CVPR Nashville JUNE 11-15, 2025

Augmenting Multimodal LLMs with Self-Reflective Tokens for Knowledge-based VQA

Federico Cocchi*[1,2], Nicholas Moratelli*[1],
Marcella Cornia[1], Lorenzo Baraldi[1], Rita Cucchiara[1].
University of Modena and Reggio Emilia, Italy
[1]name.surname@unimore.it [2]name.surname@phd.unipi.it

AImageLab
UNIMORE

Motivation

- Retrieval-augmented generation boosts LLM performance in specialized domains by integrating external knowledge.
- Highly specific visual questions are often difficult to answer accurately. Leveraging multimodal external knowledge enhances understanding capacity of the model.
- InfoSeek and Encyclopedic VQA use a Wikipedia subset as an external knowledge base.

Our Proposal: ReflectiVA

- **Zero-shot models** (LLMs and MLLMs) fall short on complex and specific Visual Question Answering tasks.
- A **multimodal retrieval** pipeline is employed to selects the most relevant knowledge passages, starting from a visual query.
- ReflectiVA, based on the image and the question provided, can **independently decide whether retrieving external knowledge** is necessary. Moreover, when external content is used, it can also **evaluate the relevance** of that information — determining whether it is useful for answering the question or should be treated as noise and not considered.
- The model's dictionary has been expanded by adding **four special tokens**: ([RET] - [NORET]) to control the need for retrieval, and ([REL] - [NOREL]) to assess the relevance of the retrieved passage.
- **Preservation of the performance** on multimodal datasets that do not require external knowledge.

MLLMs with External Knowledge

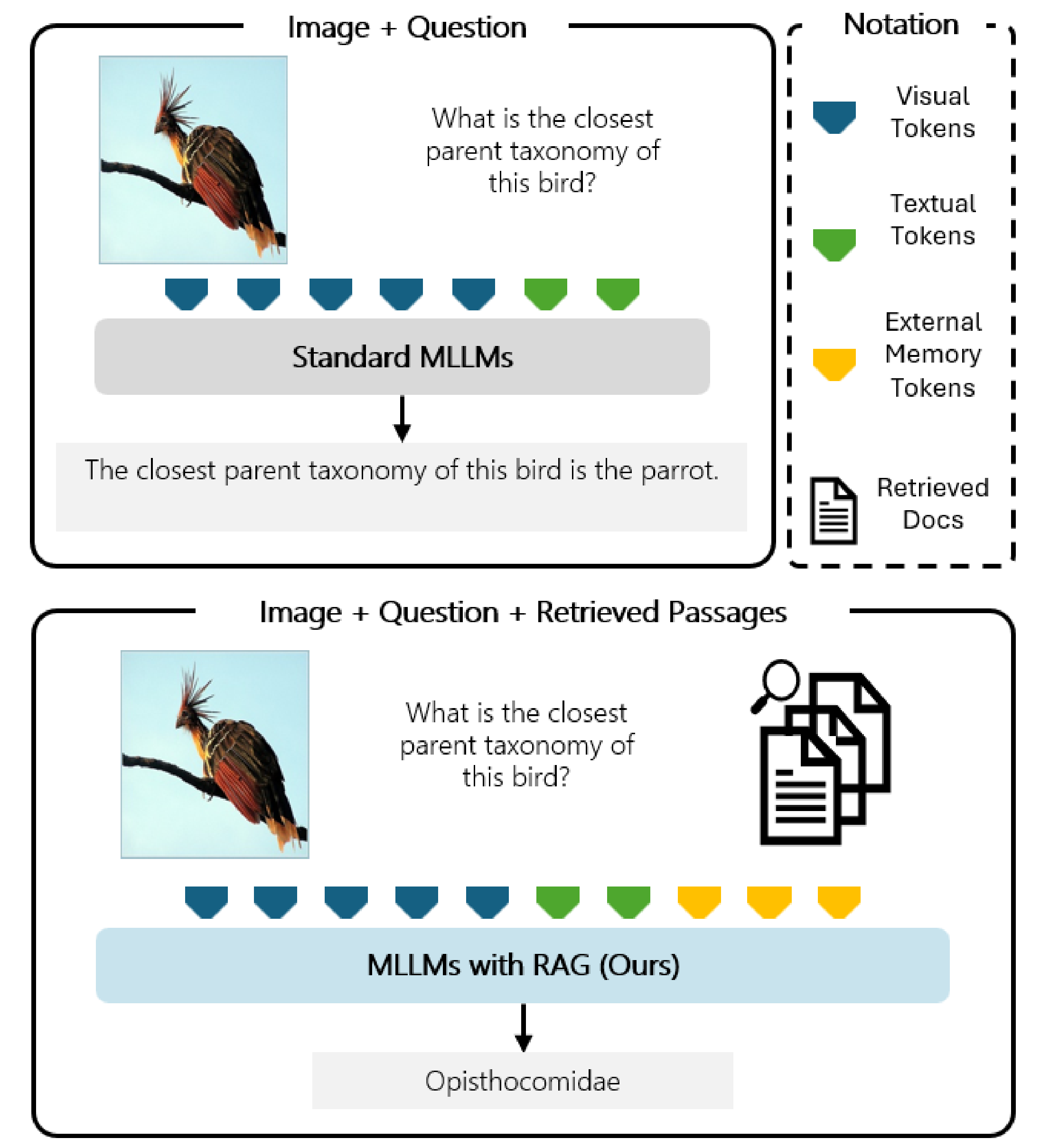
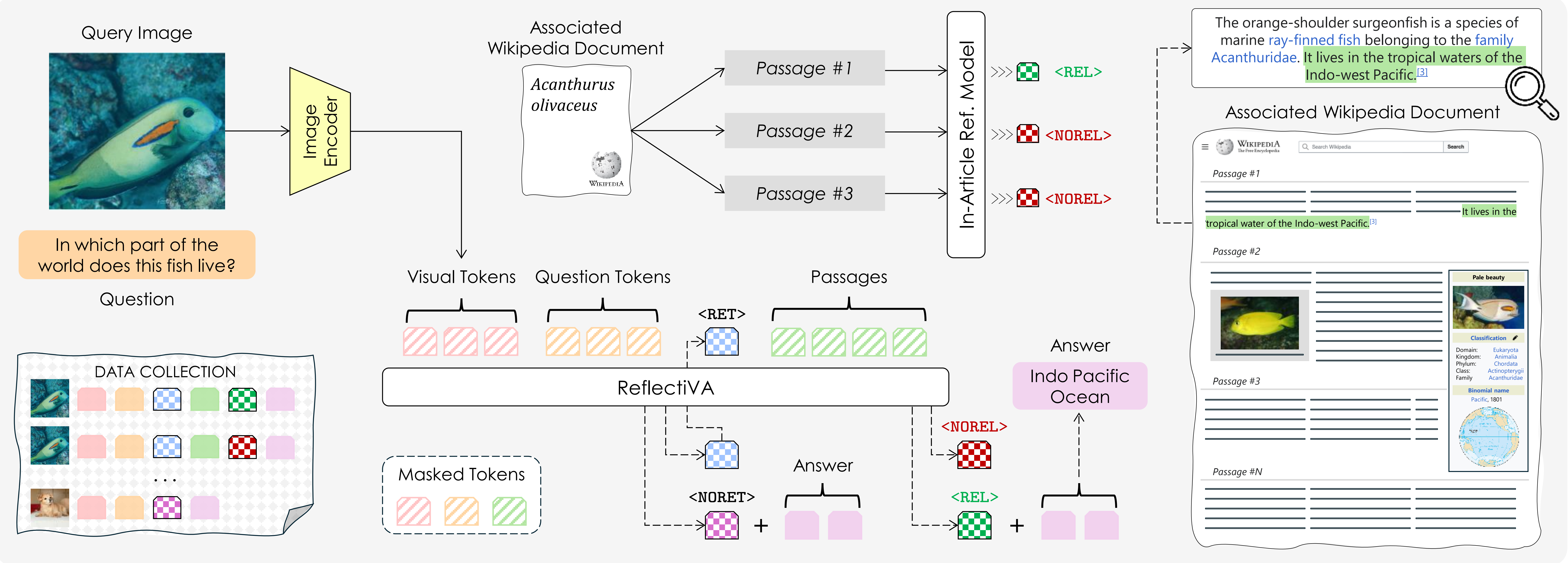


Image + Question
What is the closest parent taxonomy of this bird?
Standard MLLMs
The closest parent taxonomy of this bird is the parrot.
Notation
Visual Tokens
Textual Tokens
External Memory Tokens
Retrieved Docs
Image + Question + Retrieved Passages
What is the closest parent taxonomy of this bird?
MLLMs with RAG (Ours)
Opisthocomidae


Query Image
In which part of the world does this fish live?
Question
Image Encoder
Associated Wikipedia Document
Acanthurus olivaceus
Wikipedia
Passage #1
Passage #2
Passage #3
In-Article Ref. Model
<REL>
<NOREL>
<NOREL>
The orange-shoulder surgeonfish is a species of marine ray-finned fish belonging to the family Acanthuridae. It lives in the tropical waters of the Indo-west Pacific.[3]
Associated Wikipedia Document
Passage #1
It lives in the tropical water of the Indo-west Pacific.[3]
Passage #2
Pale beauty
Classification
Passage #3
Passage #N
DATA COLLECTION
Visual Tokens
Question Tokens
<RET>
Passages
ReflectiVA
Answer
Indo Pacific Ocean
Masked Tokens
<NORET>
Answer
<NOREL>
<REL>

Qualitative Results

Q: What is one of the traditional uses of this plant?

Wiki-LLaVA:
Food ✗
EchoSight:
Promote wound healing ✗
ReflectiVA (Ours):
Astringent ✓

Q: What is the parent organization of this building?

Wiki-LLaVA:
National Park Service ✗
EchoSight:
National Register of Historic Places ✗
ReflectiVA (Ours):
Colonial Williamsburg Foundation ✓

Q: Who designed this palace?

Wiki-LLaVA:
Johann Von Fischer ✗
EchoSight:
A team of architects, including Johan Dientzenhofer ✗
ReflectiVA (Ours):
Balthasar Neumann ✓

Q: Which road, railway or canal does this river carry?

Wiki-LLaVA:
Alp Railway ✗
EchoSight:
Railway ✗
ReflectiVA (Ours):
Albula Railway ✓

Comparison with the State of the Art

		E-VQA		InfoSeek		
Model	**LLM**	Single-Hop	All	Unseen-Q	Unseen-E	All
Zero-shot						
Text-only	Vicuna-7B	2.1	2.0	0.3	0.0	0.0
Text-only	LLaMA-3.1-8B	16.5	16.6	2.1	0.0	0.0
LLaVA-v1.5	Vicuna-7B	16.3	16.9	9.6	9.4	9.5
LLaVA-v1.5	LLaMA-3.1-8B	16.0	16.9	8.3	8.9	7.8
Retrieval-Augmented Models						
DPR_{V+T}	Multi-passage BERT	29.1	-	-	-	12.4
RORA-VLM	Vicuna-7B	-	20.3	25.1	27.3	-
Wiki-LLaVA	Vicuna-7B	17.7	20.3	30.1	27.8	28.9
Wiki-LLaVA	LLaMA-3.1-8B	18.3	19.6	28.6	25.7	27.1
EchoSight	LLaMA-3.1-8B	26.4	24.9	30.0	30.7	30.4
ReflectiVA (Ours)	LLaMA-3.1-8B	**35.5**	**35.5**	**40.4**	**39.8**	**40.1**

Preservation on Datasets without Knowledge Base

Model	**LLM**	**MMMU**	**MMB (EN)**	**POPE**	**SEED-Img**	**MME (P)**	**MME (C)**	**GQA**	**TextVQA**	**Science-QA**	**AI2D**
LLaVA-v1.5	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)	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)	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

[1] Caffagni et al. *Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs.* In CVPR W, 2024.
[2] Caffagni et al. *The Revolution of Multimodal Large Language Models: A Survey.* In ACL, 2023.
[3] Liu et al. *Visual Instruction Tuning.* NeurIPS, 2023.
[4] Chen et al. *Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions?.* In EMNLP, 2023.
[5] Mensink et al. *Encyclopedic VQA: Visual questions about detailed properties of fine-grained categories.* In ICCV, 2023.