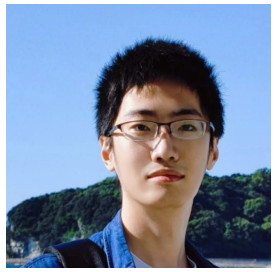
SHIELD: Suppressing Hallucinations In LVLM Encoders via Bias and Vulnerability Defense

Yiyang Huang



Liang Shi

Yitian Zhang

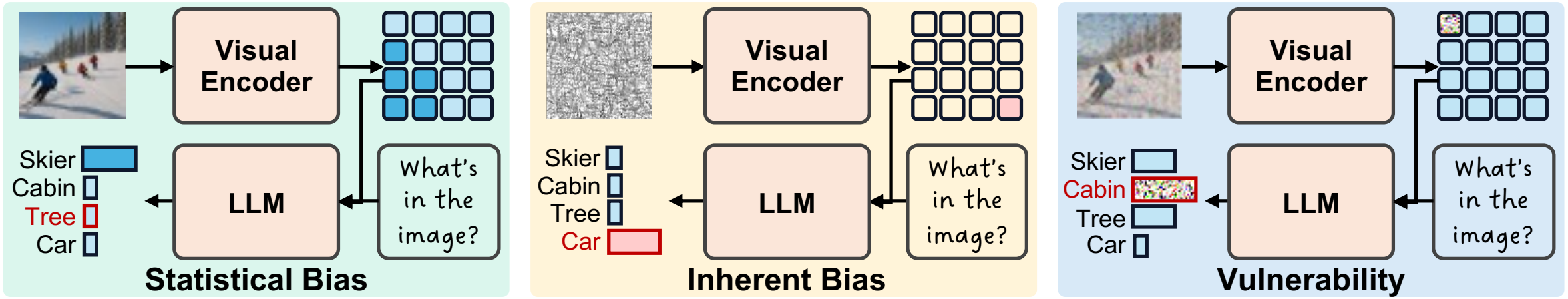
Yi Xu

Yun Raymond Fu

{huang.yiyan,shi.lia,zhang.yitian,xu.yi,y.fu}@northeastern.edu

Northeastern University

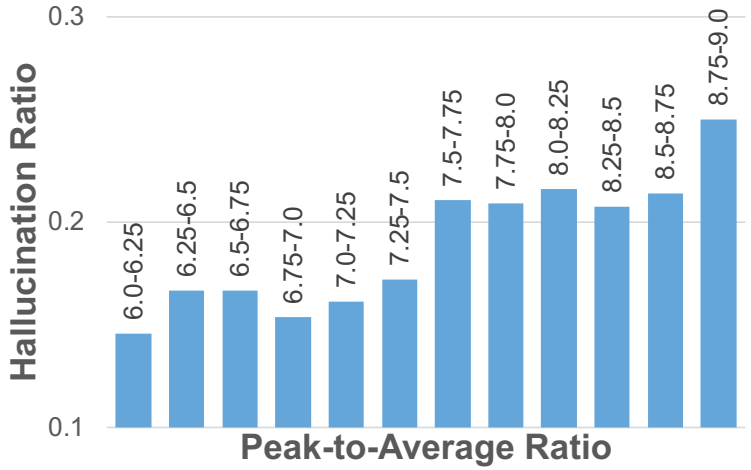
Motivation

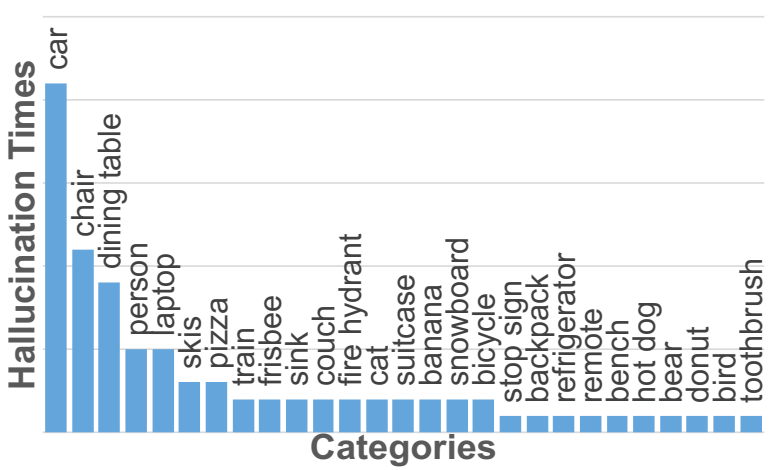


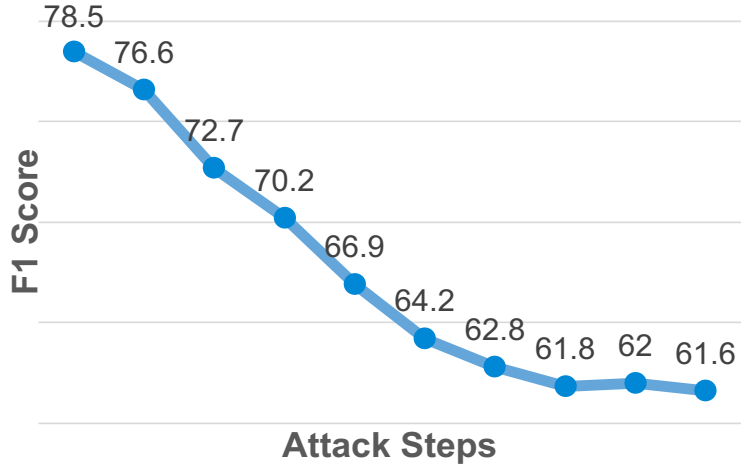
Hallucinations Stem from Visual Encoders:

- **Statistical bias:** encoder overemphasizes frequent visual patterns, distorting fine-grained perception.
- **Inherent bias:** encoder produces erroneous representations of dominant objects in the pretraining data, regardless of input.
- **Vulnerability:** encoder is sensitive to minor perturbations, yielding unreliable features.

Hallucinations Stem from Visual Encoders





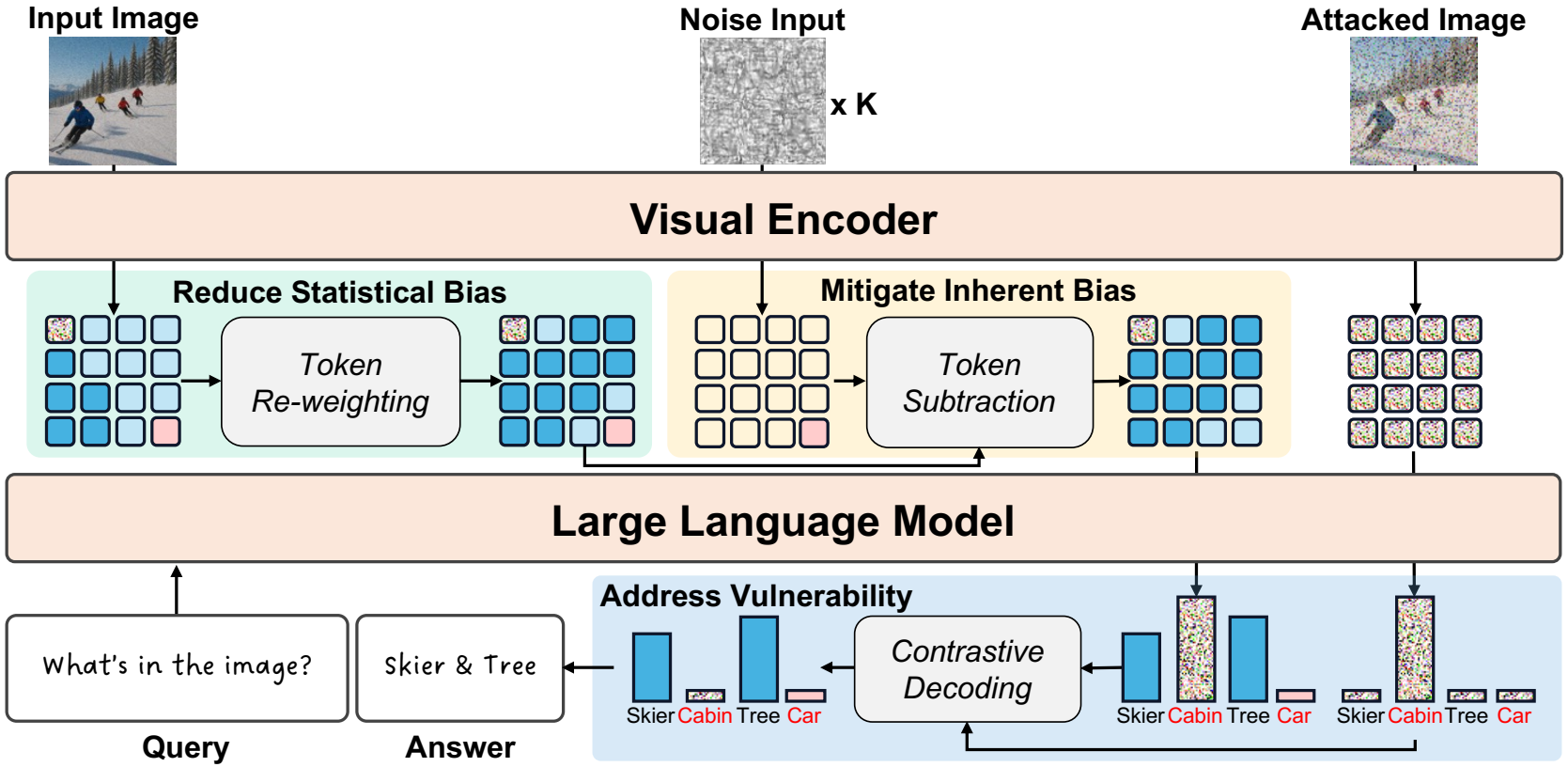


Token Overemphasis:
Stronger overemphasis leads to higher hallucination rates.

Dominant Object Error:
Hallucination occurrences under meaningless inputs. Dominant objects are more likely to be falsely perceived as present.

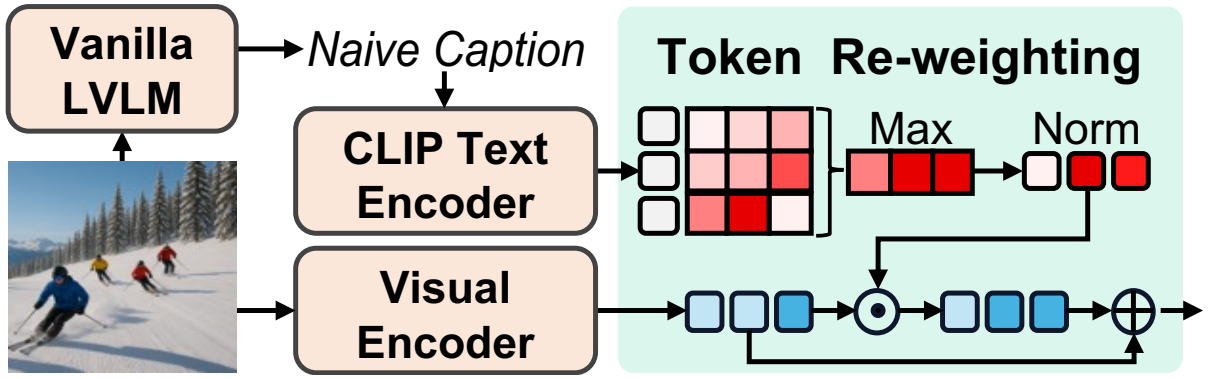
Perturbation Vulnerability:
Even small perturbations can increase hallucinations and degrade performance.

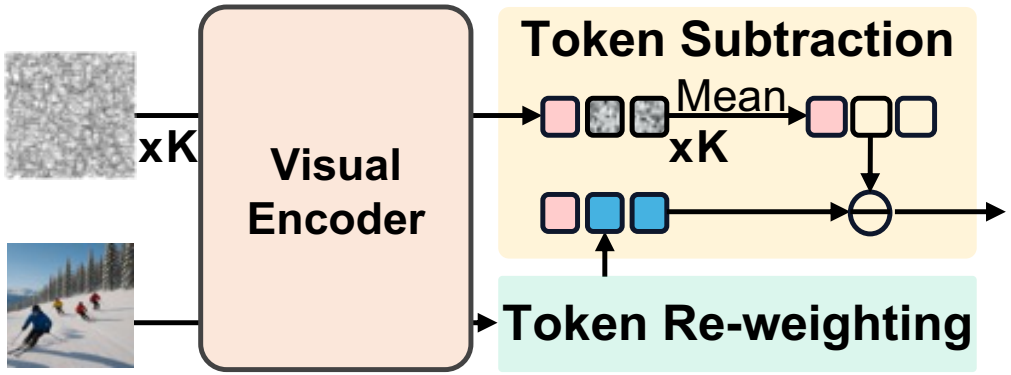
Overall Framework



- **Token Re-weighting** redistributes attention to more ground-truth-object relevant tokens to alleviate overemphasis.
- **Token Subtraction** estimates and removes erroneous representations via noise-derived tokens.
- **Contrastive Decoding** exposes inaccurate features using attacked images and suppresses corresponding outputs by contrasting them with those from the natural image.

Token Re-weighting & Subtraction

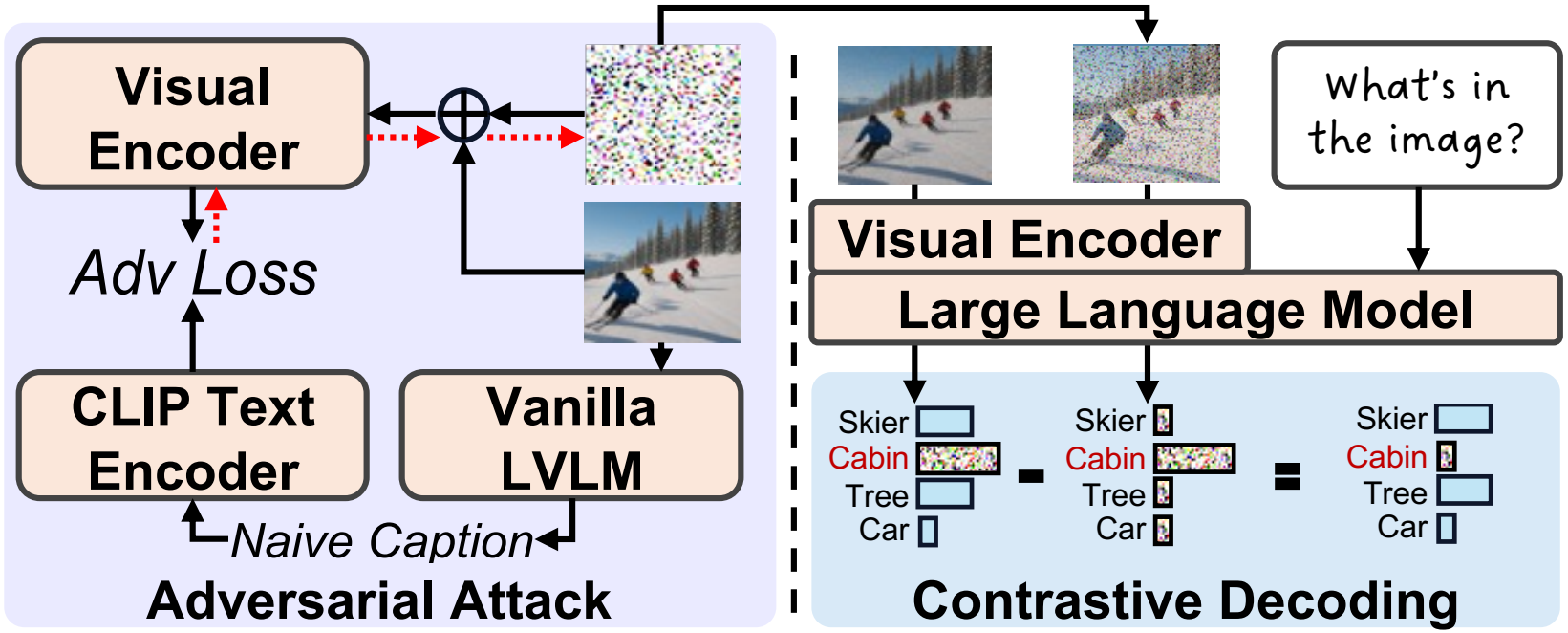




Mitigating Statistical Bias:
Visual tokens are re-weighted via a similarity matrix between visual tokens and naive caption tokens, emphasizing more ground-truth-object relevant tokens and reducing overemphasis.

Reducing Inherent Bias:
K noise inputs are used to estimate erroneous representations of dominant objects in the pre-training data, which are then removed from visual tokens via feature subtraction.

Address Vulnerability



Addressing Vulnerability:

An attack tensor constructed from the input image and its naive caption via adversarial learning is applied to reveal objects likely to be hallucinated, followed by contrastive decoding to suppress their generation.

Results on Hallucination Benchmarks

Table 1: CHAIR Hallucination Evaluation

Method	LLaVA-1.5		InstructBLIP		Qwen-VL	
	$C_S\downarrow$	$C_I\downarrow$	$C_S\downarrow$	$C_I\downarrow$	$C_S\downarrow$	$C_I\downarrow$
Vanilla	48.8	14.2	54.6	24.8	49.2	13.1
VCD	46.8	13.2	44.0	13.6	46.4	11.9
OPERA	44.6	12.8	46.4	14.2	34.6	9.5
Ours	**36.6**	**10.3**	**40.4**	**10.9**	**28.9**	**9.2**

Table 2: GPT4o-aid Hallucination Evaluation

Method	LLaVA-1.5		InstructBLIP		Qwen-VL	
	$C\uparrow$	$D\uparrow$	$C\uparrow$	$D\uparrow$	$C\uparrow$	$D\uparrow$
Vanilla	4.9	5.0	4.2	4.2	6.2	4.6
VCD	5.5	5.5	5.1	**5.5**	6.5	5.7
OPERA	5.6	6.0	5.3	5.2	6.5	5.6
Ours	**6.2**	**6.1**	**5.6**	5.3	**6.9**	**5.8**

Table 3: POPE Hallucination Evaluation on COCO subset

LVLM	Method	Random		Popular		Adversarial		Average	
		Accuracy ↑	F1 ↑	Accuracy ↑	F1 ↑	Accuracy ↑	F1 ↑	Accuracy ↑	F1 ↑
LLaVA-1.5	Vanilla	83.2	81.3	81.8	80.0	78.9	77.5	81.3	79.6
	VCD	87.7	87.1	85.3	85.0	80.8	81.3	84.6	84.4
	OPERA	89.1	89.0	86.0	86.3	79.1	80.9	84.7	85.4
	Ours	**91.3**	**91.1**	**87.4**	**87.6**	**82.5**	**83.6**	**87.0**	**87.4**
InstructBLIP	Vanilla	80.7	80.4	78.2	78.3	75.8	76.5	78.2	78.4
	VCD	84.5	83.6	81.4	81.0	79.5	79.5	81.8	81.3
	OPERA	**89.8**	**89.6**	83.4	84.0	80.7	81.8	84.6	**85.1**
	Ours	88.2	87.6	**84.6**	**84.3**	**82.2**	**82.4**	**85.0**	84.8
Qwen-VL	Vanilla	84.7	82.6	84.1	82.0	82.2	80.3	83.6	81.6
	VCD	88.6	87.8	87.1	86.4	84.2	83.9	86.6	86.0
	OPERA	86.1	84.2	85.7	83.8	83.9	82.1	85.2	83.3
	Ours	**89.2**	**88.6**	**87.6**	**87.1**	**84.3**	**84.2**	**87.0**	**86.6**

Table 4: MME Hallucination Evaluation

LVLM	Method	Object-level		Attribute-level		Total Score ↑
		Existence Score ↑	Count Score ↑	Position Score ↑	Color Score ↑	
LLaVA-1.5	Vanilla	175.6	124.6	114.0	151.0	565.3
	VCD	184.6	138.3	128.6	153.0	604.6
	OPERA	180.6	133.3	123.3	155.0	592.3
	Ours	**195.0**	**141.6**	**148.3**	**183.3**	**668.3**
InstructBLIP	Vanilla	141.0	75.3	66.6	97.3	380.3
	VCD	168.3	**92.3**	64.0	123.0	447.6
	OPERA	156.0	78.3	55.0	95.0	384.3
	Ours	**170.0**	75.0	**88.3**	**128.3**	**461.6**
Qwen-VL	Vanilla	155.0	127.6	131.6	173.0	587.3
	VCD	156.0	131.0	128.0	181.6	596.6
	OPERA	165.0	145.0	**133.3**	180.0	623.3
	Ours	**180.0**	**170.0**	128.3	**190.0**	**668.3**

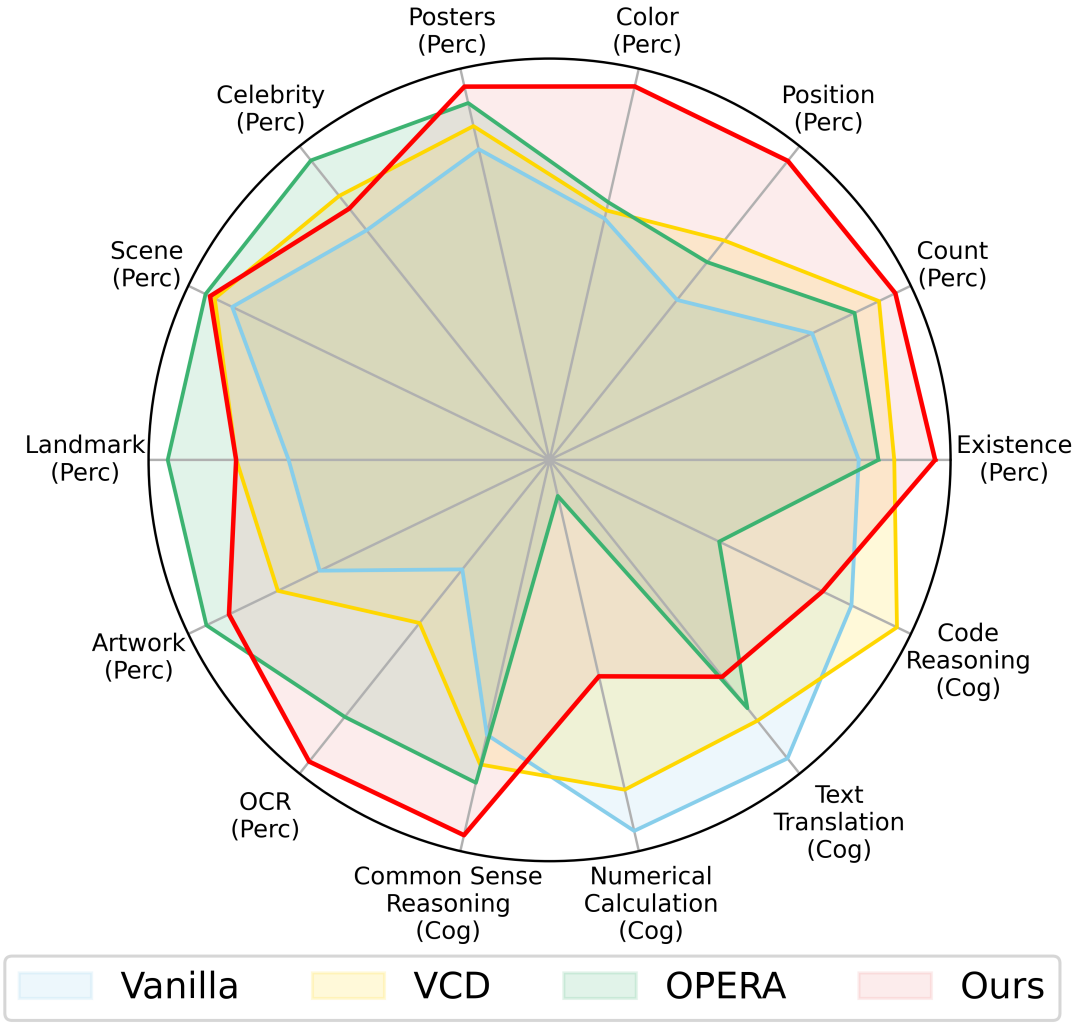
Table 5: AMBER Hallucination Evaluation

Method	**Generative Task**				**Discriminative Task**				**AMBER Score**
	CHAIR↓	Cover↑	Hallucination↓	Cognition↓	Accuracy↑	Precision↑	Recall↑	F1↑	
Vanilla	9.2	41.3	29.2	3.7	65.7	83.2	64.7	73.2	82.0
VCD	8.1	44.2	28.6	3.1	68.3	85.8	65.2	74.0	82.9
OPERA	8.3	43.1	31.2	2.9	76.0	79.2	**83.8**	81.4	86.5
Ours	**6.4**	**46.1**	**25.1**	**1.8**	**78.3**	**89.1**	76.6	**82.4**	**88.0**

Results on General Benchmark

Table 6: MME Full Set Evaluation

Method	Perception↑	Cognition↑	Total Score↑
Vanilla	1279.2	352.9	1632.1
VCD	1363.9	**353.2**	1717.1
OPERA	1413.0	304.2	1717.2
Ours	**1473.0**	337.8	**1810.8**



Ablation Studies

Table 7: Module Ablation on CHAIR

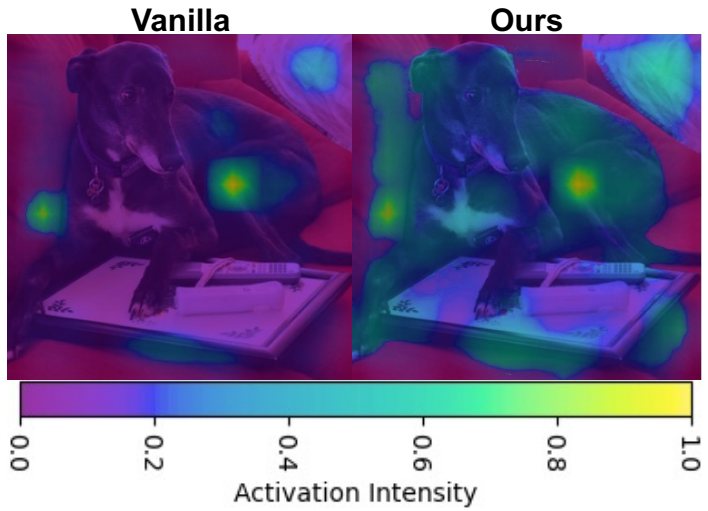
Module	$C_S\downarrow$	$C_I\downarrow$
Vanilla LLaVA-1.5	48.8	14.2
+ adaptive plausibility constraint	50.2	13.8
+ address vulnerability (Ours)	46.4	12.8
+ mitigate statistical bias (Ours)	40.4	11.0
+ reduce inherent bias (Ours)	**36.6**	**10.3**

Table 8: Efficiency Comparison on CHAIR

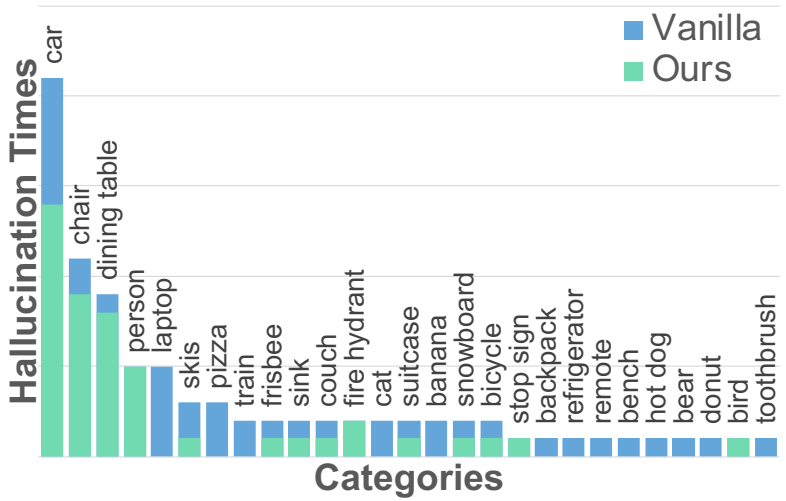
Method	$C_S\downarrow$	T (s/sample)↓	Mem↓
Vanilla	48.8	2.59	15.69GB
VCD	46.8	4.89	16.52GB
OPERA	44.6	24.01	34.88GB
Ours	**36.6**	7.34	18.17GB

Table 9: Module Efficiency on CHAIR

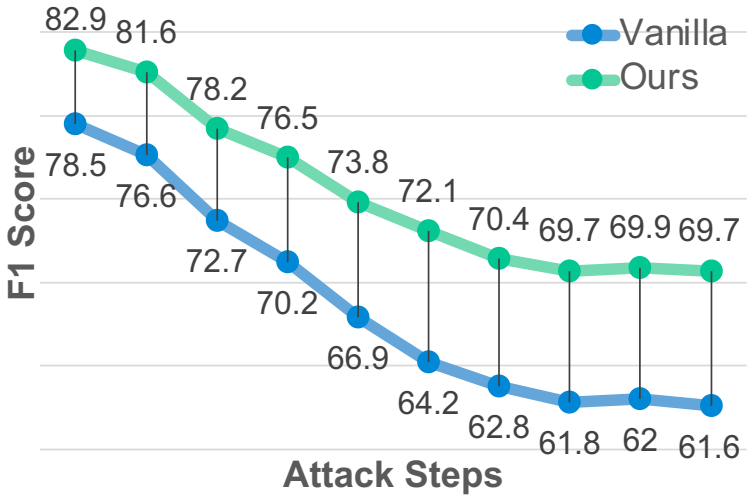
Module	T (s/sample)↓	Mem↓
Vanilla	2.59	15.69GB
w/ Mitigate Statistical Bias	4.64	16.56GB
w/ Reduce Inherent Bias	2.63	16.50GB
w/ Address Vulnerability	7.30	18.17GB
Ours (All Modules)	7.34	18.17GB



Token Re-weighting



Erroneous Representation Removal



Address Vulnerability

ICLR

Synergetic Media Learning Lab

Thank You!

Please contact: **huang.yiyan@northeastern.edu** for more questions.