Iterative Adversarial Attack on Image-guided Story Ending Generation

Youze Wang, Wenbo Hu, and Richang Hong, *Member, IEEE,*

***Abstract*—Multimodal learning involves developing models that can integrate information from various sources like images and texts. In this field, multimodal text generation is a crucial aspect that involves processing data from multiple modalities and outputting text. The image-guided story ending generation (IgSEG) is a particularly significant task, targeting on an understanding of complex relationships between text and image data with a complete story text ending. Unfortunately, deep neural networks, which are the backbone of recent IgSEG models, are vulnerable to adversarial samples. Current adversarial attack methods mainly focus on single-modality data and do not analyze adversarial attacks for multimodal text generation tasks that use cross-modal information. To this end, we propose an iterative adversarial attack method (Iterative-attack) that fuses image and text modality attacks, allowing for an attack search for adversarial text and image in a more effective iterative way. Experimental results demonstrate that the proposed method outperforms existing single-modal and non-iterative multimodal attack methods, indicating the potential for improving the adversarial robustness of multimodal text generation models, such as multimodal machine translation, multimodal question answering, etc.**

***Index Terms*—Multimodal, adversarial attack, multimodal text generation.**

I. Introduction

MULTIMODAL learning aims to build models that can process and integrate information from multiple modalities, such as image and language, which is an increasing research field with great potential for artificial general intelligence [1], [2]. The multimodal text generation task takes data from multiple modalities as input and ends up with text as output, which is considered as the basic ability of human intelligence. The potential applications of multimodal text generation are far-reaching and transformative, whose innovation and expansion have been into new fields, reshaping the way we communicate and interact with information-rich content. Typical applications include multimodal machine translation [3], [4], multimodal dialogue response generation [5], [6], multimodal question answering [7], multimodal MemexQA [8] and image-guided story ending generation [9], [10]. Among these tasks, image-guided story ending generation is a natural task for an average person to understand and generate multimodal information and represents a fundamental problem for machine intelligence. This task introduces ending-related images to the story ending generation, which can supplement the story ending with diverse visual concepts. IgSEG models requiring understanding complex relationships between text modality data and image modality data is a standard multimodal text generation task.

The deep neural networks, the backbone model of the recent IgSEG models, albeit rapidly developing, have been shown vulnerable to adversarial samples which include adding imperceptible perturbations on original images [11], [12] or modifying some words in original texts that do not affect human semantic understanding [13], [14]. Regarding multimodal tasks especially multimodal text generation models, to our knowledge, there has been no research to systematically analyze the adversarial robustness performance and design an adversarial attack solution that utilizes the cross-modal information. The previous adversarial attack methods mainly focus on the single-modal data, such as image modality adversarial attack methods [15], [16], text modality adversarial attack methods [13], [17], [18], or simply use a step-wise mechanism which first perturbs the discrete texts and then perturbs the continuous image based on the text perturbation, which is difficult to find the most vulnerable multimodal information patch pairs.

As we know, the input of multimodal text generation tasks involves inputs from multiple modalities, which is more challenging than the single-modal tasks. Simply migrating the single-modal adversarial attack methods may face performance bottlenecks since the information shift caused by the perturbation may be corrected by data for another modality. For example, Figure 1 shows the multimodal information for a story including text context (i.e. story context) and visual information (i.e. ending-related image), for which the single-modal adversarial attack methods all failed due to the complementary information between text data and image data, where the information shift caused by a single-modal adversarial attack can be corrected by another modality data. Therefore, the critical issue examined in this paper: *how to find the most vulnerable adversarial patch that could take advantage of the cross-modal information?* To tackle this issue, we need to consider the gap between the discrete text modality and the continuous image modality, it is hard to optimize the designed object function in discrete space.

In this paper, to investigate the adversarial robustness problem for the image-guided story ending generation task, we propose an iterative adversarial attack method to effectively craft an imperceptible attack for text-image pair samples.

Manuscript received 11 July 2023; revised 11 September 2023 and 15 November 2023; accepted 11 December 2023. This work was supported by the National Key Research and Development Program of China (No. 2022ZD0118802), the NSF of China Project (No. 61932009 and 62306098) and the Open Projects Program of State Key Laboratory of Multimodal Artificial Intelligence Systems. (Corresponding author: Wenbo Hu).

Y. Wang, W. Hu, R. Hong are at the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China. (E-mail: wangyouze@mail.hfut.edu.cn, {wenbohu,hongrc}@hfut.edu.cn)

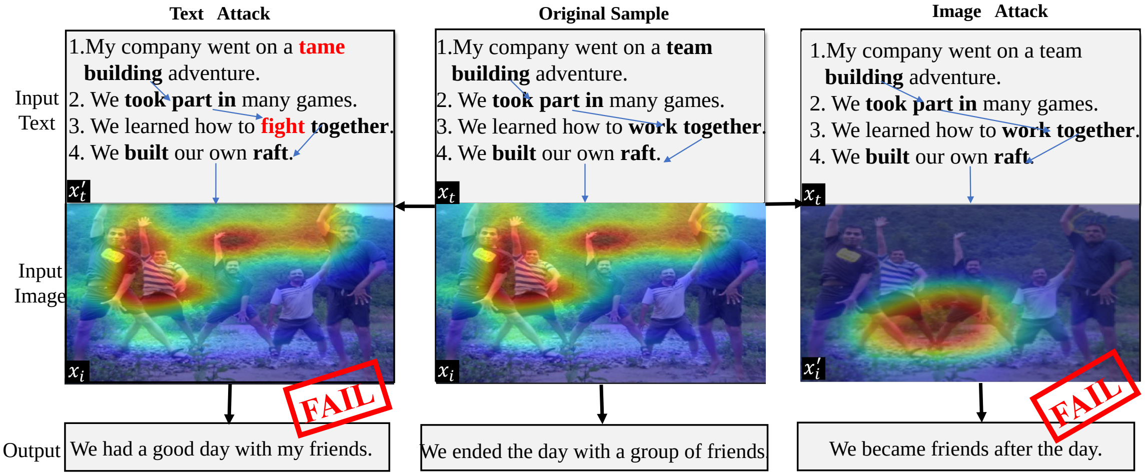

Text Attack
Original Sample
Image Attack
Input Text
1.My company went on a tame building adventure.
2. We took part in many games.
3. We learned how to fight together.
4. We built our own raft.
1.My company went on a team building adventure.
2. We took part in many games.
3. We learned how to work together.
4. We built our own raft.
1.My company went on a team building adventure.
2. We took part in many games.
3. We learned how to work together.
4. We built our own raft.
x_t'
x_t
x_t
Input Image
x_i
x_i
x_i'
FAIL
FAIL
Output
We had a good day with my friends.
We ended the day with a group of friends
We became friends after the day.

Fig. 1. An example of the bottleneck single modality adversarial attack against the multimodal text generation model. The blue arrows denote the information chain, and the heat map shows where the target model focuses on the image. When a single modality adversarial example attacks the target model, the other unperturbed modality data may provide complementary information, making the attack fail.

We fuse the image modality attack into the text modality attack and iteratively perturb image modality when every small perturbation in the original text modality is changed into continuous space, which allows for an attack search for the adversarial text and image for IgSEG models instead of independently. Experimental results show that Iterative-attack outperforms the existing single-modal adversarial attack method (kNN and WordSwap [19]) and multimodal adversarial attack method [20] in terms of success rate, and semantic similarity.

In summary, our contribution is as follows:

- To the best of our knowledge, this is the first multimodal adversarial attack on a multimodal text generation task.
- We propose a new iterative multimodal adversarial attack against IgSEG models to fuse the image modality attack into a text modality attack, which can iteratively find the most vulnerable multimodal information patch.
- We evaluate our iterative attack method on four IgSEG models with two real-world datasets, and experimental results demonstrate our approach outperforms the baseline methods.

The rest of the paper is organized as follows. Section II summarises the related works on the image-guided story ending generation task and adversarial attack including single-modal adversarial attacks and multimodal adversarial attacks. In Section III, we first demonstrate the problem definition in our proposed method. Then we introduce the proposed multimodal adversarial attack algorithm for IgSEG models in detail. In Section IV, we first conduct a multimodal adversarial attack on IgSEG models with different attack methods. Then we conduct ablation experiments to evaluate the effectiveness of the proposed Iterative-attack. After that, we present some visualization results and discussions to further analyze our method. Section V concludes the paper.

II. Related Works

A. Image-guided Story Ending Generation Task

Recently, while multimodal tasks have attracted significant attention [21], [22], Huang et al. [9] proposed a new task called Image-guided Story Ending Generation (IgSEG), which generates a story ending for the story context and an ending-related image. Huang et al. [9] first explored incorporating a context and an image to generate a story ending, which proposed a GCN-based text encoder and an LSTM-based decoder. Xue et al. [10] proposed an end-to-end Multimodal Memory Transformer that modeled and fused both contextual and visual information to obtain the multimodal dependency for IgSEG. Other multimodal text generation tasks, such as IgSEG, also take data from multiple modalities as input and end up with texts as output, which utilize complementary multimodal information. The adversarial attack against multimodal text generation tasks requires finding the most vulnerable multimodal information patch.

B. Adversarial Attack

Recently, the adversarial attack against Deep Neural Networks (DNN) has drawn the keen interest of researchers [23]. The deep neural networks are found vulnerable to adversarial samples, for which the small perturbations are added on the original inputs [13], [16], [18], [24]–[29]. That is, the attack on adversarial samples is imperceptible to human judges while they can mislead incorrect outputs of the deep neural networks. Now, there have been many works done on different data types, such as images, texts, and graphs. Based on the difference in modality, we roughly summarize existing adversarial attack models into two categories: single-modal and multimodal adversarial attacks.

1) Single-modal Adversarial Attack: The adversarial attack is first proposed in computer vision for classification tasks, which illustrates the vulnerability of deep learning models. In the image classification task, there are many algorithms based on both the architecture and the parameters of the model performing gradient-based optimization on the input and constructing adversarial examples, such as FGSM [16], PGD [25], MIM [24] and AutoMA [30]. In the text classification task, current successful adversarial attack methods adopt heuristic rules to modify the characters of a word [26], and substitute words with synonyms [27]. Gao et al. [14] applied perturbations based on word embeddings such as Glove [31], which were not strictly semantically and grammatically coordinated. Li et al. [13] turned BERT against its fine-tuned models and used BERT to generate adversarial samples for texts. In addition, Wang et al. [32] explored a white-box adversarial attack for images classifiers from the perspective of interpretable features. Shen et al. [33] applied boosting-based black-box attacks to enhance the diversity of perturbation, which can contribute to adversarial training. Naseer [34] trained a purifier network in a self-supervised manner to defend against unseen adversarial attacks. During the training, the adversarial images are generated by a self-supervised perturbation that can disrupt the deep perceptual features.

For text generation tasks, universal adversarial attack [28], a new type of attack, consists of a single snippet of text that can be added to any input sentence to mislead the neural machine translation model. Seq2Sick [18] is a white-box attack method against sequence-to-sequence models, which solves an optimization problem by gradient projection. T3 [35] is a tree-based autoencoder to embed the discrete text data into a continuous representation space, which can perform an adversarial perturbation against QA models. CLAPS [36] can mitigate the conditional text generation problem by contrasting positive pairs with negative pairs, such that the model is exposed to various valid or incorrect perturbations of the inputs, for improved generalization.

However, the input of multimodal models involving multiple modalities and the complexity of the text generation tasks make it impractical to directly employ the standard single-modal adversarial attack methods against multimodal text generation tasks.

2) Multimodal Adversarial Attack: Multimodal learning [9], [37], [38] aims to understand the current scene from multiple modalities. There are some adversarial attacks attempted on multimodal neural networks. Xu et al. [39] investigated attacking the visual question answering model by perturbing the image modality. Agrawal et al. [40] and Shah et al. [41] attempted to attack the vision-and-language model by perturbing the text modality. Yan et al. [42] showed that standard multimodal fusion models were vulnerable to single-source adversaries, and studied an adversarial robust fusion strategy and proposed a defense method. Zhou [43] evaluated the vulnerability of CLIP to the universal attack on the image-text retrieval task and the image classification task. Wang et al. [44] first explored the targeted adversarial attack against cross-modal hash retrieval, which collaborates with the semantic translator to generate adversarial examples that contain the target semantics specified by the attacker. Zhu et al. [45] proposed a query-based multi-modal knockofs-driven adversarial samples generation method to attack cross-modal hash retrieval in a black-box setting. Zhang et al. [20] studied the adversarial attack on popular vision-language pre-training models and vision-language tasks. However, regarding multimodal text generation tasks, to our knowledge, there is no relevant work to systematically analyze and design adversarial attacks. Compared with the above adversarial attack for multimodal classification tasks, the adversarial attack for multimodal text generation is more challenging.

III. Method

In this section, we detail the proposed iterative multimodal adversarial attack method (Iterative-attack) for IgSEG models. Our method is inspired by the idea: a multimodal neural network incorporates multimodal information from different modalities that can complement each other. Simply migrating the single-modal adversarial attack methods to multimodal text generation tasks would face the dilemma that the information shift caused by a single-modal adversarial attack may be corrected by another modality's information. To solve the problem, Iterative-attack fuses the image modality attack into the text modality attack to iteratively find the most vulnerable multimodal information patch. The key notations used in the paper as summarized in Table I.

A. Problem Formulation

Given a pre-trained IgSEG model $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$, which maps from the origin multimodal feature space $\mathcal{X}$ to the target text space $\mathcal{Y}$. The IgSEG model $\mathcal{F}$ generally has an DNN-based encoder-decoder structure [9], [10] and aims to maximize the story ending generation probability $p(y|x)$, where $x \in \mathcal{X}$ is the input story context x_t and ending-related image x_i in the origin multimodal space, and $y \in \mathcal{Y}$ is the ground-truth story ending in target text space, where $\mathcal{F}((x_t, x_i)) = y_p$. A successful multimodal adversarial attack against IgSEG models is to generate an adversarial context x'_t and an adversarial image x'_i, so that the BLEU score of the story ending generated by taking the adversarial sample (x'_t, x'_i) as input relative to the BLUE score of the original story ending is less than a threshold λ.

B. Multimodal Adversarial Attack Loss

By perturbing texts or images, generating adversarial examples against DNN models can fool DNN models. However, single-modal adversarial attack can't effectively maximize the attack on the output of multimodal models [20]. We address this issue by developing an effective iterative multimodal adversarial attack method against IgSEG models.

To find a more effective adversarial input (x'_t, x'_i), we try to maximize the adversarial loss of the target IgSEG model. Since the IgSEG models are trained to generate the next token of the story ending given the ending up until that token, we are looking for the adversarial text and image that can maximize the probability of wrong story ending (i.e., minimizes the probability of correct story ending) for the i-th token, given that the IgSEG model has produced the correct story ending up to step (i-1). We can calculate the adversarial loss as the following loss function:

$$\mathcal{L}_{adv} = \frac{1}{n}\sum_{i=1}^{n} \log p\left(y_i|(x'_t, x'_i), y_1, ..., y_{(i-1)}\right), \qquad (1)$$

TABLE I
THE MAIN NOTATIONS OF OUR PROPOSED METHOD.

Notation	Description
$\mathcal{F}$	the pre-trained IgSEG models
(x_t, x_i)	the original story context and ending-related image
(x'_t, x'_i)	the adversarial context and adversarial image
y	the ground-truth story ending
y_p	the story ending generated by the pre-trained IgSEG models with (x_t, x_i)
y_f	the story ending generated by the pre-trained IgSEG models with (x'_t, x'_i)
$\mathcal{X}$	the original multimodal feature space
$\mathcal{Y}$	the target text space
n	the length of the story ending
w_i	the i_{th} word in the text
Q_{wh}	the importance score of h_{th} word in the x_t
Q_x	the importance score of all words in x_t in descending order
L	the selected top-K important words from Q_x
S_c	the perturbation in character-level
S_w	the perturbation in word-level
C_{wh}	the substitutes set for the import word w_h
c_j	the j_{th} substitute word in C_{wh}
ϵ	the step size of the image attack
θ	the parameters of the target models
$S \in R^d$	a set of allowed perturbations for image attack
λ	the threshold for determining whether an attack is successful
sign(x)	a mathematical function that returns the sign of a real number ”x”

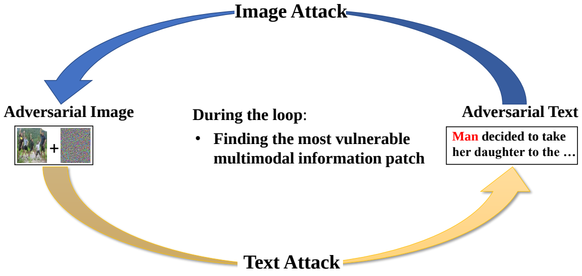


Image Attack
Adversarial Image
+
During the loop:
• Finding the most vulnerable multimodal information patch
Adversarial Text
Man decided to take her daughter to the ...
Text Attack

Fig. 2. Illustration of Iterative-attack. We fuse the image modality attack into the text modality attack to iteratively find the most vulnerable multimodal information patch, which can avoid the dilemma that the information shift caused by a single-modal adversarial attack may be corrected by another modality's information.

where n is the length of the story ending. By minimizing $\log p(\cdot)$, normalized by the sentence length n, we force the output probability vector of the IgSEG model to differ from the delta distribution on the token y_i, which may cause the predicted story ending to be wrong.

C. Adversarial Attack Algorithms

We now introduce the proposed algorithm that generates adversarial examples including adversarial text and adversarial image for IgSEG models. We first need to generate potential adversarial texts as supervised information, and then iteratively take the potential adversarial texts as input to attack the relevant image information to find the most vulnerable multimodal information patch. The flowchart of Iterative-attack is shown in Algorithm 1.

Step 1: Generate Potential Adversarial Text. Potential adversarial texts can be used as the input of Iterative-attack to find complementary adversarial images. In general, if a word is more important to the output of the target model, it is more likely to be attacked. Motivated by [13], we select the most important words in the story context which have a higher significant influence on the final output logit.

Specifically, Let $x_t = [w_0, ..., w_h, ...]$, and $\mathcal{F}((x_t, x_i)$ denotes the normal output of the IgSEG model, the importance score Q_{w_h} of word w_h in the original text x_t is defined as:

$$Q_{w_h} = \mathcal{F}((x_t, x_i)) - \mathcal{F}((x_{t@h}, x_i)) \tag{2}$$

where $x_{t@h} = [w_0, \cdots, w_{h-1}, \text{MASK}, w_{h+1}...]$ is the story context after replacing w_h with a special character [MASK]. We can obtain the important score of every word for the target model by using the Equation. 2. The important score of all words in the original text is Q_x, which is defined as (Algorithm 1, line 2 - 4) :

$$Q_x = [Q_{w_1}, \cdots, Q_{w_h}, \cdots, Q_{w_n}] \tag{3}$$

Then, we sort Q_x in descending order and select Top-K words as the important word list L based on the descending order of Q_x (Algorithm 1, line 5).

To generate candidate perturbation for every important word in list L and ensure that the generated adversarial text is visually or semantically similar to the benign one for human understanding, we combine bugs generation [17] in character-level perturbation called S_c and words replacement via BERT [13] in word-level perturbation called S_w as the text perturbation mechanism of Iterative-attack. The substitutes for the important word w_h is $C_{w_h} = S_c \oplus$[1]S_w , which will be used to generate a potential adversarial context as the input of iterative multimodal adversarial attack in step 2.

Step 2: Iterative Multimodal Adversarial Attack. To effectively attack the IgSEG models, we propose a new iterative multimodal adversarial attack solution, which fuses the image attack into the text attack, as shown in Figure 2. We iteratively attack the image and text inputs (x_t, x_i) to find the most vulnerable multimodal information patch until the story ending generation quality of the adversarial example (x'_t, x'_i) relative to the original story ending generation quality is less than a threshold λ.

For the text adversarial attack, we generate substitute words set C_{w_h} for the vulnerable word w_h obtained in Step 1 (Algorithm 1, line 7). For every substitute word c_j in C_{w_h}, we replace w_h in the original context with c_j to generate the potential adversarial text x'_t (Algorithm 1, line 10 - 11).

For the image adversarial attack, we try to perturb image information that is complementary to the perturbations in the text, and output the potential adversarial image as follows (Algorithm 1, line 12):

$$x_i^{'(a+1)} = \Pi_{x_i+S}(x_i^{'a} + \epsilon \cdot \text{sign}(\bigtriangledown_{x_i}\mathcal{L}_{adv}(\theta, (x'_t, x_i), y))) \tag{4}$$

[1] means concatenation

where a is the steps of iteration, ϵ is the step size, θ is the parameter of target models, $S \subseteq R^d$ is a set of allowed perturbations, $\mathrm{sign}(x)$ is a mathematical function that returns the sign of a real number "x". Based on potential adversarial text x_t', perturb the image x_i to find the most vulnerable multimodal information pairs (x_t', x_i') on the target model, thereby affecting the model's output that significantly differs from the ground truth label y.

When the BLEU score of the story ending generated by taking the adversarial sample (x_t', x_i') as input relative to the BLEU score of the original story ending is less than the threshold λ, which shows the multimodal adversarial attack on target models is successful as follows:

$$\frac{BLEU(\mathcal{F}((x_t^{'}, x_i{'})), y)}{BLEU(\mathcal{F}((x_t, x_i)), y)} \leq \lambda. \tag{5}$$

where $BLEU$ is a function that calculates the BLEU value [46] between the story ending outputted by the target model $\mathcal{F}$ and the ground truth label y. Equation 5 reflects the relative percentage decrease in IgSEG model performance against multimodal adversarial samples.

Otherwise, saving the loss between y_f and y, and the potential adversarial text x_t'. We continue to attack the potential adversarial text which replaces the word w_h with the next substitute word c_{j+1} (Algorithm 1, line10 - 19).

If all the substitute words in C_{wh} can not satisfy the Eq. 5, we will continue to iteratively attack the next important word w_{h+1} based on the text where the word w_h is replaced by the substitute word in C_{w_h} maximizing the adversarial loss (Algorithm 1, line 20 - 23).

IV. Experiments

In this section, we evaluate our proposed Iterative-attack method by applying it to four pre-trained IgSEG models on two datasets: VIST-E [9] and LSMDC-E [10]. For comparative analysis, three baseline methods are also employed. To assess the individual components' contribution within our method, we have conducted an ablation study. Additionally, we examine the influence of hyper-parameters on the Iterative-attack's performance by varying the number of word perturbations and the count of important words, analyzing their impact on runtime. To further analyze the effectiveness of Iterative-attack, we test our method in a multimodal machine translation dataset. Finally, we present two visualizations of multimodal adversarial samples, accompanied by an error analysis, to offer an intuitive understanding of the Iterative-attack approach.

A. Experimental Setup

Dataset. This work utilizes the VIST-E [9] and LSMDC-E [10] datasets to evaluate the task of generating image-guided story endings. The VIST-E dataset is derived from the VIST dataset and consists of 49,913 training samples, 4,963 validation samples, and 5,030 test samples. Unlike the original VIST dataset, each sample in VIST-E specifically contains the story ending, the corresponding image related to the ending, and the first four sentences that constitute the story context. To ensure uniformity, the length of each sentence is limited to a maximum of 40 words. Similarly, the LSMDC-E dataset is derived from the LSMDC 2021 dataset [47] and comprises 20,151 training samples, 1,477 validation samples, and 2,005 test samples. In LSMDC-E, the story context is constructed by selecting the first four sentences from each five-sentence story, while the last sentence is designated as the story ending. As each sentence in LSMDC-E is associated with a set of movie frames, the set corresponding to the last frame is chosen as the ending-related image set. To maintain consistency with prior work [9], a maximum sentence length of 20 words is imposed. Due to the LSMDC-E dataset is not publicly available, we constructed the dataset as described in [10]. Both of the VIST-E and LSMDC-E datasets are widely recognized benchmarks for evaluating the task of Image-guided story ending generation. They serve as effective measures for assessing a model's capacity to comprehend multimodal information and generate text.

The Multi30k [48] dataset, an expansion of the original Flickr30k [49], is a pivotal dataset in multimodal machine translation research. It comprises 30,000 images, each accompanied by textual descriptions in both English and German. The dataset consists of two primary variants: M30kT and M30KC. M30kT features each image with a single English description professionally translated into German. In contrast, M30KC provides five English and five German descriptions for each image, with the German texts sourced independently via crowdsourcing, rather than direct translations. This dataset is partitioned into training, validation, and test sets containing 29,000, 1,014, and 1,000 instances, respectively. In the attack, we present experiment results on the English-German (En-De) Test2016.

Hyper-parameters. For the perturbation on images, we apply the PGD attack [25]. The maximum perturbation ϵ is set to 4/255, and the step size is set to ϵ/10. The number of PGD iterations is set to 20. For a fair comparison, the maximum number P of the perturbed words in the text is set to 2 for all adversarial attack methods. The number of important words K is set to 10; Following the setting in [50], the threshold λ is set to 0.5. In the multimodal machine translation task, the maximum number P of the perturbed words in the text is set to 1 for all adversarial attack methods due to the texts in Multi30k dataset is shorter than texts in VIST-E and LSMDC-E datasets.

B. IgSEG Methods and Attacking Baselines

IgSEG Methods. To prove the effectiveness of the proposed multimodal adversarial attack method for IgSEG models, we select Seq2Seq [51], Transformer [52], MGCL [9], and MMT [10] as the target models. On the VIST-E dataset and LSMDC-E dataset, we reimplement four IgSEG methods based on the official open source codes or settings in the original papers, where we apply four widely-used automatic: BLEU [46], METEOR [53], CIDEr [54], and ROUGE-L [55] to evaluate the three IgSEG models. The reproducible results are reported in Table IV. Compared the actual results of running with the report results in the papers, we can observe that the results of the four IgSEG models on two datasets

Algorithm 1 Iterative-attack Algorithm

```
Input: Original sample $X = \{x_t, x_i\}$; Ground-truth ending of a story $Y$; Normal story ending $Y_p = \mathcal{F}((x_t, x_i))$; The maximum number of iterations for generating adversarial visual sample $N_{iter}$; The number of perturbing words $P$; The number of important words K; The threshold $\lambda$;
Output: A multimodal adversarial sample $X' = \{x_t', x_i'\}$;
Initialize: image_attacker : iterative multimodal adversarial attack;
for word $w_a$ in $x_t$ do
    Calculate importance score $Q_{x_a}$ according to Eq. 2
end for
create important word list $L \leftarrow [w_{top-1}, w_{top-2}, \cdots, w_{top-K}]$ according to $Q_x$
for  $w_h$ in $L$ do
    Generate substitutes set $C_{w_h}$ for word $w_h$: $C_{w_h} = S_c \oplus S_w$
    $U \leftarrow$ empty set for saving the loss between $Y_f$ and $Y$;
    $V \leftarrow$ empty set for saving the adversarial story context;
    for  $c_j$ in $C_{w_h}$ do
        $x_t' \leftarrow$ replace word $w_h$ with $c_j$
        $x_i' \leftarrow$ image_attacker $(x_t', x_i, N_{iter})$
        Generate story ending with adversarial sample $(x_t', x_i')$: $Y_f = \mathcal{F}((x_t', x_i'))$
        if $\frac{BLEU((Y_f,Y))}{BLEU(Y_p,Y)} \leq \lambda$ then
            return $(x_t', x_i')$
        else
            add $\mathcal{L}(Y_f, Y)$ to set $U$; add $x_t'$ to set $V$;
        end if
    end for
    $x_t = V[t]$, where $t$ = argmax$(U)$
    h += 1
    if $h > P$ then
        return None
    end if
end for
return  None
```

are close to the report results in the paper [9], [10], which demonstrates the IgSEG models we attack are normal. The four IgSEG models as follows:

- **Seq2Seq** [51] is an attention-based model with stacked RNNs, which utilizes two attentional mechanisms for neural machine translation, where the global attention always looks at all source positions and the local one only attends to a subset of source positions at a time. To adapt to the IgSEG task, we concatenate textual and visual features as its inputs.
- **Transformer** [52] is a parallel model based solely on the attention mechanism and is widely used in text generation tasks. To adapt to the IgSEG task, we concatenate textual and visual features as its inputs.
- **MGCL** [9] is the first method to introduce an ending-related image to explore a more informative and reasonable ending, which proposes a GCN-based text encoder and an LSTM-based decoder to build logically consistent and semantically rich story endings.
- **MMT** [10] extracts the multimodal semantic dependency for IgSEG with a multimodal transformer that can build and fuse visual and contextual information. Besides, a cross-modal attention network is used to learn cross-modal relations and fuse the fine-grained feature.

Adversarial Attack Methods. We compare our attack with Co-attack [20], which is a multimodal adversarial attack against vision-language pre-training models for classification tasks and image-text retrieval tasks in a non-iterative way; We also adapt the kNN and CharSwap in [19], a white-box untargeted attack against neural machine translation models. kNN substitutes some words with their neighbors in the embedding space; CharSwap considers swapping the characters in the target word.

Variants of Iterative-attack. The following variants of the proposed Iterative-attack are designed for comparison in the ablation experiment.

- Text-attack: the variant of Iterative-attack, which removes image adversarial attack when attacking the target models.
- Image-attack: the variant of Iterative-attack, which removes text adversarial attack when attacking the target models.
- Character-attack: the variant of Iterative-attack, which removes the word-level substitutes generation strategy for vulnerable words when attacking the target models.
- Word-attack: the variant of Iterative-attack, which removes the character-level substitutes generation strategy for vulnerable words when attacking the target models.

C. Evaluation Metrics

For evaluation, we adopt a multifaceted approach to performance metrics, similar to [50]. We assess: (1)**Attack Success Rate (ASR)**: This metric quantifies the proportion of successful adversarial examples, defined in line with [56] as those with a BLEU score for the story ending less than half that of the original. **(2) Relative Decrease in Story Ending Generation Quality (RDBLEU and RDchrF)**: We measure the degradation in story ending generation quality using BLEU score and chrF, computed respectively as RDBLEU and RDchrF [57]. We choose to compute the relative decrease in story ending generation quality so that scores are comparable across different models and datasets. **(3) Semantic Similarity (Sim.)**: Calculated between original and adversarial sentences using the universal sentence encoder [58], this metric approximates the semantic alteration introduced by the attack. **(4) Perplexity Score (Perp.)**: The fluency of adversarial examples is evaluated using the GPT-2 (large) perplexity score. The whole method is implemented in Pytorch [59], with all experiments conducted on a GeForce RTX 1080Ti GPU.

D. Quantitative Results

Table II and Table III shows the experimental results [2] of automatic metrics of attacking different IgSEG methods

[2] We discard the original samples whose BLEU score of generated story ending is zero to prevent improving the results artificially.

TABLE II
PERFORMANCE OF ADVERSARIAL ATTACK AGAINST DIFFERENT IGSEG MODELS ON VIST-E DATASET.

Dataset	Method	Attack	ASR(%) ↑	RDBLEU↑	RDchrF↑	Sim.↑	Perp.↓
VIST-E	Seq2Seq	**Iterative-attack**	**57.09**	**0.46**	**0.27**	**0.96**	**122.07**
		Co-attack	25.84	0.20	0.26	0.94	149.50
		kNN	23.25	0.19	0.18	0.94	122.28
		CharSwap	31.69	0.30	0.24	0.94	171.37
	Transformer	**Iterative-attack**	**35.04**	**0.18**	**0.06**	0.93	104.43
		Co-attack	22.80	0.12	0.05	**0.95**	159.28
		kNN	14.62	0.11	0.02	0.94	**100.82**
		CharSwap	12.63	0.08	0.02	0.91	180.56
	MGCL	**Iterative-attack**	**50.37**	**0.49**	**0.23**	**0.96**	82.95
		Co-attack	39.17	0.38	0.15	0.93	118.94
		kNN	15.45	0.19	0.14	0.93	**82.83**
		CharSwap	15.30	0.11	0.12	0.88	109.23
	MMT	**Iterative-attack**	**39.67**	**0.30**	**0.21**	**0.95**	82.54
		Co-attack	30.01	0.25	0.18	0.94	90.28
		kNN	14.89	0.12	0.08	0.93	**82.22**
		CharSwap	15.56	0.12	0.12	0.90	102.24

TABLE III
PERFORMANCE OF ADVERSARIAL ATTACK AGAINST DIFFERENT IGSEG MODELS ON LSMDC-E DATASET.

Dataset	Method	Attack	ASR(%)↑	RDBLEU↑	RDchrF↑	Sim.↑	Perp. ↓
LSMDC-E	Seq2Seq	**Iterative-attack**	**57.14**	**0.53**	**0.29**	**0.96**	**126.77**
		Co-attack	23.72	0.20	0.15	0.94	179.82
		kNN	3.40	0.03	0.03	0.96	243.42
		CharSwap	12.04	0.09	0.07	0.93	303.33
	Transformer	**Iterative-attack**	**31.72**	**0.28**	**0.20**	**0.96**	**176.30**
		Co-attack	20.75	0.12	0.06	0.95	189.45
		kNN	7.02	0.01	0.01	0.93	190.42
		CharSwap	6.04	0.01	0.01	0.93	205.33
	MGCL	**Iterative-attack**	**47.18**	**0.42**	**0.22**	**0.96**	176.31
		Co-attack	25.76	0.20	0.18	0.93	**126.98**
		kNN	21.08	0.22	0.12	0.75	167.69
		CharSwap	19.76	0.19	0.08	0.85	213.72
	MMT	**Iterative-attack**	**52.34**	**0.44**	**0.21**	**0.96**	151.75
		Co-attack	36.92	0.33	0.21	0.94	199.73
		kNN	19.04	0.17	0.10	0.90	**142.08**
		CharSwap	20.24	0.21	0.12	0.87	187.99

on VIST-E and LSMDC-E datasets. From the Table II and Table III we can draw the following main observations:

(1) Overall, The proposed Iterative-attack can decrease the BLEU score of the target model to more than 30% of the BLEU score of the original story ending for more than 39% of the stories for MMT, MGCL, and Seq2Seq on two datasets (except for the Transformer model, where the ASR is more than 30% on two datasets). Also, in all cases, semantic similarity is more than 0.95, which shows that the Iterative-attack can maintain a high level of semantic similarity with the original context.

(2) When compared to baselines like kNN and CharSwap, Iertaive-attack exhibits a superior attack success rate across different IgSEG architectures, more significantly degrading the quality of generated story endind. The lower performance of kNN and CharSwap in terms of ASR, RDBLEU, and RDchrF suggests that text-only adversarial attacks are insufficient for disrupting multimodal text generation models, highlighting the importance of attacking both modalities.

(3) Co-attack, while outperforming single-modal attack methods in multimodal text generation tasks, falls short when compared to the Iterative-attack. The reason can be attributed to its iterative process that integrates image modality attacks into text modality attacks, enhancing its disruptive capability.

(4) Multimodal attack methods (Iterative-attack and Co-attack) demonstrate superior performance over single-modal methods (kNN and CharSwap). This confirms that for multimodal text generation models, reliance on a single modality for perturbation is often ineffective due to the models' ability to compensate with complementary information from the other modality.

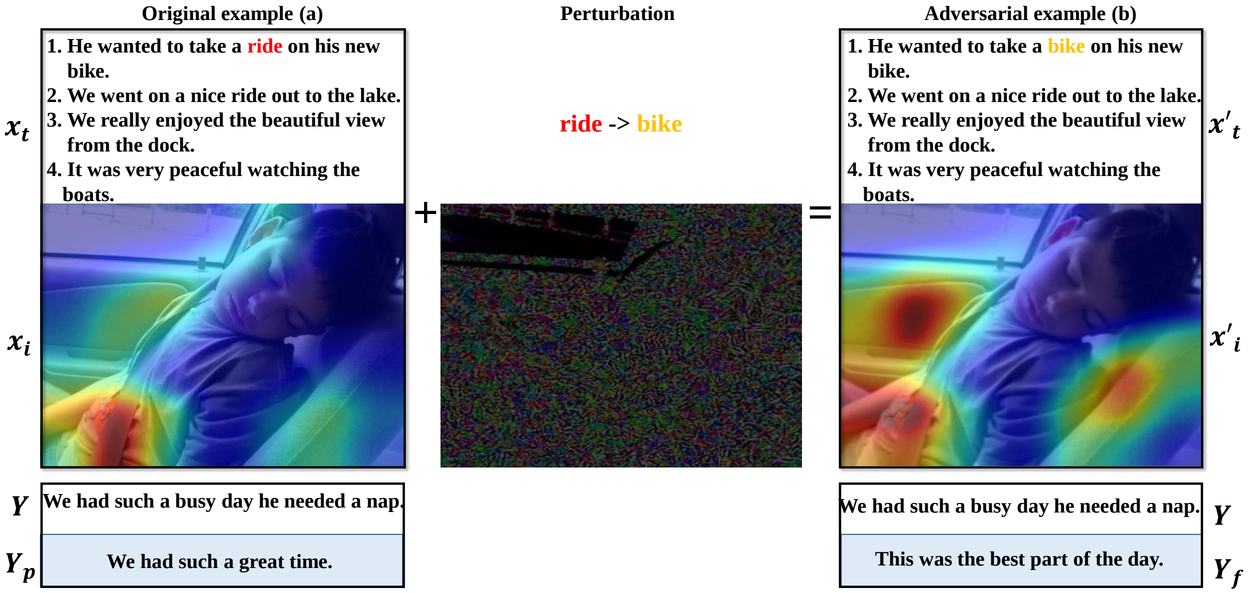
(5) The adversarial robustness of MGCL has worse performance than that of MMT on VIST-E dataset, where the reason may be that the LSTM-based decoder in MGCL is more sensitive than the transformer decoder in MMT.

(6) In summary, these results collectively indicate that Iterative-attack effectively exploits the vulnerabilities of IgSEG models, significantly degrading the quality of generated text while preserving semantic similarity of adversarial texts. The comparative analysis of attack methods highlights the necessity of targeting both text and image modalities to effectively compromise multimodal text generation systems.

TABLE IV
THE PERFORMANCE OF THE IGSEG MODELS ON THE VIST-E DATASET AND LSMDC-E DATASET.

Dataset	Method	B1	B2	B3	B4	M	C	R-L
VIST-E	Seq2Seq#	13.96	5.57	2.94	1.69	4.54	12.04	16.84
	Seq2Seq*	14.35	6.11	3.89	1.45	8.15	10.01	11.95
	Transformer#	17.18	6.29	3.07	2.01	6.91	12.75	18.23
	transformer*	18.26	5.76	4.02	1.69	11.80	12.31	13.44
	MGCL#	22.57	8.16	4.23	2.49	7.84	21.46	21.66
	MGCL*	22.36	7.94	5.55	2.33	14.30	18.96	19.32
	MMT#	22.87	8.68	4.38	2.61	15.55	25.41	23.61
	MMT*	22.65	8.64	4.41	2.53	14.93	23.17	22.12
LSMDC-E	Seq2Seq#	14.21	4.56	1.70	0.70	11.01	8.69	19.69
	Seq2Seq*	13.53	3.44	1.49	0.50	8.83	5.49	16.51
	Transformer#	15.35	4.49	1.82	0.76	11.43	9.32	19.16
	transformer*	14.11	3.71	2.21	0.65	8.88	7.09	18.94
	MGCL#	15.89	4.76	1.57	0.00	11.61	9.16	20.30
	MGCL*	14.60	3.75	1.61	0.00	9.20	6.79	17.75
	MMT#	18.52	5.99	2.51	1.13	12.87	12.41	20.99
	MMT*	16.85	5.58	2.10	0.96	11.07	13.05	18.75

* indicates the actual results of running. # indicates the results in the paper.



1. He wanted to take a ride on his new bike.
2. We went on a nice ride out to the lake.
3. We really enjoyed the beautiful view from the dock.
4. It was very peaceful watching the boats.

We had such a busy day he needed a nap.

We had such a great time.

1. He wanted to take a bike on his new bike.
2. We went on a nice ride out to the lake.
3. We really enjoyed the beautiful view from the dock.
4. It was very peaceful watching the boats.

We had such a busy day he needed a nap.

This was the best part of the day.

Fig. 3. The Grad-CAM visualizations of (a) the original example (x_t, x_i), (b) the adversarial example (x'_t, x'_i) derived by Iterative-attack against MMT on VIST-E dataset where the adversarial perturbation is obtained by $x'_i - x_i$ (pixel values of perturbation are amplified ×20 for visualization).

E. Ablation Study

In this section, we compare the variants of Iterative-attack against MGCL on VIST-E and LSMDC-E datasets as shown in Table V and Table VI, respectively. From the Table V and Table VI, we can have the following observations:

(1) Perturbing multimodal input iteratively (Iteratively-attack) is consistently stronger than perturbing any single-model input (Text-attack and Image-attack), which demonstrates that the adversarial samples generated by multimodal adversarial attacks are more dangerous than those generated by single-moda adversarial attacks.

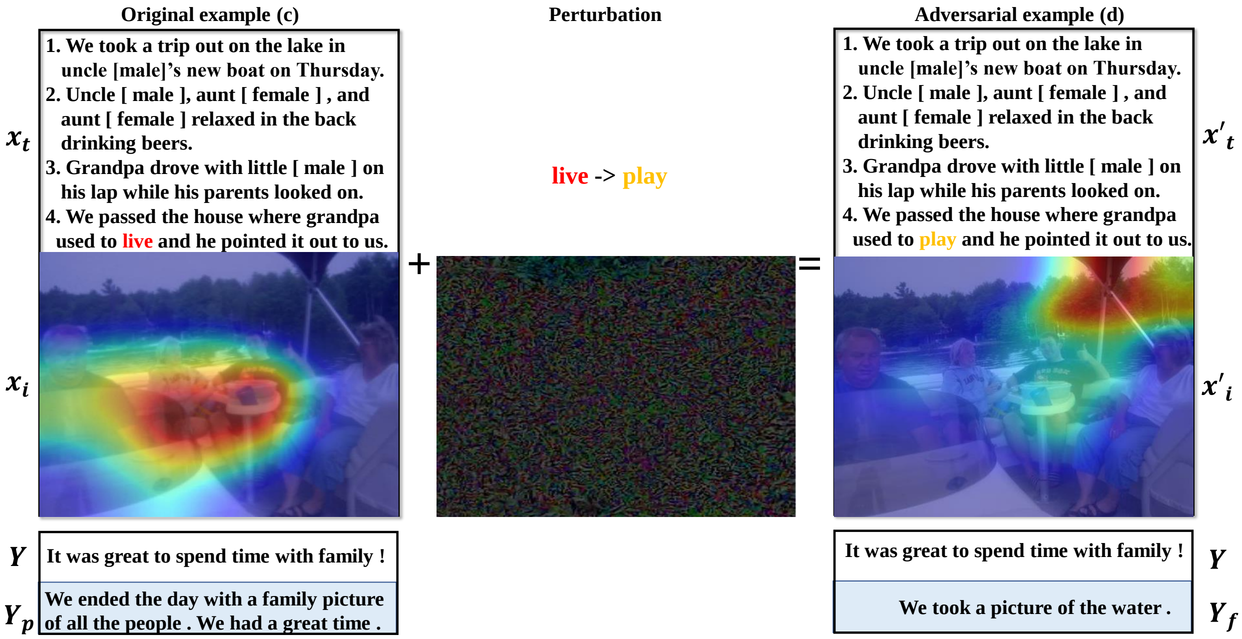
(2) The performance of Text-attack is better than that of Image-attack on two datasets, however, without the image adversarial attack perturbing the critical visual information, Text-attack performers worse than Iterative-attack, which suggests that due to the complementarity between multimodal data, the information shift caused by single-modal adversarial attack can be corrected by another modality data, thus leading to attack failure.

(3) Iterative-attack outperforms Character-attack and Word-attack, which shows that generating substitutes in character-level perturbation and word-replacement via BERT in word-level perturbation together have a greater impact than single-level text perturbation.

F. Visualization

To further understand Iterative-attack intuitively, we provide the Grad-CAM[3] [60] visualizations for Iterative-attack against MMT on the VIST-E dataset. The tokens modified by Iterative-attack are written in red in the original story contexts, and the changes in the images are shown by the heat map.

[3] https://github.com/frgfm/torch-cam



Original example (c):

1. We took a trip out on the lake in uncle [male]'s new boat on Thursday.
2. Uncle [male], aunt [female] , and aunt [female] relaxed in the back drinking beers.
3. Grandpa drove with little [male] on his lap while his parents looked on.
4. We passed the house where grandpa used to live and he pointed it out to us.

It was great to spend time with family !

We ended the day with a family picture of all the people . We had a great time .

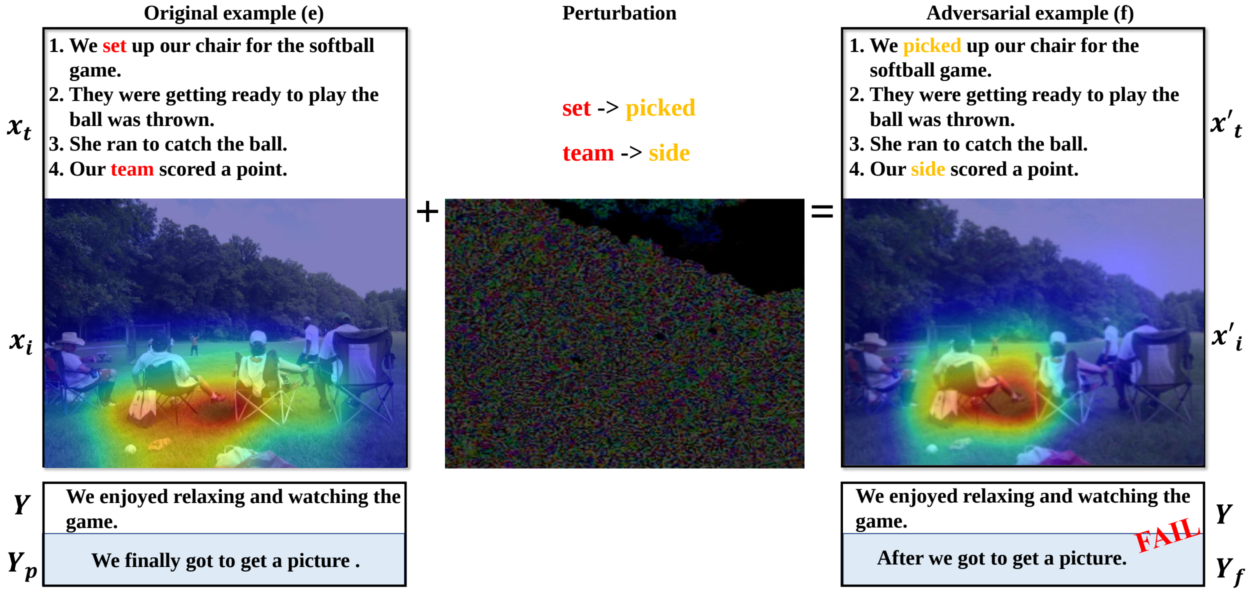
Adversarial example (d):

1. We took a trip out on the lake in uncle [male]'s new boat on Thursday.
2. Uncle [male], aunt [female] , and aunt [female] relaxed in the back drinking beers.
3. Grandpa drove with little [male] on his lap while his parents looked on.
4. We passed the house where grandpa used to play and he pointed it out to us.

It was great to spend time with family !

We took a picture of the water .

Fig. 4. The Grad-CAM visualizations of (c) the original example (x_t, x_i), (d) the adversarial example (x'_t, x'_i) derived by Iterative-attack against MMT on VIST-E dataset, where the adversarial perturbation is obtained by $x'_i - x_i$ (pixel values of perturbation are amplified ×20 for visualization).



Original example (e):

1. We set up our chair for the softball game.
2. They were getting ready to play the ball was thrown.
3. She ran to catch the ball.
4. Our team scored a point.

We enjoyed relaxing and watching the game.

We finally got to get a picture .

Adversarial example (f):

1. We picked up our chair for the softball game.
2. They were getting ready to play the ball was thrown.
3. She ran to catch the ball.
4. Our side scored a point.

We enjoyed relaxing and watching the game.

After we got to get a picture.

Fig. 5. The Grad-CAM visualizations of (e) the original example (x_t, x_i), (f) the adversarial example (x'_t, x'_i) derived by Iterative-attack against MMT on VIST-E dataset where the adversarial perturbation is obtained by $x'_i - x_i$ (pixel values of perturbation are amplified ×20 for visualization).

Compared the original image with an adversarial image in Fig. 3, Fig. 4, we can observe that the word in the adversarial samples has only slightly changed with semantics preservation, but the focus of the visual extractor in the target model on the adversarial images has changed significantly, although these perturbations are imperceptible to humans, which strongly suggests the perturbation added to the original image successfully mislead the attention of the target model. Meanwhile, the perturbed percentage for the adversarial context is also low, which can result in more semantic consistency.

To further analyze the proposed method, we added the error analysis. When analyzing instances of failed Iterative-attack attempts, a common observation is that these failures often stem from a fundamental issue: the target model's inability to comprehend the input image effectively. In other words, the ending of the story generated by the model lacks a robust semantic connection with the associated image. In Fig. 5, we provide a side-by-side comparison of the original sample and the adversarial sample. Notably, we observe that two words have been altered in the original story context, yet the focus of the visual extractor within the target model remains almost virtually unchanged when examining the adversarial images.

TABLE V
RESULTS OF COMPARISON WITH DIFFERENT MODALITY ATTACKS IN OUR METHOD AGAINST MGCL ON VIST-E AND LSMDC-E DATASET.

Dataset	Attack	ASR(%)↑	RDBLEU↑	RDchrF↑	Sim.↑	Perp.↓
VIST-E	Iterative-attack	50.37	0.49	0.23	0.96	82.95
	Text-attack	38.46	0.34	0.12	0.95	94.94
	Image-attack	22.45	0.24	0.10	1.00	79.33
LSMDC-E	Iterative-attack	47.18	0.42	0.22	0.96	176.31
	Text-attack	45.59	0.58	0.23	0.96	177.24
	Image-attack	15.06	0.16	0.05	1.00	128.17

TABLE VI
RESULTS OF COMPARISON BETWEEN DIFFERENT LEVEL TEXT PERTURBATIONS IN OUR METHOD AGAINST MGCL ON VIST-E AND LSMDC-E DATASET.

Dataset	Attack	ASR(%)↑	RDBLEU↑	RDchrF↑	Sim.↑	Perp.↓
VIST-E	Iterative-attack	50.37	0.49	0.23	0.96	82.95
	Character-attack	48.00	0.40	0.20	0.96	88.75
	Word-Attack	46.32	0.44	0.22	0.97	85.54
LSMDC-E	Iterative-attack	47.18	0.42	0.22	0.96	176.31
	Character-attack	45.21	0.57	0.20	0.96	180.67
	Word-attack	46.83	0.58	0.21	0.97	186.09

TABLE VII
THE RESULTS OF USING DIFFERENT WORDS PERTURBATION NUMBER P IN THE ATTACKING PROCESSING.

P	ASR(%)↑	Sim. ↑	Perp. ↓	Runtime↓
1	54.00	0.98	75.65	59.88
2	66.00	0.96	87.73	75.39
3	71.00	0.95	93.11	77.58
4	80.00	0.95	97.36	100.26
5	82.00	0.93	105.77	107.40

Furthermore, when comparing the story ending generated by the model to the ground truth label, it becomes evident that there is a distinct lack of correlation between the story ending and the associated images. This holds true both when the model is provided with clean samples and when it encounters adversarial samples. These findings collectively point to a deficiency in the target model's capacity to effectively comprehend multimodal information, which stands as a pivotal factor constraining the efficacy of the Iterative-attack strategy.

G. Runtime Comparison

Since kNN and CharSwap are text modality adversarial attack methods, for a fair comparison, we show the average runtime of Iterative-attack with different important word numbers K and Co-attack when attacking on MGCL in the first 100 stories of VIST-E dataset to generate an adversarial sample in Fig. 6. We can observe that the runtime of Iterative-attack is slightly higher than Co-attack, yet in the same order of magnitude. What's more, when the runtime is close between Iteration-attack and Co-attack, the ASR score of Iteration-attack is not worse than Co-attack.

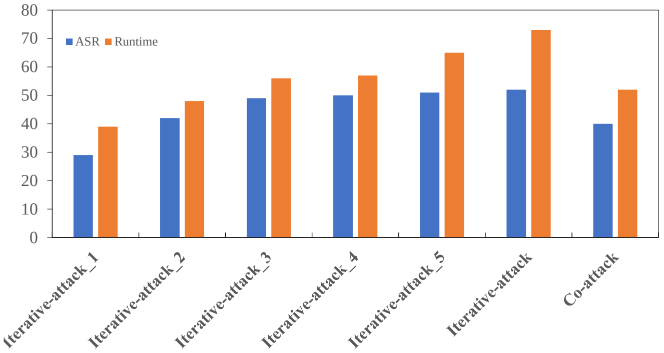


Fig. 6. The runtime comparison between Iterative-attack with different important word numbers K and Co-attack when attacking MGCL on VIST-E dataset. Iteration-attack_X means the selected important word number in Iteration-attack is X. ASR means the attack success rate. Runtime is in seconds.

H. Effect on The Number of Perturbed Words in Text

To verify the effect of using different words perturbation number P in Iterative-attack, we use the first 100 stories and vary the number of perturbing words P from 1 to 5 to attack MGCL on VIST-E dataset, as shown in Table VII. We can find that the ASR rises as P increases, but the similarity and fluency of the generated adversarial text are deteriorating, besides, the runtime of attacking is also gradually increasing, which indicates that we need to make a trade-off between the attack success rate and the number of perturbed words in the adversarial attack.

I. Multimodal Adversarial Attack against Multimodal Machine Translation

To further analyze the effectiveness of Itertaive-attack on multimodal text generation tasks, we test our proposed method with a multimodal machine translation dataset.

1) Multimodal Machine Translation Models: MTMMT [61] is a multimodal machine translation model that uses multimodal self-attention in Transformer to avoid encoding irrelevant information in images. To further investigate the performance of MTMMT on more data, the authors also train the MTMMT with the additional training

TABLE VIII
PERFORMANCE OF ADVERSARIAL ATTACKS AGAINST DIFFERENT MULTIMODAL MACHINE TRANSLATION MODELS ON MULTI30K DATASET.

Dataset	Method	Attack	ASR(%)↑	RDBLEU↑	RDchrF↑	Sim.↑	Perp.↓
Multi30K	MTMMT	**Iterative-attack**	**44.71**	**0.40**	**0.20**	0.94	428.11
		Co-attack	20.89	0.26	0.15	0.94	**407.38**
		kNN	19.90	0.25	0.14	**0.96**	479.16
		CharSwap	20.10	0.25	0.13	**0.96**	425.16
	MTMMT$_{back}$	**Iterative-attack**	**45.40**	**0.41**	**0.21**	0.94	442.34
		Co-attack	23.50	0.29	0.16	0.94	**410.09**
		kNN	19.30	0.22	0.12	**0.97**	490.53
		CharSwap	19.30	0.22	0.12	0.96	490.99

TABLE IX
RESULTS OF ABLATION EXPERIMENTS FOR THE MULTIMODAL MACHINE TRANSLATION MODELS ON MULTI30K DATASET.

Dataset	Model	Attack	ASR(%)	RDBLEU	RDchrF	Sim.	Perp.
Multi30K	MTMMT	Iterative-attack	44.71	0.40	0.20	0.94	428.11
		Text-attack	43.56	0.39	0.20	0.93	403.27
		Image-attack	0.99	0.01	0.01	1.00	180.36
		Character-attack	37.19	0.37	0.19	0.93	433.62
		Word-attack	35.37	0.33	0.17	0.95	387.55
	MTMMT$_{back}$	Iterative-attack	45.40	0.41	0.21	0.94	442.34
		Text-attack	44.80	0.39	0.20	0.94	418.65
		Image-attack	1.48	0.02	0.02	1.00	180.36
		Character-attack	37.90	0.36	0.19	0.93	470.85
		Word-attack	39.02	0.36	0.19	0.95	390.34

TABLE X
PERFORMANCE OF THE MULTIMODAL MACHINE TRANSLATION MODELS ON MULTI30K DATASET.

Dataset	Method	B1	B2	B3	B4	Meteor
Multi30K	MTMMT#	-	-	-	38.70	55.70
	MTMMT*	67.6	54.91	45.81	38.65	55.06
	MTMMT$_{back}$#	-	-	-	39.50	56.90
	MTMMT$_{back}$*	68.2	55.26	47.15	39.48	56.88

* indicates the actual results of running. # indicates the results in the paper.

data where the authors use a back-translation model [62] to translate 145k monolingual German description in M30kC into English, and the model refers to MTMMT$_{back}$. The reproducible results are reported in Table X

2) Quantitative Results: Table VIII and Table IX presents the results of attacks on two multimodal machine translation models within the Multi30K dataset. The following observations are noteworthy:

(1) Comparative Efficacy of Iterative-Attack: Our Iterative-attack method outperforms other baseline approaches in terms of attack success rate and the extent to which translation quality is reduced. This is evident from the close semantic similarity and perplexity scores. Such results underscore the superior performance of our proposed attack method in both the IgSEG task and the multimodal machine translation task.

(2) Compared the performance of multimodal attack methods (Iterative-attack and Co-attack) with single-modal attack methods (kNN and CharSwap), it is obvious that multimodal attack methods outperforms the single-modal attack methods, which proves our observation that for the multimodal text generation models, the single-modal perturbation tends to fail, due to the complementary information between text data and image

(3) Effectiveness of Multimodal Attacks in Text Generation: In multimodal text generation tasks, the integration of textual and visual data plays a significant role in producing coherent texts. Our findings suggest that combined attacks on both texts and images are more effective than attacks on a single modality.

These insights reflect a comprehensive understanding of the vulnerabilities and characteristics of multimodal machine translation models, providing a foundation for further research in this domain.

3) Ablation Study: In this section, we compare the variants of Iterative-attack against MTMMT and MTMMT$_{back}$ on Multi30K dataset as shown in Table IX. From the Table IX, we can have the following observations:

(1) Perturbing multimodal input iteratively (Iteratively-attack) is consistently stronger than perturbing any single-model input (Text-attack and Image-attack), which demonstrates that the adversarial samples generated by multimodal adversarial attacks are more dangerous than those generated by single-moda adversarial attacks in multimodal machine translation.

(2) The Iterative-attack, which consists of both character-level and word-level perturbations, significantly outperforms either Character-attack or Word-attack alone. This outcome highlights the amplified impact achieved by integrating character-level substitutes and BERT-based word-replacement strategies, as opposed to focusing on a single level of text perturbation in multimodal text generation tasks.

(3) In multimodal machine translation tasks, Text-attack performs better than Image-attack. This is attributed to the distinct role of images in this domain, primarily serving to enhance contextual understanding and translation accu-

racy. Unlike in IgSEG tasks, image perturbations in multimodal machine translation frequently fail to significantly influence the output of the target models, reflecting the limited impact of visual modifications on the translation process.

J. Discussion

To further understand why multimodal adversarial examples outperform single-modal adversarial samples, what makes the proposed method fail, and the practical implications of our research, we discuss some key concepts of adversarial attacks in this paper that encompass both empirical and theoretical elements. In detail:

- For a multimodal model, perturbing bi-modal inputs is stronger than perturbing any single-modal input [20]. In the experiment of adversarial attack on the multimodal text generation models, the single-modal perturbation tends to fail, due to the complementary information between text data and image. The multimodal adversarial attack can find the most vulnerable multimodal adversarial patches to avoid the dilemma that the information shift caused by a single-modal adversarial attack may be corrected by another modality's information.
- A robust correlation exists between the effectiveness of the Iterative-attack and the target model's capacity to interpret multimodal information holistically. In our observations, we have a conclusion that if the ending of a story produced by the target model exhibits no meaningful connection to the accompanying image, perturbations applied to the image do not exert a discernible influence on the model's output, which suggests that the target model's ability to understand multimodal information is important for the success of Iterative-attack.
- Multimodal text generation represents a pivotal challenge in artificial intelligence, requiring the integration and interpretation of diverse information sources to produce coherent textual outputs. Key applications of multimodal text generation include multimodal machine translation [3], [4], multimodal dialogue response generation [5], [6], multimodal question answering [7], multimodal MemexQA [8], and image-guided story ending generation [9], [10]. These diverse applications underscore the versatility and complexity of multimodal text generation, highlighting its significance in advancing the field of artificial intelligence. Our study contributes significantly to the unexplored domain of adversarial robustness in multimodal text generation systems. By investigating the adversarial robustness of multimodal text generation models to multimodal adversarial attacks, this research accomplishes two primary objectives: firstly, it elucidates the internal mechanisms of these complex models; secondly, it aids in the development of more robust and reliable multimodal systems.

K. Insights and important conclusions

we present critical observations and conclusions drawn from adversarial attacks against image-guided story ending generation (IgSEG) and multimodal machine translation tasks. The key conclusions are as follows:

- **Effectiveness of Iterative-Attack Methodology**: Our analysis demonstrates that Iterative-attack consistently outperforms other attack methods (such as Co-attack, kNN, and CharSwap) in achieving a higher attack success rate (ASR) across different datasets and models. This suggests that the Iterative-attack is more effective at finding and exploiting vulnerabilities in IgSEG models and multimodal machine translation models.
- **Impact of Attack Multi-modality**: In the context of multimodal text generation, the effectiveness of a single-modal attack is likely to be affected by the complementary information in texts and images. Our proposed Iterative-attack, which iteratively attacks text and image modalities, consistently shows higher ASR compared to image adversarial attacks and text adversarial attacks alone. Additionally, adversarial attacks on classification tasks seek to induce incorrect labels, exploiting vulnerabilities around the decision boundary of the classifier. In contrast, adversarial attacks on multimodal text generation tasks aim to disrupt the generation process, leading to outputs that may be grammatically correct but are contextually inappropriate or misleading. The latter is more difficult due to the complexity and variability of language generation.
- **Vulnerability of Multimodal Text Generation Models**: The varying performance of different IgSEG models (such as Seq2Seq, Transformer, MGCL, MMT) and multimodal machine translation models (like MTMMT) to various attack strategies suggests a fundamental vulnerability. Current methodologies for integrating and processing multimodal inputs (text and images) for text generation are not sufficiently robust against adversarial attacks. This indicates a need for more advanced models capable of effectively fusing multimodal information, potentially enhancing resilience to such attacks.

V. Conclusion

In this paper, we first propose an iterative multimodal adversarial attack against IgSEG models in multimodal text generation tasks. Compared with the single-modal adversarial attack methods, our method fuses the image modality attack into the text modality attack to iteratively find the most vulnerable multimodal information patch. The generated multimodal adversarial samples can avoid the dilemma that the information shift caused by a single-modal adversarial attack may be corrected by another modality's information. Our experimental results show that Iterative-attack is highly effective against multimodal text generation tasks.

Evaluating the adversarial robustness within the realm of multimodal text generation assumes paramount importance when considering the pragmatic deployment of multimodal models. By employing iterative attacks on the target model, we can systematically identify and comprehend the inherent vulnerabilities within the multimodal architecture. This, in turn, motivates us to institute comprehensive measures aimed at fortifying its robustness. In the future, we plan to establish a comprehensive and rigorous benchmark to evaluate adversarial robustness on multimodal text generation for different applications and investigate methods for defending against multimodal adversarial attacks. Besides, we plan to establish a comprehensive and rigorous benchmark to evaluate adversarial robustness on multimodal text generation for different applications and investigate a method for defending against multimodal adversarial attacks.

We hope that this study will draw attention to the adversarial robustness of multimodal text generation tasks.

REFERENCES

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[2] 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.

[3] Y. Su, K. Fan, N. Bach, C.-C. J. Kuo, and F. Huang, "Unsupervised multi-modal neural machine translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 482–10 491.

[4] P. Liu, H. Cao, and T. Zhao, "Gumbel-attention for multi-modal machine translation," *arXiv preprint arXiv:2103.08862*, 2021.

[5] Q. Sun, Y. Wang, C. Xu, K. Zheng, Y. Yang, H. Hu, F. Xu, J. Zhang, X. Geng, and D. Jiang, "Multimodal dialogue response generation," *arXiv preprint arXiv:2110.08515*, 2021.

[6] S. Wang, Y. Meng, X. Sun, F. Wu, R. Ouyang, R. Yan, T. Zhang, and J. Li, "Modeling text-visual mutual dependency for multi-modal dialog generation," *arXiv preprint arXiv:2105.14445*, 2021.

[7] H. Singh, A. Nasery, D. Mehta, A. Agarwal, J. Lamba, and B. V. Srinivasan, "Mimoqa: Multimodal input multimodal output question answering," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 5317–5332.

[8] J. Liang, L. Jiang, L. Cao, Y. Kalantidis, J. Li, and A. Hauptmann, "Focal visual-text attention for memex question answering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, preprint. [Online]. Available: https://ieeexplore.ieee.org/document/8603827

[9] Q. Huang, C. Huang, L. Mo, J. Wei, Y. Cai, H.-f. Leung, and Q. Li, "Igseg: Image-guided story ending generation," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 3114–3123.

[10] D. Xue, S. Qian, Q. Fang, and C. Xu, "Mmt: Image-guided story ending generation with multimodal memory transformer," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 750–758.

[11] 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.

[12] L. Gao, Z. Huang, J. Song, Y. Yang, and H. T. Shen, "Push & pull: Transferable adversarial examples with attentive attack," *IEEE Transactions on Multimedia*, vol. 24, pp. 2329–2338, 2021.

[13] 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.

[14] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi, "Black-box generation of adversarial text sequences to evade deep learning classifiers," in *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018, pp. 50–56.

[15] J. Chen, C. Wang, K. Wang, C. Yin, C. Zhao, T. Xu, X. Zhang, Z. Huang, M. Liu, and T. Yang, "Heu emotion: a large-scale database for multimodal emotion recognition in the wild," *Neural Computing and Applications*, vol. 33, no. 14, pp. 8669–8685, 2021.

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

[17] 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.

[18] M. Cheng, J. Yi, P.-Y. Chen, H. Zhang, and C.-J. Hsieh, "Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 3601–3608.

[19] P. Michel, X. Li, G. Neubig, and J. Pino, "On evaluation of adversarial perturbations for sequence-to-sequence models," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 3103–3114.

[20] 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.

[21] S. Liu, H. Fan, S. Qian, Y. Chen, W. Ding, and Z. Wang, "Hit: Hierarchical transformer with momentum contrast for video-text retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 915–11 925.

[22] Y. Wang, S. Qian, J. Hu, Q. Fang, and C. Xu, "Fake news detection via knowledge-driven multimodal graph convolutional networks," in *Proceedings of the 2020 International Conference on Multimedia Retrieval*, 2020, pp. 540–547.

[23] Z. Zhang, X. Wang, G. Lu, F. Shen, and L. Zhu, "Targeted attack of deep hashing via prototype-supervised adversarial networks," *IEEE Transactions on Multimedia*, vol. 24, pp. 3392–3404, 2021.

[24] 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.

[25] 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.

[26] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is bert really robust? natural language attack on text classification and entailment," *arXiv preprint arXiv:1907.11932*, vol. 2, 2019.

[27] 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.

[28] E. Wallace, M. Stern, and D. Song, "Imitation attacks and defenses for black-box machine translation systems," *arXiv preprint arXiv:2004.15015*, 2020.

[29] X. Du and C.-M. Pun, "Robust audio patch attacks using physical sample simulation and adversarial patch noise generation," *IEEE Transactions on Multimedia*, vol. 24, pp. 4381–4393, 2021.

[30] H. Yuan, Q. Chu, F. Zhu, R. Zhao, B. Liu, and N. Yu, "Automa: Towards automatic model augmentation for transferable adversarial attacks," *IEEE Transactions on Multimedia*, vol. 25, pp. 203–213, 2023.

[31] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[32] Y. Wang, J. Liu, X. Chang, R. J. Rodríguez, and J. Wang, "Di-aa: An interpretable white-box attack for fooling deep neural networks," *Information Sciences*, vol. 610, pp. 14–32, 2022.

[33] J. Shen and N. Robertson, "Bbas: Towards large scale effective ensemble adversarial attacks against deep neural network learning," *Information Sciences*, vol. 569, pp. 469–478, 2021.

[34] 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.

[35] 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.

[36] S. Lee, D. B. Lee, and S. J. Hwang, "Contrastive learning with adversarial perturbations for conditional text generation," *arXiv preprint arXiv:2012.07280*, 2020.

[37] G. Boateng, "Towards real-time multimodal emotion recognition among couples," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 748–753.

[38] Y. Long, P. Tang, H. Wang, and J. Yu, "Improving reasoning with contrastive visual information for visual question answering," *Electronics Letters*, vol. 57, no. 20, pp. 758–760, 2021.

[39] 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.

[40] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4971–4980.

[41] M. Shah, X. Chen, M. Rohrbach, and D. Parikh, "Cycle-consistency for robust visual question answering," in *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 2019, pp. 6649–6658.

[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] Z. Zhou, S. Hu, M. Li, H. Zhang, Y. Zhang, and H. Jin, "Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 6311–6320.

[44] T. Wang, L. Zhu, Z. Zhang, H. Zhang, and J. Han, "Targeted adversarial attack against deep cross-modal hashing retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[45] L. Zhu, T. Wang, J. Li, Z. Zhang, J. Shen, and X. Wang, "Efficient query-based black-box attack against cross-modal hashing retrieval," *ACM Transactions on Information Systems*, vol. 41, no. 3, pp. 1–25, 2023.

[46] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[47] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3202–3212.

[48] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30k: Multilingual english-german image descriptions," *arXiv preprint arXiv:1605.00459*, 2016.

[49] 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.

[50] S. Sadrizadeh, L. Dolamic, and P. Frossard, "Transfool: An adversarial attack against neural machine translation models," *arXiv preprint arXiv:2302.00944*, 2023.

[51] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[53] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[54] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[55] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[56] J. Ebrahimi, D. Lowd, and D. Dou, "On adversarial examples for character-level neural machine translation," *arXiv preprint arXiv:1806.09030*, 2018.

[57] M. Popović, "chrf: character n-gram f-score for automatic mt evaluation," in *Proceedings of the tenth workshop on statistical machine translation*, 2015, pp. 392–395.

[58] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung *et al.*, "Multilingual universal sentence encoder for semantic retrieval," *arXiv preprint arXiv:1907.04307*, 2019.

[59] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[60] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[61] S. Yao and X. Wan, "Multimodal transformer for multimodal machine translation," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 4346–4350.

[62] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," *arXiv preprint arXiv:1511.06709*, 2015.

Youze Wang received a B.S. and master's degree from the School of Computer Science and Information Engineering at Hefei University of Technology, Hefei, China, where he is currently working toward his Ph.D. degree. His research interests include multimodal computing and multimodal adversarial robustness in machine learning.

Wenbo Hu is an associate professor in Hefei University of Technology. He received a Ph.D. degree from Tsinghua University in 2018. His research interests lie in machine learning, especially probabilistic machine learning and uncertainty, generative AI, and AI security. He has published more than 20 peer-reviewed papers in prestigious conferences and journals, including NeurIPS, KDD, IJCAI, etc.

Richang Hong (Member, IEEE) received a Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2008. He was a Research Fellow of the School of Computing at the National University of Singapore, from 2008 to 2010. He is currently a Professor at the Hefei University of Technology, Hefei. He is also with the Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education. He has coauthored over 100 publications in the areas of his research interests, which include multimedia content analysis and social media. He is a member of the ACM and the Executive Committee Member of the ACM SIGMM China Chapter. He was a recipient of the Best Paper Award from the ACM Multimedia 2010, the Best Paper Award from the ACM ICMR 2015, and the Honorable Mention of the IEEE Transactions on Multimedia Best Paper Award. He has served as the Technical Program Chair of the MMM 2016, ICIMCS 2017, and PCM 2018. Currently, he is an Associate Editor of IEEE Transactions on Big Data, IEEE Transactions on Computational Social Systems, ACM Transactions on Multimedia Computing Communications and Applications, Information Sciences (Elsevier), Neural Processing Letter (Springer) and Signal Processing (Elsevier).