中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES
智能算法安全全国重点实验室
State Key Laboratory of AI Safety

多模态大模型对抗攻防：从可控攻击到测试时防御

张杰

中科院计算所-智能信息处理重点实验室

智能算法安全全国重点实验室

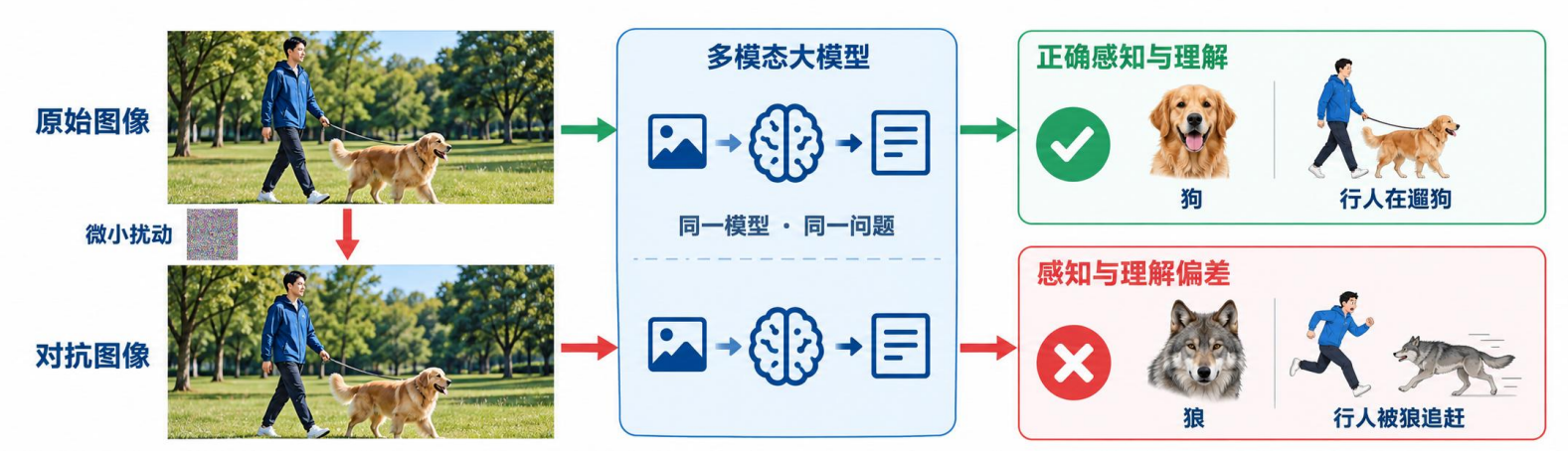
2026年10月21日

报告提纲

- 研究背景
- 基于值引导的可控黑盒对抗攻击（V-Attack）
- 高可迁移性的黑盒对抗攻击（O-Attack）
- 基于频域对比矫正的测试时对抗防御（CSR）
- 总结与展望

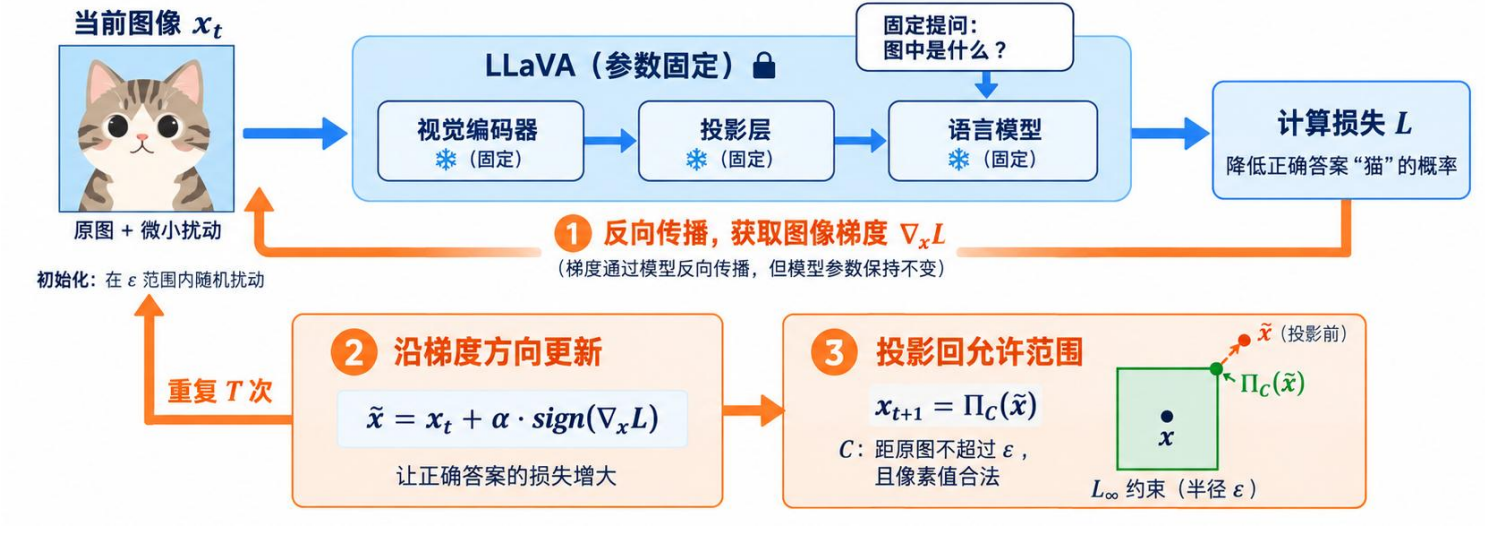
研究背景

- 多模态大模型（MLLMs）的内生安全问题
  - 对抗攻击 是指通过精心设计的输入扰动（通常人类难以察觉），诱使模型产生错误预测或非预期行为的攻击方式。



相关工作

- 白盒 对抗攻击
  - 定义：攻击者 完全掌握 目标模型的内部信息（网络结构、权重参数、梯度），据此直接计算梯度来生成扰动样本，使模型输出错误结果。



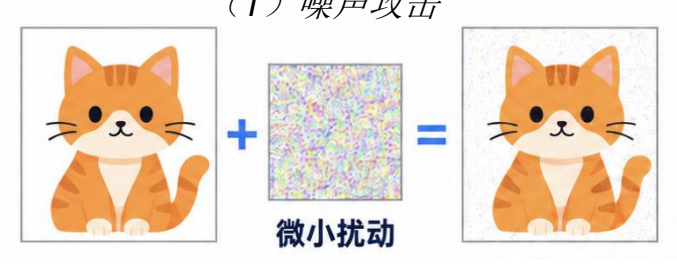
* 对抗攻击 PGD 算法示意图，如上所示

相关工作（攻击）

- 白盒 对抗攻击
  - 分类：噪声攻击、排版攻击

Method	Year	Category	Subcategory	Target model
Co-Attack (Zhang *et al.*, 2022)	2022	White-box	Invisible	ALBEF, TCL, CLIP
AdvCLIP (Zhou *et al.*, 2023)	2023	White-box	Invisible	CLIP
Typographical Attacks (Noever and Noever, 2021)	2021	White-box	Visible	CLIP
Multi-Image Typographical Attacks (Wang *et al.*, 2025e)	2025	White-box	Visible	OpenCLIP, InstructBLIP
Caption Attack (Schlarmann and Hein, 2023)	2023	White-box	Task-specific +V	OpenFlamingo
VisBreaker (Cui *et al.*, 2024b)	2023	White-box	Task-specific +V	LLaVA/BLIP-2/ InstructBLIP
CroPA (Luo *et al.*, 2024a)	2024	White-box	Cross-prompt +VL	OpenFlamingo/ BLIP-2/ Instruct BLIP
GroundBreaker (Gao *et al.*, 2024b)	2024	White-box	Task-specific +V	MiniGPT-v2
Stop-reasoning Attack (Wang *et al.*, 2024r)	2024	White-box	CoT attack+V	MiniGPT-4/ OpenFlamingo/ LLaVA

（1）噪声攻击

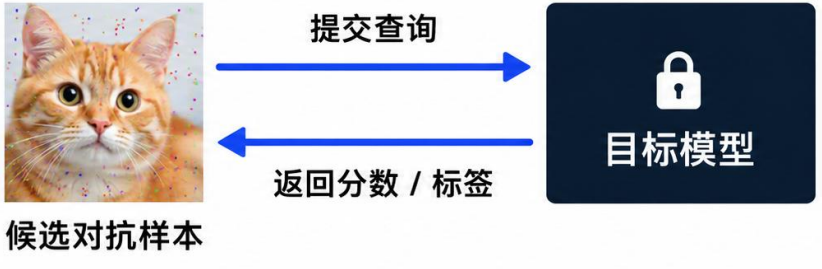


（2）排版攻击



相关工作（攻击）

- 黑盒 对抗攻击
  - 定义：攻击者**完全无法获知**目标模型的内部信息（网络结构、权重参数、梯度），只能通过 (a) 反复查询其输入输出、或 (b) 借助替代模型与迁移性来构造扰动样本，使模型输出错误结果。



根据反馈调整扰动，反复查询

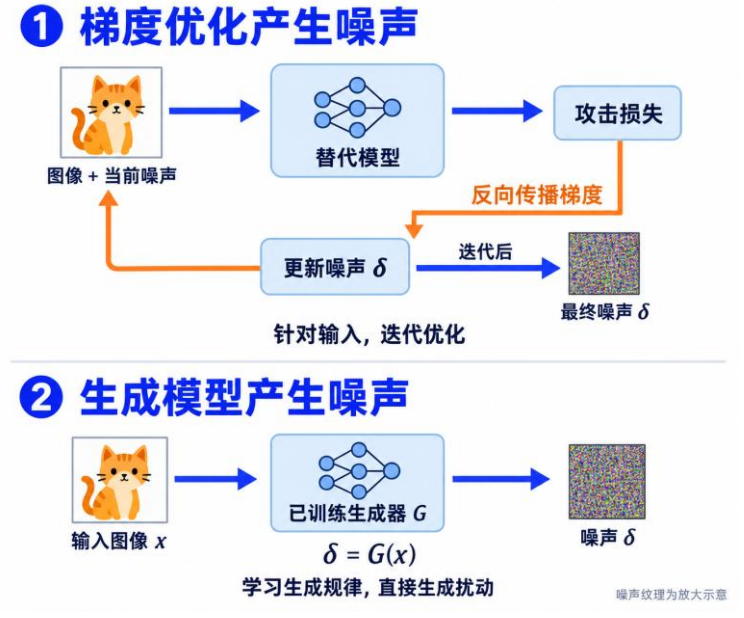


在替代模型上生成，迁移至目标模型

相关工作（攻击）

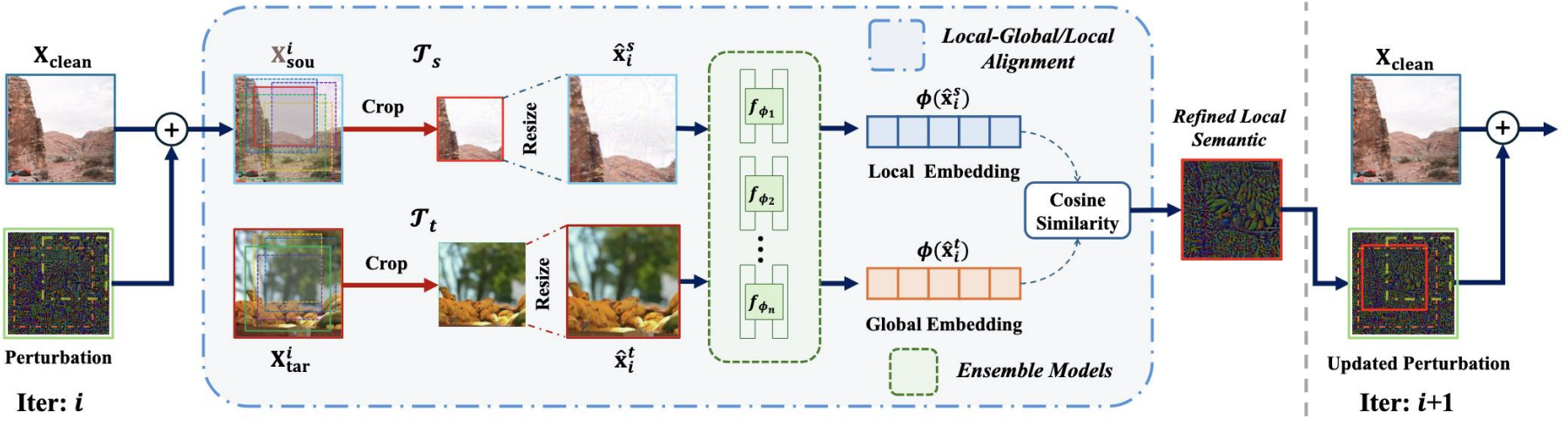
- 黑盒 对抗攻击
  - 分类：基于梯度优化噪声、用生成模型生成噪声

Method	Year	Category	Subcategory	Target models
Attack Bard (Dong *et al.*, 2023)	2023	Black-box	Transfer-based +V	Bard/GPT-4V/Bing Chat/ ERNIE Bot
AttackVLM (Zhao *et al.*, 2024f)	2024	Black-box	Transfer-based +V	BLIP/UniDiffuser/ Img2Prompt/BLIP-2/ LLaVA/MiniGPT-4
DynVLA (Gu *et al.*, 2025)	2025	Black-box	Transfer-based +VL	InstructBLIP/MiniGPT4/ LLaVA/Gemini
AdvDiffVLM (Guo *et al.*, 2024b)	2024	Black-box	Generator-based+V	MiniGPT-4/LLaVA/ UniDiffuser/MiniGPT-4/ BLIP/BLIP-2/Img2LLM
AnyAttack (Zhang *et al.*, 2024j)	2024	Black-box	Generator-based+V	CLIP/BLIP/BLIP2/ InstructBLIP/ MiniGPT-4
CAVALRY-V (Zhang *et al.*, 2025d)	2025	Black-box	Generator-based+V	GPT-4.1/Gemini/ QwenVL/InternVL/ LLaVA/Aria/ MiniCPM



黑盒攻击经典工作

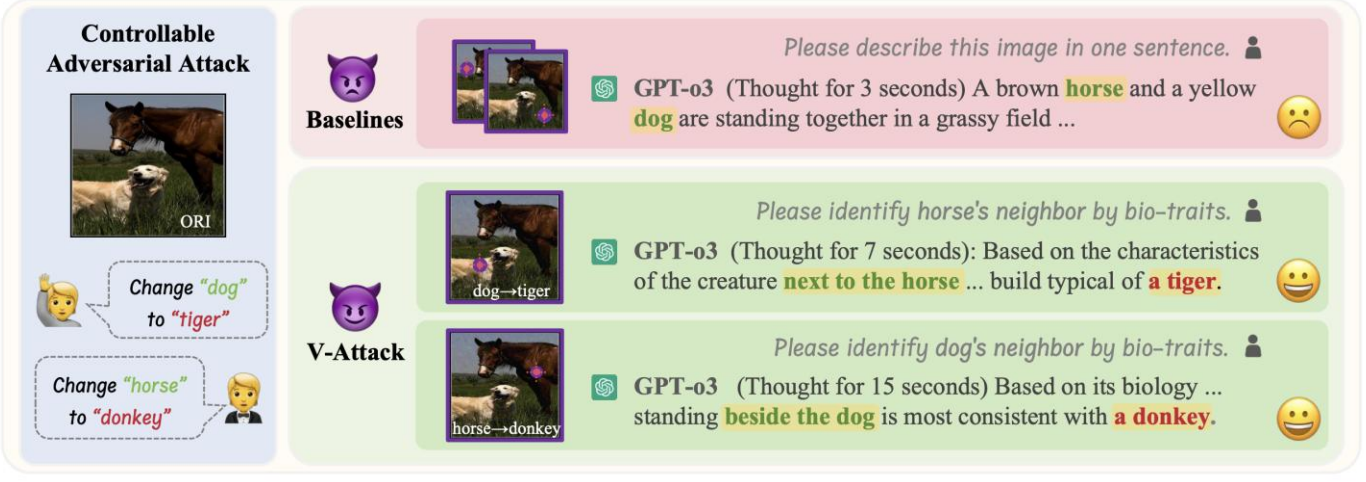
- M-Attack
  - 基于梯度优化，代理模型为多个CLIP，采用了数据增强



Li, Zhaoyi; Zhao, Xiaohan; Wu, Dong-Dong; Cui, Jiacheng; Shen, Zhiqiang. A Frustratingly Simple Yet Highly Effective Attack Baseline: Over 90% Success Rate Against the Strong Black-box Models of GPT-4.5/4o/o1. In: Advances in Neural Information Processing Systems (NeurIPS), 2025.

基于值引导的可控黑盒对抗攻击（V-Attack）

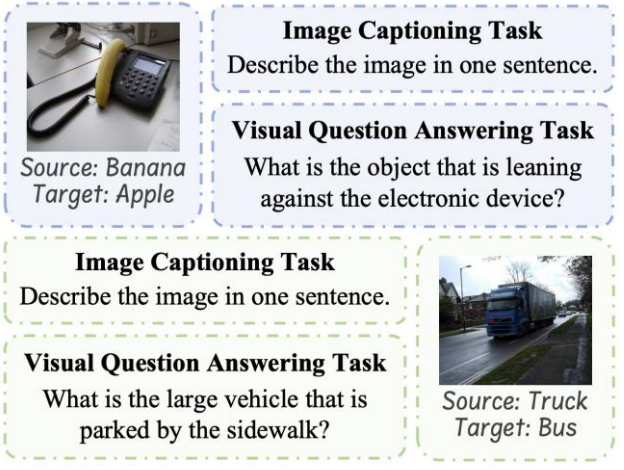
- 问题：
  - 对抗攻击的目标从"让任务模型分错类"转向"操纵 LVLM 对图像语义的理解"，但现有方法**可控性差**——无法精确改写图像中某个指定概念。



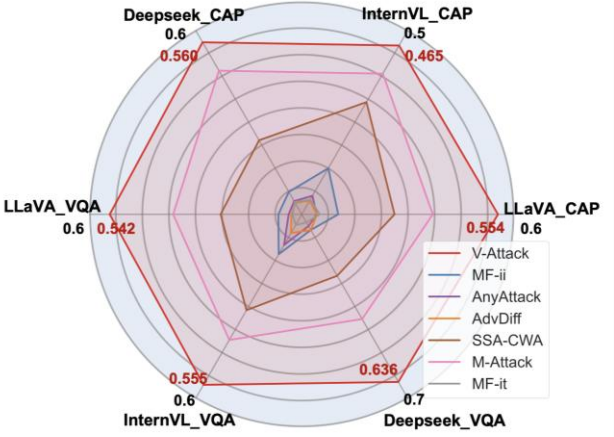
Sen Nie, Jie Zhang, Jianxin Yan, Shiguang Shan, and Xilin Chen. "V-Attack: Targeting Disentangled Value Features for Controllable Adversarial Attacks on LVLMs." *CVPR 2026*

基于值引导的可控黑盒对抗攻击（V-Attack）

- 探究实验：
  - 我们发现现有方法在精确局部语义攻击任务上表现很差



(a) Local Semantic Attack.

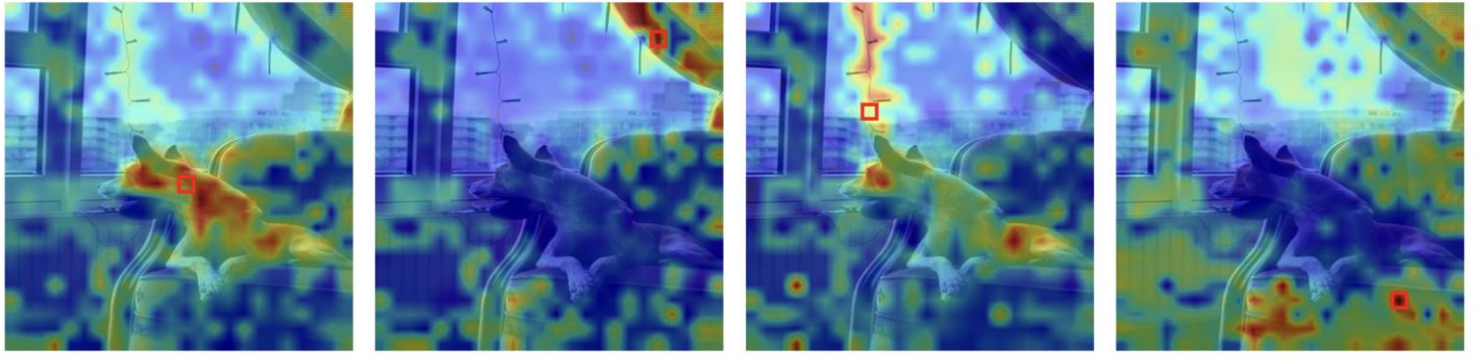


(b) Comparative performance.

Sen Nie, Jie Zhang, Jianxin Yan, Shiguang Shan, and Xilin Chen. "V-Attack: Targeting Disentangled Value Features for Controllable Adversarial Attacks on LVLMs." *CVPR 2026*

基于值引导的可控黑盒对抗攻击（V-Attack）

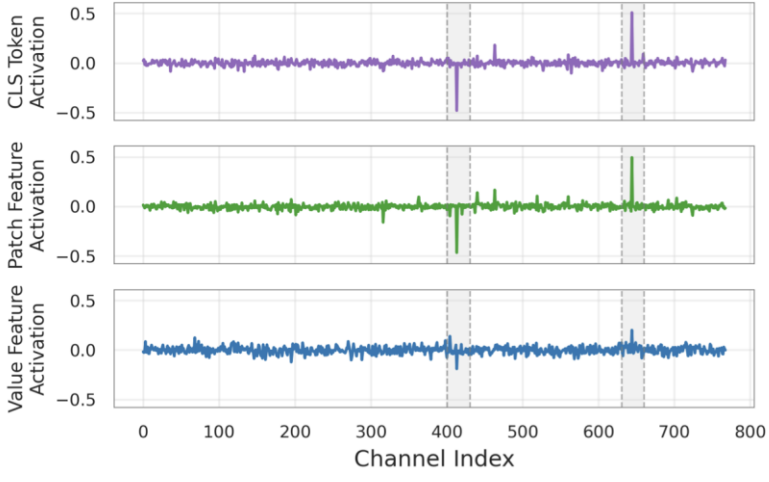
- 现有攻击方法表现差的原因分析：
  - 注意力机制使得特征出现“**语义杂糅**”现象
  - 视觉特征不适合做为对抗攻击时的优化代理特征

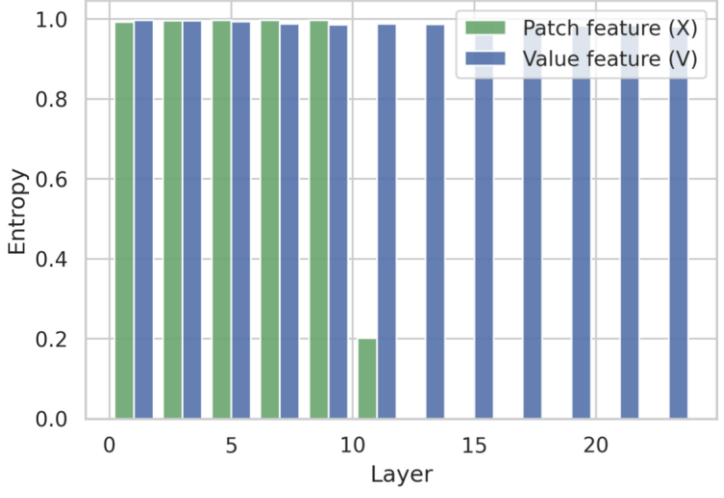


图中的注意力图为红色小块对所有块的注意力分布图

基于值引导的可控黑盒对抗攻击（V-Attack）

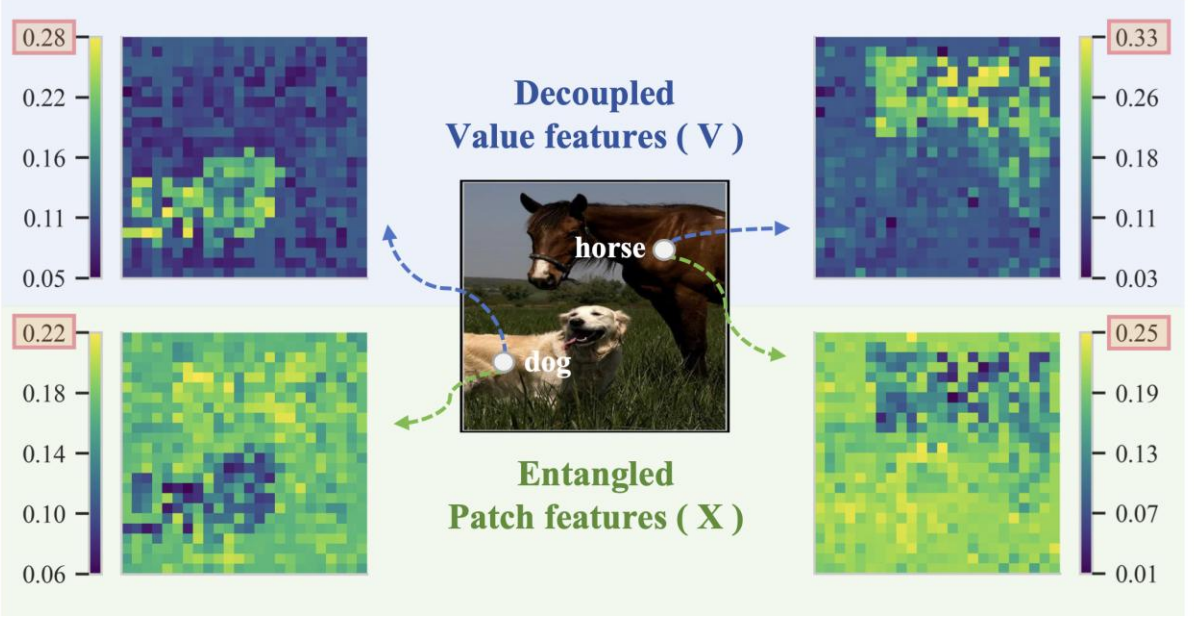
- 通道分析与熵值分析
  - 视觉特征（X）被少量高激活通道主导，且熵值在后半层开始骤降
  - 注意力机制中的值特征（V），通道值分布均匀、熵值自始至终稳定





基于值引导的可控黑盒对抗攻击（V-Attack）

- 图文对齐分析
  - V比X有更好的跨模态对齐性质、且每块的语义更为单一



- 发现一

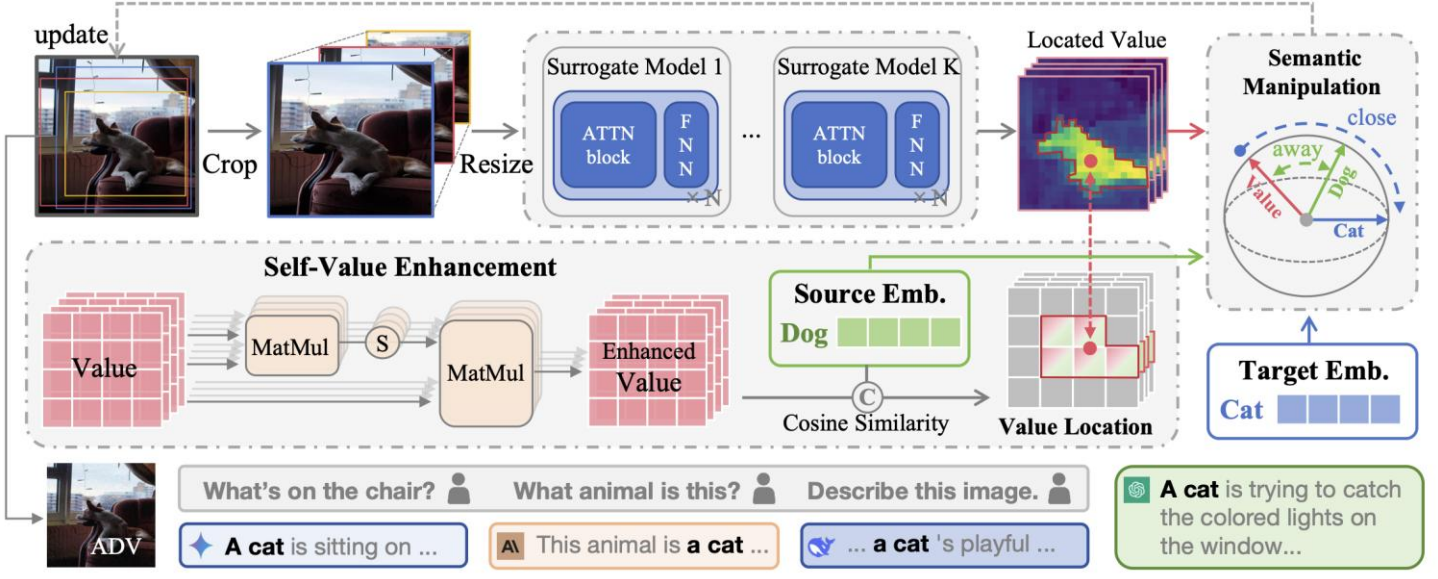
跨模态余弦相似度更高

- 发现二

对齐的空间关系更清晰

基于值引导的可控黑盒对抗攻击（V-Attack）

- 方法
  - 代理模型：用多个CLIP作为代理模型
  - 损失函数：让增强后的 Value 特征远离 源文本特征 接近 目标文本特征



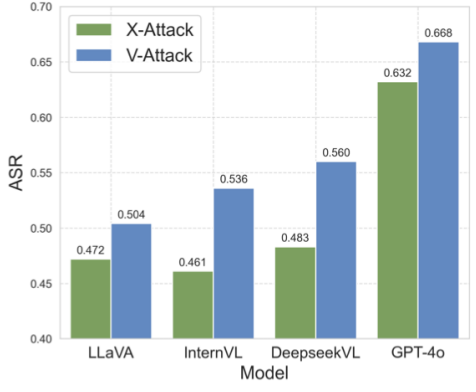
基于值引导的可控黑盒对抗攻击（V-Attack）

- 攻击效果
  - 用GPT-4o评判攻击 图片摘要任务 和 视觉问答任务 的成功率
  - V-Attack 相对 SOTA方法 相对提升36%

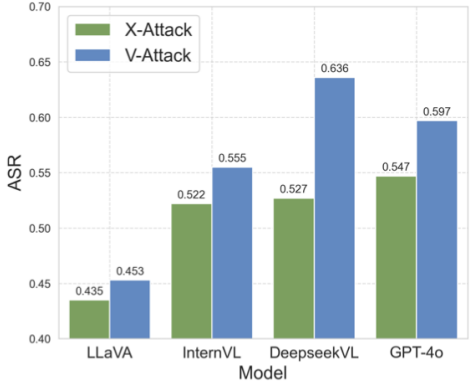
Method	Train	Aug.	LLaVA*		InternVL		DeepseekVL		GPT-4o		Avg.		Imperceptibility	
			CAP	VQA	CAP	VQA	CAP	VQA	CAP	VQA	CAP	VQA	ℓ_1 (↓)	ℓ_2 (↓)
MF-it	Sgl	–	0.051	0.026	0.040	0.035	0.040	0.031	0.028	0.033	0.046	0.031	0.036	0.041
MF-ii	Sgl	–	0.103	0.066	0.125	0.130	0.073	0.060	0.039	0.045	0.114	0.078	0.032	0.039
AnyAttack	Ens	✓	0.042	0.037	0.050	0.100	0.043	0.045	0.030	0.030	0.045	0.058	0.088	0.101
AdvDiff	Ens	✓	0.043	0.029	0.037	0.061	0.037	0.060	0.031	0.028	0.039	0.054	0.106	0.113
SSA-CWA	Ens	✓	0.262	0.229	0.304	0.312	0.241	0.233	0.285	0.270	0.273	0.270	0.085	0.091
M-Attack	Ens	✓	0.370	0.363	0.405	0.409	0.483	0.398	0.544	0.475	0.450	0.411	0.079	0.085
V-Attack(ours)	Sgl	✓	**0.554**	**0.542**	0.283	0.352	0.210	0.240	0.445	0.391	0.354	0.355	**0.030**	**0.038**
V-Attack(ours)	Ens	✓	0.504	0.453	**0.536**	**0.555**	**0.560**	**0.636**	**0.668**	**0.597**	**0.567**	**0.560**	0.074	0.079

基于值引导的可控黑盒对抗攻击（V-Attack）

- 消融分析与案例分析
  - 我们提出 除攻击特征不同外其余完全相同的 攻击方法 X-Attack
  - **V-Attack 全面领先 X-Attack**，表明我们选取Value特征的正确性



(a) Image Captioning task results.



(b) VQA task results.

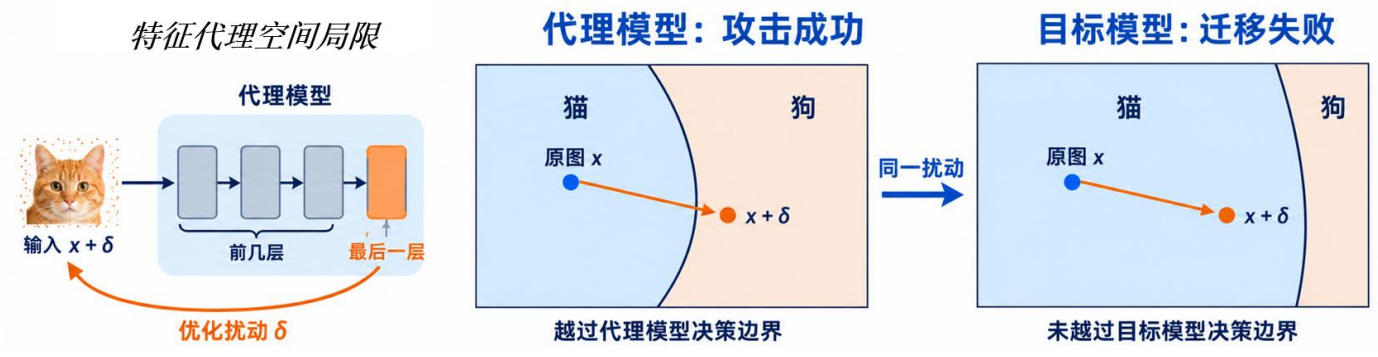




V-Attack在GPT-4o上的例子，图为对抗样本

高可迁移性的黑盒对抗攻击（O-Attack）

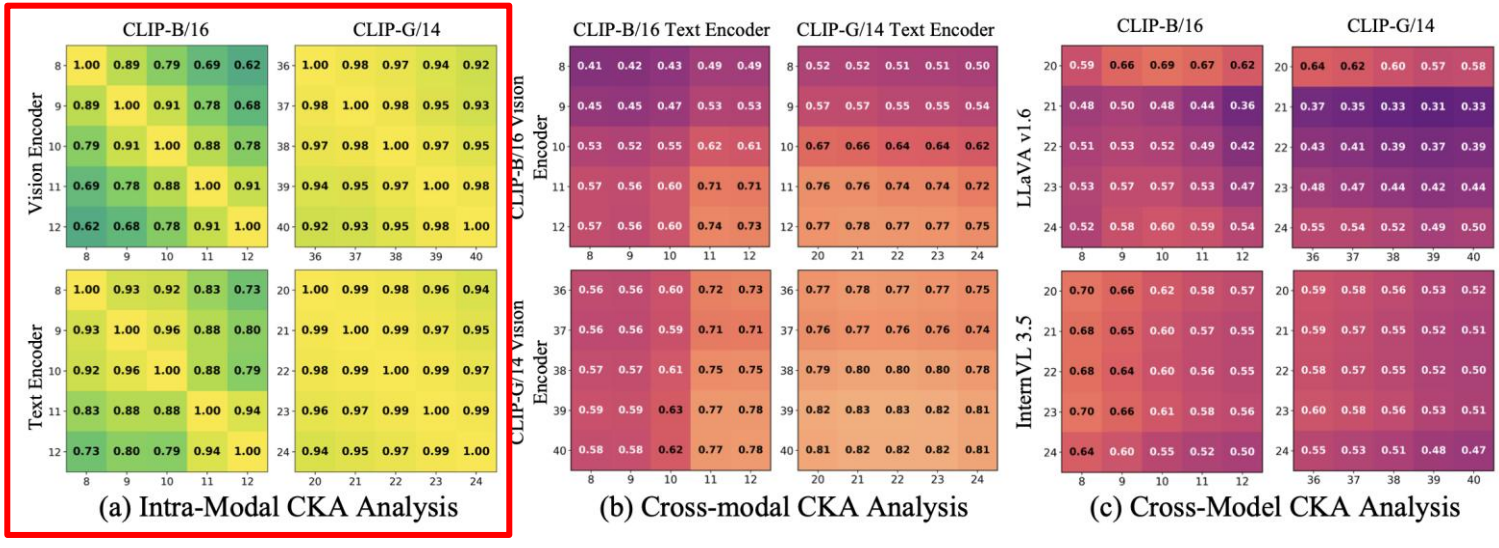
- 问题分析：
  - 现有迁移式黑盒攻击以代理模型的末层特征为代理，使得对抗噪声高度拟合代理模型的决策空间，生成的对抗样本迁移到其他模型上表现很差
  - 探索模型内部更广泛的特征空间具有提升对抗样本可迁移性的潜力



Sen Nie, Jie Zhang, Zhongqi Wang, Shiguang Shan, and Xilin Chen. “One Attack to Fool Them All: Highly Transferable Black-Box Adversarial Attacks on Frontier MLLM.” arXiv preprint arXiv:2609.33833.

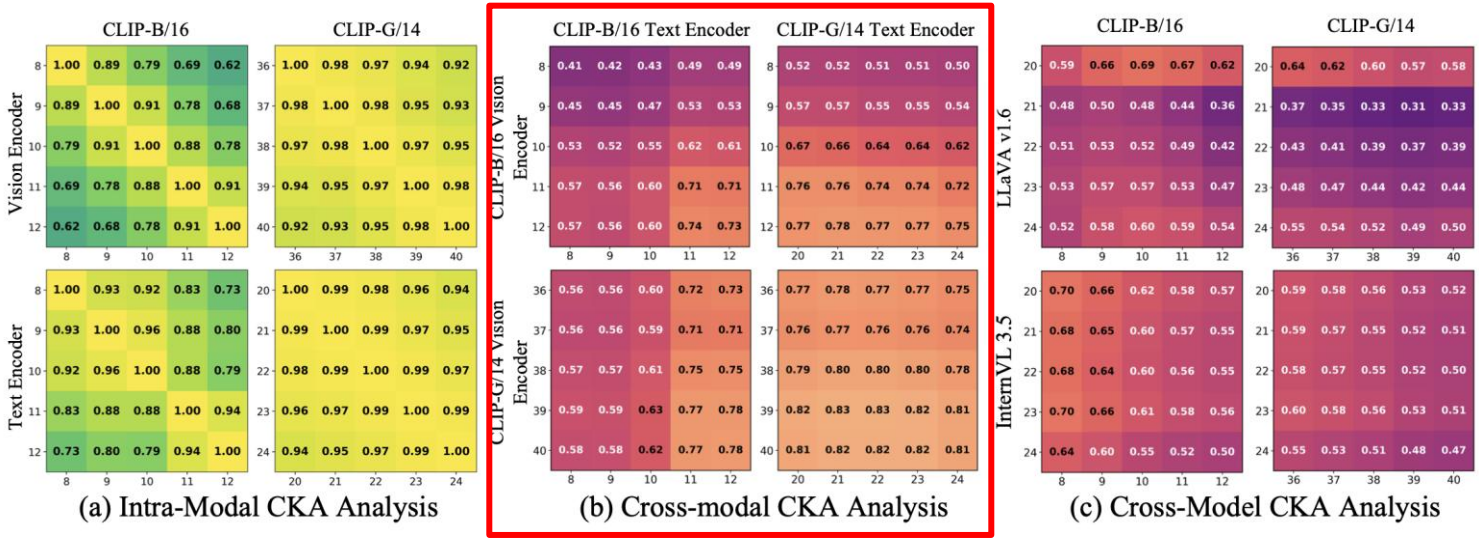
高可迁移性的黑盒对抗攻击（O-Attack)

- **模态内**CKA分析:
  - （视觉/文本）编码器内的末N层之间的CKA相似度分数大于 0.9
  - 这表明编码器的末N层特征的结构相似度很高



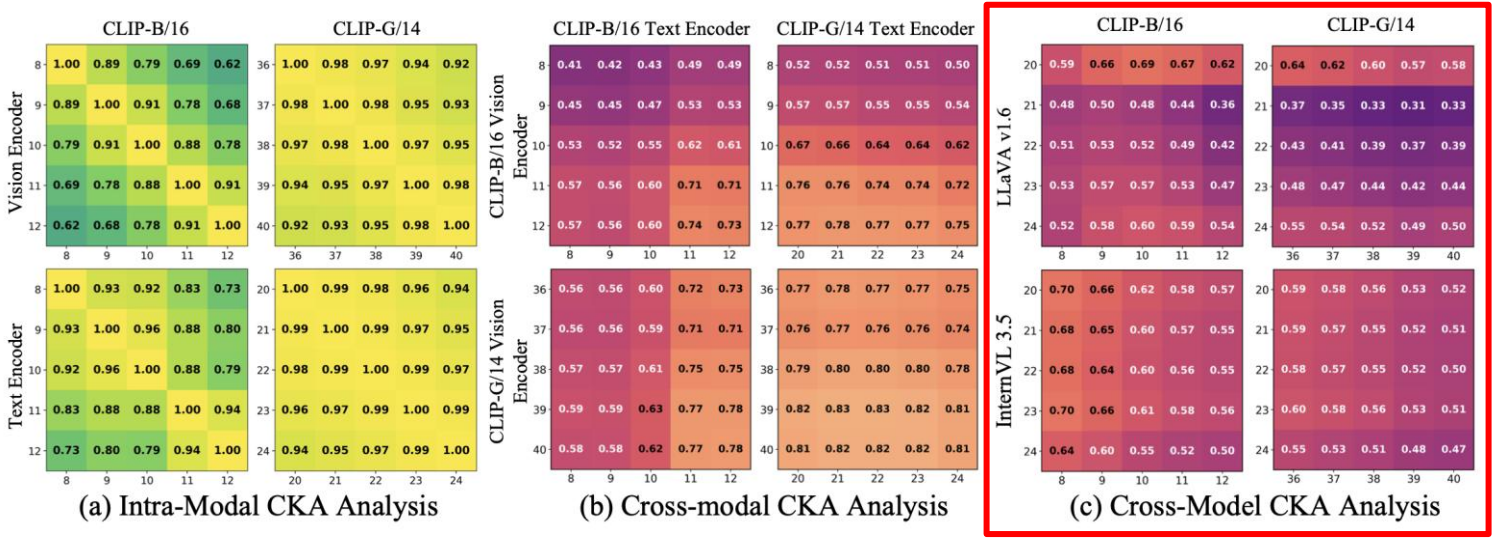
高可迁移性的黑盒对抗攻击（O-Attack）

- **模态间**CKA分析:
  - 视觉编码器的末N层和文本编码器的末N层之间CKA相似度都很高
  - 这表明跨模态对齐可能不局限于视觉/文本编码器的最后一层



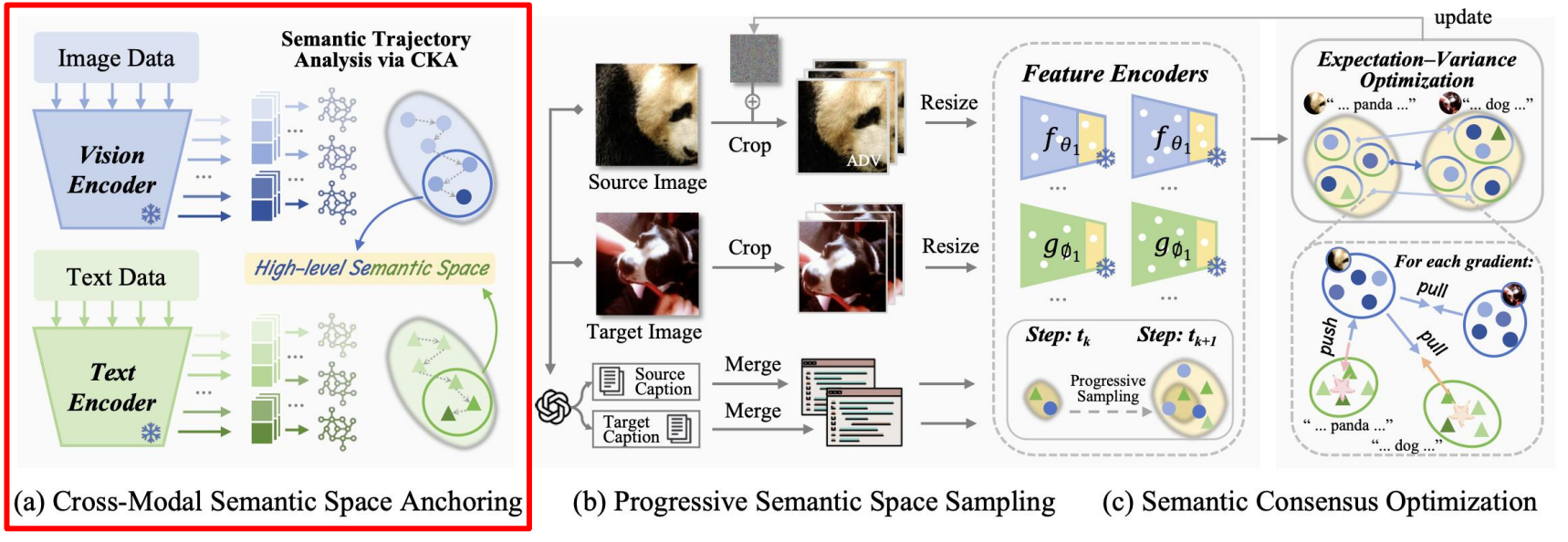
高可迁移性的黑盒对抗攻击（O-Attack）

- 代理-受害模型间 CKA分析:
  - 代理模型和受害模型的非末层CKA分数高于末层CKA分数
  - 这表明中间层特征可能在代理模型和受害模型间有更强迁移性



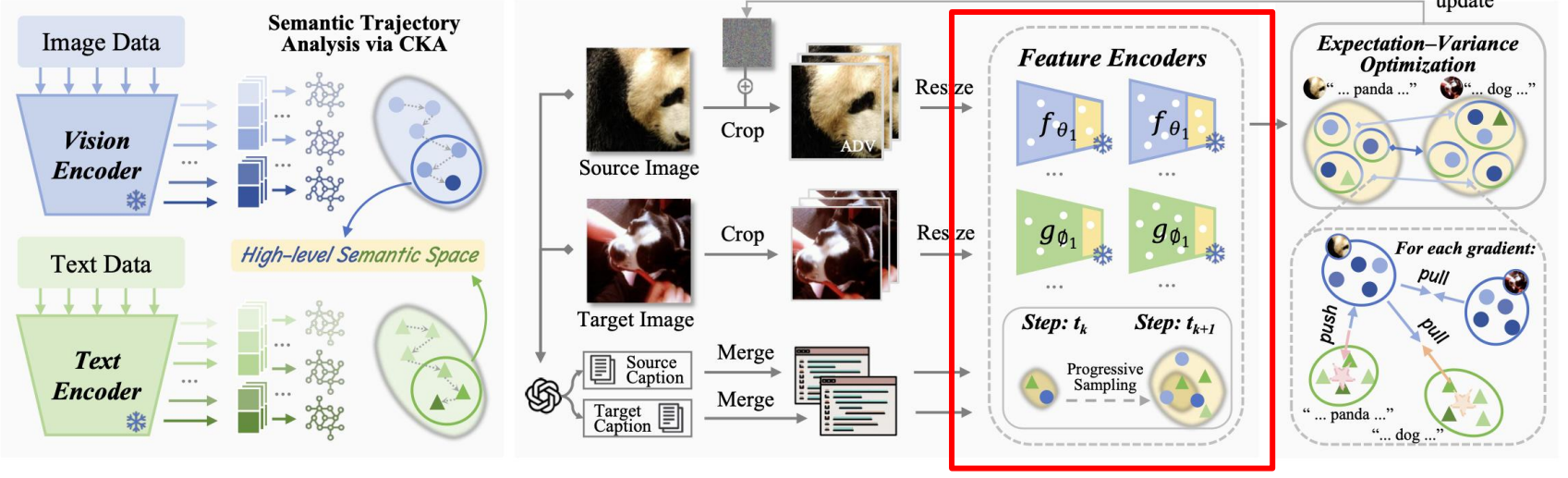
高可迁移性的黑盒对抗攻击（O-Attack）

- 模块一：CSA
  - O-Attack 用离线 CKA 轨迹分析 在每个（视觉/文本）编码器内锚定一段连续的 末N层 特征集合，作为后续损失函数计算的代理



高可迁移性的黑盒对抗攻击（O-Attack）

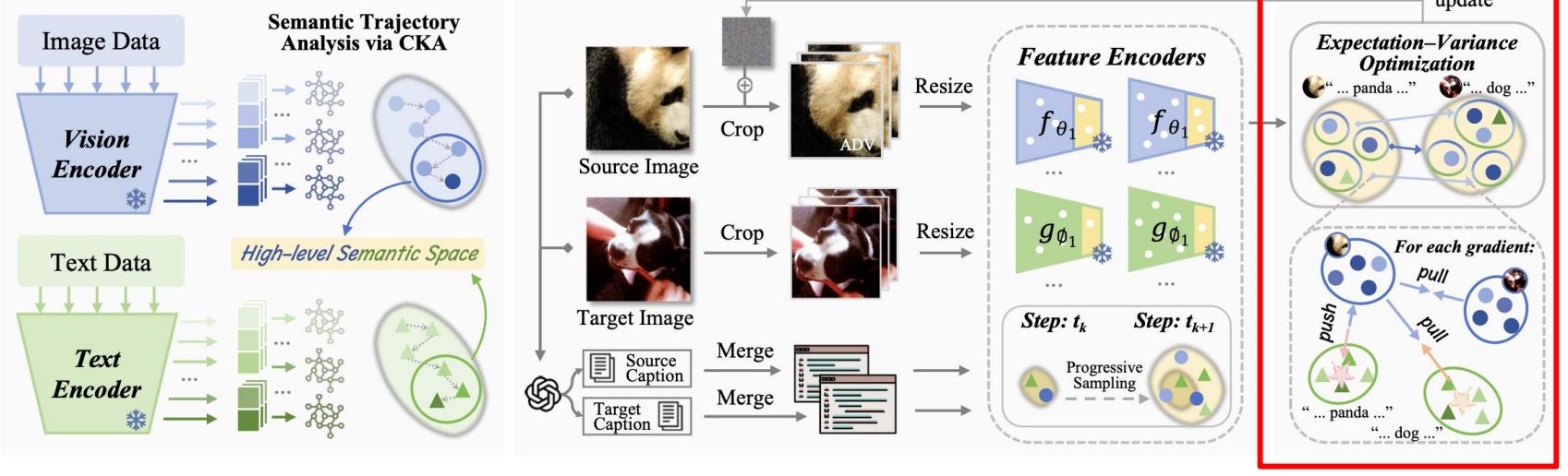
- 模块二：PSS
  - O-Attack 在这些锚定的视觉和文本末N层上做渐进式 dropout 采样以充分在这个语义空间做采样，提升对抗样本迁移性



(a) Cross-Modal Semantic Space Anchoring (b) Progressive Semantic Space Sampling (c) Semantic Consensus Optimization

高可迁移性的黑盒对抗攻击（O-Attack）

- 模块三：SCO
  - O-Attack 在采样出的特征上优化以**远离**源特征，**接近**目标特征，优化期望最大、方差最小以找到共性方向



(a) Cross-Modal Semantic Space Anchoring (b) Progressive Semantic Space Sampling (c) Semantic Consensus Optimization

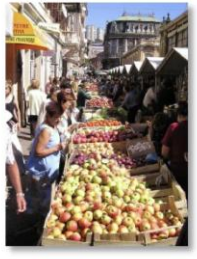
高可迁移性的黑盒对抗攻击（O-Attack）

- O-Attack生成的案例

Query image



Target image



Query Prompt : *Please describe this image in a short sentence.*

GPT 5.4: *A market vendor* stands beneath a green *awning surrounded by crates of apples and other produce*.

Claude 4.6: *A vibrant outdoor fruit market stall* displaying large quantities of *green apples and other fresh produce* under a green *canopy*.

Query Prompt : *What type of insect is shown in the picture?*

Grok 4.1: The photo shows *a bustling outdoor market stall with green apples* ... However, **no insect is clearly visible**.

Gemini 3.1: There are actually **no insects visible**. The picture shows *a crowded outdoor market scene with stalls, yellow canopies ... large quantities of produce* ...

Query Prompt : *What kinds of fruits are in the picture?*

Kimi K2.5: ... identify **apples as the prominent fruit** ... including **oranges**. *The market seems to be a fruit/vegetable market with various produce*.

Qwen 3.5: *The most prominent fruit* ... **large piles of green apples** ... there appear to be *other fruits with orange/yellow coloring*. These could potentially be **oranges, mangoes, or papayas** ...

高可迁移性的黑盒对抗攻击（O-Attack）

实验结果

Model	AnyAttack		COA		M-Attack		FOA-Attack		M-Attack-V2		MPCAttack		**O-Attack (ours)**	
	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim
10 frontier commercial MLLMs														
GPT 5.4	2.3	0.071	1.8	0.026	29.1	0.325	35.0	0.381	46.5	0.485	43.5	0.439	**77.2**	**0.650**
Claude 4.6	1.4	0.044	0.9	0.014	42.8	0.452	50.2	0.495	76.4	0.643	76.7	0.627	**81.6**	**0.693**
Gemini 3.1	1.7	0.055	0.0	0.028	38.2	0.419	48.3	0.482	66.7	0.581	77.6	0.649	**80.9**	**0.678**
Grok 4.3	1.5	0.054	0.9	0.015	37.6	0.469	50.1	0.512	79.6	0.637	66.3	0.620	**85.7**	**0.700**
Qwen 3.5	1.7	0.065	0.7	0.030	40.9	0.439	41.4	0.466	77.9	0.664	77.7	0.663	**83.2**	**0.668**
Kimi K2.5	3.9	0.064	0.0	0.024	49.5	0.504	52.2	0.548	67.8	0.628	76.2	0.632	**79.7**	**0.652**
MiniMax M3	2.6	0.078	0.5	0.031	23.9	0.314	37.6	0.420	51.4	0.516	45.5	0.452	**64.6**	**0.580**
GLM-4.6V	1.3	0.058	0.6	0.017	52.2	0.529	58.1	0.553	83.0	0.679	81.6	0.672	**84.7**	**0.688**
MiMo V2.5	2.9	0.082	0.4	0.026	31.1	0.358	40.6	0.404	59.8	0.566	53.6	0.495	**62.3**	**0.571**
Step 3.7	1.2	0.053	0.3	0.037	17.0	0.256	26.5	0.316	46.3	0.430	47.2	0.439	**70.6**	**0.607**
Avg. (10)	2.1	0.062	0.6	0.025	36.2	0.407	44.0	0.458	65.5	0.583	64.6	0.569	**77.1**	**0.648**
14 widely used MLLMs														
Llama 4	3.3	0.071	0.8	0.056	55.8	0.513	55.1	0.550	79.0	0.664	76.7	0.642	**80.5**	**0.688**
NEX N2	2.5	0.083	1.3	0.042	58.6	0.536	67.1	0.577	80.4	0.686	81.6	0.704	**92.0**	**0.722**
LLaVA v1.6	3.3	0.087	0.6	0.039	73.9	0.640	77.6	0.672	89.5	0.774	91.3	0.775	**95.5**	**0.796**
MiniCPM V4.5	3.1	0.097	0.7	0.028	63.3	0.563	63.0	0.577	79.1	0.691	80.3	0.674	**90.7**	**0.734**
Gemma 3	0.3	0.100	1.1	0.040	26.6	0.330	36.7	0.380	48.2	0.495	56.0	0.522	**71.4**	**0.628**
InternVL 3.5	4.1	0.088	0.0	0.033	61.0	0.541	64.3	0.590	84.1	0.711	85.1	0.704	**90.4**	**0.737**
DeepSeekVL 2	2.6	0.078	0.9	0.033	62.9	0.599	76.9	0.647	91.0	0.749	92.6	0.757	**95.1**	**0.763**
Molmo 2	0.7	0.063	0.2	0.018	42.5	0.434	41.2	0.440	61.5	0.591	63.8	0.611	**77.5**	**0.650**
Llama-Vision	3.9	0.096	0.1	0.035	57.5	0.563	64.5	0.589	77.0	0.701	84.7	0.728	**88.4**	**0.758**
Mistral 3.5	1.7	0.069	0.1	0.033	34.7	0.382	35.3	0.388	48.7	0.476	38.8	0.401	**66.2**	**0.601**
Qwen3 VL	4.7	0.072	0.7	0.031	57.8	0.544	66.2	0.567	86.1	0.694	80.0	0.681	**88.3**	**0.730**
Seed 2.0	1.7	0.063	0.0	0.020	12.6	0.229	16.3	0.269	15.7	0.269	23.4	0.297	**49.2**	**0.510**
Nova 2	2.2	0.075	0.6	0.030	40.4	0.457	44.0	0.511	68.7	0.630	71.0	0.635	**73.1**	**0.645**
Nemotron 2	3.1	0.084	0.3	0.022	40.0	0.459	47.8	0.482	65.0	0.612	75.3	0.667	**84.9**	**0.687**
Avg. (14)	2.7	0.080	0.5	0.033	49.1	0.485	54.0	0.517	69.6	0.625	71.5	0.628	**81.7**	**0.689**



- *安全审核场景将GPT-5.5和Claude Opus-4.8 审查率几乎降到0%*

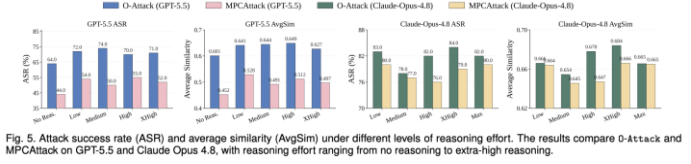


Fig. 5. Attack success rate (ASR) and average similarity (AvgSim) under different levels of reasoning effort. The results compare `O-Attack` and MPCAttack on GPT-5.5 and Claude Opus 4.8, with reasoning effort ranging from no reasoning to extra-high reasoning.

- *思维链无法缓解对抗攻击反而会加剧*

- ***在24个黑盒模型上，O-Attack仅使用3个CLIP做代理，平均攻击成功率达到80%***

高可迁移性的黑盒对抗攻击（O-Attack）

- 消融实验
  - 组件消融实验表明CSA、PSS、SCO三个组件**彼此互补**

Method	CSA		PSS	SCO	GPT 5.4		Claude 4.6		Gemini 3.1		Grok 4.3		Qwen 3.5		Kimi K2.5		Avg.(6)	
	Space	Text			ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim
M-Attack	×	×	×	×	29.1	0.325	42.8	0.452	38.2	0.419	37.6	0.469	40.9	0.439	49.5	0.504	39.7	0.435
O-Attack	×	×	×	✓	39.0	0.442	67.0	0.584	53.0	0.534	65.0	0.574	62.0	0.583	73.0	0.625	59.8	0.557
O-Attack	×	✓	×	✓	41.0	0.451	68.0	0.585	52.0	0.532	65.0	0.589	76.0	0.624	62.0	0.553	60.7	0.556
O-Attack	✓	×	✓	✓	52.0	0.520	70.0	0.625	68.0	0.593	59.0	0.579	74.0	0.617	62.0	0.529	64.2	0.577
O-Attack	✓	✓	×	✓	59.0	0.541	77.0	0.675	66.0	0.622	72.0	0.632	**76.0**	**0.661**	77.0	0.650	71.2	0.630
O-Attack	✓	✓	✓	✓	**77.2**	**0.650**	**81.6**	**0.693**	**80.9**	**0.678**	**85.7**	**0.700**	**83.2**	**0.668**	**79.7**	**0.652**	**81.4**	**0.673**

对抗防御

- 对抗防御
  - 定义：对抗防御是指通过精心设计的防护机制（如鲁棒训练或输入净化）增强模型抵御对抗扰动的能力，减少错误预测或非预期行为的防御方式
  - 分类：对抗训练/微调、测试时防御



相关工作（防御）

- 对抗训练/微调
  - 定义：加入对抗样本以训练模型，提升模型内生鲁棒性
  - **对抗微调**，用对抗样本微调 CLIP 并配辅助目标抑制灾难性遗忘（TeCoA、PMG-AFT），或用无监督学习换取更好的鲁棒性迁移（FARE、Sim-CLIP+）
  - **对抗提示调优**——冻结骨干、只优化可学习 prompt 来把文本嵌入对齐到对抗视觉特征（VPT、APT、FedAPT、AdvPT、 FAP、CoAPT）
  - 挑战：**训练开销大，且有损害良性性能或过拟合的风险**

相关工作（防御）

- 测试时防御
  - 定义：不重新训练或微调模型，只在前向推理时处理输出
  - 一条线是**文本对齐**，用基于熵的 prompt 调优或最优传输把文本端调向对抗视觉特征（TAPT、R-TPT、D-TPT、COLA）；
  - 另一条线是**视觉净化**，用扩散先验或变换集成把图像"洗"回自然流形（CLIPure、TTE、DDA、AOM、TTC）。
  - 核心挑战：（1）防御方法只针对特定任务，**泛化性不足**；（2）防御**性能不足**；（3）防御带来严重的**推理延迟**

基于频域对比矫正的测试时对抗防御（CSR）

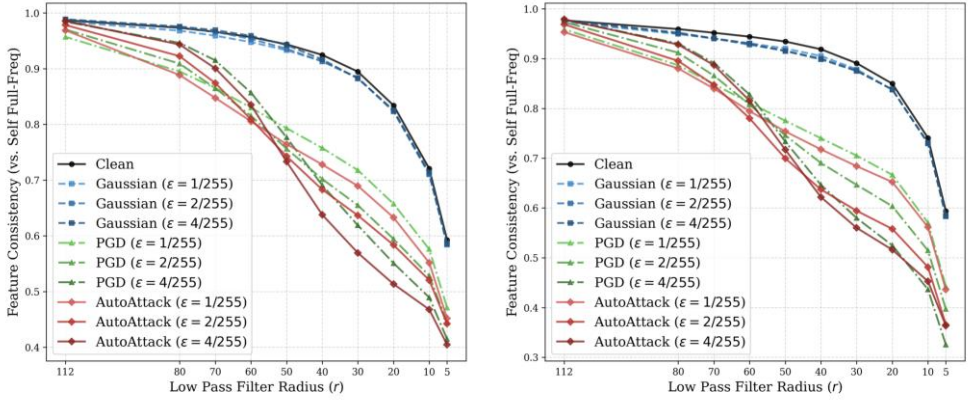
- 动机
  - MLLMs（如CLIP） 的零样本泛化很强，但对抗脆弱性极严重
  - 现有防御不理想，需要一种有效、高效、通用的 MLLMs 测试时防御



Sen Nie, Jie Zhang, Zhuo Wang, Shiguang Shan, Xilin Chen. “Contrastive Spectral Rectification: Test-Time Defense towards Zero-shot Adversarial Robustness of CLIP.” *ICML 2026*

基于频域对比矫正的测试时对抗防御（CSR）

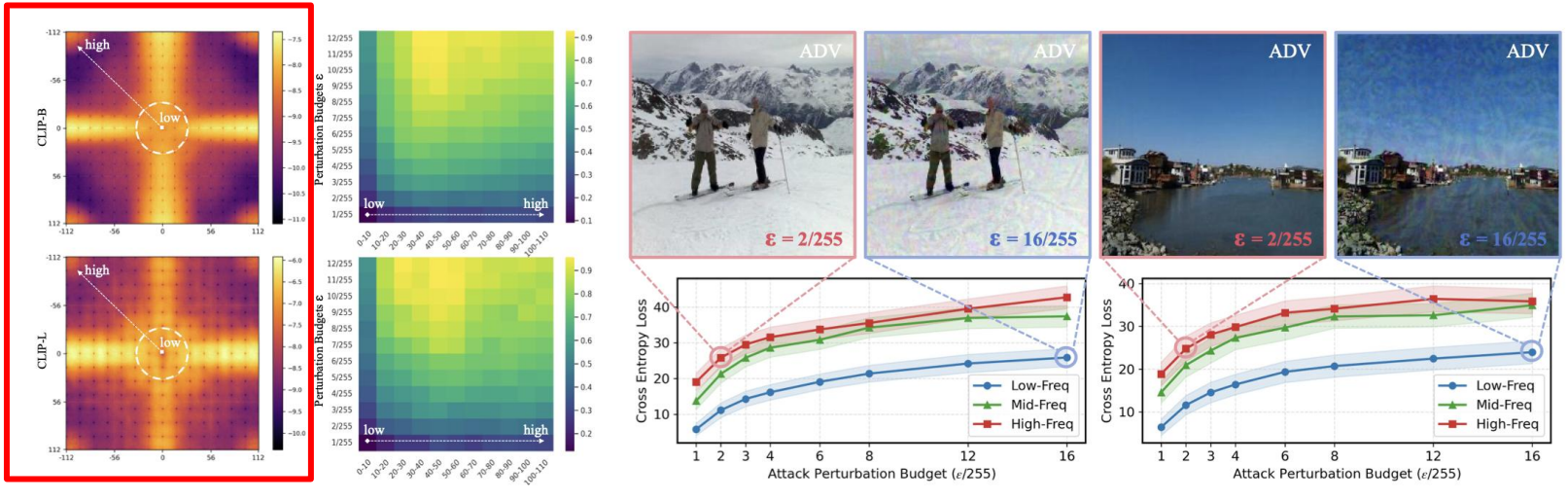
- 关键发现
  - 当频率逐渐衰减时：**良性样本**（干净图与高斯噪声图）特征高度稳定，而**对抗样本**在移除中高频分量后特征急剧崩塌



Sen Nie, Jie Zhang, Zhuo Wang, Shiguang Shan, Xilin Chen. “Contrastive Spectral Rectification: Test-Time Defense towards Zero-shot Adversarial Robustness of CLIP.” *ICML 2026*

基于频域对比矫正的测试时对抗防御（CSR）

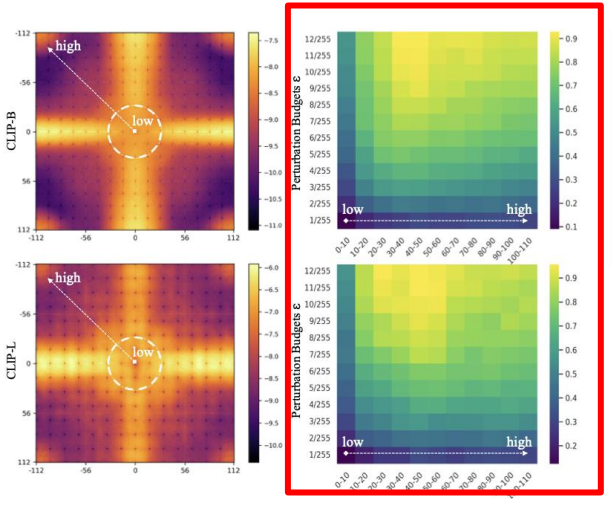
- 归因
  - **频谱偏置**——损失值对频谱幅值的梯度热图表明 CLIP 在预测时倾向于使用中高频特征，这些"非鲁棒特征"成了攻击者的捷径



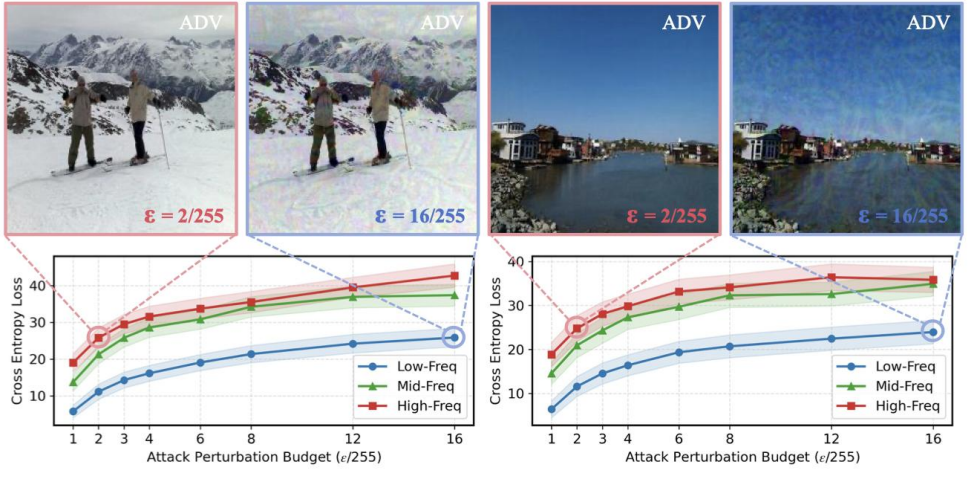
(a) Spectral Heatmaps　　*(b)* Attack Effectiveness in Different Frequency Bands

基于频域对比矫正的测试时对抗防御（CSR）

- 归因
  - **频谱超敏**——按频带分别做 PGD 对抗攻击时，低频扰动造成的特征位移很小，这表明对中高频段的扰动能让特征表示变化巨大



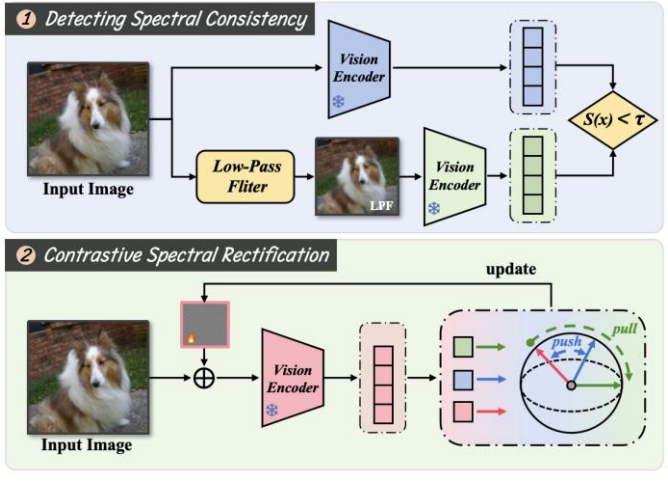
(a) Spectral Heatmaps



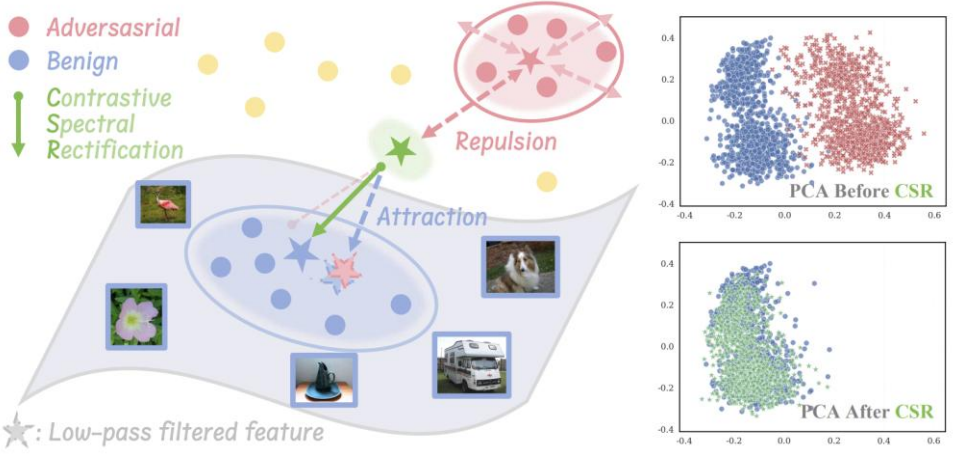
(b) Attack Effectiveness in Different Frequency Bands

基于频域对比矫正的测试时对抗防御（CSR）

- 方法
  - 第一阶段：去除图片的中高频，分析图片视觉特征的变化率
  - 第二阶段：对于识别出来的对抗样本，在其上优化一个**矫正扰动**以净化输入，优化目标为**接近**图片的低频特征、**远离**图片的原特征



方法分为两步：检测、矫正



方法第二步对比频谱矫正示意图

基于频域对比矫正的测试时对抗防御（CSR）

- 弱对抗攻击结果
  - CSR 平均鲁棒精度 58.1%，超过SOTA方法 6.9%，干净精度仅掉 0.7%

Dataset		Original		Adversarial Fine-tuning				Test-Time Defense														Δ	
		CLIP		TeCoA		FARE		R-TPT		LPF		HD		Anti-Adv		TTE		TTC		CSR (Ours)			
Type	Name	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.
General	ImageNet	63.9	0.0	56.3	53.6	48.0	46.7	**66.7**	51.0	58.1	30.5	59.7	4.1	61.5	23.9	66.2	23.2	40.9	27.8	62.5	**58.9**	-1.4	+58.9
	CIFAR10	88.1	0.5	63.7	63.8	61.4	61.2	81.6	69.2	89.0	40.4	84.1	11.8	82.8	63.5	85.5	29.8	**90.0**	28.2	87.2	**75.0**	-0.9	+74.5
	CIFAR100	59.6	0.1	37.9	37.8	36.2	36.2	51.8	36.4	63.4	19.6	57.6	7.9	51.5	34.9	60.4	14.2	**63.1**	11.1	59.4	**45.3**	-0.2	+45.2
	STL10	97.5	4.8	90.7	90.6	92.0	92.0	96.8	**92.7**	96.9	77.6	96.8	34.0	97.2	83.1	**97.6**	75.6	96.4	51.1	97.2	87.5	-0.3	+82.7
	Caltech101	83.5	1.3	77.1	76.3	83.7	**82.0**	86.1	80.9	81.3	65.6	82.6	28.4	82.3	58.2	**87.2**	61.2	75.8	31.3	83.2	75.3	-0.3	+74.0
	Caltech256	83.5	1.5	72.0	69.5	78.4	77.6	**88.0**	**80.5**	82.8	66.5	80.4	20.9	81.6	55.8	87.3	59.4	73.4	41.8	82.9	76.6	-0.6	+75.1
Fine-Gained	OxfordPets	88.9	0.0	71.3	69.3	79.3	**74.3**	85.8	68.3	79.4	41.7	84.2	4.0	86.5	36.7	84.8	12.5	78.8	27.0	**88.9**	65.9	+0.0	+65.9
	Flowers102	66.0	0.0	21.3	21.5	41.7	38.7	**65.8**	48.8	59.8	33.1	63.3	3.5	63.5	25.3	65.3	5.5	55.8	23.3	**65.8**	**53.5**	-0.2	+53.5
	Food101	84.8	0.0	30.6	29.2	47.0	44.5	**86.6**	68.9	79.6	38.6	86.0	1.1	83.9	29.3	85.3	24.8	57.5	33.2	81.9	**80.7**	-2.9	+80.7
	StanfordCars	65.2	0.0	32.9	23.6	**70.7**	**67.9**	68.6	45.4	54.0	16.9	57.8	1.4	62.5	13.7	59.0	14.4	46.6	20.2	62.7	60.9	-2.5	+60.9
Scene	SUN397	63.6	0.2	39.9	37.2	47.6	45.7	64.1	53.1	59.2	31.3	59.7	4.0	62.5	22.7	**65.4**	20.2	47.5	26.2	63.0	**66.3**	-0.6	+66.1
	Country211	17.0	0.0	3.0	2.7	4.4	3.9	**19.2**	8.9	14.7	2.5	15.1	0.0	15.2	1.9	15.8	0.3	10.2	4.4	16.6	**22.9**	-0.4	+22.9
Domain	FGVCAircraft	23.1	0.0	5.2	7.9	12.2	12.6	**23.8**	16.7	17.8	7.2	18.8	1.3	20.4	6.0	23.3	4.9	13.8	10.6	22.7	**31.2**	-0.4	+31.2
	EuroSAT	42.9	0.0	16.5	16.4	12.6	12.6	29.6	22.3	41.6	4.9	41.7	8.5	38.7	25.5	42.3	8.3	**45.6**	10.2	42.9	**40.4**	+0.0	+40.4
	DTD	42.3	0.1	29.9	29.3	35.3	34.3	**44.2**	36.2	40.5	26.7	40.4	8.5	40.6	22.1	41.9	20.6	35.7	22.1	41.9	**39.1**	-0.4	+39.1
	PCAM	48.4	7.4	48.8	48.8	51.7	**51.7**	**54.6**	39.2	48.6	48.4	48.4	36.1	48.5	48.2	44.3	4.3	48.2	23.3	48.5	50.3	+0.1	+42.9
All	Avg.	63.6	1.0	43.6	42.3	50.1	48.9	**63.3**	51.2	56.7	34.5	61.0	11.0	61.2	34.4	63.2	23.7	55.0	24.5	62.9	**58.1**	-0.7	+57.1

基于频域对比矫正的测试时对抗防御（CSR）

- 强对抗攻击结果
  - 面对强对抗攻击时，CSR优势更大，相比SOTA提升18.1%

Method	General			Fine-Grained			Scene			Domain		
	Clean	PGD	AutoAttack	Clean	PGD	AutoAttack	Clean	PGD	AutoAttack	Clean	PGD	AutoAttack
CLIP	79.4	0.0	0.0	76.2	0.0	0.0	40.3	0.0	0.0	39.2	0.0	0.0
TeCoA	66.3	65.1	65.0	39.0	35.8	35.7	21.5	20.0	19.7	25.1	25.5	25.5
FARE	66.6	65.8	65.4	59.7	55.7	55.2	26.0	24.2	24.2	28.0	27.8	27.6
R-TPT	78.5	36.2	28.4	**76.7**	46.4	42.8	**41.7**	26.7	26.4	38.1	27.5	23.2
TTE	**80.7**	13.4	11.5	73.6	1.7	0.4	40.6	1.2	1.4	38.0	13.4	11.5
LPF	78.6	37.4	30.8	68.2	11.3	8.2	37.0	8.0	6.9	37.1	18.6	16.7
HD	76.9	0.5	0.1	72.8	0.0	0.0	37.4	0.0	0.0	37.3	0.1	0.0
Anti-Adv	76.2	22.6	2.7	74.1	1.7	0.0	38.9	1.1	0.2	37.1	9.8	1.7
TTC	73.3	14.7	0.6	59.7	2.0	0.0	28.9	1.2	0.0	35.8	9.3	0.3
CSR	78.7$^{\downarrow 0.7}$	**78.0**$^{\uparrow 78.0}$	**66.4**$^{\uparrow 66.4}$	74.8$^{\downarrow 1.4}$	**75.4**$^{\uparrow 75.4}$	**60.3**$^{\uparrow 60.3}$	39.8$^{\downarrow 0.5}$	**43.8**$^{\uparrow 43.8}$	**34.2**$^{\uparrow 34.2}$	**39.0**$^{\downarrow 0.2}$	**43.2**$^{\uparrow 43.2}$	**32.6**$^{\uparrow 32.6}$

基于频域对比矫正的测试时对抗防御（CSR）

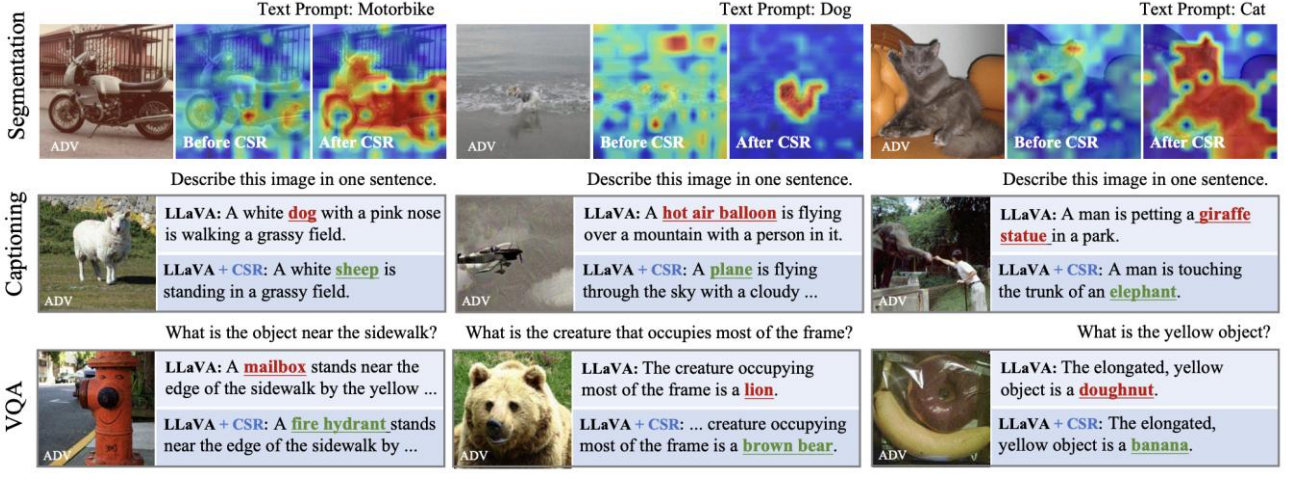
- CSR推理延迟低且在多个视觉任务取得良好防御效果

Table 3. Running time per image (ms) (↓) on single RTX 4090.

Time	CLIP	R-TPT	HD	Anti-Adv	TTE	TTC	CSR(ours)
T_{Clean}	3.37	176.45	184.43	32.39	5.43	33.24	**4.16**
$T_{Rob.}$	3.37	176.20	189.34	32.44	**5.41**	35.13	26.18
$T_{Avg.}$	3.37	176.33	189.34	32.42	**5.42**	35.19	15.17

Table 7. Results on various vision tasks. (mIoU/Accuracy)

Method	Segmentation		Captioning		VQA	
	Clean	Rob.	Clean	Rob.	Clean	Rob.
Origin Model	37.7	20.8	100.0	21.3	100.0	22.6
+TTC defense	35.1	21.9	87.4	24.1	87.7	27.3
+CSR defense	**37.5**$^{\downarrow 0.2}$	**37.5**$^{\uparrow 16.7}$	**99.2**$^{\downarrow 0.8}$	**64.3**$^{\uparrow 43.0}$	**98.7**$^{\downarrow 1.3}$	**65.6**$^{\uparrow 43.0}$



总结

- 多模态大模型的内生视觉安全问题
  - **攻击侧**，找到模型中解耦的表示，能实现精确语义理解操纵；借助更广的代理模型的特征空间，能实现更强的黑盒攻击性能；
  - **防御侧**，可以利用来自对模型自身先验来提升模型鲁棒性，如CSR利用低频稳定性先验作为锚点来净化对抗样本。
- 攻防一体，共同促进视觉内生安全

感谢聆听，敬请批评指正！

邮箱: zhangjie@ict.ac.cn
个人学术主页: https://jiezhang010.github.io

安全可信的智能算法研究组

课题组简介

在中科院先导项目、科技部青年科学家、国自然等项目的支持下，围绕智能算法内生安全问题和衍生安全问题展开基础和应用研究。通过分析算法缺陷的内在机理，构建算法安全评估体系，探索缺陷与风险防护机制，突破智能算法“可信、可管、可控”的理论和技术瓶颈，为智能算法的应用安全保驾护航。

微信扫码关注

欢迎关注课题组公众号!