Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.081490

ARTICLE

Adversarial Example Transfer Method for Vision-Language Pre-Training Models Based on Negative Sample Feature Perturbation

Zhichao Pei, Ou Ye*, Panyu Yang and Kaiwen He

College of Artificial Intelligence & Computer Science, Xi'an University of Science and Technology, Xi'an, China

*Corresponding Author: Ou Ye. Email: oye0928@xust.edu.cn

Received: 03 March 2026; Accepted: 17 April 2026; Published: 15 June 2026

ABSTRACT: To address the issue of insufficient transferability of existing adversarial example generation methods for vision-language pre-training (VLP) models, this paper proposes an adversarial example transfer method for VLP models based on negative sample feature perturbation. First, a novel cross-modal collaborative perturbation strategy is constructed. By introducing negative samples into the cross-modal perturbation mechanism, the strategy explores more perturbation directions, breaks the original modal alignment constraints and avoids the local focus of adversarial perturbations. Then, to reduce the computational cost, a dynamic threshold attack strategy is built to measure the modal similarity of the generated adversarial examples. Finally, with the help of a multi-modal fusion encoder, a cross-modal fusion semantic attack (CFSA) module is designed. This module extracts the middle-layer features of image-text pairs and improves the transfer attack effect of adversarial examples. The proposed attack method is experimentally evaluated on the Flickr30K and MSCOCO datasets. The results show that for the adversarial examples generated on the Flickr30K dataset, the attack success rate (ASR) of the proposed method reaches up to 95.3% on multiple black-box models; for those generated on the MSCOCO dataset, the maximum attack success rate on multiple black-box models reaches 70.17%. Compared with the current methods, the adversarial examples generated by the proposed method achieve better attack performance.

KEYWORDS: Vision-language pre-training model; multimodal; adversarial attack transferability; cross-modality perturbation; negative samples

1 Introduction

In recent years, the emergence of Vision–Language Pre-training (VLP) models [1] leads to significant progress in multimodal tasks within the field of artificial intelligence, including image–text retrieval and visual question answering. VLP models integrate multimodal data such as images and text and capture cross-modal semantic correlations. They not only achieve precise understanding and joint reasoning over "image content–text description" pairs, but also efficiently accommodate the demand for multi-source information fusion in real-world scenarios. However, existing studies [2] demonstrate that VLP models still exhibit security vulnerabilities and remain susceptible to adversarial perturbations. For example, in autonomous driving systems [3], the addition of subtle and imperceptible perturbations to traffic signs causes VLP models to misclassify them, potentially resulting in severe traffic accidents. In medical diagnosis [4], adversarial examples are likely to induce incorrect diagnostic outputs, thereby leading to serious medical consequences. The limited robustness caused by adversarial examples introduces substantial safety risks for the practical deployment of VLP models. Adversarial examples serve to evaluate VLP models, uncover latent

vulnerabilities, and further enhance model security. Therefore, research into adversarial example generation methods possesses considerable theoretical value and practical significance.

At present, adversarial attacks are categorized into white-box attacks and black-box attacks [5]. Black-box attacks, which do not require access to internal model parameters, more accurately reflect real-world attack scenarios. Black-box attacks are further divided into query-based attacks and transfer-based attacks. Query-based attacks [6–8] iteratively optimize adversarial examples through repeated queries to obtain model feedback; however, in practical settings, the number of permissible queries is often limited. Consequently, existing research primarily concentrates on transfer-based attacks. Transfer-based attacks [9] refer to the generation of adversarial examples on a white-box model (i.e., a surrogate model) in order to attack an unknown black-box model (i.e., a target model). The effectiveness of transfer-based attacks depends critically on the transferability of adversarial examples [10], which enables them to generalize across different models and has become a major research focus in recent years.

Transfer-based attack methods targeting VLP models are generally divided into two categories [11]: universal adversarial patch attacks and modality joint attack methods for adversarial example generation. Universal adversarial patch attacks aim to generate input-agnostic, fixed perturbations that can be superimposed on arbitrary images to mislead the target model. By contrast, modality joint attacks employ a surrogate model to generate adversarial examples and seek to achieve improved attack performance from the perspective of cross-modal interaction. Existing research on universal adversarial patch attacks remains confined to generating perturbation patches within the image modality. Compared with universal adversarial patch attacks, modality joint attacks are inherently more challenging. They require consideration of inter-modal feature interactions and achieve improved transfer performance by mutually perturbing image–text pairs. Accordingly, numerous studies investigate different strategies to enhance the transferability of modality joint attacks.

Among adversarial example generation methods based on modality joint attacks, Zhang et al. [12] investigate attack strategies against VLP models. Their results indicate that attacking the image or text modality independently produces inferior performance compared with the collaborative multimodal adversarial attack (Co-Attack). Independent attacks on images or text potentially introduce conflicting perturbation directions between modalities, thereby degrading attack effectiveness. In contrast, under the joint perturbation strategy, textual adversarial examples are generated on the basis of image adversarial examples. In the context of VLP adversarial perturbations, this joint strategy ensures that image and text perturbations follow a consistent optimization direction while simultaneously enlarging the representational distance between image–text pairs. Although Co-Attack focuses on white-box attacks, it achieves a relatively low success rate in transfer-based scenarios.

To address this limitation, Lu et al. [13] propose the Set-level Guidance Attack (SGA), which enhances transferability by expanding the sample dataset. However, during the iterative optimization process, SGA relies excessively on perturbed samples along the adversarial trajectory, resulting in a relatively low attack success rate when transferring to other VLP models. To mitigate this over-reliance on the adversarial trajectory, Gao et al. [14] introduce the Semantic-aligned Adversarial Evolution Triangle (SA-AET). The iterative strategies of SGA and SA-AET are illustrated in Fig. 1. In SGA, at each iteration, perturbations are generated by sampling from the previous adversarial point v_i, selecting a perturbation direction, and subsequently producing the next adversarial point v_{i+1} (see Fig. 1a). In contrast, SA-AET samples not only from the previous adversarial point v_i but also from the original sample v during each iteration, thereby obtaining a broader range of perturbation directions and alleviating overfitting of perturbations to a specific model (see Fig. 1b).

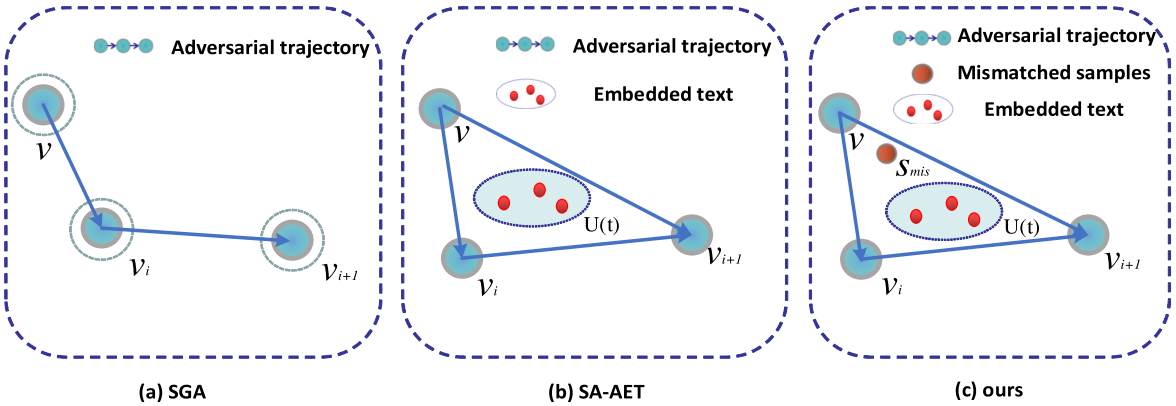


Figure 1: Comparison of different perturbation methods for vision–language pre-training (VLP) models. (**a**) Shows that the SGA method performs data augmentation primarily centered on the textual dataset. (**b**) Demonstrates that the SA-AET method constructs an adversarial triangle, in which the red data points represent the augmented textual dataset within the sub-triangle used to guide adversarial perturbations. (**c**) Presents the core idea of the proposed approach, whereby mismatched sample pairs are incorporated into the adversarial triangle to guide the generation of adversarially perturbed samples.

Overall, adversarial example generation methods based on modality joint attacks focus on aligning modal features within the semantic space and demonstrate improved transfer performance across different models. Nevertheless, existing approaches generate adversarial examples that depend heavily on the surrogate model and the originally matched modality, leading to insufficient perturbation diversity and consequently constraining both attack effectiveness and transferability.

To address the aforementioned limitations, this study proposes a transfer-based adversarial example generation method for Vision–Language Pre-training (VLP) models based on negative sample feature perturbation. The proposed approach incorporates the principle of contrastive learning and introduces a Semantic-Aligned Misguided Adversarial Evolution Triangle (SAM-AET) module. Building upon the SA-AET framework, SAM-AET incorporates negative samples [15], defined as mismatched image–text pairs that interfere with the model's semantic judgement.

During the generation of each adversarial point v_{i+1}, the proposed method samples not only from the original sample v and the previous adversarial point v_i, but also introduces negative samples that are mismatched with the current modality to guide the perturbation direction. This strategy eliminates reliance on the originally matched modality while simultaneously providing a broader range of perturbation directions for adversarial example generation (see Fig. 1c).

Subsequently, a dynamic threshold attack strategy is designed to evaluate similarity scores. Finally, a multimodal fusion encoder is employed to extract fused representations of image–text pairs, thereby guiding the disruption of fine-grained semantic alignment between images and text. Through this mechanism, the transferability of the generated adversarial examples is further enhanced.

The major contributions of this paper are as follows:

(1) A Semantic-Aligned Misguided Adversarial Evolution Triangle (SAM-AET) module is constructed. By incorporating a contrastive learning strategy, mismatched negative samples are introduced to misguide the perturbation direction of the current modality, thereby ensuring effective exploration of adversarial regions in the target model.

(2) A Cross-modal Fusion Semantic Attack (CFSA) module is designed. A dynamic threshold attack strategy is established, whereby the CFSA module is activated according to similarity scores. By integrating

intermediate-layer features of image–text pairs within a multimodal fusion encoder, fine-grained similarity scores for each modality are obtained. This mechanism optimizes perturbations between the image and text modalities, further enhances the adaptability of adversarial examples, and ensures improved transferability across different models.

(3) The proposed adversarial attack method is validated on two public datasets, Flickr30K and MSCOCO, in order to evaluate the influence of different datasets on the transferability of generated adversarial examples. Furthermore, the generalization capability of perturbations generated on generic datasets against black-box models is systematically investigated.

2 Related Works

2.1 Single-Modality Adversarial Attack Methods

The earliest adversarial attack methods originate in the field of image recognition, particularly in image classification tasks. In 2014, Szegedy et al. [16] demonstrate that adding imperceptibly small perturbations to input samples causes a model to produce incorrect outputs with high confidence, thereby introducing the concept of adversarial attacks and proposing the L-BFGS method. This discovery draws widespread attention to the vulnerability of deep learning models to subtle perturbations. Subsequently, Goodfellow et al. [17], motivated by the linearity hypothesis of deep neural networks, propose the Fast Gradient Sign Method (FGSM) to overcome the slow attack speed of L-BFGS. FGSM requires only the computation of a gradient vector, thereby reducing the time required to generate adversarial examples, although at the expense of some attack precision.

Since attackers do not always have access to the original input samples during the testing phase, adversarial attacks based on generative models are further proposed. Generative models learn from training data, estimate the overall probability distribution of samples, and generate new samples that are similar to and follow the same distribution as the training data. For instance, Baluja and Fischer [18] train a deep neural network termed the Adversarial Transformation Network (ATN), which transforms original samples into adversarial examples. Once trained, the ATN generates adversarial examples through a single forward propagation. In addition to adding noise or employing generative networks, image-based adversarial attacks also include semantic-level manipulations [19], such as rotation, occlusion, contrast adjustment, color modification, and semantic content transformation.

In the visual domain, adversarial attacks on images receive extensive investigation. Similar attack strategies are gradually extended to Natural Language Processing (NLP) [20], particularly in text classification tasks. In text classification adversarial attacks, the objective is to maximize misclassification while minimizing perceptible alterations to human readers. Analogous to modifying a small number of pixels in image-based attacks, textual adversarial attacks alter a limited number of words or characters in order to deceive the model into producing incorrect classifications with high confidence. For example, Jia and Liang [21] conduct adversarial attacks on sentiment analysis and fake news detection tasks. Chen et al. [22] propose a Fast Adversarial Watermark Attack method, which cleverly disguises perturbations within a watermark, thereby maintaining attack success rates while ensuring that adversarial images appear natural to human observers. Owing to the inaccessibility of gradient information in certain scenarios, Xu et al. [23] propose a black-box attack based on the Differential Evolution Algorithm (DEA), which attacks scene text recognition models by minimizing pixel perturbations. As a stochastic parallel direct search algorithm, DEA requires minimal assumptions regarding the optimization objective, does not depend on gradient information, and performs direct searches in high-dimensional spaces to iteratively obtain optimal solutions.

Overall, although single-modality adversarial attack methods achieve significant progress in tasks such as image classification and natural language processing, they struggle to address the challenges posed by complex real-world applications, particularly when confronting multimodal models. In such cases, the effectiveness of single-modality attacks is difficult to guarantee. Consequently, research into multimodal adversarial attacks becomes a prominent focus. By jointly considering the relationships between multiple modalities, such as images and text, multimodal attacks enhance generality and adaptability, thereby overcoming the inherent limitations of single-modality approaches.

2.2 Multimodal Adversarial Attack Methods

Since Zhang et al. [12] conduct a systematic investigation of adversarial example generation methods for Vision–Language Pre-training (VLP) models, subsequent studies [11] propose a series of adversarial generation approaches tailored to VLP architectures. These methods are broadly categorized into universal adversarial patch-based approaches and modality joint attack-based approaches.

Among universal adversarial patch-based methods, Zhou et al. [24] propose the AdvCLIP framework, which generates universal perturbation patches targeting the CLIP encoder. This framework does not rely on downstream task-specific details and effectively disrupts all downstream tasks built upon CLIP. Huang et al. [25] design a Targeted Universal Adversarial Perturbation (TUAP) method, which employs a surrogate model to obtain target gradients and generate universal adversarial patches. As a result, any image embedded with the patch is misled into aligning with a pre-defined textual representation in the CLIP embedding space. The attack also remains effective when CLIP image encoders are integrated into large-scale vision–language models. Fang et al. [15] propose a Contrastive Perturbation Generator with Cross-modal Conditions (C-PGC) based on generative adversarial networks. During training, C-PGC adopts a contrastive learning strategy by constructing mismatched image–text pairs and incorporating them into the generator training process. The generated perturbations enlarge the representational distance between originally matched image–text pairs while reducing the distance between mismatched pairs, thereby disrupting the original cross-modal alignment and producing universal perturbations for arbitrary inputs.

Overall, universal adversarial patch-based methods interfere with a wide range of downstream tasks or models without relying on specific task or architectural details. However, some approaches do not explicitly optimize perturbations according to the vulnerabilities of particular models, which leads to relatively low attack success rates.

In modality joint attack-based methods, Sep-Attack employs Projected Gradient Descent (PGD) [26] and BERT-Attack [27] to attack the image and text modalities separately when targeting VLP models, yet it does not account for interactions between the two modalities. Zhang et al. [28] propose the Co-Attack method, which enhances the robustness evaluation of VLP models under multimodal adversarial attacks. By simultaneously perturbing both image and text modalities and explicitly considering cross-modal interactions, Co-Attack disrupts shared information across modalities and increases model susceptibility to attack. Lu et al. [13] observe that Co-Attack generates adversarial examples using only a single image–text pair, thereby limiting attack diversity and effectiveness. To address this issue, they propose the Set-level Guidance Attack (SGA), which introduces set-level alignment during adversarial example generation. By expanding the corresponding modality sample set and exploiting inter-sample relationships, SGA enhances perturbation diversity. Subsequently, Gao et al. [14] identify that existing approaches rely excessively on the surrogate model's training outcomes, which easily results in overfitting to specific models. They introduce the concept of adversarial trajectory intersection regions, leveraging the intersection of adversarial trajectories to increase perturbation diversity. Wang et al. [29] further observe that internal cross-modal attention mechanisms in VLP models simultaneously influence shared modal features. Accordingly, they

propose the TMM attack framework, which utilizes cross-modal attention weights within VLP models to design guided perturbation strategies, thereby improving the transferability of adversarial examples across VLP architectures.

Overall, although these approaches evolve from single-modality attacks towards cross-modal perturbation strategies, the generation of perturbations still depends on a limited sample space. This constraint results in insufficient perturbation diversity and a tendency towards overfitting, leaving substantial room for improvement in cross-model and cross-task transferability.

3 Methodology

3.1 Overall Framework Architecture

The overall architecture of the proposed method is illustrated in Fig. 2. It consists of a dual-encoder, a Semantic-Aligned Misguided Adversarial Evolution Triangle (SAM-AET) module, and a Cross-modal Fusion Semantic Attack (CFSA) module. The primary objective is to generate, for a given original image V and its corresponding matched text T, a subtle perturbation δ that satisfies the l_∞ norm constraint $\|\delta\| \leq \varepsilon$. The perturbation direction is guided by the semantic relationships between modalities within the Vision–Language Pre-training (VLP) models. The final adversarial example is defined as $x_{adv} = x + \delta$, where x denotes an input image–text pair sampled from the training set. The generated adversarial example is intended to cause the VLP model to misjudge the semantic relationship between the image and text.

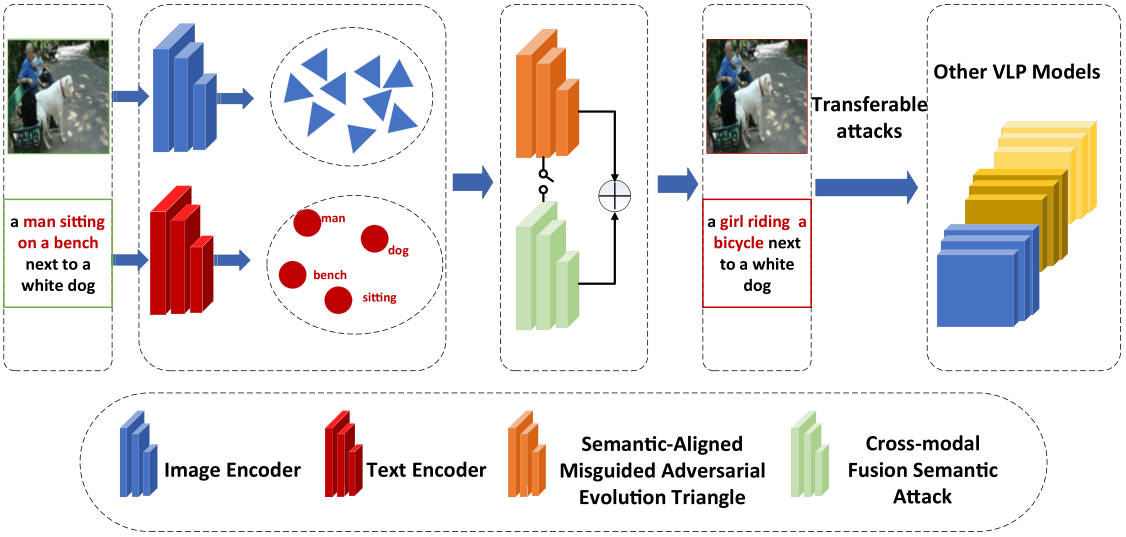


Figure 2: Proposed method framework.

The dual-encoder facilitates effective learning of modality-specific feature distributions within the VLP model. It independently encodes the image and text, and computes their similarity via the dot product between the corresponding feature vectors. Building upon the dual-encoder structure and the SA-AET method, mismatched negative modality samples are introduced to construct the SAM-AET module, which generates an initial adversarial perturbation. The generated adversarial example is then evaluated by computing the similarity score S between x_{adv} and its corresponding modality. If S falls below a predefined threshold τ, the attack is considered successful and the process terminates. Otherwise, the adversarial example proceeds to the next stage, where it is further refined using the CFSA module.

The CFSA module adopts a multimodal fusion encoder. Unlike the dual-encoder, the multimodal fusion encoder incorporates additional Transformer layers to extract joint image and text features. Through a cross-attention mechanism, it captures deep semantic associations between image regions and textual tokens. These cross-modal interactions guide the generation of refined perturbations in both image and text modalities, thereby disrupting fine-grained semantic alignment and producing the final adversarial example.

To implement this framework, ALBEF [30], TCL [31], CLIP_{VIT} [32], and CLIP_{CNN} [32] are employed as white-box surrogate models. The encoder–decoder structure is trained to generate adversarial examples of size 384 × 384, and the transferability of the generated adversarial examples is evaluated across different black-box models.

3.2 Semantic-Aligned Misguided Adversarial Evolution Triangle

In current research on adversarial attacks against Vision–Language Pre-training (VLP) models, mainstream approaches generate adversarial examples based on the similarity between image and text features. However, such methods often overlook the issue of semantic alignment between modalities, which leads to adversarial examples that fail to produce effective interference in certain scenarios and consequently exhibit limited attack strength. To enhance both the effectiveness and cross-model transferability of adversarial examples, this study designs a Semantic-Aligned Misguided Adversarial Evolution Triangle (SAM-AET) module, which introduces mismatched negative modality samples to construct a more deceptive adversarial generation mechanism.

From the perspective of VLP model characteristics, the original cross-modal alignment relationship exhibits a degree of inherent recovery capability. If only slight perturbations are applied to originally aligned samples—for example, causing the feature representation of a cat image to deviate marginally from that of the corresponding cat text—the model is capable of re-establishing cross-modal alignment by adjusting feature weights, thereby diminishing the adversarial effect. In contrast, the SAM-AET module disrupts this dependency by randomly shuffling or replacing one modality within image–text pairs. For instance, a cat image may be paired with a dog-related text description, or the cat image may be replaced with a car image to form mismatched negative samples. Under such conditions, the objective of adversarial perturbation shifts towards further reducing the semantic distance between mismatched modality pairs. This deeper level of interference not only impairs the alignment recovery capability of VLP models but also ensures that the perturbation mechanism no longer depends on model-specific alignment biases. Consequently, the generated adversarial examples remain effective across VLP models with different architectures.

Examples of adversarial samples generated by the proposed method are illustrated in Fig. 3. To effectively exploit the similarity between image and text features, the method adopts a mutual guidance strategy between modalities. Specifically, in order to ensure that the image modality moves sufficiently away from the corresponding text modality during perturbation generation, the original text modality t is expanded into a semantically similar text set U(t). This expansion guarantees that the perturbed image representation diverges from all semantically related textual descriptions, thereby strengthening the adversarial objective.

Subsequently, the image modality is employed to guide the generation of adversarial text. For each textual instance in the expanded semantic set U(t), a corresponding adversarial text set $U_{(t')} = \{t'_1, t'_2, \ldots, t'_n\}$ is generated. An image-based mismatched negative sample pair, denoted as v_{mis}, is then constructed to establish an augmented visual guidance environment for the generation of each adversarial text instance t'_i. This environment comprises both matched image–text pairs and mismatched image pairs.

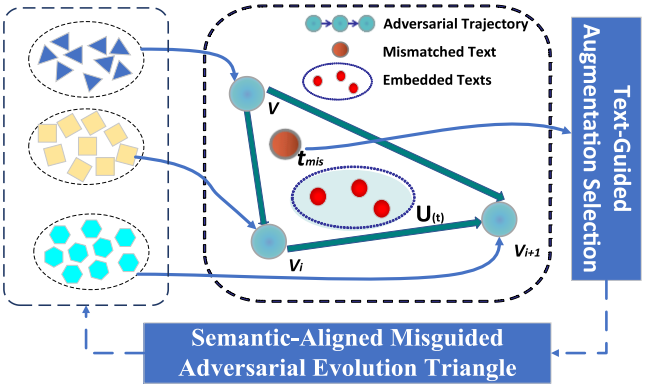


Figure 3: Semantic-aligned misguided adversarial evolution triangle (SAM-AET).

The optimisation objective is formulated to maximise the semantic distance between the adversarial text and its originally matched image, while simultaneously minimising the semantic matching loss with respect to the mismatched image v_{mis}. By solving this dual-objective optimisation problem, the adversarial text t'_i that satisfies both constraints is obtained. The generation process is formally expressed in Eq. (1).

$$t'_i = \max_{t' \in B_{[t,\varepsilon_t]}} \left(-\frac{f_T(t'_i) \cdot f_V(v)}{\|f_T(t'_i)\| \|f_V(v)\|} + \frac{f_T(t'_i) \cdot f_V(v_{mis})}{\|f_T(t'_i)\| \|f_V(v_{mis})\|} \right) \tag{1}$$

where the text encoder f_T and the image encoder f_V constitute the dual-encoder architecture of the VLP model.

Subsequently, the adversarial text t' is employed to guide the generation of the adversarial image v'. This strategy operates within a region defined by three adversarial points. Specifically, during each iterative perturbation step, the method samples not only from the previous adversarial point v_i to determine the perturbation direction, but also from the original sample v, thereby generating the next adversarial point v_{i+1}. This triangular evolution mechanism mitigates overfitting of the perturbation to a single model.

Within this triangular region, perturbation sampling is performed over the feature spaces of both image and text modalities in order to explore a broader adversarial space. For each sampled point e_k, a weighted formulation, as expressed in Eq. (2), is adopted to regulate the perturbation contribution of each adversarial point:

$$e_k = \beta \cdot v + \gamma \cdot v_i + \eta \cdot v_{i+1} \tag{2}$$

where β, γ, η denote hyperparameters controlling the perturbation weights and satisfy the constraint $\beta + \gamma + \eta = 1$.

During the generation of adversarial points, each image v_i is rescaled to multiple proportions to introduce variations in scale and viewpoint, thereby simulating the diversity encountered in real-world scenarios. Drawing upon the contrastive learning strategy adopted in VLP model training, a deliberately adversarial contrastive mechanism is incorporated. Specifically, mismatched textual negative samples t_{mis} are introduced within the SAM-AET module to guide adversarial example generation.

Ultimately, the generated adversarial text set $U_{(t')}$, together with the negative modality samples t_{mis}, jointly guides the generation of the adversarial image v', as formulated in Eq. (3):

$$v' = argmax_{v' \in B[v,\epsilon]} \left(\sum_{k=1}^{M} \frac{F_T(t_{\text{mis}})}{\| F_T(t_{\text{mis}}) \|} - \sum_{m=1}^{M} \frac{F_T(t'_m)}{\| F_T(t'_m) \|} \right) \cdot \sum_{s_i \in S} \frac{F_V(g(v_i, s_i))}{\| F_V(g(v_i, s_i)) \|} \tag{3}$$

where $g(v_i, s_i)$ denotes an anti-aliasing scaling function, which takes the image v_i and the scaling factor s_i as inputs and outputs the corresponding set of scaled images, thereby guiding the generation of the adversarial image v'.

To improve the clarity and reproducibility of the proposed method, we present the pseudocode of the SAM-AET module in Algorithm 1.

Algorithm 1: Semantic-aligned misguided adversarial evolution triangle (SAM-AET)

Input: clean image v, matched text t, perturbation bound ε, iteration number K, step size α
Output: adversarial image v'

Initialize the adversarial sample:
$v_0' \leftarrow \text{clip}(v + \text{Uniform}(-\varepsilon, \varepsilon), 0, 1)$, $v_1' \leftarrow v_0'$
Construct the semantically similar text set $U_{(t)}$ from the matched text t
Generate the adversarial text set $U_{(t')}$ according to Eq. (1)
Randomly sample mismatched text instances from the dataset to construct the negative modality t_{mis}
for iter = 1 to K do
    Construct the sampled point in the adversarial triangle according to Eq. (2):
    Compute the perturbation direction using the adversarial text set $U_{(t')}$ and the negative modality t_{mis} according to Eq. (3)
    Update the adversarial sample v'_{i+1} according to Eq. (3)
end for
return v'_K

3.3 Cross-Modal Fusion Semantic Attack

Although the perturbations generated by the SAM-AET module exhibit enhanced diversity, they are not explicitly optimized for the feature requirements of different downstream tasks, such as image–text retrieval and visual question answering. In different tasks, VLP models assign varying levels of importance to image and text modalities. For instance, retrieval tasks primarily emphasize global semantic alignment, whereas question answering tasks rely more heavily on fine-grained local feature associations. Consequently, perturbations generated solely through the dual-encoder within the SAM-AET module are insufficient to accommodate all application scenarios.

To further enhance the transferability of adversarial perturbations, this study designs a Cross-modal Fusion Semantic Attack (CFSA) module. First, in order to balance computational cost and perturbation efficiency, a dynamic threshold attack strategy is introduced. This strategy evaluates the similarity score S between the currently generated adversarial sample x_{adv} and its corresponding original matched sample y. If S exceeds a predefined threshold τ, the sample is regarded as a "hard" instance, indicating that the preliminary adversarial perturbation generated by the SAM-AET module is insufficient. In such cases, the CFSA module shifts the attack objective from the dual-encoder to the deeper multimodal fusion encoder within the VLP model.

As illustrated in Fig. 4, the CFSA module aggregates intermediate-layer features from the fusion encoder and maximizes the semantic discrepancy between perturbed and original features. This strategy prevents effective cross-modal alignment across different VLP architectures, thereby addressing the limitation of SAM-AET perturbations in adapting to heterogeneous models and further improving adversarial transferability.

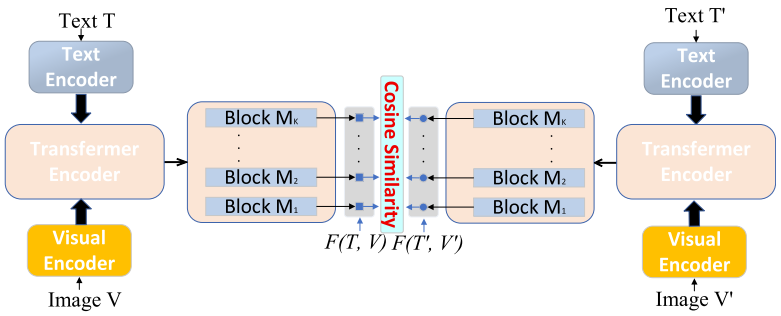


Figure 4: Cross-modal fusion semantic attack (CFSA).

The proposed method aims to impose adversarial perturbations simultaneously on both image and text modality features. Unlike conventional single-modality attacks, CFSA performs joint optimization over image feature perturbations V' and text feature perturbations T'.

The core principle of the CFSA module is to generate adversarial examples by minimizing the discrepancy between the intermediate-layer features of the two modalities within the multimodal fusion encoder. The loss function consists of two components: an image loss and a text loss, which respectively measure the deviation between the perturbed image and text features and their corresponding target features.

First, visual and textual features are extracted via the visual encoder f_V and the text encoder f_T, respectively. The input image V is transformed into an image feature representation F_V, as expressed in Eq. (4):

$$F_V = f_V(V) \tag{4}$$

Similarly, the input text T is transformed into a text feature representation F_T, as expressed in Eq. (5):

$$F_T = f_T(T) \tag{5}$$

The loss function of this module integrates both image and text losses in order to balance the influence of the two modalities during perturbation generation. To maximise adversarial effectiveness, the final optimisation objective of CFSA is to minimise the overall loss L, thereby jointly optimising perturbations applied to both image and text modalities. By minimising this loss, the generated adversarial examples introduce sufficient interference in both modalities, successfully misleading the vision–language model, as formulated in Eq. (6):

$$L = -\sum_{i=1}^{K} \left[\cos\left(F_{\mathrm{i}}\left(F_T, F_V\right), F_i\left(F_{V'}, F_{T'}\right)\right)\right] \tag{6}$$

where $F_i(F_T, F_V)$ and $F_i(F_{V'}, F_{T'})$ denote the intermediate feature representations extracted at the i layer of the multimodal fusion encoder for the original sample and the perturbed sample, respectively, and K represents the total number of layers in the cross-modal encoder.

Ultimately, by maximising the feature discrepancy between perturbed and original samples, the VLP model is rendered less capable of aligning the semantic correspondence between the perturbed image and text, thereby producing erroneous predictions.

4 Experimental Results and Analysis

4.1 Datasets and Evaluation Metrics

4.1.1 Datasets

This study employs the widely used Flickr30K [33] and MSCOCO [34] image–text datasets to evaluate model performance.

The Flickr30K dataset comprises 31,783 images, each accompanied by five textual descriptions. It is extensively utilised in tasks involving joint image–text processing, particularly image caption generation, image–text retrieval, and visual question answering.

The MSCOCO (Microsoft Common Objects in Context) dataset contains 123,287 images, with each image associated with approximately five textual descriptions. It is a widely adopted benchmark in computer vision, incorporating large-scale image and textual annotations. MSCOCO is broadly applied to vision–language tasks, object detection, image segmentation, and image caption generation, among others.

These two datasets provide rich and diverse training samples, enabling models to learn more generalisable and accurate feature representations. They also facilitate improved adaptability to data distributions across different domains and scenarios, thereby enhancing the generalisation capability of adversarial perturbations. To achieve a comprehensive evaluation of model performance, this study utilises both Flickr30K and MSCOCO to assess the transferability of the generated adversarial examples.

4.1.2 Evaluation Metrics

To evaluate the effectiveness of adversarial attacks, this study adopts the Attack Success Rate (ASR) at Rank-(K) as the primary evaluation metric. The attack success rate refers to the proportion of incorrect results among the top-(K) retrieved matches. In the image-to-text retrieval task, TR R (Text Retrieval Rate)@(K) is employed to denote the percentage of incorrect texts among the top-(K) retrieved results for all input perturbed images. Similarly, in the text-to-image retrieval task, IR R (Image Retrieval Rate)@(K) represents the percentage of incorrect images among the top-(K) retrieved results for all input perturbed texts. All adversarial perturbations are generated under a fixed setting of $\varepsilon = 2/255$, steps = 10, and step size = 0.5/255 to ensure a fair and controlled comparison.

4.2 Comparative Experiments on Transfer Success Rate

This study employs four widely adopted VLP models—ALBEF [30], TCL [31], CLIP_{CNN} [32], and CLIP_{ViT} [32]—as surrogate (white-box) and transfer (black-box) models. ALBEF is designed for large-scale vision–language representation learning and supports a variety of downstream vision–language tasks. However, its alignment mechanism focuses primarily on global image–text alignment and does not explicitly consider intra-modal alignment or fine-grained global–local alignment. TCL extends ALBEF by expanding its single alignment strategy into three distinct alignment mechanisms. CLIP is trained on large-scale contrastive image–text pairs and demonstrates strong capabilities in joint visual–textual understanding. In this study, when CLIP adopts ViT-B/16 as its image encoder backbone, it is denoted as CLIP_{ViT}; when it employs ResNet-101 as its image encoder backbone, it is denoted as CLIP_{CNN}.

Table 1 presents comparative results on the Flickr30K dataset for both text retrieval (TR) and image retrieval (IR) sub-tasks under the highest matching score search strategy across different VLP models. The experimental results demonstrate that the proposed method consistently outperforms existing multimodal attack methods in both white-box and black-box settings. Specifically, in the white-box setting, the proposed method achieves an attack success rate exceeding 98% for TR R@1 on ALBEF, and it surpasses baseline performance on IR R@1 across various VLP models. As shown in Table 1, the proposed method achieves higher attack success rates than competing approaches, with an average improvement of 1.03% over the strongest baseline, SA-AET, thereby further validating its effectiveness.

Table 1: Comparison between the proposed method and baseline methods on the Flickr30K dataset.

Flickr30K									
Target Model (→)		**ALBEF**		**TCL**		**$CLIP_{ViT}$**		**$CLIP_{CNN}$**	
Surrogate Model	**Attack Method**	**TR R@1**	**IR R@1**	**TR R@1**	**IR R@1**	**TR R@1**	**IR R@1**	**TR R@1**	**IR R@1**
ALBEF	PGD	52.45	58.65	3.06	6.79	8.96	13.21	10.34	14.65
	BERT-Attack	11.57	27.46	12.64	28.07	29.33	43.17	32.69	46.11
	Sep-Attack	65.69	73.95	17.6	32.95	31.17	45.23	32.82	45.49
	Co-Attack	77.16	83.86	15.21	29.49	23.6	36.48	25.12	38.89
	SGA	97.5	97.17	46.36	55.33	32.64	44.52	32.64	44.52
	SA-AET	96.39	96.56	49	57.71	39.23	48.46	41.38	51.66
	Ours	**98.3**	**98.37**	**54.58**	**59.61**	**40.23**	**49.39**	41.35	**54.37**
TCL	PGD	6.15	10.78	77.87	79.48	7.48	13.72	10.34	15.33
	BERT-Attack	11.89	26.82	14.54	29.17	29.69	44.49	33.46	46.07
	Sep-Attack	20.13	36.48	84.72	86.07	31.29	44.65	33.33	45.80
	Co-Attack	48.91	60.34	98.37	98.81	33.87	44.88	37.74	48.30
	SGA	48.91	48.91	48.91	48.91	48.91	48.91	48.91	48.91
	SA-AET	95.2	95.58	100	99.98	47.24	57.28	52.23	62.23
	Ours	**95.3**	95.73	99.24	99.95	**48.21**	**58.31**	52.21	**64.35**
$CLIP_{ViT}$	PGD	2.50	4.93	4.85	8.17	70.92	13.21	78.61	8.44
	BERT-Attack	9.59	27.46	11.80	28.07	28.34	43.17	30.40	46.11
	Sep-Attack	9.59	22.64	11.38	25.07	79.75	39.08	30.78	37.43
	Co-Attack	10.57	24.33	11.94	26.69	93.25	95.86	32.52	41.82
	SGA	13.40	27.22	16.23	30.76	99.08	98.94	38.76	47.79
	SA-AET	12.51	30.0	14.65	30.62	98.77	99.0	45.47	50.74
	Ours	**13.43**	**30.11**	**15.32**	**31.21**	**98.78**	**99.12**	**47.32**	**51.16**
$CLIP_{CNN}$	PGD	2.09	4.82	4.00	7.81	1.10	6.60	86.46	92.25
	BERT-Attack	8.86	23.27	12.33	25.48	27.12	37.44	30.40	40.10
	Sep-Attack	8.55	23.41	12.64	26.12	28.34	39.43	91.44	95.44
	Co-Attack	8.79	23.74	13.10	26.07	28.79	40.03	94.76	96.89
	SGA	11.42	24.80	14.91	28.82	31.24	42.12	99.24	99.49
	SA-AET	12.2	26.59	14.33	29.29	35.21	45.94	99.11	99.49
	Ours	**12.56**	**27.16**	**15.69**	**30.56**	**38.21**	**48.94**	**99.12**	99.13

Note: The bolded parts represent the best data for this indicator.

The experimental results on the MSCOCO dataset are presented in Table 2. Table 2 compares performance on the TR and IR sub-tasks under the same highest matching score search strategy across different VLP models. The results indicate that, in both white-box and black-box scenarios, the proposed method outperforms existing VLP adversarial attack methods. In particular, the transfer attack success rate improves by an average of 0.68% compared with the best-performing baseline, SA-AET.

Table 2: Comparison between the proposed method and baseline methods on the MSCOCO dataset.

MSCOCO									
Target Model (→)		**ALBEF**		**TCL**		**$CLIP_{ViT}$**		**$CLIP_{CNN}$**	
Surrogate Model	**Attack Method**	**TR R@1**	**IR R@1**	**TR R@1**	**IR R@1**	**TR R@1**	**IR R@1**	**TR R@1**	**IR R@1**
ALBEF	PGD	67.54	73.72	11.08	13.66	14.5	22.86	17.12	23.75
	BERT-Attack	35.91	47.33	35.93	35.93	53.22	63.41	56.11	65.07
	Sep-Attack	83.89	87.5	42.12	51.29	54.25	63.88	57.13	65.07
	Co-Attack	79.87	87.83	32.62	43.09	43.09	54.75	47.3	55.64
	SGA	96.75	96.95	58.56	65.38	57.06	65.25	58.95	66.52
	SA-AET	96.57	96.47	60.69	67.46	61.69	67.43	62.32	69.22
	Ours	**97.32**	**98.67**	**61.35**	**68.33**	**62.91**	**69.32**	**63.12**	**71.15**
TCL	PGD	14.2	18.41	83.33	85.61	14.46	23.05	85.61	23.51
	BERT-Attack	36.04	47.36	37.67	48	55.28	63.76	57.13	65.92
	Sep-Attack	44.83	56.18	91.85	92.75	51.62	64.72	58.11	65.93
	Co-Attack	46.08	57.09	85.38	91.39	51.62	60.46	59.92	62.49
	SGA	65.93	73.30	98.97	99.15	56.34	63.99	59.44	65.70
	SA-AET	68.06	75.86	98.99	99.02	63.3	68.59	64.24	70.58
	Ours	**70.17**	74.01	**99.37**	**99.5**	**64.79**	**69.81**	**65.68**	**71.66**
$CLIP_{ViT}$	PGD	7.12	10.48	8.2	11.78	11.78	11.42	7.85	11.25
	BERT-Attack	23.38	34.39	24.66	34.21	34.21	42.88	42.38	48.49
	Sep-Attack	23	34.3	25.05	34.8	34.8	34.6	42.09	50.04
	Co-Attack	30.28	42.67	34.41	44.69	44.69	44.69	55.05	62.51
	SGA	33.41	44.64	37.54	47.76	99.79	99.79	58.93	47.79
	SA-AET	35.96	48.0	36.32	48.56	99.66	99.7	64.41	65.83
	Ours	35.43	47.11	**37.32**	48.21	98.78	99.12	**65.32**	**66.99**
$CLIP_{CNN}$	PGD	6.65	10.4	11.4	13.65	6.11	11.4	11.25	11.25
	BERT-Attack	26.27	38.86	28.54	29.61	50.6	56.52	50.48	57.71
	Sep-Attack	26.35	38.86	25.05	39.64	51.8	53.89	50.1	56.53
	Co-Attack	29.83	42.67	32.77	44.69	53.51	60.35	55.43	56.72
	SGA	31.61	43.00	34.81	45.95	56.62	60.77	99.61	99.80
	SA-AET	33.26	45.15	33.89	46.49	59.6	64.87	99.51	99.7
	Ours	32.56	**44.16**	32.69	**47.56**	**60.21**	**65.94**	99.12	99.13

Note: The bolded parts represent the best data for this indicator.

As shown in Tables 1 and 2, although the proposed method achieves strong attack performance on both datasets, its ASR on MSCOCO is consistently lower than that on Flickr30K. This gap can be explained not only by the higher scene complexity of MSCOCO, but also by the differences in dataset scale and semantic diversity. Compared with Flickr30K, MSCOCO contains a larger number of images and covers a broader range of object categories, scene compositions, and contextual relationships. As a result, the cross-modal alignment patterns learned from MSCOCO are more dispersed, making it harder for the generated perturbations to capture a stable and transferable adversarial direction across different models. In addition, the higher semantic diversity of MSCOCO weakens the consistency of negative-sample guidance. Since mismatched image-text pairs may differ from the original sample in many possible semantic aspects, the perturbation optimization is more likely to be distributed over multiple directions, which reduces the concentration and transferability of the generated perturbations. Furthermore, MSCOCO often contains more fine-grained correspondences among multiple objects, attributes, and contextual cues, so disrupting the image-text alignment under a limited perturbation budget becomes more difficult than in Flickr30K. Nevertheless, the proposed method still maintains strong attack performance on MSCOCO, indicating that the introduced negative-sample perturbation and cross-modal fusion strategy remain effective even on a larger-scale and more diverse dataset.

4.3 Comparison of Visualisation Results

As illustrated in Figs. 5 and 6, the effectiveness of the proposed method is comprehensively validated across different scenarios.

Fig. 5 presents a visualization of adversarial examples generated for the image captioning task, demonstrating that the semantic alignment between images and captions is successfully disrupted after applying the proposed method. For instance, the original caption “a group of people standing on the lawn in front of a building” is perturbed into “a group of people standing on the lawn in front of a someone”, clearly indicating semantic distortion. This example provides intuitive evidence of the adversarial sample generation capability of the proposed approach.

Figure 5: Visualization of clean samples and perturbed samples.

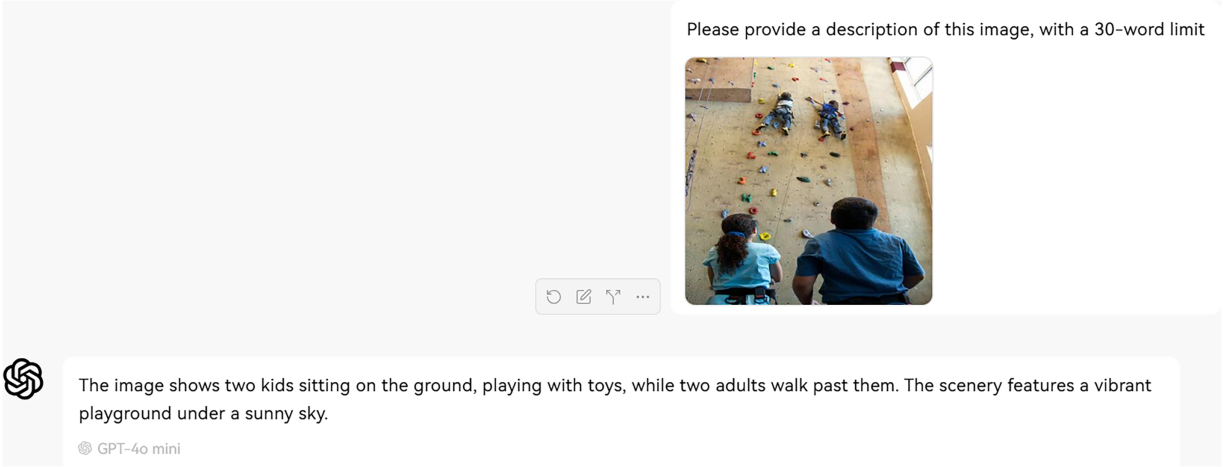


Figure 6: Attack performance on GPT-4o mini.

Fig. 6 further illustrates the adversarial effect on a large-scale model. Specifically, the perturbation samples generated on the Flickr30K dataset are fed into GPT-4o mini, together with the prompt: "Please provide a description of this image, with a 30-word limit."

As shown in Fig. 6, the description produced by the large model is inconsistent with the actual image content.

As shown in Fig. 6, the description produced by the large model is inconsistent with the actual image content. This result demonstrates that the proposed method successfully deceives mainstream large-scale vision–language models, further validating its transferability and practical attack effectiveness in digital environments. It should be noted, however, that the above results are still obtained in a digital evaluation setting. Extending the proposed perturbations to physical-world multimodal attacks remains challenging. On the one hand, physically realized perturbations may be degraded by the printing and imaging process, as well as by variations in viewpoint, distance, scale, and illumination, which can reduce the stability of the optimized pixel-level perturbations. On the other hand, unlike conventional single-modality physical attacks, multimodal attacks must not only affect the visual representation of the image, but also continuously interfere with the semantic alignment between image and text under different prompts, captions, or downstream tasks. In addition, improving physical robustness often requires stronger or more localized perturbation patterns, which may conflict with the imperceptibility constraint considered in this study. Therefore, the current experiments mainly validate the effectiveness and transferability of the proposed method in the digital domain, while its extension to physically realizable multimodal attacks will be further investigated in future work.

4.4 Ablation Study

To further investigate the key factors influencing the proposed method, ablation studies are conducted on the image–text retrieval task. In particular, the performance of the SAM-AET module is evaluated against comparative methods on the Flickr30K dataset, in order to assess the effectiveness of introducing mismatched negative modality samples for attacking VLP models.

Specifically, ALBEF is employed as the source (white-box) model, while CLIP-CNN serves as the target model on Flickr30K. The evaluation focuses on metrics such as TR and IR. The detailed experimental results are presented in Fig. 7. The proposed method achieves higher transfer attack success rates on both

text retrieval and image retrieval tasks compared with the SA-AET method, indicating that SAM-AET outperforms SA-AET.

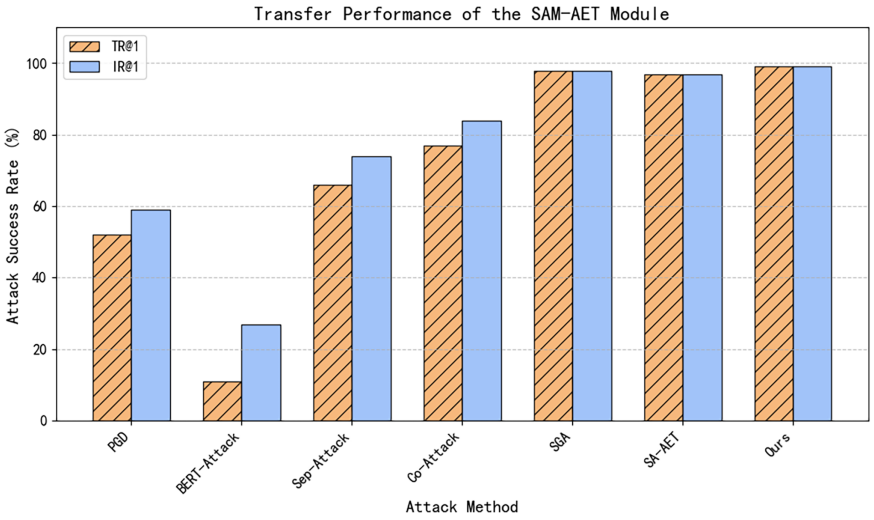


Figure 7: Transfer performance of the semantic-aligned misguided adversarial evolution triangle (SAM-AET).

These findings demonstrate that fully exploiting cross-modal interactions and introducing mismatched negative modality samples to disrupt modal correspondence significantly enhances the transferability of adversarial attacks against VLP models.

To evaluate the contribution of each module to the overall performance, we conduct ablation experiments on the transfer task from ALBEF to TCL, as shown in Table 3. In addition to R@1, we further report R@5 to provide a more comprehensive evaluation of attack effectiveness.

Table 3: Module ablation for attack success rate.

Transferability	Module Setting		Attack Success Rate			
	SAM-AET	CFSA	TR R@1	IR R@1	TR R@5	IR R@5
ALBEF to TCL	×	×	49.0	57.71	26.52	39.48
	√	×	51.23	59.33	27.32	41.23
	×	√	50.36	59.45	26.89	41.46
	√	√	54.58	59.61	28.31	42.13

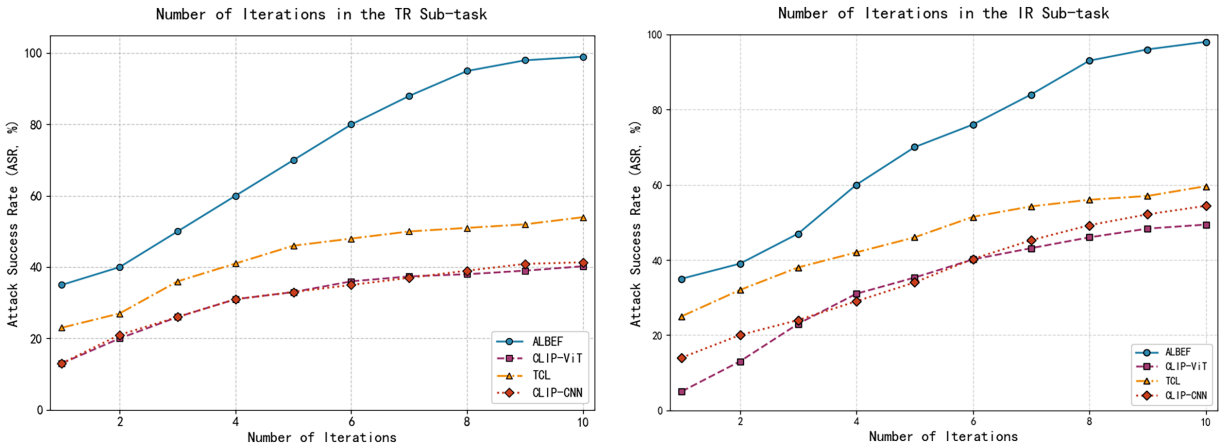
When both modules are removed, the method degenerates to the baseline, achieving 49.0% and 57.71% in TR R@1 and IR R@1, and 26.52% and 39.48% in TR R@5 and IR R@5, respectively. After introducing the SAM-AET module alone, the attack success rate increases to 51.23% (TR R@1) and 59.33% (IR R@1), as well as 27.32% (TR R@5) and 41.23% (IR R@5), indicating that incorporating mismatched negative samples effectively enhances perturbation diversity and improves transfer attack performance.

When only the CFSA module is applied, the performance improves to 50.36% and 59.45% in TR R@1 and IR R@1, and 26.89% and 41.46% in TR R@5 and IR R@5, demonstrating that cross-modal fusion semantic optimization can further refine adversarial perturbations at a fine-grained semantic level.

When both SAM-AET and CFSA are jointly employed, the model achieves the best performance, reaching 54.58% and 59.61% in TR R@1 and IR R@1, and 28.31% and 42.13% in TR R@5 and IR R@5.

These results indicate that the two modules are complementary, where SAM-AET enhances perturbation diversity and CFSA further strengthens semantic-level disruption, jointly improving the transferability of adversarial examples.

This study further analyses the relationship between the attack success rate and the number of adversarial iterations (T) in the two sub-tasks of image–text retrieval, as illustrated in Fig. 8.



(a) TR R@1 ASR (b) IR R@1 ASR

Figure 8: Iteration counts of attack success rate (ASR) on different target models. (**a**) illustrates the attack success rate curve for the text retrieval subtask as a function of the number of perturbation iterations, while (**b**) illustrates the corresponding attack success rate curve for the image retrieval subtask with respect to the number of perturbation iterations.

For both the TR and IR tasks, the attack performance becomes progressively more effective as the number of iterations increases. However, when the iteration number exceeds ($T > 8$), the gradient magnitude of adversarial updates decreases and the performance curve gradually flattens, indicating that the attack performance begins to converge. Accordingly, this study sets the iteration step to ($T = 10$) in order to balance attack success rate and computational cost.

As shown in Table 4, we analyze the sensitivity of the key hyperparameter τ in the proposed dynamic threshold strategy under the ALBEF-to-TCL transfer setting on the Flickr30K dataset. Specifically, τ is varied from 0.1 to 0.5. When τ is set to a small value, more samples are forwarded to the CFSA module for further refinement, resulting in longer runtime. As τ increases, the number of samples triggering the subsequent optimization stage decreases, leading to a gradual reduction in the average runtime.

Table 4: Sensitivity analysis of the dynamic threshold τ.

τ	TR R@1	IR R@1	Avg. Runtime (ms/sample)
0.1	54.02	59.08	168
0.2	54.61	59.83	149
0.3	**54.58**	**59.61**	**132**
0.4	54.41	59.43	124
0.5	53.12	58.21	117

Note: The bolded parts represent the best data for this indicator.

In terms of attack performance, the proposed method achieves the best results at $\tau = 0.3$, where TR R@1 and IR R@1 reach 54.58% and 59.61%, respectively. When τ is further increased, the attack performance shows a slight decline, indicating that an excessively large threshold may cause premature stopping and weaken semantic-level perturbation.

Overall, the proposed method is relatively stable within a reasonable range of τ, and $\tau = 0.3$ provides a good trade-off between attack effectiveness and computational efficiency.

To further evaluate the computational efficiency of the proposed dynamic threshold strategy, we compare the method with and without threshold control in terms of runtime, memory consumption, and CFSA trigger frequency. As shown in Table 5, without the threshold mechanism, CFSA is applied to all samples, resulting in a trigger rate of 100%, higher computational cost (180 ms/sample), and increased memory usage (1820 MB). In contrast, the proposed method selectively activates CFSA for only 63% of samples, reducing the average runtime to 132 ms/sample and memory consumption to 1650 MB.

Table 5: Efficiency analysis of the dynamic threshold strategy.

Setting	Avg. Runtime (ms/sample)	CFSA Trigger Rate (%)	Avg. Memory (MB)	TR R@1 (ALBEF→TCL)
Ours w/o threshold	180	100	1820	54.51
Ours	132	63	1650	54.58

Importantly, this efficiency gain is achieved without sacrificing attack performance. The TR@1 slightly improves from 54.51% to 54.58%, indicating that the dynamic threshold effectively avoids unnecessary optimization while maintaining, and even slightly enhancing, adversarial transferability. These results demonstrate that the proposed strategy achieves a better trade-off between computational efficiency and attack effectiveness.

5 Conclusion

This study proposes a transfer-based method, with the objective of enhancing adversarial transferability across VLP architectures. First, a Semantic-Aligned Misguided Adversarial Triangle module is designed, which introduces mismatched negative modality samples to guide adversarial example generation. Subsequently, a dynamic threshold attack strategy is constructed to determine whether further optimization is required by evaluating the similarity between adversarial image–text pairs. This mechanism reduces computational cost while improving overall efficiency. Finally, a Cross-modal Fusion Semantic Attack module based on a multimodal encoder introduces perturbations at the semantic level. By extracting and leveraging multiple intermediate-layer features, the module guides adversarial example generation and promotes the production of perturbations that transfer effectively across different VLP models. Experimental results demonstrate that the proposed method achieves maximum attack success rates of 95.3% on the Flickr30K dataset and 70.17% on the MSCOCO dataset. These findings indicate that the proposed approach exhibits strong adversarial effectiveness. This study primarily focuses on transferable adversarial attacks in the digital domain. As discussed in the experimental analysis, although the proposed method achieves strong transferability across different VLP models, its extension to physical-world multimodal attacks remains an open challenge. Future work will therefore investigate physically realizable perturbation strategies that are more robust to real-world transformations while preserving their effectiveness in disrupting cross-modal semantic alignment.

Acknowledgement: The authors are deeply grateful to all team members involved in this research.

Funding Statement: This work was partially supported by the National Natural Science Foundation of China (Grant No. 62303375). Additional support was provided by the Key Research and Development Program of Shaanxi Province (Grant Nos. 2024CY2-GJHX-43, 2024CY2-GJHX-49), the Key Scientific Research Program of Education Department of Shaanxi Province under Grant Nos. 24JR110, 24JR111, and in part by the Youth Innovation Team of Shaanxi Universities.

Author Contributions: Conceptualization: Zhichao Pei and Ou Ye; Methodology: Zhichao Pei, Ou Ye, and Panyu Yang; Formal analysis and investigation: Zhichao Pei, Ou Ye, and Panyu Yang; Writing—original draft preparation: Zhichao Pei and Ou Ye; Writing—review and editing: Zhichao Pei and Ou Ye; Funding acquisition: Ou Ye; Supervision: Ou Ye and Kaiwen He. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data available on reasonable request from the authors.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen FL, Zhang DZ, Han ML, Chen XY, Shi J, Xu S, et al. VLP: a survey on vision-language pre-training. Mach Intell Res. 2023;20(1):38–56. doi:10.1007/s11633-022-1369-5.
2. Cao M, Li S, Li J, Nie L, Zhang M. Image-text retrieval: a survey on recent research and development. arXiv:2203.14713. 2022. Available from: https://arxiv.org/abs/2203.14713.
3. Han X, Xu G, Zhou Y, Yang X, Li J, Zhang T. Physical backdoor attacks to lane detection systems in autonomous driving. In: Proceedings of the 30th ACM International Conference on Multimedia. New York, NY, USA: ACM; 2022. p. 2957–68. doi:10.1145/3503161.3548171.
4. Sun K, Xue S, Sun F, Sun H, Luo Y, Wang L, et al. Medical multimodal foundation models in clinical diagnosis and treatment: applications, challenges, and future directions. Artif Intell Med. 2025;170(1):103265. doi:10.1016/j.artmed.2025.103265.
5. Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A. Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. New York, NY, USA: ACM; 2017. p. 506–19. doi:10.1145/3052973.3053009.
6. Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: reliable attacks against black-box machine learning models. arXiv:1712.04248. 2017.
7. Cheng M, Singh S, Chen P, Chen PY, Liu S, Hsieh CJ. Sign-OPT: a query-efficient hard-label adversarial attack. arXiv:1909.10773. 2019.
8. Chen J, Jordan MI, Wainwright MJ. HopSkipJumpAttack: a query-efficient decision-based attack. In: Proceedings of the 2020 IEEE Symposium on Security and Privacy (SP); 2020 May 18–21; San Francisco, CA, USA. p. 1277–94. doi:10.1109/sp40000.2020.00045.
9. Yu Z, Ti Y, Ye O, Cong X, Zhang Y, Song HH. Targeted black-box adversarial example generation method using multi-layer heatmap mapping for industrial IoT. IEEE Internet Things J. 2025;13(4):5855–68. doi:10.1109/JIOT.2025.3634518.
10. Georgi MA, Nern LF, Raj H, Sharma Y. On transfer of adversarial robustness from pretraining to downstream tasks. In: Advances in Neural Information Processing Systems 36; 2023 Dec 10–16; New Orleans, LA, USA. p. 59206–26. doi:10.52202/075280-2584.
11. Zeng C, Ge YJ, Zhao LC. Survey of multimodal vision-language representation learning models and their adversarial examples attack and defense techniques. J Comput Res Dev. 2025;2(9):2208–32. doi:10.20944/preprints202511.1363.v1.
12. Zhang J, Huang J, Jin S, Lu S. Vision-language models for vision tasks: a survey. IEEE Trans Pattern Anal Mach Intell. 2024;46(8):5625–44. doi:10.1109/TPAMI.2024.3369699.

13. Lu D, Wang Z, Wang T, Guan W, Gao H, Zheng F. Set-level guidance attack: boosting adversarial transferability of vision-language pre-training models. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 102–11. doi:10.1109/ICCV51070.2023.00016.
14. Gao S, Jia X, Ren X, Tsang I, Guo Q. Boosting transferability in vision-language attacks via diversification along the intersection region of adversarial trajectory. In: Computer Vision—ECCV 2024. Cham, Switzerland: Springer Nature; 2024. p. 442–60. doi:10.1007/978-3-031-72998-0_25.
15. Fang H, Kong J, Yu W, Chen B, Li J, Wu H, et al. One perturbation is enough: on generating universal adversarial perturbations against vision-language pre-training models. arXiv:2406.05491. 2024.
16. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. arXiv:1312.6199. 2013.
17. Goodfellow IJ, Shlens J, Szegery C. Explaining and harnessing adversarial examples. In: Proceedings of the International Conference on Learning Representations (ICLR); 2015 May 5–7; San Diego, CA, USA.
18. Baluja S, Fischer I. Learning to attack: adversarial transformation networks. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2018 Feb 2–7; New Orleans, LA, USA.
19. Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP); 2017 May 22–26; San Jose, CA, USA. p. 39–57. doi:10.1109/SP.2017.49.
20. Jin D, Jin Z, Zhou JT, Szolovits P. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. Proc AAAI Conf Artif Intell. 2020;34(5):8018–25. doi:10.1609/aaai.v34i05.6311.
21. Jia R, Liang P. Adversarial examples for evaluating reading comprehension systems. arXiv:1707.07328. 2017.
22. Chen L, Sun J, Xu W. FAWA: fast adversarial watermark attack on optical character recognition (OCR) systems. In: Machine learning and knowledge discovery in databases. Cham, Switzerland: Springer International Publishing; 2021. p. 547–63. doi:10.1007/978-3-030-67664-3_33.
23. Xu X, Chen J, Xiao J, Wang Z, Yang Y, Shen HT. Learning optimization-based adversarial perturbations for attacking sequential recognition models. In: Proceedings of the 28th ACM International Conference on Multimedia. New York, NY, USA: ACM; 2020. p. 2802–22. doi:10.1145/3394171.3413543.
24. Zhou Z, Hu S, Li M, Zhang H, Zhang Y, Jin H. AdvCLIP: downstream-agnostic adversarial examples in multimodal contrastive learning. In: Proceedings of the 31st ACM International Conference on Multimedia. New York, NY, USA: ACM; 2023. p. 6311–20. doi:10.1145/3581783.3612454.
25. Huang H, Erfani S, Li Y, Ma X, Bailey J. X-transfer attacks: towards super transferable adversarial attacks on CLIP. arXiv:2505.05528. 2025.
26. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083. 2017.
27. Li L, Ma R, Guo Q, Xue X, Qiu X. BERT-ATTACK: adversarial attack against BERT using BERT. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 Nov 8–12; Online. p. 6193–202. doi:10.18653/v1/2020.emnlp-main.500.
28. Zhang J, Yi Q, Sang J. Towards adversarial attack on vision-language pre-training models. In: Proceedings of the 30th ACM International Conference on Multimedia. New York, NY, USA: ACM; 2022. p. 5005–13. doi:10.1145/3503161.3547801.
29. Wang H, Dong K, Zhu Z, Qin H, Liu A, Fang X, et al. Transferable multimodal attack on vision-language pre-training models. In: Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP); 2024 May 19–23; San Francisco, CA, USA. p. 1722–40. doi:10.1109/SP54263.2024.00102.
30. Li J, Selvaraju R, Gotmare A, Joty S, Xiong C, Hoi S. Align before fuse: vision and language representation learning with momen-tum distillation. Adv Neural Inf Process Syst. 2021;34:9694–705.
31. Yang J, Duan J, Tran S, Xu Y, Chanda S, Chen L, et al. Vision-language pre-training with triple contrastive learning. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 15650–9. doi:10.1109/CVPR52688.2022.01522.
32. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. arXiv:2103.00020. 2021.

33. Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S. Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV); 2015 Dec 7–13; Santiago, Chile. p. 2641–9. doi:10.1109/ICCV.2015.303.
34. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: Computer Vision—ECCV 2014. Cham, Switzerland: Springer International Publishing; 2014. p. 740–55. doi:10.1007/978-3-319-10602-1_48.