NeurIPS 2026

Where Are MLLMs Looking When They Hallucinate?

Mitigating Visual Hallucination via Gaze Steering

Yebo Wu* Han Jin* Feng Liu Changwang Zhang Jun Wang† Li Li†

University of Macau · OPPO Research Institute

* Equal contribution † Corresponding authors

A confident answer can follow the wrong visual cue



What berries are on top of the cake?

Hallucinated candidates

raspberries, strawberries

Correct answer

blueberries

Where does visual grounding begin to drift?

Tracking visual gaze across layers

1 **Score each visual token**

Aggregate attention from query tokens, averaged across heads.

2 **Fix the reference set**

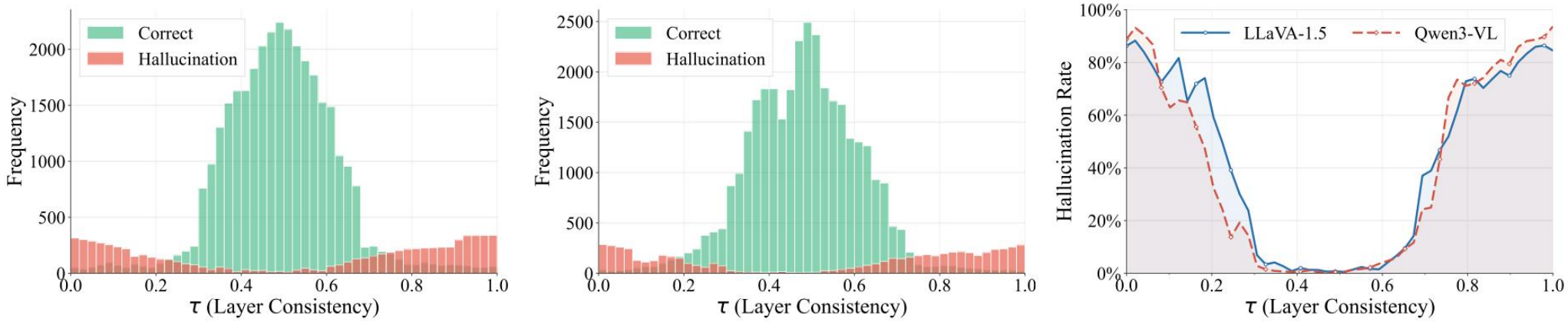
Select the previous layer's top K = 20 visual tokens.

3 **Compare their ranks in adjacent layers**

$$\tau = (\text{concordant pairs} - \text{discordant pairs}) / [K(K-1)/2]$$

"Gaze" refers to internal visual attention, used as a proxy for grounding.

Hallucinations concentrate at both extremes



(a) Sample Distribution: LLaVA-1.5. (b) Sample Distribution: Qwen3-VL. (c) Hallucination Rate Distribution.

Attention Collapse

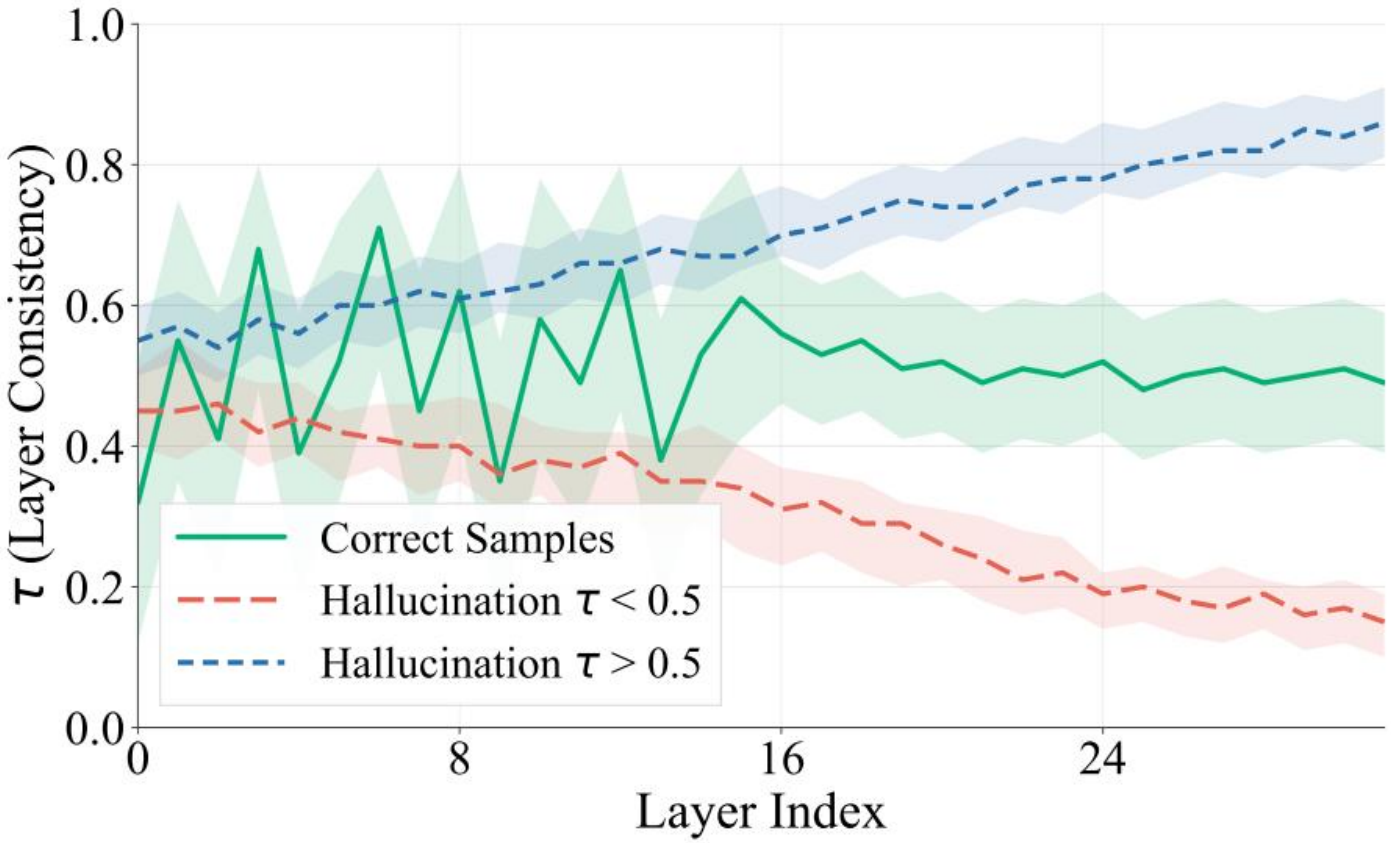
Visual priorities become unstable.

Attention Locking

A few visual anchors dominate rigidly.

Faithful responses cluster in a moderate-consistency regime.

Gaze trajectories diverge in deeper layers



Shallow layers

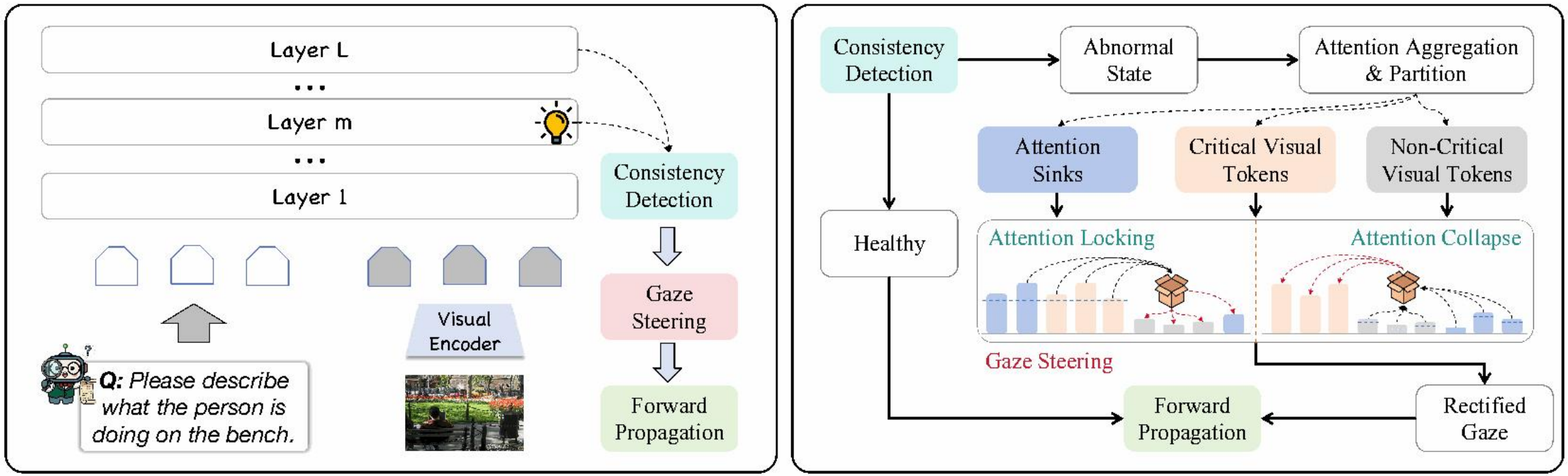
Similar early alignment

Mid-to-deep layers

Increasing rigidity or instability

REGAZE starts monitoring at layer 16.

REGAZE monitors and corrects gaze during inference



(a) Overall Framework.

(b) Workflow of Gaze Steering.

Training-free No parameter updates No extra model forward pass

Stable token groups guide attention redistribution

Aggregate the most recent W = 8 layers

Adaptive threshold: $T = \gamma \times$ mean visual intensity, with $\gamma = 1.2$

Token group	Selection	Role
Attention sinks S	Broad query coverage	Separate attention reservoir
Critical tokens K	Visual intensity ≥ T	Persistent visual anchors
Non-critical tokens U	Visual intensity < T	Peripheral visual context

Two gaze states require opposite corrections

Attention Locking

$\tau > 0.7$

Clip peaks above T across visual tokens and sinks.

Redistribute the excess to tokens below T in proportion to their attention.

Attention Collapse

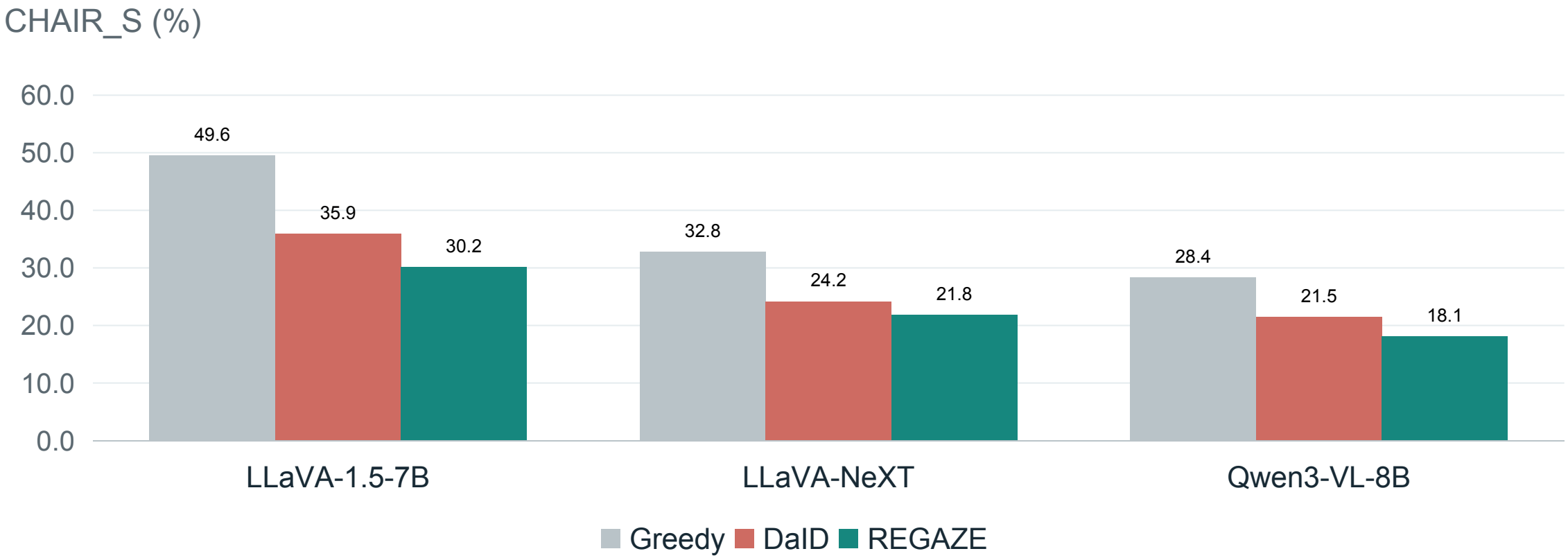
$\tau < 0.3$

Extract $\beta = 0.1$ of attention from non-critical tokens and sinks.

Allocate that mass to critical visual anchors in proportion to their saliency.

$0.3 \leq \tau \leq 0.7$: retain the original attention

REGAZE reduces sentence-level hallucination



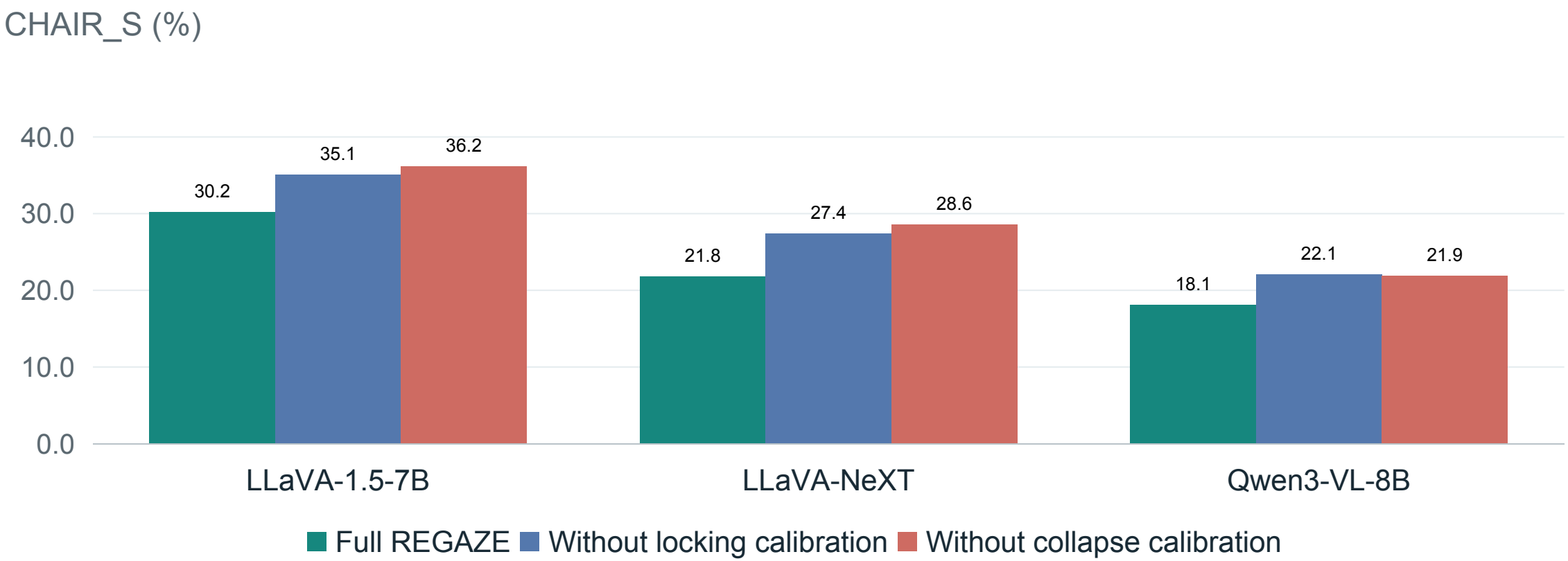
5.7 percentage points lower than DaID on LLaVA-1.5-7B

Gains extend to object probing and perception

Model	Method	POPE Acc.	POPE F1	MME total
LLaVA-1.5-7B	DaID	85.08	85.92	633.68
LLaVA-1.5-7B	**REGAZE**	**86.62**	**86.84**	**652.82**
LLaVA-NeXT	DaID	85.32	85.76	644.40
LLaVA-NeXT	**REGAZE**	**87.15**	**87.22**	**657.29**
Qwen3-VL-8B	DaID	86.78	86.42	662.90
Qwen3-VL-8B	**REGAZE**	**88.89**	**89.08**	**676.41**

+2.11 pp POPE accuracy on Qwen3-VL-8B over DaID

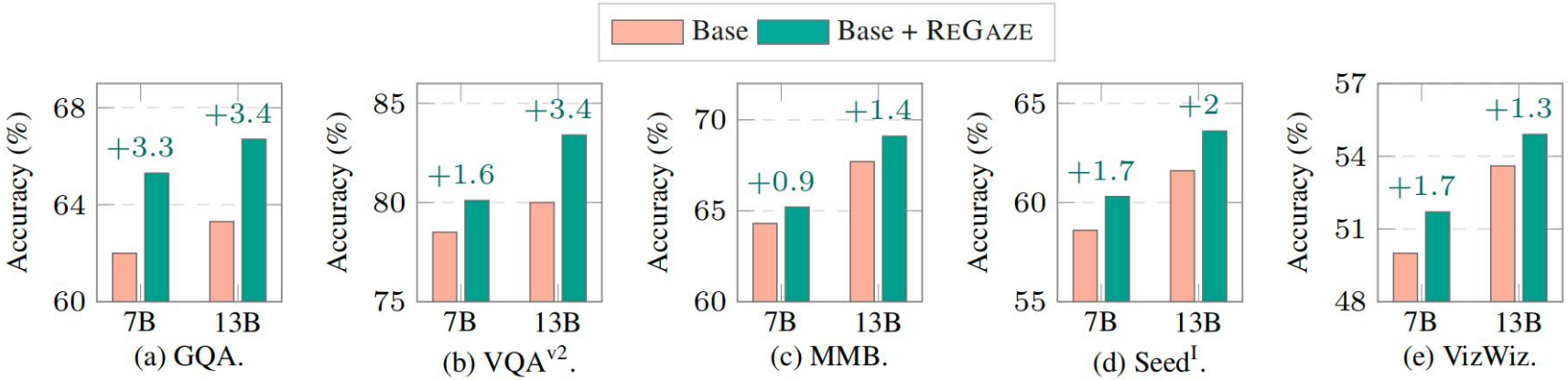
Both calibration branches contribute



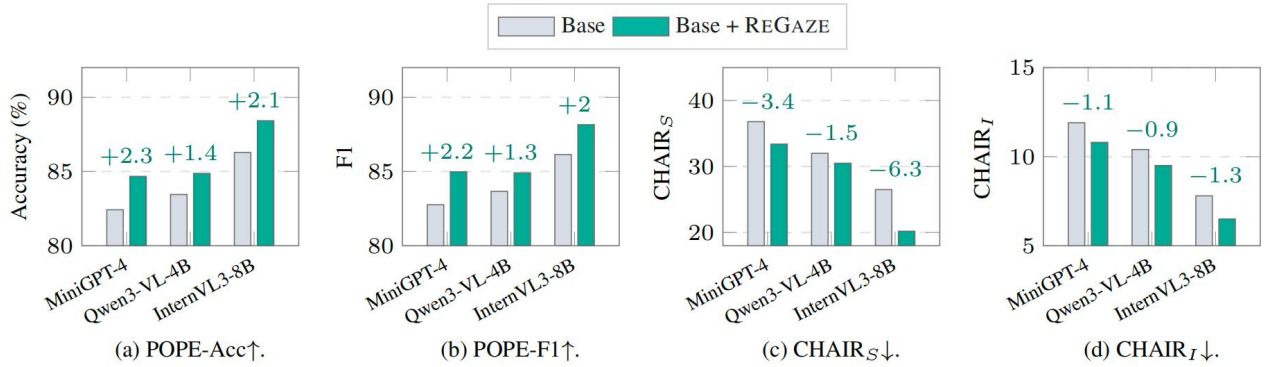
Removing either correction increases hallucination across all three models.

Benefits extend across tasks and architectures

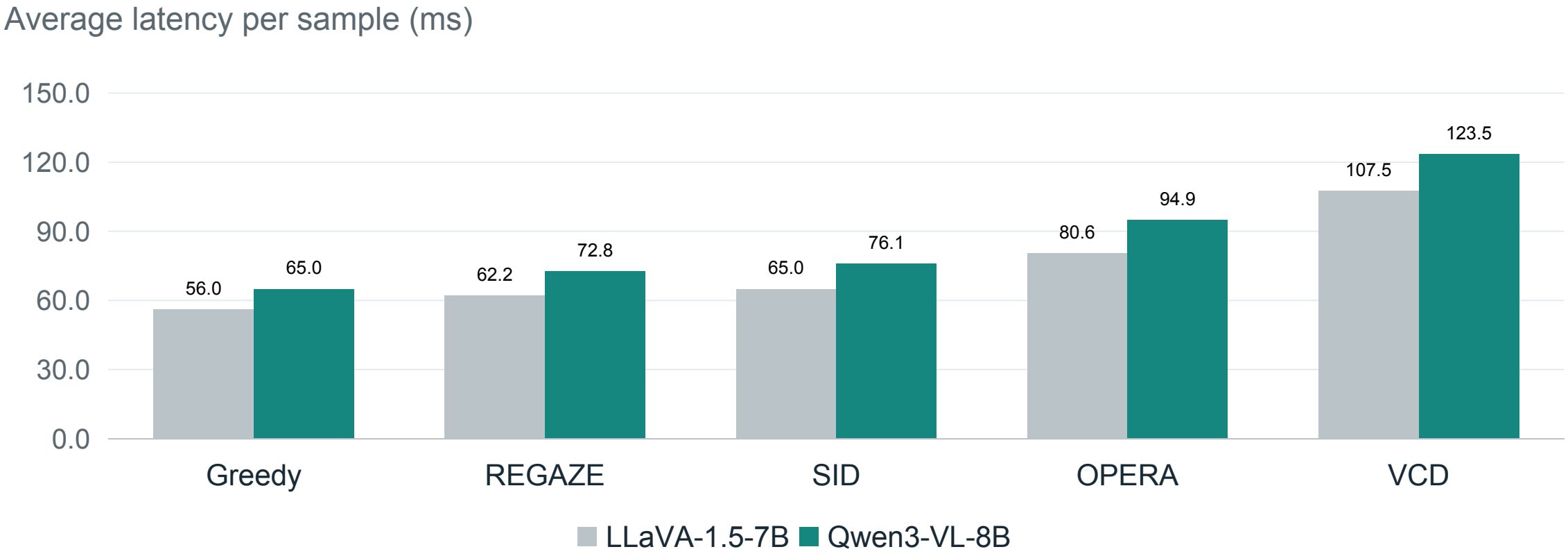
Visual reasoning accuracy gain



Additional model families



Gaze steering adds about 11–12% latency



1.11× on LLaVA-1.5-7B **1.12× on Qwen3-VL-8B**

REGAZE restores balanced visual grounding

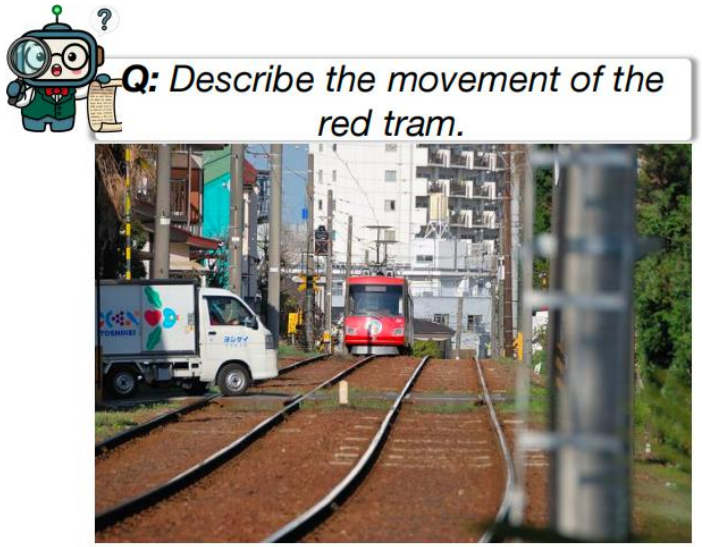
Hallucinated gaze drifts toward locking or collapse.

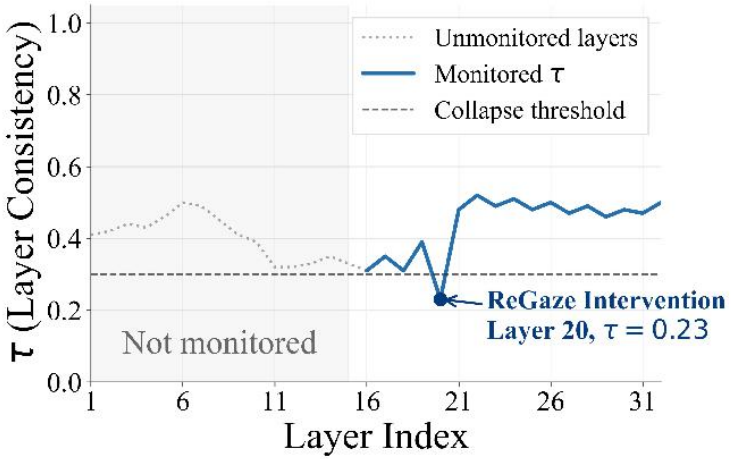
Layer-wise monitoring enables targeted, training-free correction.

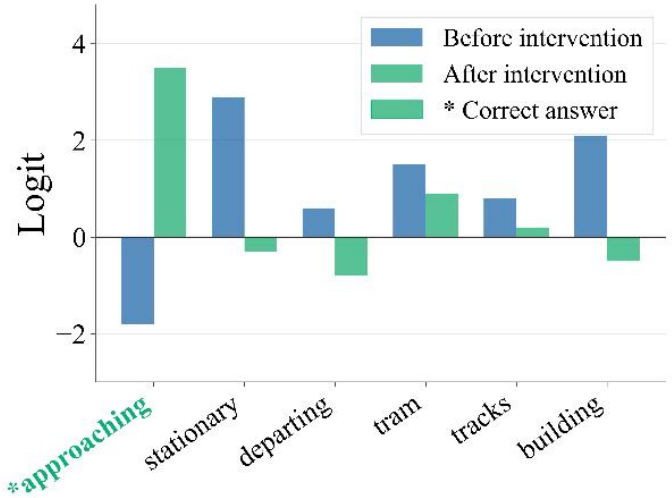
Current scope

Open models with accessible attention states. Evaluation beyond 13B and proprietary models remains open.

A collapse intervention restores the correct candidate







Layer 20: τ = 0.23 triggers collapse calibration

The correct candidate, “approaching”, becomes dominant after intervention.