Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs

D. Caffagni, F. Cocchi, N. Moratelli, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara

{name.surname}@unimore.it

AImage Lab — UNIMORE Università degli Studi di Modena e Reggio Emilia

CVPR JUNE 17-21, 2024 SEATTLE, WA

Why Retrieval-Augmented MLLMs?

⇒ Multimodal LLMs (MLLMs) [1] exhibit limitations when faced with **highly specific** queries.

In what state is this building located?

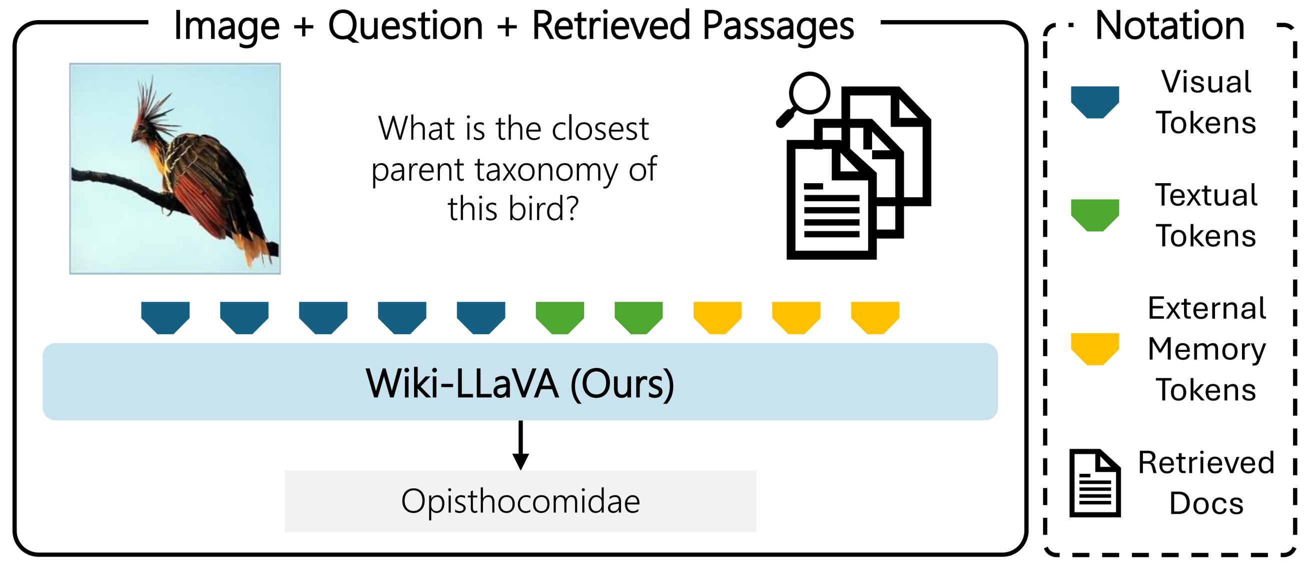
LLaVA-1.5: California ✗
BLIP-2: Florida ✗

Where this fish is found?

LLaVA-1.5: Gulf of Mexico ✗
BLIP-2: Alaska ✗

Having access to external documents at generation time can be a powerful source of information!

Architecture Overview



⇒ We enable MLLMs to answer **complex and specific questions** that cannot be resolved through image content and pre-trained knowledge alone.

⇒ The model leverages diverse information in its responses and learns to discern the **relative importance** of each retrieved document.

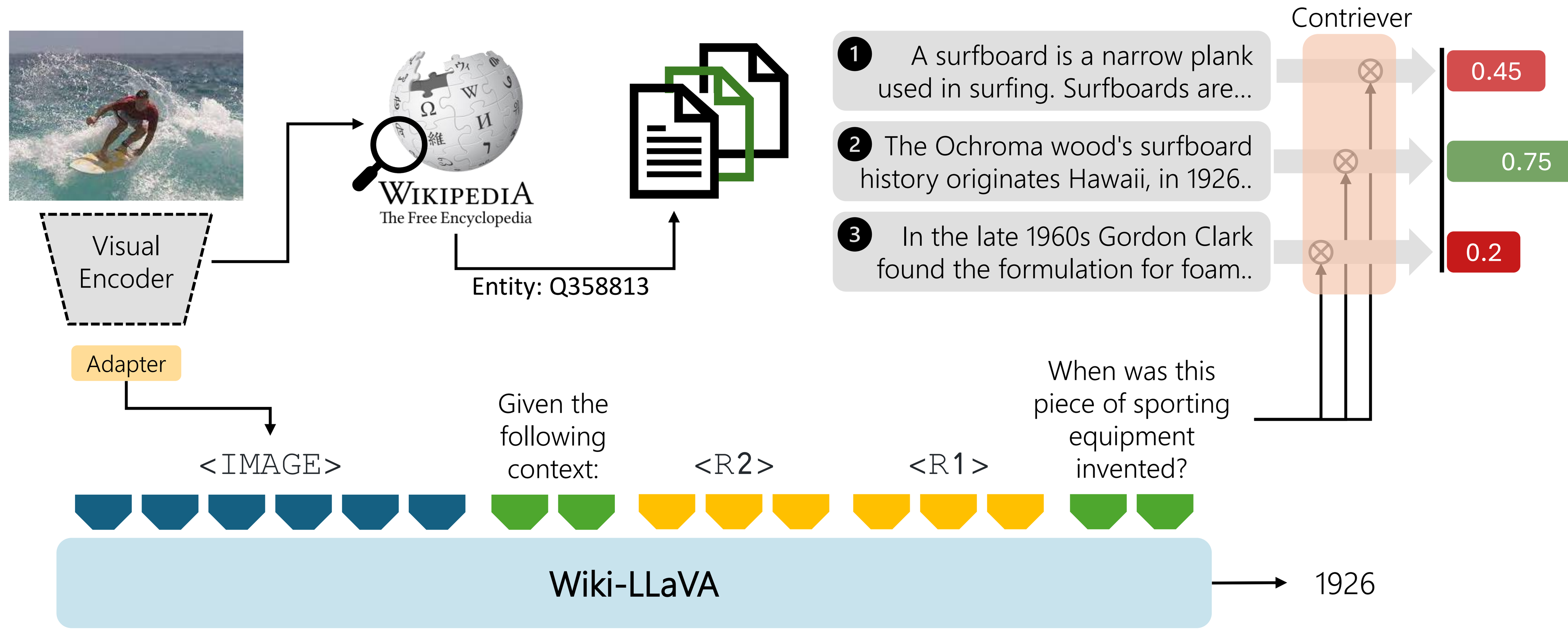
References

[1] D. Caffagni et al. *The Revolution of Multimodal Large Language Models: A Survey*, ACL Findings 2024.

[2] C. Yang et al. *Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions?*, EMNLP 2023.

[3] T. Mensink et al. *Encyclopedic VQA: Visual Questions About Detailed Properties of Fine-Grained Categories*, ICCV 2023.

Wiki-LLaVA



⇒ Wiki-LLaVA integrates **knowledge** derived from an external memory as **additional input context** into the LLaVA model

→ we do not need to change the underlying LLM architecture.

⇒ The external memory comprises (document, image, text-title) triplets **from Wikipedia** web pages [2, 3].

Hierarchical Retrieval & Results

(1) Cross-modal retrieval is performed using an input image I, k-NN is performed within the external memory, using document titles as keys

⇒ the top-k documents are retrieved.

(2) Each retrieved doc is analyzed

⇒ the Contriever model identifies the most relevant passages given the input question.

(3) The raw content of these most relevant passages is employed as additional context.

Formally, the final input context is:

$$\underbrace{v_o, v_1, \ldots, v_N}_{\text{Visual tokens}}, \quad \underbrace{w_0, w_1, \ldots, w_{t-1}}_{\text{System + user prompt}}, \quad \underbrace{e_0, e_1, \ldots, e_\tau}_{\text{External memory tokens}}$$

Model	LLM	KB	k	n	Enc-VQA Single-Hop	Enc-VQA All	InfoSeek Unseen-Q	InfoSeek Unseen-E	InfoSeek All
Zero-shot Models									
BLIP-2	Flan-T5$_{XL}$	✗	-	-	12.6	12.4	12.7	12.3	12.5
InstructBLIP	Flan-T5$_{XL}$	✗	-	-	11.9	12.0	8.9	7.4	8.1
LLaVA-1.5	Vicuna-7B	✗	-	-	16.3	16.9	9.6	9.4	9.5
Fine-tuned Models									
LLaVA-1.5	Vicuna-7B	✗	-	-	23.3	28.5	19.4	16.7	17.9
Wiki-LLaVA	Vicuna-7B	✓	1	1	21.8	26.4	26.6	24.6	25.5
Wiki-LLaVA	Vicuna-7B	✓	1	2	19.9	23.2	29.1	26.3	27.6
Wiki-LLaVA	Vicuna-7B	✓	2	1	21.3	25.4	27.8	24.6	26.1
Wiki-LLaVA	Vicuna-7B	✓	1	1	34.7	37.2	41.1	41.1	41.1
Wiki-LLaVA	Vicuna-7B	✓	1	2	39.2	40.2	49.1	46.5	47.8

Takeaways

Standard MLLMs struggle to answer questions correctly, as they rely **solely** on embedded knowledge.

⇒ Fine-tuning helps, but **retrieval is vital** to answer knowledge-intensive questions.

⇒ Training on a **mixture of datasets** preserves fluency on more general benchmarks, without sacrificing VQA performance.

But...there is large room for improvement!

⇒ **CLIP struggles** with fine-grained image-text retrieval, especially when the dataset size increases.

Dataset	KB	R@1	R@10	R@20	R@50
Encyclopedic-VQA	2M	3.3	9.9	13.2	17.5
InfoSeek	100k	36.9	66.1	71.9	78.4

⇒ Oracle entities improve accuracy but remain challenging to find answers from the correct web page.

Directly using retrieved passages to augment pre-trained MLLMs is effective, but requires a robust entity retrieval model to avoid noisy content.

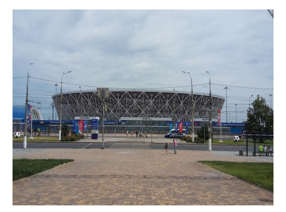
Qualitative Results

Who designed this building?

LLaVA-1.5: Architect ✗
Wiki-LLaVA: James of St.George ✓

What is the oldest age of this animal?

LLaVA-1.5: 10 years ✗
Wiki-LLaVA: 24.9 ✓



What is the name of the main club of this stadium?

LLaVA-1.5: Real Madrid ✗
Wiki-LLaVA: FC Dynamo Kyiv ✗

When was this building constructed?

LLaVA-1.5: 1970 ✗
Wiki-LLaVA: 1927 ✓