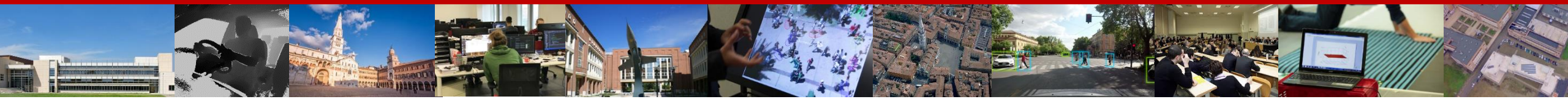
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

Towards an Italian Large Multimodal Language Model

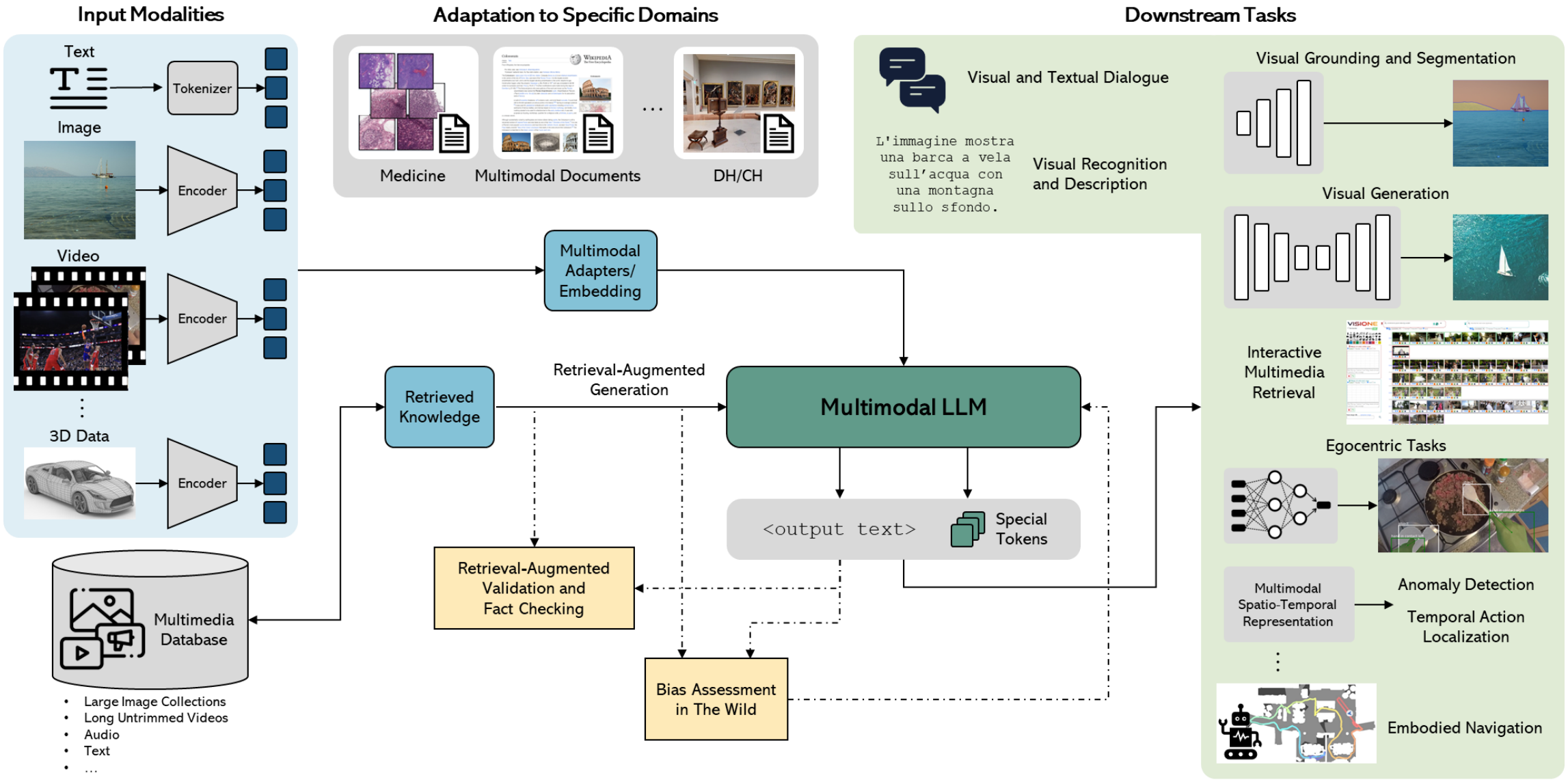


Lorenzo Baraldi, Marcella Cornia, Rita Cucchiara

Panel “IA generativa: verso un Large Language Model italiano“ Conferenza GARR 2024

Future Artificial Intelligence Research

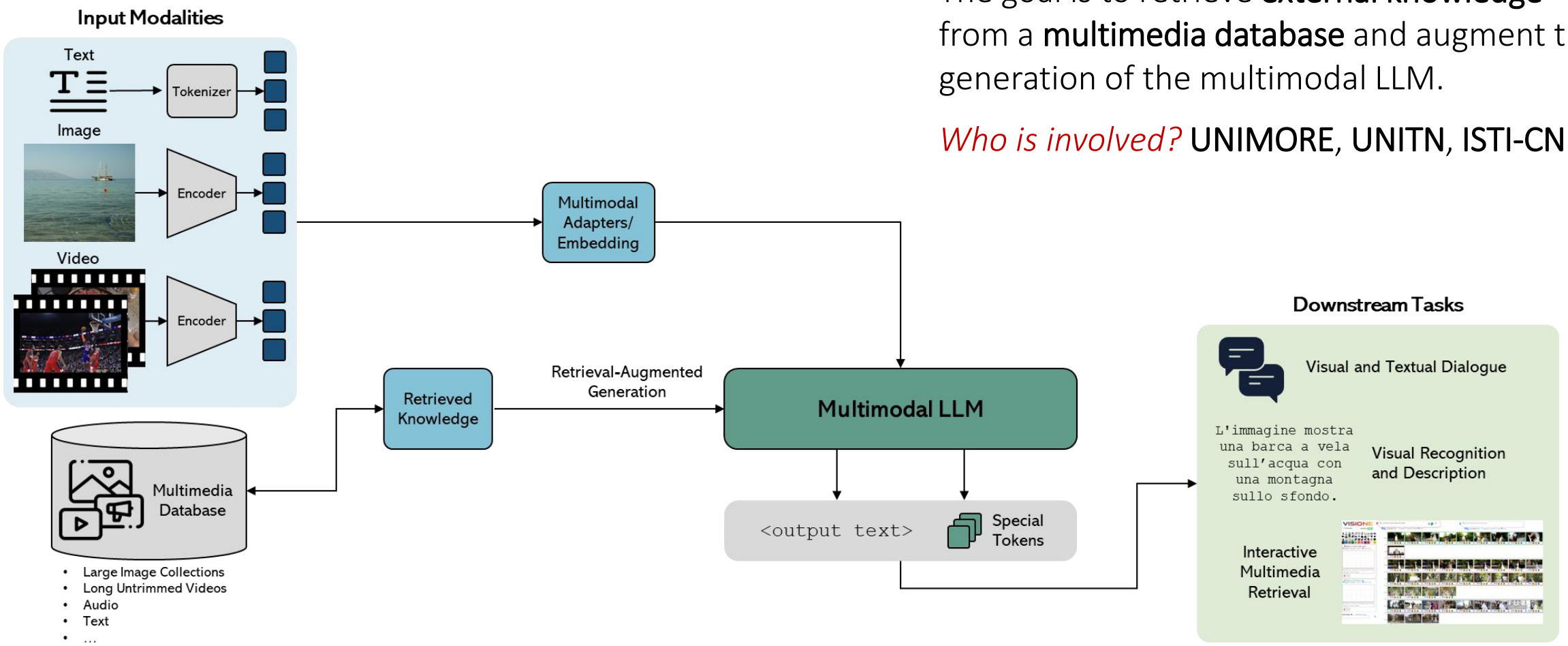
The Full Picture



What we are working on

The goal is to retrieve **external knowledge** from a **multimedia database** and augment the generation of the multimodal LLM.

Who is involved? UNIMORE, UNITN, ISTI-CNR

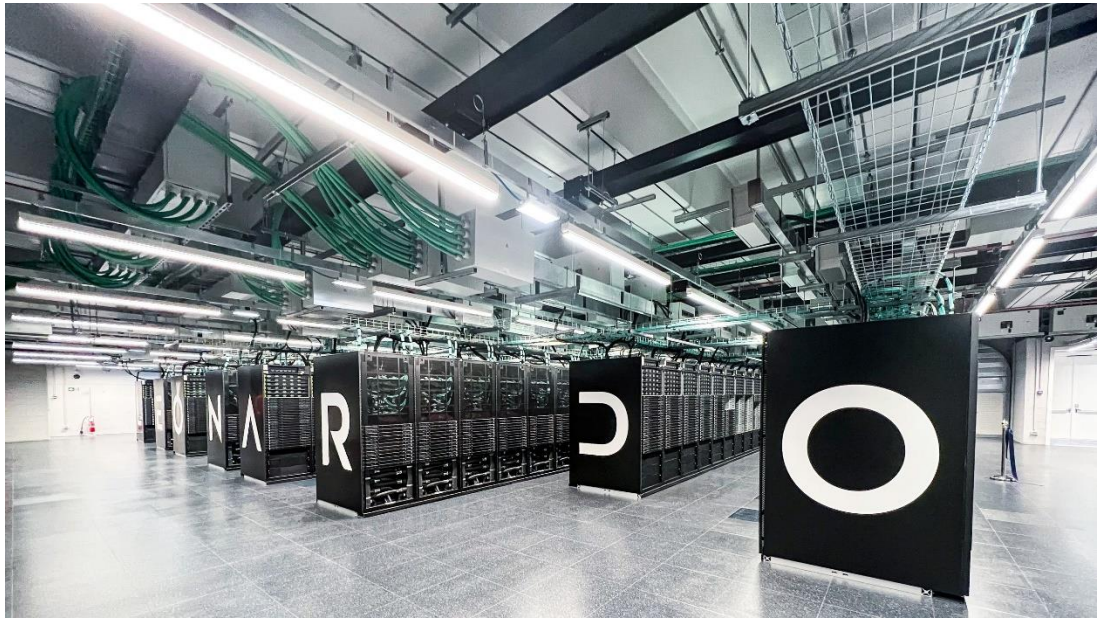


MORE - Multimodal Retrieval Augmented Vision-and-Language Model

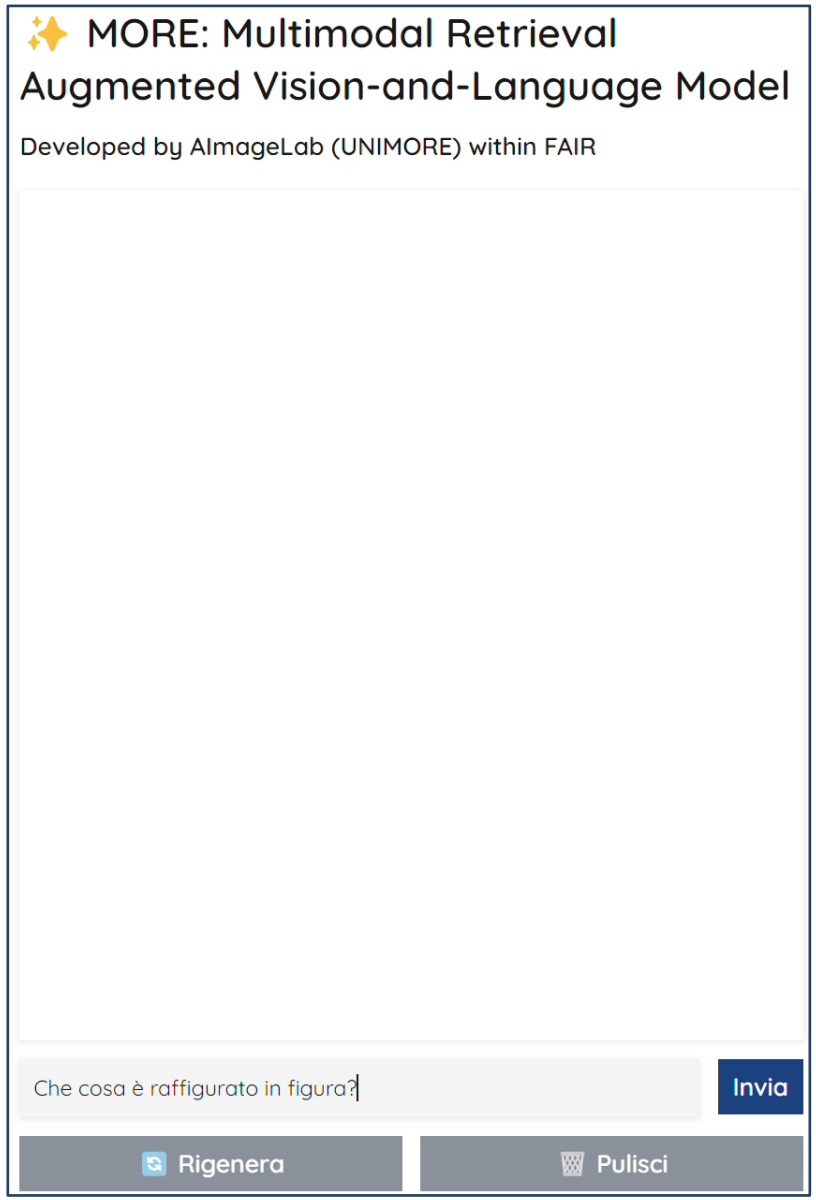
The first Italian Multimodal Language Model, endowed with retrieval capabilities

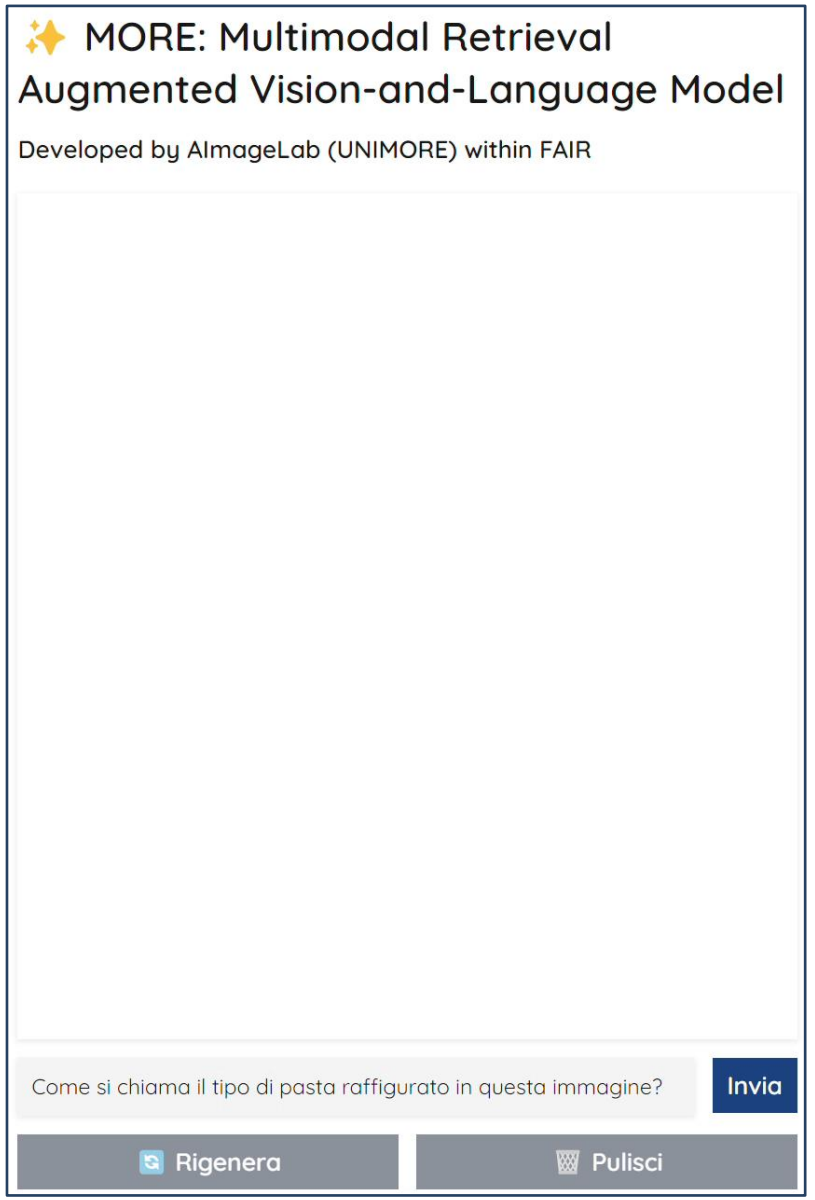
- Retrieves and exploits **multimodal knowledge** from an external source
- **Embeds input images** via an MLP-based adapter
- **Fine-tuned for Italian** on translated visual conversation datasets
- It is basis for a family of models supporting more tasks.
- *Publicly available soon!*

Proudly trained on the Leonardo Supercomputer thanks to an ISCRA-B grant.



MORE - Multimodal Retrieval Augmented Vision-and-Language Model

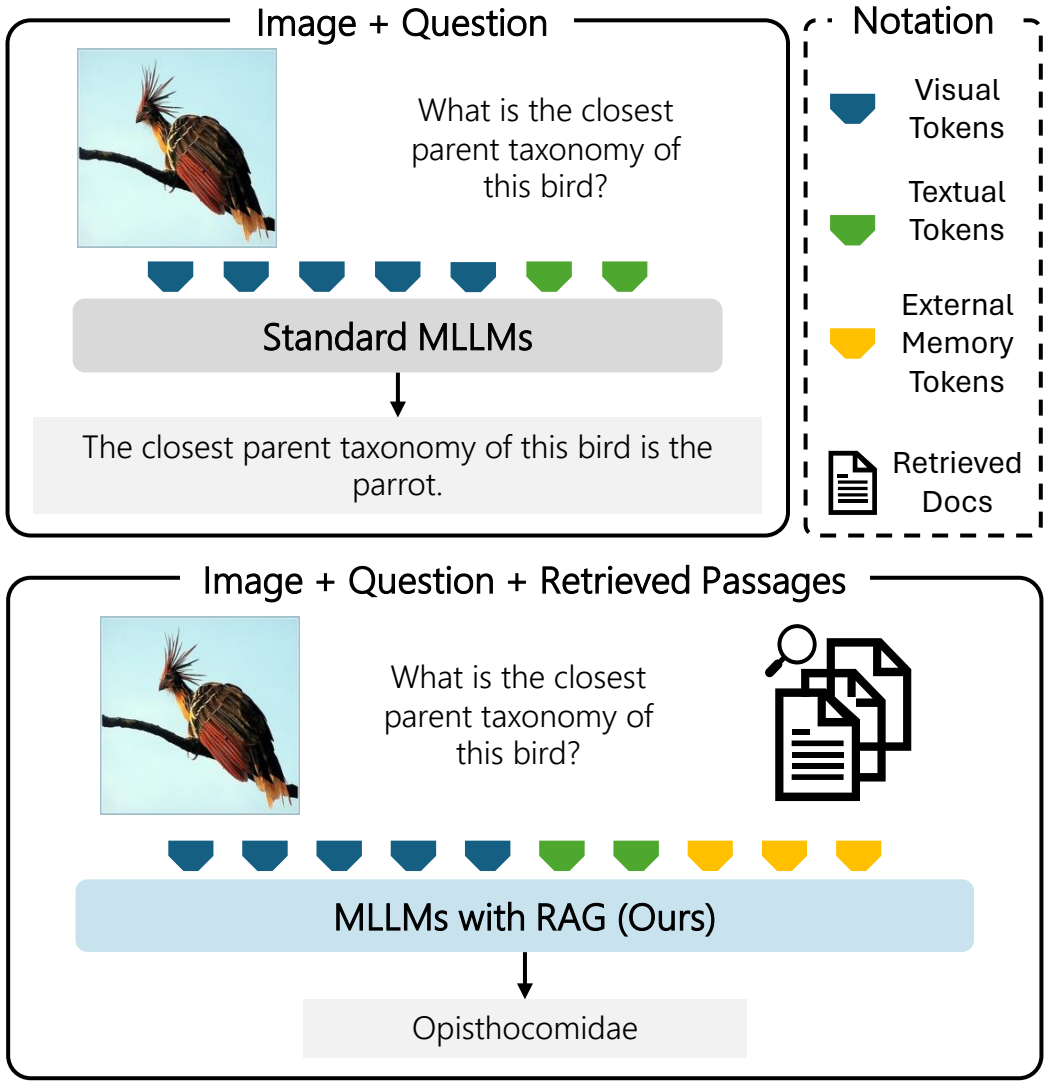




Augmenting Multimodal LLMs with World-Specific Knowledge

What are we doing now?

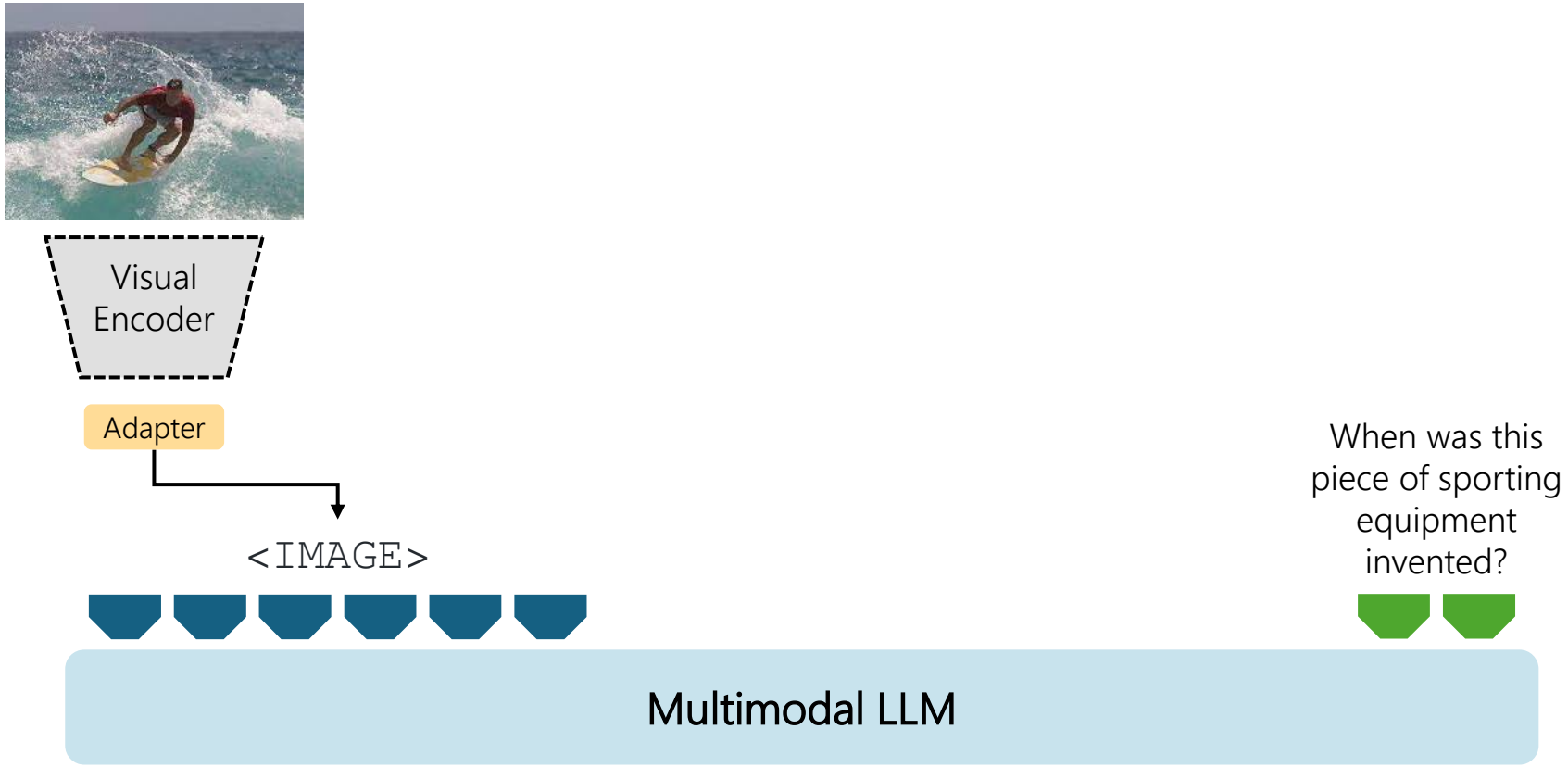
- Extending MORE to incorporate **world-specific knowledge** (*e.g.* extracted from Wikipedia).
- We design a new model that integrates knowledge retrieved from an external knowledge base of documents through a **hierarchical retrieval pipeline**.



D. Caffagni, F. Cocchi, N. Moratelli, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara "Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs"
CVPR Workshops, 2024 - What is Next in Multimodal Foundation Models?

Augmenting Multimodal LLMs with World-Specific Knowledge

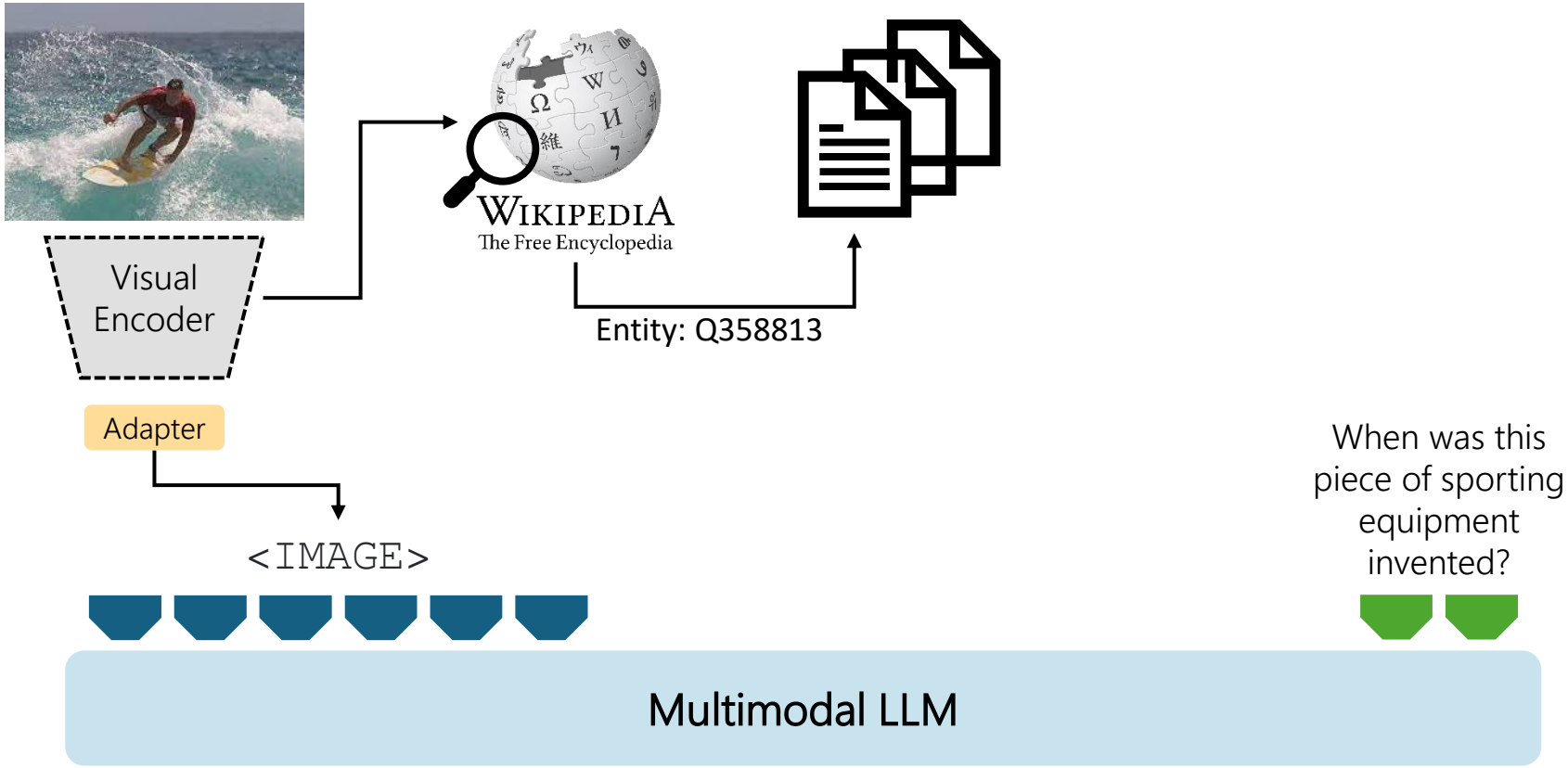
- The visual encoder is employed to **provide the MLLM with visual context** and as a query to retrieve from an external knowledge base.



D. Caffagni, F. Cocchi, N. Moratelli, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara "Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs"
CVPR Workshops, 2024 - What is Next in Multimodal Foundation Models?

Augmenting Multimodal LLMs with World-Specific Knowledge

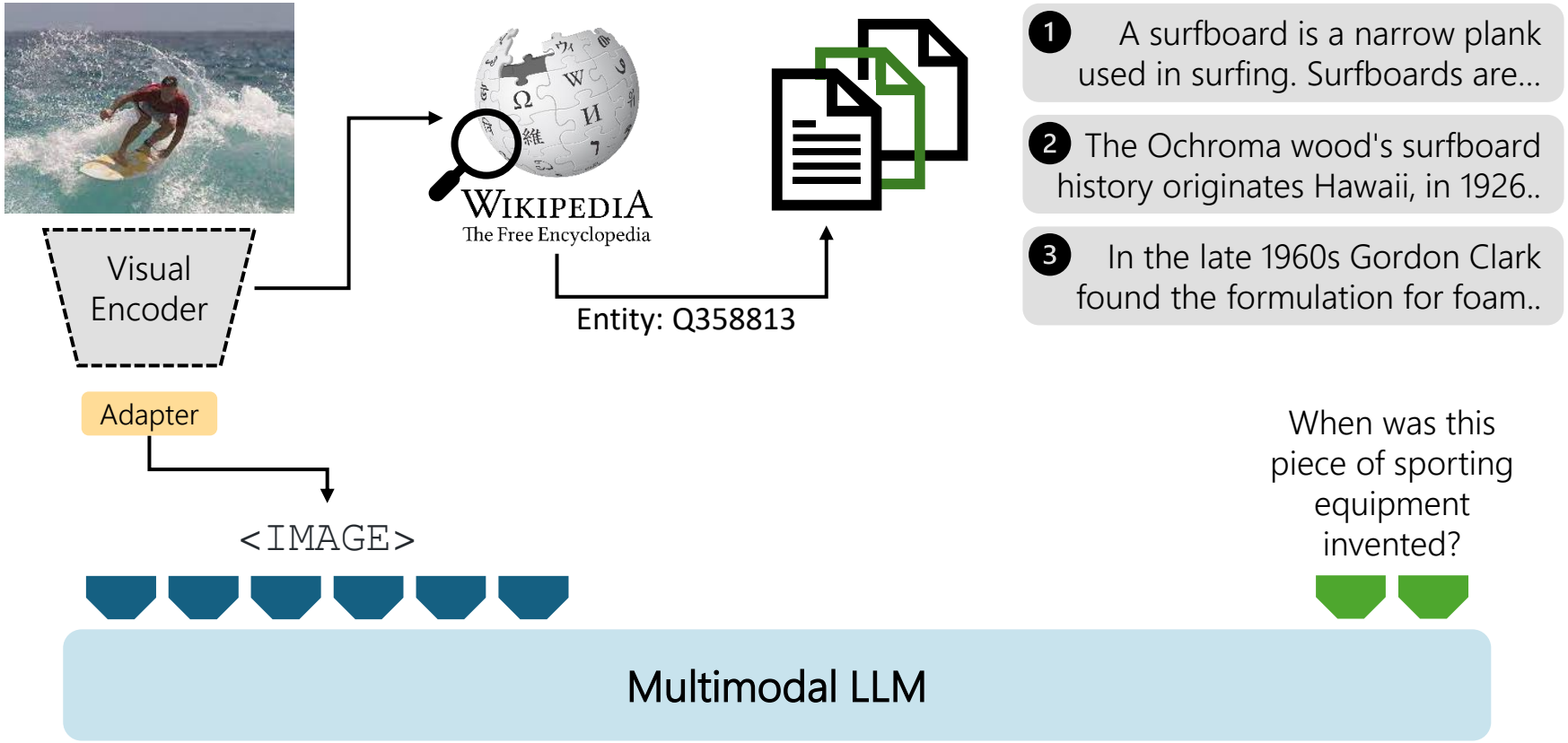
- The visual encoder is employed to provide the MLLM with visual context and **as a query to retrieve from an external knowledge base**.



D. Caffagni, F. Cocchi, N. Moratelli, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara "Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs"
CVPR Workshops, 2024 - What is Next in Multimodal Foundation Models?

Augmenting Multimodal LLMs with World-Specific Knowledge

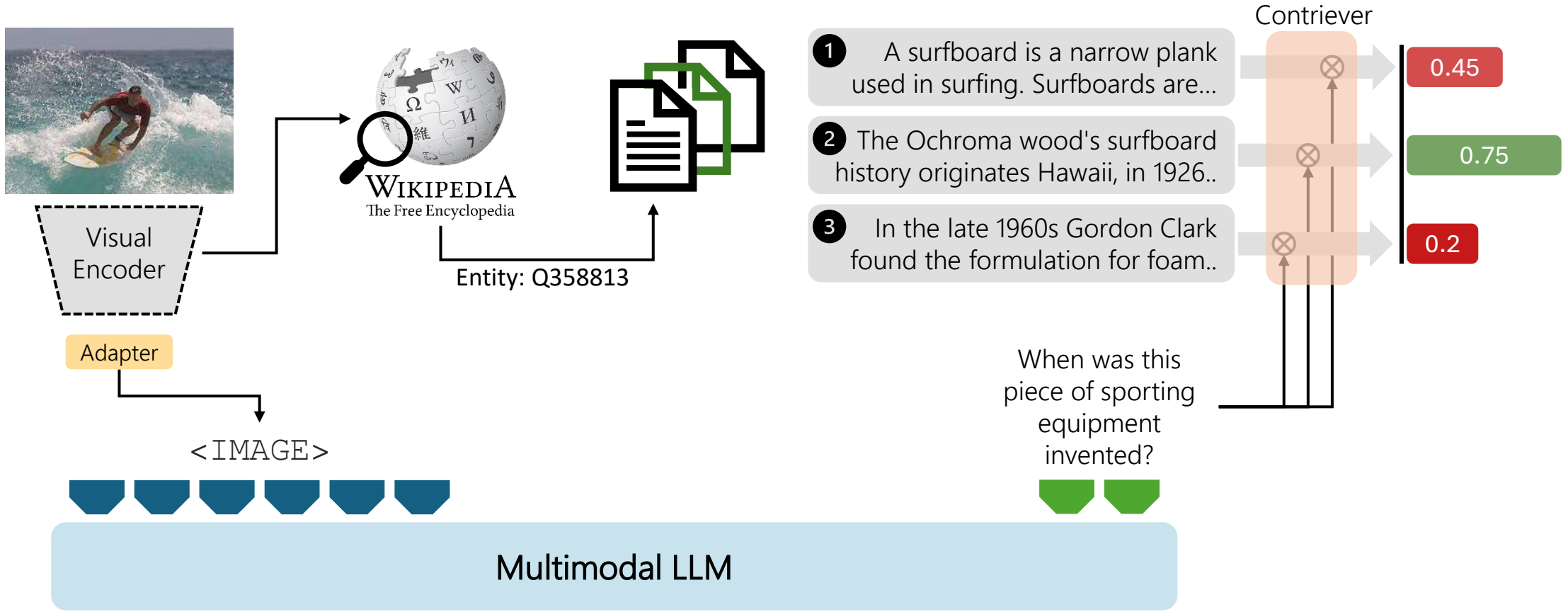
- A **hierarchical retrieval module** is designed to first find the relevant document, using a similarity score between the **CLIP-based embeddings** extracted from the input image and the Wikipedia page title.



D. Caffagni, F. Cocchi, N. Moratelli, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara "Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs"
CVPR Workshops, 2024 - What is Next in Multimodal Foundation Models?

Augmenting Multimodal LLMs with World-Specific Knowledge

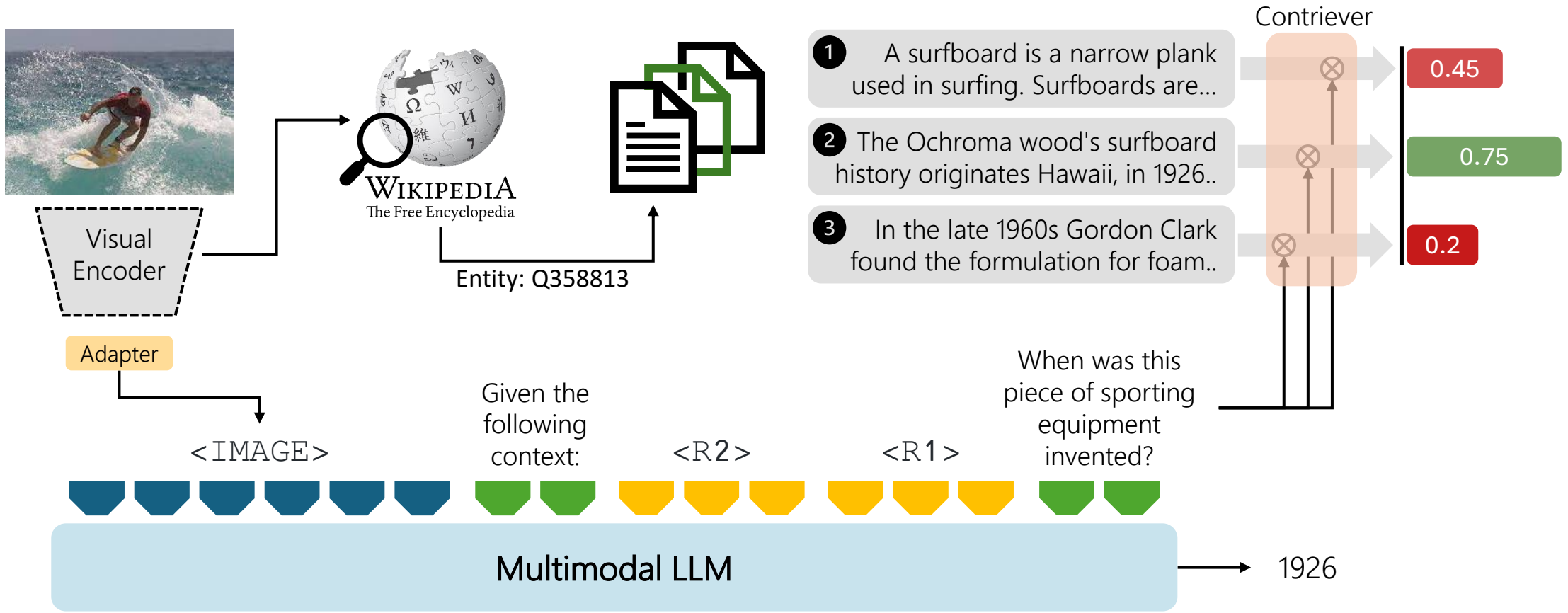
- Then, the **most relevant passages** are retrieved inside the document computing similarities between **Contriever-based textual embeddings** extracted from each passage and the given question.



D. Caffagni, F. Cocchi, N. Moratelli, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara "Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs"
CVPR Workshops, 2024 - What is Next in Multimodal Foundation Models?

Augmenting Multimodal LLMs with World-Specific Knowledge

- The retrieved passages are given as input to the multimodal LLM as **additional input context**, allowing the model to generate **more specific answers**.



D. Caffagni, F. Cocchi, N. Moratelli, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara "Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs"
CVPR Workshops, 2024 - What is Next in Multimodal Foundation Models?

Augmenting Multimodal LLMs with World-Specific Knowledge

What's next?

- **Effective retrieval strategies** are needed to improve the results of CLIP-based models, which perform poorly in retrieving the most relevant document related to a given image.
- Integration of **self-reflection techniques** inside the LLM to
  - decide when performing retrieval;
  - verify whether the retrieved knowledge is relevant or not to the given question.

When was this building constructed?

Standard MLLMs (e.g. LLaVA):
1970 ✗

MLLMs with RAG (Ours):
1927 ✓

Who designed this building?

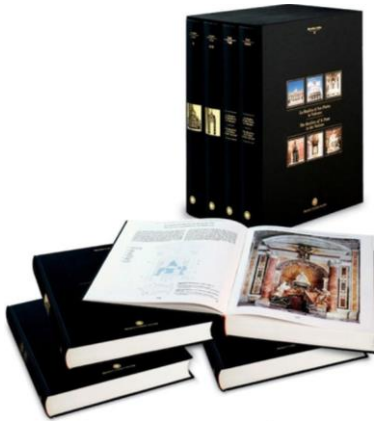
Standard MLLMs (e.g. LLaVA):
Architect ✗

MLLMs with RAG (Ours):
James of Saint George ✓

- Collection of a **new benchmark** for the task **focusing on the Italian cultural heritage**.

→ on-going collaboration with Franco Cosimo Panini to investigate the use of data from the ***Mirabilia Italiae*** **collection**.

FRANCO COSIMO PANINI



Want to Know More about Multimodal LLMs?

Check out our survey on the topic, recently accepted at ACL 2024 (Findings):

- Revises the prevalent choices for **vision encoders** and **adapter modules** that equip LLMs with cross-modal capabilities, and gives an overview of the **training process** and **data used**.
- Explores the **range of tasks** addressed by MLLMs, including visual grounding and image generation and editing.
- Offers a comparative perspective on the **performance** and **hardware requirements** of existing MLLMs.

The Revolution of Multimodal Large Language Models: A Survey

Davide Caffagni[1*], Federico Cocchi[1,2*], Luca Barsellotti[1*], Nicholas Moratelli[1*], Sara Sarto[1*], Lorenzo Baraldi[2*], Lorenzo Baraldi[1], Marcella Cornia[1], and Rita Cucchiara[1,3]
[1]University of Modena and Reggio Emilia, Italy
[2]University of Pisa, Italy
[3]IIT-CNR, Italy
[1]{name.surname}@unimore.it [2]{name.surname}@phd.unipi.it

Abstract

Connecting text and visual modalities plays an essential role in generative intelligence. For this reason, inspired by the success of large language models, significant research efforts are being devoted to the development of Multimodal Large Language Models (MLLMs). These models can seamlessly integrate visual and textual modalities, both as input and output, while providing a dialogue-based interface and instruction-following capabilities. In this paper, we provide a comprehensive review of recent visual-based MLLMs, analyzing their architectural choices, multimodal alignment strategies, and training techniques. We also conduct a detailed analysis of these models across a wide range of tasks, including visual grounding, image generation and editing, visual understanding, and domain-specific applications. Additionally, we compile and describe training datasets and evaluation benchmarks, conducting comparisons among existing models in terms of performance and computational requirements. Overall, this survey offers a comprehensive overview of the current state of the art, laying the groundwork for future MLLMs.

Figure 1: General architecture of Multimodal Large Language Models (MLLMs), composed of a visual encoder, a language model, and an adapter module that connects visual inputs to the textual space.

1 Introduction

The introduction of the attention operation and the Transformer architecture (Vaswani et al., 2017) has enabled the creation of models capable of handling various modalities on an increasingly large scale. This advancement is largely attributed to the versatility of the operator and the adaptability of the architecture. Initially, this breakthrough was leveraged for language-specific models (Devlin et al., 2018; Brown et al., 2020) but quickly extended to support diverse modalities (Li et al., 2019; Lu et al., 2019) and facilitate their integration within unified embedding spaces (Radford et al., 2021).

The surge in sophisticated Large Language Models (LLMs), particularly their capacity for in-context learning, has encouraged researchers to broaden the scope of these models to encompass multiple modalities, both as inputs and outputs. This expansion has led to the development of cutting-edge models such as GPT-4V (Achiam et al., 2023) and Gemini (Anil et al., 2023), showcasing state-of-the-art performance.

The development of Multimodal Large Language Models (MLLMs) entails merging single-modality architectures for vision and language, establishing effective connections between them through vision-to-language adapters, and devising innovative training approaches. These methodologies are crucial for ensuring modality alignment and the ability to follow instructions accurately.

In a context marked by the rapid release of new models, our goal is to offer an exhaustive overview of the MLLM landscape, with a focus on models exploiting the visual modality. This overview serves as both an update on the current state and a source of inspiration for future developments. We identify three core aspects that define these models:

D. Caffagni, F. Cocchi, L. Barsellotti, N. Moratelli, S. Sarto, L. Baraldi, L. Baraldi, M. Cornia, R. Cucchiara “The Revolution of Multimodal Large Language Models: A Survey”
ACL, 2024 (Findings)

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

Thank you!


Lorenzo Baraldi


Marcella Cornia


Rita Cucchiara


Lorenzo Baraldi


Luca Barsellotti


Davide Caffagni


Federico Cocchi


Nicholas Moratelli


Sara Sarto


Samuele Poppi


Tobia Poppi