2024 IEEE Symposium on Security and Privacy (SP) | 979-8-3503-3130-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/SP54263.2024.00102

Transferable Multimodal Attack on Vision-Language Pre-training Models

Haodi Wang[1], Kai Dong[1]*, Zhilei Zhu[3], Haotong Qin[4],
Aishan Liu[4], Xiaolin Fang[1], Jiakai Wang[2]*, Xianglong Liu[4]
[1]*Southeast University*
{*whd,dk,xiaolin*}*@seu.edu.cn*
[2]*Zhongguancun Laboratory*
wangjk@mail.zgclab.edu.cn
[3]*Data Space Research Institute of Hefei Comprehensive National Science Centre*
zhu742002988@gmail.com
[4]*Beihang University*
qinhaotong@gmail.com, liuaishan@buaa.edu.cn, xlliu@nlsde.buaa.edu.cn

***Abstract*—Vision-Language Pre-training (VLP) models have achieved remarkable success in practice, while easily being misled by adversarial attack. Though harmful, adversarial attacks are valuable in revealing the blind-spots of VLP models and promoting their robustness. However, existing adversarial attacking studies pay insufficient attention to the key roles of different modality-correlated features, leading to unsatisfactory transferable attacking performance. To tackle this issue, we propose the Transferable MultiModal (TMM) attack framework, which tailors both the modality consistency and modality discrepancy features. To promote transferability, we propose the attention-directed feature perturbation to disturb the modality-consistency features in critical attention regions. In light of the commonly employed cross-attention can represent the consistent features among diverse models, it is more possible to mislead the similar model perception for activating stronger transferability. For improving attacking ability, we proposed the orthogonal-guided feature heterogenization to guide the adversarial perturbation to contain more modality-discrepancy features in the encoded embeddings. Since VLP models rely more on aligned features among different modalities during decision-making, increasing the modality-discrepant could confuse the learned representation for better attacking ability. Extensive experiments under diverse settings demonstrate that the proposed TMM outperforms the comparisons by large margins, *i.e.*, 20.47% improvements in transferable attacking ability on average. Moreover, we highlight that our TMM also shows outstanding attacking performance on large models, such as MiniGPT-4, Otter, *etc*.**

1. Introduction

Deep neural networks (DNNs) have significantly advanced various machine learning tasks but face challenges from imperceptible adversarial examples [1], [2]. Although these adversarial examples may appear detrimental, they hold substantial value for deepening our understanding and scrutiny of the underlying mechanisms of deep learning models. Specifically, adversarial examples could serve as efficient testing cases and help us explore the blind spots of models, showing great significance [3]. For some time now, researchers have been diligently creating adversarial examples to challenge deep learning models, including white-box [4]–[6] and black-box attacks [7]–[10]. The main difference between the white-box and black-box settings is whether the adversary is familiar with the target model, *i.e.*, the white-box setting always has complete knowledge and full access, while the black-box one is the opposite.

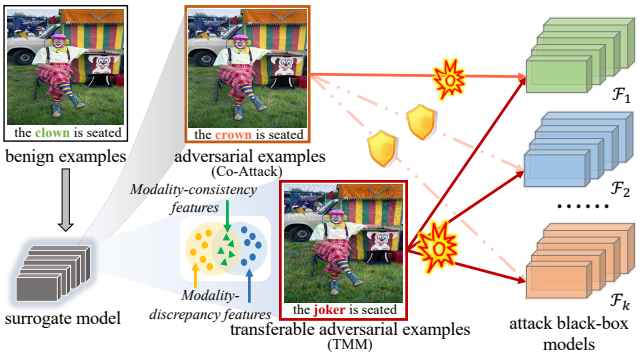


Figure 1: The proposed TMM attack framework generates adversarial examples with stronger transferable attacking ability on different VLP models.

Recently, Vision-Language Pre-training (VLP) models in the multimodal field have commanded considerable attention [11]–[15], due to their impressive performance across a range of Vision-Language (V+L) tasks. However, existing work on their adversarial robustness has shortcomings. Specifically, Co-Attack [16] pioneered the exploration of multimodal joint adversarial attacks under white-box settings, optimizing perturbations from the overall difference of fused feature probability distributions, without consider-

*Kai Dong and Jiakai Wang are corresponding authors.

DOI 10.1109/SP54263.2024.00102

ing the feature differences between modalities. In addition, SGA [17], published during the same period as our work, investigated adversarial attacks under black-box settings based on Co-Attack, but it only considered the similarity of set-level text and images, thus not fully utilizing modality interaction-related features. Given the inherent multimodality of VLP models, including a range of non-classification tasks such as cross-modal Vision-Language Retrieval, it is inappropriate to directly use existing single-modal black-box adversarial attack methods, and may even reduce performance. A plausible solution would be to mount adversarial attacks against the embedding representations instead of the downstream task labels. However, considering the nature of V+L tasks, adversarial perturbations across modalities should be jointly addressed rather than independently [16]. Therefore, we are motivated to further investigate black-box adversarial attacks on VLP models.

In fact, the utility of adversarial examples is fundamentally determined by their black-box attacking ability in practice. Unfortunately, the current study on adversarial attacks in VLP models shows unsatisfactory transferability, where the reasons behind this can be summarized as 1) the neglect of the modality-consistency features. The modality-consistency features represent decision-related characteristics shared among different modalities in various VLP models, which are highly correlated to perform transferable attacks. Specifically, these features refer to attributes that are consistently present across diverse data inputs and play a pivotal role in decision-making. For instance, both textual and visual inputs might share attributes like color or shape information. Existing literature has demonstrated that accurately identifying these shared attributes can enhance the effectiveness of VLP models [18]. 2) The neglect of the modality-discrepancy features. The modality-discrepancy features represent unique characteristics specific to each modality that the decision of the VLP models does not depend on. For instance, in textual inputs, a unique attribute could be the syntax, while in visual inputs, it could be the pixel intensity. Enhancing these unique, modality-specific attributes can potentially destabilize the decision-making process of the VLP models. This is attributed to the fact that these attributes do not contribute to the decision-making process, and their amplification can introduce ambiguity in the model's decision-making process [19].

Our work: In this study, we propose the **T**ransferable **M**ulti**M**odal (TMM) attack framework to generate transferable adversarial examples with strong attacking ability on multiple VLP models by exploiting modality-consistency and modality-discrepancy features. **To enhance transferability**, we introduce an attention-directed feature perturbation strategy. Given the prevalent use of cross-attention in numerous VLP models for facilitating cross-modal interaction [20], we believe that this shared module can exploit modality-consistency features, thereby enhancing transferable attacks. By identifying decision-related modality-consistency features of text and image under cross-attention guidance, replacing the relevant text, and allocating more perturbation budget to critical regions within the images, the attacking ability could effectively transfer among multiple models. **To amplify the attacking ability**, we devise an orthogonal-guided feature heterogenization approach. Considering that VLP models rely more on aligned features among different modalities during decision-making, we aim to heterogenize additional embedded features into discrepant ones to improve attacking abilities. By orthogonalizing the modality fused representation, we guide the adversarial perturbation to incorporate more modality-discrepancy features within the encoded embeddings, thereby promoting stronger attacks.

To evaluate the effectiveness of our proposed TMM attack framework, we conduct extensive experiments on vision-language retrieval [21], visual grounding [22] and visual entailment [23] in a black-box setting. In the main experiments, we employed ALBEF [12] as a surrogate model to evaluate the attacking ability of our adversarial example on six distinct VLP models. Additionally, given that Large generative Vision-Language Models (LVLMs) like GPT-4 [24] and BLIP-2 [15] have recently achieved remarkable results in multimodal generation tasks, and diverge from traditional VLP models in architecture and parameters (*e.g.*, ALBEF has 314M parameters, BLIP 224M, while GPT-4 has a staggering 1760B, and BLIP-2 has 6.7B), this discrepancy drives us further to investigate the effectiveness of TMM on large models. Therefore, we further test LVLMs in a black-box setting. The experimental results demonstrate that our proposed TMM achieves higher transferable attacking ability and outperforms compared baselines by large margins. Moreover, TMM has a more pronounced impact on the decision-making of LVLMs than baselines.

Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first investigation of transferable adversarial examples on VLP models by considering modality-consistency and modality-discrepancy features.
- We propose an effective framework to generate strong transferable adversarial examples against VLP models, named TMM, which compose attention-directed feature perturbation and orthogonal-guided feature heterogenization.
- We demonstrate the effectiveness of TMM under various settings, and the experimental results show that the TMM improves the attack success rate by an average of $\mathbf{20.47}\%$ compared to the baselines while maintaining good stealth.
- We conducted further explorations on LVLMs under black-box settings, preliminarily demonstrating the transferable attacking ability of our TMM against powerful LVLMs, *i.e.*, revealing their drawbacks.

In the remainder of this paper, we first provide a brief introduction to VLP models in Section 2. Next, in Section 3, we define the threat model. Then, in Section 4, we formally describe our attack framework and extensively evaluate it in Section 5. In Section 6, we provide some countermeasures. In Section 7, we discuss related work closely to our attack.

2. Background

In this section, we first provide a brief introduction to Vision-Language Pre-training Model (VLP), then delve into its structure and downstream tasks.

2.1. Vision-Language Pre-training Model

The pursuit of replicating human-like responses in machines has spurred AI research toward sophisticated tasks such as vision-language comprehension. However, the requirement for extensive labelled data has presented a substantial challenge. This obstacle has been circumvented by pre-training models based on the transformer structure, which utilizes self-supervised learning and many unlabeled data, thereby enabling fine-tuning for downstream tasks with minimal manual labelling. This strategy, initiated with BERT [25] in NLP, has been extended to various unimodal pre-training models like Vision Transformer [26] in CV and Wave2Vec [27] in speech processing, yielding improvements in downstream tasks. Likewise, the multimodal field also grapples with data labelling challenges, instigating investigations into the effective use of pre-training methods for multimodal tasks. Presently, VLP models have successfully confronted this challenge, primarily by learning the semantic correspondence between different modalities through large-scale data pre-training. Moreover, LVLMs built on VLP have shown remarkable abilities in multimodal generative tasks, such as MMGPT [28] and Otter [29].

2.2. VLP's Architecture

Existing VLP models fall into two main architectures: single-stream and dual-stream.

Single-stream architecture: The single-stream architecture facilitates the interaction between image and text by employing a deep fusion encoder with cross-attention [12], [30]–[32]. For the image-text pair $d = \{d_v, d_t\}$ sampled from a multimodal dataset, this architecture can be represented as $\mathcal{F}(d_v, d_t) = \text{Encoder}_{\text{single}}(d_v \parallel d_t)$, where $\parallel$ denotes the concatenation operation, and Encoder is the single-stream encoder. Standard training methods for fusion encoder-based models include image-text matching, masked language modelling, word-region/patch alignment, and feature regression. While these models demonstrate superior performance in vision-language classification tasks, the comprehensive encoding of all image-text pairs often results in slower inference speeds for retrieval tasks. Recently, ALBEF [12] employs transformer [25], [26] to obtain the representations of image and text. TCL [32] builds on this, adding features like intra-modal alignment and fine-grained global-local alignment. X-VLM [30] uses patch embeddings for flexible visual concept representation and alignment with texts. Similarly, ViLT [31] transforms images into patch embeddings and combines them with word embeddings for image-text interaction modeling.

Dual-stream architecture: The dual-stream architecture employs two separate transformer structures for text and visual features, learning cross-modal interactions through cross-attention [13], [14], [33]. This architecture can be represented as $\mathcal{F}(d_v, d_t) = \text{Encoder}_{\text{visual}}(d_v) \parallel \text{Encoder}_{\text{text}}(d_t)$, where $\text{Encoder}_{\text{visual}}$ and $\text{Encoder}_{\text{text}}$ are the visual and textual encoders, respectively. Customarily, image-text contrastive learning is deployed for model optimization. These dual-encoder models excel in vision-language retrieval tasks. However, their simple interaction mechanism falls short for tasks requiring complex reasoning, such as visual reasoning and visual question answering. CLIP [13] uses 400 million web-sourced (image, text) pairs to train image and text encoders, achieving top performance. Blip [14] introduces the CapFilt module to reduce data noise and improve the dataset. METER [33] explores designing and pre-training a transformer-based VLP model end-to-end to balance performance and inference speed.

2.3. VLP's Downstream Tasks

The VLP models can be applied to many tasks that require vision-language interaction. This paper mainly mentions the following three:

- **Vision-Language Retrieval (VLR) [21]:** This task involves matching visual data with corresponding textual data. It consists of two subtasks: image-to-text retrieval (TR), retrieving corresponding text for an image, and text-to-image retrieval (IR), finding the matching image for specific text.
- **Visual Entailment (VE) [23]:** This task uses images and text as conditions and hypotheses to predict whether the relationship between them is entailment, neutral, or contradiction.
- **Visual Grounding (VG) [22]:** This task aims to locate the object regions in the image corresponding to a particular textual description.

3. Threat Model

We delineate our threat model regarding the attacker's goals, constraints, capabilities, and attack scenarios.

Attacker's goals: The attacker aims to introduce adversarial perturbations into the visual and textual inputs of the VLP models, leading to incorrect outputs in downstream tasks that rely on these pre-training models. Given a benign image-text pair $d = \{d_v, d_t\}$, a VLP model is able to encode this input into a fused embedding e. An adversarial perturbation δ_v and δ_t are designed to mislead the surrogate model $\mathcal{F}$ into producing an incorrect embedding:

$$\mathcal{F}(d_v \oplus \delta_v, d_t \xrightarrow{\delta_t} d_t^{adv}) \neq e \tag{1}$$

where the operator $\oplus$ represents an addition operation of a perturbation δ_v, which is subject to a small budget ϵ_v. The symbol $\rightarrow$ denotes the modification or replacement of certain tokens in the input text d_t according to δ_t. It should be noted that δ_t is constrained to the token level ϵ_t, *i.e.*, how many tokens are modified/replaced with semantic consistency. By obtaining adversarial examples through the above

goal and directly applying them to the target VLP model, the attacker can influence the target model's decision-making.

Attacker's constraints: When deploying VLP models in real-world scenarios, it is crucial to account for the intricacy of potential adversarial attacks, which may involve unknown structures, parameters, and training details. Therefore, it is essential to simulate realistic conditions as much as possible, including white-box and black-box settings.

In the white-box setting, we use the target model that may be attacked as a surrogate model during training. This implies that the attacker has complete access to the model's information, such as its architecture and parameters. In the black-box setting, we impose stricter restrictions on the attacker, which means that the attacker has very limited knowledge about the target model. To be more practical, we consider the transfer-based attack. Specifically, we only generate adversarial examples based on a specific model and then perform the attack without any additional operations, such as fine-tuning or querying. This ensures that the success of an attack is fundamentally dependent on strong intermodel transferability and guarantees the attacker does not have access to all information of the target model in the black-box setting.

Attacker's capabilities: Most attackers performing black-box adversarial attacks can be classified into two categories: query-based and transfer-based. Query-based attacks generate adversarial perturbation by querying the target model to obtain the model's output, and then employ an optimization method for gradient estimation. This process typically requires a large number of queries, making it inefficient and challenging to implement in real-world scenarios. Consequently, in this paper, we basically follow a transfer-based attack. Specifically, the attacker needs to initially generate adversarial examples on a publicly available white-box model using a certain dataset, and then directly apply them to the target model.

Possible attacking scenarios: A paramount question for adversarial attack tasks is their practical feasibility, as the existence of potential threats or benefits determines the value and importance of attack methods. For our proposed TMM, we believe it applies to various scenes of V+L tasks.

For instance, in the VLR tasks, users may search for products on e-commerce websites or related posts on social media platforms. If attackers can successfully alter the model's prediction results through adversarial attacks, users may not find what they want, potentially affecting user experience and even causing damage to the platform's commercial interests. In the case of VE tasks, like autonomous driving and robotics, models need to infer some conclusions based on visual information, such as judging the situation of the road ahead. If attackers manipulate the model to produce incorrect inference results, it could lead to safety incidents. As for VG tasks, the model must associate language descriptions with specific areas in the image, such as robot navigation. If attackers cause the model to fail to associate the description and image correctly, it may affect the model's performance and even lead to severe consequences.

4. Transferable Multimodal Attack

4.1. Key Ideas

For improving the transferability, we propose the *attention-directed feature perturbation* (***ADFP***) strategy. The cross-attention module in various VLP models merges features from different modalities, we believe that this shared module can leverage modality-consistency features to guide the enhancement of transferable attacks. Considering the perceptual similarity of attention correlated characteristics [7], we aim to disturb modality-consistency features by replacing the relevant text and adding perturbations with varying weights to the attention-directed critical regions in the image, which represent important features related to the model's decisions. Further, building upon the observations that robust model predictions are highly related to the structural features, *e.g.*, shape features, among models [34], [35], to further activate the transferability of adversarial examples, we introduce Structural SIMilarity (SSIM) [36] to guide structural features in regions where adversarial pixel blurring is associated with modality-consistency features.

For enhancing the attacking ability, we design an *orthogonal-guided feature heterogenization* (***OGFH***) approach. Existing VLP models rely more on aligned features among different modalities during decision-making, and usually deliberately use different training strategies to ignore modality-discrepancy features that exhibit negative impacts on models [12], [20], [30]. Motivated by this observation, we exploit the idea of orthogonalization to enrich discrepant features to perform attacks. Technically, we design the cosine orthogonalization optimization strategy to guide the adversarial perturbation and heterogenize more embedded features into discrepant ones. By orthogonalizing the embedded features subject to different modalities, we could guide the adversarial perturbation to contain more modality-discrepancy features within the encoded embeddings. Therefore, VLP models will struggle to better capture aligned consistent features, thereby enhancing attacking ability.

By synthetically exploiting modality-consistency and modality-discrepancy features, we propose a framework for generating transferable adversarial examples with strong attacking ability to diverse VLP models. Fig. 2 depicts the general framework of our proposed TMM.

4.2. Attention-Directed Feature Perturbation

Previous work has pointed out that the modality-consistency features have a significant impact on the decisions of multimodal models [18], [37]. Under this perspective, we believe that the commonly-used cross-attention module, which is designed to enable cross-modal interaction in VLP models [20], is highly related to the perception of modality-consistency features. Hence, we suggest using cross-attention to guide the perturbation process, *i.e.*, replacing relevant text, and allocating a higher perturbation budget to critical attention regions in images, helping the attacks effectively transfer across multiple models. In addition, given

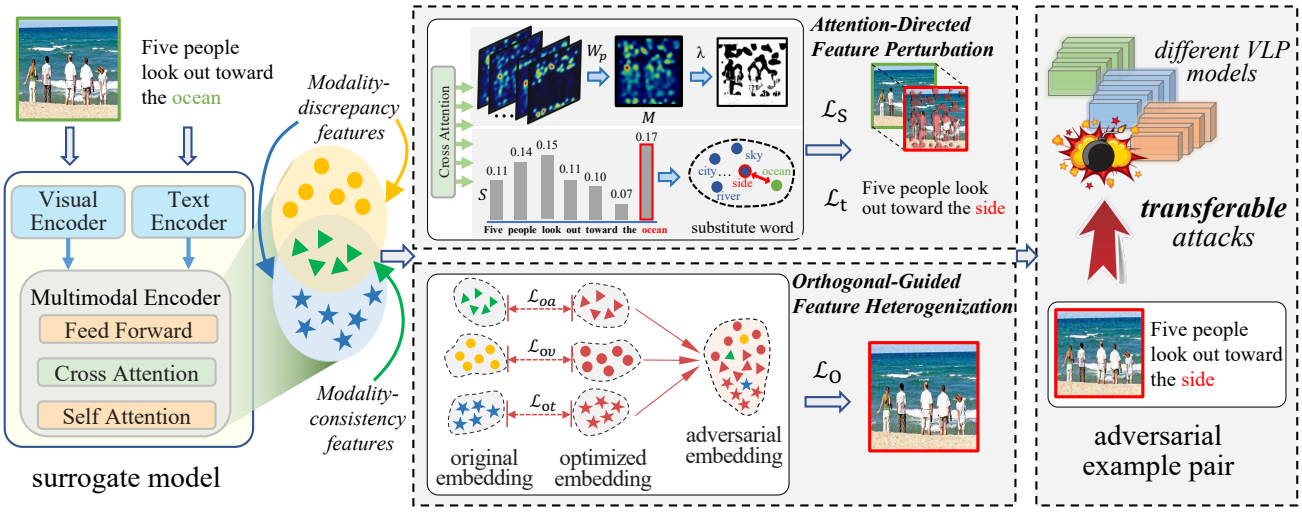


Figure 2: The framework of our proposed Transferable Multimodal (TMM) attack. We enhance the transferability of the adversarial examples via attention-directed feature perturbation strategy, and promote their attacking ability via orthogonal-guided feature heterogenization strategy. Benefiting from the modality-related tailoring, the proposed TMM is allowed to attack diverse VLP models strongly.

that structural features make sense to robust predictions [34], [35], we consider further perturbing the structural features of regions in images correlated to modality-consistency features to enhance the transferability of adversarial examples.

Specifically, for the extraction of attention features, some studies have indicated that the information in both lower and higher layers of the attention module is a mixture of local and global information [38], [39]. Therefore, in order to effectively extract information from multiple layers of the cross-attention module, we compute the word attention score and image attention map by averaging the information across these layers, which can be defined as:

$$M = \frac{\sum_{p=1}^{N} \left[(m_1 \parallel m_2 \ldots \parallel m_k) W_p \right]}{N}$$
$$m_i = \text{softmax}\left(\frac{QK^T}{\sqrt{d}} \right) V \quad i = 1, \ldots, k \tag{2}$$

where m_i denotes the attention feature vector of the i-th head, W_p is the weight matrix of different layers, Q, K, V denote the feature matrix of different modality, and $\sqrt{d}$ denotes the scaling factor for stabilizing the model.

For the attack on textual modality, we identify the top ϵ_t words with the highest word attention scores. We then employ the BertForMaskedLM [25] model, which predicts masked words in text by generating contextually appropriate replacement tokens. This is achieved by masking high-attention words in the input text to identify the words that are most semantically and contextually related to each identified word in the current textual description. Considering that alterations in text modalities can easily affect text readability, we limit the modification in textual modality d_t to a single token, that is, $\epsilon_t = 1$. Following this, the substitute word that induces the most pronounced effect on the fused representation—both pre and post textual attack—is determined by Eq. 5.

For the attack on visual modality, to transfer the attacks across multiple models effectively, we selectively apply perturbations of varying budgets to the attention-directed critical regions and non-attention-directed regions, which are distinguished by the parameter λ. Thus, the calculation of the adversarial perturbation can be defined as:

$$\delta_v = [(M > \lambda) \odot d_v] \oplus \delta_{crit} + d_v \oplus \delta_{n\_crit}$$
$$\text{s.t. } \|\delta_{crit}\|_\infty \leq \epsilon_{crit}, \ \|\delta_{n\_crit}\|_\infty \leq \epsilon_{n\_crit} \tag{3}$$

where $M > \lambda$ is a $0-1$ mask to decide the position that perturbed pixels lay, and $\odot$ is the element-wise multiplication, the perturbation budget $\epsilon_{crit} = \epsilon_v \times (1-r) \times \frac{\text{PX}_{\text{image}}}{\text{PX}_{\text{crit}}}$ and $\epsilon_{n\_crit} = \epsilon_v \times r$ for the attention-directed critical regions and others based on the same perturbation budget ϵ_v by setting the budget ratio r. The PX_{image} denotes the total number of image pixels. The PX_{crit} denotes the number of pixels in the critical regions $M > \lambda$.

Given that structural features contribute to robust predictions [34], [35], we believe that guiding structural perturbations to select critical regions of images can significantly improve the transferability of adversarial examples. To achieve this, we suggest incorporating the structure loss based on SSIM [36] into the optimization process, as it directly influences the structural features of images [40]:

$$\mathcal{L}_S = \frac{\left(2\mu_{d_v}\mu_{d_v \oplus \delta_v} + C_1\right)\left(2\sigma_{d_v(d_v \oplus \delta_v)} + C_2\right)}{\left(\mu_{d_v}^2 + \mu_{d_v \oplus \delta_v}^2 + C_1\right)\left(\sigma_{d_i^v}^2 + \sigma_{d_v \oplus \delta_v}^2 + C_2\right)} \tag{4}$$

where μ_{d_v}, $\mu_{d_v \oplus \delta_v}$ and σ_{d_v}, $\sigma_{d_v \oplus \delta_v}$ are the mean and standard deviations of d_v and $d_v \oplus \delta_v$ respectively, $\sigma_{d_v(d_v \oplus \delta_v)}$ is their covariance, $C_1 = 0.01^2$ and $C_2 = 0.03^2$ are used to avoid dividing by zero.

Using the aforementioned method, we first identified critical features in various VLP models and then applied a more focused perturbation to these regions from both textual and visual modalities. Simultaneously, we utilized structure loss on the visual modality to guide the addition of aggressive perturbations to the structural features of these regions. This significantly enhanced the transferability of the adversarial examples.

4.3. Orthogonal-Guided Feature Heterogenization

Beyond the modality-consistency features, it is also necessary to take full consideration of modality-discrepancy features [19], [37], [41]. Modality-discrepancy features in VLP models are defined as the features unique to a specific modality and not relied upon by models during decision-making. Given this understanding, we suggest that heterogenizing more embedded features into discrepant ones could make it difficult for VLP models to better capture aligned consistency features, thereby enhancing the attacking ability of adversarial examples against VLP models. As VLP models use modal fused representations as their basic features, we propose an orthogonalization training strategy to guide the adversarial perturbation to contain more modality-discrepancy features in the encoded embeddings.

Given that when two vectors are orthogonal, they form a right angle. We use the cosine function to calculate the orthogonality between multimodal fused representations, which yields a value of zero when the vectors are orthogonal. Specifically, in order to guide the adversarial perturbation to contain more modality-discrepancy features in the encoded embeddings and thus promote considerable attacks, we impose orthogonality on the original fused embedding $e^o = \mathcal{F}(d_v, d_t)$ and the adversarial text fused embedding $e^t = \mathcal{F}\left(d_v, d_t^{adv}\right)$. Similarly, we orthogonalize between e^o and the adversarial visual fused embedding $e^v = \mathcal{F}(d_v \oplus \delta_v, d_t)$. This can be formally expressed as:

$$\begin{aligned} \mathcal{L}_{ot} &= \frac{e^o \cdot e^t}{\max\left(\|e^o\|_2 \cdot \|e^t\|_2, \tau\right)} \\ \mathcal{L}_{ov} &= \frac{e^o \cdot e^v}{\max\left(\|e^o\|_2 \cdot \|e^v\|_2, \tau\right)} \end{aligned} \tag{5}$$

where τ is a constant that prevents the denominator from being zero. By orthogonalizing the embedded features across different modalities, this strategy guides the adversarial perturbations that perturb the benign images to contain abundant discrepant features.

Considering that the VLP model uses fused representation as a basic feature, we also need to calculate the orthogonality loss between the original fused embedding e^o and the adversarial fused embedding $e^a = \mathcal{F}\left(d_v \oplus \delta_v, d_t^{adv}\right)$, to further enhance the heterogeneity of the features. This can be defined as:

$$\mathcal{L}_{oa} = \frac{e^o \cdot e^a}{\max\left(\|e^o\|_2 \cdot \|e^a\|_2, \tau\right)} \tag{6}$$

Given the discrete nature of the textual modality and the limited perturbation space, our strategy enables the visual adversarial perturbations to include both visual and textual modality-discrepancy features, thus forcing the fused representation to be more indistinguishable and enhancing attacking ability in turn. Finally, the complete loss function can be defined as:

$$\mathcal{L}_{\mathrm{O}} = \mathcal{L}_{oa} + \mathcal{L}_{ot} + \mathcal{L}_{ov} \tag{7}$$

By intentionally incorporating additional modality-discrepancy features into the encoded embeddings, the orthogonalization strategy for fused representation enhances the attacking ability of the adversarial perturbation, making the adversarial example more aggressive.

4.4. Overall Training Process

To sum up, our proposed TMM framework first employs a cross-attention module to guide the perturbation process to disturb modality-consistency features, and then uses structure loss $\mathcal{L}_{\mathrm{S}}$ to guide adversarial perturbation that blurs the structural features inside the image regions correlated to modality-consistency features, thereby enhancing the transferability of adversarial examples. To increase the attacking ability of the adversarial examples, we further guide the adversarial perturbation to contain more modality-discrepancy features in the encoded embeddings through orthogonality loss $\mathcal{L}_{\mathrm{O}}$.

Specifically, given a VLP model $\mathcal{F}$, benign sample pairs $d = \{d_v, d_t\}$, visual and textual modality perturbation budget ϵ_v and ϵ_t. The overall optimization function can be formulated as the following equation:

$$\arg\min\left(\mathcal{L}_{\mathrm{O}} + \alpha \cdot \mathcal{L}_{\mathrm{S}}\right), \ s.t. \ \|\delta_v\|_\infty \leq \epsilon_v \, \|\delta_t\|_0 \leq \epsilon_t \tag{8}$$

where α is a hyperparameter to control the contributions of the $\mathcal{L}_{\mathrm{S}}$, with a default value of 10. The detailed training algorithm can be described as Algorithm 1 in Appendix.

5. Evaluation

In this section, we first demonstrate the effectiveness of the transferable adversarial examples generated by the baselines and our proposed TMM, in the context of the VLR tasks within VLP models, and provide an analysis of the experimental results. Then, we conduct a series of ablation studies to analyze the influence of different parameters on the experiment results. Subsequently, we further perform experimental analyses on other downstream tasks and surrogate models, and then discuss the effectiveness of our proposed TMM from the perspective of embedding features. Finally, we investigate the adversarial robustness of LVLMs in a black-box setting.

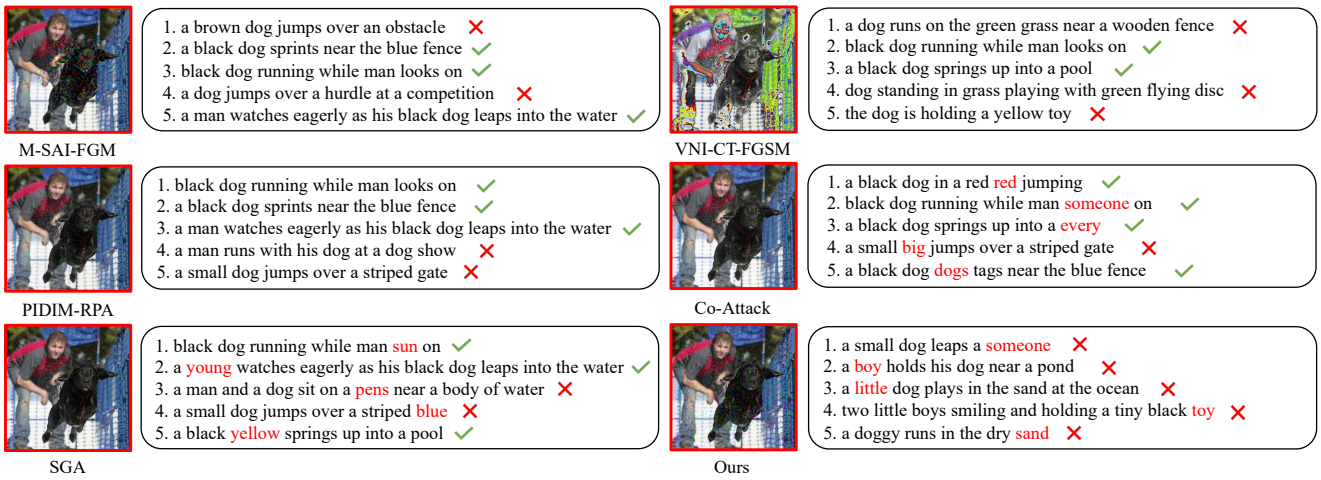


Figure 3: The adversarial example sample for image-to-text retrieval task based on Flickr30K, where the surrogate model is ALBEF and the target model is X-VLM. The right side showing the five sentences retrieved based on the image. The red square, red text, red 'X', and green checkmark represent the adversarial image, modified word, incorrectly retrieved text, and correctly retrieved text, respectively. The more irrelevant text retrieved with the image, the better the ASR performance.

5.1. Experiment Setup

Downstream tasks and datasets: Following most of the previous VLP works [12], [30]–[33], we conduct tests using the following datasets:

- **Flickr30K [42]:** This dataset, collected from the Flickr website, serves as a standard benchmark for sentence-based image descriptions and is often used in VLR tasks. The dataset includes 1,000 images and their corresponding 5,000 sentence descriptions for testing purposes.
- **MSCOCO [43]:** This dataset represents a large-scale sentence-based image description resource and serves as a benchmark dataset for VLR tasks. The dataset includes 5,000 images and 25,000 corresponding sentence descriptions for the test.
- **RefCOCO+ [44]:** This dataset is a large-scale dataset for referring expression comprehension. It is a commonly used VG task dataset containing 141,564 expressions for 19,992 images from the MSCOCO training set.
- **SNLI-VE [45]:** This dataset is specifically designed to facilitate research in VE tasks, where each data point includes a pair of sentences (a premise and a hypothesis) and an associated image. The dataset contains 1,000 images for testing.

Surrogate and victim models: For the surrogate model, we choose ALBEF [12] because it has been widely adopted in previous related work [16], [17]. For target models, considering the two dominant architectures of VLP models and feature extraction methods, we choose six typical VLP models, *i.e.*, CLIP [13], BLIP [14], METER [33] with a dual-stream structure and TCL [32], X-VLM [30], ViLT [31] with a single-stream structure. We use the corresponding pre-trained models from open-source. It should be noted that different image feature extraction modules are used in CLIP, *i.e.*, ViT-B/16 (CLIP_{ViT}) and ResNet-101 (CLIP_{CNN}), while other models adopt ViT only. Additionally, only ALBEF and TCL can handle VE tasks, while only ALBEF, TCL, and X-VLM can handle VG tasks.

Evaluation metrics: We employ the attack success rate (ASR) as the main metric for evaluating the attacking capability of the generated adversarial examples in VLP downstream tasks. This metric reflects the proportion of adversarial examples that successfully influence the model's decisions. The higher the ASR, the better the attacking ability. Due to space constraints, we only provide the mean ASR values for R@1, R@5, and R@10 in all tables for the tasks of image-to-text (TR) and text-to-image retrieval (IR), where R@N represents the top N most relevant text/image based on the image/text. Lastly, we introduce the notation used in this paper in Table 8 of the appendix.

Baselines: To demonstrate the performance improvement of our method, we compare the proposed TMM with the existing five competitive baselines as follows.

- **M-SAI-FGM [6]:** This method, guided by superpixels, adds perturbations to salient regions of the image, which not only ensures local smoothness but also modifies the image more effectively.
- **VNI-CT-FGSM [46]:** This method enhances the transferability of iterative gradient attack methods by considering the variance of the gradient from the previous iteration to adjust the current gradient during each gradient computation.
- **PIDIM-RPA [9]:** This method significantly enhances the transferability of adversarial examples based on FIA [10] by introducing patch-wise random transformations to perturb the common discriminative regions of different models.

TABLE 1: The ASR(%) results of our proposed TMM and compared baselines for VLR tasks on Flickr30K and MSCOCO. The bold numbers indicate better results, while the gray shading indicates white-box settings.

Models	Method	Flickr30K		MSCOCO	
		TR↑	IR↑	TR↑	IR↑
ALBEF [12]	M-SAI-FGM	97.11	97.03	96.33	96.82
	VNI-CT-FGSM	97.25	97.26	96.66	97.28
	PIDIM-RPA	97.20	97.18	96.47	96.87
	Co-Attack	95.64	96.52	94.63	96.59
	SGA	97.11	96.72	96.19	97.13
	Ours	**97.53**	**97.51**	**96.79**	**97.73**
TCL [32]	M-SAI-FGM	26.37	37.94	31.28	45.60
	VNI-CT-FGSM	53.11	57.35	55.87	58.97
	PIDIM-RPA	42.80	46.74	51.82	54.75
	Co-Attack	21.87	32.19	33.07	42.11
	SGA	43.95	48.83	54.04	58.82
	Ours	**64.97**	**69.60**	**70.19**	**74.02**
X-VLM [30]	M-SAI-FGM	15.63	28.94	21.13	34.09
	VNI-CT-FGSM	39.39	46.83	42.73	45.22
	PIDIM-RPA	25.67	40.89	31.75	41.66
	Co-Attack	13.43	27.45	20.43	33.35
	SGA	26.54	39.11	32.70	44.89
	Ours	**47.14**	**55.49**	**52.97**	**58.23**
CLIP_{ViT} [13]	M-SAI-FGM	31.03	44.99	39.81	48.40
	VNI-CT-FGSM	46.09	51.51	58.50	59.67
	PIDIM-RPA	38.30	45.92	48.93	55.34
	Co-Attack	21.13	33.23	39.85	48.05
	SGA	33.83	43.57	51.76	60.25
	Ours	**52.90**	**60.90**	**68.37**	**75.34**
CLIP_{CNN} [13]	M-SAI-FGM	33.67	47.27	41.62	51.77
	VNI-CT-FGSM	47.53	53.78	60.47	65.08
	PIDIM-RPA	40.22	49.04	51.32	60.14
	Co-Attack	23.43	35.26	43.47	52.11
	SGA	34.39	46.68	54.82	62.81
	Ours	**56.61**	**62.97**	**70.97**	**76.88**
BLIP [14]	M-SAI-FGM	35.07	47.85	42.24	51.78
	VNI-CT-FGSM	50.77	57.82	61.40	65.99
	PIDIM-RPA	43.10	53.71	52.90	60.89
	Co-Attack	25.39	42.84	44.31	55.61
	SGA	39.65	53.65	55.13	64.72
	Ours	**59.99**	**66.01**	**71.29**	**74.21**
ViLT [31]	M-SAI-FGM	22.00	35.57	27.71	40.93
	VNI-CT-FGSM	43.33	50.33	50.17	56.51
	PIDIM-RPA	32.94	46.76	40.16	47.62
	Co-Attack	17.67	34.81	27.01	42.34
	SGA	30.27	45.94	41.31	51.02
	Ours	**53.29**	**61.40**	**60.19**	**66.25**
METER [33]	M-SAI-FGM	19.55	32.58	-	-
	VNI-CT-FGSM	37.40	43.04	-	-
	PIDIM-RPA	31.54	38.08	-	-
	Co-Attack	15.85	29.90	-	-
	SGA	27.40	40.74		
	Ours	**49.10**	**55.49**	-	-

- **Co-Attack [16]:** This methodology pioneers exploring multimodal adversarial attacks under the VLP model. By concurrently considering both visual and textual modalities, it enhances the attacking ability of adversarial perturbations in white-box settings.
- **SGA [17]:** This frontier method builds upon Co-Attack by introducing set-level alignments to enhance the transferability of adversarial examples, but its simple interaction mechanism fails to fully utilize modality-correlated features.

Implementation details: For our proposed TMM, we have implemented three versions with different optimization methods, *i.e.*, DI-FGSM [47], DTI-FGSM [48] and DTMI-FGSM [48] (TMM-DI, TMM-DTI and TMM). For TMM-DTI, we use gaussian kernel $\mathbf{W}$ with kernel size 5×5 following [48]. For TMM-DTMI, we set the momentum decay factor γ to 1 following [49]. For visual modality, we selected 12/255 as the maximum perturbation ϵ_v, and set the step size β to 1.25. In addition, we set the perturbation ratio r, which represents the ratio between critical regions and other regions in images, to 0.4, and λ is set to 0.1 to balance the trade-off between perturbation invisibility and attacking ability. The α is set to 10. The maximum perturbation ϵ_t for text modality is set to 1 token following Co-Attack and SGA. For the baselines, it should be noted that all methods except for Co-Attack and SGA are traditional attack methods for classification tasks using cross-entropy loss, and the downstream tasks of VLP are non-classification tasks. Consequently, the same as Co-Attack, we employ the method proposed by Zhang *et al.* [50], conducting adversarial attacks by maximizing the KL divergence loss $\mathcal{L}_{\text{KL}}$ in the embedding representation, defined as:

$$\delta_v = \epsilon_v \cdot \text{sign}\left(\nabla_{d_v^{adv}} \mathcal{L}_{\text{KL}}\left(\mathcal{F}_i\left(d_v^{adv}\right), \mathcal{F}_i\left(d_v\right)\right)\right) \quad (9)$$

where $d_v^{adv} = d_v \oplus \delta_v$, $\mathcal{F}_i$ represents the image encoder output from VLP models. Furthermore, in the interest of fairness, we set the maximum perturbation budget to be the same as that of TMM, while keeping other parameters consistent with the settings in the original paper. For both white-box and black-box attacks, we faithfully follow the setting of the threat model. For VLP models, we employ open-source implementations and faithfully follow the original settings. Our experiments are conducted in a cluster of NVIDIA GeForce RTX 4090 GPUs.

5.2. Attacking performance

In this section, we conduct comprehensive experiments on Flickr30K and MSCOCO, taking VLR as a typical task. We present the performance in Table 1. The samples of the adversarial examples based on Flickr30K are shown in Fig. 3. It should be noted that we did not test on the MSCOCO dataset for the METER model due to computational resource constraints.

To sum up, our proposed TMM framework achieves significantly higher ASR values in all cases. We provide some further insights and discussions as follows:

(1) For white-box attacks, from the gray area in Table 1, it is evident that the proposed TMM framework displays significantly enhanced the attacking ability on the ALBEF model compared to the baselines. On average, TMM achieves an ASR improvement of $\mathbf{0.69}\%$ over the baselines. Specifically, it performs better than baselines by $\mathbf{0.67}\%$ and $\mathbf{0.57}\%$ on TR and IR for Flickr30K, and by $\mathbf{0.74}\%$ and $\mathbf{0.80}\%$ on TR and IR for MSCOCO, respectively. We attribute this favorable outcome to the efficacy of our TMM strategy in exploiting the modality-discrepancy features of the VLP model. The adversarial perturbation is guided by

the orthogonality training of the modal fused representation, which accentuates modality-discrepancy features and renders them negatively influential against VLP models.

(2) For black-box attacks, our TMM outperforms baselines by large margins, with an average ASR increase of $\mathbf{20.47}\%$. Specifically, on Flickr30K and MSCOCO, it achieves average increases of $+\mathbf{22.56}\%/+\mathbf{21.97}\%$ for TR and $+\mathbf{18.52}\%/+\mathbf{18.81}\%$ for IR tasks. The top-performing TMM is even up to $\mathbf{76.88}\%$ higher than the baselines. Further, despite building upon a single-stream structure, the ALBEF based TMM adversarial examples can still effectively transfer to dual-stream structure models such as CLIP, BLIP and METER. This highlights the practical effectiveness of our TMM framework, indicating the efficiency of exploiting both modality-consistency and modality-discrepancy features in generating strong transferable adversarial examples against diverse VLP models. In comparison, the Co-Attack and SGA, which are designed for multi-modality attacks, show significant defects in cross-structure transferable attacks, *i.e.*, $\mathbf{23.59}\%$ average lower than our TMM. Besides, regarding the modality characteristic, we found that some unimodal attacks are able to achieve considerable transferable attacking ability. For example, both M-SAI-FGM and PIDIM-RPA perform better than Co-Attack ($+\mathbf{7.61}\%$), and VNI-CT-FGSM even outperforms SGA ($+\mathbf{6.62}\%$). However, all the aforementioned methods exhibit weaker transferability compared with our TMM ($-\mathbf{20.47}\%$). This observation suggests that there is significant potential for multimodal tailored adversarial studies, especially transferable adversarial attacks, which have a lot of room for imagination. Further efforts to investigate modality-correlated characteristics may prove beneficial for future research.

5.3. Ablation Study

In this section, we further investigate the critical factors that influence our proposed TMM framework for generating transferable adversarial examples with strong attacking ability to VLP models.

Strategies: According to the results in Table 2, the effectiveness of the ADFP and OGFH could be demonstrated. In the white-box setting, the OGFH strategy performs best, achieving TR and IR on average $\mathbf{0.90}\%$ and $\mathbf{0.85}\%$ higher than the combined scheme. We believe this is due to the incorporation of ADFP slightly affecting the perturbation of modality-discrepancy features by adversarial pixels in OGFH. Moreover, in the black-box setting, ADFP has an average improvement of $\mathbf{11.80}\%$ and $\mathbf{11.08}\%$ in TR and IR, respectively, compared to OGFH, showing notable transferability. The results of the combined scheme demonstrate that modality-consistency and modality-discrepancy features can exploit potential vulnerabilities in VLP models, resulting in both transferability and strong attacking ability of the adversarial examples.

Optimization methods: The results of different optimization methods are shown in Table 3, which shows that TMM using DTMI outperforms TMM-DI and TMM-DTI

TABLE 2: The ASR(%) of our proposed TMM under different strategy combinations for VLR tasks on Flickr30K and MSCOCO. The complete scheme shows the best performance in all black-box settings.

Models	Strategy			Flickr30K		MSCOCO	
	Co-Attack	ADFP	OGFH	TR↑	IR↑	TR↑	IR↑
ALBEF [12]	✓			95.64	96.52	94.63	96.59
		✓		78.82	79.15	77.09	78.46
			✓	**98.21**	**98.53**	**97.91**	**98.41**
		✓	✓	97.53	97.51	96.79	97.73
TCL [32]	✓			21.87	32.19	33.07	42.11
		✓		46.04	49.24	52.10	58.09
			✓	32.90	37.45	40.02	46.53
		✓	✓	**64.97**	**69.60**	**70.19**	**74.02**
X-VLM [30]	✓			13.43	27.45	20.43	33.35
		✓		35.50	40.04	44.86	46.14
			✓	22.23	29.08	31.53	34.23
		✓	✓	**47.14**	**55.49**	**52.97**	**58.23**
CLIP_{ViT} [13]	✓			21.13	33.23	39.85	48.05
		✓		42.11	47.23	55.78	60.81
			✓	32.07	38.51	44.81	50.44
		✓	✓	**52.90**	**60.90**	**68.37**	**75.34**
CLIP_{CNN} [13]	✓			23.43	35.26	43.47	52.11
		✓		45.11	50.15	61.36	65.49
			✓	34.41	39.49	49.20	53.84
		✓	✓	**56.61**	**62.97**	**70.97**	**76.88**
BLIP [14]	✓			25.39	42.84	44.31	55.61
		✓		48.33	54.27	62.85	66.56
			✓	37.25	42.01	51.20	57.84
		✓	✓	**59.99**	**66.01**	**71.29**	**74.21**
ViLT [31]	✓			17.67	34.81	27.01	42.34
		✓		40.19	49.38	49.49	55.47
			✓	29.85	37.32	37.53	42.93
		✓	✓	**53.29**	**61.40**	**60.19**	**66.25**
METER [33]	✓			15.85	29.90	-	-
		✓		36.44	42.01	-	-
			✓	23.75	31.18	-	-
		✓	✓	**49.10**	**55.49**	-	-

in transferable attacks. This observation is consistent with previous studies [48]. Despite slightly lower performance in white-box settings compared to TMM-DI and TMM-DTI, TMM exhibits superior performance in all black-box settings, possibly due to TMM-DI and TMM-DTI focus on overfitting specific network parameters for data enhancement. However, in general, our TMM could always achieve a relatively higher ASR than the baselines.

Various ϵ_v: Fig. 4 presents the results in evaluating the performance of adversarial attacks with different ϵ_v, including 2/255, 4/255, 8/255, and 12/255, where the perturbation budget ϵ_v for other baselines are also set to the corresponding values. In white-box settings, TMM outperforms other baselines with smaller ϵ_v, and the performance becomes comparable as ϵ_v increases. In black-box settings, as ϵ_v increases, the performance of TMM and the baselines on transfer attacks gradually improves, but the degree of improvement varies. Specifically, TMM, VNI-CT-FGSM, PIDIM-RPA and SGA show significant improvements, but the adversarial perturbations of VNI-CT-FGSM are very pronounced, while other baselines show slower improvement. Overall, TMM demonstrates excellent adversarial transferability across all ϵ_v, highlighting its superiority.

Various ϵ_t: Fig. 5a shows the results of evaluating adversarial attack performance using different ϵ_t, including 1, 2,

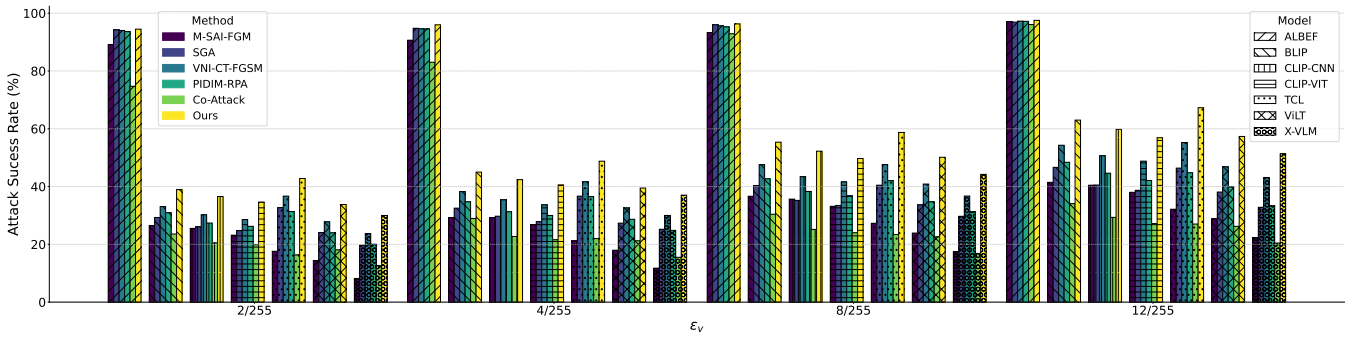
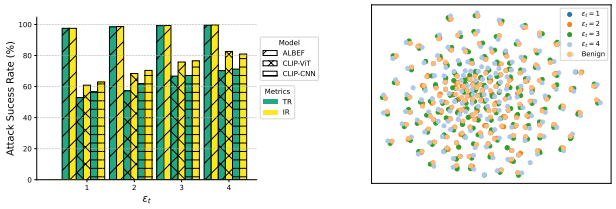


Figure 4: The attacking ability of the proposed TMM on Flickr30K under different ϵ_v settings.



(a) (b)

Figure 5: The left figure displays the attacking ability of the proposed TMM on Flickr30K under different ϵ_t settings. The right figure shows the changes in the distribution of text semantics under different ϵ_t settings.

3, and 4. In this case, the surrogate model is ALBEF, and the target models are CLIP_{CNN} and CLIP_{ViT}. In the white-box setting, the performance does not change significantly as ϵ_t increases. In the black-box setting, the transferability of TMM increases from $\epsilon_t = 1$ to $\epsilon_t = 4$, with the ASR of TR and IR respectively increasing on average by 4.78%, 7.42%, 3.8% and 7.49%, 6.78%, 5.6%. This indicates that changes in ϵ_t have a significant impact on IR.

Ratio of perturbation r: We evaluated the effect of adding perturbations to critical regions at different perturbation budget ratios (0.2 to 1.0). The $r = 1.0$ indicates adding perturbations to the entire image without distinguishing between critical regions and others. Appendix Table 7 shows the results. Under white-box settings, the ASR gradually increases and stabilizes as the ratio r increases, reaching a maximum at $r = 0.8$. Under black-box settings, the ASR decreased as the ratio r increased, demonstrating the effectiveness of our modality-consistency attack in improving transferability. Since the perturbations become more noticeable as r decreases, we choose $r = 0.4$ as the default, considering the trade-off between attacking ability and stealth.

Weighting factors α: We evaluated the results of testing different weighting factors for the combination strategy used in 8, specifically for α values of 0.01, 0.10, 1, 10, and 100, considering their numerical magnitudes. From Fig. 8 in Appendix, it can be observed that the transferability is optimal when α is 10, and as α decreases, the effect of structure loss becomes less and less. Therefore, the results are getting closer to using only the OGFH strategy, and thus the white-box attack is optimal when α is 0.01.

TABLE 3: The ASR(%) results of our proposed TMM using different optimization methods, with gray shading indicating white-box settings. The adversarial examples generated using the DTMI-FGSM exhibit better transferability.

Models	Method	Flickr30K		MSCOCO	
		TR↑	IR↑	TR↑	IR↑
ALBEF [12]	TMM-DI	**98.60**	**98.80**	**98.63**	**98.76**
	TMM-DTI	98.47	98.45	98.19	98.54
	TMM	97.53	97.51	96.79	97.73
TCL [32]	TMM-DI	56.08	58.74	66.91	68.72
	TMM-DTI	60.80	66.95	67.27	70.42
	TMM	**64.97**	**69.60**	**70.19**	**74.02**
X-VLM [30]	TMM-DI	39.27	47.42	45.94	51.95
	TMM-DTI	41.99	50.96	49.67	54.24
	TMM	**47.14**	**55.49**	**52.97**	**58.23**
CLIP_{ViT} [13]	TMM-DI	46.47	52.89	61.65	66.61
	TMM-DTI	49.65	55.61	65.89	68.96
	TMM	**52.90**	**60.90**	**68.37**	**75.34**
CLIP_{CNN} [13]	TMM-DI	47.93	56.71	63.69	70.48
	TMM-DTI	51.66	59.27	67.41	72.37
	TMM	**56.61**	**62.97**	**70.97**	**76.88**
BLIP [14]	TMM-DI	52.10	59.32	61.46	66.52
	TMM-DTI	55.17	62.52	65.11	68.36
	TMM	**59.99**	**66.01**	**71.29**	**74.21**
ViLT [31]	TMM-DI	43.88	51.37	52.08	58.24
	TMM-DTI	46.39	55.06	56.99	60.96
	TMM	**53.29**	**61.40**	**60.19**	**66.25**
METER [33]	TMM-DI	40.65	46.58	-	-
	TMM-DTI	45.16	48.68	-	-
	TMM	**49.10**	**55.49**	-	-

5.4. Analysis and Discussion

Attacks on other downstream tasks: In order to evaluate our proposed TMM framework more comprehensively, we introduced additional downstream tasks, *i.e.*, visual grounding and visual entailment, using the same parameter settings as the VLR task. The visual entailment and visual grounding results are shown in Table 4, and the samples of the adversarial examples are shown in Fig. 6. Similar

to the VLR results, our scheme outperforms the baseline under both white-box and black-box settings, which proves its outstanding performance.

Other surrogate models: To verify the generalizability of our proposed TMM across different models, we employed TCL and BLIP as surrogate models, maintaining the same parameter settings as in the main experiment. The results are presented in Table 9 and Table 10 in Appendix. Our findings indicate that the adversarial examples generated using TCL as a surrogate model have the best attack performance. We attribute this to TCL's improved ALBEF pre-training strategy, which enhances the accuracy of its cross-modal interaction module in capturing modality-consistency features. As a result, the added perturbations are more effective in improving the transferability of the adversarial examples. Moreover, we observed that adversarial examples generated using surrogate models with different architectures demonstrate similar performance in both white-box and black-box settings, underscoring the effectiveness of our proposed TMM across various surrogate models.

Embedded feature analysis: Our proposed TMM attack framework aims to perturb the multimodal embedding process by interfering with both modality-consistency and modality-discrepancy features, which can confuse the VLP model's decision-making process and make it harder to capture important features. We randomly selected 300 examples from Flickr30K and compared the t-SNE clusters of the fused embedding features before and after the attack. Fig. 7 shows that the fused embedding features of examples from TMM are perturbed more seriously, *i.e.*, the barycenter of the adversarial distribution (green points) is much farther from that of the benign distribution (orange points) compared to baselines. Additionally, Fig. 5b shows the distribution of changes in text semantics under different ϵ_t, indicating that the adversarial text does not have a huge impact on the original text semantics.

(a) Visual Grounding (b) Visual Entailment

Figure 6: The samples of visual grounding and visual entailment, and the baselines are shown in Fig. 9 and 10 in Appendix. The adversarial examples generated by TMM on ALBEF successfully make the X-VLM focus on the regions that deviate from the ground truth in the visual grounding task, *i.e.*, imprecise grounding regions, and effectively influence the predictions of TCL in the visual entailment task.

5.5. Attacks on Large Models

Large models have achieved unprecedented results in several multimodal tasks, being expected to become indispensable tools in people's daily lives. Recently, some Large generative Vision-Language Models (LVLMs) employ VLP models as encoders, such as using CLIP's visual encoder in MMGPT. Given this background, we are also curious if the proposed TMM could affect these large models that are intuitively robust. Therefore, we are motivated to conduct additional validations by introducing several LVLMs as target models. Considering the differences in training parameters between these large models and traditional models, it is challenging for individual users to train these large models locally. As a result, we contemplate evaluating our approach and the baselines on open-source and state-of-the-art LVLMs in a fully black-box setting to assess the impact of the generated adversarial examples on the LVLMs.

Specifically, we conducted black-box testing on BLIP-2 [15], MMGPT [28], Otter [29], PandaGPT [51], Visual-GLM [52] and MiniGPT-4 [53], utilizing their open-source and pre-trained weights for local deployment without any

TABLE 4: The ASR(%) results for Visual Grounding and Visual Entailment, with gray shading denoting white-box settings. The adversarial examples generated by TMM exhibit better transferability.

Models	Method	Visual Grounding		Visual Entailment
		RefCOCO+		SNLI-VE
		TestA	TestB	
ALBEF [12]	M-SAI-FGM	51.34	46.27	70.68
	VNI-CT-FGSM	63.16	55.65	89.53
	PIDIM-RPA	59.62	54.21	85.24
	Co-Attack	54.33	49.82	80.66
	SGA	55.74	50.63	86.81
	Ours	**67.14**	**59.26**	**93.36**
TCL [32]	M-SAI-FGM	39.24	32.13	37.28
	VNI-CT-FGSM	48.62	41.91	55.42
	PIDIM-RPA	45.14	39.05	51.17
	Co-Attack	40.52	34.18	40.68
	SGA	43.53	37.74	51.36
	Ours	**57.49**	**50.85**	**65.35**
X-VLM [30]	M-SAI-FGM	32.15	23.65	-
	VNI-CT-FGSM	42.49	35.32	-
	PIDIM-RPA	40.73	32.88	-
	Co-Attack	37.58	26.66	-
	SGA	39.82	32.43	-
	Ours	**53.63**	**44.18**	-

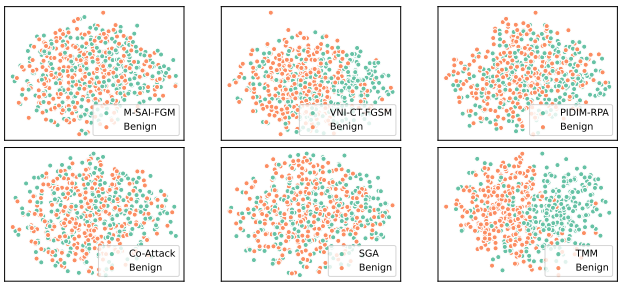


Figure 7: The feature visualization of the fused embedded vectors. The TMM generated adversarial examples have a more pronounced impact on the fused embedding output of VLP models.

TABLE 5: The results of our proposed TMM and several compared baselines for black-box attacks on LVLMs using the Flickr30K datasets, where the surrogate model is ALBEF. The total number of test examples is 5,000, with bold numbers indicating the best performance and red number indicating the ASR(%).

Method/Model	BLIP2 [15]	MMGPT [28]	Otter [29]	PandaGPT_7B [51]	PandaGPT_13B [51]	VisualGLM [52]	MiniGPT-4_7B [53]	MiniGPT-4_13B [53]
M-SAI-FGM [6]	71/1.42	404/8.08	663/13.26	88/1.76	103/2.06	80/1.60	1122/22.44	159/3.18
VNI-CT-FGSM [46]	40/0.80	340/6.80	708/14.16	288/5.76	205/4.10	59/1.18	1105/22.10	283/5.66
PIDIM-RPA [9]	127/2.54	396/7.92	554/11.08	80/1.60	95/1.90	64/1.28	1087/21.74	132/2.64
Co-Attack [16]	85/1.70	373/7.46	359/7.18	116/2.32	182/3.64	153/3.06	1062/21.24	188/3.76
SGA [17]	111/2.22	358/7.16	437/8.74	137/2.74	213/4.26	197/3.94	1116/22.32	274/5.48
Ours	**474/9.48**	**915/18.30**	**1103/22.06**	**517/10.34**	**651/13.02**	**514/10.28**	**1612/32.24**	**663/13.26**

TABLE 6: The ASR(%) results of our proposed TMM and several defense methods for VLR tasks on Flickr30K and MSCOCO. Lower ASR indicates better defense capability, the gray shading indicates white-box settings, and the red numbers indicate a performance decrease compared to the original results.

Models	Defense Method	Flickr30K		MSCOCO	
		TR↓	IR↓	TR↓	IR↓
ALBEF [12]	Bit-Red	95.73/1.80	96.29/1.22	95.93/0.86	96.76/0.97
	JPEG	96.80/0.74	96.68/0.82	96.72/0.07	97.69/0.04
	NRP	83.84/13.69	85.48/12.03	83.39/13.40	86.53/11.19
	NRP+LT	83.73/13.81	85.61/11.90	83.56/13.23	86.59/11.13
TCL [32]	Bit-Red	63.66/1.31	68.20/1.40	69.22/0.97	73.86/0.16
	JPEG	64.11/0.86	68.74/0.98	69.35/0.85	73.76/0.26
	NRP	63.03/1.94	67.43/2.17	69.00/1.19	73.31/0.71
	NRP+LT	63.10/1.87	67.38/2.22	69.41/0.78	73.54/0.48
X-VLM [30]	Bit-Red	46.21/0.93	53.89/1.59	52.66/0.31	58.03/0.20
	JPEG	46.38/0.76	54.49/1.00	52.63/0.34	57.67/0.56
	NRP	45.02/2.12	53.32/2.16	52.42/0.54	56.90/1.33
	NRP+LT	45.00/2.14	53.29/2.20	52.68/0.29	57.16/1.07
$CLIP_{ViT}$ [13]	Bit-Red	51.97/0.94	60.66/0.24	67.69/0.68	74.76/0.58
	JPEG	50.20/2.71	59.97/0.93	66.56/1.81	74.72/0.62
	NRP	52.14/0.76	60.69/0.21	65.60/2.77	74.60/0.74
	NRP+LT	52.47/0.43	60.88/0.02	65.80/2.57	74.64/0.70
$CLIP_{CNN}$ [13]	Bit-Red	52.77/3.85	61.78/1.19	70.29/0.68	76.72/0.16
	JPEG	54.05/2.56	61.76/1.21	69.05/1.92	76.12/0.76
	NRP	54.53/2.08	61.49/1.48	68.20/2.77	76.02/0.86
	NRP+LT	54.63/1.98	61.76/1.22	68.40/2.57	75.99/0.89
BLIP [14]	Bit-Red	58.55/1.44	64.93/1.08	70.82/0.47	73.95/0.26
	JPEG	58.81/1.17	65.40/0.60	70.69/0.60	73.82/0.39
	NRP	58.05/1.93	64.36/1.65	69.53/1.77	73.85/0.36
	NRP+LT	57.97/2.02	64.44/1.57	69.64/1.65	74.02/0.19
ViLT [31]	Bit-Red	52.26/1.03	60.27/1.13	59.80/0.39	65.96/0.30
	JPEG	50.79/2.50	60.01/1.39	58.99/1.19	65.61/0.64
	NRP	51.23/2.06	59.65/1.74	58.46/1.73	65.52/0.73
	NRP+LT	51.15/2.14	59.69/1.71	58.51/1.68	65.60/0.66

additional parameter tuning or fine-tuning steps. Considering the difference between VLP and LVLMs tasks, we redesigned the experiment for the VLR tasks. Specifically, we used a language template ('Does the picture depict that " ". Only answer Yes or No.') to assemble the adversarial text and combined it with the adversarial image input to ask the LVLMs model to give judgment on the question. We used the Flickr30K dataset, which includes 1k images and 5k corresponding captions, for testing. To exclude the effect of incorrect ground truth captions, we only counted the examples with different judgments between the original and adversarial examples in the results, as shown in Table 5.

The results reveal that although TMM was not specifically designed for LVLMs, it still had an impact on these large models, and TMM overall performed better than baselines. In particular, BLIP2 and PandaGPT exhibit relatively superior adversarial robustness, while the performance of MiniGPT-4, which has a performance similar to GPT-4, is not as satisfactory. Furthermore, considering that MMGPT and Otter employ CLIP's visual encoder, this reflects that the lack of robustness in LVLMs may stem from the insufficient robustness of the pre-trained model structures used, and our findings provide a new perspective for developing more robust LVLMs. Additionally, although VisualGLM and MiniGPT utilize BLIP-2's Qformer as a visual encoder, their adversarial robustness shows a significant decrease compared to BLIP-2, further suggesting that the robustness of LVLMs may also originate from their training strategies and modality interaction policies. Lastly, it is worth noting that different parameter sizes also have an impact on adversarial robustness, as MiniGPT-4_13B has better adversarial robustness compared to MiniGPT-4_7B, while PandaGPT shows the opposite trend. The adversarial robustness differences among these various LVLMs further reflect the urgent need for the research community to devote more efforts to optimizing the adversarial robustness of LVLMs.

6. Countermeasures

Facing attacks, it is necessary to discuss the adversarial defense approaches as countermeasures so that we can evade the potential negative social influence. To this end, we adopt some typical strategies to defend against our TMM attacks.

For visual modality, some research has focused on improving the robustness of adversarial examples in classification tasks [54], [55]. However, these methods require classification information, such as output probabilities and class labels, which makes them unsuitable for downstream VLP tasks. Therefore, we chose input space defense methods, including Bit-Red [56], JPEG [57], and NRP [58], which are universal adversarial attack defense methods. For textual modality, considering that traditional adversarial text defense methods are typically geared towards text classification models and require support from datasets and class labels, we opted to employ the LanguageTool[1](LT) for adversarial text correction, owing to its task- and model-agnostic nature. LT has been widely adopted in industrial settings, making it a suitable choice for our purposes.

Table 6 presents the results of various defense methods, with NRP exhibiting superior defense performance compared to Bit-Red and JPEG. Consequently, we integrated

1. https://github.com/languagetool-org/languagetool

NRP with LT for defense validation, denoted as NRP+LT in the table. The results indicate that even when LT is employed for adversarial text correction, our TMM still demonstrates strong attack capabilities (an average of $\mathbf{65.60}\%$). Upon this observation, we believe that more countermeasures should be taken when facing such adversarial attack methods like our TMM. Beyond the mentioned input space defense, it is also worth noting that adversarial training has been demonstrated as the most effective defense against adversarial attacks [59], especially in the field of self-supervised learning related to VLP [60], [61]. A series of studies have demonstrated that adversarial examples could be beneficial for model robustness via standard adversarial training scheme [50], [59], [60]. Therefore, although we are unable to conduct adversarial training due to computational resource constraints to verify its real effectiveness, we believe that it could help models to obtain considerable robustness when facing our TMM.

7. Related Work

Adversarial attack: Although deep neural networks (DNNs) have been successful in various areas such as image, text, and multimodal tasks, recent research has shown that they are vulnerable to human-imperceptible adversarial examples that can cause DNNs to make a mistake [2], [6], [9], [10], [47], [62], [63]. In general, adversarial attacks can be divided into white-box attacks and black-box attacks. In white-box attacks, the attacker has full access to the target model. For example, [1] proposed the PGD method, which is currently the strongest first-order attack. In black-box attacks, the attacker cannot directly access the target model, thus in an agnostic state to the target model. Black-box attacks can be divided into two categories, *i.e.*, query-based and transfer-based. Query-based attacks rely on the output scores or labels of the target network, which limits their applicability in the real world [64], [65]. Transfer-based attacks generate adversarial perturbations on the source model and then transfer them to an unknown target model [2], [66]. In addition, adversarial attacks in the discrete space still pose a challenge in NLP. Adversarial attacks in NLP will substitute some tokens of the input text to maximize the risk of embedded errors in the output [67]–[70].

Previous work in the field of multimodal focused on attacking multimodal models by selecting one single modality for perturbation [71]–[73]. However, these methods only target a single V+L task, and are essentially standard unimodal attacks, which have been shown to be susceptible to attack failures [16]. Regarding multimodal attacks against multimodal models, Zhang *et al.* [16] groundbreakingly investigated the robustness of VLP models to adversarial attacks and proposed a novel multimodal adversarial attack method under the white-box setting. Zhao *et al.* [74] first explored the adversarial robustness of LVLMs and proposed a query-based targeted black-box adversarial examples generation method. Lu *et al.* [17] built upon Co-Attack by introducing set-level alignments to enhance adversarial transferability. Based on the multimodal attack assumption of the multimodal pre-trained models, this paper explores the transferable attacking method on the VLP tasks from modality-correlated features.

Adversarial defense: Many adversarial defense methods have been proposed to mitigate the threat of adversarial attacks. One promising defense method is adversarial training [5], [50], [59], which injects adversarial examples into the training data to improve the model's robustness. However, adversarial training, as one of the most powerful and extensively studied defense methods, often incurs high computational costs and is difficult to scale up to large datasets and complex neural networks [75]. In addition, there are defense methods in the input space. For example, [57] uses a set of image transformations (*e.g.*, JPEG compression) on the inputs before feeding the image to the model. Xu *et al.* [56] propose two feature squeezing methods: bit reduction (Bit-Red) and spatial smoothing to detect adversarial examples. Naseer *et al.* [58] design a neural representation purifier (NRP) model that learns to purify the adversarially perturbed images based on the automatically derived supervision. In addition, defenses against textual adversarial attacks can be broadly categorized into passive and active defenses. Passive methods detect adversarial inputs during inference [76], [77], while active methods typically enhance the robustness of the model during training [78], [79].

8. Conclusion

In this work, we generate transferable adversarial examples against VLP models by taking advantage of both modality-consistency and modality-discrepancy features. Specifically, we first design an attention-directed feature perturbation strategy to promote transferability. Further, we elaborate an orthogonal-guided feature heterogenization strategy to amplify the attacking ability. Extensive experiments have been conducted under both white and black-box settings, demonstrating that the proposed TMM attack framework outperforms the baselines by large margins ($\mathbf{20.47}\%$ ASR improvement on average). Additionally, by introducing LVLMs into evaluation, we reveal the potential adversarial risks of these popular and burgeoning models. To mitigate potential misuse of TMM, we explored data-end defenses and suggested adversarial training [50], [59] for improved robustness. Furthermore, we only partially considered semantic relevance in the textual adversarial attack, which exists instances that result in adversarial text not matching the original text semantically. Moreover, a small proportion of images with minimal critical regions may result in compromised stealth for adversarial examples, necessitating further research and improvement. Beyond the limitations, we highlight that the modality-consistency and modality-discrepancy features do play important roles in VLP models, which can be considered as further exploration directions for enhancing the adversarial robustness of VLP models. We hope our observations could inspire more future research to explore better training strategies and model architectures for reliable VLP models, including large models.

Acknowledgments

This work was supported by the Zhongguancun Laboratory, the National Natural Science Foundation of China (Grant No. 62072098, 62022009, 62372102, 62061146001), Jiangsu Key R&D Program Grant No. BE2022065-5, the State Key Laboratory of Software Development Environment (SKLSDE-2022ZX-23), Key Laboratory of Computer Network and Information Integration of Ministry of Education of China (No. 93K-9), the Jiangsu Provincial Key Laboratory of Network and Information Security (BM2003201), Collaborative Innovation Center of Novel Software Technology and Industrialization, the MindSpore, the CANN, and the Ascend AI Processor.

References

[1] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[2] Y. Long, Q. Zhang, B. Zeng, L. Gao, X. Liu, J. Zhang, and J. Song, "Frequency domain model augmentation for adversarial attack," in *European Conference on Computer Vision*, pp. 549–566, 2022.

[3] D. Wu, S.-T. Xia, and Y. Wang, "Adversarial weight perturbation helps robust generalization," *Neural Information Processing Systems*, 2020.

[4] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," *IEEE Symposium on Security and Privacy*, 2017.

[5] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv: Machine Learning*, 2014.

[6] X. Dong, J. Han, D. Chen, J. Liu, H. Bian, Z. Ma, H. Li, X. Wang, W. Zhang, and N. Yu, "Robust superpixel-guided attentional adversarial attack," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12895–12904, 2020.

[7] J. Wang, A. Liu, Z. Yin, S. Liu, S. Tang, and X. Liu, "Dual attention suppression attack: Generate adversarial camouflage in physical world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8565–8574, 2021.

[8] Y. Wang, J. Wang, Z. Yin, R. Gong, J. Wang, A. Liu, and X. Liu, "Generating transferable adversarial examples against vision transformers," in *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5181–5190, 2022.

[9] Y. Zhang, Y.-a. Tan, T. Chen, X. Liu, Q. Zhang, and Y. Li, "Enhancing the transferability of adversarial examples with random patch," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 1672–1678, 2022.

[10] Z. Wang, H. Guo, Z. Zhang, W. Liu, Z. Qin, and K. Ren, "Feature importance-aware transferable adversarial attacks," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7639–7648, 2021.

[11] A. Jain, M. Guo, K. Srinivasan, T. Chen, S. Kudugunta, C. Jia, Y. Yang, and J. Baldridge, "Mural: multimodal, multitask retrieval across languages," *arXiv preprint arXiv:2109.05125*, 2021.

[12] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.

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

[14] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*, pp. 12888–12900, PMLR, 2022.

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

[16] J. Zhang, Q. Yi, and J. Sang, "Towards adversarial attack on vision-language pre-training models," in *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5005–5013, 2022.

[17] D. Lu, Z. Wang, T. Wang, W. Guan, H. Gao, and F. Zheng, "Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models," *arXiv preprint arXiv:2307.14061*, 2023.

[18] Y. Gou, T. Ko, H. Yang, J. Kwok, Y. Zhang, and M. Wang, "Leveraging per image-token consistency for vision-language pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19155–19164, 2023.

[19] M. Ye, X. Lan, Q. Leng, and J. Shen, "Cross-modality person re-identification via modality-aware collaborative ensemble learning," *IEEE Transactions on Image Processing*, pp. 9387–9399, 2020.

[20] F.-L. Chen, D.-Z. Zhang, M.-L. Han, X.-Y. Chen, J. Shi, S. Xu, and B. Xu, "Vlp: A survey on vision-language pre-training," *Machine Intelligence Research*, vol. 20, no. 1, pp. 38–56, 2023.

[21] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, "A comprehensive survey on cross-modal retrieval," *arXiv preprint:1607.06215*, 2016.

[22] R. Hong, D. Liu, X. Mo, X. He, and H. Zhang, "Learning to compose and reason with language tree structures for visual grounding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 2, pp. 684–696, 2019.

[23] N. Xie, F. Lai, D. Doran, and A. Kadav, "Visual entailment: A novel task for fine-grained image understanding," *arXiv preprint arXiv:1901.06706*, 2019.

[24] OpenAI, "Gpt-4 technical report," *ArXiv*, vol. abs/2303.08774, 2023.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *North American Chapter of the Association for Computational Linguistics*, 2018.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[27] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[28] T. Gong, C. Lyu, S. Zhang, Y. Wang, M. Zheng, Q. Zhao, K. Liu, W. Zhang, P. Luo, and K. Chen, "Multimodal-gpt: A vision and language model for dialogue with humans," 2023.

[29] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu, "Otter: A multi-modal model with in-context instruction tuning," 2023.

[30] Y. Zeng, X. Zhang, and H. Li, "Multi-grained vision language pre-training: Aligning texts with visual concepts," *arXiv preprint arXiv:2111.08276*, 2021.

[31] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International Conference on Machine Learning*, pp. 5583–5594, PMLR, 2021.

[32] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, and J. Huang, "Vision-language pre-training with triple contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15671–15680, 2022.

[33] Z.-Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, P. Zhang, L. Yuan, N. Peng, *et al.*, "An empirical study of training end-to-end vision-and-language transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18166–18176, 2022.

[34] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," *arXiv preprint arXiv:1811.12231*, 2018.

[35] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. Shahbaz Khan, and M.-H. Yang, "Intriguing properties of vision transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23296–23308, 2021.

[36] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," *Asilomar Conference on Signals, Systems and Computers*, 2003.

[37] X. Lan, M. Ye, R. Shao, B. Zhong, P. C. Yuen, and H. Zhou, "Learning modality-consistency feature templates: A robust rgb-infrared tracking system," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9887–9897, 2019.

[38] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, vol. 8, no. 3, pp. 331–368, 2022.

[39] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12116–12128, 2021.

[40] M. Z. Hameed and A. Gyorgy, "Perceptually constrained adversarial attacks," *arXiv preprint arXiv:2102.07140*, 2021.

[41] Z. Zhao, B. Liu, Q. Chu, Y. Lu, and N. Yu, "Joint color-irrelevant consistency learning and identity-aware modality adaptation for visible-infrared cross modality person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 3520–3528, 2021.

[42] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

[43] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," *Lecture Notes in Computer Science*, 2014.

[44] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," *Cornell University - arXiv*, 2016.

[45] N. Xie, F. Lai, D. Doran, and A. Kadav, "Visual entailment: A novel task for fine-grained image understanding," *arXiv: Computer Vision and Pattern Recognition*, 2019.

[46] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1924–1933, 2021.

[47] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2730–2739, 2019.

[48] F. Liu, C. Zhang, and H. Zhang, "Towards transferable unrestricted adversarial examples with minimum changes," in *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 327–338, IEEE, 2023.

[49] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," *Cornell University - arXiv*, 2017.

[50] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International conference on machine learning*, pp. 7472–7482, PMLR, 2019.

[51] Y. Su, T. Lan, H. Li, J. Xu, Y. Wang, and D. Cai, "Pandagpt: One model to instruction-follow them all," *arXiv preprint arXiv:2305.16355*, 2023.

[52] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang, "Glm: General language model pretraining with autoregressive blank infilling," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 320–335, 2022.

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

[54] S. Dathathri, S. Zheng, T. Yin, R. M. Murray, and Y. Yue, "Detecting adversarial examples via neural fingerprinting," *arXiv preprint arXiv:1803.03870*, 2018.

[55] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *Advances in neural information processing systems*, vol. 32, 2019.

[56] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," *arXiv preprint arXiv:1704.01155*, 2017.

[57] C. Guo, M. Rana, M. Cisse, and L. Van Der Maaten, "Countering adversarial images using input transformations," *arXiv preprint arXiv:1711.00117*, 2017.

[58] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, "A self-supervised approach for adversarial robustness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 262–271, 2020.

[59] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[60] L. Fan, S. Liu, P.-Y. Chen, G. Zhang, and C. Gan, "When does contrastive learning preserve adversarial robustness from pretraining to finetuning?," *Advances in neural information processing systems*, vol. 34, pp. 21480–21492, 2021.

[61] Z. Jiang, T. Chen, T. Chen, and Z. Wang, "Robust pre-training by adversarial contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 16199–16210, 2020.

[62] X. Wang, S. Lin, H. Zhang, Y. Zhu, and Q. Zhang, "Interpreting attributions and interactions of adversarial attacks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1095–1104, 2021.

[63] Z. Wang, H. Guo, Z. Zhang, W. Liu, Z. Qin, and K. Ren, "Feature importance-aware transferable adversarial attacks," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7639–7648, 2021.

[64] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, "Simple black-box adversarial attacks," in *International Conference on Machine Learning*, pp. 2484–2493, PMLR, 2019.

[65] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *International conference on machine learning*, pp. 2137–2146, PMLR, 2018.

[66] S. Fang, J. Li, X. Lin, and R. Ji, "Learning to learn transferable attack," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 571–579, 2022.

[67] L. Li, R. Ma, Q. Guo, X. Xue, and X. Qiu, "Bert-attack: Adversarial attack against bert using bert," *arXiv preprint arXiv:2004.09984*, 2020.

[68] B. Liang, H. Li, M. Su, P. Bian, X. Li, and W. Shi, "Deep text classification can be fooled," *arXiv preprint arXiv:1704.08006*, 2017.

[69] M. Cheng, J. Yi, P.-Y. Chen, H. Zhang, and C.-J. Hsieh, "Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 3601–3608, 2020.

[70] J. Li, S. Ji, T. Du, B. Li, and T. Wang, "Textbugger: Generating adversarial text against real-world applications," *arXiv preprint arXiv:1812.05271*, 2018.

[71] T. Kim and J. Ghosh, "On single source robustness in deep fusion models," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[72] M. Shah, X. Chen, M. Rohrbach, and D. Parikh, "Cycle-consistency for robust visual question answering," *arXiv: Computer Vision and Pattern Recognition*, 2019.

[73] K. Yang, W.-Y. Lin, M. Barman, F. Condessa, and Z. Kolter, "Defending multimodal fusion models against single-source adversaries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3340–3349, 2021.

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

[75] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[76] M. Mozes, P. Stenetorp, B. Kleinberg, and L. D. Griffin, "Frequency-guided word substitutions for detecting textual adversarial examples," *arXiv preprint arXiv:2004.05887*, 2020.

[77] Y. Zhou, J.-Y. Jiang, K.-W. Chang, and W. Wang, "Learning to discriminate perturbations for blocking adversarial attacks in text classification," *arXiv preprint arXiv:1909.03084*, 2019.

[78] D. Kang, T. Khot, A. Sabharwal, and E. Hovy, "Adventure: Adversarial training for textual entailment with knowledge-guided examples," *arXiv preprint arXiv:1805.04680*, 2018.

[79] Y. Zhou, X. Zheng, C.-J. Hsieh, K.-w. Chang, and X. Huang, "Defense against adversarial attacks in nlp via dirichlet neighborhood ensemble," *arXiv preprint arXiv:2006.11627*, 2020.

Appendix A.

Due to space limitations, in this section, we provide a detailed training algorithm for TMM in Algorithm 1, experimental results under various rations r and α settings in Table 7 and Fig. 8, respectively. Furthermore, we present the notation used throughout this paper in Table 8. The adversarial sample results for visual grounding and visual entailment are respectively shown in Fig. 9 and Fig. 10, as well as experimental results using TCL and BLIP models as surrogate models in Tables 9 and 10, respectively.

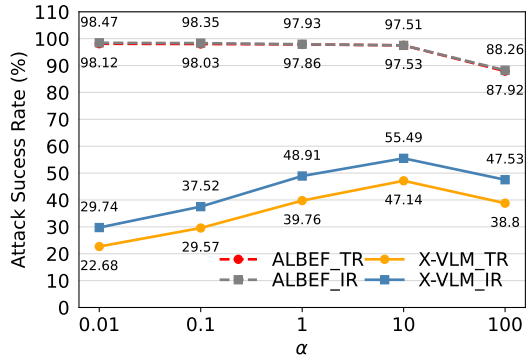


Figure 8: The performance of adversarial attacks on Flickr30K with various α.

Algorithm 1: Transferable Multimodal Attack

Input: The original clean image d_v and clean text d_t pair, max perturbation ϵ_v and ϵ_t, perturbation ratio r, step size β, decay factor γ.

Output: The adversarial image d_v^{adv} and adversarial text d_t^{adv} pair.

Initialize: $\delta_v = 0$, $\delta_t = \emptyset$, $\boldsymbol{m}_0 = 0$

Get the word attention score and image attention map by Eq. (2) ;

Get loss function $\mathcal{L}$ by Eq. (8);

Get the textual adversarial perturbation δ_t and adversarial text d_t^{adv} through ADFP 4.2;

Update d_v^{adv} by momentum iterative method:

for $i = 0 \rightarrow I - 1$ **do**

$$\boldsymbol{m}_{i+1} = \gamma \cdot \boldsymbol{m}_i + \frac{\mathbf{W} * \nabla_{\boldsymbol{d}_{v,i}^{adv}} \mathcal{L}}{\left\| \mathbf{W} * \nabla_{\boldsymbol{d}_{v,i}^{adv}} \mathcal{L} \right\|_1}$$

$$\boldsymbol{d}_{v,i+1}^{adv} = \mathrm{Clip}_{\boldsymbol{d}_v, \epsilon_v} \left(\boldsymbol{d}_{v,i}^{adv} + \beta \cdot \mathrm{sign}\left(\boldsymbol{m}_{i+1} \right) \right)$$

end

return d_v^{adv}, d_t^{adv}

TABLE 7: The experimental results of studying the effectiveness of various rations r, with gray shading denoting white-box settings.

r	Models	Flickr30K		MSCOCO	
		TR↑	IR↑	TR↑	IR↑
0.2	ALBEF [12]	96.67	96.81	96.10	97.04
	TCL [32]	68.49	73.37	74.04	78.03
	X-VLM [30]	49.61	58.49	55.87	61.39
	CLIP_{ViT} [13]	55.06	64.20	72.12	78.80
	CLIP_{CNN} [13]	58.96	66.39	74.86	81.05
	BLIP [14]	63.22	69.58	75.20	78.23
	ViLT [31]	56.97	64.72	63.49	69.84
0.4	ALBEF [12]	97.53	97.51	96.79	97.73
	TCL [32]	64.97	69.60	70.19	74.02
	X-VLM [30]	47.14	55.49	52.97	58.23
	CLIP_{ViT} [13]	52.90	60.90	68.37	75.34
	CLIP_{CNN} [13]	56.61	62.97	70.97	76.88
	BLIP [14]	59.99	66.01	71.29	74.21
	ViLT [31]	53.29	61.40	60.19	66.25
0.6	ALBEF [12]	97.53	97.67	96.95	97.90
	TCL [32]	62.61	67.25	67.71	71.54
	X-VLM [30]	45.41	53.56	51.00	56.22
	CLIP_{ViT} [13]	50.25	58.81	65.94	71.64
	CLIP_{CNN} [13]	53.84	60.82	68.46	74.31
	BLIP [14]	57.76	63.77	68.77	71.72
	ViLT [31]	50.41	59.29	58.00	64.01
0.8	ALBEF [12]	97.87	98.01	97.29	98.24
	TCL [32]	60.44	64.37	65.59	68.65
	X-VLM [30]	43.04	50.68	49.71	53.34
	CLIP_{ViT} [13]	47.96	55.93	63.81	68.76
	CLIP_{CNN} [13]	51.59	57.94	66.36	72.43
	BLIP [14]	55.54	60.88	65.92	68.84
	ViLT [31]	48.09	56.41	55.79	61.12
1	ALBEF [12]	97.85	97.99	97.27	98.22
	TCL [32]	54.49	59.31	60.48	64.38
	X-VLM [30]	40.87	46.17	48.28	50.67
	CLIP_{ViT} [13]	45.51	51.21	60.77	65.69
	CLIP_{CNN} [13]	49.92	54.14	64.22	69.91
	BLIP [14]	53.79	57.96	64.31	70.55
	ViLT [31]	45.75	54.67	52.07	58.15

TABLE 8: Summary of Notations

Symbol	Meaning
d, d_v, d_t	Benign image-text pair, image data, text data
δ, ϵ	Adversarial perturbation, perturbation budget
$\mathcal{F}$, e	The VLP model, fused embedding
m, M, W	Attention feature vector, attention score, weight matrix
Q, K, V	Feature matrix of cross-attention module
$\sqrt{d}$, λ	Scaling factor, parameter to distinguished critical regions
μ, δ, σ	Mean value, standard deviations, covariance
$\mathbf{W}$, γ, β,	Gaussian kernel, momentum decay factor, step size
r, α	The ratio between critical regions and other regions in images, hyperparameter
ASR, R@N	Attack success rate, the top N most relevant text/image based on the image/text
VLR, TR, IR	Vision-language retrieval, image-to-text retrieval, text-to-image retrieval

Figure 9: The visual grounding samples on baseline methods, where the surrogate model is ALBEF, the target model is X-VLM, both SGA and VNI-CT-FGSM demonstrate better performance compared to other baselines.

Figure 10: The visual entailment samples on baseline methods, where the surrogate model is ALBEF, the target model is TCL, both PIDIM-RPA and VNI-CT-FGSM demonstrate better performance compared to other baselines.

TABLE 9: The ASR(%) results of our proposed TMM and compared baselines for VLR tasks on Flickr30K and MSCOCO, where the surrogate model is TCL. The bold numbers indicate the best performance, while the gray shading indicates white-box settings.

Models	Method	Flickr30K		MSCOCO	
		TR↑	IR↑	TR↑	IR↑
TCL [32]	M-SAI-FGM	97.27	97.17	96.52	97.03
	VNI-CT-FGSM	97.43	97.48	96.86	97.47
	PIDIM-RPA	97.34	97.37	96.57	97.37
	Co-Attack	95.97	96.71	94.82	96.80
	SGA	97.37	96.95	96.73	97.32
	Ours	**97.87**	**97.60**	**97.00**	**97.92**
ALBEF [12]	M-SAI-FGM	27.61	39.38	33.77	48.25
	VNI-CT-FGSM	55.68	59.57	61.92	69.09
	PIDIM-RPA	48.96	52.05	53.39	58.01
	Co-Attack	22.88	33.40	34.65	44.56
	SGA	47.69	56.04	58.29	68.80
	Ours	**68.10**	**72.30**	**73.62**	**78.38**
X-VLM [30]	M-SAI-FGM	18.12	30.10	22.09	35.46
	VNI-CT-FGSM	41.26	48.71	44.68	48.03
	PIDIM-RPA	28.44	44.47	35.86	44.34
	Co-Attack	14.52	28.54	21.38	35.76
	SGA	30.09	43.67	37.04	47.81
	Ours	**49.17**	**57.71**	**54.56**	**60.56**
$CLIP_{ViT}$ [13]	M-SAI-FGM	32.82	46.97	41.18	50.52
	VNI-CT-FGSM	48.02	53.75	62.44	65.24
	PIDIM-RPA	43.20	50.21	51.38	58.60
	Co-Attack	22.63	35.40	41.63	51.20
	SGA	36.71	47.32	54.62	62.36
	Ours	**54.63**	**63.52**	**70.60**	**78.98**
$CLIP_{CNN}$ [13]	M-SAI-FGM	35.64	50.04	43.03	54.09
	VNI-CT-FGSM	49.76	56.15	64.11	66.92
	PIDIM-RPA	45.55	52.34	54.88	62.39
	Co-Attack	24.04	39.09	44.97	54.33
	SGA	38.20	49.32	57.63	64.95
	Ours	**58.87**	**65.71**	**73.27**	**80.21**
BLIP [14]	M-SAI-FGM	37.08	49.99	43.68	54.11
	VNI-CT-FGSM	53.26	60.35	63.41	67.92
	PIDIM-RPA	48.14	56.20	55.93	62.33
	Co-Attack	29.78	44.78	45.66	57.17
	SGA	43.30	55.76	56.85	65.15
	Ours	**62.69**	**68.87**	**73.60**	**77.43**
ViLT [31]	M-SAI-FGM	23.63	37.22	28.70	42.82
	VNI-CT-FGSM	45.60	53.23	51.84	59.27
	PIDIM-RPA	38.68	49.86	42.62	51.89
	Co-Attack	24.16	36.42	27.99	44.29
	SGA	39.28	48.19	43.28	52.80
	Ours	**55.71**	**64.07**	**62.16**	**69.16**

TABLE 10: The ASR(%) results of our proposed TMM and compared baselines for VLR tasks on Flickr30K and MSCOCO, where the surrogate model is BLIP. The bold numbers indicate the best performance, while the gray shading indicates white-box settings.

Models	Method	Flickr30K		MSCOCO	
		TR↑	IR↑	TR↑	IR↑
BLIP [14]	M-SAI-FGM	97.01	96.84	96.24	96.19
	VNI-CT-FGSM	97.16	97.07	96.56	96.65
	PIDIM-RPA	97.02	96.87	96.46	96.58
	Co-Attack	95.51	96.33	94.54	95.96
	SGA	97.11	96.56	96.12	96.32
	Ours	**97.27**	**97.31**	**96.70**	**97.11**
ALBEF [12]	M-SAI-FGM	24.42	36.84	29.35	43.00
	VNI-CT-FGSM	46.99	53.84	53.32	58.12
	PIDIM-RPA	40.12	46.93	48.82	50.87
	Co-Attack	22.33	32.09	32.36	41.92
	SGA	42.01	46.71	49.95	54.75
	Ours	**55.12**	**63.40**	**63.95**	**69.47**
TCL [32]	M-SAI-FGM	22.96	35.10	27.70	45.69
	VNI-CT-FGSM	44.63	51.42	50.19	56.37
	PIDIM-RPA	38.04	43.89	46.78	49.37
	Co-Attack	18.96	27.74	30.40	45.89
	SGA	39.37	43.81	47.81	49.17
	Ours	**49.87**	**59.31**	**57.76**	**63.86**
X-VLM [30]	M-SAI-FGM	14.22	26.66	19.36	33.29
	VNI-CT-FGSM	36.12	42.23	40.00	42.18
	PIDIM-RPA	24.31	37.75	31.60	37.73
	Co-Attack	12.89	21.86	19.47	30.03
	SGA	25.43	37.88	30.79	41.27
	Ours	**43.06**	**51.44**	**48.51**	**55.36**
$CLIP_{ViT}$ [13]	M-SAI-FGM	29.05	42.77	36.91	46.97
	VNI-CT-FGSM	43.99	47.96	54.48	57.24
	PIDIM-RPA	35.27	42.70	45.06	49.78
	Co-Attack	19.65	29.97	35.81	44.62
	SGA	31.92	42.24	44.56	57.18
	Ours	**49.16**	**54.89**	**63.76**	**72.91**
$CLIP_{CNN}$ [13]	M-SAI-FGM	30.12	44.25	36.29	48.20
	VNI-CT-FGSM	41.30	50.44	54.01	63.77
	PIDIM-RPA	34.91	47.74	45.29	58.12
	Co-Attack	18.50	30.60	36.89	42.46
	SGA	34.17	44.09	48.43	61.37
	Ours	**51.25**	**59.17**	**67.88**	**74.45**
ViLT [31]	M-SAI-FGM	17.38	33.15	24.10	35.22
	VNI-CT-FGSM	38.42	47.16	45.22	50.80
	PIDIM-RPA	29.63	40.81	36.11	43.84
	Co-Attack	14.70	32.42	23.45	40.91
	SGA	26.85	39.94	39.85	46.37
	Ours	**48.05**	**57.68**	**54.64**	**61.51**

Appendix B.
Meta-Review

The following meta-review was prepared by the program committee for the 2024 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

B.1. Summary

The paper proposes a transferable multimodal (TTM) attack framework for improving the transferability of black-box adversarial examples against vision-language pre-trained models. TMM uses attention-directed feature perturbations to target the modality-consistency features for improving transferability and it uses orthogonal-guided feature heterogenization to target modality-discrepancy features to improve attack success. This work shows strong empirical results, outperforming existing black-box attacks across several models.

B.2. Scientific Contributions

- Provides a Valuable Step Forward in an Established Field
- Establishes a New Research Direction

B.3. Reasons for Acceptance

1) This paper provides a valuable step forward in the literature of adversarial examples. The multi-modal attack algorithm presented here outperforms existing black-box attacks significantly across many models. There are strong empirical results: experiments are performed across 8 models (7 in the black-box setting, with one model used as the surrogate) and two datasets; further, the attack shows significant success improvements (often by 10-20 percentage points).
2) This paper establishes new ideas for multi-modal attacks. This paper proposes and demonstrates that modality-consistency and modality-discrepancy have significant impact on multi-modal adversarial examples. Ablations verify the importance of each for transferability and attack success, respectively.