Computer Modeling in
Engineering & Sciences

Tech Science Press

Doi:10.32604/cmes.2025.067658

ARTICLE

AMA: Adaptive Multimodal Adversarial Attack with Dynamic Perturbation Optimization

Yufei Shi, Ziwen He[*], Teng Jin, Haochen Tong and Zhangjie Fu

Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing, 210044, China

*Corresponding Author: Ziwen He. Email: ziwen.he@nuist.edu.cn

Received: 09 May 2025; Accepted: 23 July 2025; Published: 31 August 2025

ABSTRACT: This article proposes an innovative adversarial attack method, AMA (Adaptive Multimodal Attack), which introduces an adaptive feedback mechanism by dynamically adjusting the perturbation strength. Specifically, AMA adjusts perturbation amplitude based on task complexity and optimizes the perturbation direction based on the gradient direction in real time to enhance attack efficiency. Experimental results demonstrate that AMA elevates attack success rates from approximately 78.95% to 89.56% on visual question answering and from 78.82% to 84.96% on visual reasoning tasks across representative vision-language benchmarks. These findings demonstrate AMA's superior attack efficiency and reveal the vulnerability of current visual language models to carefully crafted adversarial examples, underscoring the need to enhance their robustness.

KEYWORDS: Adversarial attack; visual language model; black-box attack; adaptive multimodal attack; disturbance intensity

1 Introduction

In recent years, visual language tasks have made significant progress in the fields of computer vision, natural language processing, and information security [1–5]. Especially with the advancement of pre-trained models including ViLT [6], CLIP [7], BLIP [8], OFA [9] and UniTAB [10], they have exhibited strong performance and generalization capabilities in tasks such as visual question answering (VQA), visual entailment, visual reasoning, referring expression comprehension, image captioning, and image classification. However, recent studies have raised concerns about the vulnerability of these models to adversarial attacks. As shown in Fig. 1, vision language models can be destabilized by relatively small cross-modal perturbations. This undermines the reliability of their semantic alignment mechanisms and indicates that adversarial perturbations can cause the model to make misjudgments. This leads to the core research question of this article: the challenge of creating adaptive and covert adversarial attacks that maintain effectiveness across multiple architectures and tasks. The goal is to design more efficient, generalized and covert adversarial attack methods to enhance the robustness evaluation of visual language models and promote the safe application of these models.

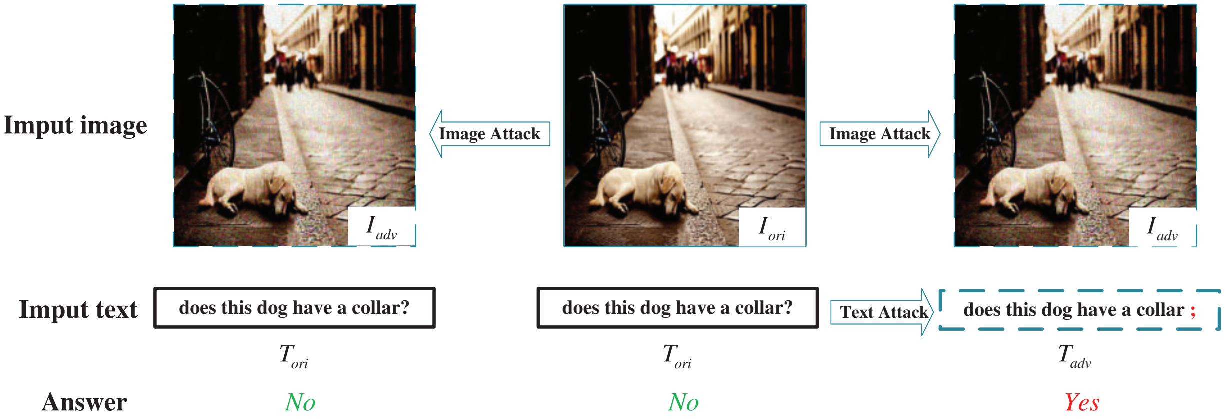

Figure 1: An example of adversarial attack on VQA. The middle shows the original image and text, the left shows the adversarial image generated through image attacks, and the right shows both the adversarial image and text generated through multimodal attack. The incorrect answers after successful attack are marked in red

Most existing research focuses on white box attack settings [11,12], where attackers can access the gradient information of the model. In a black-box attack environment, attackers can only access publicly available pre-trained models without prior knowledge of fine-tuning models for downstream tasks [13]. Implementing effective attacks in this setting remains an important challenge. In recent years, significant research efforts have been devoted to black-box adversarial attacks in visual language tasks. The primary focus has been on enhancing the transferability and effectiveness of these attacks. Zhao et al. [14] systematically evaluated the adversarial robustness of visual language models in black-box environments for the first time. Han et al. [15] proposed the OTAttack method, which overcomes the problem of poor transferability between different models in traditional attack methods by optimizing the mapping relationship between images and text through optimal transport and improves the effectiveness of black-box attacks. Lu et al. [16] proposed the set-level guidance attack (SGA) method, which improves the interaction between modals to improve the transitivity of adversarial samples and performs well in multiple downstream tasks, especially in black box settings, successfully increasing the success rate of attacks. In addition, Yin et al. [17] proposed the VLATTACK method, which is a black-box attack method that utilizes pre-trained models to generate multimodal adversarial samples. By integrating visual and textual perturbations, it successfully improves the transferability of adversarial attacks and achieves efficient attack effects in multiple visual language tasks. Although they have made many contributions to the adversarial attack of visual language models, these methods still have significant limitations: insufficient attack efficiency and difficulty in generating effective adversarial samples in complex tasks; The cross model generalization ability is weak, and the attack effect fluctuates greatly between different architecture models. These issues constrain the further development of adversarial attack research.

To address the above limitations, this paper introduces a black-box adversarial attack approach called Adaptive Multimodal Attack (AMA). This approach enhances previously employed frameworks to increase the attack success rate. First, we introduce a Dynamic Adaptive Perturbation Strength (DAPS) strategy and a Stepwise Refinement Optimization (SRO) strategy. DAPS dynamically adjusts the disturbance strength based on the importance scores of features at each layer of the model, while SRO iteratively optimizes the perturbation direction to maximize its impact on model predictions. Second, in the multimodal attack stage, we propose a Task Difficulty Adaptive Adjustment (TDAA) strategy and an Adaptive Feedback Mechanism (AFM). TDAA dynamically adjusts the intensity of image and text perturbations based on the task difficulty, which is evaluated by the model's confidence level. AFM, on the other hand, dynamically adjusts the perturbation strategies by evaluating the impact of image and text perturbations on the prediction results after each iteration. These strategies enable AMA to adaptively optimize the attack process according to the specific characteristics of different tasks and models, thereby improving the attack success rate.

We conducted extensive experiments on mainstream pre-trained visual language models and demonstrated that AMA has significantly higher attack success rates than existing methods in visual question-answering and visual reasoning tasks. Our experimental results not only highlight the effectiveness of AMA, but also emphasize the urgent need to enhance the robustness of current visual language models to complex adversarial attacks.

2 Related Work

2.1 Single-Modal Adversarial Attack

2.1.1 Image-Attack

Image-based adversarial attacks are one of the earliest studied forms of unimodal adversarial attacks. Attackers mislead deep learning models into producing incorrect classification results by adding carefully designed perturbations to images, which are often difficult for humans to detect. Common adversarial attack methods include gradient-based methods such as FGSM [18] and PGD [19], which calculate the gradient of the input image relative to the output of the model, add perturbations in the gradient direction and shift the model output in the wrong direction. There are also optimization-based methods, such as L-BFGS [20], C&W attacks [21], and Deepfool [22], which model the process of generating adversarial samples as an optimization problem. By optimizing the objective function to find the optimal perturbation that causes the model to fail, although multiple iterations and complex optimization processes are required, the generated adversarial samples often have a high success rate for attacks. In addition, methods based on Generative Adversarial Networks (GANs) use GAN generators to generate adversarial samples [23,24]. The generator takes noise as input and learns through training how to generate adversarial samples that can deceive the discriminator (that is, the target model). Finally, score-based methods indirectly infer gradient information from attackers using the output scores of the target model (such as softmax probability) to guide the search for adversarial perturbations, such as Physical One-Pixel Attack [25] and Time-aware Perception Attack [26].

2.1.2 Text-Attack

Text Adversarial Attack refers to the modification of text that affects its semantics, credibility, authenticity, and other aspects, resulting in errors in the classification or prediction results of deep learning models, while preserving its semantics and not affecting human understanding. The target of the attack is to make small perturbations of the raw data to maximize the prediction error results. Research on adversarial text attacks has made certain progress, and various attack methods have emerged. Li et al. proposed the BERT-Attack [27], which utilizes the Mask Language Model (MLM) feature of the BERT model to generate adversarial samples by replacing words in the text. HotFlip is a character-level white-box attack method that generates adversarial samples through atomic flipping operations [28]. In addition, some studies draw on the Jacobian-based Saliency Map Attack (JSMA) algorithm in the field of imaging [29], using computational graph unfolding techniques to evaluate forward derivatives related to the embedding input of word sequences, constructing Jacobian matrices and combining the idea of FGSM to calculate adversarial perturbations. DeepWordBug [30] is a black-box character level attack method that uses a new scoring strategy to identify key characters and sort them, changing the original classification by simply replacing the character. TEXTFOOLER [31] has successfully implemented black-box attacks on pre-trained

BERT models, convolutional neural networks, and recurrent neural networks in two types of tasks: text classification and text embedding. It has shown excellent performance in attack effectiveness, computational efficiency, and semantic and syntactic integrity. Yang et al. [32] proposed a prompt-based adversarial sample generation technique (PAT), which generates natural, smooth, and diverse adversarial samples by constructing malicious prompt templates. This method outperforms traditional search methods in terms of the naturalness and diversity of generated adversarial samples, and significantly improves the robustness of the model in adversarial training.

Single-modal attack methods have limitations in disrupting multimodal representations, as the decisions of visual language models rely on the joint representation of text and visual features. Therefore, single-modal attack methods cannot fully disrupt the multimodal representation of visual language models, thereby limiting their effectiveness on these models.

2.2 Multimodal Adversarial Attack

In the field of multimodal attacks, researchers have proposed various innovative methods to address the complexity of visual language models. For example, Co-Attack [11] successfully deceived multiple visual language models by co-perturbing image and text modalities. However, Co-Attack lacks adaptability to varying input semantics, which limits its effectiveness in complex scenarios. On this basis, M-Attack [33] further optimized the attack strategy, especially in black-box scenarios, significantly improving the transferability of adversarial samples through local semantic aggregation perturbation and model integration strategies, making it perform well on multiple commercial visual language models including GPT-4.5, GPT-4o, and o1. At the same time, MMA-Diffusion [34] targets the text-to-image diffusion model and successfully circumvents existing defense mechanisms by combining text and visual-modality attack strategies, revealing potential security vulnerabilities in text-to-image technology. However, the complexity of multimodal attacks also brings new challenges, such as the cross-modal transferability of attacks and the multimodal robustness of models. For the VQA task, the Fool-VQA [35] side iteratively adds pixel-level perturbations to the image to achieve the attack. For image text retrieval tasks, CMLA [36] and AACH [37] increase the Hamming distance between image and text hash codes by adding perturbations, resulting in incorrect image text matching results. In addition, Yin et al. [17] proposed the VLATTACK method, which is a black-box attack method that uses pre-trained models to generate multimodal adversarial samples. Although VLAttack achieves some success by integrating visual and textual perturbations, it applies fixed perturbation strategies that limit its flexibility, especially when dealing with diverse model architectures and tasks.

These methods often demand more samples or a larger perturbation budget to succeed in attacks. AMA surpasses them by employing dynamic adjustment and adaptive feedback. It adjusts to different tasks and models more effectively, markedly cutting the samples needed for success and the perturbation budget. Thus, AMA is more efficient and potent in creating adversarial samples than prior methods.

3 Methodology

3.1 Overview of Adaptive Multimodal Attack (AMA)

Adaptive Multimodal Attack (AMA) tackles efficiency and generalization issues in traditional adversarial attacks on visual language models. Its framework, illustrated in Fig. 2, uses a visual VQA task as an example. AMA generates adversarial samples via two stages: single-modal and multimodal attacks. In the single-modal stage, it employs DAPS and SRO strategies to dynamically adjust the perturbation strength based on the importance of the model feature layer and iteratively optimize the perturbation direction. In the multimodal stage, the TDAA and AFM mechanisms adjust the intensity of the image and text perturbation according to the difficulty of the task and refine strategies using iteration results. This dynamic approach improves attack efficiency and success rates compared to traditional methods with fixed disturbance intensity or single optimization strategies.

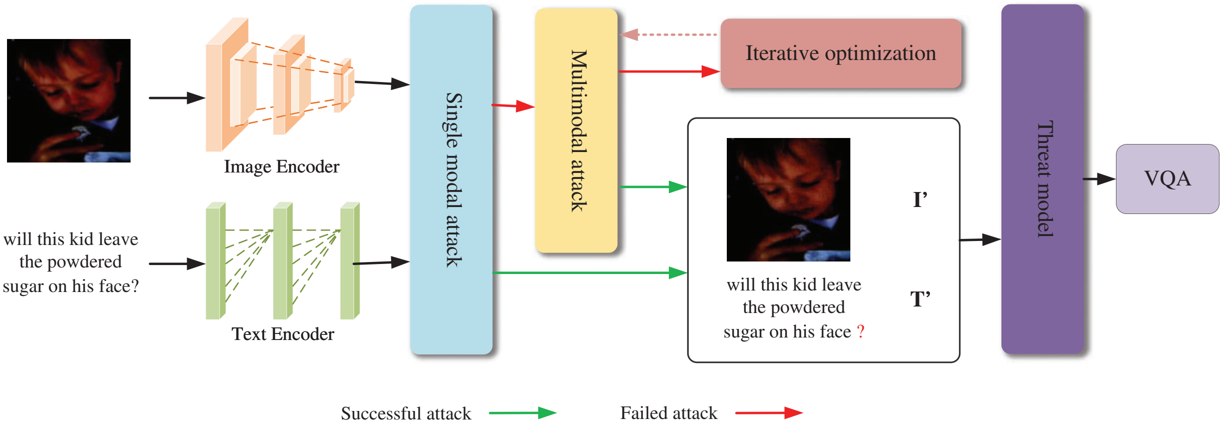

Figure 2: The overall illustration of the proposed AMA

In practice, the single-modal attack first perturbs the image and text separately. If the attack succeeds, the final adversarial samples are achieved; otherwise, the process progresses to the multimodal attack. In the multimodal attack, image and text perturbations are iteratively updated and optimized using TDAA and AFM until adversarial samples that can mislead the model are generated. This dynamic process contrasts sharply with traditional fixed adversarial attack methods, offering greater flexibility. Specific details will be presented in Sections 3.2 and 3.3.

3.2 Single-Modal Adversarial Attack

The goal of the single-modal attack stage is to perturb the image and text modalities separately, trying to alter the model prediction results using one modality individually. The core of single-modal attacks lies in how to efficiently generate perturbations while ensuring that the strength of the perturbations can maximize the attack effect. In single-modal attacks, we introduce two strategies: dynamic adaptive perturbation intensity and gradual refinement optimization. The following will provide a detailed explanation of these two strategies.

3.2.1 Dynamic Adaptive Perturbation Strength

In single-modal attacks, we propose a Dynamic Adaptive Perturbation Strength (DAPS) strategy, which dynamically adjusts the perturbation strength for different layers based on their feature importance. Unlike existing approaches where the perturbation strength is fixed across all layers, DAPS assigns stronger perturbations to critical layers (e.g., high-level semantic features) and weaker perturbations to less influential layers (e.g., low-level texture features). Specifically, under a gray-box setting where we have access to an auxiliary white-box model structurally similar to the target, we compute the importance of each layer by summing the absolute gradient values of the layer parameters with respect to the loss:

$$\text{Importance}\left(L_i\right) = \sum_i \left|\nabla_{\theta_{L_i}} \mathcal{L}(x, y)\right|, \tag{1}$$

where L_i represents the i-th layer of the model, θ_L is the parameter of this layer, $\mathcal{L}$ is the cross entropy loss, $|\cdot|$ represents taking the absolute value.

Based on the feature importance score, the perturbation strength for each layer is then dynamically adjusted to:

$$\varepsilon_L = \varepsilon_0 \times (1 + \alpha \times \text{Importance}(L_i)), \tag{2}$$

where ε_0 is the initial perturbation strength, and α is a hyper-parameter, that controls the magnitude of the disturbance intensity.

This approach enables the attack to focus on layers that most influence the model's decision, improving both the attack success rate and its stealthiness. In strict black-box settings where gradients are not accessible, this importance can be approximated using gradient estimation techniques such as finite differences, though in our implementation we rely on the white-box auxiliary model for efficient importance computation.

3.2.2 Stepwise Refinement Optimization

Existing attack methods often have relatively rough optimization processes, which can lead to certain perturbations that unnecessarily impact the robustness of the model. To address this, AMA introduces the Stepwise Refinement Optimization (SRO) strategy.

The SRO optimization process begins with the initialization of the perturbation δ. In each subsequent iteration, the strategy evaluates the direction of the perturbation by calculating the cosine similarity between the joint representation of perturbed image features and problem text features and the correct answer text features. The direction of the perturbation is dynamically adjusted using the following formula.

$$\delta \leftarrow \delta + \beta \times \nabla_\delta \text{Cos}(F(x + \delta, q), F(y)), \tag{3}$$

where β is the learning rate. $F(x + \delta, q)$ represents the joint feature between the perturbed image and the problem input q while $F(y)$ represents the textual feature of the correct answer.

The iterations will cease when either a predefined maximum number of iterations is reached or the cosine similarity between the perturbed features and the target category stabilizes. This dual stopping criterion ensures that the optimization process is both efficient and effective, preventing unnecessary computations while achieving the desired level of convergence.

This distinctive approach allows SRO to iteratively refine the perturbation in a direction that significantly enhances its effect on the model's predictions, setting it apart from other attack methods that may only focus on magnitude adjustments or lack such a dynamic directional refinement mechanism. By gradually refining SRO, AMA can more accurately adjust the direction and amplitude of disturbances, thereby improving the attack success rate.

After the image modality attack, if the perturbed image does not fool the model, a text modality attack is initiated. We choose to use BERT-attack [27] to attack the text. For each perturbed text T', the cosine similarity γ_i between T' and the original text T is calculated. If γ_i exceeds the text perturbation budget σ_{txt}, T' is added to the perturbation list. If the perturbed text T' fools the model, it is returned as a result. We conclude the single-modal attack in Algorithm 1.

Algorithm 1: Single-modal attack

```
Require: A local substitute model $S$, clean image-text pair $(I, T)$, groud-truth label $y$, perturbation budget $\sigma_{\text{img}}$ on $I$, perturbation budget $\sigma_{\text{txt}}$ on $T$, initial perturbation strength $\varepsilon_0$
Ensure: Perturbed image-text pair that fools the model $S$ or None if no such pair is found
Initialize $\delta \sim \mathcal{N}(0,1)$, $\delta$ is constrained to the interval $\left[-0.5 \times \sigma_{\text{img}}, 0.5 \times \sigma_{\text{img}}\right]$
Initialize $I' = I + \delta$, $T' = T$
Initialize perturbation list $T_{\text{perturbations}} = [\,]$
//Image Modality Attack
for t = 1 to 40 do
    Adjust $\varepsilon_{\text{img}}$ by DAPS strategy using the Eq. (2), Update the $\delta$ using the Eq. (3)
    Generate perturbed image $I' = I + \varepsilon_{\text{img}} \times \delta$
    if $S(I', T) \neq y$ then
        return $(I', T)$
    end if
end for
// If image attack fails, start text attack.
//Text Modality Attack
for perturbed text $T'$ in BERT-attack
    Calculate cosine similarity $\gamma_i$ between $T'$ and $T$
    if $\gamma_i > \sigma_{\text{txt}}$ then
        Add $T'$ to $T_{\text{perturbations}}$
        if $S(I, T') \neq y$ then
            return $(I, T')$
        end if
    end if
end for
```

3.3 *Multimodal Adversarial Attack*

If the single-modal attack does not change the prediction results of the model, it enters the multimodal attack stage. Multimodal attacks generate adversarial samples by iteratively updating image and text perturbations. We introduced Task Difficulty Adjustment (TDAA) and Adaptive Feedback Mechanism (AFM) to optimize attack strategies. These improvements enable AMA to tackle complex visual language tasks more effectively.

3.3.1 Task Difficulty Adaptive Adjustment

In the multimodal attack stage, we adopt the Task Difficulty Adaptive Adjustment (TDAA) strategy to dynamically adjust the perturbation intensity of image and text modalities based on task difficulty. The task difficulty is quantified by the model's confidence in its prediction, where higher uncertainty implies a more complex task. Specifically, we define the task difficulty using the following formula.

$$\text{Difficulty} = 1 - \mathrm{C}\left(y_{\mathbf{pred}}\right), \tag{4}$$

where $\mathrm{C}\left(y_{\mathbf{pred}}\right)$ is the confidence score of the model in its predicted result $y_{\mathbf{pred}}$.

Based on this difficulty measure, we linearly adjust the perturbation intensities as follows:

$$\varepsilon_{\text{img}} = \varepsilon_{\text{base}} \times (1 + \gamma \times \text{Difficulty}), \tag{5}$$

$$\varepsilon_{\text{txt}} = \varepsilon_{\text{base}} \times (1 - \gamma \times \text{Difficulty}), \tag{6}$$

where ε_0 is the initial perturbation strength, and γ is the adjustment parameter.

This linear adjustment offers a simple yet effective way to shift perturbation emphasis based on task type: when the task is hard, more image perturbation is applied to ensure sufficient feature disruption, while the text perturbation is restrained to maintain semantic plausibility. The linear form is chosen for its interpretability, numerical stability, and empirical effectiveness, which we validated via ablation studies (see Fig. 5) showing consistent performance improvement over fixed-weight methods. Similar confidence-based perturbation modulation has been explored in prior works [15,17] supporting the validity of our approach.

3.3.2 Adaptive Feedback Mechanism

The Adaptive Feedback Mechanism (AFM) optimizes multimodal attacks by dynamically adjusting the direction and intensity of perturbations based on the task prediction results after each perturbation update. Unlike existing methods such as VLAttack [17] that use fixed alternating updates for image and text perturbations, AFM evaluates their impact on predictions and adjusts strategies accordingly. The process starts with initializing the image and text perturbations. In each iteration, the predicted results after disturbance are calculated. Then, the disturbance direction is adjusted on the basis of these results.

Specifically, the perturbation for the image is updated using the following formula.

$$\delta_{\text{img}} \leftarrow \delta_{\text{img}} + \beta \times \nabla_{\delta_{\text{img}}} \mathcal{L}\big(F(x + \delta_{\text{img}}, q), y\big), \tag{7}$$

where $\mathcal{L}$ is the loss function, $F(x + \delta_{\text{img}}, q)$ is the predicted model results after applying the current perturbations to the image inputs, y is the correct prediction.

For text perturbations, we start by generating a range of candidate text perturbations using the BERT-Attack [27]. For each candidate text perturbation, we compute its semantic similarity to the original text. Only candidates with a semantic similarity above a preset threshold are kept. For the selected candidate text perturbations, calculate the loss function $\mathcal{L}(F(x, q + \delta_{\text{txt}}), y)$ and select the optimal text perturbation based on the loss value. Through this approach, the text attack component can effectively generate adversarial samples while maintaining semantic consistency.

Through an adaptive feedback mechanism, the attack process of AMA is more flexible and can be optimized based on the specific requirements of the task, thus improving the success rate of the attack. We conclude the multimodal attack in Algorithm 2.

Algorithm 2: Multimodal attack

```
Require: Perturbation list $T_{\text{perturbations}}$, fine-tuned model $S$, clean image $I$, clean text $T$, initial perturbation strength $\varepsilon_{\text{base}}$, adjustment parameter $\gamma$, learning rate $\beta$, number of top samples $K$
Ensure: Perturbed image-text pair $(I', T')$ that fools the model $S$ or None if no such pair is found
Rank $T_{\text{perturbations}}$ and select top-K samples
Initialize $\delta_{\text{img}}$ and $\delta_{\text{txt}}$
Calculate initial Difficulty
for $k = 1$ to $K$ do
```

(Continued)

Algorithm 2 (continued)

```
    Adjust $\varepsilon_{\text{img}}$ and $\varepsilon_{\text{txt}}$ using Eqs. (5) and (6)
    Update $\delta_{\text{img}}$ using Eq. (7)
    Generate perturbed image $I'_k = I + \varepsilon_{\text{img}} \times \delta_{\text{img}}$
    Generate perturbed text $T'_k$ by applying the top-$k$ text perturbation from $T_{\text{perturbations}}$
    **if** $S(I'_k, T'_k) \neq y$ **then**
        **return** $(I'_k, T'_k)$
    **end if**
    **for** each candidate $T'_i$ **do**
        Calculate semantic similarity $\gamma_i$
        **if** $\gamma_i > \sigma_{\text{txt}}$ **then**
            Update $\delta_{\text{txt}}$ based on loss gradient // Feedback-based update
        **end if**
    **end for**
    Select top adversarial text $T'_k$
    **if** $S(I'_{k+1}, T'_k) \neq y$ **then**
        **return** $(I'_{k+1}, T'_k)$
    **end if**
    Update Difficulty
**end for**
**return** None
```

4 Experiments

In order to comprehensively evaluate the effectiveness of our proposed AMA in visual language tasks, we designed a series of experiments that followed strict scientific methods and the principles of controlled variable.

4.1 Experimental Setting

Pre-trained VL Models and Tasks In this study, we selected four mainstream pre-trained visual language (VL) models to evaluate the effectiveness of our proposed AMA method: ViLT (Vision-and-Language Transformer Without Convolution or Region Supervision) [6], BLIP [6], Unitab [10], and OFA (One-For-All) [9]. These models were chosen because of their strong performance and generalization capabilities in various visual language tasks. The experiments were carried out on their original architecture and parameter settings to ensure fairness and comparability. We focus on two primary visual language tasks: Visual Question Answering (VQA): This task involves answering questions about images. We used the VQAv2 dataset, which contains a large number of pairs of image questions and corresponding answers. We selected 5000 of them for the experiment. Visual Reasoning (VR): This task requires inferring the relationship between images and textual descriptions. For this purpose, we selected 5000 sets of image text pairs from the NLVR2 [38] dataset for experimentation.

Baseline Methods To establish a comprehensive baseline, we compared AMA with several state-of-the-art adversarial attack methods. For single-modal attacks, we included methods that target both image and text patterns. Specifically, for image pattern attacks, we compared AMA with DR [39] and Block-wise Similarity Attack (BSA) [17], which are representative methods for generating adversarial perturbations in visual data. For text pattern attacks, we used BERT-Attack (B&A) [40] and R&R [41], which are effective in modifying textual inputs to mislead deep learning models. Furthermore, for multimodal attacks, we compared AMA with Co-attack [11] and VLAttack [17], which are designed to exploit interactions between images and text to enhance attack effectiveness. These baselines provide a detailed comparison to validate the superiority and robustness of AMA in different attack scenarios.

Perturbation Budget Configuration To ensure a fair comparison across different adversarial attack methods, the maximum perturbation budgets were standardized across all experiments. Specifically, the image perturbation budget σ_{img} was set to 0.2, and the text perturbation budget σ_{txt} was set to 0.15. In AMA, the initial perturbation strength ε_0 for DAPS was set to 0.125, and the hyper-parameter α was fixed at 0.3. For multimodal perturbations, the base perturbation strength ε_{base} for TDAA was set to 0.1, while the adjustment parameter γ was defined as 0.3. These hyper-parameters were determined through careful experimentation and validation to achieve a balance between attack effectiveness and perturbation magnitude. They ensure a dynamic and fair comparison between different adversarial attack methods while maintaining the flexibility required for task-specific adaptations.

Evaluation Metrics The performance of AMA was evaluated using the Attack Success Rate (ASR), which measures the percentage of successful attacks. Higher ASR indicates better attack performance. ASR is calculated as the ratio of successfully attacked samples to the total number of samples attempted.

4.2 Experimental Results

4.2.1 Single-Modal Attack Performance

The performance of AMA in single-modal attacks was rigorously evaluated against several state-of-the-art baseline methods across different models and tasks. The results presented in Table 1 demonstrate that AMA significantly outperforms all baseline methods in single-modal attacks. Specifically, AMA achieves an attack success rate (ASR) of 37.70% on the VQA task using the BLIP model, which is 12.26% higher than the next best method (DR, 25.4%). Similarly, on the ViLT model for the VQA task, AMA achieves an ASR of 76.8%, outperforming the baseline methods by a substantial margin. These results highlight the effectiveness of AMA's dynamic adaptive perturbation strength (DAPS) and stepwise refinement optimization (SRO) strategies in generating adversarial samples that are both robust and transferable.

Table 1: Comparison of AMA single-modal attacks and baseline on ViLT, BLIP, Unitab and OFA for different tasks under identical experimental settings. All results are displayed by ASR (%)

Pre-trained model	Task	Dataset	Image only		Text only		Ours
			DR	BSA	B&A	R&R	
BLIP	VQA	VQAv2	7.04	25.4	21.04	2.94	37.70
	VR	NLVR2	6.66	27.6	33.08	16.92	48.34
ViLT	VQA	VQAv2	23.89	65.20	17.24	8.69	76.8
	VR	NLVR2	21.58	52.17	33.08	16.69	61.3
Unitab	VQA	VQAv2	22.8	48.40	14.20	5.48	59.66
OFA	VQA	VQAv2	25.06	54.05	10.22	2.34	67.03

The superior performance of AMA can be attributed to its ability to dynamically adjust perturbation strength based on feature importance, ensuring that the perturbations are optimally targeted to maximize their impact on model predictions. Additionally, the stepwise refinement optimization strategy allows AMA to iteratively refine perturbations, further enhancing attack efficiency and success rates. These findings underscore the importance of exploiting feature hierarchy and perturbation optimization in adversarial attacks on visual-language models.

We also evaluated the CLIP image classification task in the SVHN dataset. Specifically, we use CLIP's image encoder as a pre-trained model, and then fine tune the SVHN dataset after adding linear classification heads. For the CLIP image encoder, we selected ViT-B/16 and ResNet-50, and tested the attack performance using 5000 correctly predicted samples. All results are shown in Table 2. Due to the fact that this task only accepts images as input, we compared our method with other baselines for image attack. As shown in the table, our proposed unimodal method still maintains the best ASR using different image encoder structures, clearly demonstrating its effectiveness.

Table 2: CLIP model evaluation on SVHN under identical experimental settings. All results are displayed by ASR (%)

Dataset	Method	CLIP-ViT/16	CLIP-RN50
SVHN	DR	3.32	71.62
	SSP	6.36	84.26
	FDA	6.20	83.52
	BSA	15.74	84.98
	Ours	24.86	85.78

4.2.2 Multimodal Attack Performance

The performance of AMA in multimodal attacks was also fully evaluated compared to baseline methods. The results presented in Table 3 show that AMA achieves significantly higher attack success rates in multimodal settings compared to existing methods. For example, on the BLIP model for the VQA task, AMA achieves an ASR of 61.82%, which is 13.12% higher than Co-attack (14.2%) and 13.12% higher than VLAttack (48.7%). Similarly, on the ViLT model for the VQA task, AMA achieves an ASR of 89.56%, outperforming Co-attack (35.13%) and VLAttack (78.05%) by a substantial margin.

Table 3: Comparison of AMA with baselines on ViLT, BLIP, Unitab and OFA for different tasks under identical experimental settings. All results are displayed by ASR (%)

Pre-trained model	Task	Dataset	Co-attack	VLAttack	Ours
BLIP	VQA	VQAv2	14.2	48.7	61.82
	VR	NLVR2	8.7	52.66	73.58
ViLT	VQA	VQAv2	35.13	78.05	89.56
	VR	NLVR2	42.04	66.65	84.96
Unitab	VQA	VQAv2	33.87	62.20	73.49
OFA	VQA	VQAv2	51.16	78.82	90.36

These results demonstrate the effectiveness of AMA's TDAA and AFM in optimizing perturbation strategies for complex visual-language tasks. By dynamically adjusting perturbation intensity based on task difficulty and using feedback from previous iterations, AMA is able to generate adversarial samples that are highly effective in deceiving visual-language models. This highlights the importance of cross-modal interaction and adaptive optimization in multimodal adversarial attacks.

To further validate the stability and reliability of AMA, we conducted five independent runs for each experimental setting. Table 4 provides the mean ASR and standard deviation, while Fig. 3 shows the box plot of ASR results across these runs. This analysis confirms the robustness and consistency of the proposed method under varying conditions.

Table 4: Comparison of mean ASR and standard deviation of different models. MA means mean ASR, SD means standard deviation

Model	MA (%)	SD (%)
BLIP	59.48	1.65
ViLT	87.65	1.33
Unitab	72.15	1.11
OFA	88.55	1.25

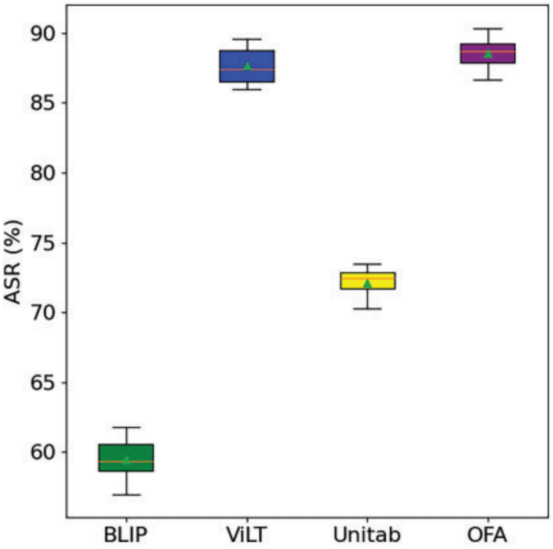

Figure 3: Comparison of average ASR and standard deviation of different models

4.3 Semantic and Perceptual Fidelity Evaluation

To further assess the semantic and perceptual fidelity of the adversarial samples generated by AMA, we performed additional experiments using metrics for both the text and the image modalities. For the text, we used BERTScore and BiLingual Evaluation Understudy—4-gram (BLEU-4) to assess whether the perturbations altered the intended meaning of the text. For images, we used SSIM (Structural Similarity Index) and LPIPS (Learned Perceptual Image Patch Similarity) to quantify the perceptual similarity between the original and perturbed images. Table 5 presents the results of these supplementary metrics on the VQAv2 dataset using the ViLT model. SSIM scores greater than 0.9 are generally considered acceptable for perceptual similarity, indicating that AMA preserves high visual fidelity. LPIPS values lower than 0.15 are indicative of minimal perceptual distortion. Furthermore, BERTScore values above 0.8 and BLEU-4 scores above 0.4 suggest that the semantic content of the text is well-preserved. We evaluated 1000 samples with confidence intervals calculated at a 95% significance level to ensure the robustness of the results.

Table 5: Semantic and perceptual fidelity evaluation metrics on VQAv2 (BLIP model)

Modal	Metric	VLAttack	AMA
Text	BERT-Score (↑)	0.815	0.862
	BLEU-4 (↑)	0.386	0.412
Image	SSIM (↑)	0.890	0.921
	LPIPS (↓)	0.156	0.126

4.4 Ablation Experiment

To validate the effectiveness of each component in AMA, we performed ablation experiments to evaluate the contributions of DAPS, SRO, TDAA, and AFM. The results are illustrated in Figs. 4 and 5.

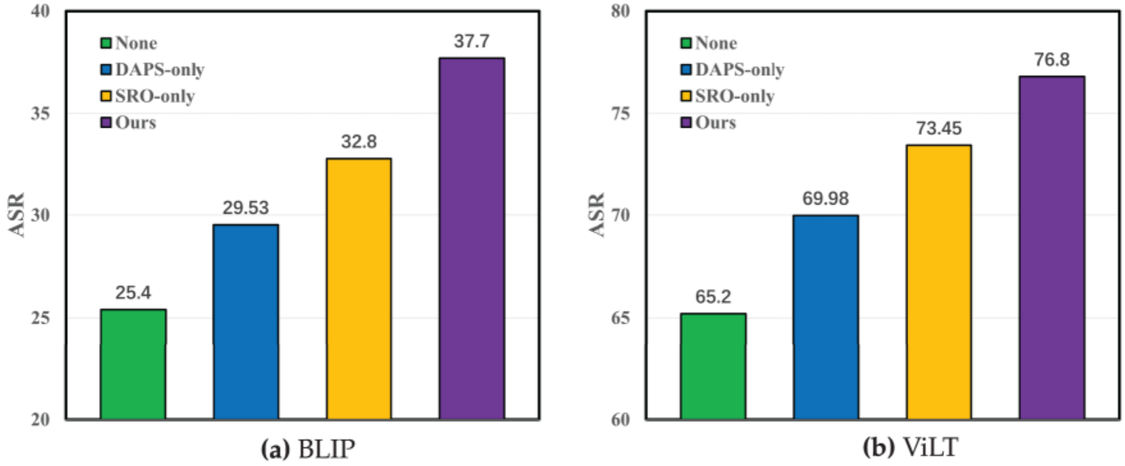

Figure 4: Single-modal attack performance under different settings (1000 samples, $\sigma_{\text{img}} = 0.15$, $\sigma_{\text{txt}} = 0.1$) on BLIP and ViLT models. All results are displayed by ASR (%)

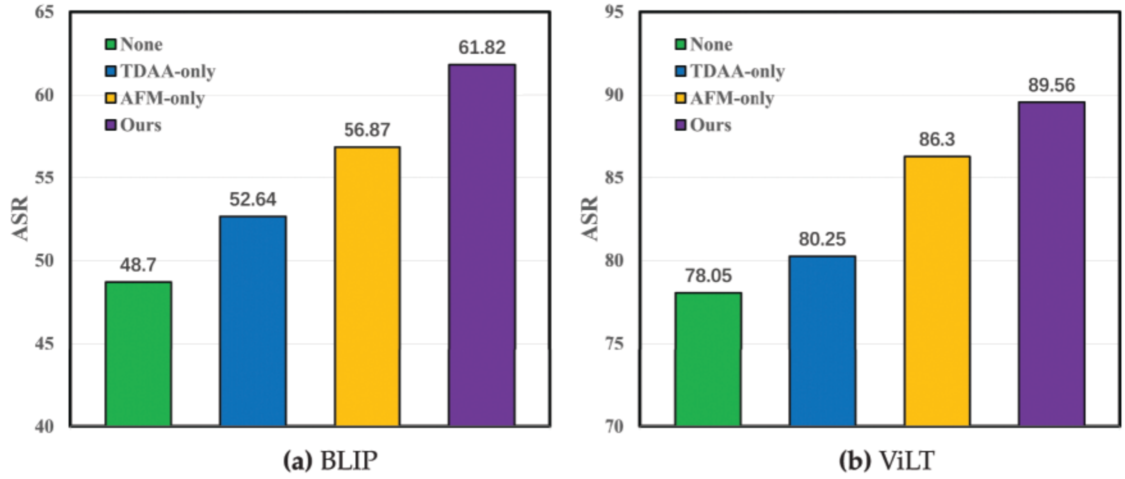

Figure 5: Multimodal attack performance under different settings (1000 samples, $\sigma_{\text{img}} = 0.15$, $\sigma_{\text{txt}} = 0.1$) on BLIP and ViLT models. All results are displayed by ASR (%)

Fig. 4a,b presents the performance of single-modal attacks in different settings on BLIP and ViLT models. The results demonstrate that the incorporation of DAPS and SRO significantly enhances the attack success rate compared to using only basic perturbation strategies. Specifically, DAPS allows for more effective utilization of the feature hierarchy in images and text, while SRO ensures that each perturbation has a maximized impact on the model's prediction. This combination leads to a substantial improvement in the efficiency of attack and the success rate.

Fig. 5a,b shows the performance of multimodal attacks under different settings. The results indicate that TDAA and AFM further enhance the effectiveness of the attack by dynamically adjusting perturbation strategies based on task difficulty and feedback from previous iterations. TDAA ensures that the perturbation intensity is optimized according to the task's complexity, while AFM allows for real-time adjustments based on the model predictions. This makes the attack process more flexible and efficient. These improvements highlight the importance of these components in achieving higher attack success rates across various visual language tasks.

In general, the ablation experiments confirm that each component of AMA plays a crucial role in improving the effectiveness and robustness of the attack. By integrating these strategies, AMA achieves superior performance in both single-modal and multimodal attack scenarios.

4.5 Case Study

To further validate the effectiveness of the AMA method, we conducted a case study that demonstrated the attack effects of AMA in Visual Question Answering (VQA) and Visual Reasoning (VR) tasks. Fig. 6 shows the comparison between normal samples and adversarial samples, as well as the prediction results of the model under these samples. The disturbed part and the original prediction are displayed in red and blue, respectively. We show the predictions after the adversarial attack with an underscore. Through these case studies, we can clearly see the ability of the AMA method in generating adversarial samples, as well as the significant impact of these samples on the prediction results of visual language models.

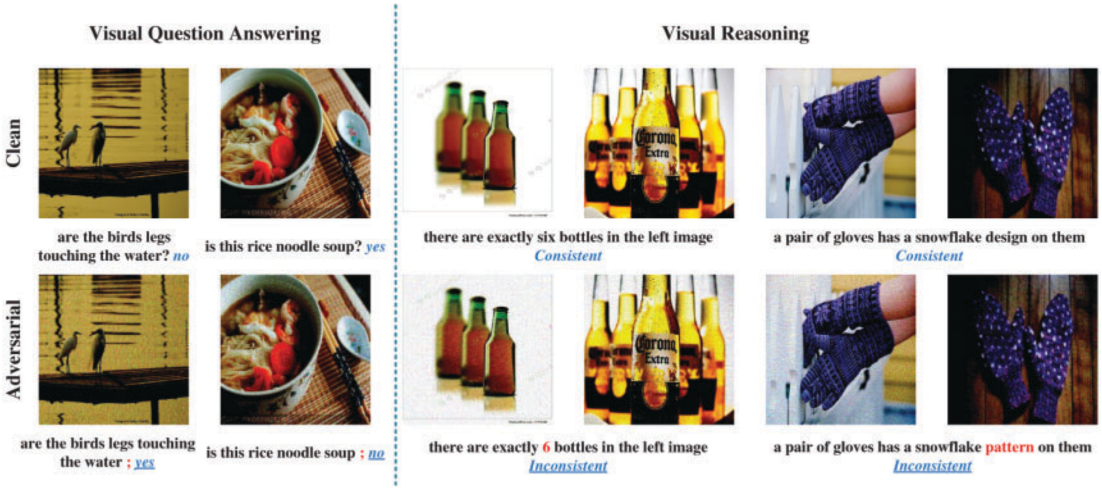
Visual Question Answering
Visual Reasoning
Clean
are the birds legs touching the water? no
is this rice noodle soup? yes
there are exactly six bottles in the left image
Consistent
a pair of gloves has a snowflake design on them
Consistent
Adversarial
are the birds legs touching the water ; yes
is this rice noodle soup ; no
there are exactly 6 bottles in the left image
Inconsistent
a pair of gloves has a snowflake pattern on them
Inconsistent

Figure 6: Successful case study of multimodal attacks on VQA and VR tasks. Perturbed word tokens and original answers are displayed in red and blue, respectively. We show the answers after the adversarial attack with underline

4.6 Discussion

The experimental results reveal that current pre-trained visual language models are vulnerable to adversarial attacks. AMA's effectiveness in generating adversarial samples across various tasks underscores the need for more robust models. Future research should focus on enhancing model robustness against sophisticated attacks like AMA. This is crucial for ensuring the security and reliability of visual-language systems in real-world applications.

While AMA performs well in multiple visual language tasks, we recognize its theoretical limitations. For instance, the feedback loop in AFM might cause overfitting to surrogate models. However, we have mitigated this risk by carefully selecting diverse surrogate models and applying regularization techniques. Our experiments also confirm that AMA generalizes well to unseen models, demonstrating its practical effectiveness despite theoretical concerns.

Moreover, we acknowledge the adversarial attacks present significant ethical challenges, particularly with their potential for malicious misuse. To mitigate these risks, we advocate for the development of effective defense mechanisms, including adversarial input filtering and defense-aware fine-tuning. Adversarial input filtering can detect and block adversarial samples before they impact model predictions, using methods such as input randomization, noise injection, or JPEG compression. Moreover, defense-aware fine-tuning, where models are trained with adversarial examples, can help improve model robustness against such attacks. We also advocate for the establishment of clear ethical guidelines and usage policies to ensure that these methods are employed responsibly. Future research should continue to explore how to strengthen models against complex adversarial attacks while balancing the need for innovation with ethical considerations.

5 Conclusion

This article introduces a novel method, Adaptive Multimodal Attack (AMA), to improve adversarial attack success rates (ASR) in visual language tasks while enhancing the robustness evaluation of visual language models. In the single-modal attack stage, our method dynamically adjusts the perturbation intensity based on the importance of the model layer and feature sensitivity, improving both the attack efficiency and adversarial sample transferability. In multimodal attacks, iterative optimization and adaptive feedback mechanisms further enhance attack effectiveness, particularly in complex tasks. The experimental results show that AMA achieves an ASR of 61.82% on the BLIP model and an impressive 89.56% ASR on the ViLT model for the VQA task, outperforming the baseline methods 13.12% and 11.51%.

AMA not only generates effective adversarial samples for image classification, visual question answering, and image description generation, but it also reveals the vulnerability of current pre-trained visual language models to adversarial attacks. These findings underscore the urgent need for further research to enhance the robustness of these models. The substantial improvements in ASR validate the effectiveness of our method and emphasize its ability to adaptively generate adversarial samples in diverse settings. For future research, enhancing adversarial sample quality to make perturbations more natural and less detectable, developing more sophisticated attack strategies that are adaptable to various model architectures and tasks, and strengthening visual language models' robustness against advanced adversarial attacks are critical areas for exploration. These efforts will ensure the security, reliability, and robustness of visual language models, contributing to the development of more trustworthy visual language systems.

Acknowledgement: The authors thank the Natural Science Foundation of Jiangsu Province and National Natural Science Foundation of China for supporting this study.

Funding Statement: This study is funded by the Natural Science Foundation of Jiangsu Province (Program BK20240699) and National Natural Science Foundation of China (Program 62402228).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Yufei Shi; Methodology, Yufei Shi; Validation, Yufei Shi; Formal analysis, Yufei Shi; Investigation, Yufei Shi; Writing—original draft preparation, Yufei Shi; Writing—review and editing, Yufei Shi, Ziwen He, Teng Jin, Haochen Tong, Zhangjie Fu; Visualization, Yufei Shi; Supervision, Ziwen He, Zhangjie Fu; Project administration, Zhangjie Fu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in VQAv2 at https://visualqa.org/download.html (accessed on 22 July 2025) and NLVR2 at https://lil.nlp.cornell.edu/nlvr/ (accessed on 22 July 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. 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.
2. Yang Y, Zhang X, Xu J, Han W. Empowering vision-language models for reasoning ability through large language models. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2024 Apr 14–19; Seoul, Republic of Korea. p. 10056–60.
3. Ghosh A, Acharya A, Saha S, Jain V, Chadha A. Exploring the frontier of vision-language models: a survey of current methodologies and future directions. arXiv:2404.07214. 2024.
4. Lin Y, Xie Z, Chen T, Cheng X, Wen H. Image privacy protection scheme based on high-quality reconstruction DCT compression and nonlinear dynamics. Expert Syst Appl. 2024;257(5):124891. doi:10.1016/j.eswa.2024.124891.
5. Liao Y, Lin YT, Xing Z, Yuan XC. Privacy image secrecy scheme based on chaos-driven fractal sorting matrix and fibonacci Q-matrix. Vis Comput. 2025;41(9):6931–41. doi:10.1007/s00371-025-04014-4.
6. Kim W, Son B, Kim I. Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning; 2021. p. 5583–94.
7. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning; 2021. p. 8748–63.
8. Li J, Li D, Xiong C, Hoi S. Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning; 2022 Jul 17–23; Baltimore, MD, USA. p. 12888–900.
9. Wang P, Yang A, Men R, Lin J, Bai S, Li Z, et al. Ofa: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: International Conference on Machine Learning; 2022 Jul 17–23; Baltimore, MD, USA. p. 23318–40.
10. Yang Z, Gan Z, Wang J, Hu X, Ahmed F, Liu Z, et al. Unitab: unifying text and box outputs for grounded vision-language modeling. In: European Conference on Computer Vision; 2022 Oct 23–27; Tel Aviv, Israel. p. 521–39.
11. 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; 2022 Oct 10–14; Lisboa, Portugal. p. 5005–13.
12. Zhou Y, Wu J, Wang H, He J. Adversarial robustness through bias variance decomposition: a new perspective for federated learning. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management; 2022 Oct 17–21; Atlanta, GA, USA. p. 2753–62.
13. Xu L, Zhai J. DCVAE-adv: a universal adversarial example generation method for white and black box attacks. Tsinghua Sci Technol. 2023;29(2):430–46. doi:10.26599/tst.2023.9010004.
14. Zhao Y, Pang T, Du C, Yang X, Li C, Cheung NMM, et al. On evaluating adversarial robustness of large vision-language models. Adv Neural Inf Process Syst. 2024;36:54111–38.
15. Han D, Jia X, Bai Y, Gu J, Liu Y, Cao X. Ot-attack: enhancing adversarial transferability of vision-language models via optimal transport optimization. arXiv:2312.04403. 2023.

16. 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 IEEE/CVF International Conference on Computer Vision; 2023 Oct 1–6; Paris, France. p. 102–11.
17. Yin Z, Ye M, Zhang T, Du T, Zhu J, Liu H, et al. Vlattack: multimodal adversarial attacks on vision-language tasks via pre-trained models. Adv Neural Inf Process Syst. 2024;36:52936–56.
18. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv:1412.6572. 2014.
19. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083. 2017.
20. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. arXiv:1312.6199. 2013.
21. Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP); 2017 May 22–26; San Jose, CA, USA. p. 39–57.
22. Moosavi-Dezfooli SM, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 2574–82.
23. Zhao G, Zhang M, Liu J, Wen JR. Unsupervised adversarial attacks on deep feature-based retrieval with GAN. arXiv:1907.05793. 2019.
24. Benaddi H, Jouhari M, Ibrahimi K, Benslimane A, Amhoud EM. Adversarial attacks against iot networks using conditional gan based learning. In: GLOBECOM 2022-2022 IEEE Global Communications Conference; 2022 Dec 4–8; Rio de Janeiro, Brazil: IEEE. p. 2788–93.
25. Huang J, Jiang X, Xia YF. Deceiving traffic sign recognition with physical one-pixel attacks. In: IRC-SET 2022: Proceedings of the 8th IRC Conference on Science, Engineering and Technology. Singapore: Springer; 2023. p. 135–45.
26. Lu Y, Ren H, Chai W, Velipasalar S, Li Y. Time-aware and task-transferable adversarial attack for perception of autonomous vehicles. Patt Recognit Lett. 2024;178(9):145–52. doi:10.1016/j.patrec.2024.01.010.
27. Li J, Ji S, Du T, Li B, Wang T. Textbugger: generating adversarial text against real-world applications. arXiv:1812.05271. 2018.
28. Ebrahimi J, Rao A, Lowd D, Dou D. Hotflip: white-box adversarial examples for text classification. arXiv:1712.06751. 2017.
29. Papernot N, McDaniel P, Swami A, Harang R. Crafting adversarial input sequences for recurrent neural networks. In: MILCOM 2016-2016 IEEE Military Communications Conference; 2016 Nov 1–3; Baltimore, MD, USA. p. 49–54.
30. Gao J, Lanchantin J, Soffa ML, Qi Y. Black-box generation of adversarial text sequences to evade deep learning classifiers. In: 2018 IEEE Security and Privacy Workshops (SPW); 2018 May 24; San Francisco, CA, USA. p. 50–6. doi:10.1109/spw.2018.00016.
31. Jin D, Jin Z, Zhou JT, Szolovits P. Is bert really robust? A strong baseline for natural language attack on text classification and entailment. In: Proceedings of the AAAI conference on artificial intelligence; 2020 Feb 7–12; New York, NY, USA. Vol. 34 p. 8018–25.
32. Yang Y, Huang P, Cao J, Li J, Lin Y, Ma F. A prompt-based approach to adversarial example generation and robustness enhancement. Front Comput Sci. 2024;18(4):184318. doi:10.1007/s11704-023-2639-2.
33. Li Z, Zhao X, Wu DD, Cui J, Shen Z. A frustratingly simple yet highly effective attack baseline: over 90% success rate against the strong black-box models of GPT-4.5/4o/o1. arXiv:2503.10635. 2025.
34. Yang Y, Gao R, Wang X, Ho TY, Xu N, Xu Q. Mma-diffusion: multimodal attack on diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 16–22; Seattle, WA, USA. p. 7737–46.
35. Xu X, Chen X, Liu C, Rohrbach A, Darrell T, Song D. Fooling vision and language models despite localization and attention mechanism. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 4951–61.
36. Li C, Gao S, Deng C, Xie D, Liu W. Cross-modal learning with adversarial samples. Adv Neural Inf Process Syst. 2019;32.

37. Li C, Gao S, Deng C, Liu W, Huang H. Adversarial attack on deep cross-modal hamming retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 11–17; Montreal, BC, Canada. p. 2218–27.
38. Suhr A, Zhou S, Zhang A, Zhang I, Bai H, Artzi Y. A corpus for reasoning about natural language grounded in photographs. arXiv:1811.00491. 2018.
39. Lu Y, Jia Y, Wang J, Li B, Chai W, Carin L, et al. Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–1; Seattle, WA, USA. p. 940–9.
40. Li L, Ma R, Guo Q, Xue X, Qiu X. Bert-attack: adversarial attack against bert using bert. arXiv:2004.09984. 2020.
41. Xu L, Cuesta-Infante A, Berti-Equille L, Veeramachaneni K. R&R: metric-guided adversarial sentence generation. arXiv:2104.08453. 2021.