Building Open-World Multi-Modal AI Assistant

Ziwei Liu

Nanyang Technological University

AI Assistant with Scene Graph

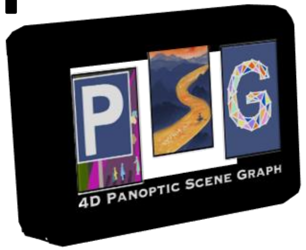
Neuro-Symbolic Assistant

PANOPTIC SCENE GRAPH

Panoptic Scene Graph [ECCV'22]

PANOPTIC VIDEO SCENE GRAPH

Panoptic **Video** Scene Graph [CVPR'23]



4D Panoptic Scene Graph [NeurIPS'23]

Jingkang Yang

Bo Li

Yuanhan Zhang

MMLab @NTU

AI Assistant with VLMs

OTTER

Deep diving into large Multi-modality Models

End-to-End Assistant

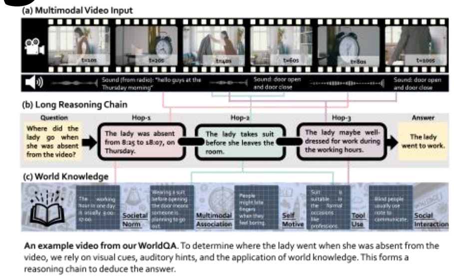
OTTER-HD

A High-Res Multi-modality Model

LLaVA-NeXT

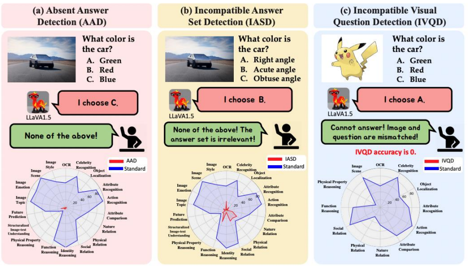
Benchmarking AI Assistants

LMM-Eval



WorldQA

FunQA



Unsolvable Problem Detection

Embodied AI Assistant

Octopus

AI Assistant with Scene Graph

Neuro-Symbolic Assistant

Panoptic Scene Graph [ECCV'22]

Panoptic **Video** Scene Graph [CVPR'23]

4D Panoptic Scene Graph [NeurIPS'23]

Jingkang Yang

Bo Li

Yuanhan Zhang

MMLab @NTU

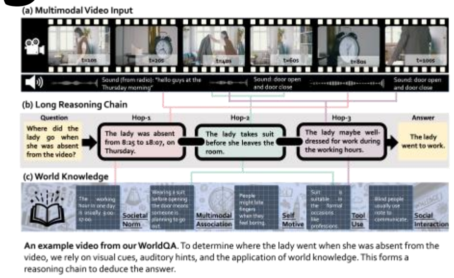
AI Assistant with VLMs

End-to-End Assistant

LLaVA-NeXT

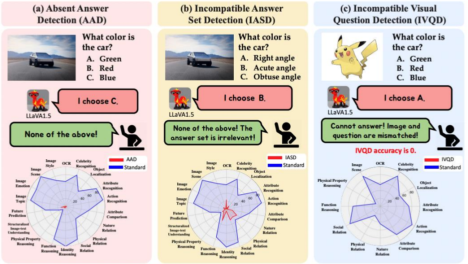
Benchmarking AI Assistants

LMM-Eval



WorldQA

FunQA



Unsolvable Problem Detection

Embodied AI Assistant

Octopus


April 7, 2024
11:06:54
Give way to her
Your Contact
Fangzhou
Your Contact
Yushi
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

[ECCV'22] Panoptic Scene Graph

2D Video

[CVPR'23] Panoptic **Video** Scene Graph

3D video

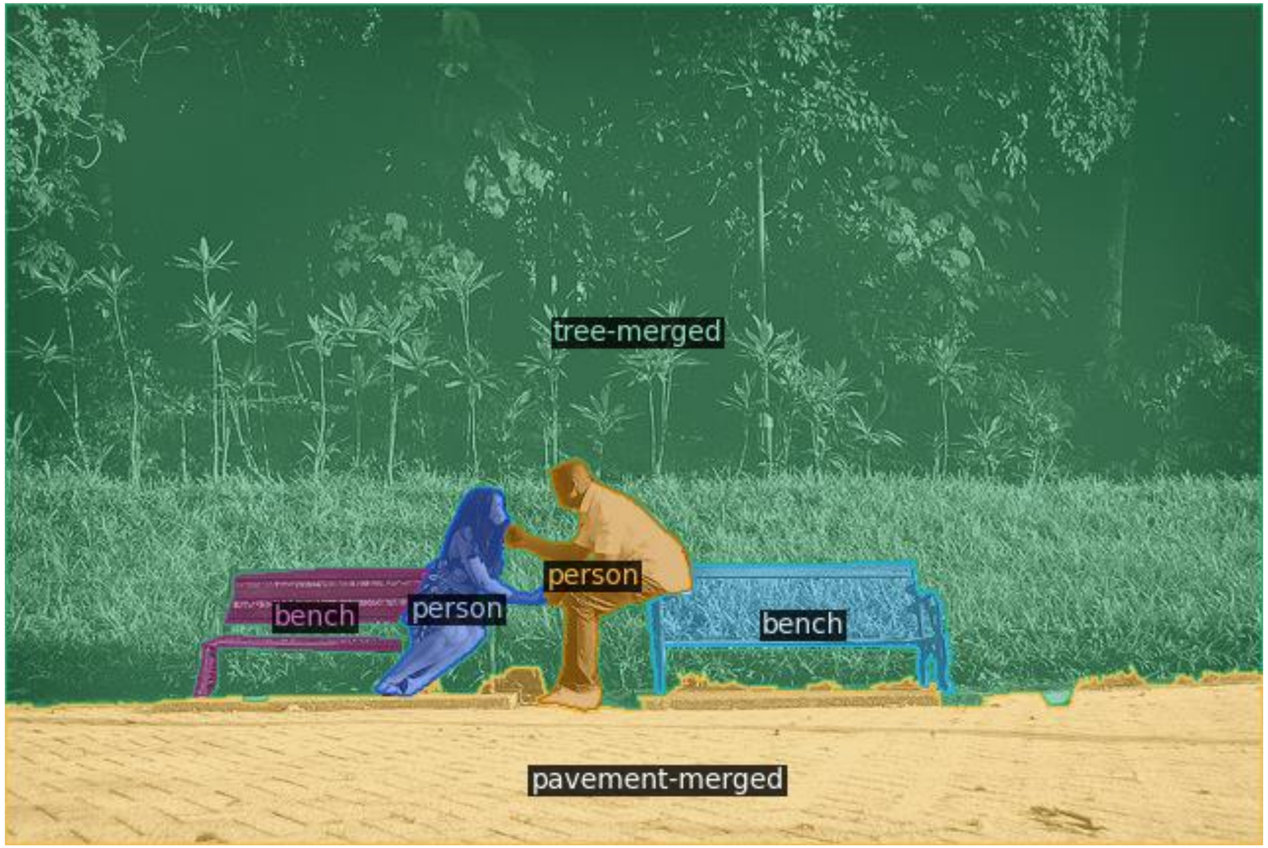
[NeurIPS'23] **4D** Panoptic Scene Graph

Beyond Object Recognition

Q: What is in the image?

Q: What happened in the image?

Beyond Object Recognition

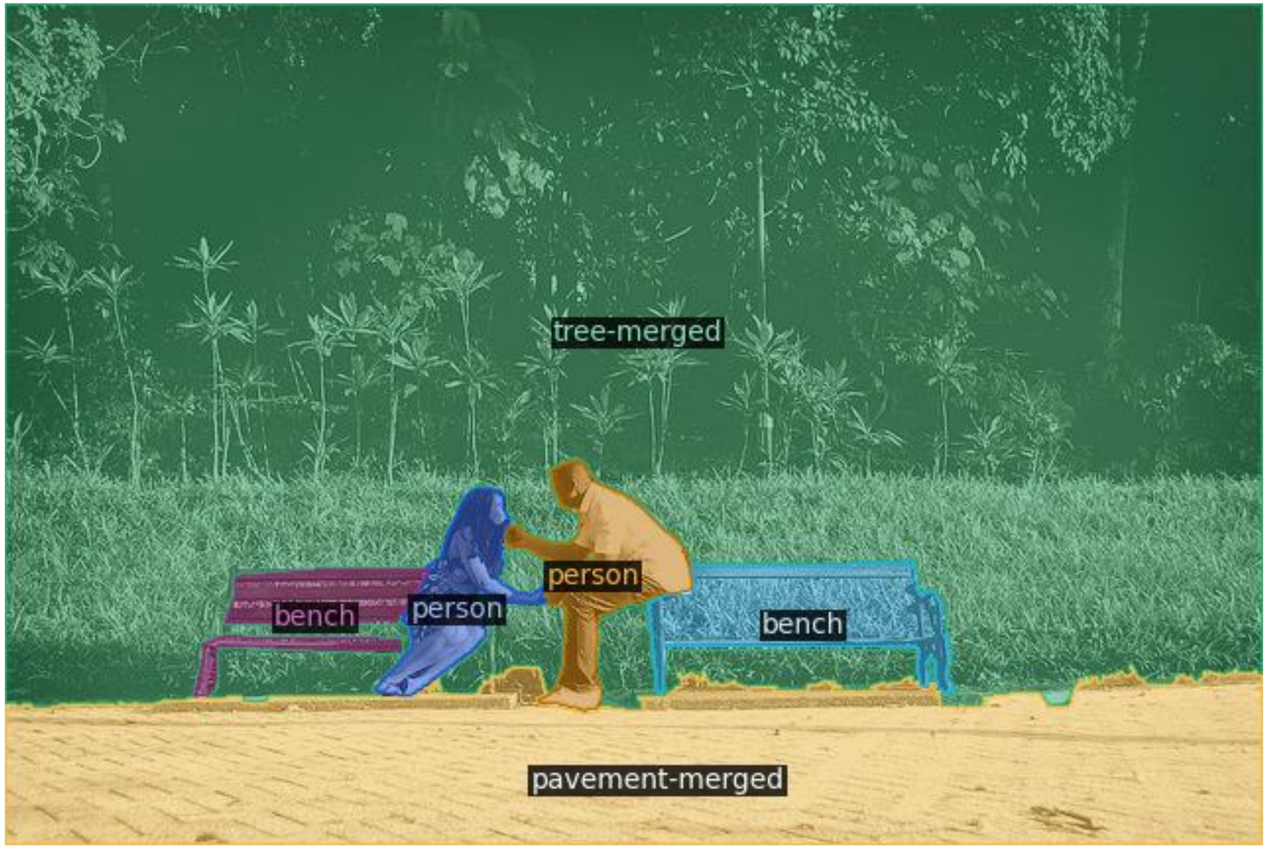


Q: What is in the image?

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

Q: What happened in the image?

Beyond Object Recognition

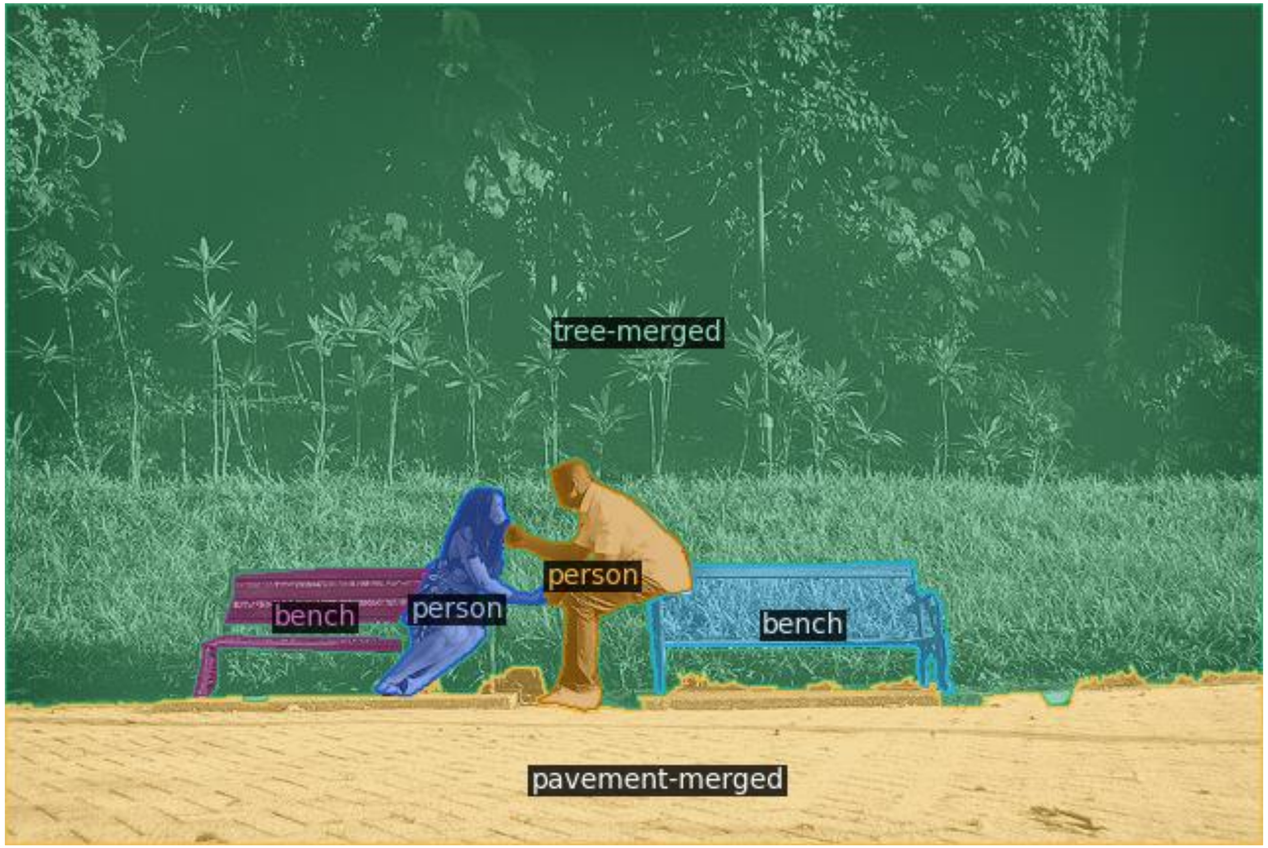


Q: What is in the image?

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

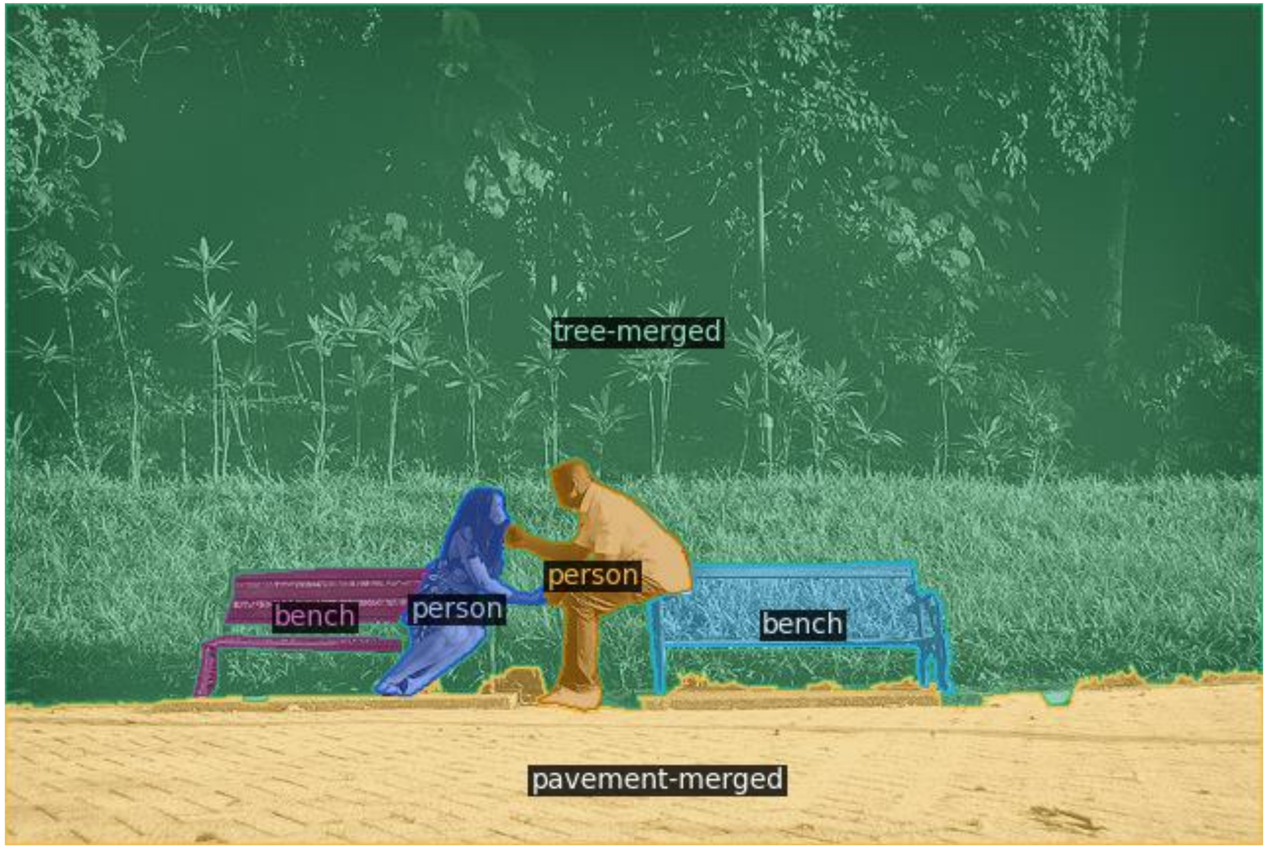
Q: What happened in the image?

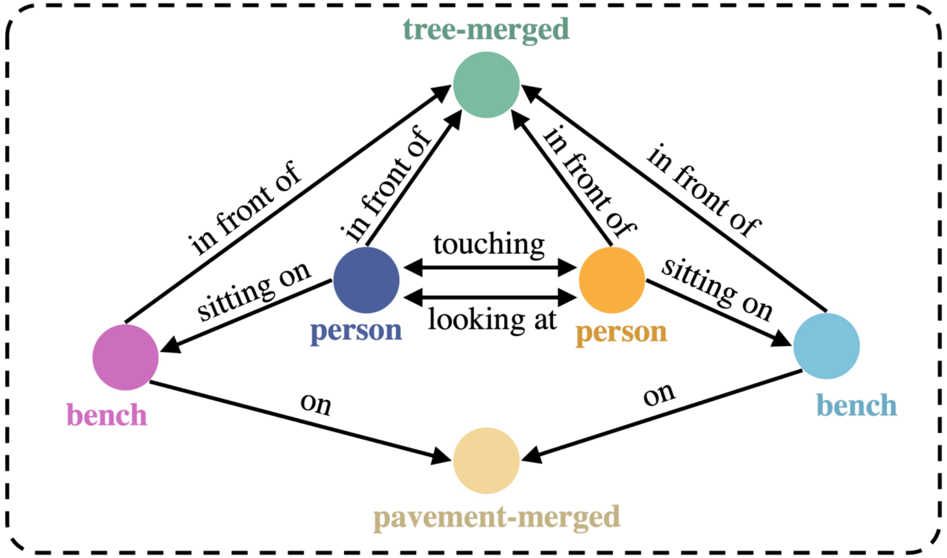
Q: What happened in the image?



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**.

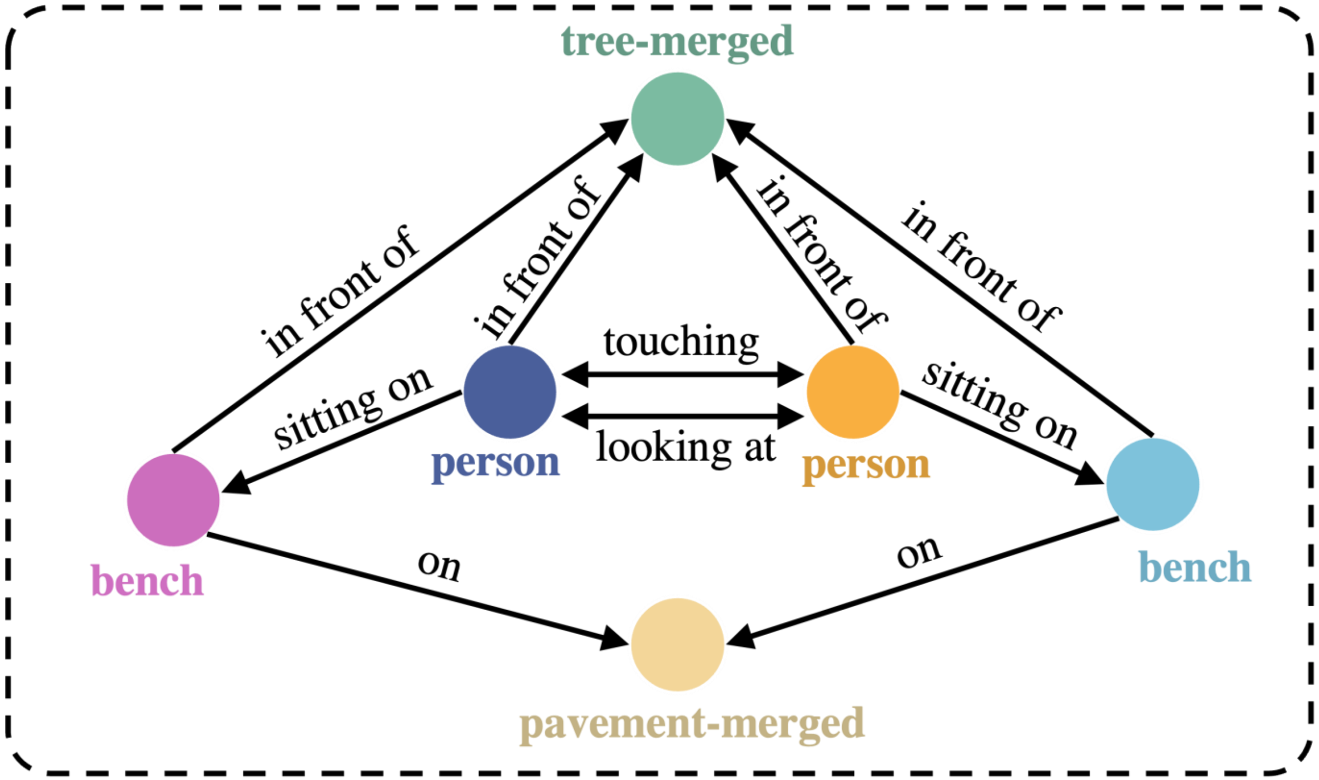
Q: What happened in the image?





Scene Graph

Beyond Object Recognition



Using scene graph:

Q: What happened in the image?

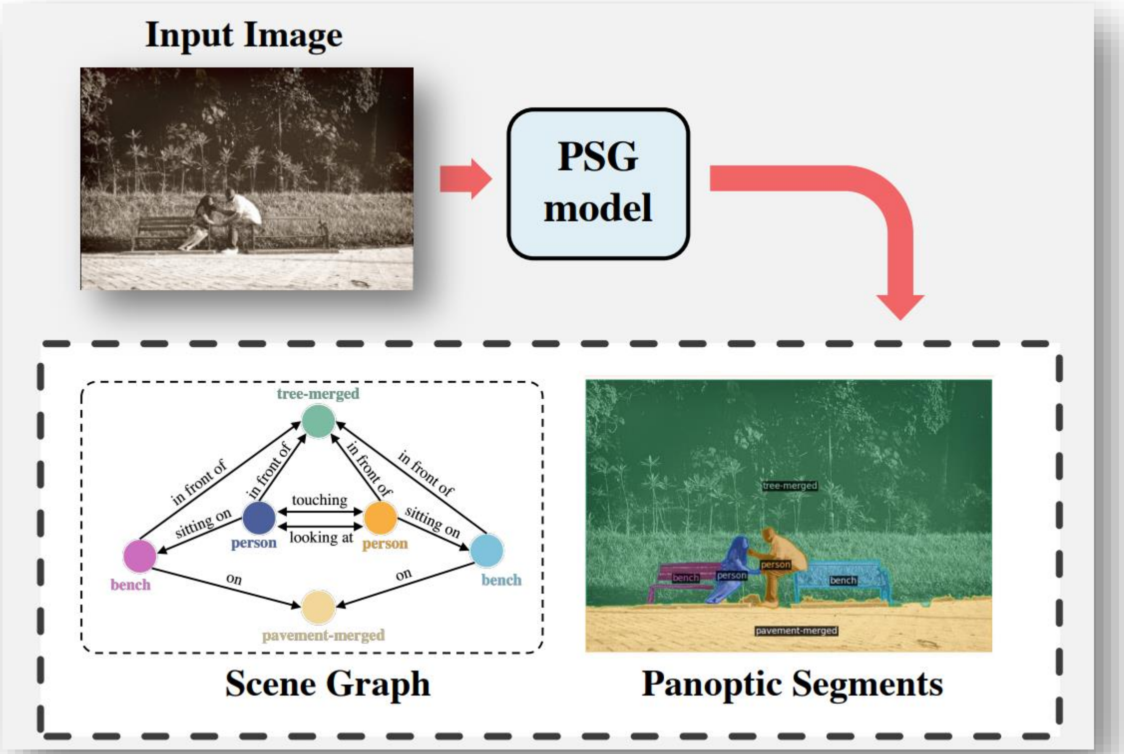
Q: Where is **the man** sitting on?

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

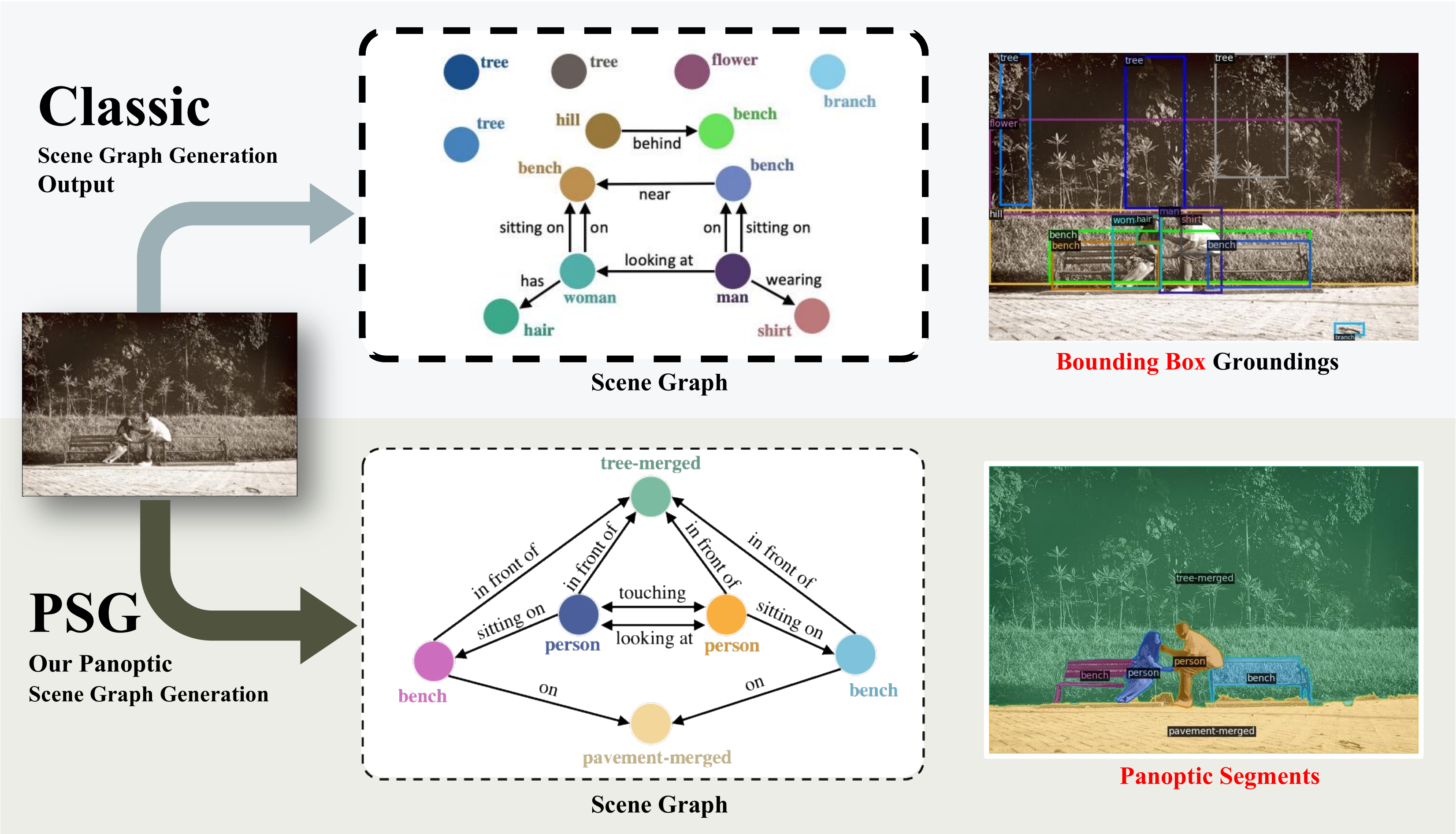
Output:

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



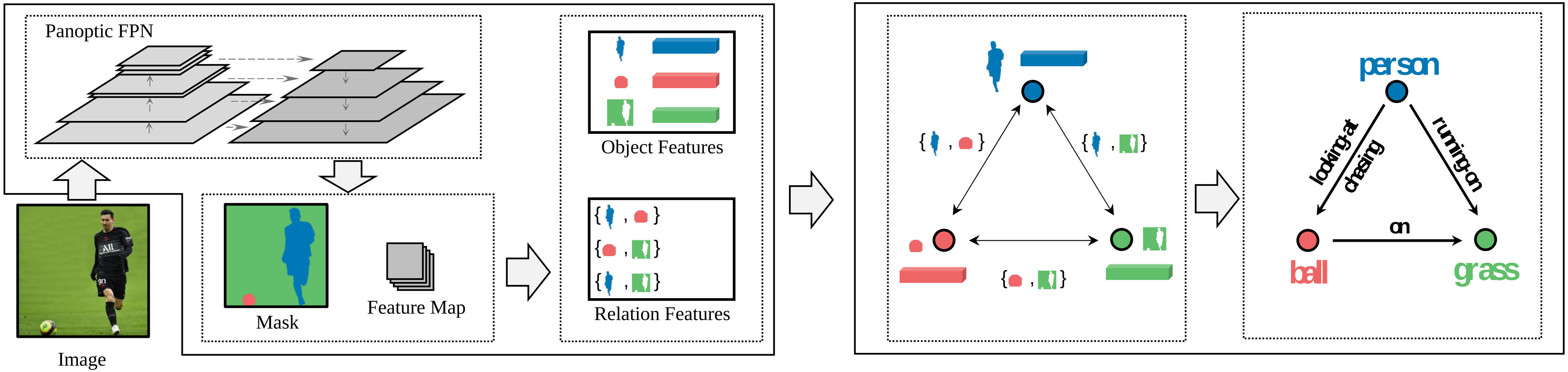
PSG Dataset

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

http://psgdataset.org/explore.html

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

Two Stage Methods

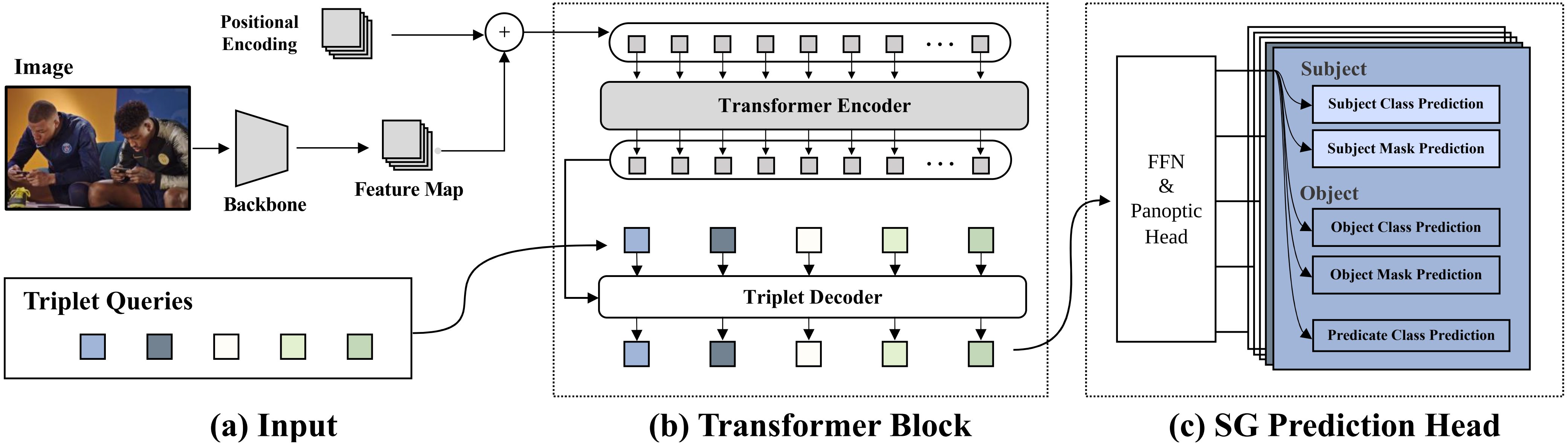


(a) Stage-1: Segment Feature Extractor

(b) Stage-2: Scene Graph Prediction

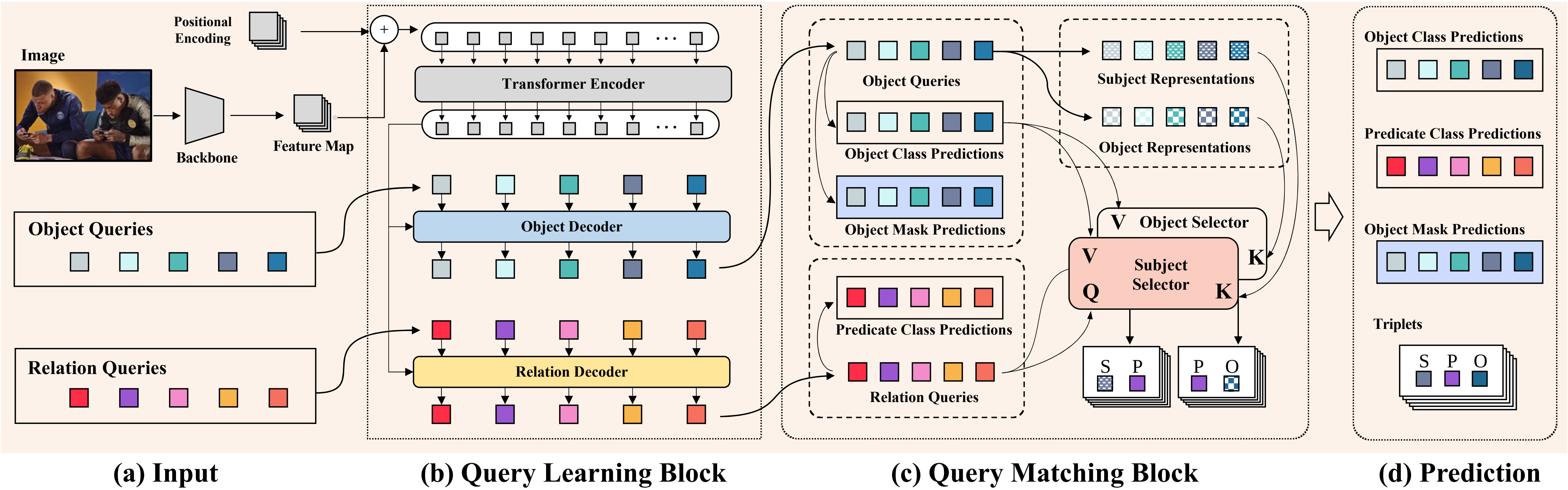
+ Fast, Simple, Easy to use
+ Support Classic Methods
- Heavily Rely on Detectors

One Stage Method (PSGTR)



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

One Stage Method (PSGFormer)



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

PSG: Panoptic Scene Graph

Prediction Result from the PSG Model

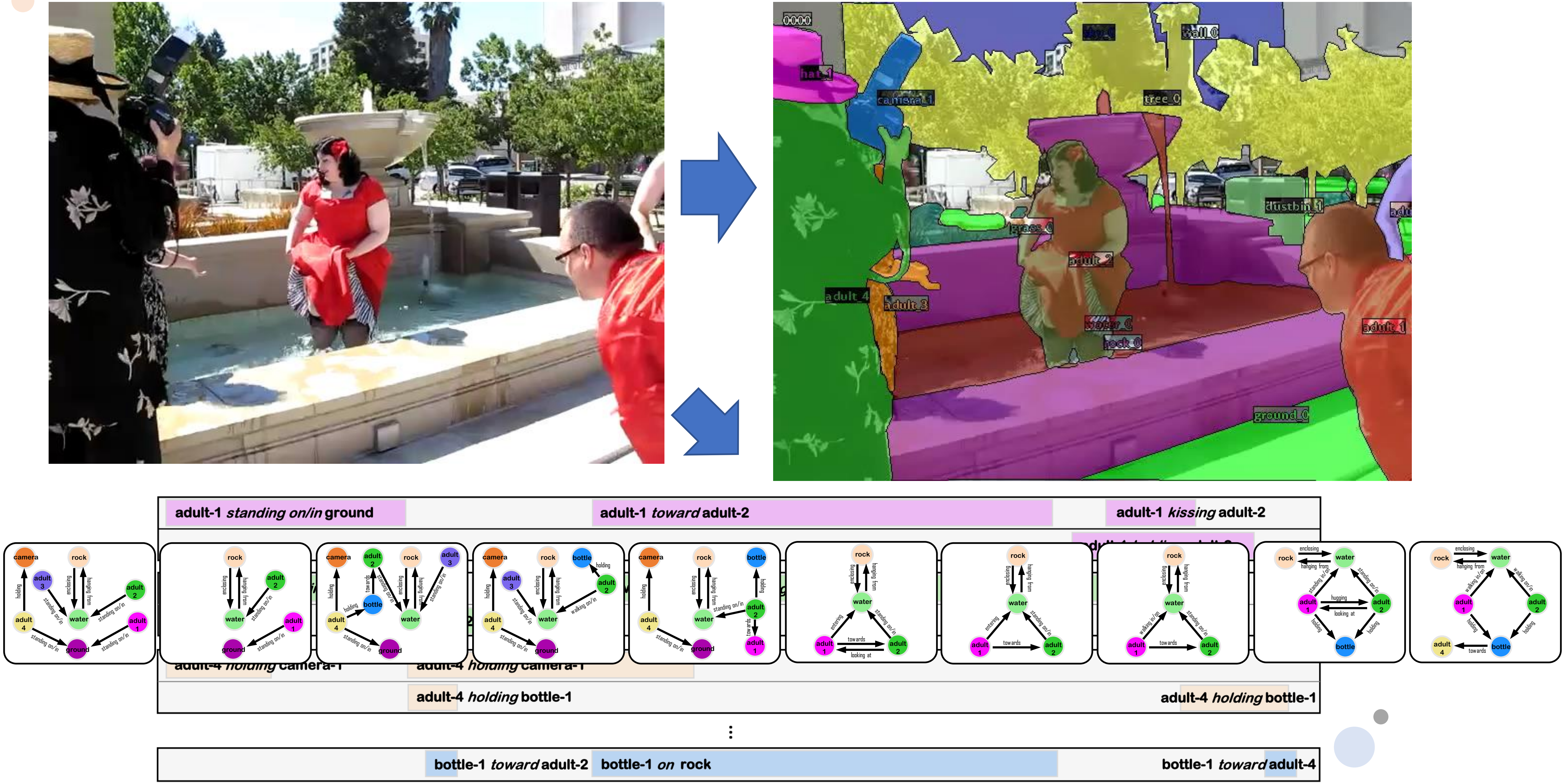
Q: Who is wearing a fancy bag in the photo?

Q: Where is the man with fancy bag standing?

Panoptic Video Scene Graph

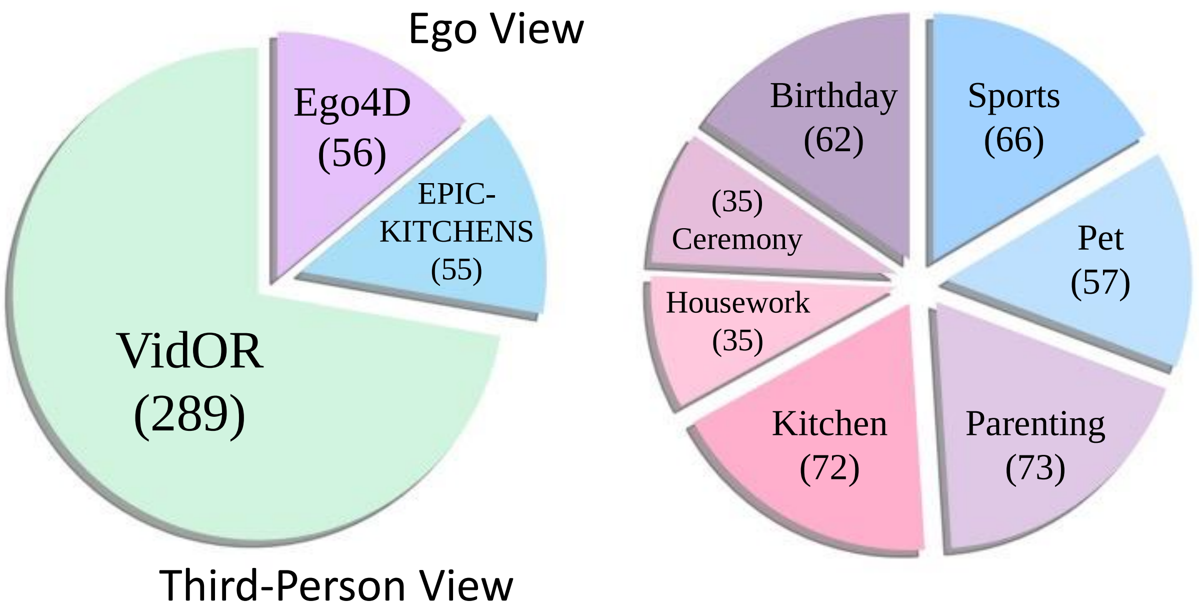
PSG + Video = PVSG

PVSG - Panoptic Video Scene Graph Generation



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

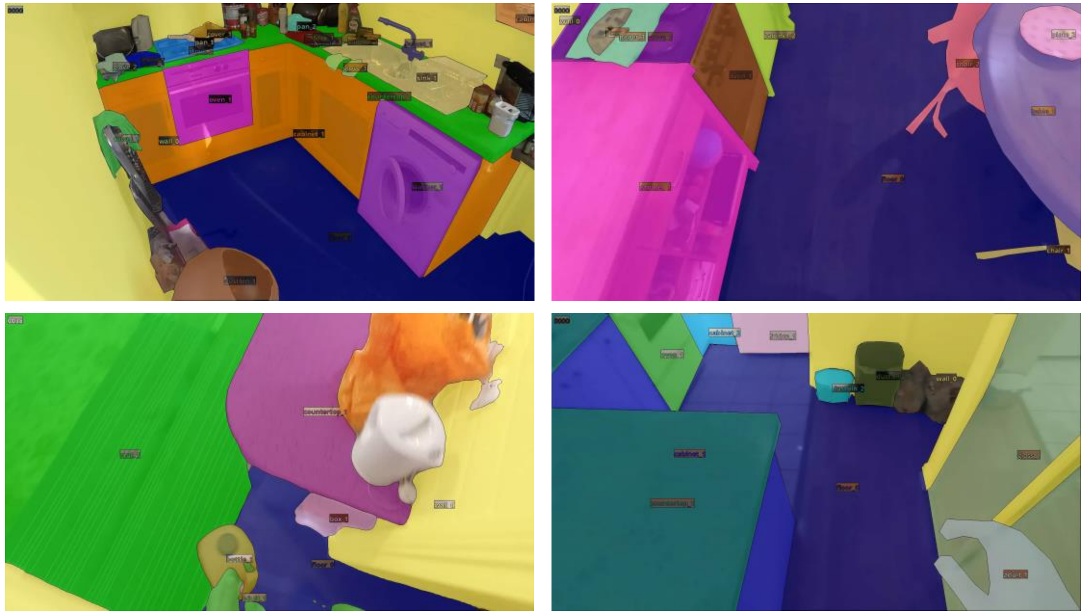
0000-0018: The **little boy (child-1)** passed through **the television (tv-1)** to pick up **a gift (gift-1).**

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)**.

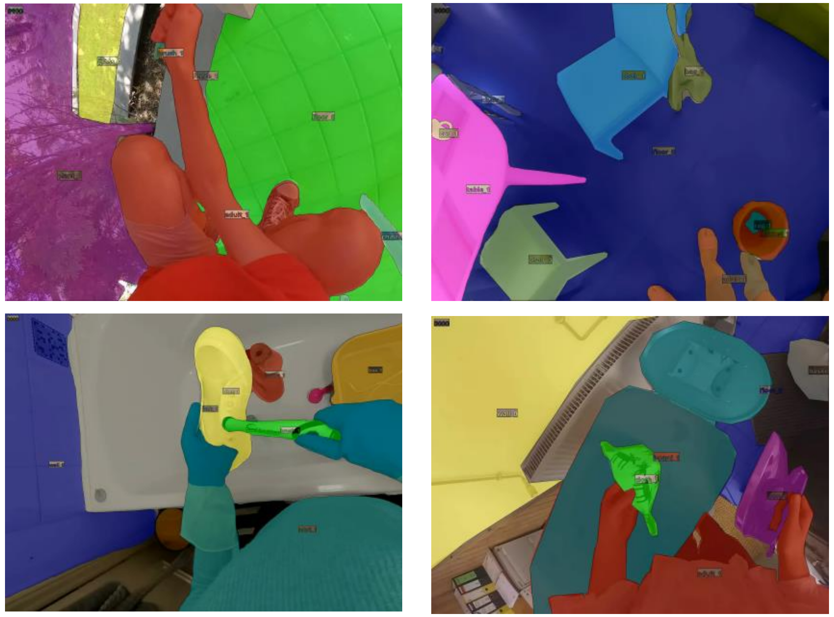
Dense QA

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)

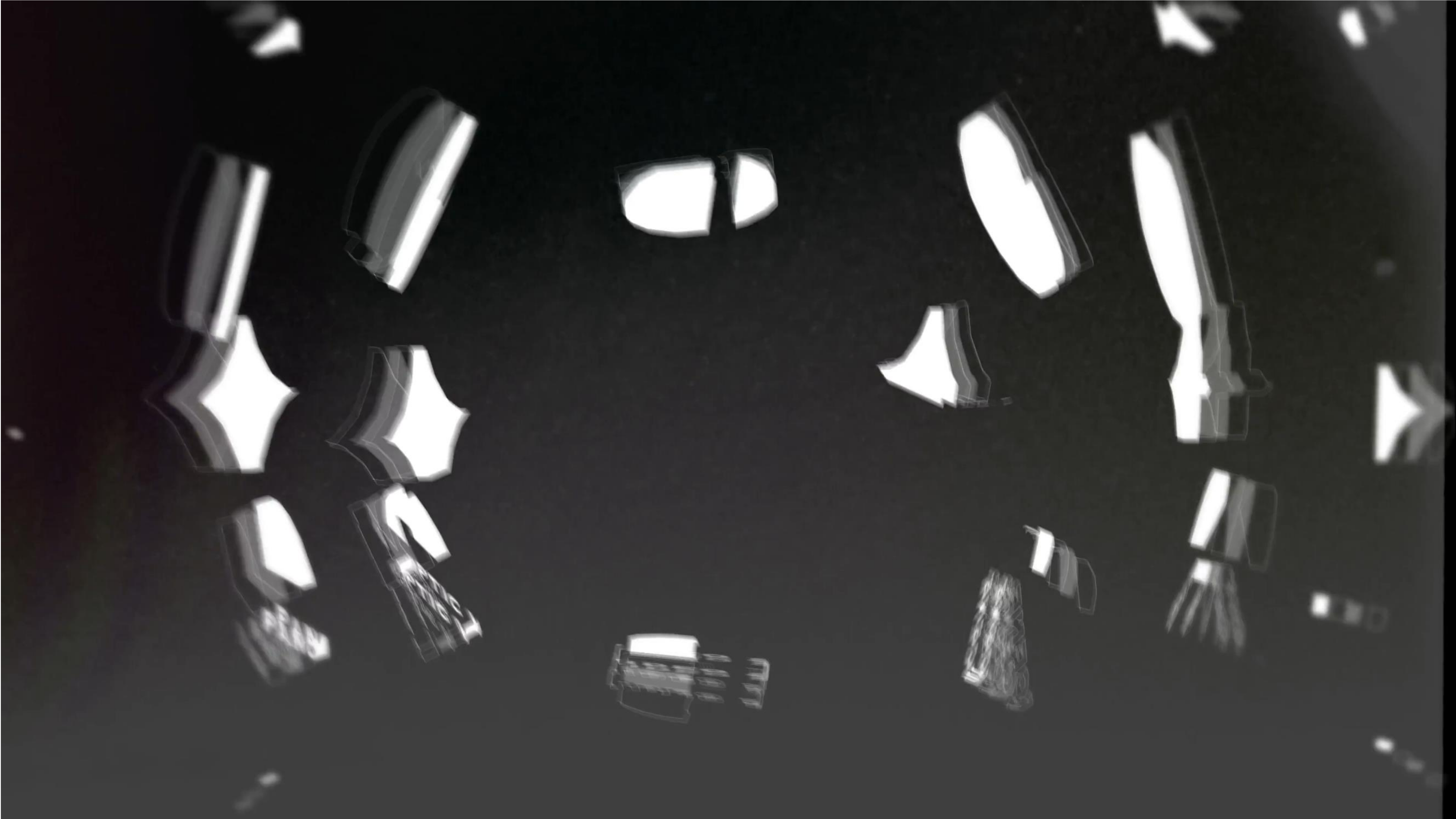


EpicKitchen (55 videos)

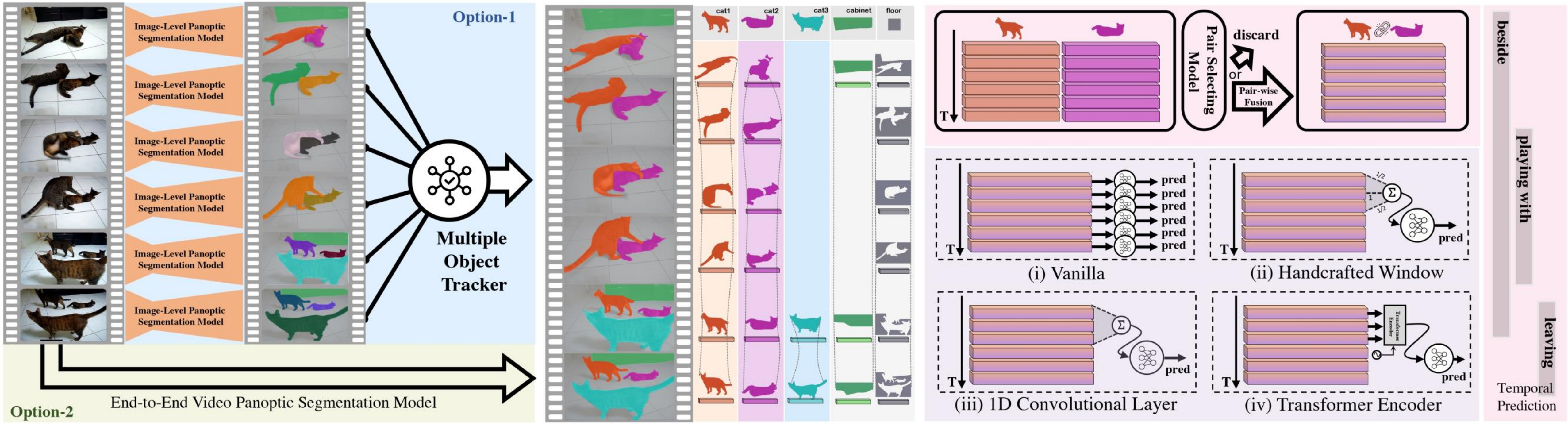


Ego4D (56 videos)

Towards Comprehensive Egocentric Video Scene Understanding



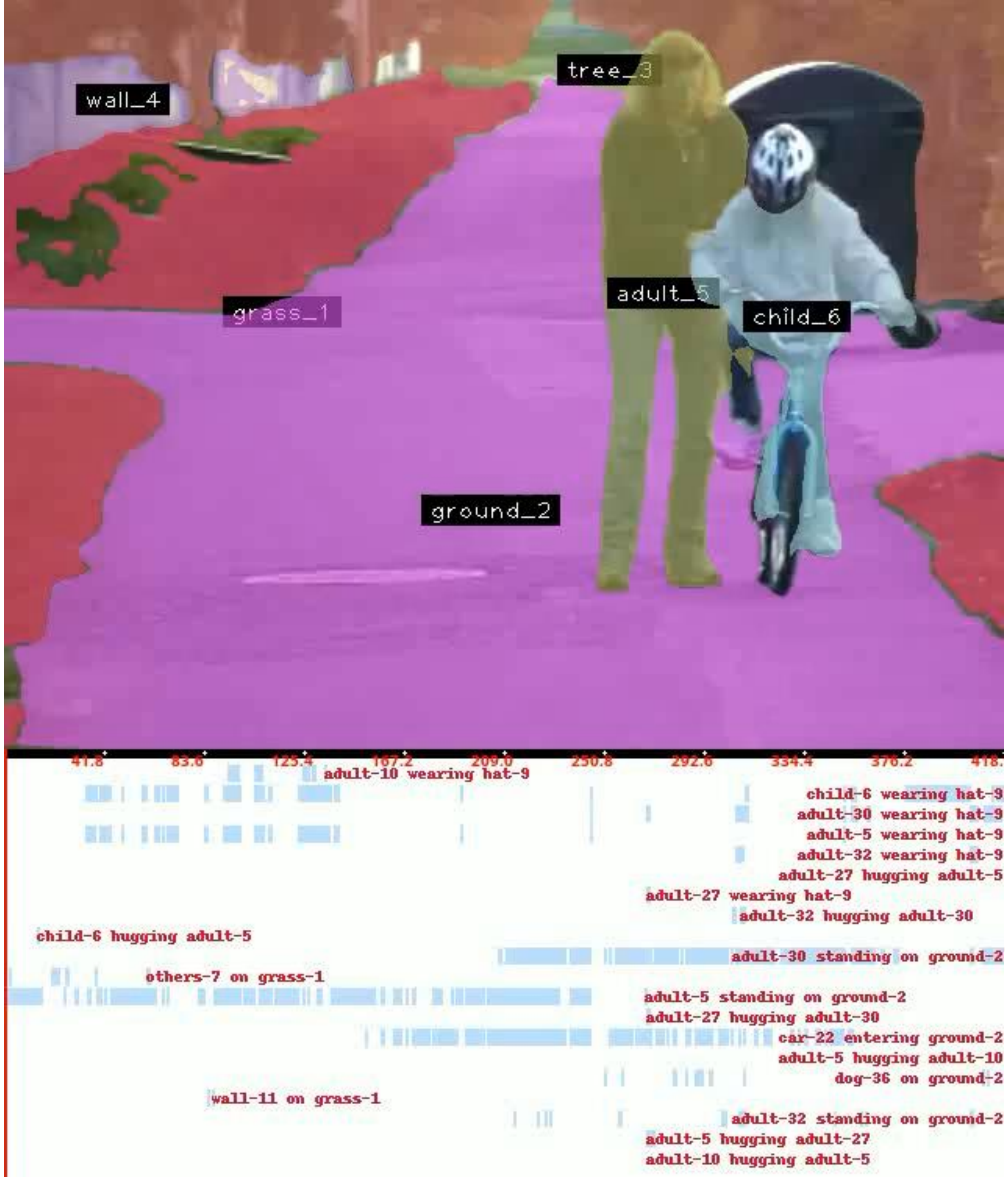
Model



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

(b) Stage 2: Relation Prediction

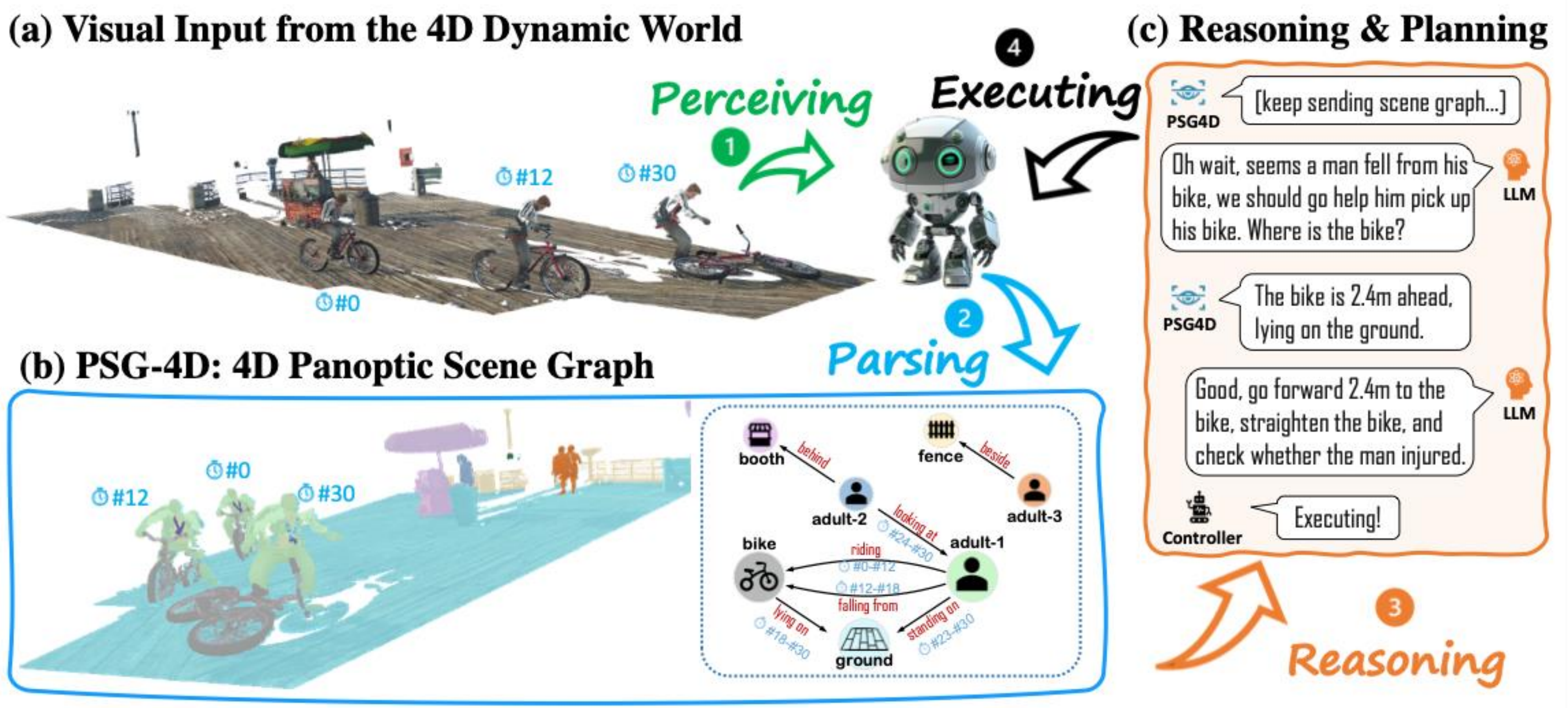
Demo Results



4D Panoptic Scene Graph

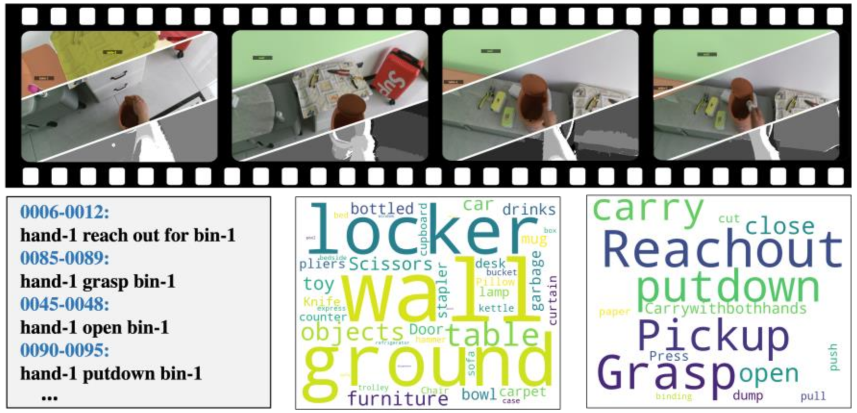
PSG + Video + 3D = PSG4D

PSG4D: AI Assistant in 4D world



PSG4D Dataset

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



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

Source: Grand Theft Auto V

67 videos (avg. 84s)

28K RGB-D images

35 object classes, 43 relations

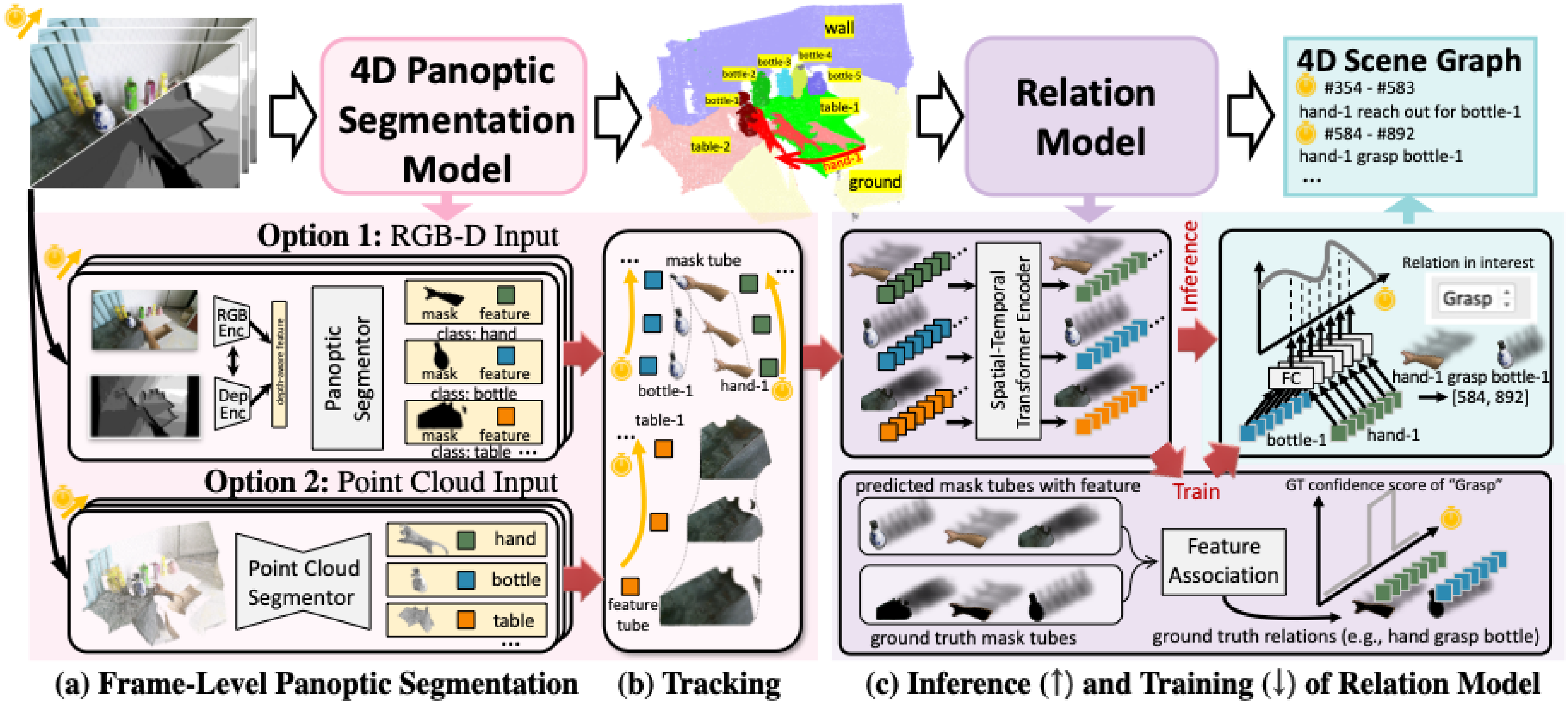
Source: HOI-4D

2973 videos (avg. 20s), Egocentric

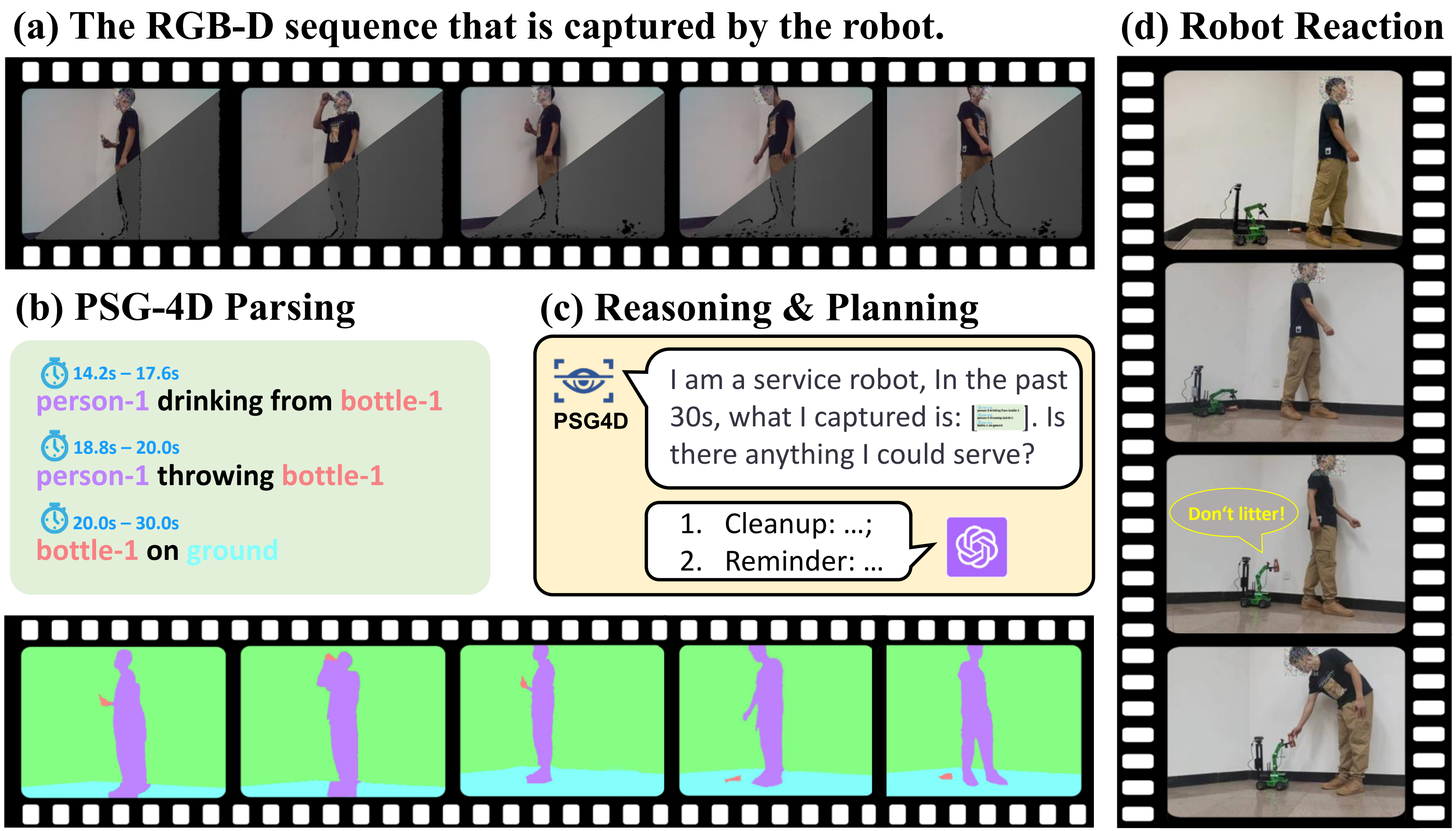
891K RGB-D images

46 object classes, 15 relations

PSG4D Pipeline



PSG4D Real-World Application



+ **Collecting SG with Real World Assistant**

+ **Providing LLM with Eyes (Scene Graph)**

? **Can we directly feed LLM with Images?**

→ **Towards VLM**

AI Assistant with SG

2D Image

2D Video

3D video

[ECCV'22] Panoptic Scene Graph

[CVPR'23] Panoptic **Video** Scene Graph

[NeurIPS'23] **4D** Panoptic Scene Graph

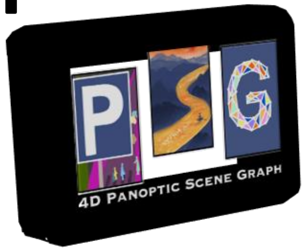
Multimodal input -> Scene Graph -> LLM Prompts

AI Assistant with Scene Graph

Neuro-Symbolic Assistant

Panoptic Scene Graph [ECCV'22]

Panoptic **Video** Scene Graph [CVPR'23]



4D Panoptic Scene Graph [NeurIPS'23]

Jingkang Yang

Bo Li

Yuanhan Zhang

MMLab @NTU

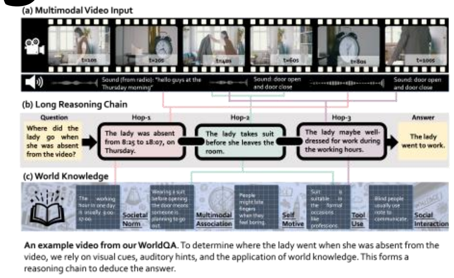
AI Assistant with VLMs

End-to-End Assistant

LLaVA-NeXT

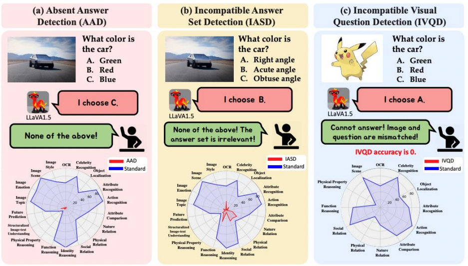
Benchmarking AI Assistants

LMM-Eval



WorldQA

FunQA



Unsolvable Problem Detection

Embodied AI Assistant

Octopus

The pathway: From Language Models to Language Assistant

GPT-2 OpenAI

GPT-3 OpenAI

GPT-3.5 OpenAI

ChatGPT OpenAI

Industrial

Open-source

BERT

LLaMA/T5

T5

Vicuna/Flan-T5

FLAN-T5

Open Assistant

Zero-shot learning

Zero-shot learning
In-context learning

Zero-shot learning
In-context learning
Instruct following

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

CLIP

OpenAI

Flamingo

Google DeepMind

Industrial

Open-source

OpenCLIP

LAION

OpenFlamingo

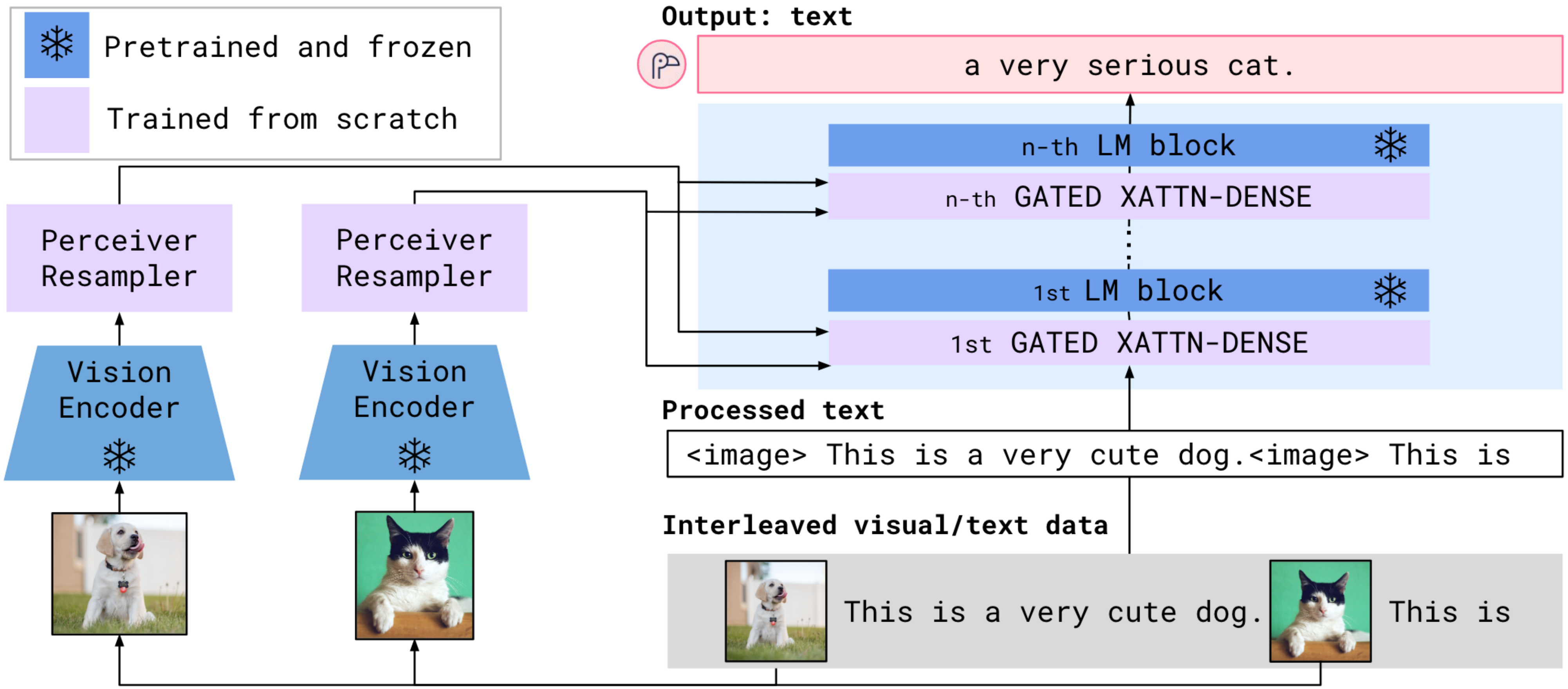
LAION

Otter

Zero-shot learning

Zero-shot learning
In-context learning

Flamingo: a Visual Language Model for Few-Shot Learning



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

Flamingo ≠ Multi-modal Assistants

OpenFlamingo simply completes the next reasonable sentence.

he 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

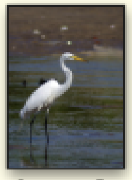
Flamingo ≠ Multi-modal Assistants

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

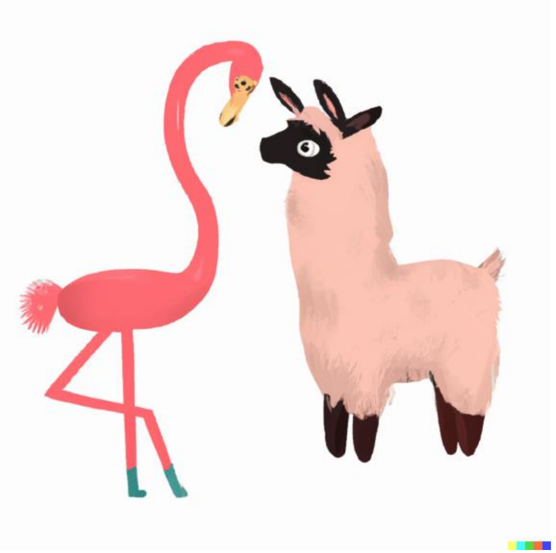
[..., "Check out Shane Driscoll's take on sustainable communities and how his photograph fits this year's Green Cities theme.", ...,  ,"Man-made platforms like the one pictured here allow these fish-eating birds of prey to thrive in developed coastal areas.", "A city surrounded by mountains.", "I took 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

MMC4
(interleaved pretraining)



OpenFlamingo

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

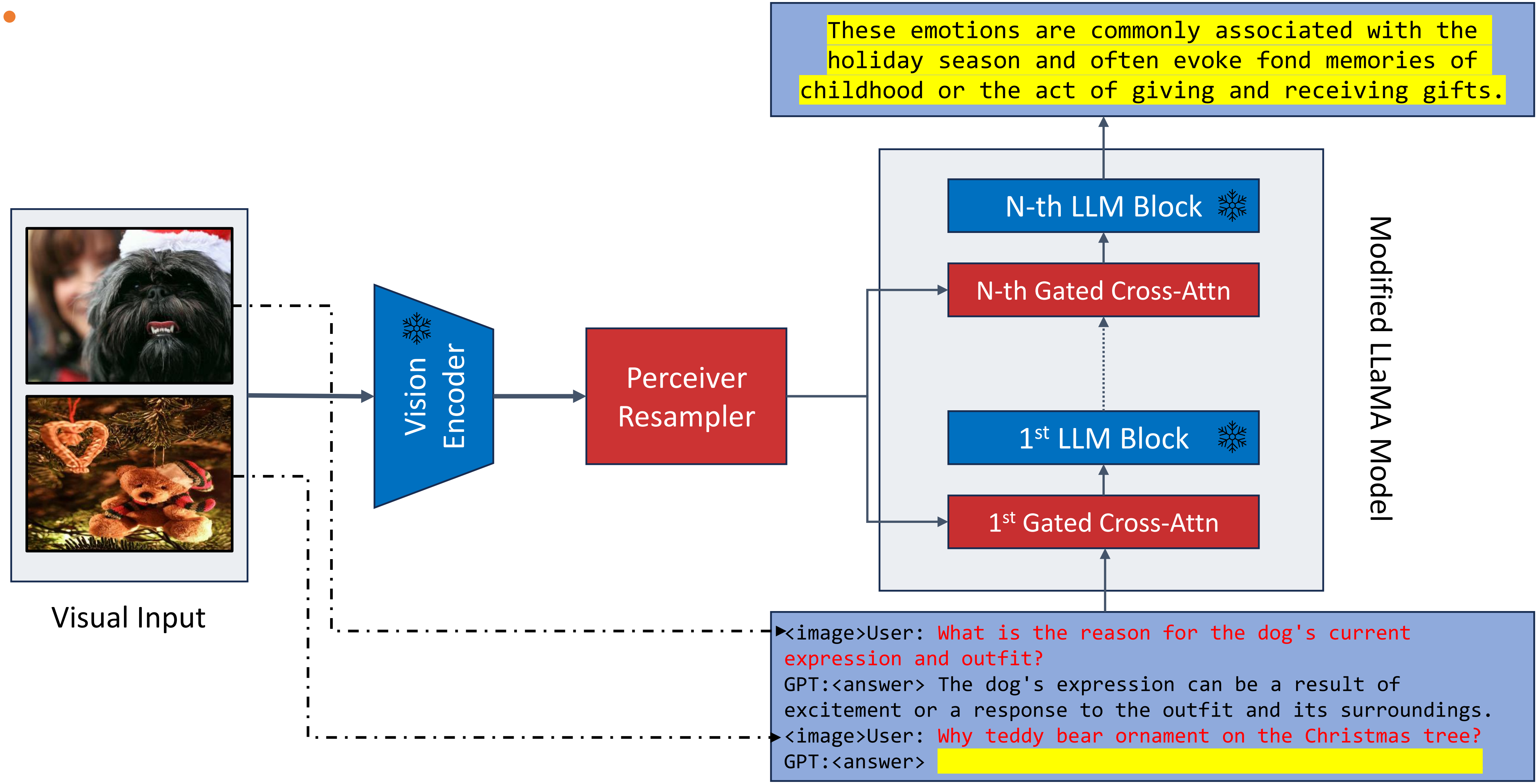
Otter

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.

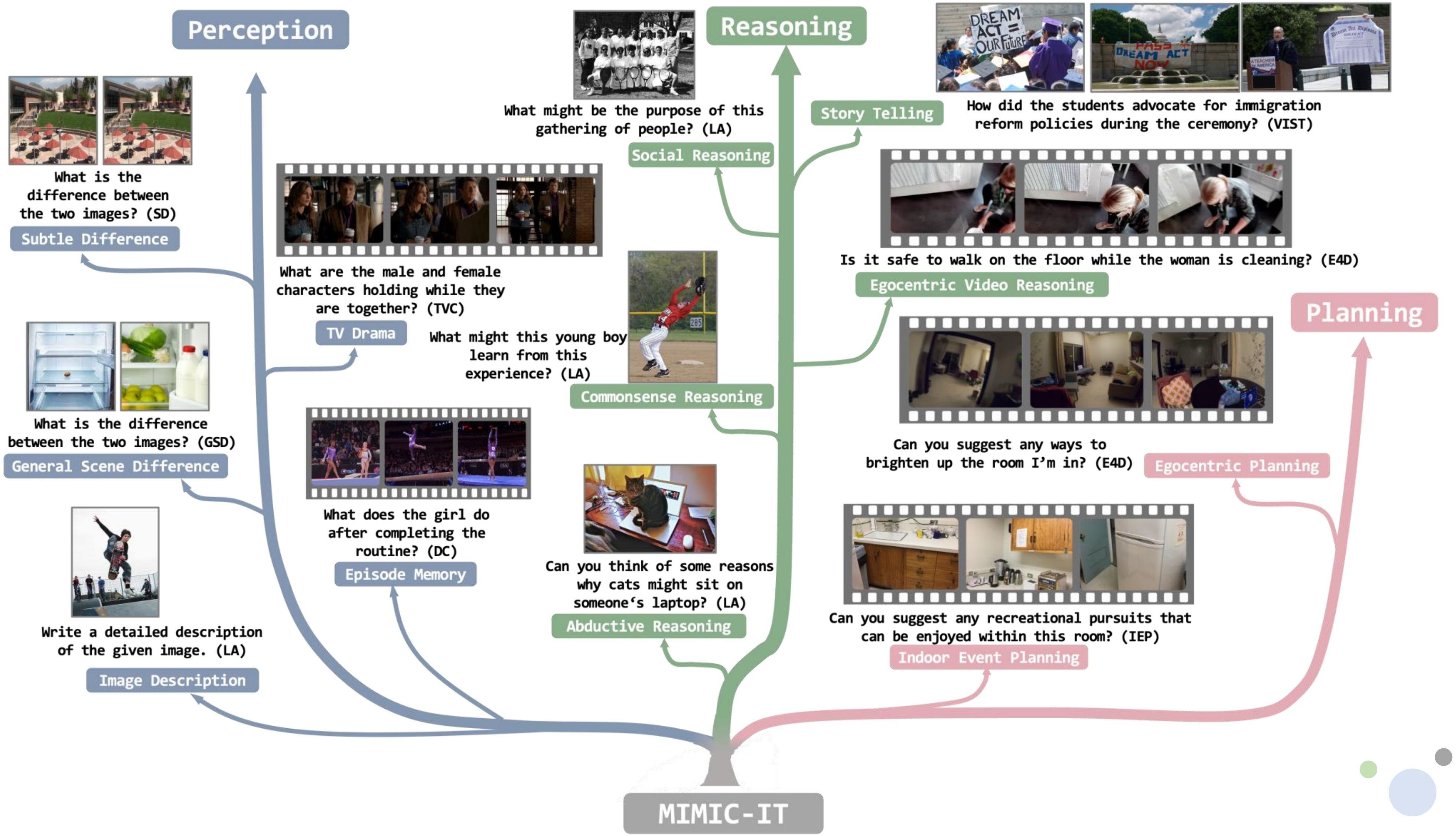
Otter: A Multi-Modal In-context Instruction Tuned Model



MIMIC-IT Dataset

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

S-LAB FOR ADVANCED INTELLIGENCE



MIMIC-IT Dataset

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.

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)

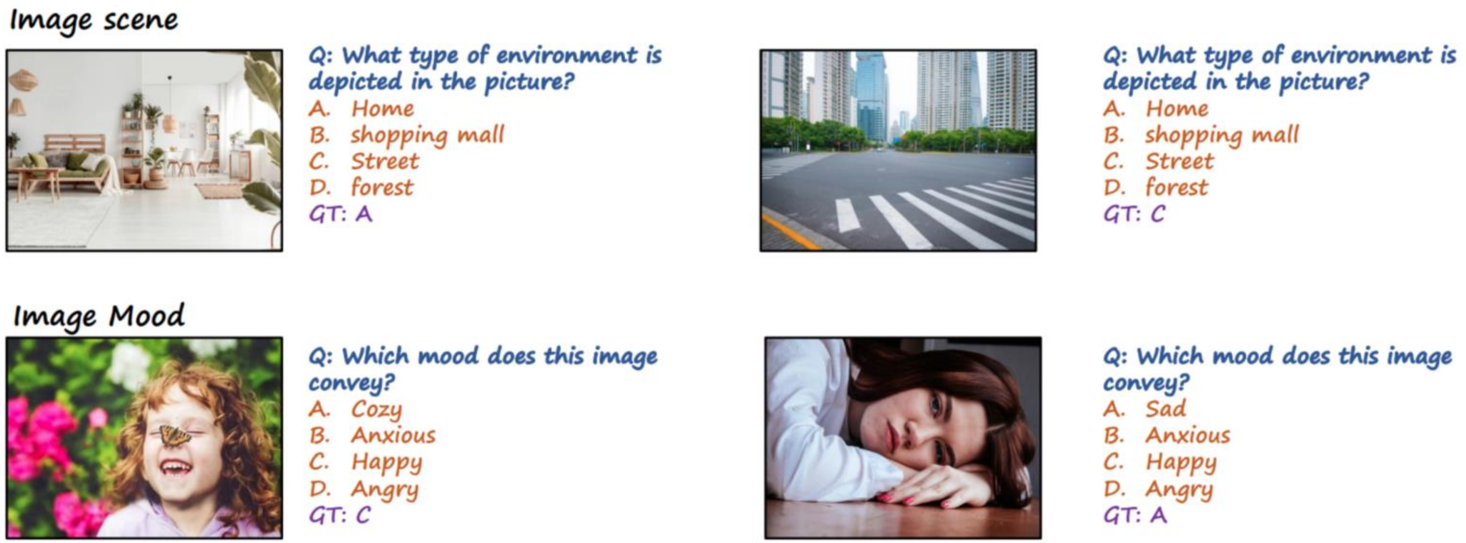
Instruction: Hey Otter, do you think it looks cool?

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

Otter’s Benchmark Performance

MMBench

- Category
  - Perception (1860 Questions)
  - Reasoning (1114 Questions)
- Example Question
- Eval Method
  - Multiple-Choice with GPT-4 as Judge.
  - Circular Evaluation
- Otter’s Performance (Test set)
  - 48.3 – 55.9 %
- Analysis
  - Formatting is important.
  - Padding with in-context examples to enable model’s better formatting.



Liu, Yuan, et al. "MMBench: Is Your Multi-modal Model an All-around Player?." *arXiv preprint arXiv:2307.06281* (2023).

Method	Parameters	Language Model	Vision Model	Overall	LR	AR	RR	FP-S	FP-C	CP
JiuTian	12.3B	FLANT5-XXL	EVA-G	61.8	39.8	76.5	56.8	62.0	51.2	68.7
mPLUG-Owl	7.2B	LLaMA 7B	CLIP ViT-L/14	61.6	36.4	75.8	62.0	64.8	49.2	66.2
LMEye	4.4B	Flan-XL	CLIP ViT-L/14	61.3	36.9	73.0	55.4	60.0	58.0	68.9
Shikra	7.2B	Vicuna 7B	CLIP ViT-L/14	60.2	33.5	69.6	53.1	61.8	50.4	71.7
Otter-I	9B	LLaMA 7B	CLIP ViT-L/14	48.3	22.2	63.3	39.4	46.8	36.4	60.6
G2PT	7B	LLaMA 7B	ViT-G	39.8	14.8	46.7	31.5	41.8	34.4	49.8
LLaMA-Adapter-v2	7.2B	LLaMA 7B	CLIP ViT-L/14	38.9	7.4	45.3	19.2	45.0	32.0	54.0
LLaVA	7.2B	LLaMA 7B	CLIP ViT-L/14	36.2	15.9	53.6	28.6	41.8	20.0	40.4
InstructBLIP	8B	Vicuna 7B	EVA-G	33.9	21.6	47.4	22.5	33.0	24.4	41.1
VisualGLM	8B	ChatGLM 6B	EVA-CLIP	33.5	11.4	48.8	27.7	35.8	17.6	41.5
PandaGPT	14B	Vicuna 13B	ImageBind ViT-H/14	30.6	15.3	41.5	22.0	20.3	20.4	47.9
MiniGPT-4	8B	Vicuna 7B	EVA-G	23.0	13.6	32.9	8.9	28.8	11.2	28.3
MMGPT	9B	LLaMA 7B	CLIP ViT-L/14	16.0	1.1	23.8	20.7	18.3	5.2	18.3
OpenFlamingo v2	9B	MPT 7B	CLIP ViT-L/14	5.7	11.4	12.8	1.4	5.5	0.8	4.0
OpenFlamingo	9B	LLaMA 7B	CLIP ViT-L/14	4.3	9.1	11.4	3.3	2.5	1.6	1.5

OTTER-HD

A High-Res Multi-modality Model

- Multimodal model fine-tuned from Fuyu
- Fine-grained interpretation of high-resolution visual input
- Eliminates the need for a vision encoder
- Supports flexible input sizes at test time
- Adaptable to diverse inference budgets

Perception & Recognition Demonstrations

Query Prompt:
What's inside this painting?

Model Response:
The image depicts a detailed, traditional East Asian scene, possibly of Chinese origin. A large, intricately designed building with multiple tiers and a sloping roof stands prominently in the left-center. The architecture, with its wooden beams, intricate carvings, and elevated structure, is reminiscent of ancient Chinese temples or palaces.

Query Prompt:
How many camels are inside this image?

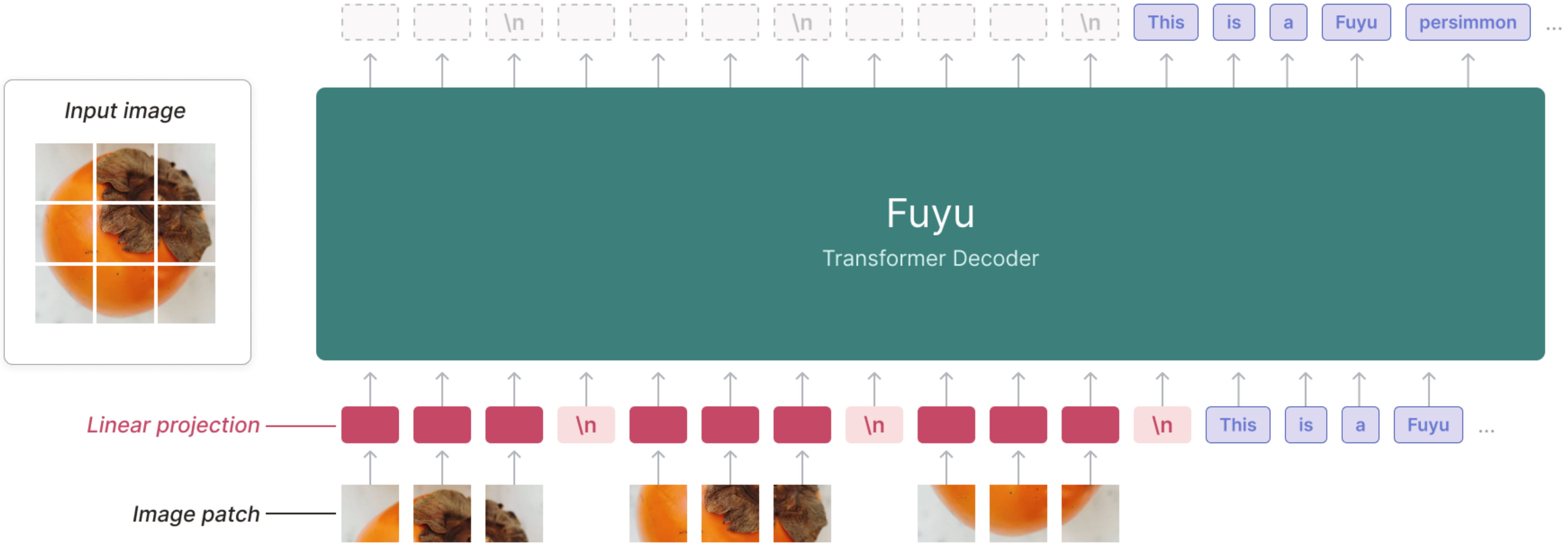
Model Response:
It is hard to examine the camels in such intricate image. However, I find there are three camels visible in the image.

Figure 1: Perception and recognition demonstration of OtterHD-8B. The image is a traditional Chinese painting from the Song Dynasty, *Along the River During the Qingming Festival*. This is a part of the entire artwork, with a resolution of 2466×1766.

Table 1: Performance comparison of OtterHD-8B with prominent open-sourced LMMs, detailing instruction/response data pairs, training, and evaluation resolutions. The term *Dynamic* refers to training with varied resolutions as elaborated in Sec. 4.2. The term *Original* indicates evaluations using each image's resolution without any resizing operation, whereas other models undergo evaluations with images resized to a consistent square resolution at denoted in `Eval Res`. Details on metrics are provided in Sec. 4.1.

Models	I/R Pairs	Train Res.	Eval Res.	MagBench		MME[1]		POPE	MM-V	MMB	M-Vista
				Multi.	FF.	Cog.	Percep.				
Idefics-9B$_{instruct}$ [24]	1M	224	224	20.8	13.4	187.9	1165.0	74.6	23.7	45.5	19.8
Otter-9B [25]	150K	224	224	25.7	15.8	306.4	1292.3	72.5	24.7	48.3	19.7
InstructBLIP-7B [13]	1.2M	224	224	5.6	15.2	-	-	-	26.2	36.0	-
InstructBLIP-13B [13]	1.2M	224	224	3.8	16.3	291.8	1212.8	78.9	25.6	33.9	**25.3**
LLaVA-7B$_{1.5}$ [30]	3.6M[2]	336	336	26.8	24.7	-	**1510.7**	85.9	**30.5**	<u>59.5</u>	-
Qwen-VL-7B$_{chat}$ [4]	1.4B	448	448	14.5	15.9	**360.7**	1487.5	-	-	**61.8**	-
Fuyu-8B [5]	-	-	*Original*	29.3	15.2	237.5	728.6	74.1	21.4	10.7	20.6
OtterHD-8B	370K	512	512	33.5	31.4	289.8	<u>1359.3</u>	**86.1**	25.1	58.5	22.3
		1024	1024	37.8	37.2	288.5	1313.7	81.5	19.8	53.6	17.3
		Dynamic	*Original*	**42.7**	**39.9**	<u>331.4</u>	1223.4	<u>86.0</u>	<u>26.3</u>	58.3	<u>23.5</u>

OtterHD



- Improved with FlashAttention-2 and other fused operators
  - Fused layernorm
  - Fused square ReLU
  - Fused rotary positional embedding
- Optimizing Fuyu-8B's HuggingFace implementation

MagnifierBench

Small Object Localization

Q: Where is the laptop?
A. On the green sofa
B. On the square dining table
C. On the grey round stool
D. On the brown round table

Correct Answer: C
GPT-4V Answer: A

Small Object Counting

Q: How many training cones are on the grass?
A. 3
B. 4
C. 5
D. 6

Correct Answer: C
GPT-4V Answer: A

Small Color Identifying

Q: What color is the lid of the small bottle closest to the green lid of the seasoning jar?
A. Green,
B. Yellow,
C. White,
D. Purple

Correct Answer: D
GPT-4V Answer: C

OtterHD

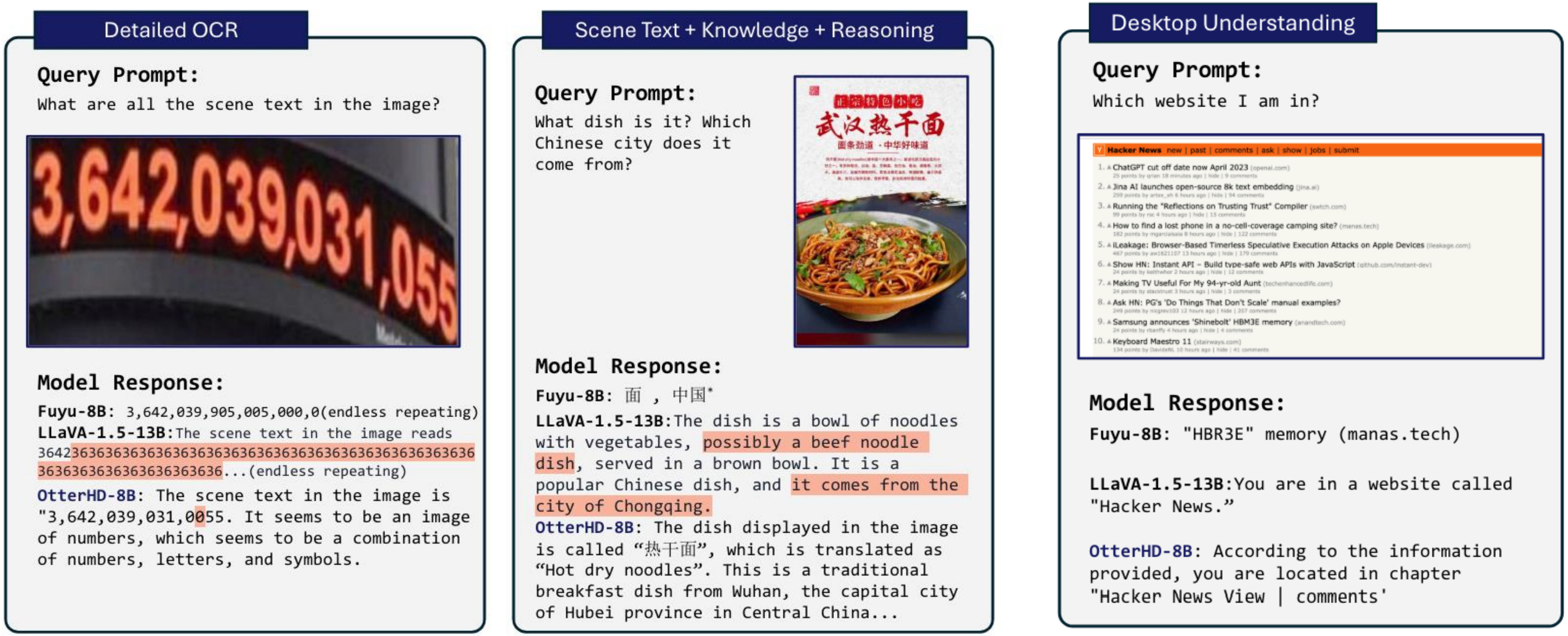
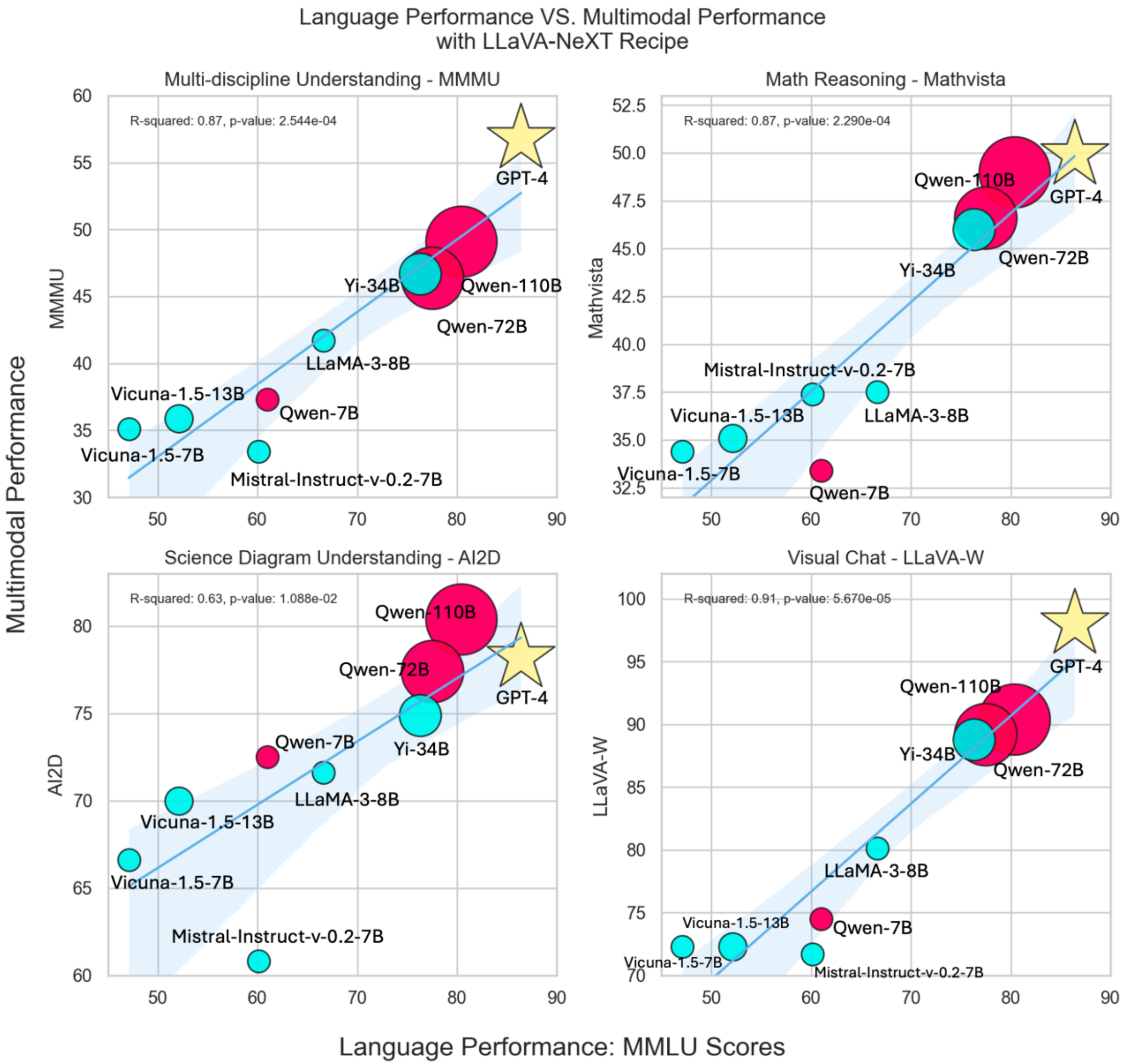


Figure 6: Desktop understanding abilities comparison between LMMs. Incorrect parts are noted.

LLaVA-NeXT



Stronger LLMs Supercharge Multimodal Capabilities in the Wild

- **We expand current LLaVA-NeXT to 3x model size, up to 72B and 110B model.**
- **We found that, under the same training strategy, stronger LLM naturally brings stronger multimodal performance.**

LLaVA-NeXT

Benchmark Results

Results with LMMs-Eval				GPT4-V	LLaVA-NeXT (2024-05 Release)		
Datasets	Split	Metric	Instances		Qwen1.5-110B	Qwen1.5-72B	LLaMA3-8B
AI2D*	test	Acc.	3088	78.2	**80.4**	77.4	71.6
ChartQA*	test	RelaxedAcc.	2500	78.5	**79.7**	77.0	69.5
DocVQA*	val	ANLS	5349	-	**85.7**	84.4	78.2
MathVista	test	Acc.	1000	**49.9**	49.0	46.6	37.5
MMBench	dev	Acc.	4377	75.0	**80.5**	**80.5**	72.1
MME-Cognition	test	Total Score	2374	**517.1**	453.9	459.6	367.8
MME-Perception	test	Total Score	2374	1409.4	**1746.5**	1699.3	1603.7
MMMU	val	Acc.	900	**56.8**	50.1	49.9	41.7
RealWorldQA	test	Acc.	765	61.4	63.1	**65.4**	60.0
LLaVA-W**	test	GPT4-Eval	60	**98.0**	90.4	89.2	80.1
LLaVA-Bench (Wilder)	Small	GPT4V-Eval	120	**71.5**	70.5	71.2	62.5
	Medium	GPT4V-Eval	1020	**78.5**	72.5	73.4	63.1

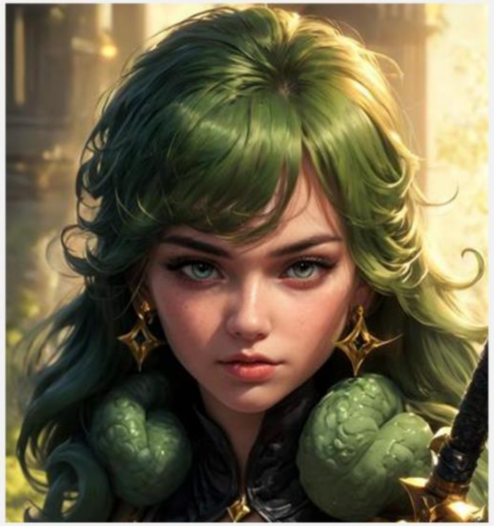
**Train split observed during SFT stage.*

***We report the evaluation results with GPT-4-0613 on LLaVA-W.*

- **LLaVA-NeXT reaches to GPT4-V level multimodal performance.**
- **Our largest model, LLaVA-NeXT-110B model only trained on 128-H800 for 18 hours.**

LLaVA-NeXT

Daily-life: Prompt Interpretation (the images in the right are generated via DALLE-3)

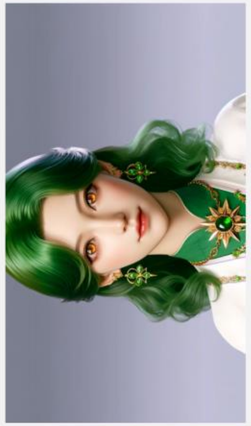


User Instruction: write an image prompt for this character without details about the surrounding. Include color and details about all of these variables: age, eyes, hair, skin, expression, clothes

LLaVA-NeXT-110B: This character exudes a youthful and vibrant energy, with a radiant complexion that suggests a youthful age. The eyes are a striking shade of blue, full of life and curiosity. The hair is a lush green, cascading in soft waves that hint at a playful and spirited personality. The skin is smooth and fair, reflecting a healthy and energetic lifestyle. The expression is one of determination and confidence, with a slight hint of mischief in the eyes. The clothes are a mix of dark and light tones, with a gold and green color scheme that suggests a regal or mystical status. The character is adorned with intricate jewelry and accessories, adding to the overall impression of someone of high importance or magical ability.

Qwen-VL-Max: This is a portrait of a young woman with vibrant green hair and matching eyes. Her skin is fair, and she has a serious expression on her face. She is dressed in a white and black outfit, adorned with a green gemstone necklace.

GPT4-V: Create an image of a young adult female character with striking features. She should appear to be in her early twenties. Her eyes are a captivating amber hue, with a sharp gaze that conveys a sense of determination and curiosity. Her hair is a lush, vibrant green with soft waves cascading around her shoulders, giving her an otherworldly charm.\n\nHer skin is fair with a smooth, porcelain-like quality, and her expression is confident yet approachable, with a subtle hint of a smile playing on her lips. She is adorned in unique attire that includes a white top with a plunging neckline, accented with emerald green and gold jewelry that complements her hair. The jewelry is bold and fantastical, featuring large, ornate earrings shaped like stars and a prominent, jewel-encrusted brooch at her collar.



Intern-VL-1.5: image prompt:title: \"the enigmatic maiden\"description: a digital illustration of a young woman with striking features. her age appears to be in her late teens to early twenties. she has vibrant green eyes that seem to hold a hint of mystery, and her skin is fair with a soft, natural glow. her hair is a rich, emerald green, styled in loose, wavy curls that cascade around her face and shoulders. she has a serene yet intense expression, with her lips slightly parted as if she is about to speak or is lost in thought.she is adorned in a fantasy-inspired outfit that combines elements of elegance and combat readiness. the top is a white, flowing garment with a plunging neckline, revealing a hint of her d\u00e9colletage. over this, she wears a black leather corset that accentuates her waist and adds a touch of edginess to her appearance. the corset is embellished with gold accents and a large, green gemstone at the center, which matches the color of her hair and the green holly-like pendant she wears around her neck.her accessories include large, gold earrings with a star-like design that complement her overall look. in her right hand, she holds a sword with a golden hilt and a green gemstone embedded in the pommel, suggesting that she is a warrior or a guardian of some sort. the sword's design is intricate, with gold and green accents that echo the colors found in her hair, eyes, and jewelry.the background is blurred, drawing focus to the character, but it appears to be a natural setting with hints of sunlight filtering through, possibly a forest or a garden, which adds to the mystical and otherworldly atmosphere of the image.

LLaVA-NeXT

LLaVA-Wilder Qualitative Example

User Instruction: 请针对于这幅画写一首中文古诗。

GPT4V: 好的，这幅画中的水獭看起来非常有趣和可爱，我尝试写一首中文古诗来描述它们：

湖光倒影水中映，
双獭共读故纸情。
萍水相逢最是缘，
浮生若梦伴诗行。

这首诗试图捕捉画面中水獭共享阅读的温馨场景，以及它们在自然环境中的和谐生活。希望您喜欢这首诗！

LLaVA-NeXT-110B:
好的，下面是我创作的中文古诗：
《两河獭，共读书》
水波荡漾，字字珠玑。
河清海晏，岁月静好，
读书河獭，传颂千古。

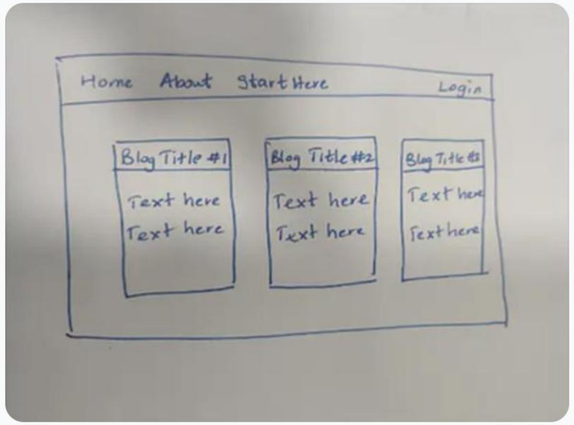
LLaVA-NeXT-72B:
两海獭水中读，
手捧书卷乐融融。
水波荡漾映日光，
知识海洋任遨游。

Qwen-VL-Max: 两只水獭在水中戏笑，手中拿着一本书。
它们似乎在协商，眼神专注而认真。
水波荡漾着它们的影子，伴随它们自在。
这是一幅多么美妙的画面，让人感到无比的温馨和平静。

Intern-VL-1.5: 好的，下面是我创作的中文古诗：
水边双影坐，
书卷共读时。
风起涟漪动，
心静意自怡。

LLaVA-NeXT

Chatbot



Here's a design for blogging website. Provide the working source code for the website using HTML, CSS and JavaScript as required.

The image you've provided is a hand-drawn sketch of a website layout, which includes a navigation bar with links to "Home," "About," "Start Here," and "Login," as well as three blog post sections with titles and text content.

To create a website based on this design, you would typically use HTML for the structure, CSS for the styling, and JavaScript for any interactive elements. Here's a simplified example of how you might code this in HTML and CSS:

```
<!DOCTYPE html>
<html lang="en">
<head>
<meta charset="UTF-8">
<meta name="viewport" content="width=device-width, initial-scale=1.0">
<title>Blog Website</title>
<style>
  body {
    font-family: Arial, sans-serif;
```

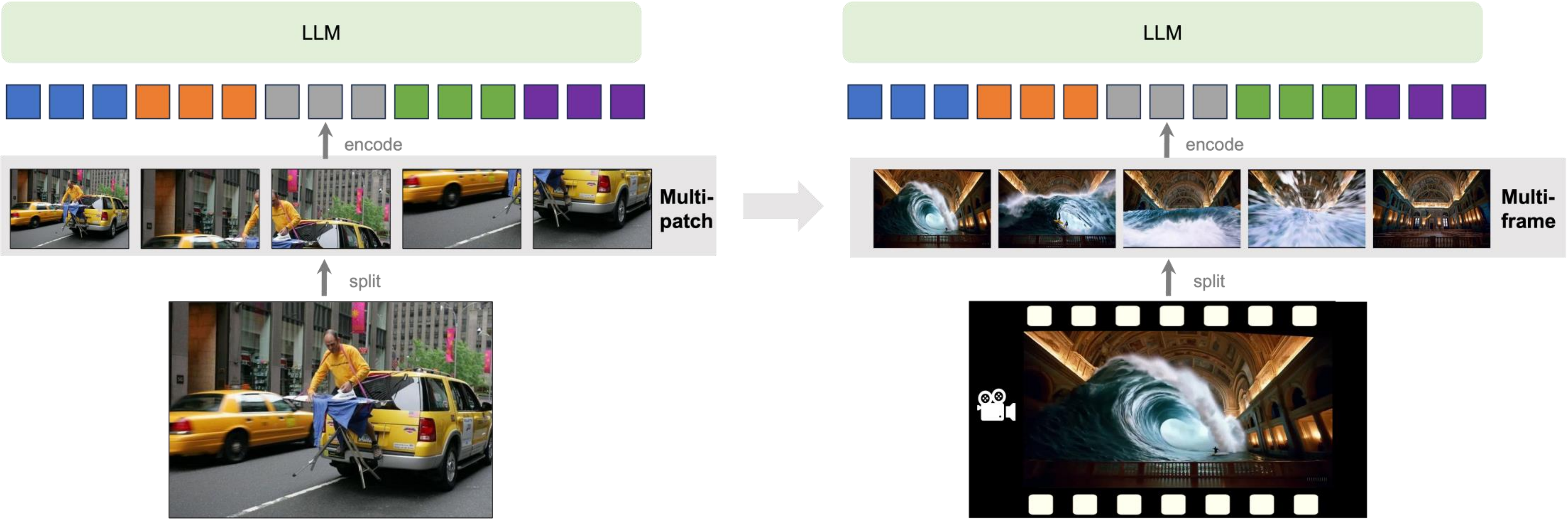
Enter message or upload file...

None | Up-Vote | Down-Vote | Regenerate | Send

Please visit our demos to try more
https://llava-next.lmms-lab.com/

LLaVA-NeXT Video

Takeaway message 1: A unified view of images and videos is important for leveraging high-quality image data in video understanding.



LLaVA-NeXT Video digests a set of image patches/frames as a sequence of concatenated visual tokens, allowing unified view of image and video input.

LLaVA-NeXT Video

Takeaway message 2: Direct Preference Optimization (DPO) from AI Feedback Direct Preference Optimization from AI Feedback allows an LMM to reduce hallucination and follow the instructions better.

User: Please provide a detailed description of the video, focusing on the main subjects, their actions, the background scenes, and the temporal transitions.

LLaVA-NeXT-Video-DPO

The video depicts a dynamic and surreal scene where the action of surfing is juxtaposed with an indoor setting. In the foreground, there are two surfers riding waves, one slightly ahead of the other, maintaining their balance on surfboards. The surfer in the foreground is skillfully navigating the crest of a wave, while the one behind is positioned on the face of the wave, both displaying athletic postures typical of surfing. The waves are large and crashing, with the wave in the background creating a dramatic arc that suggests motion and power.

The background is an opulent interior space with high vaulted ceilings, ornate detailing, and classical architecture, including arches and what appears to be marble or stone work. The lighting is warm, with a golden hue that contrasts with the cool tones of the water. The architecture suggests a grand hall or a space designed for public gatherings or performances, possibly inspired by classical or neoclassical styles.

The temporal transition in the video is interesting as it captures a moment of action (the surfers riding the wave) within a stationary, non-moving environment (the interior space). This creates a surreal effect, blending the dynamic motion of the ocean with the static elegance of a building interior. The image is a blend of two disparate subjects—the exhilarating sport of surfing and the serene grandeur of architectural design—into a single, captivating composition.

LLaVA-NeXT-Video

The video depicts a large wave crashing inside a building, with a surfer riding the wave. The surfer is positioned in the center of the frame, skillfully navigating the wave's crest. The wave is massive, filling the majority of the space from the top to the bottom of the frame, with its white foam and water spray creating a dynamic and powerful visual effect. The background of the scene is a grand hall with high ceilings and ornate architecture, featuring arches and columns that add a sense of grandeur to the setting. The lighting is dim, with the focus on the surfer and the wave, casting shadows and highlighting the contours of the wave's movement. The camera captures the surfer from a low angle, emphasizing the height of the wave and the surfer's skill in maintaining balance. As the wave begins to break, the surfer is seen riding it towards the right side of the frame, creating a sense of motion and direction. The video transitions from a close-up of the surfer to a wider shot, showing the full scale of the wave and the grand hall. The sound of the crashing wave is prominent, adding to the immersive experience. The video concludes with a shot of the surfer disappearing from view as the wave subsides, leaving the grand hall empty and still.

We illustrate two examples to demonstrate the superiority of DPO. Texts of interest are highlighted in blue, while parts that might contain hallucinations are marked in red

LLaVA-NeXT Video

Takeaway message 3: Our LLaVA-NeXT-Video 34B model achieves SoTA performance on the recently proposed, most comprehensive diagnosis benchmark: Video-MME.

Models	LLM Params	Short (%)		Medium (%)		Long (%)		Overall (%)	
		w/o subs	w/ subs	w/o subs	w/ subs	w/o subs	w/ subs	w/o subs	w/ subs
Open & Closed-source Image MLLMs									
Qwen-VL-Chat [5]	7B	46.4	47.1	38.1	39.8	38.0	38.3	40.9	41.7
Qwen-VL-Max [5]	-	56.5	58.3	**49.9**	49.8	**49.0**	46.9	**51.8**	51.7
InternVL-Chat-V1.5 [9]	20B	**61.2**	**62.4**	47.3	**50.0**	46.0	**47.0**	51.5	**53.2**
Open-source Video MLLMs									
Video-LLaVA [28]	7B	45.9	47.1	38.1	40.2	37.3	39.6	40.4	42.3
VideoChat2 [24]	7B	38.2	41.6	33.2	34.3	29.7	31.9	33.7	35.9
ST-LLM [33]	7B	47.0	49.9	36.9	42.2	31.8	37.3	38.6	43.2
Chat-UniVi-V1.5 [19]	7B	46.3	51.4	40.3	45.2	36.9	42.3	41.2	46.3
LLaVA-NeXT-Video [68]	34B	**63.1**	**66.4**	**51.1**	**53.2**	**44.6**	**48.7**	**52.5**	**56.0**
Closed-source MLLMs									
GPT-4V [45]	-	71.4	74.5	56.5	59.3	54.2	57.2	60.7	63.7
GPT-4o [46]	-	77.1	77.5	62.1	63.0	59.2	56.7	66.2	65.8
Gemini 1.5 Pro [51]	-	**82.3**	**84.7**	**75.3**	**82.6**	**67.5**	**76.3**	**75.7**	**81.6**

Info for LLaVA-NeXT Video:
1. Technical report: https://llava-vl.github.io/blog/2024-04-30-llava-next-video/
2. Model and code: https://github.com/LLaVA-VL/LLaVA-NeXT?tab=readme-ov-file

AI Assistant with Scene Graph

Neuro-Symbolic Assistant

PANOPTIC SCENE GRAPH

Panoptic Scene Graph [ECCV'22]

PANOPTIC VIDEO SCENE GRAPH

Panoptic **Video** Scene Graph [CVPR'23]

4D PANOPTIC SCENE GRAPH

4D Panoptic Scene Graph [NeurIPS'23]

Jingkang Yang

Bo Li

Yuanhan Zhang

MMLab @NTU

AI Assistant with VLMs

End-to-End Assistant

OTTER

Deep diving into large Multi-modality Models

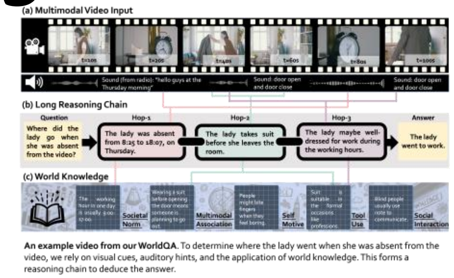
OTTER-HD

A High-Res Multi-modality Model

LLaVA-NeXT

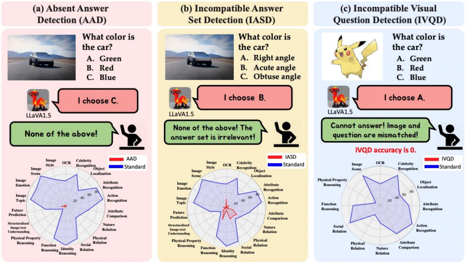
Benchmarking AI Assistants

LMM-Eval



WorldQA

FunQA



Unsolvable Problem Detection

Embodied AI Assistant

Octopus

LMMs-Eval

Holistic Evaluation of LMMs

1. Standardized process of evaluation for **(easily)** reproducible results.
2. Holistic evaluation to reflect model performance across 20-30 benchmarks.

Integration to LMMs Development Life Cycle

1. Efficient evaluation during development (100+ of production checkpoints)
2. Easy (<10 minutes) integration of new (updated) datasets/models.
3. Analysis tools for developers to aware of drawbacks of current models.

Comprehensive Benchmarks Evaluation Guides AI

Holistic evaluation is necessary

- *More evaluations from different dimensions could better reflect model's overall performance.*

Method	FT	Shot	OKVQA (I)	VQAv2 (I)	COCO (I)	MSVDQA (V)	VATEX (V)	VizWiz (I)	Flick30K (I)	MSRVTTQA (V)	iVQA (V)	YouCook2 (V)	STAR (V)	VisDial (I)	TextVQA (I)	NextQA (I)	HatefulMemes (I)	RareAct (V)
Zero/Few shot SOTA	✗	(X)	[34] 43.3 (16)	[114] 38.2 (4)	[124] 32.2 (0)	[58] 35.2 (0)	-	-	-	[58] 19.2 (0)	[135] 12.2 (0)	-	[143] 39.4 (0)	[79] 11.6 (0)	-	-	[85] 66.1 (0)	[85] 40.7 (0)
Flamingo-3B	✗	0	41.2	49.2	73.0	27.5	40.1	28.9	60.6	11.0	32.7	55.8	39.6	46.1	30.1	21.3	53.7	58.4
	✗	4	43.3	53.2	85.0	33.0	50.0	34.0	72.0	14.9	35.7	64.6	41.3	47.3	32.7	22.4	53.6	-
	✗	32	45.9	57.1	99.0	42.6	59.2	45.5	71.2	25.6	37.7	76.7	41.6	47.3	30.6	26.1	56.3	-
Flamingo-9B	✗	0	44.7	51.8	79.4	30.2	39.5	28.8	61.5	13.7	35.2	55.0	41.8	48.0	31.8	23.0	57.0	57.9
	✗	4	49.3	56.3	93.1	36.2	51.7	34.9	72.6	18.2	37.7	70.8	**42.8**	50.4	33.6	24.7	62.7	-
	✗	32	51.0	60.4	106.3	47.2	57.4	44.0	72.8	29.4	40.7	77.3	41.2	50.4	32.6	28.4	63.5	-
Flamingo	✗	0	50.6	56.3	84.3	35.6	46.7	31.6	67.2	17.4	40.7	60.1	39.7	52.0	35.0	26.7	46.4	**60.8**
	✗	4	57.4	63.1	103.2	41.7	56.0	39.6	75.1	23.9	44.1	74.5	42.4	**55.6**	36.5	30.8	68.6	-
	✗	32	**57.8**	**67.6**	**113.8**	**52.3**	**65.1**	**49.8**	**75.4**	**31.0**	**45.3**	**86.8**	42.2	**55.6**	**37.9**	**33.5**	**70.0**	-
Pretrained FT SOTA	✔	(X)	54.4 [34] (10K)	80.2 [140] (444K)	143.3 [124] (500K)	47.9 [28] (27K)	76.3 [153] (500K)	57.2 [65] (20K)	67.4 [150] (30K)	46.8 [51] (130K)	35.4 [135] (6K)	138.7 [132] (10K)	36.7 [128] (46K)	75.2 [79] (123K)	54.7 [137] (20K)	25.2 [129] (38K)	79.1 [62] (9K)	-

Table 1: **Comparison to the state of the art.** A *single* Flamingo model reaches the state of the art on a wide array of image **(I)** and video **(V)** understanding tasks with few-shot learning, significantly outperforming previous best zero- and few-shot methods with as few as four examples. More importantly, using only 32 examples and without adapting any model weights, Flamingo *outperforms* the current best methods – fine-tuned on thousands of annotated examples – on seven tasks. Best few-shot numbers are in **bold**, best numbers overall are underlined.

Flamingo model was (at 2022) a state-of-the-art multimodal model on multiple datasets across image and video modalities.

Comprehensive Benchmarks Evaluation Guides AI

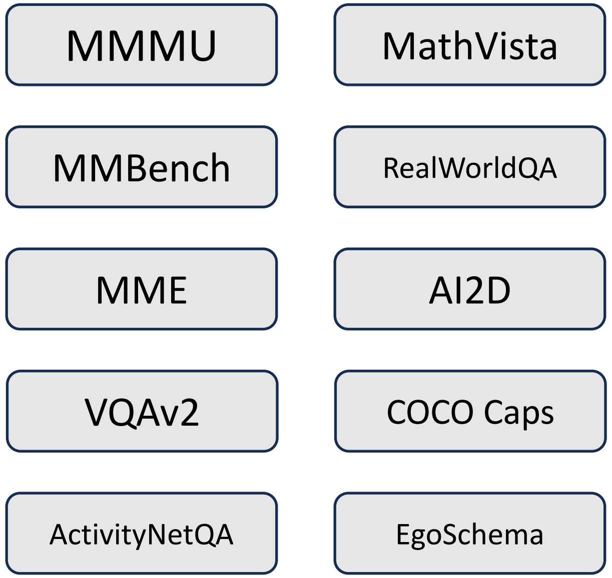
Standardized and reproducible evaluation is necessary

MMMU	MathVista
MMBench	RealWorldQA
MME	AI2D
VQAv2	COCO Caps
ActivityNetQA	EgoSchema

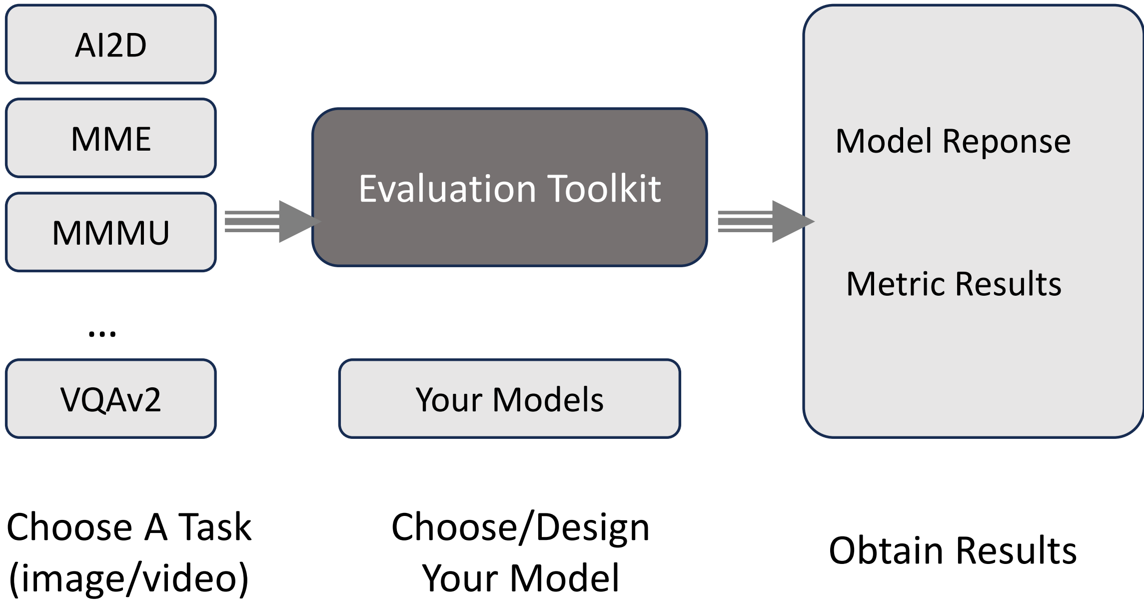
Previously, developers collect and evaluate on each task.
Non-standardized tasks make reproducibility and comparison with prior work expensive (and sometimes impossible)

Comprehensive Benchmarks Evaluation Guides AI

Benchmarks emerge fast. Everyday, we have new benchmarks.

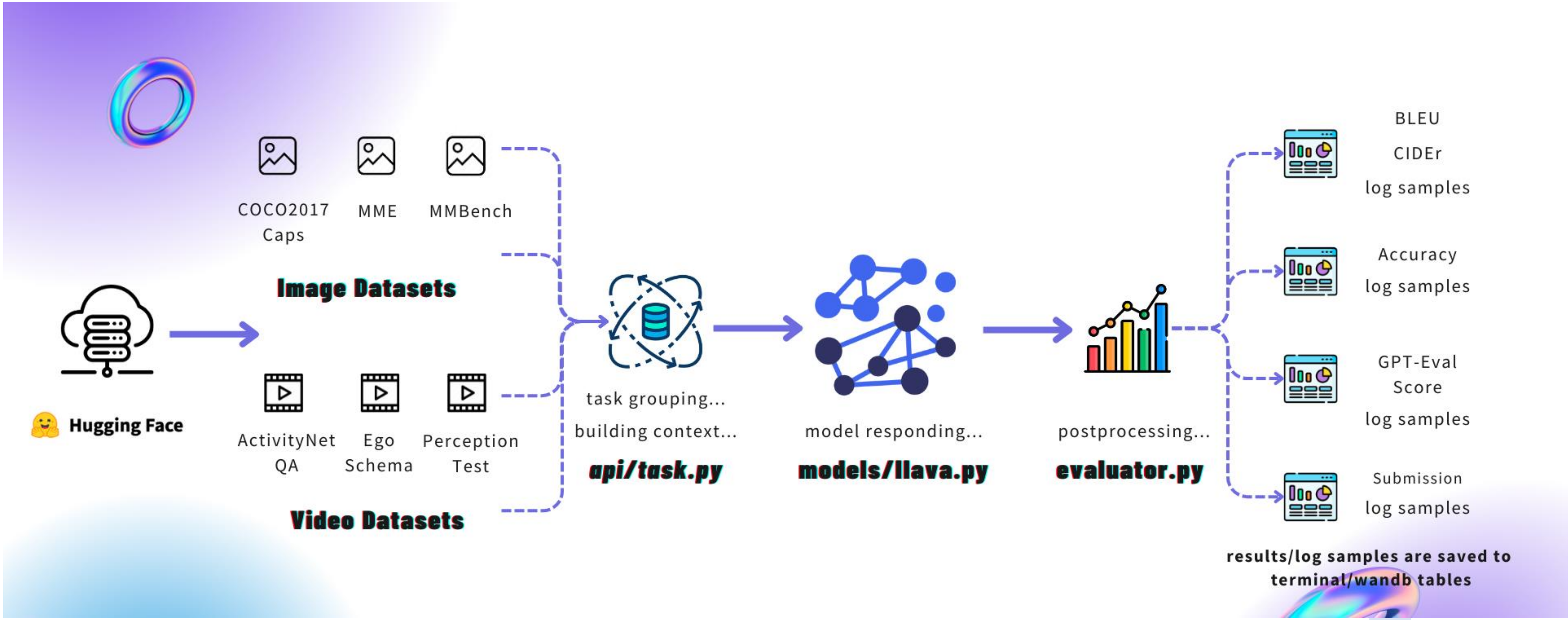


Previously, developers collect and evaluate on each task.
Non-standardized tasks make reproducibility and comparison with prior work expensive
(and sometimes impossible)



We wish to have a unified evaluation toolkit that defines the interface for multimodal tasks and models.

LMMs-Eval is an **efficient, standardized and reproducible** evaluation framework for accelerating the development for large-scale multimodal models



LMMs-Eval: Unified interface for multimodal tasks and models.

- activitynetqa
- ai2d
- chartqa
- cmmmu
- coco_cap
- conbench
- cvrr
- docvqa
- egoschema
- ferret
- flickr30k
- gqa
- hallusion_bench
- iconqa
- ii_bench
- infovqa
- llava-bench-coco
- llava-in-the-wild
- llava_wilder
- longvideobench
- mathverse
- mathvista
- mmbench
- mme
- mmmu
- mmupd
- mmvet
- multidocvqa
- multilingual-llava-bench-in-th
- nextqa
- nextqa
- nocaps
- ocrbench
- ok_vqa
- olympiadbench
- perceptiontest
- pope
- qbench
- realworldqa
- refcoco+
- refcoco
- refcocog
- scienceqa
- screenspot
- seedbench
- stvqa
- synthdog
- tempcompass
- textcaps
- textvqa
- vatex
- vcr_wiki
- video_detail_description
- videochatgpt
- videomme
- vizwiz_vqa
- vqav2
- websrc
- worldqa
- youcook2

(a) Tasks

- batch_gpt4.py
- claude.py
- from_log.py
- fuyu.py
- gemini_api.py
- gpt4v.py
- idefics2.py
- instructblip.py
- internvl.py
- llama_vid.py
- llava.py
- llava_hf.py
- minicpm_v.py
- mplug_owl_video.py
- phi3v.py
- qwen_vl.py
- qwen_vl_api.py
- reka.py
- video_chatgpt.py
- video_llava.py
- xcomposer2_4KHD.py

(b) Models

LMMs-Eval integrates **50+ image tasks, 10+ video tasks** and **~10 state-of-the-art level LMM models.**

FunQA: Surprising Video Understanding

FunQA

HumorQA Demo

FunQA
4.3K Video Clips
312K Free-Text QA Pairs
34.2 average answer word length

HumorQA
1,769 Video Clips
142.1K Free-Text QA Pairs
28.2 average answer word length

CreativeQA
927 Video Clips
78.7K Free-Text QA Pairs
59.1 average answer word length

MagicQA
1,672 Video Clips
92.1K Free-Text QA Pairs
22.7 average answer word length

In what period the magic took place?
The magic moment is from frame 71 to 119.

Can you describe the magical segment of this video?
The man reached out his hand, took one of the clouds from the sky, and put it into his mouth.

How can you tell that the video showcases magic?
What's commonsense-violating is that, people standing on the ground can't touch the clouds, so the clouds can't be taken down by the man.

What method do you think the magician used to accomplish the magic trick?
The magician used video editing technology to combine a video of reaching for the cloud with a video of eating a marshmallow to create the magic effect.

FunQA
4.3K Video Clips
312K Free-Text QA Pairs
34.2 average answer word length

HumorQA
1,769 Video Clips
142.1K Free-Text QA Pairs
28.2 average answer word length

CreativeQA
927 Video Clips
78.7K Free-Text QA Pairs
59.1 average answer word length

MagicQA
1,672 Video Clips
92.1K Free-Text QA Pairs
22.7 average answer word length

In what period the creative took place?
The creative moment is from frame 71 to 1800.

Can you describe the content of this video?
A group of swimmers appears to jump into the water for a swimming competition. However, these swimmers, rather than jump into the water, run to a board in front of them. In the board there are five lanes of water looks like swimming lanes. Each lane has a floating swimmer toy to represent the swimmer. The swimmers need to drink the water. The faster they drink, the faster the floating toy approaching to the finish line.

Can you explain why the video is creative?
...

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)

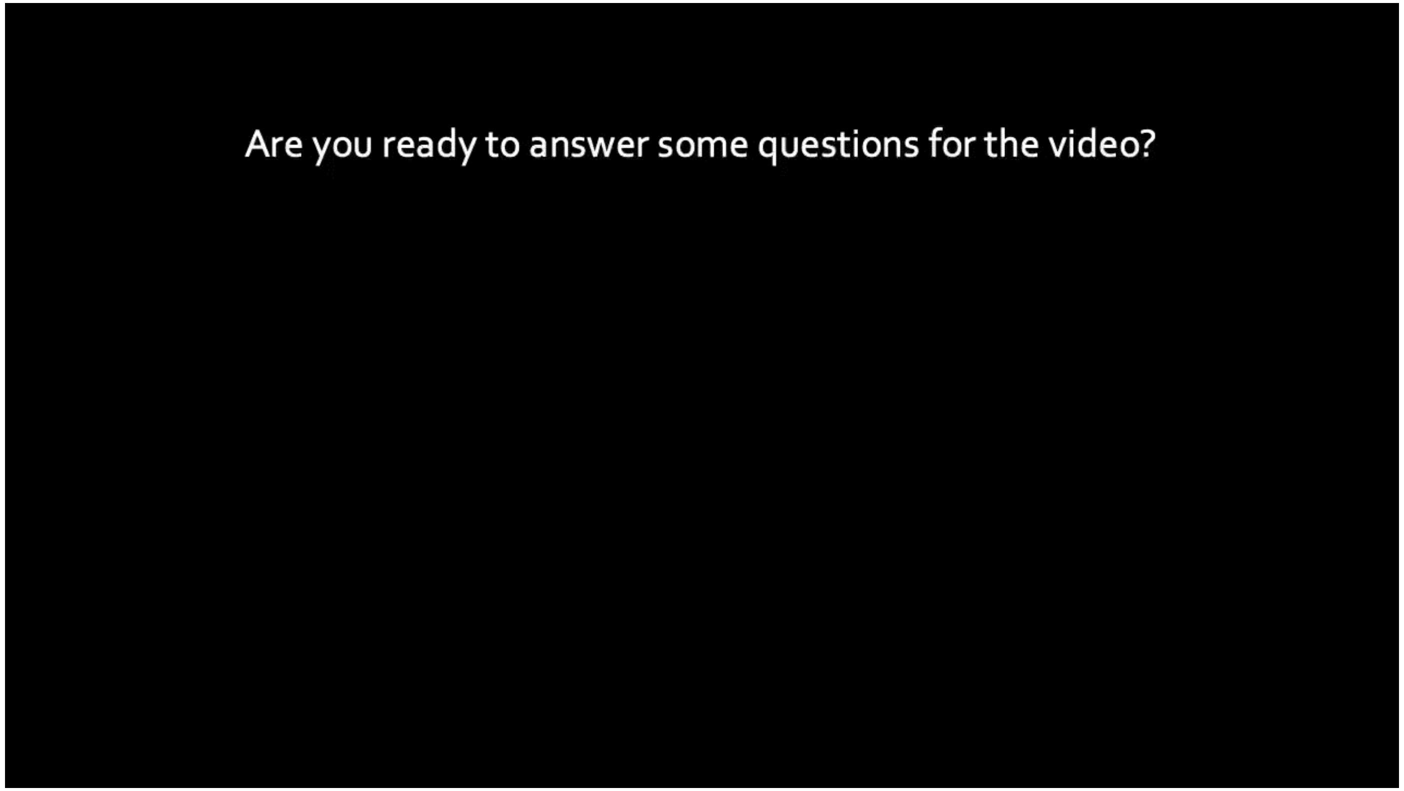
Otter (trained on FunQA)

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.

WorldQA



WorldQA

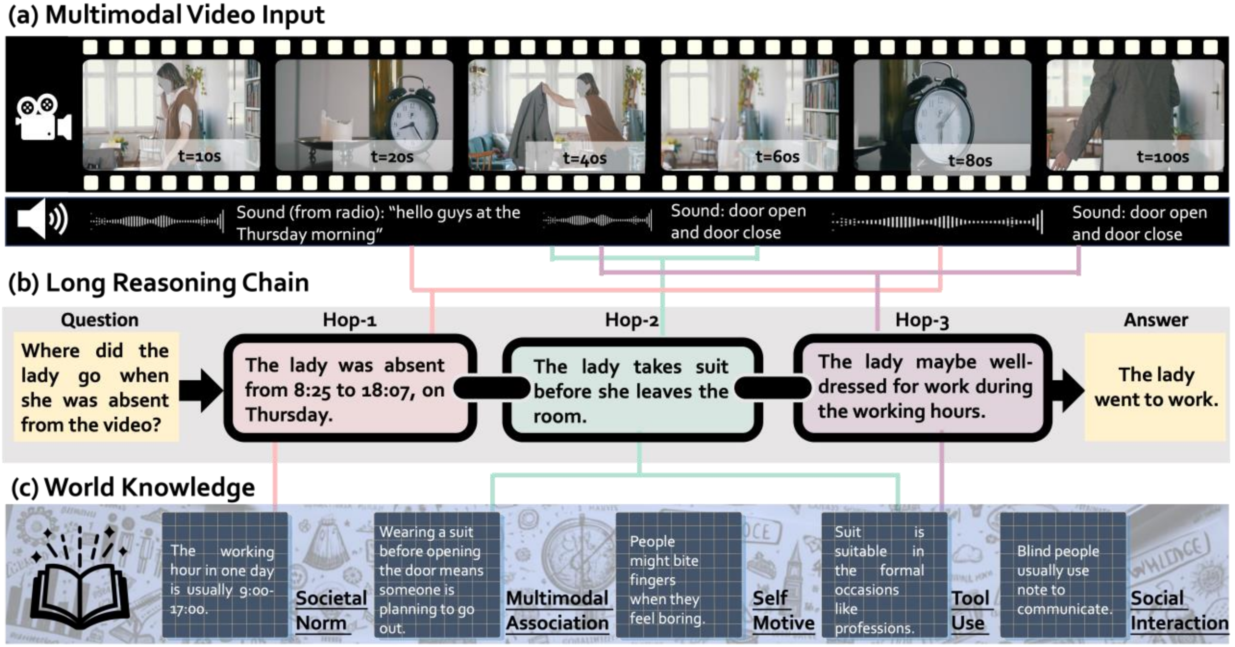


Table 1. **Dataset comparisons.** Reason. stands for reasoning. Avg. stands for average. Q/A stands for question and answer.

Dataset	Multi modal?	World knowledge?	Avg. reason. steps	Avg. length (Q/A)
MSVD-QA [37]	✗	✗	1.40	6.6/1.0
TGIF-QA [15]	✗	✗	1.71	8.7/2.1
TVQA [16]	✗	✗	1.91	13.4/5.3
ActivityNet-QA [46]	✗	✗	1.62	8.7/1.2
NExT-QA [35]	✗	✗	1.31	11.6/2.9
Social-IQ [47]	✓	✗	1.98	11.7/11.4
WorldQA	✓	✓	**4.45**	**14.2/24.3**

1. To determine where the lady went when she was absent from the video, we rely on **visual cues**, **auditory hints**, and the **application of world knowledge**. This forms a reasoning chain to deduce the answer.
2. WorldQA comprises 1**007 question-answer pairs and 303 videos**, spanning five types of world knowledge. On average, the reasoning chain consists of 4.45 steps.

1. The questions in WorldQA require a complex reasoning process, which is evident in the answer lengths: while answers in other VideoQA datasets average below five words, those in WorldQA average 24.3 words.
2. To our knowledge, WorldQA is the first VideoQA dataset that incorporates questions necessitating world knowledge.v

WorldQA: Multimodal World Knowledge in Videos through Long-Chain Reasoning

WorldQA

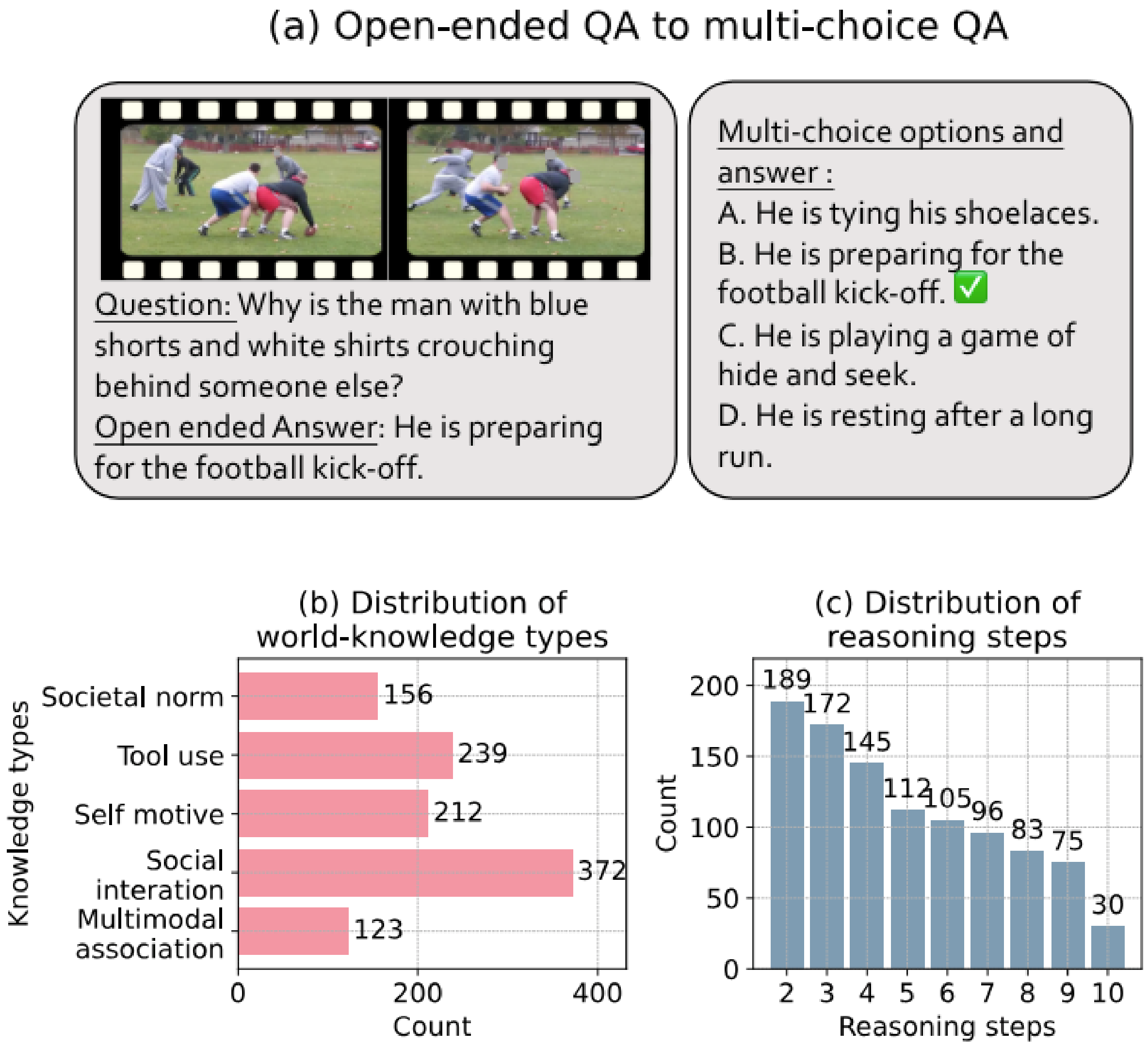


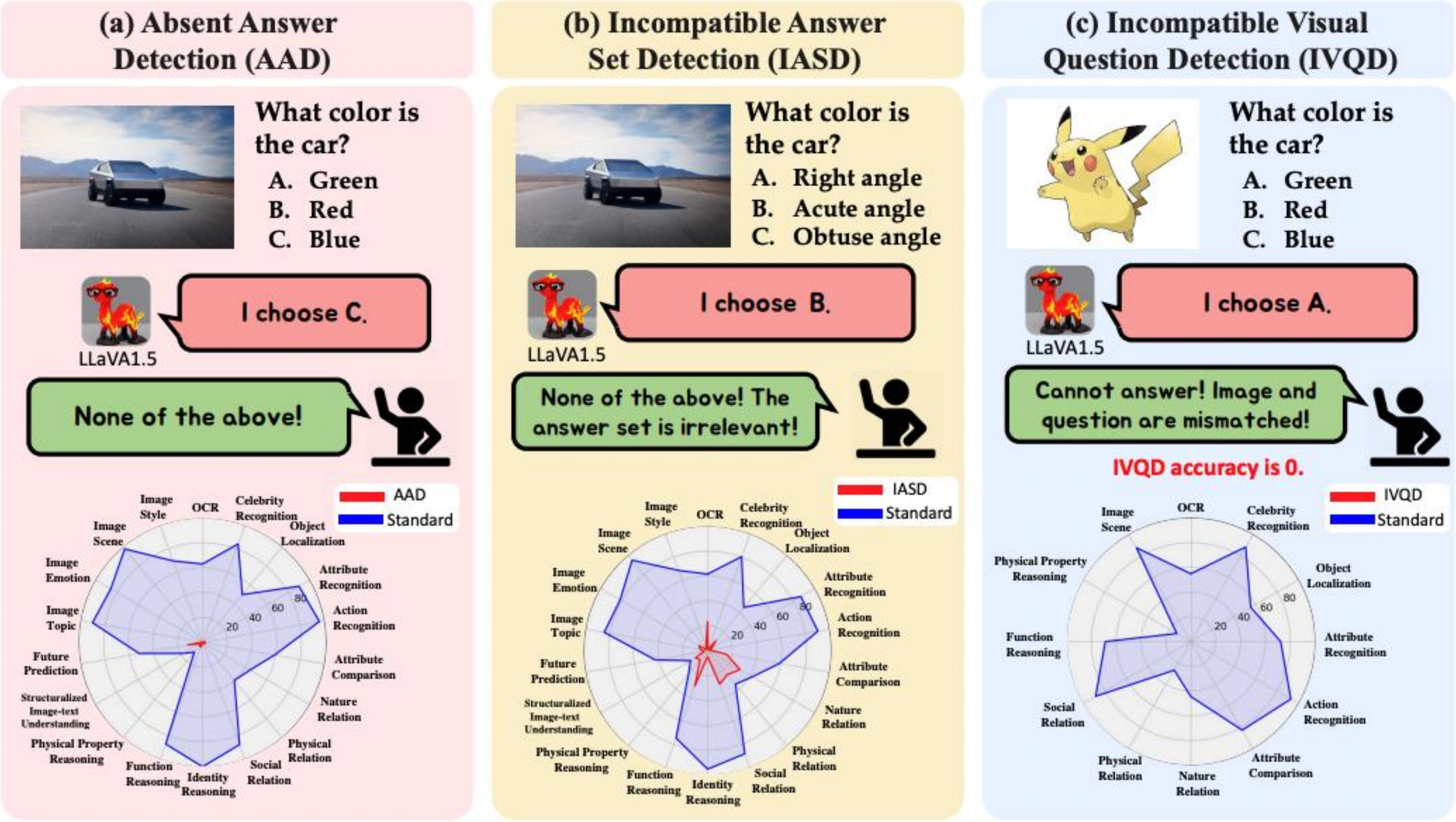
Table 2. **Evaluation of Large Multimodal Models (LMMs), Language Models (LLMs), and WorldRetriver in WorldQA.** We introduce two upperbounds for comparison: LLMs augmented with huamn-annotated event descriptions, labeled as (Aug.) X, and the human performance. The best model in each group is highlighted in blue, while the overall top performer in all tasks is marked in red. Different types of inputs include: Q for Question, V for Video, I for Image, and V_d for Video Description. Langauge model param. indicates the parameter of the language model.

Model (language model param.)	Input	Open ended ↑	Multi choice ↑
Large Multimodal Models (LMMs)-Video			
FrozenBiLM (900M) [40]	Q,V	8.21	0.32
Otter-Video (7B) [17]	Q,V	24.22	6.11
VideoChat (7B) [23]	Q,V	24.43	1.29
LLaMA-Adapter (7B) [11]	Q,V	25.87	12.04
Video-LLaMA (7B) [48]	Q,V	26.80	4.81
Video-ChatGPT (7B) [18]	Q,V	28.51	13.25
mPLUG-Owl (7B) [44]	Q,V	31.89	0.75
GPT-4V(ision) (-) [42]	Q,V	35.37	32.83
Large Multimodal Models (LMMs)-Image			
Qwen-VL (7B) [4]	Q,I	24.04	12.80
LLaVA (7B) [19]	Q,I	31.31	0.30
Large Language Models (LLMs)			
Vicuna (7B) [49]	Q	22.44	0.00
ChatGPT (20B) [24]	Q	24.24	0.00
GPT-4 (-) [24]	Q	28.73	0.00
LLM Agent			
Ours (ChatGPT as LLM) (20B)	Q,V	36.38	36.59
Upper Bound with Human Transcription			
(Aug.) Vicuna (7B)	Q, V_d	38.71	23.90
(Aug.) ChatGPT (20B)	Q, V_d	46.50	46.06
(Aug.) GPT-4 (-)	Q, V_d	48.46	56.06
Human-Level Performance			
Human	Q,V	72.43	88.79

WorldQA: Multimodal World Knowledge in Videos through Long-Chain Reasoning

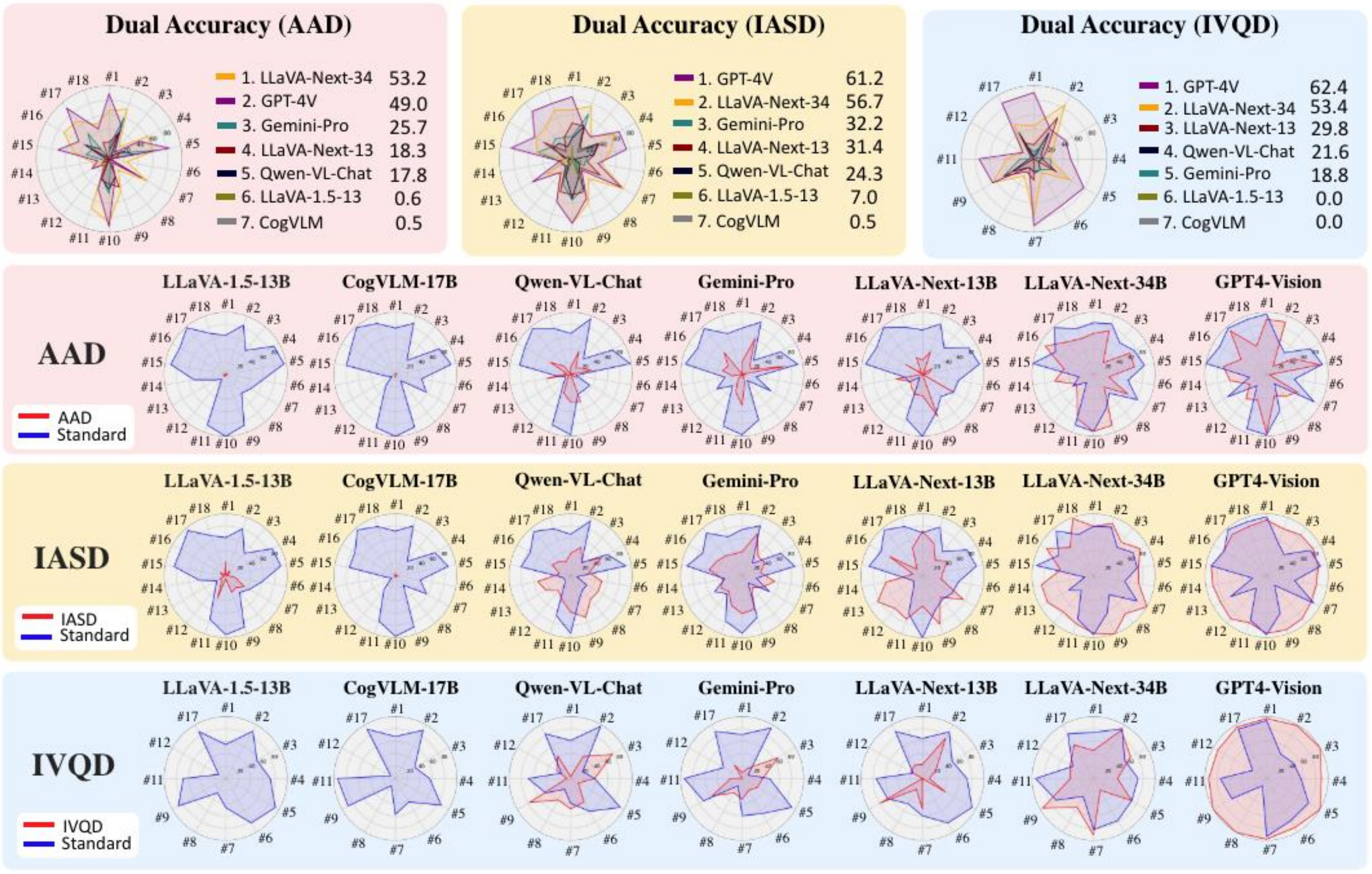
Unsolvable Problem Detection

Evaluating Trustworthiness of Vision Language Models



Unsolvable Problem Detection

Evaluating Trustworthiness of Vision Language Models



#1: OCR #2: Celebrity Recognition #3: Object Localization #4: Attribute Recognition #5: Action Recognition #6: Attribute Comparison
#7: Nature Relation #8: Physical Relation #9: Social Relation #10: Identity Reasoning #11: Function Reasoning #12: Physical Property Reasoning
#13: Structuralized Image-text Understanding #14: Future Prediction #15: Image Topic #16: Image Emotion #17: Image Scene #18: Image Style

AI Assistant with Scene Graph

Neuro-Symbolic Assistant

PANOPTIC SCENE GRAPH

Panoptic Scene Graph [ECCV'22]

PANOPTIC VIDEO SCENE GRAPH

Panoptic **Video** Scene Graph [CVPR'23]

4D PANOPTIC SCENE GRAPH

4D Panoptic Scene Graph [NeurIPS'23]

Jingkang Yang

Bo Li

Yuanhan Zhang

MMLab @NTU

AI Assistant with VLMs

End-to-End Assistant

OTTER

Deep diving into large Multi-modality Models

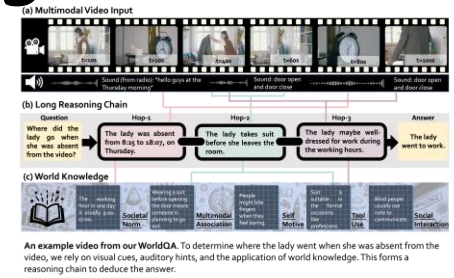
OTTER-HD

A High-Res Multi-modality Model

LLaVA-NeXT

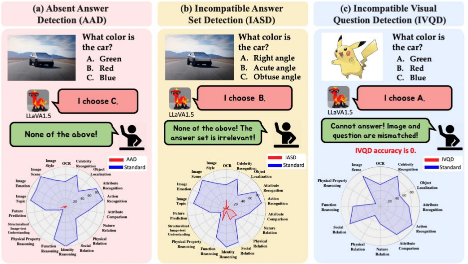
Benchmarking AI Assistants

LMM-Eval



WorldQA

FunQA



Unsolvable Problem Detection

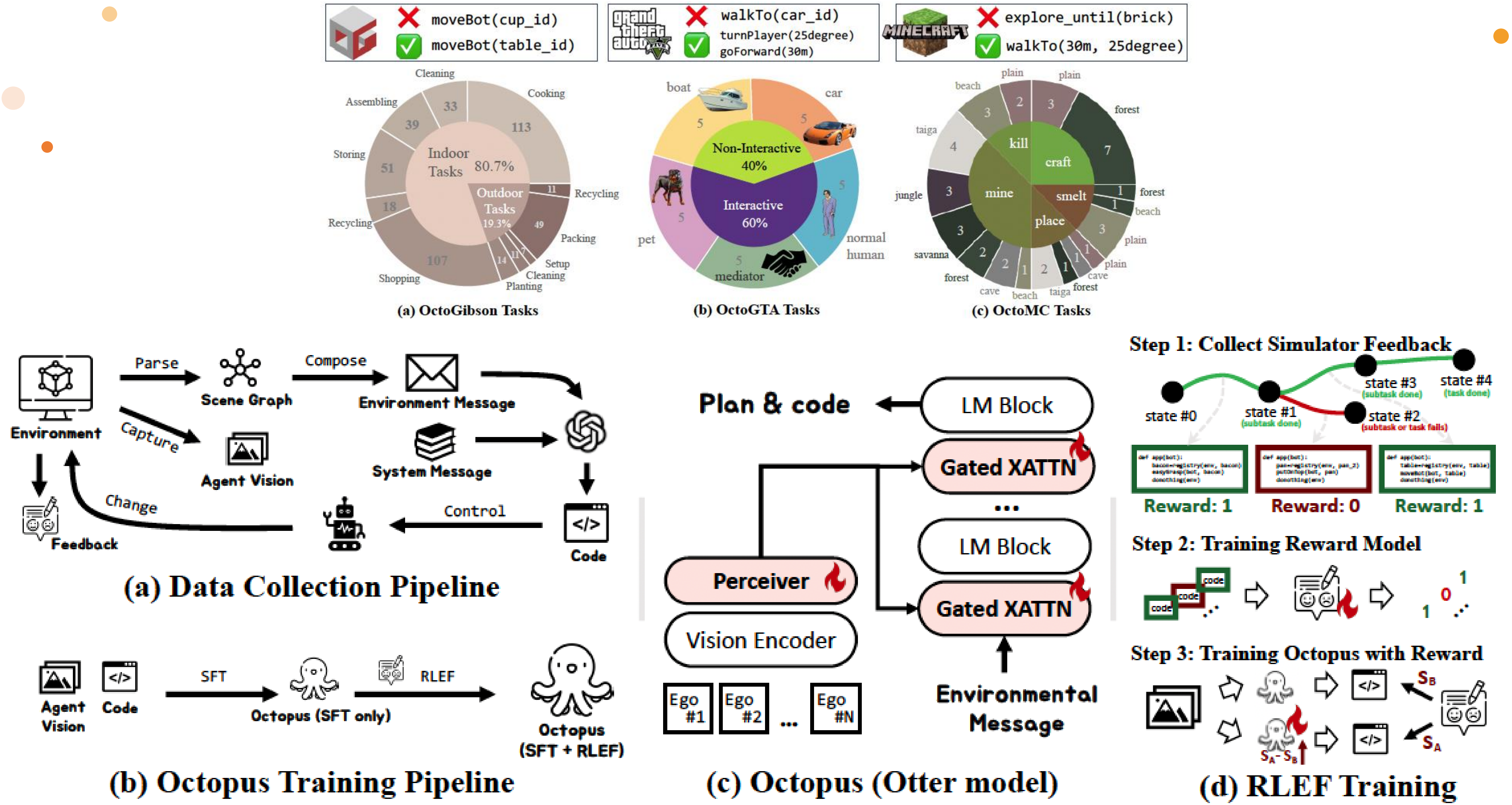
Embodied AI Assistant

Octopus

Towards Building
Multi-Modal AI Assistant

↓

Towards Building
Embodied Multi-Modal AI Assistant



Octopus: Embodied Vision-Language Programmer from Environmental Feedback

AI Assistant with Scene Graph

Neuro-Symbolic Assistant

Panoptic Scene Graph [ECCV'22]

Panoptic **Video** Scene Graph [CVPR'23]

4D Panoptic Scene Graph [NeurIPS'23]

Jingkang Yang

Bo Li

Yuanhan Zhang

MMLab @NTU

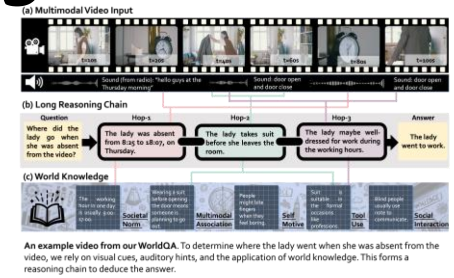
AI Assistant with VLMs

End-to-End Assistant

LLaVA-NeXT

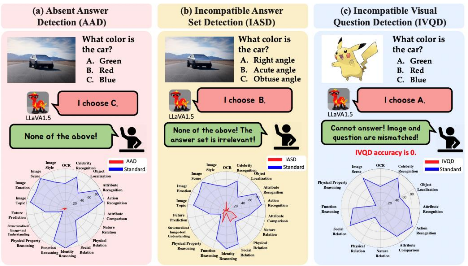
Benchmarking AI Assistants

LMM-Eval



WorldQA

FunQA



Unsolvable Problem Detection

Embodied AI Assistant

Octopus