Ask, Attend, Attack: An Effective Decision-Based Black-Box Targeted Attack for Image-to-Text Models

Qingyuan Zeng[1], Zhenzhong Wang[2], Yiu-ming Cheung[3], and Min Jiang[*,4]

[1]Institute of Artificial Intelligence, Xiamen University
[2]Department of Computing, The Hong Kong Polytechnic University
[3]Department of Computer Science, Hong Kong Baptist University
[4]School of Informatics, Xiamen University

Abstract

While image-to-text models have demonstrated significant advancements in various vision-language tasks, they remain susceptible to adversarial attacks. Existing white-box attacks on image-to-text models require access to the architecture, gradients, and parameters of the target model, resulting in low practicality. Although the recently proposed gray-box attacks have improved practicality, they suffer from semantic loss during the training process, which limits their targeted attack performance. To advance adversarial attacks of image-to-text models, this paper focuses on a challenging scenario: decision-based black-box targeted attacks where the attackers only have access to the final output text and aim to perform targeted attacks. Specifically, we formulate the decision-based black-box targeted attack as a large-scale optimization problem. To efficiently solve the optimization problem, a three-stage process *Ask, Attend, Attack*, called *AAA*, is proposed to coordinate with the solver. *Ask* guides attackers to create target texts that satisfy the specific semantics. *Attend* identifies the crucial regions of the image for attacking, thus reducing the search space for the subsequent *Attack*. *Attack* uses an evolutionary algorithm to attack the crucial regions, where the attacks are semantically related to the target texts of *Ask*, thus achieving targeted attacks without semantic loss. Experimental results on transformer-based and CNN+RNN-based image-to-text models confirmed the effectiveness of our proposed *AAA*.

1 Introduction

Image-to-text models, referring to generating descriptive and accurate textual descriptions of images, have received increasing attention in various applications, including image-captioning [1, 2], visual-question-answering [3, 4], and image-retrieval [5, 6]. Despite the remarkable progress, they are vulnerable to deliberate attacks, giving rise to concerns about the reliability and trustworthiness of these models in real-world scenarios. For example, one may mislead models to output harmful content such as political slogans and hate speech by making imperceptible perturbations to images [7, 8, 9, 10, 11].

To gain insight into the reliability and trustworthiness of the image-to-text models, a series of adversarial attack methods have been proposed to poison the outputted textual descriptions of given

---

* The corresponding author: Min Jiang, minjiang@xmu.edu.cn

Min Jiang and Qingyuan Zeng are with the Department of Artificial Intelligence, Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, School of Informatics, Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

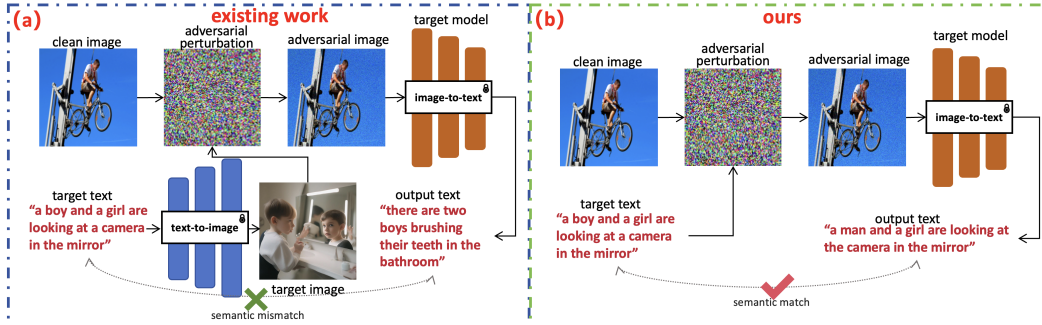



Figure 1: The semantic loss problem is existing in existing gray-box targeted attack methods.

images [7, 12, 13, 10]. Specifically, based on the attacker's level of access to information about the target model, they can be divided into three categories: white-box attacks [7, 14, 12], gray-box attacks [13, 10], and black-box attacks [15, 16, 17, 18, 19]. The white-box attacks can obtain target models' information including the entire architecture, parameters, gradients of both the image encoder and text decoder, and probability of each word of the output text. Gray-box attacks can only access the architecture, parameters, and gradients of the image encoder, while black-box attacks cannot access any internal information of the target model, but only the output text of the model. Furthermore, black-box attacks can be divided into score-based and decision-based attacks. Score-based black-box attacks can access the probability of each word of the output text [16], while decision-based black-box attacks can only access the output text [15, 20, 21]. Because less information about the target models is provided, decision-based black-box attacks are more challenging than other categories [21]. Additionally, these attack methods can be categorized based on whether the attacker is able to specify the incorrect output text, dividing them into two types: targeted and untargeted attacks [22, 13].

Although numerous adversarial attack methods for image-to-text models have been proposed, to our best knowledge, the study on black-box attacks is under-explored, especially decision-based black-box targeted attacks. This kind of attack is more challenging due to the following reasons. Firstly, less information on the target model can be accessed. Specifically, only the output text instead of gradients, architectures, parameters, and the probability of each word in the output text is available. Secondly, the attackers not only cause the target model to output incorrect text, but also outputs the specified target text. Existing attacks easily suffer from the loss of semantics, resulting in the inability to effectively output the specified target text. Figure 1 (a) show that transfer+query [10] fabricates one target text to poison the target image-to-text model, leading to this model outputting an incorrect text. However, the output text could mismatch the original semantics of the target text, as the target image-to-text model may focus on secondary information while ignoring the crucial semantics of the target text behind the target image, resulting in semantic loss. More examples are in Appendix B.1.

To narrow the research gap, we propose a decision-based black-box targeted attack approach for image-to-text models. In our work, only the output text of the target model can be accessed, which is closer to the real-world cases [15]. Additionally, Figure 1 (b) demonstrates our targeted attack method, which optimizes against the target text directly under the decision-based black-box conditions, preventing semantic loss and maintaining semantic consistency with the target text.

Perturbing pixels in the image can change the output text. Therefore, the objective of the targeted attack can be considered to find the imperceptible pixel modification to make the output text similar to the target text. In this manner, the targeted attack can be formulated as a large-scale optimization problem, where pixels are decision variables and the optimization objective is to poison the output text. Inspired by the distinctive competency of evolutionary algorithms for solving large-scale optimization problems [23, 24, 25, 26, 27], we develop a dedicated evolutionary algorithm-based framework for decision-based black-box targeted attacks on image-to-text models. However, directly applying evolutionary algorithms to solve this large-scale optimization problem could suffer from low search efficiency, due to the numerous pixels and their wide range of values. To address the issue, we embed

Taiwan, Ministry of Culture and Tourism, Xiamen University, Xiamen 361005, Fujian, P.R. China (e-mail: minjiang@xmu.edu.cn; 36920221153145@stu.xmu.edu.cn).

2

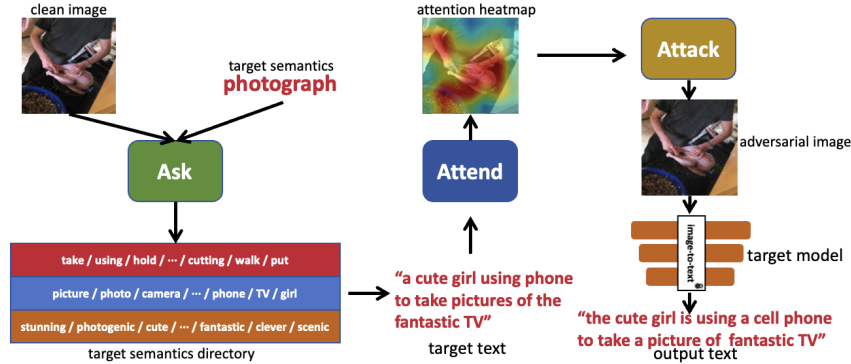



Figure 2: Diagram of our decision-based black-box targeted attack method *Ask, Attend, Attack*.

three-step processes, i.e., *Ask, Attend, Attack*, into the proposed evolutionary algorithm-based attacks. As shown in Figure 2, during the *Ask* stage, attackers can arbitrarily specify words related to certain semantics, such as *photograph*. Then, candidate words (e.g., *camera*, *scenic*, and *phone*) that are related to certain semantics are searched. Meanwhile, these words are close to the clean image in the feature space of the target image-to-text model. By selecting words from the candidate words, the target text (e.g., *a cute girl using a phone to take pictures of the fantastic TV*) related to the attacker's specified semantics can be formed to poison the target model. Subsequently, based on the attention mechanism, *Attend* identifies the crucial regions of the clean image (e.g. attention heatmap) [28, 29], thus reducing the search space for the subsequent *Attack*. Lastly, *Attack* uses a differential evolution strategy to impose imperceptible adversarial perturbations on the crucial regions, where the optimization objective is to minimize the discrepancy between the target text in *Ask* stage and the output text of the target model. Our contributions can be summarized as follows:

1. We first propose a decision-based black-box targeted attack *Ask, Attend, Attack* (*AAA*) for image-to-text models. Specifically, our method achieves targeted attacks without losing semantics while only the model's output text can be accessed.

2. We designed a target semantic directory to guide attackers in creating target text and utilized attention heatmaps to significantly reduce search space. This improves the search efficiency of evolutionary algorithms in adversarial attacks and makes attacks difficult to perceive.

3. We conducted extensive experiments on the Transformer-based VIT-GPT2 model and CNN+RNN-based Show-Attend-Tell model, which are the two most-used image-to-text models in HuggingFace, and surprisingly found that our decision-based black-box method has stronger attack performance than existing gray-box methods.

2 Related work

2.1 White-box Attack

In white-box attacks, the attacker has full access to all parameters, gradients, architecture of the target model, and the probability of each word of the output text. The authors in [7] add invisible perturbations to the image to make the image-to-text model produce wrong or targeted text outputs. The authors in [30] add global or local perturbations to the image to make the vision and language models unable to correctly locate and describe the content of the image. The authors in [14] modify the content of the image at the semantic level to make the image-to-text model output text that is inconsistent with the original image. The authors in [31] craft adversarial examples with semantic embedding of targeted captions as perturbation in the complex domain. The authors in [32] preserve the accuracy of non-target words while effectively removing target words from the generated captions. The authors in [33] generate coherent and contextually rich story endings by integrating textual narratives with relevant visual cues. The authors in [12] add limited-area perturbations to the image to make the image-to-text model fail to correctly describe the content of the perturbed area. The above methods require complete information of the image-to-text target model, including architecture, gradients, parameters, and probability distribution of the output text, which limits their practicality.

3

2.2 Gray-box Attack

To improve the practicality of adversarial attacks for image-to-text models, recent research explores how to attack with partial knowledge of the target model. All existing gray-box attack studies [34, 22, 13, 10] assume full access to the image encoder of the image-to-text model. The basic idea of gray-box targeted attacks is to reduce the distance between the adversarial image and the target image generated based on the target text in the image encoder's feature space. The authors in [34] generate adversarial images to mimic the feature representation of original images. The authors in [22] use a generative model to destroy the image encoder's features, achieving the untargeted attack. The authors in [13] minimize the feature distance in the image encoder between the adversarial image and the target image, thereby using gradient back-propagation to optimize the adversarial image and achieve the targeted attack. The authors in [10] combine existing gray-box method [13] with pseudo gradient estimation method [35] to achieve better performance in targeted attack. It is worth noting that they [10] call their method a black-box attack, but since they use the image encoder of the target model as the surrogate model, we classify their method as a gray-box attack. These gray-box attacks on image-to-text models are more practical than white-box attacks, but it is still unrealistic to assume that attackers can access the image encoder of the image-to-text model. Moreover, existing gray-box methods may have poor targeted attack performance due to the semantic loss mentioned above.

3 Methodology

3.1 Problem Formulation

The image-to-text model $\mathcal{G} : \mathcal{X} \rightarrow \mathcal{Y}$ maps the image domain $\mathcal{X}$ to the text domain $\mathcal{Y}$. A well-trained model should be able to accurately describe the content of the image using grammatically correct and contextually coherent text. Given a target text y_t, the attacker's goal is to find an adversarial image $\mathbf{x_{adv}}$ that is visually similar to clean image $\mathbf{x}$ and can generate an adversarial text y_{adv} that is semantically similar to y_t. We formalize the optimization problem for black-box targeted attack as:

$$\arg \max_{\mathbf{x}_{adv}} S(\mathcal{G}(\mathbf{x_{adv}}), y_t) \ \text{ s.t. } \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{x_{adv}}(i) - \mathbf{x}(i)\| \leq \epsilon, \tag{1}$$

where $S(\cdot, \cdot)$ represents the semantic similarity function between two texts, ϵ is the threshold for the average perturbation size per pixel, $\mathbf{x_{adv}}(i)$ and $\mathbf{x}(i)$ represents the value of the i-th pixel in the adversarial and clean images. n is the total number of pixels in all channels of the image.

3.2 Overview

To enhance the efficiency and stealth of decision-based black-box attacks, we propose the *Ask, Attend, Attack* (*AAA*) framework as shown in Figure 2. *Ask*: We compile a semantic dictionary from words within the input image's search space that align with the attacker's specified semantics. This facilitates targeted text generation, meeting the attacker's target semantics while simplifying the search process. *Attend*: We employ attention visualization and a surrogate model to generate an attention heatmap for the target text on the image, narrowing the search to significant decision variables and enhancing perturbation stealth. *Attack*: We use the differential evolution in the reduced search space to find the optimal solution that can mislead the target model to output target text. The framework's pseudo-code is detailed in Appendix A.1.

3.3 Ask Stage

According to the target semantics, the goal of *Ask* is to find words in the feature space of the target model to form a target semantic dictionary. These words should be closer to the input image. Firstly, we treat each pixel in each channel of image $\mathbf{x}$ as a variable, which means the search space size is the product of length, width, and number of channels. And then generate NP (number of population) individuals to form a population based on the following formula:

$$\mathbf{x}_j(i) = \mathbf{x}(i) + rand(-1, 1) \cdot \eta, \tag{2}$$

where $\mathbf{x}(i)$ is the i-th variable of clean image $\mathbf{x}$, $\mathbf{x}_j(i)$ is the i-th variable of the j-th individual in the population, η is a hyperparameter about the maximum search range, *rand*(-1,1) is a random number from the range of -1 to 1.

4

Secondly, for each variable, random mutation occurs between different individuals. The mutation for the i-th variable of the j-th individual $\mathbf{x}_j(i)$ is as follows:

$$\mathbf{v}_j^g(i) = \mathbf{x}_{r1}^g(i) + F * (\mathbf{x}_{r2}^g(i) - \mathbf{x}_{r3}^g(i)), \tag{3}$$

where $\mathbf{v}_j^g(i)$ is the mutated variable for mutation in the g-th generation of $\mathbf{x}_j(i)$. $\mathbf{x}_{r1}^g(i)$, $\mathbf{x}_{r2}^g(i)$, and $\mathbf{x}_{r3}^g(i)$ are three randomly selected individuals from the current population who are different from each other, F is the scaling factor.

Thirdly, each individual crossovers with the mutated individuals with a certain probability of generating candidate individuals. The formula is as follows:

$$\mathbf{u}_j^g(i) = \begin{cases} \mathbf{v}_j^g(i), & \text{if rand } (0,1) \leq CR, \\ \mathbf{x}_j^g(i), & \text{otherwise,} \end{cases} \tag{4}$$

where CR is crossover probability factor, $\mathbf{u}_j^g(i)$ is the i-th variable of the candidate individual in the g-th generation of the j-th individual in the population.

Fourthly, we use WordNet [36], a synonym dictionary, to measure the similarity between the target semantics and each individual's output text. WordNet groups words with the same semantics into synonyms, each representing a basic concept. We use WordNet to count the same semantic words m between each individual's output text and the target semantics. We calculate the $Precison=(m/t)$ and $Recall=(m/r)$, where t is the output text word count and r is the target semantics word count. Then, we calculate semantic similarity using the following formula:

$$S_{sem} = \frac{(1 - \gamma(\frac{ch}{m})^\theta)(\alpha^2 + 1) \cdot Precision \cdot Recall}{\alpha^2 \cdot Precision + Recall}, \tag{5}$$

where S_{sem} is the semantic similarity between the individual's output text and the target semantics [36], α balances the precision and recall weights, γ and θ control the penalty factor strength, ch is the number of consecutive word sets that match between the output text and the target semantics, with fewer chunks meaning more consistent word order.

Ultimately, we select offspring based on S_{sem}, choosing the current and candidate individuals that match the target semantics better as the next generation:

$$\mathbf{x}_j^{g+1} = \begin{cases} \mathbf{u}_j^g, & S_{sem}(\mathcal{G}(\mathbf{u}_j^g), TS) \geq S_{sem}(\mathcal{G}(\mathbf{x}_j^g), TS), \\ \mathbf{x}_j^g, & \text{otherwise,} \end{cases} \tag{6}$$

where TS is the attacker's target semantics. We extract nouns, adjectives, and verbs from the output texts of all the more semantically relevant and preserved individuals of each generation, expanding the target semantic dictionary. We use $\mathcal{G}(\mathbf{x}_j^{g+1}) = \{w_1, w_2, \cdots, w_n\}$ to represent the target model $\mathcal{G}$'s output text for the next generation of individuals, where w_i is the i-th word of the text and n is the word count. Then we use the following formula to extract important words and make a dictionary:

$$\mathbf{D}_j^{g+1} = \{w \in \mathcal{G}(\mathbf{x}_j^{g+1}) \mid w \text{ is noun, adjective or verb}\}, \tag{7}$$

where $\mathbf{D}_j^{g+1}$ is the dictionary for the preserved individual. We combine the dictionaries of each preserved individual in each generation to get the target semantic dictionary $\mathbf{D} = \mathbf{D}_1^2 \cup \mathbf{D}_2^2 \cup \cdots \cup \mathbf{D}_{\text{NP}}^m$, where m is the total number of generation. The attacker selects words from dictionary $\mathbf{D}$ that match the specified semantics to make the target text y_t. Words in dictionary $\mathbf{D}$ near input image $\mathbf{x}$ in feature space enhance searchability, enabling more efficient targeted attacks.

3.4 Attend Stage

The goal of *Attend* is to calculate the target text's attention area on the image $\mathbf{x}$. Because we do not have access to the internal information of the target model, we can only calculate the Grad-CAM attention heatmap [37] with the help of surrogate model f (such as ResNet trained in ImageNet). The surrogate model's sole purpose is to the compute attention heatmap. Since different models produce similar heatmaps for the same target text and input image, selecting a well-established visual model suffices [38]. The calculation formula of attention heatmap $\mathbf{A}$ is as follows:

$$\mathbf{A}(i,j) = \text{MAX}\left(0, \frac{1}{Z}\sum_k \sum_i \sum_j \cdot \frac{\partial y^{c^*}}{\partial \mathcal{F}_k(i,j)} \cdot \mathcal{F}_k(i,j)\right), \tag{8}$$

5

where $\mathbf{A}(i,j)$ is the decision-making contribution of the image to the target text at pixel (i,j), $\mathcal{F}_k(i,j)$ is the pixel (i,j) of the feature map of the k-th convolution kernel of the last convolutional layer of the surrogate model f, Z is the feature map's pixel count, y^{c^*} is the probability that f predicts that the image $\mathbf{x}$ belongs to class c^*. We use $\mathbf{C} = \{c_1, c_2, \cdots, c_{1000}\}$ for the ImageNet category names, where c_i is the i-th category name. We make the category text $y_{c_i} =$ "*a photo of*" $+ c_i$ from the category name c_i. We calculate the category c^* as:

$$c^* = \underset{c_i \in C}{argmax} \, \frac{E(y_t) \cdot E(y_{c_i})}{\|E(y_t)\|_2 \|E(y_{c_i})\|_2},\tag{9}$$

where E is the text encoder of the pre-trained CLIP model, and c^* is the closest category to the target text. We substitute c^* into Formula 8 to get the target text's attention heatmap $\mathbf{A}$. $\mathbf{A}(i,j)$ is the pixel (i,j)'s contribution to the target text. 1 means more contribution, and 0 means less contribution.

Table 1: Performance comparison (%) of different attack methods.

ϵ	Attack Methods	VIT-GPT2				Show-Attend-Tell			
		METEOR	BLEU	CLIP	SPICE	METEOR	BLEU	CLIP	SPICE
	Clean Sample	0.201±0.11	0.24±0.11	0.64±0.07	0.156±0.07	0.21±0.11	0.229±0.13	0.646±0.09	0.179±0.08
25	transfer (black)	0.206±0.11	0.246±0.11	0.639±0.07	0.165±0.07	0.211±0.12	0.225±0.14	0.648±0.09	0.185±0.11
	transfer+query (black)	0.221±0.16	0.264±0.15	0.651±0.18	0.167±0.07	0.219±0.11	0.231±0.14	0.654±0.05	0.187±0.14
	transfer (gray)	0.414±0.23	0.396±0.14	0.821±0.09	0.32±0.16	0.382±0.26	0.348±0.17	0.782±0.11	0.299±0.17
	transfer+query (gray)	0.433±0.21	0.411±0.12	0.832±0.13	0.35±0.09	0.401±0.21	0.355±0.15	0.794±0.11	0.311±0.13
	AAA (w/o *Attend*)	0.541±0.25	0.519±0.19	0.854±0.24	0.477±0.11	0.642±0.19	0.564±0.19	0.841±0.06	0.455±0.14
	AAA (w/o *Ask*)	0.398±0.21	0.384±0.18	0.795±0.25	0.412±0.13	0.364±0.21	0.322±0.19	0.754±0.08	0.376±0.13
	AAA	**0.696±0.21**	**0.658±0.22**	**0.952±0.29**	**0.634±0.15**	**0.855±0.15**	**0.799±0.21**	**0.964±0.04**	**0.786±0.14**
15	transfer (black)	0.204±0.09	0.241±0.15	0.627±0.18	0.164±0.07	0.232±0.13	0.236±0.14	0.643±0.08	0.187±0.09
	transfer+query (black)	0.211±0.14	0.256±0.15	0.644±0.15	0.181±0.09	0.245±0.13	0.246±0.11	0.656±0.06	0.203±0.09
	transfer (gray)	0.398±0.24	0.381±0.15	0.816±0.11	0.325±0.16	0.361±0.24	0.359±0.17	0.778±0.11	0.296±0.16
	transfer+query (gray)	0.408±0.19	0.399±0.11	0.824±0.15	0.341±0.13	0.375±0.19	0.368±0.15	0.784±0.11	0.311±0.13
	AAA (w/o *Attend*)	0.461±0.21	0.423±0.15	0.808±0.11	0.375±0.09	0.438±0.15	0.434±0.16	0.827±0.04	0.422±0.14
	AAA (w/o *Ask*)	0.378±0.25	0.361±0.17	0.768±0.15	0.356±0.15	0.341±0.15	0.337±0.18	0.749±0.07	0.365±0.13
	AAA	**0.556±0.31**	**0.504±0.26**	**0.851±0.12**	**0.44±0.17**	**0.617±0.25**	**0.574±0.22**	**0.913±0.05**	**0.553±0.14**

3.5 Attack Stage

The goal of *Attack* is to search for the best individual (adversarial sample) that outputs the target text y_t in the smaller search space reduced by the attention heatmap. Firstly, we copy the attention heatmap $\mathbf{A}$ three times in the channel dimension to match the shape of the image $\mathbf{x}$. We generated NP (number of population) individuals as a population with this formula:

$$\mathbf{x}_j(i) = \mathbf{x}(i) + rand(-\mathbf{A}(i), \mathbf{A}(i)) \cdot \eta,\tag{10}$$

where $\mathbf{x}(i)$ is the i-th variable of clean image $\mathbf{x}$, $\mathbf{x}_j(i)$ is the i-th variable of the j-th individual in the population, $\mathbf{A}(i)$ is the contribution of the i-th variable to the target text, and $rand(-\mathbf{A}(i), \mathbf{A}(i))$ is a random number in the range from $-\mathbf{A}(i)$ to $\mathbf{A}(i)$. The value of $\mathbf{A}$ is less than 1, and its mean and median are about [0.3,0.4]. The search space volume from the attention heatmap is much smaller than a hypersphere with radius η, because the radius and volume have an exponential relationship. This improves the search efficiency and concealment of adversarial perturbation.

Secondly, in order to accelerate convergence and better find the global optimal solution, we use the following CurrentToBest mutation [39]:

$$\begin{aligned}\mathbf{v}_j^g(i) = \mathbf{x}_j^g(i) &+ F * (\mathbf{x}_{r1}^g(i) - \mathbf{x}_{r2}^g(i)) \\ &+ F * (\mathbf{x}_{best}^g(i) - \mathbf{x}_j^g(i)),\end{aligned}\tag{11}$$

where $\mathbf{x}_j^g(i)$ is the i-th variable of the j-th individual in the g-th generation, $\mathbf{v}_j^g(i)$ is the mutated variable, $\mathbf{x}_{best}^g$ is the best fitness individual in the g-th generation population, $\mathbf{x}_{r1}^g$ and $\mathbf{x}_{r2}^g$ are two randomly selected individuals in the g-th generation population, and F is the scaling factor. The main advantage of this mutation strategy is that it combines the information of the current individual $\mathbf{x}_j$ and the best fitness individual $\mathbf{x}_{best}$, which can better guide the search process towards the direction of the optimal solution.

Thirdly, we use Formula 4 to calculate the candidate individual $\mathbf{u}_j^g(i)$. We design the following formula to calculate the deep feature similarity S_{clip} between two texts (u and v):

$$S_{clip} = 1 - \frac{E(u) \cdot E(v)}{\|E(u)\|_2 \|E(v)\|_2},\tag{12}$$

6

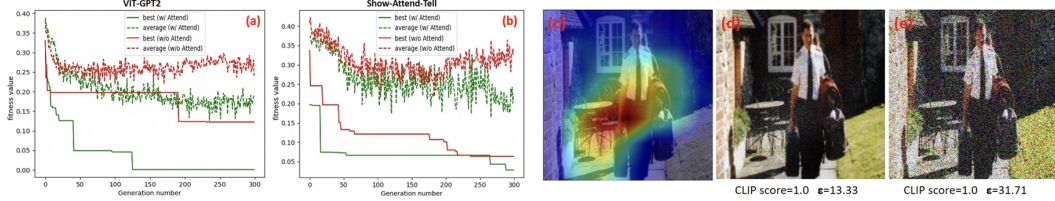



Figure 3: We compared the convergence curves of populations with and without *Attend* under the same perturbation size ϵ in (a-b). The fitness function is S_{clip} in Formula 12, where lower values mean stronger attacks. The dashed line is the average fitness value, and the solid line is the best fitness value. The green line is *AAA* and the red line is *AAA w/o Attend*. (c) shows the attention heatmap. (d) and (e) show the visual effects of adversarial image with and without *Attend*, with minimal perturbation of 100% attack success rate.

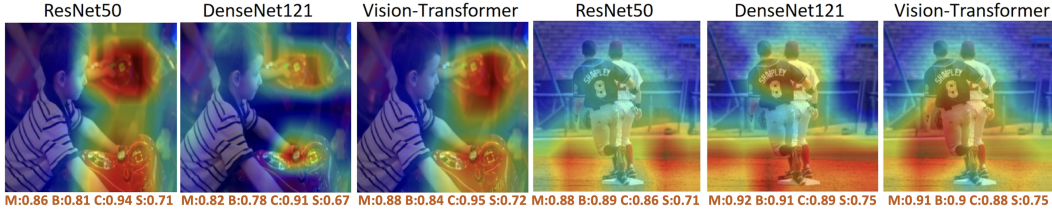



Figure 4: Grad-CAM attention heatmaps of different surrogate models for the same target text *a woman is holding a pair of shoes*. M is METEOR, B is BLEU, C is CLIP, S is SPICE.

where E is the text encoder of the pre-trained CLIP model. Text is discrete and complex, so it cannot calculate the distance directly [40]. Therefore, we use the CLIP text encoder E to extract the deep features of the texts, and then calculate the feature distance to obtain the similarity S_{clip} between the texts. The closer S_{clip} is to 0, the higher the similarity between the two texts u and v.

Ultimately, we select offspring using the following formula:

$$\mathbf{x}_j^{g+1} = \begin{cases} \mathbf{u}_j^g, & S_{clip}(\mathcal{G}(\mathbf{u}_j^g), y_t) \leq S_{clip}(\mathcal{G}(\mathbf{x}_j^g), y_t), \\ \mathbf{x}_j^g, & \text{otherwise}, \end{cases} \quad (13)$$

where $\mathbf{x}_j^{g+1}$ is the next individual with closer feature distance between the output text and the target text y_t. After performing the above evolutionary calculations multiple times, the optimal solution (adversarial sample) for outputting the target text is found.

4 Evaluation and Results

4.1 Experiment setups

Model and dataset We experimented with the two most-used image-to-text models on HuggingFace: VIT-GPT2 (Transformer-based) [41] and Show-Attend-Tell (CNN+RNN-based) [42]. VIT-GPT2 was trained on ImageNet-21k. Show-Attend-Tell was trained on MSCOCO-2014. We only used the target model's output text, not its internal information like gradients, parameters, or word probability. Following this work [13], we used Flick30k as our dataset, which has 31783 images and 5 caption texts each. We removed samples with less than 0.7 similarities between predicted text and truth text to ensure the target model's accuracy on clean images.

Evaluation metrics We used these evaluation metrics in our experiments: (1) BLEU(#4), an early machine translation metric that measures text precision [43]. 1 means similar, and 0 means dissimilar. (2) METEOR, a more comprehensive metric that considers synonyms, stems, word order, etc [36]. 1 means similar, and 0 means dissimilar. (3) CLIP, the distance between the CLIP text encoder's deep features for two texts [44]. 1 means similar, and 0 means dissimilar. (4) SPICE, an evaluation metric tailored for image-to-text models [45]. 1 means similar, and 0 means dissimilar. (5) ϵ, the mean perturbation size of each pixel of the adversarial sample [13].

7

Figure 5: Performance of adversarial image attacks varies with perturbation size ϵ. The ϵ of (a) and (f) is 25, ϵ of (b) and (g) is 15, ϵ of (c) and (h) is 10, ϵ of (d) and (i) is 5. (e) is our attention heatmap of the target text on the image. (j) is the target image generated based on the target text used in existing works. M is METEOR score, B is BLEU score, and C is CLIP score.

4.2 Experiment results

Comparison experiment of existing gray-box attacks. We evaluate state-of-the-art gray-box attacks [13, 10] on image-to-text models. We designate the gray-box attack [13] as transfer (gray) and the one [10] as transfer+query (gray). To simulate a black-box environment, we adapted these gray-box attacks by employing the CLIP model's image encoder in lieu of the target model's encoder, resulting in "transfer (black)" and "transfer+query (black)" variants. As depicted in Table 1, adversarial samples generated by the original gray-box attacks exhibit a marked increase in textual similarity to the target text when compared to clean samples. Conversely, the black-box adaptations maintain a similarity level akin to that of clean samples, indicating a significant loss of attack capability upon changing the image encoder. This underscores the dependency of gray-box attacks on the target model's image encoder. Our proposed method *AAA* demonstrates superior attack performance in black-box scenarios compared to the existing methods in their native gray-box settings. This is attributed to the semantic loss inherent in existing gray-box attacks, which constrains their attacking potential. It is noteworthy that our work represents the first black-box attack on image-to-text models. So we can only compare our approach with existing gray-box attacks. We have adapted these gray-box attacks into a black-box version solely to demonstrate their ineffectiveness in a black-box scenario.

Ablation experiment of our black-box attack. We conducted ablation experiments on our *AAA* method. *AAA* (w/o *Attend*) means no attention heatmap to reduce the search space, but the proportional reduction of the search range. *AAA* (w/o *Ask*) means the target text is not from the target semantic dictionary, but random words. Table 1 shows that losing any module decreases our attack performance. In addition, *Ask* performs worse than *AAA* (w/o *Attend*), indicating that finding a target text with lower search difficulty contributes relatively more to the performance of our targeted attack.

Qualitative experiment of attention. We presented the optimization curves of *AAA* and *AAA* (w/o *Attend*) in Figure 3. Figure 3 (a) and (b) illustrate the best and average fitness values during *AAA* and *AAA* (w/o *Attend*) optimization of VIT-GPT2 and Show-Attend-Tell. It is evident that the inclusion of *Attend* expedites and enhances the convergence of the population, with an equivalent perturbation size. Consequently, *AAA* exhibits more effective concealment in adversarial perturbations, maintaining the same level of attack efficacy, as depicted in Figures 3 (d) and (e). Furthermore, we evaluated the impact of selecting different surrogate models during *Attend*. Notably, the sole function of the surrogate model is to compute the attention heatmap. Figure 4 demonstrates that, despite significant structural variances among several surrogate models, they produce strikingly similar attention heatmaps for the same target text and input images. This similarity arises from mapping the target text to the most pertinent category within the surrogate model's label space (as Formula 9).

8

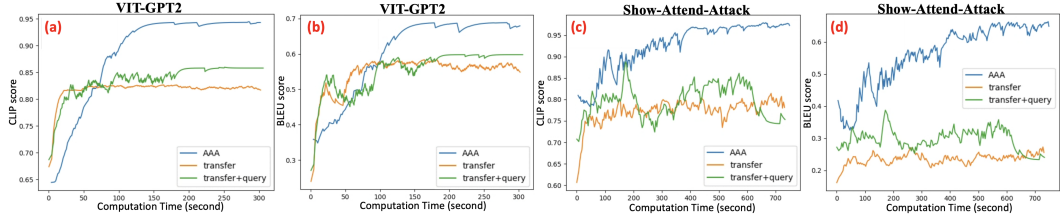



Figure 6: Comparison of computation time for generating a single adversarial sample using different adversarial attack methods. The y-axis is a measure of similarity between the generated text and the target text, with higher values indicating better target attack performance. The x-axis represents the computation time, and the shorter the time required to find a stable solution, the better.

The position of the same category of objects on the same picture is constant, and the model needs to focus on the object first, no matter what structure it is [38]. Performance comparisons, as shown in Figure 4, indicate that the similarity in attention heatmaps across different surrogate models leads to similar final attack performances. Therefore, we opted for a stable, well-established, pre-trained model, such as ResNet-50, to serve as our surrogate model.

Qualitative experiment of different perturbation sizes.   We used the words *mirror*, *cell phone*, *man*, *looking at* from the target semantic dictionary (as shown in Appendix B.2) to make the target text *a man is looking at a cell phone in a mirror*. We compared output texts of our black-box method *AAA* and the existing gray-box method [10] for adversarial samples with different ϵ, the average pixel perturbation size, in Figure 5. The same conclusion drawn from both methods is that bigger perturbation causes worse concealment and better attack performance; too small perturbation causes attack failure. Moreover, (f) and (j) in Figure 5 show that the existing methods have a semantic loss that limits their attack performance. Subjectively, target image (j) accurately draws the semantics of the target text, and the output text of adversarial image (f) perfectly describes the content of the target image (j). However the adversarial sample (f)'s output text does not have the semantics of the target text. Our method does not have semantic loss, so our black-box method *AAA* does a better targeted attack than the existing gray-box method. More examples of semantic loss are in Appendix B.1.

Comparison experiment on computation time.   We evaluated the computational efficiency of various attack methodologies for generating adversarial samples in image-to-text models. As depicted in Figure 6, our black-box attack method *AAA*, demonstrates a longer computation time to reach an optimal solution compared to existing gray-box attacks. For instance, the transfer approach [13] illustrated in Figure 6 (a) produces an adversarial sample with a CLIP score of 0.82 within a mere 29 seconds, while the transfer+query approach [10] achieves a CLIP score of 0.85 in just 97 seconds. Conversely, our *AAA* method requires 151 seconds to generate an adversarial sample with a superior CLIP score of 0.951. The shorter computation times of the existing gray-box methods are expected due to their ability to access real gradients, which significantly expedites the optimization process. Given that adversarial attacks are not time-sensitive operations and considering that our *AAA* method delivers a more potent attack capability and is applicable in a broader range of realistic black-box scenarios, the trade-off for a higher computational cost is deemed acceptable. Additional experiments on similarity measurements are included in the Appendix B.5.

Further analyses.   Firstly, we show the impact of different forms of target semantics TS in *Ask* on the target semantic dictionary, as shown in Appendix B.2. More ambiguous target semantics can enrich the target semantic dictionary, which also means that the attacker has more choices when designing y_t. Secondly, we show the effect of different word selection strategies of y_t based on target semantic dictionary on the final attack effect, as shown in Appendix B.3. Thirdly, we compare the convergence curves of different population sizes and choose a population size of 40 based on the trade-off of attack performance and convergence efficiency, as shown in Appendix B.4. Furthermore, we compare the effects of different evolutionary algorithms on attack performance and convergence efficiency, as shown in Appendix B.6. Additionally, to better observe the attack effect of our framework, we show more examples of attention heatmaps $\mathbf{A}$, optimization convergence curves, target text y_t, and output text, as shown in Appendix B.7. Lastly, we discuss the limitations of our framework, defense strategies, and future work in Appendix C.

9

5 Conclusion

In our research, we introduce a novel and practical approach for adversarial attacks on image-to-text models. We propose the *Ask, Attend, Attack* (*AAA*) framework, a decision-based black-box attack method that achieves targeted attacks without semantic loss, even with access limited to the target model's output text. Our framework uses the target semantic directory to guide the creation of target text and attention heatmap to reduce the search space, thereby improving the efficiency of evolutionary algorithms and making our attack harder to detect. Our extensive experiments on the Transformer-based VIT-GPT2 model and the CNN+RNN-based Show-Attend-Tell model demonstrate that our decision-based black-box method outperforms existing gray-box methods in targeted attack performance. These findings highlight the vulnerabilities in current image-to-text models and underscore the need for more robust defense mechanisms, significantly contributing to the field of adversarial machine learning and enhancing the security of vision-language systems.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276222 and the Public Technology Service Platform Project of Xiamen City, Grant No.3502Z20231043.

References

[1] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 12888–12900. PMLR, 2022.

[2] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2425–2433. IEEE, 2015.

[4] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5583–5594. PMLR, 2021.

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 67–74. IEEE, 2018.

[6] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Proceedings of the Neural Information Processing Systems (NIPS)*, 32, 2019.

[7] Hongge Chen, Huan Zhang, Pinyu Chen, Jinfeng Yi, and Cho-Jui Hsieh. Attacking visual language grounding with adversarial examples: A case study on neural image captioning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2587–2597. Association for Computational Linguistics, 2018.

[8] Baoer Liu, Qingyuan Zeng, Jianbin Huang, et al. Ivim using convolutional neural networks predicts microvascular invasion in hcc. *European Radiology*, 32(10):7185–7195, 2022.

[9] Yunpeng Gong, Liqing Huang, and Lifei Chen. Person re-identification method based on color attack and joint defence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4313–4322, 2022.

[10] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In *Proceedings of the Neural Information Processing Systems (NIPS)*, 2023.

10

[11] Yunpeng Gong, Chuangliang Zhang, Yongjie Hou, Lifei Chen, and Min Jiang. Beyond dropout: Robust convolutional neural networks based on local feature masking. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024.

[12] Hyun Kwon and SungHwan Kim. Restricted-area adversarial example attack for image captioning model. In *Wireless Communications and Mobile Computing (WCMC)*. Hindawi, 2022.

[13] Raz Lapid and Moshe Sipper. I see dead people: Gray-box adversarial attack on image-to-text models. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, 2023.

[14] Anand Bhattad, Minjin Chong, Kaizhao Liang, Bo Li, and D. A. Forsyth. Unrestricted adversarial examples via semantic manipulation. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR, 2020.

[15] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. 2019.

[16] Yucheng Shi, Yahong Han, Qinghua Hu, Yi Yang, and Qi Tian. Query-efficient black-box adversarial attack with customized iteration and sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2), 2022.

[17] Yunpeng Gong, Yongjie Hou, Zhenzhong Wang, Zexin Lin, and Min Jiang. Adversarial learning for neural pde solvers with sparse data. *arXiv preprint arXiv:2409.02431*, 2024.

[18] Yunpeng Gong, Qingyuan Zeng, Dejun Xu, et al. Cross-modality attack boosted by gradient-evolutionary multiform optimization. *arXiv preprint arXiv:2409.17977*, 2024.

[19] Yunpeng Gong, Zhun Zhong, Zhiming Luo, Yansong Qu, Rongrong Ji, and Min Jiang. Cross-modality perturbation synergy attack for person re-identification. *arXiv preprint arXiv:2401.10090*, 2024.

[20] Shuai Jia, Yibing Song, Chao Ma, and Xiaokang Yang. Iou attack: Towards temporally coherent black-box adversarial attack for visual object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[21] Kaixun Jiang, Zhaoyu Chen, Hao Huang, Jiafeng Wang, Dingkang Yang, Bo Li, Yan Wang, and Wenqiang Zhang. Efficient decision-based black-box patch attacks on video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4379–4389, 2023.

[22] Hanjie Wu, Yongtuo Liu, Hongmin Cai, and Shengfeng He. Learning transferable perturbations for image captioning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 18(2), 2022.

[23] Mohammad Nabi Omidvar, Xiaodong Li, and Yi Mei. Cooperative co-evolution with differential grouping for large scale optimization. *IEEE Transactions on evolutionary computation*, 18(3):378–393, 2013.

[24] Zhenzhong Wang, Haokai Hong, Kai Ye, Guangen Zhang, Min Jiang, and Kay Chen Tan. Manifold interpolation for large-scale multiobjective optimization via generative adversarial networks. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):4631–4645, 2023.

[25] Zhenzhong Wang, Lulu Cao, Liang Feng, et al. Evolutionary multitask optimization with lower confidence bound-based solution selection strategy. *IEEE Transactions on Evolutionary Computation*, 2024.

[26] Zhenzhong Wang, Dejun Xu, Min Jiang, et al. Spatial-temporal knowledge transfer for dynamic constrained multiobjective optimization. *IEEE Transactions on Evolutionary Computation*, 2024.

[27] Zhenzhong Wang, Qingyuan Zeng, Wanyu Lin, et al. Multi-view subgraph neural networks: Self-supervised learning with scarce labeled data. *arXiv preprint arXiv:2404.12569*, 2024.

11

[28] Qingyuan Zeng and Wu Zhou. An attention based deep learning model for direct estimation of pharmacokinetic maps from dce-mri images. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2368–2375. IEEE, 2021.

[29] Qingyuan Zeng, Baoer Liu, Yikai Xu, et al. An attention-based deep learning model for predicting microvascular invasion of hepatocellular carcinoma using an intra-voxel incoherent motion model of diffusion-weighted magnetic resonance imaging. *Physics in Medicine & Biology*, 66(18):185019, 2021.

[30] Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. Fooling vision and language models despite localization and attention mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

[31] Shaofeng Zhang, Zheng Wang, Xing Xu, Xiang Guan, and Yang Yang. Fooled by imagination: Adversarial attack to image captioning via perturbation in complex domain. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020.

[32] Jiayi Ji, Xiaoshuai Sun, Yiyi Zhou, Rongrong Ji, and Fuhai Chen. Attacking image captioning towards accuracy-preserving target words removal. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*. ACM, 2020.

[33] Qingbao Huang, Chuan Huang, Linzhang Mo, Jielong Wei, Yi Cai, Ho-fung Leung, and Qing Li. Igseg: Image-guided story ending generation. In *Findings of the ACL: International Journal of Conference on Natural Language Processing*, 2021.

[34] Akshay Chaturvedi and Utpal Garain. Mimic and fool: A task-agnostic adversarial attack. *IEEE Transactions on Neural Networks and Learning Systems*, 32(4):1801–1808, 2020.

[35] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

[36] Satanjeev Banerjee and Alon 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 (ACL WIEEMMTS)*, pages 65–72, 2005.

[37] Ramprasaath R. Selvaraju, Michael Cogswell, and Abhishek Das. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[38] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8565–8574, 2021.

[39] Jingqiao Zhang and Arthur C Sanderson. Jade: adaptive differential evolution with optional external archive. *IEEE Transactions on evolutionary computation*, 13(5):945–958, 2009.

[40] Jyun-Yu Jiang, Mingyang Zhang, Cheng Li, Michael Bendersky, Nadav Golbandi, and Marc Najork. Semantic text matching for long-form documents. In *Proceedings of the World Wide Web Conference (WWW)*, pages 795–806, 2019.

[41] NLP Connect. vit-gpt2-image-captioning (revision 0e334c7). `https://huggingface.co/nlpconnect/vit-gpt2-image-captioning`, 2022.

[42] Kelvin Xu, Jimmy Ba, and Kiros Jamie. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2048–2057. PMLR, 2015.

[43] Kishore Papineni and Salim Roukos. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

12

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agar-wal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.

[45] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 382–398. Springer, 2016.

[46] Zhenzhong Wang, Qingyuan Zeng, Wanyu Lin, Min Jiang, and Kaychen Tan. Generating diagnostic and actionable explanations for fair graph neural networks. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2024.

[47] Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11:341–359, 1997.

[48] Wael Khatib and Peter J. Fleming. The stud ga: a mini revolution? In *Proceedings of the International Conference on Parallel Problem Solving from Nature (PPSN)*, pages 683–691. Springer Berlin Heidelberg, 1998.

13

Overview

In this appendix, we describe implementation details, additional experiment results and analyses to support the methods proposed in the main paper. In addition, we show more examples of black-box adversarial attacks using *AAA*, each of which includes clean image, attention heatmap, adversarial image, optimization curve, target text, output text, and attack performance.

Reproducibility

Our **source code** and **data** are included in the supplemental material and uploaded, and we will publish the code on GitHub after the paper is accepted. We provide concise and understandable pseudo-code below.

Contents

1   Introduction   **1**

2   Related work   **3**

   2.1   White-box Attack   3

   2.2   Gray-box Attack   4

3   Methodology   **4**

   3.1   Problem Formulation   4

   3.2   Overview   4

   3.3   Ask Stage   4

   3.4   Attend Stage   5

   3.5   Attack Stage   6

4   Evaluation and Results   **7**

   4.1   Experiment setups   7

   4.2   Experiment results   8

5   Conclusion   **10**

A   Additional implementation details   **16**

   A.1   Pseudo code of our proposed framework   16

   A.2   Basic setups   16

   A.3   Standard deviation in the experiments   16

   A.4   Evaluation metrics   17

B   Additional experiments   **17**

   B.1   Analysis of semantic loss   17

   B.2   Comparison experiment of target semantic dictionary   18

   B.3   Word selection strategies for target semantic dictionaries   19

   B.4   Comparison experiment on population size   20

   B.5   Comparison experiment on computation time   20

14

B.6   Comparison experiment of optimization algorithms   21

B.7   Visualization of more adversarial samples   22

C   Discussion   **22**

C.1   Limitation   22

C.2   Future work   22

A Additional implementation details

A.1 Pseudo code of our proposed framework

Algorithm 1 Ask, Attend, Attack (AAA) Framework

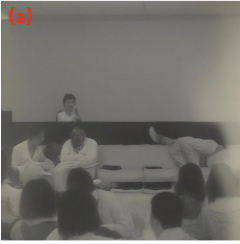
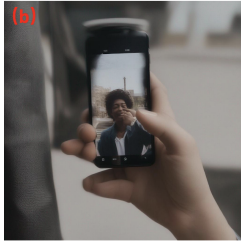
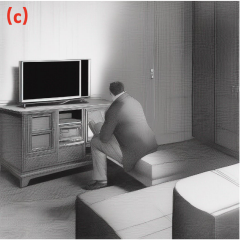
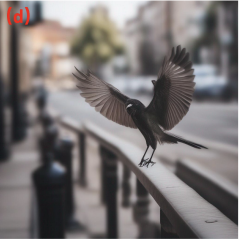
1: **Input:** Image $\mathbf{x}$, Target text y_t, Target semantics TS, Surrogate model f, Pre-trained CLIP model E
2: **Output:** Adversarial image $\mathbf{x_{adv}}$ that generates text y_{adv} semantically similar to y_t
3: Initialize hyperparameters: population size NP, mutation factor F, crossover probability CR, perturbation threshold ϵ, maximum search range η
4: Initialize target semantic dictionary $\mathbf{D} \leftarrow \emptyset$
5: **function** ASK($\mathbf{x}$, TS)
6:     Generate initial population with perturbations using Eq. (2)
7:     **for** each generation g **do**
8:         Perform mutation using Eq. (3)
9:         Perform crossover using Eq. (4)
10:        Calculate semantic similarity S_{sem} using Eq. (5)
11:        Select offspring based on S_{sem} using Eq. (6)
12:        Update $\mathbf{D}$ with relevant words from $\mathcal{G}(\mathbf{x}_j^{g+1})$ using Eq. (7)
13:    **end for**
14:    **return D**
15: **end function**
16: **function** ATTEND($\mathbf{x}$, y_t, f)
17:     Determine the category c^* closest to y_t using Eq. (9)
18:     Attention heatmap $\mathbf{A}$ is calculated by surrogate model f using Eq. (8)
19:     **return A**
20: **end function**
21: **function** ATTACK($\mathbf{x}$, y_t, $\mathbf{A}$)
22:     Generate initial population with attention-guided perturbations using Eq. (10)
23:     **for** each generation g **do**
24:         Perform CurrentToBest mutation using Eq. (11)
25:         Perform crossover using Eq. (4)
26:         Calculate deep feature similarity S_{clip} using Eq. (12)
27:         Select offspring based on S_{clip} using Eq. (13)
28:     **end for**
29:     **return** Best individual as $\mathbf{x_{adv}}$
30: **end function**
31: $\mathbf{D} \leftarrow$ ASK($\mathbf{x}$, TS)
32: $y_t \leftarrow$ The attacker create a sentence from the dictionary $\mathbf{D}$
33: $\mathbf{A} \leftarrow$ ATTEND($\mathbf{x}$, y_t, f)
34: $\mathbf{x_{adv}} \leftarrow$ ATTACK($\mathbf{x}$, y_t, $\mathbf{A}$)

A.2 Basic setups

We set the population size NP to 40, scaling factor F to 0.7, cross probability factor CR to 0.7, γ to 0.5, α to 1, and θ to 3, and η to ϵ required in the experiment divided by the average of attention heatmap $\mathbf{A}$. Our device uses three GPUs of RTX2080ti with 11GB memory, and a CPU of Intel(R) Core(TM) i5-10400F. Our operating system is linux, the evolutionary algorithm framework uses the Geatpy library, and the deep learning framework uses Pytorch.

A.3 Standard deviation in the experiments

In the quantitative experiment of our paper, experiments were repeated for 10 times, and the optimal performance was obtained for each experiment, and the mean value and standard deviation were finally obtained.

16

Target Image:	Target Image:	Target Image:	Target Image:

Target Text:
a group of people sitting down

Output Text:
a woman in a white dress is talking to a man in a white dress

Similarity:
M:0.04 B:0.05 C:0.71 B:0.08

Target Text:
a picture of a man in a cell phone

Output Text:
a person is taking a picture of a person on a cell phone

Similarity:
M:0.73 B:0.51 C:0.88 B:0.4

Target Text:
a man is watching tv

Output Text:
a man sitting on a chair next to a fire hydrant

Similarity:
M:0.18 B:0.17 C:0.65 B:0.18

Target Text:
a bird is sitting on a bird flying near a street

Output Text:
a bird flying over a ledge with a bird perched on top

Similarity:
M:0.43 B:0.5 C:0.84 B:0.33

Figure 7: More examples of semantic loss of existing gray-box targeted attacks. The target text is the error-generated text of the image-to-text model that the attacker wants to obtain. The target image is the image generated by using the text-to-image model (Stable Diffusion) based on the target text. The output text is based on the target image using the image-to-text target model (VIT-GPT2/Show-Attend-Tell). similarity indicates the similarity between the target text and the output text. We also show the similarity between the target text and the output text. M stands for METEOR score, B for BLEU score, C for CLIP score, and S for SPICE score.

A.4 Evaluation metrics

(1) iteration, the number of iterations for the differential evolution algorithm in *Attack* to find the optimal solution (no more fitness convergence). Fewer iterations mean fewer queries and faster attack. (2) ϵ, the mean perturbation size of each pixel of the adversarial sample. Smaller value means higher concealment of adversarial perturbation. (3) diversity, the number of words in the target semantic dictionary from *Ask*. More words mean more diversity. (4) correlation, the average CLIP score between each word in the target semantic dictionary and the target semantics. The higher correlation, the more relevant the words in the target semantic dictionary are to the target semantics.

B Additional experiments

B.1 Analysis of semantic loss

We show more examples of the semantic loss phenomenon, as shown in Figure 7. In order to realize the targeted attack with the existing gray-box methods, it is necessary to convert the target text into the target image with the help of text-to-image model (such as Stable Diffusion). Then the distance between the adversarial image and the target image is narrowed, so that the text decoder of the image-to-text target model mistakes the adversarial image as the target image and outputs the description of the target image incorrectly. The target image often contains more semantic information than the target text, and the image-to-text target model may focus on the semantic information that is not specified by the attacker, which leads to semantic loss. For example, in Figure 7 (c), the text-to-image model generates the target image corresponding to the target text (*a man is watching tv*) very accurately, and the image-to-text target model also generates the output text (*a man sitting on a chair next to a fire hydrant*) of the target image very accurately, but the output text and the target text are very different. This means that even if there is a gray-box method that can completely make the features of the adversarial image identical to the features of the target image, the image-to-text target model can only generate the output text after semantic loss, and the targeted attack performance is limited by semantic loss.

17

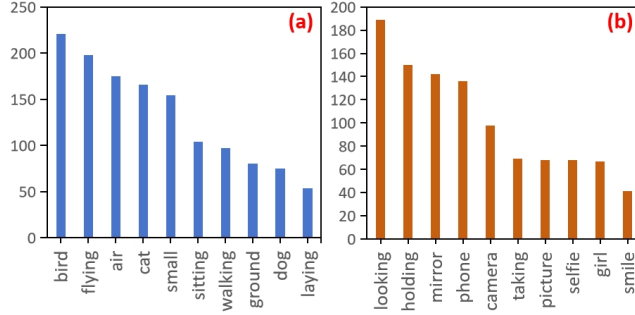



Figure 8: The top ten words in the target semantic dictionary for different target semantics, with the word frequency on the vertical axis. (a) is for *animal*; (b) is for *photograph*.

B.2 Comparison experiment of target semantic dictionary

We showed the target semantic dictionary's diversity and correlation for different perturbations in Table 2. More perturbation means more word choices for the target text. The correlation between dictionaries and target semantics is not affected by the size of perturbations. We also see that one vague word for target semantic makes more diversity and relevance in the dictionary than the detailed sentences. This is because a word has vague semantics, resulting in more words that are closer to the input image in the feature space being added to the dictionary. So we suggest using simple words as target semantics, as attackers can get richer dictionaries to make target text.

Table 2: Target semantic dictionaries for different semantics. *animal* word means the vague word *animal*, while *animal* sentence means *a dog is running after a cat*. *photograph* word means the vague word *photograph*, while *photograph* sentence means *a photo of a parking lot*.

semantic	*animal* word			*animal* sentence			*photograph* word			*photograph* sentence		
ϵ	10	15	25	10	15	25	10	15	25	10	15	25
diversity↑	50.6	65.4	90.1	38.9	54.1	79.6	51.7	62.5	87.7	43.1	52.6	75.5
correlation (%)↑	0.746	0.742	0.744	0.653	0.65	0.654	0.842	0.841	0.843	0.765	0.761	0.758

Table 3: Output text under different word selection strategies.

Strategy	Target Text	Output Text	Similarity
A	a bird is flying through air	a bird is flying through the air	great
A	a girl is taking pictures by camera	a girl is using a camera to take pictures	great
B	a camera is flying through the air	a man is holding a camera	medium
C	a giraffe is eating grass	a person is cutting a piece of food	bad
C	a boy is capturing a beautiful moment	a man is looking at his cell phone	bad
D	the helicopter is hovering in the sky	a man is holding a knife in his hand	bad

18

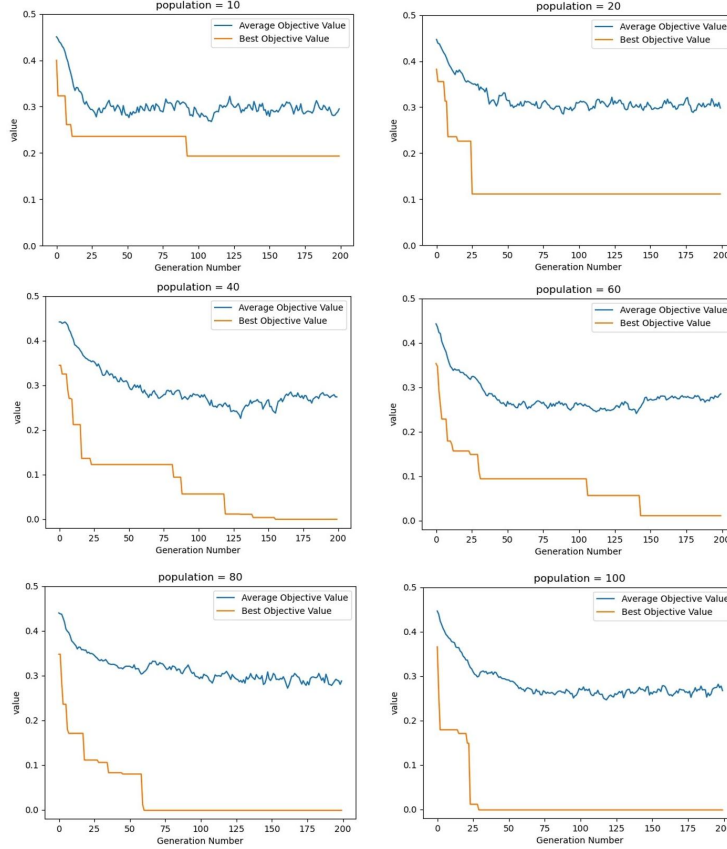



Figure 9: The best fitness curve and average fitness curve of the population under the same target text and different population sizes.

Table 4: The time (s) required for one optimization iteration under different population sizes.

population	10	20	40	60	80	100
iteration time	0.41	0.65	1.14	1.65	2.14	2.56

B.3 Word selection strategies for target semantic dictionaries

We showed the words and frequencies in the target semantic dictionary for different semantics in Figure 8. We compared different word selection strategies for targeted attacks with these dictionaries. The results show that: (1) Words in the dictionary do better when the semantics are similar, while words outside may fail; (2) Words from two dictionaries in one sentence decrease the performance.

We used four word selection strategies based on two dictionaries in Figure 8 to compare how different target texts y_t affect our method: (A) All words in y_t are from the same dictionary; (B) Some words in y_t are from each of the two dictionaries; (C) y_t is artificially created with the target semantics (*animal* or *photograph*), but without any words from the target semantic dictionary; (D) y_t is artificially

Table 5: Performance (%) of different evolutionary algorithms and average number of iterations to find the optimal solution.

	CTB-DE	R-DE	S-GA
iteration↓	46.47±37.11	57.35±43.62	**15.41±11.52**
METEOR↑	**0.696±0.209**	0.538±0.264	0.327±0.254
BLEU↑	**0.658±0.219**	0.546±0.218	0.279±0.172
CLIP↑	**0.95±0.291**	0.871±0.112	0.748±0.096

19

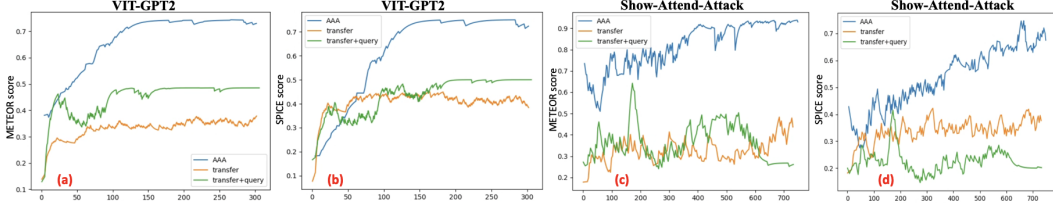



Figure 10: Comparison of computation time for generating a single adversarial sample using different adversarial attack methods. The y-axis is a measure of similarity between the generated text and the target text, with higher values indicating better target attack performance. The x-axis represents the computation time, and the shorter the time required to find a stable solution, the better.

created with different semantics from both target semantics (*animal* and *photograph*), and without any words from either target semantic dictionary. The output texts of the adversarial images obtained from different y_t word selection strategies are shown in Table 3.

The first row of Table 3 shows that strategy (A) can achieve a strong targeted attack, making the output text very similar to the target text. This is because words in the same dictionary are close to each other in the feature space. Strategy (B) selects the words *flying* and *air* from dictionary *animal* in Figure 8 (a), and *camera* from dictionary *photograph* in Figure 8 (b), to form the target text. The third row of Table 3 shows that the output text and the target text y_t are not very similar. The output text only contains the word *camera* in dictionary (b). This is because the feature distance between the two dictionaries is large, even though they are both close to the input image and easy to search in the feature space. It is hard to optimize the target text y_t that contains words from both target semantic dictionaries. Strategy (C) randomly creates y_t based on the *animal* and *photograph* semantics, without using any words from dictionary (a) and (b). For example, *giraffe* is an *animal*, but not in dictionary (a), and *capture beautiful moment* is related to *photograph*, but not in dictionary (b). The output text and the target text y_t are totally different, indicating a failed targeted attack. Strategy (D) randomly creates y_t with different semantics from both target semantics, and without any words from either target semantic dictionary. The targeted attack also fails. Therefore, we recommend selecting words from one target semantic dictionary for the target text y_t, which will greatly improve the success rate of our method's targeted attack.

B.4 Comparison experiment on population size

We show convergence curves with the same target text but different population sizes NP to observe how they affect the optimization iteration process of *Attack*. Figure 9 shows that when NP is 10 and 20, the best fitness values are 0.2 and 0.1, corresponding to CLIP scores of 0.8 and 0.9 for the output texts and target texts, respectively. When NP is larger than 40, the output text and the target text are completely consistent (CLIP score = 1). This means that a larger NP can find better solutions with fewer iterations [24, 46]. However, a larger NP also increases the computation time per iteration, as Table 4 shows. Moreover, as this is a large-scale optimization problem with 196608 decision variables per individual, a larger NP demands more hardware resources [47, 39]. Considering all factors, we set the population size NP to 40.

B.5 Comparison experiment on computation time

In Figure 6 of the main paper, we show the computational efficiency of two metrics, CLIP score and BLEU score. In this part, we will supplement the other two metrics, METEOR score and SPICE score. As shown in Figure 10, the computation time of the existing gray-box attack methods to find the optimal solution is still shorter than that of our black-box attack method. For example, the transfer approach [13] illustrated in Figure 10(a) produces an adversarial sample with a METEOR score of 0.34 within a mere 62 seconds, while the transfer+query approach [10] achieves a METEOR score of 0.49 in just 119 seconds. Conversely, our *AAA* method requires 179 seconds to generate an adversarial sample with a superior METEOR score of 0.75. Because our method is more practical and performs better, the additional computation time is acceptable.

20

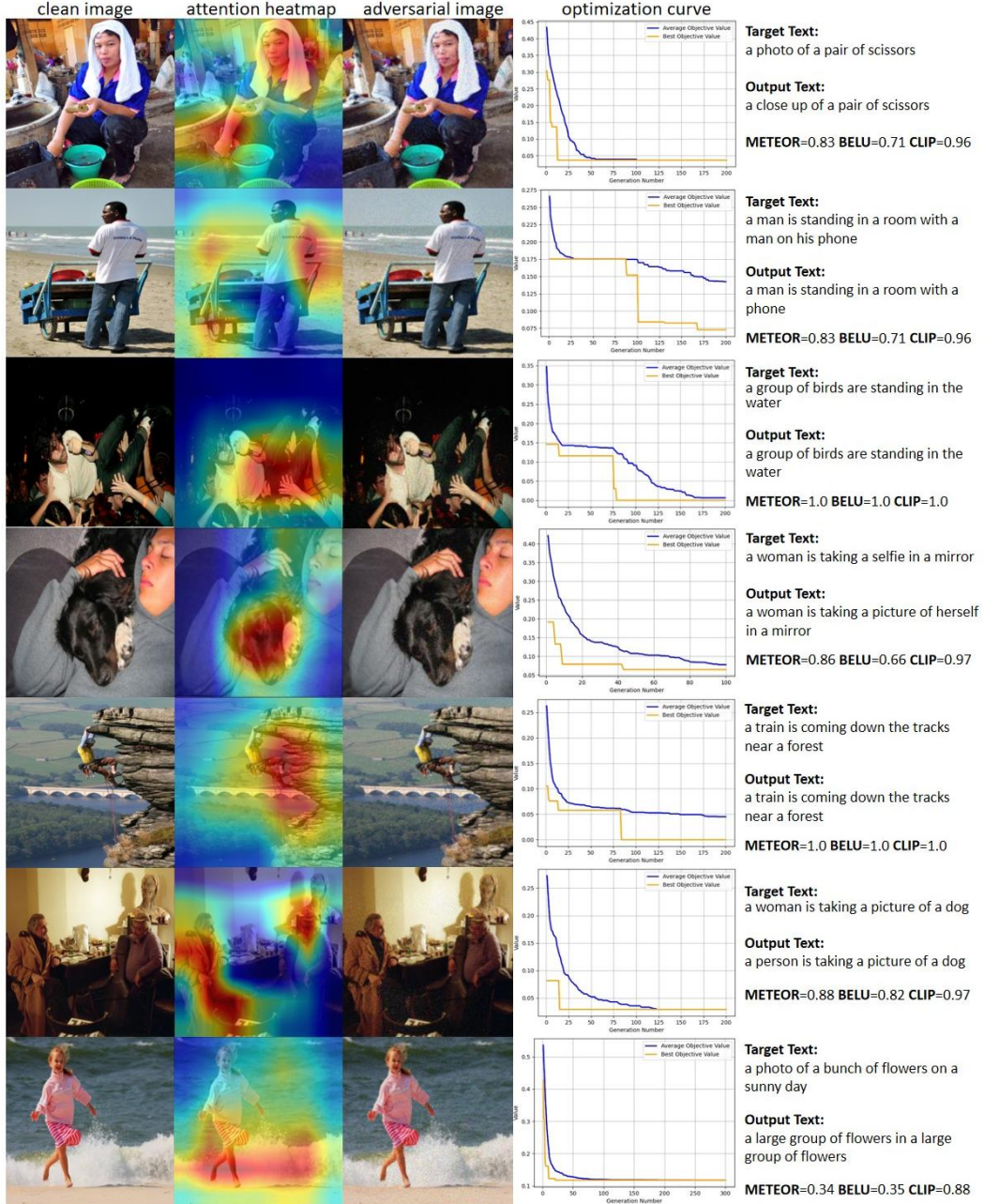



Figure 11: Attention heatmaps, optimization convergence curves, target text, output text and attack performance for more adversarial samples.

B.6 Comparison experiment of optimization algorithms

We compared different optimization strategies in *Attack*: CurrentToBest Differential Evolution (CTB-DE) [39], Rand Differential Evolution (R-DE) [47], and Stud Genetic Algorithm (S-GA) [48]. Table 5 shows that the genetic algorithm needs the fewest iterations, but easily gets stuck in local optima, leading to poor attack performance. Differential evolution needs more iterations but finds better solutions. Also, the CurrentToBest mutation does better and faster than the random mutation. So we adopted the CurrentToBest differential evolution strategy in *Attack*.

21

B.7 Visualization of more adversarial samples

We presented attention heatmaps **A**, optimization convergence curves, target text y_t, and output text for more adversarial samples, as shown in Figure 11.

C Discussion

C.1 Limitation

Our work represents the first black-box targeted attack on image-to-text models, with the core idea utilizing evolutionary algorithms to solve a large-scale optimization problem. The drawbacks of evolutionary algorithms, which are also the limitations of our work, include: (1) **Low optimization efficiency**. Gradient-based algorithms use the gradient information of the objective function, which is a powerful guide regarding the optimization direction. Evolutionary algorithms do not directly use gradient information but search through random mutation and crossover operations. Compared to gradient optimization algorithms, evolutionary algorithms require more iterations to find the optimal solution. (2) **High number of queries**. Each individual in the population requires access to the target model in every iteration, and the service provider of the image-to-text target model can simply set a limit on the number of accesses to defend against our attack.

C.2 Future work

Our black-box targeted attack framework *Ask, Attend, Attack* on image-to-text models employs classic evolutionary algorithms. In our future work, we will explore how our framework *AAA* can be combined with the current state-of-the-art (SOTA) evolutionary algorithms, which have the fastest convergence efficiency, to mitigate the limitations mentioned above.

22

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope. We have clearly stated our novel methodology and its implications in the abstract and introduction, and these are further elaborated upon and validated in the main body of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Our experiments in Section 4.2 and our conclusions in Section 5 describe the limitations of our approach.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

23

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our paper presents a framework for adversarial attacks based on evolutionary algorithms. The underlying techniques employed, such as differential evolution and similarity computation, are well-established with their theoretical foundations extensively proven in existing literature.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide experimental Settings in section 4.1, we use public data sets, and we provide pseudocode and source code in supplementary materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

24

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have packaged our source code and data into a zip file and uploaded it to the system as supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide experimental Settings in section 4.1, and we provide pseudocode and source code in supplementary materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: In our performance comparison experiments, we report the results of 10 repeated experiments as mean ± standard deviation. This provides a measure of the variability of the results, and gives a better understanding of the range of performance we can expect from our method.

25

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide the GPU models in Section 4.1 and a comparison experiment on computation time in Section 4.2

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: Our research adheres to the NeurIPS Code of Ethics in all respects. We have ensured fairness and transparency in all our experiments, respected the rights of all participants, and our research does not violate any laws or regulations.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

26

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Our research may have potential negative societal impacts. Malicious actors could potentially use our black-box adversarial attack techniques to generate adversarial samples to deceive commercial image-to-text models, thereby damaging the reputation of the models. An effective solution could be to set a threshold for the number of accesses.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The original intention of our proposed adversarial attack technique is to provide a baseline for researchers in the robustness of deep neural networks, but it could still potentially be misused by malicious actors. To prevent these actors from using our method to attack commercial image-to-text models, we have proposed a defensive measure, which is to set a limit on the number of accesses.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

27

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We used open data sets and correctly referenced the source code papers.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We have introduced new assets in the form of source code in our paper. This code is well documented, with comments explaining the functionality of different sections. We have included this code in the supplementary material zip file that we have submitted alongside our paper.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

28

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

29