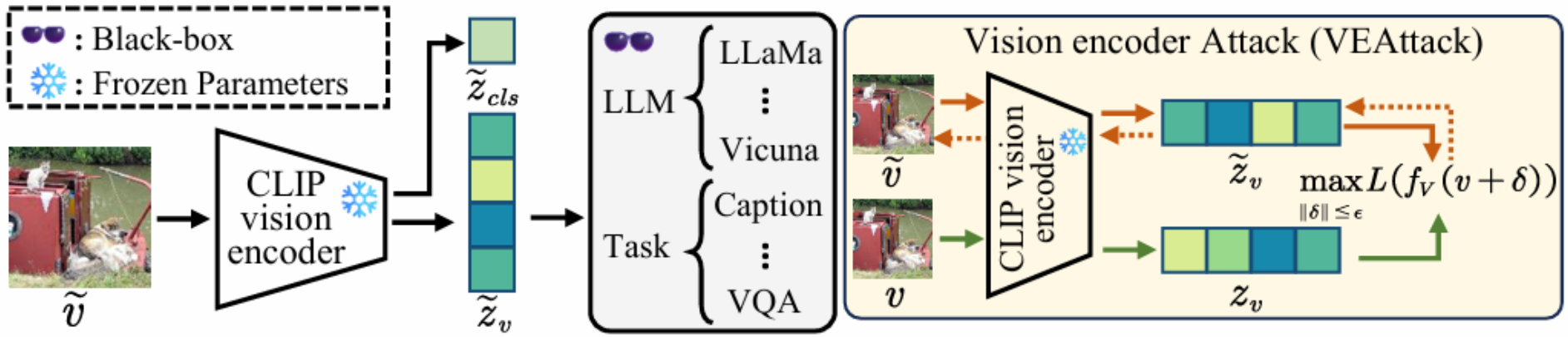
PA-Attack: Guiding Gray-Box Attacks on LVLM Vision Encoders with Prototypes and Attention

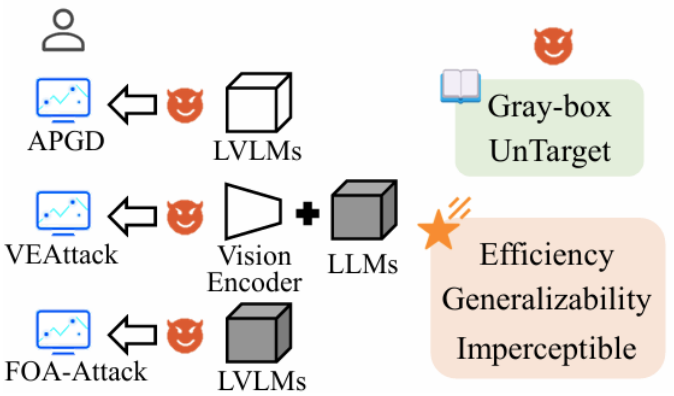
Hefei Mei[1], Zirui Wang[1], Chang Xu[2], Jianyuan Guo[1], Minjing Dong[1*]

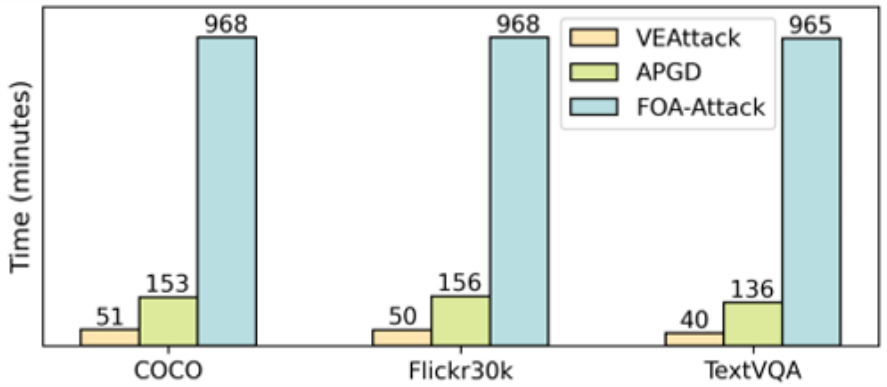
[1]City University of Hong Kong [2]The University of Sydney

{hefeimei2-c, zrwang23-c}@my.cityu.edu.hk
c.xu@sydney.edu.au, {jianyguo, minjdong}@cityu.edu.hk

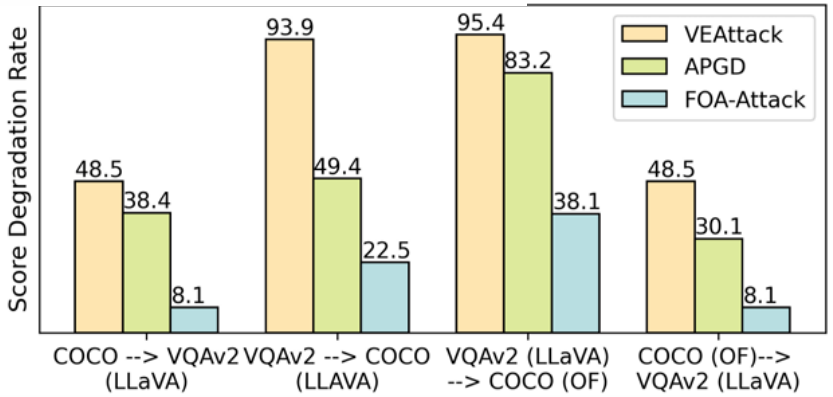
VEAttack







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



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

Most LVLMs, such as LLaVA, Yi-VL, are built by pairing a shared vision backbone (e.g. CLIP) with different LLMs, which makes the vision encoder a common component across diverse LVLMs.

The vision encoder typically contains significantly fewer parameters than LLM modules, thereby boosting efficiency.

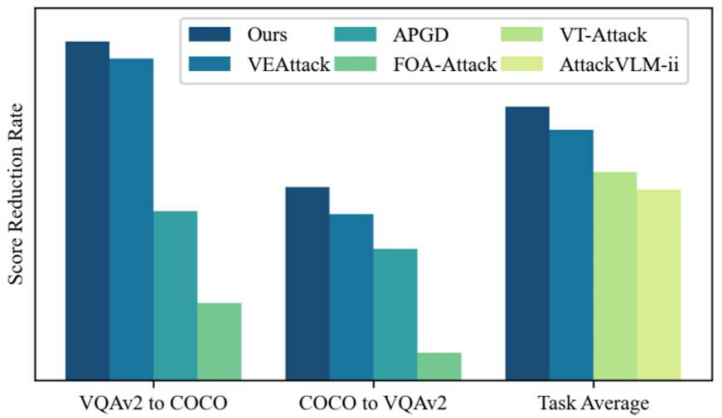
Perturbations crafted to disrupt the vision encoder could be more generalizable across LVLM tasks.

PA-Attack

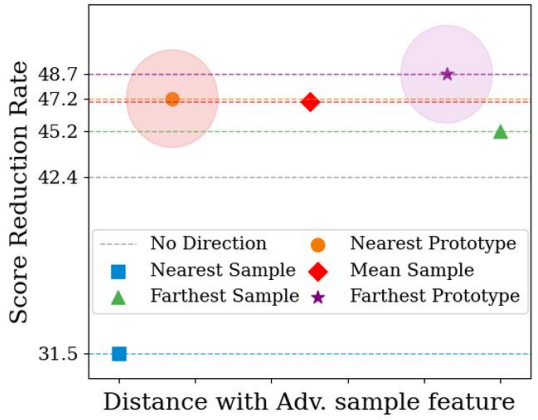
With lower attack iterations and perturbations, it is difficult for existing gray-box attacks to attack LVLMs in different tasks.

We show significant token redundancy in performance impact.

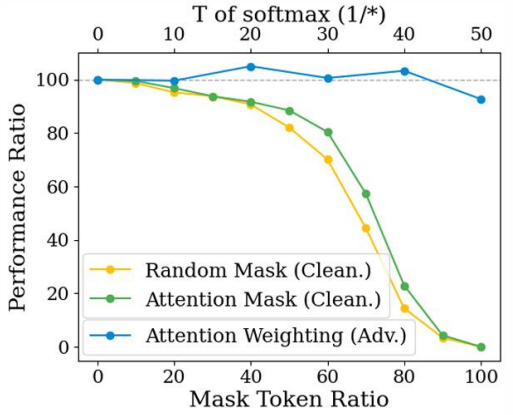
Preserving higher attention maintains better performance than random one.



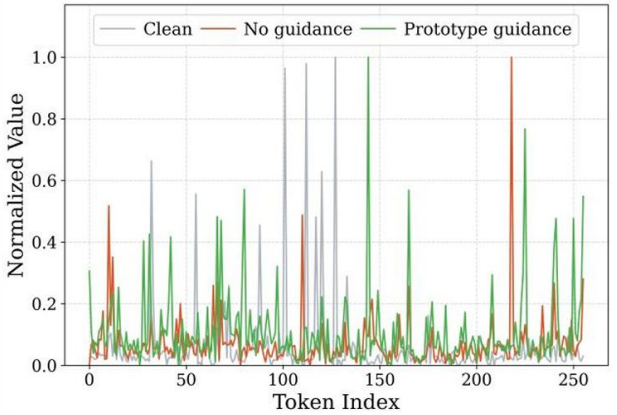
(a) Attack performance with task transfer



(b) Attack with direction samples



(c) Clean and Adv. performance changes



(d) Attention value before and after attack

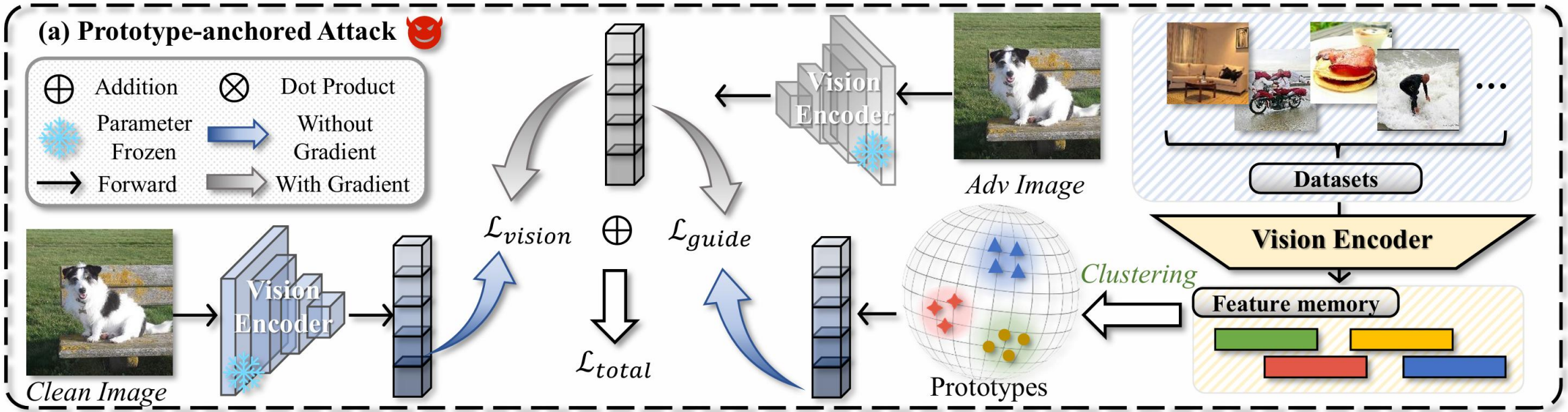
The furthest prototype achieves the best attack effect by generating more general guidance.

A few specific tokens dominate the attack.

The generated adversarial example can hardly generalize to tasks that focus on different visual attributes.

PA-Attack

$$\mathcal{L} = -\frac{1}{N}\sum_j \boldsymbol{w}_j \cdot [-\cos(\boldsymbol{v}_j, \boldsymbol{v'}_j) + \lambda \cdot \cos(\boldsymbol{v'}_j, \mathbf{p}_j^{k^*})].$$



$$\boldsymbol{x'}_{i+1} \leftarrow \boldsymbol{x'}_i + \alpha \cdot \mathrm{sign}(\nabla_{\boldsymbol{x'}_i}\mathcal{L}(f(\boldsymbol{x}), f(\boldsymbol{x'}_i), \boldsymbol{w}_{s1}, \mathbf{p}^{k^*})),$$

$$\boldsymbol{x'}_{r+1} \leftarrow \boldsymbol{x'}_r + \alpha \cdot \mathrm{sign}(\nabla_{\boldsymbol{x'}_r}\mathcal{L}(f(\boldsymbol{x}), f(\boldsymbol{x'}_r), \boldsymbol{w}_{s2}, \mathbf{p}^{k^*})).$$

PA-Attack

Algorithm 1 PA-Attacking Procedure

Require: clean image $\boldsymbol{x}$, perturbation budget ϵ, stage-one iterations S_1, stage-two iterations S_2, loss function $\mathcal{L}$, prototype $\mathcal{P}$ from guidance dataset $\mathbb{D}_{guide}$, step size α, random initialization range $\boldsymbol{\eta} \in [0, \epsilon]$.

$\boldsymbol{x'}_0 \leftarrow \boldsymbol{x} + \text{Uniform}(-\boldsymbol{\eta}, \boldsymbol{\eta})$ ▷ Random start

$k^* = \arg\min_k \cos\left(\boldsymbol{v}, \mathbf{p}^k\right), \mathbf{p}^k \in \mathcal{P}$ ▷ Index in Eq. (5)

$\boldsymbol{w}_{s1} = \text{softmax}(\boldsymbol{a}^l \leftarrow f(\boldsymbol{x}))$ ▷ Token weights Eq. (9)

for $i = 0$ to $S_1 - 1$ **do**

  Compute $\mathcal{L}(f(\boldsymbol{x}), f(\boldsymbol{x'}_i), \boldsymbol{w}_{s1}, \mathbf{p}^{k^*})$ in Eq. (10)

  $\boldsymbol{x'}_{i+1} \leftarrow \boldsymbol{x'}_i + \alpha \cdot \text{sign}(\nabla_{\boldsymbol{x'}_i}\mathcal{L})$ ▷ Eq. (11)

  $\boldsymbol{x'}_{i+1} \leftarrow \text{clip}(\boldsymbol{x'}_{i+1}, \boldsymbol{x} - \epsilon, \boldsymbol{x} + \epsilon)$ ▷ To ℓ_∞-ball

end for

$\boldsymbol{w}_{s2} = \text{softmax}(\boldsymbol{a}^l \leftarrow f(\boldsymbol{x'}_{S_1-1}))$ ▷ Token weights

for $r = S_1$ to $S_1 + S_2 - 1$ **do**

  Compute $\mathcal{L}(f(\boldsymbol{x}), f(\boldsymbol{x'}_r), \boldsymbol{w}_{s2}, \mathbf{p}^{k^*})$ in Eq. (10)

  $\boldsymbol{x'}_{r+1} \leftarrow \boldsymbol{x'}_r + \alpha \cdot \text{sign}(\nabla_{\boldsymbol{x'}_r}\mathcal{L})$ ▷ Eq. (12)

  $\boldsymbol{x'}_{r+1} \leftarrow \text{clip}(\boldsymbol{x'}_{r+1}, \boldsymbol{x} - \epsilon, \boldsymbol{x} + \epsilon)$ ▷ To ℓ_∞-ball

end for

return $\boldsymbol{x}_{\text{adv}}$; ▷ $\boldsymbol{x'}_{S_1+S_2-1} \rightarrow \boldsymbol{x}_{\text{adv}}$

Model	Attack Perturbation (ϵ)	COCO ↓ 2/255	COCO ↓ 4/255	Flickr30k ↓ 2/255	Flickr30k ↓ 4/255	TextVQA ↓ 2/255	TextVQA ↓ 4/255	VQAv2 ↓ 2/255	VQAv2 ↓ 4/255	POPE ↓ 2/255	POPE ↓ 4/255	Average SRR ↑ 2/255	Average SRR ↑ 4/255
LLaVa1.5-7B	Clean	115.5		77.5		37.1		74.5		84.5		0.0	
	MIX.Attack [40]	67.5	55.4	42.1	35.9	24.6	19.1	59.4	57.9	72.0	69.1	31.2	38.9
	VT-Attack [43]	66.8	29.5	40.0	19.9	26.9	16.1	60.1	**28.3**	67.1	58.5	31.6	59.6
	AttackVLM-ii [51]	41.3	26.3	30.2	20.3	19.7	14.0	55.1	50.4	69.0	61.6	43.3	54.5
	VEAttack [33]	10.8	7.1	10.7	6.3	13.8	10.1	42.9	38.4	47.5	42.8	65.2	71.2
	PA-Attack (ours)	**6.1**	**4.1**	**4.7**	**3.3**	**8.3**	**5.1**	**33.9**	32.5	**29.6**	**33.6**	**77.1**	**79.0**
OF-9B	Clean	79.7		60.1		23.8		48.5		65.7		0.0	
	MIX.Attack [40]	45.9	25.4	33.7	18.0	13.4	8.8	39.8	36.0	59.0	53.3	31.6	49.2
	VT-Attack [43]	49.7	31.6	38.6	23.0	16.1	13.2	39.7	39.5	65.1	62.7	25.0	37.9
	AttackVLM-ii [51]	27.7	10.1	20.1	9.3	10.0	7.4	37.8	33.3	53.3	48.1	46.1	59.8
	VEAttack [33]	7.5	3.7	8.7	**3.2**	12.5	5.7	34.0	32.8	60.6	59.6	52.3	61.5
	PA-Attack (ours)	**4.7**	**3.6**	**5.4**	3.7	**6.6**	**4.8**	**33.0**	**32.6**	**47.6**	**45.9**	**63.4**	**66.4**
LLaVa1.5-13B	Clean	119.2		77.1		39.0		75.6		84.1		0.0	
	MIX.Attack [40]	73.8	60.1	41.0	32.2	22.8	19.7	59.8	58.0	74.2	68.2	31.8	39.9
	VT-Attack [43]	68.6	34.0	40.7	25.2	28.7	18.4	59.8	**27.6**	70.9	67.9	30.5	54.9
	AttackVLM-ii [51]	46.6	25.3	27.5	17.9	19.6	14.7	55.0	49.9	63.8	57.8	45.3	56.6
	VEAttack [33]	11.2	6.5	9.1	5.7	12.4	8.6	41.5	37.6	47.6	44.7	67.1	72.4
	PA-Attack (ours)	**5.4**	**3.9**	**4.1**	**2.0**	**7.0**	**6.5**	**36.5**	34.4	**31.4**	**27.9**	**77.3**	**79.8**

Method	Qwen3-VL-8B RealWorldQA	Qwen3-VL-8B ODinW-13	InternVL2-8B RealWorldQA	InternVL2-8B POPE	SRR ↑
Clean	73.3	35.5	62.5	88.6	0.0
VT-Attack	68.2	33.9	45.4	77.4	12.9
VEAttack	58.2	14.4	46.4	**63.8**	33.4
PA-Attack	**57.0**	**9.8**	**43.4**	64.8	**38.0**

Guidance	Distribution	COCO	Flickr30k	TextVQA	SRR ↑
Clean	In	115.5	77.5	37.1	0.0
COCO	In	4.1	3.3	5.1	92.8
RVL-CDIP	Out	4.1	3.7	6.8	91.1
ScienceQA	Out	4.2	3.6	5.8	92.0

CVPR JUNE 3-7, 2026 DENVER COLORADO

CityU 香港城市大學 City University of Hong Kong

THE UNIVERSITY OF SYDNEY

Thank you for your listening!