

Grounding Language Models for Visual Entity Recognition



Vicente Ordonez
Sept 30th, 2024

Zilin Xiao¹



Ming Gong²



Paola Cascante-Bonilla¹



Xingyao Zhang²



Jie Wu²



Vicente Ordonez¹



Introduction

➤ Task: Visual Entity Recognition

- Input: an image and a query question
- Output: a Wikipedia entity from a pre-defined set



Text Query

Who manufactured the plane?



McDonnell Douglas

McDonnell Douglas was a major American aerospace manufacturing corporation and defense contractor formed by ...



What piece of equipment is placed on the animal in the image?

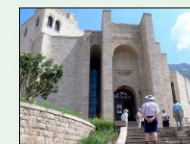


Bridle

A bridle is a piece of equipment used to direct a horse. As defined in the Oxford English Dictionary, the "bridle" includes both the headstall that...



What is this building called?



Skanderbeg Museum

The National History Museum "Gjergj Kastrioti Skënderbeu" (Albanian: Muzeu Historik Kombëtar), also known as the Skanderbeg Museum...

Input

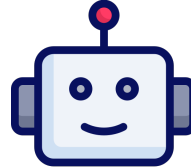
Output

Introduction

- How it differs with Visual Question Answering?
 - Answers grounded to Wikipedia Entities.



What is the **model** of this aircraft?



Boeing 767

The Boeing 767 is an American wide-body airliner developed and manufactured by Boeing Commercial Airplanes. The aircraft was launched...

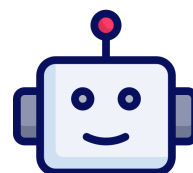
Introduction

➤ How it differs with Visual Question Answering?

- Answers grounded to Wikipedia Entities.



What is the **model** of this aircraft?



Boeing 767

The Boeing 767 is an American wide-body airliner developed and manufactured by Boeing Commercial Airplanes. The aircraft was launched...

- Fine-grained Recognition Capability Required.



Boeing 717



Douglas DC-8



Boeing 777



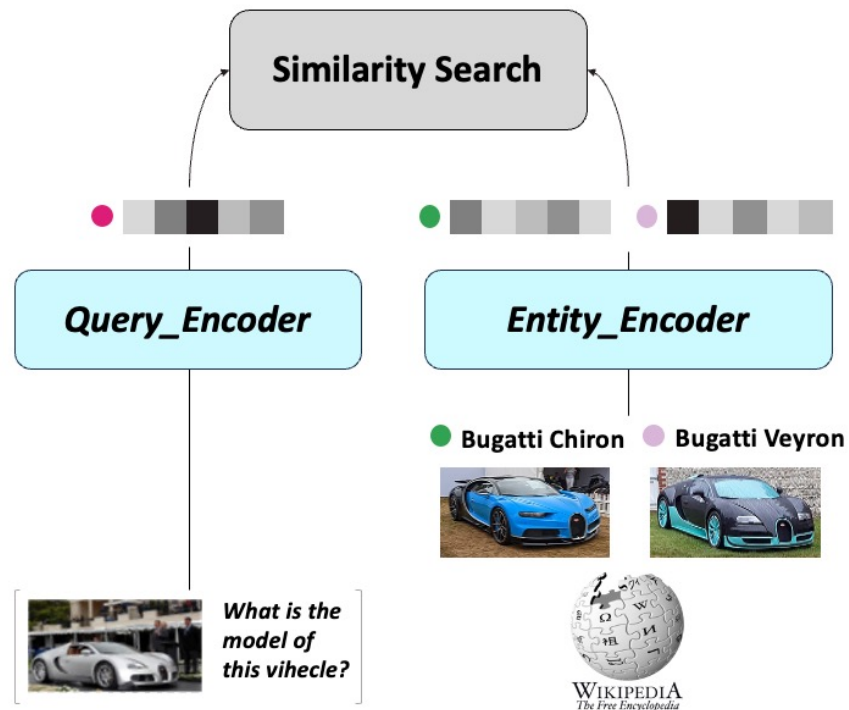
Airbus A320



Boeing 767

Prior Approach: Similarity Search

- Similarity search: cross-modal retrieval
 - Problem: hard to reason about spatial relations



What is the yellow item on top of truck?



Surfboard

A surfboard is a narrow plank used in surfing. Surfboards are relatively light, but are strong enough to support an individual standing on them while riding an ocean wave...

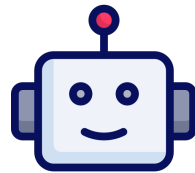
Prior Approach: Multimodal LLM

- Multimodal LLM: LLM with Visual Instruction Tuning
 - Problem 1: sometimes hallucinates a lot
 - Problem 2: natural language response grounding to Wikipedia entities are not trivial



Which **category** of bird is shown in the image?

The bird in the image is likely to be an **Australasian darter**...



Heron

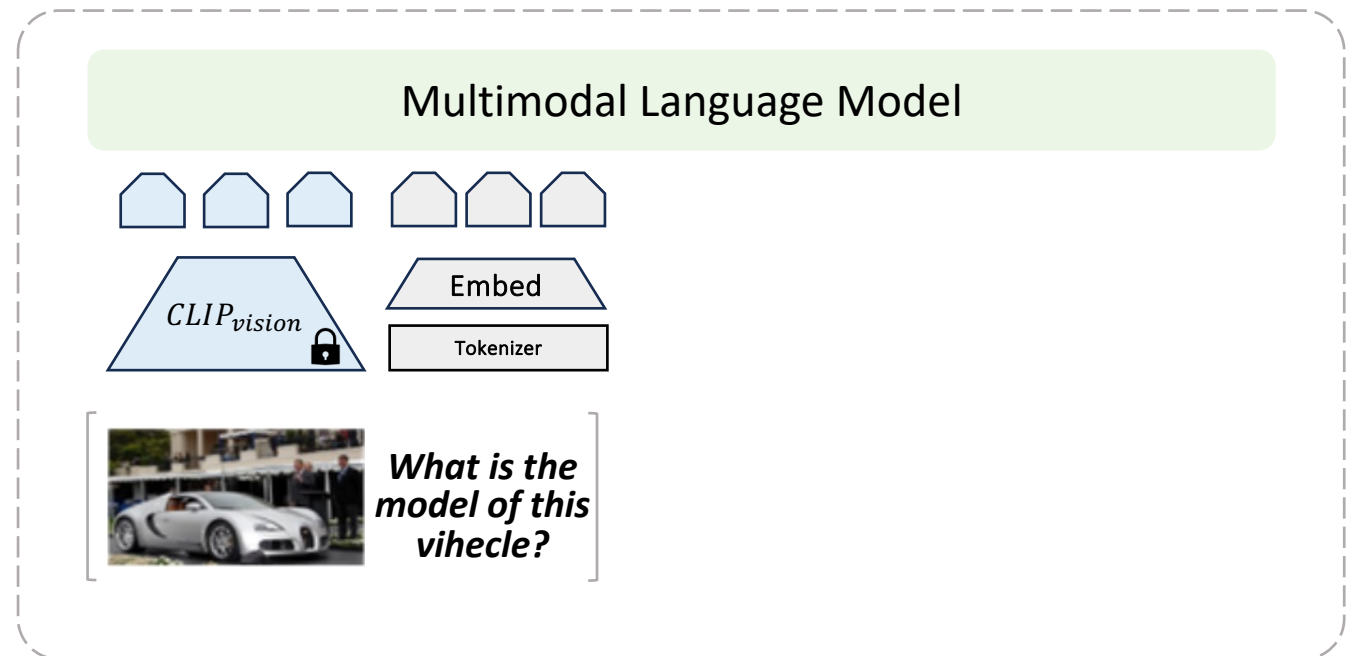
Herons are long-legged, long-necked, freshwater and coastal birds in the family Ardeidae, with 72 recognized species, some of which are referred...

Our Approach

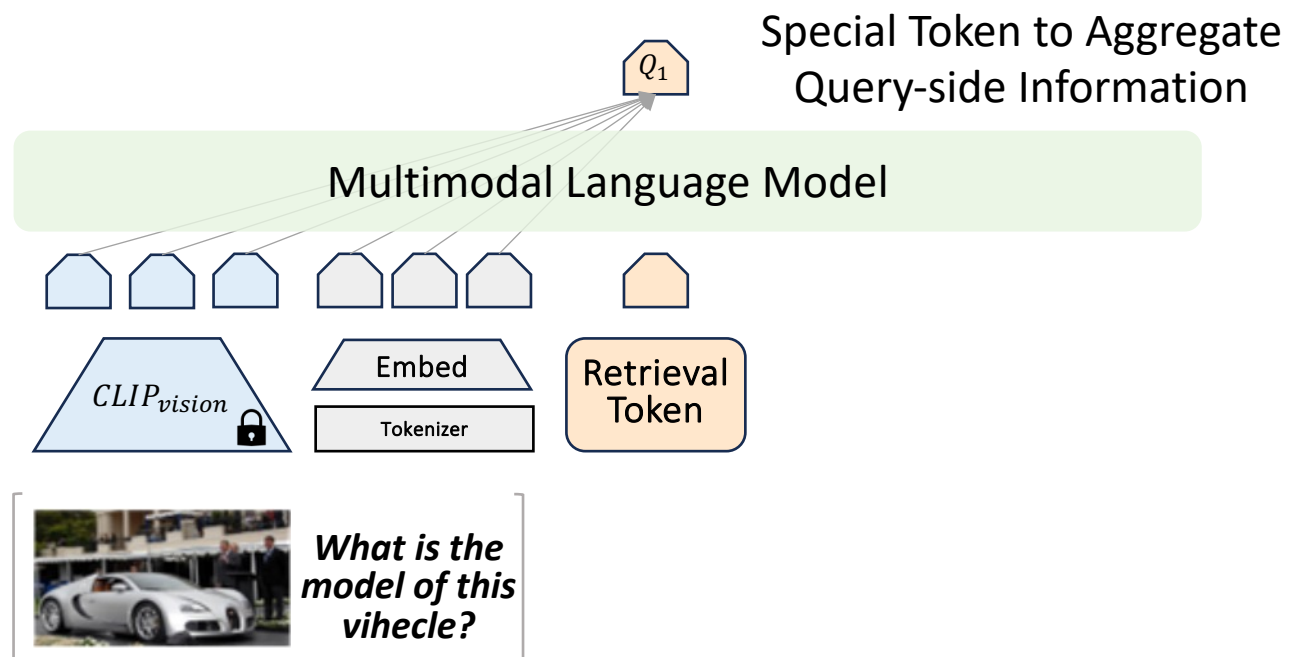
- Can we take advantage of both similarity search and MLLM inherent reasoning ability?
 - Yes!
- We can use similarity search to do a coarse **retrieval**, narrowing down the candidates.
- Then ask an Multimodal LLM to further “**re-ranking**” the candidates.

Our Approach: AutoVER

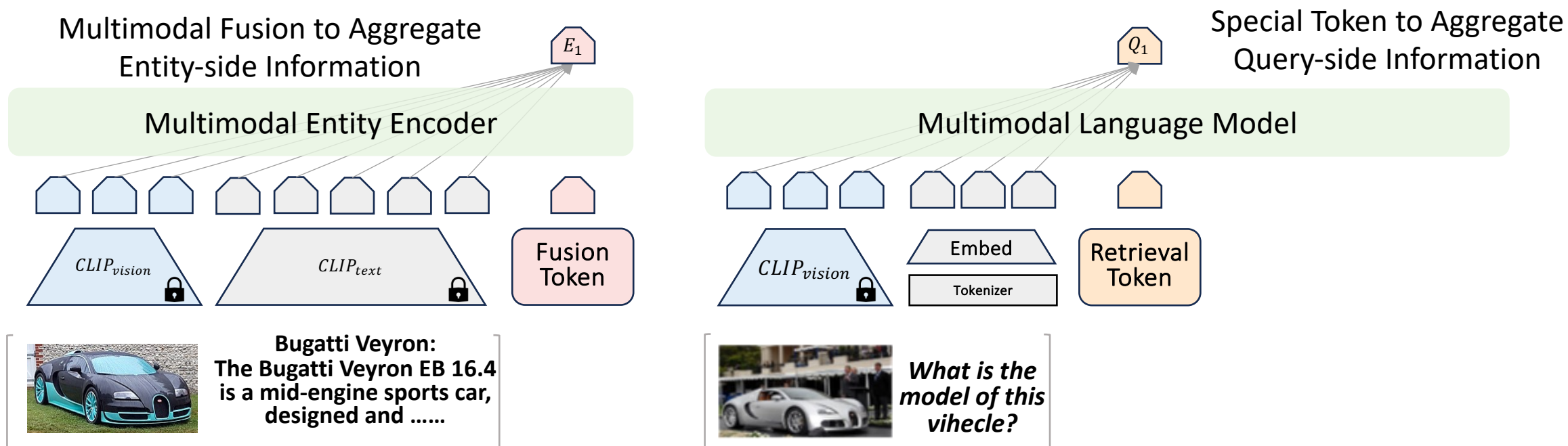
Standard LLaVA design



Our Approach: AutoVER

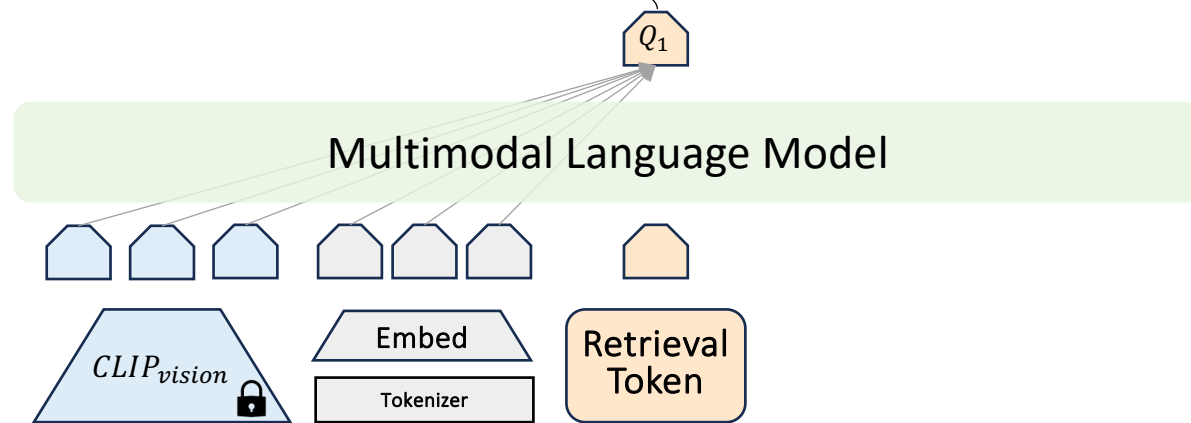
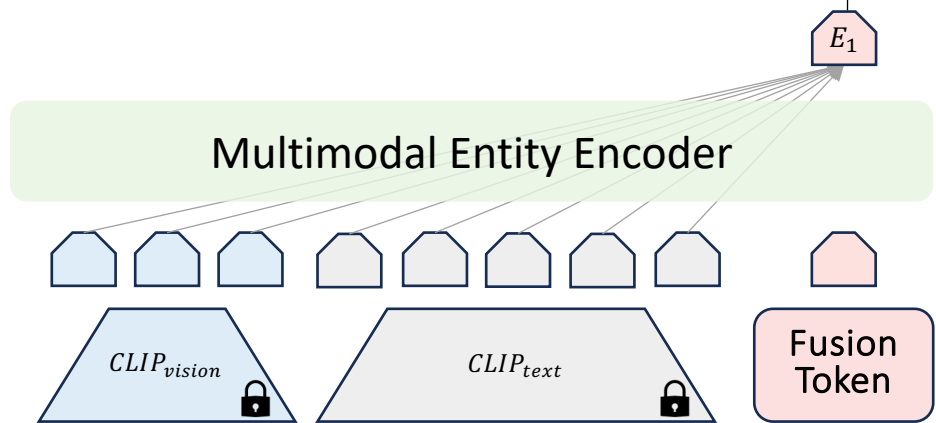
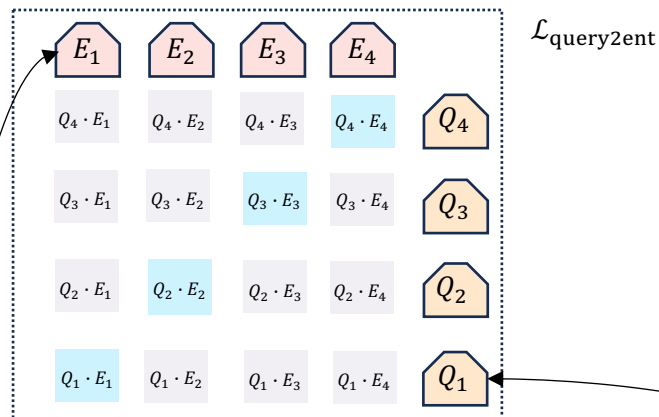


Our Approach: AutoVER



Our Approach: AutoVER

In-Batch Contrastive Training

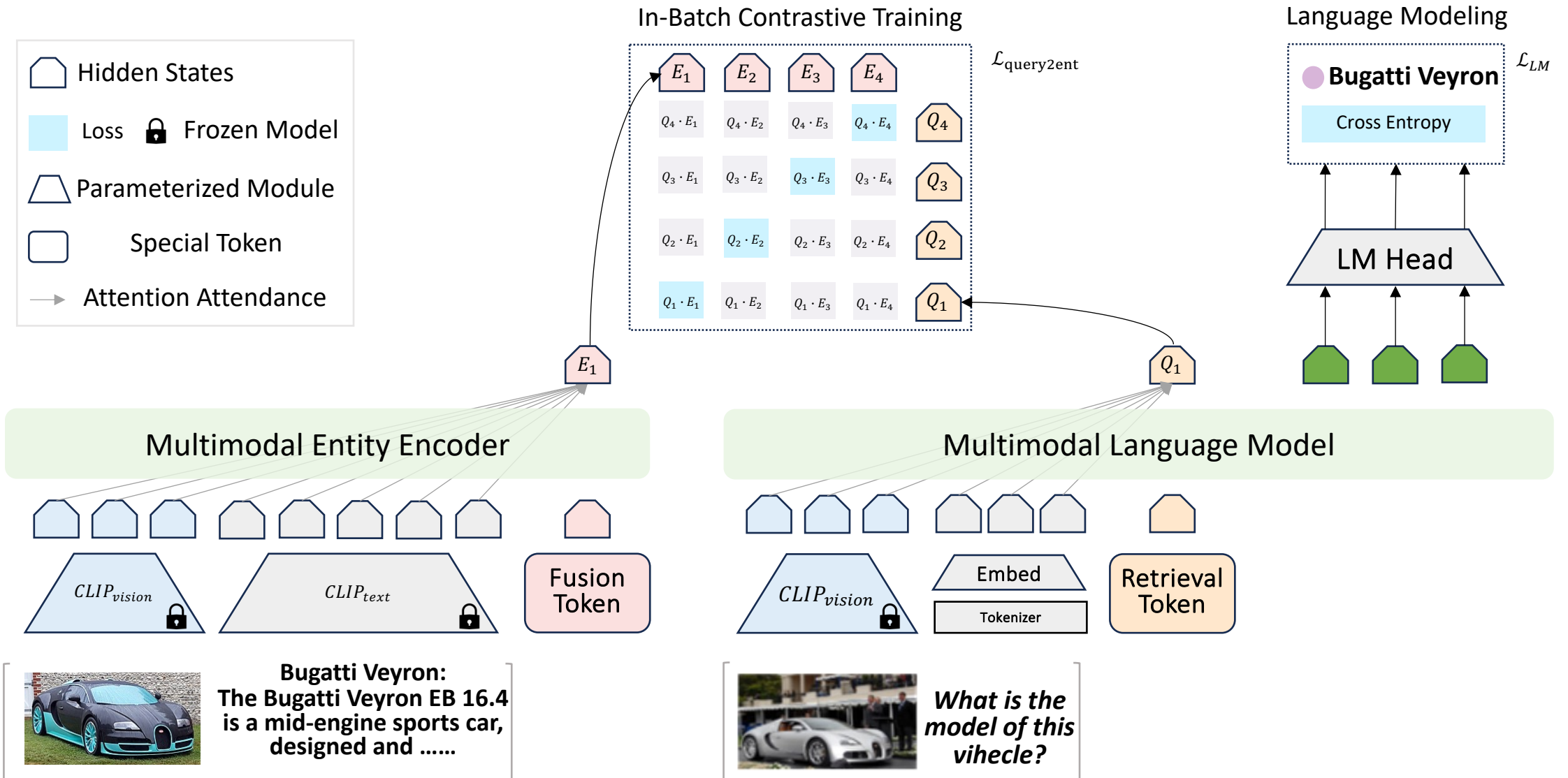


Bugatti Veyron:
The Bugatti Veyron EB 16.4 is a mid-engine sports car, designed and

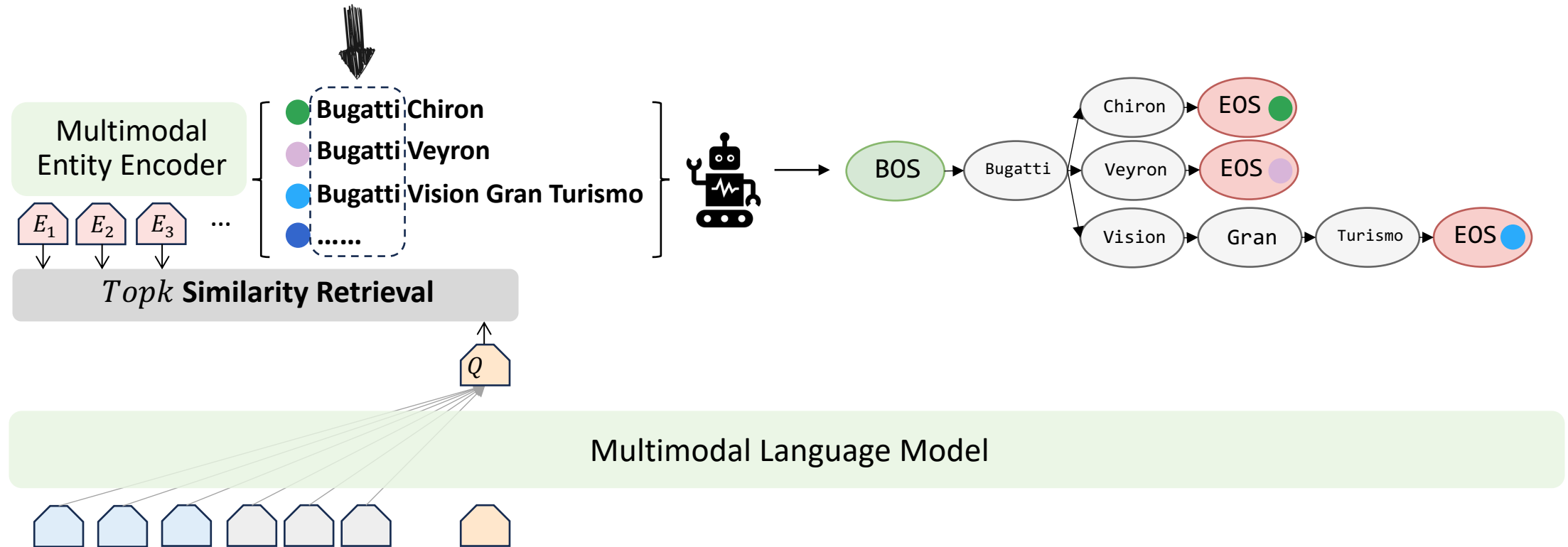


What is the model of this vihecle?

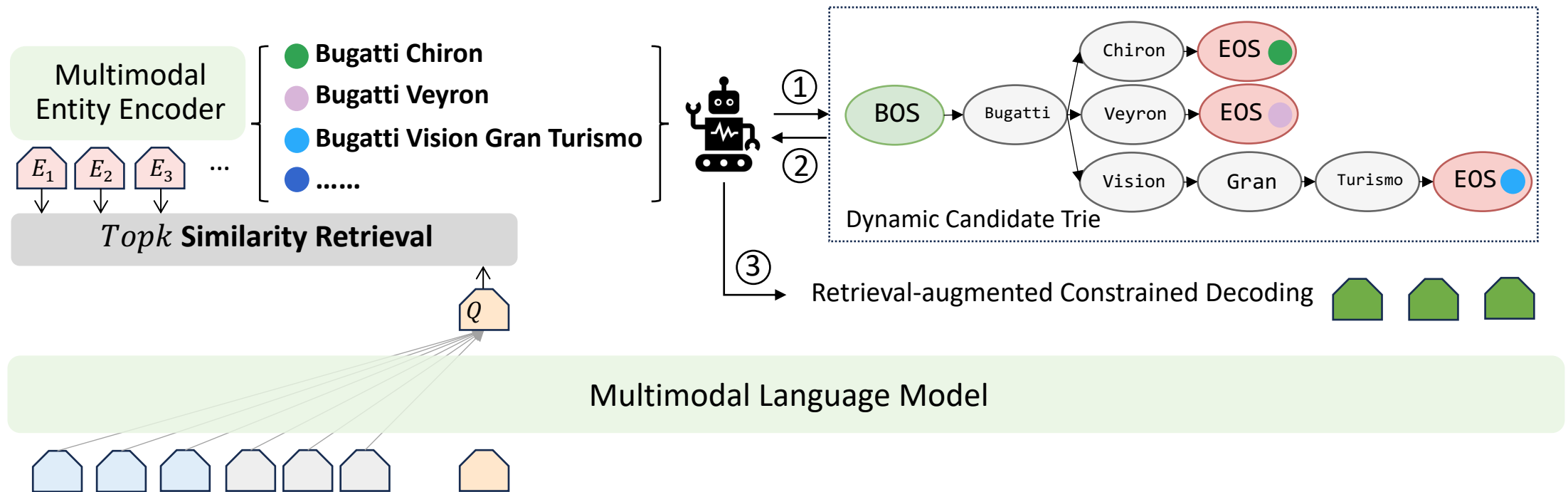
Our Approach: AutoVER



At Inference-Time: Retrieve then Generate with Guidance



At Inference-Time: Retrieve then Generate with Guidance
















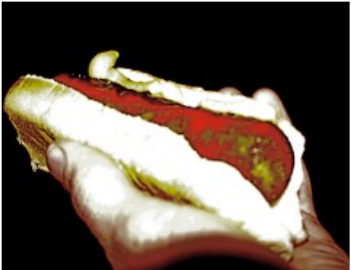










Quantitative Results

- We trained AutoVER 7B & 13B on 2.5M+ image-query pairs from OVEN-Train dataset and tested the model on val and test splits.
- It demonstrates a consistent improvement in all data splits and subsets.
 - with the exception for falling short of zero-shot GPT-4V on Query Unseen Split.

Category	Method	Entity Split			Query Split			Overall
		SEEN	UNSEEN	HM	SEEN	UNSEEN	HM	HM
Discriminative	CLIP _{ViTL14}	5.4	5.3	5.4	0.8	1.4	1.0	1.7
	CLIP Fusion _{ViTL14}	32.7	4.3	7.7	33.4	2.2	4.2	5.4
	CLIP2CLIP _{ViTL14}	12.6	10.1	11.2	4.1	2.1	2.8	4.4
Generative	PaLI-3B	21.6	6.6	10.1	33.2	14.7	20.4	13.5
	PaLI-17B	30.6	12.4	17.6	44.2	22.4	29.8	22.1
Zero-shot	BLIP-2 _{F1an-T5-XXL}	8.6	3.4	4.9	24.6	17.7	20.6	7.9
	GPT-4V	29.8	19.3	23.4	56.5	52.7	54.5	32.9
Ours	AUTOVER-7B	61.5	21.7	32.1	69.0	31.4	43.2	36.8
	AUTOVER-13B	63.6	24.5	35.6	68.6	32.3	43.9	39.2

Qualitative Results

- We visualize some retrieved entity candidates and the decision made by MLLM.
- The model adeptly captures slight variations in the query text and retrieves entirely different entity candidates.

Input	Retrieved Candidates	MLLM Decision
		
What is the model of this aircraft?	 Boeing 717	 Boeing 767
What is the manufacturer of this aircraft?	 Boeing	 Boeing
	 Douglas DC-8	
	 Boeing 777	
	 Airbus A320	
	 Boeing 767	
	 Fairchild Aircraft	
	 Cessna Cessna	
	 British Aerospace	
	 Suzuki	
		
What is the topping on the hot dog called?	 Green sauce	 Relish
What kind of food is it?	 Cheeseburger	 Hot dog
	 Taco	
	 Curry	
	 Omelette	
	 Marmite	
	 Ham sandwich	
	 Pizza	

Ablation Study

- Ablation experiments on a subset of training data shows that:
- Retrieve-then-Generate paradigm heavily boost the performance on UNSEEN split of test dataset.
- With constrained decoding design, AutoVER suffers fewer hallucination brought by ungrounded response.

Table 5: Ablation study of AUTOVER-7B-0.1 on OVEN-Wiki ENTITY Split_(val).

Method	SEEN	UNSEEN	HM
AUTOVER-7B-0.1	48.9	19.0	27.4
+ w/o retrieval	50.7	0.6	1.2
+ w/o constrained decoding	46.8	0.6	1.2
+ w/ LoRA	43.5	2.8	5.3