Sara **Sarto**

PhD Student · Research Scientist · Computer Vision

Modena, ITALY

(+39) 3661239610 | sara.sarto@unimore.it | sarasarto | sara-sarto-4520a4209 | Sara Sarto

Summary

I am a PhD Student with 3+ years experience on deep learning architectures and tools. My current research focuses on cutting-edge **multimodal architectures** and their integration with **advanced retrieval techniques**. I have extensive experience with the **NLP** taks, **foundation models**, such as CLIP, and have worked extensively with **vision-and-language** architectures primarily focusing on their **evaluation** and addressing the problem of **hallucination**. Recently, I have been working on training and development of **multimodal large language models**, including LLaVA and its derivatives.

Education

PhD Student in Artificial Intelligence and Computer Vision

Modena, Italy

AImageLab, University of Modena and Reggio Emilia — *2022 - Present*

- Computer Vision & Deep Learning | Multimodal LLM and Foundation Models | Vision & Language Evaluation Metrics | RAG
- Supervisor: Prof. Rita Cucchiara

M.S in Artificial Intelligence

Modena, Italy

University of Modena and Reggio Emilia — *2020 - 2022*

- Thesis title: Retrieval-Augmented Transformer for Image Captioning – Final grade: 110/110 cum laude

B.S. in Computer Science and Engineering

Modena, Italy

University of Modena and Reggio Emilia — *2017 - 2020*

Work Experience

Applied Scientist Intern

London, UK

Amazon, Prime Video Team — *November 2024 - May 2025*

- Focus on Multimodal Foundation Models (Video, Text and Audio) and merging modalities.

PhD Student

Modena, Italy

AImageLab, University of Modena and Reggio Emilia — *November 2022 - November 2025*

- Research on Multimodal tasks (Image Captioning, Multimodal Large Language Models applications) using deep learning techniques.
- Integration of Retrieval-augmentation into Multimodal Models (CLIP, LLaVa, etc)
- Focus on Multimodal Foundation Models Evaluation and understanding of the Hallucination problem.

Research Scientist

Modena, Italy

AImageLab, University of Modena and Reggio Emilia — *July 2022 - Novemeber 2022*

- Research on Vision & Language task using Deep Learning techniques. Image Captioning architectures with Retrieval.

Research Intern

Modena, Italy

AImageLab, University of Modena and Reggio Emilia — *January 2022 - July 2022*

- Development of multi-modal architectures for the integration of visual and textual features, based on contrastive and/or regressive learning.
- Conducting experiments with tensor computing libraries in distributed queueing computing environment.

Publications

Recurrence-Enhanced Vision-and-Language Transformers for Robust Multimodal Document Retrieval

CVPR

Davide Caffagni*, **Sara Sarto***, M. Cornia, L. Baraldi, R. Cucchiara — *2025*

- An approach that allows for multimodal queries – composed of both an image and a text – and can search within collections of multimodal documents. To allow for multi-level and cross-modal understanding and feature extraction, we employ a novel Transformer-based recurrent cell that integrates both textual and visual features at different layers.

LLaVA-MORE : A Comparative Study of LLMs and Visual Backbones for Enhanced Visual Instruction Tuning

ICCV Workshop

Federico Cocchi*, Nicholas Moratelli*, Davide Caffagni*, **Sara Sarto***, M. Cornia, L. Baraldi, R. Cucchiara — *2025*

- A new family of MLLMs that integrates recent language models with diverse visual backbones.

Image Captioning Evaluation in the Age of Multimodal LLMs: Challenges and Future Perspectives

IJCAI

Sara Sarto, M. Cornia, R. Cucchiara — *2025*

- An overview of image captioning evaluation, analyzing metric evolution, strengths, and limitations. We assess challenges from longer MLLM-generated captions and the adaptability of current metrics.

Semantically Conditioned Prompts for Visual Recognition under Missing Modality Scenarios

WACV

V. Pipoli, F. Bolelli, **Sara Sarto**, M. Cornia, L. Baraldi, C. Grana, R. Cucchiara, E. Ficarra — *2025*

- A paper on multimodal prompting for visual recognition. The model exploits the semantic representation of available modalities to query a learnable memory bank, which allows the generation of prompts based on the semantics of the input.

Positive-Augmented Contrastive Learning for V&L Evaluation and Training

IJCV

Sara Sarto, N. Moratelli, M. Cornia, L. Baraldi, R. Cucchiara — *2025*

- A contrastive-based evaluation metric for image captioning, not only used as simple metrics but integrated into the fine-tuning stage of a captioning model resulting in semantically richer captions with fewer repetitions and grammatical errors.

BRIDGE: Bridging Gaps in Image Captioning Evaluation with Stronger Visual Cues

ECCV

Sara Sarto, M. Cornia, L. Baraldi, R. Cucchiara — *2024*

- A learnable and reference-free image captioning metric that employs a novel module to map visual features into dense vectors and integrates them into multi-modal pseudo-captions. It demonstrate stronger alignment with human judgement and ability to detect hallucinated objects.

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

CVPR Workshop

D. Caffagni*, F. Cocchi*, N. Moratelli*, **Sara Sarto***, M. Cornia, L. Baraldi, R. Cucchiara — *2024*

- Integration of an external knowledge source of documents in a Multimodal Large Language Model through a hierarchical retrieval approach.

The REvolution of Multimodal Large Language Models: A Survey

ACL Findings

D. Caffagni*, F. Cocchi*, L.Barsellotti*, N. Moratelli*, **Sara Sarto***, L.Baraldi*, M. Cornia, L. Baraldi, Cucchiara — *2024*

- A comprehensive review of recent visual-based MLLMs, analyzing their architectural choices, multimodal alignment strategies, and training techniques.

Towards Retrieval-Augmented Architectures for Image Captioning

ACM TOMM

Sara Sarto, M. Cornia, L. Baraldi, R. Cucchiara — *2024*

- A novel approach towards developing image captioning models that utilize an external kNN memory to improve the generation process.

Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation

CVPR

Sara Sarto, M. Barraco, M. Cornia, L. Baraldi, R. Cucchiara — *2023*

- Highlight Paper.
- A contrastive-based evaluation metric for image captioning, that in a novel way unifies the learning of a contrastive visual-semantic space with the addition of generated images and text on curated data.

With a Little Help from your own Past: Prototypical Memory Networks for Image Captioning

ICCV

M. Barraco*, **Sara Sarto***, M. Cornia, L. Baraldi, R. Cucchiara — *2023*

- A network which perform attention over activations obtained while processing other training samples, through a prototypical memory model.

Retrieval-Augmented Transformer for Image Captioning

CBMI

Sarto Sara, M. Cornia, L. Baraldi, R. Cucchiara — *2022*

- Best Paper Award
- An image captioning approach with a kNN memory, with retrieval from an external corpus to aid the generation process.

Video Surveillance and Privacy: A Solvable Paradox?

Computer Society

R. Cucchiara, L. Baraldi, M. Cornia, **Sarto Sara** — *2023*

Honors & Awards

2025	**Doctoral Consortium**, received at CVPR 2025.	*Nashville, Texas*
2024	**Travel Award for Workshop "Women in Computer Vision"**, received for the paper: "BRIDGE: Bridging Gaps in Image Captioning Evaluation with Stronger Visual Cues".	*ECCV, Milan*
2023	**Best Poster Award**, received for the poster: "Augmented Architectures for Vision and Language"	*Summer School VISMAC, Italy*
2023	**Travel Award for Workshop "Women in Computer Vision"**, received for the paper: "With a Little Help from your own Past: Prototypical Memory Networks for Image Captioning".	*ICCV, France*
2023	**Master Thesis Award**, Premio alla Memoria Davide Rabotti	*Modena, Italy*
2023	**Travel Award for Workshop "Women in Computer Vision"**, received for the paper: "Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation"	*CVPR, Canada*
2022	**Best Student Paper Award**, received for the paper: "Retrieval-augmented Transformer for Image Captioning"	*CBMI, Austria*

Reviewer Activities

2025	**International Conference on Computer Vision**, ICCV
2025	**Conference on Computer Vision and Pattern Recognition**, CVPR
2024	**European Conference on Computer Vision**, ECCV
2024	**ACM Multimedia** , ACM MM
2024	**International Conference on Geometric Modeling and Processing**, GMP
2023	**Pattern Recognition Letters**, PRL

Skills

Programming and CV	Python, Pytorch, OpenCV, LaTeX
Research Tools	Wandb, GitHub, SLURM
Languages	English, Italian, Spanish