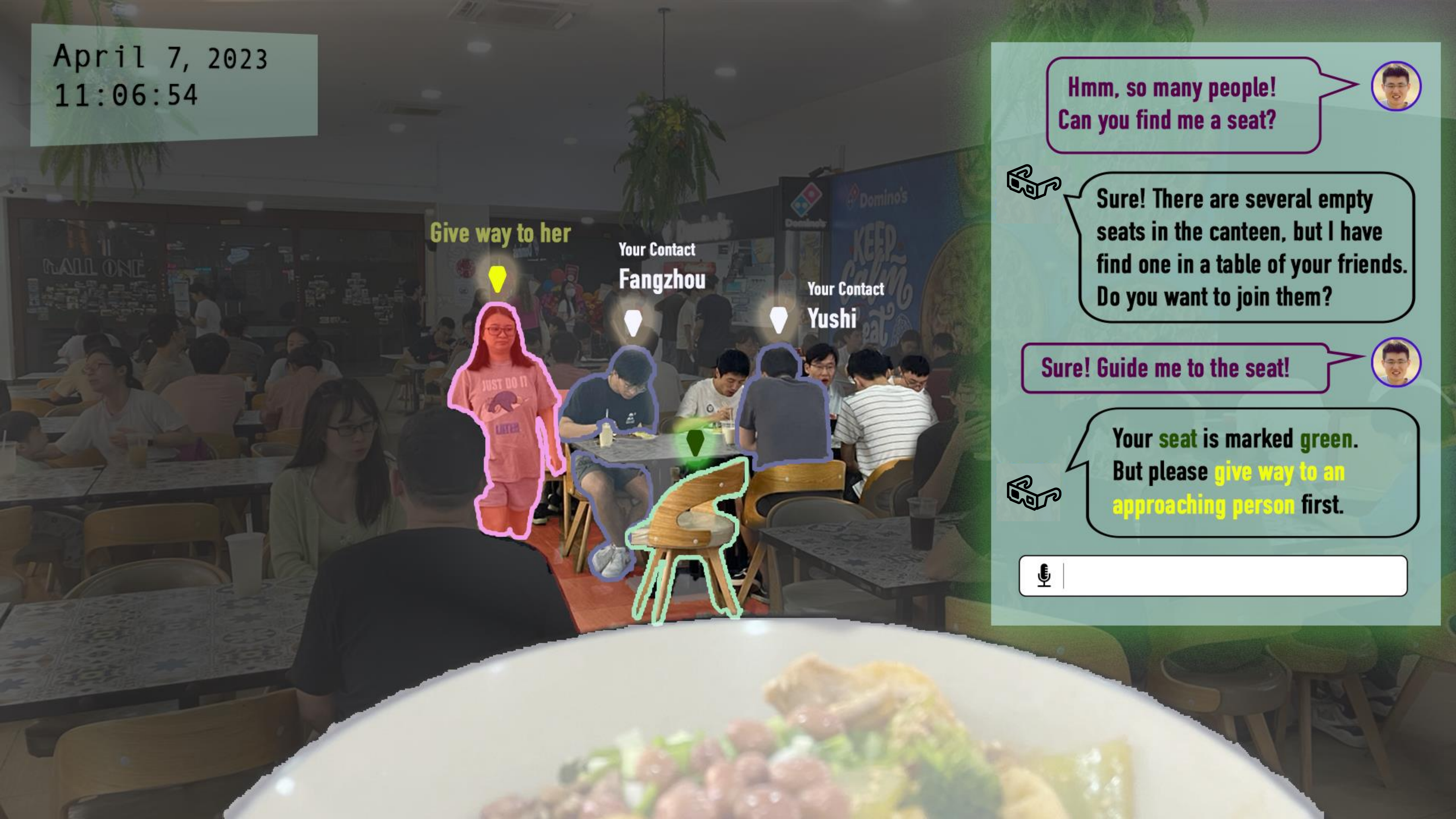
Towards Building Practical AI Assistant

Ziwei Liu

Nanyang Technological University



April 7, 2023
11:06:54

Hmm, so many people! Can you find me a seat?

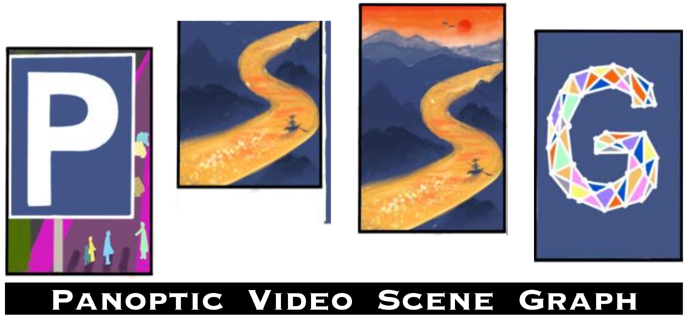
Sure! There are several empty seats in the canteen, but I have find one in a table of your friends. Do you want to join them?

Sure! Guide me to the seat!

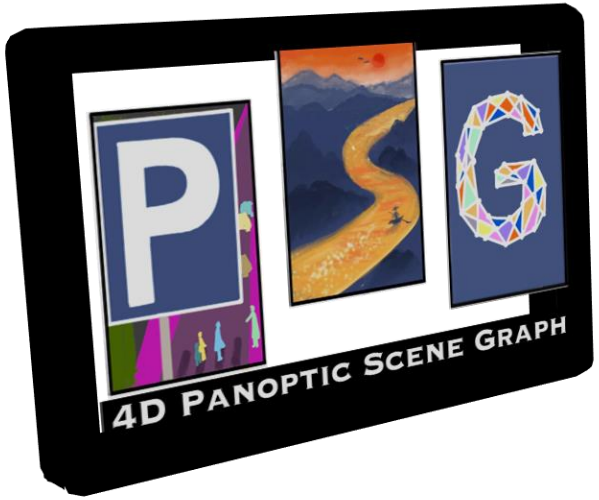
Your seat is marked green. But please give way to an approaching person first.

AI Assistant with Scene Graph

2D Image

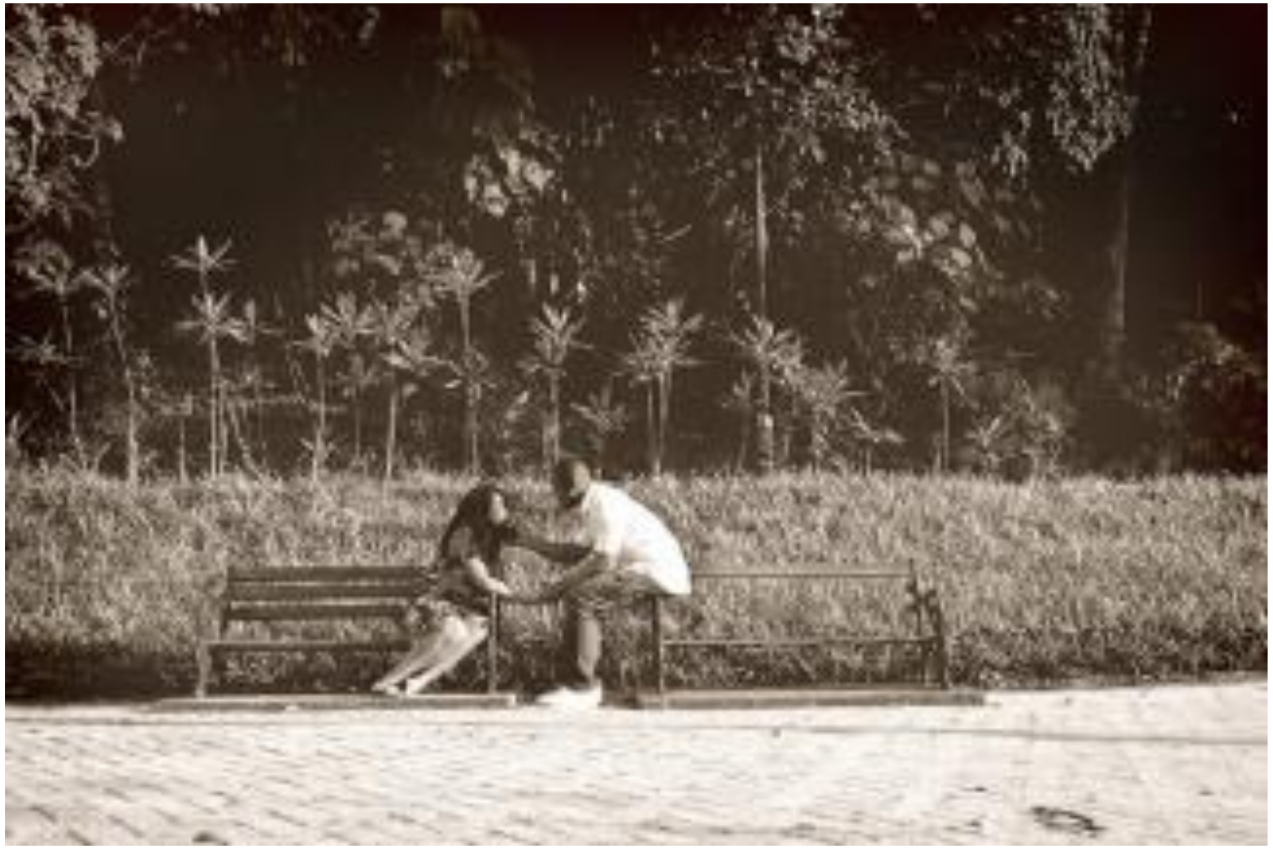


2D Video



3D video

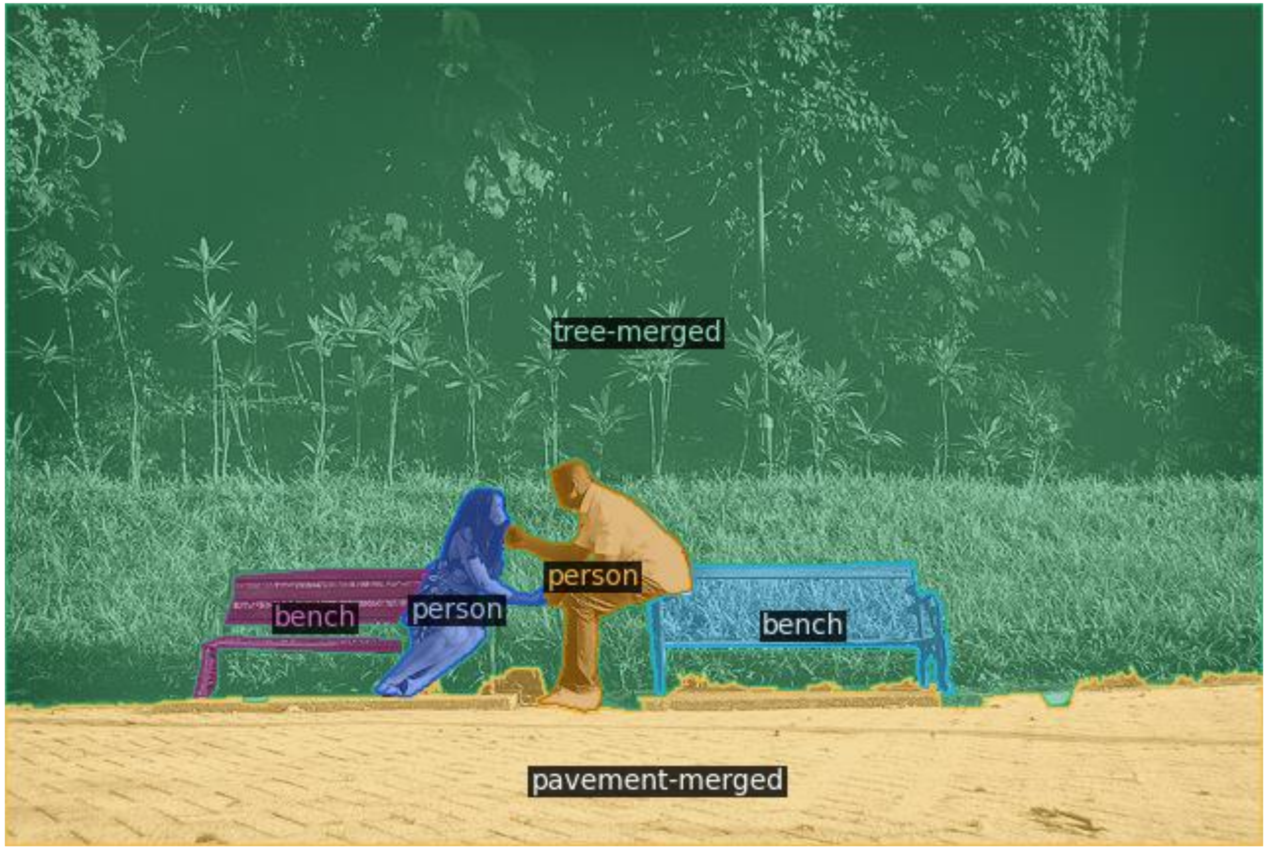
Beyond Object Recognition



Q: What is in the image?

Q: What happened in the image?

Beyond Object Recognition

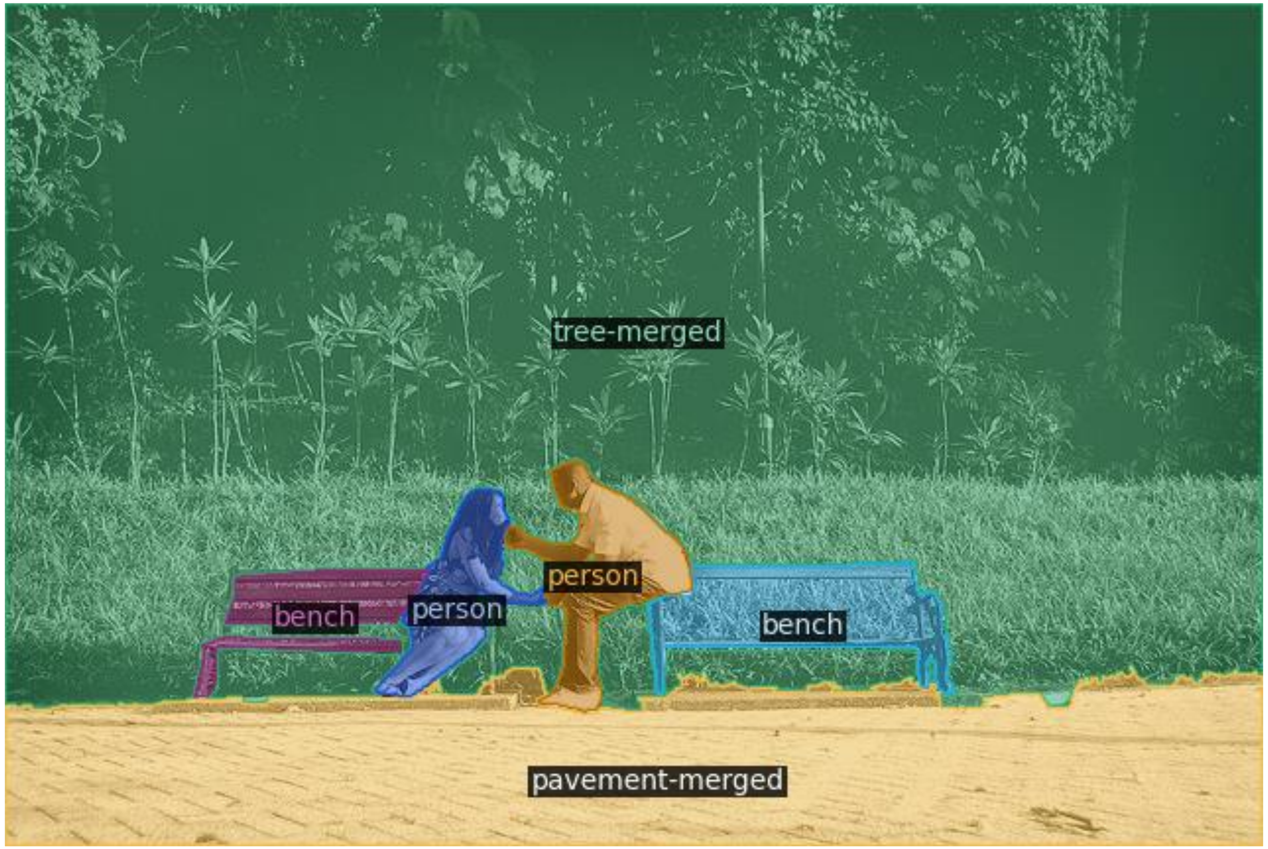


Q: What is in the image?

A: 2 x person, 2 x bench, tree, and pavement

Q: What happened in the image?

Beyond Object Recognition

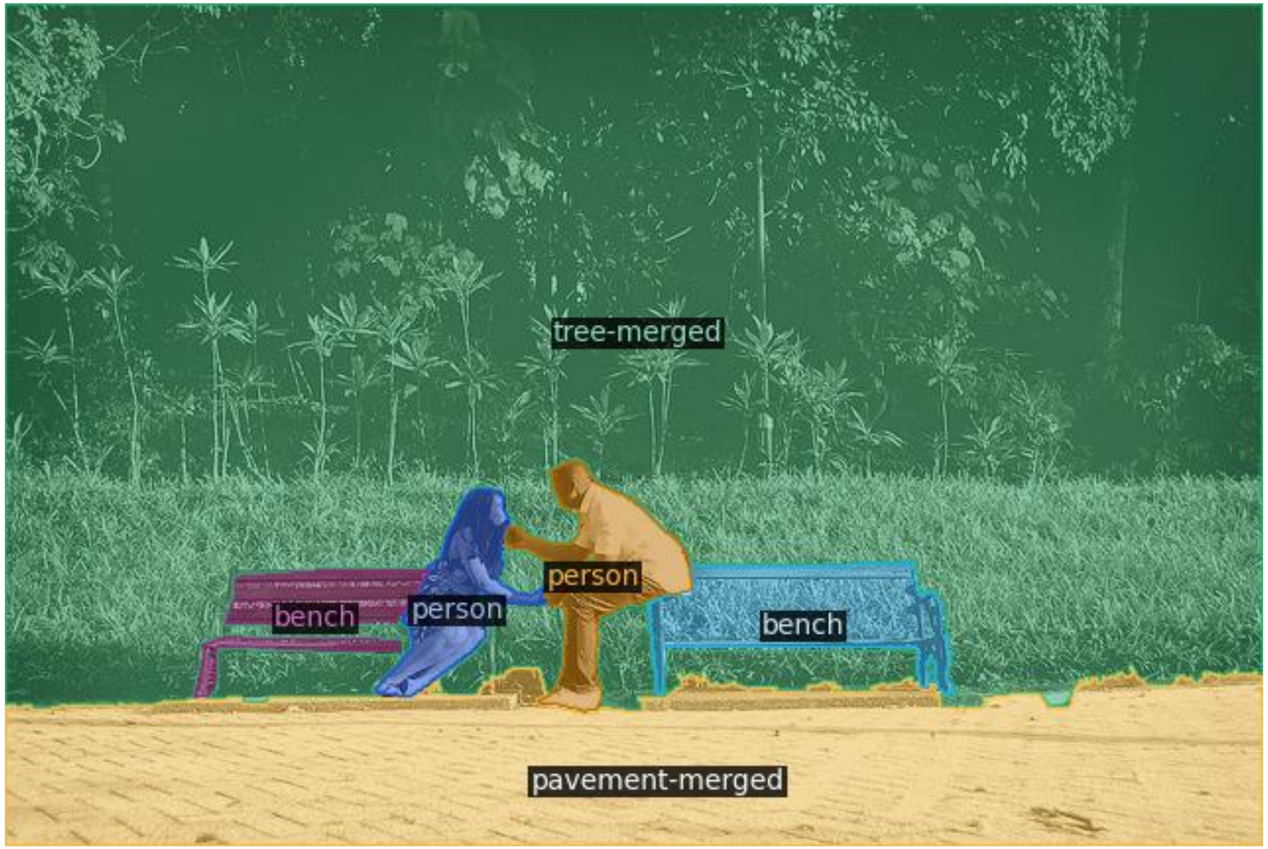


Q: What is in the image?

A: <u>2 x person, 2 x bench, tree, and pavement</u>

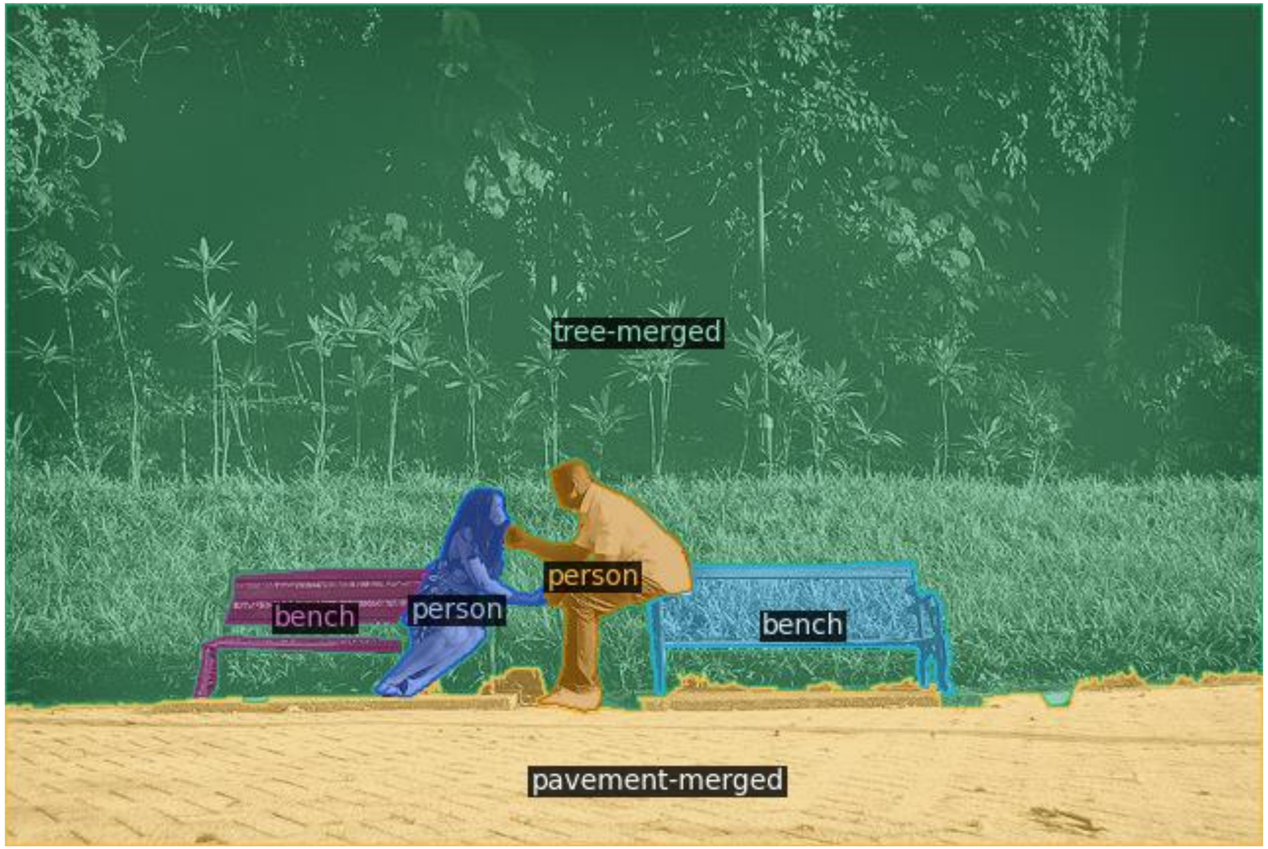
Q: What happened in the image?

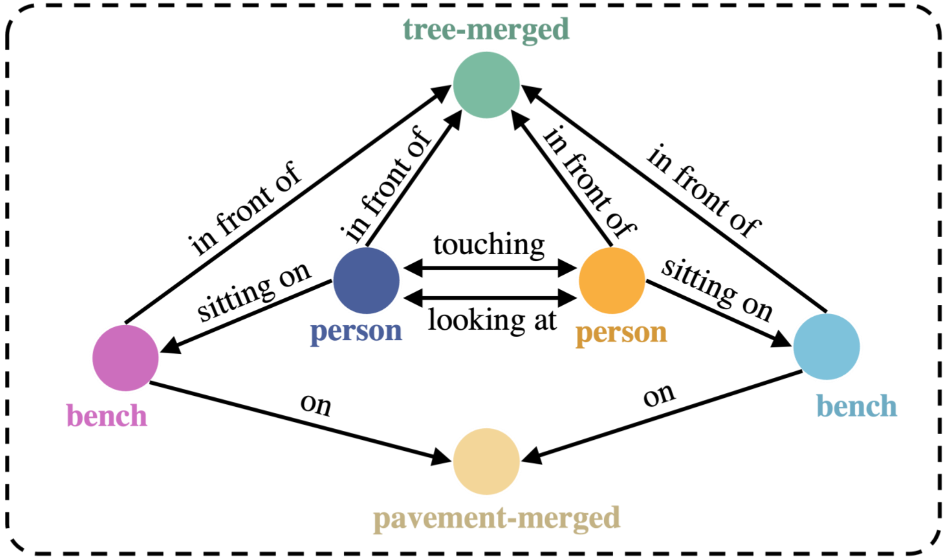
Q: What happened in the image?



A woman and **a man** touching and looking at each other. The woman is sitting on **the bench on the left**, and the man is sitting on the **right bench**. They are in front of **many trees**.

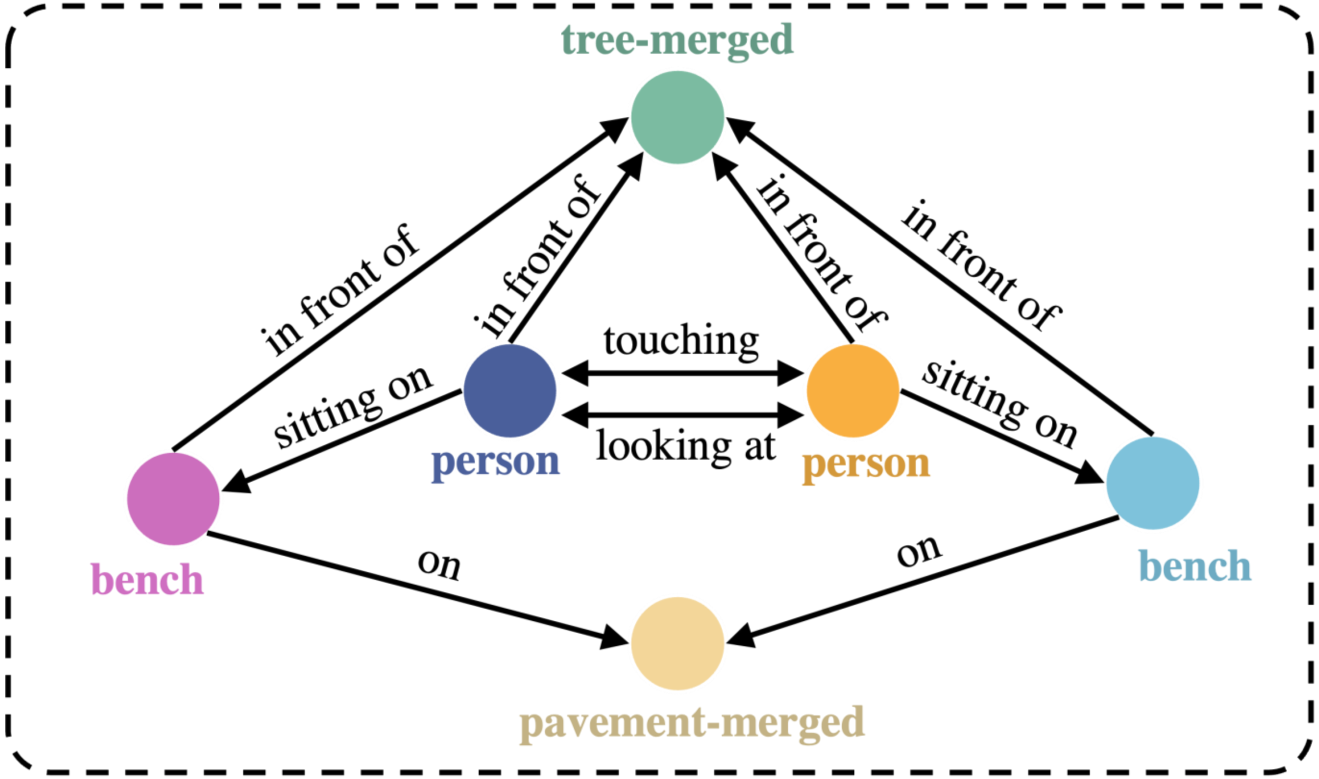
Q: What happened in the image?





Scene Graph

Beyond Object Recognition



Using scene graph:

Q: What happened in the image?

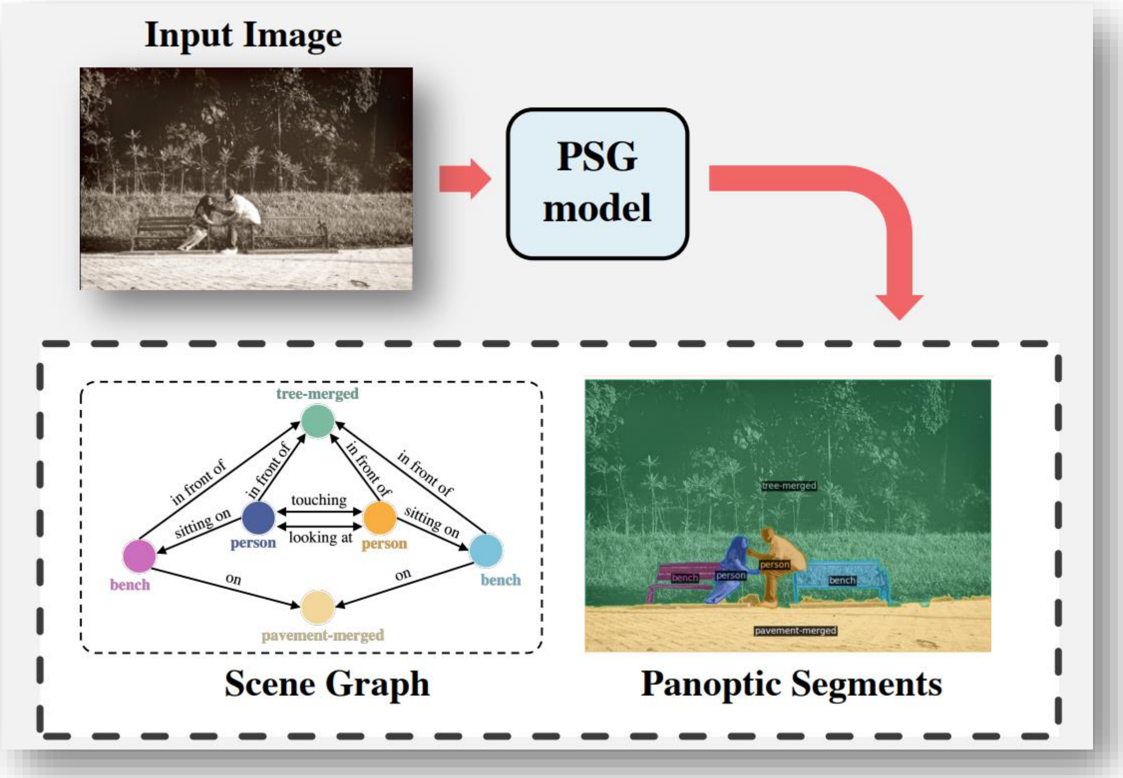
Q: Where is **the man** sitting on?

Add some commonsense:

Q: What are **the man** and **the woman** doing?

Q: What is the relation between **the man** and **the woman**?

PSG: Panoptic Scene Graph



Input:

An image with complex scene

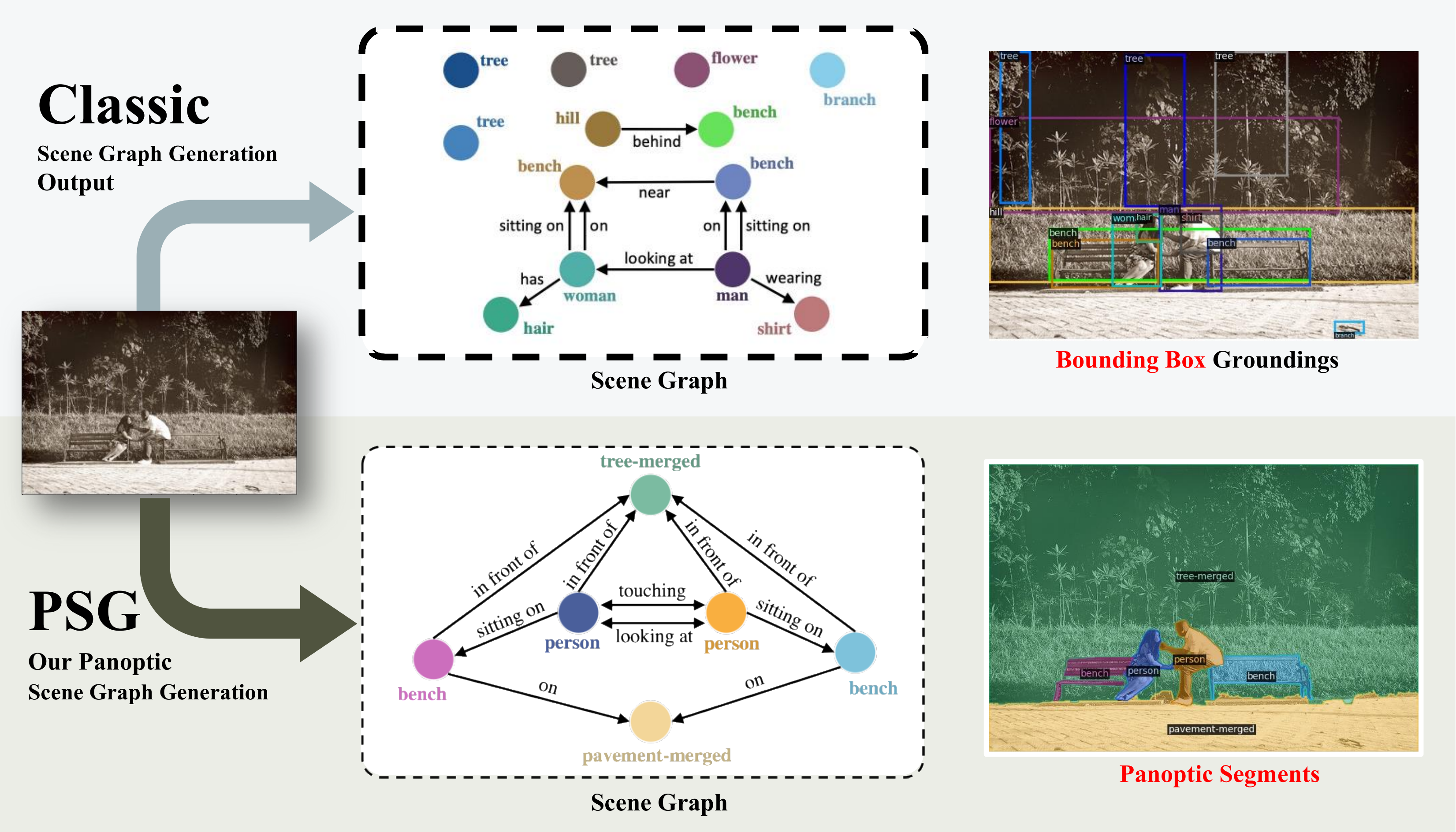
Output:

A scene graph
with panoptic segments

J Yang, et al. Panoptic Scene Graph Generation, ECCV 2022

PSG: Panoptic Scene Graph

+ Accurate Grounding + Proper Class Granularity + Able to involve Background



Classic

Scene Graph Generation Output

Scene Graph

Bounding Box Groundings

PSG

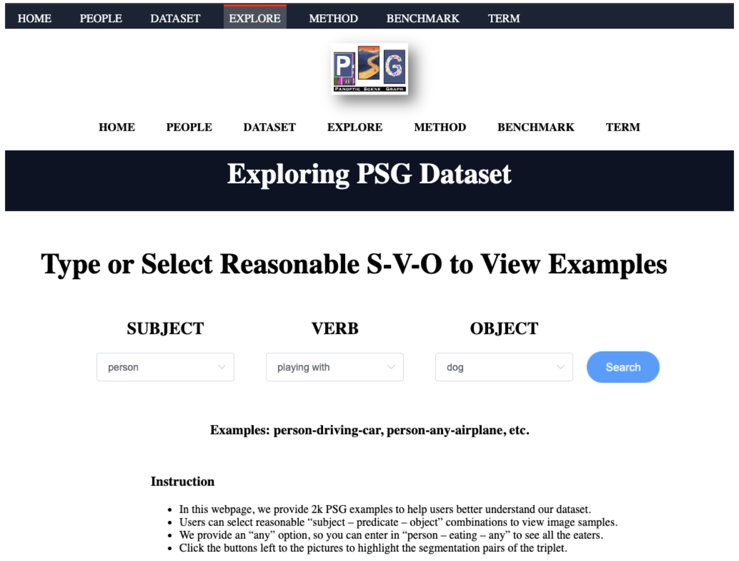
Our Panoptic Scene Graph Generation

Scene Graph

Panoptic Segments

PSG Dataset

- 49K images
- 133 object classes (80 objects and 53 stuff)
- COCO + VG
- 56 predicate classes.
- **Careful Predicate Design and Annotate**


HOME PEOPLE DATASET EXPLORE METHOD BENCHMARK TERM
PSG
HOME PEOPLE DATASET EXPLORE METHOD BENCHMARK TERM

Exploring PSG Dataset

Type or Select Reasonable S-V-O to View Examples

SUBJECT	VERB	OBJECT
person	playing with	dog

Examples: person-driving-car, person-any-airplane, etc.

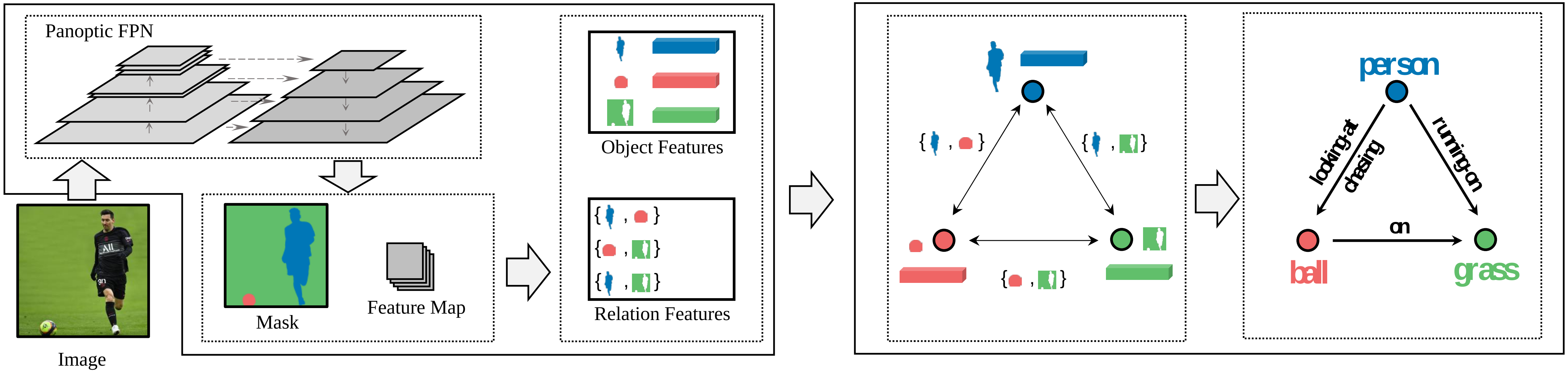
Instruction

- In this webpage, we provide 2k PSG examples to help users better understand our dataset.
- Users can select reasonable “subject – predicate – object” combinations to view image samples.
- We provide an “any” option, so you can enter in “person – eating – any” to see all the eaters.
- Click the buttons left to the pictures to highlight the segmentation pairs of the triplet.

http://psgdataset.org/explore.html

J Yang, et al. Panoptic Scene Graph Generation, ECCV 2022

Two Stage Methods

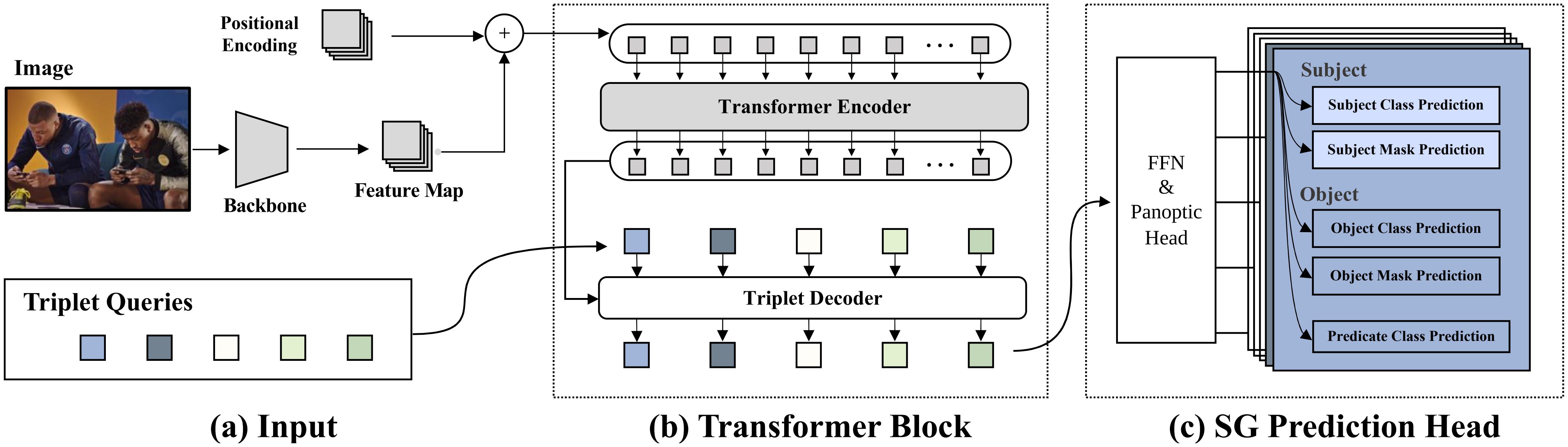


(a) Stage-1: Segment Feature Extractor

(b) Stage-2: Scene Graph Prediction

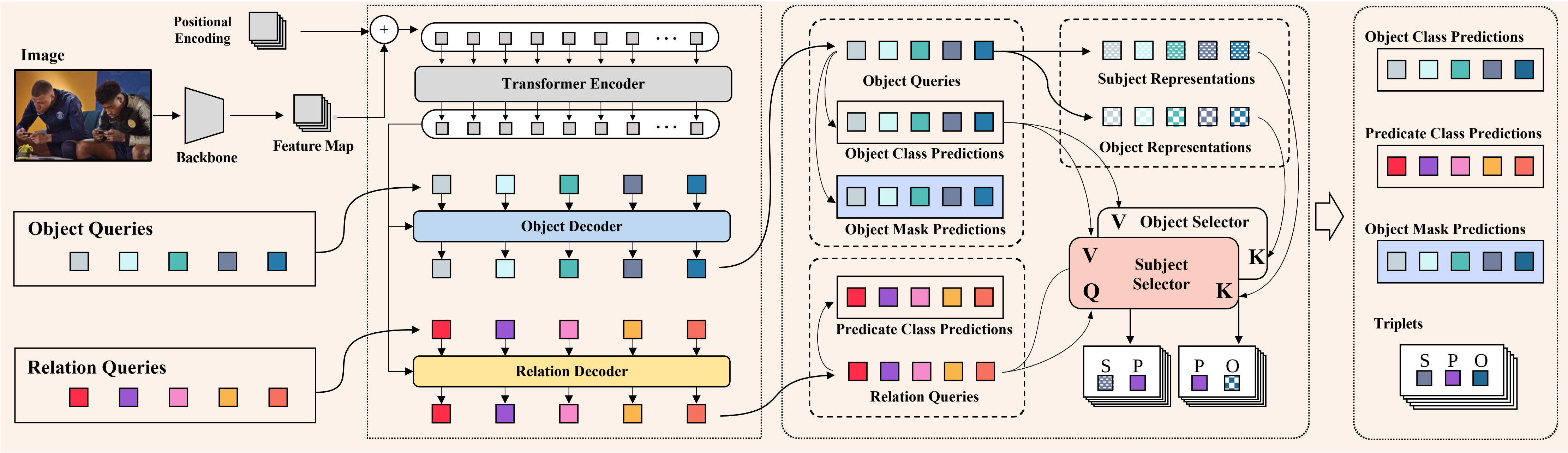
+ Fast, Simple, Easy to use
+ Support Classic Methods
- Heavily Rely on Detectors

One Stage Method (PSGTR)



+ Focus on Vision
+ Direct Training
- Need Long Time to Learn
- Conflict with PanSeg

One Stage Method (PSGFormer)



(a) Input (b) Query Learning Block (c) Query Matching Block (d) Prediction

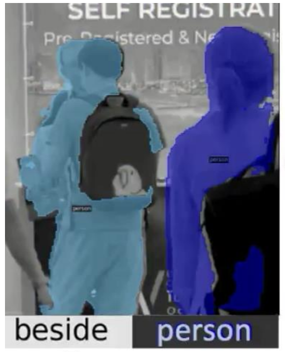
+ Explicit Relation Model
+ Fun Query Matching
+ Quick Converge
- Larger model

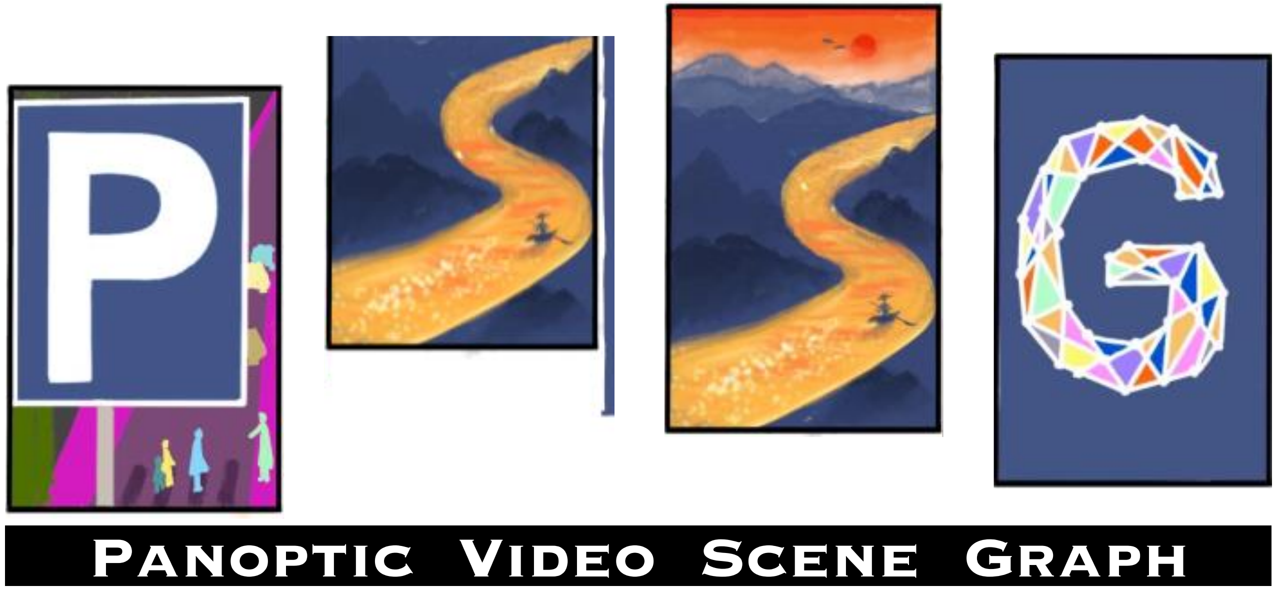
PSG: Panoptic Scene Graph

Prediction Result from the PSG Model

Q: Who is wearing a fancy bag in the photo?

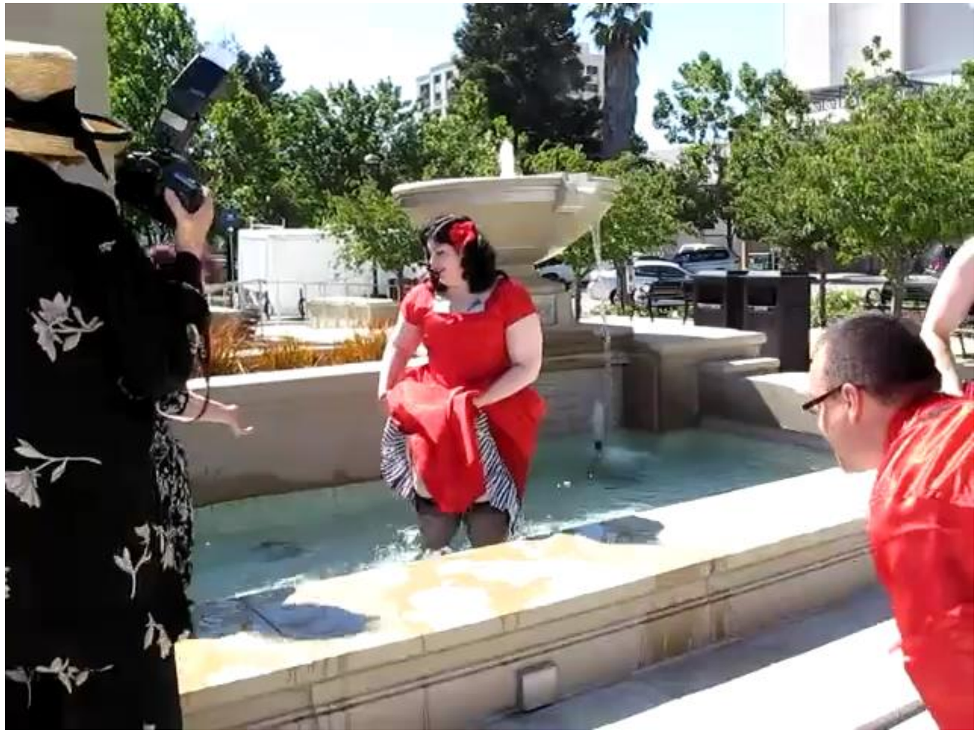
Q: Where is the man with fancy bag standing?

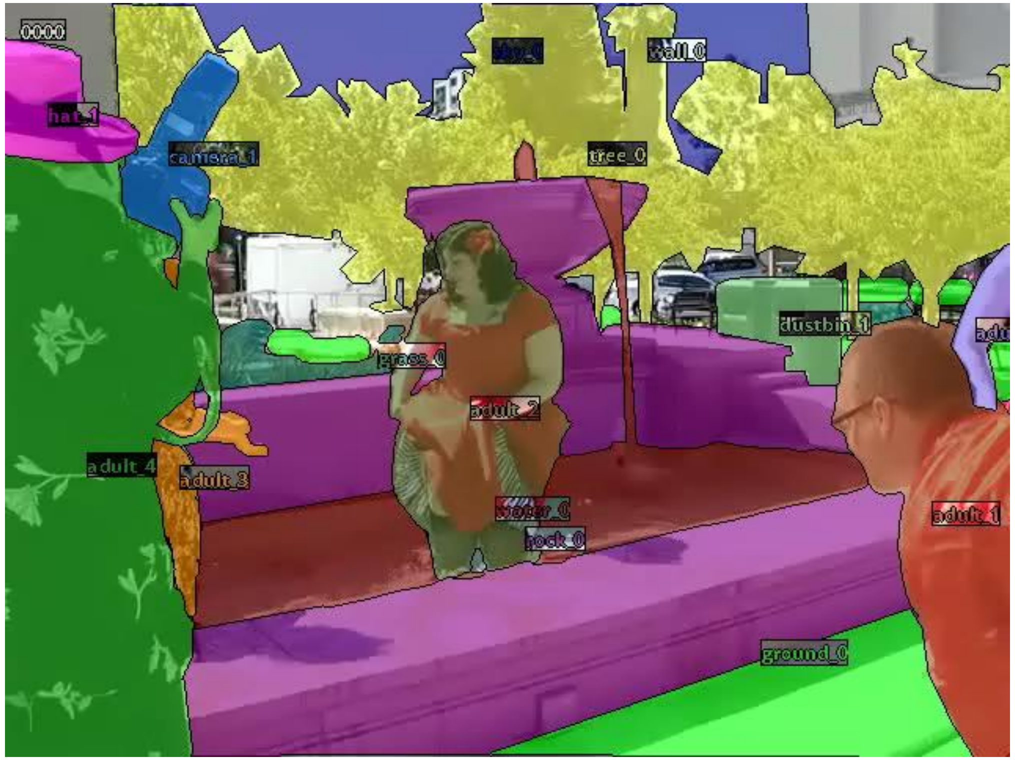


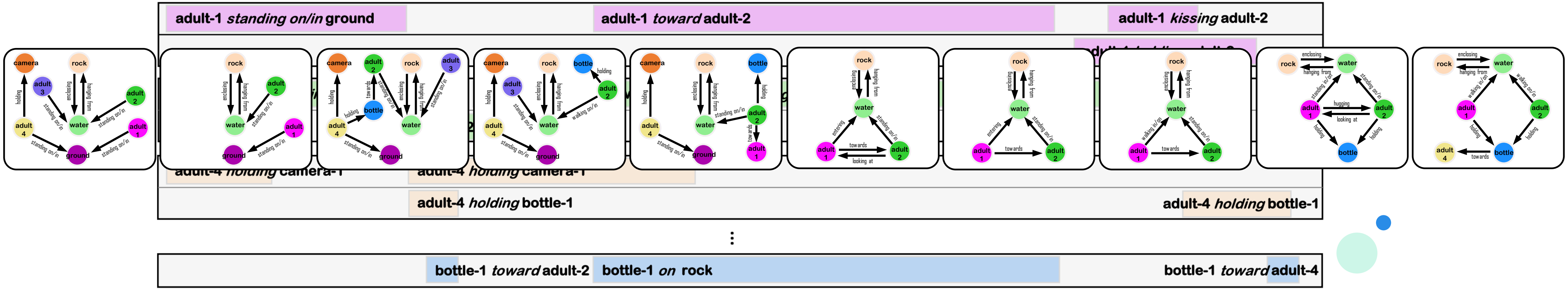


PSG + Video = PVSG

PVSG - Panoptic Video Scene Graph Generation

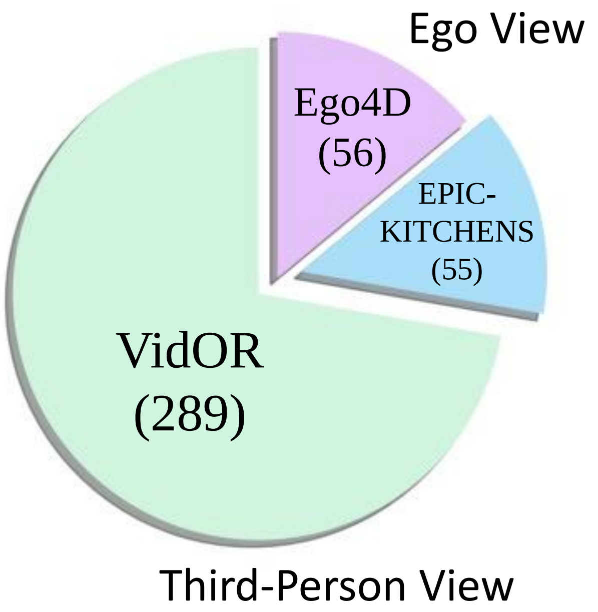


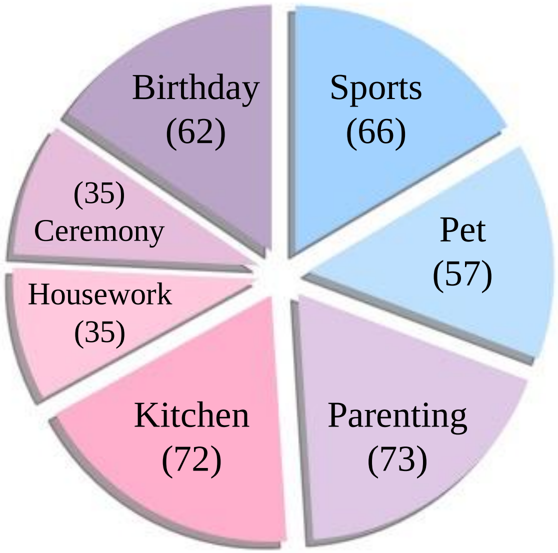



adult-1 *standing on/in* ground
adult-1 *toward* adult-2
adult-1 *kissing* adult-2
adult-4 *holding* camera-1
adult-4 *holding* camera-1
adult-4 *holding* bottle-1
adult-4 *holding* bottle-1
⋮
bottle-1 *toward* adult-2
bottle-1 *on* rock
bottle-1 *toward* adult-4
camera
rock
adult 3
adult 2
adult 4
water
adult 1
ground
bottle
holding
enclosing
hanging from
standing on/in
towards
walking on/in
entering
looking at
hugging

PVSG Dataset

A long-video, multiple perspectives, dense annotated, long-term dependent VidSGG dataset





400 videos, 9 hours

77s long in average

3rd + egocentric

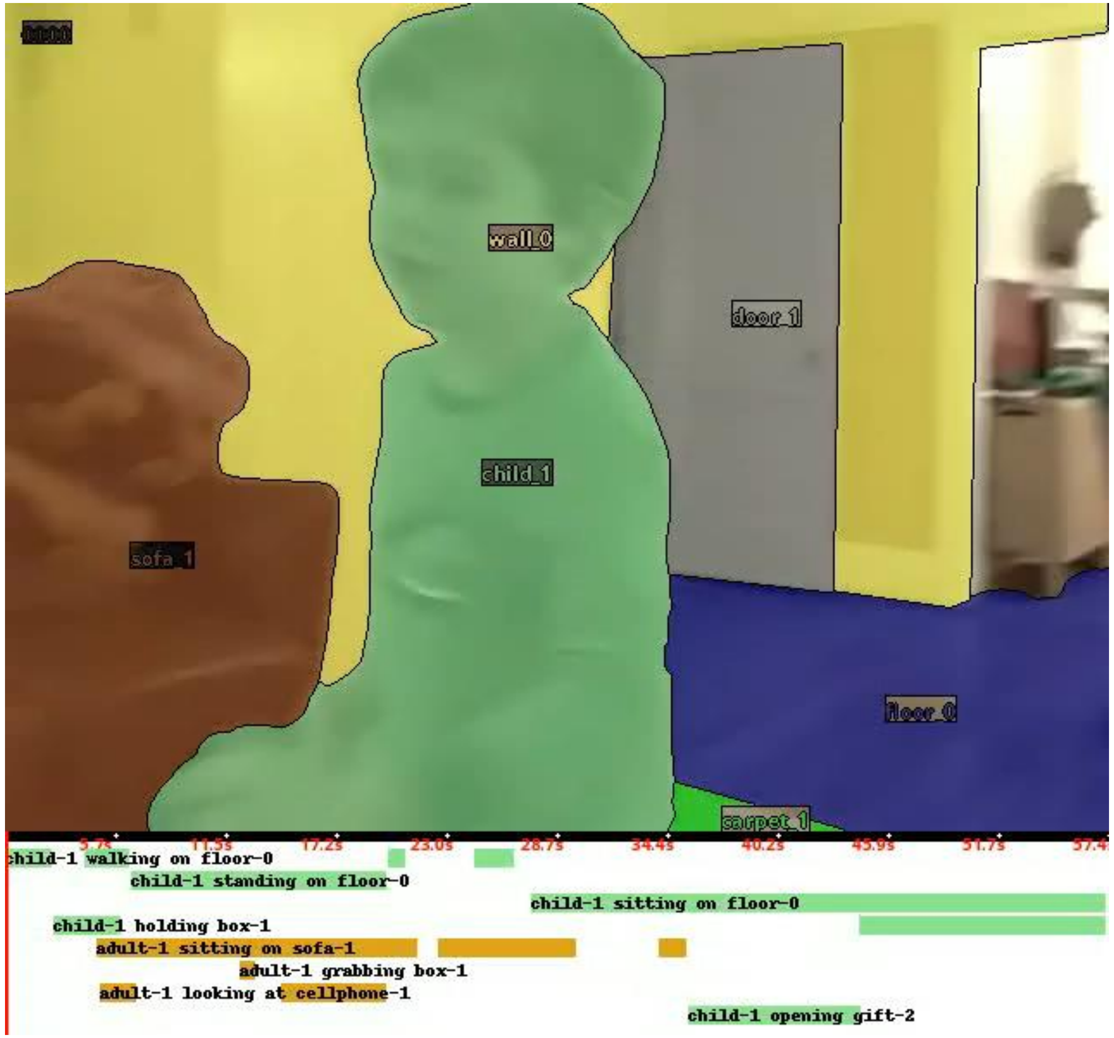
150K Panoptic Seg.

Dynamic Scene Graph

Dense Captioning

Commonsense QA

PVSG Dataset



Video Description

The scene depicts the boy receiving, giving, and unwrapping gifts on the holiday.

Dense Description

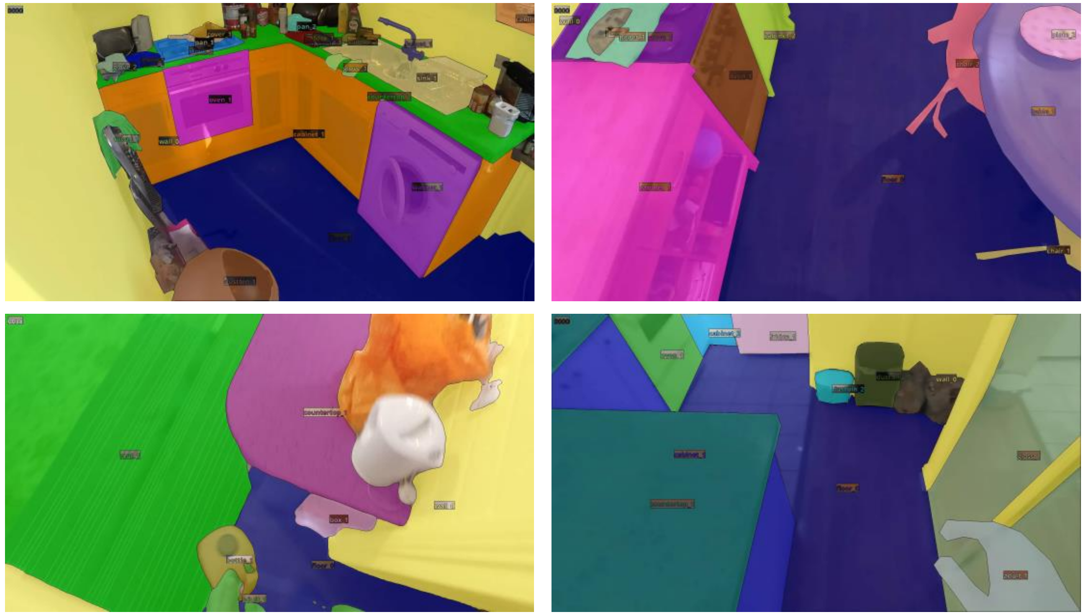
0000-0018: The **little boy (child-1)** passed through **the television (tv-1)** to pick up **a gift (gift-1).**

0018-0045: The **little boy (child-1)** handed the **gift (gift-1)** to a **woman (adult-1),** who appears to be **his mother (adult-1)**.

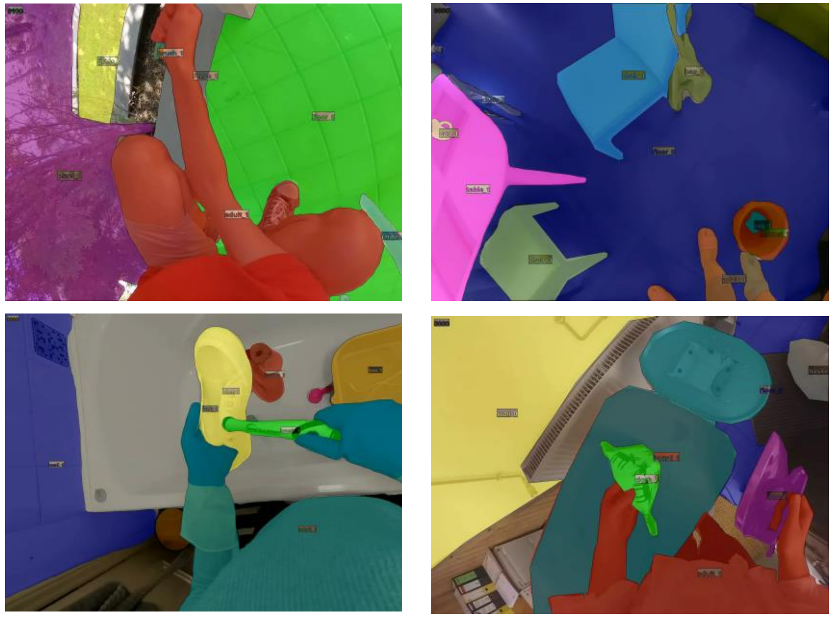
Dense QA

At Frame 0035: Q: Why **did the little boy (child-1)** give **the gift (gift-1)** to the **woman (adult-1)?**
A: It might be a gift exchange moment, and the gift is for **the woman (adult-1).**

PVSG Dataset (Egocentric)



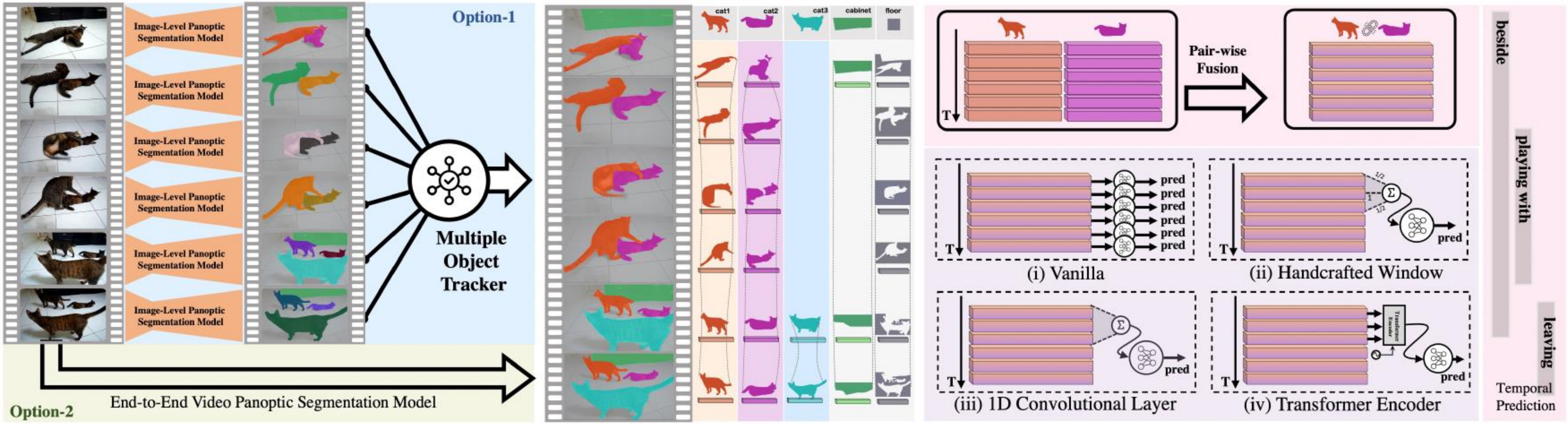
EpicKitchen (55 videos)



Ego4D (56 videos)

Towards Comprehensive Egocentric Video Scene Understanding

PVSG Method



(a) Stage-1: For Feature Tube and Mask Tube Output

(b) Stage 2: Relation Prediction

Bottleneck on Tracking

PVSG
Panoptic Video Scene Graph

Panoptic Video Scene Graph Generation

Jingkang Yang, Wenxuan Peng, Xiangtai Li,
Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma,
Kaiyang Zhou, Wayne Zhang, Chen Change Loy, Ziwei Liu

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE | 商汤 sensetime

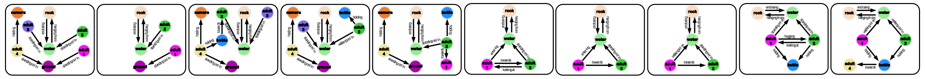
What is Panoptic Scene Graph Generation?

Input: a video sequence

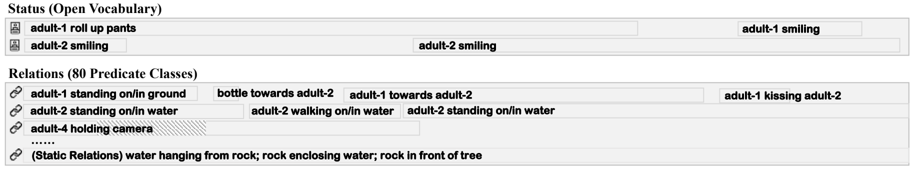
(a) An example video from the PVSG dataset. The PVSG dataset carefully selects 400 first/third-person long videos (avg. 1 min) with clear storyline.

Output: a frame-level panoptic segmentation & video-level scene graph

(b) The mask annotation of the example video. The PVSG dataset has dense (5 fps) and accurate video panoptic segmentation annotation.



(c) The frame-level scene graph representation of the example video. Nodes represent object category and its status. Edges represent relations.



Status (Open Vocabulary)

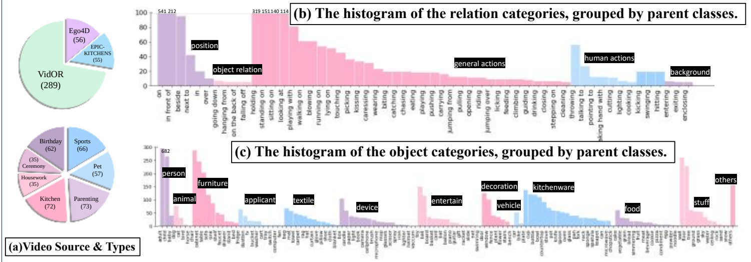
adult-1 roll up pants
adult-1 smiling
adult-2 smiling
adult-2 smiling

Relations (80 Predicate Classes)

adult-1 standing on/in ground
bottle towards adult-2
adult-1 towards adult-2
adult-1 kissing adult-2
adult-2 standing on/in water
adult-2 walking on/in water
adult-2 standing on/in water
adult-4 holding camera
……
(Static Relations) water hanging from rock; rock enclosing water; rock in front of tree

(d) The video-level status and relation annotation, which contains interchangeable information of frame-level scene graph in (c).

The PVSG Dataset



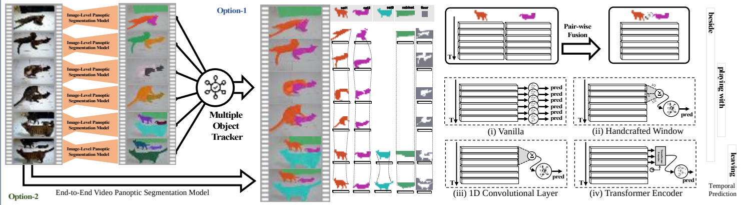
(a)Video Source & Types

(b) The histogram of the relation categories, grouped by parent classes.

(c) The histogram of the object categories, grouped by parent classes.

Large-Scale
150K frames
Long Videos
Logical, avg. 77s
Dense Annotation
PanSeg + SG
Multiple Views
289 TPV
111 ego-view

Solution: The PVSG Pipeline



(a) Stage-1: For Feature Tube and Mask Tube Output

(b) Stage 2: Relation Prediction

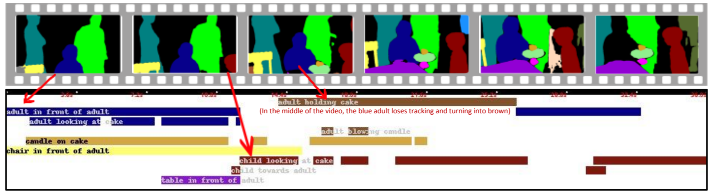
Results and Conclusion

1. "Frame-Level PanSeg + Tracking" better than "end-to-end VPS"

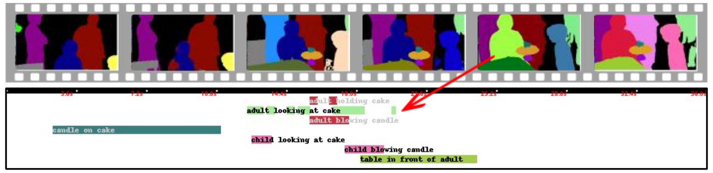
2. Transformer-based relation model obtains optimal results.

3. VPS fails due to inconsistent tracking.

Method		PVSG Metrics		
Stage-1	Stage-2	R/mR@20	R/mR@50	R/mR@100
IPS+T [3, 44]	Vanilla	2.35 / 1.22	2.71 / 1.31	2.94 / 1.45
	Handcrafted Window	2.56 / 1.24	2.78 / 1.35	3.05 / 1.54
	1D Convolution	2.79 / 1.24	2.80 / 1.47	3.10 / 1.59
	Transformer Encoder	**4.02 / 1.75**	**4.41 / 1.86**	**4.88 / 2.03**
VPS [3, 25]	Vanilla	0.52 / 0.24	0.60 / 0.24	0.63 / 0.24
	Handcrafted Window	0.54 / 0.27	0.61 / 0.29	0.62 / 0.29
	1D Convolution	0.60 / 0.27	0.73 / 0.28	0.76 / 0.29
	Transformer Encoder	**0.75 / 0.36**	**0.91 / 0.39**	**0.94 / 0.40**

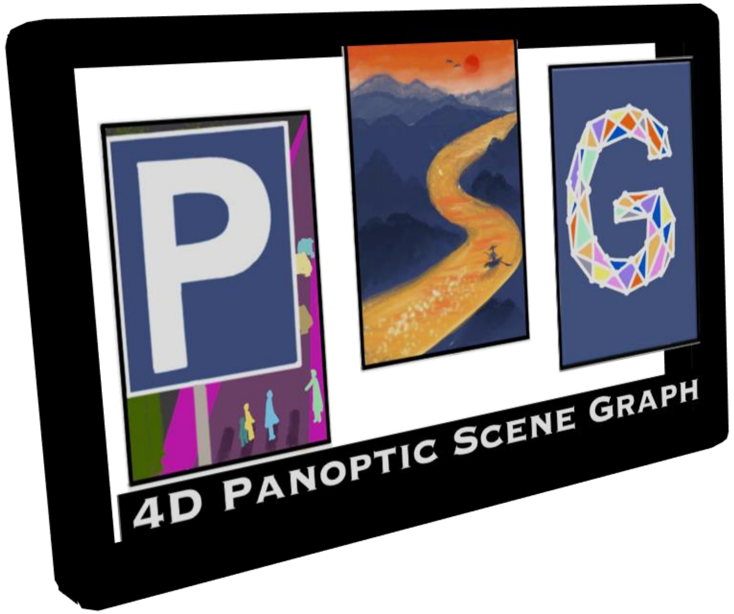


(a) The visualization result with the IPS+T method in the first stage and Transformer Encoder in the second stage.



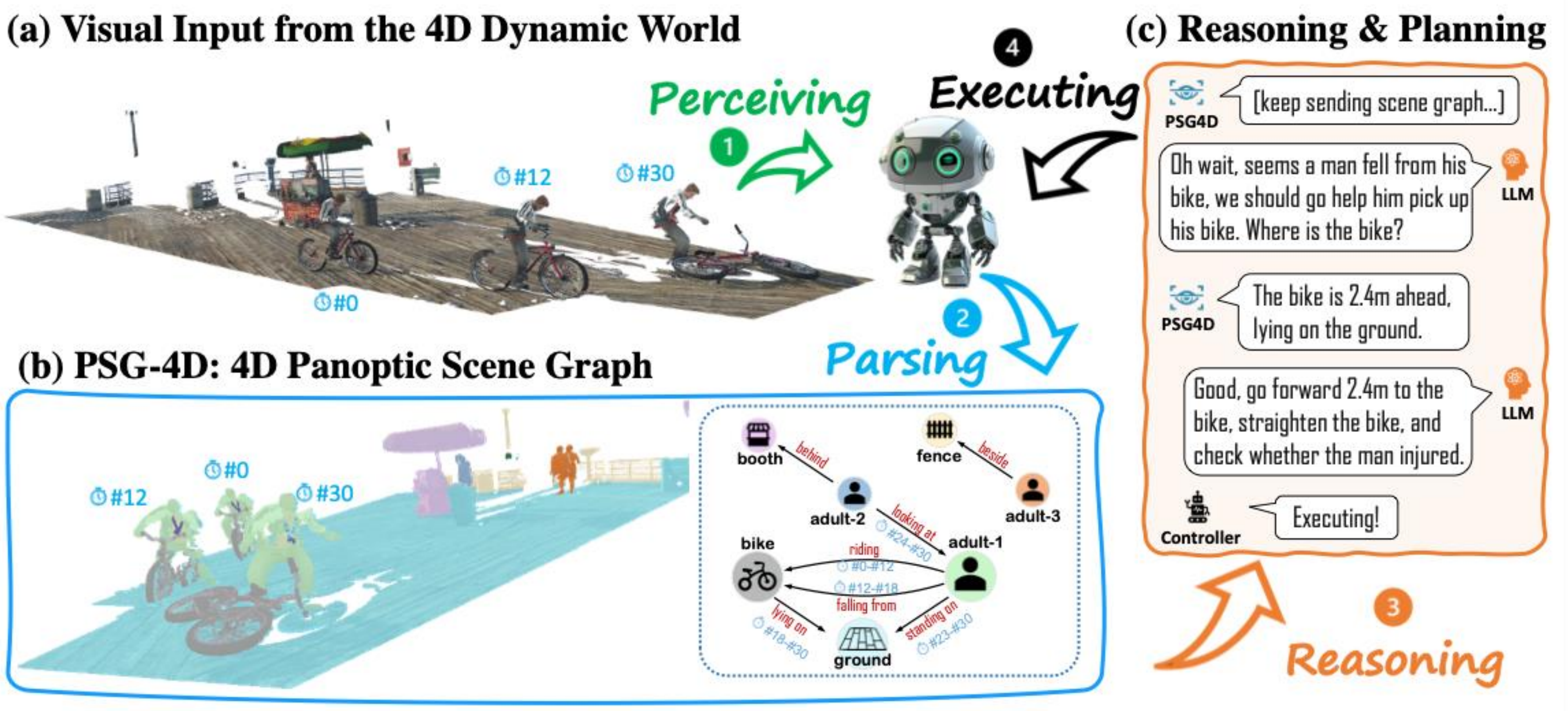
(b) The visualization result with the VPS method in the first stage and Transformer Encoder in the second stage.

CVPR 2023, Thursday morning, West Building Exhibit Halls ABC 210



PSG + Video + 3D = PSG4D

PSG4D: AI Assistant in 4D world



PSG4D: [keep sending scene graph...]

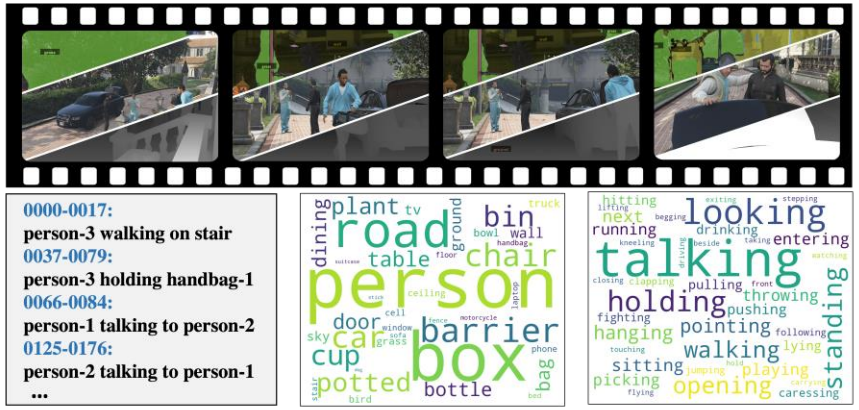
LLM: Oh wait, seems a man fell from his bike, we should go help him pick up his bike. Where is the bike?

PSG4D: The bike is 2.4m ahead, lying on the ground.

LLM: Good, go forward 2.4m to the bike, straighten the bike, and check whether the man injured.

Controller: Executing!

PSG4D Dataset



0000-0017:
person-3 walking on stair
0037-0079:
person-3 holding handbag-1
0066-0084:
person-1 talking to person-2
0125-0176:
person-2 talking to person-1
...

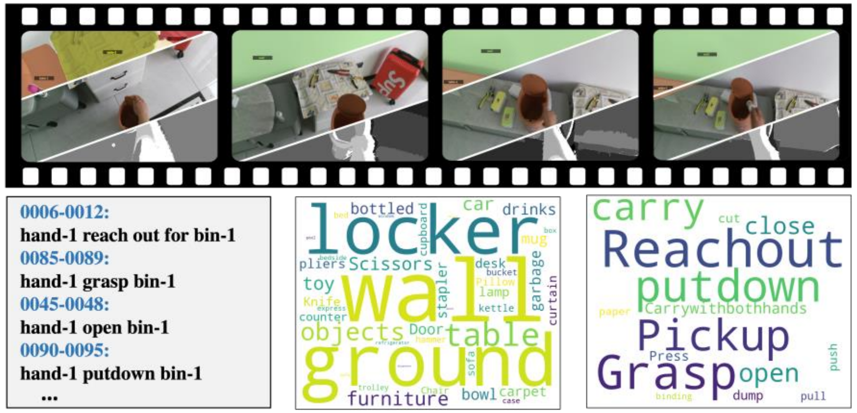
(a) **PSG4D-GTA** (Synthetic, Third-Person View)

Source: Grand Theft Auto V

67 videos (avg. 84s)

28K RGB-D images

35 object classes, 43 relations



0006-0012:
hand-1 reach out for bin-1
0085-0089:
hand-1 grasp bin-1
0045-0048:
hand-1 open bin-1
0090-0095:
hand-1 putdown bin-1
...

(b) **PSG4D-HOI** (Real-World, Egocentric)

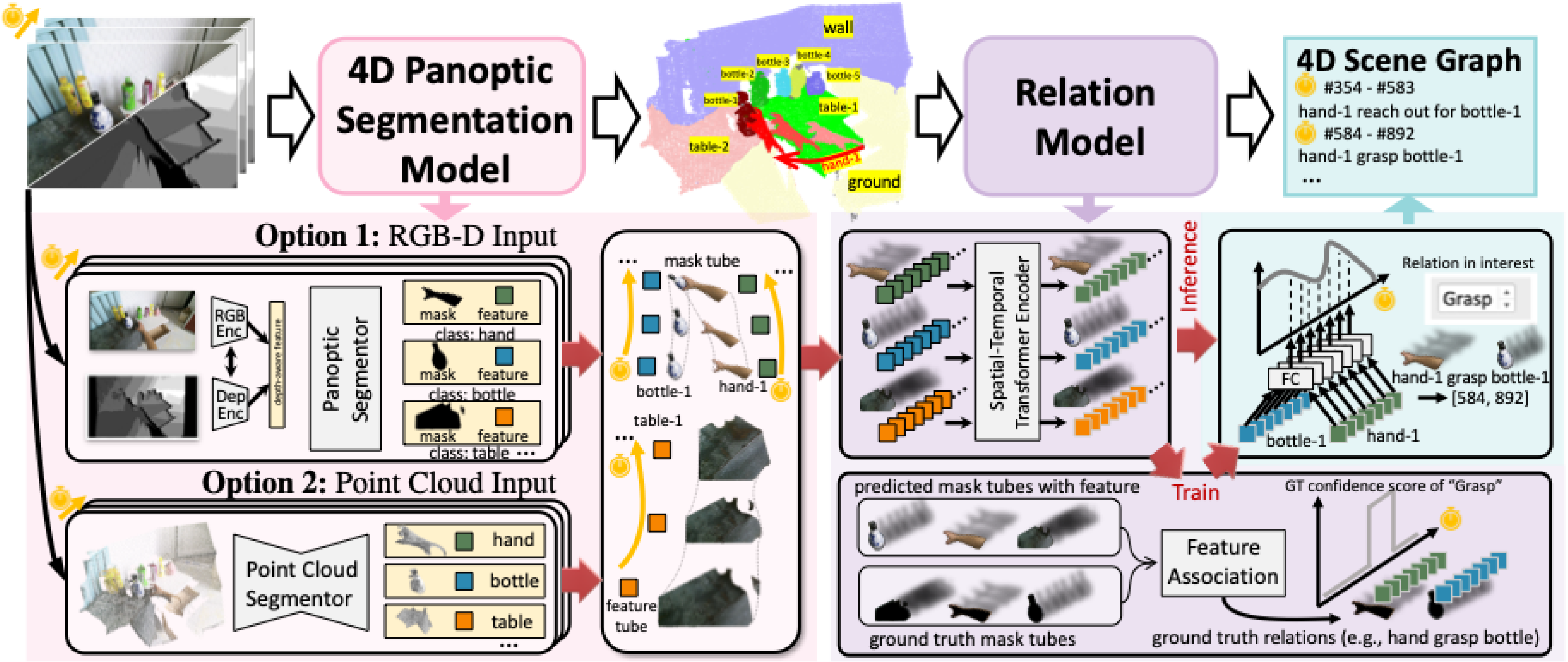
Source: HOI-4D

2973 videos (avg. 20s), Egocentric

891K RGB-D images

46 object classes, 15 relations

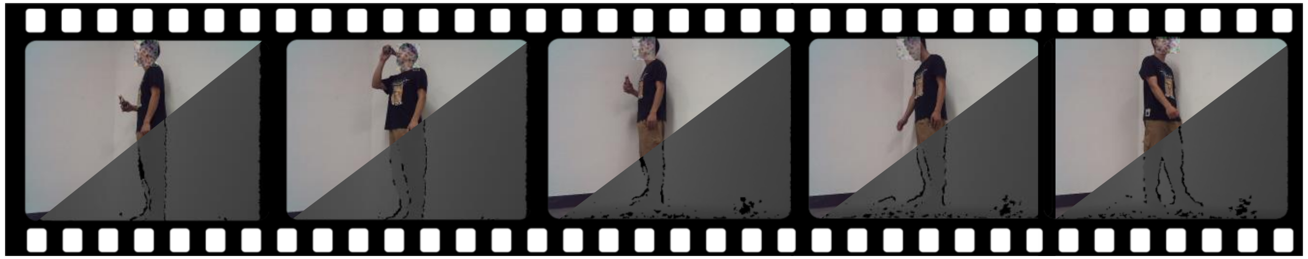
PSG4D Pipeline



(a) Frame-Level Panoptic Segmentation (b) Tracking (c) Inference (↑) and Training (↓) of Relation Model

PSG4D Real-World Application

(a) The RGB-D sequence that is captured by the robot.

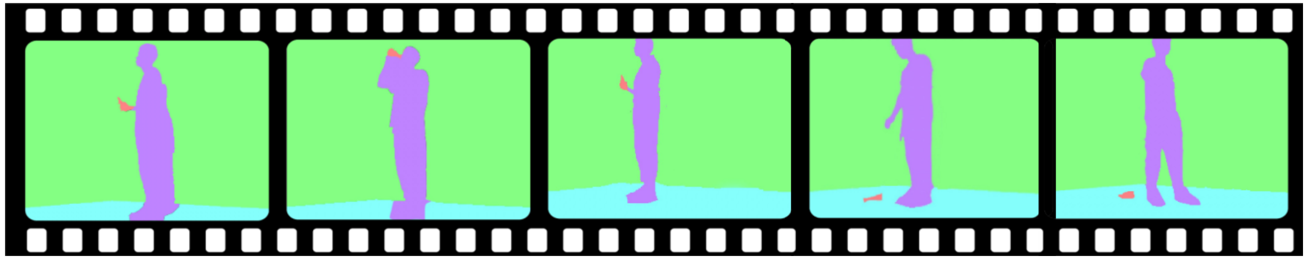


(b) PSG-4D Parsing

14.2s – 17.6s
person-1 drinking from bottle-1

18.8s – 20.0s
person-1 throwing bottle-1

20.0s – 30.0s
bottle-1 on ground



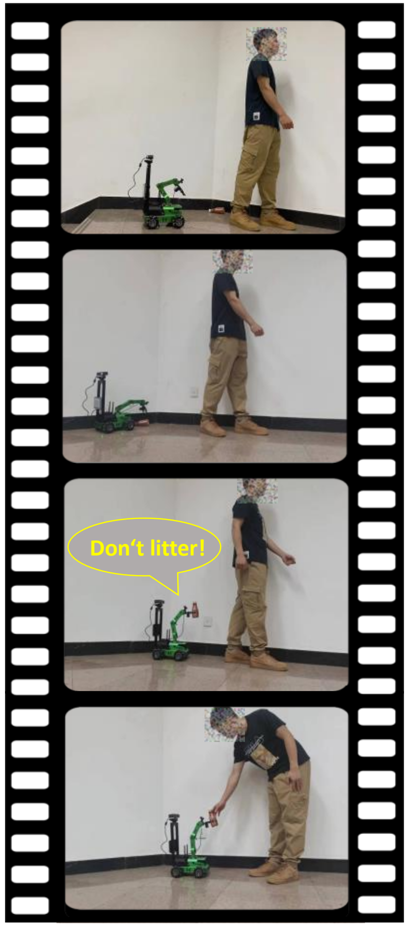
(c) Reasoning & Planning

PSG4D

I am a service robot, In the past 30s, what I captured is: [ ]. Is there anything I could serve?

1. Cleanup: …;
2. Reminder: …

(d) Robot Reaction



The pathway: From Language Models to Language Assistant

GPT-2 OpenAI

GPT-3 OpenAI

GPT-3.5 OpenAI

ChatGPT OpenAI

Industrial

Open-source

BERT

Zero-shot learning

LLaMA/T5

T5

Zero-shot learning
In-context learning

Vicuna/Flan-T5

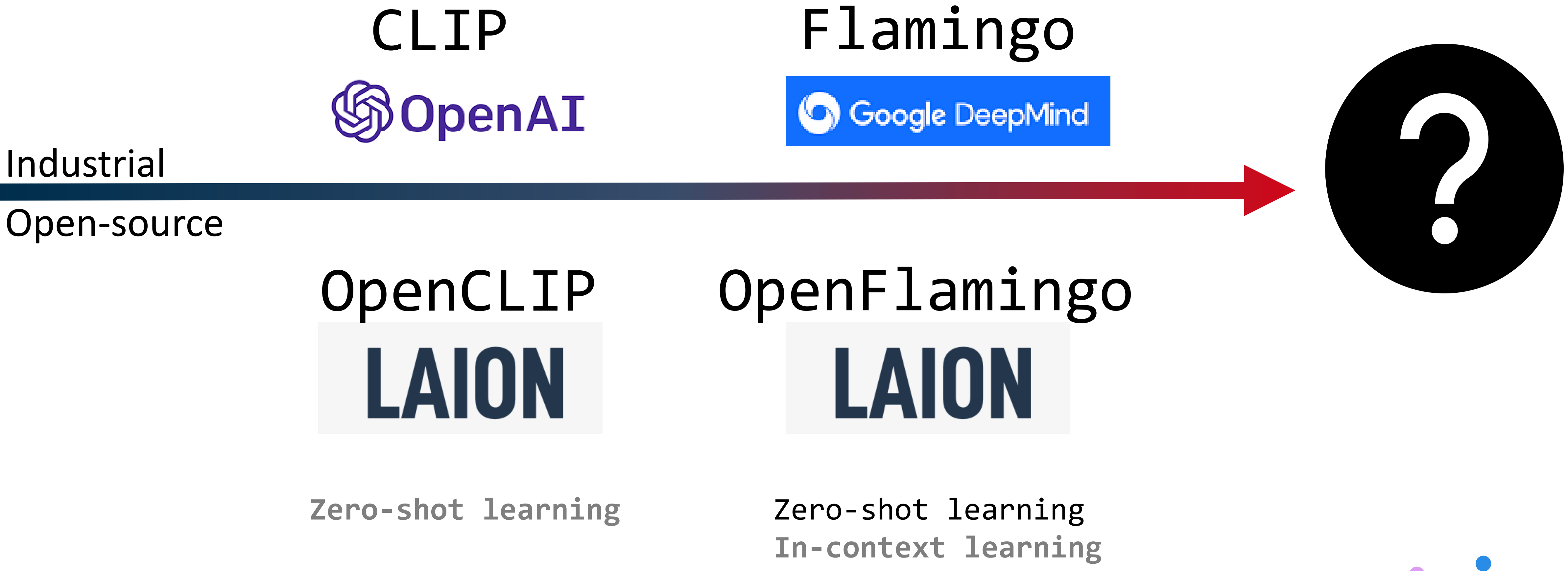
FLAN-T5

Zero-shot learning
In-context learning
Instruct following

Open Assistant

Zero-shot learning
In-context learning
Instruct following
Human alignment

The pathway: From Multi-modal Models to Multi-modal Assistants



CLIP

OpenAI

Flamingo

Google DeepMind

Industrial

Open-source

OpenCLIP

LAION

OpenFlamingo

LAION

Zero-shot learning

Zero-shot learning
In-context learning

?

The pathway: From Multi-modal Models to Multi-modal Assistants

Industrial

CLIP

OpenAI

Zero-shot learning

Flamingo

Google DeepMind

Zero-shot learning
In-context learning

Open-source

OpenCLIP

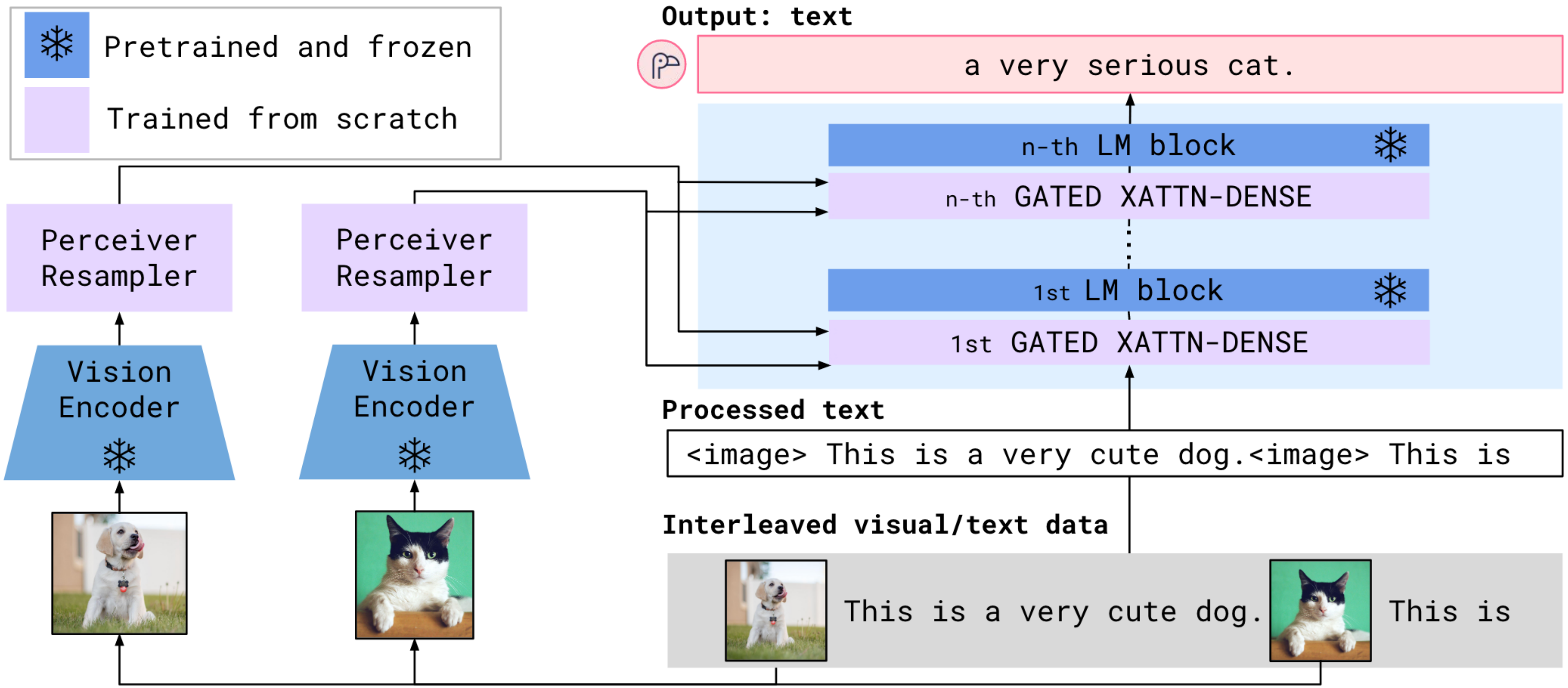
LAION

OpenFlamingo

LAION

Otter

Flamingo: a Visual Language Model for Few-Shot Learning



❄ Pretrained and frozen

▨ Trained from scratch

Output: text

a very serious cat.

n-th LM block ❄

n-th GATED XATTN-DENSE

1st LM block ❄

1st GATED XATTN-DENSE

Perceiver Resampler

Perceiver Resampler

Vision Encoder ❄

Vision Encoder ❄

Processed text

<image> This is a very cute dog.<image> This is

Interleaved visual/text data

This is a very cute dog. This is

Alayrac et. al. Flamingo: a visual language model for few-shot learning. 2022

Perceiver: versatile to multiple images and in-context examples

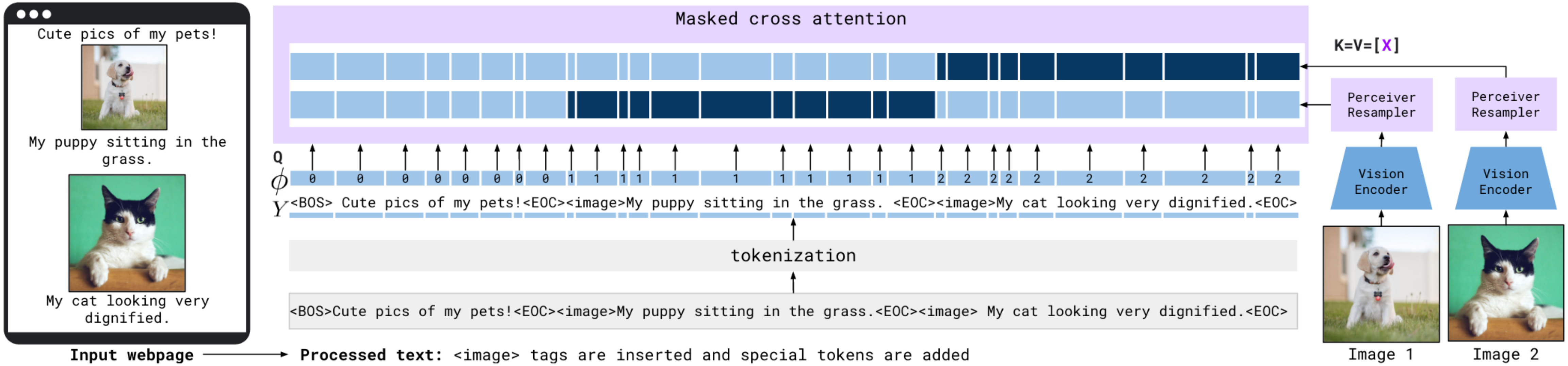

Cute pics of my pets!
My puppy sitting in the grass.
My cat looking very dignified.
Input webpage
Masked cross attention
K=V=[X]
Perceiver Resampler
Perceiver Resampler
Vision Encoder
Vision Encoder
Image 1
Image 2
Q
ϕ
0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2
Y
<BOS> Cute pics of my pets!<EOC><image>My puppy sitting in the grass. <EOC><image>My cat looking very dignified.<EOC>
tokenization
<BOS>Cute pics of my pets!<EOC><image>My puppy sitting in the grass.<EOC><image> My cat looking very dignified.<EOC>
Processed text: <image> tags are inserted and special tokens are added

Image-Text Pairs dataset
[N=1, T=1, H, W, C]

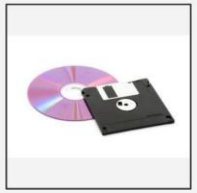
Video-Text Pairs dataset
[N=1, T>1, H, W, C]

Multi-Modal Massive Web (M3W) dataset
[N>1, T=1, H, W, C]

Flamingo Application

Zero-shot learning

Input Prompt

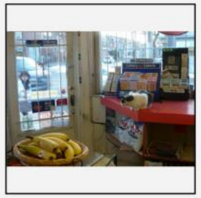


Question: What do you think the capacities of these are? Answer:

Completion

The floppy disk is 1.44MB and the CD is 700MB.

Input Prompt



Question: What nutrients is the fruit in the image rich in? Answer:

Completion

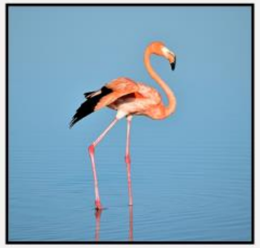
Vitamin C, potassium, and fiber.

In-context learning

Input Prompt

This is a chinchilla. They are mainly found in Chile.

This is a shiba. They are very popular in Japan.

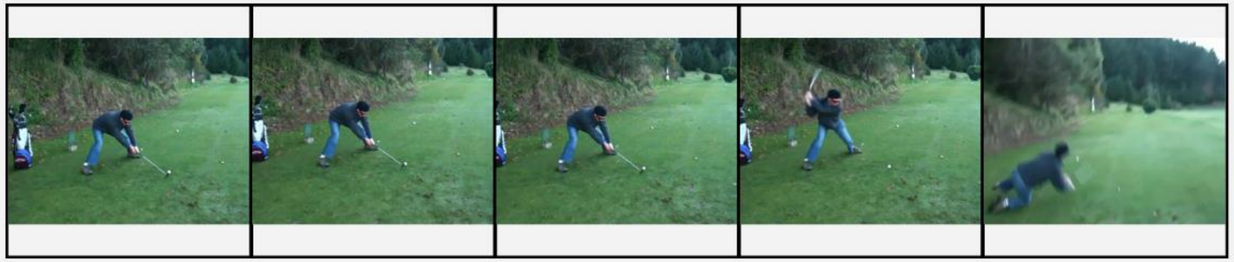


This is →

Completion

a flamingo. They are found in the Caribbean and South America.

Video Understanding

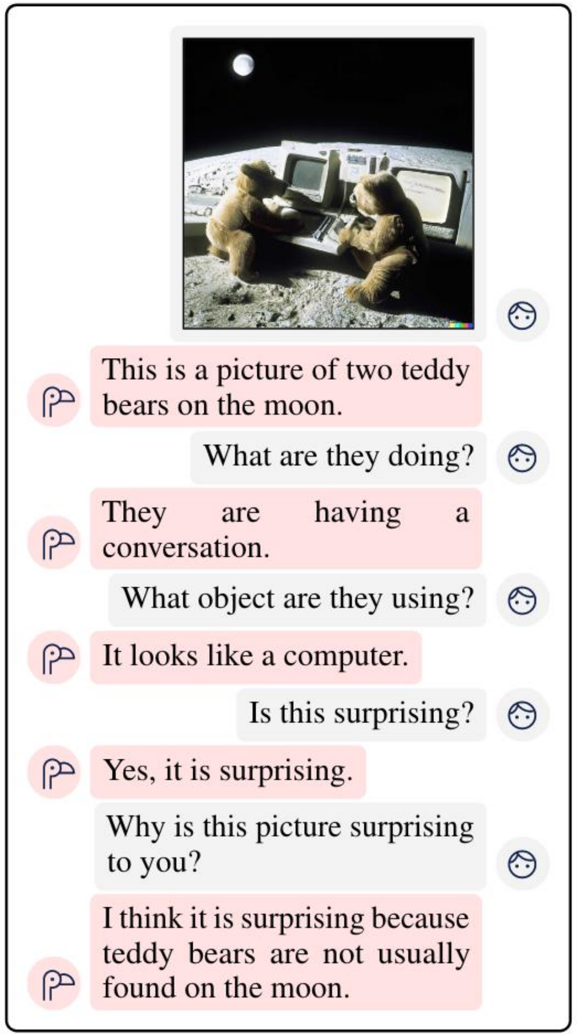


What happens to the man after hitting the ball? Answer: →

he falls down.

Flamingo Application

multi-image visual dialogue



This is a picture of two teddy bears on the moon.

What are they doing?

They are having a conversation.

What object are they using?

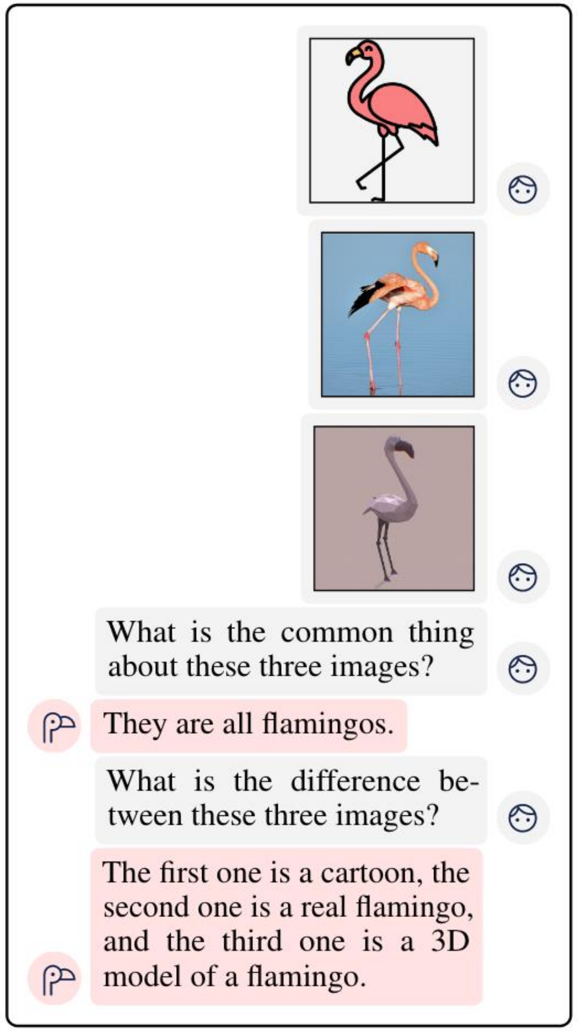
It looks like a computer.

Is this surprising?

Yes, it is surprising.

Why is this picture surprising to you?

I think it is surprising because teddy bears are not usually found on the moon.

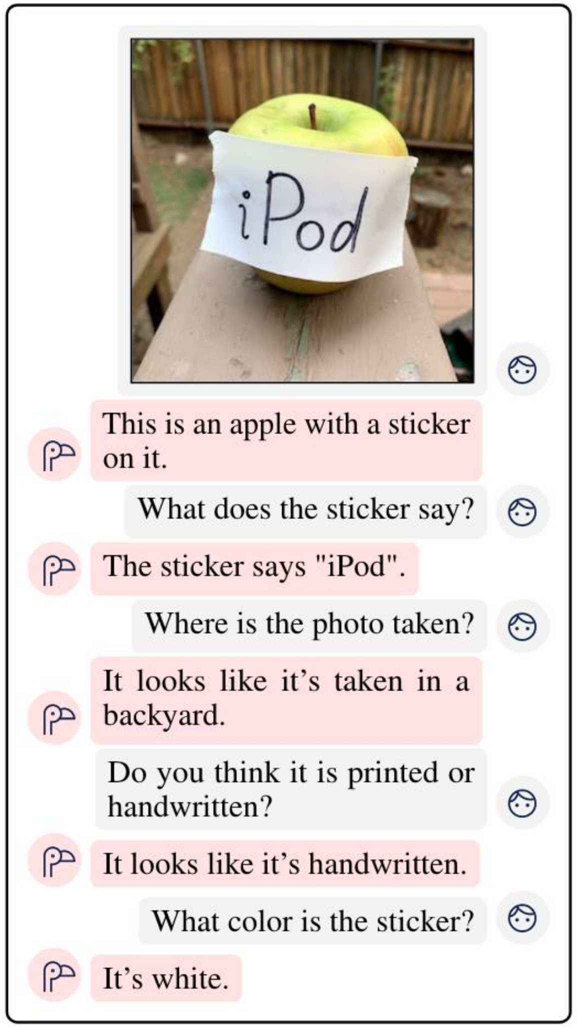


What is the common thing about these three images?

They are all flamingos.

What is the difference between these three images?

The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.



This is an apple with a sticker on it.

What does the sticker say?

The sticker says "iPod".

Where is the photo taken?

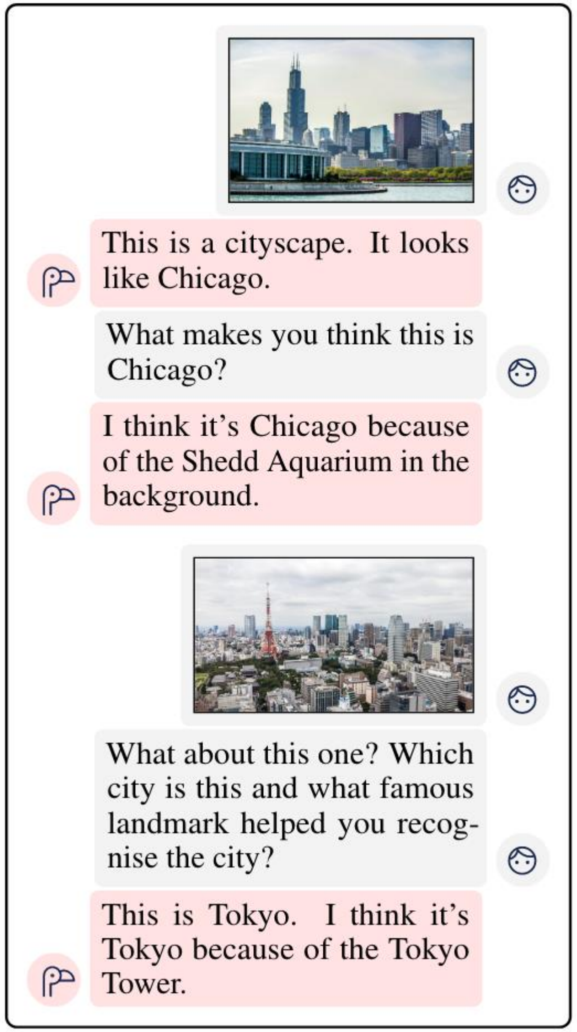
It looks like it's taken in a backyard.

Do you think it is printed or handwritten?

It looks like it's handwritten.

What color is the sticker?

It's white.



This is a cityscape. It looks like Chicago.

What makes you think this is Chicago?

I think it's Chicago because of the Shedd Aquarium in the background.

What about this one? Which city is this and what famous landmark helped you recognise the city?

This is Tokyo. I think it's Tokyo because of the Tokyo Tower.

Flamingo ≠ Multi-modal Assistants

OpenFlamingo simply completes the next reasonable sentence.

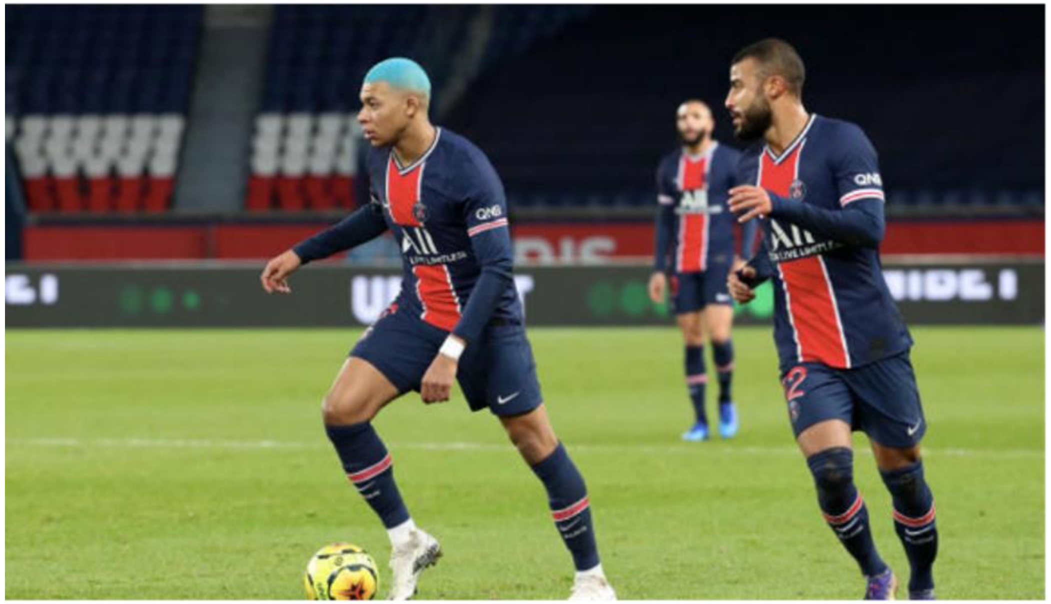
e danger of this sport?

OpenFlamingo*: What is the danger of playing baseball? What is the danger of this sport? What might be the danger of this sports?

Flaming (trained in the SSL manner) are not aligned with user intent and serve as a Chatbot.

*OpenFlamingo is the open-source version of Flamingo, enabling community research with a strong interleaved data pretrained model

Flamingo ≠ Multi-modal Assistants

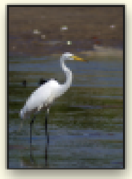


Question: What is the danger of this sport?

Human Expected: The sport involves players running and trying to catch the ball while others are standing in the grass, which can lead to collisions or accidents.

Flaming (trained in the SSL manner) are not aligned with user intent and serve as a Chatbot.

MMC4: Image-text interleaved data for OpenFlamingo Pretraining

[..., "Check out Shane Driscoll's take on sustainable communities and how his photograph fits this year's Green Cities theme.", ...,  ,"Man-made platforms like the one pictured here allow these fish-eating birds of prey to thrive in developed coastal areas.", "A city surrounded by mountains.", "I took this photo in October on a hike in New Hampshire.", , "It is looking at Mt. Chicora from the middle sister mountain.", "Getting people out into beautiful places like this is becoming more and more popular, and each time we bring a little piece of nature back with us that inspires us to make our cities better.", ...]

Diverse and large-scale, but lack of Instruct-following scenario

Zhu et. al. Multimodal C4: An open, billion-scale corpus of images interleaved with text. 2023

From interleaved data pretraining to multi-modal In-context instruction tuning

you'll have tender, juicy, flavourful chicken thighs

Open the memory card cover…

Paris has a reputation for being…

Bryant and the Lakers encountered difficulties.

MMC4
(interleaved pretraining)

OpenFlamingo

MIMIC-IT
(Multi-Modal In-Context Instruction Tuning)

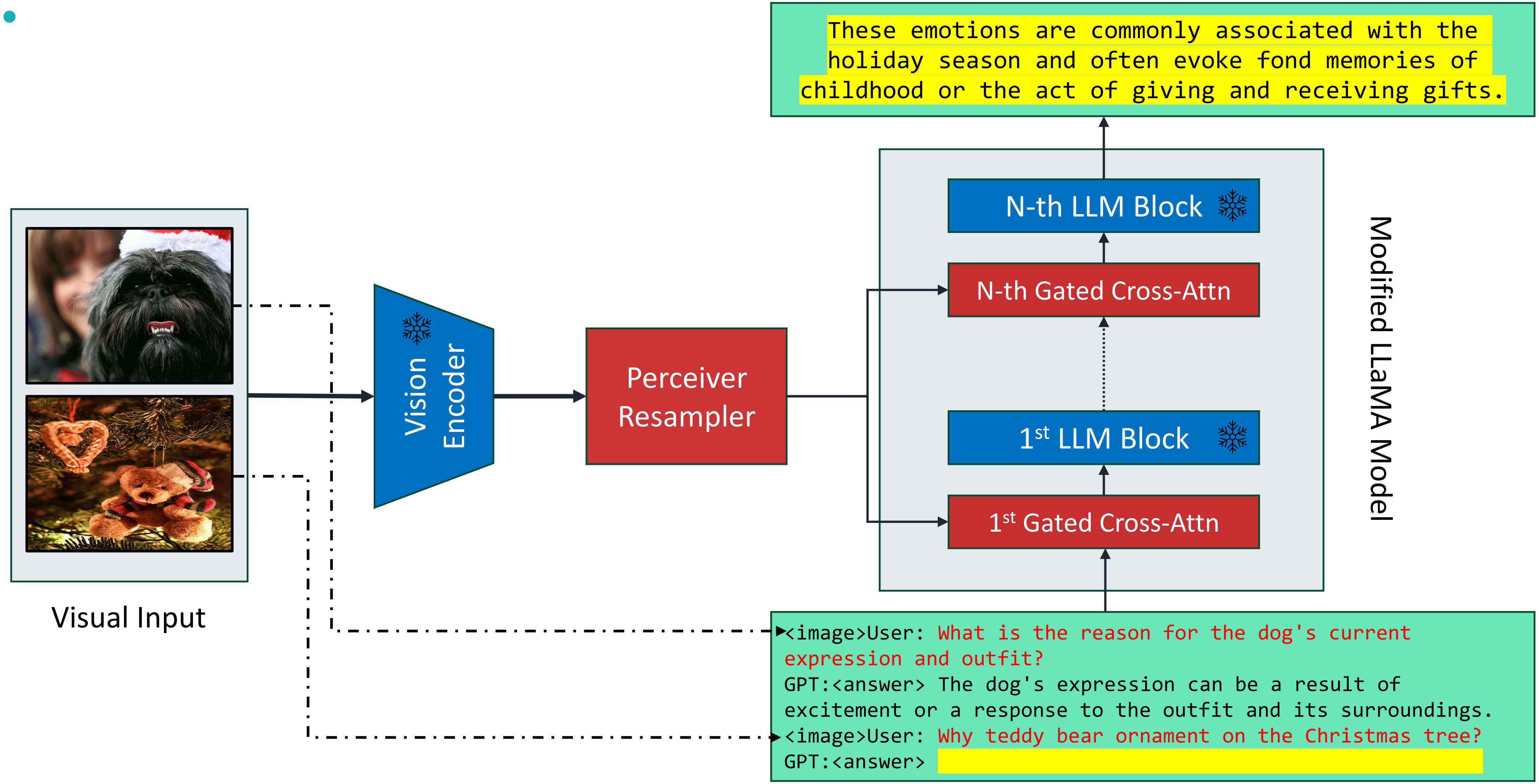
Otter

From interleaved data pretraining to multi-modal In-context instruction tuning

🦦 Otter enhances OpenFlamingo’s capabilities, including:

- Instruction following: aligning with user intent
- Stronger in-context learning ability.
- Fine-grained understanding: spot the difference between images.
- Vision reasoning and even planning: It can write story for a series of images and even suggestion how to clean your room from your room tour video.

Otter: A Multi-Modal In-context Instruction Tuned Model



Visual Input

Vision Encoder

Perceiver Resampler

N-th LLM Block

N-th Gated Cross-Attn

1st LLM Block

1st Gated Cross-Attn

Modified LLaMA Model

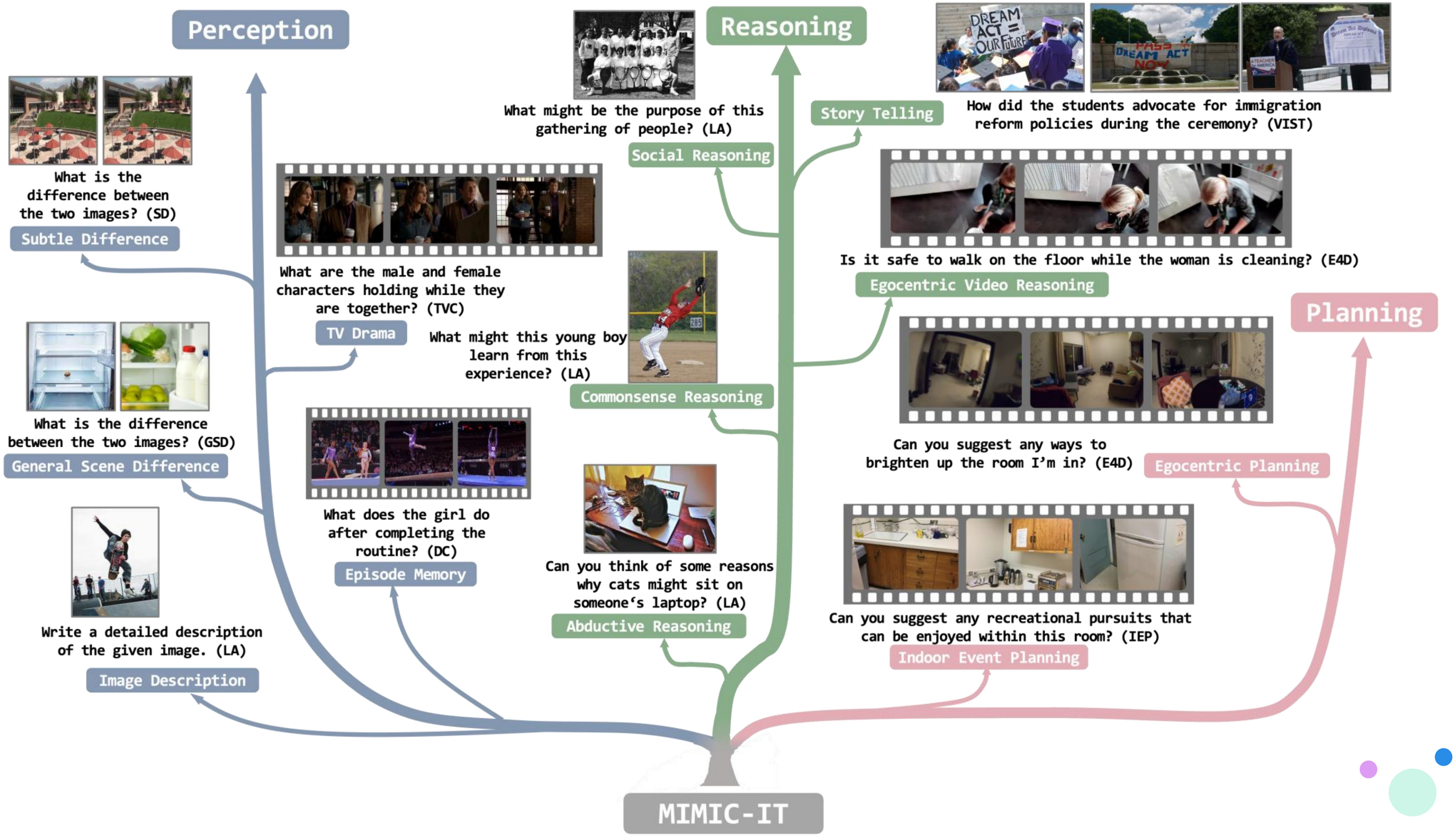
These emotions are commonly associated with the holiday season and often evoke fond memories of childhood or the act of giving and receiving gifts.

```
<image>User: What is the reason for the dog's current
expression and outfit?
GPT:<answer> The dog's expression can be a result of
excitement or a response to the outfit and its surroundings.
<image>User: Why teddy bear ornament on the Christmas tree?
GPT:<answer>
```

MIMIC-IT Dataset

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

S-LAB
FOR ADVANCED INTELLIGENCE



Perception

What is the difference between the two images? (SD)
Subtle Difference

What is the difference between the two images? (GSD)
General Scene Difference

Write a detailed description of the given image. (LA)
Image Description

What are the male and female characters holding while they are together? (TVC)
TV Drama

What does the girl do after completing the routine? (DC)
Episode Memory

Reasoning

What might be the purpose of this gathering of people? (LA)
Social Reasoning

What might this young boy learn from this experience? (LA)
Commonsense Reasoning

Can you think of some reasons why cats might sit on someone's laptop? (LA)
Abductive Reasoning

How did the students advocate for immigration reform policies during the ceremony? (VIST)
Story Telling

Is it safe to walk on the floor while the woman is cleaning? (E4D)
Egocentric Video Reasoning

Planning

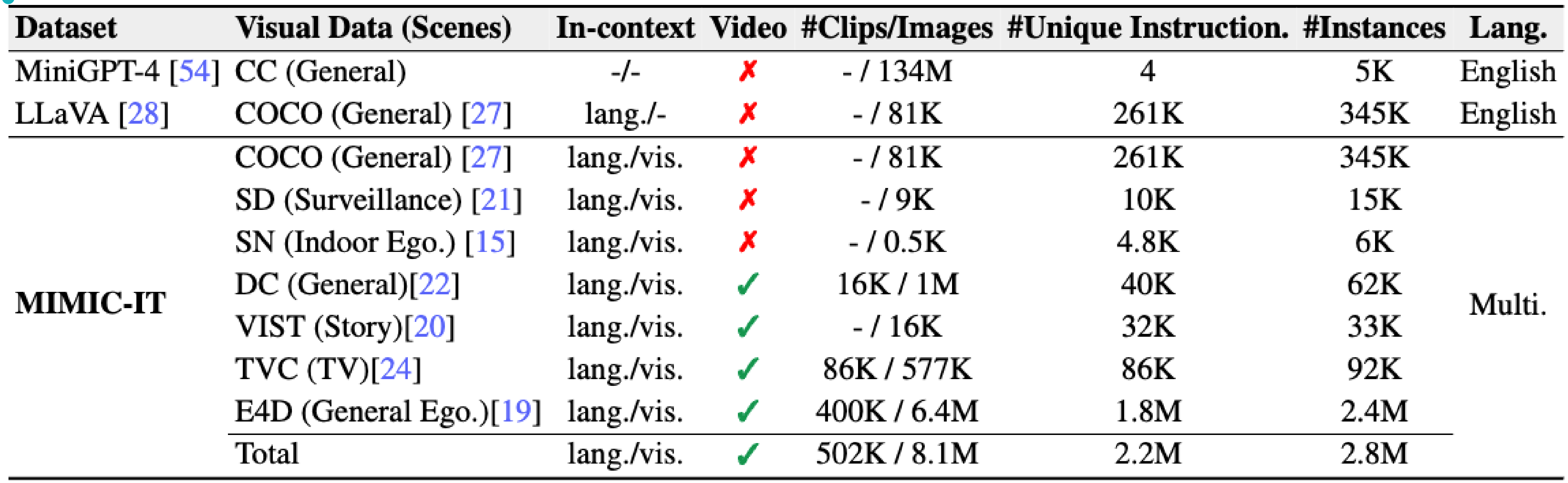
Can you suggest any ways to brighten up the room I'm in? (E4D)
Egocentric Planning

Can you suggest any recreational pursuits that can be enjoyed within this room? (IEP)
Indoor Event Planning

MIMIC-IT

MIMIC-IT Dataset

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE



Dataset	Visual Data (Scenes)	In-context	Video	#Clips/Images	#Unique Instruction.	#Instances	Lang.
MiniGPT-4 [54]	CC (General)	-/-	✗	- / 134M	4	5K	English
LLaVA [28]	COCO (General) [27]	lang./-	✗	- / 81K	261K	345K	English
MIMIC-IT	COCO (General) [27]	lang./vis.	✗	- / 81K	261K	345K	Multi.
	SD (Surveillance) [21]	lang./vis.	✗	- / 9K	10K	15K	
	SN (Indoor Ego.) [15]	lang./vis.	✗	- / 0.5K	4.8K	6K	
	DC (General)[22]	lang./vis.	✓	16K / 1M	40K	62K	
	VIST (Story)[20]	lang./vis.	✓	- / 16K	32K	33K	
	TVC (TV)[24]	lang./vis.	✓	86K / 577K	86K	92K	
	E4D (General Ego.)[19]	lang./vis.	✓	400K / 6.4M	1.8M	2.4M	
	Total	lang./vis.	✓	502K / 8.1M	2.2M	2.8M	

2.8M Instructions

Our dataset has 2.8M multimodal instruction-response pairs, with 2.2M unique instruc- tions derived from images and videos. Each pair is accompanied by multi-modal in-context information, forming conversational contexts aimed at empowering VLMs in perception, reasoning, and planning.

Multi-Modal In-context

Discover the first multi-modal in-context instruction dataset, a integrated compilation that seamlessly blends videos and images, spanning a diverse array of scenes.

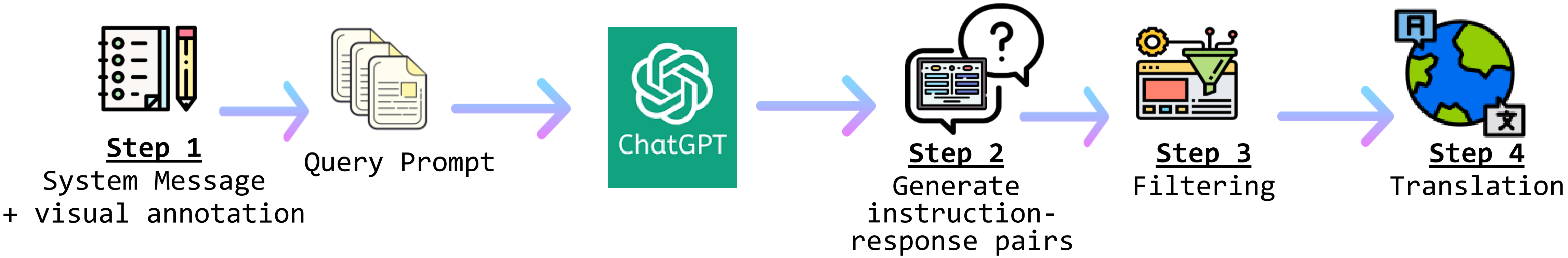
Multi-Lingual

Featuring 8 languages: English, Chinese, Korean, Japanese, German, French, Spanish, and Arabic, thereby allowing a larger global audience to altogthter enjoy from the convenience brought about by advancements in artificial intelligence.

Data Generation Process

- Design System Message and In-context Examples
  - e.g. Youtube Video Watcher (for Dense Caption task)
- Collect source data and add annotations
  - Obtain original annotations.
  - On EGO4D, we also get more annotations from off-shelf detection/caption model.
- Organize annotations and then query ChatGPT/GPT-4 to generate instruction-response pairs

Data Generation Process



Step 1
System Message
+ visual annotation

Query Prompt

Step 2
Generate
instruction-
response pairs

Step 3
Filtering

Step 4
Translation

Data Generation Process

System Message

You are an AI assistant that can understand videos, you can observe multiple frames in the given video. Your are provided with the timestamps and corresponding descriptions in following format:

timestamps: [[0, 19], [17, 60]]
sentences: ["A young woman is seen standing in a room and leads into her dancing."," The girl dances around the room while the camera captures her movements."]

The first description "A young woman is seen standing in a room and leads into her dancing." is happening from 0 to 19 seconds, while the second description is happening from 17 to 60 seconds.

Design a conversation between you and a person asking about this video. 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 video…

In-context Examples

timestamps: [[0, 91], [73, 174], [168, 212]]

sentences:
[“Several shots of boats are shown riding around as well as people riding on the boats and speaking to one another.”,
" Several shots are then shown of people sitting on the water as well as life under the water.",
" The camera pans around old cars under water as well as people in the area."]

Question: What is the main theme of this video?

Answer: The main theme of this video is centered around water activities, featuring scenes of boats, people on boats, underwater life, and exploring underwater areas with old cars.

Question: What are people doing on the boats?
Answer: People on the boats are riding around and engaging in conversations with one another.

Generated Instruction-Response Pairs

Query Inputs (Original Data)

timestamps: [[0, 7], [7, 18], [19, 37]]

sentences:
["We see a hallway with a wooden floor.",
" A dog in socks walks slowly out onto the floor as a lady films him.", " The dog turns around and goes back to the other room."]

GPT Response (Our Data)

question: What is the main subject of the video?
answer: The main subject of the video is a dog walking on a wooden floor while wearing socks.

question: What is the dog wearing on its feet?
answer: The dog is wearing socks on its feet.

question: Where is the dog walking?
answer: The dog is walking on a wooden floor in a hallway.

MIMIC-IT: Learning with In-context Examples

`LLaVA Complex Reasoning T2T (LACR_T2T)`

In-context examples

Instruction: What factors contribute to the impressive performance of the dog in this image?

Response: Several factors contribute to the impressive performance of the dog jumping high in the air to catch the frisbee…

Instruction: What can be inferred about the dog and its owner from the image?

Response: From the image, it can be inferred that the dog is well-trained, active, and enjoys playing with its owner..

Query

Instruction: What benefits might the dog's owner gain from this activity?

Response: Engaging the dog in an activity like playing with a frisbee and running in a field offers multiple benefits for both the dog and its owner…

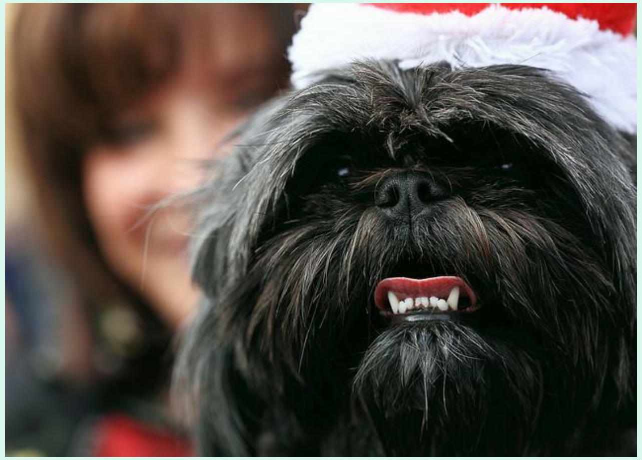
Liu et. al. Visual instruction tuning. 2023

MIMIC-IT: Learning with In-context Examples

LLaVA I2I (LACR_I2I)

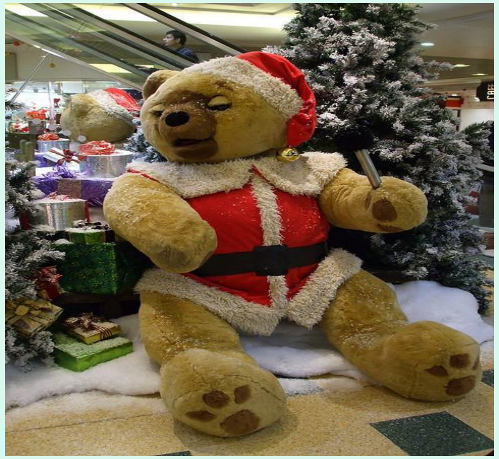
In-context examples

Instruction: What might be the possible reason for the dog's current expression and outfit?



Response: The dog seems to be wearing a festive Christmas hat and possibly panting with a slight smile, which could be attributed to the holiday season celebration…

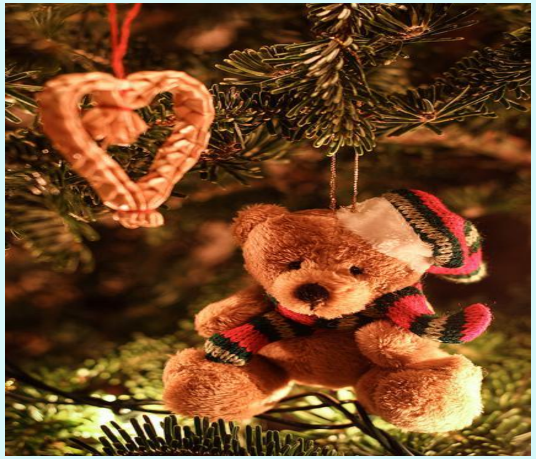
Instruction: What purpose does the large teddy bear dressed as Santa serve in this setting?



Response: The large teddy bear dressed as Santa serves the purpose of attracting and entertaining holiday shoppers in the department store setting…

Query

Instruction: What does the yteddy bear ornament on the Christmas tree signify?

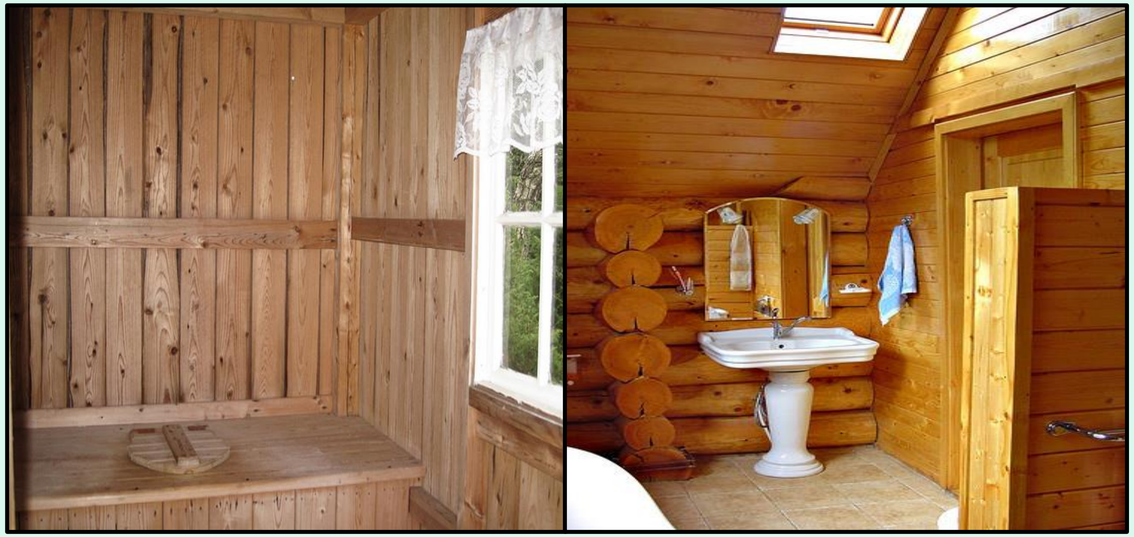


Response: The teddy bear ornament hanging on the Christmas tree conveys a sense of warmth, love, and nostalgia...

MIMIC-IT: Spot The Difference

Spot-The-Difference (General Difference)

Query



Instruction: What is the difference between these two images?
Response: The first one shows an outhouse with a toilet, while the second one shows…

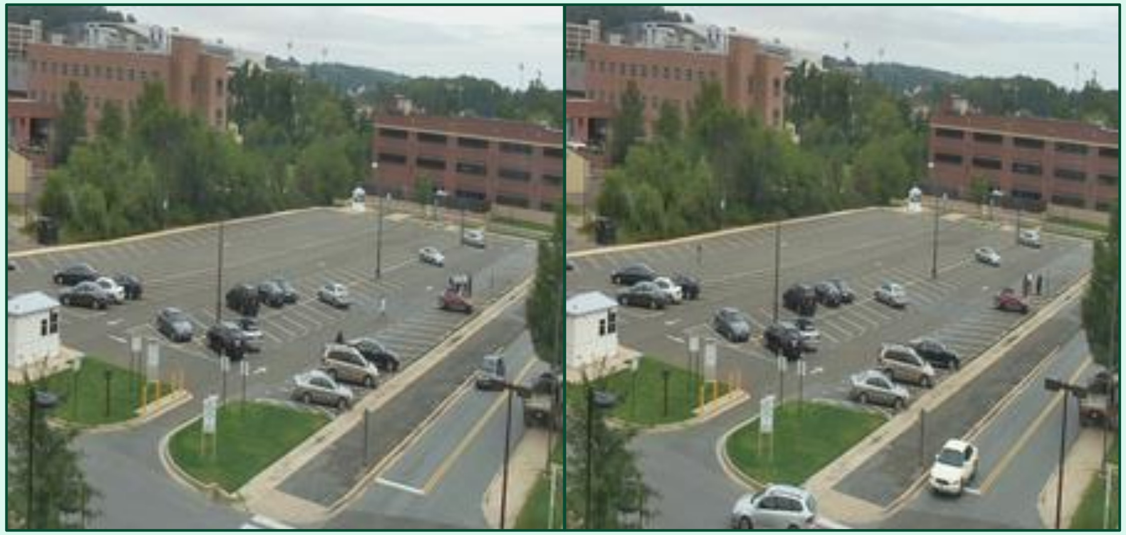
Instruction: Are there any objects in image two but not in image one?
Response: Yes, there are additional objects in image two that are not present in image one. Image two has a white sink, a bathtub, a mirror, and other bathroom items, which are not mentioned in image one.

Lin et. al. Microsoft coco: Common objects in context. 2014

MIMIC-IT: Spot The Difference

Spot-The-Difference (Subtle Difference)

Query



Instruction: What are the new vehicles that appear in the second image?

Response: There are two new vehicles in the second image, a silver hatchback and a white sedan on the road.

Jhamtani et. al. Learning to describe differences between pairs of similar images. 2018

MIMIC-IT: Dense Captions

Dense Caption (DC)

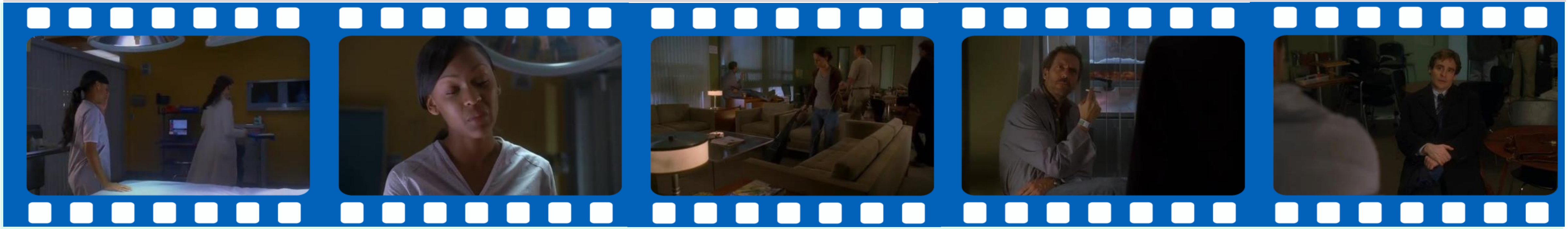
Instruction: What is the main activity shown in this video?
Response: The main activity shown in this video is jogging and running.

Instruction: Where is the man jogging?
Response: The man is jogging on the side of the road, along mountainsides, and around a neighborhood.

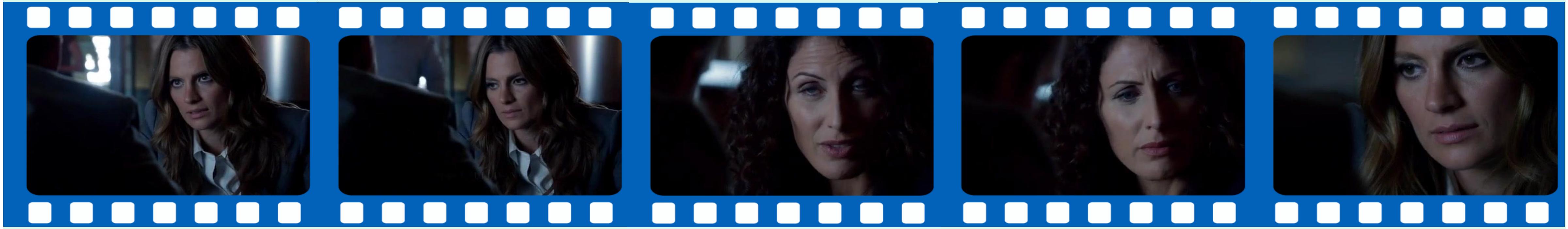
Krishna et. al. Dense- captioning events in videos. 2017

MIMIC-IT: TV Captions

TV Captions (TVC)



Instruction: What task is the woman performing in the room?
Response: The woman is vacuuming the rug area between chairs and couches, suggesting that she is cleaning the living space.

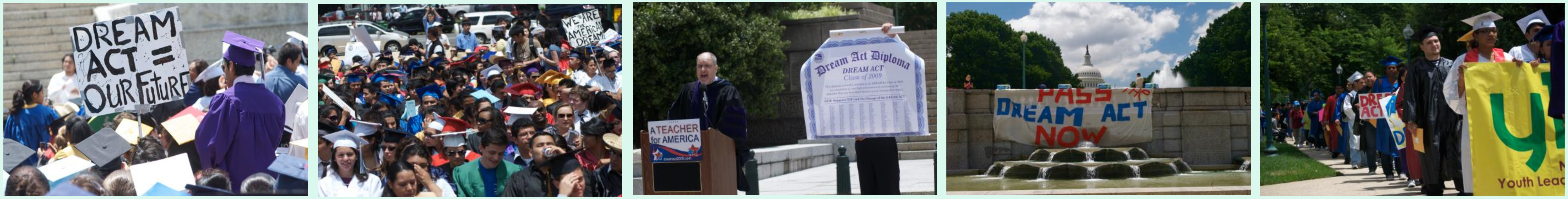


Instruction: How does the female character respond to the male character while he is talking to her?
Response: The female character shakes her head, which could indicate confusion, disagreement, or disapproval of what the male character is saying.

Lei et. al. TVR: A large-scale dataset for video-subtitle moment retrieval. 2020

MIMIC-IT: Visual Story Telling

Visual Story Telling (VIST)



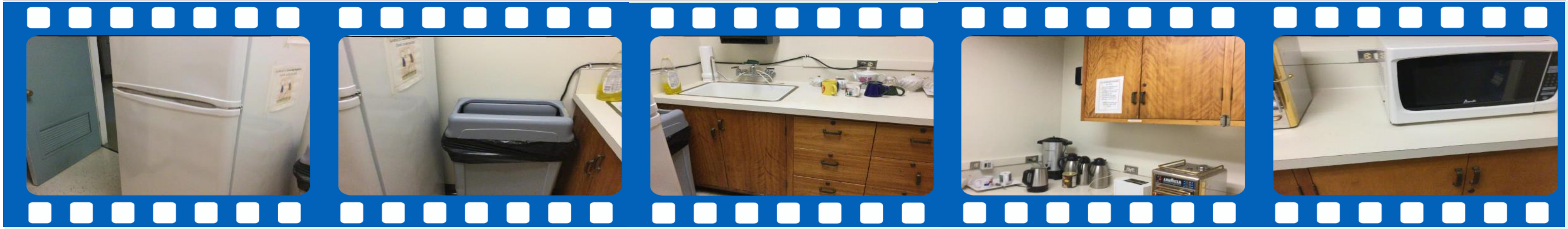
<u>Instruction:</u> How did the students advocate for immigration reform policies during the ceremony?

<u>Response:</u> The students who participated in the National DREAM Graduation ceremony advocated for immigration reform policies by fully dressing in graduation attire and rallying together to show their support for the DREAM Act. They also carried signs and banners promoting immigration reform and comprehensive immigration policies that would create a pathway to citizenship, keep families together, and promote the social, economic, and political integration of all immigrants.

Huang et. al. Visual storytelling. 2016

MIMIC-IT: Indoor Event Planning

Indoor Event Planning (SN)



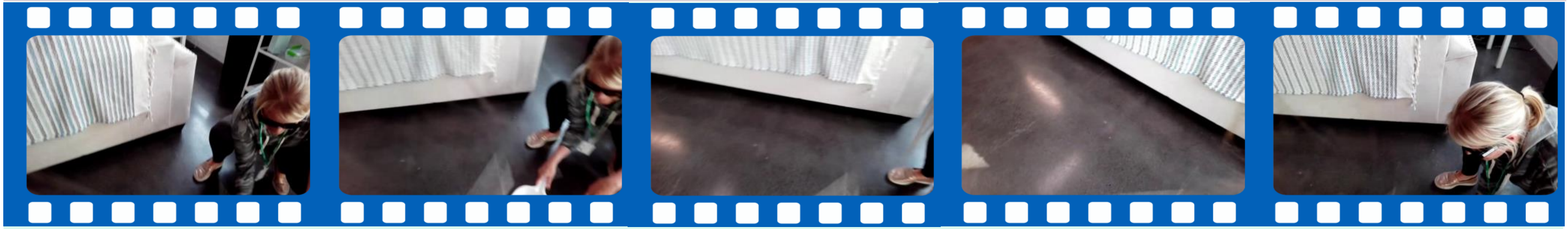
Instruction: Can you suggest any recreational pursuits that can be enjoyed within this room?
Response: I think Home Brewing is a terrific choice. The kitchen area in the room seems like an ideal spot for a home brewing activity, given the presence of a sink, a refrigerator, and a kitchen counter

Instruction: Is there enough space in the room to set up a home brewing station?
Response: The room seems to have enough space to set up a home brewing station, but you might need to rearrange some furniture to create more space. For example, you could move the brown cabinet with the coffee pot and cups to another location to create more counter space. You could also move the blue bin away from the door to create more space for movement

Dai et. al. Scannet: Richly-annotated 3d reconstructions of indoor scenes. 2017

MIMIC-IT: Egocentric Video Understanding

EGO4D (E4D)



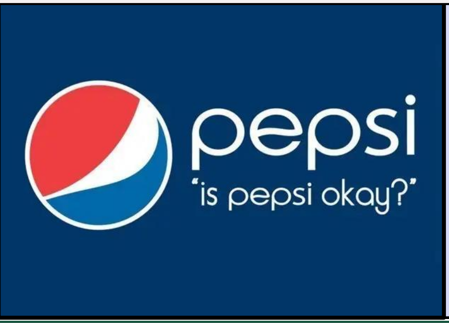
Instruction: Is it safe to walk on the floor while the woman is cleaning?

Response: Based on what I saw, it's best to avoid walking on the floor while it's being cleaned to prevent slipping or accidentally disrupting the cleaning process. Would you like me to suggest ways to stay safe while the floor is being cleaned?

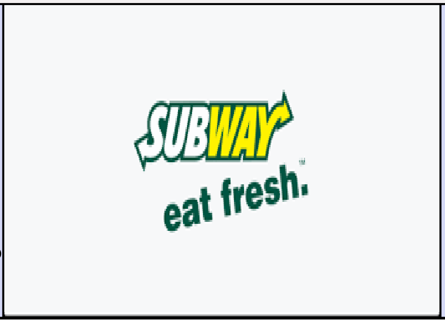
Grauman et. al. Ego4d: Around the world in 3,000 hours of egocentric video. 2022

Otter’s Capacities Preview

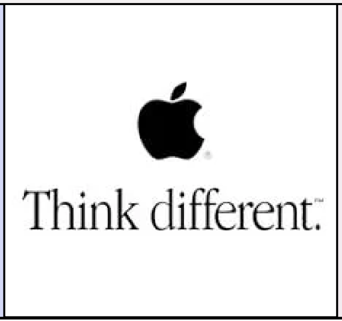
Learning with In-context Examples



Instruction:
What’s written on this image?
Otter: pepsi, is pepsi okay?

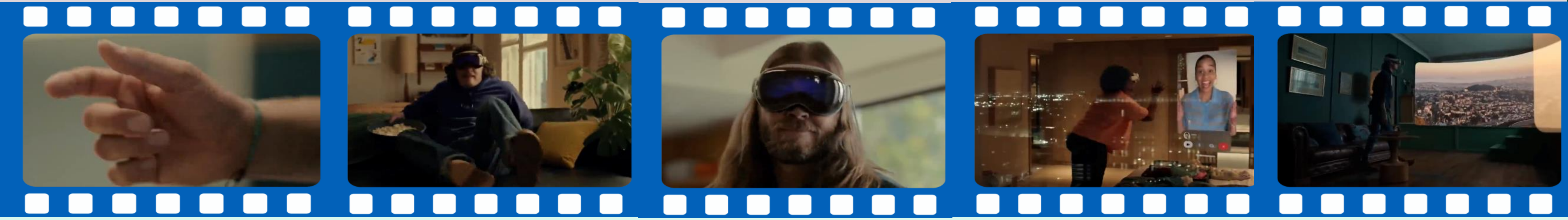


Instruction:
What’s written on this image?
Otter: SUBWAY, eat fresh.



Instruction:
What’s written on this image?
(wo/ICE) *Otter:* “Think Different” is written on this image.
(w/ICE) *Otter:* Think Different

Understanding Videos (trained on MIMIC-IT-DC)



Instruction: Hey Otter, do you think it looks cool?

Otter: Yes, I do think it is cool that the man is playing video games while wearing a head-mounted display.

FunQA: Surprising Video Understanding

FunQA

HumorQA Demo

FunQA: Surprising Video Understanding

Counter-intuitive Timestamp Localization Task

Counter-intuitiveness Reasoning Task

Detailed Description Task

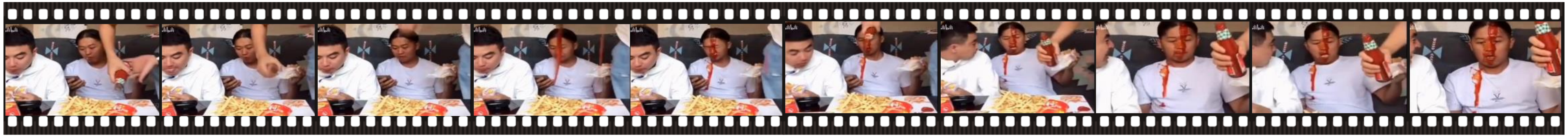
Higher Level Tasks

HumorQA:
1.8K videos, avg. 7s
141.3K QA pairs

CreativeQA:
0.9K videos, avg. 48s
78.7K QA pairs

MagicQA:
1.6K videos, avg. 10s
91.9K QA pairs

FunQA: Surprising Video Understanding



H2: **Describe** the comedic scene in the video

The comedic scenes depict a man acting as a messenger, pretending to have a seizure and falling to the ground, and a man pretending his face is on fire while trying to eat a French fry. ✗

One person was multitasking by eating while using their phone, and another person accidentally flicked ketchup onto their face. ✓

H3: **Can you explain why the content is humorous?**

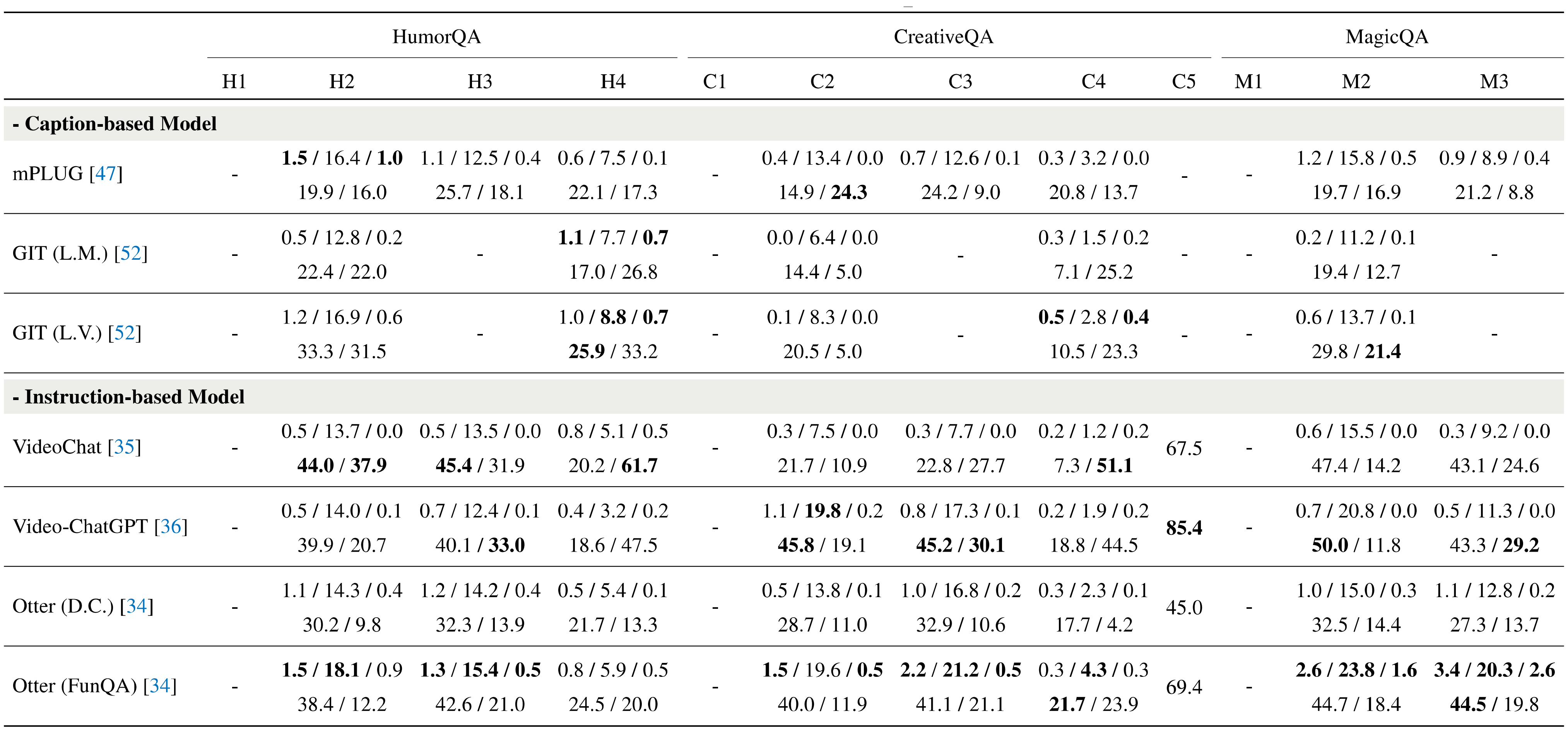
The video' s humorous elements include the man's over-the-top reaction to eating the fries, his attempts to eat the fry with his mouth open, and the messenger's role-playing as a seizing man. ✗

The most amusing aspect of the video is that the man had barely started his meal when he had to rush to answer the phone. He had to hold the burger in one hand and use the other to drag himself away, which was quite humorous. ✗

Otter (Trained on Dense Caption)

Otter (trained on FunQA)

FunQA: Surprising Video Understanding



	HumorQA				CreativeQA					MagicQA		
	H1	H2	H3	H4	C1	C2	C3	C4	C5	M1	M2	M3
- Caption-based Model												
mPLUG [47]	-	**1.5** / 16.4 / **1.0** 19.9 / 16.0	1.1 / 12.5 / 0.4 25.7 / 18.1	0.6 / 7.5 / 0.1 22.1 / 17.3	-	0.4 / 13.4 / 0.0 14.9 / **24.3**	0.7 / 12.6 / 0.1 24.2 / 9.0	0.3 / 3.2 / 0.0 20.8 / 13.7	-	-	1.2 / 15.8 / 0.5 19.7 / 16.9	0.9 / 8.9 / 0.4 21.2 / 8.8
GIT (L.M.) [52]	-	0.5 / 12.8 / 0.2 22.4 / 22.0	-	**1.1** / 7.7 / **0.7** 17.0 / 26.8	-	0.0 / 6.4 / 0.0 14.4 / 5.0	-	0.3 / 1.5 / 0.2 7.1 / 25.2	-	-	0.2 / 11.2 / 0.1 19.4 / 12.7	-
GIT (L.V.) [52]	-	1.2 / 16.9 / 0.6 33.3 / 31.5	-	1.0 / **8.8** / **0.7** **25.9** / 33.2	-	0.1 / 8.3 / 0.0 20.5 / 5.0	-	**0.5** / 2.8 / **0.4** 10.5 / 23.3	-	-	0.6 / 13.7 / 0.1 29.8 / **21.4**	-
- Instruction-based Model												
VideoChat [35]	-	0.5 / 13.7 / 0.0 **44.0** / **37.9**	0.5 / 13.5 / 0.0 **45.4** / 31.9	0.8 / 5.1 / 0.5 20.2 / **61.7**	-	0.3 / 7.5 / 0.0 21.7 / 10.9	0.3 / 7.7 / 0.0 22.8 / 27.7	0.2 / 1.2 / 0.2 7.3 / **51.1**	67.5	-	0.6 / 15.5 / 0.0 47.4 / 14.2	0.3 / 9.2 / 0.0 43.1 / 24.6
Video-ChatGPT [36]	-	0.5 / 14.0 / 0.1 39.9 / 20.7	0.7 / 12.4 / 0.1 40.1 / **33.0**	0.4 / 3.2 / 0.2 18.6 / 47.5	-	1.1 / **19.8** / 0.2 **45.8** / 19.1	0.8 / 17.3 / 0.1 **45.2** / **30.1**	0.2 / 1.9 / 0.2 18.8 / 44.5	**85.4**	-	0.7 / 20.8 / 0.0 **50.0** / 11.8	0.5 / 11.3 / 0.0 43.3 / **29.2**
Otter (D.C.) [34]	-	1.1 / 14.3 / 0.4 30.2 / 9.8	1.2 / 14.2 / 0.4 32.3 / 13.9	0.5 / 5.4 / 0.1 21.7 / 13.3	-	0.5 / 13.8 / 0.1 28.7 / 11.0	1.0 / 16.8 / 0.2 32.9 / 10.6	0.3 / 2.3 / 0.1 17.7 / 4.2	45.0	-	1.0 / 15.0 / 0.3 32.5 / 14.4	1.1 / 12.8 / 0.2 27.3 / 13.7
Otter (FunQA) [34]	-	**1.5** / **18.1** / 0.9 38.4 / 12.2	**1.3** / **15.4** / **0.5** 42.6 / 21.0	0.8 / 5.9 / 0.5 24.5 / 20.0	-	**1.5** / 19.6 / **0.5** 40.0 / 11.9	**2.2** / **21.2** / **0.5** 41.1 / 21.1	0.3 / **4.3** / 0.3 **21.7** / 23.9	69.4	-	**2.6** / **23.8** / **1.6** 44.7 / 18.4	**3.4** / **20.3** / **2.6** **44.5** / 19.8

Otter gains significant improvement after training on FunQA training set, but the FunQA benchmark is still very challenging,

Otter’s Interactive Demo

- Otter Video: Otter Chat Video (cliangyu.com)
- Otter Image: Otter Chat Image (cliangyu.com)

Let's build Otter, or your multi-modal model through "MIMIC-IT"

luodian/otter-9b-dc-hf
luodian/otter-9b-hf

GitHub

luodian/otter

Acknowledgement

Jingkang Yang

Bo Li

Yuanhan Zhang