VEAttack: Downstream-agnostic Vision Encoder Attack against Large Vision Language Models

Paper ID: 363

Hefei Mei[1], Zirui Wang[1], Shen You[1], Minjing Dong[1], Chang Xu[2]

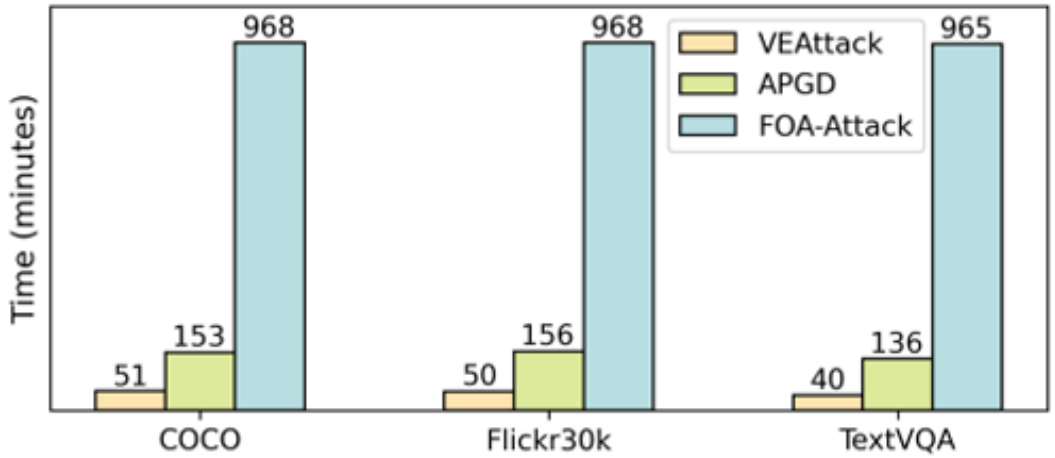
[1]City University of Hong Kong

[2]University of Sydney

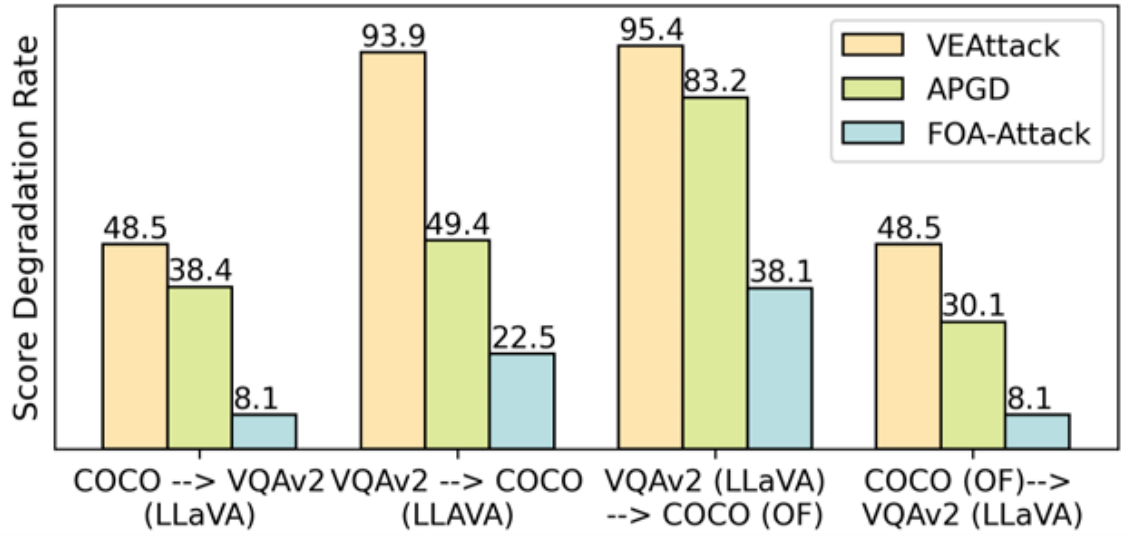
hefeimei2-c@my.cityu.edu.hk

Background

- LVLMs typically serve as foundation models for diverse multimodal tasks.
- Two stubborn challenges for attacking LVLMs lie in efficiency and task generalization.
- Given that traditional adversarial attacks focus on specific tasks and the increasing parameter scale of LVLMs, VEAttack aims to alleviate task-specific dependency and reduce the computational overhead.
- Black-box attack FOA-Attack struggles to achieve a high score reduction rate (SRR) of LVLM performance, despite employing larger perturbations.



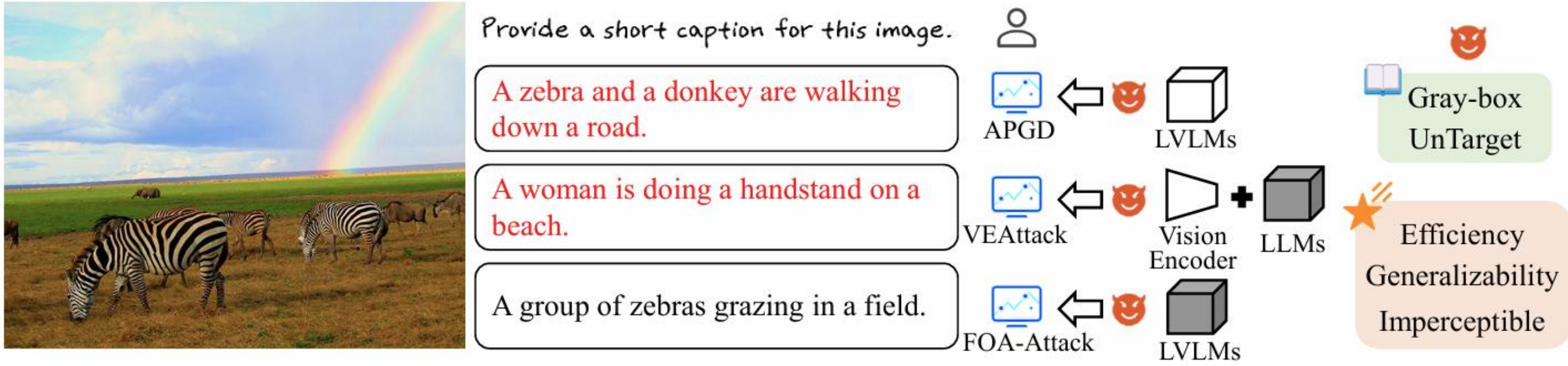
(a) Time consumption of VEAttack, APGD and FOA-Attack.



(b) Performance degradation ratio after VEAttack, APGD and FOA-Attack.

Motivation

- A gray-box attack targeting a foundational LVLM module provides better trade-offs among efficiency, transferability, and practicality since only partial parameters are required and overfitting to a specific LVLM task is relaxed.
- Drawing from traditional vision tasks, where a strong vision backbone enhances downstream task success, and recognizing the critical role of the vision encoder in LVLMs for overall performance, we target the vision encoder of LVLMs for attack.



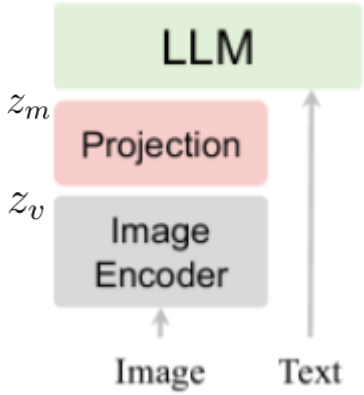
Redefine adversarial objective

- White-box Attack:

$$\max_{\|\delta\| \leq \epsilon} \mathcal{L}\left(v+\delta, t; \theta\right) = \max_{\|\delta\| \leq \epsilon} -\log p\left(y \mid f_A\left(f_V(v+\delta)\right); tokenizer(t); \theta\right),$$

$$\text{multi-task } T = \{T^1, \cdots, T^m\}$$

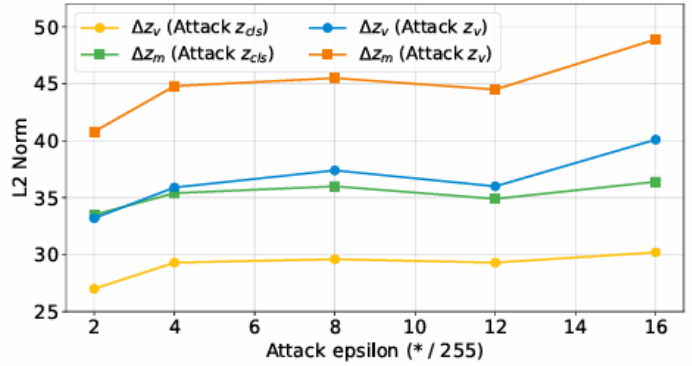
$$\max_{\|\delta^{T^i}\| \leq \epsilon} \mathcal{L}\left(\tilde{v}^{T^i}, t^{T^i}; \theta\right) = \max_{\|\delta^{T^i}\| \leq \epsilon} -\log p\left(y^{T^i} \mid f_A\left(f_V(\tilde{v}^{T^i})\right); tokenizer(t^{T^i}); \theta\right),$$



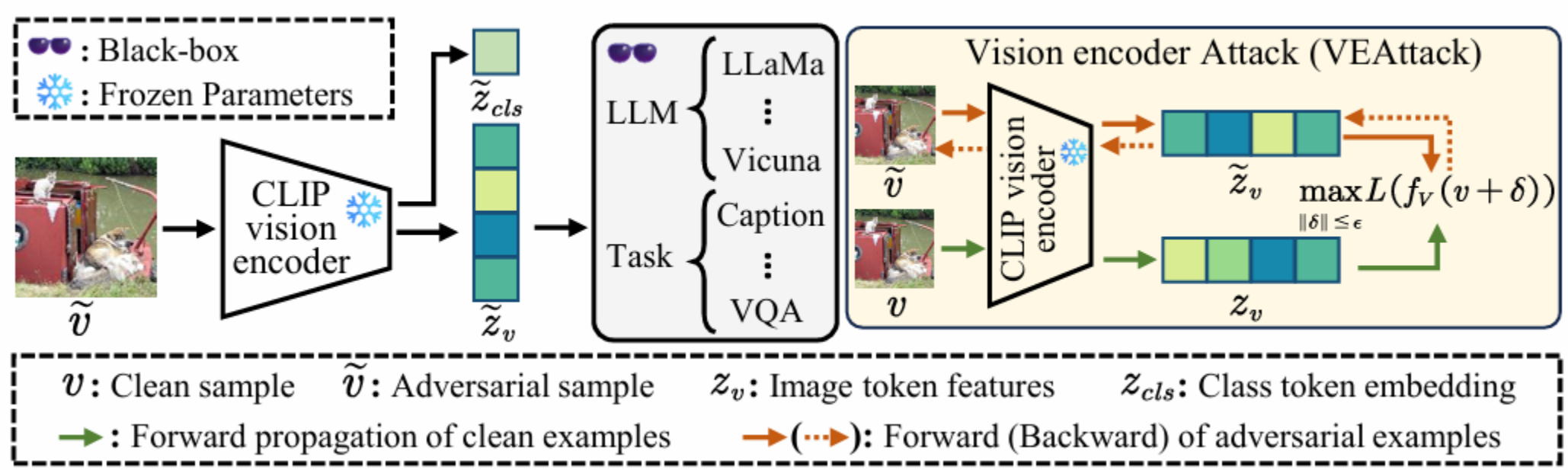
Proposition 2 *For LLaVa [57] with a linear alignment layer, let $\Delta z_v = \tilde{z}_v - z_v$ denote the difference between the image tokens output by the vision encoder CLIP before and after the perturbation, $\|\Delta z_v\|_F \geq \Delta$, W_a is the weight of projection layer, σ_{min} denotes the minimum singular value of W_a. Assume $\sigma_{min}(W_a) > 0$, then the difference between aligned features $z_m = f_A(z_v)$ and $\tilde{z}_m = f_A(\tilde{z}_v)$ for downstream LLMs will satisfy $\|\Delta z_m\|_F \geq \sigma_{min}(W_a)\Delta > 0$.*

- Redefined objective:

$$\max_{\|\delta^{T^i}\| \leq \epsilon} \mathcal{L}\left(v^{T^i} + \delta^{T^i}, t^{T^i}; \theta\right) \longrightarrow \max_{\|\delta\| \leq \epsilon} \mathcal{L}(v+\delta; \theta_V) = \max_{\|\delta\| \leq \epsilon} \mathcal{L}(f_{CLIP}(v+\delta)),$$



Framework



- we can treat the attacks in adversarial training algorithms proposed by black-box attack and robust CLIP work:

$$\text{AttackVLM-ii}: \ \tilde{v} = \arg\max_{\|\delta\| \le \epsilon} -\cos\left(\tilde{z}_{cls}, z_{cls}\right) \qquad L_2 \text{ Attack}: \ \tilde{v} = \arg\max_{\|\delta\| \le \epsilon} \|\tilde{z}_{cls} - z_{cls}\|_2^2,$$

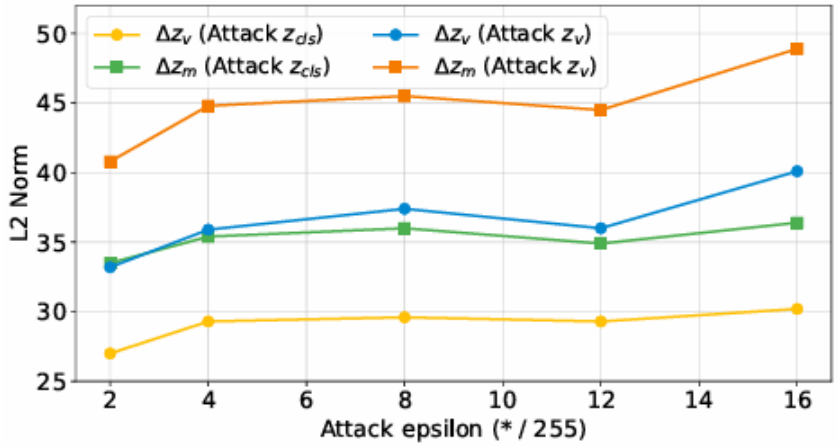
Vision encoder attack

Proposition 3 *Consider two kinds of attack targets: 1) If the perturbation is introduced to the single class token z_{cls} and propagates through $z_{cls} \xrightarrow{backward} v \xrightarrow{forward} z_v$. Let $\Delta z_{cls}(z_{cls})$, $\Delta z_v(z_{cls})$, $\Delta z_m(z_{cls})$ denote the perturbations on features z_{cls}, z_v and z_m through the first attack, respectively. 2) If the perturbation is directly introduced to the image token features z_v, let $\Delta z_m(z_v)$ denote the perturbation on the aligned feature z_m through the second attack. Assume the same degree of perturbation on z_{cls} and z_v during the two attacks, then the ratio of the effect is given by:*

$$\frac{\|\Delta z_m(z_{cls})\|_F}{\|\Delta z_m(z_v)\|_F} = \frac{\|\Delta z_v(z_{cls})\|_F}{\|\Delta z_{cls}(z_{cls})\|_2} \leq \frac{3+\epsilon_V}{\sqrt{n_v}}, \ \epsilon_V \ll 1. \tag{32}$$

- VEAttack Objective:

$$\tilde{v} = \underset{\|\delta\| \leq \epsilon}{\arg\max} - \cos\left(f_V(v+\delta), f_V(v)\right) = \underset{\|\delta\| \leq \epsilon}{\arg\max} - \cos(z_v + \delta, z_v).$$



Experiments

Task	LVLMs / Attack	OpenFlamingo-9B $\epsilon = 2/255$	OpenFlamingo-9B $\epsilon = 4/255$	LLaVa1.5-7B $\epsilon = 2/255$	LLaVa1.5-7B $\epsilon = 4/255$	LLaVa1.5-13B $\epsilon = 2/255$	LLaVa1.5-13B $\epsilon = 4/255$	Average $\epsilon = 2/255$	Average $\epsilon = 4/255$
COCO	Clean	79.7		115.5		119.2		104.8(↓0.0%)	
	MIX.Attack Tu et al. (2023)	45.9	25.4	67.5	55.4	73.8	60.1	62.4(↓40.5)	47.0(↓55.1)
	VT-Attack Wang et al. (2024c)	38.9	21.6	50.8	12.2	58.2	20.1	49.3(↓52.9)	18.0(↓82.8)
	AttackVLM-ii Zhao et al. (2023)	24.4	10.3	40.9	25.8	42.7	27.5	36.0(↓65.6)	21.2(↓79.8)
	VEAttack	**7.5**	**3.7**	**10.8**	**7.1**	**11.2**	**6.5**	**9.8(↓90.6)**	**5.8(↓94.5)**
Flickr30k	Clean	60.1		77.5		77.1		71.6	
	MIX.Attack Tu et al. (2023)	33.7	18.0	42.1	35.9	41.0	32.2	38.9(↓45.7)	28.7(↓59.9)
	VT-Attack Wang et al. (2024c)	27.7	13.8	35.1	12.3	34.0	14.6	32.3(↓54.9)	13.6(↓81.0)
	AttackVLM-ii Zhao et al. (2023)	18.1	9.9	29.9	19.8	29.7	21.6	25.9(↓63.8)	17.1(↓76.1)
	VEAttack	**8.7**	**3.2**	**10.7**	**6.3**	**9.1**	**5.7**	**9.5(↓86.7)**	**5.1(↓92.9)**
TextVQA	Clean	23.8		37.1		39.0		33.3	
	MIX.Attack Tu et al. (2023)	13.4	8.8	24.6	19.1	22.8	19.7	20.3(↓39.0)	15.9(↓52.3)
	VT-Attack Wang et al. (2024c)	15.3	10.5	23.7	**10.0**	24.9	10.0	21.3(↓36.0)	10.2(↓69.4)
	AttackVLM-ii Zhao et al. (2023)	**12.1**	7.6	19.7	11.9	19.8	14.2	17.2(↓48.3)	11.2(↓66.4)
	VEAttack	12.5	**5.7**	**13.8**	10.1	**12.4**	**8.6**	**12.9(↓61.3)**	**8.1(↓75.7)**
VQAv2	Clean	48.5		74.5		75.6		66.2	
	MIX.Attack Tu et al. (2023)	39.8	36.0	59.4	57.9	59.8	58.0	53.0(↓19.9)	50.6(↓23.6)
	VT-Attack Wang et al. (2024c)	38.5	37.0	53.6	**21.4**	55.2	**21.0**	49.1(↓25.8)	**26.5(↓59.9)**
	AttackVLM-ii Zhao et al. (2023)	37.5	35.8	54.1	49.7	56.2	49.6	49.3(↓25.5)	45.0(↓32.0)
	VEAttack	**34.0**	**32.8**	**42.9**	38.4	**41.5**	37.6	**39.5(↓40.3)**	36.3(↓45.2)
POPE	Clean	65.7		84.5		84.1		78.1	
	MIX.Attack Tu et al. (2023)	59.0	53.3	72.0	69.1	74.2	68.2	68.4(↓12.4)	63.5(↓18.7)
	VT-Attack Wang et al. (2024c)	63.6	63.5	64.0	60.1	65.5	66.4	64.4(↓17.5)	63.3(↓18.9)
	AttackVLM-ii Zhao et al. (2023)	53.3	48.1	69.0	61.6	63.8	57.8	49.3(↓36.9)	45.0(↓42.4)
	VEAttack	**60.6**	**59.6**	**47.5**	**42.8**	**47.6**	**44.7**	**51.9(↓33.5)**	**49.0(↓37.3)**

Table 11: Ablation of VEAttack targets on diverse tasks with cosine similarity as the loss metric.

Objective z_v	Objective z_{cls}	COCO (ϵ)↓ $2/255$	COCO (ϵ)↓ $4/255$	VQAv2 (ϵ)↓ $2/255$	VQAv2 (ϵ)↓ $4/255$	POPE (ϵ)↓ $2/255$	POPE (ϵ)↓ $4/255$
✗	✓	43.6	25.5	56.0	50.0	69.4	59.1
✓	✓	22.0	10.5	46.4	42.0	59.2	46.0
✓	✗	**10.8**	**7.1**	**42.9**	**38.4**	**47.5**	**42.8**

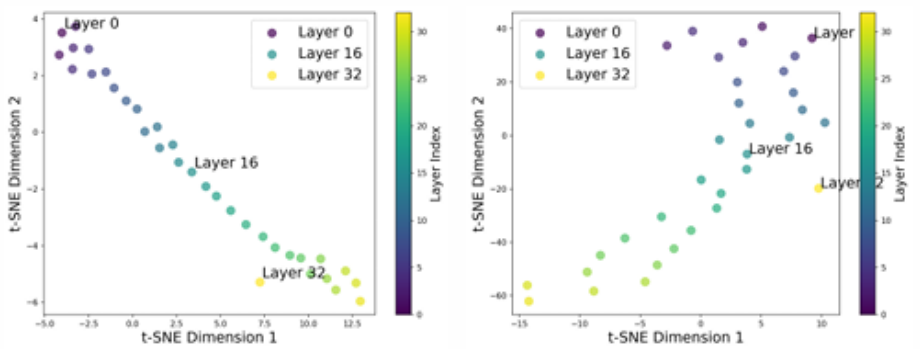
Table 12: Ablation of loss metrics in attack objective.

Measurement	COCO (ϵ)↓ $2/255$	COCO (ϵ)↓ $4/255$	VQAv2 (ϵ)↓ $2/255$	VQAv2 (ϵ)↓ $4/255$	POPE (ϵ)↓ $2/255$	POPE (ϵ)↓ $4/255$
Euclidean	46.9	39.6	56.1	53.6	62.1	64.3
K-L divergence	70.0	34.3	59.4	**33.2**	70.0	60.5
Cosine similarity	**10.8**	**7.1**	**42.9**	38.4	**47.5**	**42.8**

Observation

Observation 1 *Even though the LLMs and tasks are downstream-agnostic, attacks on output of the vision encoder can lead to variations in the hidden layer features of LLMs.*

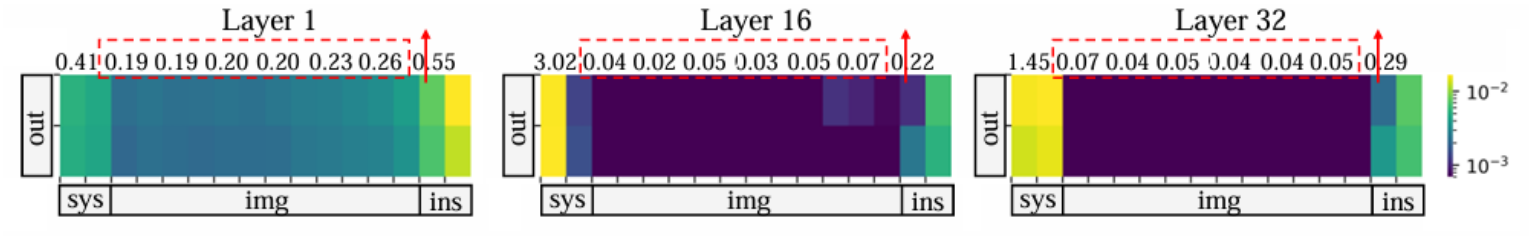
Observation 2 *LVLMs will pay more attention to image tokens in the image caption task. Conversely, LVLMs focus more on user instruction tokens in the VQA task.*



(a) t-SNE of the hidden layers when the input is a clean image

(b) t-SNE of the hidden layers when the input is an adversarial image

Figure 11: t-SNE visualization of the first visual tokens across hidden layers in the LLM model for clean and adversarial image inputs.



(a) The attention score between output token and different tokens when the task is COCO caption.

(b) The attention score between output token and different tokens when the task is VQAv2.

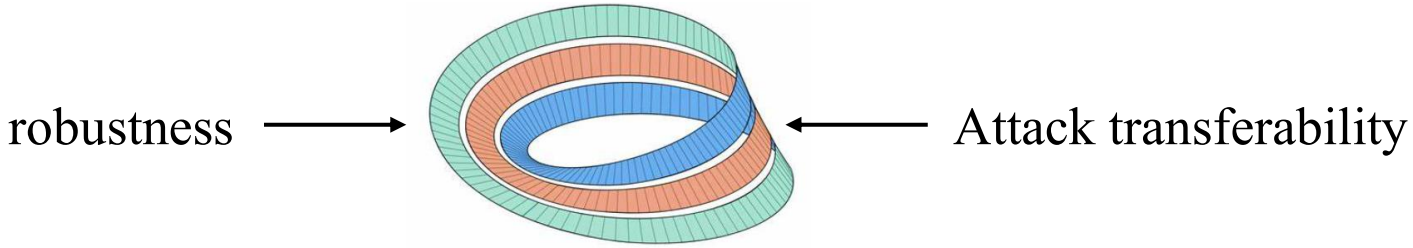
Figure 12: Illustration of attention maps across different layers of the language model for two tasks. The attention scores shown represent the relationship between the output token and three input token types (out: output tokens, sys: system tokens, img: image tokens, ins: instruction tokens).

Observation

Observation 3 *The transfer attack on LVLMs normally resembles a Möbius band, where robustness and vulnerability of LVLMs intertwine as a single, twisted continuum. For defenders, using a more robust vision encoder can enhance the ability of LVLM to resist attacks. Conversely, for attackers, adversarial samples obtained by attacking a more robust vision encoder usually have higher attack transferability on diverse LVLMs.*

Table 2: Transfer Attack of original CLIP and robust CLIP after adversarial training on the COCO dataset. Bold indicates the best attack performance, while underlined indicates the second best.

Source \ Target	LLaVa1.5-7B			OpenFlamingo-9B			MiniGPT-4	mPLUG-Owl2	Qwen-VL
	CLIP	TeCoA	FARE	CLIP	TeCoA	FARE	BLIP-2	MplugOwl	CLIP-bigG
clean	115.5	88.3	102.4	79.7	66.9	74.1	96.7	132.3	138.5
CLIP Radford et al. (2021)	**3.7**	92.2	100.4	**3.6**	70.7	78.1	55.2	124.9	131.2
TeCoA Mao et al. (2023)	22.1	**25.3**	<u>25.1</u>	17.7	**20.8**	<u>22.6</u>	**11.0**	<u>96.4</u>	<u>102.7</u>
FARE Schlarmann et al. (2024)	<u>38.7</u>	<u>54.5</u>	**48.6**	<u>26.3</u>	<u>42.7</u>	**36.4**	<u>8.9</u>	**88.0**	**101.4**



Observation

Observation 4 *Reducing attack steps does not significantly impair the effectiveness of VEAttack, while increasing the perturbation budget beyond a threshold within an imperceptible range does not continuously degrade the performance of LVLMs significantly.*

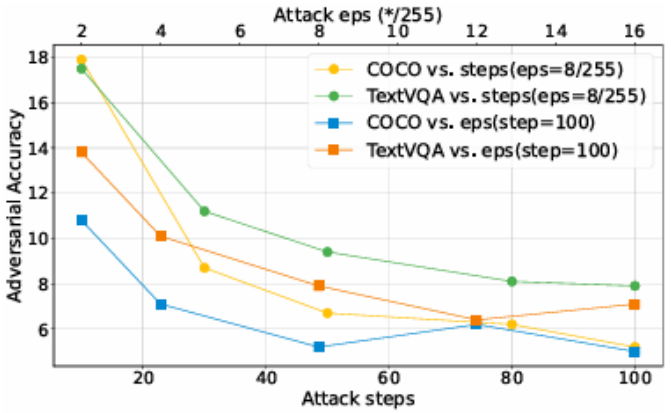


Figure 13: VEAttack performance under different attack steps and perturbation budgets on COCO and TextVQA datasets (top x-axis: perturbation budget, bottom x-axis: attack steps).

Table 3: Comparison of the effectiveness and efficiency between VEAttack and other white-box and gray-box attacks. Flops count the computation of forward once.

Attack	Version	Flops	COCO	Time (h)	Flickr30k	Time (h)	TextVQA	Time (h)
clean	None	99.3G	115.5	1.33	77.5	0.3	37.1	0.42
APGD Croce & Hein (2020)	White-box	9.93T	13.1	25.5	9.5	5.2	8.1	22.7
Ensemble Schlarmann & Hein (2023)	White-box	9.93T	**3.1**	41.9	**1.2**	14.3	**0.0**	37.8
VT-Attack Wang et al. (2024c)	Gray-box	3.04T	12.2	65.0	12.3	13.3	10.0	64.8
VEAttack ($\epsilon = 4/255, t = 100$)	Gray-box	2.59T	7.1	8.5	6.3	1.7	10.1	6.7
VEAttack ($\epsilon = 8/255, t = 50$)	Gray-box	2.59T	5.5	**5.3**	4.5	**1.1**	9.4	**4.1**

ICLR

香港城市大學
City University of Hong Kong

THE UNIVERSITY OF SYDNEY

Thank you for your listening