When RAG Hurts: Diagnosing and Mitigating Attention Distraction in Retrieval-Augmented LVLMs

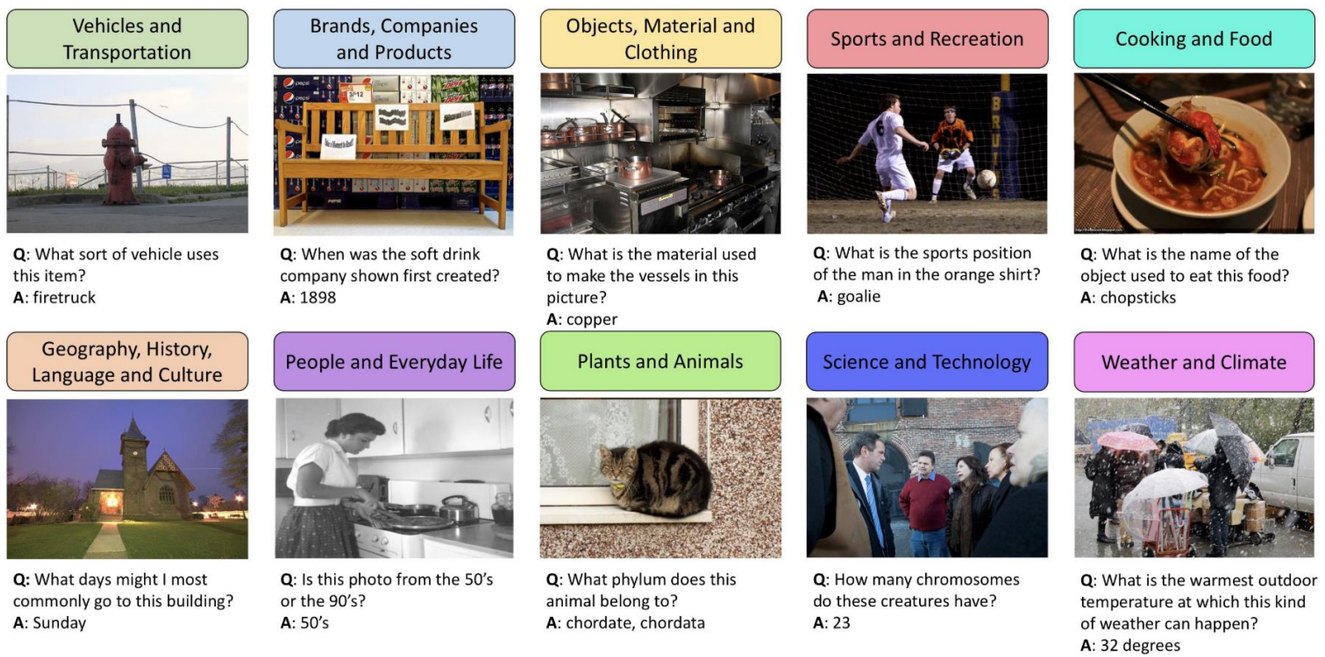
Authors: Beidi Zhao, Wenlong Deng, Xinting Liao, Yushu Li, Nazim Shaikh, Yao Nie*, Xiaoxiao Li*

Presenter: Beidi Zhao

Scan for paper

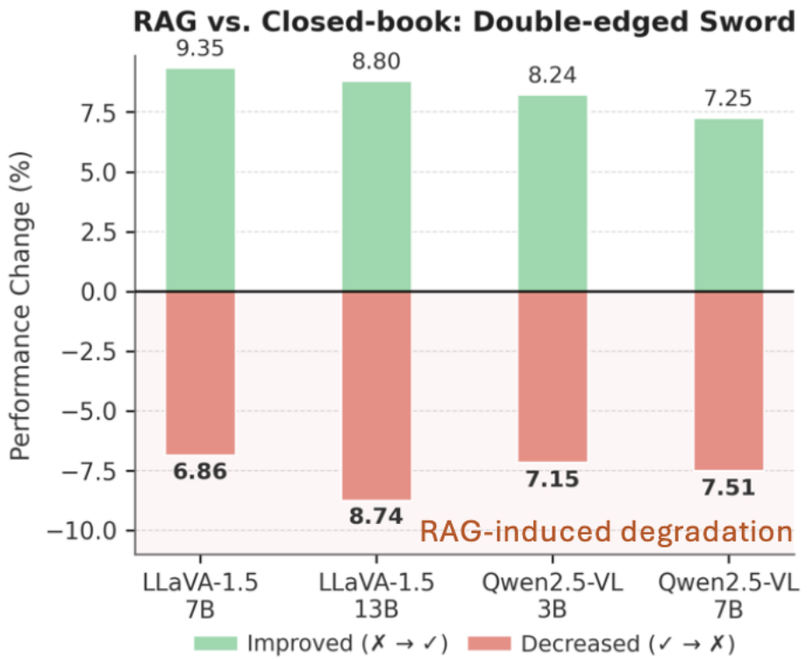
Background: Knowledge-Based VQA (KB-VQA)

- **Challenge:** LVLMs (Large Vision-Language Models) require external factual knowledge to answer specific queries, as parametric knowledge is often insufficient.
- **Solution:** Retrieval-Augmented Generation (RAG).
  - Retrieves external documents to supplement the model input at inference time.
  - Standard Input: [Image, Question, Context].



Motivation: RAG as a Double-Edged Sword

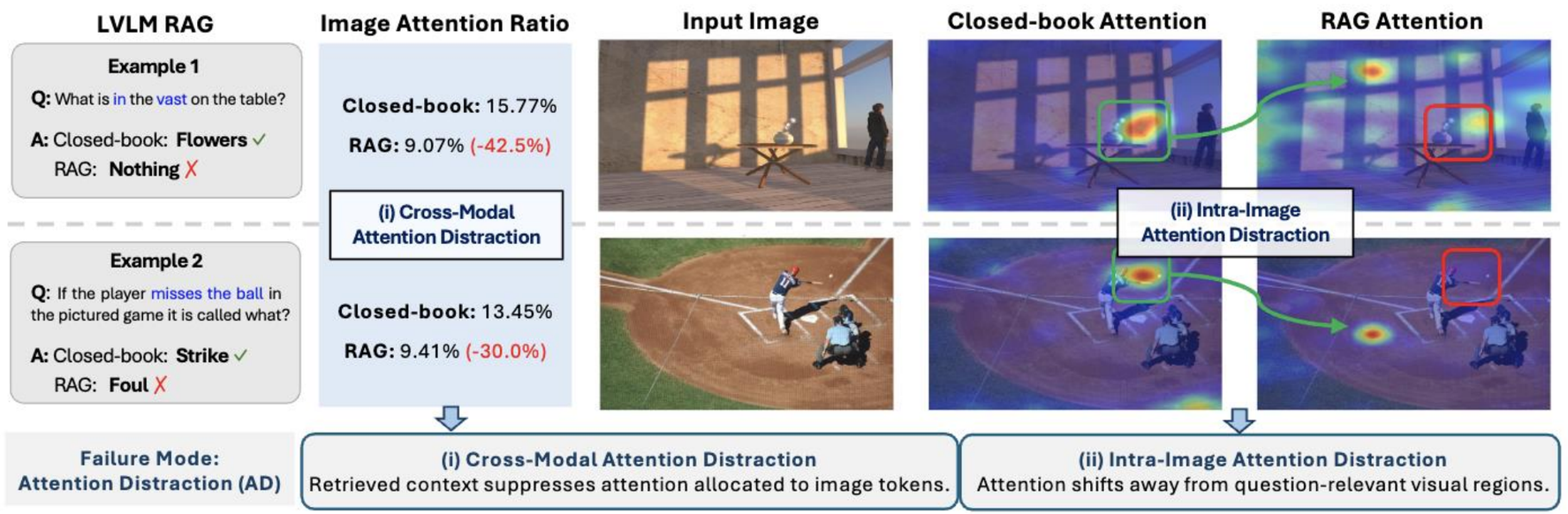
- **The Paradox**
  - While RAG improves average performance, it **degrades** a non-trivial subset of instances.
  - **Phenomenon:** The model answers correctly in a "Closed-book" setting (no context) but answers **incorrectly** when provided with *correct* and *relevant* retrieved context.
  - **Conclusion:** This is not a retrieval quality issue, but an inference dynamics issue.



What explains this paradox?
Why does accurate retrieval cause failures on questions answerable from parametric knowledge alone?

Diagnosing the Failure: Attention Distraction (AD)

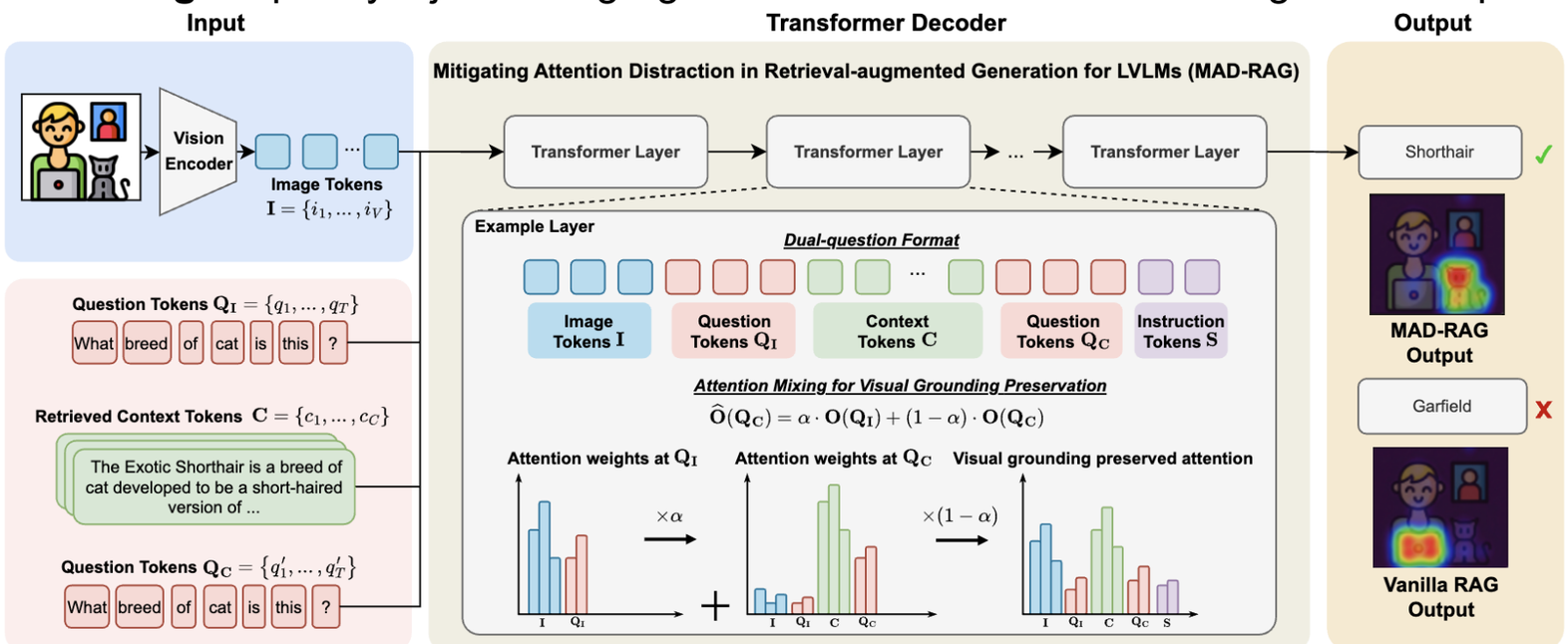
- **Definition: Attention Distraction (AD)**
  - A ***failure mode*** where retrieved text disrupts the model's attention mechanisms during multimodal generation in retrieval-augmented LVLMs.
- **Two Types of AD:**
  - Cross-Modal Attention Distraction
  - Intra-Image Attention Distraction



Proposed Method: MAD-RAG

MAD-RAG: Mitigating Attention Distraction in RAG

- **Goal:** Restore visual grounding without discarding retrieved knowledge.
- **Nature:** Training-free, inference-time intervention.
- **Key Components:**
  - **Dual-Question Input Formulation:** Decouples visual perception from context integration.
  - **Attention Mixing:** Explicitly injects image-grounded attention back into the generation process.



Proposed Method: MAD-RAG

Dual-Question Formulation

- $\mathbf{X}_{\mathrm{MAD-RAG}} = [\mathbf{X}_{\mathbf{V}}, \mathbf{X}_{\mathbf{Q_I}}, \mathbf{X}_{\mathbf{C}}, \mathbf{X}_{\mathbf{Q_C}}]$
- $\mathbf{Q_I}$ (Image-Question): Placed immediately after the image. Its role is to focus purely on visual information.
- $\mathbf{Q_C}$ (Context-Question): Placed after the context. Its role is to integrate external knowledge.

Attention Mixing for Visual Grounding Preservation

- Due to the 'recency bias' of Transformers [1,2], the context will still dominate $\mathbf{Q_C}$.
- At every decoding layer, we force $\mathbf{Q_C}$ to 'borrow' the visual focus from $\mathbf{Q_I}$.

$$\widehat{\mathbf{O}}(\mathbf{Q_C}) = \alpha \cdot \mathbf{O}(\mathbf{Q_I}) + (1 - \alpha) \cdot \mathbf{O}(\mathbf{Q_C})$$

- **Mechanism:**
  - Transfers image-grounded attention from $\mathbf{Q_I}$ to $\mathbf{Q_C}$.
  - Preserves the context awareness of $\mathbf{Q_C}$.

[1] Press et al., Train short, test long: Attention with linear biases enables input length extrapolation, ICLR 2022.
[2] Zhao et al., Calibrate before use: Improving few-shot performance of language models. PMLR 2021.

Experimental Setup

- **KB-VQA datasets:** OK-VQA, E-VQA, InfoSeek
- **Model families and scales:** LLaVA-1.5-7B, LLaVA-1.5-13B, Qwen2.5-VL-3B, Qwen2.5-VL-7B, Qwen3-VL-4B, Qwen3-VL-8B
- **Baselines:**
  - RAG-oriented methods: CAD (Shi et al., 2024), ADA-CAD (Wang et al., 2025), ALFAR (An et al., 2025)
  - VLM hallucination mitigation methods: DoLa (Chuang et al., 2023), VCD (Leng et al., 2024), OPERA (Huang et al., 2024), VHR (He et al., 2025), SPIN (Sarkar et al., 2025)
- **Retriever:** Oracle chunks, CLIP-L/14-336

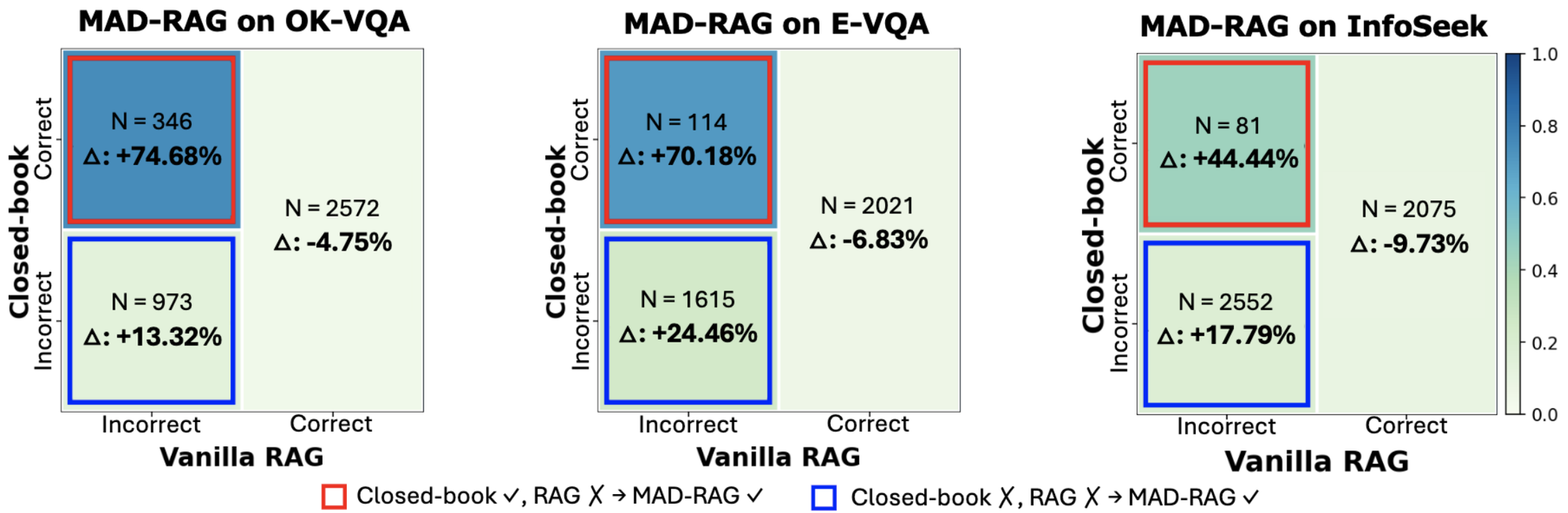
Main Results: State-of-the-Art Performance

- MAD-RAG achieves State-of-the-Art performance, up to 9.2% improvement.

Method	LLaVA-1.5-7B			LLaVA-1.5-13B			Qwen2.5-VL-3B			Qwen2.5-VL-7B		
	OK	E-VQA	Info	OK	E-VQA	Info	OK	E-VQA	Info	OK	E-VQA	Info
Closed-book (parametric)	61.65	17.57	8.01	64.32	18.80	8.14	59.97	22.93	17.88	61.92	25.15	19.58
Vanilla RAG	65.46	53.89	44.01	66.76	61.23	46.33	62.24	75.63	51.44	63.80	82.21	45.71
I. RAG-Oriented Methods												
CAD Shi et al. (2024)	68.44	54.88	39.06	69.38	60.19	39.06	66.24	74.11	47.98	**65.45**	77.89	44.12
AdaCAD Wang et al. (2025b)	67.01	54.69	46.45	66.15	59.47	46.45	59.11	73.04	52.04	58.84	75.65	46.33
ALFAR An et al. (2025a)	65.32	53.20	38.70	67.66	60.36	45.11	62.63	75.97	51.44	63.55	82.40	48.09
II. VLM Hallucination Mitigation Methods (Visual-Grounding-based)												
DoLA Chuang et al. (2023)	65.65	51.01	43.01	66.02	58.21	44.69	62.09	72.40	52.25	63.77	75.50	47.28
VCD Leng et al. (2024)	64.48	53.20	43.86	65.55	57.92	43.61	60.81	71.57	51.68	63.17	74.99	46.98
OPERA† Huang et al. (2024)	66.47	48.56	43.03	67.55	57.47	44.03	-	-	-	-	-	-
VHR‡ He et al. (2025)	61.99	53.04	38.06	64.13	58.88	40.19	34.67	63.52	28.16	42.91	58.67	24.04
SPIN† Sarkar et al. (2025)	61.55	48.72	40.97	65.58	57.95	45.35	-	-	-	-	-	-
MAD-RAG (ours)	**70.22**	**63.09**	**50.19**	**71.32**	**69.07**	**51.59**	**66.44**	**80.11**	**54.06**	64.91	**83.31**	**48.43**

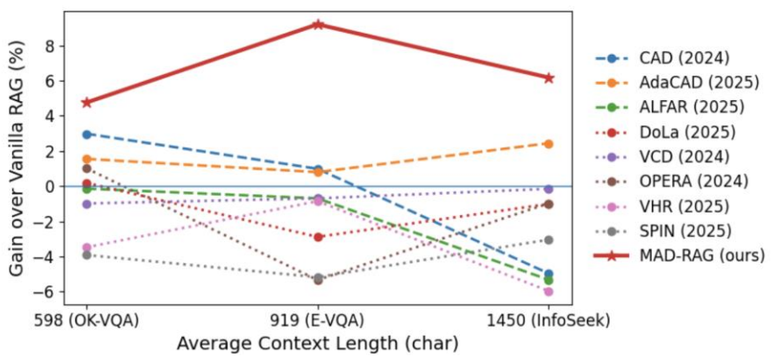
Effectiveness Analysis: Recovering Failures

- We analyze cases where **Closed-book was Correct** but **Vanilla RAG was Incorrect** (The AD failure mode). (red box)
- **Recovery Rate:** OK-VQA: **74.68%**; E-VQA: **70.18%**; InfoSeek: **44.44%** recovered.
- MAD-RAG also improves performance on 'hard' cases where both failed, suggesting it effectively combines noisy context with visual cues. (blue box)

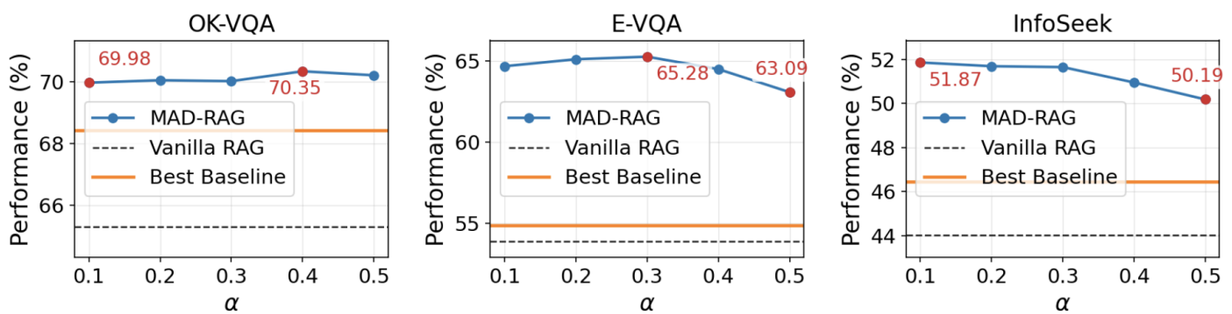


Robustness & Sensitivity

- **Robustness**
  - Context length
  - Retriever and number of chunks
- **Sensitivity**
  - Not sensitive to the value of α



Retriever	Num.	Accuracy	
		Closed-book	Vanilla RAG	MAD-RAG (ours)
Oracle	1	8.01	43.31	**50.19**
	3		44.01	**51.08**
	5		40.46	**52.00**
CLIP-L	Top-1		24.21	**27.32**
	Top-3		15.68	**22.81**
	Top-5		9.86	**21.20**



Efficiency Analysis

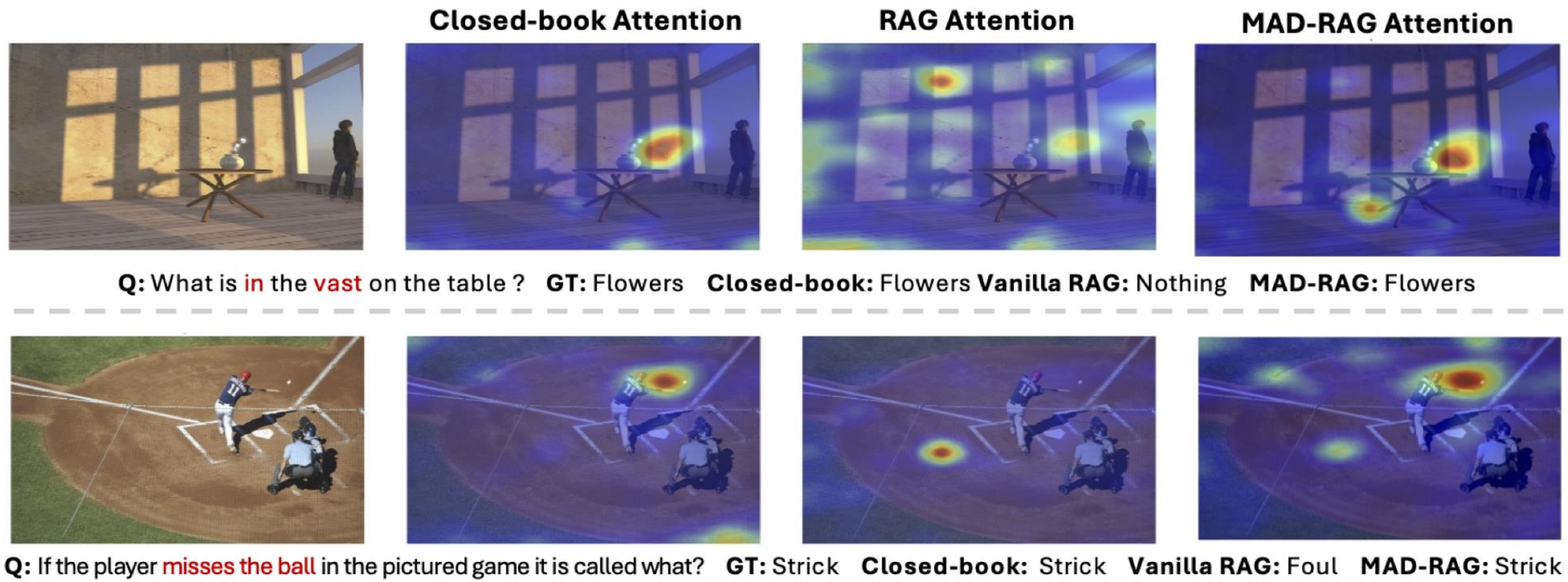
- MAD-RAG only adds about **9-11%** overhead compared to Vanilla RAG.

Table 2: **Inference time comparison with LLaVA-1.5-13B (s/sample).**

Method	OKVQA	E-VQA	Infoseek
Vanilla RAG	0.30 (1.00x)	0.29 (1.00x)	0.31 (1.00x)
CAD (2024)	0.94 (3.12x)	1.29 (4.45x)	1.26 (4.05x)
AdaCAD (2025)	0.49 (1.62x)	0.57 (1.95x)	0.51 (1.66x)
ALFAR (2025)	0.41 (1.36x)	0.40 (1.38x)	0.44 (1.41x)
DOLA (2023)	0.42 (1.39x)	0.53 (1.81x)	0.47 (1.52x)
VCD (2024)	0.48 (1.59x)	0.46 (1.59x)	0.49 (1.59x)
OPERA (2024)	0.92 (3.06x)	0.92 (3.18x)	1.04 (3.37x)
VHR (2025)	0.35 (1.15x)	0.37 (1.27x)	0.38 (1.22x)
SPIN (2025)	0.47 (1.55x)	0.58 (1.97x)	0.56 (1.78x)
MAD-RAG (ours)	**0.34 (1.10x)**	**0.33 (1.11x)**	**0.34 (1.09x)**

Qualitative Result

- The attention is back to the question-related regions



Takeaways

- **Identified Attention Distraction (AD):** a distinct failure mode in retrieval-augmented LVLMs.
  - Cross-Modal AD
  - Intra-Image AD
- **Proposed MAD-RAG:** a training-free, plug-and-play intervention applied at inference time.
  - Dual-Question Formulation
  - Attention Mixing
- **Key Results & Impact**
  - **State-of-the-Art:** Achieved consistent performance gains across OK-VQA (**+4.76%**), E-VQA (**+9.20%**), and InfoSeek (**+6.18%**)
  - **Failure Recovery:** Successfully rectified up to **74.68%** of RAG-induced error cases
  - **Efficiency:** Incurred negligible computational overhead (**~9-11%**) compared to existing decoding strategies

Thanks for your Attention!

Scan for paper