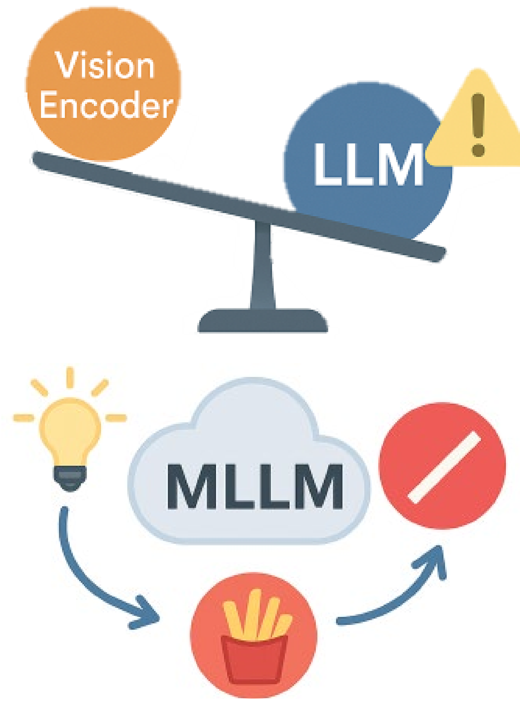
Poison as Cure: Visual Noise for Mitigating Object Hallucinations in LVMs

Kejia Zhang, Keda Tao, Jiasheng Tang, Huan Wang*

Code | Paper

Introduction & Motivation

1. Why LVMs Hallucinate: Parametric Bias



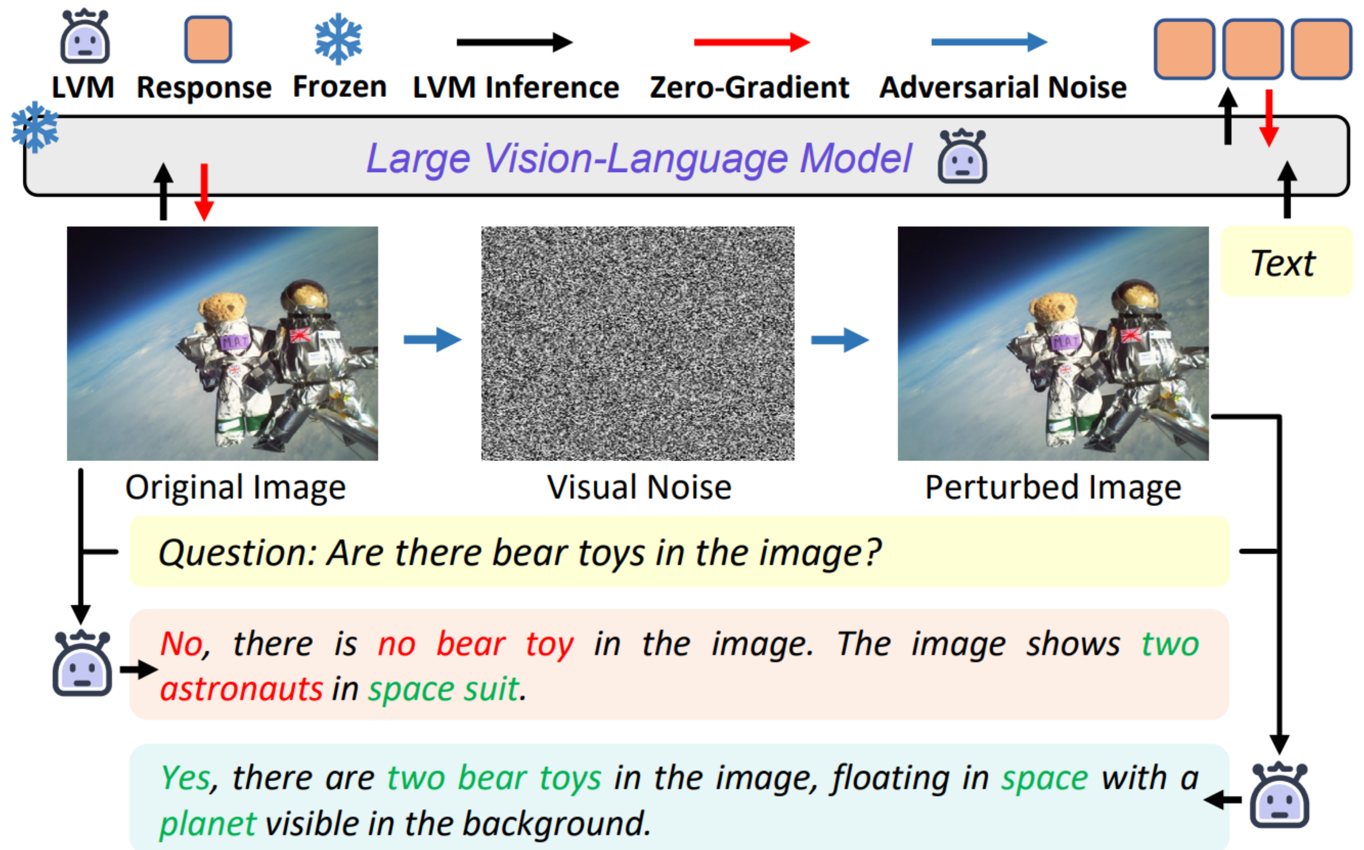
Bias 1: Dominance Bias

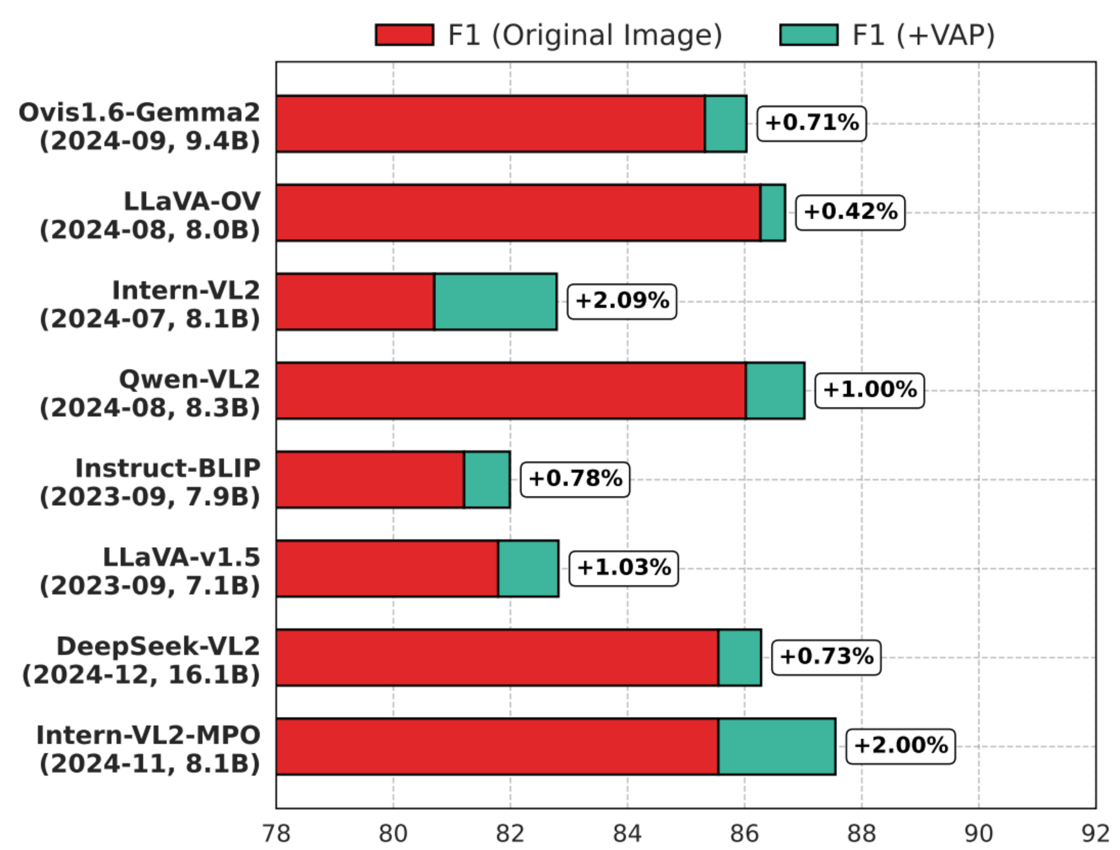
- **LLMs** dominate due to massive scale compared to vision encoders.
- Causes models to **over-rely on language priors**, ignoring visual grounding.

Bias 2: Parametric Knowledge Bias

- MLLMs encode **biased world knowledge** from pretraining (e.g., **long-tail data imbalance**), overfit to frequency exists objects.
- Model outputs reinforce **superficial correlations.**

2. Key Idea: Turning Parametric Bias into Beneficial Visual Noise



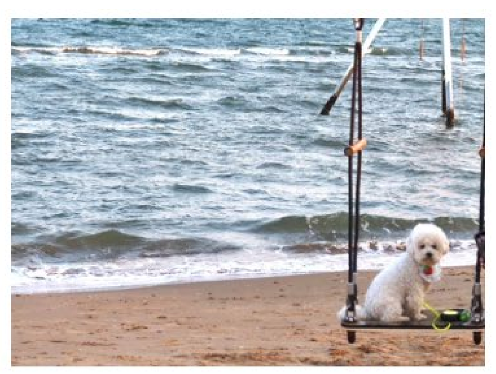


➢**A new perspective view:** Hallucination stems from *input conditioning*, not model incapability.

➢**Our approach:** Reformulate parametric bias as an adversarial optimization problem — inject strategic visual noise to shift model attention toward actual image evidence and away from text-prior biases.

Examples of Vision-Question-Answer (VQA)

(a) open-ended evaluation under original image



Provide a brief description of the given image?

The image shows a small child sitting on a swing set near the ocean. The swing set is located on a sandy bench. The child appears to be enjoying the ride on the swing, while a kit floats gently above the sand, and a large sandcastle stands nearby.

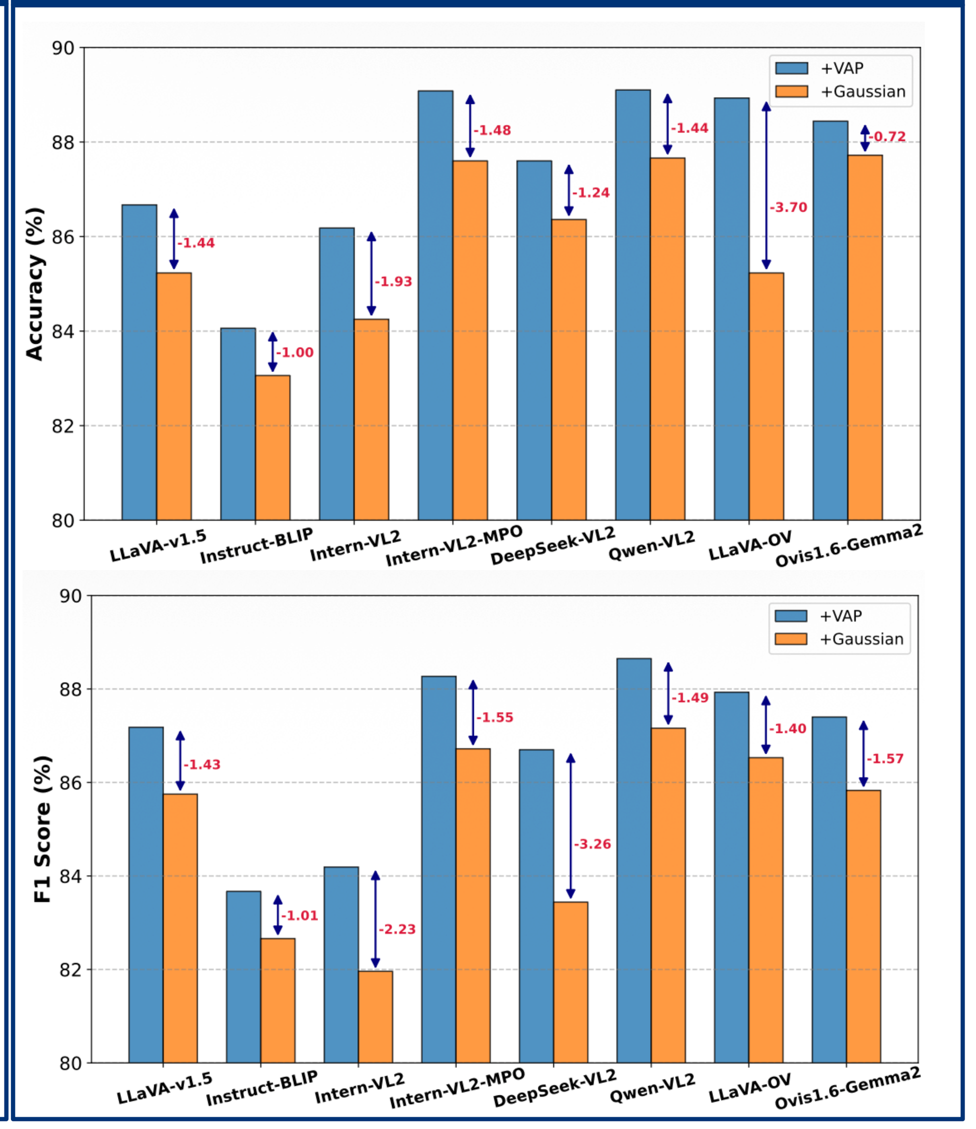
(b) open-ended evaluation under noise image (+ VAP)

Provide a brief description of the given image?

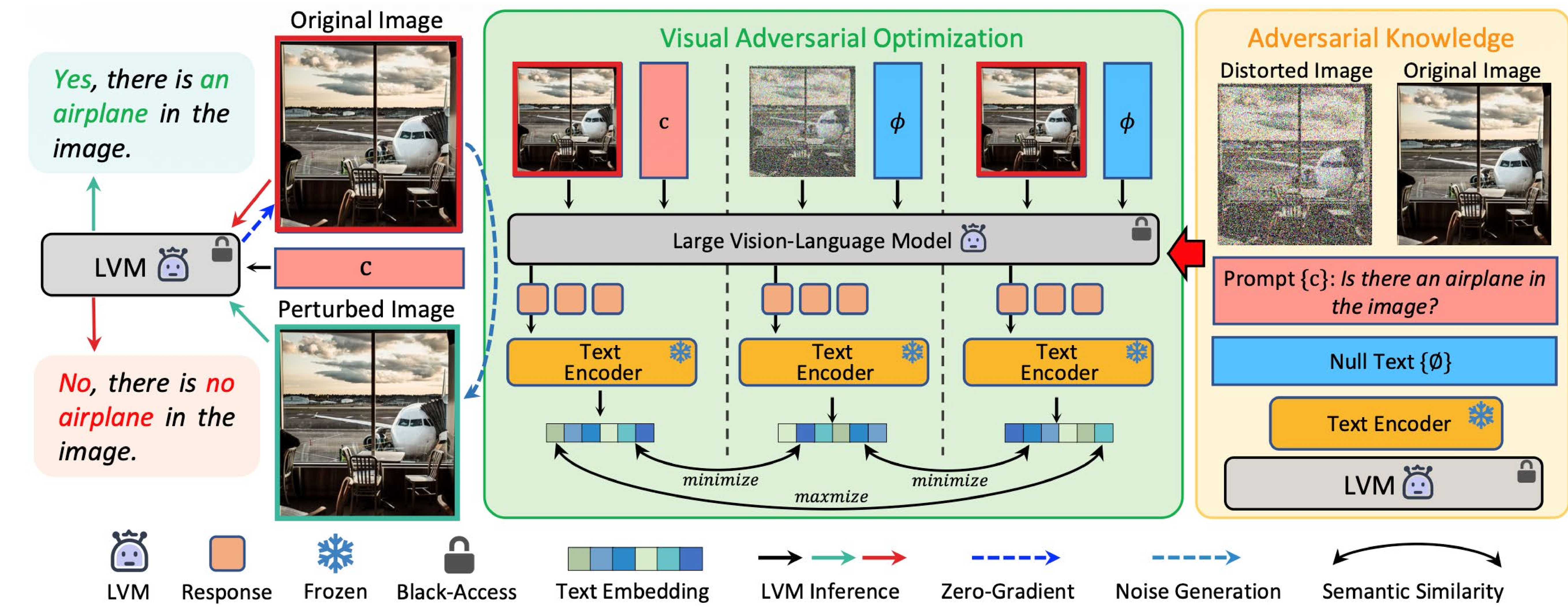
The image shows a small white dog sitting on a swing set near the ocean. The swing set is located on a sandy bench, and the ocean waves can be seen in the background. The dog appears to be enjoying the ride on the swing.

Originally captured by Ms. Xinjun Lin in Aranya, this little dog now travels with us to San Diego for NeurIPS 2025.

VAP vs. Gaussian Noise



Method



Goal: Inject noise δ → enforce visual grounding → suppress parametric bias.

(1) Visual–Text Alignment

- Match $f(x+\delta, c)$ with $f(x+\delta, \emptyset)$
- Reduce prompt-driven hallucination

(2) Bias Exposure via Distorted Views

- Contract $x+\delta$ with distorted image $\bar{x}$
- Push apart $f(x+\delta, \cdot)$ and $f(\bar{x}, \cdot)$

(3) Zero-Gradient Noise Optimization

- Solve $\delta = \arg\max \mathcal{L}(x, \delta)_{\|\delta\| \leq \epsilon}$
- Training-free, model-agnostic

Experiments

1. POPE & BEAF Benchmark: Consistent Hallucination Reduction Across 8 LVMs

LVM	Vision Input	POPE-Popular		POPE-Random		POPE-Adversarial		BEAF	
		Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
LLaVA-v1.5	Original	85.57	86.19	88.97	89.09	79.80	81.79	79.99	74.06
	+VAP	**86.67** +1.10	**87.18** +0.99	**90.00** +1.03	**90.07** +0.98	**80.97** +1.17	**82.82** +1.03	**80.36** +0.37	**74.35** +0.29
Instruct-BLIP	Original	83.30	82.85	88.13	87.18	81.33	81.21	81.91	73.55
	+VAP	**84.06** +0.76	**83.67** +0.82	**89.00** +0.87	**88.12** +0.99	**82.03** +0.70	**81.99** +0.78	**82.07** +0.16	**73.96** +0.41
Intern-VL2	Original	84.11	81.64	85.14	82.60	82.00	80.70	88.38	79.10
	+VAP	**86.18** +2.07	**84.19** +2.00	**86.30** +1.16	**84.08** +1.48	**84.81** +2.81	**82.79** +2.09	**88.69** +0.31	**79.72** +0.62
Intern-VL2-MPO	Original	87.51	86.53	88.68	87.58	86.28	85.55	89.21	82.56
	+VAP	**89.08** +1.57	**88.27** +1.74	**90.20** +1.52	**89.30** +1.72	**88.13** +1.85	**87.55** +2.00	**89.63** +0.42	**82.72** +0.18
DeepSeek-VL2	Original	86.80	85.86	88.70	87.64	86.47	85.55	89.39	82.51
	+VAP	**87.60** +0.80	**86.70** +0.84	**89.30** +0.60	**88.31** +0.67	**87.13** +0.66	**86.28** +0.73	**89.72** +0.33	**83.12** +0.61
Qwen-VL2	Original	88.13	87.68	90.60	89.99	86.27	86.02	87.96	81.13
	+VAP	**89.10** +0.97	**88.65** +0.97	**91.16** +0.56	**90.54** +0.55	**87.30** +1.03	**87.02** +1.00	**88.39** +0.43	**81.57** +0.44
LLaVA-OV	Original	88.30	87.33	89.53	88.51	87.17	86.27	90.76	84.53
	+VAP	**88.93** +0.63	**87.93** +0.60	**89.87** +0.34	**88.83** +0.32	**87.76** +0.59	**86.69** +0.42	**91.07** +0.33	**85.01** +0.48
Ovis1.6-Gemma2	Original	87.96	86.88	88.96	87.87	86.22	85.32	90.12	83.04
	+VAP	**88.44** +0.48	**87.40** +0.52	**89.59** +0.65	**88.54** +0.67	**86.85** +0.63	**86.03** +0.71	**90.91** +0.79	**84.53** +1.49

2. Proxy-Based VAP: 1/8× Cost with Comparable Gains

Metric	Source: Intern-VL2-1B			Source: Qwen-VL2-2B	
	⇒ Intern-VL2-1B	⇒ Intern-VL2-4B	⇒ Intern-VL2-8B	⇒ Qwen-VL2-2B	⇒ Qwen-VL2-7B
Accuracy	81.69/**83.28**	81.55/**82.56**	82.00/**84.07**	84.47/**85.42**	86.27/**86.87**
Precision	89.72/**92.13**	85.65/**87.21**	87.40/**90.97**	83.98/**84.85**	87.21/**88.03**
Recall	70.94/**72.34**	75.05/**75.90**	72.24/**75.50**	84.04/**85.26**	84.87/**85.33**
F1 Score	79.23/**81.04**	80.00/**81.16**	80.70/**82.52**	84.01/**85.05**	86.02/**86.66**
Inference Cost Reduction	**1×**	**1/3×**	**1/8×**	**1×**	**1/5×**