NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2024

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

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

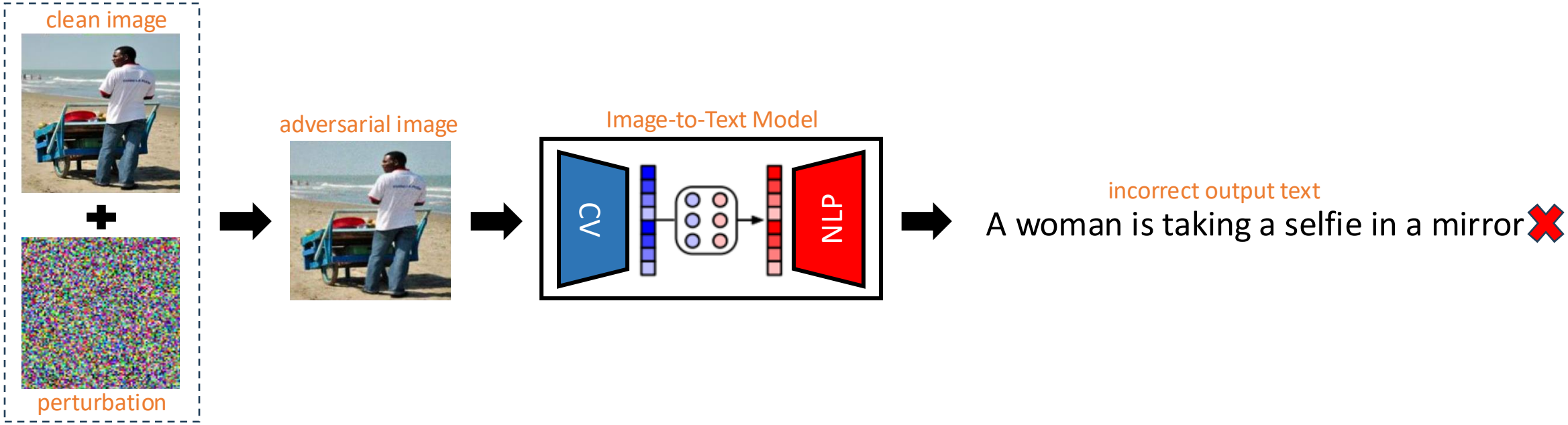
Institute of Artificial Intelligence, Xiamen University[1]
Department of Computing, The Hong Kong Polytechnic University[2]
Department of Computer Science, Hong Kong Baptist University[3]
School of Informatics, Xiamen University[4]

Presenter Name: Qingyuan Zeng

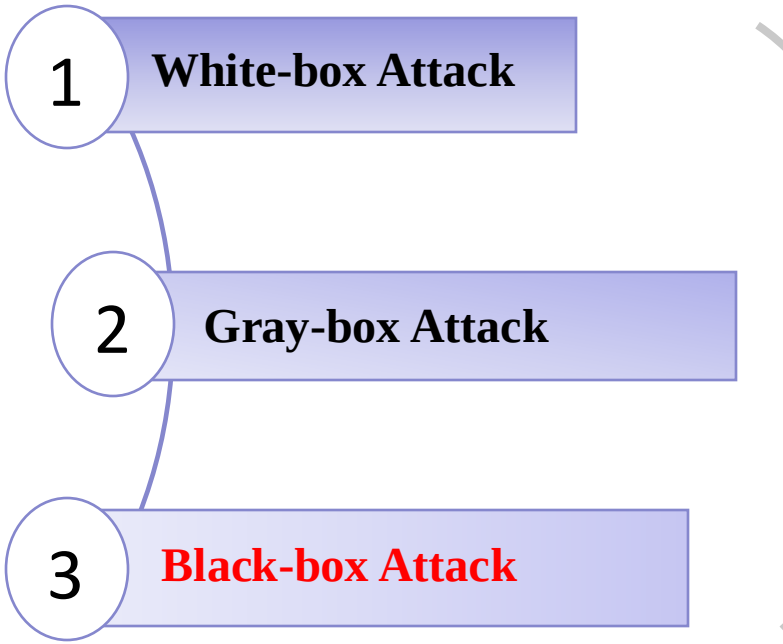
The corresponding author*: Min Jiang, minjiang@xmu.edu.cn

Background

- Image-to-text models, referring to generating descriptive and accurate textual descriptions of images, have received increasing attention in various applications.
- 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.



Background



- Attackers have full access to the image-to-text model's architecture, parameters, gradients, and output probabilities
- Attackers have access to the architecture, parameters, and gradients of the image encoder in the image-to-text model
- Attackers have no access to the model's internal information and can only observe the model's output text.

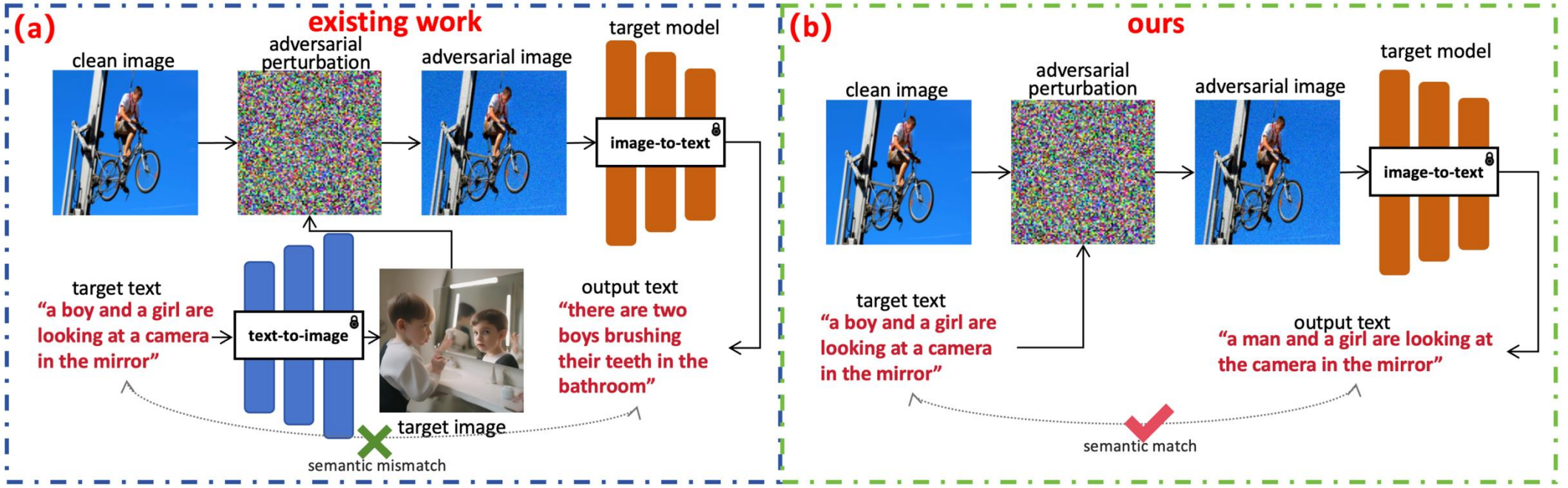
Why are black-box targeted attacks the most practical and difficult?

- attackers have minimal information, with only the output text available
- the goal is not only to mislead the model into producing incorrect text but also to output specific text designated by the attacker

Motivation

Why is it necessary to develop a black-box targeted attack against image-to-text models?

Gray-box attacks on image-to-text models have been overlooked by AI service providers due to the unavailability of model internals and the limitations imposed by semantic loss. However, our research reveals that attackers can effectively control the output with just the model's output text, highlighting the need for more secure image-to-text models.

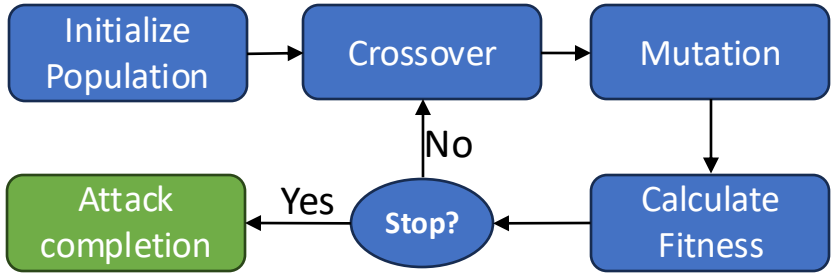


Methodology

Black-box targeted attack can be formulated as a large-scale optimization problem

- Decision variables： each pixel of the input image
- Constraint： the grayscale value of each pixel does not exceed a certain threshold ε
- Individual： An input image with perturbed pixel points
- Population： A collection of input images with various perturbations.
- Optimization objective： To select individuals with output text that is more similar to the target text

Evolutionary Algorithm Flowchart

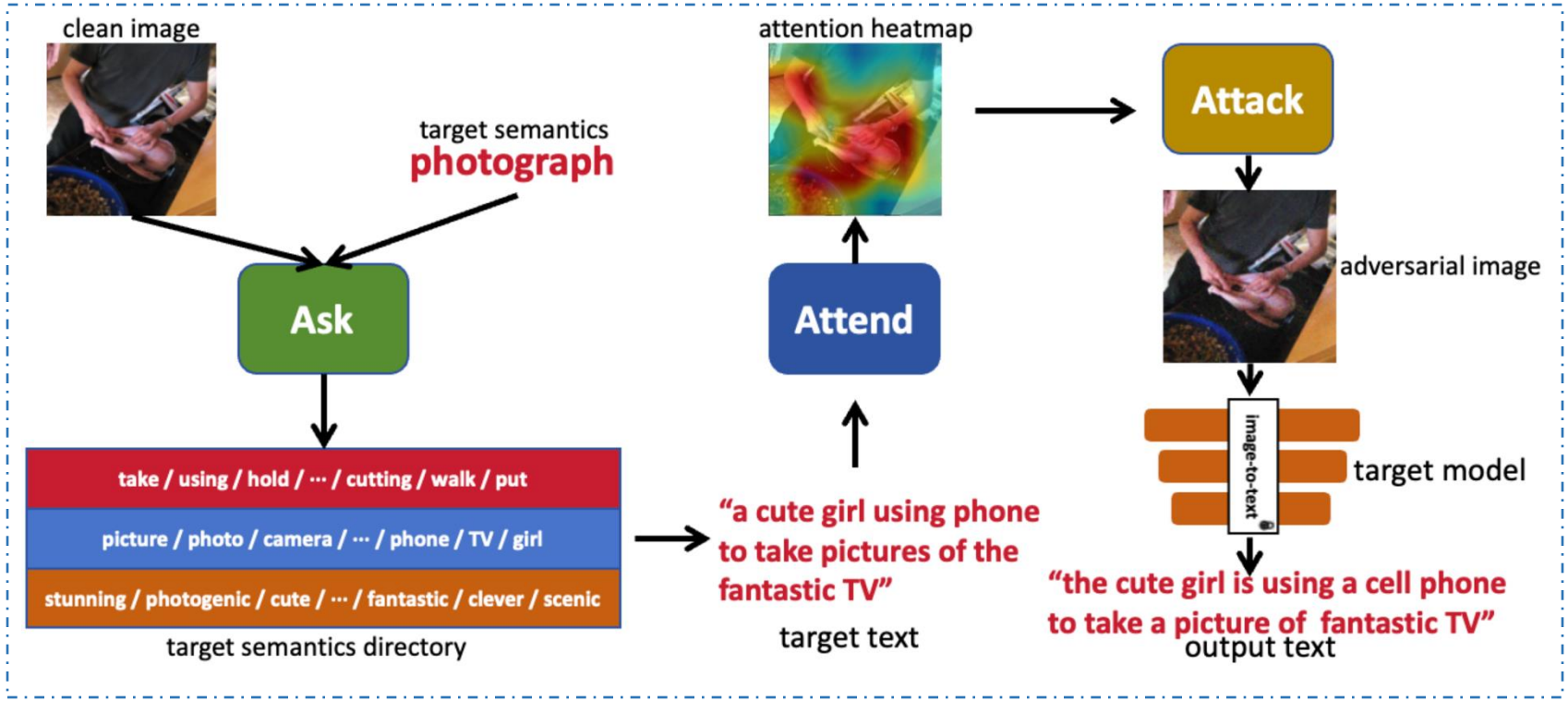


Challenges:

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

Methodology

Framework of Our Method AAA (Ask, Attend, Attack)



To enhance the search efficiency of evolutionary algorithms in image-to-text model black-box targeted attacks, we propose a three-step framework: *Ask*, *Attend*, *Attack*

Methodology

- **Ask:**

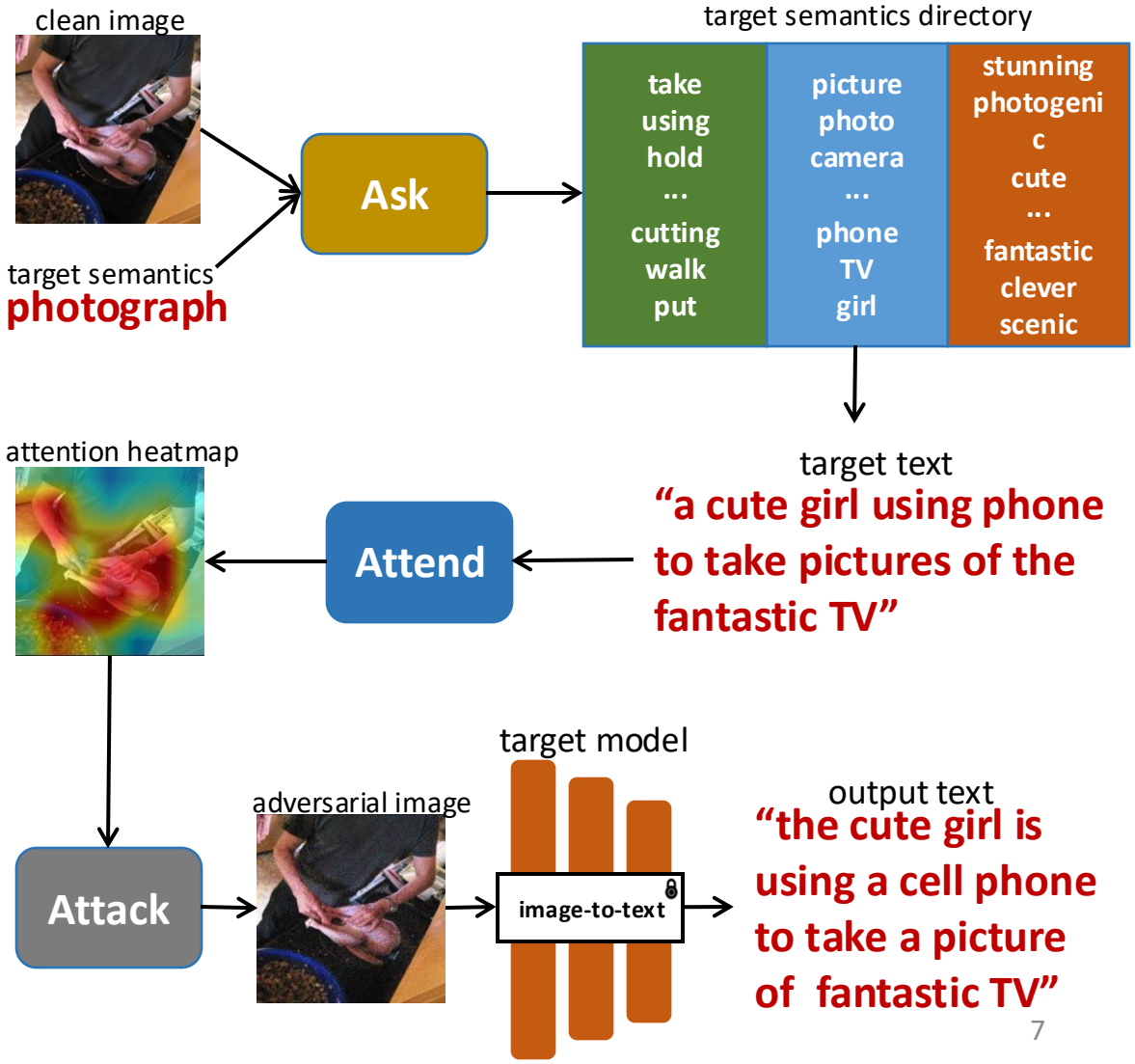
Search a target semantic dictionary that aligns with the attacker's intended semantics, enabling the attacker to craft target texts that are more easily searchable

- **Attend:**

Reduce the decision variable range for areas of the input image with low relevance to the target text, thereby reducing the search space and improving search efficiency

- **Attack:**

Based on the target text and reduced search space obtained from the previous steps, we employ differential evolution algorithms to find adversarial images that output text most similar to the target text



NEURAL INFORMATION PROCESSING SYSTEMS

Experiment

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**

- Our proposed method AAA demonstrates superior attack performance in black-box scenarios compared to the existing methods in their native gray-box settings
- Ablation experiment shows that losing any module will decrease our attack performance

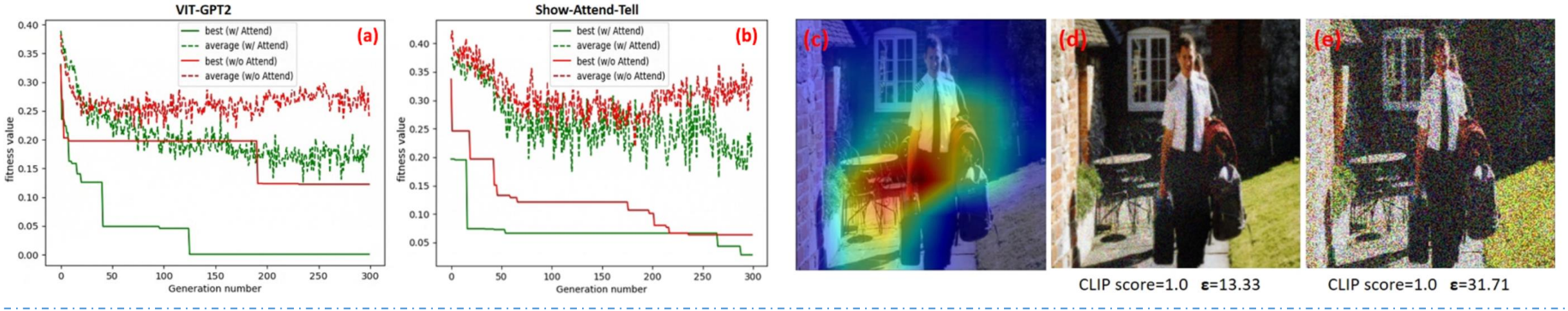
Experiment

Performance of adversarial image attacks varies with perturbation size ϵ

- Bigger perturbation causes worse concealment and better attack performance; too small perturbation causes attack failure
- Figure (f) and (j) show that the existing methods have a semantic loss that limits their attack performance
- *AAA* does not have semantic loss, so *AAA* does a better targeted attack than the existing gray-box method

Experiment

Qualitative Experiments of the Attend Module



- As shown in Figure (a) and (b), the inclusion of *Attend* module expedites and enhances the convergence of the population, with an equivalent perturbation size
- As shown in Figure (d) and (e), *Attend* module exhibits more effective concealment in adversarial perturbations, maintaining the same level of attack performance

Discussion

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* 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.

Limitation

- Low optimization efficiency
- High number of queries

Future work

- In our future work, we will explore how our framework AAA can be combined with the current state-of-the-art evolutionary algorithms, which have the fastest convergence efficiency, to mitigate the limitations mentioned above

Thank you for your attention.