Grounding Language Models for Visual Entity Recognition

[1] RICE   vislang   [2] Microsoft

Vicente Ordonez

Sept 30th, 2024

Zilin Xiao[1]

Ming Gong[2]

Paola Cascante-Bonilla[1]

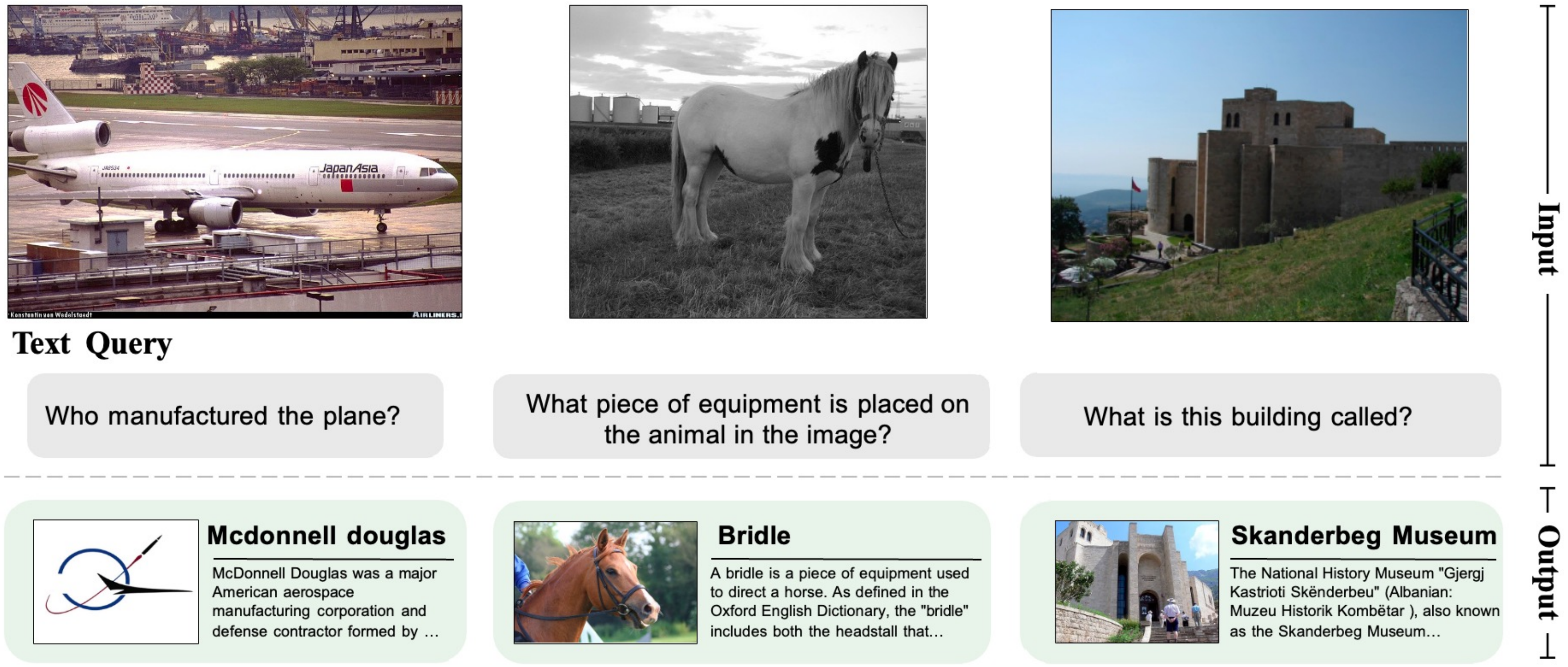
Xingyao Zhang[2]

Jie Wu[2]

Vicente Ordonez[1]

Introduction

- **Task:** Visual Entity Recognition
  - Input: an image and a query question
  - Output: a Wikipedia entity from a pre-defined set

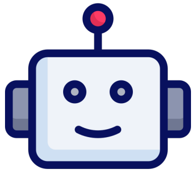


Introduction

➢ How it differs with Visual Question Answering?

- Answers grounded to Wikipedia Entities.

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



Boeing 767

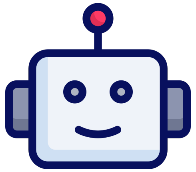
The Boeing 767 is an American wide-body airliner developed and manufactured by Boeing Commercial Airplanes. The aircraft was launched…

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

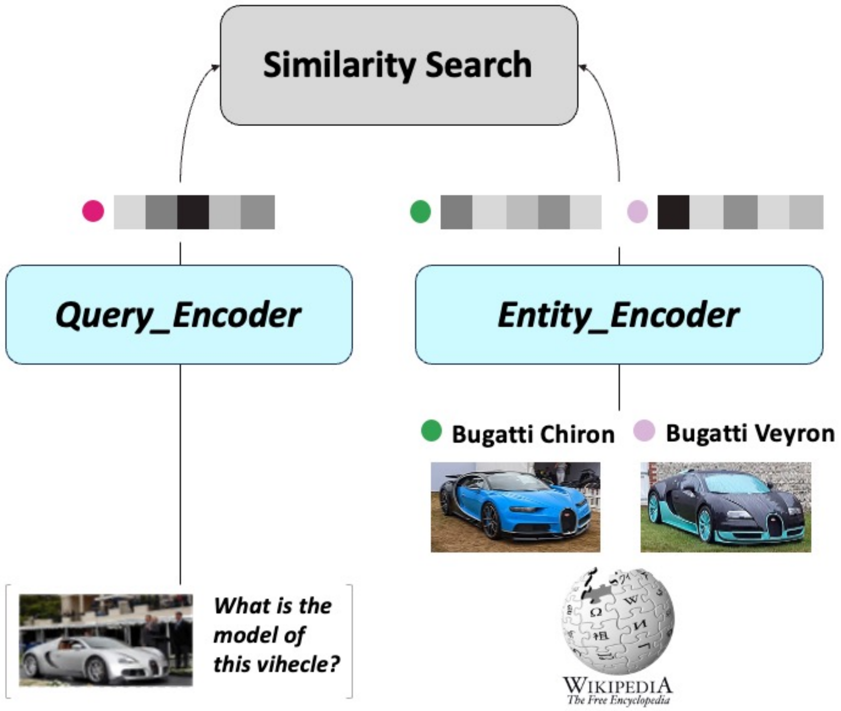
Boeing 777

Airbus A320

Boeing 767

Prior Approach: Similarity Search

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

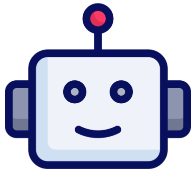


Prior Approach: Multimodal LLM

- 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

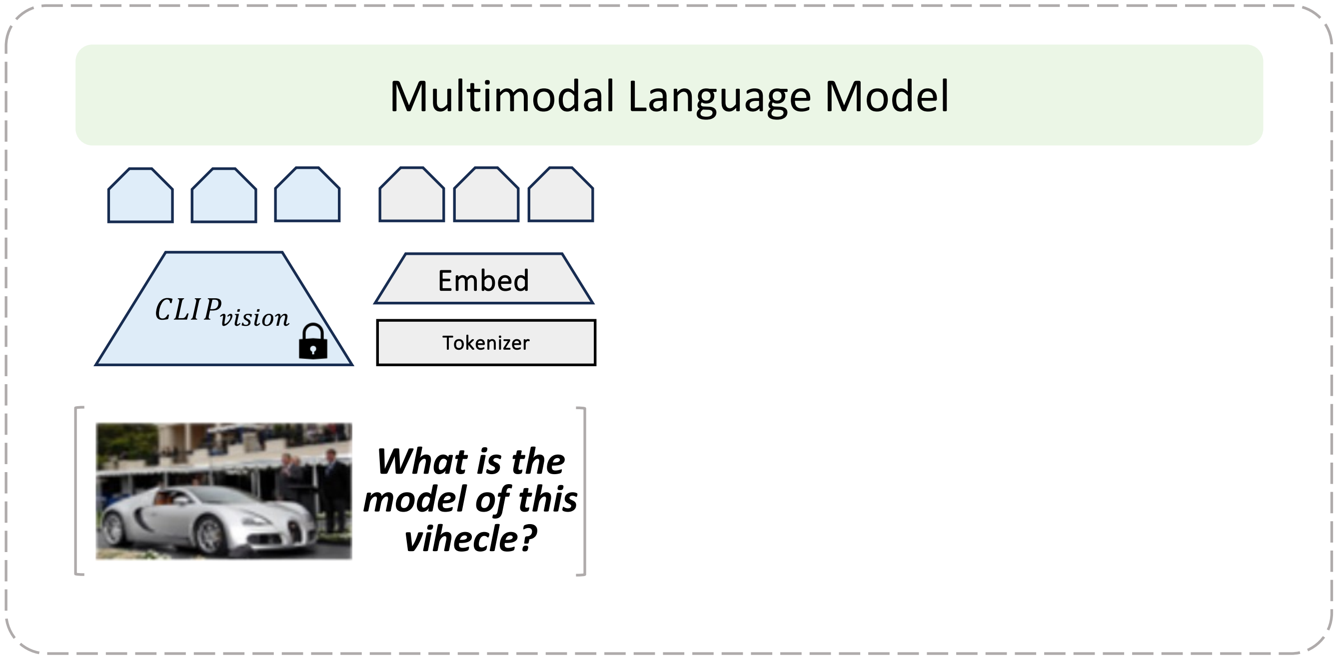
Herons are long-legged, long-necked, freshwater and coastal birds in the family Ardeidae, with 72 recognized species, some of which are referred…

Our Approach

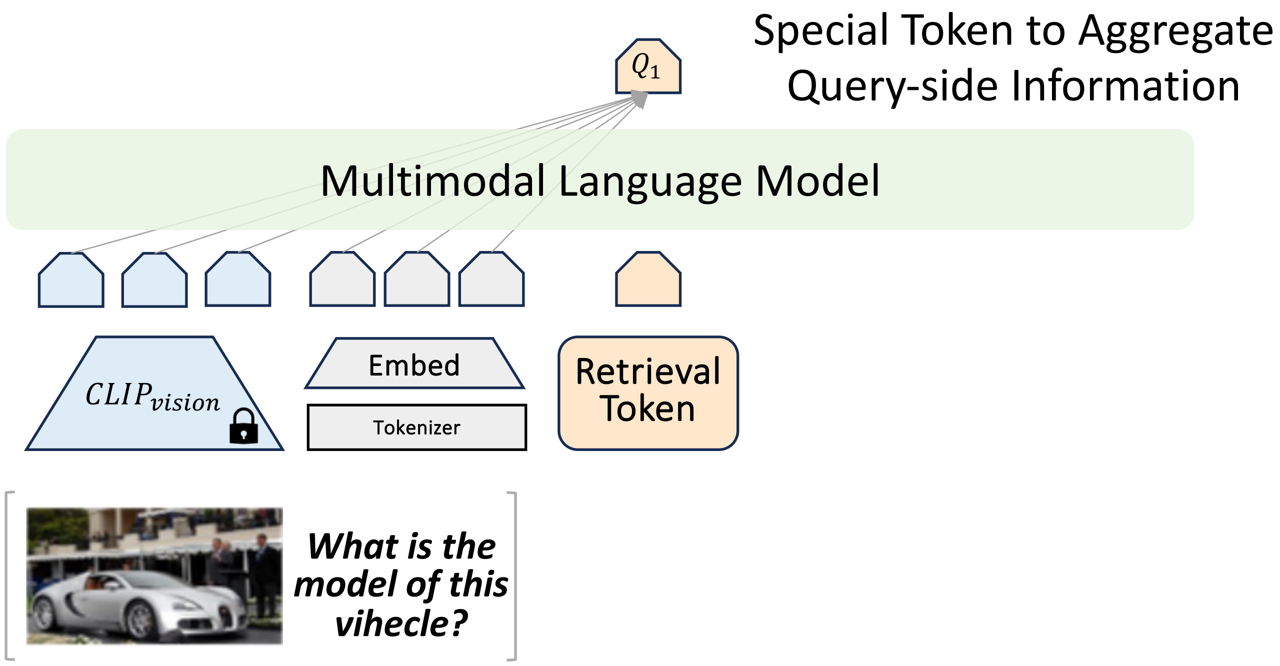
- Can we take advantage of both similarity search and MLLM inherent reasoning ability?
  - Yes!
- We can use similarity search to do a coarse **retrieval**, narrowing down the candidates.
- Then ask an Multimodal LLM to further "**re-ranking**" the candidates.

Our Approach: AutoVER

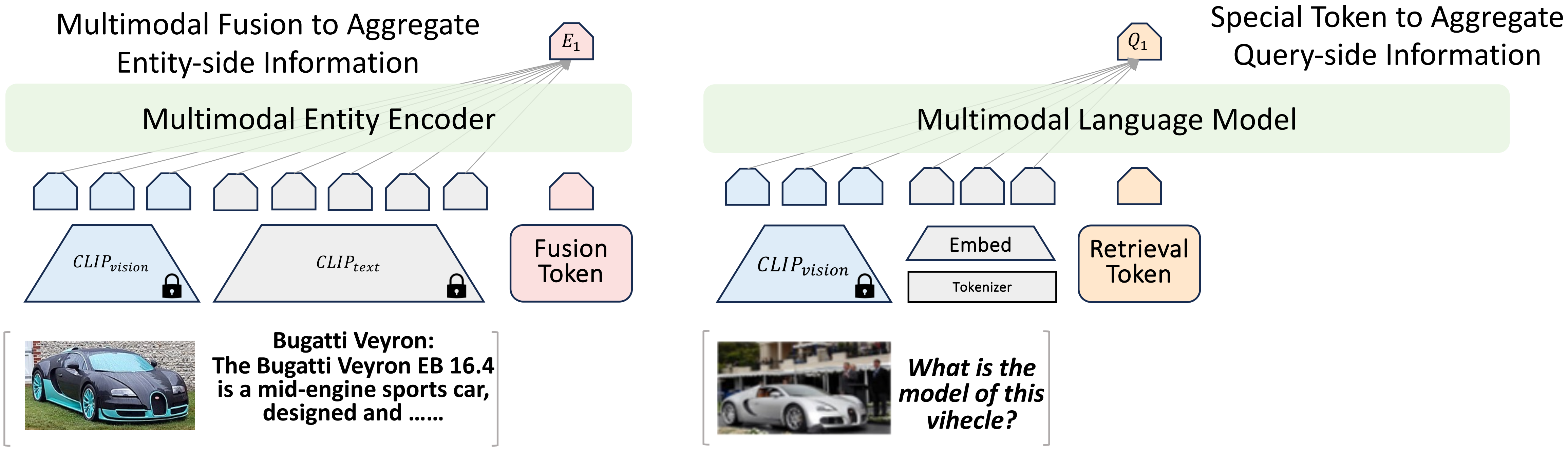
Stardard LLaVA design



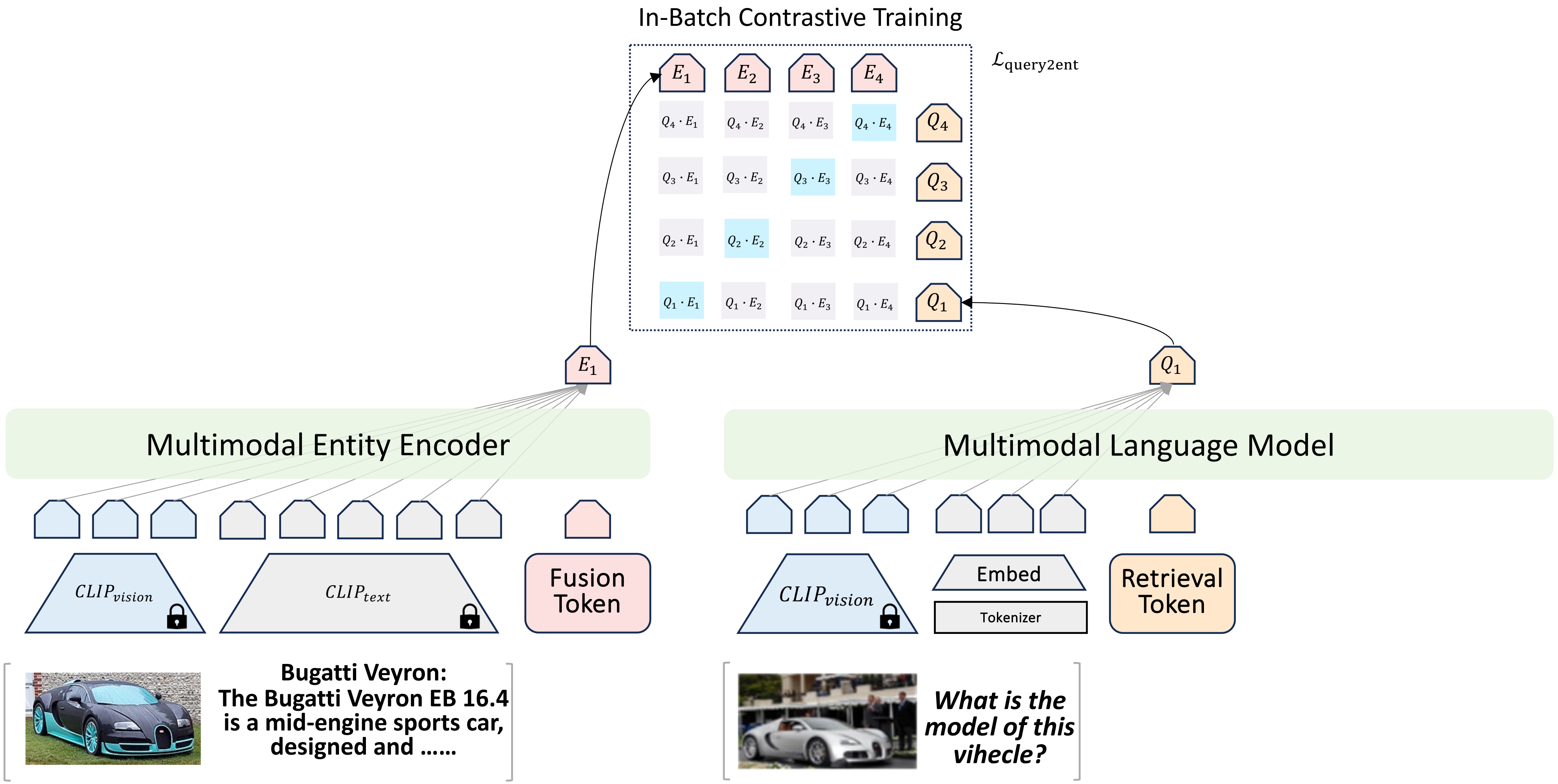
Our Approach: AutoVER



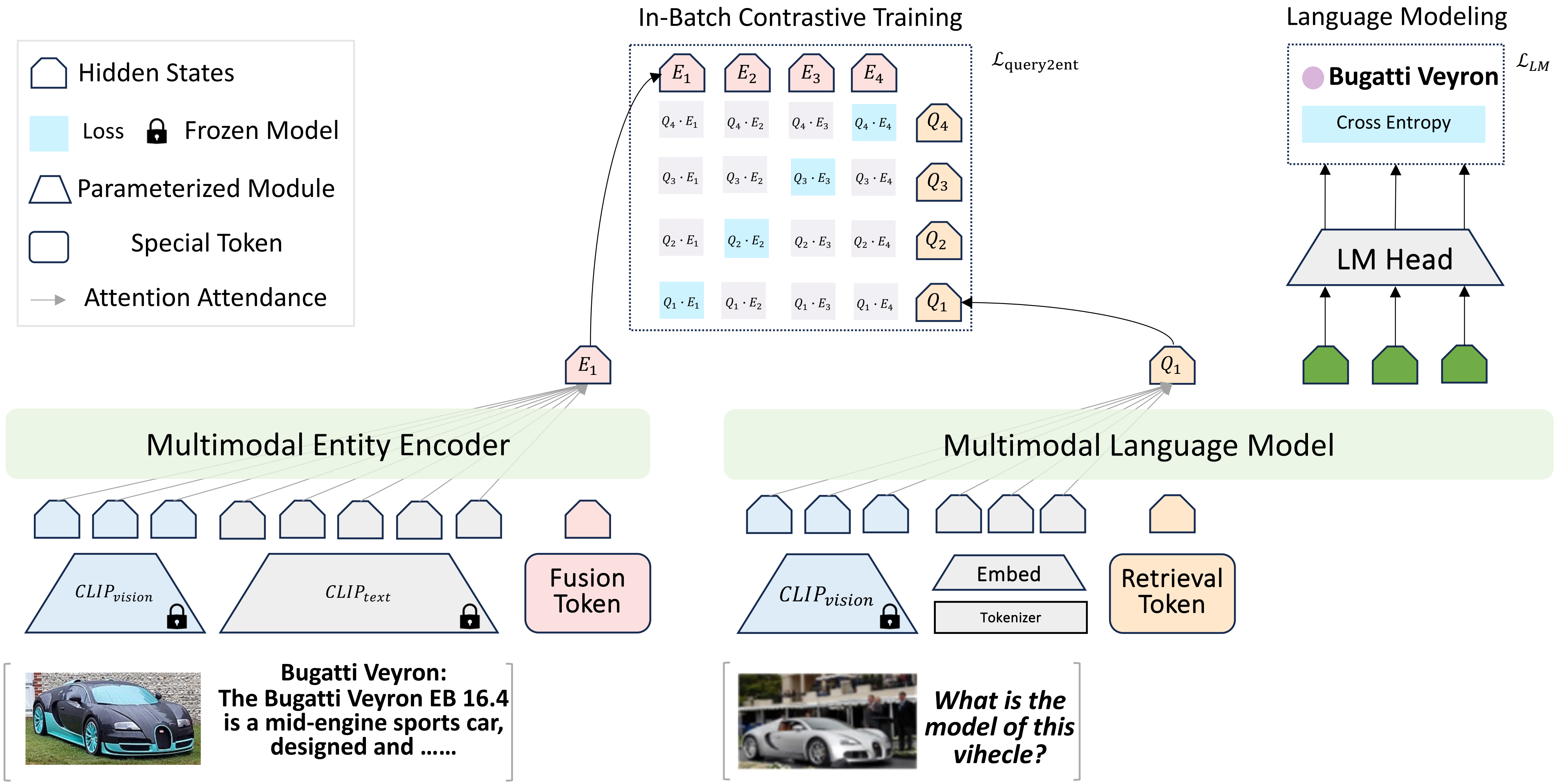
Our Approach: AutoVER



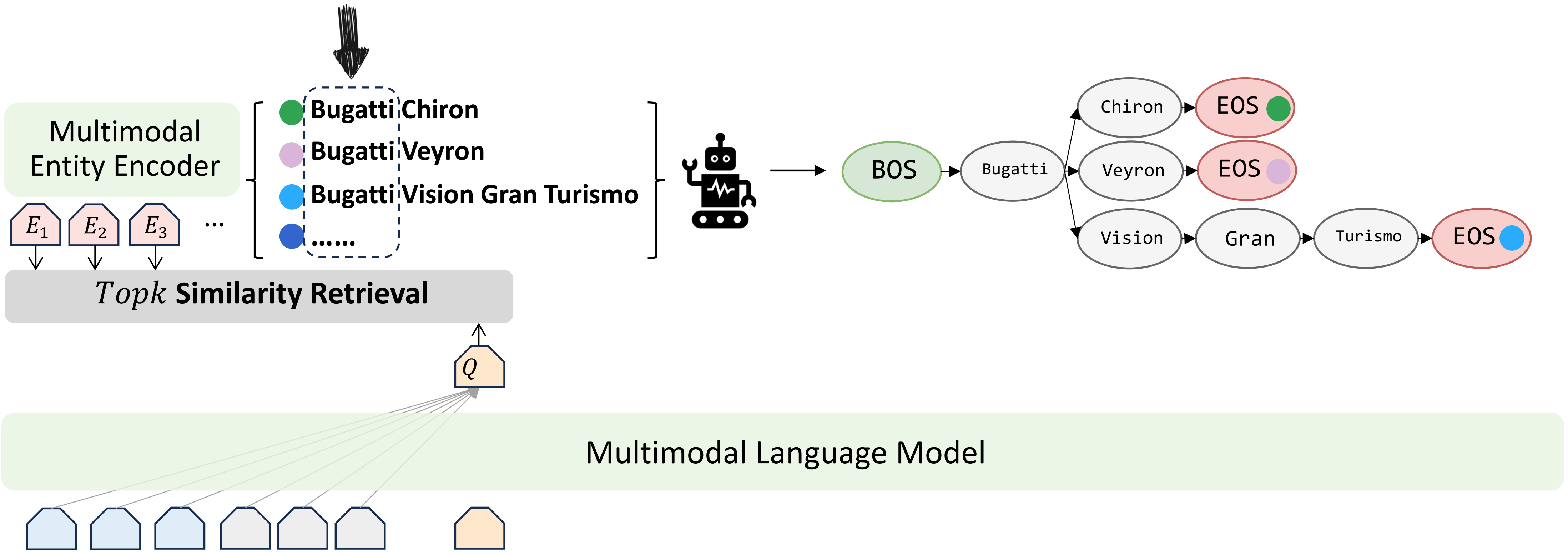
Our Approach: AutoVER



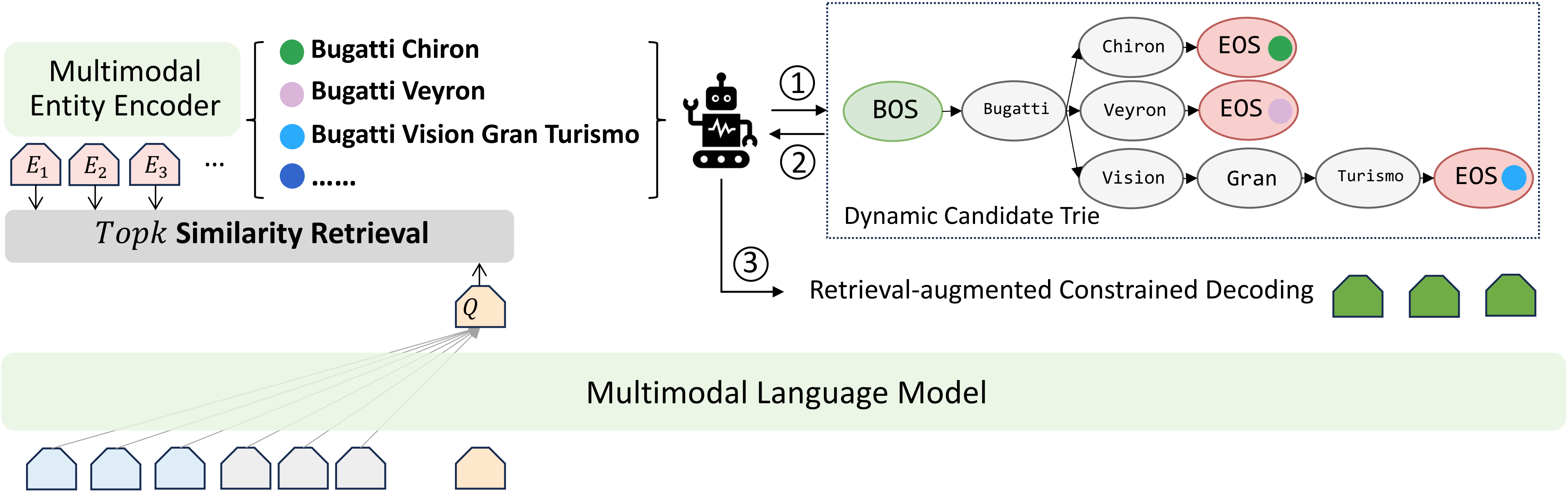
Our Approach: AutoVER



At Inference-Time: Retrieve then Generate with Guidance



At Inference-Time: Retrieve then Generate with Guidance



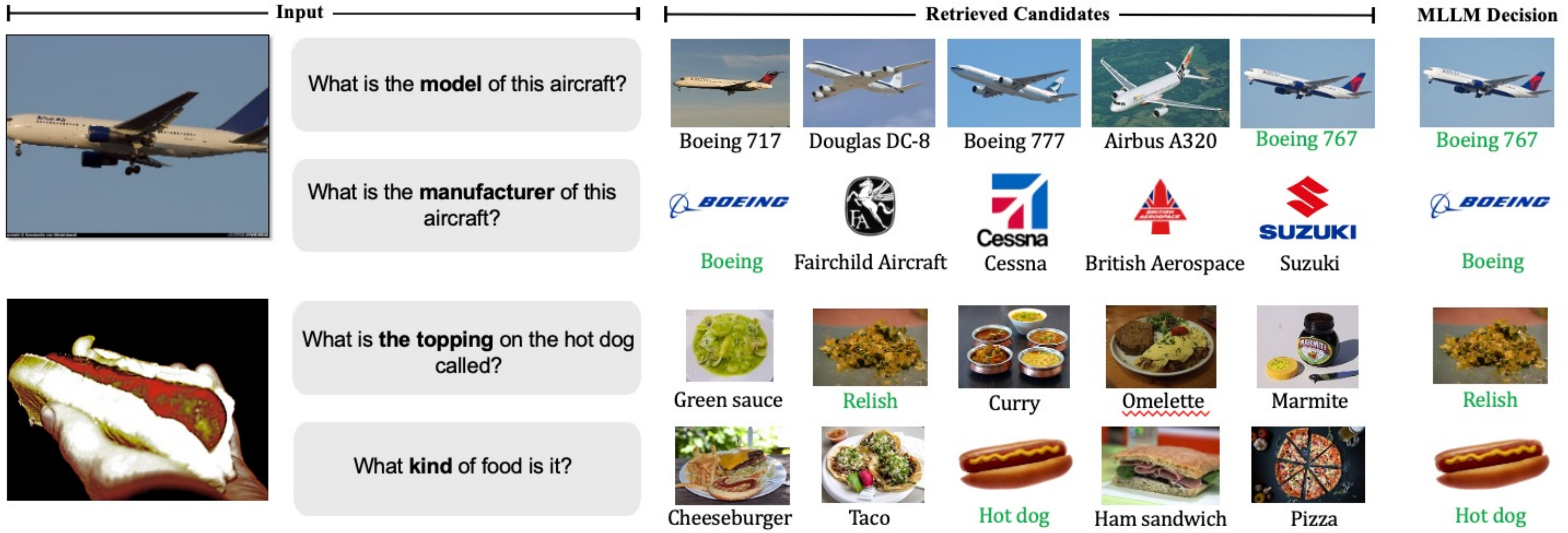
Quantitative Results

- 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.

		Entity Split			Query Split			Overall
Category	Method	SEEN	UNSEEN	HM	SEEN	UNSEEN	HM	HM
Discriminative	$\text{CLIP}_{\text{ViTL14}}$	5.4	5.3	5.4	0.8	1.4	1.0	1.7
	$\text{CLIP Fusion}_{\text{ViTL14}}$	32.7	4.3	7.7	33.4	2.2	4.2	5.4
	$\text{CLIP2CLIP}_{\text{ViTL14}}$	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	$\text{BLIP-2}_{\text{Flan-T5-XXL}}$	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.



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$_{(\mathtt{Val})}$.

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