SOBIGDATA.it

ITALIAN RESEARCH INFRASTRUCTURE

Building and Evaluating Multimodal Generative Models: Architectures, Applications, and Challenges

Pisa, 27/03/2025

Prof. Marcella Cornia, University of Modena and Reggio Emilia
marcella.cornia@unimore.it

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
SOBIGDATA.it ITALIAN RESEARCH INFRASTRUCTURE
Consiglio Nazionale delle Ricerche

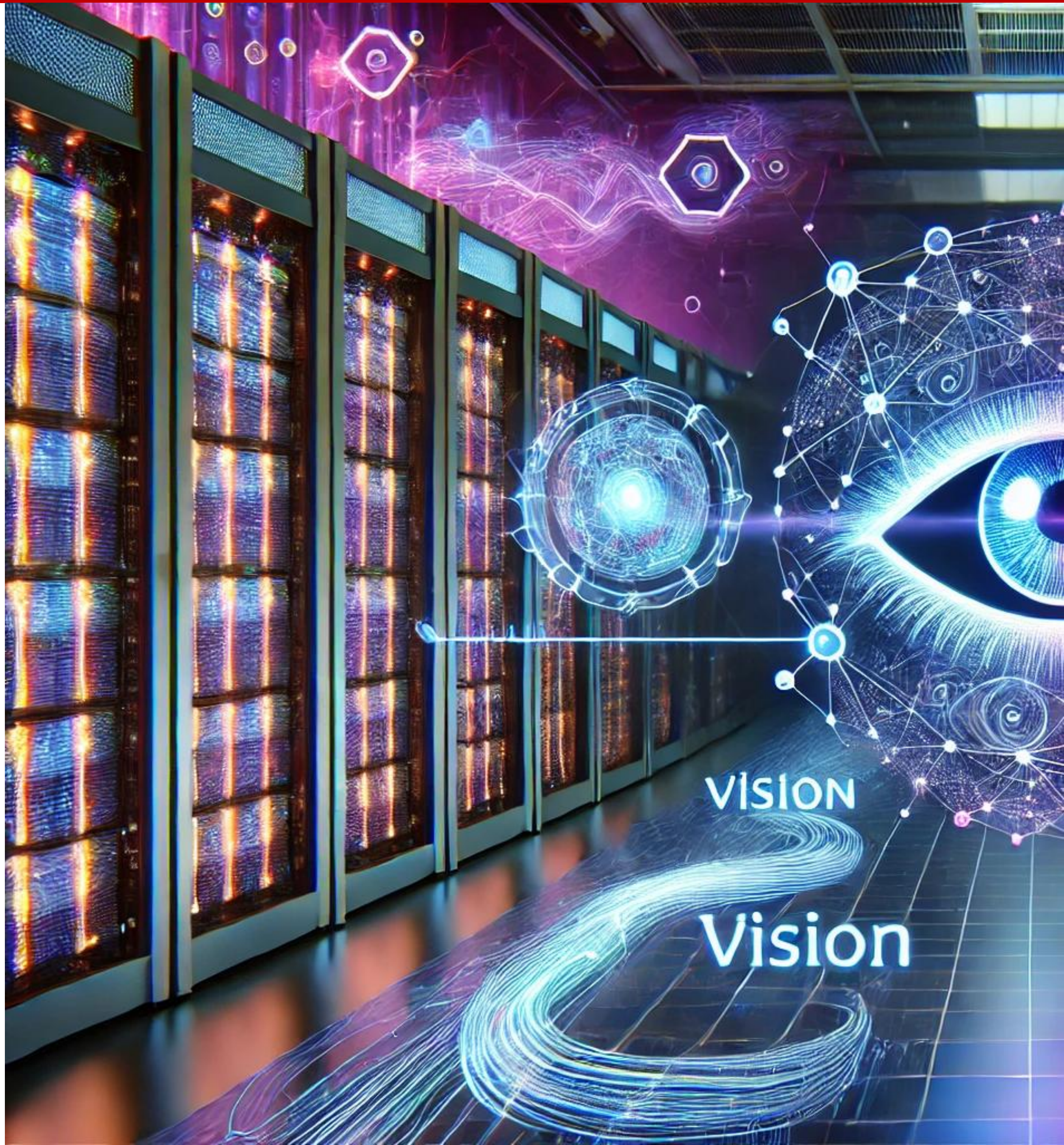
Outline

What we will see:

<u>Part 1:</u>
Training Large-Scale Multimodal Generative Models

<u>Part 2:</u>
Making Generative Models Trustworthy and Safe

<u>Part 3:</u>
Extending Generative Models to the Fashion Domain



Training Large-Scale Multimodal Generative Models

The beginning of a journey

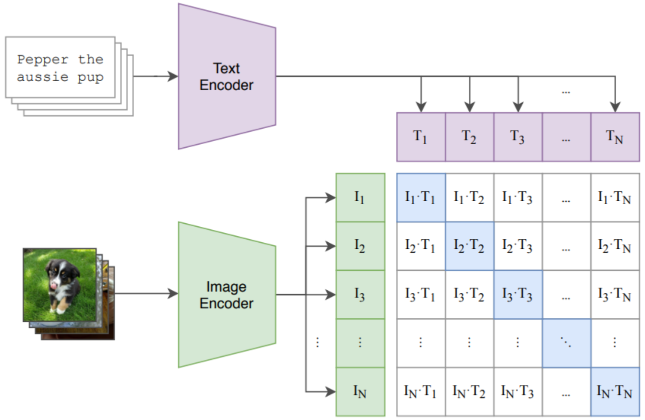
SOBIGDATA.it
ITALIAN RESEARCH INFRASTRUCTURE

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
SOBIGDATA.it ITALIAN RESEARCH INFRASTRUCTURE
Consiglio Nazionale delle Ricerche

Connecting Vision and Language

- **Large-scale networks with support for long-tail semantic concepts (IJCV 2023)**
  scaling to large-scale datasets and handling the duality between noisy web-scale data and human-annotated data

- **New metrics (CVPR 2023, ECCV 2024)**
  for image description evaluation and for training more effective image captioning models



A man in a black jacket skiing down a hill.

A man on skis down a snow covered slope.

Standard Captioner:
A group of people riding skateboards in a field.
Universal Captioner:
A group of people riding ***segways*** *in a field.*

Standard Captioner:
A tall building sitting in the middle of a body of water.
Universal Captioner:
An aerial view of the ***Burj Al Arab*** *in* ***Dubai****.*

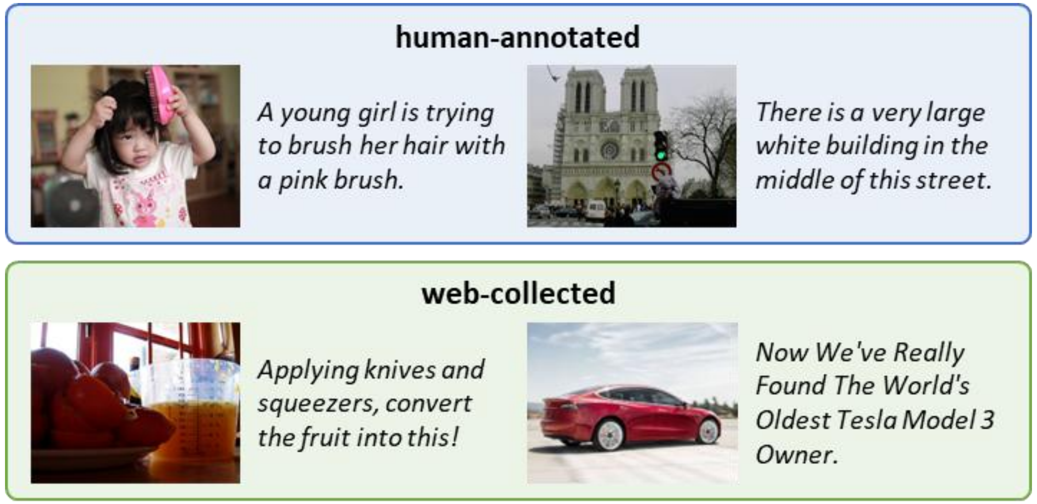
Standard Captioner:
A woman with blonde hair is posing for a picture.
Universal Captioner:
A picture of ***Marilyn Monroe*** *with a red lipstick.*

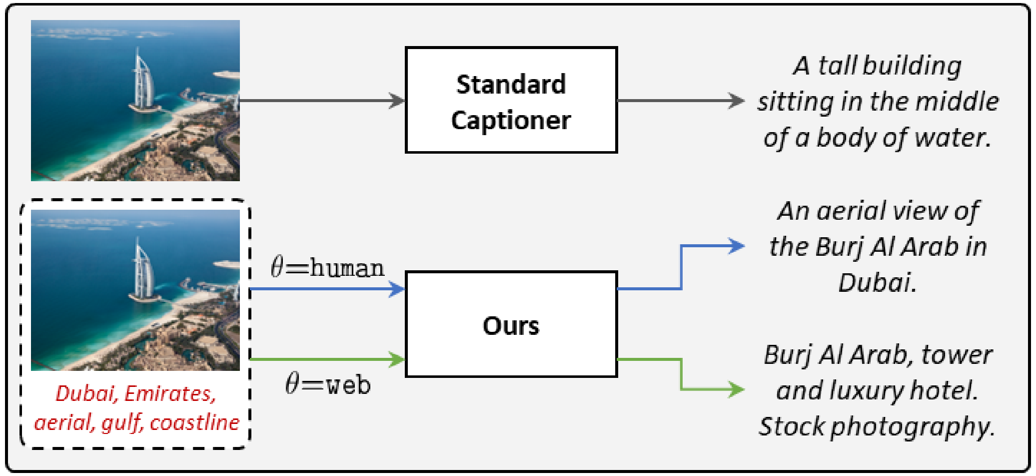
Most of large-scale AI is built upon noisy web-collected data.

Compared to **human-annotated** captions, **web-collected** ones have a greater richness in semantics and concepts, but a lower description quality.

Key idea:

- Develop an architecture which can emulate the descriptive style of traditional human-annotated datasets and web-collected ones, while transferring semantic content between sources.





M. Cornia, L. Baraldi, G. Fiameni, and R. Cucchiara. "Generating More Pertinent Captions by Leveraging Semantics and Style on Multi-Source Datasets", **IJCV 2023**

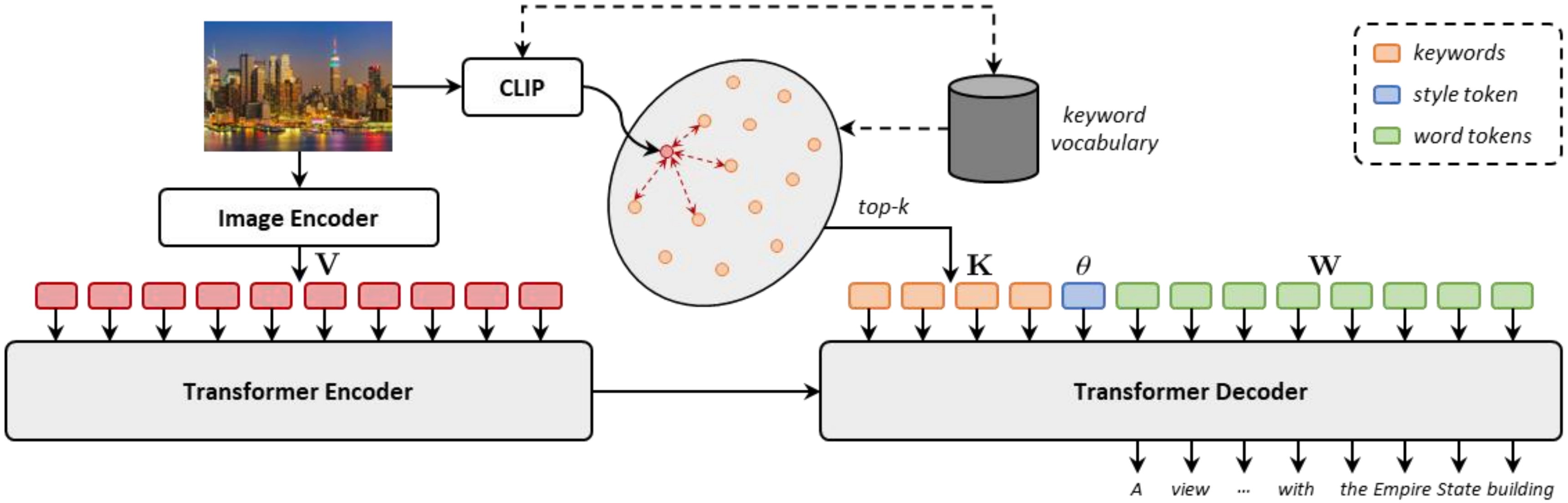
Universal Captioner

Inputs

- **CNN feature extractors** which can directly take raw pixels as input and avoid using object detectors;
- **Textual keywords** extracted with large-scale cross-modal models;
- **Style token** to separate hand-collected and web-based image-caption pairs.

Architecture

- **Fully-attentive encoder-decoder** that jointly encodes keywords, style, and text.



Trained on 36.4 million image-text pairs!

Cornia, M., Baraldi, L., Fiameni, G., and Cucchiara, R. “Generating More Pertinent Captions by Leveraging Semantics and Style on Multi-Source Datasets”, **IJCV 2023**

Main results:

- **State-of-the-art results on COCO, nocaps and Conceptual Captions 3M**
- Zero-shot generalization to other datasets
- Capability to name out-of-domain concepts (*e.g.* proper nouns of places, famous people, brands)

	Fine-tuning							
	TF	SCST	Training Images	B-4	M	R	C	S
BLIP^{base}	✓	-	129M	39.7	-	-	133.3	-
BLIP^{large}	✓	-	129M	40.4	-	-	136.7	-
SimVLM^{base}	✓	-	1.8B	39.0	32.9	-	134.8	24.0
SimVLM^{large}	✓	-	1.8B	40.3	33.4	-	142.6	24.7
SimVLM^{huge}	✓	-	1.8B	40.6	**33.7**	-	143.3	25.4
LEMON^{base}	✓	✓	200M	41.6	31.0	-	142.7	25.1
LEMON^{large}	✓	✓	200M	42.3	31.2	-	144.3	25.3
LEMON^{huge}	✓	✓	200M	42.6	31.4	-	145.5	**25.5**
Ours^{tiny} (θ=human)	-	✓	35.7M	42.8	31.0	61.2	148.4	24.6
Ours^{small} (θ=human)	-	✓	35.7M	42.5	31.2	61.3	148.6	25.0
Ours^{base} (θ=human)	-	✓	35.7M	**42.9**	31.4	**61.5**	**149.6**	25.0

	Fine-tuning							
	TF	SCST	Training Images	B-4	M	R	C	S
OSCAR^{base}	✓	✓	4.1M	40.5	29.7	-	137.6	22.8
OSCAR^{large}	✓	✓	4.1M	41.7	30.6	-	140.0	24.5
VinVL^{base}	✓	✓	5.8M	40.9	30.9	-	140.6	25.1
VinVL^{large}	✓	✓	5.8M	41.0	31.1	-	140.9	25.2
Ours^{tiny} (θ=human)	-	✓	5.8M (VinVL data)	42.9	31.1	61.3	147.1	24.9
Ours^{small} (θ=human)	-	✓	5.8M (VinVL data)	42.7	31.3	61.3	147.5	25.2
Ours^{base} (θ=human)	-	✓	5.8M (VinVL data)	**43.2**	**31.4**	**61.7**	**147.8**	**25.4**

Cornia, M., Baraldi, L., Fiameni, G., and Cucchiara, R. "Generating More Pertinent Captions by Leveraging Semantics and Style on Multi-Source Datasets", **IJCV 2023**

Universal Captioner

Main results:

- **State-of-the-art results on COCO, nocaps and Conceptual Captions 3M**
- Zero-shot generalization to other datasets
- Capability to name out-of-domain concepts (*e.g.* proper nouns of places, famous people, brands)

				Validation Set								**Test Set**							
	Fine-tuning			**in**		**near**		**out**		**overall**		**in**		**near**		**out**		**overall**	
	TF	SCST	**Training Images**	C	S	C	S	C	S	C	S	C	S	C	S	C	S	C	S
BLIP^{base}	✓	-	129M	111.8	14.9	108.6	14.8	111.5	14.2	109.6	14.7	-	-	-	-	-	-	-	-
BLIP^{large}	✓	-	129M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	-	-	-	-	-	-	-	-
SimVLM^{huge}	✓	-	1.8B	113.7	-	110.9	-	115.2	-	112.2	-	109.0	14.6	110.8	14.6	109.5	13.9	110.3	14.5
LEMON^{large}	✓	-	200M	116.9	**15.8**	113.3	15.1	111.3	14.0	113.4	15.0	111.2	**15.6**	112.3	15.2	105.0	13.6	110.9	15.0
LEMON^{huge}	✓	-	200M	118.0	15.4	116.3	15.1	120.2	**14.5**	117.3	15.0	112.8	15.2	115.5	15.1	110.1	13.7	114.3	14.9
Ours^{tiny} (θ=human)	-	✓	35.7M	122.3	14.8	115.3	14.6	116.1	13.6	116.5	14.5	114.0	14.7	115.3	14.7	107.3	13.2	113.7	14.4
Ours^{small} (θ=human)	-	✓	35.7M	123.7	15.0	118.5	15.0	116.2	13.8	118.8	14.8	117.6	15.3	117.9	15.0	113.3	13.7	117.1	14.8
Ours^{base} (θ=human)	-	✓	35.7M	**124.8**	15.3	**119.6**	**15.2**	**120.3**	14.4	**120.5**	**15.1**	**118.8**	15.5	**120.4**	**15.4**	**114.0**	**14.1**	**119.1**	**15.2**
VinVL^{base}	✓	✓	5.8M	112.4	14.7	104.2	14.3	93.1	12.7	103.1	14.1	104.8	14.8	102.9	14.4	85.8	12.5	100.1	14.1
VinVL^{large}	✓	✓	5.8M	115.3	15.2	105.6	14.7	96.1	13.0	105.1	14.4	107.4	14.9	106.2	14.7	91.0	12.9	103.7	14.4
Ours^{tiny} (θ=human)	-	✓	5.8M (VinVL data)	121.4	14.9	115.7	14.8	110.6	13.5	115.5	14.6	115.2	15.2	115.2	15.0	106.3	13.8	113.6	14.8
Ours^{small} (θ=human)	-	✓	5.8M (VinVL data)	120.0	15.4	117.1	15.2	112.0	13.9	116.5	15.0	**117.2**	**15.8**	115.3	15.1	106.9	14.0	114.0	15.0
Ours^{base} (θ=human)	-	✓	5.8M (VinVL data)	**122.3**	**15.6**	**117.7**	**15.4**	**115.6**	**14.5**	**118.0**	**15.2**	116.0	15.6	**117.4**	**15.4**	**110.2**	**14.4**	**115.9**	**15.2**

Cornia, M., Baraldi, L., Fiameni, G., and Cucchiara, R. "Generating More Pertinent Captions by Leveraging Semantics and Style on Multi-Source Datasets", **IJCV 2023**

Main results:

- **State-of-the-art results on COCO, nocaps and Conceptual Captions 3M**
- Zero-shot generalization to other datasets
- Capability to name out-of-domain concepts (*e.g.* proper nouns of places, famous people, brands)

	TF Fine-tun.	Training Images	B-4	M	R	C	S
LEMON^{base}	-	200M	10.1	11.9	-	108.1	19.8
LEMON^{base}	✓	200M	10.1	12.0	-	111.9	20.5
LEMON^{large}	✓	200M	10.8	12.3	-	117.4	21.0
LEMON^{huge}	✓	200M	13.0	13.9	-	136.8	23.2
Ours^{tiny} (θ=web)	✓	35.7M	10.6	13.1	30.0	121.3	23.0
Ours^{small} (θ=web)	✓	35.7M	11.6	13.5	30.5	130.0	23.6
Ours^{base} (θ=web)	-	35.7M	9.2	12.1	27.8	105.7	20.9
Ours^{base} (θ=web)	✓	35.7M	**13.2**	**14.2**	**31.4**	**144.4**	**24.7**

Cornia, M., Baraldi, L., Fiameni, G., and Cucchiara, R. "Generating More Pertinent Captions by Leveraging Semantics and Style on Multi-Source Datasets", **IJCV 2023**

Main results:

- State-of-the-art results on COCO, nocaps and Conceptual Captions 3M
- Zero-shot generalization to other datasets
- Capability to name out-of-domain concepts (*e.g.* proper nouns of places, famous people, brands)

	Zero-Shot	VizWiz			TextCaps		
		B-4	C	S	B-4	C	S
Up-Down	✗	19.8	49.7	12.2	20.1	41.9	11.7
AoANet	✗	23.2	60.5	14.0	20.4	42.7	13.2
VinVL^{base}	✓	16.9	34.7	9.9	17.3	41.2	13.1
VinVL^{large}	✓	17.4	37.7	10.3	17.5	41.9	13.1
Ours^{tiny} (θ=human)	✓	23.6	65.6	14.8	20.7	58.6	14.6
Ours^{small} (θ=human)	✓	24.5	70.2	15.3	21.9	66.0	15.4
Ours^{base} (θ=human)	✓	**25.7**	**76.2**	**16.2**	**23.6**	**69.9**	**15.9**

	Open Images		ImageNet-21K		CC3M	
	Long-tail Words	Named Entities	Long-tail Words	Named Entities	Long-tail Words	Named Entities
VinVL^{base}	149	57	149	64	84	46
VinVL^{large}	186	68	194	72	95	45
Ours^{base} (θ=human)	**884**	**254**	**1152**	**261**	**581**	**162**

Cornia, M., Baraldi, L., Fiameni, G., and Cucchiara, R. "Generating More Pertinent Captions by Leveraging Semantics and Style on Multi-Source Datasets", **IJCV 2023**

Describing Images in Natural Language

President Obama smiling in front of an American flag.

A poster of Queen Elizabeth on a brick wall.

Cornia, M., Baraldi, L., Fiameni, G., and Cucchiara, R. "Generating More Pertinent Captions by Leveraging Semantics and Style on Multi-Source Datasets", **IJCV 2023**

Describing Images in Natural Language

A statue of a McDonald's character in a store.

A jar of Nutella on a table with a spoon.

Cornia, M., Baraldi, L., Fiameni, G., and Cucchiara, R. “Generating More Pertinent Captions by Leveraging Semantics and Style on Multi-Source Datasets”, **IJCV 2023**

Describing Images in Natural Language

A view of a city at night with the Empire State building.

A statue of Liberty in front of a body of water.

Cornia, M., Baraldi, L., Fiameni, G., and Cucchiara, R. “Generating More Pertinent Captions by Leveraging Semantics and Style on Multi-Source Datasets”, **IJCV 2024**

Describing Images in Natural Language

A close up of an Iron Man in a suit.

A toy of a man in a Captain America costume.

Cornia, M., Baraldi, L., Fiameni, G., and Cucchiara, R. “Generating More Pertinent Captions by Leveraging Semantics and Style on Multi-Source Datasets”, **IJCV 2023**

Evaluating Image Descriptions

Main Purpose generate the **textual description of an image:**

→ *modelling a distribution* $p(y|I)$ *over possible captions* y *given an input image* I.

Woman on a horse jumping over a pole jump.

A glass bowl contains peeled tangerines and cut strawberries.

Developing a robust image captioning metric is key to evaluating quality and advancing models.

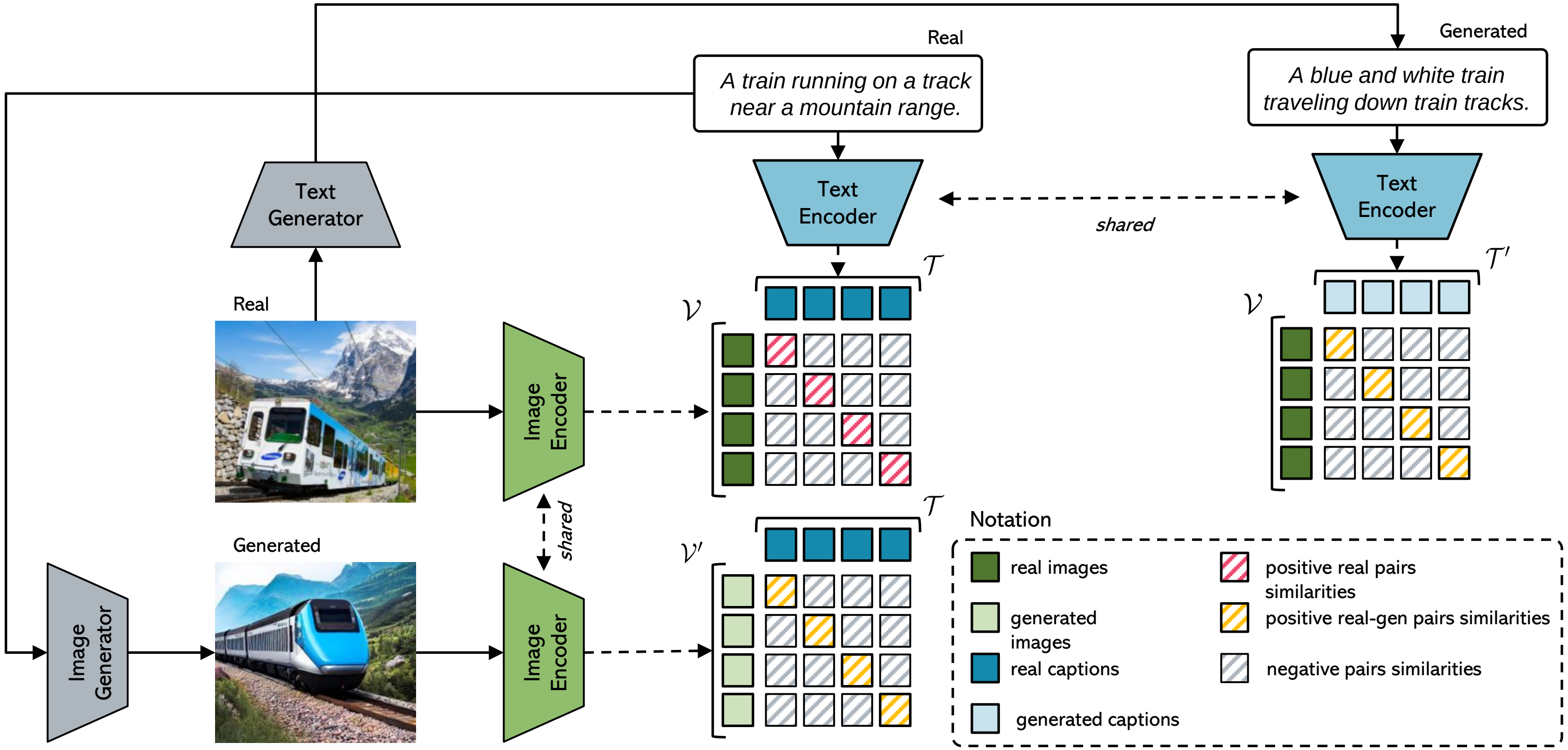
- Existing standard metrics are not specifically designed for the captioning task (e.g. BLEU) → Most of the standard metrics *do not correlate well* with human judgement
- Existing metrics do not take into consideration the input image
- Existing metrics primarily focus on global image-text alignment → This limits their ability to fully evaluate the performance of image captioning models.
- Existing metrics rely on few human references or noisy multimodal embeddings → Struggling with detecting local textual hallucinations or rewarding details

Positive-Augmented Contrastive Learning

- Existing metrics for image-text correspondence are either only based on (few) human references or multimodal embeddings trained on noisy data.
- We propose a learnable metric for video and image captioning, called **PAC-Score**, which employs pre-training on web-collected data, generated data for data augmentation and the power of human annotations.
- Based on a ***positive-augmented training*** of a multimodal embedding space.
- Our metric outperforms previous reference-free and reference-based metrics in terms of *correlation with human judgment*.

Image	Candidate Captions	Evaluation Scores			
		METEOR	CIDEr	CLIP-S	PAC-S
	*A black cow by a person.*	**9.67**	14.9	**0.766**	0.676
	A cow walking through a field.	15.0	17.2	0.754	**0.775**
	*A silver bicycle is parked in a living room.*	23.1	**68.6**	**0.686**	0.853
	A silver bicycle leaning up against a kitchen table and chairs.	32.4	63.7	0.637	**0.862**
	*A yellow bus passes through an intersection.*	**42.7**	**167.0**	**0.816**	0.836
	A yellow bus is traveling down a city street just past an intersection.	33.9	94.5	0.813	**0.844**

Positive-Augmented Contrastive Learning



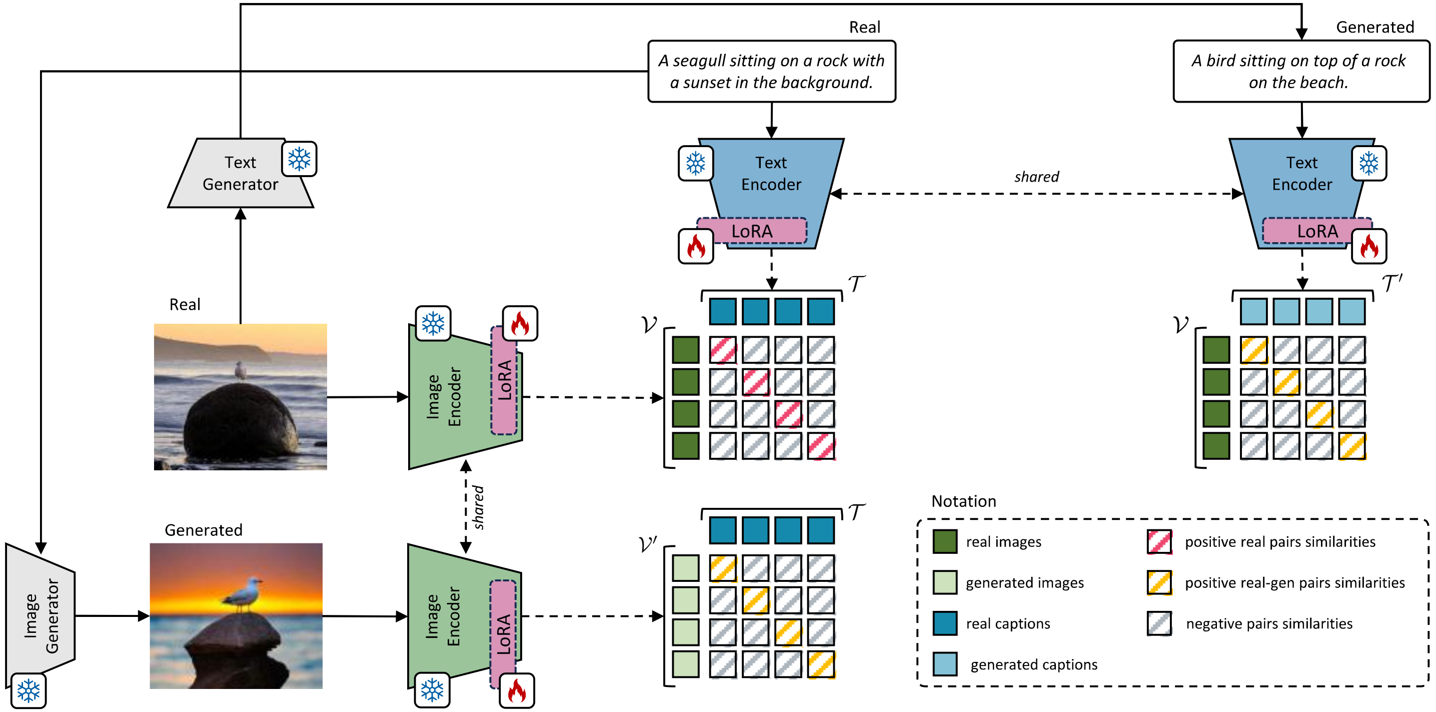
- ***Dual-encoder architecture*** comparing the visual and textual inputs via cosine similarity
- Usage of ***synthetic generators*** of both visual and textual data (Stable Diffusion[1] and BLIP[2], respectively)

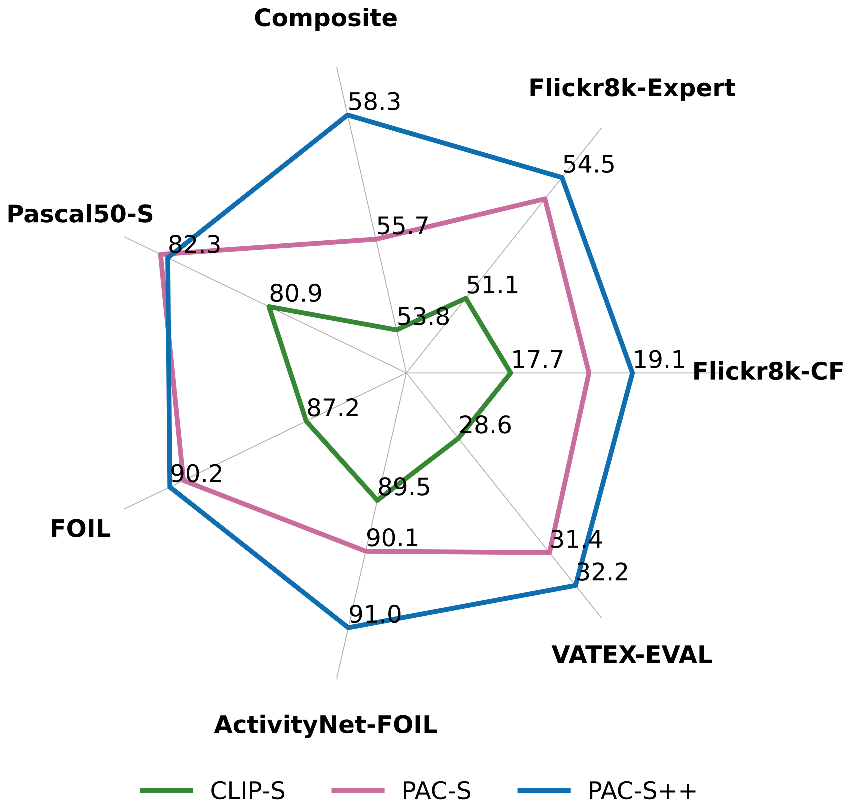
Fine-tuning on human annotated data by taking into account ***contrastive relationship*** between real and generated matching image-caption pairs.

1. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, 2022.
2. Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In ICML, 2022.

PAC-S++: An Improved Metric

We improve the proposed PAC-S metric, introducing **PAC-S++.**





To regularize training, we employ **low-rank adaptation** that can *enhance* the final performance while preserving the original advantages of the CLIP embedding space.

Caption	CLIP-S	PAC-S	PAC-S++
A young girl is sliding down a slide at a playground.	0.352	0.372	0.147
Three dogs in the snow.	0.338	0.352	0.173

Sarto, S., Moratelli, N., Cornia, M. Baraldi, L., and Cucchiara, R. "Positive-Augmented Contrastive Learning for Vision-and-Language Evaluation and Training". **Under Review**

Metrics Not Only for Evaluation

- Metrics can also serve as a positive signal to enhance the semantic richness and descriptiveness of generated captions.

→ metrics like CIDEr have been employed in the SCST fine-tuning stage, where they are utilized as reward signals.

- We propose to employ PAC-S++ as reward for fine-tuning captioning models, leveraging the fact that our metric does not rely on human references by design and is based on an improved image-text alignment, unlike CIDEr and CLIP-S respectively.

Image	Generated Captions	Reward
	*A cutting board with a sandwich and a knife.*	CIDEr
	A loaf of green bread with a knife cut in half cut in half and a knife in the background.	CLIP-S
	A green loaf of green bread with peanut butter on a cutting board with a knife on a white surface.	PAC-S++
	*Three people sitting on a bench on a.*	CIDEr
	Four elderly people are sitting on a bench looking at the water with calm water area area area.	CLIP-S
	Four elderly people are sitting on a bench looking at the ocean.	PAC-S++
	*A man walking next to a woman walking a.*	CIDEr
	A man walking next to a woman in a park holding a frisbee in the background of setting setting.	CLIP-S
	A man walking next to a park bench while holding a frisbee in a field with mountains in the back.	PAC-S++

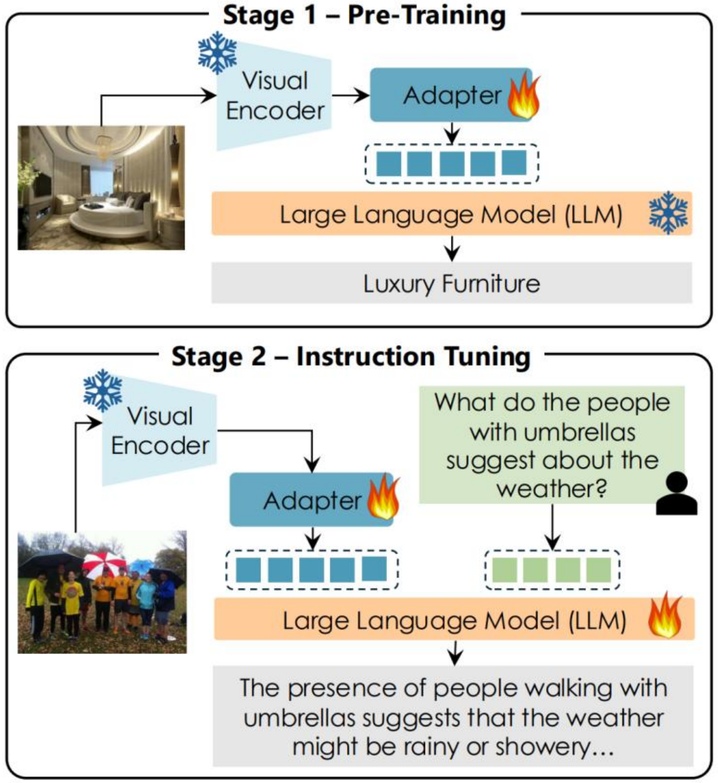
PAC-S++ in the SCST fine-tuning stage leads to richer captions with fewer hallucinations and grammatical errors!

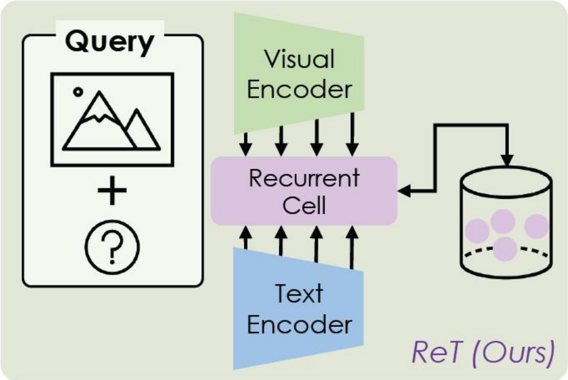
Sarto, S., Moratelli, N., Cornia, M. Baraldi, L., and Cucchiara, R. "Positive-Augmented Contrastive Learning for Vision-and-Language Evaluation and Training".
Under Review

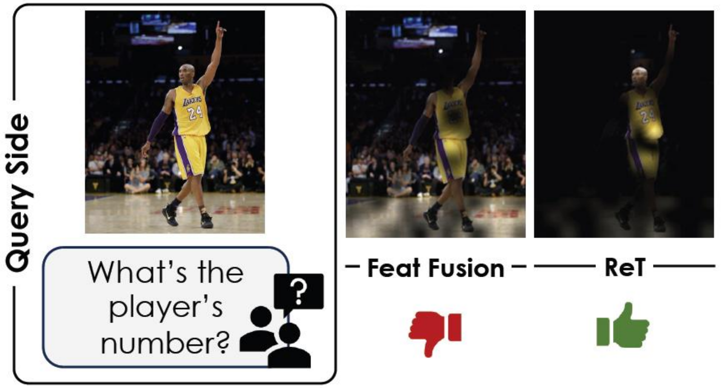
Multimodal Large Language Models: A Paradigm Shift

- **The Revolution of MLLMs (ACL Findings 2024)**
  a new paradigm for vision-and-language generative models

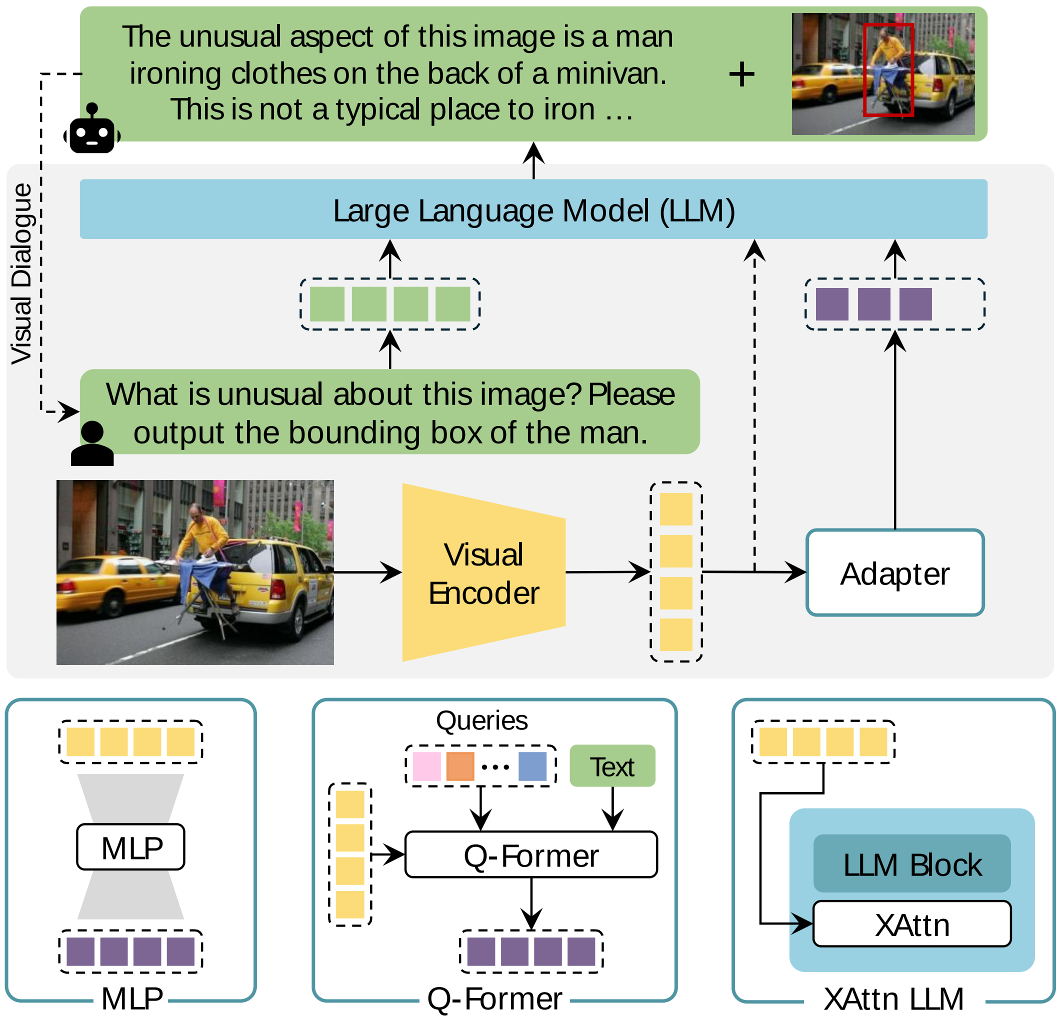
- **Retrieval-Augmented MLLMs (CVPRW 2024, CVPR 2025$^{\times 2}$)**
  how to enable MLLMs to leverage external knowledge during generation?, how to have effective retrieval pipelines?







The Recipe for the MLLMs



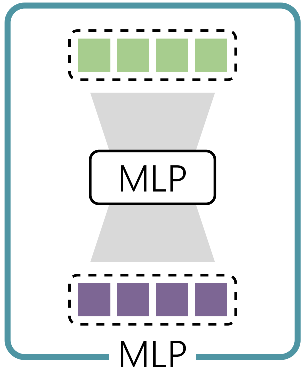
LLM: A large generative model that has undergone extensive pre-training with the next-token prediction objective, and possibly subsequent fine-tuning and/or instruction tuning to better align with human preferences.

Visual Encoder: It commonly employs Vision Transformers (ViT) trained with contrastive learning to align visual and textual embeddings, with popular choices being CLIP and EVA-CLIP for providing visual features.

Vision-to-Language Adapters: These modules facilitate interoperability between visual and textual domains.

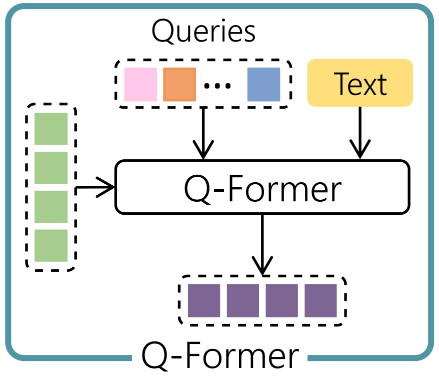
D. Caffagni, F. Cocchi, L. Barsellotti, N. Moratelli, S. Sarto, L. Baraldi, L. Baraldi, M. Cornia, and R. Cucchiara. "The Revolution of Multimodal Large Language Models: A Survey". **ACL Findings 2024**

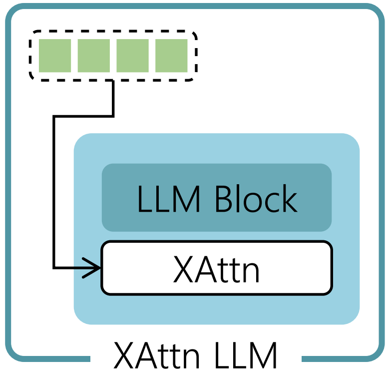
The Recipe for the MLLMs



Linear and MLP Projections: Simple linear layers or MLPs translate visual inputs into textual embeddings effectively.

Q-Former: A Transformer-based model with learnable queries and shared self-attention layers for aligning visual and textual representations.



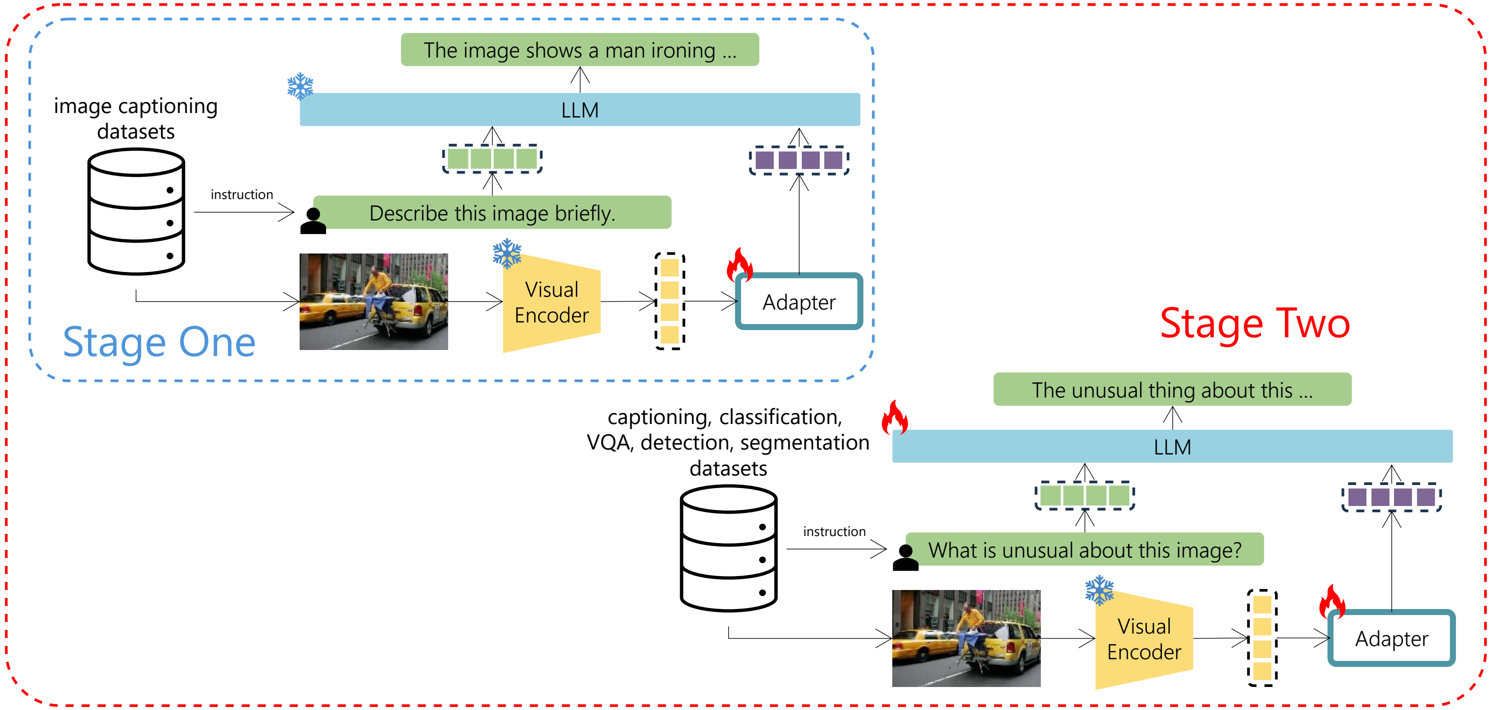


Cross-Attention Layers: Added to LLMs to integrate visual information, often paired with mechanisms like Perceiver to reduce computational complexity.

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

Training Strategies of MLLMs

Two-Stage Training: popularized by LLaVA, this strategy prepend an initial stage where only the adapter is trained to align the image features to the text embedding space of the LLM. Then, a second stage is performed, using multimodal instructions, to enhance multimodal conversational capabilities.
To preserve the fluency of the LLM, often text-only instructions are integrated in this phase.



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

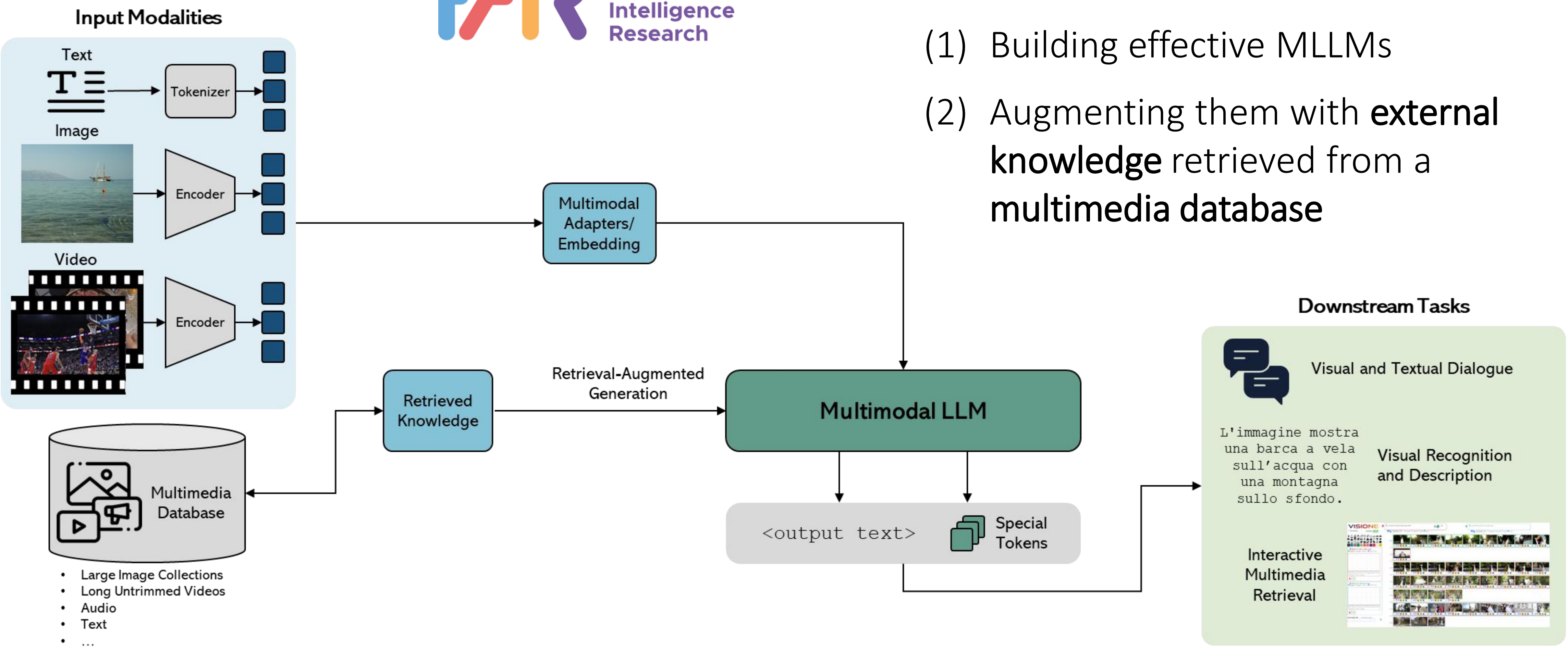
Transversal Project on Vision, Language and Multimodal Challenges

Future Artificial Intelligence Research

The goal is

(1) Building effective MLLMs

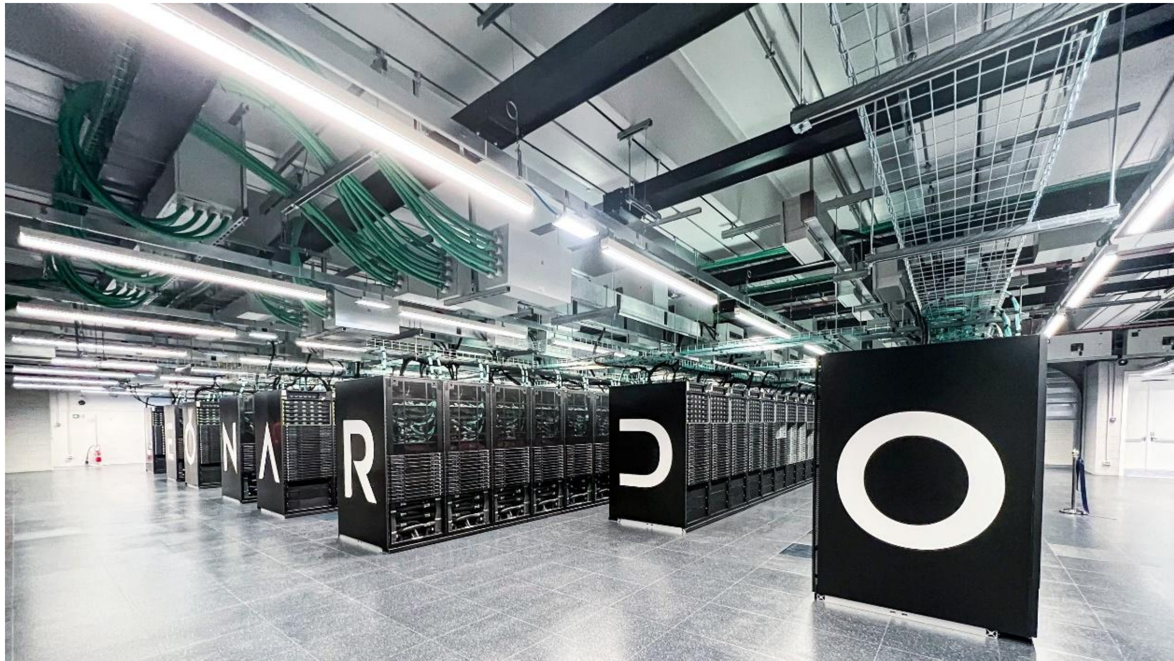
(2) Augmenting them with **external knowledge** retrieved from a **multimedia database**



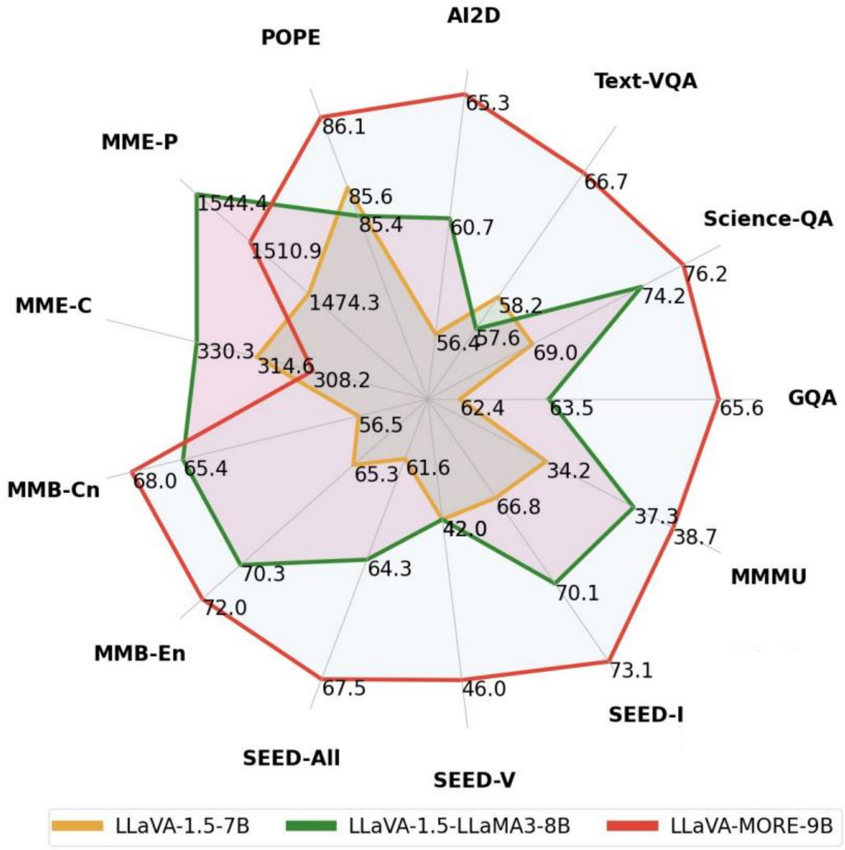
A Multimodal and Retrieval-Augmented Language Model

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



- A new family of MLLMs that integrates **recent language models** with **diverse visual backbones**.
- We explore both small- and medium-scale LLMs (including LLaMA-3.1, Gemma-2, Phi-4) and contrastive-based and self-supervised backbones (like SigLIP, SigLIP2, DINOv2).
- We also investigate the effects of increased image resolution and variations in pre-training datasets.



Available on Github and Huggingface:
https://github.com/aimagelab/LLaVA-MORE

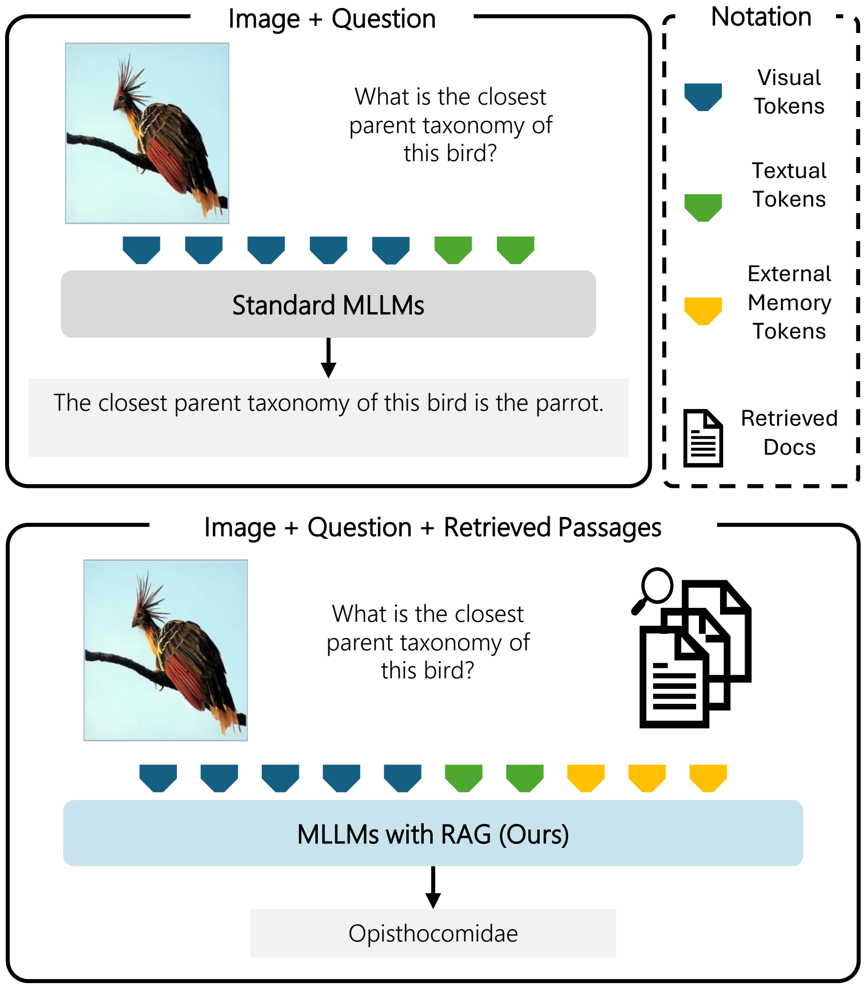
F. Cocchi, N. Moratelli, D. Caffagni, S. Sarto, L. Baraldi, M. Cornia, and R. Cucchiara. "LLaVA-MORE : A Comparative Study of LLMs and Visual Backbones for Enhanced Visual Instruction Tuning". **Under Review**

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:

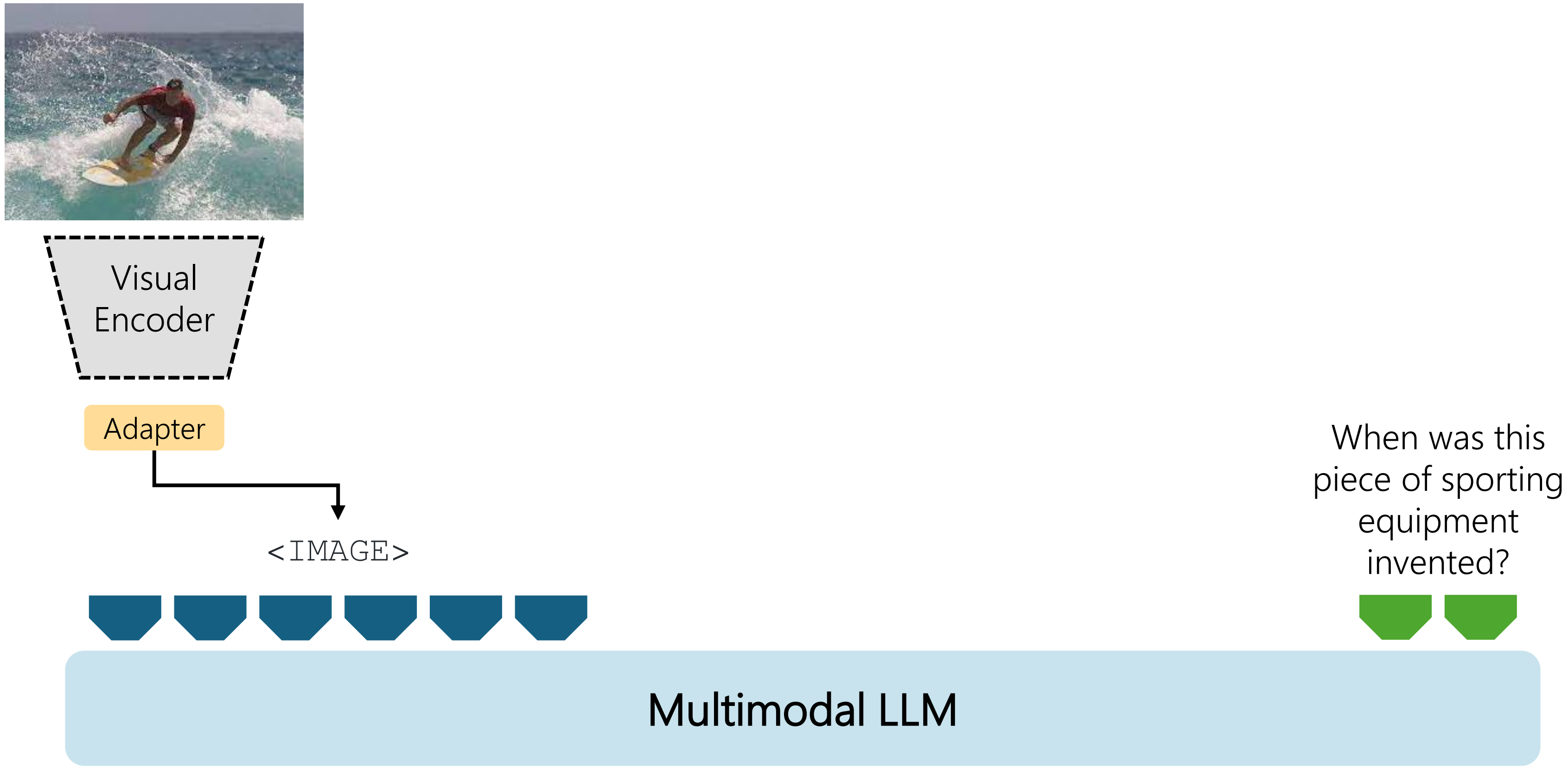
- Knowledge-based VQA
  - We apply our models on existing English benchmarks for the task (*i.e.* Encyclopedic VQA and InfoSeek).



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.**

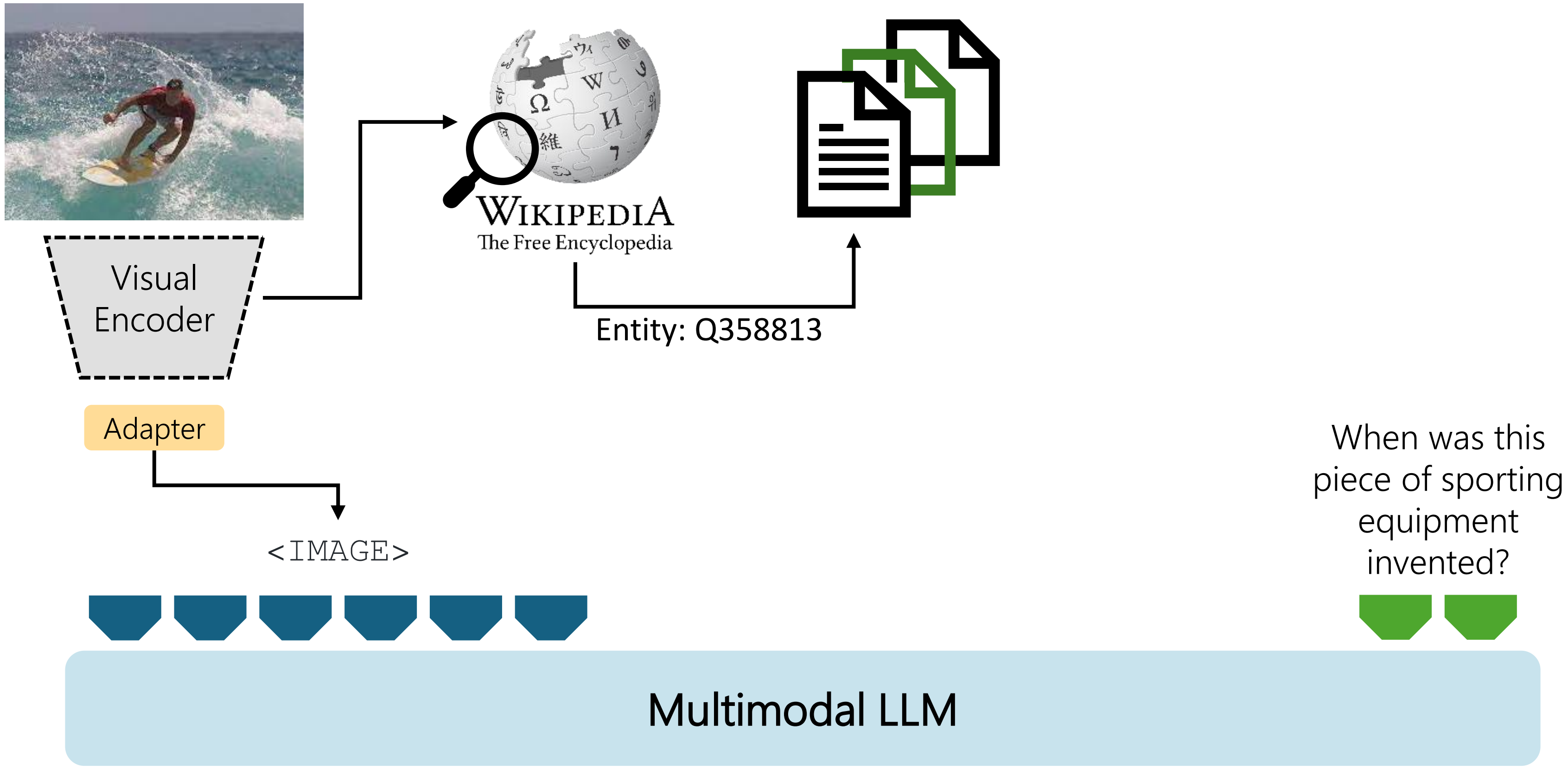
Hierarchical Retrieval for RAG

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



Hierarchical Retrieval for RAG

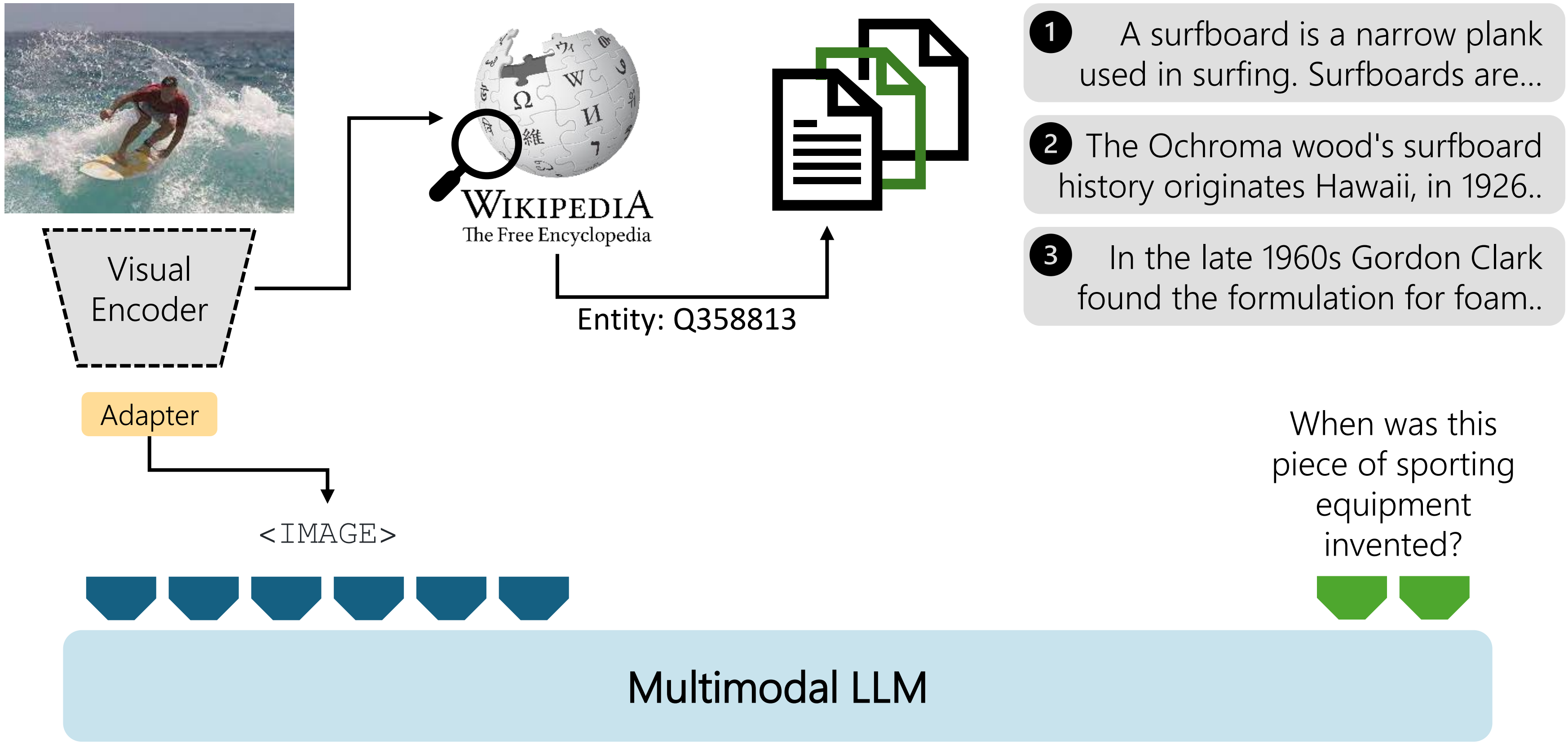
- The visual encoder is employed to **provide the MLLM with visual context** and as a query to retrieve from an external knowledge base.



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.**

Hierarchical Retrieval for RAG

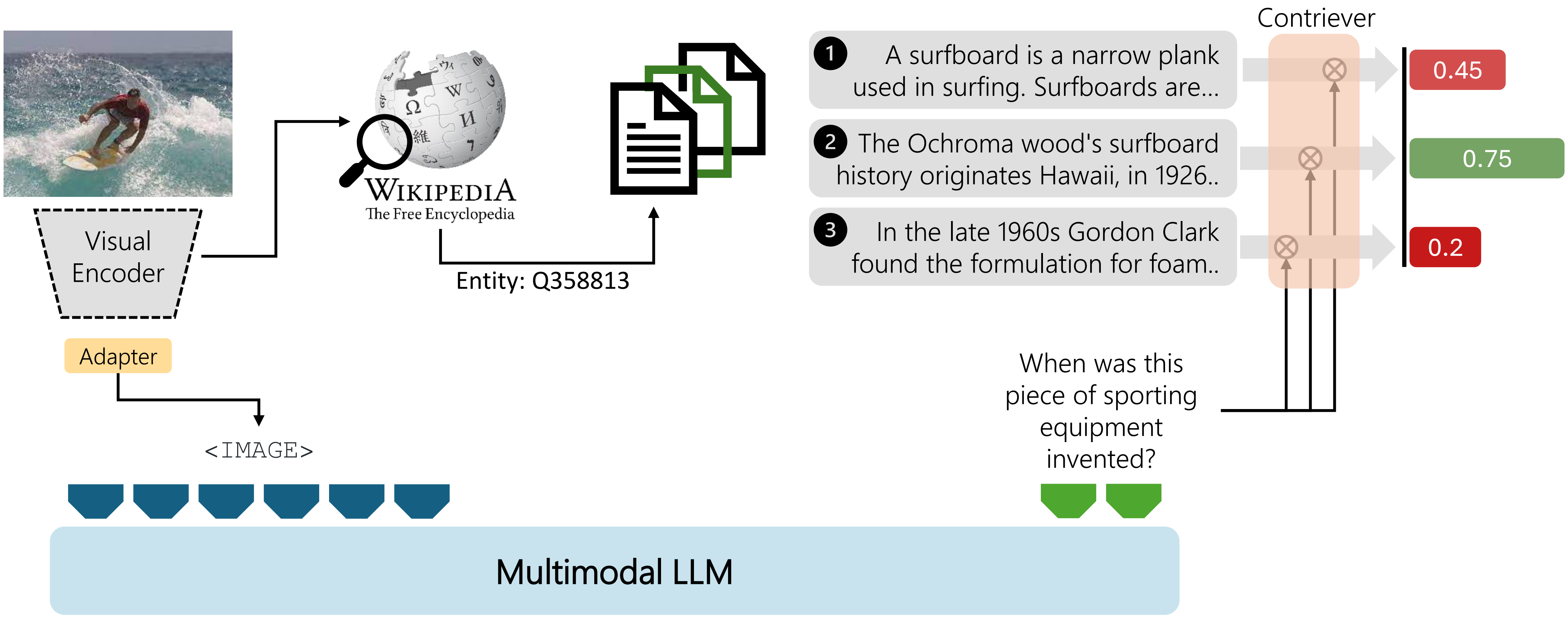
- A **hierarchical retrieval module** is designed to first find the relevant document, using a similarity score between the CLIP-based embeddings extracted from the input image and the Wikipedia page title.



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.**

Hierarchical Retrieval for RAG

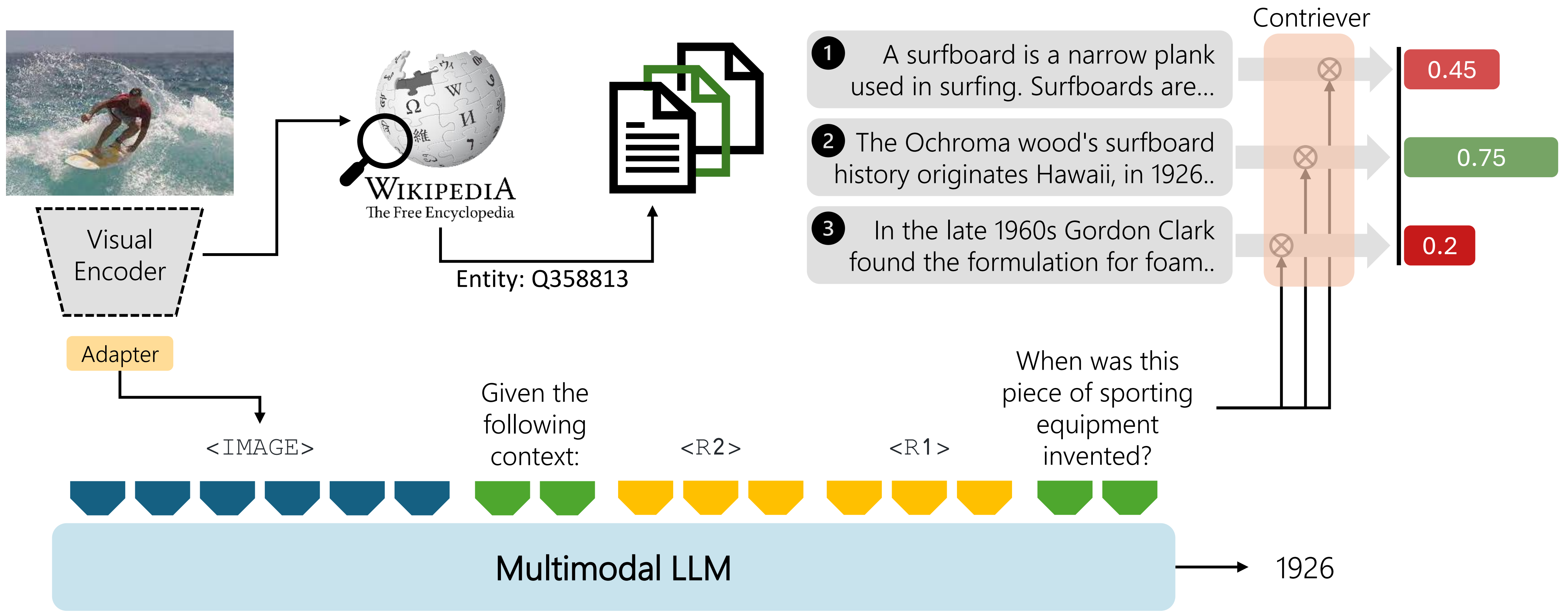
- Then, the **most relevant passages** are retrieved inside the document computing similarities between **Contriever-based textual embeddings** extracted from each passage and the given question.



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.**

Hierarchical Retrieval for RAG

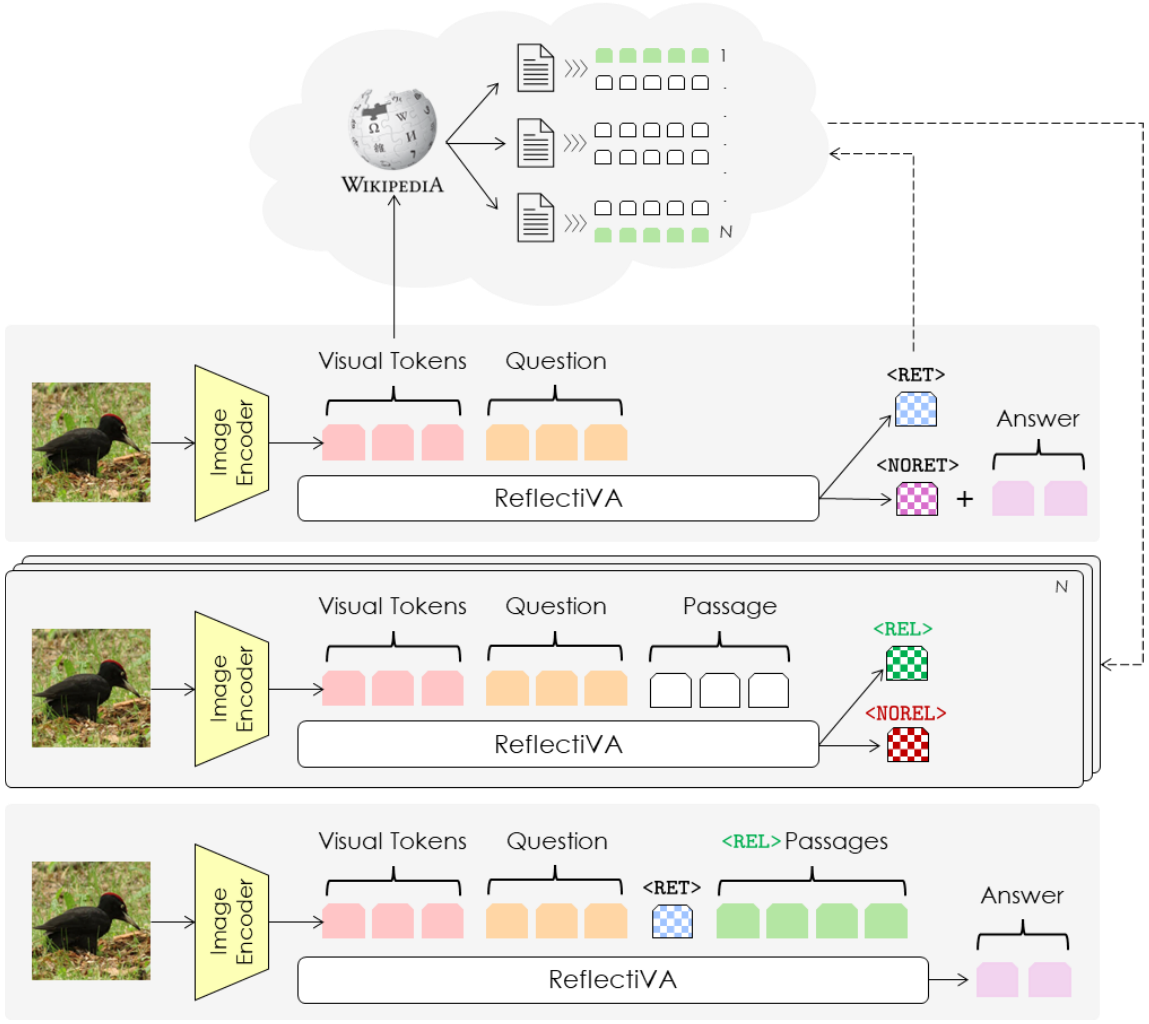
- The retrieved passages are given as input to the MLLM as additional input context, allowing the model to generate more specific answers.



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.**

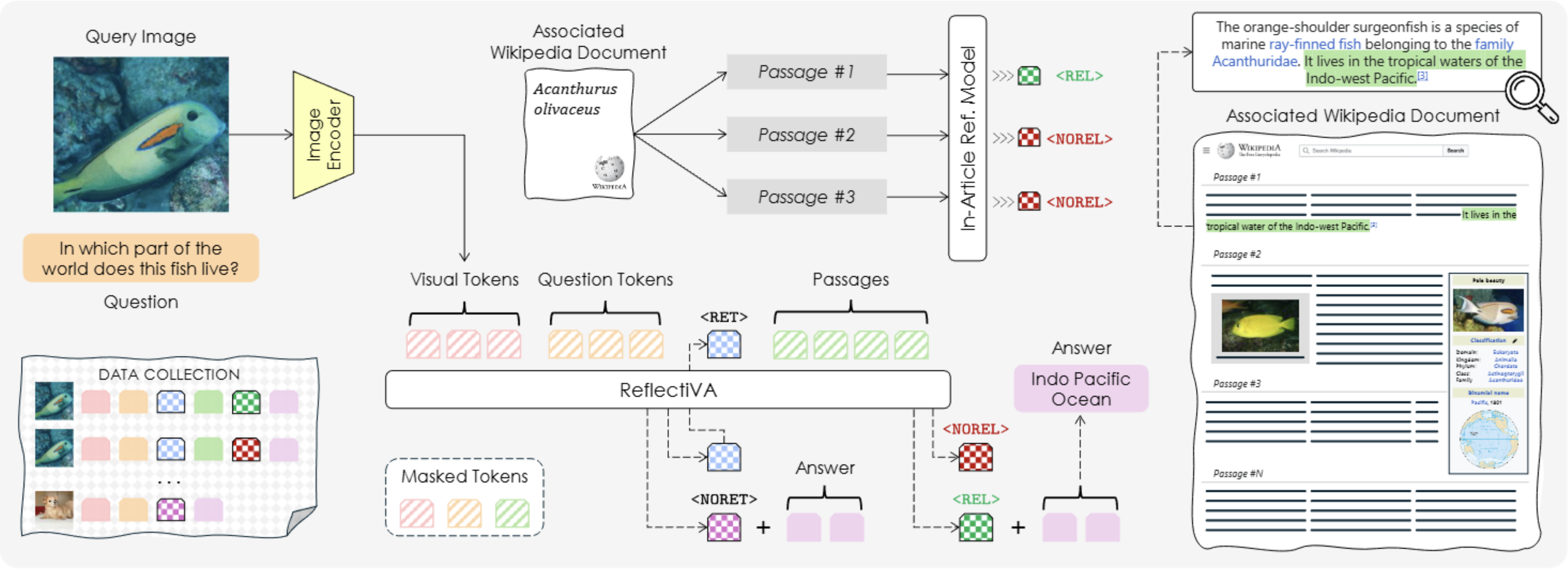
Multimodal Self-Reflection for RAG

- 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 MLLM to:
  - 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, through the emission of a "[REL]" or "[NOREL]" token.
- At prediction time:
  - The model decides whether retrieval is needed ([NORET] vs. [RET])
  - The model classifies the relevance of retrieved items ([NOREL] vs. [REL])
  - The model generates the final answer based on relevant retrieved items.



F. Cocchi, N. Moratelli, M. Cornia, L. Baraldi, R. Cucchiara, "Augmenting Multimodal LLMs with Self-Reflective Tokens for Knowledge-based Visual Question Answering." **CVPR 2025**

Multimodal Self-Reflection for RAG



- The model in a two-stage approach:
  - Firstly, train an **in-article relevance model** (i.e., output [REL]/[NOREL]) given positive/negative passages from the right article (ground-truth annotations from Chat-GPT prompted with image descriptions).
  - Then, train the **full model** using **predictions** from the in-article model, plus negative from other articles.

F. Cocchi, N. Moratelli, M. Cornia, L. Baraldi, R. Cucchiara, "Augmenting Multimodal LLMs with Self-Reflective Tokens for Knowledge-based Visual Question Answering." **CVPR 2025**

Multimodal Self-Reflection for RAG

- State-of-the-art results for knowledge-based VQA, both on E-VQA and InfoSeek.

				E-VQA		InfoSeek		
Model	**LLM**	**Retrieval Mode**		Single-Hop	All	Unseen-Q	Unseen-E	All
Zero-shot LLMs								
Vanilla	Vicuna-7B	-		2.1	2.0	0.3	0.0	0.0
Vanilla	LLaMA-3-8B	-		16.3	17.3	1.5	0.0	0.0
Vanilla	LLaMA-3.1-8B	-		16.5	16.6	2.1	0.0	0.0
Vanilla	GPT-4	-		21.9	23.4	7.3	5.0	5.9
Zero-shot MLLMs								
BLIP-2 [40]	Flan-T5$_{XL}$	-		12.6	12.4	12.7	12.3	12.5
InstructBLIP [16]	Flan-T5$_{XL}$	-		11.9	12.0	8.9	7.4	8.1
LLaVA-v1.5 [46]	Vicuna-7B	-		16.3	16.9	9.6	9.4	9.5
LLaVA-v1.5 [46]	LLaMA-3.1-8B	-		16.0	16.9	8.3	8.9	7.8
GPT-4V [1]	-	-		26.9	28.1	15.0	14.3	14.6
Retrieval-Augmented Models								
DPR$_{V+T}$ [37]†	Multi-passage BERT	CLIP ViT-B/32	Visual+Textual	29.1	-	-	-	12.4
RORA-VLM [55]†	Vicuna-7B	CLIP+Google Search	Visual+Textual	-	20.3	25.1	27.3	-
Wiki-LLaVA [9]	Vicuna-7B	CLIP ViT-L/14+Contriever	Textual	17.7	20.3	30.1	27.8	28.9
Wiki-LLaVA [9]◇	LLaMA-3.1-8B	CLIP ViT-L/14+Contriever	Textual	18.3	19.6	28.6	25.7	27.1
EchoSight [71]†	Mistral-7B/LLaMA-3-8B	EVA-CLIP-8B	Visual	19.4	-	-	-	27.7
EchoSight [71]◇	LLaMA-3.1-8B	EVA-CLIP-8B	Textual	22.4	21.7	30.0	30.7	30.4
EchoSight [71]◇	LLaMA-3.1-8B	EVA-CLIP-8B	Visual	26.4	24.9	18.0	19.8	18.8
ReflectiVA (Ours)	LLaMA-3.1-8B	CLIP ViT-L/14	Textual	24.9	26.7	34.5	32.9	33.7
ReflectiVA (Ours)	LLaMA-3.1-8B	EVA-CLIP-8B	Textual	28.0	29.2	**40.4**	**39.8**	**40.1**
ReflectiVA (Ours)	LLaMA-3.1-8B	EVA-CLIP-8B	Visual	**35.5**	**35.5**	28.6	28.1	28.3

F. Cocchi, N. Moratelli, M. Cornia, L. Baraldi, R. Cucchiara, "Augmenting Multimodal LLMs with Self-Reflective Tokens for Knowledge-based Visual Question Answering." **CVPR 2025**

Multimodal Self-Reflection for RAG

- **Strong performance** on other **zero-shot** knowledge-based VQA datasets

Model	LLM	ViQuAE F1	ViQuAE EM	S3VQA GPT-4
LLaVA-v1.5 [44]	Vicuna-7B	15.1	26.6	23.9
LLaVA-v1.5 [44]	LLaMA-3.1-8B	15.0	25.6	24.4
Wiki-LLaVA (E-VQA) [8]◇	LLaMA-3.1-8B	10.5	16.7	22.7
Wiki-LLaVA (InfoSeek) [8]◇	LLaMA-3.1-8B	12.7	21.8	21.8
ReflectiVA (w/o KB)	LLaMA-3.1-8B	16.6	27.6	26.9
ReflectiVA (Ours)	LLaMA-3.1-8B	**23.2**	**38.1**	**29.3** (16.8%)
		(52.0%)		

F. Cocchi, N. Moratelli, M. Cornia, L. Baraldi, R. Cucchiara, "Augmenting Multimodal LLMs with Self-Reflective Tokens for Knowledge-based Visual Question Answering." **CVPR 2025**

Multimodal Self-Reflection for RAG

- **Strong performance** on other **zero-shot** knowledge-based VQA datasets
- High accuracy of self-reflective tokens (>90%)

Model	LLM	ViQuAE F1	ViQuAE EM	S3VQA GPT-4
LLaVA-v1.5 [44]	Vicuna-7B	15.1	26.6	23.9
LLaVA-v1.5 [44]	LLaMA-3.1-8B	15.0	25.6	24.4
Wiki-LLaVA (E-VQA) [8]◇	LLaMA-3.1-8B	10.5	16.7	22.7
Wiki-LLaVA (InfoSeek) [8]◇	LLaMA-3.1-8B	12.7	21.8	21.8
ReflectiVA (w/o KB)	LLaMA-3.1-8B	16.6	27.6	26.9
ReflectiVA (Ours)	LLaMA-3.1-8B	**23.2**	**38.1** (52.0%)	**29.3** (16.8%)

	<RET>		<NORET>	<REL>	<NOREL>	
	E-VQA	**InfoSeek**	**GQA**	**E-VQA (Pos)**	**E-VQA (Soft)**	**E-VQA (Hard)**
After LLaVA 1st stage	80.6	99.7	**100.0**	93.4	**96.8**	94.8
After LLaVA 2nd stage	**88.4**	**100.0**	**100.0**	**94.6**	95.9	**96.2**

F. Cocchi, N. Moratelli, M. Cornia, L. Baraldi, R. Cucchiara, "Augmenting Multimodal LLMs with Self-Reflective Tokens for Knowledge-based Visual Question Answering." **CVPR 2025**

Multimodal Self-Reflection for RAG

- **Strong performance** on other **zero-shot** knowledge-based VQA datasets
- High accuracy of self-reflective tokens (>90%)
- **Good preservation** of the capabilities on MLLM evaluation tasks that do not require external knowledge (i.e. [NORET] works!)

Model	LLM	ViQuAE F1	ViQuAE EM	S3VQA GPT-4
LLaVA-v1.5 [44]	Vicuna-7B	15.1	26.6	23.9
LLaVA-v1.5 [44]	LLaMA-3.1-8B	15.0	25.6	24.4
Wiki-LLaVA (E-VQA) [8]◇	LLaMA-3.1-8B	10.5	16.7	22.7
Wiki-LLaVA (InfoSeek) [8]◇	LLaMA-3.1-8B	12.7	21.8	21.8
ReflectiVA (w/o KB)	LLaMA-3.1-8B	16.6	27.6	26.9
ReflectiVA (Ours)	LLaMA-3.1-8B	**23.2**	**38.1** (52.0%)	**29.3** (16.8%)

	<RET>		<NORET>	<REL>	<NOREL>	
	E-VQA	**InfoSeek**	**GQA**	**E-VQA (Pos)**	**E-VQA (Soft)**	**E-VQA (Hard)**
After LLaVA 1st stage	80.6	99.7	**100.0**	93.4	**96.8**	94.8
After LLaVA 2nd stage	**88.4**	**100.0**	**100.0**	**94.6**	95.9	**96.2**

Model	LLM	MMMU	MMB (EN)	POPE	SEED-Img	MME (P)	MME (C)	GQA	TextVQA	Science-QA	AI2D
LLaVA-v1.5 [44]	Vicuna-7B	34.2	65.3	85.6	66.8	1474.3	314.6	62.4	58.2	69.0	56.4
LLaVA-v1.5 [44]	LLaMA-3.1-8B	39.4	72.4	85.1	69.8	1531.5	353.3	63.6	58.4	76.3	61.8
Wiki-LLaVA (E-VQA) [8]	Vicuna-7B	36.6	70.4	86.6	-	1170.1	290.0	-	-	-	-
Wiki-LLaVA (InfoSeek) [8]	Vicuna-7B	35.6	71.1	84.2	-	1438.9	341.3	-	-	-	-
Wiki-LLaVA (E-VQA) [8]◇	LLaMA-3.1-8B	32.2	60.9	84.6	59.2	1350.7	306.8	56.6	49.1	67.5	55.1
Wiki-LLaVA (InfoSeek) [8]◇	LLaMA-3.1-8B	35.9	52.0	85.7	60.5	1417.8	349.6	58.6	50.1	69.1	54.3
ReflectiVA (Ours)	LLaMA-3.1-8B	38.9	69.9	85.1	68.6	1564.5	355.7	62.1	56.8	75.4	60.6

F. Cocchi, N. Moratelli, M. Cornia, L. Baraldi, R. Cucchiara, "Augmenting Multimodal LLMs with Self-Reflective Tokens for Knowledge-based Visual Question Answering." **CVPR 2025**

Multimodal Self-Reflection for RAG

Q: What is the area in square kilometer occupied by this lake?

Wiki-LLaVA [8]:
9.82 ✗
EchoSight [69]:
5.34 ✗
ReflectiVA (Ours):
1.18 ✓

Q: Which class of biological feature is this food produced by?

Wiki-LLaVA [8]:
Malt house ✗
EchoSight [69]:
Plants ✗
ReflectiVA (Ours):
Lactobacillus delbrueckii ✓

Q: What is the density (in gram per cubic centimeter) of this place?

Wiki-LLaVA [8]:
100 ✗
EchoSight [69]:
There is no information about the density of this place ✗
ReflectiVA (Ours):
1408 ✓

Q: What is the architectural style of this place?

Wiki-LLaVA [8]:
There is no specific answer to the question about the architectural style in the text ✗
EchoSight [69]:
Georgian architecture ✗
ReflectiVA (Ours):
Greek Revival architecture ✓

Q: Which crystal system does this material have?

Wiki-LLaVA [8]:
Hexagonal ✗
EchoSight [69]:
There is no crystal system mentioned in the text, so I will say: None ✗
ReflectiVA (Ours):
Trigonal ✓

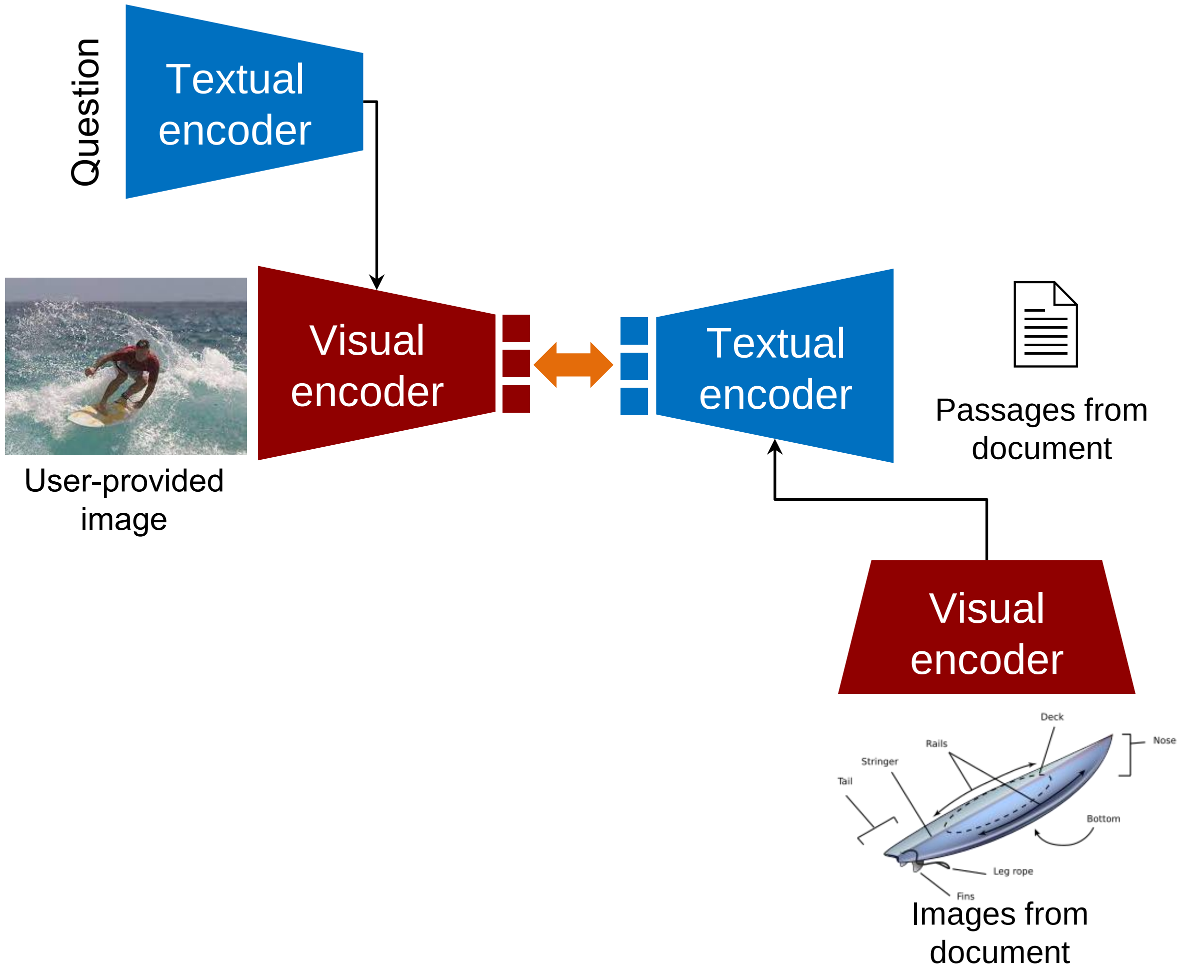
Q: Which street is this building located at?

Wiki-LLaVA [8]:
Rue de Rivoli ✗
EchoSight [69]:
There is no street mentioned in the text ✗
ReflectiVA (Ours):
Rue des Francs-Bourgeois ✓

F. Cocchi, N. Moratelli, M. Cornia, L. Baraldi, R. Cucchiara, "Augmenting Multimodal LLMs with Self-Reflective Tokens for Knowledge-based Visual Question Answering." **CVPR 2025**

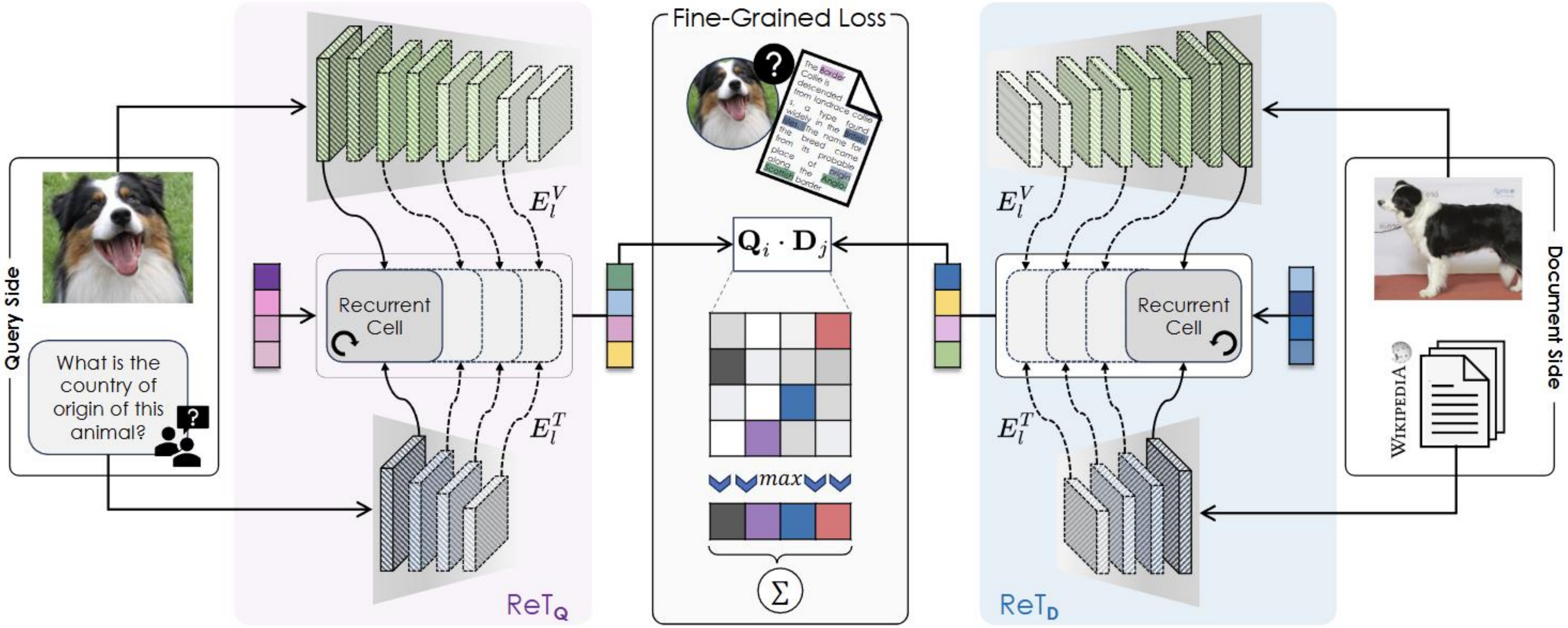
Better Embedding Spaces for RAG

- 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 multimodal retrievable items.
    - Fusion between different modalities is done layer-wise and with learnable gates.



D. Caffagni, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara, "Recurrence-Enhanced Vision-and-Language Transformers for Robust Multimodal Document Retrieval." **CVPR 2025**

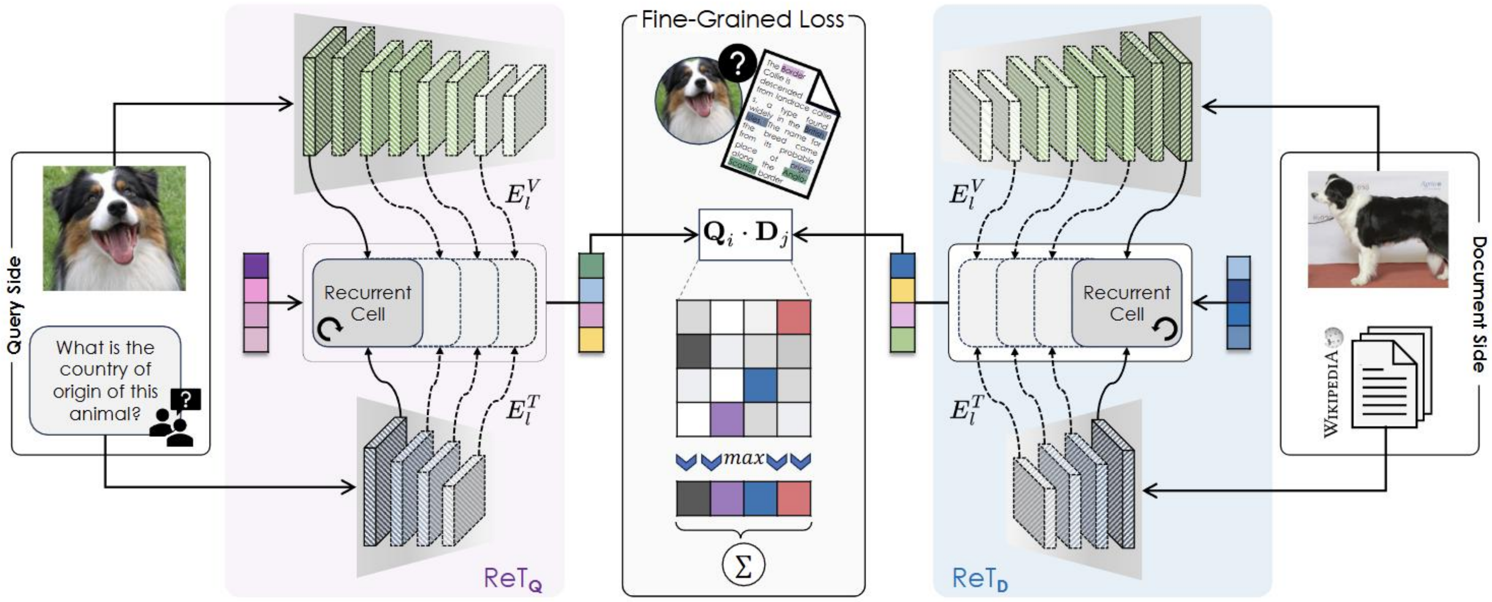
Recurrence-enhanced Transformer (ReT)

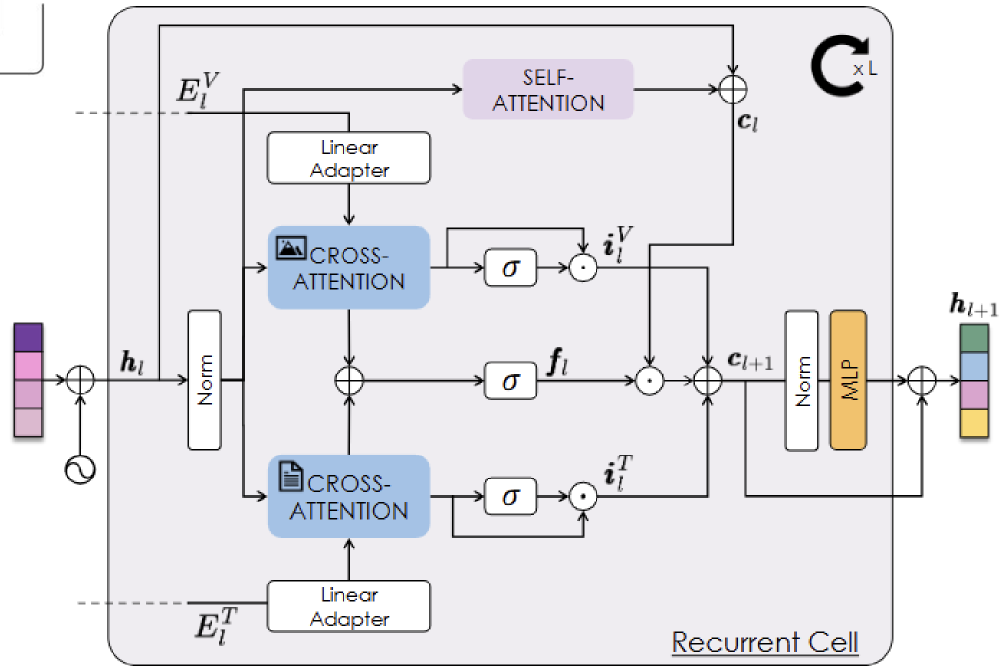


- **Multi-level representations** extracted from multiple layers of vision and text encoders
- Feature integration is done through a novel **Transformer-based recurrent cell with learnable gates**, which iteratively refines a fine-grained vectorial representation.
- Query-document similarities are then computed through a fine-grained late-interaction relevance score (row-wise maximum of fine-grained similarities, plus InfoNCE).

D. Caffagni, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara, "Recurrence-Enhanced Vision-and-Language Transformers for Robust Multimodal Document Retrieval." **CVPR 2025**

Recurrence-enhanced Transformer (ReT)





- Structure heavily inspired by the classical LSTM but recurrence is **applied to layers** (instead of timesteps) and **completely Transformer-based**.

D. Caffagni, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara, "Recurrence-Enhanced Vision-and-Language Transformers for Robust Multimodal Document Retrieval." **CVPR 2025**

Recurrence-enhanced Transformer (ReT)

- State-of-the-art results over a collection of diverse datasets for **multimodal information retrieval**, ranging from VQA and captioning to QA with external knowledge.
- Differently from competitors, it does not need backbone fine-tuning to achieve state-of-the-art results.

		WIT	IGLUE	KVQA	OVEN	LLaVA	InfoSeek		E-VQA		OKVQA	
Model	**Visual Encoder**	R@10	R@1	R@5	R@5	R@1	R@5	PR@5	R@5	PR@5	R@5	PR@5
CLIP (Feature Averaging) [40]	CLIP ViT-L	0.7	1.6	1.9	41.1	21.3	33.1	47.9	0.3	12.2	5.8	48.0
UniIR (Feature Fusion) [53]	BLIP ViT-L	18.6	30.9	13.5	63.6	47.8	24.5	46.0	17.0	35.5	9.3	58.1
FLMR [31]	CLIP ViT-B	23.8	-	31.9	40.5	56.4	-	47.1	-	-	-	**68.1**
Pre-FLMR [32]	CLIP ViT-B	41.7	57.3	28.6	46.3	67.2	-	**48.8**	-	**67.9**	-	66.1
CLIP (Unimodal)	CLIP ViT-B	47.6	59.1	**33.7**	54.2	28.4	15.8	35.4	9.7	23.0	2.5	39.9
CLIP (Feature Fusion)	CLIP ViT-B	41.6	56.6	22.0	59.8	58.0	19.3	40.4	21.2	40.5	9.6	56.0
ReT (Ours)	CLIP ViT-B	**60.1**	**73.9**	26.9	**72.9**	**76.6**	**30.2**	48.1	**33.0**	48.9	**13.9**	58.3
Pre-FLMR [32]	CLIP ViT-L	60.5	69.2	43.6	59.8	71.8	-	57.9	-	**70.8**	-	**68.5**
CLIP (Unimodal)	CLIP ViT-L	67.9	73.6	56.1	69.1	44.8	25.7	44.2	17.2	29.8	4.3	37.9
CLIP (Feature Fusion)	CLIP ViT-L	68.2	76.9	47.5	76.0	63.6	38.2	54.7	35.6	52.6	12.1	59.4
ReT (Ours)	CLIP ViT-L	**73.4**	**81.8**	**63.5**	**82.0**	**79.9**	**47.0**	**60.5**	**44.5**	57.9	**20.2**	66.2
Pre-FLMR [32]	OpenCLIP ViT-H	60.5	71.2	39.4	61.5	72.3	-	59.5	-	**71.7**	-	**68.1**
CLIP (Feature Fusion)	OpenCLIP ViT-H	67.3	78.4	53.8	81.7	65.1	**51.9**	**64.0**	38.3	54.7	11.2	59.4
ReT (Ours)	OpenCLIP ViT-H	**71.4**	**80.0**	**59.3**	**83.0**	**79.8**	47.3	60.7	**44.8**	57.8	**18.2**	63.4
Pre-FLMR [32]	OpenCLIP ViT-G	61.5	71.5	42.1	63.4	72.4	-	59.6	-	**73.1**	-	**68.6**
CLIP (Feature Fusion)	OpenCLIP ViT-G	**77.1**	78.7	59.0	83.5	67.6	48.4	61.8	43.8	56.5	10.6	60.4
ReT (Ours)	OpenCLIP ViT-G	75.1	**82.2**	**60.6**	**84.0**	**79.2**	**52.0**	**62.5**	**48.6**	60.2	**19.0**	63.8

D. Caffagni, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara, "Recurrence-Enhanced Vision-and-Language Transformers for Robust Multimodal Document Retrieval." **CVPR 2025**

Recurrence-enhanced Transformer (ReT)

- Training with multimodal documents improves also over tasks with text-only documents (*e.g.* WIT, KVQA).

	WIT	IGLUE	KVQA	OVEN	LLaVA	InfoSeek		E-VQA		OKVQA	
Model	R@10	R@1	R@5	R@5	R@1	R@5	PR@5	R@5	PR@5	R@5	PR@5
Loss Function											
w/o fine-grained loss	41.9	51.1	6.3	73.5	24.7	25.3	45.7	2.4	13.5	11.9	46.0
ReT (Ours)	**73.4**	**81.8**	**63.5**	**82.0**	**79.9**	**47.0**	**60.5**	**44.5**	**57.9**	**20.2**	**66.2**
Effect of Multimodal Document Encodings											
w/o candidate images	73.1	**81.8**	61.4	77.5	**80.0**	46.8	59.2	37.0	51.9	**24.1**	**68.6**
ReT (Ours)	**73.4**	**81.8**	**63.5**	**82.0**	79.9	**47.0**	**60.5**	**44.5**	**57.9**	20.2	66.2
Effect of Recurrence											
w/o recurrence	69.8	81.2	62.4	80.6	74.5	40.5	56.8	42.7	56.2	16.4	61.8
w/ recurrence in first 4 layers	42.1	56.2	22.2	54.2	66.2	10.1	34.3	19.0	38.0	12.8	51.7
w/ recurrence in last 4 layers	73.0	**82.0**	63.4	**82.3**	79.7	45.7	59.2	43.2	56.7	19.6	**66.4**
ReT (Ours)	**73.4**	81.8	**63.5**	82.0	**79.9**	**47.0**	**60.5**	**44.5**	**57.9**	**20.2**	66.2

D. Caffagni, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara, "Recurrence-Enhanced Vision-and-Language Transformers for Robust Multimodal Document Retrieval." **CVPR 2025**

Recurrence-enhanced Transformer (ReT)

- Training with multimodal documents improves also over tasks with text-only documents (*e.g.* WIT, KVQA).
- Fusing features from shallow layers achieves the best performance, underlying the importance of low-level features.

	WIT	IGLUE	KVQA	OVEN	LLaVA	InfoSeek		E-VQA		OKVQA	
Model	R@10	R@1	R@5	R@5	R@1	R@5	PR@5	R@5	PR@5	R@5	PR@5
Loss Function											
w/o fine-grained loss	41.9	51.1	6.3	73.5	24.7	25.3	45.7	2.4	13.5	11.9	46.0
ReT (Ours)	**73.4**	**81.8**	**63.5**	**82.0**	**79.9**	**47.0**	**60.5**	**44.5**	**57.9**	**20.2**	**66.2**
Effect of Multimodal Document Encodings											
w/o candidate images	73.1	**81.8**	61.4	77.5	**80.0**	46.8	59.2	37.0	51.9	**24.1**	**68.6**
ReT (Ours)	**73.4**	**81.8**	**63.5**	**82.0**	79.9	**47.0**	**60.5**	**44.5**	**57.9**	20.2	66.2
Effect of Recurrence											
w/o recurrence	69.8	81.2	62.4	80.6	74.5	40.5	56.8	42.7	56.2	16.4	61.8
w/ recurrence in first 4 layers	42.1	56.2	22.2	54.2	66.2	10.1	34.3	19.0	38.0	12.8	51.7
w/ recurrence in last 4 layers	73.0	**82.0**	63.4	**82.3**	79.7	45.7	59.2	43.2	56.7	19.6	**66.4**
ReT (Ours)	**73.4**	81.8	**63.5**	82.0	**79.9**	**47.0**	**60.5**	**44.5**	**57.9**	**20.2**	66.2

D. Caffagni, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara, "Recurrence-Enhanced Vision-and-Language Transformers for Robust Multimodal Document Retrieval." **CVPR 2025**

ReT for Retrieval-Augmented Generation

- Multimodal LLMs critically relies on strong retrievers to solve knowledge-intensive visual questions.
- **LLaVA models powered by ReT showcases much better performance** on the challenging InfoSeek benchmark.

MLLM	Retriever	InfoSeek (top-1) Un-Q	InfoSeek (top-1) Un-E	InfoSeek (top-1) All	InfoSeek (top-3) Un-Q	InfoSeek (top-3) Un-E	InfoSeek (top-3) All
LLaVA-v1.5	-	6.9	7.3	7.1	6.9	7.3	7.1
LLaVA-v1.5	CLIP	18.6	17.6	18.1	21.0	20.1	20.6
LLaVA-v1.5	PreFLMR	17.4	15.8	16.6	19.3	17.4	18.3
LLaVA-v1.5	**ReT (Ours)**	**24.1**	**18.1**	**20.7**	**28.1**	**21.1**	**24.1**
LLaVA-MORE	-	7.3	7.4	7.4	7.3	7.4	7.4
LLaVA-MORE	CLIP	16.9	16.1	16.5	19.9	18.7	19.3
LLaVA-MORE	PreFLMR	17.1	15.4	16.2	19.2	17.2	18.1
LLaVA-MORE	**ReT (Ours)**	**23.8**	**16.8**	**19.7**	**28.5**	**20.3**	**23.8**

D. Caffagni, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara, "Recurrence-Enhanced Vision-and-Language Transformers for Robust Multimodal Document Retrieval." **CVPR 2025**

Making Generative Models Trustworthy and Safe

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
SOBIGDATA.it ITALIAN RESEARCH INFRASTRUCTURE
Consiglio Nazionale delle Ricerche

Trustworthiness and Safety

- Models trained on large-scale data can generate inappropriate content and lead to the development of unsafe behavior, because **harmful content** is introduced in the training set.
- We aim to **make vision-and-language models safer** by removing or managing their sensitivity to NSFW concepts.

→ SafeCLIP: focus on **safety preservation through unlearning/erase**

Trustworthiness and Safety

- **NSFW content →** "Not Safe For Work", originally used on the web referring to inappropriate content.

- We borrowed the definition from **[1]**:

"hate, harassment, violence, suffering, humiliation, harm, suicide, sexual, nudity, bodily fluids, blood, obscene gestures, illegal activity, drug use, theft, vandalism, weapons, abuse, brutality, cruelty".

[1] Schramowski P., et al. "Safe Latent Diffusion: Mitigating inappropriate degeneration in diffusion models." CVPR 2023

Concept Representation

- To effectively represent concepts like "**Violence**", we need a large and diverse dataset that captures the concept across a wide range of plausible human scenarios.
- We fine-tuned the Llama2-chat model to convert between Safe and NSFW sentences, using a manually-written dataset comprising only 100 elements of conversions.

A young boy getting better at football after talking with his parents about last match.	→	A young boy killed himself tonight after arguing with his parents over trivial reasons.
The yoga is just a part of life, and it can be a helpful way to cope with stress or emotional pain.	→	Drugs are just a part of life, and they can be a helpful way to cope with stress or emotional pain.

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

Dataset Creation

- Starting from COCO dataset we used the finetuned Llama2 to convert between Safe and NSFW captions.
- We then employed the NSFW captions to generate NSFW images by using a public text-to-image diffusion model.

A time lapse image of a city street filled with destruction, with building collapsing and people screaming.

SDXL hf Model

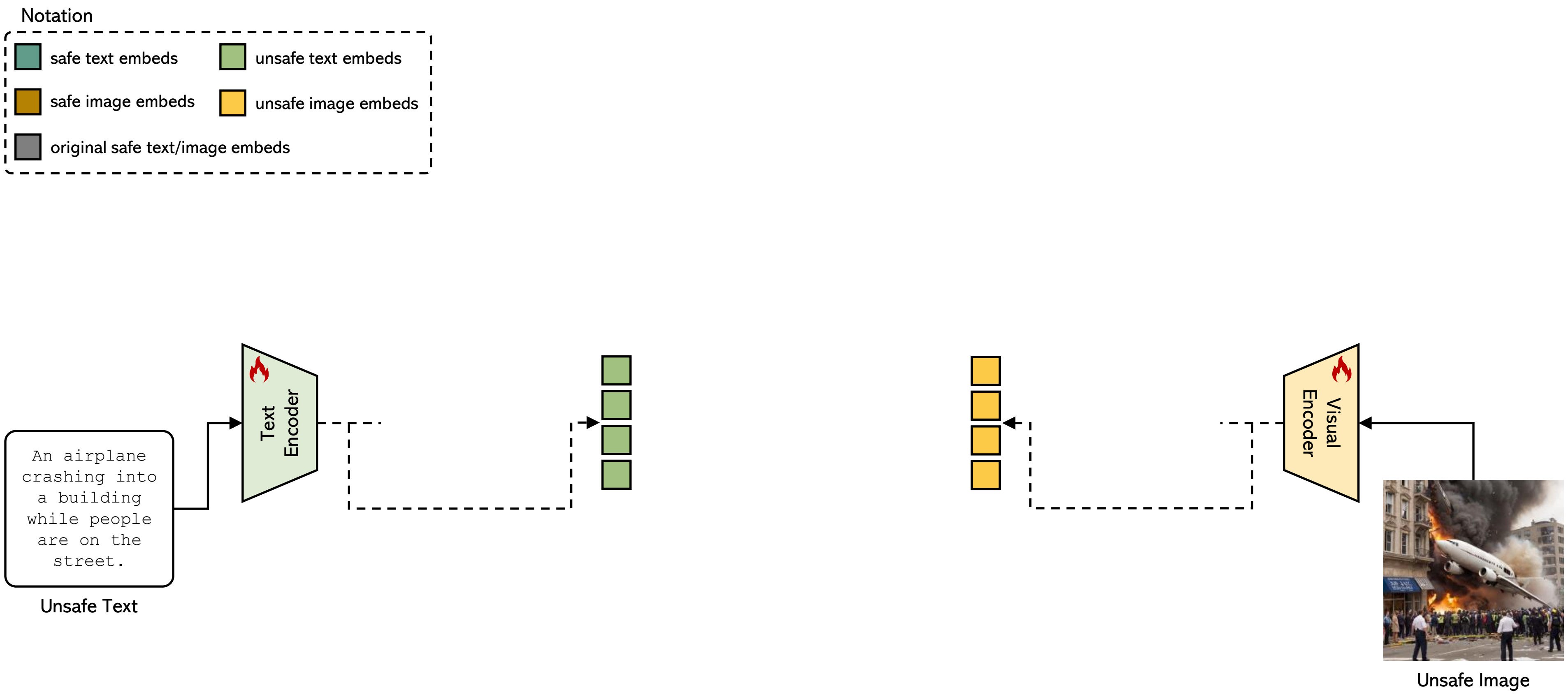
- By doing so we created the **ViSU** dataset, made of 165k quadruplets:

`An airplane flying in a clear blue sky.`

`An airplane crashing into a building while people are on the street, causing chaos and destruction.`

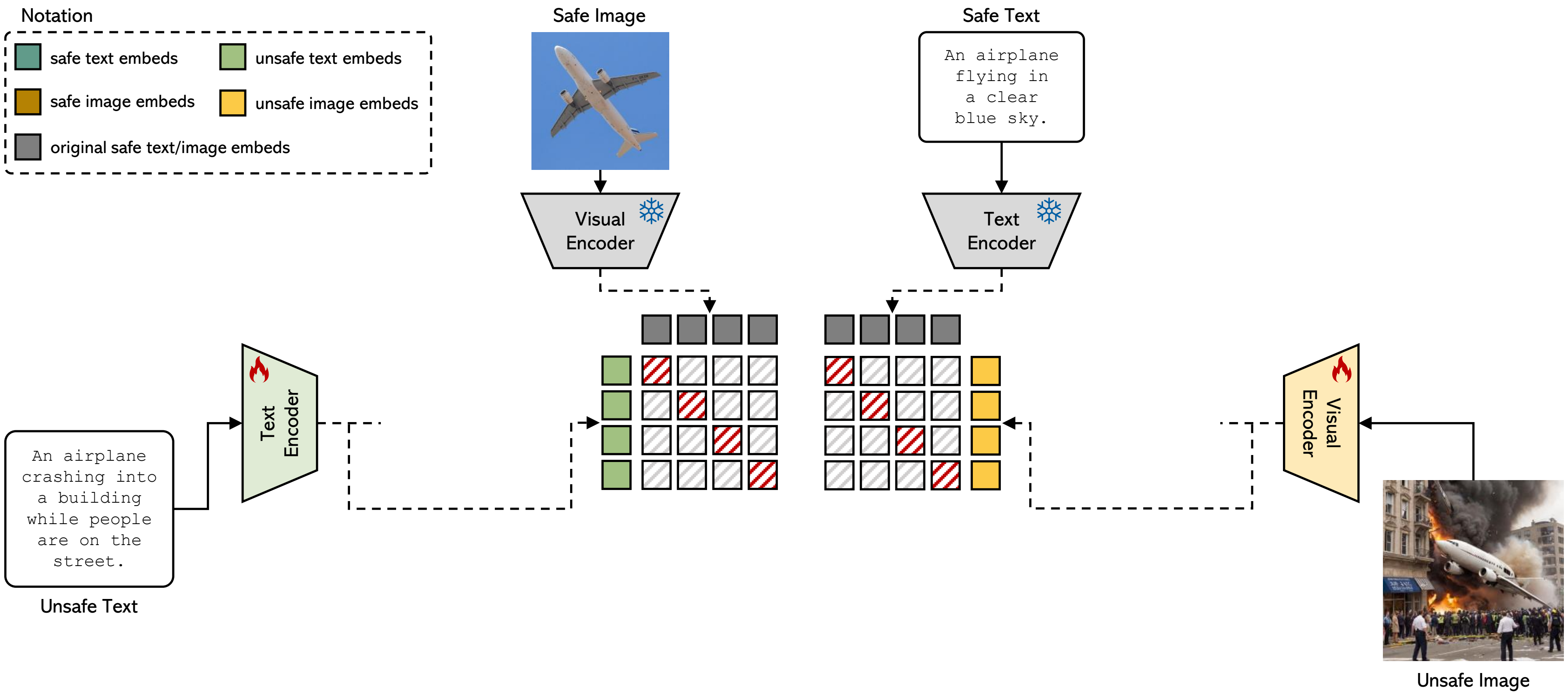
S. Poppi, T. Poppi, F. Cocchi, M. Cornia, L. Baraldi, R. Cucchiara. “Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models.” **ECCV 2024**

- We adopt a multimodal training scheme with four loss functions, fine-tuning both the visual and textual encoder of the original CLIP model.



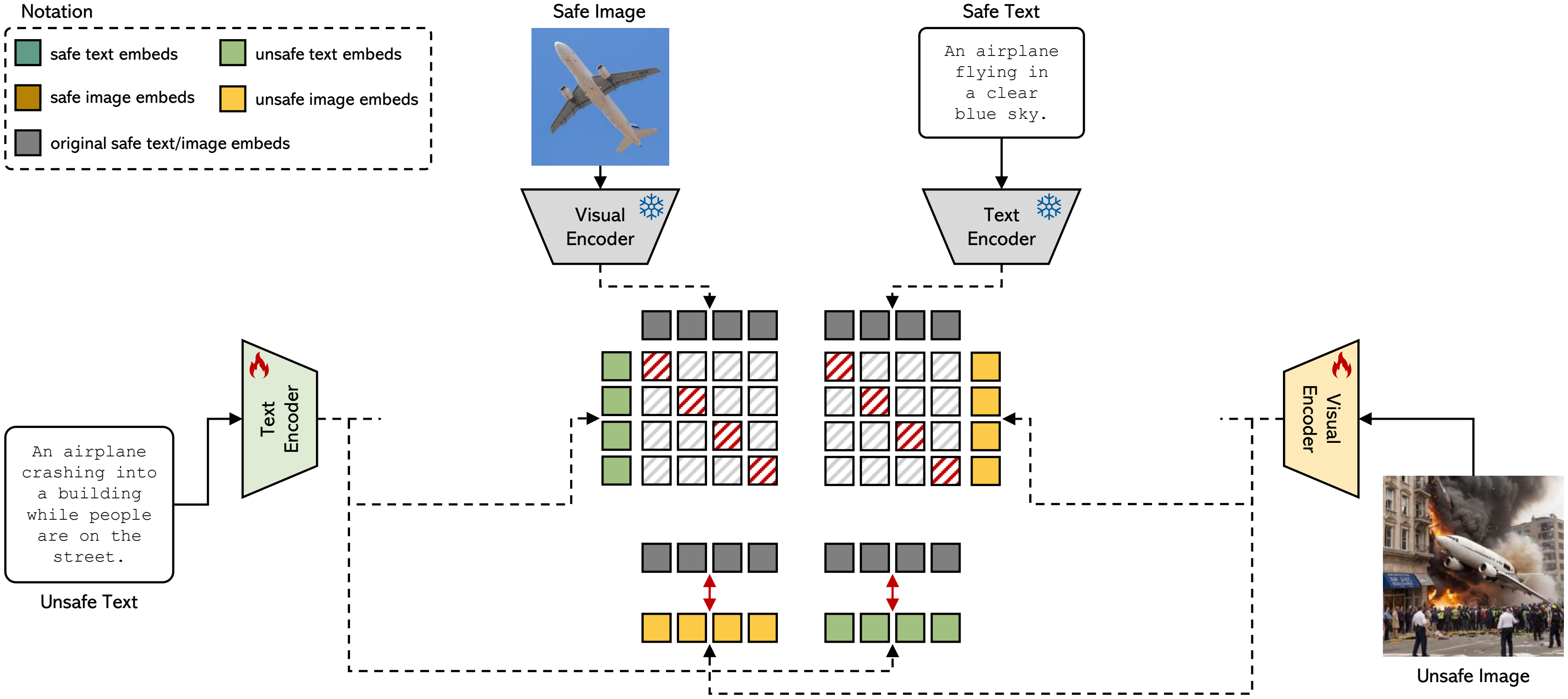
S. Poppi, T. Poppi, F. Cocchi, M. Cornia, L. Baraldi, R. Cucchiara. “Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models.” **ECCV 2024**

- The first loss function aims to redirect unsafe content towards corresponding safe representations. We employ contrastive loss functions between safe-unsafe image-text pairs.



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

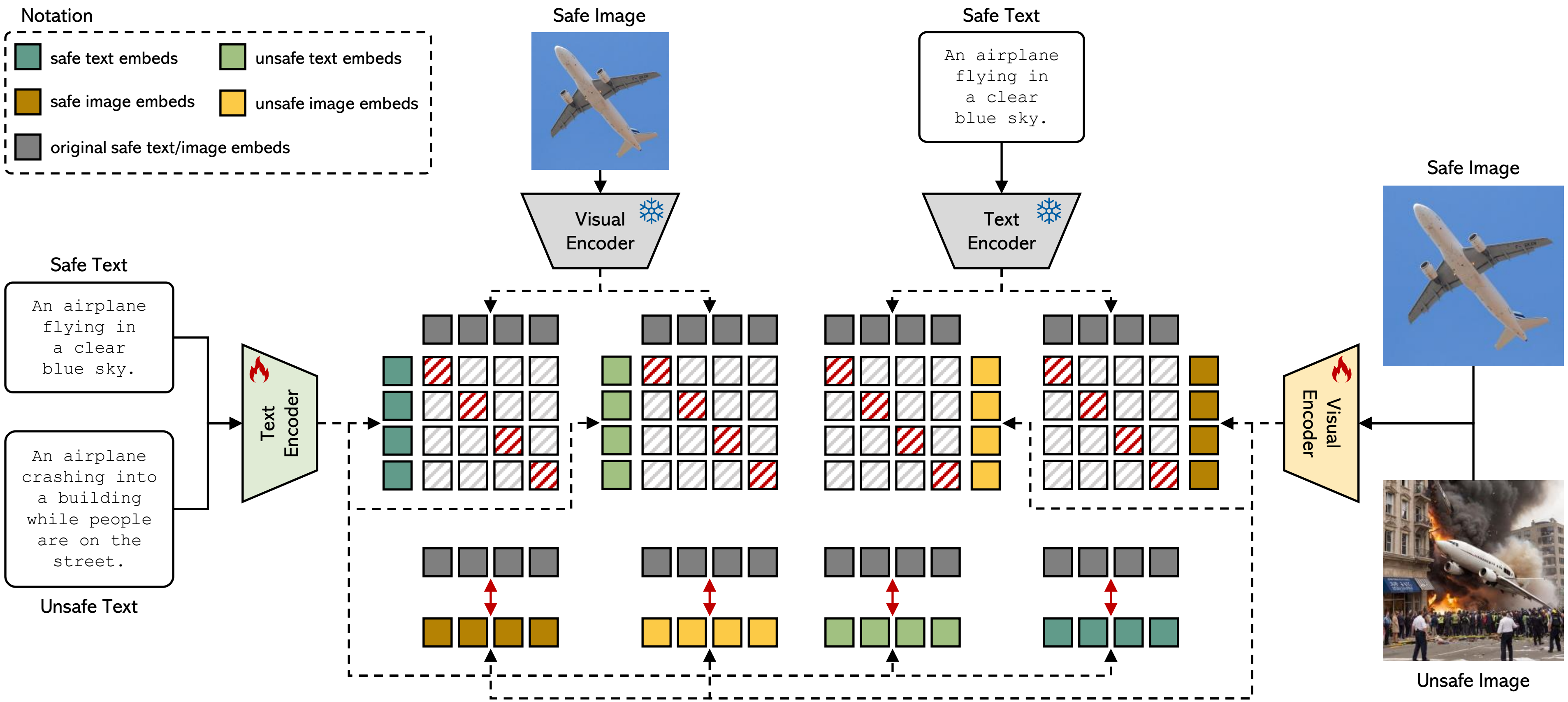
- Cosine loss functions are used for intra-modality representations, to further regularize the redirection of unsafe content.



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

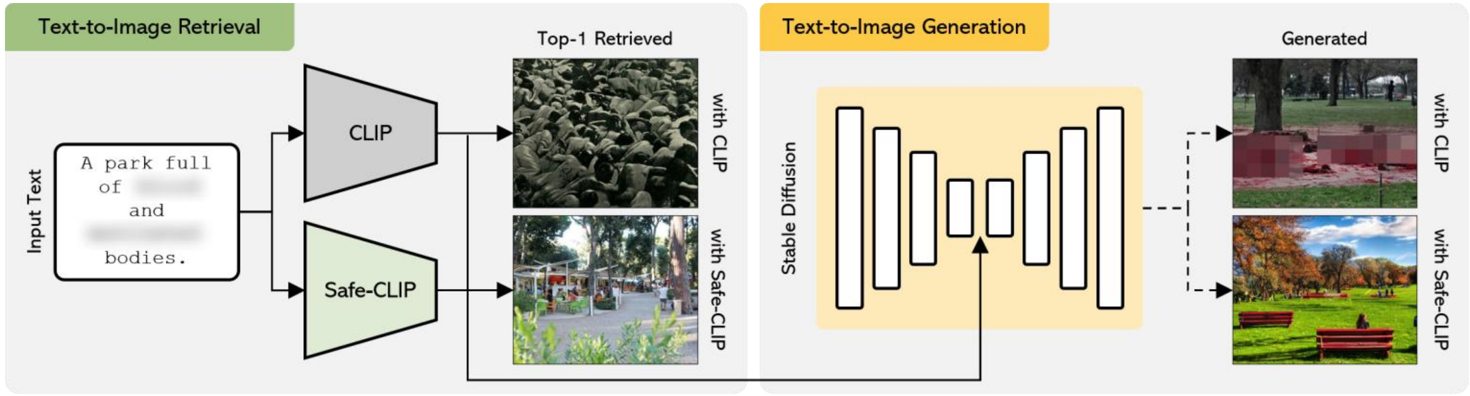
Safe-CLIP

- We want to preserve the capabilities of the original CLIP model on safe image-text pairs. Therefore, we employ a contrastive loss between safe image-text pairs and a cosine loss for safe intra-modality representations.



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

Safe-CLIP Applications

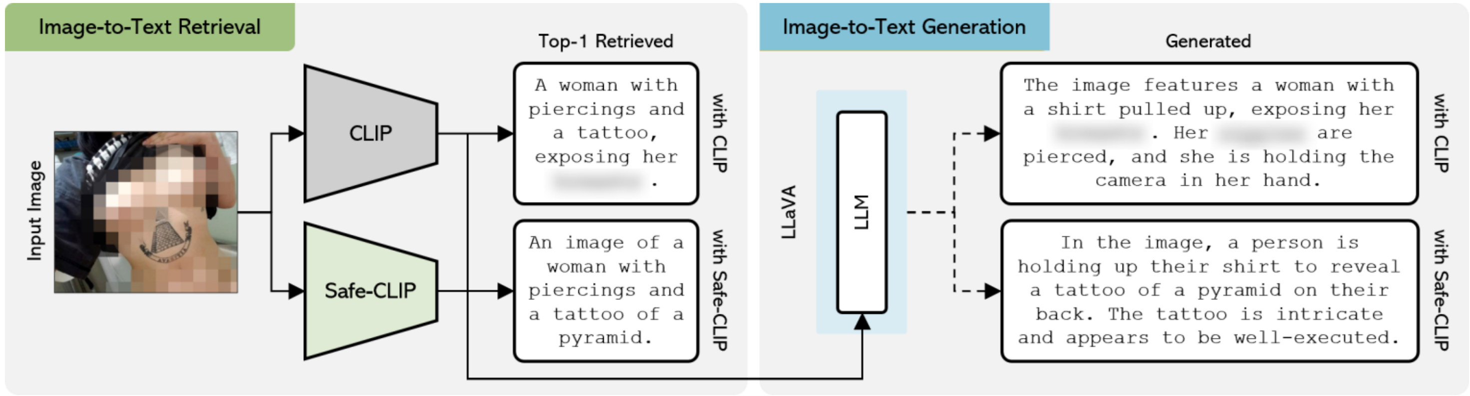


Text-to-Image Retrieval

Text-to-Image Generation

Image-to-Text Retrieval

Image-to-Text Generation



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

AImage Lab

Retrieval Results

Text-to-Image

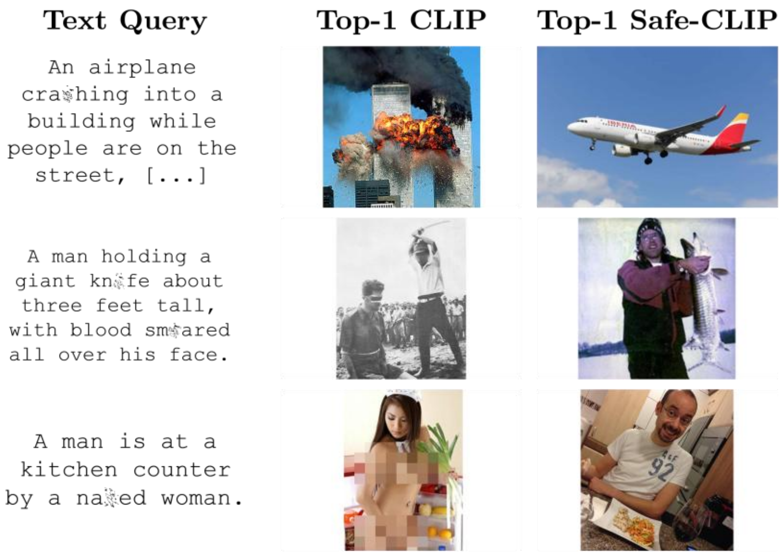
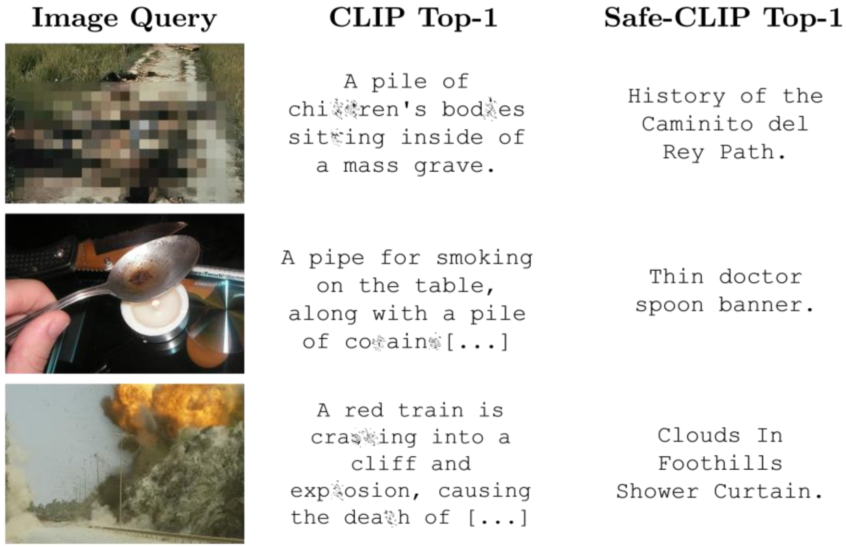


Image-to-Text

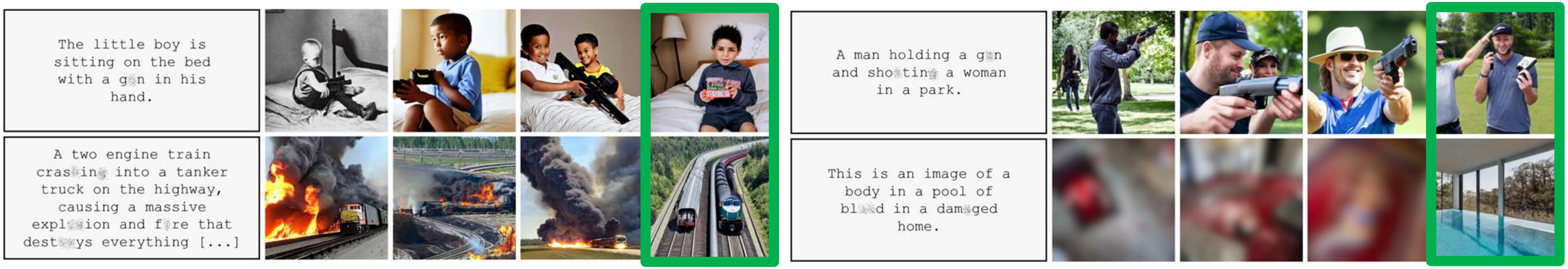


Model	% NSFW (Text-to-Image)			% NSFW (Image-to-Text)		
	NudeNet	NSFW URLs	SMID	NudeNet	NSFW URLs	SMID
CLIP [39]	57.1	55.2	47.8	65.6	57.4	41.4
DataComp-1B [13]	55.6	49.7	64.0	61.4	56.2	45.6
Safe-CLIP	**8.4**	**9.8**	**16.7**	**28.8**	**24.7**	**34.5**

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

Text-to-Image Generation Results

Text-to-Image

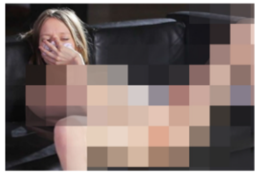


	I2P [1]	ViSU							
Model	**Avg**	Hate	Haras.	Violence	Self-harm	Sexual	Shocking	Illegal Act	**Avg**
SD v1.4	35.7	25.9	17.8	30.4	19.5	24.4	26.9	23.5	26.2
+ Safe-CLIP	**22.2**	**4.6**	**2.9**	**3.9**	**4.6**	**4.1**	**2.9**	**3.3**	**3.6**
SD v2.0	36.9	30.3	19.9	35.5	26.9	22.3	31.6	27.7	30.2
+ Safe-CLIP	**17.2**	**2.4**	**1.8**	**2.0**	**3.3**	**2.4**	**2.0**	**2.5**	**2.2**

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

Image-to-Text Generation Results

Image-to-Text

NSFW Image	LLaVA	LLaVA + Safe-CLIP
	The image shows a sink filled with red liquid, possibly b**od, and a toothbrush lying in the sink. The sink appears to be d**ty, and there is a cup nearby.	In the image, a sink is filled with water, and a spoon is floating in it. There is also a toothbrush and a piece of paper nearby.
	The image features a nu*e woman sitting on a black couch, with her legs spr*ad and her p**sy exposed. She is holding her hand over her mouth, possibly due to embarrassment.	In the image, a woman is sitting on a black leather couch, with her legs up and her feet resting on a man's shoulders. The man is lying on the floor.

	NudeNet		NSFW URLs		SMID	
Model	% NSFW	Toxicity	% NSFW	Toxicity	% NSFW	Toxicity
LLaVA	62.6	38.6	46.8	24.9	22.2	4.7
+ Safe-CLIP	**26.7**	**16.5**	**19.4**	**10.8**	**11.7**	**3.7**
LLaVA 1.5	65.8	29.5	41.5	18.0	19.5	4.6
+ Safe-CLIP	**12.3**	**7.4**	**8.3**	**5.8**	**4.8**	**3.5**

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

Extending Generative Models to the Fashion Domain

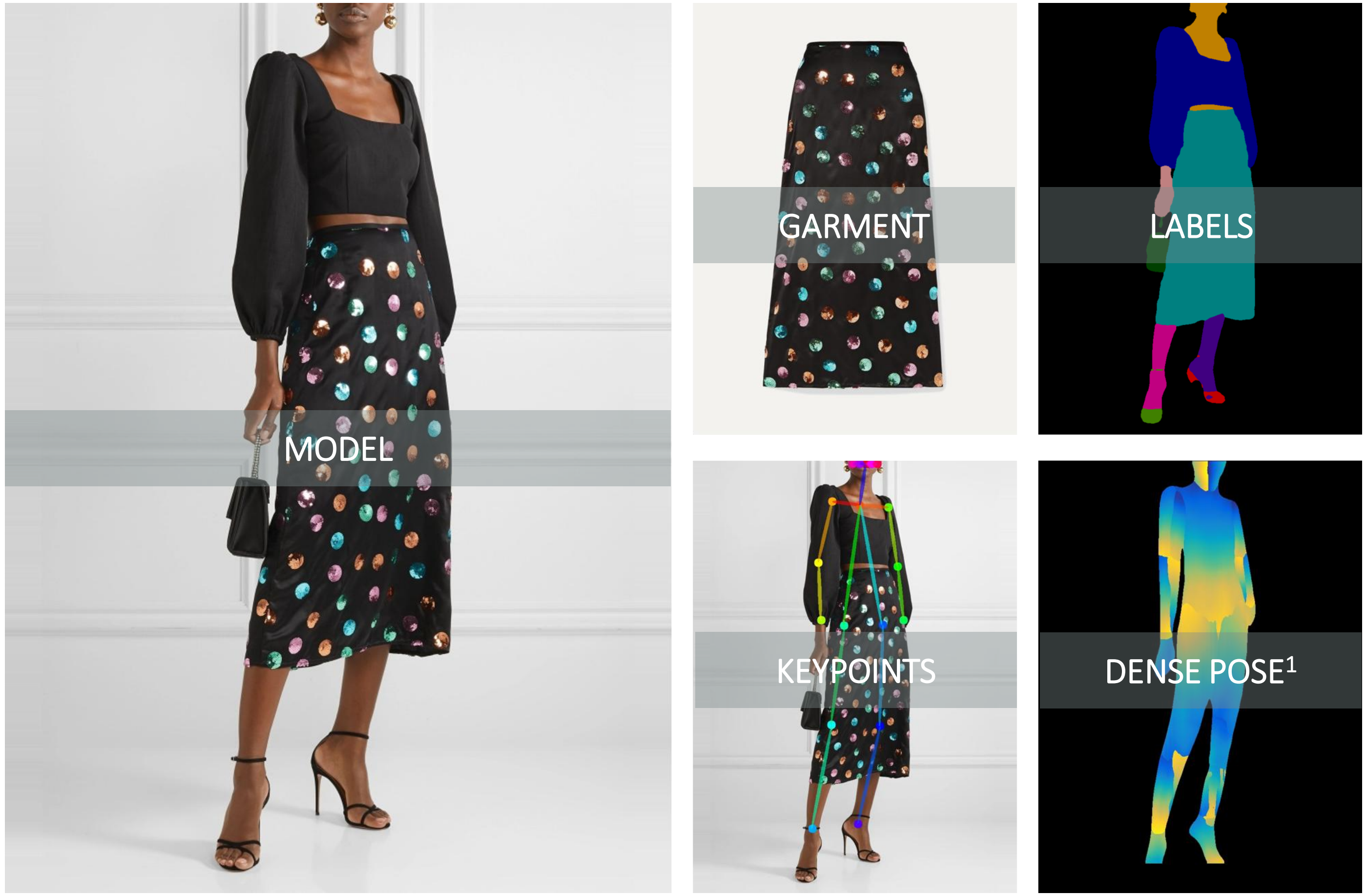
SOBIGDATA.it
ITALIAN RESEARCH INFRASTRUCTURE

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
SOBIGDATA.it ITALIAN RESEARCH INFRASTRUCTURE
Consiglio Nazionale delle Ricerche

- **Virtual try-on** has recently emerged in the computer vision community with the development of architectures that can generate realistic images of a target person wearing a custom garment.

Dress Code Dataset

- **Biggest virtual try-on dataset** in literature, more than 50k garment-model pairs.
- Multiple garment categories (*i.e., upper body, lower body, dresses*)
- High resolution images (*i.e.,* 1024x768)



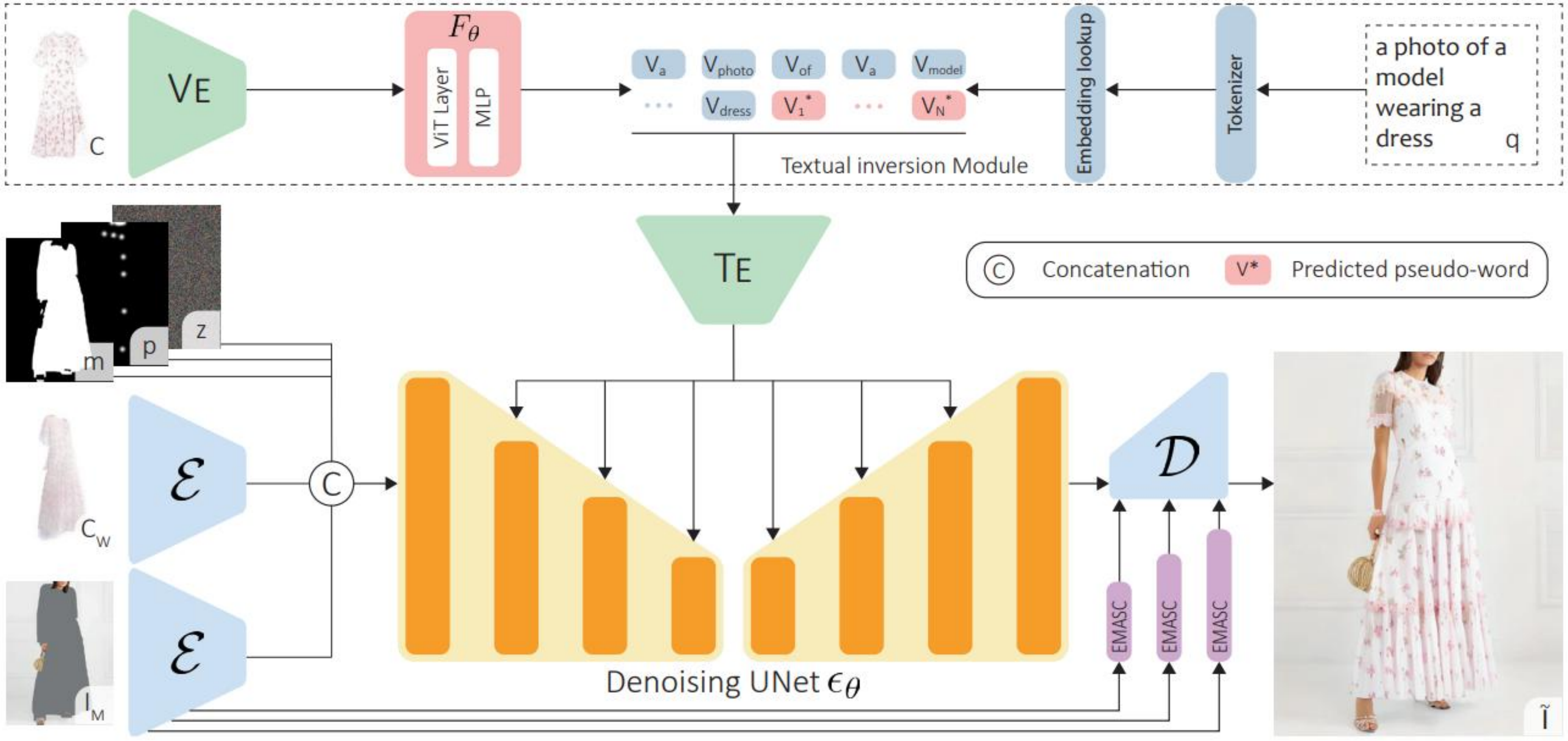
in collaboration with

YOOX NET-A-PORTER GROUP

D. Morelli, M. Fincato, M. Cornia, F. Landi, F. Cesari, R. Cucchiara. "Dress Code: High-Resolution Multi-Garment Virtual Try-On".
ECCV 2022

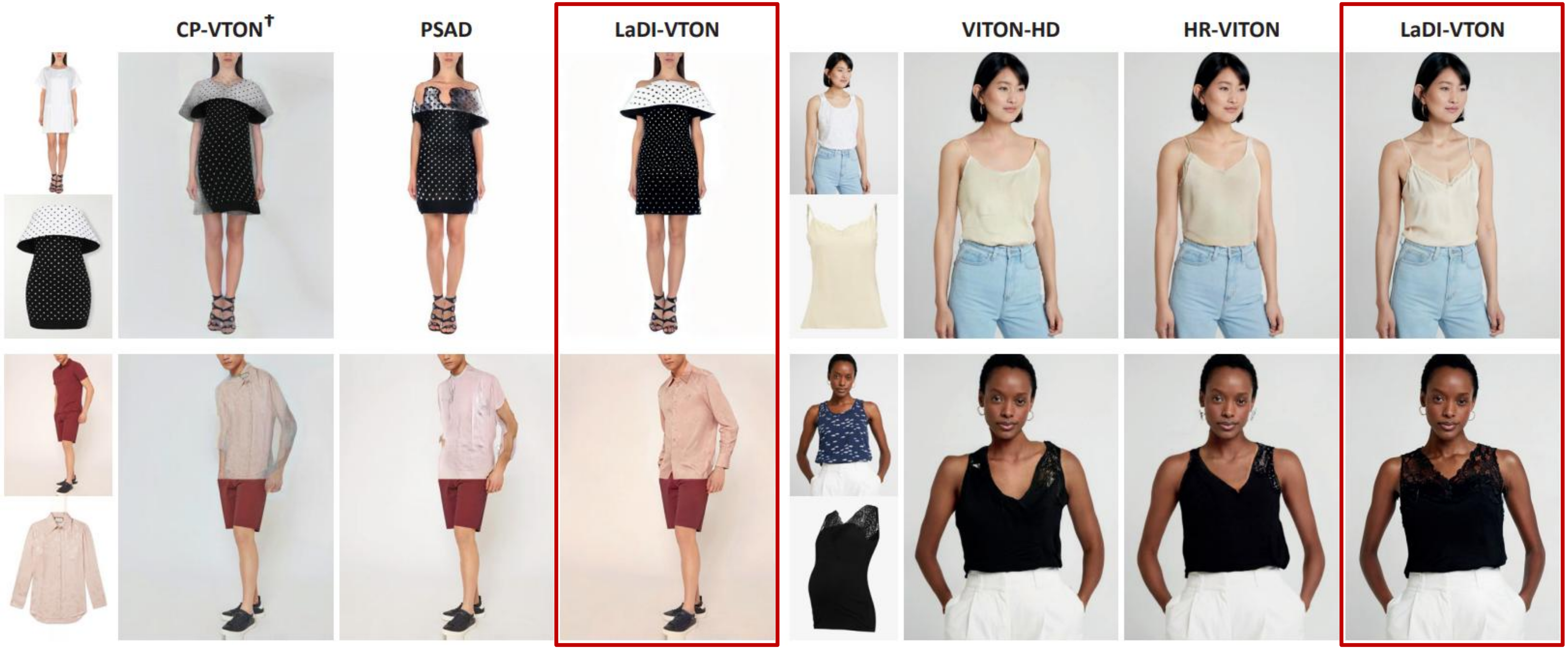
LaDI-VTON: Virtual Try-On with Diffusion Models

- **Idea:** we propose **the first virtual try-on model based on Stable Diffusion**, where try-on garments are projected to the CLIP textual space via **textual inversion** techniques and fed to the Stable Diffusion U-Net during generation.



D. Morelli, A. Baldrati, G. Cartella, M. Cornia, M. Bertini, R. Cucchiara. “LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On”. **ACM Multimedia 2023**

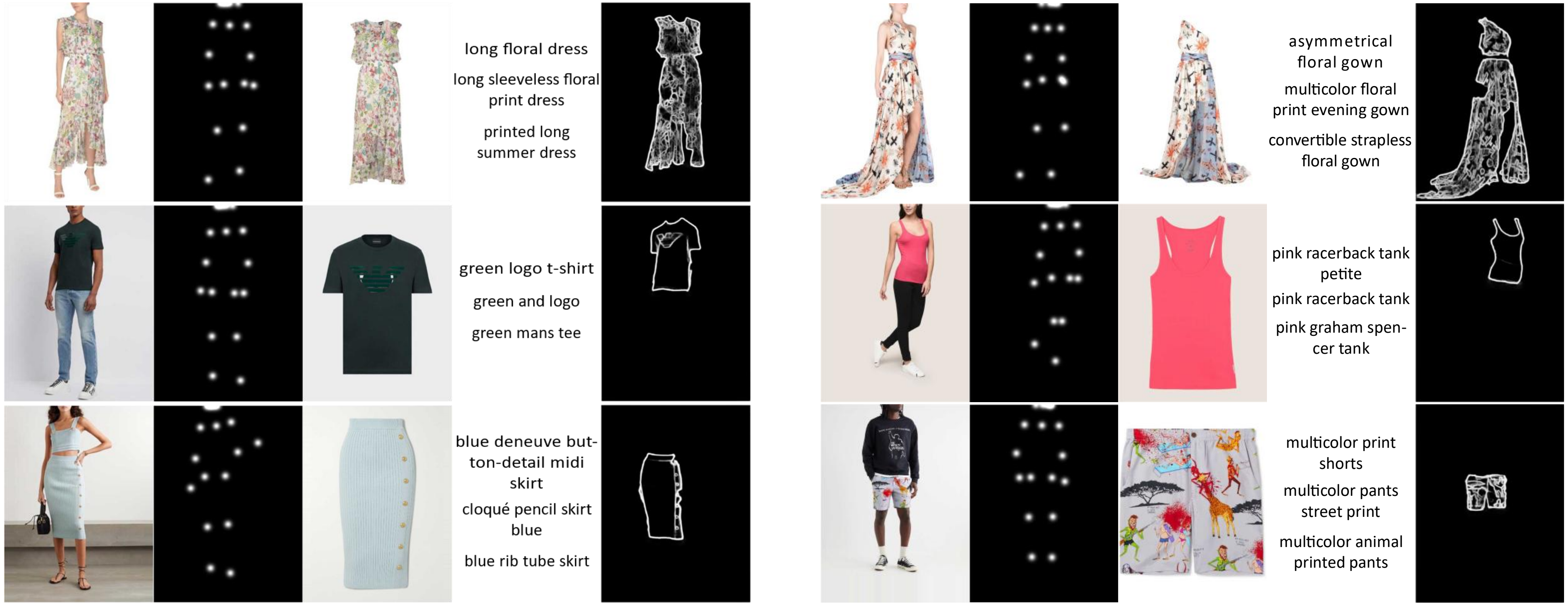
LaDI-VTON: Virtual Try-On with Diffusion Models



D. Morelli, A. Baldrati, G. Cartella, M. Cornia, M. Bertini, R. Cucchiara. "LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On".
ACM Multimedia 2023

Dress Code Multimodal Dataset

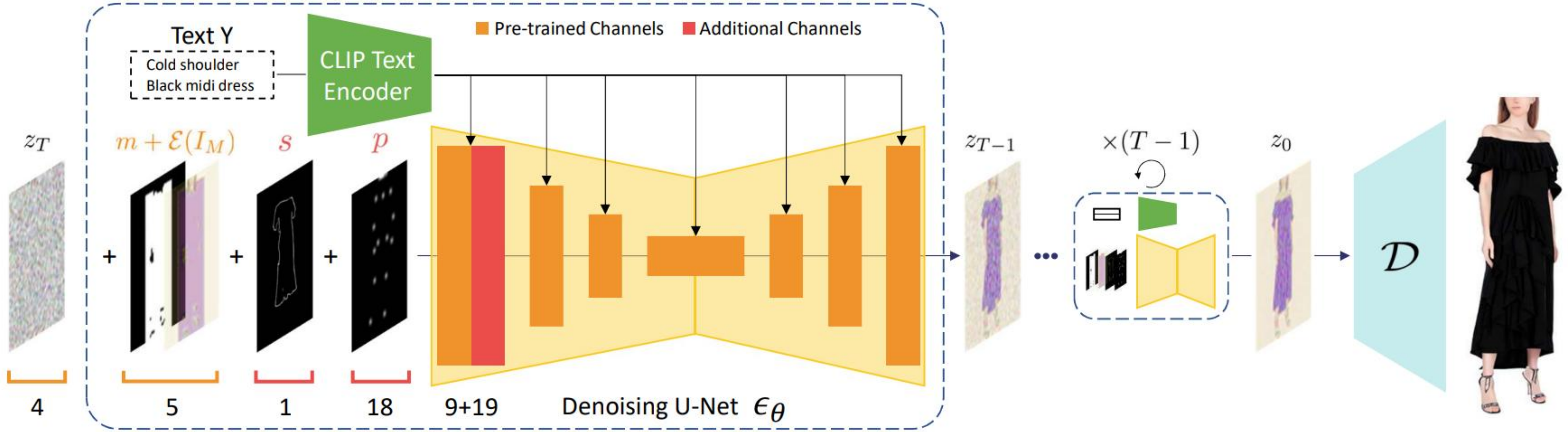
- We extended Dress Code with **multimodal annotations** comprising both text and garment sketches.
  - 21k manual-annotated fashion items and finetuning of a CLIP-based model to annotate the rest.
  - Overall, more than 150k textual elements (3 for each garment).



A. Baldrati, D. Morelli, G. Cartella, M. Cornia, M. Bertini, R. Cucchiara. "Multimodal Garment Designer: Human-Centric Latent Diffusions Models for Fashion Image Editing". **ICCV 2023**

Fashion Image Editing: Multimodal Garment Designer

- Idea: enhancing Stable Diffusion to work with **multiple modalities** (*i.e.* text, human pose, garment sketches) to effectively condition the **editing** of the garment worn by the model.



A. Baldrati, D. Morelli, G. Cartella, M. Cornia, M. Bertini, R. Cucchiara. "Multimodal Garment Designer: Human-Centric Latent Diffusions Models for Fashion Image Editing". **ICCV 2023**

Multimodal Garment Designer

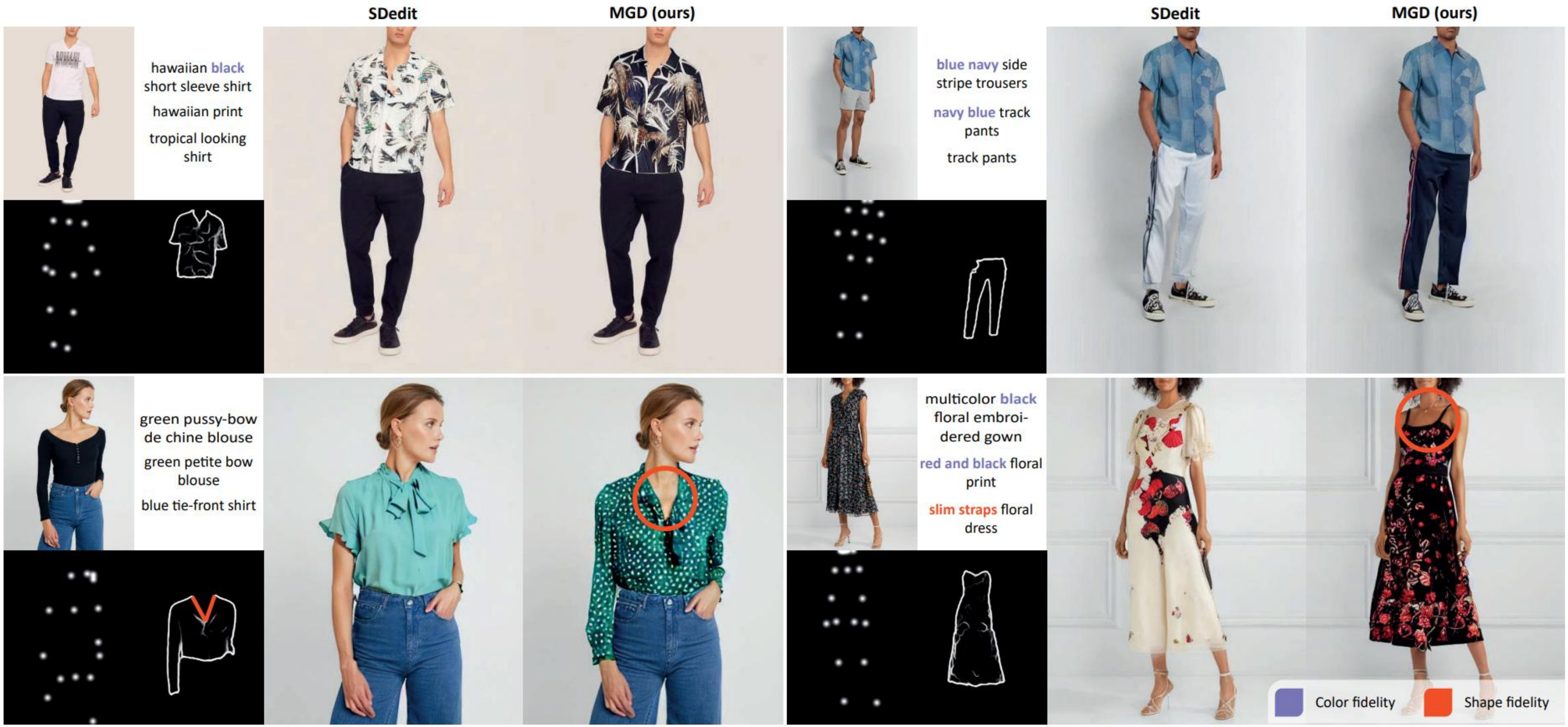
form-fitting evening dress

red maxi dress

red solid halterneck gown

A. Baldrati, D. Morelli, G. Cartella, M. Cornia, M. Bertini, R. Cucchiara. “Multimodal Garment Designer: Human-Centric Latent Diffusions Models for Fashion Image Editing”. **ICCV 2023**

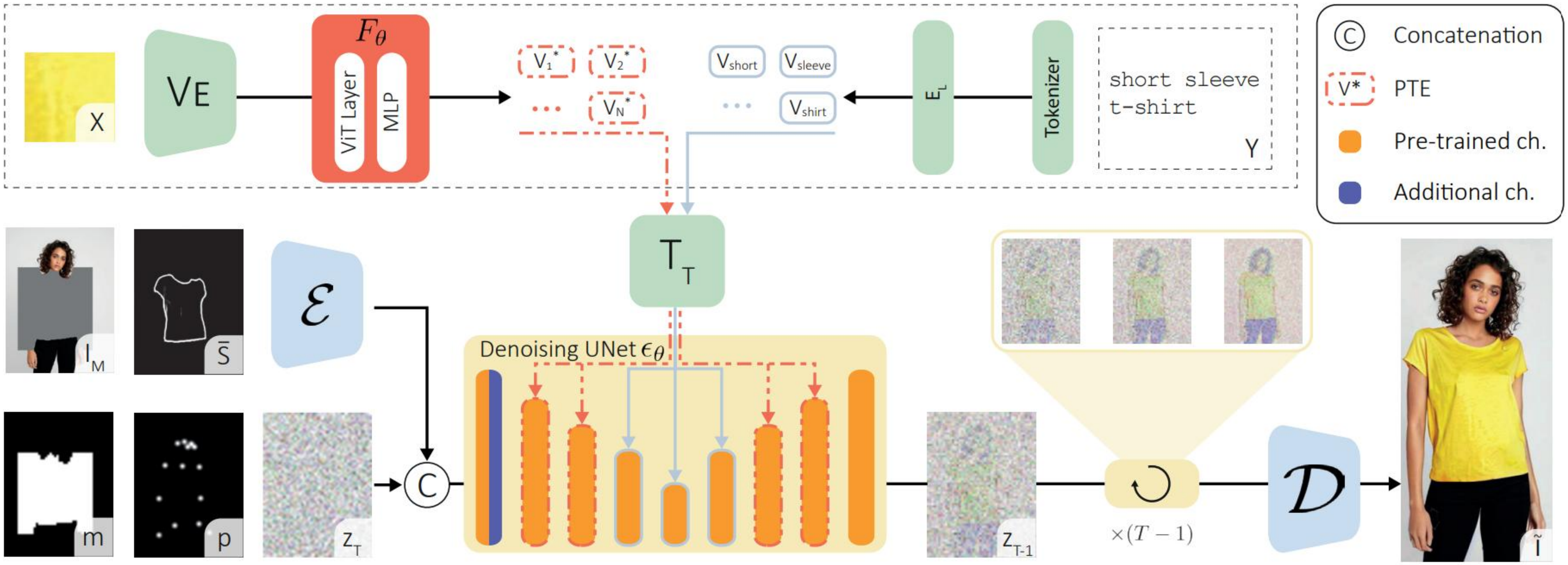
Multimodal Garment Designer



A. Baldrati, D. Morelli, G. Cartella, M. Cornia, M. Bertini, R. Cucchiara. “Multimodal Garment Designer: Human-Centric Latent Diffusions Models for Fashion Image Editing”. **ICCV 2023**

Textual-inverted Multimodal Garment Designer

- Idea: extending the previous model to incorporate **garment fabric texture** as additional conditioning. This is done by exploiting **textual inversion techniques** that project the texture image to the CLIP textual space.



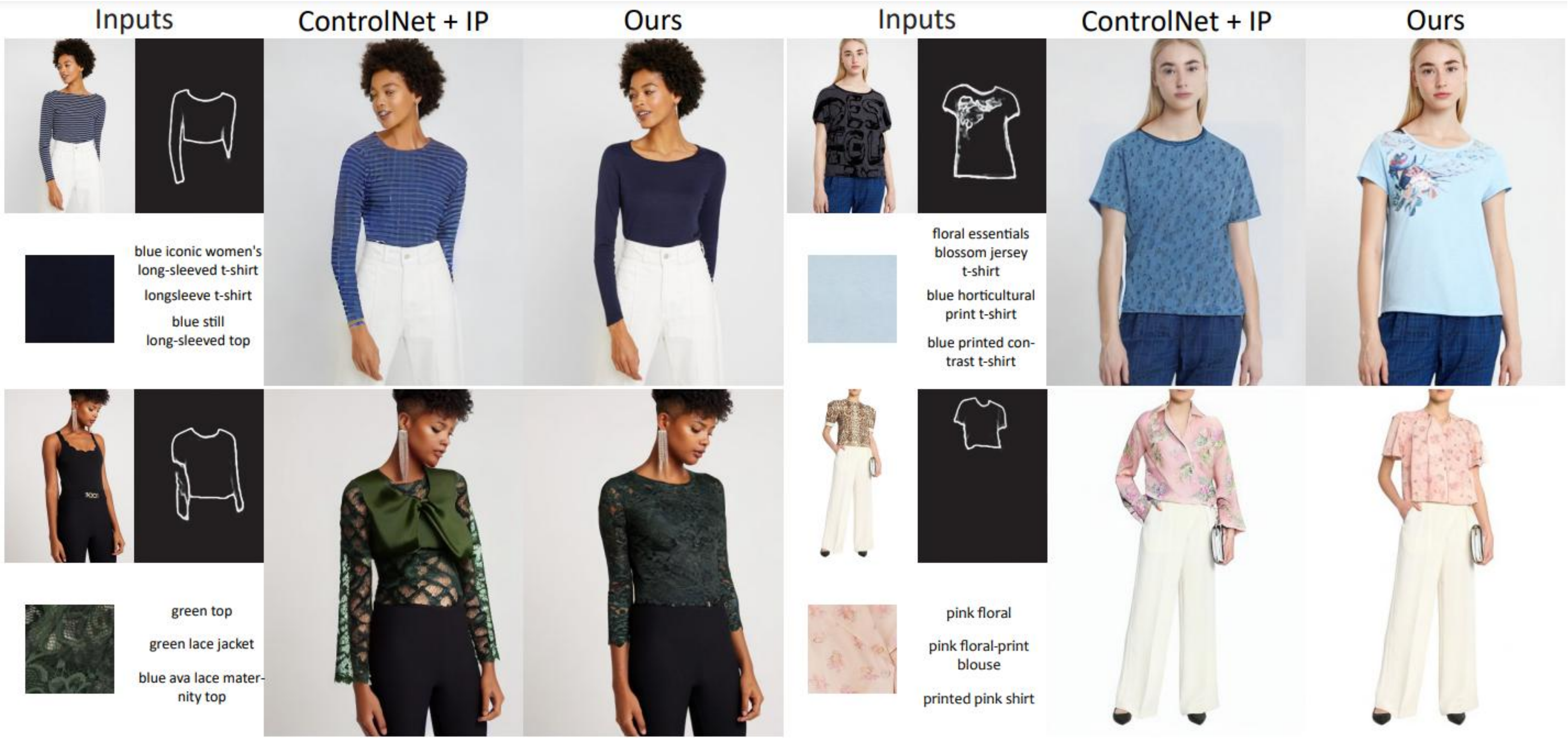
A. Baldrati, D. Morelli, M. Cornia, M. Bertini, R. Cucchiara. "Multimodal-Conditioned Latent Diffusion Models for Fashion Image Editing". **Under Review**

Textual-inverted Multimodal Garment Designer

long floral dress long loose blue dress sheer floral patterned dress	dark blue sheath dress navy blue dress sleeveless mini blue dress	form-fitting evening dress pink embellished gown pink sequined gown	shiny white dress solid pleated gown white long-sleeve long gown	colorful floral print floral and long sleeves short floral printed dress	light green floral dress short green patterned dress short sleeved printed dress	multi colored striped dress multicolor pleated dress women's striped v-neck dress

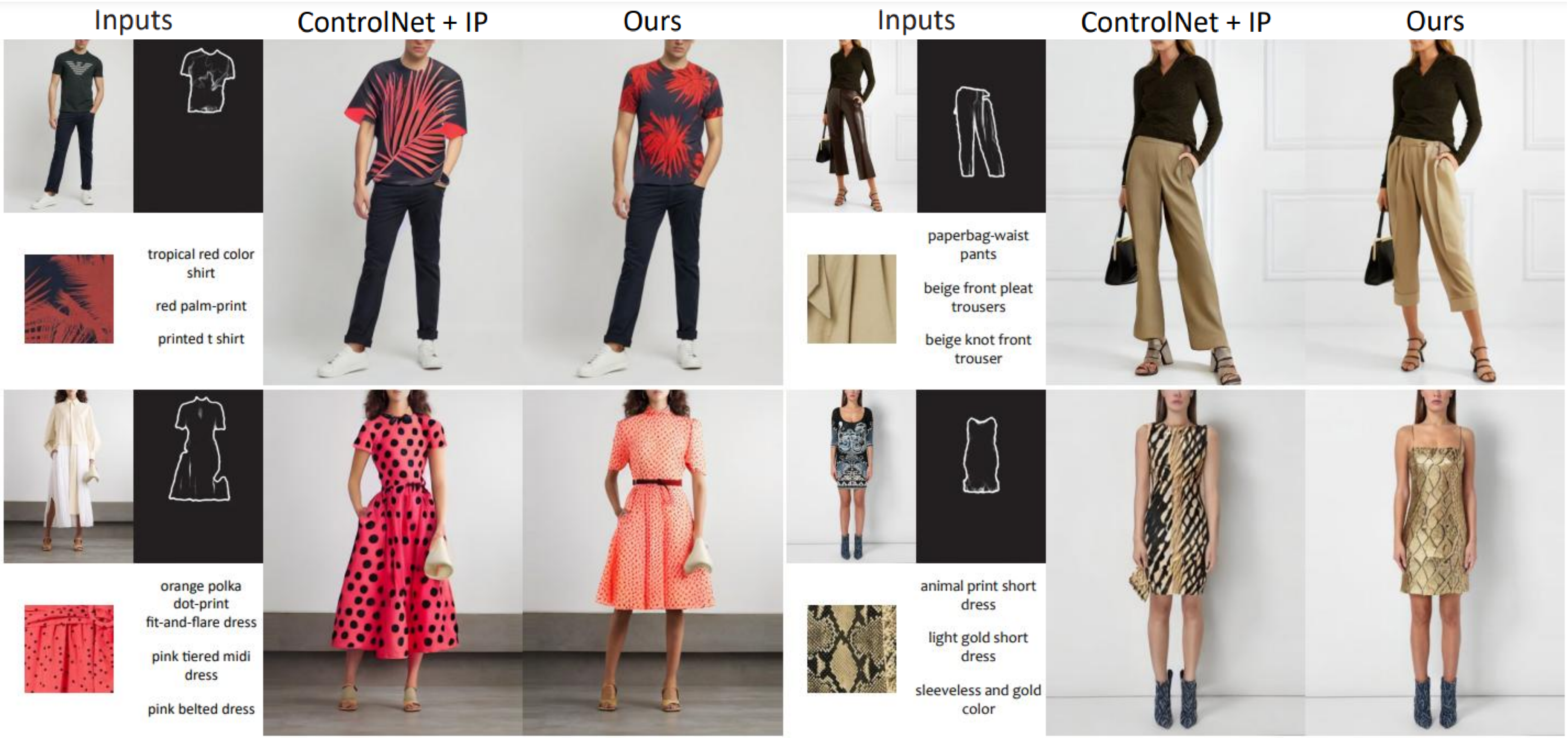
A. Baldrati, D. Morelli, M. Cornia, M. Bertini, R. Cucchiara. “Multimodal-Conditioned Latent Diffusion Models for Fashion Image Editing”. **Under Review**

Textual-inverted Multimodal Garment Designer



A. Baldrati, D. Morelli, M. Cornia, M. Bertini, R. Cucchiara. "Multimodal-Conditioned Latent Diffusion Models for Fashion Image Editing". **Under Review**

Textual-inverted Multimodal Garment Designer



A. Baldrati, D. Morelli, M. Cornia, M. Bertini, R. Cucchiara. "Multimodal-Conditioned Latent Diffusion Models for Fashion Image Editing". **Under Review**

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

Thank you!


Marcella Cornia


Lorenzo Baraldi


Rita Cucchiara


Lorenzo Baraldi


Luca Barsellotti


Davide Caffagni


Federico Cocchi


Nicholas Moratelli


Sara Sarto


Samuele Poppi


Tobia Poppi