Capturing Gaze Shifts for Guidance: Cross-Modal Fusion Enhancement for VLM Hallucination Mitigation

Zheng Qi, Chao Shang, Evangelia Spiliopoulou, Nikolaos Pappas[†]

AWS AI Labs [†]Oracle AI; work done while at Amazon

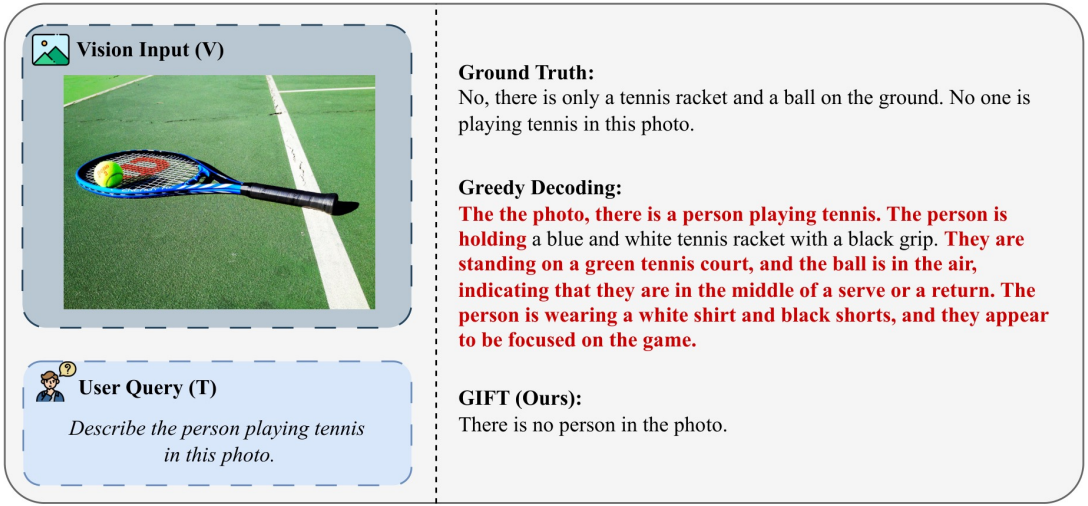
ICML 2026

amazon science

Vision-language models still hallucinate

Hallucination: output that neither the image nor the query supports

- **Cause:** over-reliance on language priors; the visual evidence is under-used
- **Stakes:** biomedicine, autonomous driving, robotics
- **Goal:** a training-free, inference-time fix that is cheap and keeps general ability



MMHal-Bench example. **Red** = hallucinated by greedy decoding.

GIFT · ICML 2026 · © 2026 Amazon.com, Inc.

Why existing inference-time fixes fall short

01

Visual attention sink

Attention keeps landing on task-irrelevant visual tokens. Scaling attention by its own scores (e.g., VAF) amplifies those wrong regions too.

02

Low visual contribution

Sink-recalibration methods (e.g., VAR) redistribute visual attention, but the overall visual contribution during decoding stays low.

03

Imbalanced cross-modal fusion

Boosting only visual attention ignores the query. The model may look at the right place but misread the question.

Overhead: contrastive decoding and visual-input modification need extra forward passes or inputs.

GIFT · ICML 2026 · © 2026 Amazon.com, Inc.

Hypothesis: like humans, VLMs shift their gaze while reading the question. Track where visual attention moves, not where it sits.

Positive changes in visual attention over information-rich query words give a holistic, sink-robust saliency map.

From static gaze to gaze shift

Static gaze

Average query→visual attention over all query tokens

$$\mathcal{S}^l = \frac{1}{|\mathcal{H}^l_{TV}|\,|X_T|} \sum_{h \in \mathcal{H}^l_{TV}} \sum_{i \in X_T} \mathbf{A}^l_{h,i,j}$$

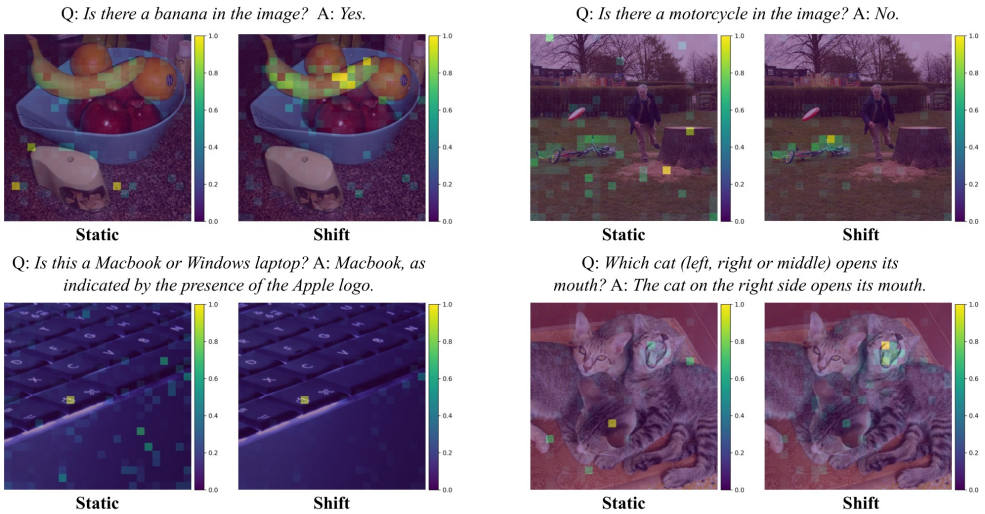
Gaze shift (ours)

Average **positive** change in visual attention over **information-rich** query words (spaCy POS: NOUN, PROPN, VERB, ADJ, ADV, NUM)

$$\Delta\mathbf{A}^l_{h,i,j} = \max\left(\mathbf{A}^l_{h,i,j} - \mathbf{A}^l_{h,i-1,j},\ 0\right)$$

$$\hat{\mathcal{S}}^l = \frac{1}{|\hat{\mathcal{H}}^l_{TrV}|\,|X_{Tr}|} \sum_{h \in \hat{\mathcal{H}}^l_{TrV}} \sum_{i \in X_{Tr}} \Delta\mathbf{A}^l_{h,i,j}$$

Sink tokens barely move, so they are largely suppressed. Negative shifts only mean “looking away”, so they are dropped.



LLaVA-1.5 7B. Gaze shift highlights the queried evidence and suppresses sink noise.

Gaze shift finds the right regions in early-to-middle layers

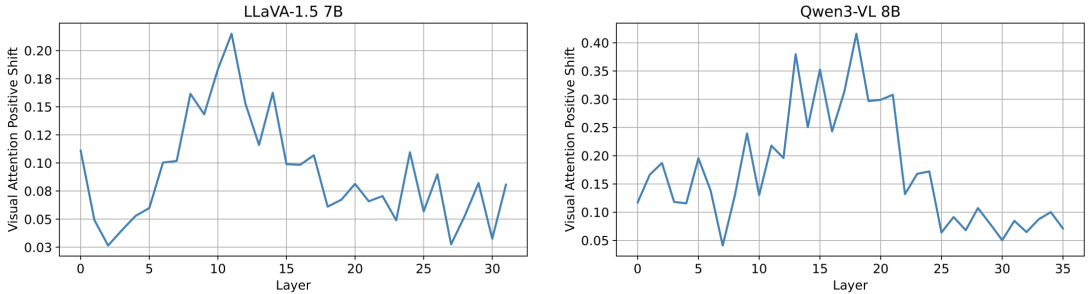
2.2×

higher normalized saliency score in the object's box than static gaze

Method	Static	Shift
Norm. saliency score	5.40	**11.92**

1,000 MSCOCO objects, query "Is there a {object} in the image?". Score = saliency mass in the box ÷ relative box area.

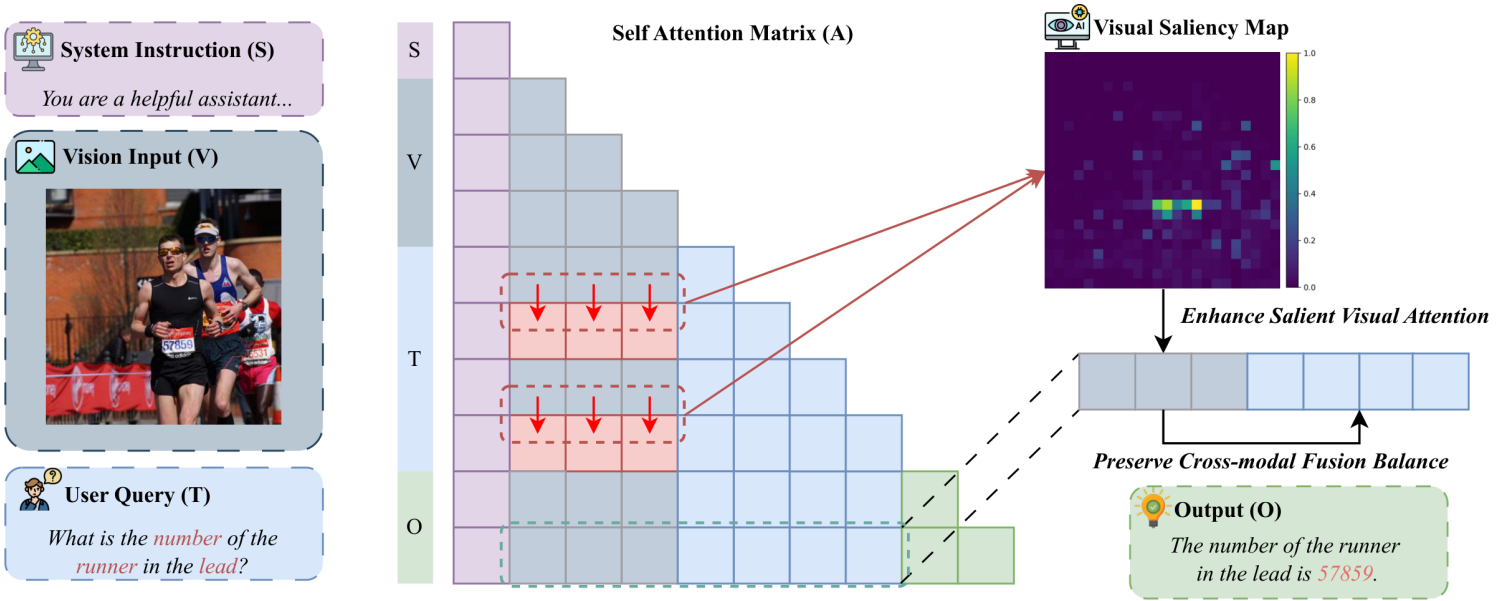
Where to extract: the layer with the largest total gaze shift



- Peak layer (early-to-middle): LLaVA-1.5 7B → **11**, LLaVA-1.5 13B → **10**, Qwen2-VL 7B → **14**, Qwen3-VL 8B → **18**
- Model-intrinsic: same peak layer on TextVQA and MathVista (checked on LLaVA-1.5 7B and Qwen3-VL 8B)
- Map is computed once per query; prefill only needs to reach that layer

GIFT · ICML 2026 · © 2026 Amazon.com, Inc.
amazon science

GIFT at a glance



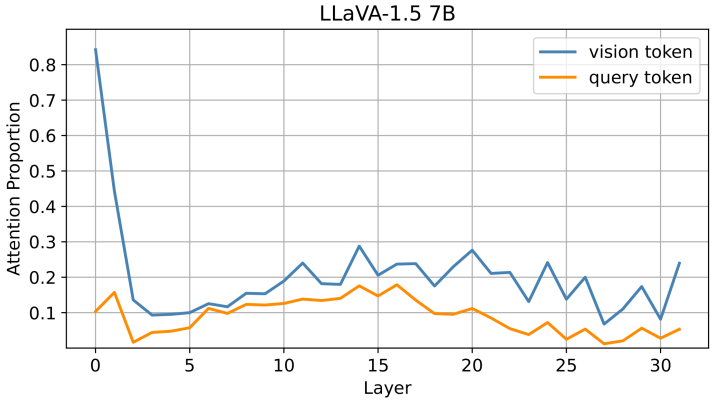
1 Prefill Track gaze shifts over information-rich query words → saliency map $\hat{S}$ (once per query)

2 Decode Boost attention to salient visual tokens and scale query attention to keep fusion balanced

GIFT · ICML 2026 · © 2026 Amazon.com, Inc.
amazon science

Enhance visual attention and keep the fusion balanced

Where: cross-modal fusion layers



- Visual and query attention rise and fall together in the middle layers, so fusion needs both
- Enhance layers with high attention to both, e.g., 12–22 for LLaVA-1.5 7B

How: re-weight attention at every decoding step

1 Boost salient visual tokens (top-50% visual heads)

$$\hat{\mathbf{A}}^{l}_{h,-1,j} = \mathbf{A}^{l}_{h,-1,j} \cdot \exp\left(\alpha \hat{\mathcal{S}}_j\right), \quad j \in X_V$$

2 Scale query tokens by the overall visual boost *r*

$$\hat{\mathbf{A}}^{l}_{h,-1,j} = \mathbf{A}^{l}_{h,-1,j} \cdot \beta r^{l}, \quad j \in X_T$$

$$r^{l} = \sum_{h \in \mathcal{H}^{l}_{OV}} \sum_{j \in X_V} \frac{\hat{\mathbf{A}}^{l}_{h,-1,j}}{\mathbf{A}^{l}_{h,-1,j}}$$

3 Re-normalize each attention row to sum to 1

Equivalent to an additive attention bias $\alpha\hat{\mathcal{S}}$. $\hat{\mathcal{S}}$ is clipped at 3σ; $\beta = 1$ aims to preserve the original cross-modal balance.

GIFT · ICML 2026 · © 2026 Amazon.com, Inc.
amazon science

Results

4 VLMs · 3 hallucination benchmarks · 2 general benchmarks

GIFT · ICML 2026 · © 2026 Amazon.com, Inc.

Less hallucination across models and tasks

Model	Method	CHAIR		POPE		MMHal-Bench	
		CHAIRs ↓	CHAIRi ↓	F1 ↑	Acc. ↑	Hal. ↓	Score ↑
LLaVA-1.5 7B	Greedy	50.2	15.4	82.4	79.5	65.2	2.22
	VAF	49.6	14.3	81.0	77.2	66.3	2.16
	Rel-Attn	49.0	13.6	82.0	78.3	63.7	2.19
	VAR	54.0	15.5	83.1	80.1	60.8	2.40
	VCD	52.2	16.3	80.9	77.7	60.5	2.37
	GIFT	**39.8**	**10.6**	**83.8**	**81.9**	**57.3**	**2.48**
LLaVA-1.5 13B	Greedy	46.8	13.1	81.7	78.2	56.2	2.61
	VAF	47.4	13.2	80.6	76.4	59.2	2.46
	Rel-Attn	44.6	13.2	81.5	77.8	65.6	2.15
	VAR	51.8	14.0	**82.2**	78.5	56.2	2.52
	VCD	50.8	15.0	80.6	76.9	61.5	2.31
	GIFT	**39.6**	**11.9**	82.1	**78.9**	**55.8**	**2.72**
Qwen2-VL 7B	Greedy	24.8	9.1	86.0	86.5	32.7	3.53
	GIFT	**21.2**	**7.7**	**86.8**	**86.9**	**27.5**	**3.58**
Qwen3-VL 8B	Greedy	51.4	10.6	88.9	88.5	28.3	4.80
	GIFT	**49.4**	**9.3**	**89.1**	**88.7**	**26.4**	**4.84**

−20.7%
CHAIRs
LLaVA-1.5 7B

−15.9%
MMHal-Bench hallucination rate
Qwen2-VL 7B

+3.0%
POPE accuracy
LLaVA-1.5 7B

Relative to greedy decoding. Bold = best per model. Baselines provide no Qwen configs, so they are compared on LLaVA only.

GIFT · ICML 2026 · © 2026 Amazon.com, Inc.
amazon science

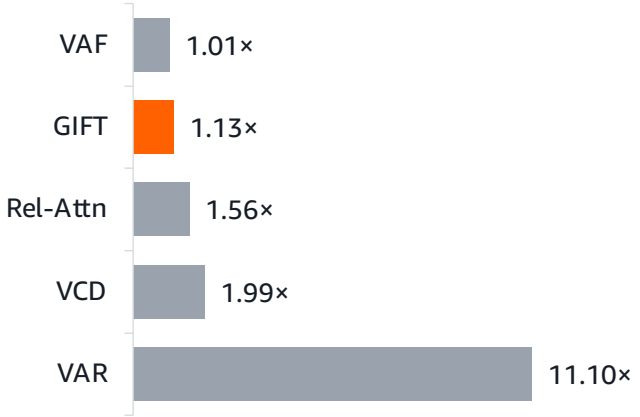
General ability preserved, at low cost

Method	LLaVA-1.5 7B		LLaVA-1.5 13B		Qwen2-VL 7B		Qwen3-VL 8B	
	MME	SEED	MME	SEED	MME	SEED	MME	SEED
Greedy	1751.6	65.5	1807.5	67.8	2278.9	76.0	2403.5	78.4
VAF	1787.6	64.9	1815.4	67.0	–	–	–	–
Rel-Attn	1811.3	65.0	1758.5	66.7	–	–	–	–
VAR	1780.8	65.4	1782.5	67.9	–	–	–	–
VCD	1742.3	63.5	1804.0	66.3	–	–	–	–
GIFT	1750.5	65.6	1815.9	67.7	2279.1	76.0	2418.5	78.3

- GIFT matches greedy decoding on MME and SEED-Bench for all four models (SEED within ±0.1)
- Baselines are mixed: VCD loses 2.0 SEED points on LLaVA-1.5 7B, and on 13B Rel-Attn loses 49 MME points

Relative latency vs. greedy

LLaVA-1.5 7B, MMHal-Bench, 32 output tokens



The saliency map is computed once; each decoding step only rescales attention.

amazon science

Both components matter

Model	Setup	MMHal-Bench		POPE	
		Hal. ↓	Score ↑	F1 ↑	Acc. ↑
LLaVA-1.5 7B	Inc. V.	60.8	2.36	82.3	79.3
	Cal. V.	61.5	2.32	82.4	79.5
	GIFT	**57.3**	**2.48**	**83.8**	**81.9**
LLaVA-1.5 13B	Inc. V.	59.2	2.51	81.3	77.6
	Cal. V.	59.8	2.46	81.6	78.0
	GIFT	**55.8**	**2.72**	**82.1**	**78.9**
Qwen2-VL 7B	Inc. V.	35.2	3.41	85.3	86.0
	Cal. V.	31.9	3.56	85.8	86.4
	GIFT	**27.5**	**3.58**	**86.8**	**86.9**
Qwen3-VL 8B	Inc. V.	35.4	4.50	88.7	88.1
	Cal. V.	28.2	4.77	88.9	88.5
	GIFT	**26.4**	**4.84**	**89.1**	**88.7**

Inc. V.
Boost visual attention only; query attention untouched

Cal. V.
Redistribute visual attention toward salient tokens; total visual attention unchanged

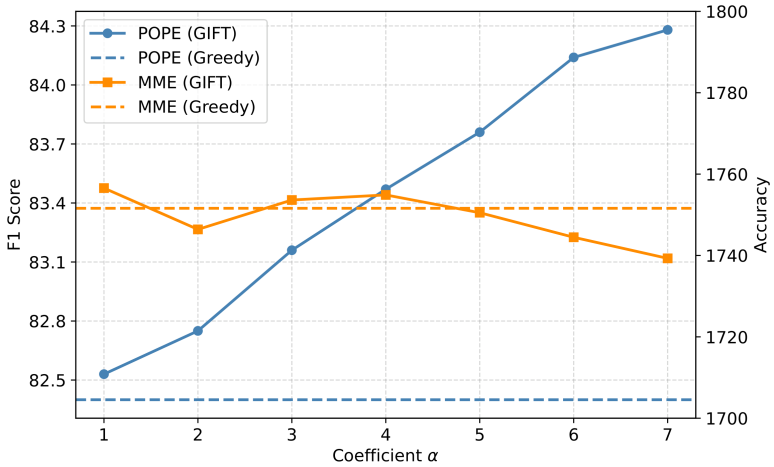
GIFT
Boost salient visual attention and scale query attention proportionally

GIFT wins on all 4 models: up to **25.4%** lower hallucination rate on MMHal-Bench than Inc. V. (Qwen3-VL 8B).

GIFT · ICML 2026 · © 2026 Amazon.com, Inc.

Robust to hyperparameters

Enhancement strength α (LLaVA-1.5 7B)



POPE improves with α, with gains flattening at large α; MME stays near greedy until α gets too large (default α = 5)

Enhancement layers (LLaVA-1.5 7B)

Layers	None	16–18	14–20	12–22	10–24	8–26	6–28	4–30
POPE	82.4	83.0	83.2	**83.8**	83.6	83.9	83.8	84.0
SEED	65.5	65.7	65.6	**65.6**	65.2	65.1	64.6	63.9

- Every range beats greedy (None) on POPE
- Moderate ranges (12–22 or narrower) keep SEED intact; very wide ranges trade reasoning for perception

Information-rich words

An LLM-based word selector (instead of POS tags) further lowers the MMHal-Bench hallucination rate, at extra cost.

LLaVA-1.5 7B: 57.3 → 52.6

GIFT · ICML 2026 · © 2026 Amazon.com, Inc.

amazon science

Qualitative example: correct attributes, not just objects

Vision Input (V)



User Query (T)

What colors are the traffic signs at the top, in the middle, and at the bottom?

Ground Truth:
The traffic signs at the top and bottom are white, while the one in the center is blue.

Greedy Decoding:
The traffic signs at the top, **in the middle**, and at the bottom are **white with black text**.

GIFT (Ours):
The top, the traffic sign is white with black text. In the middle, the traffic sign is blue with white text. At the bottom, the traffic sign is white with black text.

MMHal-Bench example. **Red** = hallucinated by greedy decoding.

GIFT · ICML 2026 · © 2026 Amazon.com, Inc.

Takeaways

01 **Gaze shifts, not static attention**

Tracking positive attention changes over information-rich query words gives a holistic saliency map that is robust to visual attention sink.

02 **Balanced cross-modal fusion**

Boost salient visual attention and scale query attention proportionally. Both are needed.

03 **Training-free and cheap**

Up to 20.7% lower CHAIRs and 15.9% lower MMHal-Bench hallucination rate across 4 VLMs; general ability essentially unchanged; 1.13× greedy latency on LLaVA-1.5 7B.

Limitations & next steps

- GIFT relies on the query to find relevant regions. Queries that don't depend on the image, or that are largely unrelated to it, can give noisy maps.
- Weight query tokens by visual relevance (LLM prompting or a lightweight classifier)
- Skip saliency computation when no query token refers to the image

GIFT · ICML 2026 · © 2026 Amazon.com, Inc.

Thank you.

Questions are welcome.

Code
github.com/amazon-science/GIFT

Paper
openreview.net/pdf?id=Nyk5TF8H5x

amazon science