CoGA: A Collaborative Gray-Box Adversarial Attack for Multimodal Language Models

Tong Wu, Feng Lin, Gaojian Wang, Tiantian Liu, Zhibo Wang, Weizhi Meng, Ajian Liu, and Kui Ren

***Abstract*—Multimodal language models (LMs) have shown significant potential for applications across various domains but remain vulnerable to adversarial attacks. Current research in white-box or black-box settings generally struggles with unrealistic attack assumptions and limited efficacy of targeted attacks. This paper introduces CoGA, a novel gray-box collaborative adversarial attack method for multimodal LMs. Under our gray-box settings, attackers have access only to the victim model's input encoders. With the guidance of different modalities, we perturb the embedding representations from encoders to disrupt the semantic alignment across modalities, ultimately causing inaccurate outputs on various downstream tasks. Specifically, we integrate text embeddings into the loss calculations of the image attack and utilize image embeddings to guide the ranking of vulnerable words and the selection of final samples. Extensive experiments demonstrate that our method achieves superior attack performance across diverse models and tasks, suggesting the shared vulnerability of multimodal LMs in confronting adversarial challenges. Our work provides new insights into the security of multimodal LMs, facilitating the deployment of more robust and secure models in practical applications.**

***Index Terms*—Adversarial Attack, Multimodal, Gray-box Attack, Semantic Alignment, Language Model.**

I. Introduction

THE recent success of multimodal language models (LMs) has attracted broad attention from both academia and industry [1]–[3]. These advanced models process and align heterogeneous data (e.g., text, images, and audio) for cross-modal understanding and generation [4], [5], enabling sophisticated real-world applications such as intelligent assistants [6], [7], autonomous driving [8], [9], and healthcare [10], [11]. However, imperceptible adversarial examples still pose a significant challenge to multimodal LMs.

The collaborative relationship between multiple types of inputs increases the vulnerability of multimodal LMs, and even small adversarial perturbations might cause the model to produce erroneous outputs [12]. Such failures in critical downstream tasks (e.g., visual-language retrieval [13], visual entailment [14], visual grounding [15], and visual question-answering [16]) could trigger severe security incidents that directly threaten the safety of public life and property, ultimately undermining society's trust in the model [17], [18]. For instance, in the visual entailment tasks, models aim to draw inferences based on visual information, such as giving a diagnosis based on medical images. If attackers manipulate the model to produce incorrect inferences, it could result in serious medical malpractice.

Recently, the academic community has initiated a series of studies within this field. The majority of them have concentrated on white-box attack methodologies, where detailed structural information and parameters of the target model are explicitly known [19], [20]. Besides, there has been a growing interest in exploring black-box attack strategies as well [21], [22]. These methods typically leverage a white-box surrogate model to create adversarial examples, optimize the transferability through critical modules or parameters of the model, and subsequently evaluate the attack performance on a black-box model. Generally speaking, the above attacks require complete and precise structural information and parameters of at least one exposed model (e.g., open-source models). However, the rapid progress of multimodal LMs makes this assumption increasingly unavailable [23], [24].

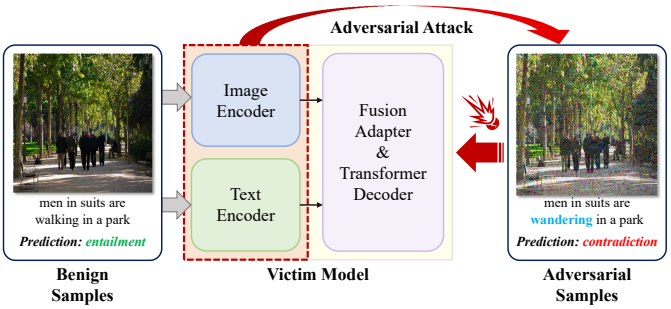

Fig. 1. An example of the adversarial attack for the visual entailment task. CoGA implements adversarial attacks on encoder generated embedding representations. The perturbed image-text pair could alter the inference of the model from *entailment* to *contradiction*. (The perturbation of the image is amplified by 10 times for visualization.)

In this paper, we propose CoGA, a **Co**llaborative **G**ray-box **A**dversarial attack method for Multimodal LMs. Unlike approaches that require full model knowledge, our method generates adversarial samples using only the input encoders of the victim model. An illustrative example of our work is shown in Figure 1. The key idea of the attack is to push the perturbed embeddings away from the original embeddings while destroying the consistency between multiple modalities, thus leading to inaccurate model outputs. Compared with white-box attacks, our approach lowers the barrier for attackers and enhances the generalizability of the attack. This is particularly relevant, as many models do not fully disclose their architectural details but often rely on widely used open-source encoders as foundational components. Different from black-box attacks, our method is more targeted, focusing on specific models and downstream tasks to achieve superior attack performance within the given constraints.

To enhance the efficacy of the attack, we take into account the collaborative effects among various data. As shown in Figure 2, attacks that work on one type of data often do not

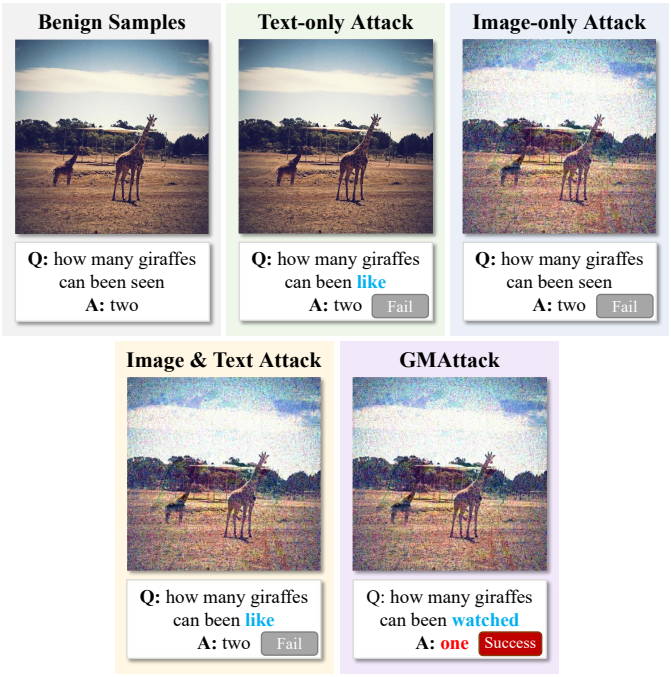

Fig. 2. An example of the adversarial attack for the visual question-answering task. (1) Single-modality adversarial perturbations exhibit limited effectiveness in multimodal settings. (2) Naively combining single-modality attacks does not necessarily improve the attack performance, contradicting the expected "$1 + 1 > 1$" effect. (3) CoGA takes into account the collaborative effects among various data to construct more effective adversarial samples. These findings highlight the necessity of modeling cross-modality interactions for effective attacks.

succeed when facing multimodal settings. Interestingly, the combination of multiple single-modality attacks fails to produce an enhancement in practice, highlighting the particularity of nonlinear interaction between modalities in multimodal adversarial attacks. In this work, the construction of adversarial samples for each modality is guided by data from other modalities, not just a simple combination of single-modality attacks. Specifically, we integrate text embeddings into the loss calculations of the image attack based on projected gradient descent. Besides, we improve the unimodal text attack by utilizing image embeddings to guide the ranking of vulnerable words and the selection of final samples. By exploiting the flexibility of multimodal data, attackers can easily execute information injection in the real world. For example, attackers could spread malicious images or audio files through phishing emails, guiding the victim to input these items into a multimodal LM.

We conduct extensive experiments on four visual-language models, i.e., ALBEF, TCL, CLIP, and BLIP. The results indicate that our method has an effective attack performance on different downstream tasks, including visual-language retrieval, visual entailment, visual grounding, and visual question-answering. Moreover, because we implement adversarial attacks at the embedding representation level, our strategy is not limited by data format or encoder type. By changing the corresponding encoders, our method can be seamlessly applied to other types of models (e.g., audio-language models), demonstrating its generalization.

Our main contributions can be listed as follows:

- We introduce CoGA, a gray-box adversarial attack method for multimodal LMs. In the gray-box scenario, attackers know the structure and parameters of the victim model's encoders. Compared with white-box and black-box attacks, our method reduces attack assumptions and enables better-targeted attacks.
- Our proposed attack solution exploits the inherent vulnerability of multimodal LMs, i.e., the collaborative relationship among different types of data. Instead of a simple combination of single-modality attacks, we break the collaborative consistency between various modalities to achieve stronger adversarial attacks.
- Extensive experiments demonstrate that the proposed method achieves superior attack performance on different downstream tasks. The results validate the effectiveness and advancement of our attack strategy, providing valuable insights for future research on enhancing the robustness of multimodal systems.

II. Related Works

A. Single-modal Adversarial Attack Methods

1) Image Attack: Adversarial attack was initially proposed in the field of computer vision, where the input is subjected to imperceptible perturbations, causing the model to make inaccurate predictions [25]. As the initial work, Fast Gradient Sign Method (FGSM) [26] used the gradient of the cost function with respect to the input to generate perturbed images, which inspired subsequent research to construct adversarial samples based on gradient [27]–[30]. Projected Gradient Descent (PGD) [29] utilized the projected gradient within the feasible region to minimize the loss function over multiple iterations, which is currently the strongest first-order attack. In recent years, substantial research has concentrated on enhancing the transferability of adversarial attacks. For instance, Xie et al. [30] introduced random transformations to input images at each iteration to promote input diversity and enhance feature representation, thereby improving the generalization of the method. However, these existing attacks typically optimize objectives tied to the pre-softmax or softmax layers of the network. Given this shortcoming, Ganeshan et al. [31] proposed Feature Destructive Attacks (FDA) to destroy the high-level semantic information of the clean samples by perturbing feature representations at each layer of the network. Naseer et al. [32] introduced Self-Supervised Perturbation (SSP), which optimized the perceptual feature distance between the clean and perturbed images without requiring label information to construct transferable task-agnostic adversaries. Lu et al. [33] further investigated the adversarial transferability across diverse real-world vision tasks and proposed Dispersion Reduction (DR) to generate adversarial images independent of labeling systems or task-specific loss functions.

2) Text Attack: Unlike images in continuous space, the discrete nature of text makes adversarial attack a difficult problem in the natural language processing (NLP) field. Adversarial attacks on NLP tasks focus on word-level and sentence-level perturbations [34]. Word-level perturbation replaces words

with synonyms that have similar word embeddings and contextual information [35]–[38]. Based on the synonym substitution strategy, Ren et al. [36] introduced a word replacement order determined by both the word saliency and the classification probability, and proposed a greedy algorithm called probability weighted word saliency (PWWS). BERT-Attack [38] used pre-trained masked language models (e.g., BERT) to generate adversarial samples by identifying and replacing vulnerable words with semantically similar and grammatically correct words. In contrast, sentence-level perturbation focuses on the logical structure of the text by paraphrasing or adding unrelated sentence fragments [39]–[41]. Xu et al. [40] presented a rewrite and rollback (R&R) framework for text adversarial attacks. It improves the quality of adversarial examples by optimizing critical scores that combine fluency, similarity, and misclassification metrics. Zou et al. [41] generated adversarial suffixes automatically through a combination of greedy and gradient-based search techniques. These adversarial suffixes can disrupt the alignment mechanism of the model, causing impermissible outputs.

B. Multimodal Adversarial Attack Methods

Yang et al. [42] investigated the robustness of multimodal neural networks against adversarial perturbations on a single modality, showing that standard multimodal fusion models are vulnerable to single-source adversaries. As one of the pioneering works, Fooling-VQA [43] was proposed especially for the visual question-answering task, which iteratively adds pixel-level disturbance on images to implement the attack. Zhang et al. [19] analyzed the performance of vision-language pre-training models (VLPMs) under different attack settings and, for the first time, combined image and text attacks to develop a collaborative multimodal adversarial attack solution (Co-Attack) under the white-box setting. This work proved that cross-mode perturbation might be more effective than a single source. Lu et al. [44] conducted the first systematic evaluation of the adversarial transferability of VLPMs and developed a highly transferable Set-level Guidance Attack (SGA). The primary idea is to enhance the variety of image-text pairs while maintaining image-text alignment. Yin et al. [20] proposed VLATTACK, which generated adversarial samples by fusing perturbations of images and texts from both single-modal and multimodal levels to attack fine-tuned models on different downstream tasks. Recently, Wang et al. [22] proposed the Transferable MultiModal (TMM) attack framework, which can improve the attack capability and transferability from both the modality consistency and modality discrepancy features.

C. Differences from CoGA

In prior work, adversarial attacks on multimodal models have typically assumed either a white-box or a black-box scenario. In a white-box setting, attackers have full access to a model's architecture and parameters, whereas black-box attacks rely on surrogate models to generate adversarial examples. Both paradigms assume complete knowledge of at least one model. In contrast, CoGA introduces a collaborative gray-box attack framework for multimodal LMs. Our method requires access only to the victim model's encoders to generate adversarial samples. Compared with white-box attacks, CoGA reduces reliance on detailed model information while preserving high attack efficacy. This assumption is practical because many real-world models conceal their full architecture but rely on standard open-source encoder modules. Different from black-box attacks, CoGA offers greater precision by directly targeting specific models and downstream tasks, yielding stronger performance under limited knowledge. Furthermore, we further observe that the collaborative relationships among modalities significantly influence attack performance. Instead of treating each modality independently, CoGA leverages cross-modal interactions to exacerbate the semantic misalignment among various modalities. This collaborative gray-box strategy balances feasibility and effectiveness, distinguishing CoGA from existing methods.

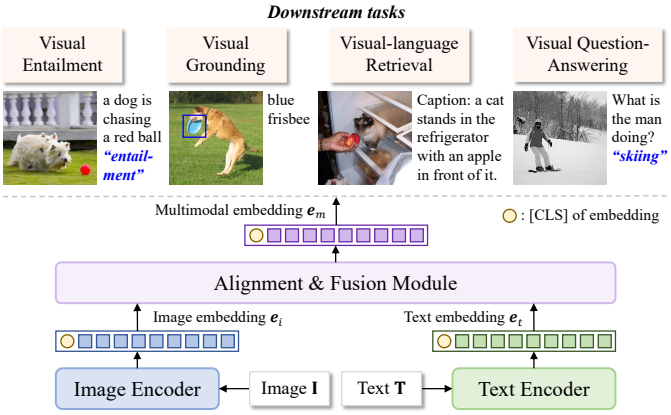

Fig. 3. Architecture overview of the visual-language multimodal LM. It consists of an image encoder and a text encoder, which independently extract visual and textual features. These features are then aligned and fused in a dedicated module to further learn high-level joint semantic information. The resulting embedding representations support diverse downstream tasks through task-specific decoding, including visual entailment, visual grounding, visual-language retrieval, and visual question answering.

III. Method

A. Design Ideas

The goal of our attack is to induce incorrect responses from multimodal LMs across various downstream tasks. Unlike traditional adversarial attack techniques in computer vision, our approach targets tasks that are not standard classification problems and often lack clear class labels. As a result, conventional adversarial methods [26] cannot be directly applied to multimodal LMs. Since these models transform heterogeneous data types into feature vectors for subsequent inference, perturbing these embedding representations presents a promising attack strategy [19], [20], [22], [44], providing adaptability across both tasks and models.

The fundamental architecture of a multimodal LM integrates encoders tailored to the respective input data types [45], [46]. To handle multiple data types, the model typically employs connection modules and adapters that align the output of each encoder with the model's unified input space.

An important insight is that encoders can independently generate embeddings for various modalities, bypassing the

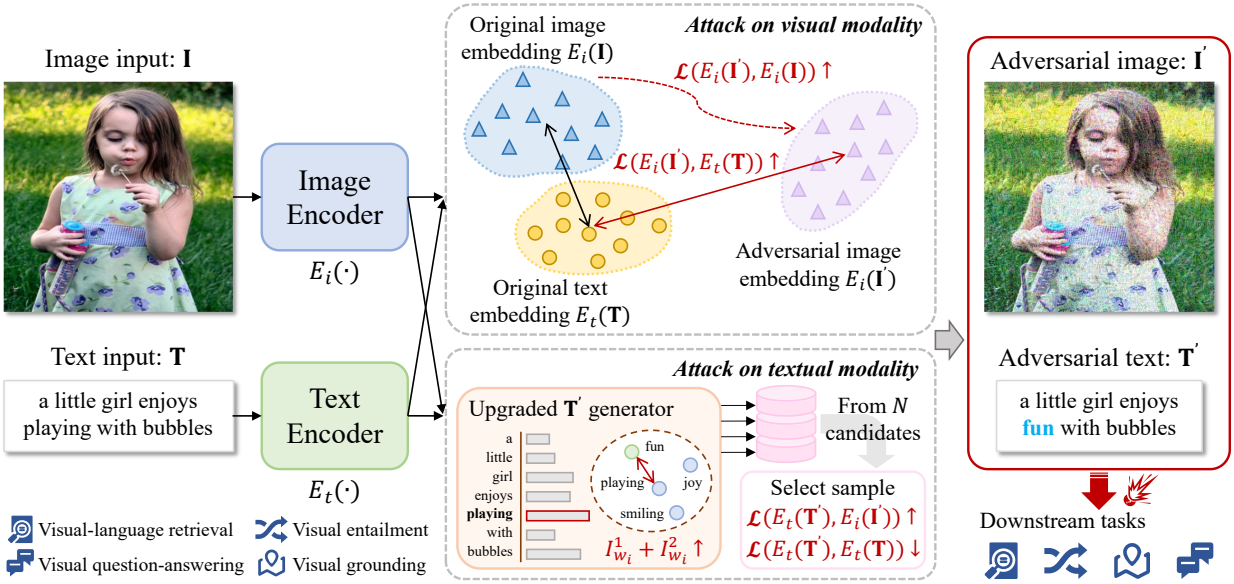

Fig. 4. Overview of CoGA. (1) For the image modality, we perturb the embedding representations from the image encoder, encouraging them to diverge from the text embeddings while remaining distinct from the original image embeddings. (2) For the text modality, we suggest an image-guided strategy based on BERT-Attack, generating several candidate adversarial texts and selecting the final sample according to the discrepancy between the adversarial text and the adversarial image, and the similarity between the adversarial text and the original text. The produced image-text adversarial samples could impact the performance of the models (i.e., ALBEF, TCL, CLIP, and BLIP) on downstream tasks: visual-language retrieval, visual entailment, visual grounding, and visual question-answering.

alignment module or the LM itself, as illustrated in Figure 3. This observation motivates our exploration of a gray-box adversarial attack, where the attacker only needs access to the encoder details of the multimodal LM, without knowledge of the entire model structure or parameters. By perturbing the embedding representations of each modality, we can create adversarial multimodal samples.

Moreover, to effectively attack the embedding representations, it is crucial to consider perturbations across multiple modalities simultaneously. Unlike traditional text-generation models, multimodal LMs are inherently more complex due to the interaction of various data types. The availability of multiple modalities allows attackers to increase the complexity and efficacy of their attacks. However, these modalities also influence each other during training and inference phase, amplifying the model's susceptibility to adversarial perturbations.

Therefore, treating each modality's perturbations independently is insufficient before combining them into adversarial samples. Instead, the interaction between modalities must be strategically considered to maximize the attack's impact. For specific multimodal downstream tasks, semantic consistency is typically maintained across inputs. We argue that effective adversarial attacks target and disrupt this inter-modal consistency, leading to incorrect model outputs. At the embedding level, this disruption is manifested as an increased distance between embeddings in the feature space, which weakens the model's ability to align information across modalities.

Building on these insights, we propose CoGA, a gray-box multimodal collaborative adversarial attack method. Under the gray-box assumption, our approach attempts to break the consistency between modalities, achieving successful adversarial attacks. To illustrate our method, we focus on visual-language models, detailing the construction of adversarial samples for both image and text modalities. The overall attack framework is shown in Figure 4, and the algorithmic flow is presented in Algorithm 1.

B. Threat Model

This work adopts a *gray-box threat model* in which the adversary can access the victim model's input encoders (e.g., CLIP ViT-B/16 for vision, BERT for text) but has no knowledge of its fusion, adapter, or task-specific decoding components. This scenario reflects a widely observed security boundary in contemporary multimodal systems: front-end encoders are exposed while downstream components remain proprietary. Encoder-level access can occur through several realistic avenues: (1) open-source pretrained checkpoints, (2) commercial API services (e.g., `/embed` endpoints) that expose intermediate features, and (3) public documentation sufficient to reconstruct the encoder. These interfaces are commonly encountered in real-world systems, making our threat model both practical and broadly applicable.

As highlighted in recent surveys [1], [2], many state-of-the-art multimodal LMs follow a modular architecture that decouples general-purpose encoders from downstream fusion or decoding layers. For instance, LLaVA [47] and MiniGPT-4 [48] use the CLIP ViT-B/16 encoder as a frozen visual backbone, integrating it with custom alignment and language modules. InstructBLIP [49] similarly freezes the CLIP-based vision encoder while fine-tuning only the subsequent layers. These designs illustrate a prevailing trend: standardized encoders are publicly accessible, whereas full model pipelines remain proprietary. Our threat model aligns with this architecture by assuming transparency at the encoder level without

requiring privileged access to proprietary modules. Operating within these boundaries, our approach strikes a compelling balance between feasibility and adversarial effectiveness in real-world multimodal systems.

C. Preliminaries

In recent years, end-to-end encoders based on transformer architectures have been widely applied in most models, serving as a unified solution for processing heterogeneous inputs.

Given an original text sequence $\mathbf{T} = [t_1, t_2, ..., t_n]$ where t_i denotes the i-th word and n is the total word count, the input is first tokenized into a token sequence $\mathbf{W} = [w_1, w_2, ..., w_{N_\mathrm{T}}] \in \mathcal{V}^{N_\mathrm{T}}$, where $\mathcal{V}$ is a predefined vocabulary. This process may expand the sequence length (i.e., $N_\mathrm{T} \geq n$) due to subword fragmentation (e.g., "unhappy" $\rightarrow$ ["un", "##happy"]). During tokenization, special control tokens such as $[CLS]$ are also appended to $\mathbf{W}$. Each token $w_i \in \mathbf{W}$ is then mapped to a d-dimensional vector $x_i^{(\mathrm{T})}$, forming an embedding matrix:

$$\mathbf{X}_\mathrm{T} = [x_1^{(\mathrm{T})}, x_2^{(\mathrm{T})}, ..., x_{N_\mathrm{T}}^{(\mathrm{T})}] \in \mathbb{R}^{N_\mathrm{T} \times d}, \tag{1}$$

where $x_i^{(\mathrm{T})} \in \mathbb{R}^d$ and d is the model's hidden dimension.

Given an original image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$, where H, W, and C respectively denote the height, width, and channel dimensions, the input is first divided into $m = \lfloor H/q \rfloor \times \lfloor W/q \rfloor$ non-overlapping patches of size $q \times q$ pixels. Similar to the preprocessing paradigm of textual modality, special control tokens are prepended to the patch sequence, resulting in an augmented sequence $\mathbf{P} = [p_1, p_2, ..., p_{N_\mathrm{I}}] \in \mathbb{R}^{N_\mathrm{I} \times q^2 \times C}$. Each patch $p_j \in \mathbf{P}$ is then linearly projected into a d-dimensional space $x_j^{(\mathrm{I})}$, producing an embedding matrix:

$$\mathbf{X}_\mathrm{I} = [x_1^{(\mathrm{I})}, x_2^{(\mathrm{I})}, ..., x_{N_\mathrm{I}}^{(\mathrm{I})}] \in \mathbb{R}^{N_\mathrm{I} \times d}. \tag{2}$$

The matrixes from both modalities are combined with positional encodings $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{k \times d}$, where k denotes the sequence length (i.e., $k = N_\mathrm{T}$ for text or $k = N_\mathrm{I}$ for image) to form the initial hidden states:

$$\mathbf{H}^{(0)} = \mathbf{X} + \mathbf{E}_{\text{pos}}, \tag{3}$$

The transformer-based encoder then processes $\mathbf{H}^{(0)}$ through L layers, each applying multi-head self-attention (MHSA) and feed-forward networks (FFN):

$$\begin{aligned} \mathbf{A}^{(l)} &= \text{MHSA}(\text{LayerNorm}(\mathbf{H}^{(l-1)})), \\ \mathbf{Z}^{(l)} &= \mathbf{H}^{(l-1)} + \mathbf{A}^{(l)}, \\ \mathbf{F}^{(l)} &= \text{FFN}(\text{LayerNorm}(\mathbf{Z}^{(l)})), \\ \mathbf{H}^{(l)} &= \mathbf{Z}^{(l)} + \mathbf{F}^{(l)}, \end{aligned} \tag{4}$$

where $\mathbf{H}^{(l)}$ is the output of the l-th layer with $l \in [1, 2, ..., L]$. The final $[CLS]$ embedding is extracted as the global semantic representation:

$$e_i = \text{AvgPool}(\mathbf{H}_\mathrm{I}^{(L)}[:, [CLS]]), \tag{5}$$

$$e_t = \text{AvgPool}(\mathbf{H}_\mathrm{T}^{(L)}[:, [CLS]]), \tag{6}$$

where e_i and e_t respectively represent the target image embedding and text embedding.

Algorithm 1 CoGA

Input: A clean image-text pair $(\mathbf{I}, \mathbf{T})$
Parameter: The image encoder of the victim model $E_i(\cdot)$, the text encoder of the victim model $E_t(\cdot)$, perturbation budget ϵ_i on $\mathbf{I}$, the number of adversarial image iteration M, the number of candidate adversarial texts N
Output: The adversarial image-text pair $(\mathbf{I}^{'}, \mathbf{T}^{'})$

```
Initialize 𝒯 = ∅, ε_t = 0.
// Calculate I' using Equation 8 and 9 for M steps.
I' = ImageAttack(I, T, E_i(·), E_t(·), ε_i, M)
// Generate adversarial text library 𝒯.
while |𝒯| < N do
    ε_t = ε_t + 1
    𝒮 = TextAttack(T, I', E_i(·), E_t(·), ε_t)
    𝒯 = 𝒯 + 𝒮
end while
// Select the adversarial text from 𝒯.
Sort 𝒯 using Equation 13 in descending order to get list {T̂_1, ..., T̂_N}.
T' = T̂_1
Return (I', T').
```

In this work, we mainly focus on the encoder-generated embedding representations rather than the internal mechanisms. For clarity, we formalize the above process as: $E_i(\cdot)$ denotes the image encoder of the victim model, an input image $\mathbf{I}$ is encoded into the image embedding e_i by the image encoder $E_i(\cdot)$, i.e., $e_i = E_i(\mathbf{I})$. $E_t(\cdot)$ denotes the text encoder of the victim model, an input text $\mathbf{T}$ is encoded into the text embedding e_t by the text encoder $E_t(\cdot)$, i.e., $e_t = E_t(\mathbf{T})$. Also, our attack strategy can be formulated as a function $G(\cdot, \cdot)$, which injects adversarial perturbations into the embeddings based on multimodal collaborative guidance and generates adversarial image-text pair $(\mathbf{I}', \mathbf{T}') = G(e_i, e_t)$.

The key idea of our attack is to systematically disrupt cross-modal alignment through embedding-space perturbations. The theoretical foundation lies in minimizing the mutual information (MI) between visual-textual embeddings, thereby artificially severing statistical dependencies across modalities. Prior research has established that mutual information can be approximated using the Kullback-Leibler (KL) divergence:

$$\begin{aligned} I(e_i, e_t) &= \iint p(e_i, e_t) \log \frac{p(e_i, e_t)}{p(e_i)p(e_t)} de_i de_t \\ &= D_{\text{KL}}(p(e_i, e_t) || p(e_i)p(e_t)), \end{aligned} \tag{7}$$

where $D_{\text{KL}}(\cdot||\cdot)$ represents the KL divergence which quantifies the discrepancy between the joint distribution $p(e_i, e_t)$ and the product of marginal distributions $p(e_i)p(e_t)$. Directly estimating $p(e_i, e_t)$ is intractable due to high dimensionality; thus, we choose to optimize the surrogate KL divergence loss $\mathcal{L}(\cdot, \cdot)$ between different embeddings, i.e., $\mathcal{L}(e_i, e_t)$, to achieve the attack objective.

D. Attacks on Image Modality

For the image modality, we perturb the embedding representations from the image encoder, encouraging them to

Algorithm 2 TextAttack

Input: Benign text $\mathbf{T} = [t_1, t_2, ..., t_n]$, adversarial image $\mathbf{I}^{'}$
Parameter: The image encoder of the victim model $E_i(\cdot)$, the text encoder of the victim model $E_t(\cdot)$, perturbation degree ϵ_t on $\mathbf{T}$
Output: The adversarial text set $\mathcal{S}$

Initialize $\mathcal{S} = \emptyset$.
Tokenize $\mathbf{T}$ using BERT tokenizer to get $\mathbf{W} = [w_1, w_2, ..., w_{N_\mathrm{T}}]$.
// Rank words in order of importance.
for w_i in $\mathbf{W}$ **do**
    Calculate importance scores $I^1_{w_i}$ and $I^2_{w_i}$ using Equation 11 and 12.
    $I_{w_i} = I^1_{w_i} + I^2_{w_i}$
end for
Sort $\mathbf{W}$ using I_{w_i} in descending order to get list $\mathbf{L} = [u_1, u_2, ..., u_{N_\mathrm{T}}]$.
Generate top-K candidates for $u_i \in \mathbf{L}$ using BERT and get $\mathcal{P}^{\in N_\mathrm{T} \times K} =$

$$\begin{bmatrix} c_1^{u_1} & c_2^{u_1} & \cdots & c_K^{u_1} \\ c_1^{u_2} & c_2^{u_2} & \cdots & c_K^{u_2} \\ \vdots & \vdots & \ddots & \vdots \\ c_1^{u_{N_\mathrm{T}}} & c_2^{u_{N_\mathrm{T}}} & \cdots & c_K^{u_{N_\mathrm{T}}} \end{bmatrix}.$$

// Generate adversarial texts based on token substitution.
$\mathbf{H} = [h_1, h_2, ..., h_n] \leftarrow \mathbf{T} = [t_1, t_2, ..., t_n]$
for u_j in $\mathbf{L}$ **do**
    Get candidates $\mathcal{C}$ = Filter($\mathcal{P}^j$).
    for c_k in $\mathcal{C}$ **do**
        $\mathbf{H}' = [h_1, .., h_{j-1}, c_k, h_{j+1}, ..., h_n]$
        if $|\mathbf{H}' - \mathbf{T}| == \epsilon_t$ **then**
            $\mathcal{S} = \mathcal{S} + \mathbf{H}'$
        else
            $\mathbf{H} = \arg\max_{\mathbf{H}'}(||E_t(\mathbf{H}') - E_t(\mathbf{T})||)$
        end if
    end for
end for
Return $\mathcal{S}$.

diverge from the text embeddings while remaining distinct from the original image embeddings. Note that image and text embeddings can be viewed as independent objects in the feature space that are close together, rather than a unified representation after further fusion.

To generate adversarial image perturbations δ_i corresponding to input image $\mathbf{I}$, we apply l_∞-norm constrained perturbation, where δ_i satisfies $||\delta_i||_\infty \leq \epsilon_i$ with ϵ_i being the maximum perturbation magnitude. Since most visual-language downstream tasks are non-classified, we adopt the method that maximizes the KL divergence loss $\mathcal{L}$ of embedding-wise representation [50]. Thus, the image adversarial perturbations δ_i are given by:

$$\delta_i = \epsilon_i \cdot \mathrm{sign}(\nabla_{\mathbf{I}'} \mathcal{L}(E_i(\mathbf{I}^{'}), E_i(\mathbf{I})). \tag{8}$$

In this work, we employ PGD [29] for gradient-based adversarial attacks.

We perform the adversarial attack on image modality as follows:

$$\max\ \mathcal{L}(E_i(\mathbf{I}^{'}), E_i(\mathbf{I})) + \alpha \cdot \mathcal{L}(E_i(\mathbf{I}^{'}), E_t(\mathbf{T})), \tag{9}$$

where α is a hyper-parameter that controls the contributions of the second term. The adversarial attack on the image modality needs to maximize the difference between the adversarial image embedding $E_i(\mathbf{I}^{'})$ and the original image embedding $E_i(\mathbf{I})$, while also increasing the distance between $E_i(\mathbf{I}^{'})$ and the original text embedding $E_t(\mathbf{T})$, thereby inducing inconsistency in the multimodal space.

E. Attacks on Text Modality

Adversarial attacks in discrete space pose a challenge in the field of NLP. Two common strategies are adding crafted substrings to the text or modifying specific tokens according to semantics and grammar. The former may introduce unreadable characters into the sentence, limiting the legibility of the text and, to some extent, reducing the practicality of the attack (e.g., security-conscious users may notice the anomalies of the input). This drawback motivates us to follow the latter strategy to develop a novel text attack method.

For the text modality, adversarial examples $\mathbf{T}^{'}$ is generated via strategic token manipulations of the input text $\mathbf{T}$. Then the text adversarial perturbations δ_t are given by:

$$\delta_t = \arg\max_{\mathbf{T}'}(||E_t(\mathbf{T}^{'}) - E_t(\mathbf{T})||) - \mathbf{T}, \tag{10}$$

where the maximum perturbation ϵ_t is constrained to the token level, i.e., how many tokens are modified or replaced in the original text. In this work, we improve BERT-Attack [38] for text adversarial attacks.

The basic BERT-Attack method does not explicitly consider the effect of image modality on the text modality; hence, we suggest an image-guided, upgraded BERT-Attack-based strategy for constructing adversarial texts. Our method introduces two key improvements: First, we use image embeddings to evaluate the importance of vulnerable words. Second, for each original text, we generate N candidate adversarial texts and select the final sample according to the difference between modalities, where N is a hyper-parameter controlling the size of the adversarial text library.

Specifically, we propose $I^1_{w_i}$ and $I^2_{w_i}$ as the importance score of the words,

$$I^1_{w_i} = \mathcal{L}(E_t(\mathbf{T}_{\backslash w_i}), E_t(\mathbf{T})), \tag{11}$$

$$I^2_{w_i} = \mathcal{L}(E_t(\mathbf{T}_{\backslash w_i}), E_i(\mathbf{I}^{'})), \tag{12}$$

where $\mathbf{T}_{\backslash w_i} = [w_0, ..., w_{i-1}, [MASK], w_{i+1}..., w_n]$ is the sentence after replacing w_i with $[MASK]$. When ranking the vulnerability of each word, we retain BERT's strengths in assessing syntax and semantics. Additionally, we incorporate the image modality as a reference, selecting the key words based on the distance in feature space between the text and image embeddings after masking a token. For a particular perturbation degree ϵ_t, the adversarial text generation algorithm is shown in Algorithm 2.

TABLE I
THE ASR(%) RESULTS OF OUR PROPOSED METHOD AND COMPARED BASELINES FOR VE TASK ON SNLI-VE DATASET, THE BOLD NUMBERS INDICATE BETTER RESULTS.

Model	Attack					
	B&A	PGD	DR	SSP	Co-Attack	Ours
ALBEF	13.34	15.59	6.36	18.71	22.56	**36.51**
TCL	14.74	9.04	3.38	12.18	21.82	**26.47**

TABLE II
THE ASR(%) RESULTS OF OUR PROPOSED METHOD AND COMPARED BASELINES FOR VQA TASK ON VQAV2 DATASET, THE BOLD NUMBERS INDICATE BETTER RESULTS.

Model	Attack					
	B&A	PGD	DR	SSP	Co-Attack	Ours
ALBEF	60.26	41.30	20.42	49.68	46.50	**66.86**
BLIP	21.04	15.89	7.04	11.84	14.24	**44.10**

TABLE III
THE ASR(%) RESULTS OF OUR PROPOSED METHOD AND COMPARED BASELINES FOR VG TASK ON REFCOCO+ DATASET, THE BOLD NUMBERS INDICATE BETTER RESULTS. THE TESTING SET CONTAINS TWO SUBSETS OF PEOPLE (TESTA) AND ALL OTHER OBJECTS (TESTB).

Model	Attack	Val	TestA	TestB
ALBEF	B&A	10.42	14.17	6.51
	PGD	11.70	12.28	9.76
	DR	7.76	10.15	3.69
	SSP	12.42	16.17	6.88
	Co-Attack	15.06	17.85	10.46
	Ours	**26.26**	**30.98**	**19.29**

The overall attack strategy is as follows: Set the perturbation degree to 1 (i.e., $\epsilon_t = 1$) and generate adversarial texts using Algorithm 2. Add the generated samples to the adversarial text library, then increase the perturbation degree by 1 and repeat the above process until the library contains N adversarial texts. Finally, select the sample with the highest adversarial score from the library:

$$\begin{aligned} score = \beta \cdot \mathcal{L}(E_t(\mathbf{T}^{'}), E_i(\mathbf{I}^{'})) \\ - (1-\beta) \cdot \mathcal{L}(E_t(\mathbf{T}^{'}), E_t(\mathbf{T})), \end{aligned} \tag{13}$$

where β is a hyper-parameter that controls the contributions of the two parts. The second term in Equation 13 is preceded by a negative sign to maximize the difference between the adversarial text embedding $E_t(\mathbf{T}^{'})$ and the adversarial image embedding $E_i(\mathbf{I}^{'})$, while preserving the semantic information of the input text, i.e., minimizing the difference between $E_t(\mathbf{T}^{'})$ and the original text embedding $E_t(\mathbf{T})$.

IV. EXPERIMENTS

A. Experiment Setups

1) Downstream tasks: Following most of the previous works in related fields [19], [20], [22], [43], we consider four general visual-language downstream tasks:

- **Visual-language Retrieval (VR)**: This task retrieves the image/text that is most relevant to a given text/image from a repository. It contains two sub-tasks: image-to-text retrieval (TR) and text-to-image retrieval (IR).
- **Visual Entailment (VE)**: This task requires visual reasoning to determine whether the relationship between an image and a text is entailment, neutral, or contradiction.
- **Visual Grounding (VG)**: This task localizes the region in an input image based on the description of the corresponding input text.
- **Visual Question-Answering (VQA)**: This task entails an understanding of vision, language and commonsense knowledge to answer questions based on the given image.

2) Datasets: For the datasets, Flickr30K [51] and MSCOCO [52] are used to evaluate VR task, SNLI-VE [14] is used to evaluate VE task, RefCOCO+ [15] is used to evaluate VG task, and VQAv2 [16] is used to evaluate VQA task. These datasets are selected based on common practices in related research, ensuring comprehensive coverage of benchmark datasets for each downstream task.

3) Victim models: We choose ALBEF [4], TCL [53], CLIP [5] and BLIP [54] to evaluate the adversarial attack performance on different downstream tasks. Note that CLIP can only handle VR task, BLIP can handle VQA task, TCL can handle VE and VR tasks, and ALBEF can deal with all the above tasks.

The image encoder of the above models is implemented by the 12-layer visual transformer ViT-B/16. The text encoders of ALBEF and TCL are implemented by a 6-layer transformer, while the text encoders of CLIP and BLIP are implemented by a 12-layer transformer.

4) Evaluation metrics: We employ the attack success rate (ASR) to evaluate the efficacy of the generated adversarial examples for downstream tasks, which is the percentage of adversarial examples that successfully influence the model's decisions. Higher ASR signifies better attacking ability. In particular, we provide the ASR values of R@1, R@5, and R@10 for the image-to-text (TR) and text-to-image retrieval (IR) tasks, where R@N represents the probability of including relevant results within the top N retrieved items.

5) Baselines: We select five representative relevant studies for comparison, including text, image, and multimodal adversarial attack methods. For attacks on the text modality, we take BERT-Attack (B&A) [38] as the baseline. For attacks on the image modality, we take PGD [29], DR [33], and SSP [32] as baselines. We also compare `CoGA` with Co-Attack [19], which pioneers exploring multimodal adversarial attacks for visual-language models.

6) Implementation details: For the adversarial attack on image modality, the maximum perturbation ϵ_i is set to 8/255, the step size of PGD is set to 1.25, the number of iterations M is set to 10, and the hyper-parameter α in Equation 9 is set to 5. For the adversarial attack on text modality, the number of the candidate adversarial texts N is set to 10, and the hyper-parameter β in Equation 13 is set to 0.7. Our experiments are

TABLE IV
THE ASR(%) RESULTS OF OUR PROPOSED METHOD AND COMPARED BASELINES FOR VR TASK ON FLICKR30K DATASET, THE BOLD NUMBERS INDICATE BETTER RESULTS.

Model	Attack	TR@1	TR@5	TR@10	TRmean	IR@1	IR@5	IR@10	IRmean	Mean
ALBEF	B&A	5.00	0.90	0.40	2.10	12.84	8.30	6.04	9.06	5.58
	PGD	59.80	49.20	45.20	51.40	58.14	56.52	52.32	55.66	53.53
	DR	24.80	11.70	7.80	14.77	24.08	13.78	9.72	15.86	15.31
	SSP	43.50	28.70	22.40	31.53	37.76	27.34	21.28	28.79	30.16
	Co-Attack	64.00	51.20	43.70	52.97	63.08	58.94	54.30	58.73	55.87
	Ours	**92.60**	**92.80**	**91.20**	**92.20**	**81.96**	**91.42**	**91.18**	**88.19**	**90.19**
TCL	B&A	5.10	0.50	0.40	2.00	14.04	9.46	6.82	10.11	6.05
	PGD	75.30	68.10	61.70	68.37	66.44	66.74	61.80	64.99	66.68
	DR	24.80	11.70	8.60	15.03	24.98	15.16	10.70	16.95	15.99
	SSP	39.10	23.10	17.90	26.70	31.52	20.74	15.06	22.44	24.57
	Co-Attack	80.20	73.20	68.00	73.80	73.02	74.08	69.22	72.11	72.95
	Ours	**94.00**	**97.50**	**96.70**	**96.07**	**83.08**	**94.36**	**94.80**	**90.75**	**93.41**
CLIP	B&A	12.90	8.10	4.20	8.40	17.32	15.24	12.36	14.97	11.69
	PGD	55.60	49.00	42.50	49.03	43.36	47.78	43.42	44.85	46.94
	DR	10.80	6.70	4.00	7.17	9.14	6.94	5.62	7.23	7.20
	SSP	14.50	9.00	5.60	9.70	11.26	8.76	7.10	9.04	9.37
	Co-Attack	74.50	79.30	75.80	76.53	58.10	75.90	77.52	70.51	73.52
	Ours	**81.10**	**95.40**	**96.70**	**91.07**	**61.94**	**85.12**	**90.88**	**79.31**	**85.19**

TABLE V
THE ASR(%) RESULTS OF OUR PROPOSED METHOD AND COMPARED BASELINES FOR VR TASK ON MSCOCO DATASET, THE BOLD NUMBERS INDICATE BETTER RESULTS.

Model	Attack	TR@1	TR@5	TR@10	TRmean	IR@1	IR@5	IR@10	IRmean	Mean
ALBEF	B&A	13.24	7.74	4.18	8.39	17.30	15.93	13.25	15.49	11.94
	PGD	58.94	62.64	59.68	60.42	46.82	58.50	59.22	54.85	57.63
	DR	30.14	21.98	16.38	22.83	22.12	19.63	15.86	19.20	21.02
	SSP	46.06	37.72	30.18	37.99	27.82	27.18	22.73	25.91	31.95
	Co-Attack	63.90	66.60	62.42	64.31	51.00	63.47	63.19	59.22	61.76
	Ours	**76.58**	**91.66**	**93.62**	**87.29**	**59.38**	**81.57**	**86.74**	**75.88**	**81.58**
TCL	B&A	14.08	7.22	4.80	8.70	17.64	15.94	13.06	15.55	12.12
	PGD	59.72	62.98	59.62	60.77	46.42	56.95	56.74	53.37	57.07
	DR	27.98	20.38	15.06	21.14	20.76	18.00	14.38	17.71	19.43
	SSP	37.24	28.76	22.56	29.52	23.29	21.08	17.29	20.55	25.04
	Co-Attack	67.12	73.76	71.44	70.77	52.80	67.80	68.87	63.16	66.97
	Ours	**74.28**	**88.58**	**90.16**	**84.34**	**57.37**	**78.83**	**83.53**	**73.24**	**78.79**
CLIP	B&A	20.56	20.34	17.44	19.45	14.65	20.34	20.73	18.57	19.01
	PGD	41.72	52.48	53.52	49.24	26.22	40.41	44.23	36.95	43.10
	DR	13.72	12.96	10.78	12.49	6.89	8.56	7.90	7.78	10.14
	SSP	16.84	15.82	12.72	15.13	8.06	10.14	9.76	9.32	12.22
	Co-Attack	50.92	72.16	77.36	66.81	32.09	55.13	64.22	50.48	58.65
	Ours	**52.38**	**76.16**	**83.94**	**70.83**	**32.97**	**57.93**	**68.43**	**53.11**	**61.97**

conducted on a single NVIDIA RTX A6000 GPU card with 48GB memory.

B. Result Analysis

1) Overall performance: Table I–V present the attack results across multiple downstream tasks. Our method consistently achieves superior attack performance, demonstrating its ability to effectively disrupt multimodal reasoning under gray-box conditions. Compared with single-modality baselines, multimodal attacks such as CoGA and Co-Attack yield significantly stronger results. This highlights the advantage of jointly exploiting visual and textual features during perturbation rather than treating each modality in isolation.

The results also reveal clear differences in each task's sensitivity to adversarial perturbations. Text-only attacks are particularly effective on the VQA task, where obtaining the correct answer depends heavily on a precise understanding of the question. In such cases, even minor textual perturbations can significantly mislead the model. In contrast, tasks like visual entailment and cross-modal retrieval demand balanced reasoning across both modalities. Their robustness to single-modality attacks suggests that perturbing only one input modality is often insufficient. These observations reflect our deeper insight: the effectiveness of adversarial perturbations is closely tied to a task's modality dependency. Therefore, attack strategies should be adapted accordingly. For example, textual noise should be prioritized for language-intensive tasks, whereas visual signals should be the focus for perception-driven tasks.

Figure 5 visualizes the embedding space before and after the attack. The increased distances among and within modalities confirm that our method effectively disrupts the original

TABLE VI
IMPERCEPTIBILITY AND ALIGNMENT METRICS OF CLEAN VS. ADVERSARIAL SAMPLES. REPORTED METRICS INCLUDE PSNR, SSIM, AND EMBEDDING COSINE SIMILARITY FOR IMAGES; BERT-SCORE AND EMBEDDING COSINE SIMILARITY FOR TEXTS; AND CROSS-MODAL IMAGE-TEXT EMBEDDING COSINE SIMILARITY BEFORE AND AFTER ATTACK.

Model	Task	CosSim(I)	PSNR(I)	SSIM(I)	CosSim(T)	BERTScore(T)	CosSim(Ori.)	CosSim(Adv.)
ALBEF	VE	0.9990 ±0.0010	34.1205 ±0.1983	0.9178 ±0.0609	0.8746 ±0.1269	0.8230 ±0.1601	0.0283 ±0.0268	-0.0147 ±0.0253
	VG	0.9993 ±0.0006	34.1780 ±0.2137	0.9271 ±0.0319	0.8420 ±0.1271	0.7564 ±0.1500	0.0186 ±0.0273	-0.0210 ±0.0311
CLIP	VR(F)	0.9989 ±0.0030	34.2692 ±0.1990	0.9398 ±0.0345	0.9526 ±0.0655	0.9439 ±0.0652	0.3169 ±0.0346	-0.0409 ±0.0780
	VR(M)	0.9994 ±0.0007	34.2978 ±0.1611	0.9376 ±0.0290	0.9383 ±0.0709	0.9270 ±0.0681	0.3075 ±0.0321	-0.0146 ±0.0744
BLIP	VQA	0.9993 ±0.0013	34.1577 ±0.2406	0.9123 ±0.0359	0.9100 ±0.0628	0.8574 ±0.0897	0.0585 ±0.0325	-0.0368 ±0.0319

TABLE VII
THE ASR(%) RESULTS OF ADVERSARIAL EXAMPLES GENERATED ON CLIP WITH VIT-B/16 ENCODER AND TRANSFERRED TO CLIP VARIANTS WITH DIFFERENT ENCODERS (VIT-B/32, VIT-L/14, RESNET-50, AND RESNET-101) FOR VR TASK.

Encoder	Flickr30K							MSCOCO						
	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10	Mean	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10	Mean
ViT-B/32	12.30	5.90	4.30	13.42	11.30	9.14	9.39	13.66	14.18	11.78	9.47	13.91	13.92	12.82
ViT-L/14	11.00	4.80	2.60	14.68	11.74	9.74	8.44	15.10	13.64	11.96	12.52	15.71	15.60	14.09
RN50	15.80	8.70	4.80	16.92	15.70	13.92	12.64	18.62	18.78	17.02	10.78	16.15	17.08	16.41
RN101	16.00	8.40	5.30	16.66	15.24	12.12	12.29	17.12	17.38	15.84	12.61	17.69	17.92	16.43

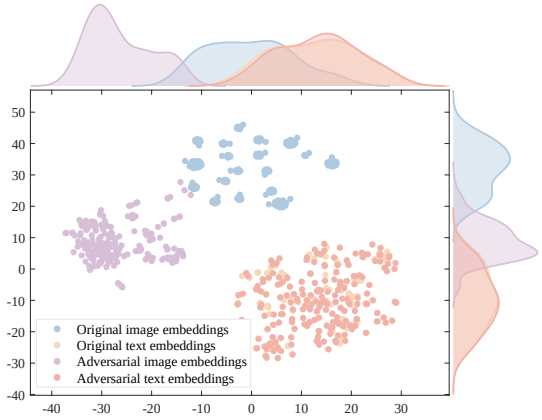

Fig. 5. The visualized t-SNE results of original and adversarial image/text embeddings.

alignment. For image embeddings, the original samples form compact clusters corresponding to natural semantic groupings. The adversarial image embeddings shift substantially in feature space, no longer forming the well-separated clusters seen before. By contrast, text embeddings shift more subtly because our method preserves semantic meaning to ensure imperceptibility. Even so, we observe a clear directional drift, with adversarial text embeddings moving away from both the original text embeddings and the perturbed image embeddings. These observations collectively underscores the asymmetric vulnerability across modalities and confirms that our method effectively disrupts cross-modal alignment without introducing perceptible semantic distortion.

2) Imperceptibility evaluation: Table VI presents quantitative results evaluating the imperceptibility of our attack. For the image modality, we report the following metrics between clean and adversarial samples:

- **CosSim(I)**: cosine similarity of image embeddings, measuring perturbation magnitude in the representation space;
- **PSNR(I)**: peak signal-to-noise ratio (in dB), quantifying pixel-level distortion; values above 30 dB indicate that the adversarial image is visually indistinguishable from the original;
- **SSIM(I)**: structural similarity index, assessing the preservation of structural information.

For the text modality, we evaluate:

- **CosSim(T)**: cosine similarity of text embeddings, reflecting semantic drift in the representation space;
- **BERTScore(T)**: a contextual semantic similarity metric that captures semantic preservation at the sentence level.

The results show that both modalities retain high fidelity after perturbation, confirming the strong imperceptibility of our attack and its practical viability in real-world scenarios.

Furthermore, we measure cross-modality alignment using cosine similarity between image and text embeddings: **CosSim(Ori.)** for the original image-text pair and **CosSim(Adv.)** for the adversarial image-text pair. As expected, the alignment score drops to a negative value after the attack, validating our core design principle: *disrupting inter-modal consistency while preserving intra-modal imperceptibility.*

3) Sensitivity to encoder architectures: To evaluate the sensitivity of CoGA to encoder architectures, we conduct a cross-encoder transfer experiment for VR task. Specifically, adversarial image-text pairs are generated using CLIP with a ViT-B/16 image encoder, and then evaluated on four CLIP variants employing different encoders: ViT-B/32, ViT-L/14, ResNet-50, and ResNet-101. As shown in Table VII, the ASR decreases when transferring to dissimilar encoders. This behavior is expected under the gray-box setting: our attack optimizes perturbations in the embedding space of a specific encoder, whose geometric properties (e.g., feature dimensionality, normalization, and semantic abstraction) differ significantly across architectures (e.g., transformer-based ViTs vs. convolutional ResNets). Consequently, perturbations tailored to one encoder may not align with the vulnerability directions of another.

TABLE VIII
THE ASR(%) RESULTS OF ADVERSARIAL EXAMPLES GENERATED ON ALBEF AND TRANSFERRED TO OTHER VISION-LANGUAGE MODELS (TCL, CLIP, BLIP) ACROSS MULTIPLE DOWNSTREAM TASKS.

Model	Task	ASR	Model	Task	Flickr30		MSCOCO	
					TRmean	IRmean	TRmean	IRmean
TCL	VE	18.03	TCL	VR	28.93	37.85	46.52	43.71
BLIP	VQA	42.20	CLIP		9.23	15.46	17.45	17.92

Rather than a weakness, this sensitivity validates the *targeted nature* of our gray-box assumption. The performance drop under encoder mismatch precisely delineates the intended operational boundary of such attacks. Interestingly, adversarial examples transfer slightly better to ResNet-based CLIP variants than to other ViTs (e.g., ViT-B/32, ViT-L/14), suggesting that architectural similarity alone does not guarantee transferability. Instead, transfer efficacy depends on how perturbation patterns interact with the target encoder's inductive biases. For instance, ViT-B/32's coarser patch embedding may attenuate fine-grained adversarial signals, while ViT-L/14's deeper architecture enhances robustness through stronger feature abstraction. In contrast, ResNet's local receptive fields may inadvertently amplify certain perturbations, yielding relatively higher ASR. This highlights that cross-encoder transferability is governed by a nuanced interplay of spatial resolution, model capacity, and training dynamics.

4) Transferability across models: To evaluate the black-box transferability of CoGA, we generate adversarial image-text pairs on ALBEF and directly apply them to other models, including TCL, CLIP, and BLIP. As shown in Table VIII, the ASR generally decreases compared to the gray-box setting, which is expected given the absence of target-model encoder information during optimization.

However, transfer remains effective when the source and target models share design similarities. For example, ALBEF and TCL both adopt momentum-based contrastive learning with ViT-B/16 image encoders and 6-layer text transformers, resulting in relatively high ASR upon transfer. In contrast, models with distinct alignment mechanisms or encoder depths (e.g., CLIP with 12-layer text transformer) exhibit lower transfer performance. This indicates that CoGA exploits *shared vulnerabilities in multimodal alignment design*, rather than model-specific artifacts. The results demonstrate that cross-modal misalignment is a generalizable attack vector, particularly among models following similar encoder–fusion paradigms.

C. Ablation Study

1) Effect of hyper-parameters: We conduct the ablation experiments to investigate the effect of α in Equation 9, β in Equation 13 and N (i.e., the number of candidate adversarial texts).

Figure 9 shows a significant performance improvement when $\alpha > 0$, with comparable results as α increases. It indicates the efficacy of our proposed image attack method. The guidance of text embeddings assists to generate powerful adversarial images.

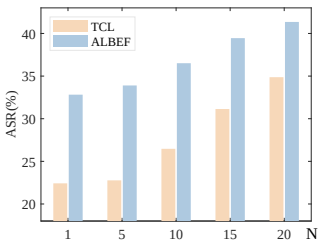

Fig. 6. The ASR(%) results for VE task with different N.

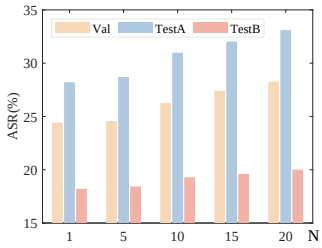

Fig. 7. The ASR(%) results for VG task with different N.

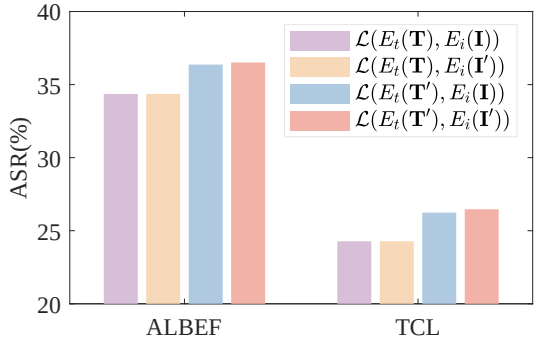

Fig. 8. The ASR(%) results for VE task with different parameter combinations in the first term of Equation 13.

Figure 10 shows a gradual rise in ASR as β increases. When $\beta = 1$, Equation 13 simply retains the first term, which selects the final sample based solely on the distance between the adversarial image and text. Such a scoring mechanism disregards the semantics and readability of the texts. Additionally, an increase in N also leads to a higher ASR, as shown in Figure 6-7. When N becomes too large, the candidate library will contain a large number of adversarial samples with $\delta_t > 1$. While these samples may have a stronger impact on the model's behavior, they often deviate significantly from the original semantics and may include confusing expressions, which are not conducive to the realistic deployment of the attack. Based on these insights and taking into account the overall computational overhead of the attack, we end up setting β at 0.7 and N at 10.

2) Why attack image before text: Due to the inherent differences between image and text spaces, existing adversarial attack methods for multimodal LMs typically attack each modality separately [19], [20], [22], [43], [44]. Such strategies do not impose a requirement on the attack sequence. Particularly for text attacks, most work relies on existing single-modality attack methods (e.g., BERT-Attack) to generate adversarial samples. In contrast, our text attacks leverage the results of the image attacks (i.e., adversarial images $\mathbf{I}^{'}$), as detailed in Equations 12 and 13. This dependency between modalities directly affects the attack sequence and motivates the design of ablation experiments to investigate the optimal attack scheme.

To illustrate the intuitive impact of attack sequence, we categorically discuss the variation of parameters in the specific equations. The first term of Equation 13 has four possible combinations as shown in Figure 8. $\mathcal{L}(E_t(\mathbf{T}), E_i(\mathbf{I}^{'}))$ and

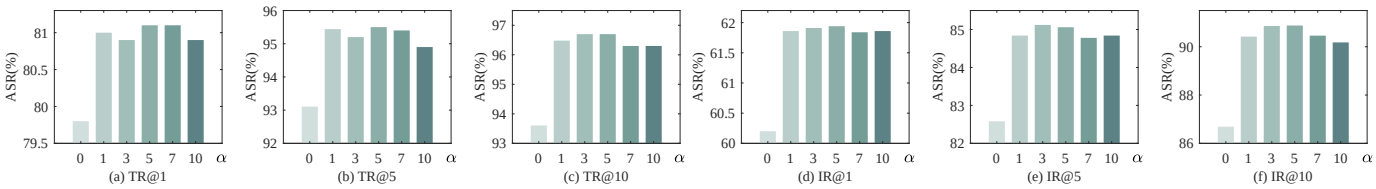

Fig. 9. The ASR(%) results of CLIP for VR task on Flickr30K dataset with different α.

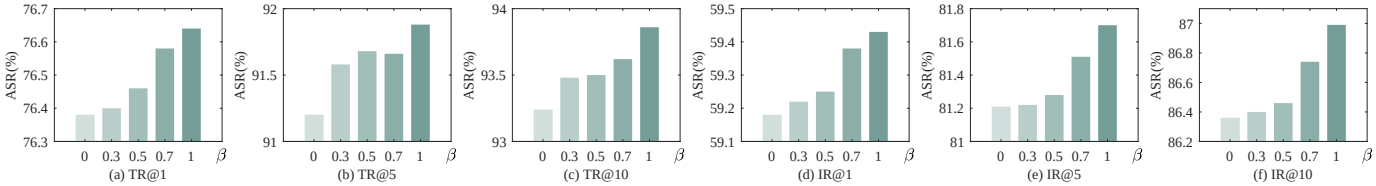

Fig. 10. The ASR(%) results of ALBEF for VR task on MSCOCO dataset with different β.

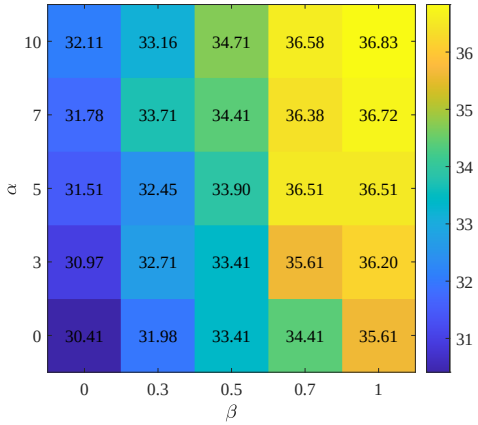

Fig. 11. The ASR(%) results of ALBEF for VE task under different combinations of hyper-parameters α and β.

$\mathcal{L}(E_t(\mathbf{T}^{'}), E_i(\mathbf{I}^{'}))$ corresponds to the original Equation 12, i.e., attacking image modality first and using $\mathbf{I}^{'}$ for subsequent text attacks. $\mathcal{L}(E_t(\mathbf{T}), E_i(\mathbf{I}))$ and $\mathcal{L}(E_t(\mathbf{T}^{'}), E_i(\mathbf{I}))$ corresponds to modifying Equation 12 to $\mathcal{L}(E_t(\mathbf{T}_{\setminus w_i}), E_i(\mathbf{I}))$, i.e., attacking the text modality first. The results show that our proposed strategy achieves optimal attack performance and further confirm that effective utilization of the collaborative relationship between modalities aids in uncovering the vulnerability of multimodal LMs.

3) Interaction effects between parameters: To further understand the interaction between the two key hyper-parameters α and β, we conduct a joint ablation study on the VE task using ALBEF. As illustrated in Figure 11, we vary α from 0 to 10, and β from 0 to 1, and record the corresponding ASR.

The results reveal a clear synergistic effect: ASR generally increases with larger values of both α and β, confirming that stronger image perturbations and more aggressive cross-modal misalignment guidance in text selection jointly enhance attack efficacy. Notably, β exerts a slightly stronger influence than α, suggesting the critical role of the image-guided adversarial text selection strategy in overall performance.

Furthermore, the most significant gains occur in the low-parameter regime (i.e., $\alpha < 5$, $\beta < 0.7$), while further increases yields relatively smaller returns. This suggests that beyond a certain threshold, additional perturbation strength or alignment disruption provides limited marginal benefit, likely due to saturation in the embedding space or constraints imposed by the imperceptibility requirement.

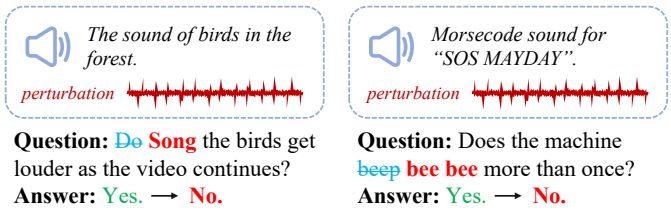

Fig. 12. Examples of CoGA on the Pengi for AQA task.

Fig. 13. Representative failure cases for VE and VQA tasks. Despite high overall ASR, certain samples resist adversarial manipulation due to (a,b,d,e) insufficient perturbation strength under imperceptibility constraints, (c) high-fidelity perceptual attributes (e.g., color), or (f) inherent ambiguity in image-text semantics or weak relevance.

V. Case Study

A. Expand to Audio-Language Model

We extend our method to the audio-language model and evaluate its performance on the audio question-answering (AQA) task [55] using Pengi [56], achieving an ASR greater than 30%. Figure 12 shows several attack examples on the AQA task. The perturbations have no effect on human comprehension of the audio, and the adversarial texts remain understandable, where the modifications are more like typos or personal idioms. However, the model produces incorrect responses. The results demonstrate that CoGA is not only applicable to image-language models but also transfers well to other kinds of multimodal LMs. Our strategy effectively identifies and amplifies the inconsistencies across various types of data to accomplish a successful adversarial attack.

B. Failure Case Analysis

While CoGA achieves high ASR across diverse tasks, certain samples remain resistant to adversarial manipulation. To better understand the boundary of our method, we analyze representative failure cases from the VE and VQA tasks, where both inputs and model decisions are highly interpretable. As shown in Figure 13, we categorize these failure cases into several typical patterns.

First, in cases (a), (b), (d), and (e), the perturbations fail to cross the model's decision boundary. This reflects a fundamental trade-off between stealth and efficacy: when imperceptibility is prioritized, perturbations remain too subtle to disrupt highly confident predictions. Second, in cases like (c), perceptual features such as object color or shape create strong visual-textual correspondences. These low-ambiguity anchors resist perturbations, preventing effective disruption of alignment. Third, cases like (f) involve semantic ambiguity, where the model may be uncertain due to factors such as affective interpretation (e.g., "sad" vs. "angry") or weak image-text relevance (e.g., questions about unseen objects). In these scenarios, the model's uncertainty reduces the effectiveness of adversarial manipulation. Furthermore, some models exhibit intrinsic robustness due to attention smoothing or feature aggregation mechanisms. These mechanisms help the model compensate for minor misalignments, thus preserving output consistency despite adversarial perturbations.

Collectively, these patterns highlight the conditions under which CoGA is most effective: when samples have moderate semantic alignment that is neither over-determined by salient attributes nor undermined by inherent ambiguity.

VI. Defense Strategies

While adversarial attacks pose a security threat to multimodal language models, various defense mechanisms have been explored to enhance model robustness. Broadly, these can be categorized into **proactive** approaches (e.g., adversarial training during model fine-tuning) and **reactive** approaches (e.g., input filtering or anomaly detection at inference time). To assess the resilience of CoGA under realistic defense scenarios, we evaluate both paradigms in this section.

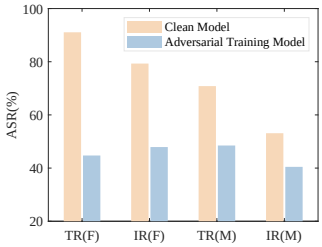

Fig. 14. The ASR(%) results of CLIP for VR tasks after implementing adversarial training defense scheme on Flickr30K (F) and MSCOOC (M) dataset. The TR/IR indicators report the average values of R@1, R@5, and R@10.

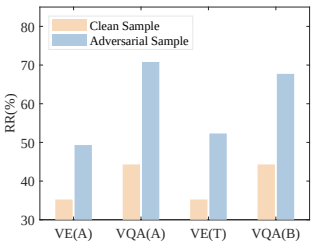

Fig. 15. The sample rejection rate (RR%) results of models (ALBEF, TCL, and BLIP) on different tasks (VE and VQA) after implementing the modal consistency detection defense scheme.

A. Adversarial Training Defense

We perform lightweight adversarial fine-tuning on CLIP using a limited set of CoGA-generated adversarial examples. Specifically, we randomly sample 5,000 clean image-text pairs from the Flickr30K and MSCOCO training sets, respectively, and generate corresponding adversarial samples using our method. The model is then fine-tuned for 3 epochs with a combined loss:

$$\mathcal{L}_{\text{total}} = \lambda \cdot \mathcal{L}_{\text{clean}} + (1 - \lambda) \cdot \mathcal{L}_{\text{adv}}, \tag{14}$$

where $\mathcal{L}_{\text{clean}}$ and $\mathcal{L}_{\text{adv}}$ denote the contrastive losses on clean and adversarial samples, respectively, and $\lambda = 0.5$ is chosen to preserve clean accuracy. The learning rate is set to 1×10^{-6} to avoid catastrophic forgetting.

As shown in Figure 14, the decline in ASR caused by adversarial training confirms its effectiveness in improving robustness. Nevertheless, CoGA still achieves an ASR of approximately 40% on both datasets after defense. This residual vulnerability is expected given the small scale of the adversarial training set, which limits the model's exposure to diverse attack patterns. Notably, this result suggests that CoGA exploits fundamental weaknesses in cross-modal alignment that cannot be easily mitigated by data augmentation alone. This highlights the need for more comprehensive defenses, particularly when encoder-level access is assumed.

B. Cross-Modal Consistency Detection

We further explore a test-time detection mechanism based on cross-modal semantic consistency. Using the pre-trained CLIP-ViT-B/16 model, we compute the cosine similarity between image and text embeddings as a consistency score. We construct a binary classifier on 5,000 clean and 5,000 CoGA-generated samples from Flickr30K training sets, and select the optimal decision threshold at the point maximizing Youden's J statistic (i.e., $J = \text{TPR} - \text{FPR}$) on the ROC curve.

As shown in Figure 15, the detector successfully identifies a considerable fraction of CoGA-generated adversarial samples across multiple models and downstream tasks. However, it also incurs a non-negligible false positive rate on clean inputs. This limitation stems from two factors: first, the decision threshold is calibrated on Flickr30K data and may not generalize well

to other data distributions or attack configurations; second, it reflects the *high imperceptibility* of CoGA. The perturbed image-text pairs preserve semantic coherence and structural fidelity, causing their cross-modal consistency scores to closely overlap with those of benign samples. This inherent ambiguity not only challenges simple consistency-based detectors but also validates the stealthiness of our attack: its perturbations are subtle enough to evade detection while still effectively disrupting model predictions.

Together, these experiments demonstrate that while existing defenses can partially mitigate CoGA, they face inherent limitations under gray-box assumptions. The persistence of moderate ASR after adversarial training and the ambiguity in consistency-based detection, jointly underscore the challenge of defending against attacks that precisely target cross-modal alignment without introducing perceptible artifacts.

VII. Discussions

In this paper, we propose CoGA, a novel gray-box collaborative multimodal adversarial attack strategy. By perturbing the embedding space of end-to-end encoders, our method disrupts cross-modal alignment, effectively lowering the model's performance across downstream tasks. Compared to existing approaches, our work offers several notable advantages. First, in terms of attack assumptions, CoGA only requires access to the target model's encoder information, thereby avoiding the overly strong and often unrealistic knowledge requirements of white-box attacks. Second, in terms of attack effectiveness, our method achieves more precise and effective adversarial attacks than traditional black-box methods. Third, we systematically consider the collaborative relationship between modalities, with particular improvements to the often-overlooked text attack component. While most previous works focus primarily on optimizing attacks for visual modality and often simply transfer single-modality text attack solutions, we innovatively enhance the classical BERT-Attack by introducing image embeddings to guide vulnerable word selection and by incorporating a multi-sample election strategy. This establishes a truly bidirectional multimodal-guided attack paradigm, where each modality's adversarial perturbation is informed by the other.

This work provides an in-depth exploration of adversarial robustness in multimodal LMs, shedding light on their potential vulnerabilities. As multimodal LMs are increasingly deployed in critical applications such as intelligent assistants, healthcare, and autonomous driving, the security challenges around these models are becoming ever more urgent. Particularly concerning is the possibility of covert information injection through non-text channels (e.g. images or audio), making multimodal adversarial attacks far less detectable than traditional text-only attacks. An attacker could pre-construct adversarial samples and exploit unsuspecting users as unwitting transmission vectors. For instance, attackers might lure users into visiting a webpage containing a maliciously crafted image or send a phishing email with a booby-trapped audio file, then rely on the user to input these adversarial multimodal samples into the victim model. Such scenarios illustrate how our findings on cross-modal vulnerabilities have serious implications for real-world security.

TABLE IX
The ASR(%) results of our proposed method and other multimodal attack methods for VR task on Flickr30K and MSCOCO datasets. The white shading indicates the white-box attack scenario, the gray shading indicates the gray-box attack scenario, and the yellow shading indicates the black-box attack scenario.

Model	Attack	Flickr30K		MSCOCO	
		TRmean	IRmean	TRmean	IRmean
ALBEF	Co-Attack	52.97	58.73	64.31	59.22
	SGA	97.11	96.72	96.19	97.13
	TMM	97.53	97.51	96.79	97.73
	Ours	92.20	88.19	87.29	75.88
TCL	Co-Attack	73.80	72.11	70.77	63.16
	SGA	43.95	48.83	54.04	58.82
	TMM	64.97	69.60	70.19	74.02
	Ours	96.07	90.75	84.34	73.24
CLIP	Co-Attack	76.53	70.51	66.81	50.48
	SGA	33.83	43.57	51.76	60.25
	TMM	52.90	60.90	68.37	75.34
	Ours	91.07	79.31	70.83	53.11

Naturally, we must acknowledge the limitations of our current work. Our attack operates in a gray-box setting, where the victim model's encoders are assumed to be accessible [57]. Within this assumption, we focus on disrupting the semantic alignment between multimodal inputs at the encoder representation level. However, it is crucial to clarify that the attack performance is closely tied to the transparency of the model, which suggests that gray-box attacks inherently face performance ceilings due to the limited knowledge available to the attacker. As shown in Table IX, state-of-the-art transferable black-box attacks, such as SGA [44] and TMM [22], achieve slightly better performance than CoGA on white-box surrogate models (i.e., entries with white shading). Yet, under black-box conditions (i.e., entries with yellow shading), their performance experiences a significant decline. Moreover, Co-Attack, which has access to similar encoder information as CoGA, performs worse due to its insufficient utilization of multimodal interactions.

It is essential to emphasize the fundamental differences between CoGA and prior transferable multimodal attack frameworks such as TMM [22]. First, from the perspective of attacker knowledge, TMM requires a surrogate model that includes both the encoders and the fusion module, whereas CoGA only requires access solely to the victim model's encoders, with no need for a surrogate model. Second, in terms of optimization goals, TMM primarily focuses on optimizing transferability, aiming to find adversarial examples that generalize to unknown target models. In contrast, CoGA directly injects semantic misalignment into the victim's encoder to disrupt cross-modal alignment. Third, from a practical deployment standpoint, CoGA's gray-box assumption is more aligned with real-world scenarios: many systems reuse open-source encoders but keep their fusion and decoding modules proprietary. This makes it far more feasible for attackers to replicate or obtain the encoder, compared to constructing a complete surrogate model as required by TMM.

Additionally, we encountered challenges when extending

our attack to the latest large-scale multimodal LMs. Our method suffers a performance drop when transferred to advanced systems such as MiniGPT-4 [45] and LLaVa [58]. This decline may stem from two factors: (1) more sophisticated internal alignment and anti-noise mechanisms in cutting-edge large-scale LMs, and (2) stronger prior knowledge acquired through large-scale training. Currently, developing effective black-box attacks against such powerful multimodal LMs remains a major challenge in the field. However, our work represents an important step in this direction by relaxing the attack's knowledge assumptions and highlighting encoder-level vulnerabilities that could inspire future research.

Looking ahead, we identify several promising directions for future research on multimodal adversarial attacks. First, as multimodal LMs continue to evolve toward full integration of diverse modalities, attack strategies should expand beyond vision–language models to encompass audio, video, and other data modalities. Although we have taken initial steps to demonstrate the feasibility of our method on audio–language models, substantial further research remains to be done. Second, achieving targeted adversarial attacks presents another crucial challenge. Most current attacks aim to cause general model failures, but the ability to precisely manipulate model outputs (e.g., inducing a specific incorrect answer in a visual question answering task) would significantly increase the real-world threat potential. Lastly, we hope that the insights provided by this work will spur the development of stronger defense techniques, ultimately fostering a more secure and reliable ecosystem for multimodal artificial intelligence.

VIII. Conclusion

In this paper, we explore a gray-box collaborative adversarial attack scenario for multimodal LMs and introduce CoGA, a novel adversarial attack method. By perturbing the embedding representations from the encoders, our method effectively disrupts the consistency of the input data, thereby degrading the model's performance on downstream tasks. Through comprehensive evaluation, we demonstrate the efficacy of the proposed attack method and the universality of the attack concept, suggesting the consistent vulnerability of multimodal LMs in the face of adversarial attacks. Overall, our work represents a significant step forward for the adversarial robustness of multimodal LMs and provides a basis for developing more robust and secure models in the future.

References

[1] F. Li, H. Zhang, Y.-F. Zhang, S. Liu, J. Guo, L. M. Ni, P. Zhang, and L. Zhang, "Vision-language intelligence: Tasks, representation learning, and large models," *arXiv preprint arXiv:2203.01922*, 2022.

[2] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *arXiv preprint arXiv:2306.13549*, 2023.

[3] S. Huang, H. Zhang, Y. Gao, Y. Hu, and Z. Qin, "From image to video, what do we need in multimodal llms?" *arXiv preprint arXiv:2404.11865*, 2024.

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

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

[6] B. Liu, L. He, Y. Liu, T. Yu, Y. Xiang, L. Zhu, and W. Ruan, "Transformer-based multimodal infusion dialogue systems," *Electronics*, vol. 11, no. 20, p. 3409, 2022.

[7] C. Li, Z. Gan, Z. Yang, J. Yang, L. Li, L. Wang, J. Gao *et al.*, "Multimodal foundation models: From specialists to general-purpose assistants," *Foundations and Trends® in Computer Graphics and Vision*, vol. 16, no. 1-2, pp. 1–214, 2024.

[8] Z. Huang, X. Mo, and C. Lv, "Multi-modal motion prediction with transformer-based neural network for autonomous driving," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2605–2611.

[9] A. Singh, "Transformer-based sensor fusion for autonomous driving: A survey," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3312–3317.

[10] S. Liu, X. Wang, Y. Hou, G. Li, H. Wang, H. Xu, Y. Xiang, and B. Tang, "Multimodal data matters: language model pre-training over structured and unstructured electronic health records," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 1, pp. 504–514, 2022.

[11] H.-Y. Zhou, Y. Yu, C. Wang, S. Zhang, Y. Gao, J. Pan, J. Shao, G. Lu, K. Zhang, and W. Li, "A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics," *Nature biomedical engineering*, vol. 7, no. 6, pp. 743–755, 2023.

[12] E. Bagdasaryan, T.-Y. Hsieh, B. Nassi, and V. Shmatikov, "Abusing images and sounds for indirect instruction injection in multi-modal llms," 2023.

[13] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

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

[15] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 69–85.

[16] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[17] C. Schlarmann and M. Hein, "On the adversarial robustness of multi-modal foundation models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3677–3685.

[18] E. Shayegani, M. A. A. Mamun, Y. Fu, P. Zaree, Y. Dong, and N. Abu-Ghazaleh, "Survey of vulnerabilities in large language models revealed by adversarial attacks," *arXiv preprint arXiv:2310.10844*, 2023.

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

[20] Z. Yin, M. Ye, T. Zhang, T. Du, J. Zhu, H. Liu, J. Chen, T. Wang, and F. Ma, "Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models," *arXiv preprint arXiv:2310.04655*, 2023.

[21] Y. Zhao, T. Pang, C. Du, X. Yang, C. Li, N.-M. M. Cheung, and M. Lin, "On evaluating adversarial robustness of large vision-language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[22] H. Wang, K. Dong, Z. Zhu, H. Qin, A. Liu, X. Fang, J. Wang, and X. Liu, "Transferable multimodal attack on vision-language pre-training models," in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2024, pp. 102–102.

[23] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[24] I. D. Raji, E. M. Bender, A. Paullada, E. Denton, and A. Hanna, "Ai and the everything in the whole wide world benchmark," *arXiv preprint arXiv:2111.15366*, 2021.

[25] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[26] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[27] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.

[28] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," *arXiv preprint arXiv:1908.06281*, 2019.

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

[30] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2730–2739.

[31] A. Ganeshan, V. BS, and R. V. Babu, "Fda: Feature disruptive attack," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8069–8079.

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

[33] Y. Lu, Y. Jia, J. Wang, B. Li, W. Chai, L. Carin, and S. Velipasalar, "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, pp. 940–949.

[34] B. Wang, C. Xu, S. Wang, Z. Gan, Y. Cheng, J. Gao, A. H. Awadallah, and B. Li, "Adversarial glue: A multi-task benchmark for robustness evaluation of language models," *arXiv preprint arXiv:2111.02840*, 2021.

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

[36] S. Ren, Y. Deng, K. He, and W. Che, "Generating natural language adversarial examples through probability weighted word saliency," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 1085–1097.

[37] B. Wang, C. Xu, X. Liu, Y. Cheng, and B. Li, "Semattack: Natural textual attacks via different semantic spaces," *arXiv preprint arXiv:2205.01287*, 2022.

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

[39] B. Wang, H. Pei, B. Pan, Q. Chen, S. Wang, and B. Li, "T3: Tree-autoencoder constrained adversarial text generation for targeted attack," *arXiv preprint arXiv:1912.10375*, 2019.

[40] L. Xu, A. Cuesta-Infante, L. Berti-Equille, and K. Veeramachaneni, "R&r: Metric-guided adversarial sentence generation," *arXiv preprint arXiv:2104.08453*, 2021.

[41] A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.

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

[43] X. Xu, X. Chen, C. Liu, A. Rohrbach, T. Darrell, and D. Song, "Fooling vision and language models despite localization and attention mechanism," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4951–4961.

[44] D. Lu, Z. Wang, T. Wang, W. Guan, H. Gao, and F. Zheng, "Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 102–111.

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

[46] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Salmonn: Towards generic hearing abilities for large language models," *arXiv preprint arXiv:2310.13289*, 2023.

[47] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[48] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.

[49] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 49 250–49 267, 2023.

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

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

[52] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[53] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, and J. Huang, "Vision-language pre-training with triple contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 671–15 680.

[54] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[55] S. Lipping, P. Sudarsanam, K. Drossos, and T. Virtanen, "Clotho-aqa: A crowdsourced dataset for audio question answering," in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 1140–1144.

[56] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: An audio language model for audio tasks," *Advances in Neural Information Processing Systems*, vol. 36, pp. 18 090–18 108, 2023.

[57] X. Wang, Z. Ji, P. Ma, Z. Li, and S. Wang, "Instructta: Instruction-tuned targeted attack for large vision-language models," *arXiv preprint arXiv:2312.01886*, 2023.

[58] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 296–26 306.