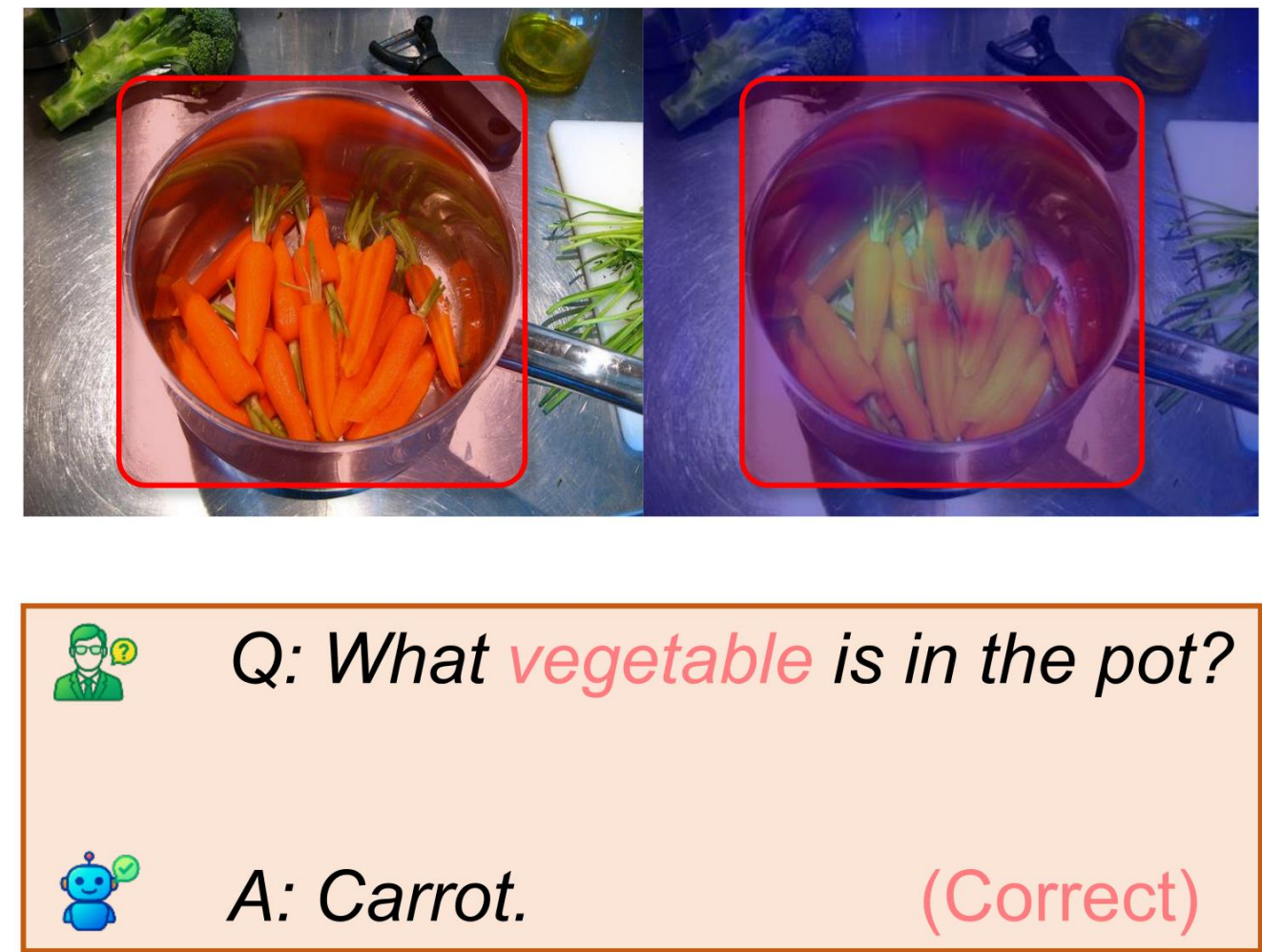


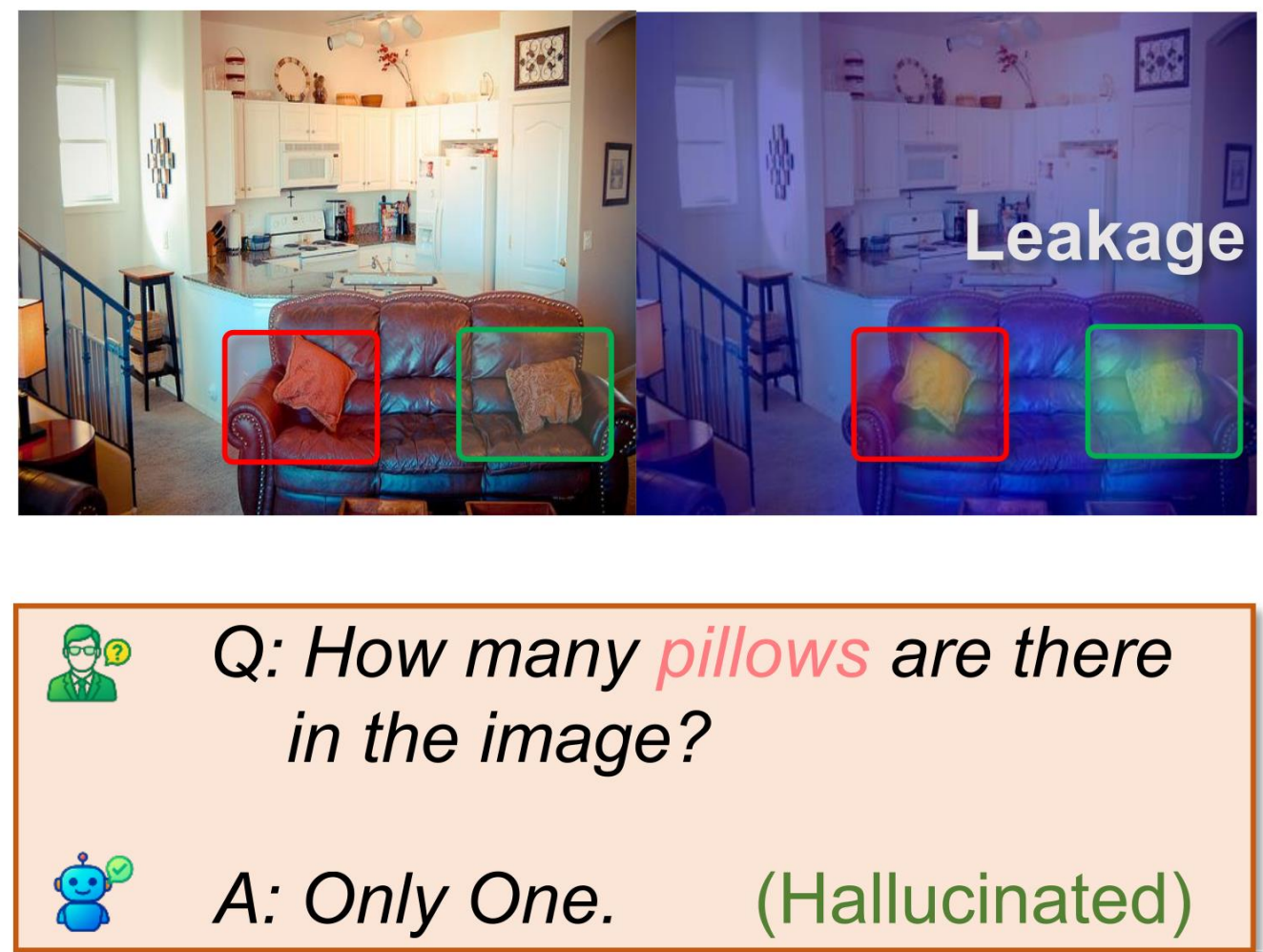
Attention can reach the right pixels and still fail to deliver useful visual evidence. FVR measures and repairs that flow in a training-free manner.

FVR asks whether visual evidence reaches the logits, not only where attention points.

01 / PROBLEM



(a) Correct Sample



(b) Hallucinated sample

High attention weight $\neq$ useful visual contribution.
Hallucination appears when the routed value is weak.

Visual Flow Leakage (VFL)

Attention is routed to visual tokens, but the value paths deliver weak or inconsistent evidence to the residual stream. The model “sees” an object without grounding the next token in it.

VALUE-AWARE SIGNAL

$$N_i^{\wedge}(\ell, h) = \sqrt{\sum_j (w_{ij}^{\wedge}(\ell, h) \cdot \|v_j^{\wedge}(\ell, h)\|_2)^2}$$

FVR averages this gain-weighted routing signal at the current decoding position to score each attention head.

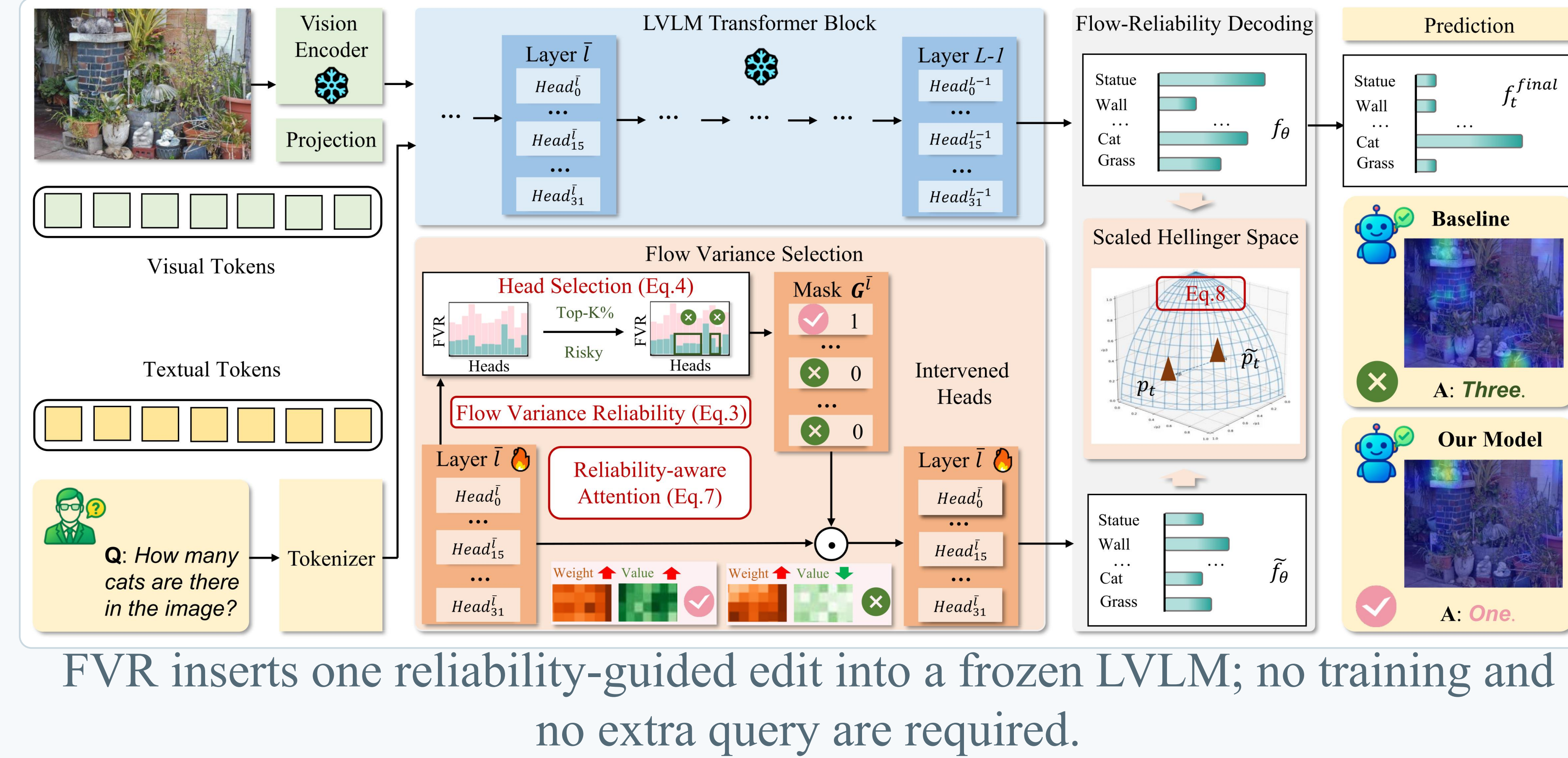
RESEARCH QUESTION

Can we measure and correct visual evidence flow before language priors complete the wrong answer?

What this work adds

- A new mechanism: Visual Flow Leakage
- A value-aware, per-head reliability score: FVR
- A training-free intervention with negligible overhead

02 / METHOD



- SCORE** Combine attention weights with projected value magnitudes to estimate visual delivery for every head.
- SELECT** Mute the lowest-FVR $K = 20\%$ heads in one layer; preserve heads with reliable value paths.
- DECODE** Blend base and edited logits only when the flow is reliable and distributionally consistent.

Qualitative evidence · object existence and spatial relations



Input Image

Text Query:
What fruit is in the left part of the fridge?

Baseline Answer:
Two types of fruit in the left part of the fridge strawberries and blueberries.

Ours Answer:
In the left part of the fridge, there is a plastic box of strawberries.

Input & Output Text



Attention Map



Input Image

Text Query:
Is the lamp on the left side of the sofa?

Baseline Answer:
No, the lamp is on the right side of the sofa.

Ours Answer:
Yes, the lamp is on the right side of the blue sofa.

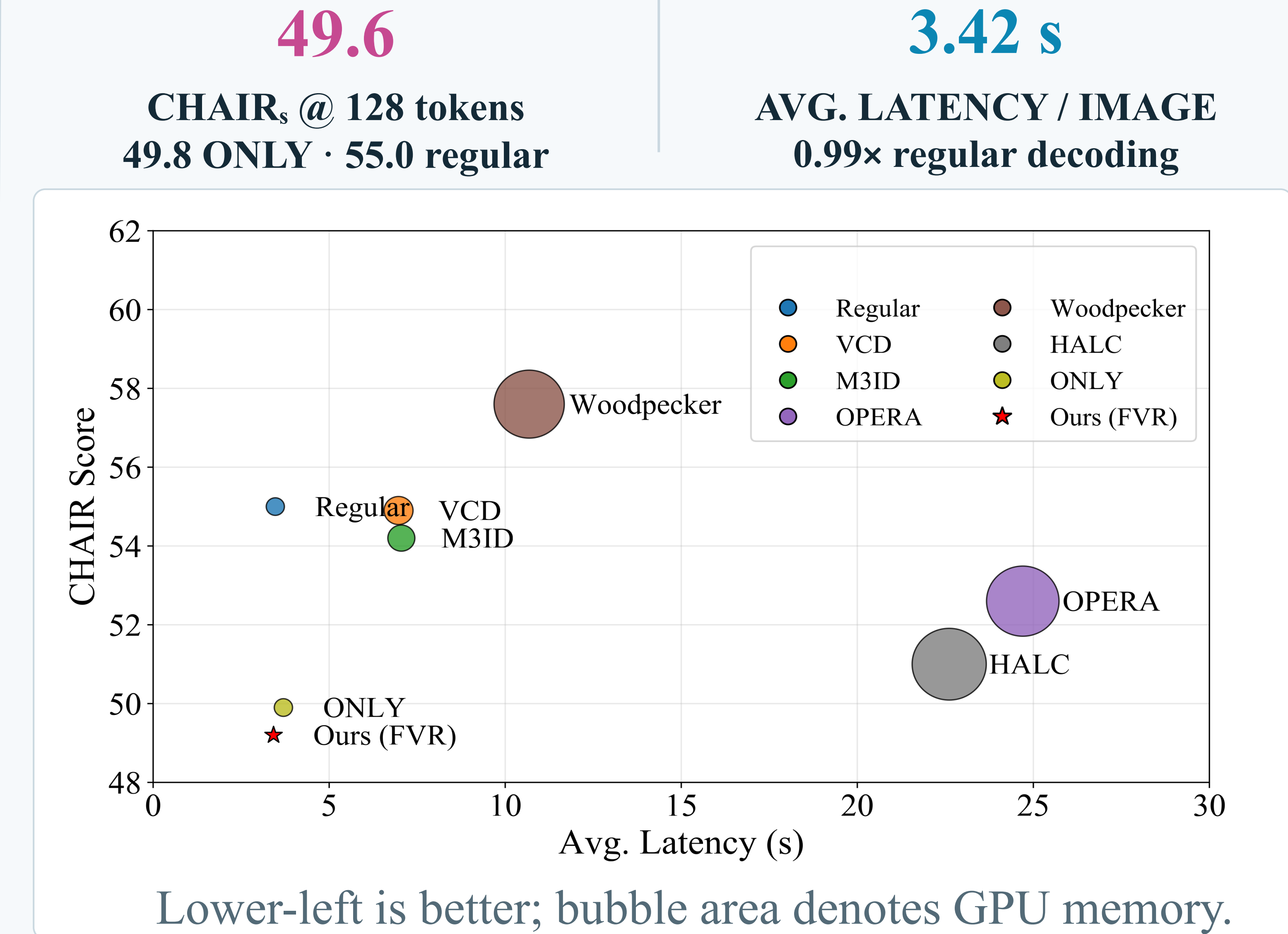
Input & Output Text



Attention Map

FVR suppresses leakage to irrelevant regions and restores query-relevant grounding.

03 / EVIDENCE



Best or strongest results across benchmarks

Setting	FVR result	Change
POPE · A-OKVQA Adv. · Qwen-VL	83.27 Acc.	+1.20
CHAIR · Qwen-VL · 64 tok.	15.0 / 7.0	-4.4 / -1.2
MME-Hallucination · LLaVA	645.75	+4.09
Efficiency · LLaVA · 128 tok.	3.42 s	0.99× regular

Both components are necessary

Variant	POPE Acc. ↑	CHAIR _s ↓
Baseline	80.42	26.2
w/o FVR	84.08	23.5
w/o FRD	83.86	24.0
Full FVR	85.97	19.6

3 LVLs × 3 benchmarks:
POPE · CHAIR · MME
LLaVA · InstructBLIP · Qwen-VL
fewer hallucinations with nearly no inference overhead.