Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Hierarchical Retrieval-Augmented Generation and Validation for Multimodal LLMs

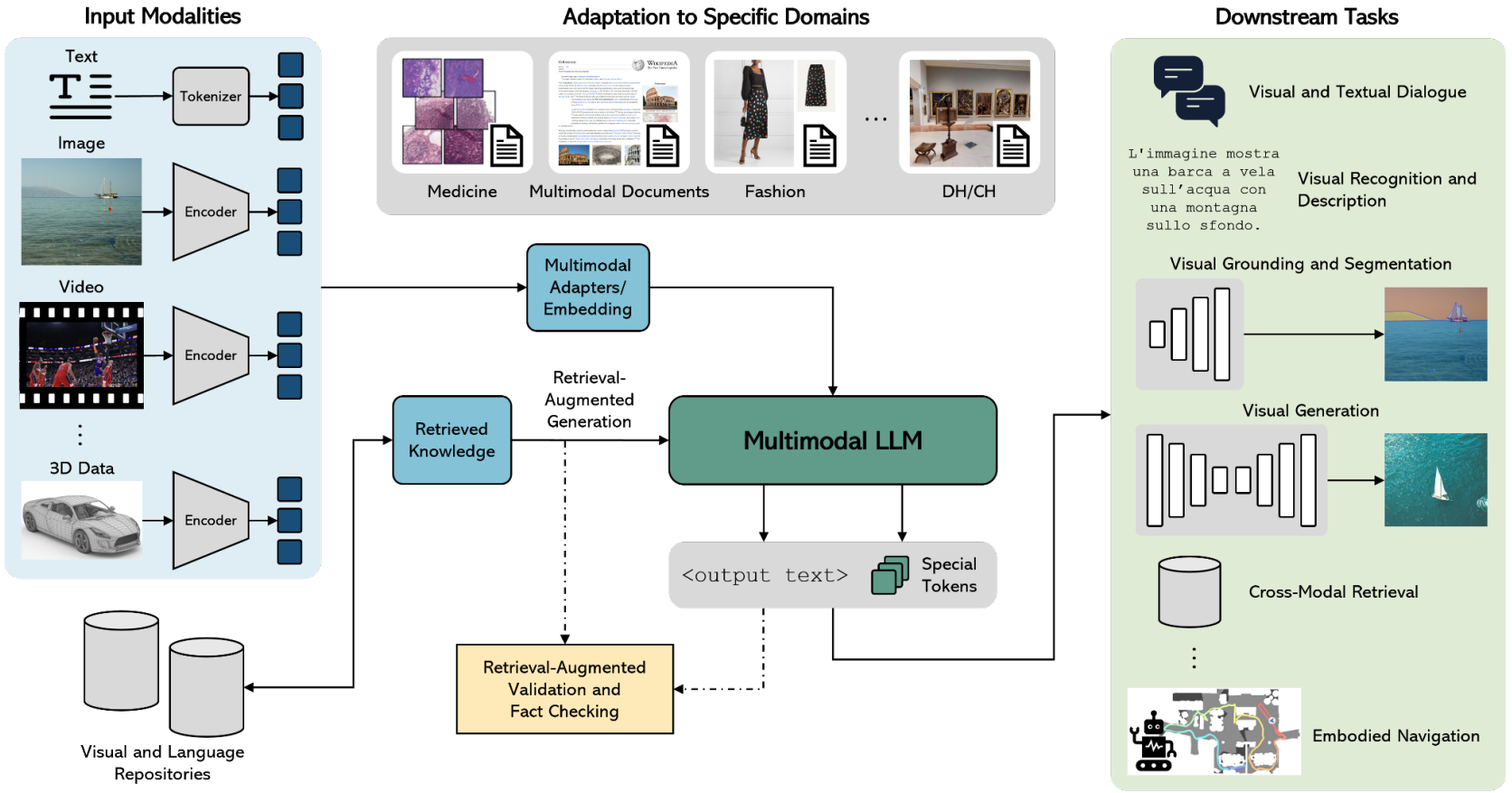
D. Caffagni*, F. Cocchi*, N. Moratelli*, S. Sarto*, M. Cornia*, L. Baraldi*, R. Cucchiara†,*

* Università degli Studi di Modena e Reggio Emilia
† Consiglio Nazionale delle Ricerche

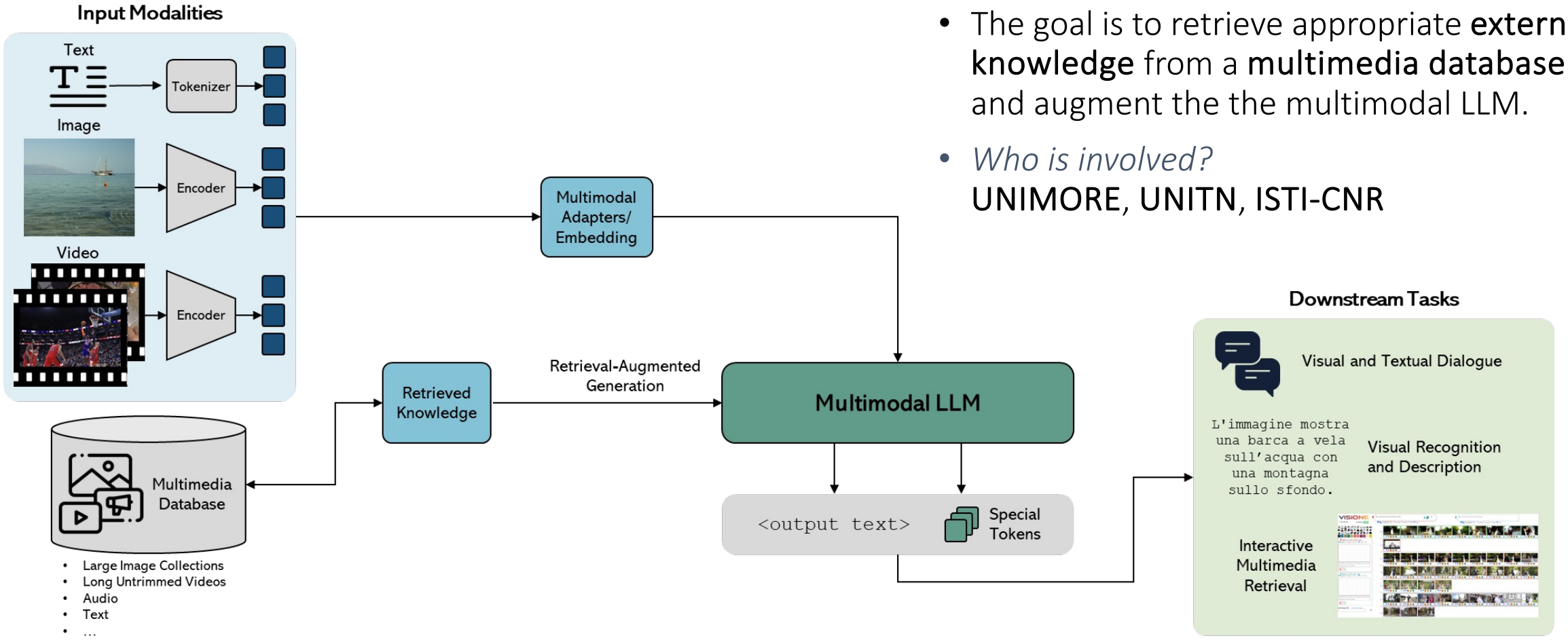
Naples, 23/09/2024

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Transversal Project on Vision, Language and Multimodal Challenges



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research



- The goal is to retrieve appropriate **external knowledge** from a **multimedia database** and augment the the multimodal LLM.
- *Who is involved?*
  UNIMORE, UNITN, ISTI-CNR

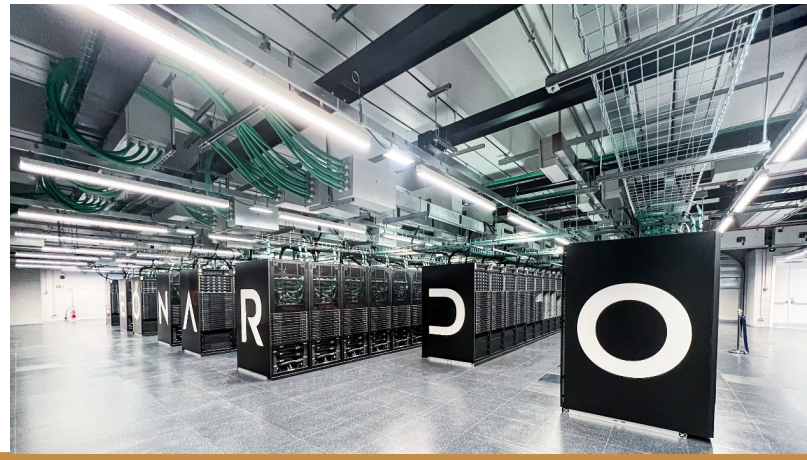
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

A Multimodal and Retrieval-Augmented Language Model

Main features

- 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

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



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

A Multimodal and Retrieval-Augmented Language Model

Cosa vedi nell'immagine?

L'immagine presenta un'intervista con un'immagine di un cantiere con una gru sullo sfondo. L'intervista sembra essere su un argomento di costruzione o ingegneria, con un'immagine di un cantiere con una gru sullo sfondo.

Quante corsie del telepass ci sono nell'immagine in input?

Ci sono due corsie del telepass nell'immagine

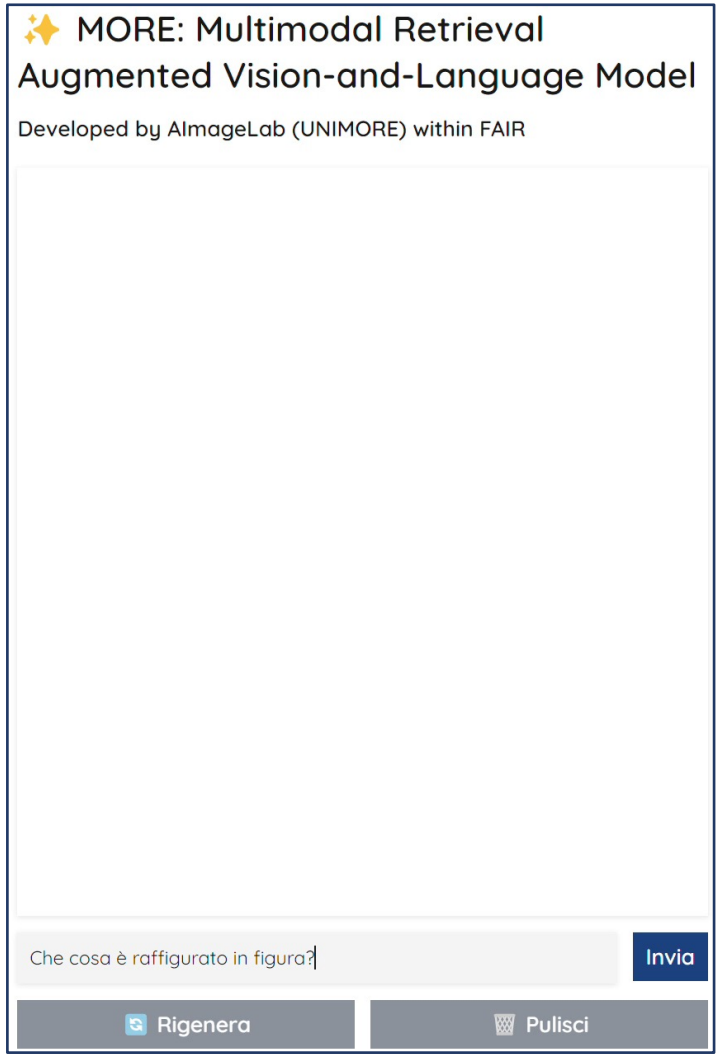
- Long-tail concept understanding
- Visual reasoning capabilities
- Reading/OCR capabilities
- Q&A Capabilities

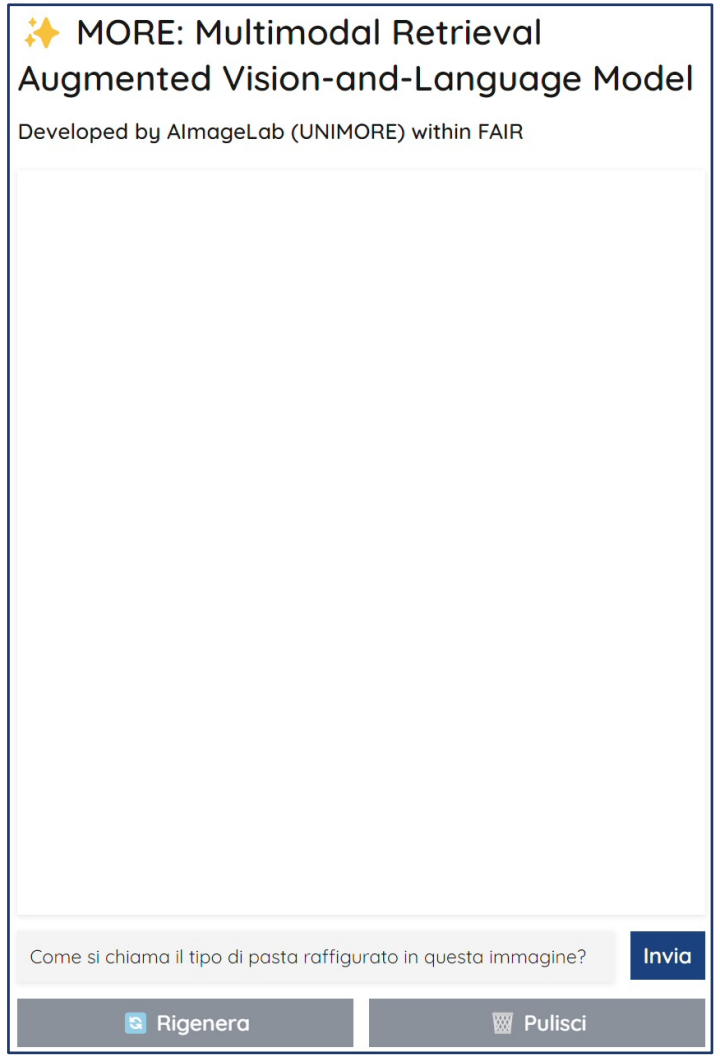
Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

FAIR Future Artificial Intelligence Research





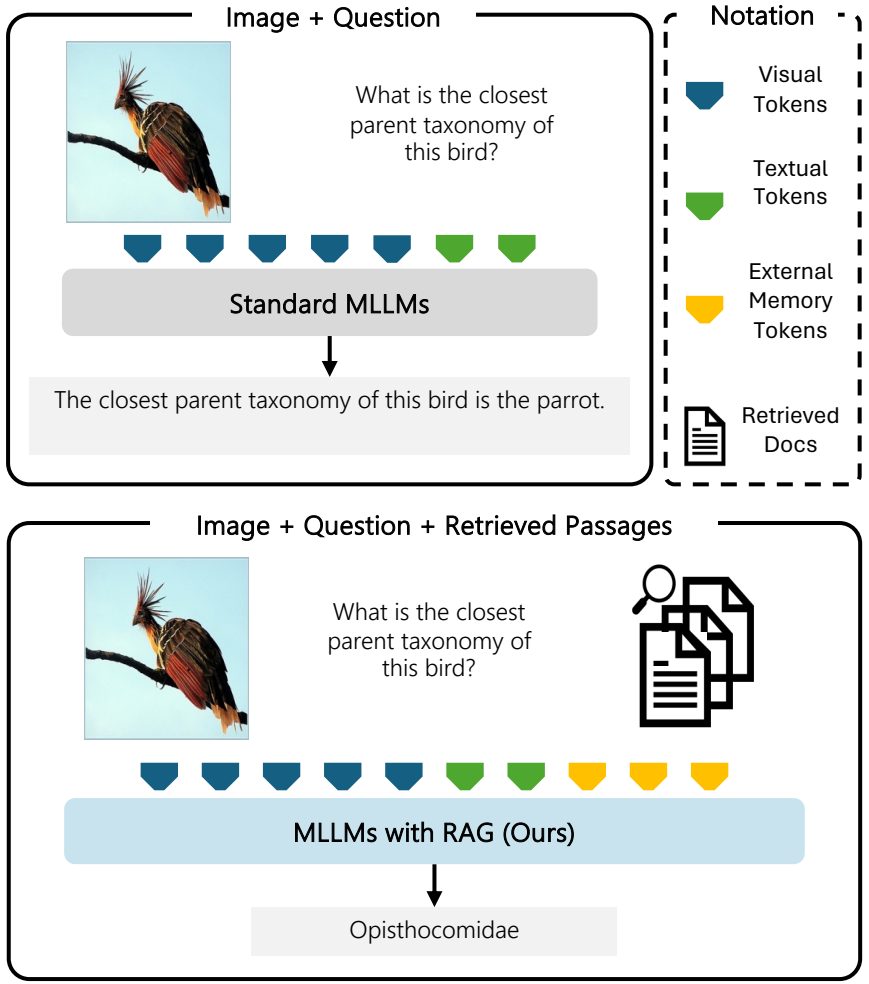
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Hierarchical Retrieval for RAG

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

Downstream task:

- Knowledge-based VQA
  - For now, existing English benchmarks (*i.e.* Encyclopedic VQA and InfoSeek) are considered.

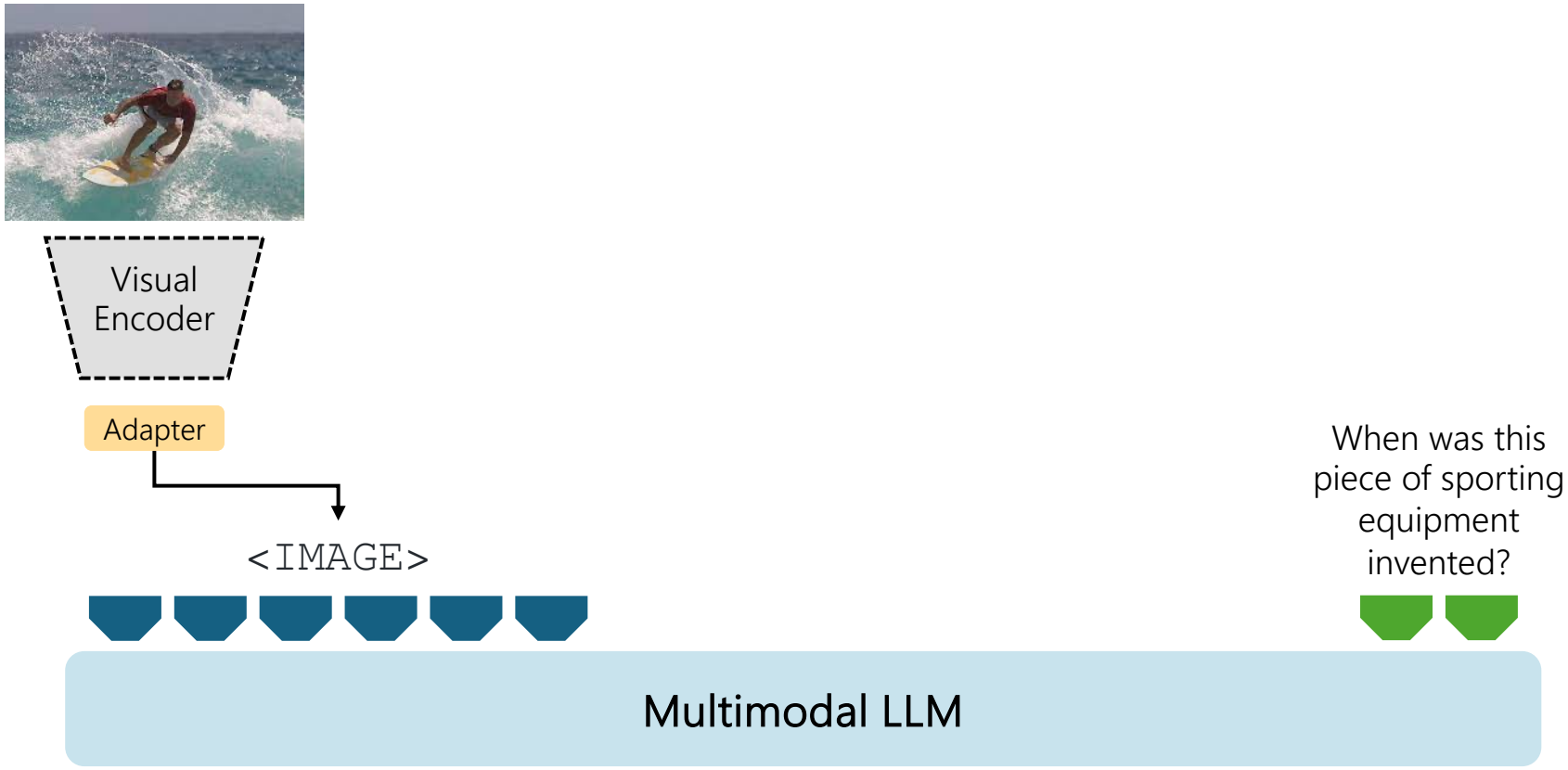


D. Caffagni, F. Cocchi, N. Moratelli, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara, "Wiki-LLaVA:Hierarchical Retrieval-Augmented Generation for Multimodal LLMs." CVPR Workshops 2024.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Hierarchical Retrieval for RAG

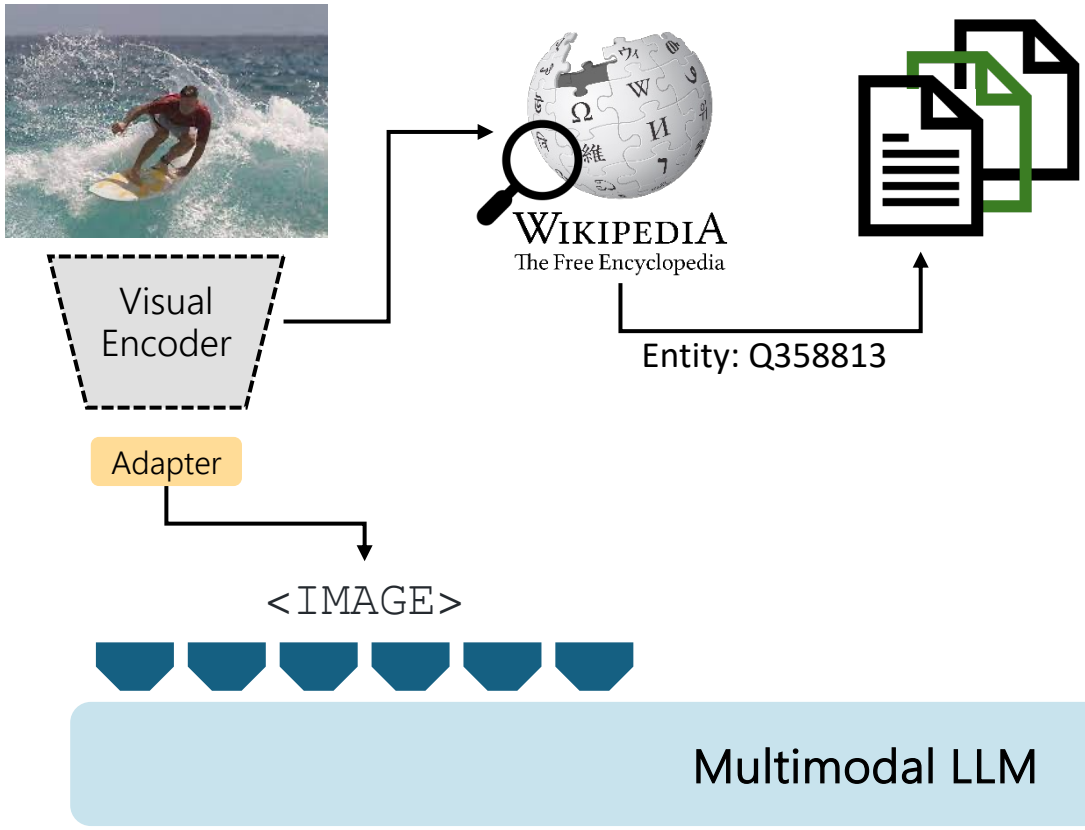
The visual encoder is employed to **provide the MLLM with visual context** and as a query to retrieve from an external knowledge base.



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Hierarchical Retrieval for RAG

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



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Hierarchical Retrieval for RAG

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.

Visual Encoder

Adapter

<IMAGE>

WIKIPEDIA
The Free Encyclopedia

Entity: Q358813

1. A surfboard is a narrow plank used in surfing. Surfboards are...
2. The Ochroma wood's surfboard history originates Hawaii, in 1926..
3. In the late 1960s Gordon Clark found the formulation for foam..

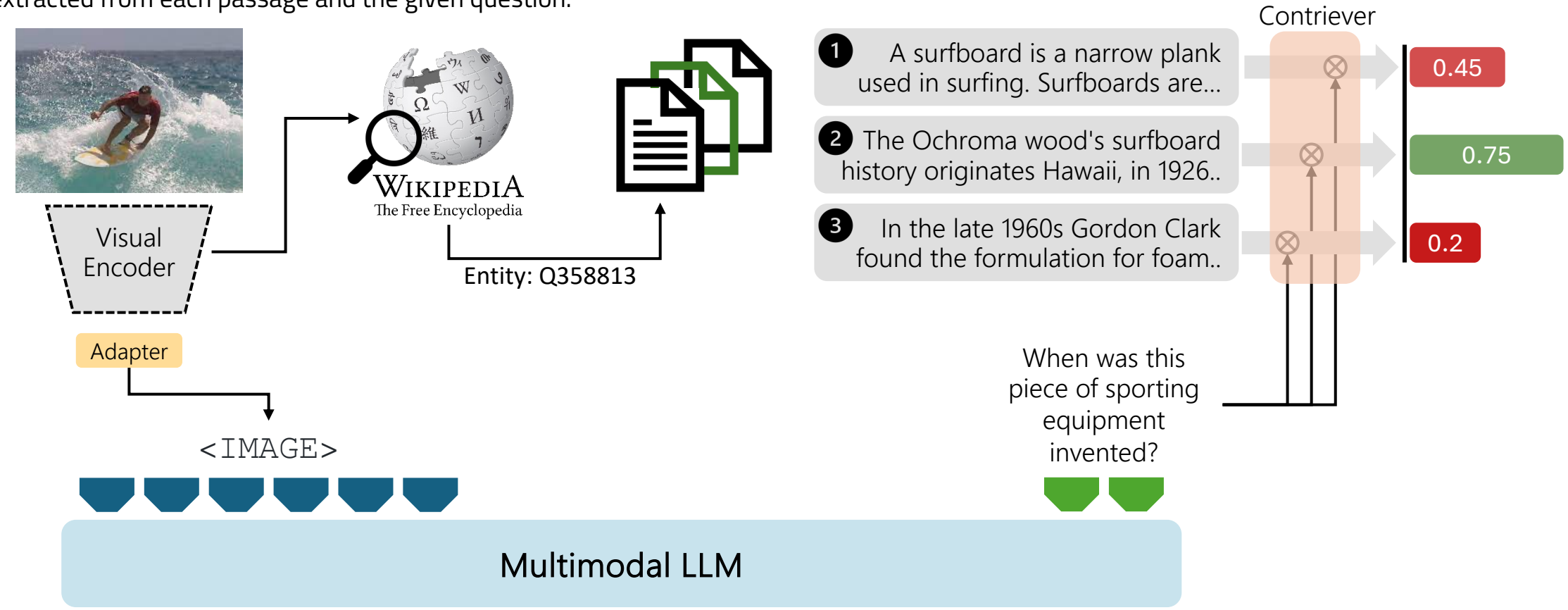
When was this piece of sporting equipment invented?

Multimodal LLM

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Hierarchical Retrieval for RAG

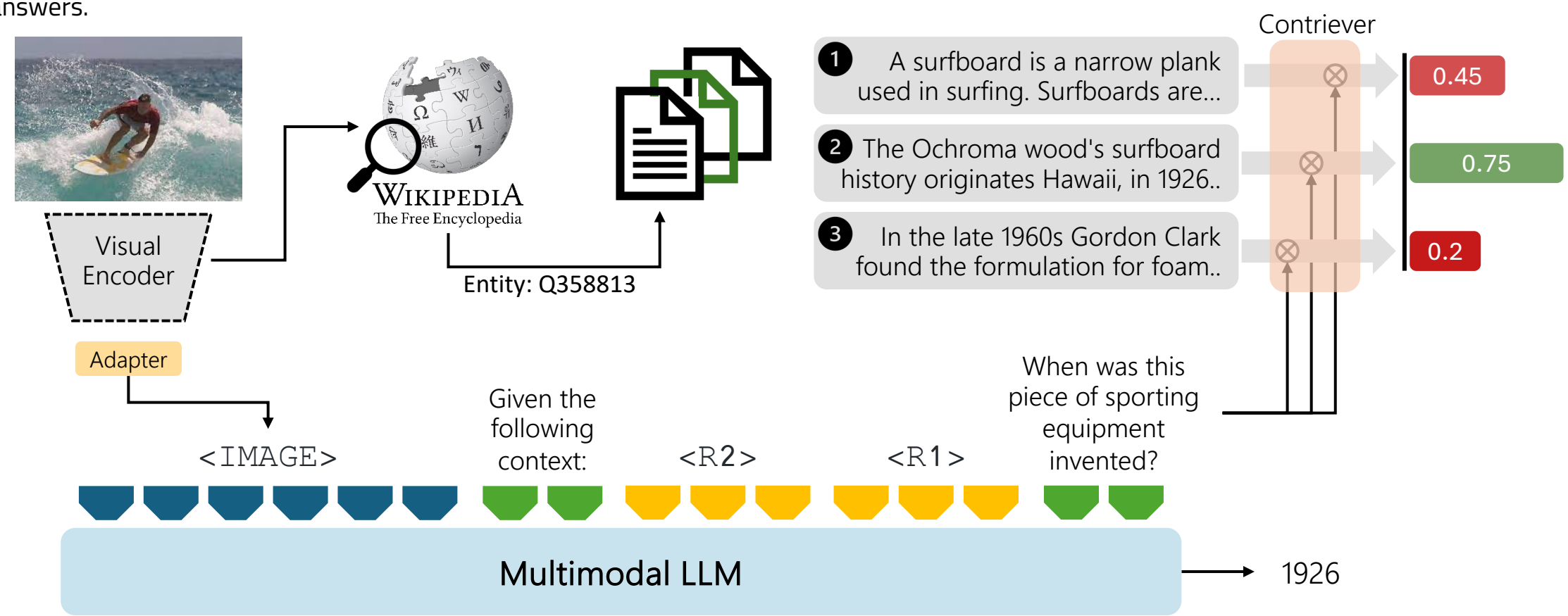
Then, the **most relevant passages** are retrieved inside the document computing similarities between **Contriever-based textual embeddings** extracted from each passage and the given question.



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Hierarchical Retrieval for RAG

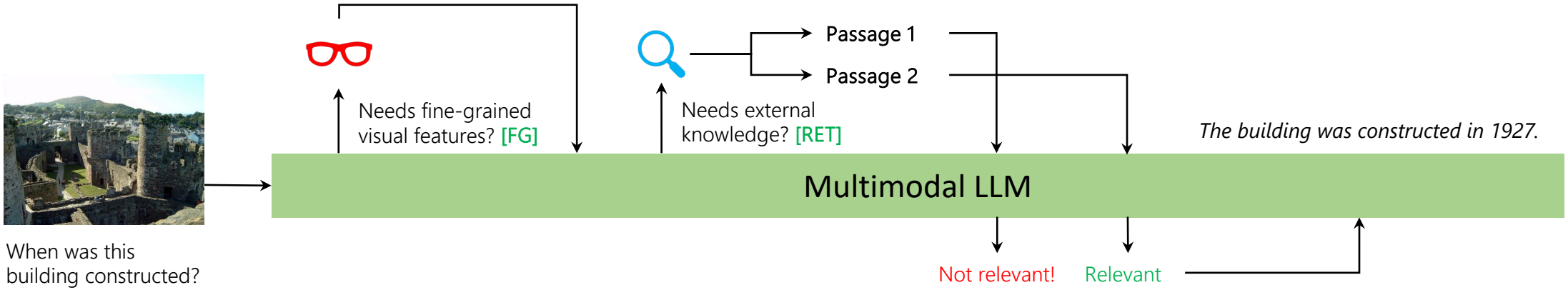
The retrieved passages are given as input to the multimodal LLM as additional input context, allowing the model to generate more specific answers.



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Multimodal Self-reflection for RAG

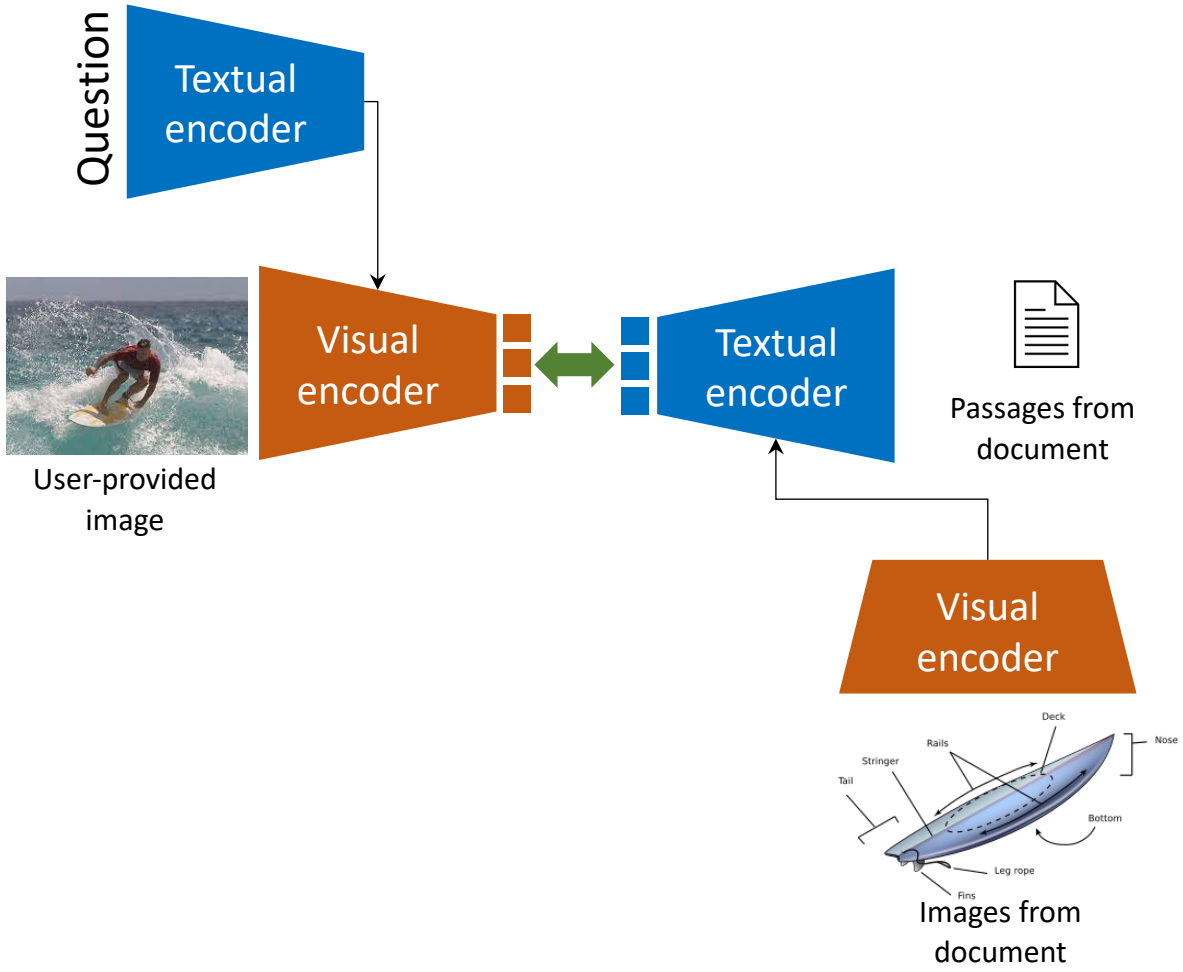
- Effective strategies are needed to manage retrieved items and to improve CLIP-based models, which perform poorly in retrieving the most relevant document related to a given image.
- **Current focus:** Integration of **self-reflection and re-ranking techniques** inside the M-LLM to:
  - Decide if global visual features are sufficient or if fine-grained visual features are needed ("[FG]" token).
  - Decide when retrieval is needed, through the emission of a "[RET]" dedicated token.
  - Verify whether the retrieved knowledge is relevant or not to the given question (→ re-ranking of retrieved items).



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Better Embedding spaces for RAG

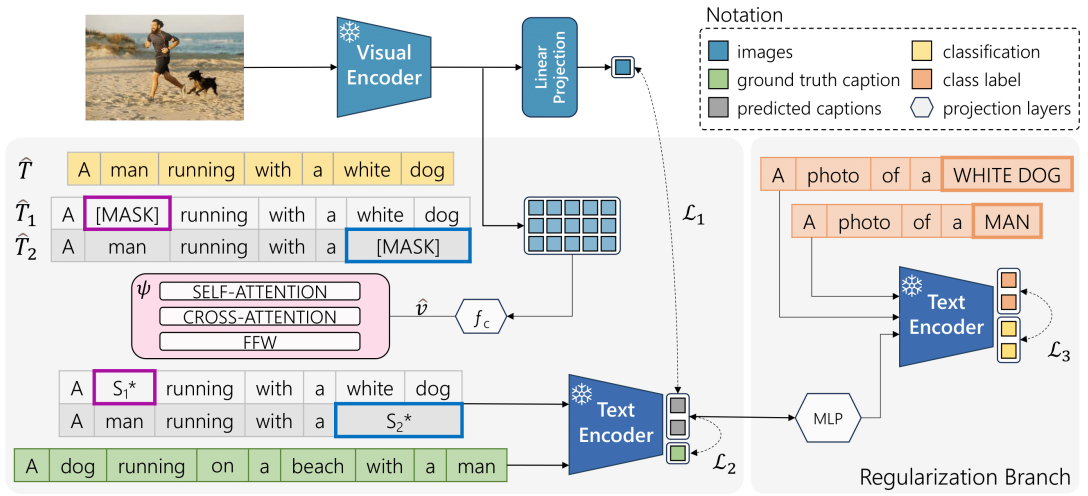
- Most embedding spaces for multimodal RAG (i.e. CLIP) consider single-modality queries and values (e.g. images or text), limiting their encoding capabilities.
- **Current focus:** Design of embedding spaces for **RAG which support multimodal queries and documents** (e.g. image + question):
  - Textual features from the question guide the extraction of fine-grained features from the input image
  - Images are fused into the text of external documents, creating multi-modal retrievable items
  - Fusion between different modalities is done layer-wise and with learnable gates.

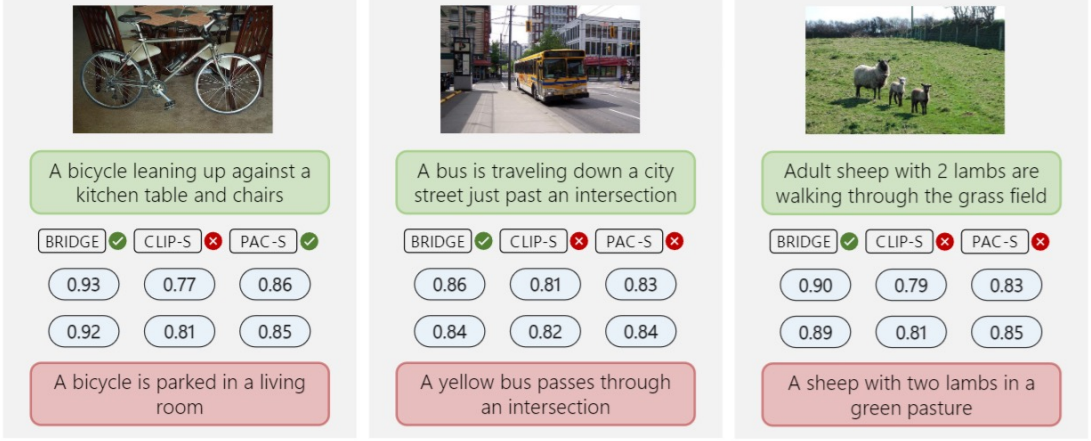


Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Evaluation of Multimodal LLMs

- Aligning machine-generated outputs with human judgment is of crucial importance.
- Previous work: training global image-text embedding spaces dedicated to evaluation.
- Current work: **BRIDGE**, a new learnable and reference-free image captioning metric that:
  - considers fine-grained visual features and maps them into dense multimodal vectors
  - employs multi-modal pseudo-captions built during the evaluation process to measure image-text correspondences at the detail level.
- SoTA on five human evaluation datasets.



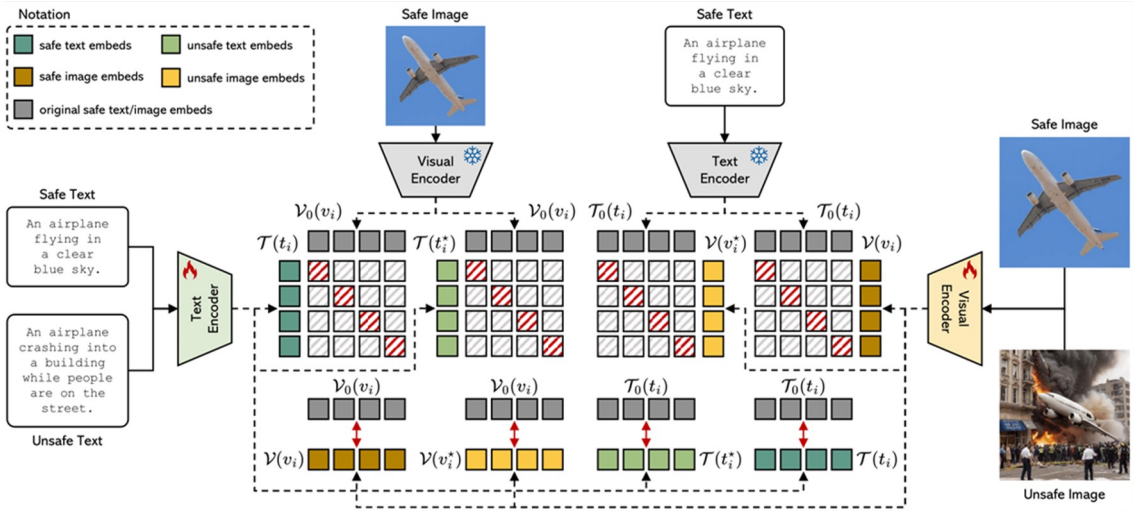


S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara, "BRIDGE: Bridging Gaps in Image Captioning Evaluation with Stronger Visual Cues." ECCV 2024.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Trustworthiness and Safety

- Models trained on large-scale data can generate inappropriate content and lead to the development of unsafe behavior.
- We aim to **make Vision-and-Language models safer** by removing their sensitivity to NSFW concepts.
- By fine-tuning Llama2 to convert between safe and unsafe sentences, we created our ViSU dataset made of quadruplets of real safe text/image pairs and generated unsafe pairs.
- We then employ the ViSU dataset to **fine-tune the CLIP space by redirecting unsafe content while preserving the CLIP structure**. Applying our method to downstream tasks like retrieval and generation improves the safety of retrieved and generated content by enhancing vulnerable user's protection.
- Current focus: hyperbolic spaces for safety preservation.

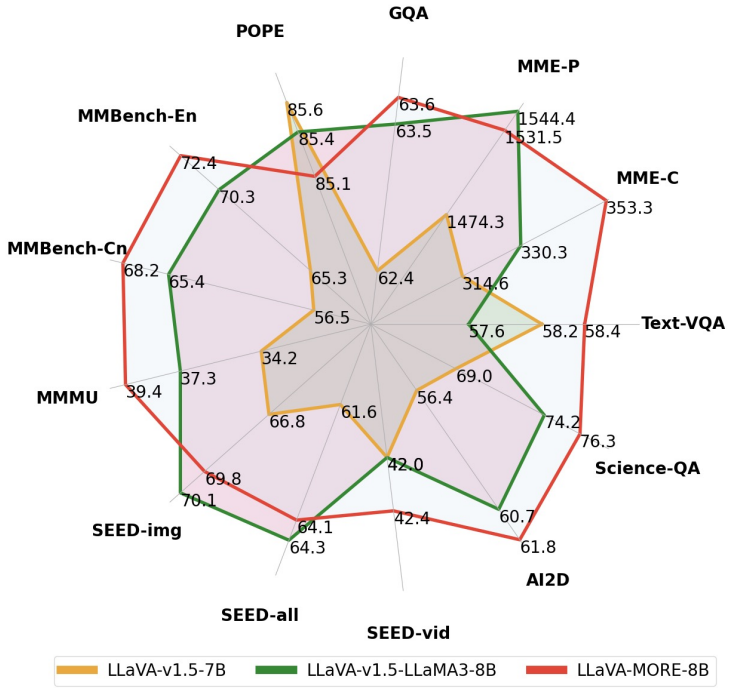


S. Poppi, T. Poppi, F. Cocchi, M. Cornia, L. Baraldi, R. Cucchiara, "Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models." ECCV 2024.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

LLaVA-MORE: Enhancing Visual Instruction Tuning with LLaMA 3.1

- The first LLaVa-based family of models that integrates LLaMa 3.1 as core language model.
- We train and release both pre-training (V&L feature alignment) and instruction fine-tuning stages.
- With a variety of visual backbones and multi-resolution encoding methodologies (OpenAI CLIP, SigLIP, S2).
- Available on Github and Huggingface: https://github.com/aimagelab/LLaVA-MORE

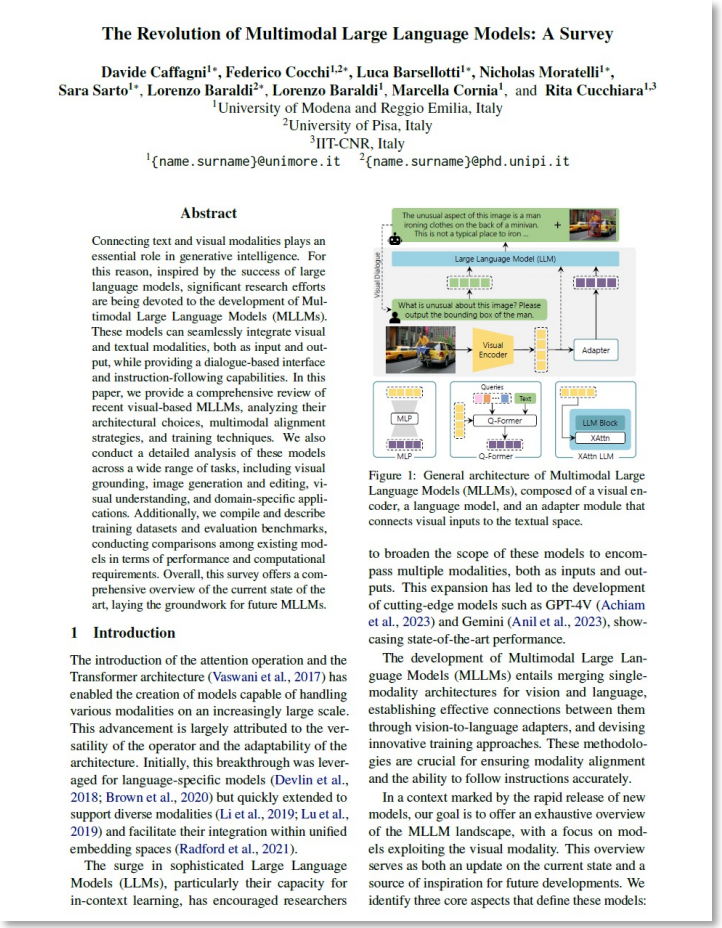


Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Finally…

Check out our survey on Multimodal LLMs, that:

- 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.

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

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.

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)