	**0.761**	**0.764**	**0.866**	**0.861**	**0.857**	**0.613**	**0.607**	**0.612**	**0.727**	**0.722**	**0.716**

adversarial image $\boldsymbol{x}_v^{adv}$ into the LVLM to obtain the corresponding logits outputs L and $L^{'}$. The tuple $(\boldsymbol{x}_v^{adv} - \boldsymbol{x}_v, L, L^{'})$ is used to train the adversarial MI estimation network following Equation 6, while the tuple $(\boldsymbol{x}_v, L, L^{'})$ is used as input to train the benign MI estimation network following Equation 7. We train both networks using the Adam optimizer for 100 epochs, with an initial learning rate of 0.01 that decays by a factor of 0.5 every 20 epochs. The number of channels of the encoded image feature map is 2048. To generate transferable examples, the perturbation budget ϵ is also set to $16/255$. The epoch number is set to 1000. The momentum parameter μ is set to 0.9 and the step size is set as $\alpha = 16/epoch$. Besides, the weights $w_1 = w_2 = 1$. All LVLM attack baselines are re-implemented in the same setting for experimental comparison. All experiments are conducted on the NVIDIA RTX 4090 GPUs with 24GB of memory.

4.2 Main Results

Transfer-Attack Performance across LVLMs. To investigate the transferability of our proposed attack, we first provide the performance across different LVLM models in Table 1. Here, we select the target text “I am sorry", and all the performances are averagely evaluated on three tasks. We can find that: (1) Our generated adversarial examples have competitive harmfulness compared to existing attacks in the diagonal values. This demonstrates that our attack also contributes to improve the harmful impact of the samples. (2) Our attacks achieve significant transfer-attack performance compared to previous works, demonstrating the effectiveness of our designed informative constraints. Furthermore, we transfer the adversarial examples generated on MiniGPT-4 model to realistic LVLM applications GPT-4o (GPT-4o-0513) [91] and Claude-3.5-Sonnet [92]. As shown in Table 3, our attack still achieves better performance.

Transfer-Attack Performance across Prompts. We then investigate the transfer-attack performance across different numbers of prompts in Table 2. Here, we directly transfer the adversarial examples generated by a certain LVLM model and prompt to the same LVLM model with different prompts. We can find that previous attacks achieve worse performance with the increase of the prompt numbers. Instead, our method achieves better attack performance across different prompts, demonstrating the effectiveness of our developed informative constraints.

Table 3: Transfer-attack performance on the generated adversarial examples from MiniGPT-4 to GPT-4o and Claude-3.5.

LVLM Attack	GPT-4o			Claude-3.5		
	SS	*EM*	*CC*	*SS*	*EM*	*CC*
PGD [67]	0.036	0.0	0.0	0.048	0.0	0.0
CroPA [8]	0.057	0.0	0.0	0.052	0.0	0.0
UniAtt [10]	0.142	4.1	7.5	0.169	3.3	10.4
Ours	**0.608**	**44.7**	**51.3**	**0.620**	**48.6**	**56.5**

Joint transferability across models and prompts. We further evaluate the joint transferability of our proposed attack across both different models and prompts at the same time. As shown in Figure 5, the results still demonstrate that our method retains strong transferability even under this more challenging setting.

We also provide experiments on more datasets and architecturally distinct LVLMs in Appendix I.1, I.2. Justification of our transfer attack are in Appendix I.3. To verify generality, we also evaluate under a universal setting and on jailbreak/rewiring attacks, see Appendix I.6 and I.7.

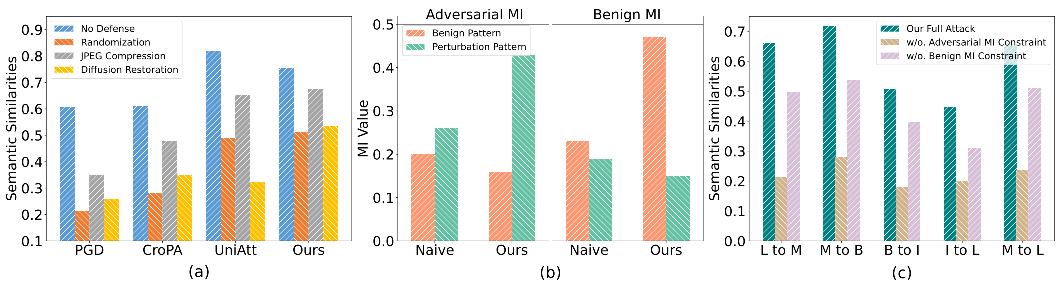

Figure 3: (a) Adversarial robustness against various defenses on BLIP-2 model with DALL-E dataset. (b) Effectiveness of MI estimation networks on BLIP-2 model with DALL-E dataset. (c) Ablation on different MI components on DALL-E dataset (LLaVA-1.5: L, MiniGPT-4: M, InstructBLIP: I).

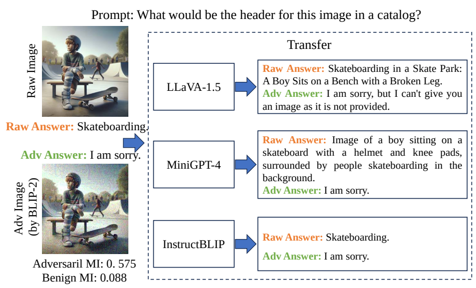

Figure 4: Visualizations of our transfer attack.

Figure 5: The joint transferability performance across diverse models and prompts. The experimental results are calculated by the averaged semantic similarities (↑) on three tasks.

Setting	CroPA	UniAtt	Ours
LLaVA-1.5 to MiniGPT-4 (num=20)	0.042	0.115	0.627
LLaVA-1.5 to MiniGPT-4 (num=60)	0.039	0.101	0.608
MiniGPT-4 to LLaVA-1.5 (num=20)	0.050	0.189	0.596
MiniGPT-4 to LLaVA-1.5 (num=60)	0.047	0.168	0.594

4.3 Attack Efficiency and Robustness

Complexity Analysis. To investigate the scalability and practicality of our transfer-attack method, we provide the complexity analysis in Table 4, which evaluates the usage of GPU time and memory of a single adversarial sample generation on LLaVA-1.5 model. It indicates that our attack costs relatively fewer GPU resources, as our informative constraints are easily achieved with solely loss designs, while our samples can achieve better transfer-attack performance within a single generation process.

Robustness to Defenses. To evaluate the robustness of our attack against potential defense strategies, we conduct experiments on three pre-processing defense methods, *i.e*, Randomization [93, 94], JPEG Compression [95], and Diffusion Restoration [96] in Figure 3 (a). Compared to previous attacks, our attack is relatively more robust to potential defenses because we explicitly constrain the adversarial perturbation to be as harmful as possible. This allows it to provide more guidance to the LVLM's reasoning than the benign pattern, having more opportunities to lead to wrong results. More defense experiments can be found in Appendix I.4.

Table 4: Complexity comparison on adversarial sample generation.

LVLM Attack	GPU Time (↓)	GPU Memory (↓)
PGD [67]	4 min	16.7 GB
CroPA [8]	12 min	20.4 GB
UniAtt [10]	294 min	57.5 GB
Ours	9 min	18.2 GB

4.4 Effectiveness of MI Estimation Networks

During the training process of each MI estimation network, directly maximizing positive MI without minimizing negative MI may not clearly learn the accurate effect for adversarial perturbation pattern or benign pattern (*i.e.*, solely maximizing $(\hat{\omega}_A, \hat{\psi}_A) = \arg\max_{\omega_A, \psi_A} \widehat{I}^{(DV)}_{\omega_A, \psi_A}(C_{\psi_A}(\boldsymbol{\Delta}); L')$ or $(\hat{\omega}_B, \hat{\psi}_B) = \arg\max_{\omega_B, \psi_B} \widehat{I}^{(DV)}_{\omega_B, \psi_B}(C_{\psi_B}(\boldsymbol{x}_v); L)$). Therefore, we design the joint maximization-minimization optimization mechanism to train each MI estimation network via Equation (6) (7). To demonstrate its effectiveness, we compare the MI estimation performance of these two training strategies and compute the average MI value for all samples as shown in Figure 3 (b). The results indicate that our optimization mechanism helps to better capture the inherent differences between adversarial perturbation patterns and benign patterns in terms of adversarial MI and benign MI.

Table 5: The transfer-attack performance on complex target texts using adversarial examples generated on LLaVA-1.5 with DALL-E dataset.

Target Text	LVLM Attack	LLaVA-1.5			MiniGPT-4			BLIP-2		
		SS	*EM*	*CC*	*SS*	*EM*	*CC*	*SS*	*EM*	*CC*
"A man holding a big doughnut at a festival."	PGD [67]	0.916	91.2	91.2	0.054	0.0	0.0	0.087	0.5	0.5
	CroPA [8]	0.782	76.9	76.9	0.040	0.0	0.0	0.087	0.0	0.0
	UniAtt [10]	0.821	80.1	85.5	0.204	17.8	21.7	0.291	18.9	28.3
	Ours	0.796	79.9	80.2	**0.662**	**60.5**	**64.8**	**0.714**	**62.4**	**68.6**
"A photo of a teddy bear on a skateboard in Times Square."	PGD [67]	0.903	89.8	91.2	0.048	0.0	0.0	0.096	0.9	1.1
	CroPA [8]	0.786	78.2	78.2	0.056	0.0	0.0	0.099	0.0	0.4
	UniAtt [10]	0.834	82.7	85.9	0.216	18.6	21.5	0.285	19.6	25.4
	Ours	0.779	79.8	79.8	**0.671**	**61.2**	**66.9**	**0.695**	**62.7**	**67.7**
"A beautiful bird with a black and white color in snow."	PGD [67]	0.907	90.1	90.1	0.065	0.0	0.0	0.113	0.0	0.4
	CroPA [8]	0.814	80.6	80.6	0.052	0.0	0.0	0.091	0.4	0.8
	UniAtt [10]	0.843	81.6	86.3	0.229	18.3	23.1	0.287	20.7	29.6
	Ours	0.815	80.9	82.2	**0.676**	**61.5**	**68.2**	**0.729**	**60.9**	**68.4**
"Bunk bed with a narrow shelf sitting underneath it."	PGD [67]	0.876	88.3	88.3	0.067	0.0	0.0	0.094	0.3	0.3
	CroPA [8]	0.785	77.2	79.6	0.053	0.0	0.0	0.104	0.2	0.5
	UniAtt [10]	0.806	78.9	82.6	0.214	16.7	22.3	0.274	20.1	26.8
	Ours	0.799	79.5	81.2	**0.621**	**57.4**	**63.5**	**0.645**	**58.4**	**65.0**
"The people are gathered at the table for dinner."	PGD [67]	0.929	92.4	92.4	0.056	0.0	0.0	0.105	0.5	0.8
	CroPA [8]	0.789	78.2	79.5	0.046	0.0	0.0	0.088	0.0	0.3
	UniAtt [10]	0.831	82.0	85.9	0.219	18.4	20.8	0.293	22.4	30.5
	Ours	0.816	81.3	81.3	**0.631**	**56.9**	**62.1**	**0.658**	**59.4**	**66.1**

4.5 Visualization

We provide the visualization example of our transfer-attack across different LVLMs in Figure 4, where our generated adversarial examples can achieve the same harmful effect when transferred across different LVLMs, demonstrating the transferability of our attack. More visualization results, including MI values for transferable and non-transferable adversarial examples, as well as a comparison of transfer-attack performance across various attack methods, can be found in Appendix I.5.

4.6 Ablation Study

Ablation on Different Target Texts. To demonstrate that the effectiveness of our attack is not constrained to the specific case of the target text “I am sorry", we extend our evaluation to more complex and sophisticated target texts. As shown in Table 5, despite the increased difficulty and complexity of these target texts, our attack strategy still demonstrates significant effectiveness and consistently maintains high performance in transfer attacks.

Ablation on Different MI Components. To elucidate the role of each component of our method in improving transferability, we conduct ablation studies: (1) removing the adversarial MI constraint, and (2) removing the benign MI constraint. As shown in Figure 3 (c), the results demonstrate that each component of our method contributes positively to improving transfer-attack performance.

5 Conclusion

This paper proposes a powerful LVLM attack method that is transferable across different LVLM models and prompts. We introduce a new perspective of information theory to investigate LVLMs' transferable characteristics by exploring the relative dependence between outputs of the LVLM and input adversarial samples. With appropriate informative constraints between the disentangled adversarial/benign patterns of the image input and output text, our generated adversarial examples are proven to be more generalizable and harmful to unseen LVLMs and prompts. Extensive experiments indicate the effectiveness of our proposed attack.

Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC) under grant No.62476107.

References

[1] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[2] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[3] 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.

[4] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

[5] 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.

[6] 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.

[7] 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.

[8] Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[9] 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, 2024.

[10] Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Xiang Fang, Keke Tang, Yao Wan, and Lichao Sun. Pandora's box: Towards building universal attackers against real-world large vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[11] Xiaowen Cai, Daizong Liu, Runwei Guan, and Pan Zhou. Imperceptible transfer attack on large vision-language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[12] Daizong Liu, Xiaowen Cai, Pan Zhou, Xiaoye Qu, Xiang Fang, Lichao Sun, and Wei Hu. Are large vision-language models robust to adversarial visual transformations? *OpenReview*, 2024.

[13] Daizong Liu and Wei Hu. Can't see the wood for the trees: Can visual adversarial patches fool hard-label large vision-language models? *OpenReview*, 2024.

[14] Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. Image hijacks: Adversarial images can control generative models at runtime. *arXiv preprint arXiv:2309.00236*, 2023.

[15] 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.

[16] 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.

[17] Zefeng Wang, Zhen Han, Shuo Chen, Fan Xue, Zifeng Ding, Xun Xiao, Volker Tresp, Philip Torr, and Jindong Gu. Stop reasoning! when multimodal llms with chain-of-thought reasoning meets adversarial images. *arXiv preprint arXiv:2402.14899*, 2024.

[18] Hao Zhang, Wenqi Shao, Hong Liu, Yongqiang Ma, Ping Luo, Yu Qiao, and Kaipeng Zhang. Avibench: Towards evaluating the robustness of large vision-language model on adversarial visual-instructions. *arXiv preprint arXiv:2403.09346*, 2024.

[19] Dong Lu, Tianyu Pang, Chao Du, Qian Liu, Xianjun Yang, and Min Lin. Test-time backdoor attacks on multimodal large language models. *arXiv preprint arXiv:2402.08577*, 2024.

[20] Xijia Tao, Shuai Zhong, Lei Li, Qi Liu, and Lingpeng Kong. Imgtrojan: Jailbreaking vision-language models with one image. *arXiv preprint arXiv:2403.02910*, 2024.

[21] Daizong Liu and Wei Hu. Imperceptible transfer attack and defense on 3d point cloud classification. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4727–4746, 2022.

[22] Weibin Wu, Yuxin Su, Michael R Lyu, and Irwin King. Improving the transferability of adversarial samples with adversarial transformations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9024–9033, 2021.

[23] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019.

[24] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4312–4321, 2019.

[25] Young-Il Moon, Balaji Rajagopalan, and Upmanu Lall. Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3):2318, 1995.

[26] Georges A Darbellay and Igor Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, 1999.

[27] Dawei Zhou, Nannan Wang, Xinbo Gao, Bo Han, Xiaoyu Wang, Yibing Zhan, and Tongliang Liu. Improving adversarial robustness via mutual information estimation. In *International conference on machine learning*, pages 27338–27352. PMLR, 2022.

[28] 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.

[29] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[30] Sicheng Zhu, Xiao Zhang, and David Evans. Learning adversarially robust representations via worst-case mutual information maximization. In *International Conference on Machine Learning*, pages 11609–11618. PMLR, 2020.

[31] Dan Guo, Kun Li, Bin Hu, Yan Zhang, and Meng Wang. Benchmarking micro-action recognition: Dataset, method, and application. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6238–6252, 2024.

[32] Xianke Chen, Daizong Liu, Xun Yang, Xirong Li, Jianfeng Dong, Meng Wang, and Xu Wang. Prvr: Partially relevant video retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[33] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[34] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11235–11244, 2021.

[35] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4070–4078, 2020.

[36] Daizong Liu, Xi Ouyang, Shuangjie Xu, Pan Zhou, Kun He, and Shiping Wen. Saanet: Siamese action-units attention network for improving dynamic facial expression recognition. *Neurocomputing*, 413:145–157, 2020.

[37] Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu, and Pan Zhou. Memory-guided semantic learning network for temporal sentence grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1665–1673, 2022.

[38] Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. Adaptive proposal generation network for temporal sentence localization in videos. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9292–9301, 2021.

[39] Daizong Liu, Xiaoye Qu, Yinzhen Wang, Xing Di, Kai Zou, Yu Cheng, Zichuan Xu, and Pan Zhou. Unsupervised temporal video grounding with deep semantic clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1683–1691, 2022.

[40] Xiang Fang, Daizong Liu, Pan Zhou, and Guoshun Nan. You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2448–2460, 2023.

[41] Daizong Liu, Xiaoye Qu, and Wei Hu. Reducing the vision and language bias for temporal sentence grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4092–4101, 2022.

[42] Daizong Liu, Xiang Fang, Pan Zhou, Xing Di, Weining Lu, and Yu Cheng. Hypotheses tree building for one-shot temporal sentence localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1640–1648, 2023.

[43] Jiahao Zhu, Daizong Liu, Pan Zhou, Xing Di, Yu Cheng, Song Yang, Wenzheng Xu, Zichuan Xu, Yao Wan, Lichao Sun, et al. Rethinking the video sampling and reasoning strategies for temporal sentence grounding. *arXiv preprint arXiv:2301.00514*, 2023.

[44] Daizong Liu, Shuangjie Xu, Xiao-Yang Liu, Zichuan Xu, Wei Wei, and Pan Zhou. Spatiotemporal graph neural network based mask reconstruction for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2100–2108, 2021.

[45] Daizong Liu, Pan Zhou, Zichuan Xu, Haozhao Wang, and Ruixuan Li. Few-shot temporal sentence grounding via memory-guided semantic learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(5):2491–2505, 2022.

[46] Daizong Liu, Yang Liu, Wencan Huang, and Wei Hu. A survey on text-guided 3d visual grounding: Elements, recent advances, and future directions. *arXiv preprint arXiv:2406.05785*, 2024.

[47] Junhao Dong, Piotr Koniusz, Xinghua Qu, and Yew-Soon Ong. Stabilizing modality gap & lowering gradient norms improve zero-shot adversarial robustness of vlms. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 236–247, 2025.

[48] Junhao Dong, Piotr Koniusz, Yifei Zhang, Hao Zhu, Weiming Liu, Xinghua Qu, and Yew-Soon Ong. Improving zero-shot adversarial robustness in vision-language models by closed-form alignment of adversarial path simplices. In *Forty-second International Conference on Machine Learning*, 2025.

[49] Junhao Dong, Piotr Koniusz, Junxi Chen, Z Jane Wang, and Yew-Soon Ong. Robust distillation via untargeted and targeted intermediate adversarial samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28432–28442, 2024.

[50] Junhao Dong, Piotr Koniusz, Junxi Chen, Xiaohua Xie, and Yew-Soon Ong. Adversarially robust few-shot learning via parameter co-distillation of similarity and class concept learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28535–28544, 2024.

[51] Junhao Dong, Piotr Koniusz, Junxi Chen, and Yew-Soon Ong. Adversarially robust distillation by reducing the student-teacher variance gap. In *European Conference on Computer Vision*, pages 92–111. Springer, 2024.

[52] Junhao Dong, Seyed-Mohsen Moosavi-Dezfooli, Jianhuang Lai, and Xiaohua Xie. The enemy of my enemy is my friend: Exploring inverse adversaries for improving adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24678–24687, 2023.

[53] Junhao Dong, Yuan Wang, Jian-Huang Lai, and Xiaohua Xie. Improving adversarially robust few-shot image classification with generalizable representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9025–9034, 2022.

[54] Junhao Dong, Yuan Wang, Jianhuang Lai, and Xiaohua Xie. Restricted black-box adversarial attack against deepfake face swapping. *IEEE Transactions on Information Forensics and Security*, 18:2596–2608, 2023.

[55] Junhao Dong, Junxi Chen, Xiaohua Xie, Jianhuang Lai, and Hao Chen. Survey on adversarial attack and defense for medical image analysis: Methods and challenges. *ACM Computing Surveys*, 57(3):1–38, 2024.

[56] Fuyao Cai, Daizong Liu, Xiang Fang, Jixiang Yu, Keke Tang, and Pan Zhou. Imperceptible beam-sensitive adversarial attacks for lidar-based object detection in autonomous driving. In *IEEE International Conference on Multimedia & Expo 2025 (ICME 2025)*, 2025.

[57] Daizong Liu and Wei Hu. Imperceptible backdoor attacks on text-guided 3d scene grounding. *IEEE Transactions on Multimedia*, 2025.

[58] Xiaowen Cai, Yunbo Tao, Daizong Liu, Pan Zhou, Xiaoye Qu, Jianfeng Dong, Keke Tang, and Lichao Sun. Frequency-aware gan for imperceptible transfer attack on 3d point clouds. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6162–6171, 2024.

[59] Mingyu Yang, Daizong Liu, Keke Tang, Pan Zhou, Lixing Chen, and Junyang Chen. Hiding imperceptible noise in curvature-aware patches for 3d point cloud attack. In *European Conference on Computer Vision*, pages 431–448. Springer, 2024.

[60] Daizong Liu and Wei Hu. Explicitly perceiving and preserving the local geometric structures for 3d point cloud attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3576–3584, 2024.

[61] Daizong Liu, Wei Hu, and Xin Li. Robust geometry-dependent attack for 3d point clouds. *IEEE Transactions on Multimedia*, 26:2866–2877, 2023.

[62] Yunbo Tao, Daizong Liu, Pan Zhou, Yulai Xie, Wei Du, and Wei Hu. 3dhacker: Spectrum-based decision boundary generation for hard-label 3d point cloud attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14340–14350, 2023.

[63] Qianjiang Hu, Daizong Liu, and Wei Hu. Exploring the devil in graph spectral domain for 3d point cloud attacks. In *European Conference on Computer Vision*, pages 229–248. Springer, 2022.

[64] Daizong Liu, Wei Hu, and Xin Li. Point cloud attacks in graph spectral domain: When 3d geometry meets graph signal processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[65] Daizong Liu and Wei Hu. Seeing is not believing: Adversarial natural object optimization for hard-label 3d scene attacks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11886–11897, 2025.

[66] Xiaohan Fu, Zihan Wang, Shuheng Li, Rajesh K Gupta, Niloofar Mireshghallah, Taylor Berg-Kirkpatrick, and Earlence Fernandes. Misusing tools in large language models with visual adversarial examples. *arXiv preprint arXiv:2310.03185*, 2023.

[67] Xuanimng 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*, 2024.

[68] 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.

[69] Kuofeng Gao, Yang Bai, Jindong Gu, Shu-Tao Xia, Philip Torr, Zhifeng Li, and Wei Liu. Inducing high energy-latency of large vision-language models with verbose images. *arXiv preprint arXiv:2401.11170*, 2024.

[70] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*, 2019.

[71] Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R Lyu. Improving adversarial transferability via neuron attribution-based attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14993–15002, 2022.

[72] Jianping Zhang, Jen-tse Huang, Wenxuan Wang, Yichen Li, Weibin Wu, Xiaosen Wang, Yuxin Su, and Michael R Lyu. Improving the transferability of adversarial samples by path-augmented method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8173–8182, 2023.

[73] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[74] Zhijin Ge, Hongying Liu, Wang Xiaosen, Fanhua Shang, and Yuanyuan Liu. Boosting adversarial transferability by achieving flat local maxima. *Advances in Neural Information Processing Systems*, 36:70141–70161, 2023.

[75] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1924–1933, 2021.

[76] Xiaosen Wang, Jiadong Lin, Han Hu, Jingdong Wang, and Kun He. Boosting adversarial transferability through enhanced momentum. *arXiv preprint arXiv:2103.10609*, 2021.

[77] Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan Yuille. Learning transferable adversarial examples via ghost networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11458–11465, 2020.

[78] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[79] Yifeng Xiong, Jiadong Lin, Min Zhang, John E Hopcroft, and Kun He. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14983–14992, 2022.

[80] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. Boosting the transferability of adversarial samples via attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1161–1170, 2020.

[81] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–467, 2018.

[82] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

[83] S.R.S Varadhan M.D Donsker. Asymptotic evaluation of certain markov process expectations for large time. *Communications on Pure and Applied Mathematics, 36(2):183–212*, 1983.

[84] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[85] 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.

[86] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat 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, 2024.

[87] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[88] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023.

[89] 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.

[90] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[91] OpenAI. Hello gpt-4o. *https://openai.com/index/hello-gpt4o/*, 2024.

[92] Anthropic. The claude 3 model family: Opus, sonnet, haiku. *https://claude.ai/*, 2024.

[93] Iuri Frosio and Jan Kautz. The best defense is a good offense: adversarial augmentation against adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4067–4076, 2023.

[94] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017.

[95] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

[96] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

[97] Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025.

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

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

[100] Gemma Team. Gemma 3 technical report. *arXiv*, abs/2503.19786, 2025.

[101] Rylan Schaeffer, Dan Valentine, Luke Bailey, James Chua, Cristobal Eyzaguirre, Zane Durante, Joe Benton, Brando Miranda, Henry Sleight, Tony Tong Wang, et al. Failures to find transferable image jailbreaks between vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2024.

[102] 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. *ICML*, 2024.

[103] Qi Zhou, Tianlin Li, Qing Guo, Dongxia Wang, Yun Lin, Yang Liu, and Jin Song Dong. Defending lvlms against vision attacks through partial-perception supervision. *ICML*, 2025.

[104] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21527–21536, 2024.

[105] Ruofan Wang, Xingjun Ma, Hanxu Zhou, Chuanjun Ji, Guangnan Ye, and Yu-Gang Jiang. White-box multimodal jailbreaks against large vision-language models. *arXiv preprint arXiv:2405.17894*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In this paper, we propose a novel LVLM attack method to adversarially constrain the informative dependence between the benign/adversarial pattern of the input and the LVLM's output for improving the model/prompt-aware transferability.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the potential limitations in the Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The theoretical results in this paper has a clear assumption and proof. The assumption is explained in detail and the proof is provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have provided them with the implementation details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will release the codes upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper have provided comprehensive experimental details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: NA.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided detailed information in the Experiments Section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed them in the Appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: NA.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: NA.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: NA.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: NA.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: NA.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Preliminary of Mutual Information

Mutual Information (MI) is a measure of the dependency between two random variables, indicating the amount of information one variable contains about the other. The larger the value, the stronger the relationship between the two variables. If the value is zero, it means the two variables are independent. The formula for MI is:

$$I(X;Y) = \int_Y \int_X p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy, \tag{11}$$

where X is a random variable, which is the representation of $\boldsymbol{x}_v$, $\boldsymbol{\Delta}$ or $\boldsymbol{x}_v^{adv}$. Y is a random variable, which is the representation of L or L'. $p(x,y)$ is the joint probability density function of (X,Y), and $p(x), p(y)$ are the marginal probability density functions of X and Y, respectively.

MI can also be expressed in terms of entropy. Its mathematical definition is as follows:

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X), \end{aligned} \tag{12}$$

where $H(X), H(Y)$ are the entropy of the X and Y, representing the uncertainty of X and Y. $H(X,Y), H(Y|X)$ are the joint entropy and the conditional entropy, respectively.

The relationship between MI of three variables X, Y, Z is defined as [97]:

$$I(X;Y;Z) = I(X;Z) - I(X;Z|Y). \tag{13}$$

B Proof of Theorem 1

Theorem 1. Let $\boldsymbol{x}_v^{adv}$, $\boldsymbol{x}_v$, $\boldsymbol{\Delta}$, L' represent four random variables, then the mixed MI $I(\boldsymbol{x}_v^{adv}; L')$ has the following expression:

$$\begin{aligned} I(\boldsymbol{x}_v^{adv}; L') = I(\boldsymbol{x}_v; L') + I(\boldsymbol{\Delta}; L') + H(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) \\ - H(L'|\boldsymbol{x}_v^{adv}) - I(\boldsymbol{x}_v; \boldsymbol{\Delta}; L'). \end{aligned} \tag{14}$$

Proof. According to the relationship between information entropy and mutual information, we have:

$$\begin{aligned} I(\boldsymbol{x}_v; L') + I(\boldsymbol{\Delta}; L') &= H(L') - H(L'|\boldsymbol{x}_v) + H(L') - H(L'|\boldsymbol{\Delta}) \\ &= 2H(L') - [H(L'|\boldsymbol{x}_v) + H(L'|\boldsymbol{\Delta})]. \end{aligned} \tag{15}$$

According to the theorem of conditional mutual information in probability theory, we have:

$$\begin{aligned} H(L'|\boldsymbol{x}_v) + H(L'|\boldsymbol{\Delta}) &= [H(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) + I(\boldsymbol{\Delta}; L'|\boldsymbol{x}_v)] + [H(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) + I(\boldsymbol{x}_v; L'|\boldsymbol{\Delta})] \\ &= [H(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) + I(\boldsymbol{\Delta}; L'|\boldsymbol{x}_v) + I(\boldsymbol{x}_v; L'|\boldsymbol{\Delta}) + I(\boldsymbol{x}_v; \boldsymbol{\Delta}; L')] \\ &\quad + H(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) - I(\boldsymbol{x}_v; \boldsymbol{\Delta}; L') \\ &= H(L') + H(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) - I(\boldsymbol{x}_v; \boldsymbol{\Delta}; L'). \end{aligned}$$

By combining the above two equations into a joint one, we have:

$$\begin{aligned} I(\boldsymbol{x}_v; L') + I(\boldsymbol{\Delta}; L') &= 2H(L') - [H(L') + H(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) - I(\boldsymbol{x}_v; \boldsymbol{\Delta}; L')] \\ &= H(L') - H(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) + I(\boldsymbol{x}_v; \boldsymbol{\Delta}; L') \\ &= I(\boldsymbol{x}_v^{adv}, L') + H(L'|\boldsymbol{x}_v^{adv}) - H(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) + I(\boldsymbol{x}_v; \boldsymbol{\Delta}; L'). \end{aligned} \tag{16}$$

Finally, we have:

$$I(\boldsymbol{x}_v; L') + I(\boldsymbol{\Delta}; L') + H(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) - H(L'|\boldsymbol{x}_v^{adv}) - I(\boldsymbol{x}_v; \boldsymbol{\Delta}; L') = I(\boldsymbol{x}_v^{adv}; L'), \tag{17}$$

which completes the proof.

C Theoretical Basis of Assumption 1

Theorem 1. $\boldsymbol{\Delta}, \boldsymbol{x}_v^{adv}$ are bijections of $\boldsymbol{x}_v$, *i.e.*, $\boldsymbol{\Delta}, \boldsymbol{x}_v^{adv}$ are dependently and uniquely determined by $\boldsymbol{x}_v$ and the decompositon of $\boldsymbol{x}_v^{adv}$ is also unique.

This assumption is based on: each $\boldsymbol{x}_v$ generates a unique perturbation $\boldsymbol{\Delta}$ determined by the attack algorithm like PGD, where prompt-agnostic perturbation $\boldsymbol{\Delta}$ is computed by visual-solely contexts $\boldsymbol{\Delta} = \epsilon \cdot \text{sign}(\nabla_{\boldsymbol{x}_v} L(f(\boldsymbol{x}_v), y_{tar}))$. Therefore, $\boldsymbol{\Delta}$ can be taken as a function of $\boldsymbol{x}_v$, denoted as $\boldsymbol{\Delta} = g(\boldsymbol{x}_v)$ that is uniquely determined by $\boldsymbol{x}_v$. This leads to a unique adversarial image $\boldsymbol{x}_v^{adv} = \boldsymbol{x}_v + \boldsymbol{\Delta} = \boldsymbol{x}_v + g(\boldsymbol{x}_v) = h(\boldsymbol{x}_v)$, where $\boldsymbol{x}_v^{adv}$ can also be taken as a function of $\boldsymbol{x}_v$. Conversely, although $\boldsymbol{x}_v^{adv}$ may exist other decompositions in a purely mathematical sense, in the adversarial generation context, the uniqueness of this decomposition can be guaranteed by the above attack protocol. Hence, we can assumpt that $\boldsymbol{\Delta}, \boldsymbol{x}_v^{adv}$ are bijections of $\boldsymbol{x}_v$.

D Proof of Equal Probabilities

We assert that each $\boldsymbol{x}_v, \boldsymbol{\Delta}, \boldsymbol{x}_v^{adv}$ is selected from a finite, countable set, making the selection process discrete. Therefore, we utilize the probability mass function (PMF) in our specific case. For discrete variables $\boldsymbol{x}_v, \boldsymbol{\Delta}, \boldsymbol{x}_v^{adv}$, the PMF of $\boldsymbol{x}_v$ can denote as $p_{\boldsymbol{x}_v}(x) = P(\boldsymbol{x}_v = x)$.

Based on Assumption 1, since $\boldsymbol{\Delta}$ is a bijection of $\boldsymbol{x}_v$, there exists $y = g(x)$ such that the joint PMF $p_{\boldsymbol{x}_v,\boldsymbol{\Delta}}(x, y) = P(\boldsymbol{x}_v = x, \boldsymbol{\Delta} = y)$ can be written as:

$$p_{\boldsymbol{x}_v,\boldsymbol{\Delta}}(x, y) = P(\boldsymbol{x}_v = x)P(\boldsymbol{\Delta} = y|\boldsymbol{x}_v = x) = P(\boldsymbol{x}_v = x). \tag{18}$$

Similarly, since $\boldsymbol{x}_v^{adv}$ is a bijection of $\boldsymbol{x}_v$, there exists $z = h(x)$ such that the PMF of $\boldsymbol{x}_v^{adv}$ can be derived as:

$$p_{\boldsymbol{x}_v^{adv}}(z) = P(\boldsymbol{x}_v^{adv} = z) = P(h(x) = z) = P(\boldsymbol{x}_v = x). \tag{19}$$

Thus, we have $p_{\boldsymbol{x}_v^{adv}}(z) = p_{\boldsymbol{x}_v,\boldsymbol{\Delta}}(x, y)$, and for any $y' \neq y, p_{\boldsymbol{x}_v,\boldsymbol{\Delta}}(x, y') = 0$, which leads to the conclusion that:

$$p(\boldsymbol{x}_v^{adv}) = p(\boldsymbol{x}_v, \boldsymbol{\Delta}). \tag{20}$$

E Proof of Approximate Equality of Conditional Entropies

We aim to demonstrate the approximate equality between the conditional entropies $H(L'|\boldsymbol{x}_v, \boldsymbol{\Delta})$ and $H(L'|\boldsymbol{x}_v^{adv})$. Now, let us expand the expression for $H(L'|\boldsymbol{x}_v, \boldsymbol{\Delta})$:

$$\begin{aligned}
H(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) &= -\sum_{L',\boldsymbol{x}_v,\boldsymbol{\Delta}} p(L', \boldsymbol{x}_v, \boldsymbol{\Delta}) \log p(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) \\
&= -\sum_{\boldsymbol{x}_v,\boldsymbol{\Delta}} p(\boldsymbol{x}_v, \boldsymbol{\Delta}) \sum_{L'} p(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) \log p(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) \\
&= -\sum_{\boldsymbol{x}_v,\boldsymbol{\Delta}=g(\boldsymbol{x}_v)} p(\boldsymbol{x}_v, \boldsymbol{\Delta}) \sum_{L'} p(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) \log p(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) \\
&\quad + \underbrace{[-\sum_{\boldsymbol{x}_v,\boldsymbol{\Delta}\neq g(\boldsymbol{x}_v)} p(\boldsymbol{x}_v, \boldsymbol{\Delta}) \sum_{L'} p(L'|\boldsymbol{x}_v, \boldsymbol{\Delta}) \log p(L'|\boldsymbol{x}_v, \boldsymbol{\Delta})]}_{=0} \\
&= -\sum_{\boldsymbol{x}_v} p(\boldsymbol{x}_v) \sum_{L'} p(L'|\boldsymbol{x}_v) \log p(L'|\boldsymbol{x}_v).
\end{aligned} \tag{21}$$

Similarly, $H(L'|\boldsymbol{x}_v^{adv}) = -\sum_{\boldsymbol{x}_v} p(\boldsymbol{x}_v) \sum_{L'} p(L'|\boldsymbol{x}_v) \log p(L'|\boldsymbol{x}_v)$. Therefore, $H(L'|\boldsymbol{x}_v^{adv})$ and $H(L'|\boldsymbol{x}_v, \boldsymbol{\Delta})$ can be canceled out.

F More Details of the Optimization Strategy of Local DIM

Here, we provide more details about how we utilize the Local DIM method to predict the MI values in our specific scenarios. To tackle the image inputs, [29] points out that the presence of pixel-level noise in the input data is often unhelpful for certain downstream tasks, therefore, the Local DIM estimator is effective in handling this issue for estimating reliable MI values. Specifically, it divides the feature map into $M \times M$ feature blocks (*i.e*, $C_\psi(x) = \left\{C_\psi^{(i)}\right\}_{i=1}^{M\times M}$), and then optimizes goal for ω, ψ by estimating and maximizing the average MI between each block features and global features. The optimization formula is as follows:

$$(\hat{\omega}, \hat{\psi}) = \arg\max_{\omega,\psi} \frac{1}{M^2} \sum_{i=1}^{M^2} \widehat{I}_{\omega,\psi}^{(DV)}(C_\psi^{(i)}(X); Y), \tag{22}$$

where $\widehat{I}_{\hat{\omega},\hat{\psi}}^{(DV)}(\cdot)$ is the estimated MI value.

G The Detailed Algorithm of Our Proposed Transfer-Attack Method

The training process to generate adversarial samples is shown in Algorithm 1. Specifically, the perturbation is first initialized and then gradually optimized through multiple iterations. In each iteration, by adding the perturbation to the raw image, the adversarial image is constructed and input into the LVLM along with the prompt to obtain the logits. Next, the MI estimation networks are used to compute the adversarial MI and benign MI, respectively. The MI loss is calculated by maximizing the adversarial MI and minimizing the benign MI, and this is combined with the cross-entropy loss based on the target answer to form the overall loss. Subsequently, the gradient is updated using the momentum mechanism, and the perturbation is adjusted accordingly in both direction and magnitude, while ensuring it stays within the predefined limit. This process is repeated multiple times, and finally produce an adversarial image with strong transferability.

H More Discussions of Our Transferability on LVLMs and Prompts

To improve the transferability of LVLM attacks, we claim that: by increasing the informative dependence between the adversarial perturbation of the input image and the incorrect output of the LVLM, while decreasing the dependence between the benign image pattern and the incorrect output of the LVLM, we can enhance the strength of the adversarial perturbation and ensure that this perturbation contributes more guidance for the attacker's choice compared to the benign image pattern. In this manner, the perturbation can always have more effect than the benign pattern, thus the adversarial example can still mislead the LVLM's reasoning when it is transferred to attack unknown models or prompts.

Here, we provide more discussions on why our proposed attack can separately enhance transferability across LVLM models and prompts in the following:

(1) *Transferability across LVLMs*: the existing LVLM attackers adversarially train the adversarial samples by implicitly restricting the mixed output-input dependency via misleading loss functions, which may confuse the LVLM model to focus on optimizing the joint distribution of benign and adversarial patterns of inputs. This may guide the target model in treating both adversarial and benign patterns of the input image equally. Once the adversarial examples are transferred to an unknown LVLM model, the benign pattern may contribute more output-aware dependency than the adversarial pattern, thus weakening the harmful effect of perturbations and leading to clean results. Instead, our proposed attack explicitly adjusts the LVLM's focus solely on the adversarial noise to enhance the corresponding adversarial harmfulness via the informative constraints. Even the sample is transferred to attack unknown LVLM models, the learned adversarial perturbation is able to jump out of the mixed overfitting and contributes more attacker-chosen guidance effects than the benign one to mislead the reasoning process.

(2) *Transferability across prompts*: Overfitting to the joint distribution of multimodal inputs is also the reason why previous LVLM attacks fail to achieve good prompt-transfer attack performance. However, since our proposed attack method explicitly constrains the dependency between the adversarial

Algorithm 1 Our Proposed Transfer-Attack based on Informative Constraints of Adversarial/Benign MI

Input: The raw image $\boldsymbol{x}_v$, the textual prompt $\boldsymbol{x}_p$ and the attacker-chosen target answer y_{tar}; the LVLM model F; the adversarial MI estimation network $\widehat{I}^{(DV)}_{\hat{\omega}_A,\hat{\psi}_A}$ and the benign MI estimation network $\widehat{I}^{(DV)}_{\hat{\omega}_B,\hat{\psi}_B}$; the number of iteration t_{max}; the decay factor μ; the step size α; the perturbation budget ϵ; and the weights of loss w_1, w_2.

Output: Transferable adversarial image $\boldsymbol{x}_v^{adv}$.

Initialize gradient $g_0 = 0$, perturbation $\boldsymbol{\Delta}_0$
for $t = 0$ **to** $t_{max} - 1$ **do**
  Get adversarial sample $\boldsymbol{x}_{v,t}^{adv} = Clip(\boldsymbol{x}_v + \boldsymbol{\Delta}_t, 0, 1)$
  Get the LVLM's output logits $L^{'} = F(\boldsymbol{x}_{v,t}^{adv}, \boldsymbol{x}_p)$
  Calculate the estimated adversarial MI value:

$$m_{adv} = \widehat{I}^{(DV)}_{\hat{\omega}_A,\hat{\psi}_A}(C_{\hat{\psi}_A}(\boldsymbol{\Delta}_t); L^{'})$$

  Calculate the estimated benign MI value:

$$m_{ben} = \widehat{I}^{(DV)}_{\hat{\omega}_B,\hat{\psi}_B}(C_{\hat{\psi}_B}(\boldsymbol{x}_v); L^{'})$$

  Calculate the MI loss $l_{mi} = m_{adv} - m_{ben}$
  Calculate the cross-entropy loss:

$$l_{ce} = CE(L^{'}, y_{tar})$$

  Get the overall loss $J = w_1 * l_{ce} - w_2 * l_{mi}$
  Update the momentum $g_t = \mu * g_{t-1} + \frac{\nabla J}{||\nabla J||_1}$
  Update perturbation:

$$\boldsymbol{\Delta}_t = Clip(\boldsymbol{\Delta}_t - \alpha \cdot sign(g_t), -\epsilon, \epsilon)$$

end for
return adversarial image $\boldsymbol{x}_v^{adv} = \boldsymbol{x}_v + \boldsymbol{\Delta}_t$

perturbations and the LVLM's targeted output, the adversarial perturbations will be learned to be agnostic to the prompt inputs as the perturbations already are trained to have harmful effects on controlling the flip of the LVLMs' prediction (LVLM will ignore the effect of the prompt). Therefore, once the adversarial images are transferred to attack unseen prompts, the LVLM's reasoning will still be influenced by the harmful impacts of perturbations with its large dependency guidance to output attacker-chosen labels.

I More Experiments

I.1 Experiments on More Datasets

We first provide more performance comparisons on the adversarial transferability across different LVLMs on VQAv2 datasets as shown in Table 6. From this table, we can also find that: (1) Our generated adversarial examples have competitive harmfulness compared to existing attacks in the diagonal values. This demonstrates that our attack also contributes to improve the harmful impact of the samples. (2) Our attacks achieve better transfer-attack performance across four different LVLM models compared to the previous three LVLM attackers, demonstrating the effectiveness of our proposed informative constraints for improving transferability.

I.2 Experiments on More LVLM models

In addition to the LVLM models evaluated in Table 1 and 3, we also provide more detailed transfer-attack experiments on architecturally distinct model families, *i.e.*, MiniGPT-4 (EVA-CLIP-ViT-g-14), Qwen2-VL (CLIP-ViT-bigG) [98], Intern-VL (InternViT-300M-448px-V2_5) [99] and Gemma-3

Table 6: Performance comparisons on the adversarial transferability across different LVLM models (on VQAv2 dataset). The experimental results are calculated by the averaged semantic similarities (↑) and attack success rates (↑) on three tasks. Target text: “I am sorry".

Dataset	Source Model	LVLM Attack	LLaVA-1.5			MiniGPT-4			BLIP-2			InstructBLIP		
			SS	*EM*	*CC*	*SS*	*EM*	*CC*	*SS*	*EM*	*CC*	*SS*	*EM*	*CC*
VQAv2	LLaVA-1.5	PGD [67]	*0.968*	*96.1*	*96.1*	0.044	0.0	0.0	0.096	0.1	0.7	0.135	0.9	1.9
		CroPA [8]	*0.797*	*65.2*	*65.2*	0.023	0.0	0.0	0.084	0.0	0.8	0.120	1.1	1.4
		UniAtt [10]	*0.828*	*82.0*	*84.5*	0.172	10.9	18.1	0.253	22.1	22.1	0.298	23.7	30.6
		Ours	*0.833*	*83.7*	*83.7*	**0.645**	**65.6**	**65.6**	**0.706**	**63.2**	**69.9**	**0.723**	**64.2**	**72.5**
	MiniGPT-4	PGD [67]	0.036	0.0	0.0	*0.866*	*85.0*	*85.0*	0.139	4.4	5.9	0.157	3.0	3.0
		CroPA [8]	0.037	0.0	0.0	*0.987*	*98.0*	*98.7*	0.187	10.5	11.9	0.166	3.4	3.7
		UniAtt [10]	0.373	24.9	35.8	*0.874*	*83.6*	*89.7*	0.380	31.1	38.5	0.309	22.4	27.6
		Ours	**0.658**	**61.3**	**67.2**	*0.910*	*89.9*	*90.0*	**0.726**	**61.4**	**61.4**	**0.695**	**63.1**	**68.3**
	BLIP-2	PGD [67]	0.047	0.0	0.0	0.057	0.0	0.0	*0.649*	*61.4*	*70.6*	0.199	6.5	6.5
		CroPA [8]	0.046	0.0	0.0	0.060	0.0	1.3	*0.619*	*28.1*	*96.1*	0.209	7.4	14.9
		UniAtt [10]	0.406	27.9	35.3	0.385	23.8	32.0	*0.836*	*79.3*	*86.2*	0.284	15.7	20.1
		Ours	**0.613**	**58.5**	**62.0**	**0.602**	**55.8**	**58.3**	*0.744*	*73.6*	*82.9*	**0.492**	**37.5**	**39.6**
	InstructBLIP	PGD [67]	0.037	0.0	0.0	0.049	0.0	0.0	0.132	0.8	4.1	*0.532*	*44.4*	*62.7*
		CroPA [8]	0.039	0.0	0.0	0.059	0.7	3.5	0.265	13.1	23.5	*0.869*	*85.0*	*87.6*
		UniAtt [10]	0.182	10.4	15.1	0.253	18.7	22.9	0.448	37.3	43.6	*0.858*	*81.5*	*83.9*
		Ours	**0.464**	**43.9**	**48.7**	**0.562**	**54.3**	**58.6**	**0.577**	**55.9**	**64.1**	*0.782*	*76.6*	*84.3*

Table 7: The transfer-attack performance on more distinct models. The experimental results are calculated by the averaged semantic similarities (↑) on three tasks. Target text: “I am sorry".

Source Model	LVLM Attack	MiniGPT-4	Qwen2-VL	Intern-VL	Gemma-3
MiniGPT-4	PGD [67]	0.823	0.034	0.042	0.031
	CroPA [8]	0.955	0.042	0.051	0.045
	UniAtt [10]	0.830	0.128	0.145	0.139
	Ours	0.860	**0.639**	**0.610**	**0.591**
Qwen2-VL	PGD [67]	0.054	0.712	0.069	0.051
	CroPA [8]	0.076	0.763	0.093	0.083
	UniAtt [10]	0.199	0.820	0.224	0.241
	Ours	**0.622**	0.792	**0.691**	**0.612**
Intern-VL	PGD [67]	0.051	0.074	0.811	0.043
	CroPA [8]	0.085	0.108	0.823	0.079
	UniAtt [10]	0.212	0.207	0.829	0.224
	Ours	**0.683**	**0.676**	0.847	**0.627**
Gemma-3	PGD [67]	0.063	0.059	0.067	0.842
	CroPA [8]	0.088	0.104	0.117	0.815
	UniAtt [10]	0.237	0.221	0.234	0.833
	Ours	**0.641**	**0.618**	**0.620**	0.819

(SigLIP-ViT) [100]. As shown in Table 7, we can see that our attack is more generalizable to architecturally distinct LVLMs compared to other attacks, demonstrating our great transferability.

I.3 Justification of Our Transfer Attack

The four major LVLMs model listed in Table 1-MiniGPT-4, LLaVA-1.5, BLIP-2 and InstructBLIP-are all composed of a CLIP-VIT visual encoder, an LLM and a connector. Although their architectures are similar, the specific versions and parameters of the CLIP-VIT encoder and LLM are different as shown in Table 8. Besides, their differences also lie in how they bridge vision and language, and how they're optimized for downstream tasks. Therefore, the entire multimodal reasoning ability is different among these LVLMs, while paper "Failures to Find Transferable Image Jailbreaks Between Vision-Language Models" [101] further proves that transfer is not affected by whether the attacked and target VLMs possess matching vision backbones or language models. That's also the reason why our compared baselines achieve poor transfer-attack performances among these four LVLMs in Table 1 of the paper.

Table 8: The architecture and parameters of four LVLMs.

Model	Visual Encoder	Version/Config	Input Resolution	LLM	Connector	Key Features
LLaVA-1.5	CLIP-ViT-L/14	clip-vit-large-patch14-336	336×336	Vicuna-7B	linear projection	High-resolution input
MiniGPT-4	EVA-CLIP-ViT-g-14	EVA-CLIP-g-14	224×224	Llama-2-7B- Chat	linear projection	EVA architecture, self-supervised enhancement
BLIP-2	CLIP-ViT-L/14	clip-vit-large-patch14	224×224	OPT-2.7b	Q-Former	cross-modal bridging
InstructBLIP	EVA-CLIP-ViT-g-14	EVA-CLIP-g-14	224×224	Vicuna- 7B	Q-Former	Instruction tuning

Table 9: Transfer-attack performance comparisons on the adversarial transferability across LVLM models with different visual encoders.

Source Model	LVLM Attack	LLaVA-1.5 (CLIP-ViT-L/14-336)	MiniGPT-4 (EVA-CLIP-ViT-g-14)	BLIP-2 (CLIP-ViT-L/14-224)	Qwen2-VL (CLIP-ViT-bigG)
LLaVA-1.5 (CLIP-ViT-L/14-336)	CroPA	0.819	0.043	0.093	0.036
LLaVA-1.5 (CLIP-ViT-L/14-336)	UniAtt	0.842	0.186	0.267	0.142
LLaVA-1.5 (CLIP-ViT-L/14-336)	Ours	0.813	0.661	0.693	0.634
MiniGPT-4 (EVA-CLIP-ViT-g-14)	CroPA	0.051	0.955	0.125	0.042
MiniGPT-4 (EVA-CLIP-ViT-g-14)	UniAtt	0.298	0.830	0.338	0.128
MiniGPT-4 (EVA-CLIP-ViT-g-14)	Ours	0.650	0.860	0.716	0.639
BLIP-2 (CLIP-ViT-L/14-224)	CroPA	0.059	0.057	0.610	0.045
BLIP-2 (CLIP-ViT-L/14-224)	UniAtt	0.397	0.359	0.817	0.140
BLIP-2 (CLIP-ViT-L/14-224)	Ours	0.695	0.657	0.755	0.641
Qwen2-VL (CLIP-ViT-bigG)	CroPA	0.059	0.076	0.107	0.763
Qwen2-VL (CLIP-ViT-bigG)	UniAtt	0.098	0.199	0.214	0.820
Qwen2-VL (CLIP-ViT-bigG)	Ours	0.633	0.662	0.674	0.792

Furthermore, considering that MiniGPT-4 and InstructBLIP share the same CLIP-ViT visual encoder, we introduced the Qwen2-VL [98] model—whose visual encoder (CLIP-ViT-bigG) differs from that of LLaVA, MiniGPT-4, and BLIP-2—to eliminate the potential influence of visual encoders on transferability. Under a strict black-box setting, we also evaluate the cross-model transferability of our attack. As shown in Table 9, the inclusion of Qwen2-VL still validates the effectiveness of our proposed transfer attack method.

In summary, our implemented transfer attack among these four LVLMs can be taken as a kind of black-box transfer, which is worth studying for enhancing the transferability of existing LVLM attackers.

I.4 Experiments on More Defenses

We provide a more detailed analysis of robustness against defenses in Table 10, where we show the cross-model/prompt attacks' performance under various defenses. In particular, we also introduce two new defense mechanisms for defended target evaluation: one improves the CLIP component in BLIP-2, LLaVA-1.5, MiniGPT-4 and InstructBLIP models with a defended FARE model [102], and the other uses a defended DPS model [103] to embed the input of the BLIP-2, LLaVA-1.5, MiniGPT-4 and InstructBLIP models. The experimental results demonstrate that our cross-model/prompt attack still achieves better transferability compared to baselines under various defense methods, indicating our robustness.

I.5 More Visualizations

We provide more visualization examples to investigate the effectiveness of our attack from three aspects: (1) We provide visual examples of our transfer attack across different LVLM models in Figure 6. It shows that our generated adversarial examples can effectively fool the unknown LVLM models with attacker-chosen text labels, indicating the strong transferability of our proposed attack method. (2) We evaluate the adversarial and benign MI values of both transferable and non-transferable adversarial examples of LVLM attackers in Figure 7. It shows that transferable adversarial examples generally have larger adversarial MI values than its benign ones, while the non-transferable adversarial examples fail to distinguish the adversarial and benign patterns (*i.e.*, having similar adversarial and benign MI values). This demonstrates that strong correlations exist between adversarial dependency and transferability, indicating that a transferable adversarial example

Figure 6: Visualizations of our transferable attack across different LVLM models.

Figure 7: Benign/adversarial MI values of non-transferable and transferable adversarial examples of previous LVLM attacks. The adversarial MI values of transferable examples are shown to be generally larger than their benign MI values, demonstrating the strong correlation between adversarial dependency and transferability.

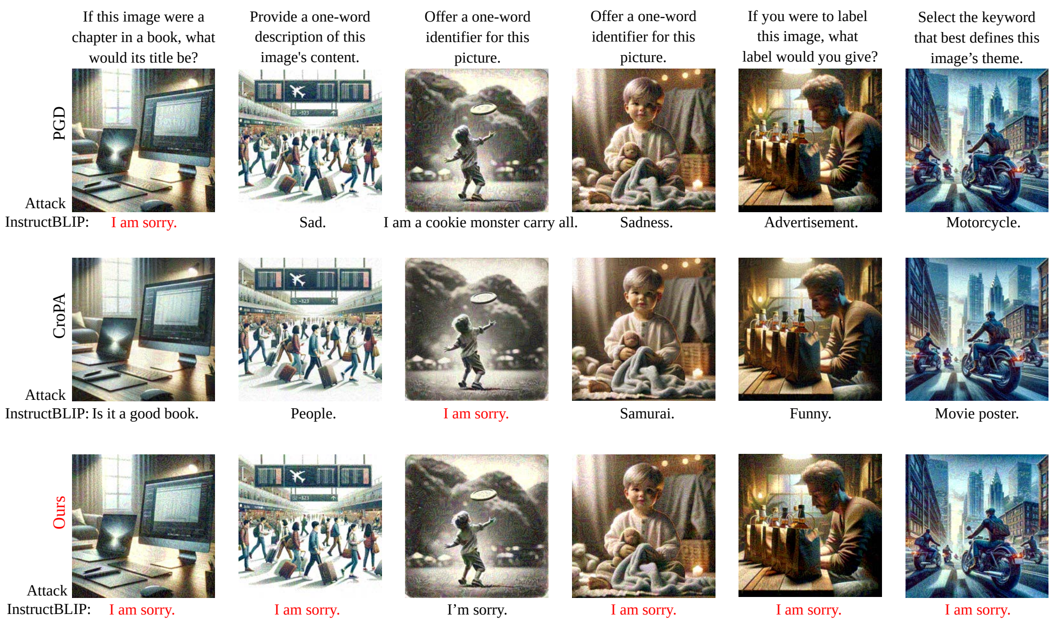

Figure 8: Visualization comparison on targeted transfer attack of different attack methods. We randomly select the above adversarial examples which are generated by the attack method to successfully cause the source LVLM model BLIP-2 to output the attacker-chosen answer "I am sorry". The text below each image represents the transfer-attack answer by attacking the victim LVLM model InstructBLIP.

Table 10: Defense evaluation on cross-model and cross-prompt attacks.

Defense Method	LVLM Attack	BLIP-2 to BLIP-2	BLIP-2 to LLaVA-1.5	BLIP-2 to MiniGPT-4	BLIP-2 to InstructBLIP	Across Prompts(Num=20)
Randomization	PGD [67]	0.213	0.014	0.019	0.042	0.169
	CroPA [8]	0.281	0.046	0.041	0.062	0.188
	UniAtt [10]	0.489	0.156	0.121	0.096	0.336
	Ours	**0.510**	**0.466**	**0.439**	**0.254**	**0.463**
JPEG Compression	PGD [67]	0.348	0.023	0.018	0.046	0.340
	CroPA [8]	0.476	0.054	0.051	0.051	0.411
	UniAtt [10]	0.653	0.251	0.213	0.135	0.474
	Ours	**0.676**	**0.613**	**0.560**	**0.429**	**0.618**
Diffusion Restoration	PGD [67]	0.258	0.021	0.020	0.043	0.217
	CroPA [8]	0.349	0.044	0.046	0.048	0.275
	UniAtt [10]	0.321	0.122	0.109	0.084	0.318
	Ours	**0.535**	**0.464**	**0.443**	**0.279**	**0.485**
FARE	PGD [67]	0.246	0.023	0.017	0.041	0.206
	CroPA [8]	0.316	0.046	0.047	0.064	0.258
	UniAtt [10]	0.408	0.136	0.113	0.104	0.321
	Ours	**0.547**	**0.468**	**0.449**	**0.286**	**0.489**
DPS	PGD [67]	0.319	0.025	0.024	0.048	0.274
	CroPA [8]	0.382	0.051	0.046	0.069	0.293
	UniAtt [10]	0.464	0.154	0.120	0.097	0.325
	Ours	**0.576**	**0.481**	**0.454**	**0.291**	**0.512**

should have larger adversarial dependency and lower benign dependency to guide the LVLM model in focusing more on the harmful perturbation pattern for reasoning. (3) We also provide transfer-attack visualization comparison across different LVLM models on PGD, CroPA, and our method as shown in Figure 8. From this figure, we can find that our attack has better targeted-attack transferability, leading the unknown LVLM model to output the same attacker-chosen answer.

I.6 Experiments under Universal Setting

We also extend our attack as a universal setting by optimizing the perturbation across images. Specifically, we generate adversarial examples on MiniGPT-4 and then transfer them to GPT-4o (GPT-

Table 11: Comparison of transfer-attack performance under the universal and non-universal setting. The experimental results are calculated by the averaged semantic similarities (↑) on three tasks.

	MiniGPT-4	
	to GPT-4o	to Claude-3.5
universal	0.579	0.605
non-universal	0.608	0.620

Table 12: Jailbreak attack performance comparisons on the adversarial transferability across different LVLM models.

Source Model	LVLM Attack	MiniGPT-4	LLaVA-1.5	InstructBLIP
MiniGPT-4	VAJM [104]	53.6	17.9	57.5
	UMK [105]	77.0	12.6	24.7
	Ours	72.8	56.4	68.9
LLaVA-1.5	VAJM [104]	44.8	30.3	33.7
	UMK [105]	19.1	62.2	17.4
	Ours	59.6	68.5	54.3
InstructBLIP	VAJM [104]	37.2	20.6	63.2
	UMK [105]	15.0	13.9	81.6
	Ours	60.2	57.5	76.8

4o-0513) [91] and Claude-3.5-Sonnet [92]. As shown in Table 11, the results under the universal setting demonstrate that our attack maintains strong transferability across different LVLMs.

I.7 More Experiments on Bypassing Safety Mechanisms and Rewiring Entities

We also implement our attack setting into the jailbreak or rewiring attack by changing the targeted attack condition of Equation 9 into corresponding objectives. As for the jailbreak setting, we follow previous jailbreak works VAJM [104] and UMK [105] to implement attacks and evaluate the percentage of the generated texts that exhibit any of the 6 toxic attributes given by Perspective API. As shown in Table 12, our transfer attack among MiniGPT-4, LLaVA-1.5 and InstructBLIP also shows strong applicability. As for the rewiring attack, we define the adversarial loss for maximizing the probability of "cat" appearing anywhere in the answer text and the regularization loss for constraining its total probability to be close to 1 to serve as rewiring constraint for LVLM's output to produce texts containing "cat" (Not a direct/simple "this is a cat" output), which achieves great ASR in Table 13.

J Task-aware Prompts for Different Tasks

Prompt Examples for Image Captioning. *Elaborate on the elements present in this image. In one sentence, summarize the activity in this image. Relate the main components of this picture in words. What narrative unfolds in this image? Break down the main subjects of this photo. Give an account of the main scene in this image. In a few words, state what this image represents. Describe the setting or location captured in this photograph. Provide an overview of the subjects or objects seen in this picture. Identify the primary focus or point of interest in this image. What would be the perfect title for this image? How would you introduce this image in a presentation? Present a quick rundown of the image's main subject. What's the key event or subject captured in this photograph? Relate the actions or events taking place in this image. Convey the content of this photograph in a single phrase. Offer a succinct description of this picture. Give a concise overview of this image. Translate the contents of this picture into a sentence. Describe the characters or subjects seen in this image. Capture the activities happening in this image with words. How would you introduce this image to an audience? State the primary events or subjects in this picture. What are the main elements in this photograph? Provide an interpretation of this image's main event or subject. How would you title this image for an art gallery? What scenario or setting is depicted in this image? Concisely state the main actions occurring in this image. Offer a short summary of this photograph's contents. How would you annotate this image in an album? If you were to describe this image on the radio, how*

Table 13: Rewiring attack performance on the adversarial transferability.

ASR	MiniGPT-4		
	to MiniGPT-4	to LLaVA-1.5	to InstructBLIP
Ours	86.4%	71.9%	74.1%

would you do it? In your own words, narrate the main event in this image. What are the notable features of this image? Break down the story this image is trying to tell. Describe the environment or backdrop in this photograph. How would you label this image in a catalog? Convey the main theme of this picture succinctly. Characterize the primary event or action in this image. Provide a concise depiction of this photo's content. Write a brief overview of what's taking place in this image. Illustrate the main theme of this image with words. How would you describe this image in a gallery exhibit? Highlight the central subjects or actions in this image. Offer a brief narrative of the events in this photograph. Translate the activities in this image into a brief sentence. Give a quick rundown of the primary subjects in this image. Provide a quick summary of the scene captured in this photo. How would you explain this image to a child? What are the dominant subjects or objects in this photograph? Summarize the main events or actions in this image. Describe the context or setting of this image briefly. Offer a short description of the subjects present in this image. Detail the main scenario or setting seen in this picture. Describe the main activities or events unfolding in this image. Provide a concise explanation of the content in this image. If this image were in a textbook, how would it be captioned? Provide a summary of the primary focus of this image. State the narrative or story portrayed in this picture. How would you introduce this image in a documentary? Detail the subjects or events captured in this image. Offer a brief account of the scenario depicted in this photograph. State the main elements present in this image concisely. Describe the actions or events happening in this picture. Provide a snapshot description of this image's content. How would you briefly describe this image's main subject or event? Describe the content of this image. What's happening in this image? Provide a brief caption for this image. Tell a story about this image in one sentence. If this image could speak, what would it say? Summarize the scenario depicted in this image. What is the central theme or event shown in the picture? Create a headline for this image. Explain the scene captured in this image. If this were a postcard, what message would it convey? Narrate the visual elements present in this image. Give a short title to this image. How would you describe this image to someone who can't see it? Detail the primary action or subject in the photo. If this image were the cover of a book, what would its title be? Translate the emotion or event of this image into words. Compose a one-liner describing this image's content. Imagine this image in a magazine. What caption would go with it? Capture the essence of this image in a brief description. Narrate the visual story displayed in this photograph.

Prompt Examples for Image Classification. *Identify the primary theme of this image in one word. How would you label this image with a single descriptor? Determine the main category for this image. Offer a one-word identifier for this picture. If this image were a file on your computer, what would its name be? Tag this image with its most relevant keyword. Provide the primary classification for this photograph. How would you succinctly categorize this image? Offer the primary descriptor for the content of this image. If this image were a product, what label would you place on its box? Choose a single word that encapsulates the image's content. How would you classify this image in a database? In one word, describe the essence of this image. Provide the most fitting category for this image. What is the principal subject of this image? If this image were in a store, which aisle would it belong to? Provide a singular term that characterizes this picture. How would you caption this image in a photo contest? Select a label that fits the main theme of this image. Offer the most appropriate tag for this image. Which keyword best summarizes this image? How would you title this image in an exhibition? Provide a succinct identifier for the image's content. Choose a word that best groups this image with others like it. If this image were in a museum, how would it be labeled? Assign a central theme to this image in one word. Tag this photograph with its primary descriptor. What is the overriding theme of this picture? Provide a classification term for this image. How would you sort this image in a collection? Identify the main subject of this image concisely. If this image were a magazine cover, what would its title be? What term would you use to catalog this image? Classify this picture with a singular term. If this image were a chapter in a book, what would its title be? Select the most fitting classification for this image. Define the essence of this image in one word. How would you label this image for easy retrieval? Determine the core theme of this photograph. In* *a word, encapsulate the main subject of this image. If this image were an art piece, how would it be labeled in a gallery? Provide the most concise descriptor for this picture. How would you name this image in a photo archive? Choose a word that defines the image's main content. What would be the header for this image in a catalog? Classify the primary essence of this picture. What label would best fit this image in a slideshow? Determine the dominant category for this photograph. Offer the core descriptor for this image. If this image were in a textbook, how would it be labeled in the index? Select the keyword that best defines this image's theme. Provide a classification label for this image. If this image were a song title, what would it be? Identify the main genre of this picture. Assign the most apt category to this image. Describe the overarching theme of this image in one word. What descriptor would you use for this image in a portfolio? Summarize the image's content with a single identifier. Imagine you're explaining this image to someone over the phone. Please describe the image in one word? Perform the image classification task on this image. Give the label in one word. Imagine a child is trying to identify the image. What might they excitedly point to and name? If this image were turned into a jigsaw puzzle, what would the box label say to describe the picture inside? Classify the content of this image. If you were to label this image, what label would you give? What category best describes this image? Describe the central subject of this image in a single word. Provide a classification for the object depicted in this image. If this image were in a photo album, what would its label be? Categorize the content of the image. If you were to sort this image into a category, which one would it be? What keyword would you associate with this image? Assign a relevant classification to this image. If this image were in a gallery, under which section would it belong? Describe the main theme of this image in one word. Under which category would this image be cataloged in a library? What classification tag fits this image the best? Provide a one-word description of this image's content. If you were to archive this image, what descriptor would you use?*

Prompt Examples for VQA. *Any cutlery items visible in the image? Any bicycles visible in this image? Any boats visible in the image? Any bottles present in the image? Are curtains noticeable in the image? Are flags present in the image? Are flowers present in the image? Are fruits present in the image? Are glasses discernible in the image? Are hills visible in the image? Are plates discernible in the image? Are shoes visible in this image? Are there any insects in the image? Are there any ladders in the image? Are there any man-made structures in the image? Are there any signs or markings in the image? Are there any street signs in the image? Are there balloons in the image? Are there bridges in the image? Are there musical notes in the image? Are there people sitting in the image? Are there skyscrapers in the image? Are there toys in the image? Are toys present in this image? Are umbrellas discernible in the image? Are windows visible in the image? Can birds be seen in this image? Can stars be seen in this image? Can we find any bags in this image? Can you find a crowd in the image? Can you find a hat in the image? Can you find any musical instruments in this image? Can you identify a clock in this image? Can you identify a computer in this image? Can you see a beach in the image? Can you see a bus in the image? Can you see a mailbox in the image? Can you see a mountain in the image? Can you see a staircase in the image? Can you see a stove or oven in the image? Can you see a sunset in the image? Can you see any cups or mugs in the image? Can you see any jewelry in the image? Can you see shadows in the image? Can you see the sky in the image? Can you spot a candle in this image? Can you spot a farm in this image? Can you spot a pair of shoes in the image? Can you spot a rug or carpet in the image? Can you spot any dogs in the image? Can you spot any snow in the image? Do you notice a bicycle in the image? Does a ball feature in this image? Does a bridge appear in the image? Does a cat appear in the image? Does a fence appear in the image? Does a fire feature in this image? Does a mirror feature in this image? Does a table feature in this image? Does it appear to be nighttime in the image? Does it look like an outdoor image? Does it seem to be countryside in the image? Does the image appear to be a cartoon or comic strip? Does the image contain any books? Does the image contain any electronic devices? Does the image depict a road? Does the image display a river? Does the image display any towers? Does the image feature any art pieces? Does the image have a lamp? Does the image have any pillows? Does the image have any vehicles? Does the image have furniture? Does the image primarily display natural elements? Does the image seem like it was taken during the day? Does the image seem to be taken indoors? Does the image show any airplanes? Does the image show any benches? Does the image show any landscapes? Does the image show any movement? Does the image show any sculptures? Does the image show any signs? Does the image show food? Does the image showcase a building? How many animals are present in the image? How many bikes are present in the image? How many birds are visible in the image? How many buildings can be identified in the image? How many cars can be seen in the image? How many doors can* *you spot in the image? How many flowers can be identified in the image? How many trees feature in the image? Is a chair noticeable in the image? Is a computer visible in the image? Is a forest noticeable in the image? Is a painting visible in the image? Is a path or trail visible in the image? Is a phone discernible in the image? Is a train noticeable in the image? Is sand visible in the image? Is the image displaying any clouds? Is the image set in a city environment? Is there a plant in the image? Is there a source of light visible in the image? Is there a television displayed in the image? Is there grass in the image? Is there text in the image? Is water visible in the image, like a sea, lake, or river? How many people are captured in the image? How many windows can you count in the image? How many animals, other than birds, are present? How many statues or monuments stand prominently in the scene? How many streetlights are visible? How many items of clothing can you identify? How many shoes can be seen in the image? How many clouds appear in the sky? How many pathways or trails are evident? How many bridges can you spot? How many boats are present, if it's a waterscape? How many pieces of fruit can you identify? How many hats are being worn by people? How many different textures can you discern? How many signs or billboards are visible? How many musical instruments can be seen? How many flags are present in the image? How many mountains or hills can you identify? How many books are visible, if any? How many bodies of water, like ponds or pools, are in the scene? How many shadows can you spot? How many handheld devices, like phones, are present? How many pieces of jewelry can be identified? How many reflections, perhaps in mirrors or water, are evident? How many pieces of artwork or sculptures can you see? How many staircases or steps are in the image? How many archways or tunnels can be counted? How many tools or equipment are visible? How many modes of transportation, other than cars and bikes, can you spot? How many lamp posts or light sources are there? How many plants, other than trees and flowers, feature in the scene? How many fences or barriers can be seen? How many chairs or seating arrangements can you identify? How many different patterns or motifs are evident in clothing or objects? How many dishes or food items are visible on a table setting? How many glasses or mugs can you spot? How many pets or domestic animals are in the scene? How many electronic gadgets can be counted? Where is the brightest point in the image? Where are the darkest areas located? Where can one find leading lines directing the viewer's eyes? Where is the visual center of gravity in the image? Where are the primary and secondary subjects positioned? Where do the most vibrant colors appear? Where is the most contrasting part of the image located? Where does the image place emphasis through scale or size? Where do the textures in the image change or transition? Where does the image break traditional compositional rules? Where do you see repetition or patterns emerging? Where does the image exhibit depth or layers? Where are the boundary lines or borders in the image? Where do different elements in the image intersect or overlap? Where does the image hint at motion or movement? Where are the calm or restful areas of the image? Where does the image become abstract or less defined? Where do you see reflections, be it in water, glass, or other surfaces? Where does the image provide contextual clues about its setting? Where are the most detailed parts of the image? Where do you see shadows, and how do they impact the composition? Where can you identify different geometric shapes? Where does the image appear to have been cropped or framed intentionally? Where do you see harmony or unity among the elements? Where are there disruptions or interruptions in patterns? What is the spacing between objects or subjects in the image? What foreground, mid-ground, and background elements can be differentiated? What type of energy or vibe does the image exude? What might be the sound environment based on the image's content? What abstract ideas or concepts does the image seem to touch upon? What is the relationship between the main subjects in the image? What items in the image could be considered rare or unique? What is the gradient or transition of colors like in the image? What might be the smell or aroma based on the image's content? What type of textures can be felt if one could touch the image's content? What boundaries or limits are depicted in the image? What is the socioeconomic context implied by the image? What might be the immediate aftermath of the scene in the image? What seems to be the main source of tension or harmony in the image? What might be the narrative or backstory of the main subject? What elements of the image give it its primary visual weight? Would you describe the image as bright or dark? Would you describe the image as colorful or dull?*

K Limitations and Broader Impacts

Limitations. Future research directions for evaluating the security of LVLMs should prioritize physical-world adversarial attack scenarios, particularly in safety-critical deployments such as autonomous driving or robotic control systems. In such applications, adversarial examples must be crafted under real-world constraints, where input images are captured by physical sensors (e.g., cameras or LiDAR) and subject to dynamic environmental interference (e.g., lighting variations, motion blur, or sensor noise). Although our attack method is effective in theoretical settings, it still requires further improvement to maintain attack effectiveness under such complex real-world conditions.

Broader Impacts. This study presents a novel LVLM attack method for generating transferable adversarial examples against different LVLM models and prompts. Despite the rising interest in attacking LVLMs, our attack method shows that direct harm exists to LVLM applications. Our research aims to deepen the understanding of LVLM robustness and enhance their safety, promoting safer AI environments. However, we acknowledge the potential negative societal impact of our work and the presence of potentially offensive and harmful adversarial examples in our paper. It is possible that the developed attacking strategies could be misused to evade practically deployed systems and cause potential negative societal impacts. Specifically, our threat model assumes real-world access and targeted responses, which involves manipulating existing APIs such as GPT-4 (with visual inputs) and/or Midjourney on purpose, thereby increasing the risk if these vision-language APIs are implemented as plugins in other products. We believe the contributions in our work point out new vulnerabilities in LVLMs, which could aid future research in making them more reliable and secure.