Bits of Language: Multimodal Learning

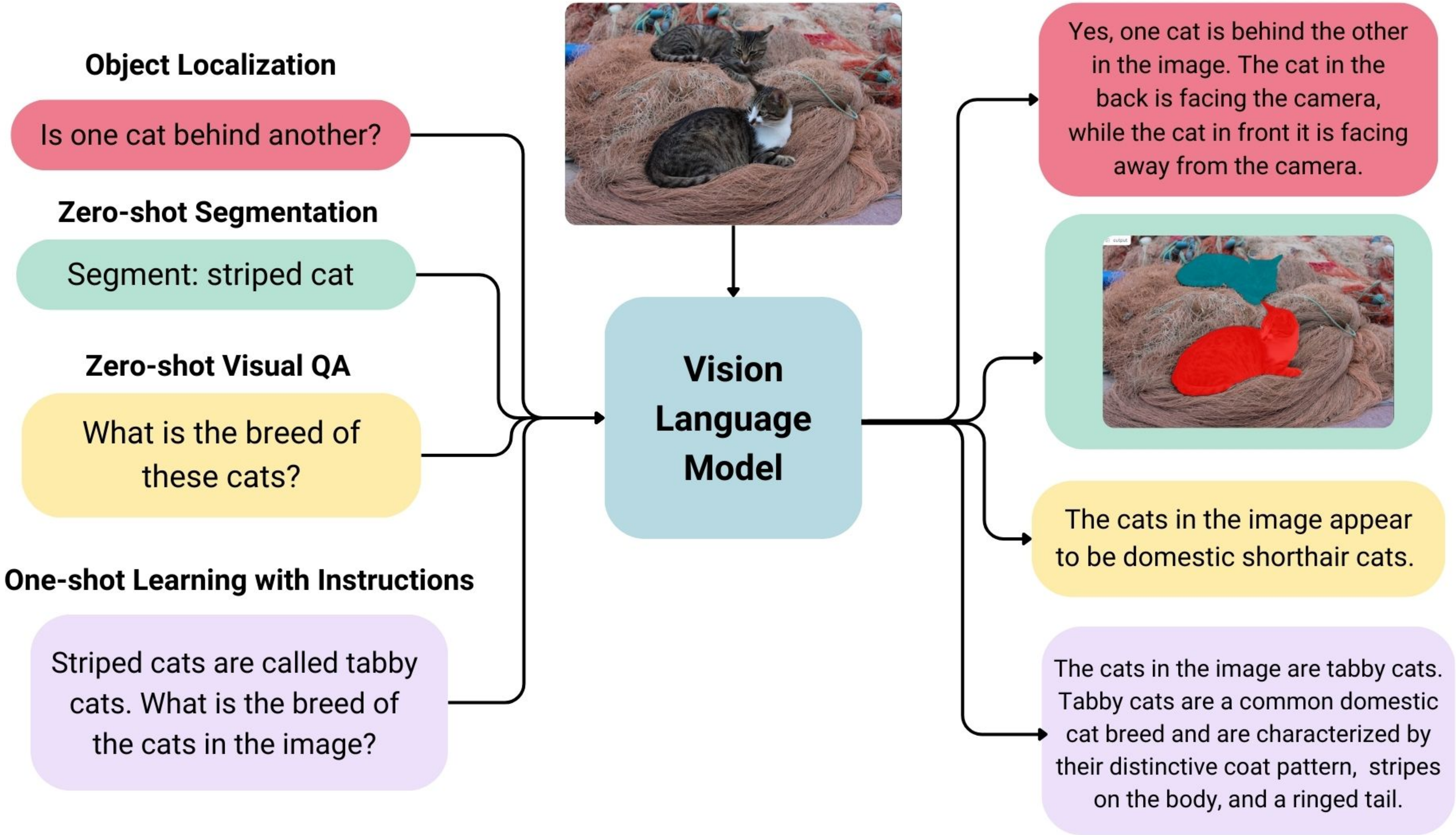
Overview

- **Last week.** Deep Learning for Text processing
  - Tokenization
  - Architectures
  - Word2Vec, BERT, GPT

- **This week.** Further developments
  - Multimodal Models
  - Post-Processing of Language Models

Overview

- **Today.** Multimodal Learning – the case of vision + language



Object Localization

Is one cat behind another?

Zero-shot Segmentation

Segment: striped cat

Zero-shot Visual QA

What is the breed of these cats?

One-shot Learning with Instructions

Striped cats are called tabby cats. What is the breed of the cats in the image?

Vision Language Model

Yes, one cat is behind the other in the image. The cat in the back is facing the camera, while the cat in front it is facing away from the camera.

The cats in the image appear to be domestic shorthair cats.

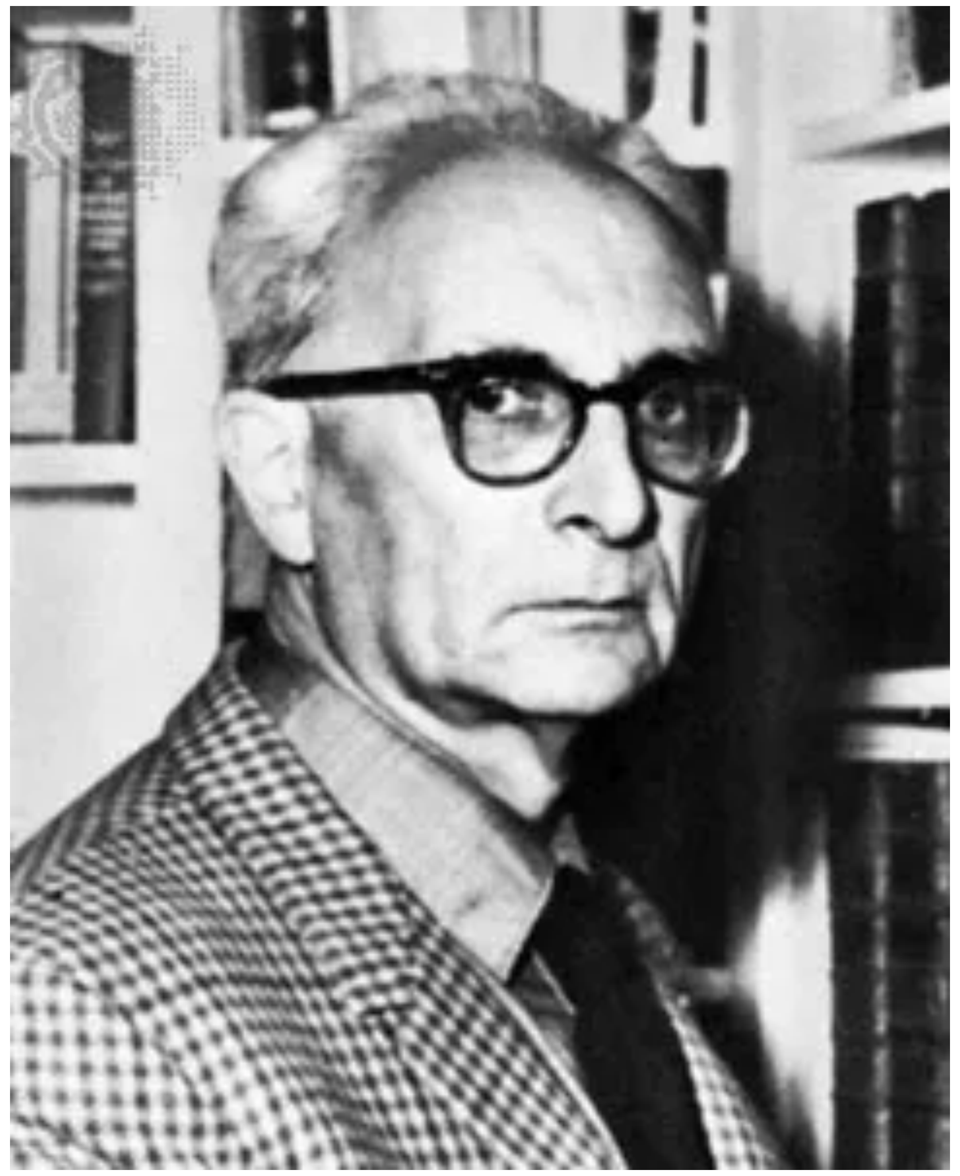
The cats in the image are tabby cats. Tabby cats are a common domestic cat breed and are characterized by their distinctive coat pattern, stripes on the body, and a ringed tail.

Overview

- Let LLMs be our central interface for thinking & reasoning
- **Why?** Language shapes how we think (or at least we believe so)

George Boole

"That language is an instrument of human reason, and not merely a medium for the expression of thought, is a truth generally admitted."

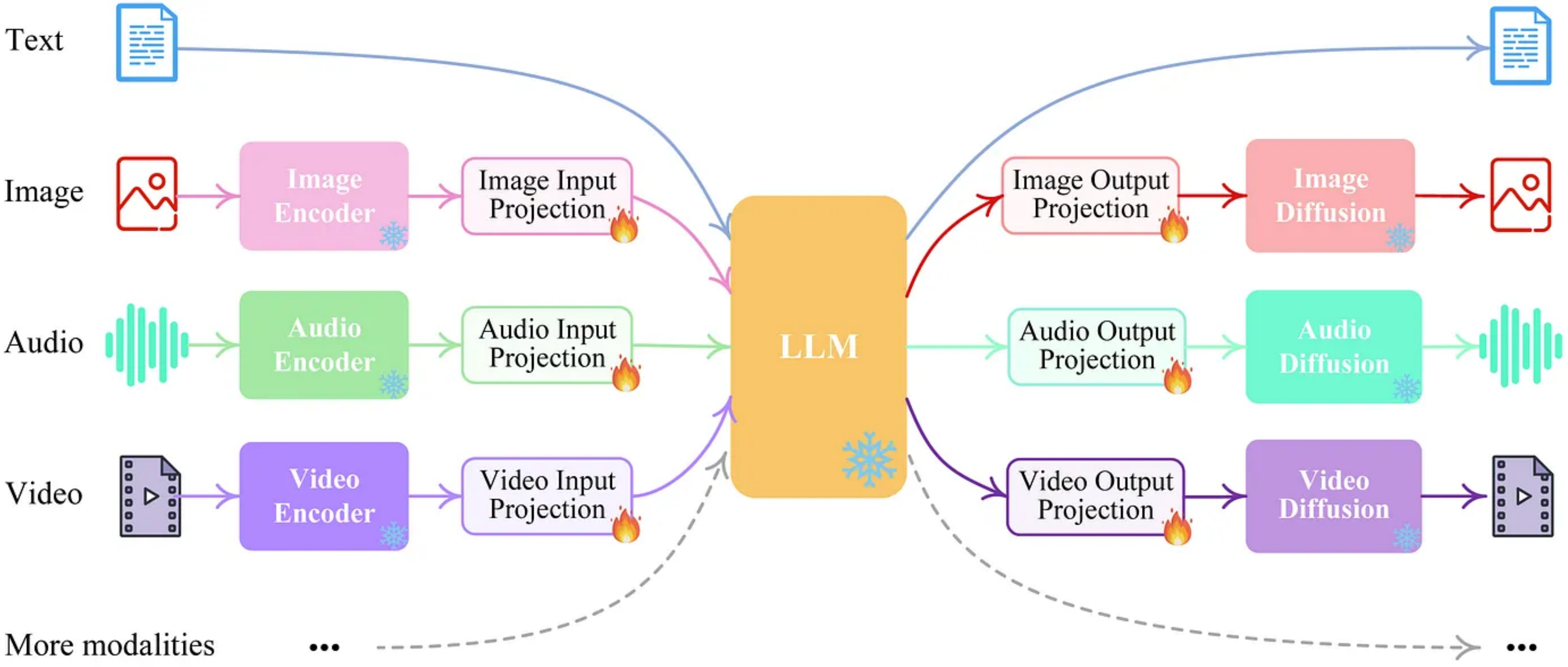


Claude Lévi-Strauss

"Language is a form of human reason, which has its internal logic of which man knows nothing."

Overview

- To let LLMs process multimodal information, we need:
- **Input.** Various modalities encoded into a form that LLMs can understand
- **Output.** Acquire LLMs with tools that can be queried with text



Overview

- **Scope.** How to make LLMs process visual inputs
  - Architecture. Vision Transformer
  - Encoder Training. CLIP
  - Pipeline. LLaVA
  - Further applications. Robotic Control

Vision Transformers

Vision transformers

- **Question.** Can we use transformers to process visual inputs?
- Hope#1. Transformers are scalable
  - Performance gets better, seemingly without limit, with larger scales
- Hope#2. Handling text and image within a unified architecture

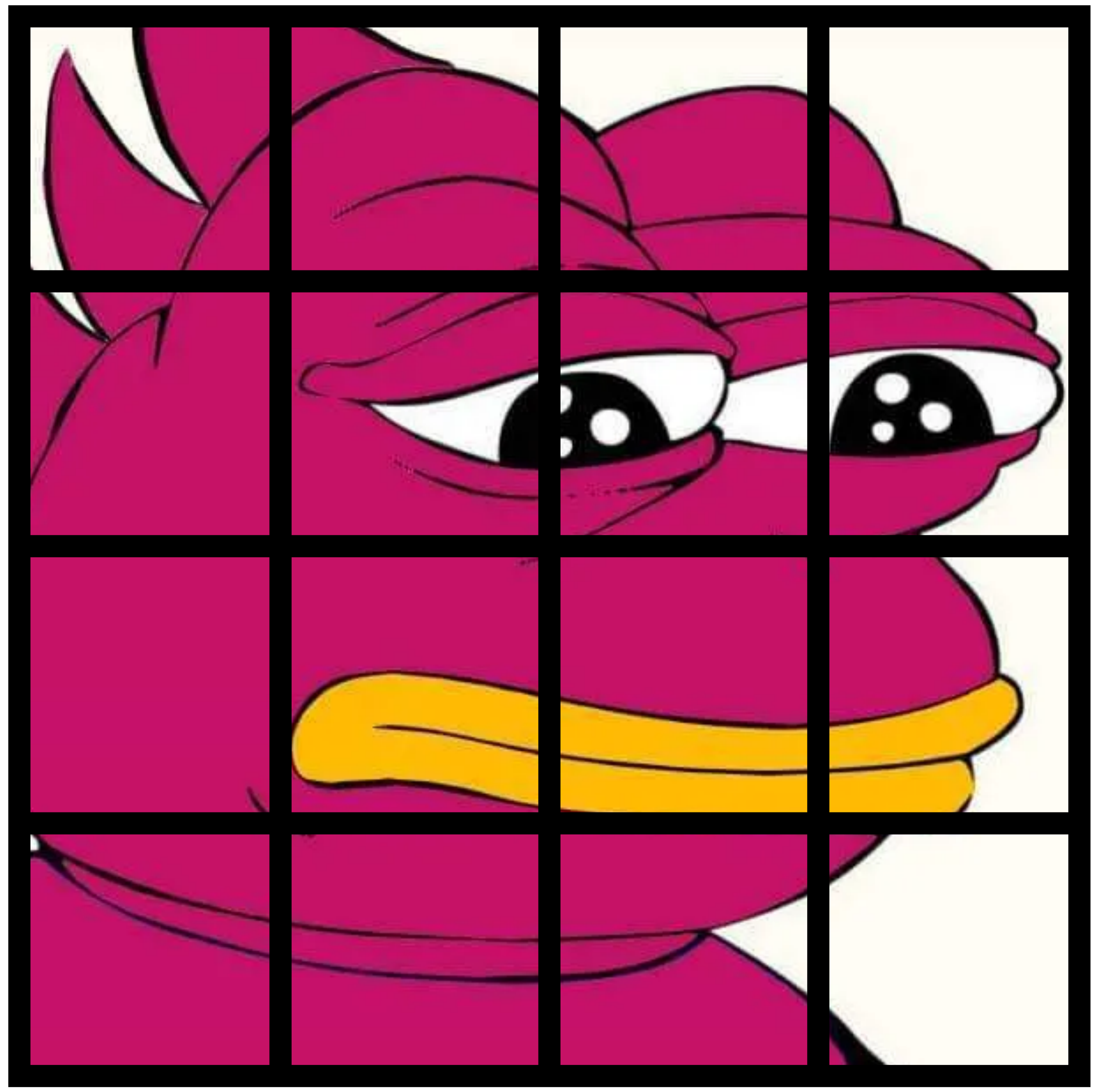
Test Loss
7
6
5
4
3
2
10^{-9}
10^{-7}
10^{-5}
10^{-3}
10^{-1}
10^{1}
$L = (C_{\min}/2.3 \cdot 10^{8})^{-0.050}$
Compute
PF-days, non-embedding
4.2
3.9
3.6
3.3
3.0
2.7
10^{8}
10^{9}
$L = (D/5.4 \cdot 10^{13})^{-0.095}$
Dataset Size
tokens
5.6
4.8
4.0
3.2
2.4
10^{5}
10^{7}
10^{9}
$L = (N/8.8 \cdot 10^{13})^{-0.076}$
Parameters
non-embedding

Idea

- Break image down into a sequence of low-res patches (token)
  - Typically 14x14 or 16x16

Image

$\mathbf{x} \in \mathbb{R}^{H \times W \times C}$



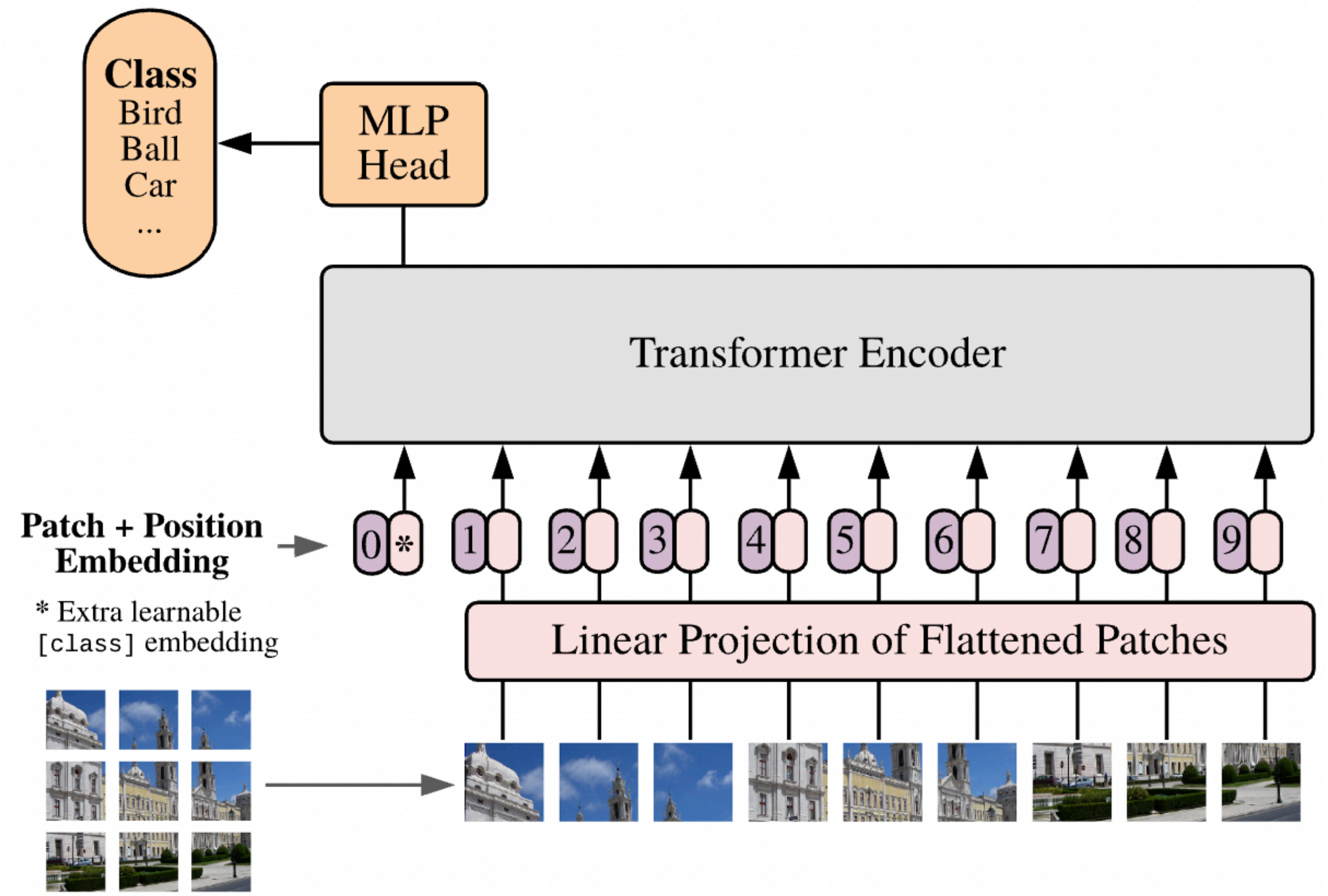
Sequence of Patches

$\mathbf{x}_i \in \mathbb{R}^{P \times P \times C}$

(total HW/P^2 patches)

Idea

- Then, we (1) embed these tokens
  (2) process with transformers
  - Jointly train a class token

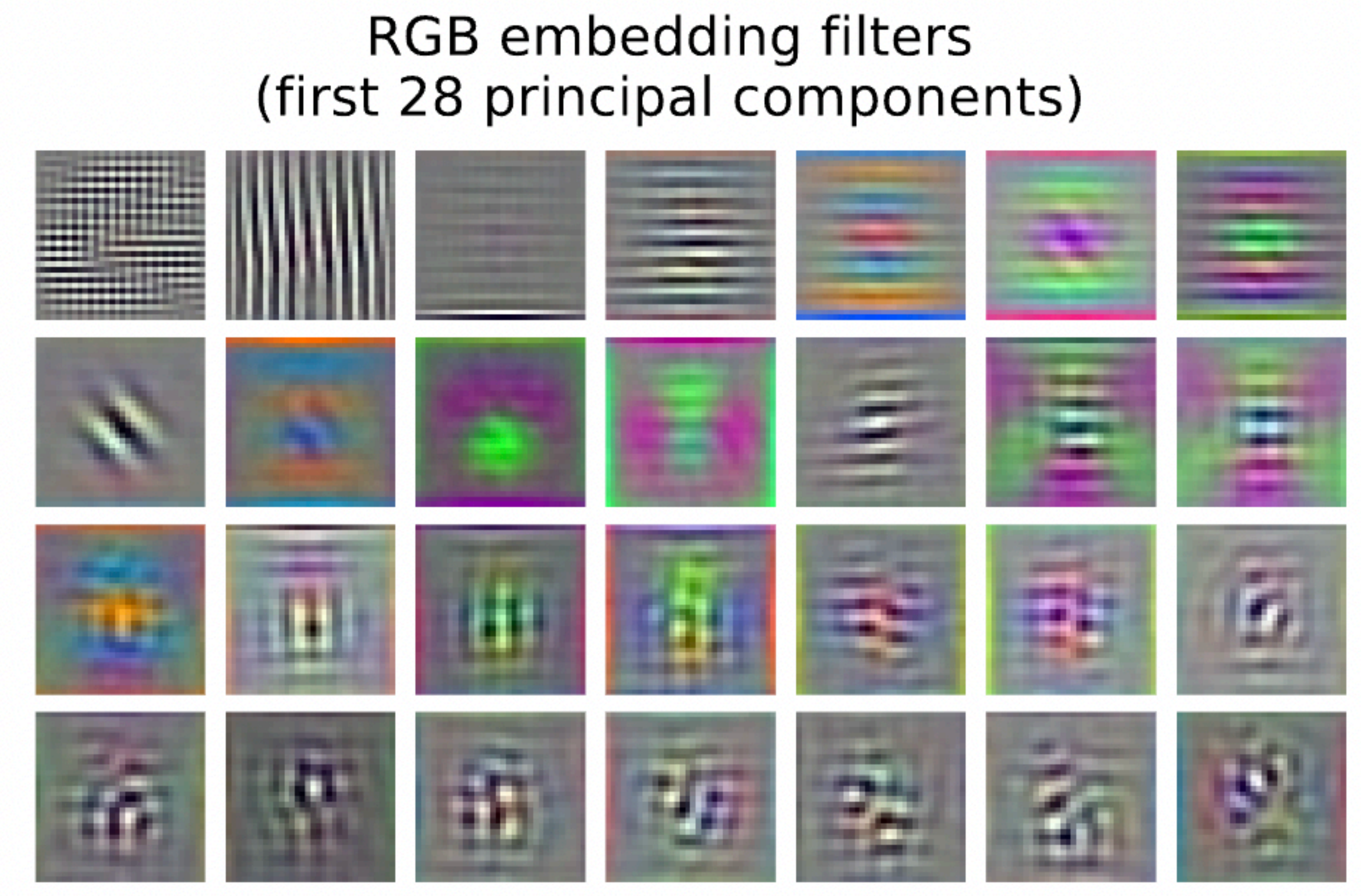

Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
0
*
1
2
3
4
5
6
7
8
9
* Extra learnable [class] embedding
Linear Projection of Flattened Patches

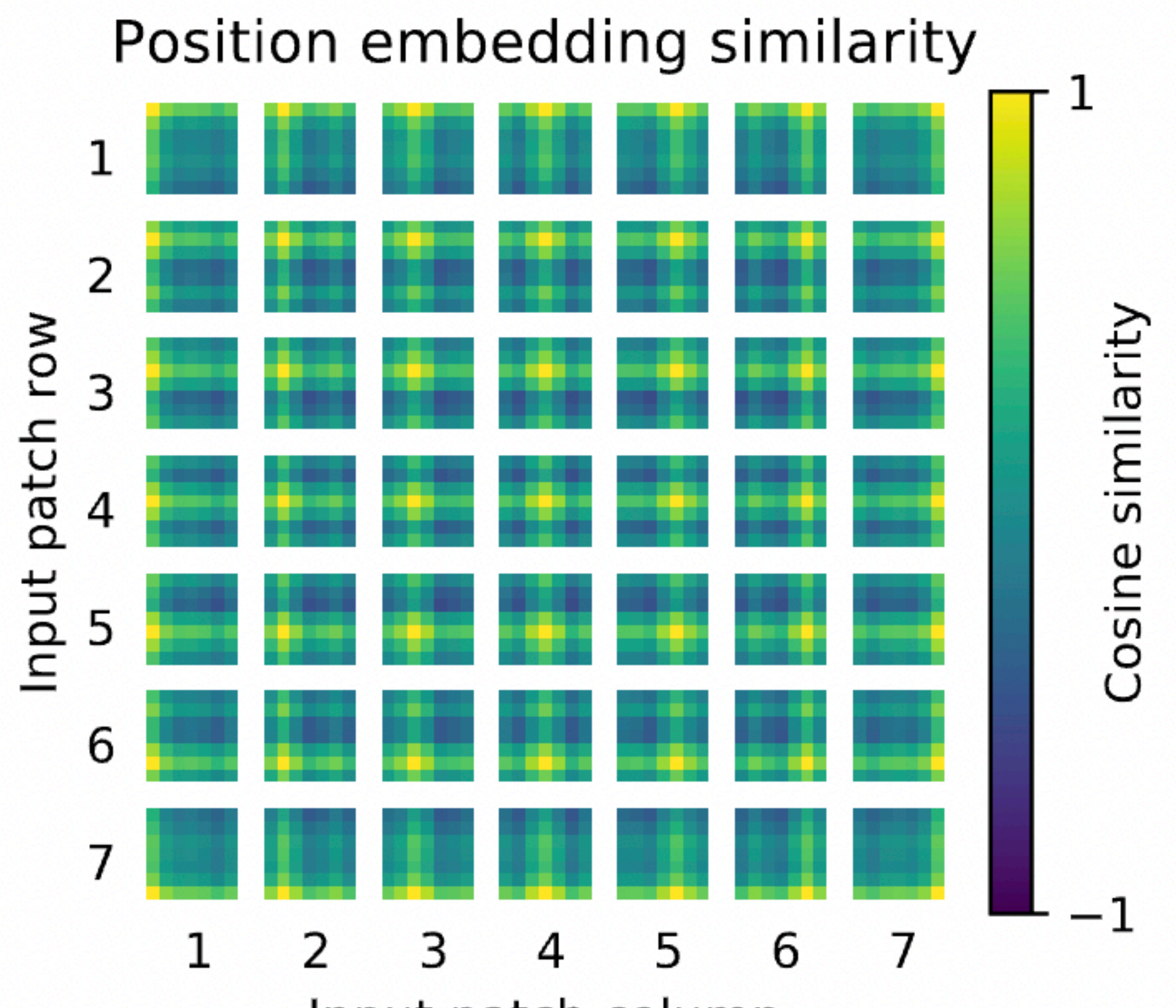
Embedding

- We train both (1) linear embedding, and (2) 1D positional encoding

$$\mathbf{z}_i = \mathbf{W}_{\mathrm{emb}}\,\mathrm{flatten}(\mathbf{x}_i) + \mathbf{e}_i$$

- **LE.** Depends only on how patch looks $\mathbf{W}_{\mathrm{emb}} \in \mathbb{R}^{d\times(P^2C)}$
- **PE.** Depends only on the location of the patch $\mathbf{e}_1, \ldots, \mathbf{e}_{HW/P^2} \in \mathbb{R}^d$


RGB embedding filters
(first 28 principal components)


Position embedding similarity

Drawback

- Transformers are known to be sample-inefficient
  - **ConvNet.** Impose some architectural constraints (inductive bias)
    - Strong locality
    - Translation-equivariance
  - **ViT.** No strong constraint
    - More flexible, but needs more data
- **Solution.**
  - Distilling ConvNet priors (DeiT)
  - Self-supervised pretraining (MAE)
  - Hybrid architectures (ConvViT)

Drawback

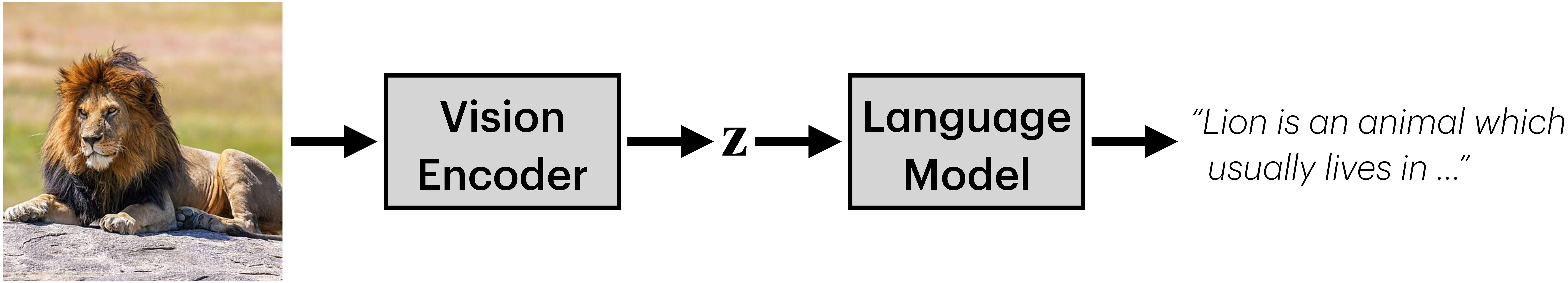
- Transformers are known to be computation-inefficient
  - Requires larger number of parameters for similar performance
  - Quadratic growth of computations w.r.t. resolution

- **Solution.**
  - Model compression
  - Lightweight architectures (MobileViT)

CLIP

CLIP

- **Question.** How can we let LLMs utilize features from visual inputs?
  - We may need some good translators
    - Process images into tokens that LLMs can understand
    - Large-scale training needed, presumably



Vision Encoder

$\mathbf{z}$

Language Model

"Lion is an animal which usually lives in …"

CLIP: Contrastive Language-Image Pretraining

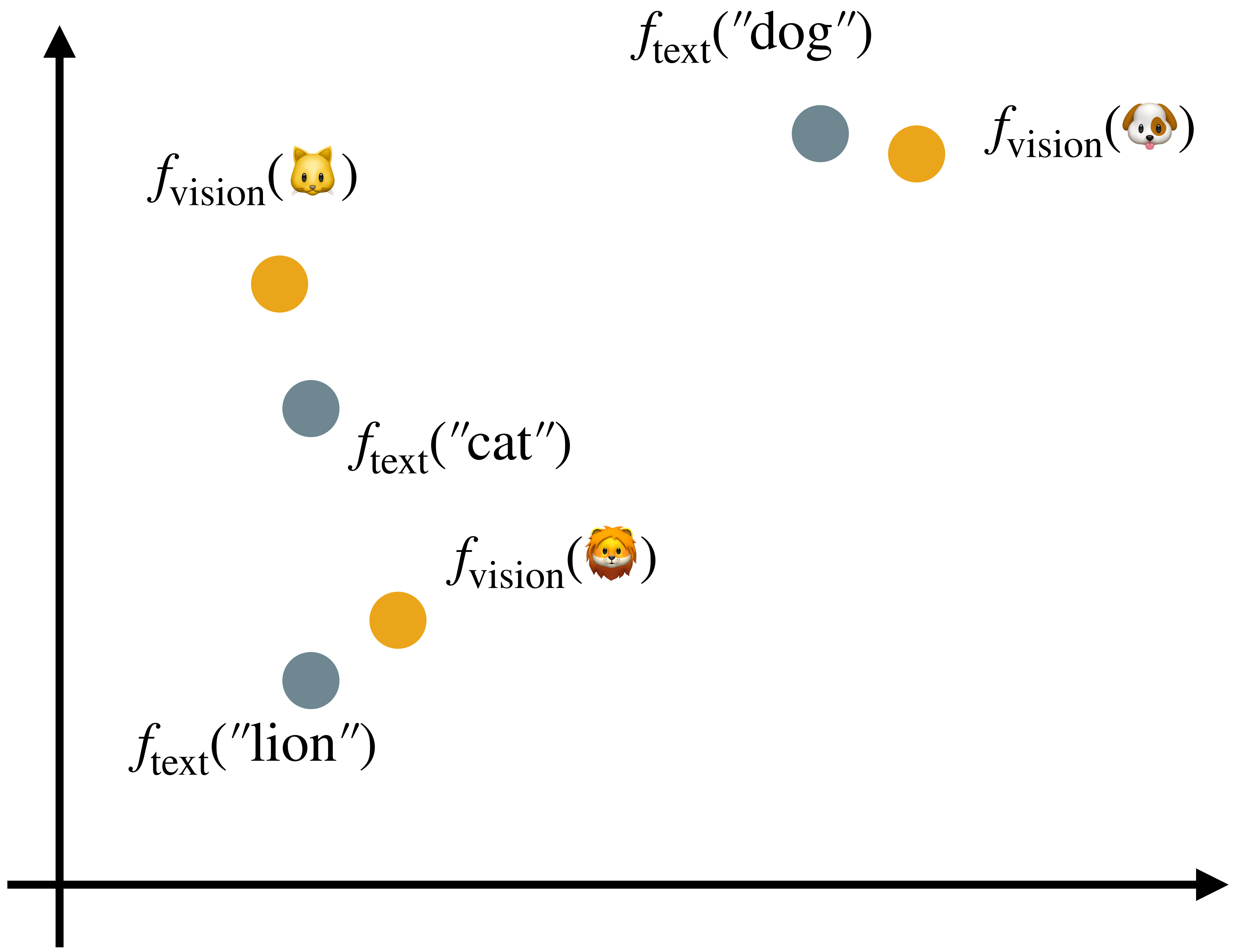
- Large-scale pre-training algorithm
  - Constructing a shared feature space of image and texts
- **Idea.** Encourage the "lion" and 🦁 to be close in this feature space, i.e.,

$$f_{\text{text}}(''\text{lion}'') \approx f_{\text{vision}}(\text{🦁})$$

- Such feature map can be used as:
  - Off-the-shelf classifier, by finding

$$\min_{\mathbf{x}} \text{dist}\left(f_{\text{text}}(\mathbf{x}), f_{\text{image}}(\text{🐯})\right)$$

  - Vision encoder for LLMs

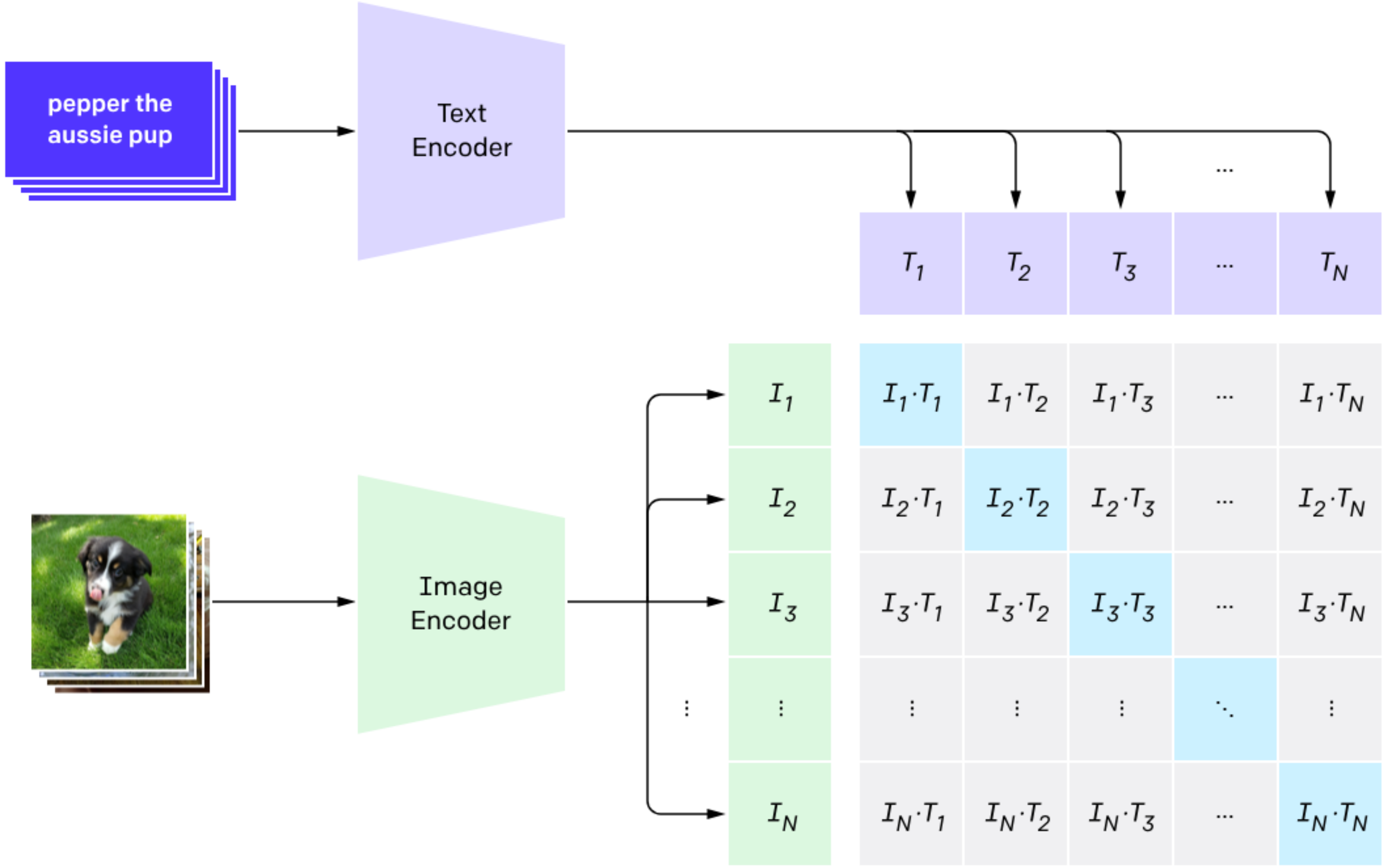

$f_{\text{text}}(''\text{dog}'')$
$f_{\text{vision}}(\text{🐶})$
$f_{\text{vision}}(\text{🐱})$
$f_{\text{text}}(''\text{cat}'')$
$f_{\text{vision}}(\text{🦁})$
$f_{\text{text}}(''\text{lion}'')$

Training

- Done by contrastive pretraining
  - Draw N image-text pairs
  - Generate image & text embeddings

$$(I_1, T_1), \ldots, (I_N, T_N)$$

  - **Text.** Transformer
    - <EOS> token output
  - **Image.** ViT or ResNet
    - <CLS> token output



pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 … T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ … $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ … $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ … $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ … $I_N \cdot T_N$

Training

- Trained with InfoNCE loss
  - Simply an N-way classification loss

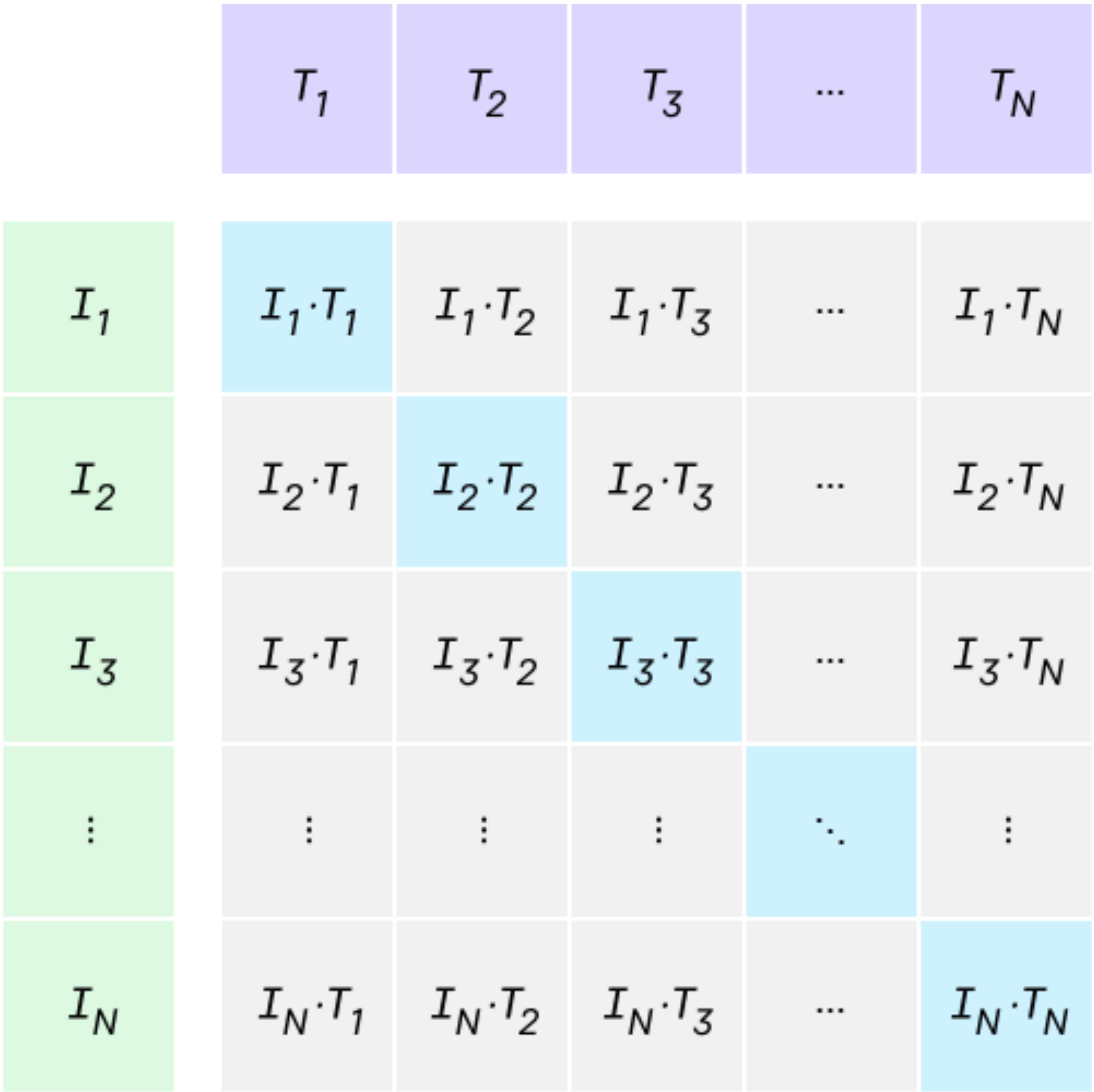
$$L = \frac{1}{N}\sum_{i=1}^{N}\frac{\ell(I_i) + \ell(T_i)}{2}$$

- The losses are:

$$\ell(I_i) = -\log\frac{\exp(I_i^\top T_i/\tau)}{\sum_j \exp(I_i^\top T_j/\tau)}$$

$$\ell(T_i) = -\log\frac{\exp(I_i^\top T_i/\tau)}{\sum_j \exp(I_j^\top T_i/\tau)}$$

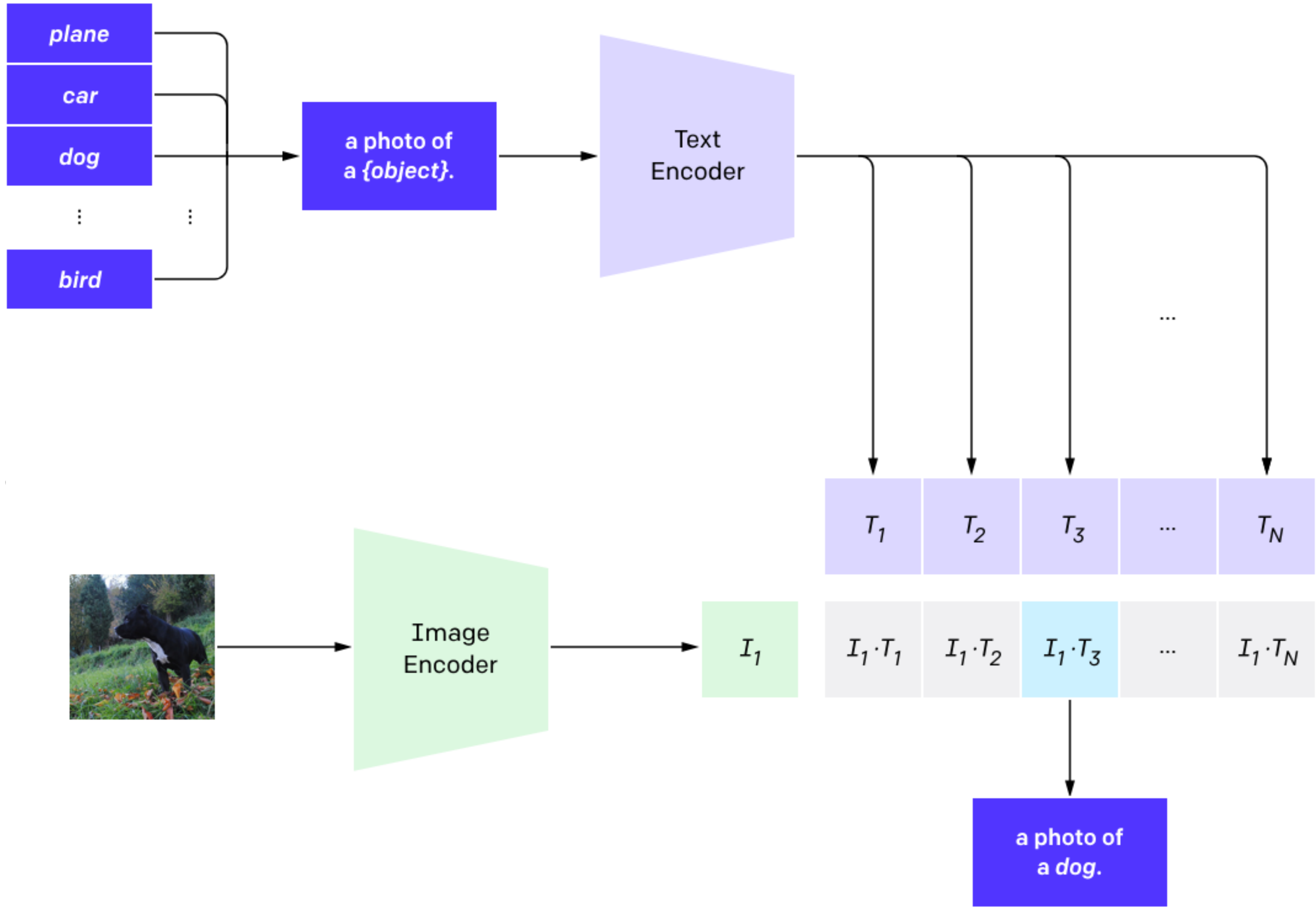
  - τ is some “temperature” hyperparameter



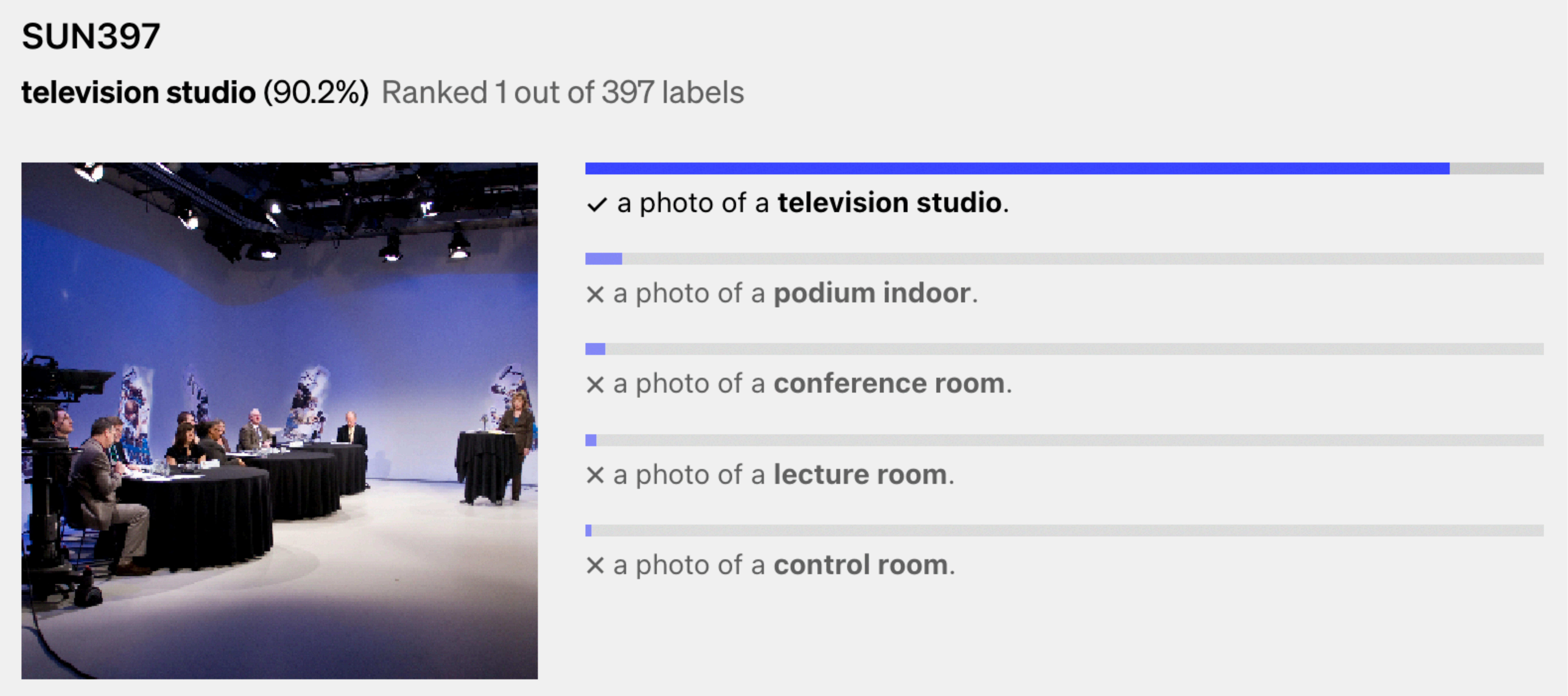
	T_1	T_2	T_3	…	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	…	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	…	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	…	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	…	$I_N \cdot T_N$

Off-the-shelf inference

- For an off-the-shelf inference, measure relevance of each image with prompt + classes
    - "This is an image of a/an {cat}"
- Allows open-set classification
    - Classification with unlimited set of target classes, configured by natural language
- **Note.** Prompt quality matters



Off-the-shelf inference



SUN397

television studio (90.2%) Ranked 1 out of 397 labels

✓ a photo of a **television studio**.

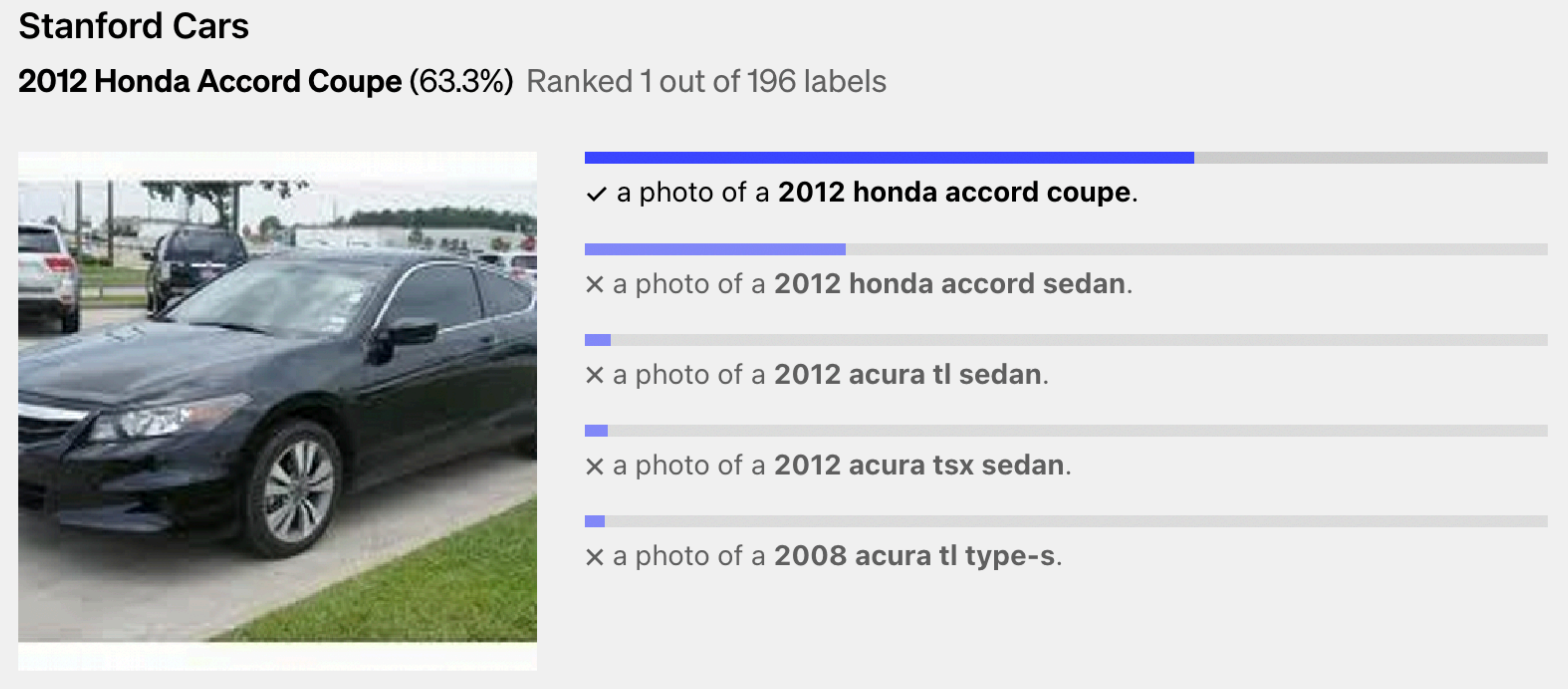
× a photo of a **podium indoor**.

× a photo of a **conference room**.

× a photo of a **lecture room**.

× a photo of a **control room**.

Off-the-shelf inference



Stanford Cars

2012 Honda Accord Coupe (63.3%) Ranked 1 out of 196 labels

✓ a photo of a **2012 honda accord coupe**.

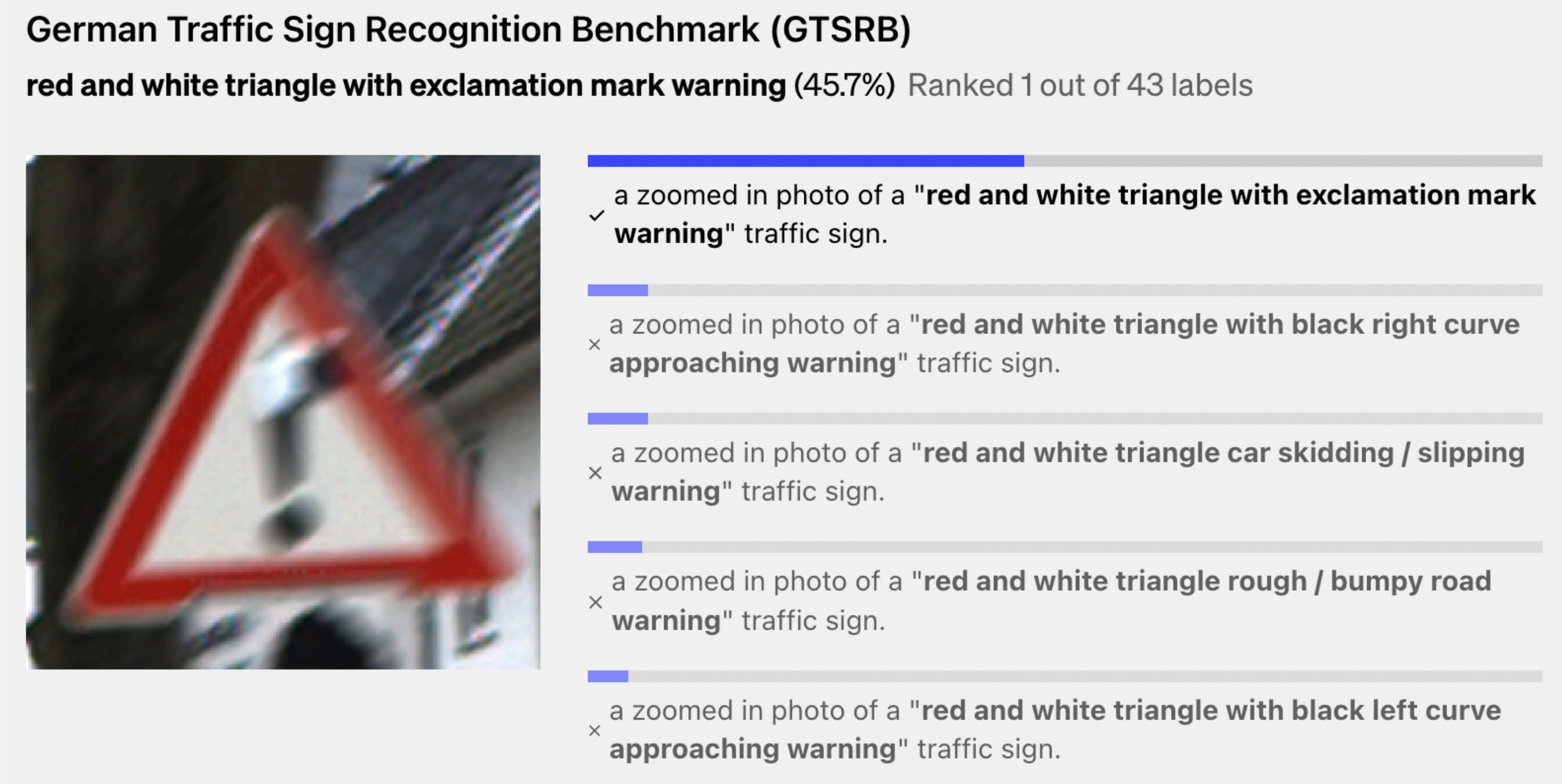
× a photo of a **2012 honda accord sedan**.

× a photo of a **2012 acura tl sedan**.

× a photo of a **2012 acura tsx sedan**.

× a photo of a **2008 acura tl type-s**.

Off-the-shelf inference



German Traffic Sign Recognition Benchmark (GTSRB)

red and white triangle with exclamation mark warning (45.7%) Ranked 1 out of 43 labels

✓ a zoomed in photo of a "**red and white triangle with exclamation mark warning**" traffic sign.

× a zoomed in photo of a "**red and white triangle with black right curve approaching warning**" traffic sign.

× a zoomed in photo of a "**red and white triangle car skidding / slipping warning**" traffic sign.

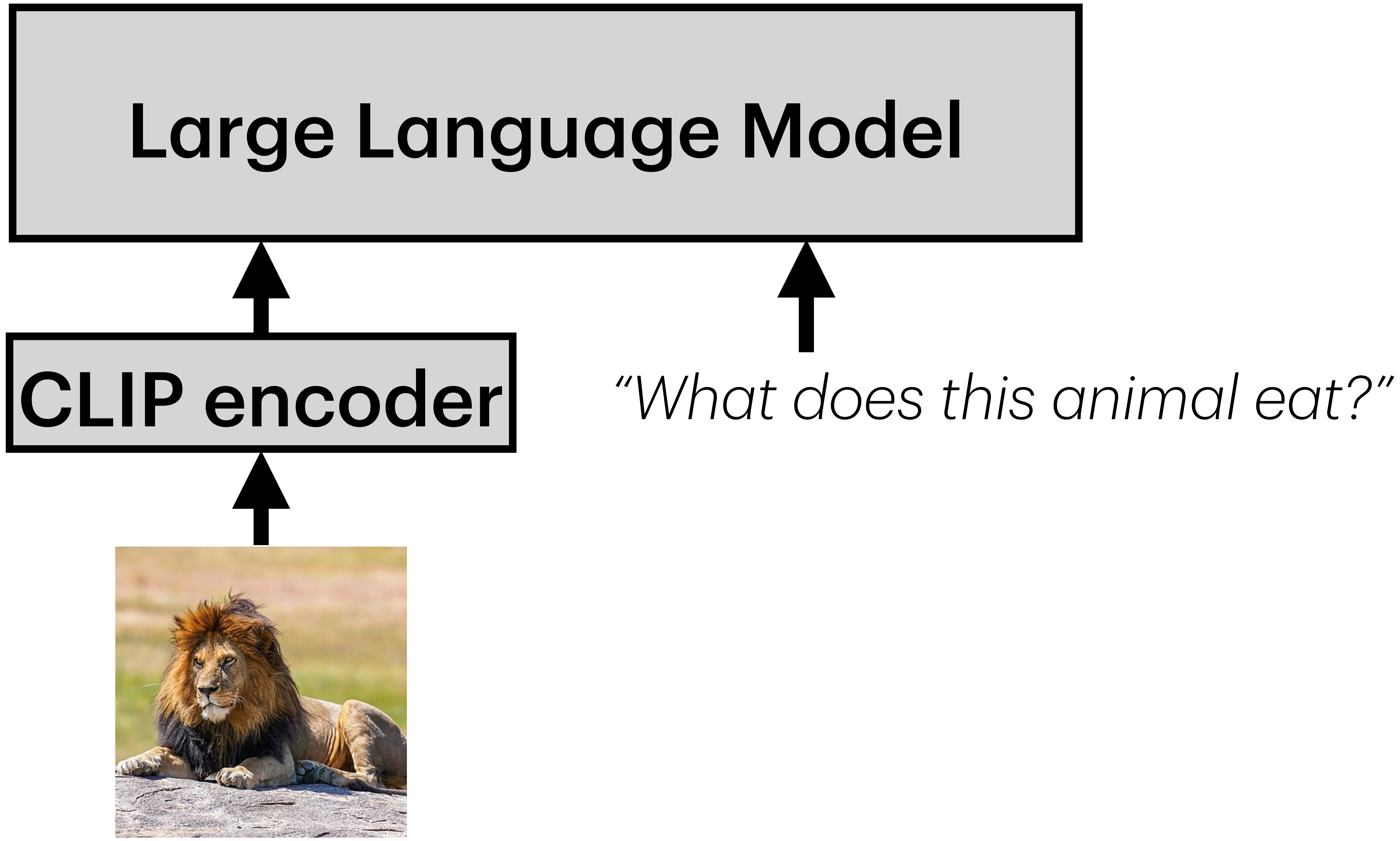
× a zoomed in photo of a "**red and white triangle rough / bumpy road warning**" traffic sign.

× a zoomed in photo of a "**red and white triangle with black left curve approaching warning**" traffic sign.

LLaVA

LLaVA

- **Idea.** Simply use the outputs of pretrained vision encoder as a prefix for the prompt of a pretrained LLM
  - Use CLIP as the vision encoder
- **Problem.** (1) LLMs features are not well-aligned with CLIP
  (2) LLMs have not be trained to do a visual Q&A



Large Language Model

CLIP encoder

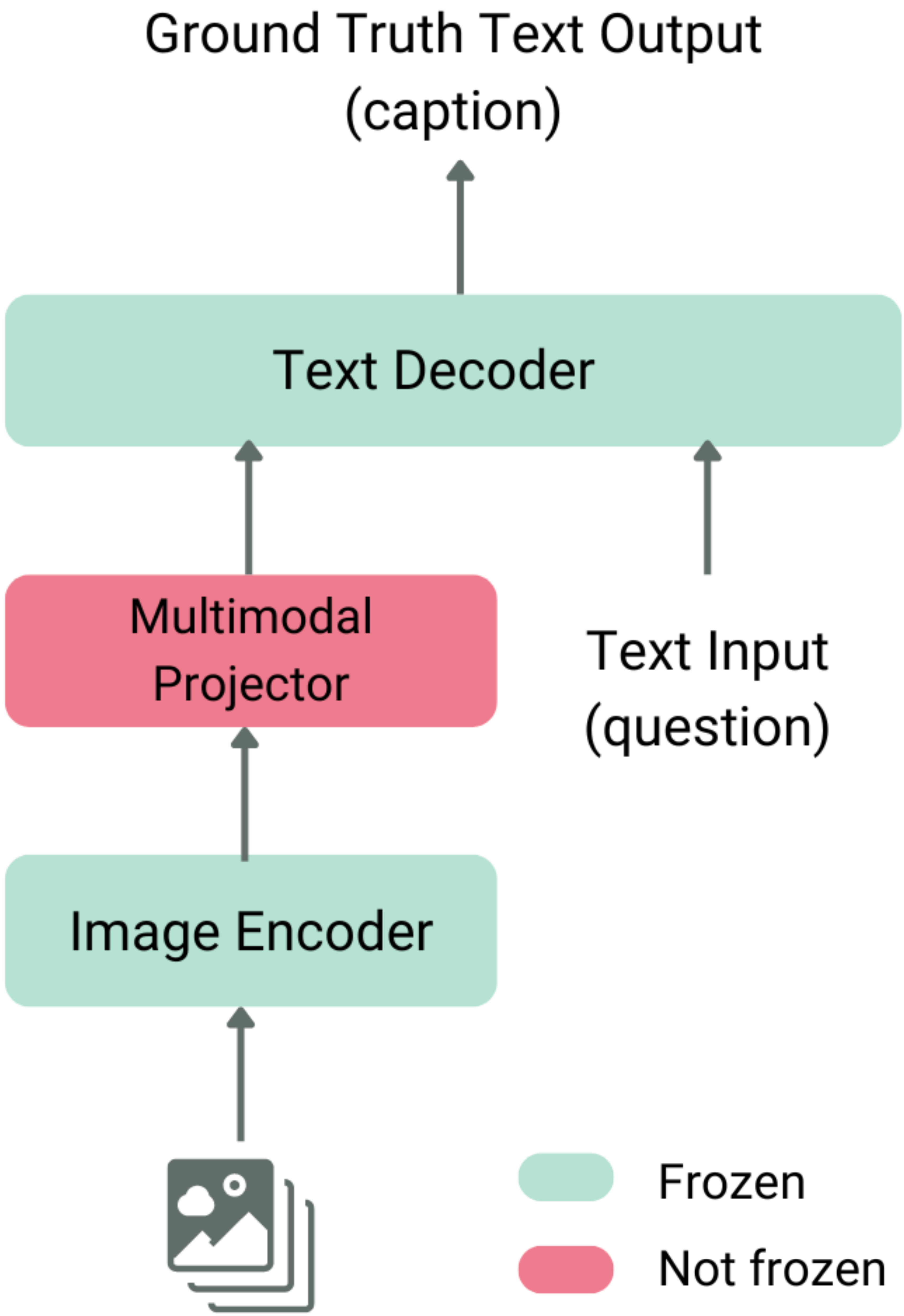
"What does this animal eat?"

LLaVA

Solution 1: Aligning

- Add a trainable projection layer between between CLIP & LLM
    - A single linear layer

- Pretrain with visual question-answering datasets (will be described shortly)
    - CLIP & LLM frozen



Ground Truth Text Output (caption)

Text Decoder

Multimodal Projector

Text Input (question)

Image Encoder

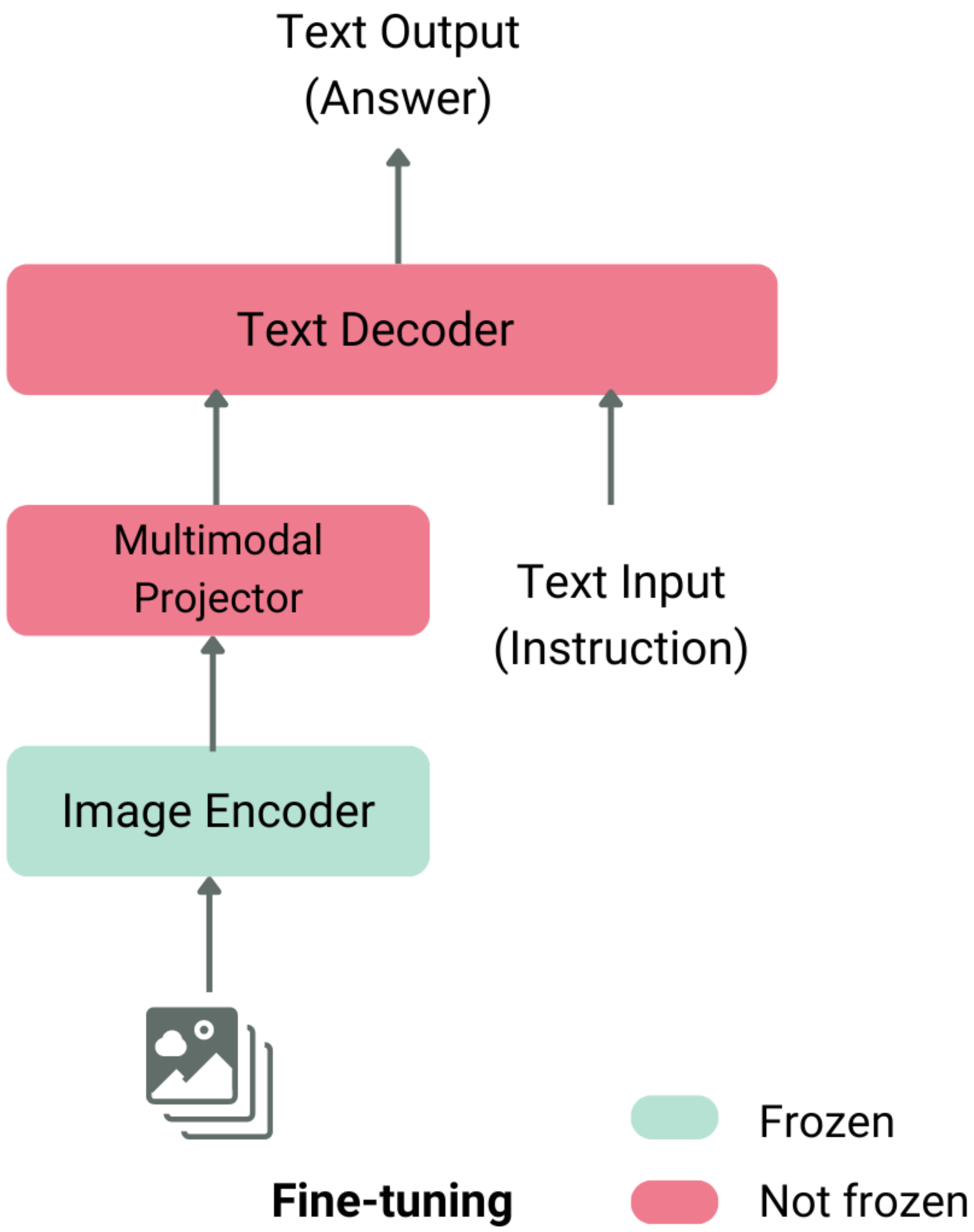
🟩 Frozen

🟥 Not frozen

LLaVA

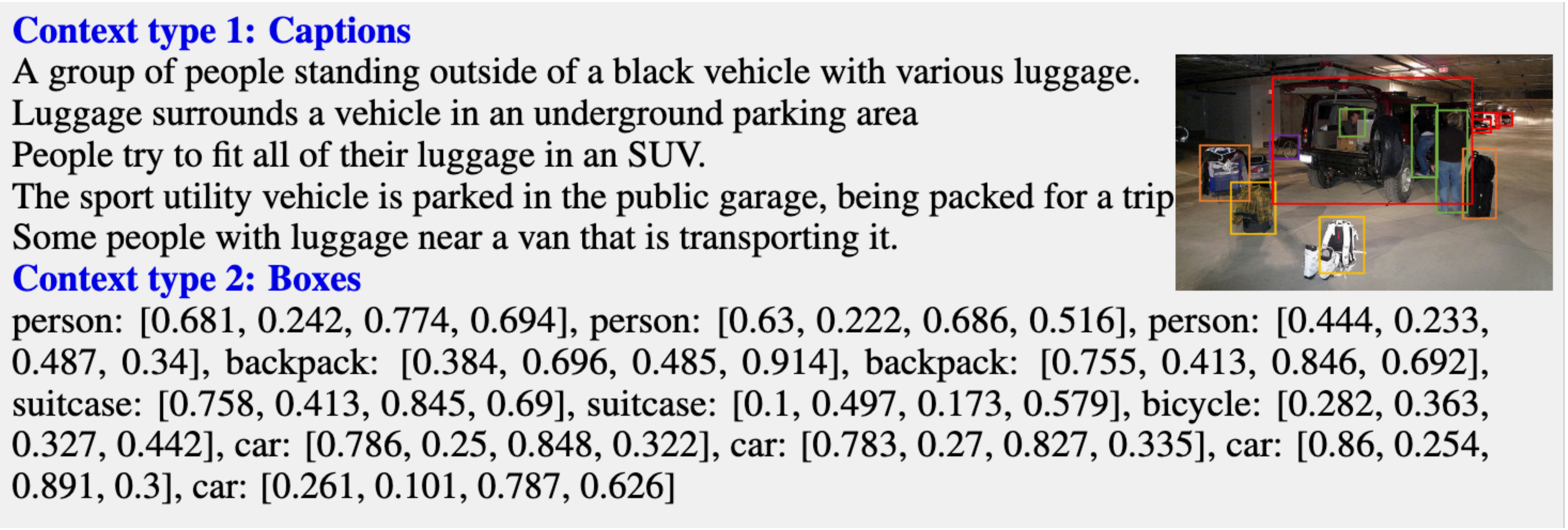
Solution 2: Fine-tuning

- Fine-tune LLM with visual Q&A datasets
  - Shorter training
    - Preserve LLM knowledge
    - Less training cost



VQA Dataset

- Visual instruction tuning dataset is collected using text-only GPT
  - GPT is provided with textual descriptions of an image
    - Captions and bounding boxes



Context type 1: Captions

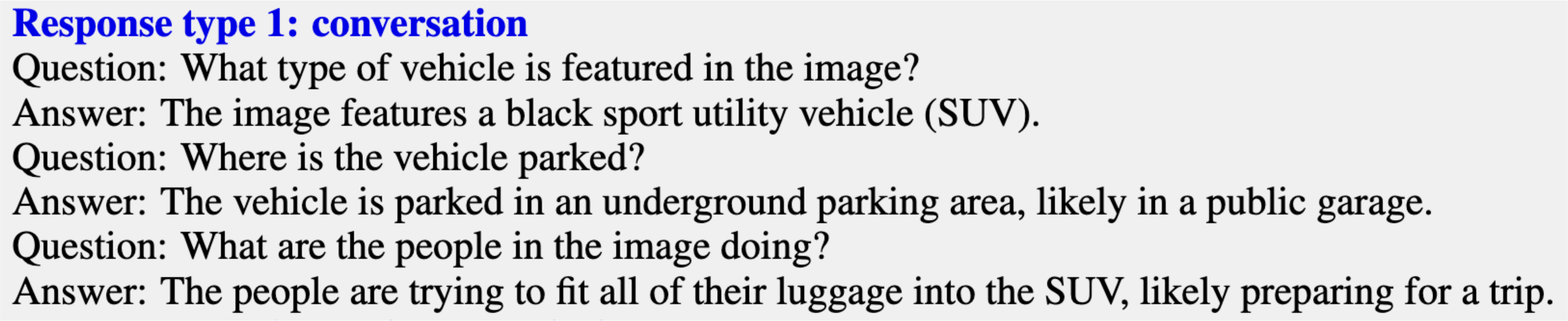
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

VQA Dataset

- Given this context, GPT is prompted to generate three responses:
- **1. Q&A Conversation.** GPT simulates both the person who asks and person who answers to generate multi-turn conversation

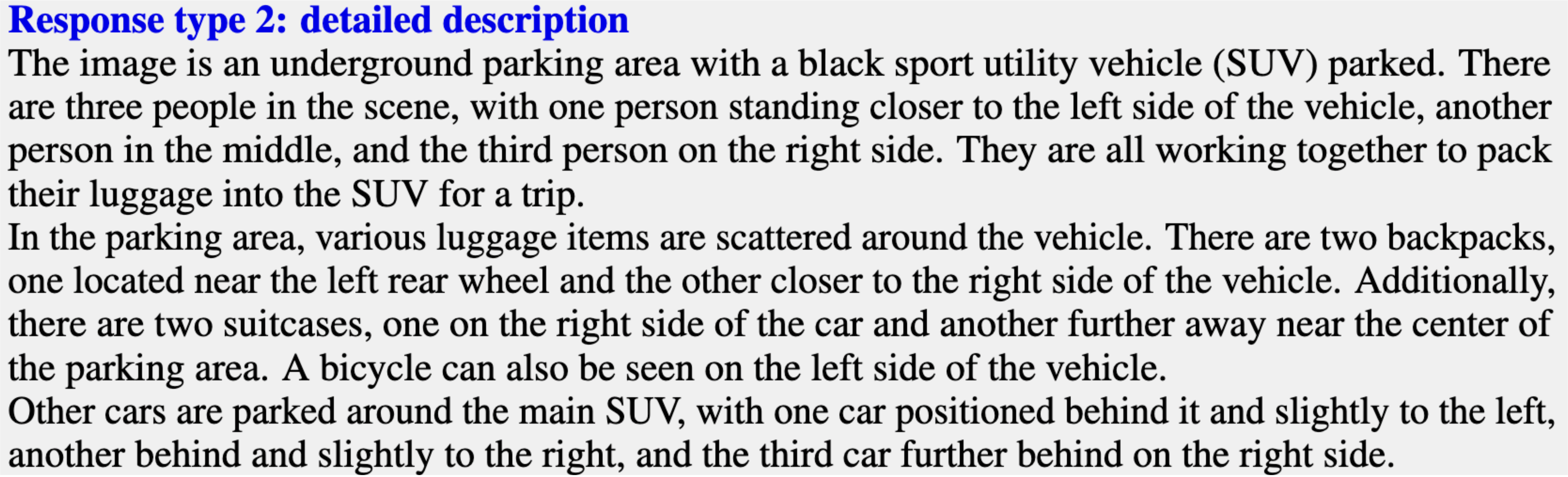


Response type 1: conversation

Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

VQA Dataset

- **2. Detailed description.** GPT generates a very detailed description of the image, using the bounding box information to fill in the details



Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip.

In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.

Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

VQA Dataset

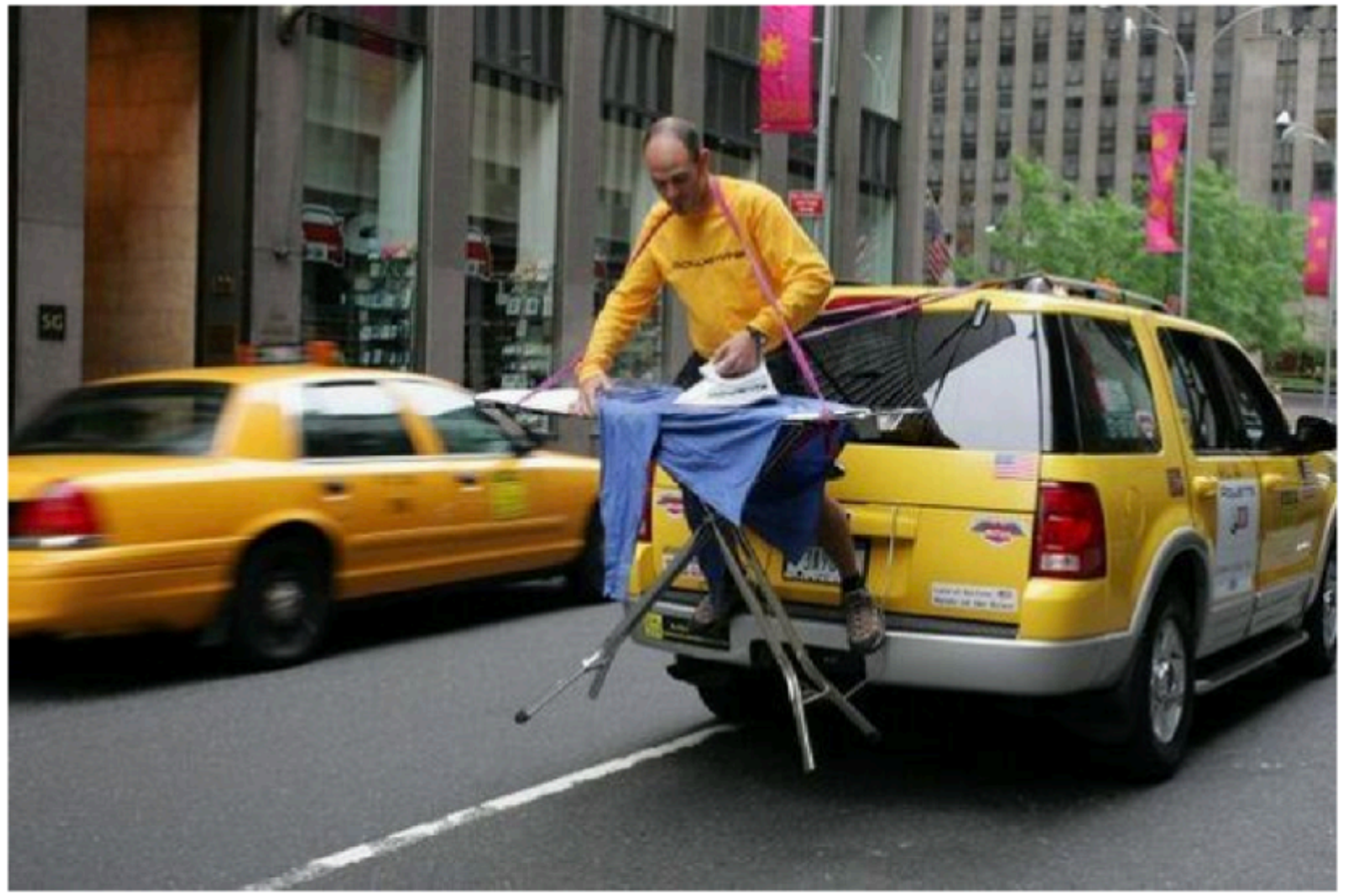
- **3. Complex Reasoning.** GPT generates both the question and the answer that needs in-depth understanding of the content of the image

Response type 3: complex reasoning

Question: What challenges do these people face?

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.

Final output



Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

Application: Robotic Control

Robotic Control

- A more recent application of VLM is the robotic control
  - Multimodal intelligence can lead to a general-purpose robot

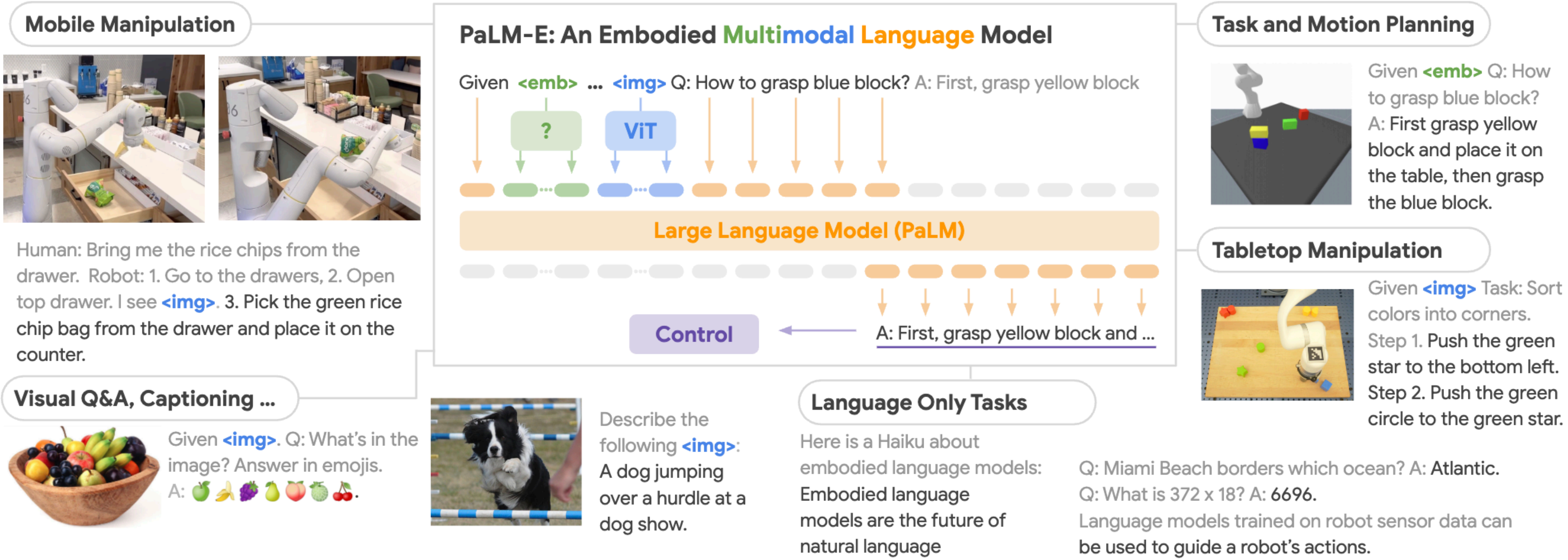
Helix: A Vision-Language-Action Model for Generalist Humanoid Control

Robotic Control

- There are many different ways to achieve this goal
  - VLM for high-level planning
  - VLA for end-to-end generation of action sequences
  - Video generation and World Models

High-level planning

- Given sensory inputs and high-level instructions, generate detailed instructions for the controller
  - **Example.** PaLM-E


Mobile Manipulation
PaLM-E: An Embodied Multimodal Language Model
?
ViT
Large Language Model (PaLM)
Control
Task and Motion Planning
Tabletop Manipulation
Visual Q&A, Captioning ...
Language Only Tasks

Given <emb> ... <img> Q: How to grasp blue block? A: First, grasp yellow block

A: First, grasp yellow block and ...

Human: Bring me the rice chips from the drawer. Robot: 1. Go to the drawers, 2. Open top drawer. I see <img>. 3. Pick the green rice chip bag from the drawer and place it on the counter.

Given <emb> Q: How to grasp blue block? A: First grasp yellow block and place it on the table, then grasp the blue block.

Given <img> Task: Sort colors into corners. Step 1. Push the green star to the bottom left. Step 2. Push the green circle to the green star.

Given <img>. Q: What's in the image? Answer in emojis. A: 🍏🍌🍇🍐🍑🍈🍒.

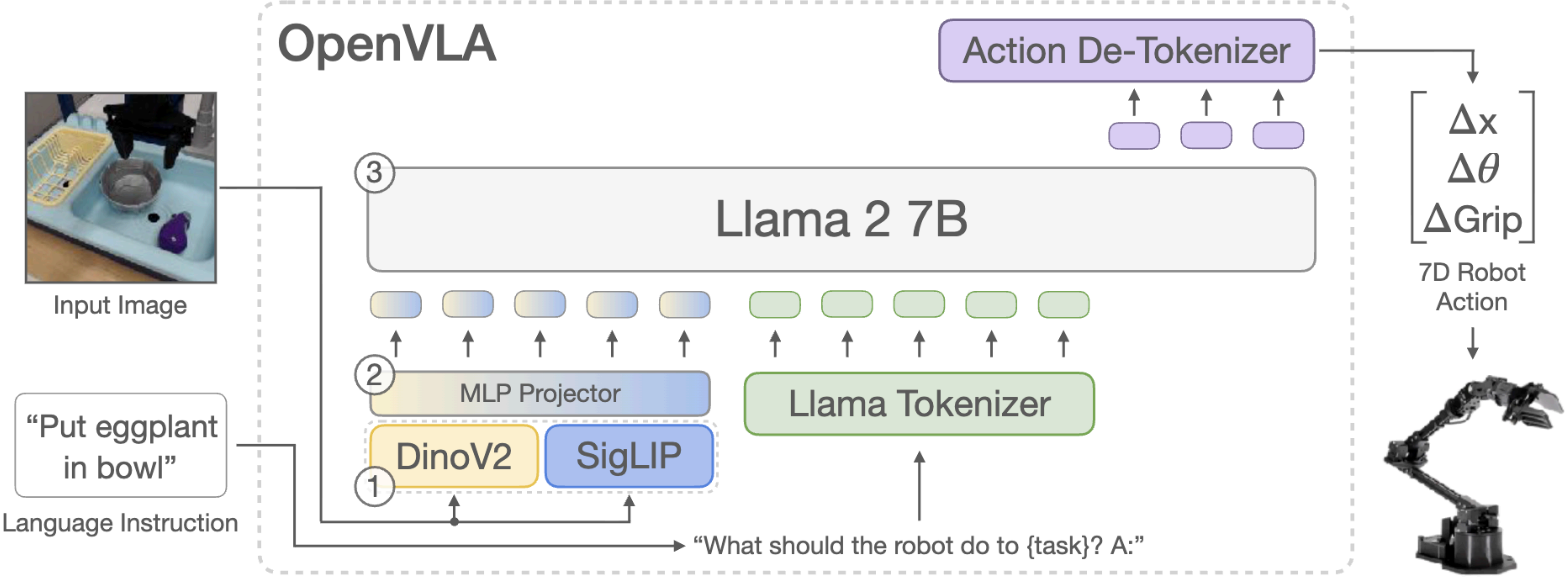
Describe the following <img>: A dog jumping over a hurdle at a dog show.

Here is a Haiku about embodied language models: Embodied language models are the future of natural language

Q: Miami Beach borders which ocean? A: Atlantic. Q: What is 372 x 18? A: 6696. Language models trained on robot sensor data can be used to guide a robot's actions.

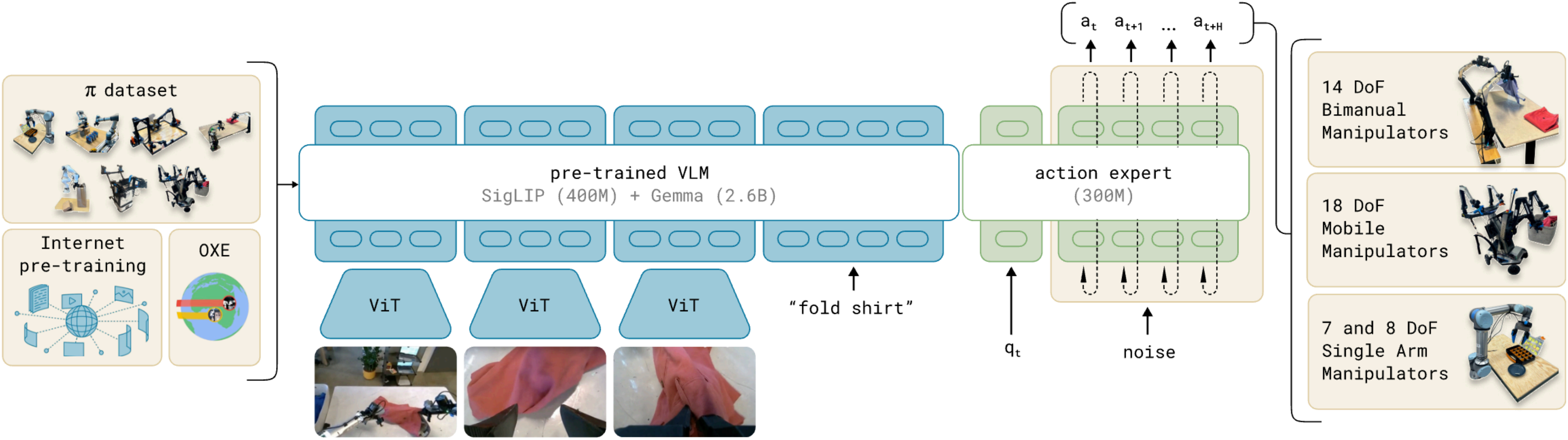
Vision-Language-Action Models

- Generate action sequences directly as model outputs
- **Example.** RT-2, OpenVLA, Pi-0/0.5/0.6



Vision-Language-Action Models

- More frequently, uses diffusion models as an “action expert”
    - More refined way to generate actions (e.g., velocities / torques)


π dataset
Internet pre-training
OXE
pre-trained VLM
SigLIP (400M) + Gemma (2.6B)
ViT
ViT
ViT
“fold shirt”
action expert
(300M)
a_t a_{t+1} ... a_{t+H}
q_t
noise
14 DoF Bimanual Manipulators
18 DoF Mobile Manipulators
7 and 8 DoF Single Arm Manipulators

Next class

- Post-training of LLMs
  - Retrieval-augmented generation
  - Reasoning models
  - Alignment
  - Acceleration
  - Agentic AI

</lecture 21>