MMLM

Multi-Modal Language Models

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.

https://huggingface.co/blog/vlms

Video-Chat

Query:
What did the person on the left and the person on the right say, respectively?

Response:

(The person on the left says, "Welcome everyone to the show."
The person on the right says,
"I'm very happy that everyone can watch our show.")

Text-Chat

Query:
Mother's Day is coming soon. Can you help me write a message for my mom?

Response:

("Mom, Mother's Day is almost here, and I just wanted to tell you how much you mean to me. You've always been there for me, no matter what. Your love has shaped who I am today. Thank you for everything you've done for me. I hope this Mother's Day brings you so much joy and happiness.")

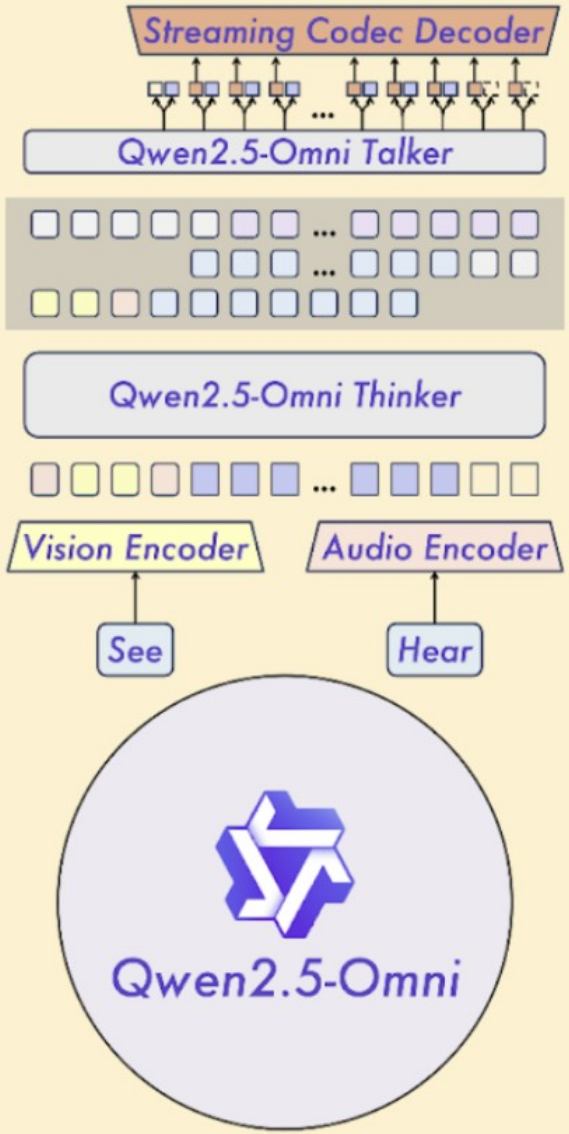
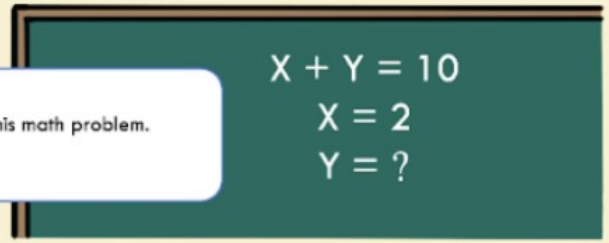


Image-Chat

Query:
Please help me solve this math problem.



Response:

(Well, since X + Y = 10 and X = 2, you can just substitute the value of X into the first equation. So it becomes 2 + Y = 10. Then, if you subtract 2 from both sides, you get Y = 8.So, Y equals 8.)

Audio-Chat

Query:
Please provide a brief description of this music.

Response:

(it's a pop song in A major with a 4/4 time signature. The chord progression mainly alternates between A major and D major. It has a tempo of about 90 BPM)

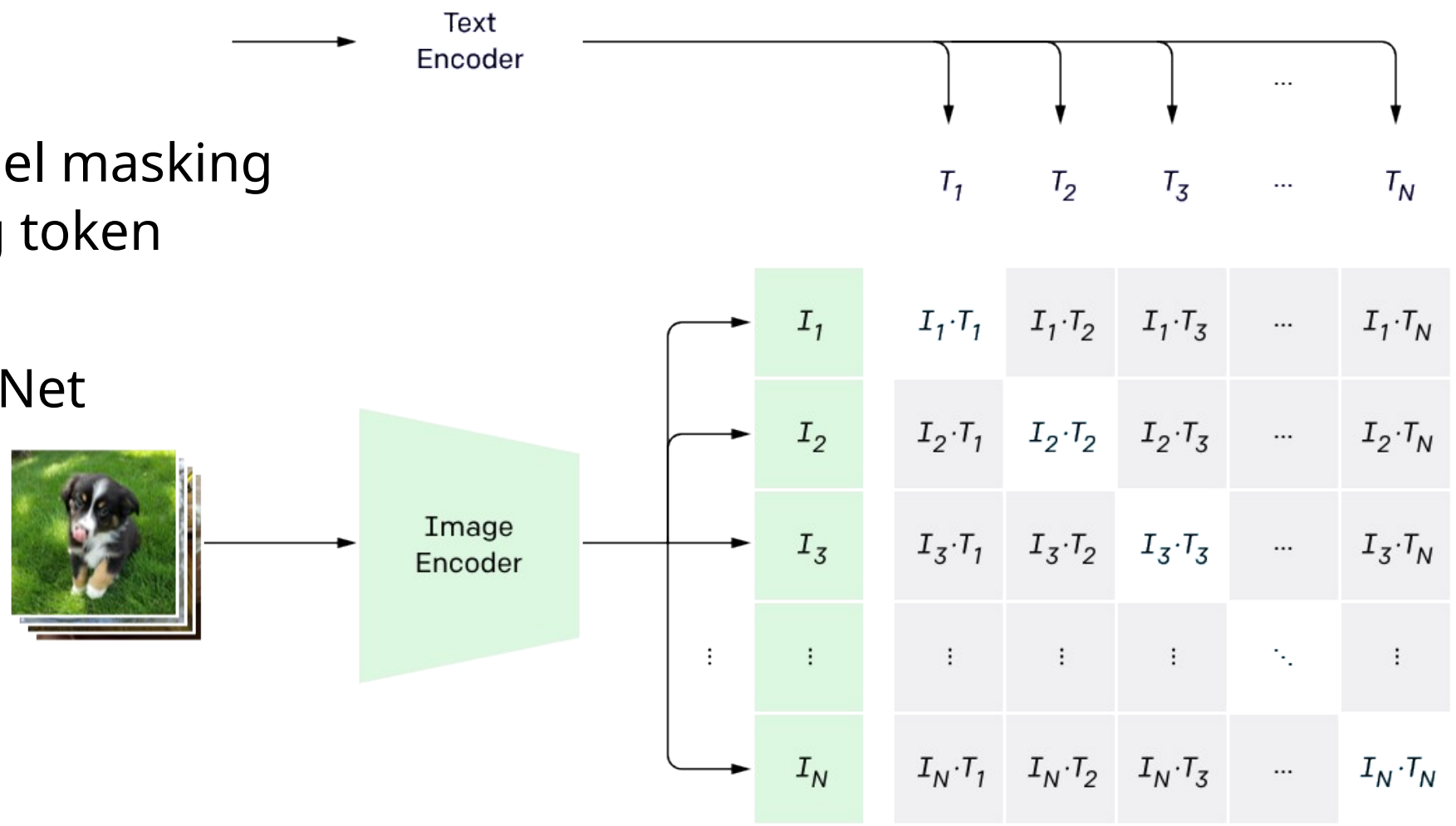
https://huggingface.co/blog/vlms-2025

CLIP model https://openai.com/index/clip/

- CLIP (*Contrastive Language–Image Pre-training*)
- Learning Transferable Visual Models From Natural Language Supervision, ICML, 2021
- Image-text alignment model
  - Not an image generation model
  - Not a text generation model
  - Image classification;
- Pretraining dataset
  - 400M (image, text) pairs from Web

Contrastive pre-training

- Text encoder
  - Transformer
  - Language Model masking
  - Use the Ending token
- Image
  - a modified ResNet



Contrastive (Self-Supervised) Learning

- Self-supervised learning
  - using the data itself to generate supervisory signals
  - BERT, GPT, auto-encoder
- Training data
  - Positive: same image
  - Negative: different images
- Contrastive loss
  - minimize the distance between positive sample pairs
  - maximizing the distance between negative sample pairs

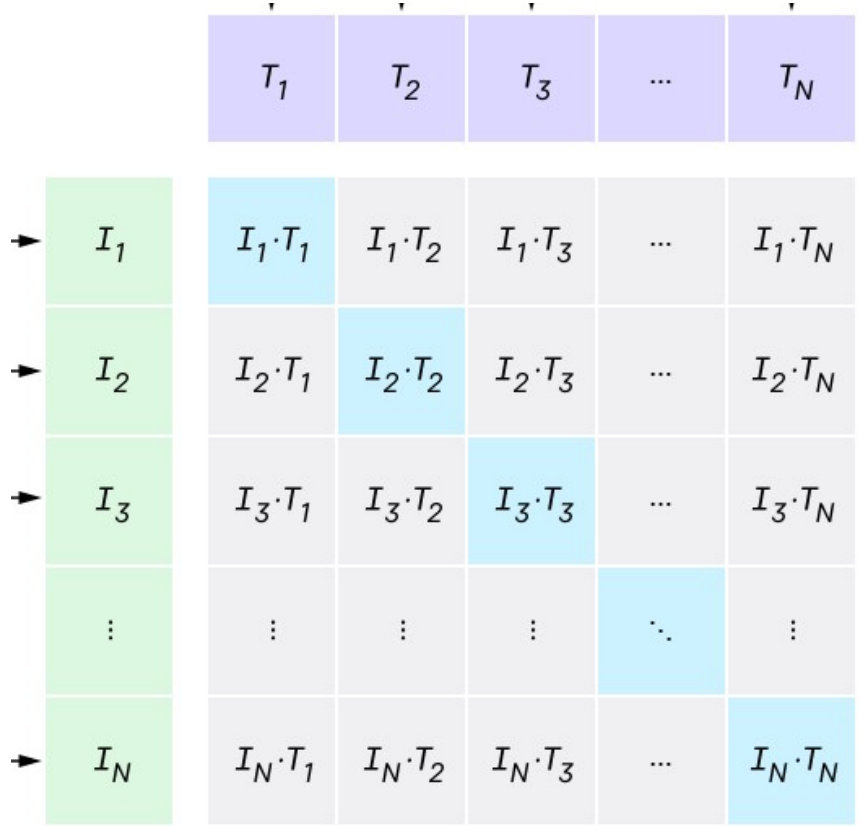
InfoNCE (Noise-Contrastive Estimation)

- *query representation*
- *:* positive key
- *:* negative key
- *:* temperature, controls the sharpness of the similarity distribution
- sim: similarity

Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." *arXiv preprint arXiv:1807.03748* (2018).

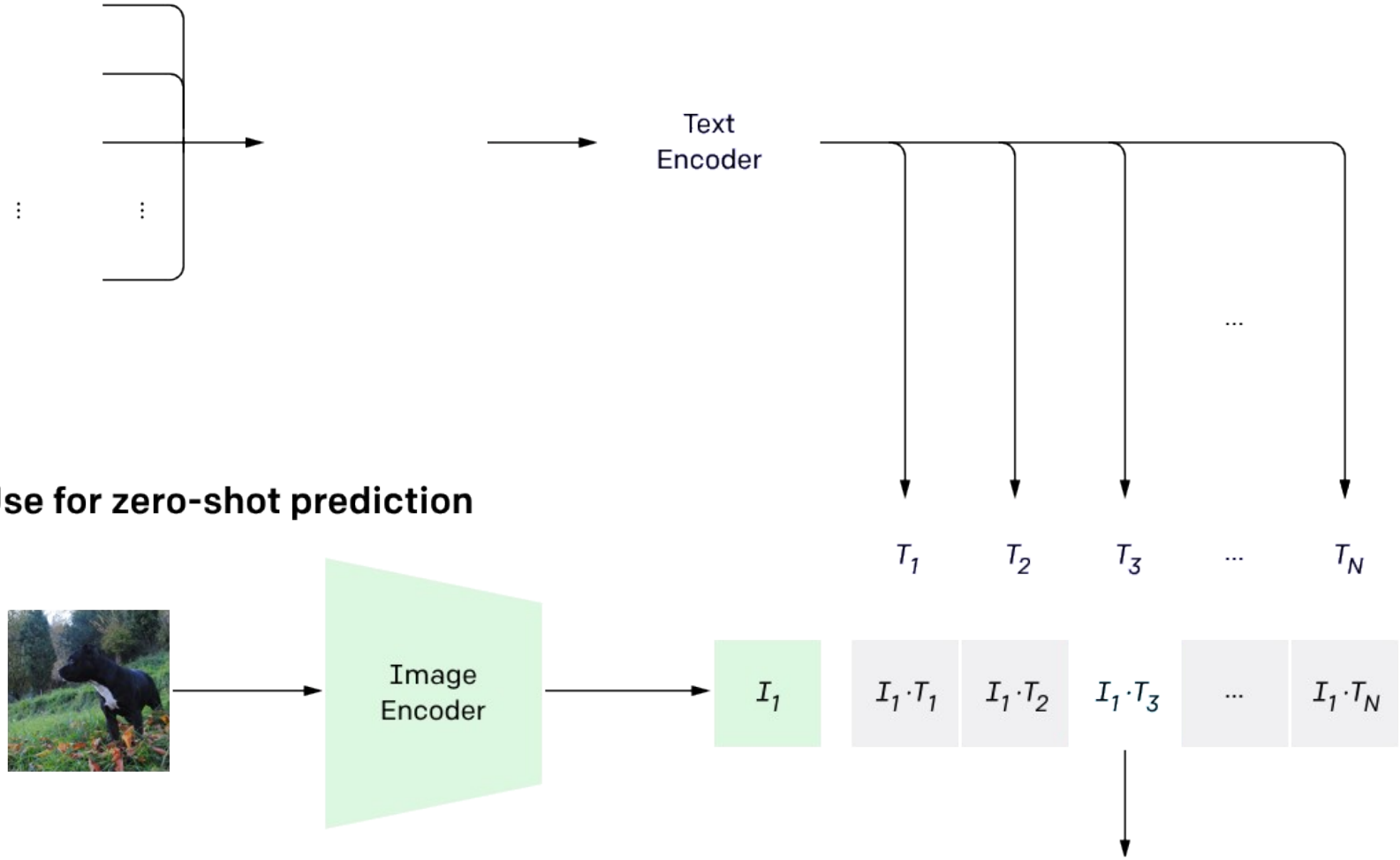
Contrastive pre-training

- Image to text
- Text to image
- Final loss: mean



Zero-shot Classification

- Generate text embeddings for class names or descriptions
- Image align with potential caption



LLaVA model

https://llava-vl.github.io/

- LLaVA (Large Language and Vision Assistant)
- Visual Instruction Tuning, NeurIPS, 2023
- visual and language understanding
  - Not an image generation model
  - Is a text generation model
  - Vision QA

LLaVA model

- Input
  - Visual Input
  - Textual Instructions: questions or commands
- Output
  - Descriptive Text: describe the visual content
  - Answers to Questions
  - Follow-up Actions: editing, retrieving, etc

1. Visual Instruction Data Generation: GPT-4
2. Visual Instruction Tuning

Visual Instruction Data Generation

```
messages = [ {"role":"system", "content": f"""You are an AI visual assistant, and you are seeing a single image. What you see are provided with five sentences, describing the same image you are looking at. Answer all questions as you are seeing the image.

Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers.

Include questions asking about the visual content of the image, including the object types, counting the objects, object actions, object locations, relative positions between objects, etc. Only include questions that have definite answers:
(1) one can see the content in the image that the question asks about and can answer confidently;
(2) one can determine confidently from the image that it is not in the image. Do not ask any question that cannot be answered confidently.

Also include complex questions that are relevant to the content in the image, for example, asking about background knowledge of the objects in the image, asking to discuss about events happening in the image, etc. Again, do not ask about uncertain details. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary."""}
]
for sample in fewshot_samples:
    messages.append({"role":"user", "content":sample['context']})
    messages.append({"role":"assistant", "content":sample['response']} )
messages.append({"role":"user", "content":'\n'.join(query)})
```

visual conversation data

Captions
There is a movie theater that displays the show times above the doors.
A red fire hydrant is deep in the snow.
The fire hydrant is in the snow near a recently plowed sidewalk.
This city has had a very hard winter with snow.
A hotel for dogs in the snow in winter.

Conversation
Question:
What color is the fire hydrant in the image?
===
Answer:
The fire hydrant in the image is red.
===
Question:
Is the fire hydrant fully covered by snow or partially visible?
===
Answer:
The fire hydrant is partially visible, with a significant portion of it covered by the deep snow.
===
Question:
What is the condition of the sidewalk near the fire hydrant?
===
Answer:
The sidewalk near the fire hydrant has been recently plowed, so it is relatively clear of snow compared to the surrounding area.
===
Question:
What does the image suggest about the winter season in the city?
===
Answer:
The image suggests that the winter season in the city has been quite harsh and challenging. The significant snow accumulation seen in the image indicates that the city has experienced heavy snowfall, which can lead to various difficulties and disruptions for its residents.

For instance, the deep snow covering the fire hydrant could potentially hinder its accessibility in case of emergencies, which poses safety concerns. The recently plowed sidewalk implies that the city's maintenance crews have been actively working to keep the walkways clear and safe for pedestrians, but the sheer amount of snowfall might make it difficult to keep up with the continuous snow removal efforts.

Furthermore, such extreme winter conditions can impact transportation, leading to traffic issues, delays in public transit, and increased risks of accidents due to slippery roads. It can also cause problems for businesses and schools, as people might face challenges commuting to work or attending classes. Additionally, the heavy snow can put extra strain on infrastructure, such as roofs and power lines, increasing the likelihood of structural damage or power outages.

In conclusion, the image of the red fire hydrant deep in the snow and the recently plowed sidewalk suggest that the city has faced a particularly severe winter season, with substantial snowfall that has likely caused various challenges and disruptions for its residents and infrastructure.

instruction-following data

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.

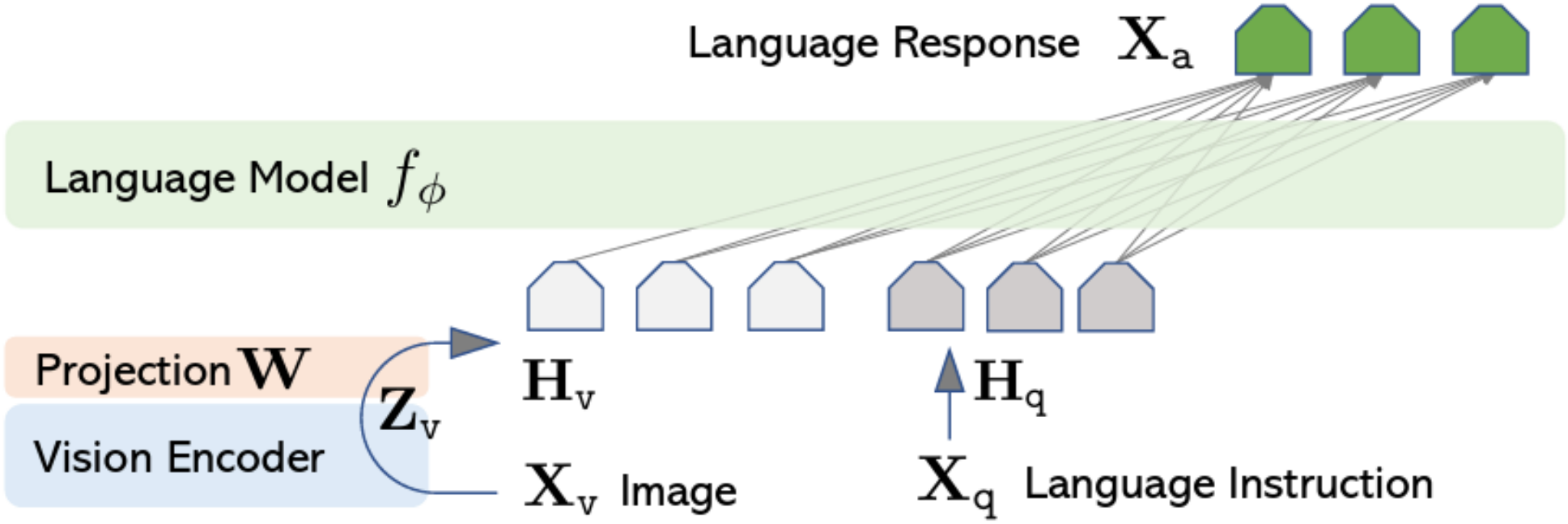
Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 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]

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.

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.

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.

Visual Instruction Tuning



- Vision encoder: clip
- W: trainable matrix to convert visual feature into language embedding tokens

$\mathbf{X}_{\texttt{system-message}}$ <STOP>
Human : $\mathbf{X}^1_{\texttt{instruct}}$ <STOP> Assistant: $\mathbf{X}^1_{\texttt{a}}$ <STOP>
Human : $\mathbf{X}^2_{\texttt{instruct}}$ <STOP> Assistant: $\mathbf{X}^2_{\texttt{a}}$ <STOP> $\cdots$

Table 2: The input sequence used to train the model. Only two conversation turns are illustrated here; in practice, the number of turns varies based on the instruction-following data. In our current implementation, we follow Vicuna-v0 [9] to set the system message $\mathbf{X}_{\texttt{system-message}}$ and we set <STOP> = ###. The model is trained to predict the assistant answers and where to stop, and thus only green sequence/tokens are used to compute the loss in the auto-regressive model.

- Pre-training for Feature Alignment
  - filtering a large dataset to a refined set of image-text pairs
  - visual encoder and LLM weights frozen (only train W)
  - Caption generation task
- Fine-tuning End-to-End
  - Visual encoder weights frozen
  - Update LLM and W
  - Chatbot and Science QA