CVPR
JUNE 17-21, 2024
SEATTLE, WA

Large Multimodal Models

Towards Building General-Purpose Multimodal Assistant

June 2024

Chunyuan Li
ByteDance
https://chunyuan.li

❑ Outline

① Overview

- Basics of Large Multimodal Model
- Year 2023 & Current Status

② Prototype: LLaVA-1.0

- Visual Instruction Tuning / LLaVA

③ Recent Advances: LLaVA Family

- Performance: LLaVA-1.5 & LLaVA-NeXT
- Applications: LLaVA-Med
- New Capabilities: LLaVA-Interactive & LLaVA-Plus

① Overview:

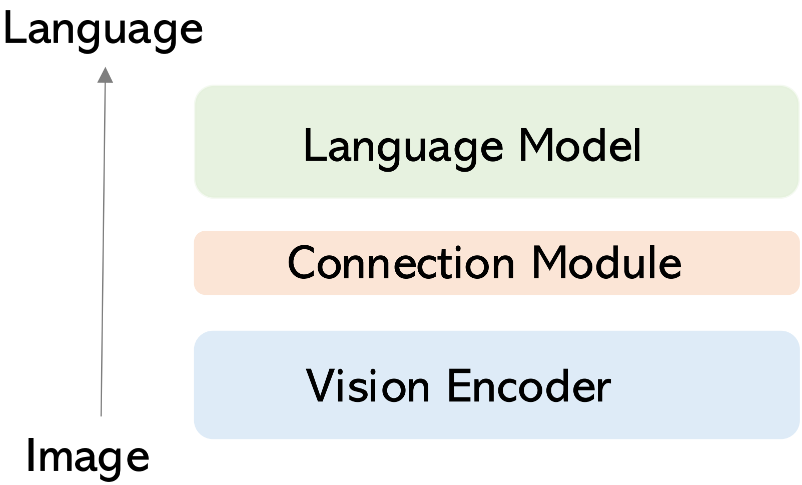
Large Multimodal Models (LMM)

Large Multimodal Models: Image-to-Text Generative Models

Model Architectures

- (Pre-trained) Image Encoder and Language Models
- Trainable modules to connect to two modalities

A dog lying on the grass next to a frisbee



Language

Language Model

Connection Module

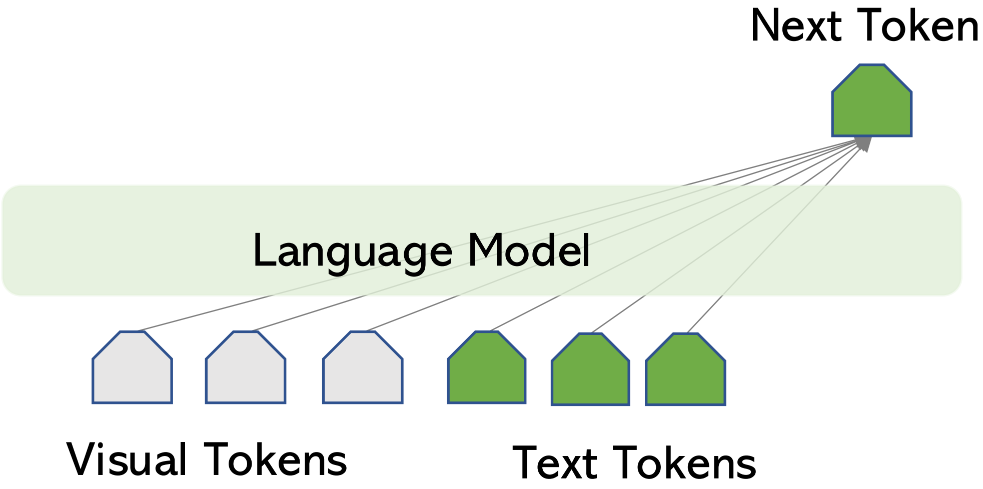
Vision Encoder

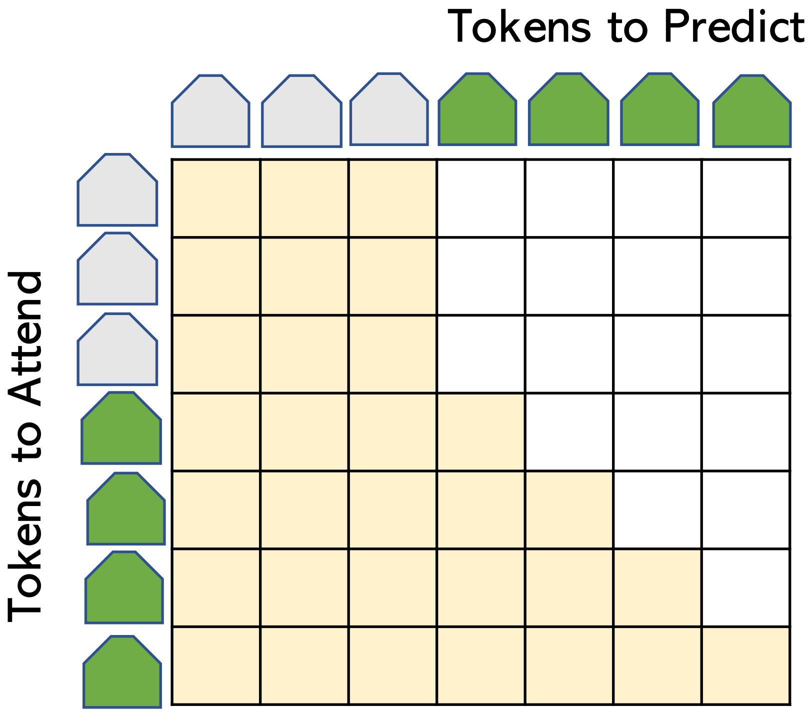
Image

Large Multimodal Models: Image-to-Text Generative Models

❑ Training Objective

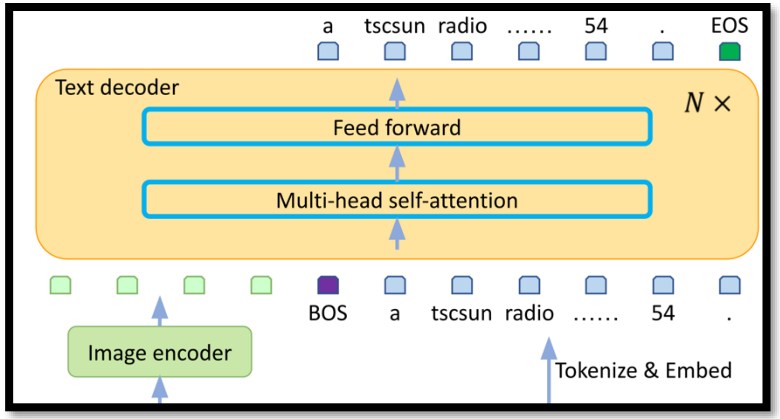
- Cross-Attended Image-to-Text Generation
- Autoregressive loss on language output



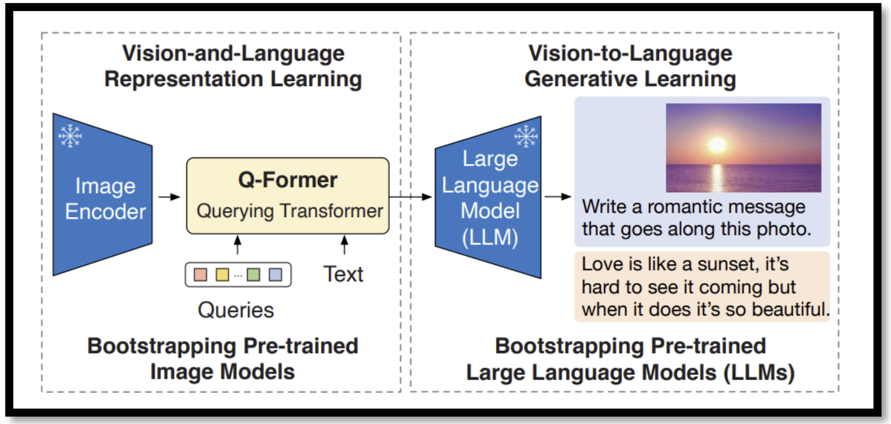


Example 1: LMM with Image-Text Pairs

- GIT



- BLIP2


Vision-and-Language
Representation Learning
Vision-to-Language
Generative Learning
Image
Encoder
Q-Former
Querying Transformer
Queries
Text
Large
Language
Model
(LLM)
Write a romantic message that goes along this photo.
Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.
Bootstrapping Pre-trained
Image Models
Bootstrapping Pre-trained
Large Language Models (LLMs)

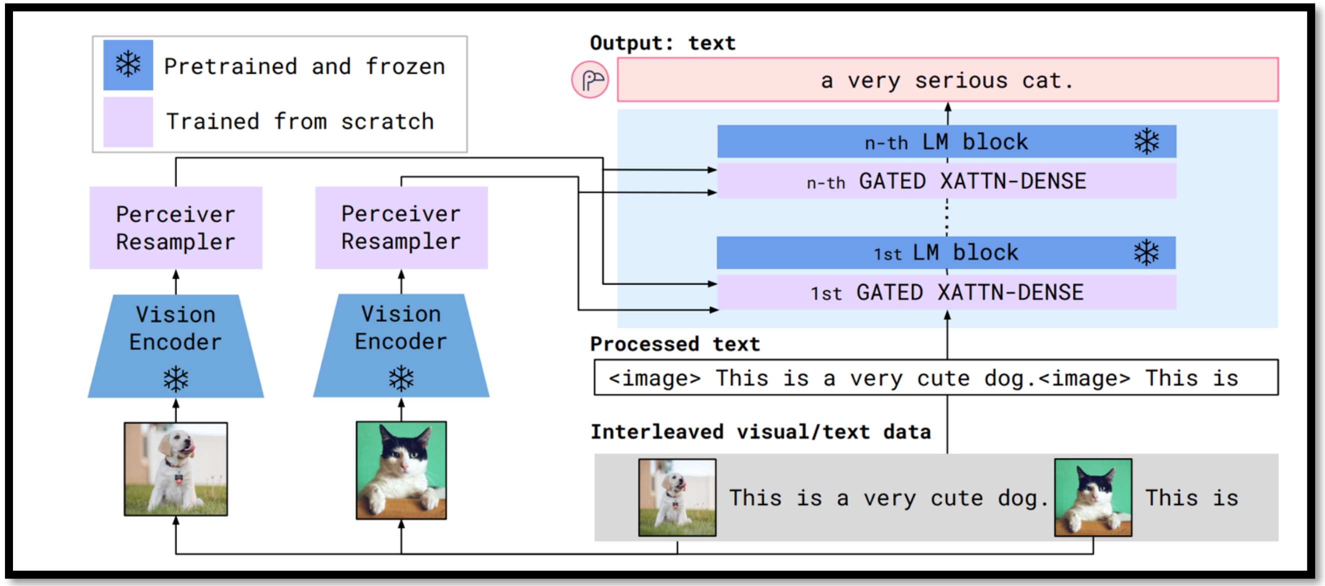
Language Model	From Scratch	Pre-trained: FLAN-T5/OPT
Connection Module		Q-Former: Lightweight Querying Transformer
Vision Encoder	Contrastive pre-trained: Florence/CLIP	Contrastive pre-trained: EVA/CLIP

GIT: A Generative Image-to-text Transformer for Vision and Language

Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models

Example 2: LMM with Interleaved Image-Text Data

- **Flamingo:**

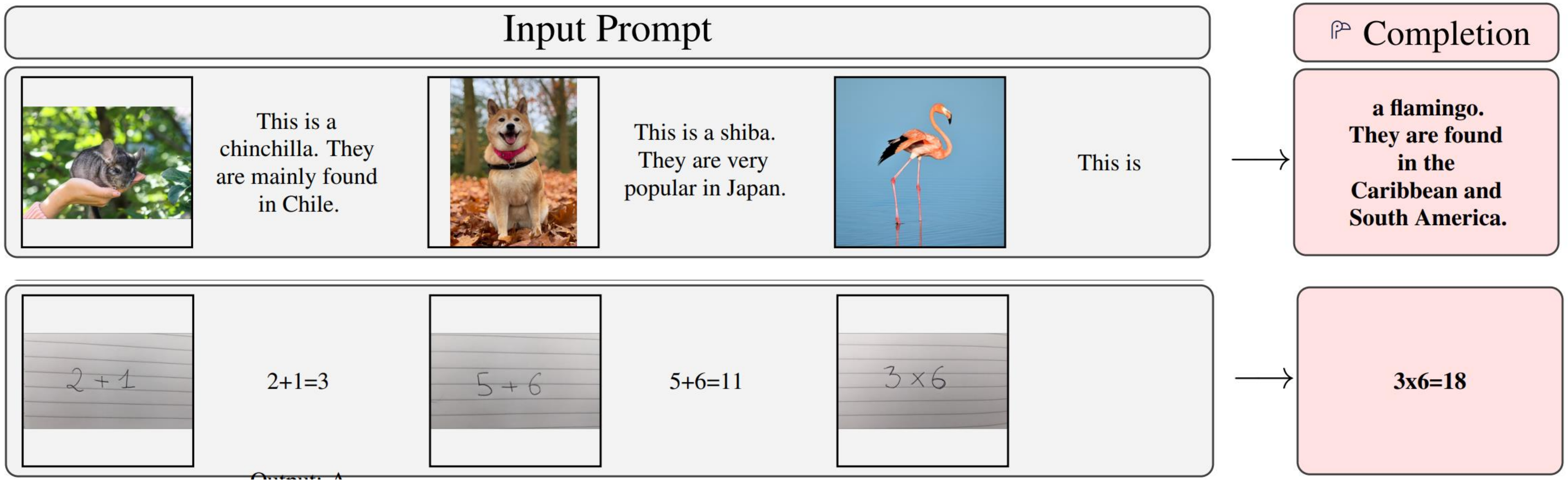


Language Model	Pre-trained: 70B Chinchilla
Connection Module	Perceiver Resampler Gated Cross-attention + Dense
Vision Encoder	Pre-trained: Nonrmalizer-Free ResNet (NFNet)

Example 2: LMM with Interleaved Image-Text Data

- Flamingo: Multimodal In-Context-Learning

Emerging Property



Input Prompt

Completion

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

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

This is

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

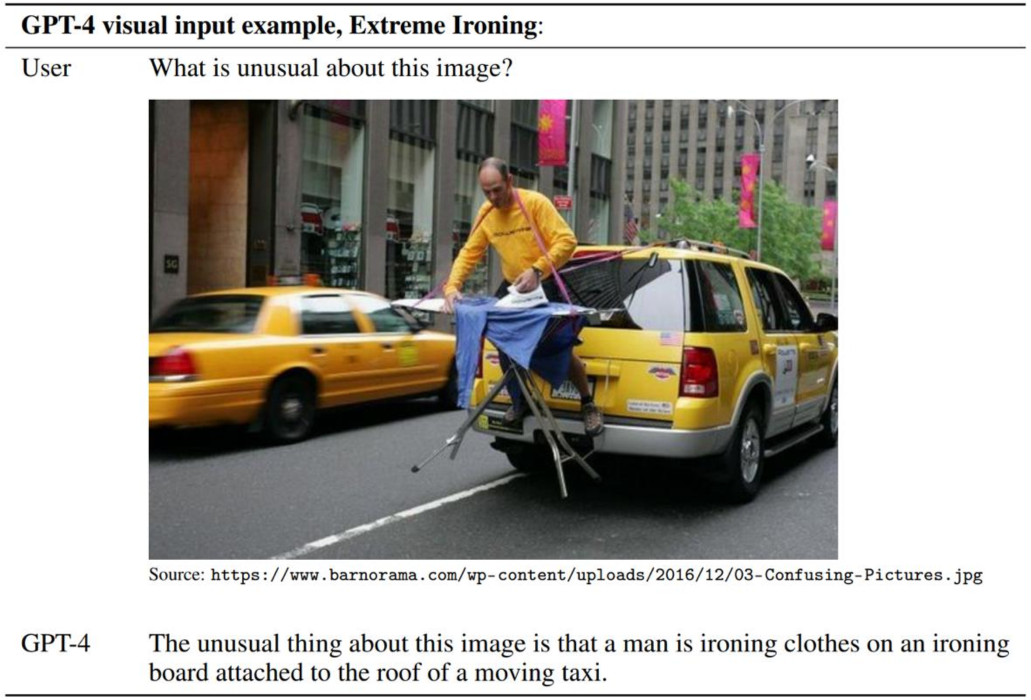
2+1=3

5+6=11

→ **3x6=18**

OpenAI GPT-4V

- Model Details: Unknown
- Capability: Strong zero-shot visual understanding & reasoning on many user-oriented tasks in the wild
- Year 2023: How can we build GPT-4V like models?

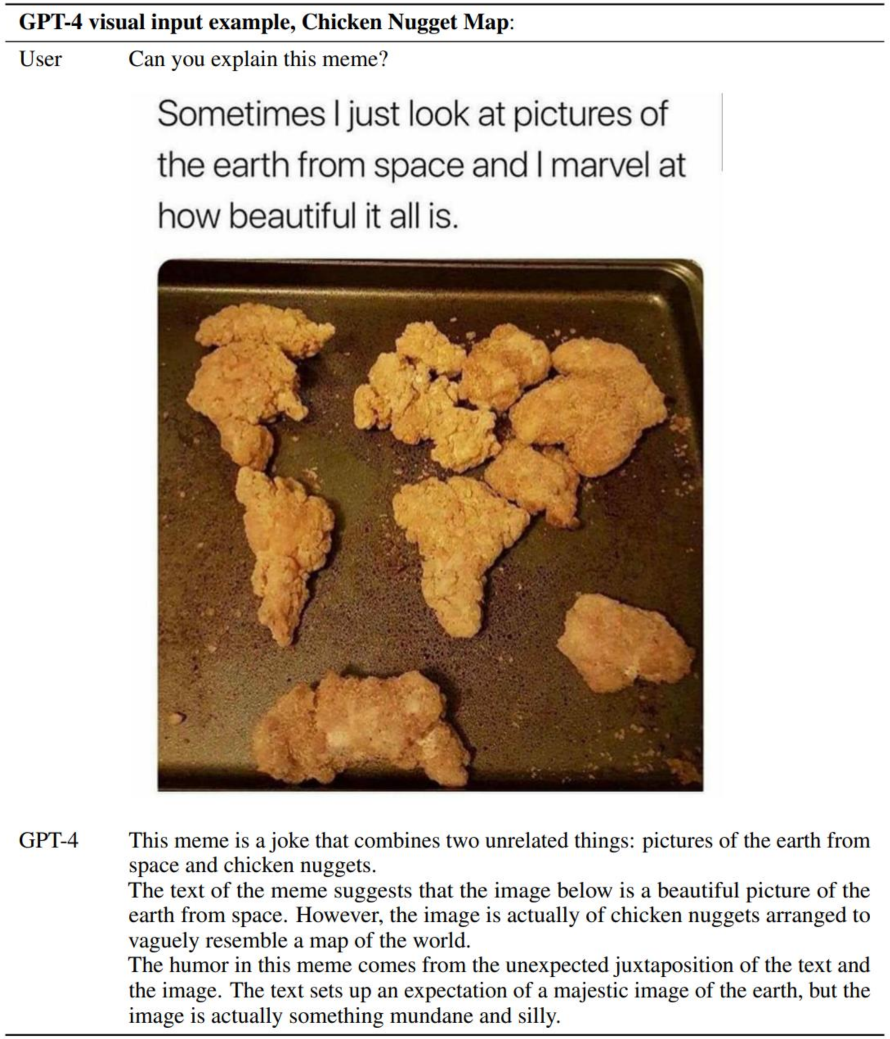


GPT-4 visual input example, Extreme Ironing:

User What is unusual about this image?

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

GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.



GPT-4 visual input example, Chicken Nugget Map:

User Can you explain this meme?

Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.

GPT-4 This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets.
The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world.
The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.

GPT-4 Technical Report, OpenAI

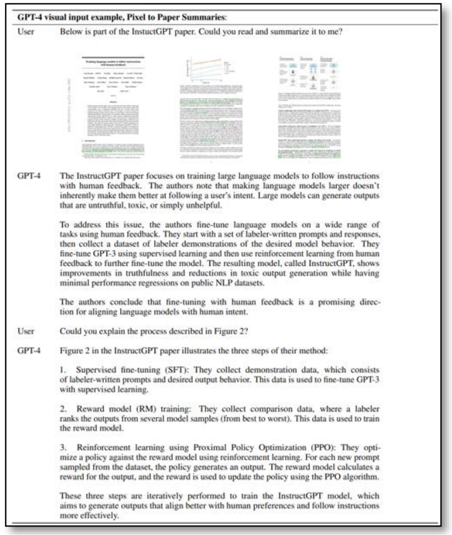
Year 2023

1. How to build GPT-4V?

In terms of prototyping new capabilities

2. A large gap remains…

In terms of scaling the capability



GPT-4 visual input example, Pixel to Paper Summaries:

User Below is part of the InstuctGPT paper. Could you read and summarize it to me?

GPT-4 The InstructGPT paper focuses on training large language models to follow instructions with human feedback. The authors note that making language models larger doesn't inherently make them better at following a user's intent. Large models can generate outputs that are untruthful, toxic, or simply unhelpful.

To address this issue, the authors fine-tune language models on a wide range of tasks using human feedback. They start with a set of labeler-written prompts and responses, then collect a dataset of labeler demonstrations of the desired model behavior. They fine-tune GPT-3 using supervised learning and then use reinforcement learning from human feedback to further fine-tune the model. The resulting model, called InstructGPT, shows improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets.

The authors conclude that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

User Could you explain the process described in Figure 2?

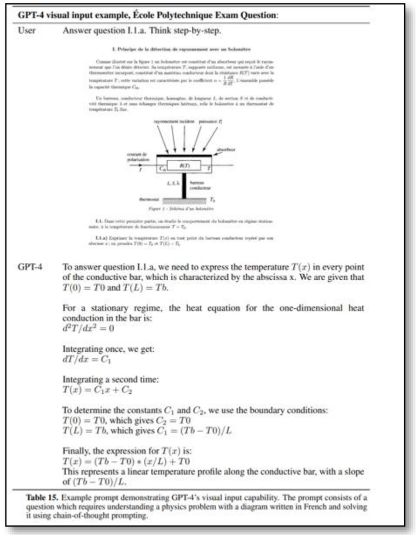
GPT-4 Figure 2 in the InstructGPT paper illustrates the three steps of their method:

1. Supervised fine-tuning (SFT): They collect demonstration data, which consists of labeler-written prompts and desired output behavior. This data is used to fine-tune GPT-3 with supervised learning.

2. Reward model (RM) training: They collect comparison data, where a labeler ranks the outputs from several model samples (from best to worst). This data is used to train the reward model.

3. Reinforcement learning using Proximal Policy Optimization (PPO): They optimize a policy against the reward model using reinforcement learning. For each new prompt sampled from the dataset, the policy generates an output. The reward model calculates a reward for the output, and the reward is used to update the policy using the PPO algorithm.

These three steps are iteratively performed to train the InstructGPT model, which aims to generate outputs that align better with human preferences and follow instructions more effectively.



GPT-4 visual input example, École Polytechnique Exam Question:

User Answer question I.1.a. Think step-by-step.

GPT-4 To answer question I.1.a, we need to express the temperature $T(x)$ in every point of the conductive bar, which is characterized by the abscissa x. We are given that $T(0) = T0$ and $T(L) = Tb$.

For a stationary regime, the heat equation for the one-dimensional heat conduction in the bar is:
$d^2T/dx^2 = 0$

Integrating once, we get:
$dT/dx = C_1$

Integrating a second time:
$T(x) = C_1 x + C_2$

To determine the constants C_1 and C_2, we use the boundary conditions:
$T(0) = T0$, which gives $C_2 = T0$
$T(L) = Tb$, which gives $C_1 = (Tb - T0)/L$

Finally, the expression for $T(x)$ is:
$T(x) = (Tb - T0) * (x/L) + T0$
This represents a linear temperature profile along the conductive bar, with a slope of $(Tb - T0)/L$.

Table 15. Example prompt demonstrating GPT-4's visual input capability. The prompt consists of a question which requires understanding a physics problem with a diagram written in French and solving it using chain-of-thought prompting.

Year 2024

1. The general recipes ``converge''

In terms of modeling and training data

2. 90% performance on benchmarks

But gaps remains as a general-purpose assistant

Recap on Language Modeling: Large Language Models (LLM)

Language Space

GPT-3 → Close Source

In-context-learning
Chain-of-thoughts (CoT)

Close Source

GPT-4 (In-context-learning
Chain-of-thoughts (CoT)
Instruction-Following)

Open Source

LLaMA Mixtral
Qwen Yi

Multimodal Space

Flamingo →

Close Source

Gemini GPT-4V GPT-4o Qwen-VL-Max
Grok-V Claude-3 Reka MM1

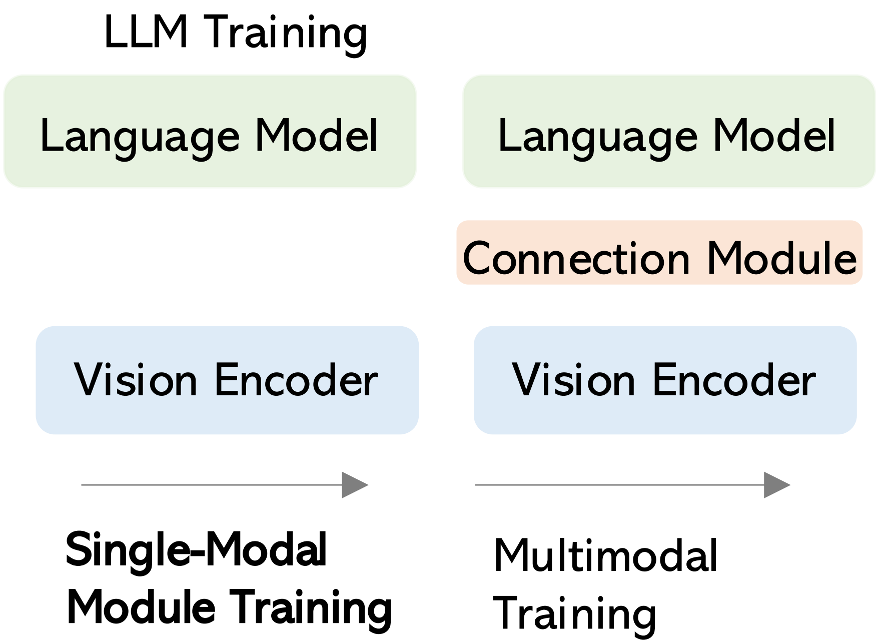
Open Source

CogVLM Yi-VL Intern-VL Emu2
LLaVA MiniGPT4 MiniCPM Idefics VILA

A debate

- Stage-wise Training

(GPT-4V)



LLM Training

Language Model

Language Model

Connection Module

Vision Encoder

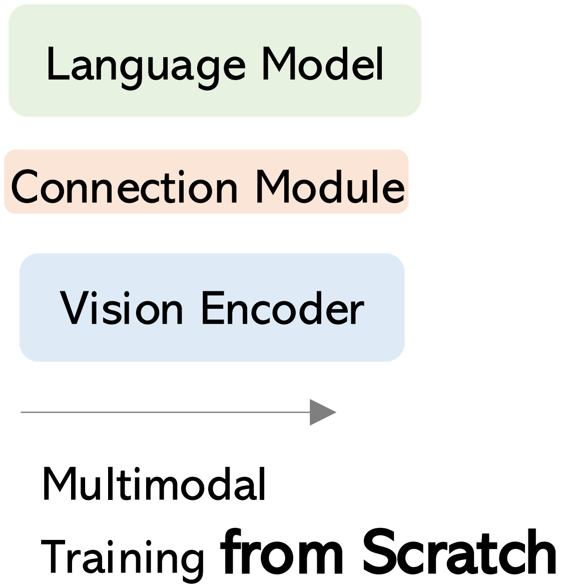
Vision Encoder

Single-Modal Module Training

Multimodal Training

- Native Multimodal Training

(Gemini, GPT-4o)



Language Model

Connection Module

Vision Encoder

Multimodal Training **from Scratch**

② Large Multimodal Models

-- Building gpt4v with open-source resources

LLaVA as a running example in this lecture

- Data
- Model
- Performance

Visual Instruction Tuning with GPT-4

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

Haotian Liu*, Chunyuan Li*, Qingyang Wu, Yong Jae Lee (* Equal contribution)

Self-Instruct with Strong Teacher LLMs

	LLaMA	Alpaca	Vicuna
Teacher		GPT-3.5	ShareGPT (Human & GPT)
Instruction-following Data	None	52K	700K (70 conversions)

But No Teacher is available on multiGPT4?

GPT-4-LLM	LLaVA
GPT-4 (text-only)	GPT-4 (text-only)

- 158K multimodal instruction following data (First & High Quality)

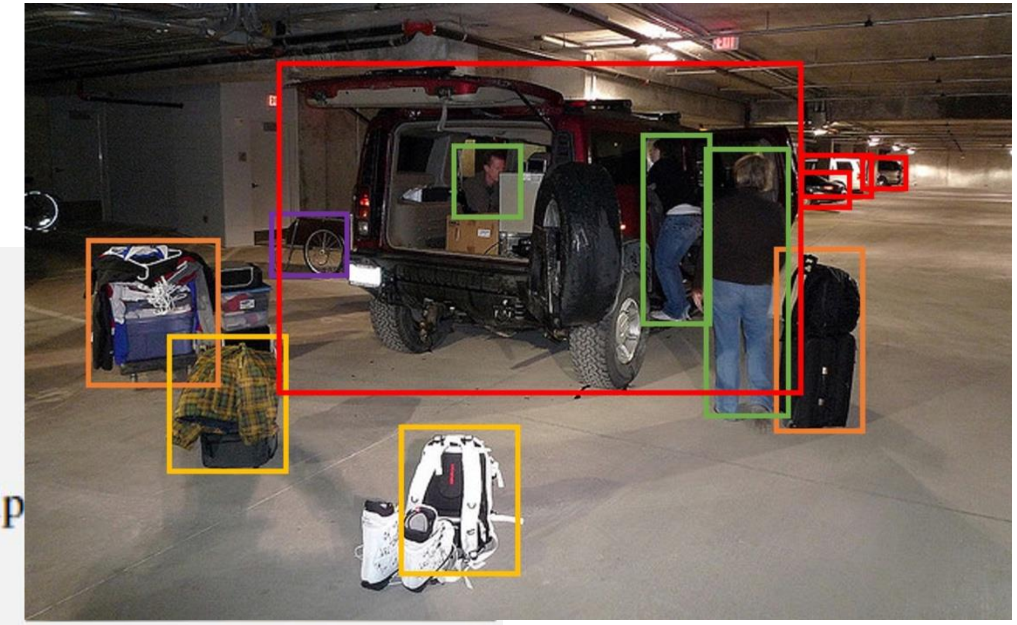
→ Multimodal Chatbot

Large Language and Vision Assistant

GPT-assisted Visual Instruction Data Generation

- Rich Symbolic Representations of Images
- In-context-learning with a few manual examples

→ Text-only GPT-4



Context type 1: Captions

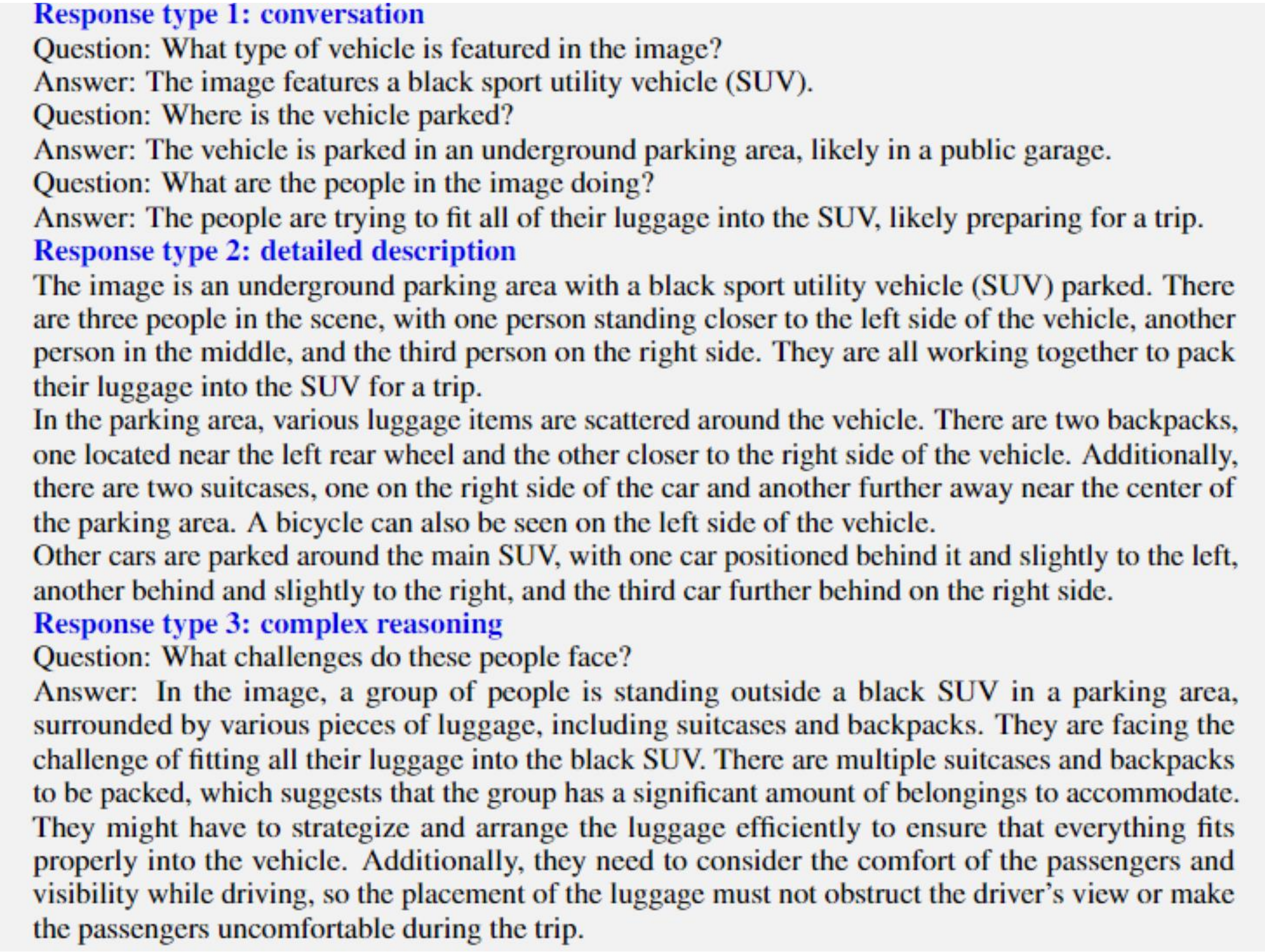
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van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

Context type 2: Boxes

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

GPT-assisted Visual Instruction Data Generation

Three type of instruction-following responses



Response type 1: conversation

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

Response type 2: detailed description

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

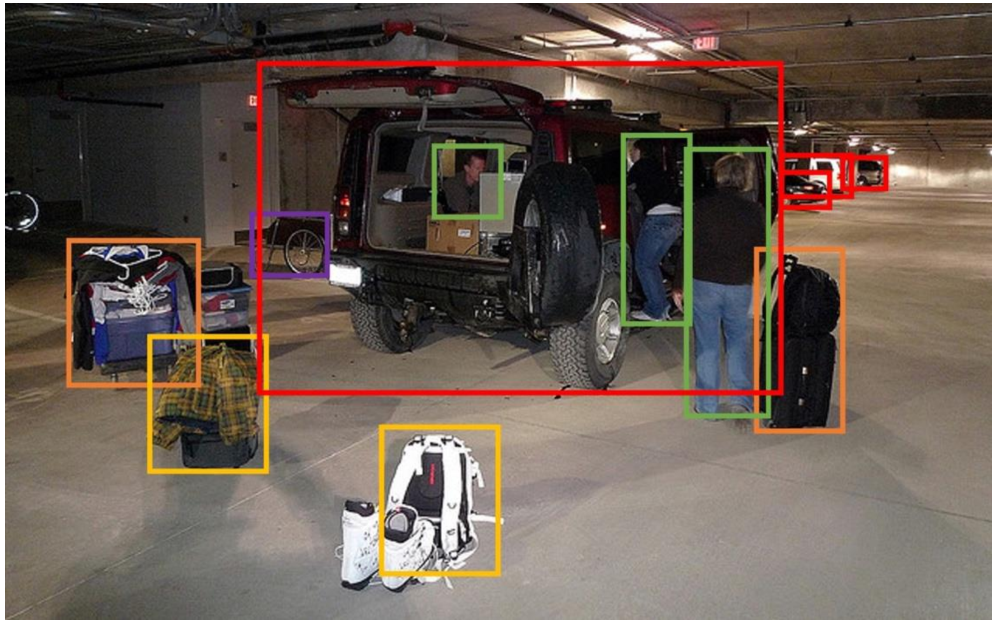
In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center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

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

Response type 3: complex reasoning

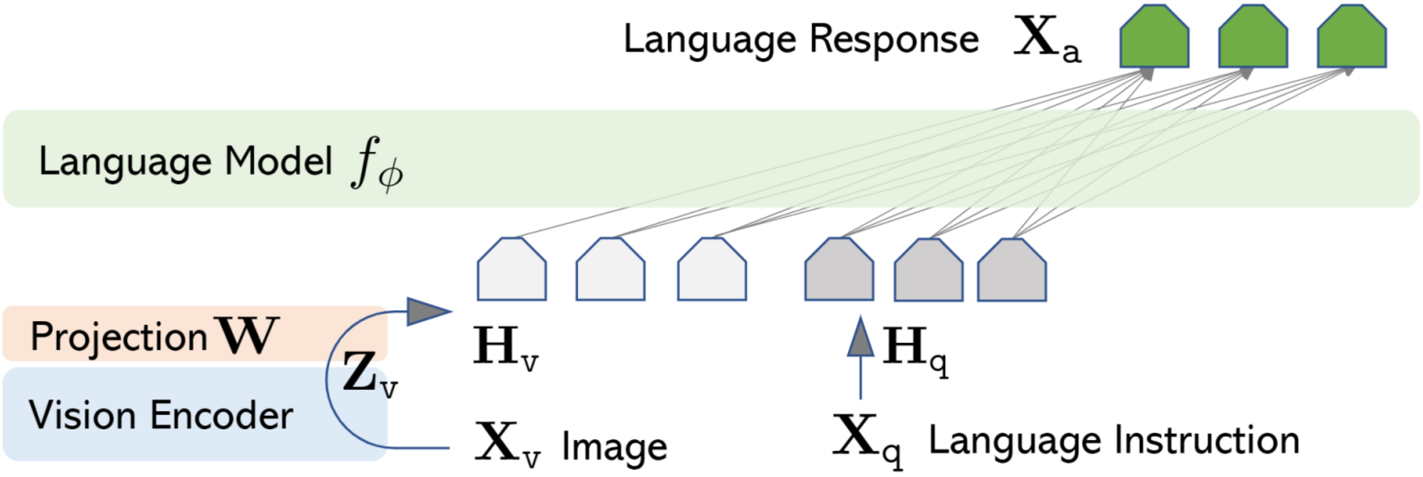
Question: What challenges do these people face?

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



LLaVA: Large Language-and-Vision Assistant

❑ Architecture



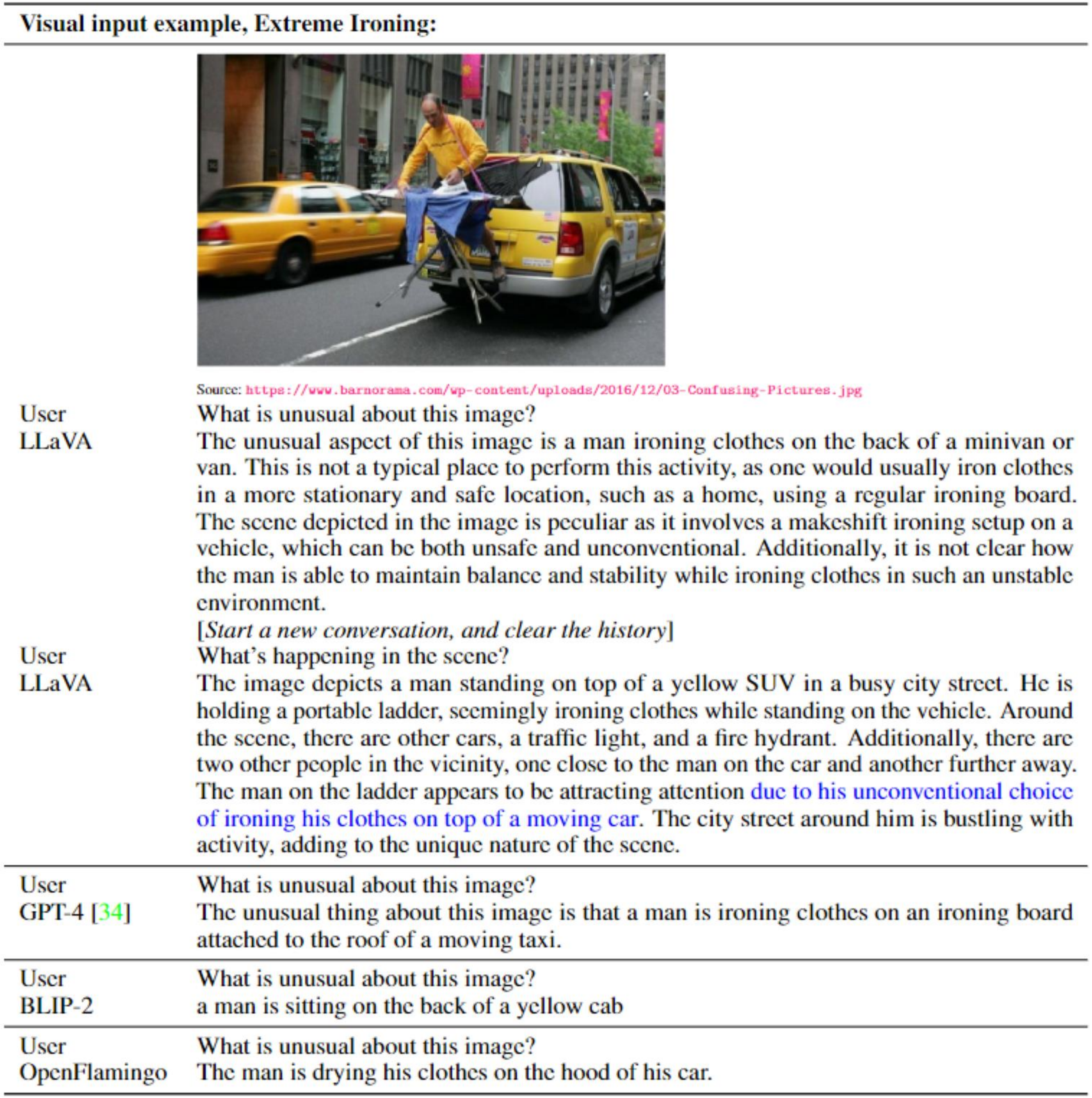
❑ Two-stage Training

- **Stage 1: Pre-training for Feature Alignment**.

  Only the projection matrix is updated, based on a subset of CC3M.
- **Stage 2: Fine-tuning End-to-End.** Both the projection matrix and LLM are updated
  - **Visual Chat**: Our generated multimodal instruction data for daily user-oriented applications.
  - **Science QA**: Multimodal reasoning dataset for the science domain.

Strong Visual Reasoning Ability

Example 1: Extreme Ironing



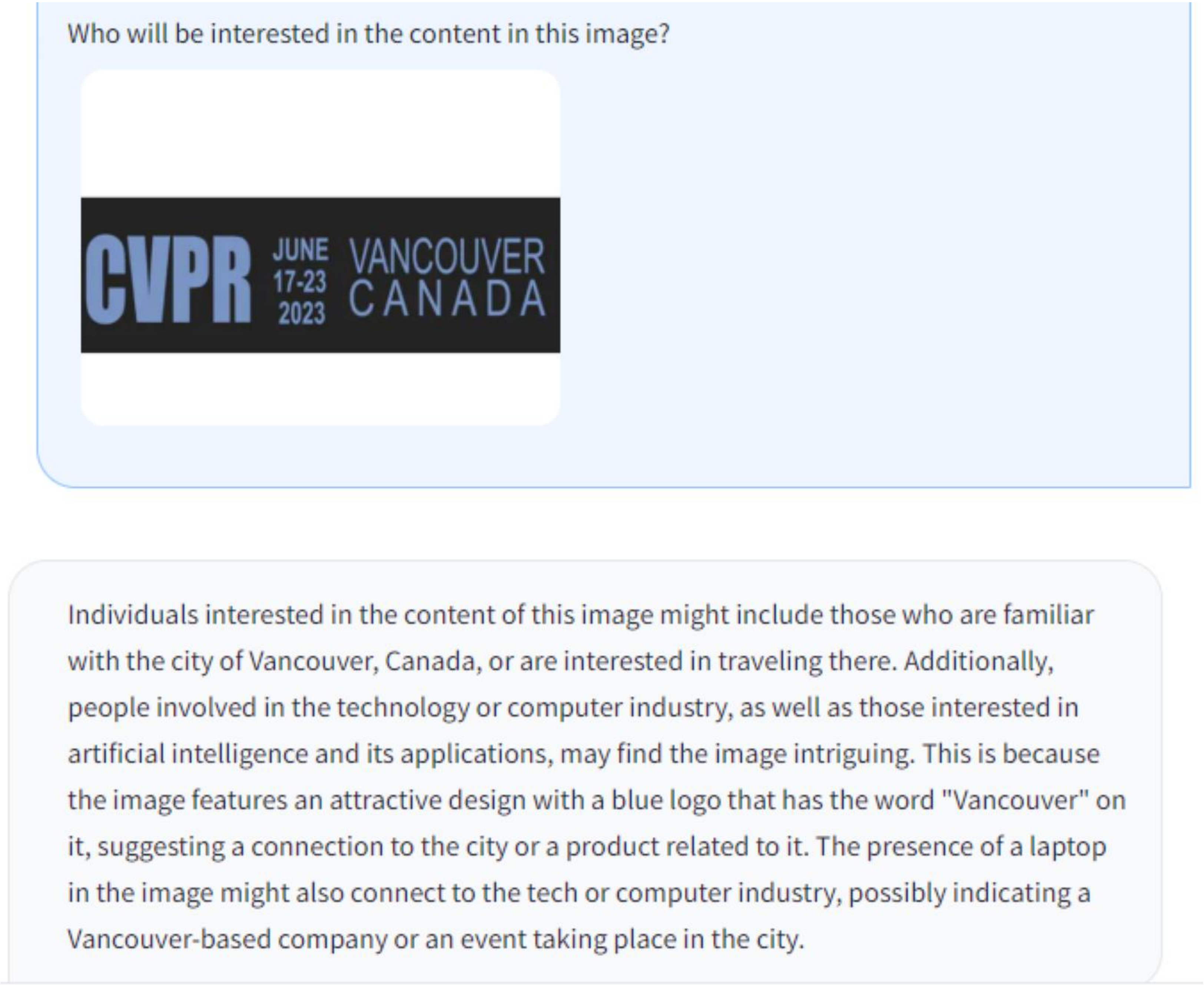
Visual input example, Extreme Ironing:

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

User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
	[*Start a new conversation, and clear the history*]
User	What's happening in the scene?
LLaVA	The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.
User	What is unusual about this image?
GPT-4 [34]	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User	What is unusual about this image?
BLIP-2	a man is sitting on the back of a yellow cab
User	What is unusual about this image?
OpenFlamingo	The man is drying his clothes on the hood of his car.

Strong OCR Ability

Example 1: CVPR & Vancouver

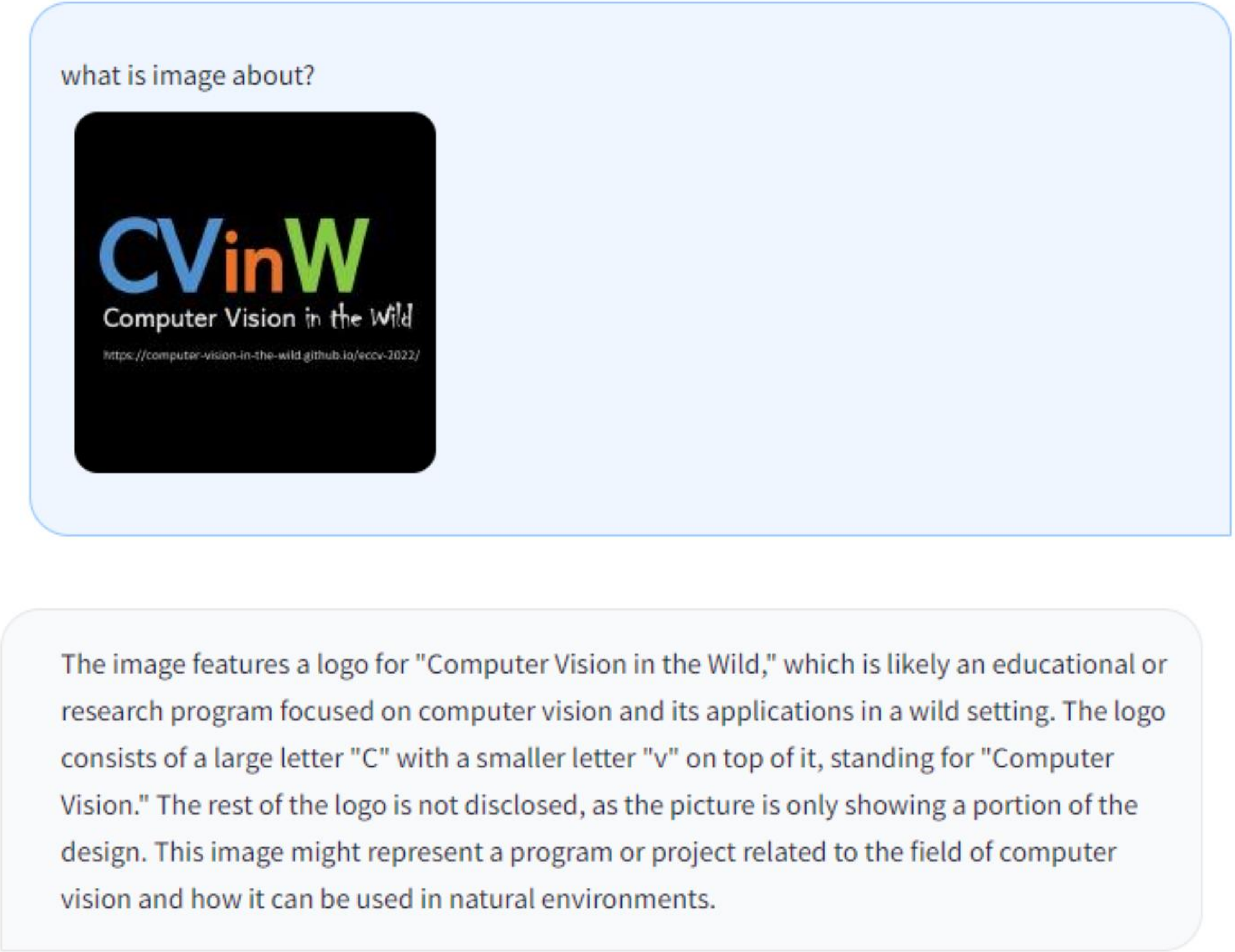


Who will be interested in the content in this image?

Individuals interested in the content of this image might include those who are familiar with the city of Vancouver, Canada, or are interested in traveling there. Additionally, people involved in the technology or computer industry, as well as those interested in artificial intelligence and its applications, may find the image intriguing. This is because the image features an attractive design with a blue logo that has the word "Vancouver" on it, suggesting a connection to the city or a product related to it. The presence of a laptop in the image might also connect to the tech or computer industry, possibly indicating a Vancouver-based company or an event taking place in the city.

Strong OCR Ability

Example 2: CVinW workshop



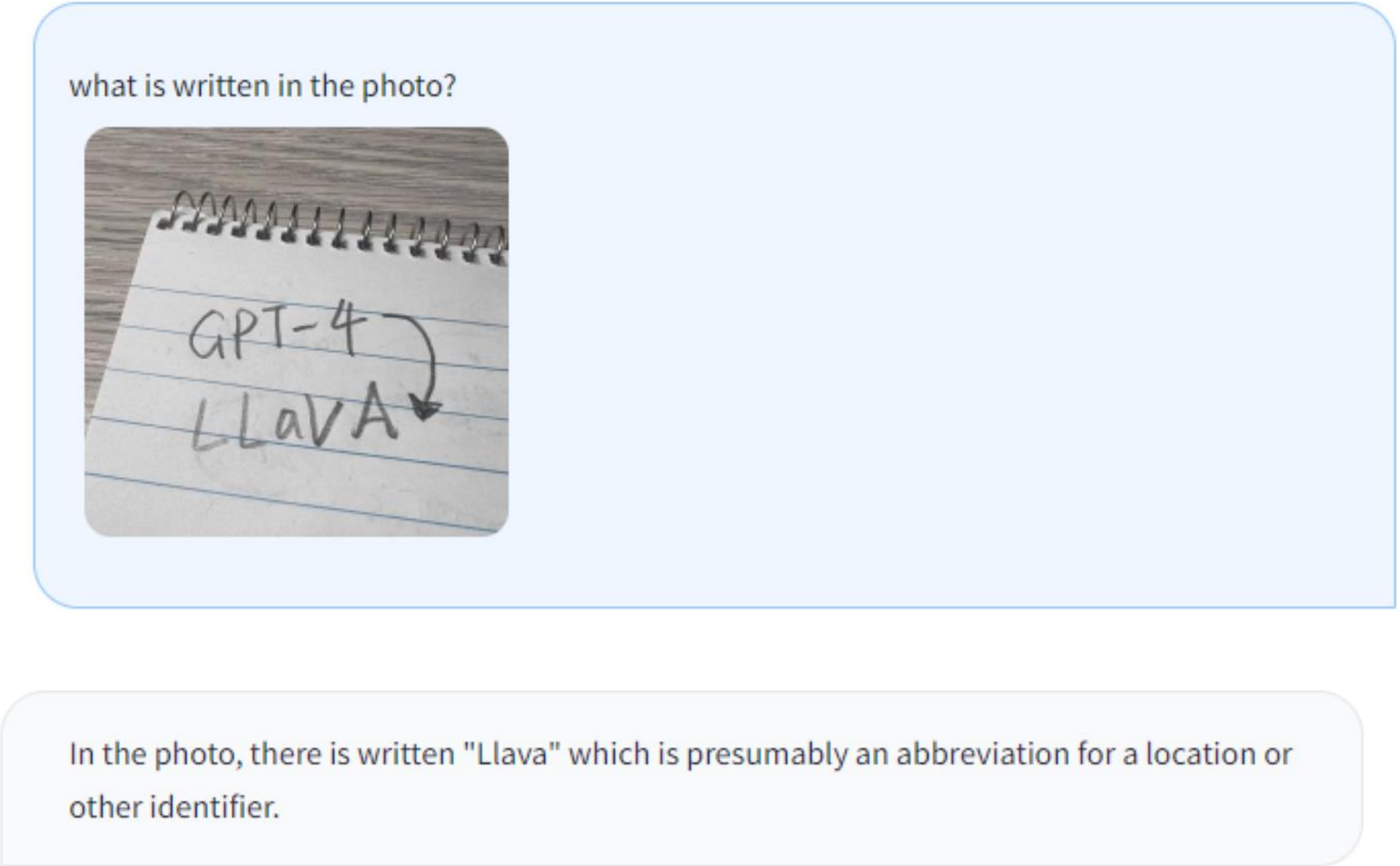
CVinW
Computer Vision in the Wild
https://computer-vision-in-the-wild.github.io/eccv-2022/

what is image about?

The image features a logo for "Computer Vision in the Wild," which is likely an educational or research program focused on computer vision and its applications in a wild setting. The logo consists of a large letter "C" with a smaller letter "v" on top of it, standing for "Computer Vision." The rest of the logo is not disclosed, as the picture is only showing a portion of the design. This image might represent a program or project related to the field of computer vision and how it can be used in natural environments.

Strong OCR Ability

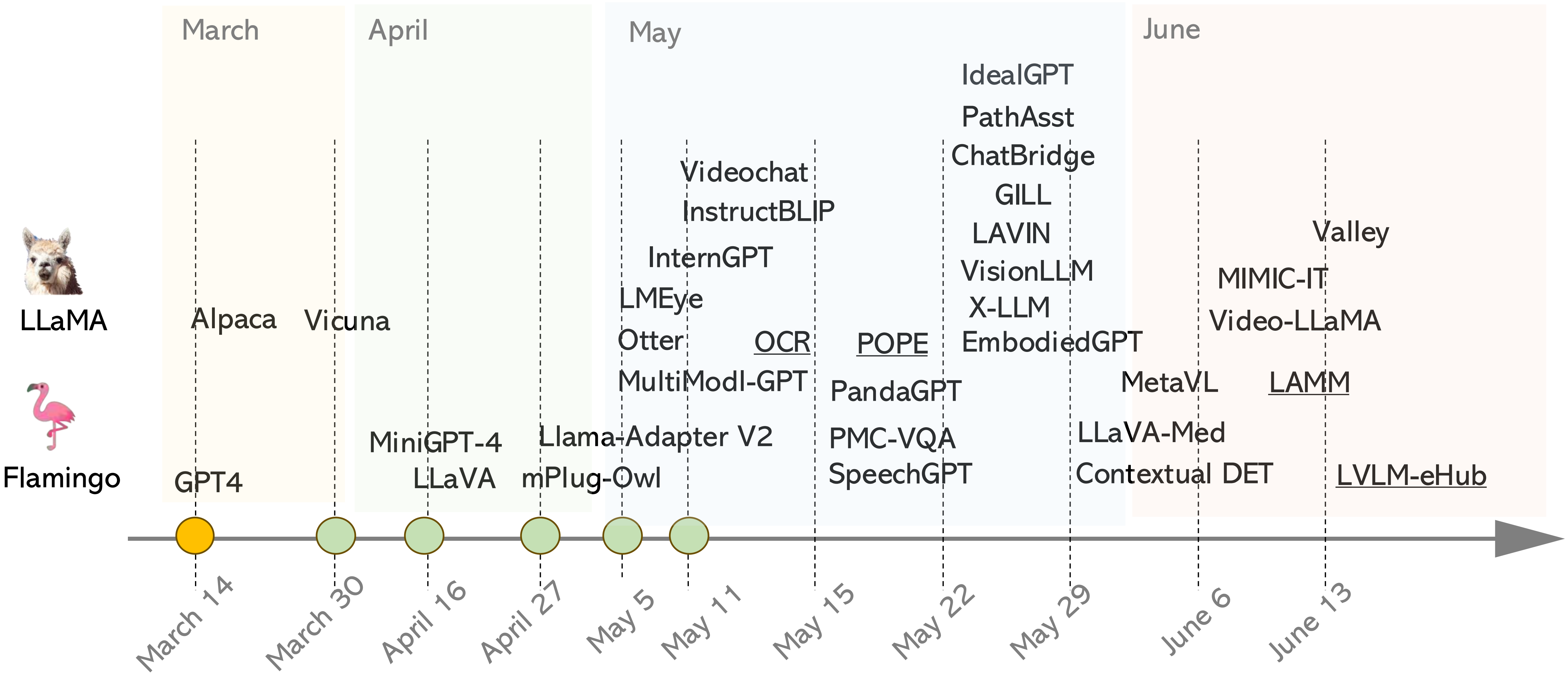
Example 3: LLaVA



what is written in the photo?

In the photo, there is written "Llava" which is presumably an abbreviation for a location or other identifier.

Emerging Topics in 2023



LLaMA

Flamingo

March

GPT4

Alpaca

Vicuna

April

MiniGPT-4

LLaVA

Llama-Adapter V2

mPlug-Owl

May

Otter

MultiModl-GPT

LMEye

InternGPT

Videochat

InstructBLIP

OCR

POPE

PandaGPT

PMC-VQA

SpeechGPT

IdealGPT

PathAsst

ChatBridge

GILL

LAVIN

VisionLLM

X-LLM

EmbodiedGPT

LLaVA-Med

Contextual DET

June

MetaVL

MIMIC-IT

Video-LLaMA

Valley

LAMM

LVLM-eHub

March 14

March 30

April 16

April 27

May 5

May 11

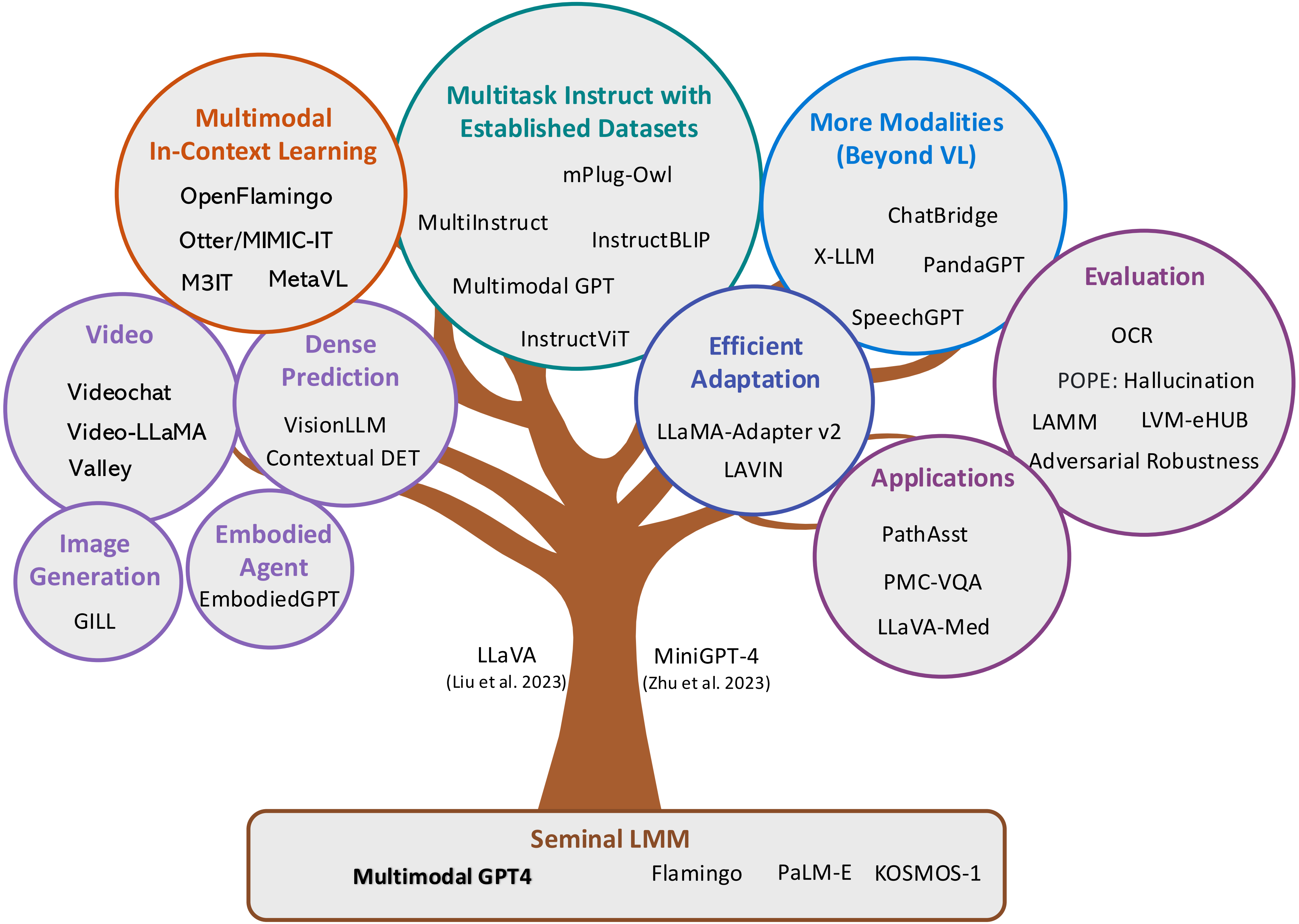
May 15

May 22

May 29

June 6

June 13


Multimodal In-Context Learning
OpenFlamingo
Otter/MIMIC-IT
M3IT
MetaVL
Multitask Instruct with Established Datasets
mPlug-Owl
MultiInstruct
InstructBLIP
Multimodal GPT
InstructViT
More Modalities (Beyond VL)
ChatBridge
X-LLM
PandaGPT
SpeechGPT
Evaluation
OCR
POPE: Hallucination
LAMM
LVM-eHUB
Adversarial Robustness
Video
Videochat
Video-LLaMA
Valley
Dense Prediction
VisionLLM
Contextual DET
Efficient Adaptation
LLaMA-Adapter v2
LAVIN
Applications
PathAsst
PMC-VQA
LLaVA-Med
Image Generation
GILL
Embodied Agent
EmbodiedGPT
LLaVA
(Liu et al. 2023)
MiniGPT-4
(Zhu et al. 2023)
Seminal LMM
Multimodal GPT4
Flamingo
PaLM-E
KOSMOS-1

③

LLaVA Family

2023 & Beyond

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

Large Language and Vision Assistant

LLaVA (llava-vl.github.io)

LLaVA is the first open-source project to build GPT-4V like model, inspiring dozens of projects
- 1.7K+ citations, and 17.5K+ GitHub stars, in ~14 months
- A cost-efficient open-source recipe to GPT-4V (1 day on 8 A-100 GPUs for a 7B model)

LLaVA

April 17

LLaVA-Med

June 1

LLaVA-RLHF

RLHF

Sep 26

LLaVA-1.5

New SoTA

Oct 5

LLaVA-Interactive

Oct

LLaVA-Plus

Tool Use

Nov

LLaVA-NeXT

New SoTA

Jan

LLaVA-NeXT

Stronger

Ablations

April

LLaVA-NeXT

Video

SoTA Video

April

LLaVA-NeXT

Interleave

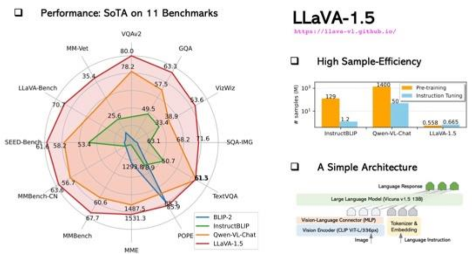
Multi-image

June

Performance

- **LLaVA-1.5**
- **LLaVA-NeXT**

New SoTA among open LMMs



❑ Performance: SoTA on 11 Benchmarks

LLaVA-1.5

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

❑ High Sample-Efficiency

❑ A Simple Architecture

Applications

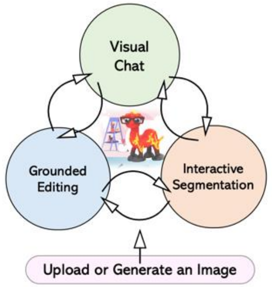
- **LLaVA-Med**
- **LLaVA-Rad**

The first open-source healthcare LMM

New Capabilities

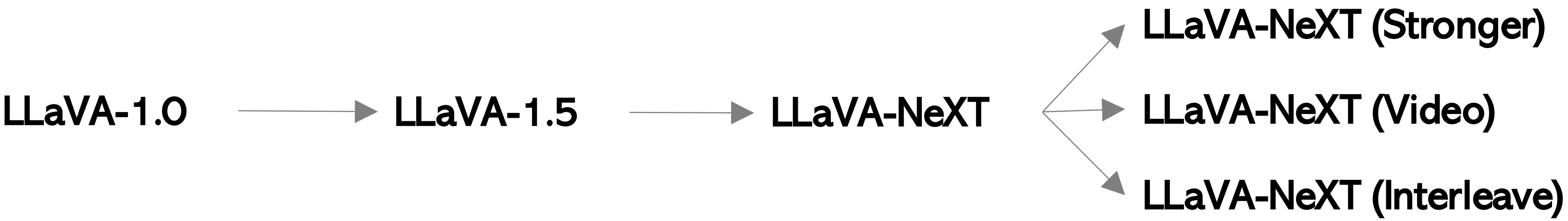
- **LLaVA Interactive**
- **LLaVA-Plus**

Tool use to expand capabilities



LLaVA Series:

3.3 Improved Performance


LLaVA-1.0 → LLaVA-1.5 → LLaVA-NeXT

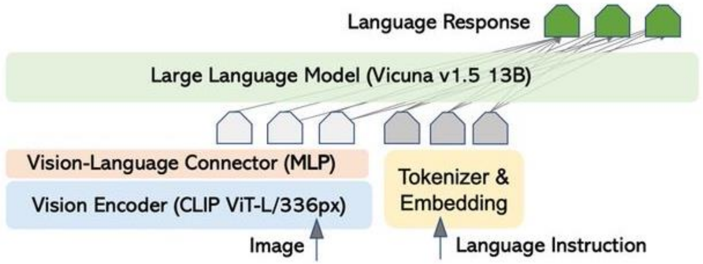
LLaVA-NeXT (Stronger)

LLaVA-NeXT (Video)

LLaVA-NeXT (Interleave)

LLaVA-1.5: Improved Baselines with Visual Instruction Tuning

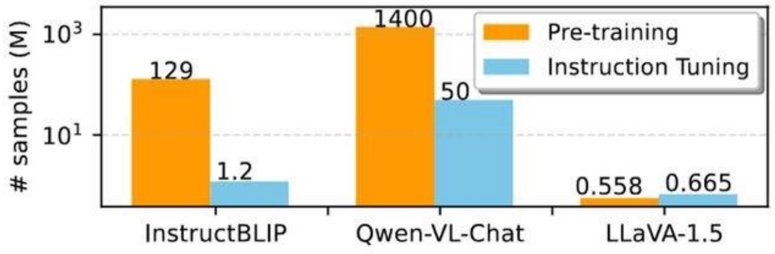
❑ A Simple Architecture


Language Response
Large Language Model (Vicuna v1.5 13B)
Vision-Language Connector (MLP)
Vision Encoder (CLIP ViT-L/336px)
Tokenizer & Embedding
Image
Language Instruction

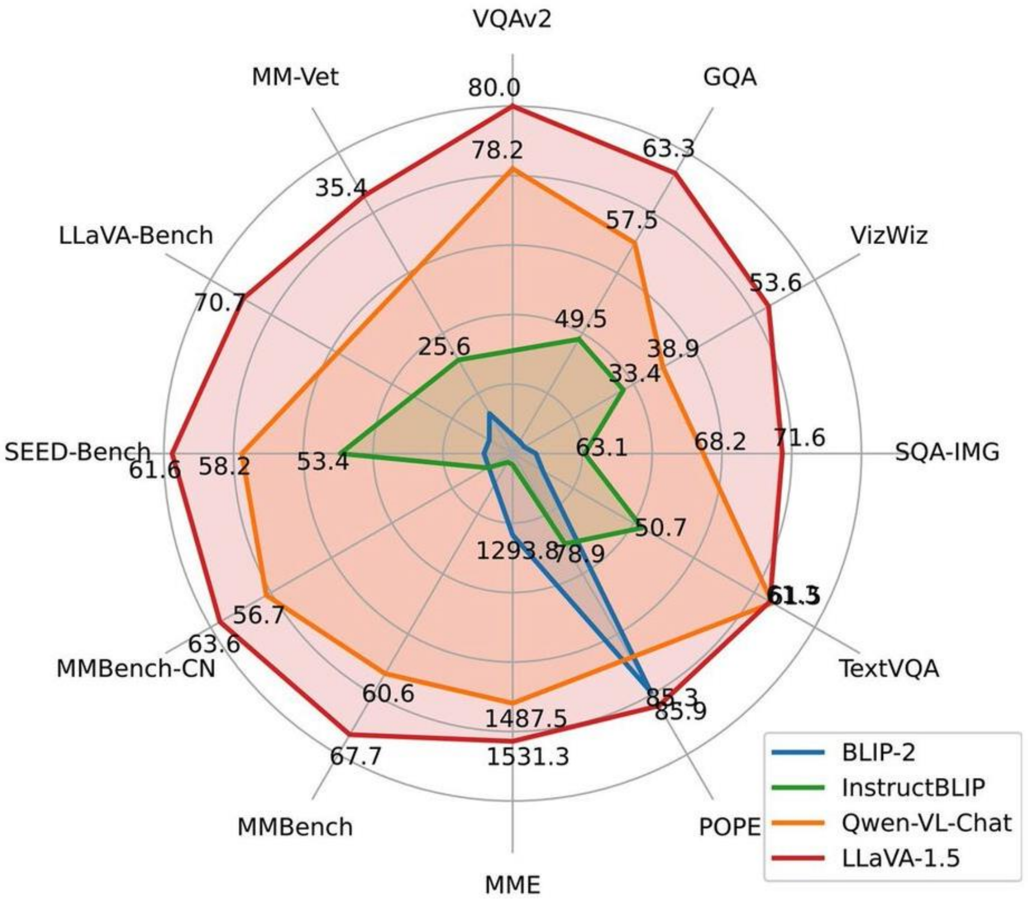
❑ Data Mixture

- Academic task-oriented data
- Response formatting prompts

❑ High Sample-Efficiency



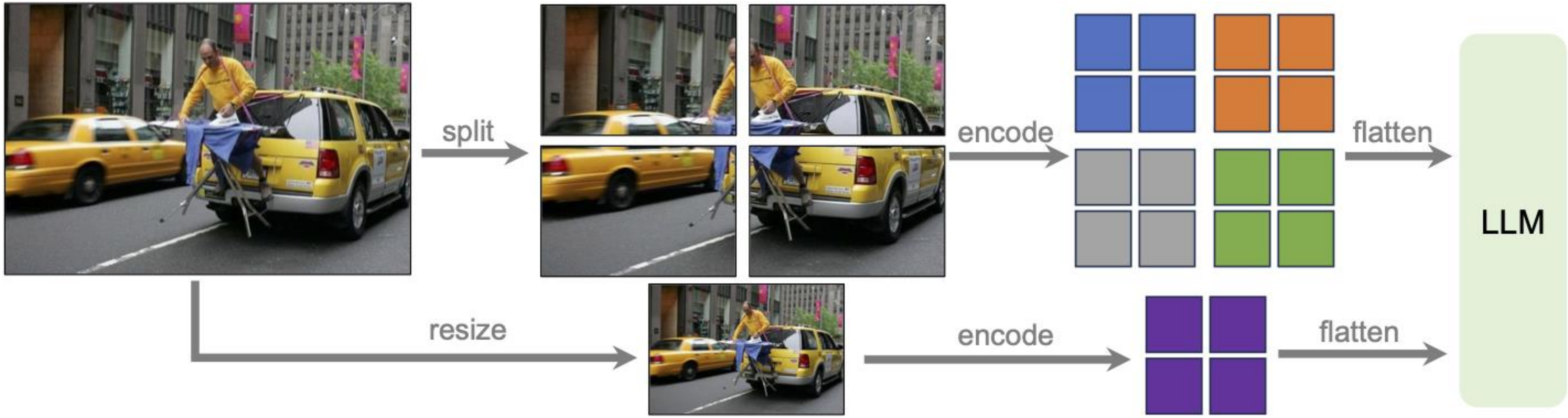
❑ Performance: SoTA on 11 Benchmarks



LLaVA-NeXT: Improved reasoning, OCR, and world knowledge

https://llava-vl.github.io/blog/2024-01-30-llava-next/

(1) AnyRes: Dynamic High Resolution



LLaVA-NeXT: Improved reasoning, OCR, and world knowledge

https://llava-vl.github.io/blog/2024-01-30-llava-next/

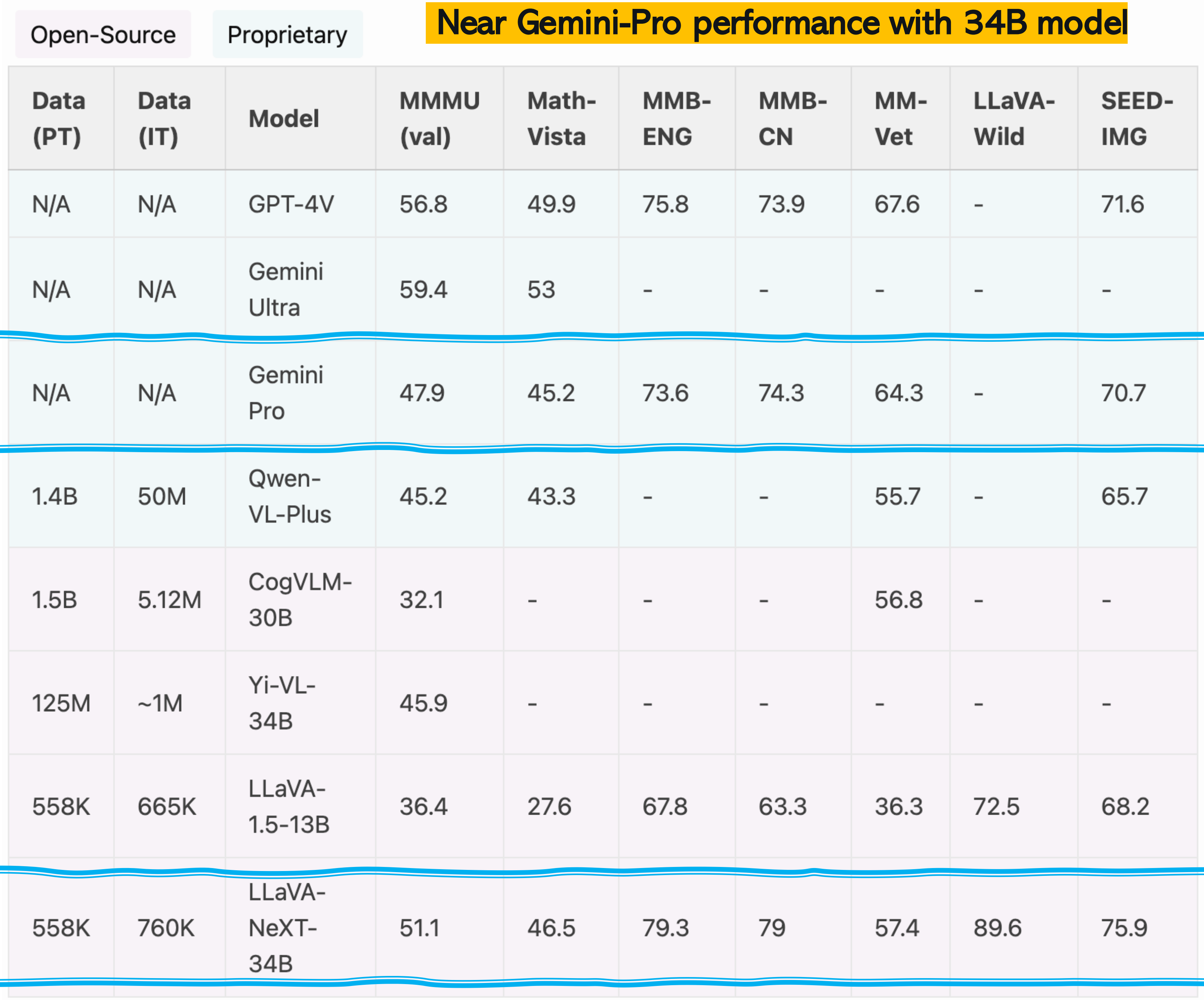
(2) Data Mixture

- High-quality User Instruct Data.
- Multimodal Document/Chart Data.

(3) Scaling LLM backbone

- Mistral-7B, Vicuna-7B & 13B, Yi-34B

Near Gemini-Pro performance with 34B model



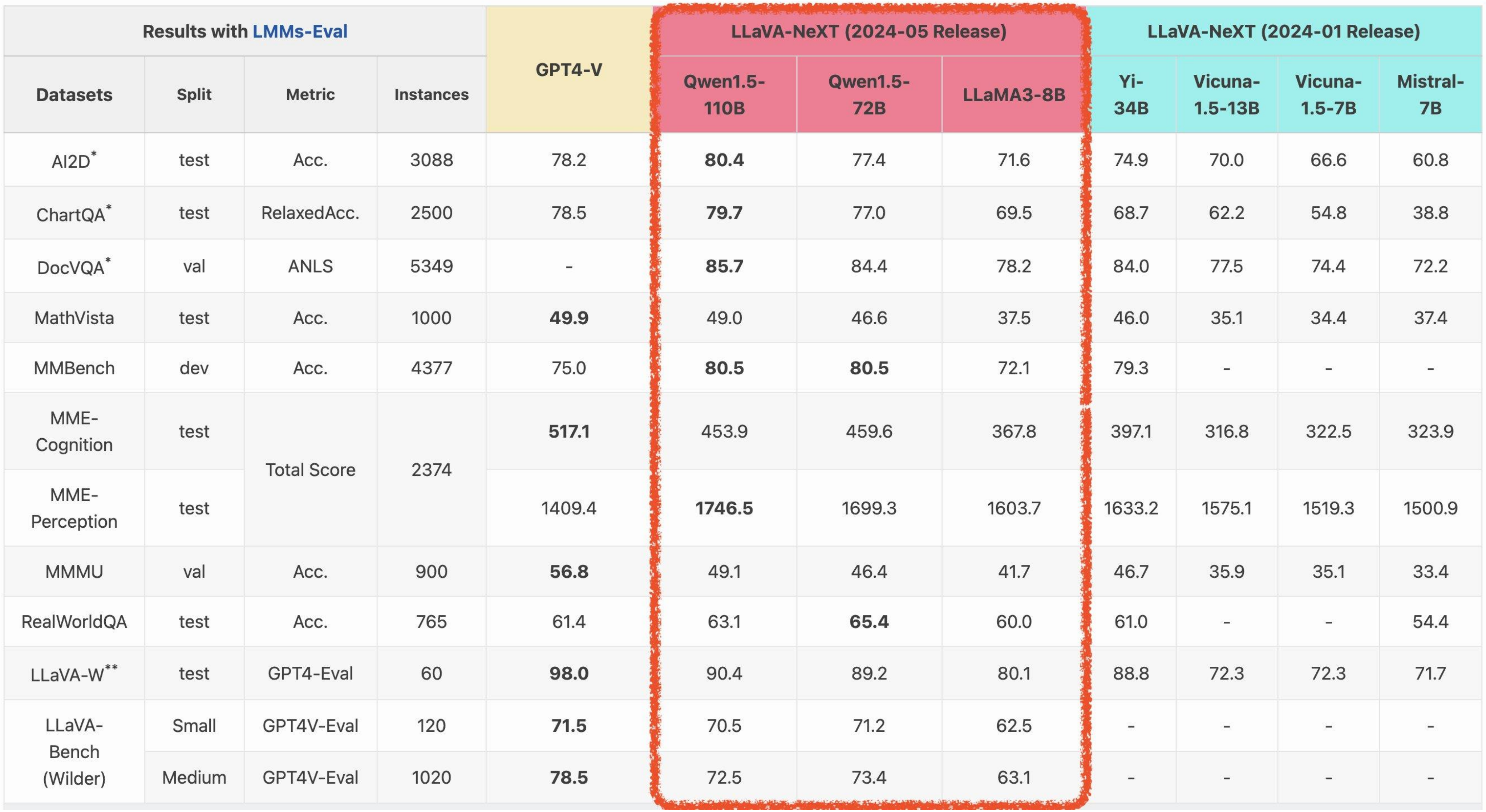
Open-Source Proprietary

Data (PT)	Data (IT)	Model	MMMU (val)	Math-Vista	MMB-ENG	MMB-CN	MM-Vet	LLaVA-Wild	SEED-IMG
N/A	N/A	GPT-4V	56.8	49.9	75.8	73.9	67.6	-	71.6
N/A	N/A	Gemini Ultra	59.4	53	-	-	-	-	-
N/A	N/A	Gemini Pro	47.9	45.2	73.6	74.3	64.3	-	70.7
1.4B	50M	Qwen-VL-Plus	45.2	43.3	-	-	55.7	-	65.7
1.5B	5.12M	CogVLM-30B	32.1	-	-	-	56.8	-	-
125M	~1M	Yi-VL-34B	45.9	-	-	-	-	-	-
558K	665K	LLaVA-1.5-13B	36.4	27.6	67.8	63.3	36.3	72.5	68.2
558K	760K	LLaVA-NeXT-34B	51.1	46.5	79.3	79	57.4	89.6	75.9

Jan 2024

~24 hours with 32 A100

Near GPT4-V performance with 110B model



Results with LMMs-Eval				GPT4-V	LLaVA-NeXT (2024-05 Release)			LLaVA-NeXT (2024-01 Release)			
Datasets	**Split**	**Metric**	**Instances**		**Qwen1.5-110B**	**Qwen1.5-72B**	**LLaMA3-8B**	**Yi-34B**	**Vicuna-1.5-13B**	**Vicuna-1.5-7B**	**Mistral-7B**
AI2D*	test	Acc.	3088	78.2	**80.4**	77.4	71.6	74.9	70.0	66.6	60.8
ChartQA*	test	RelaxedAcc.	2500	78.5	**79.7**	77.0	69.5	68.7	62.2	54.8	38.8
DocVQA*	val	ANLS	5349	-	**85.7**	84.4	78.2	84.0	77.5	74.4	72.2
MathVista	test	Acc.	1000	**49.9**	49.0	46.6	37.5	46.0	35.1	34.4	37.4
MMBench	dev	Acc.	4377	75.0	**80.5**	**80.5**	72.1	79.3	-	-	-
MME-Cognition	test	Total Score	2374	**517.1**	453.9	459.6	367.8	397.1	316.8	322.5	323.9
MME-Perception	test			1409.4	**1746.5**	1699.3	1603.7	1633.2	1575.1	1519.3	1500.9
MMMU	val	Acc.	900	**56.8**	49.1	46.4	41.7	46.7	35.9	35.1	33.4
RealWorldQA	test	Acc.	765	61.4	63.1	**65.4**	60.0	61.0	-	-	54.4
LLaVA-W**	test	GPT4-Eval	60	**98.0**	90.4	89.2	80.1	88.8	72.3	72.3	71.7
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	-	-	-	-

~18 hours
with 128 H800

~24 hours
with 32 A100

https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/

Stronger LLMs Supercharge Multimodal Capabilities in the Wild

- **Improved Language Capability:**

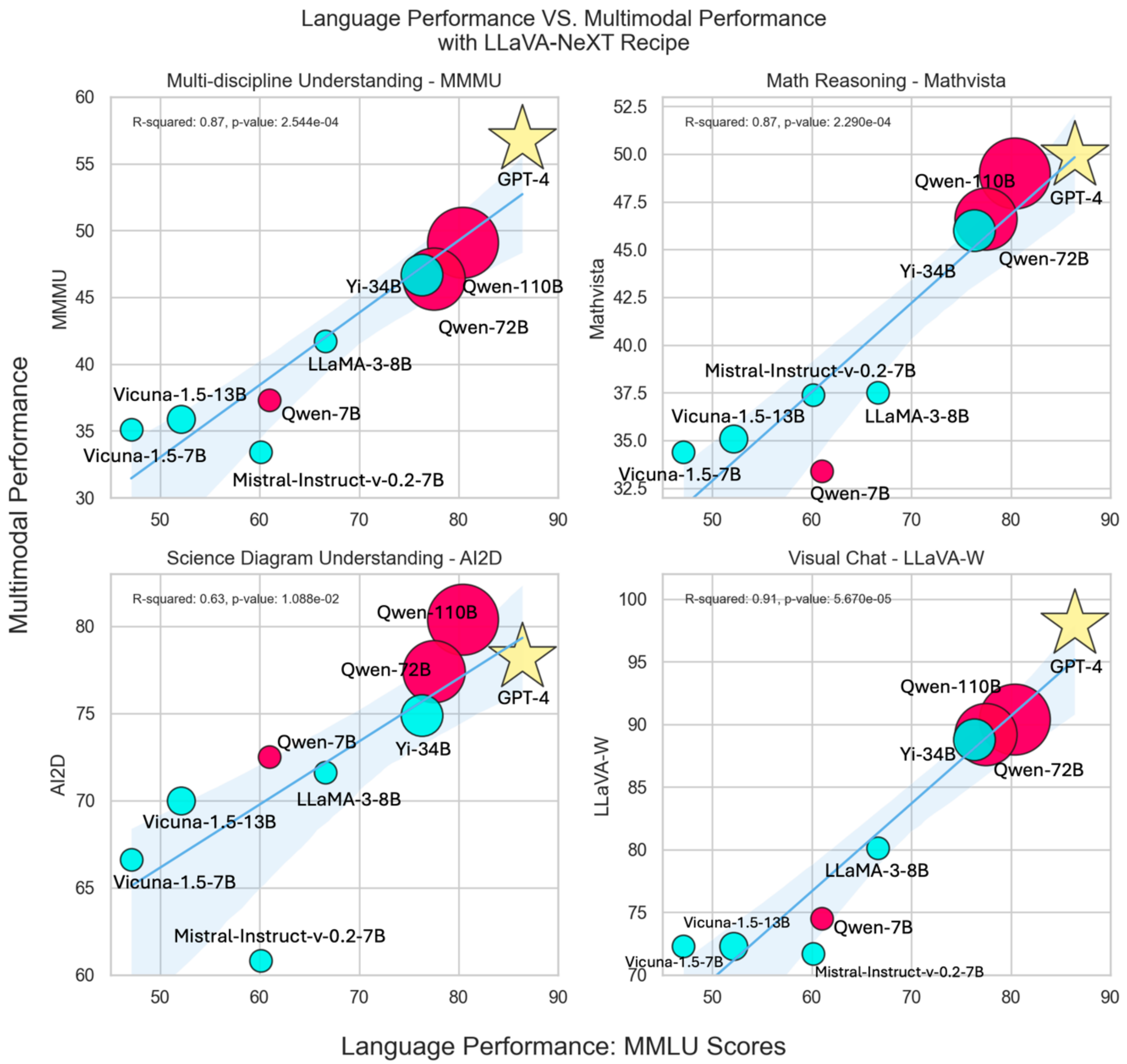
Across LLMs of comparable sizes (e.g., 7B Mistral/Vicuna, 7B Qwen, 8B LLaMa3)

Higher language proficiency corresponds to improved multimodal capabilities.

- **Influence of Model Size:**

Within the same LLM family (e.g., Qwen LLM: 7B, 72B, 110B)

Larger models consistently demonstrate superior performance on multimodal benchmarks



https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/

LLaVA-NeXT (Stronger)

What Else Influences Visual Instruction Tuning Beyond Data?

https://llava-vl.github.io/blog/2024-05-25-llava-next-ablations/

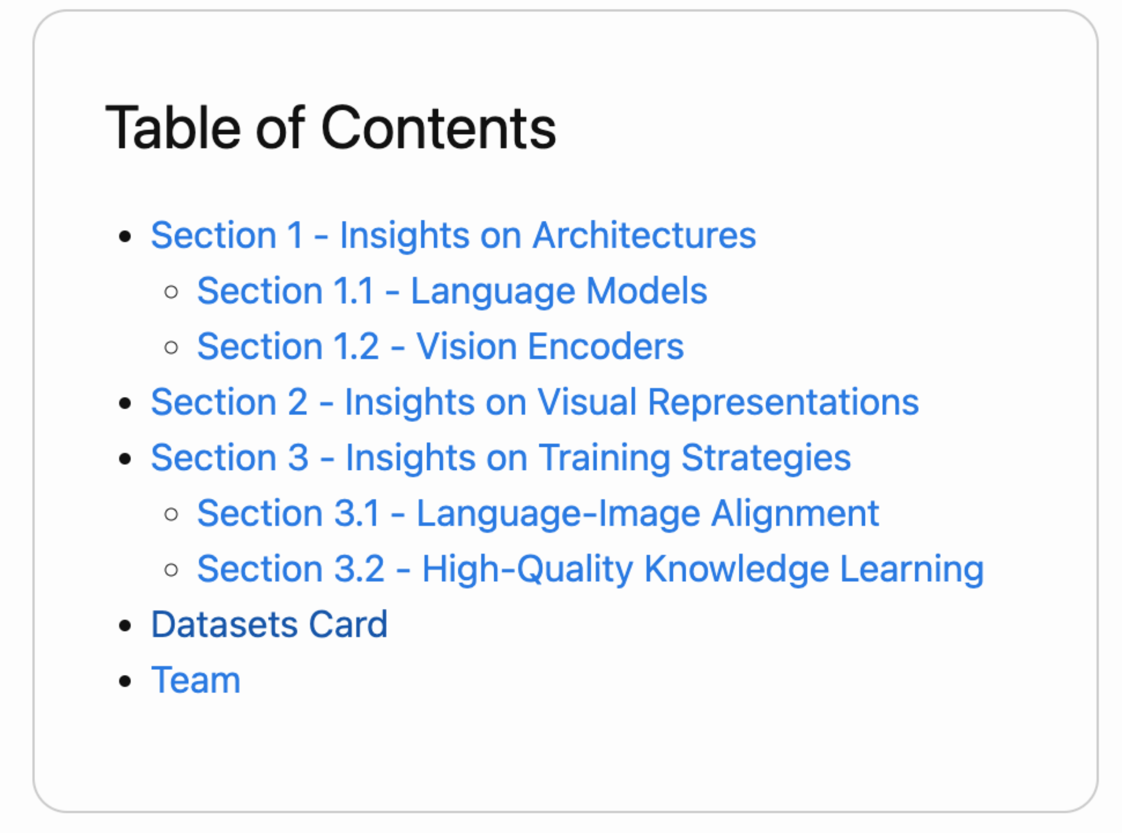
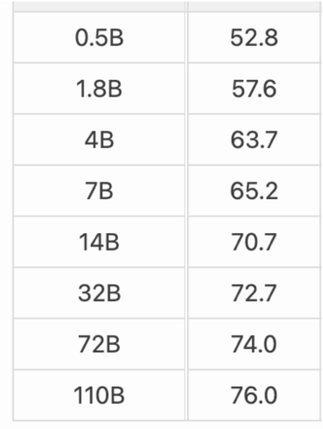


Table of Contents

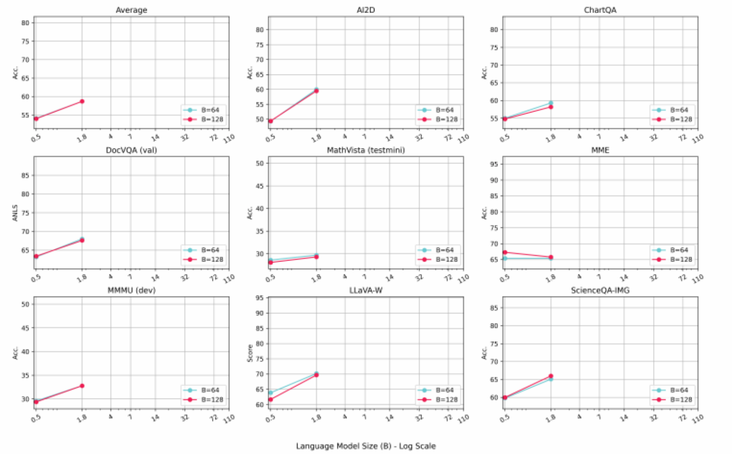
- Section 1 - Insights on Architectures
  - Section 1.1 - Language Models
  - Section 1.2 - Vision Encoders
- Section 2 - Insights on Visual Representations
- Section 3 - Insights on Training Strategies
  - Section 3.1 - Language-Image Alignment
  - Section 3.2 - High-Quality Knowledge Learning
- Datasets Card
- Team

Insight 1: Architectures (LLM, Vision Encoder)

❑ Scaling Language Models (0.5B – 110B)

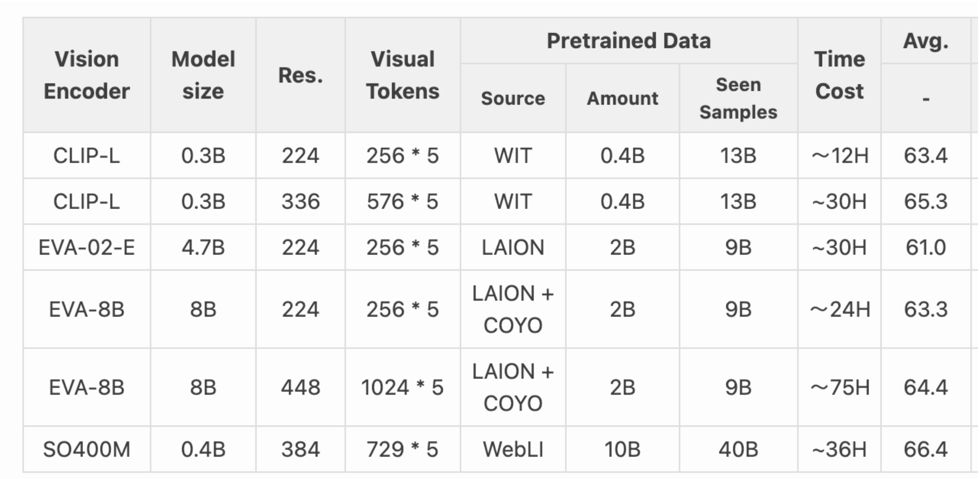


LLM	Average
0.5B	52.8
1.8B	57.6
4B	63.7
7B	65.2
14B	70.7
32B	72.7
72B	74.0
110B	76.0



Model size scaling is effective

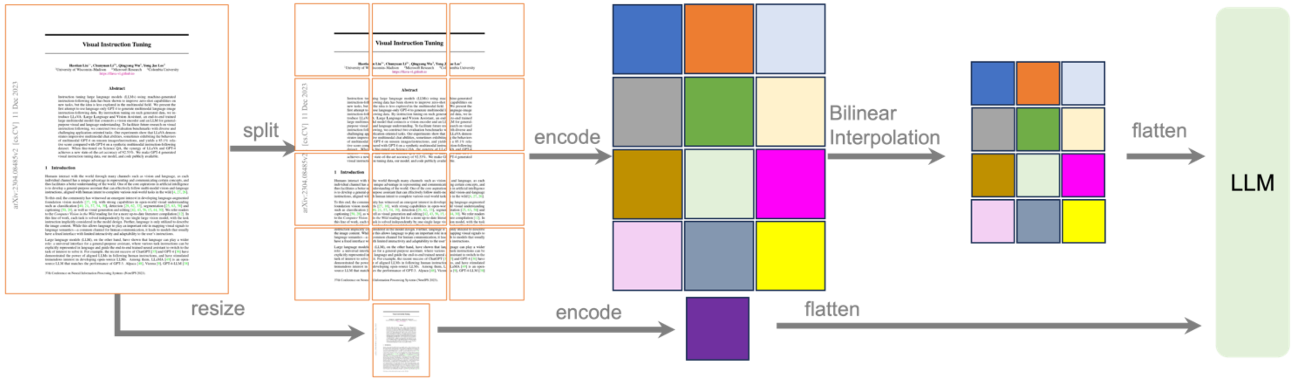
❑ Scaling Vision Encoders (0.3B – 10B)



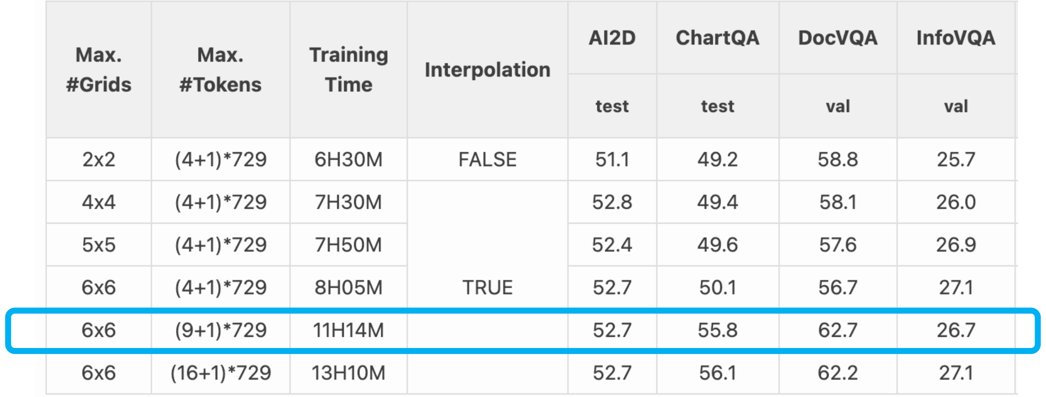
Vision Encoder	Model size	Res.	Visual Tokens	Pretrained Data			Time Cost	Avg.
				Source	Amount	Seen Samples		-
CLIP-L	0.3B	224	256 * 5	WIT	0.4B	13B	~12H	63.4
CLIP-L	0.3B	336	576 * 5	WIT	0.4B	13B	~30H	65.3
EVA-02-E	4.7B	224	256 * 5	LAION	2B	9B	~30H	61.0
EVA-8B	8B	224	256 * 5	LAION + COYO	2B	9B	~24H	63.3
EVA-8B	8B	448	1024 * 5	LAION + COYO	2B	9B	~75H	64.4
SO400M	0.4B	384	729 * 5	WebLI	10B	40B	~36H	66.4

Model size scaling is less effective than others (training data, visual representation)

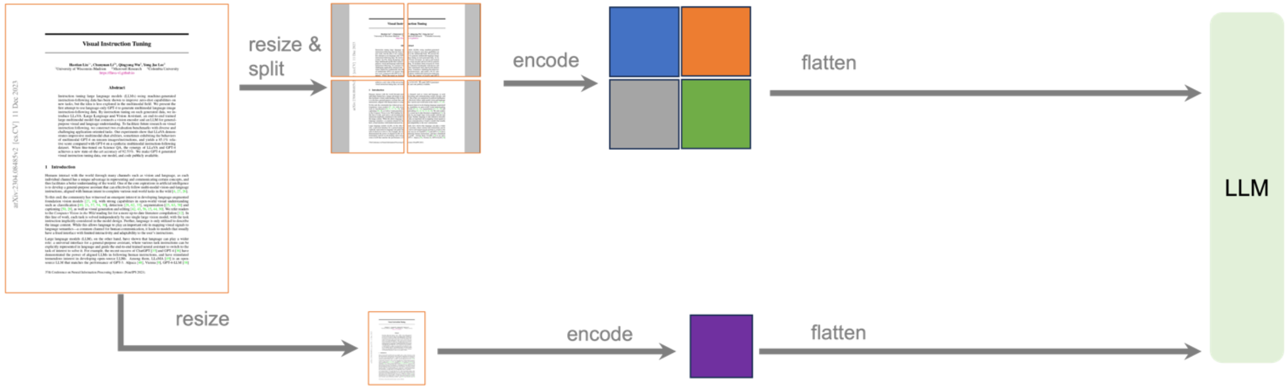
Insight 2: Visual Representations (Resolution, #Tokens)



(a) Higher-AnyRes with Bilinear Interpolation



Max. #Grids	Max. #Tokens	Training Time	Interpolation	AI2D	ChartQA	DocVQA	InfoVQA
				test	test	val	val
2x2	(4+1)*729	6H30M	FALSE	51.1	49.2	58.8	25.7
4x4	(4+1)*729	7H30M	TRUE	52.8	49.4	58.1	26.0
5x5	(4+1)*729	7H50M		52.4	49.6	57.6	26.9
6x6	(4+1)*729	8H05M		52.7	50.1	56.7	27.1
6x6	(9+1)*729	11H14M		52.7	55.8	62.7	26.7
6x6	(16+1)*729	13H10M		52.7	56.1	62.2	27.1

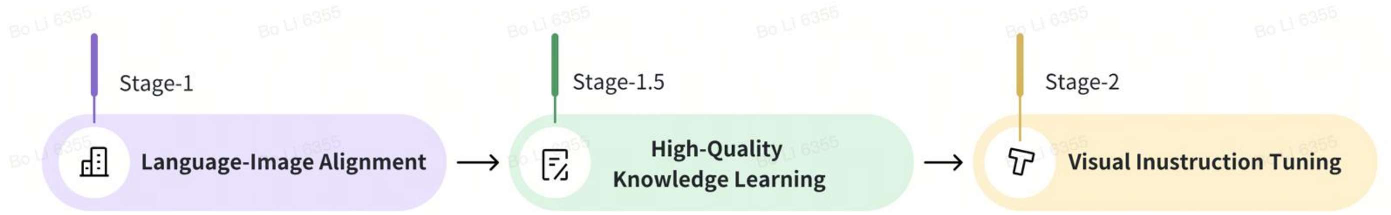


(b) The original AnyRes

- Scaling up both (resolution, #tokens)
- The best configuration for a balance of performance and cost

Insight 3: Training Strategies

(Trainable Module, Data)



Bo Li 6355

Trainable Module

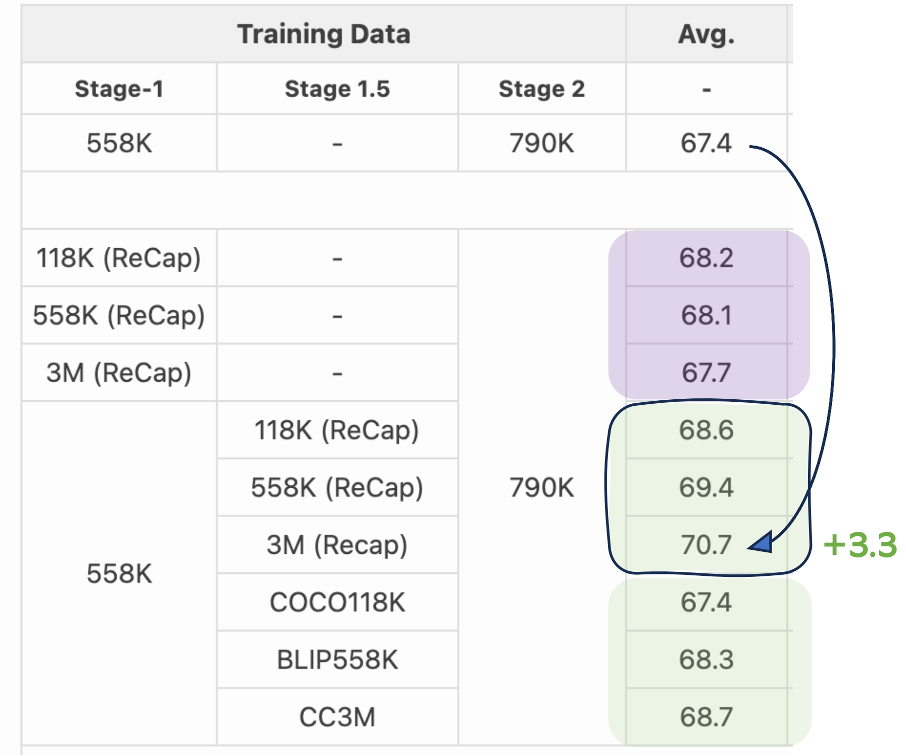
Projector

Full Model

Full Model

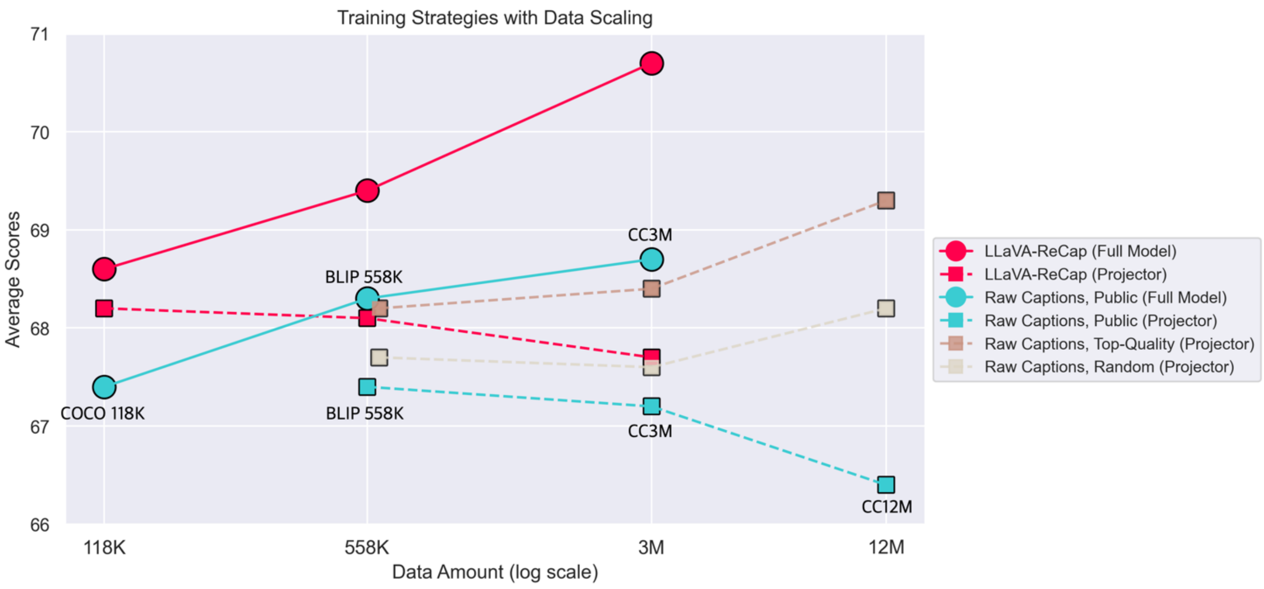
- The best training configuration (Full model, High-quality recap data)

Data



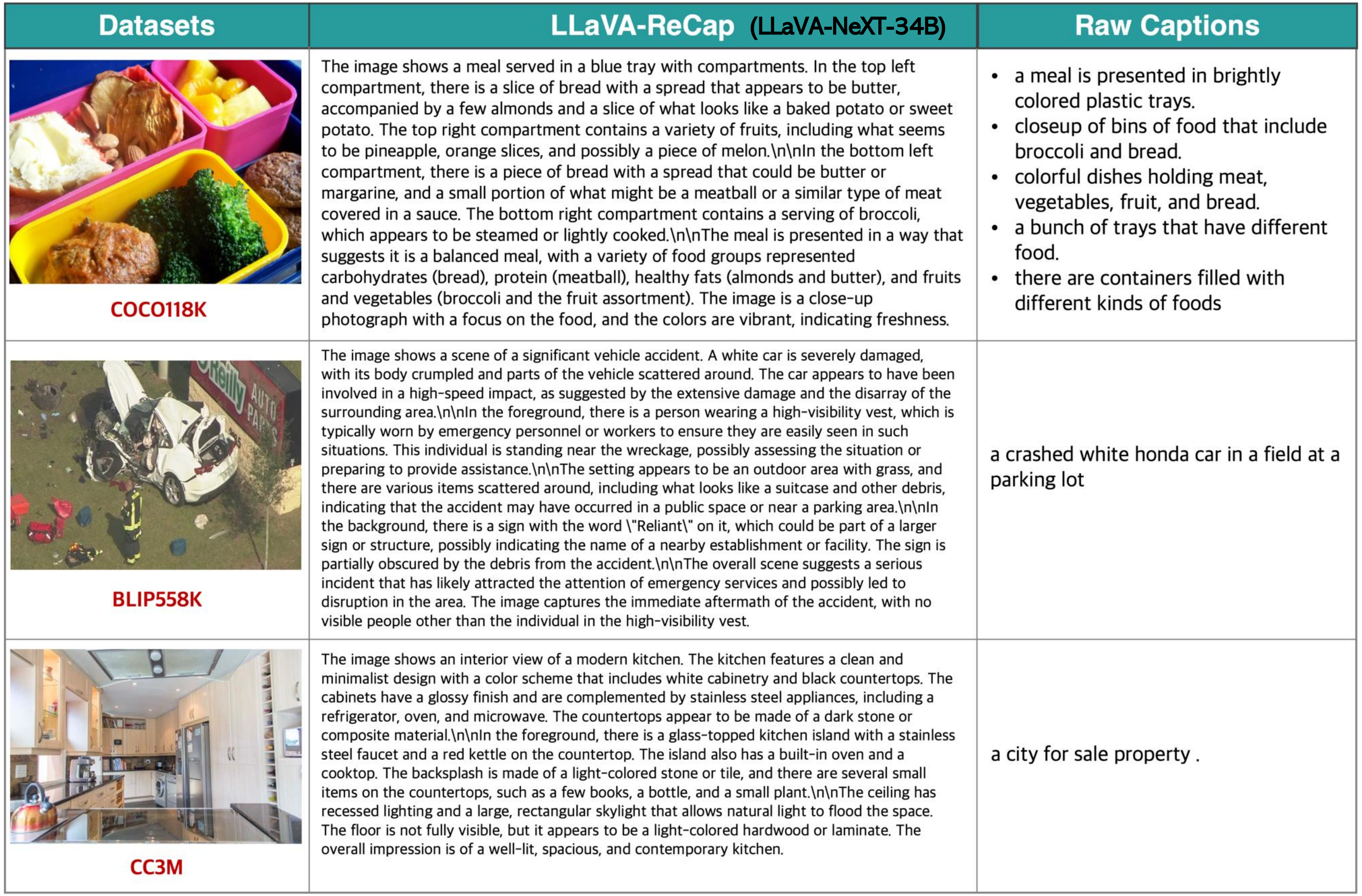
Training Data			Avg.
Stage-1	Stage 1.5	Stage 2	-
558K	-	790K	67.4
118K (ReCap)	-	790K	68.2
558K (ReCap)	-		68.1
3M (ReCap)	-		67.7
558K	118K (ReCap)		68.6
	558K (ReCap)		69.4
	3M (Recap)		70.7
	COCO118K		67.4
	BLIP558K		68.3
	CC3M		68.7

+3.3



Open-Source Recap High-Quality Data

https://llava-vl.github.io/blog/2024-05-25-llava-next-ablations/



Datasets	LLaVA-ReCap (LLaVA-NeXT-34B)	Raw Captions
COCO118K	The image shows a meal served in a blue tray with compartments. In the top left compartment, there is a slice of bread with a spread that appears to be butter, accompanied by a few almonds and a slice of what looks like a baked potato or sweet potato. The top right compartment contains a variety of fruits, including what seems to be pineapple, orange slices, and possibly a piece of melon.\n\nIn the bottom left compartment, there is a piece of bread with a spread that could be butter or margarine, and a small portion of what might be a meatball or a similar type of meat covered in a sauce. The bottom right compartment contains a serving of broccoli, which appears to be steamed or lightly cooked.\n\nThe meal is presented in a way that suggests it is a balanced meal, with a variety of food groups represented carbohydrates (bread), protein (meatball), healthy fats (almonds and butter), and fruits and vegetables (broccoli and the fruit assortment). The image is a close-up photograph with a focus on the food, and the colors are vibrant, indicating freshness.	• a meal is presented in brightly colored plastic trays. • closeup of bins of food that include broccoli and bread. • colorful dishes holding meat, vegetables, fruit, and bread. • a bunch of trays that have different food. • there are containers filled with different kinds of foods
BLIP558K	The image shows a scene of a significant vehicle accident. A white car is severely damaged, with its body crumpled and parts of the vehicle scattered around. The car appears to have been involved in a high-speed impact, as suggested by the extensive damage and the disarray of the surrounding area.\n\nIn the foreground, there is a person wearing a high-visibility vest, which is typically worn by emergency personnel or workers to ensure they are easily seen in such situations. This individual is standing near the wreckage, possibly assessing the situation or preparing to provide assistance.\n\nThe setting appears to be an outdoor area with grass, and there are various items scattered around, including what looks like a suitcase and other debris, indicating that the accident may have occurred in a public space or near a parking area.\n\nIn the background, there is a sign with the word \"Reliant\" on it, which could be part of a larger sign or structure, possibly indicating the name of a nearby establishment or facility. The sign is partially obscured by the debris from the accident.\n\nThe overall scene suggests a serious incident that has likely attracted the attention of emergency services and possibly led to disruption in the area. The image captures the immediate aftermath of the accident, with no visible people other than the individual in the high-visibility vest.	a crashed white honda car in a field at a parking lot
CC3M	The image shows an interior view of a modern kitchen. The kitchen features a clean and minimalist design with a color scheme that includes white cabinetry and black countertops. The cabinets have a glossy finish and are complemented by stainless steel appliances, including a refrigerator, oven, and microwave. The countertops appear to be made of a dark stone or composite material.\n\nIn the foreground, there is a glass-topped kitchen island with a stainless steel faucet and a red kettle on the countertop. The island also has a built-in oven and a cooktop. The backsplash is made of a light-colored stone or tile, and there are several small items on the countertops, such as a few books, a bottle, and a small plant.\n\nThe ceiling has recessed lighting and a large, rectangular skylight that allows natural light to flood the space. The floor is not fully visible, but it appears to be a light-colored hardwood or laminate. The overall impression is of a well-lit, spacious, and contemporary kitchen.	a city for sale property .

Public Available

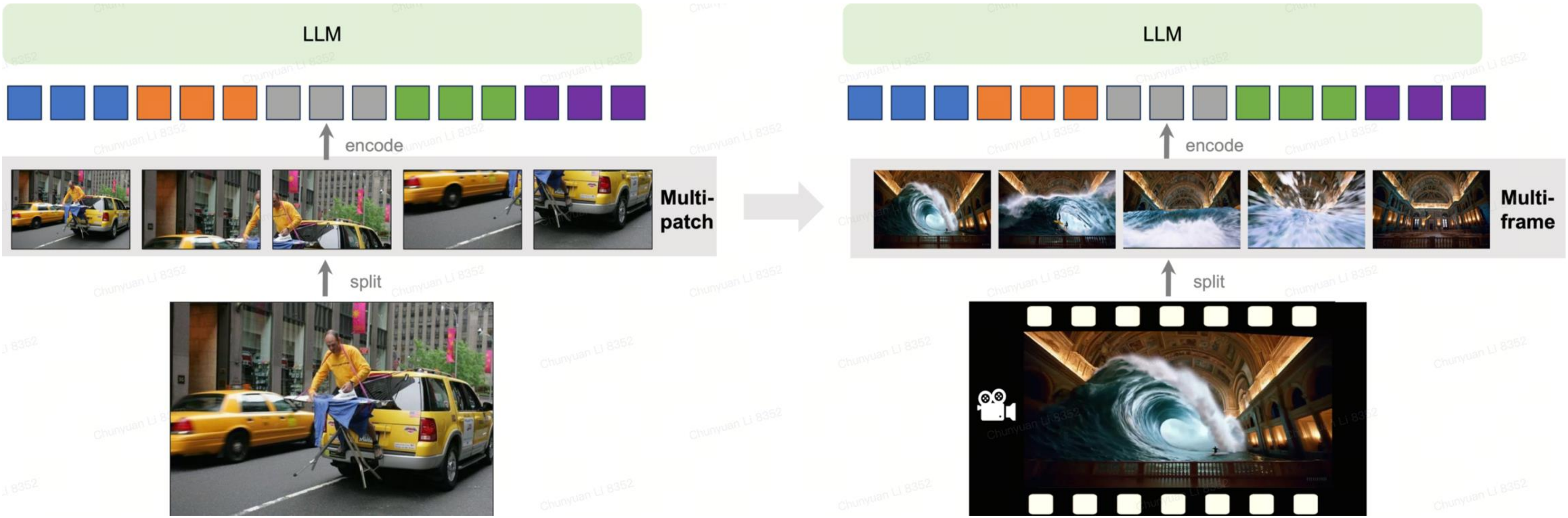
Hugging Face

LLaVA-NeXT (Video):

A Strong Zero-Shot Video Understanding Model

https://llava-vl.github.io/blog/2024-05-25-llava-next-ablations/

(1) AnyRes: From multi-patch to multi-frame



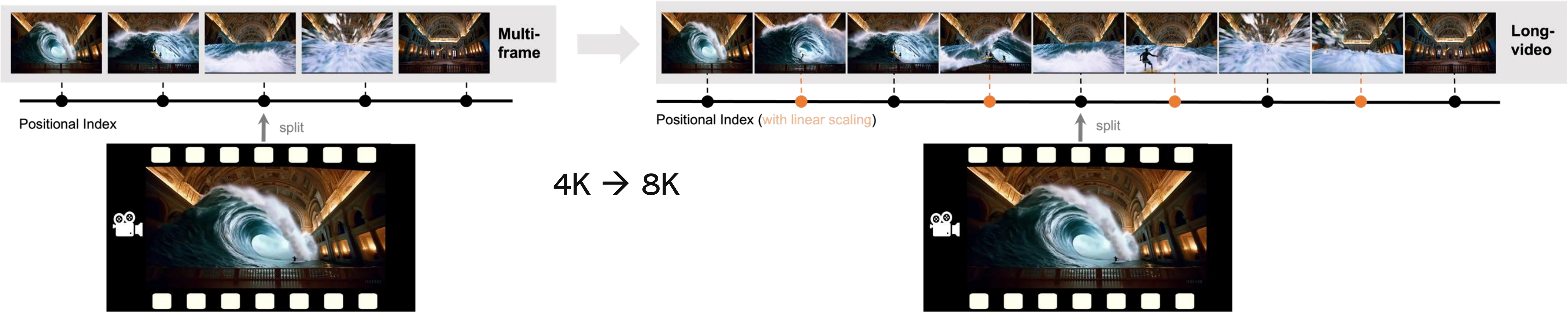
Chunyuan Li 8352

LLaVA-NeXT (Video):

A Strong Zero-Shot Video Understanding Model

https://llava-vl.github.io/blog/2024-05-25-llava-next-ablations/

(2) Length generalization: From multi-frame to long-video.



4K → 8K

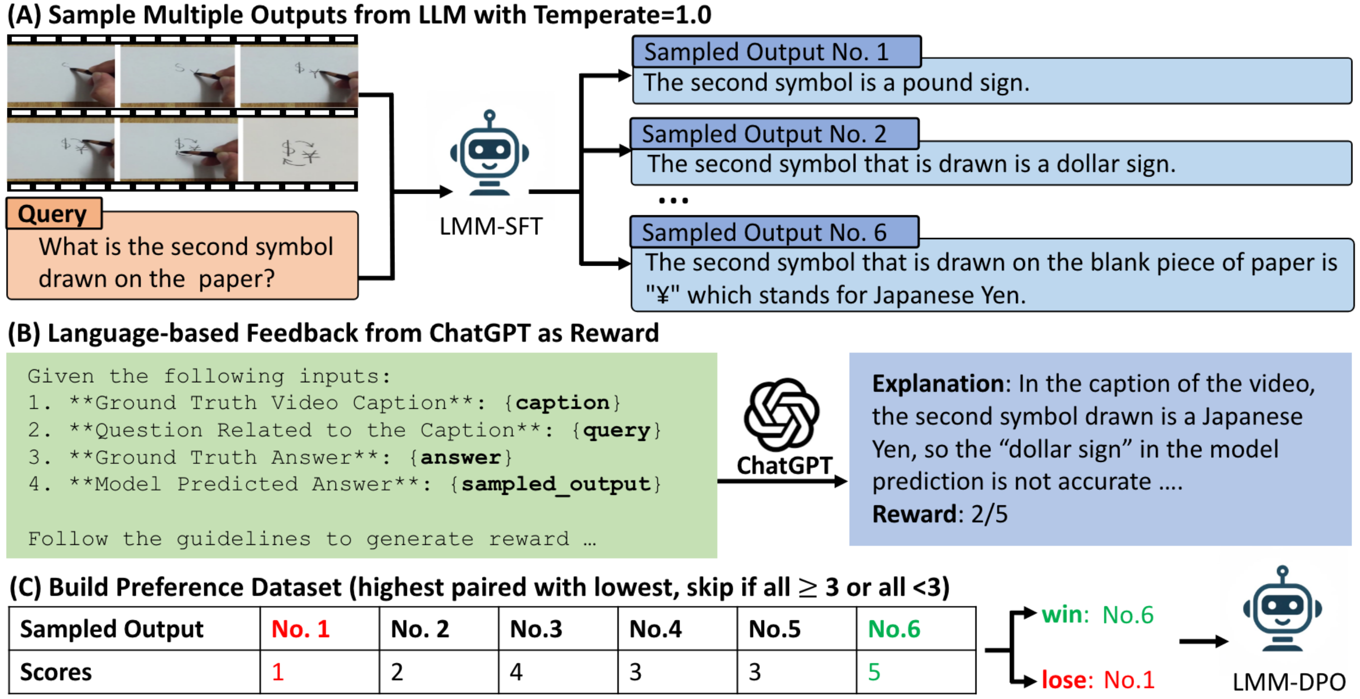
LLaVA-NeXT (Video):

A Strong Zero-Shot Video Understanding Model

https://llava-vl.github.io/blog/2024-05-25-llava-next-ablations/

(3) Direct Preference Optimization (DPO)

- AI Feedback from ChatGPT is collected
- Only the most confident pars are considered



(A) Sample Multiple Outputs from LLM with Temperate=1.0

Query
What is the second symbol drawn on the paper?

Sampled Output No. 1
The second symbol is a pound sign.

Sampled Output No. 2
The second symbol that is drawn is a dollar sign.

…

Sampled Output No. 6
The second symbol that is drawn on the blank piece of paper is "¥" which stands for Japanese Yen.

(B) Language-based Feedback from ChatGPT as Reward

```
Given the following inputs:
1. **Ground Truth Video Caption**: {caption}
2. **Question Related to the Caption**: {query}
3. **Ground Truth Answer**: {answer}
4. **Model Predicted Answer**: {sampled_output}

Follow the guidelines to generate reward …
```

Explanation: In the caption of the video, the second symbol drawn is a Japanese Yen, so the “dollar sign” in the model prediction is not accurate ….
Reward: 2/5

(C) Build Preference Dataset (highest paired with lowest, skip if all ≥ 3 or all <3)

Sampled Output	**No. 1**	**No. 2**	**No.3**	**No.4**	**No.5**	**No.6**
Scores	1	2	4	3	3	5

win: No.6
lose: No.1

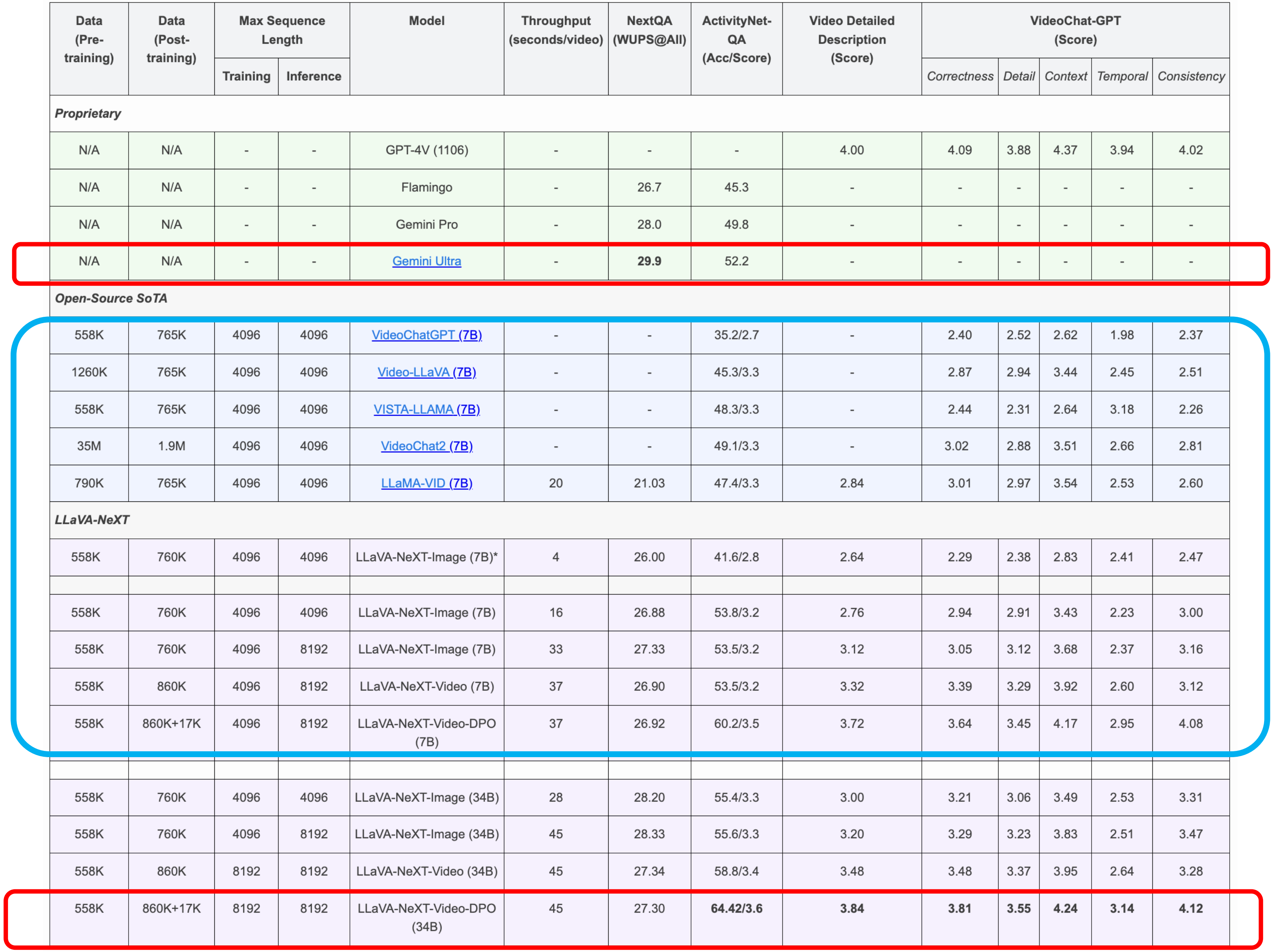
Direct Preference Optimization of Video Large Multimodal Models from Language Model Reward, 2024

7B

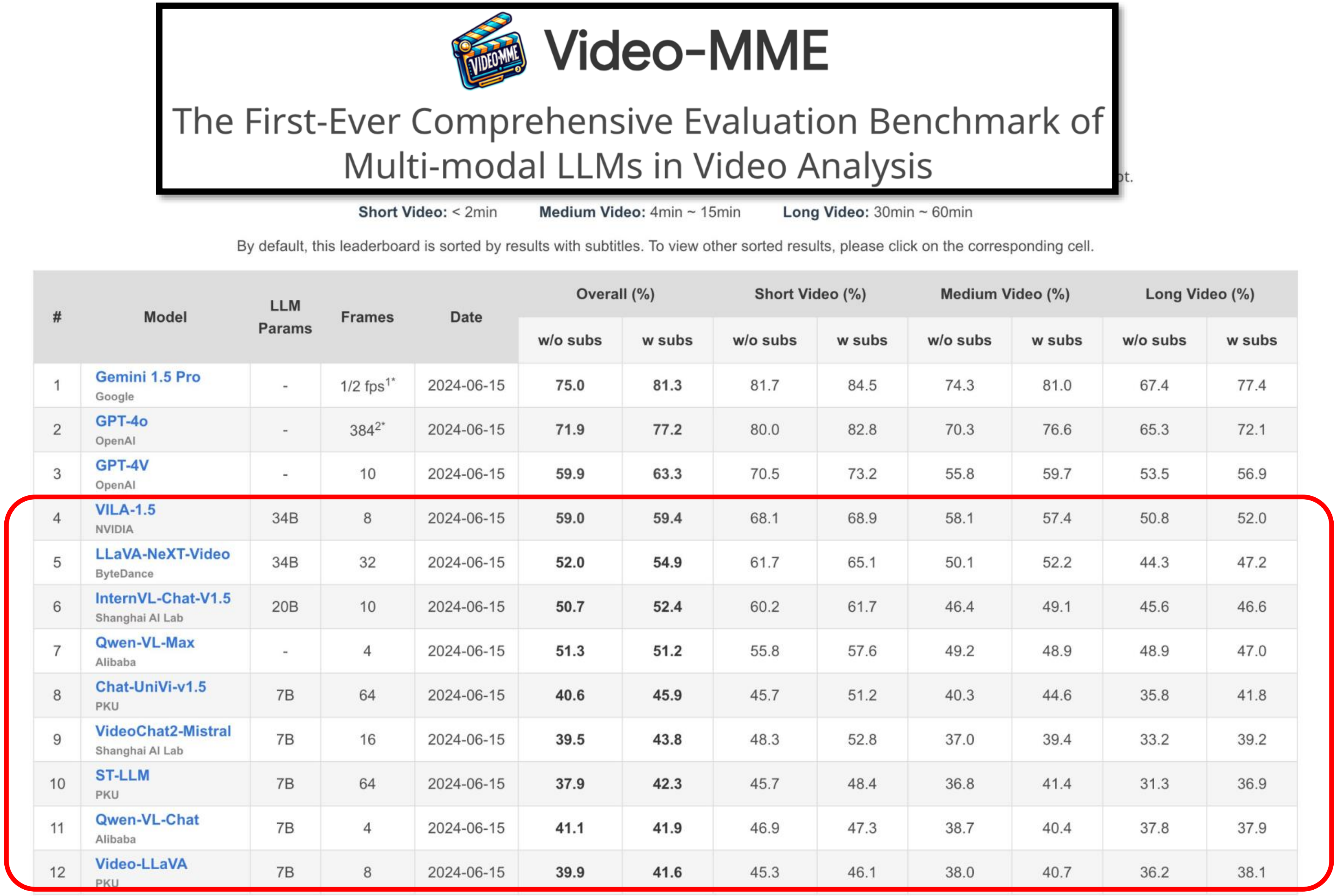
- Zero-shot of LLaVA-NeXT is better than prior Open-source SoTA
- DPO on video improves significantly

34B

LLaVA-NeXT catch up Gemini on selected benchmarks



Data (Pre-training)	Data (Post-training)	Max Sequence Length		Model	Throughput (seconds/video)	NextQA (WUPS@All)	ActivityNet-QA (Acc/Score)	Video Detailed Description (Score)	VideoChat-GPT (Score)				
		Training	Inference						*Correctness*	*Detail*	*Context*	*Temporal*	*Consistency*
Proprietary													
N/A	N/A	-	-	GPT-4V (1106)	-	-	-	4.00	4.09	3.88	4.37	3.94	4.02
N/A	N/A	-	-	Flamingo	-	26.7	45.3	-	-	-	-	-	-
N/A	N/A	-	-	Gemini Pro	-	28.0	49.8	-	-	-	-	-	-
N/A	N/A	-	-	Gemini Ultra	-	**29.9**	52.2	-	-	-	-	-	-
Open-Source SoTA													
558K	765K	4096	4096	VideoChatGPT (7B)	-	-	35.2/2.7	-	2.40	2.52	2.62	1.98	2.37
1260K	765K	4096	4096	Video-LLaVA (7B)	-	-	45.3/3.3	-	2.87	2.94	3.44	2.45	2.51
558K	765K	4096	4096	VISTA-LLAMA (7B)	-	-	48.3/3.3	-	2.44	2.31	2.64	3.18	2.26
35M	1.9M	4096	4096	VideoChat2 (7B)	-	-	49.1/3.3	-	3.02	2.88	3.51	2.66	2.81
790K	765K	4096	4096	LLaMA-VID (7B)	20	21.03	47.4/3.3	2.84	3.01	2.97	3.54	2.53	2.60
LLaVA-NeXT													
558K	760K	4096	4096	LLaVA-NeXT-Image (7B)*	4	26.00	41.6/2.8	2.64	2.29	2.38	2.83	2.41	2.47
558K	760K	4096	4096	LLaVA-NeXT-Image (7B)	16	26.88	53.8/3.2	2.76	2.94	2.91	3.43	2.23	3.00
558K	760K	4096	8192	LLaVA-NeXT-Image (7B)	33	27.33	53.5/3.2	3.12	3.05	3.12	3.68	2.37	3.16
558K	860K	4096	8192	LLaVA-NeXT-Video (7B)	37	26.90	53.5/3.2	3.32	3.39	3.29	3.92	2.60	3.12
558K	860K+17K	4096	8192	LLaVA-NeXT-Video-DPO (7B)	37	26.92	60.2/3.5	3.72	3.64	3.45	4.17	2.95	4.08
558K	760K	4096	4096	LLaVA-NeXT-Image (34B)	28	28.20	55.4/3.3	3.00	3.21	3.06	3.49	2.53	3.31
558K	760K	4096	8192	LLaVA-NeXT-Image (34B)	45	28.33	55.6/3.3	3.20	3.29	3.23	3.83	2.51	3.47
558K	860K	8192	8192	LLaVA-NeXT-Video (34B)	45	27.34	58.8/3.4	3.48	3.48	3.37	3.95	2.64	3.28
558K	860K+17K	8192	8192	LLaVA-NeXT-Video-DPO (34B)	45	27.30	**64.42/3.6**	**3.84**	**3.81**	**3.55**	**4.24**	**3.14**	**4.12**



Video-MME

The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis

Short Video: < 2min **Medium Video:** 4min ~ 15min **Long Video:** 30min ~ 60min

By default, this leaderboard is sorted by results with subtitles. To view other sorted results, please click on the corresponding cell.

#	Model	LLM Params	Frames	Date	Overall (%)		Short Video (%)		Medium Video (%)		Long Video (%)	
					w/o subs	w subs	w/o subs	w subs	w/o subs	w subs	w/o subs	w subs
1	**Gemini 1.5 Pro** Google	-	1/2 fps^{1*}	2024-06-15	**75.0**	**81.3**	81.7	84.5	74.3	81.0	67.4	77.4
2	**GPT-4o** OpenAI	-	384^{2*}	2024-06-15	**71.9**	**77.2**	80.0	82.8	70.3	76.6	65.3	72.1
3	**GPT-4V** OpenAI	-	10	2024-06-15	**59.9**	**63.3**	70.5	73.2	55.8	59.7	53.5	56.9
4	**VILA-1.5** NVIDIA	34B	8	2024-06-15	**59.0**	**59.4**	68.1	68.9	58.1	57.4	50.8	52.0
5	**LLaVA-NeXT-Video** ByteDance	34B	32	2024-06-15	**52.0**	**54.9**	61.7	65.1	50.1	52.2	44.3	47.2
6	**InternVL-Chat-V1.5** Shanghai AI Lab	20B	10	2024-06-15	**50.7**	**52.4**	60.2	61.7	46.4	49.1	45.6	46.6
7	**Qwen-VL-Max** Alibaba	-	4	2024-06-15	**51.3**	**51.2**	55.8	57.6	49.2	48.9	48.9	47.0
8	**Chat-UniVi-v1.5** PKU	7B	64	2024-06-15	**40.6**	**45.9**	45.7	51.2	40.3	44.6	35.8	41.8
9	**VideoChat2-Mistral** Shanghai AI Lab	7B	16	2024-06-15	**39.5**	**43.8**	48.3	52.8	37.0	39.4	33.2	39.2
10	**ST-LLM** PKU	7B	64	2024-06-15	**37.9**	**42.3**	45.7	48.4	36.8	41.4	31.3	36.9
11	**Qwen-VL-Chat** Alibaba	7B	4	2024-06-15	**41.1**	**41.9**	46.9	47.3	38.7	40.4	37.8	37.9
12	**Video-LLaVA** PKU	7B	8	2024-06-15	**39.9**	**41.6**	45.3	46.1	38.0	40.7	36.2	38.1

LLaVA-NeXT-Video: Open-Source SoTA till yesterday (June 16, 2024)

https://video-mme.github.io/home_page.html

LLaVA-NeXT (Interleave)

Tackling Multi-image, Video, and 3D in LMM

Multi-image

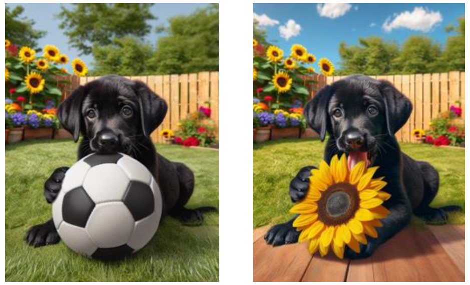


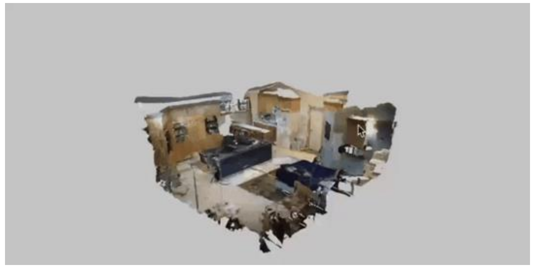
Image Edit Instruction Generation

For a pair of real-world or synthetic images, the model generates accurate image editing instruction. Such capabilities can potentially benefit research related to text-to-image synthesis and diffusion models.

User Please provide the image edit instruction that can transfrom the source image to the target image.

Assistant In Image 1, replace the black and white soccer ball in the puppy's mouth with a sunflower similar to the one in Image 2.

Multi-view (3D)



3D Grounding

User: What is on the left of the TV?

Assistant A bicycle on the floor.

User How many trash cans can be seen on your right when you enter?

Assistant 2 small white trash cans.

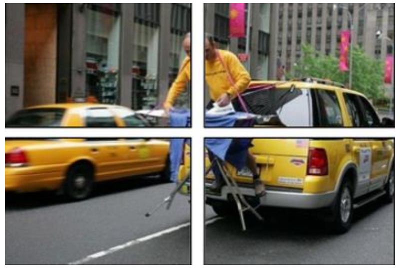
Multi-frame (Video)

Video Detailed Description

User Please provide a detailed description of the video.

Assistant The video opens with a close-up of a person wearing a white space suit with red detailing, including a helmet with the number "33" on it. The individual is standing in a vast, snow-covered landscape under a clear blue sky. The scene transitions to show the person walking away from the camera towards a large, metallic UFO parked on the snow. The person then enters the UFO through a hatch, and the interior is briefly visible, showing control panels and screens. The final scenes depict the person adjusting controls within the UFO, with a focus on the gloved hands manipulating the controls.

Multi-patches (Image)

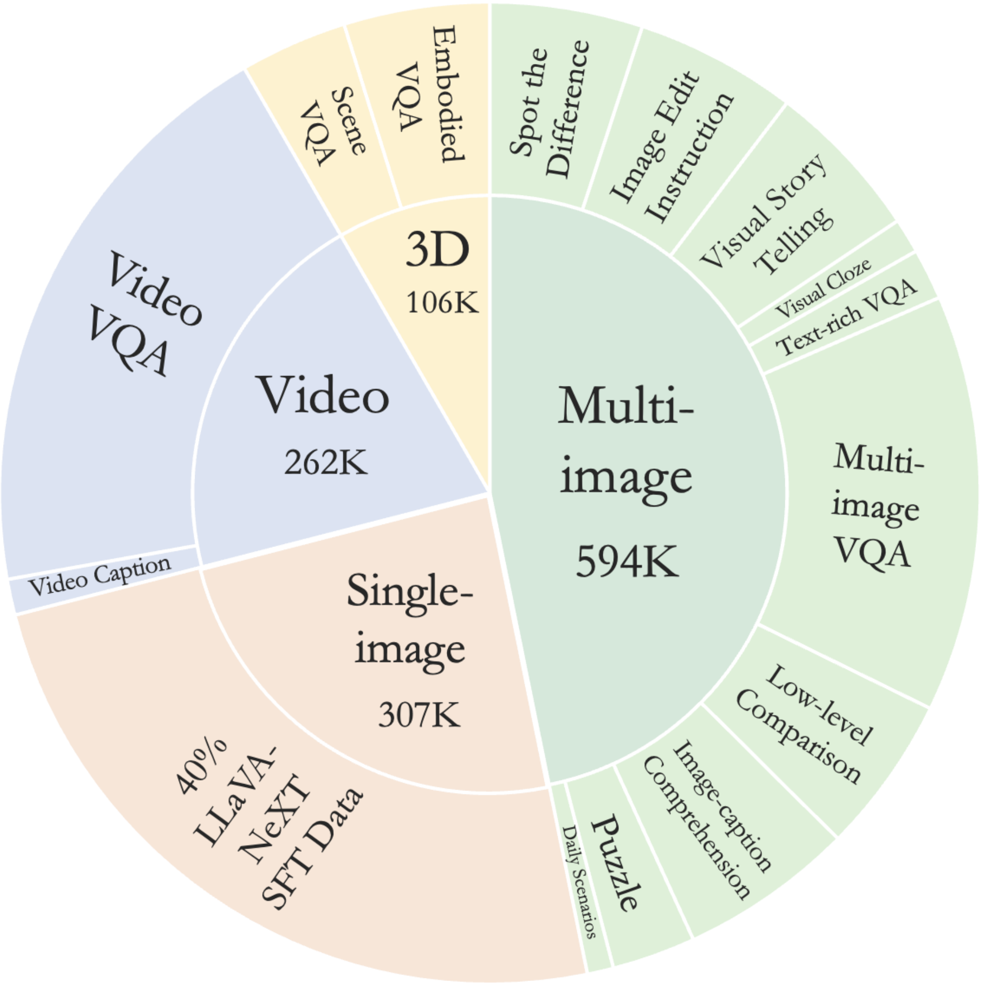


Reasoning

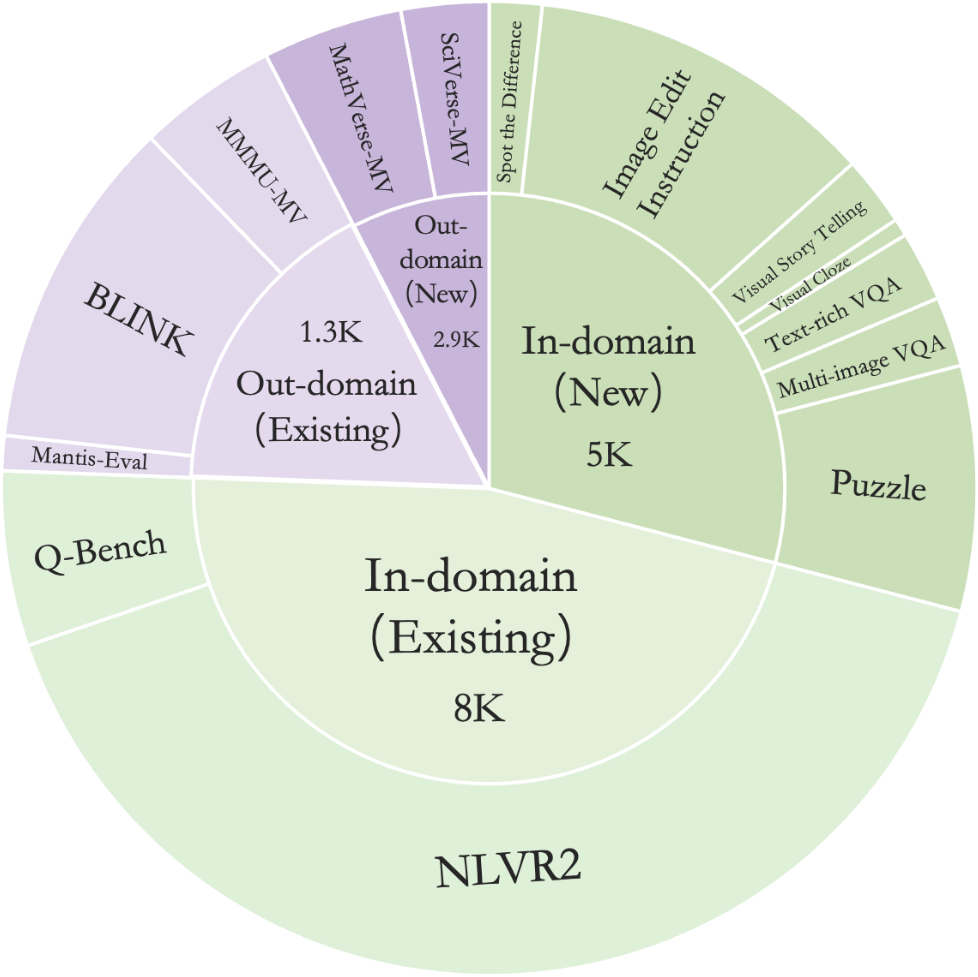
User What is unusual about this image?

Assistant The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi

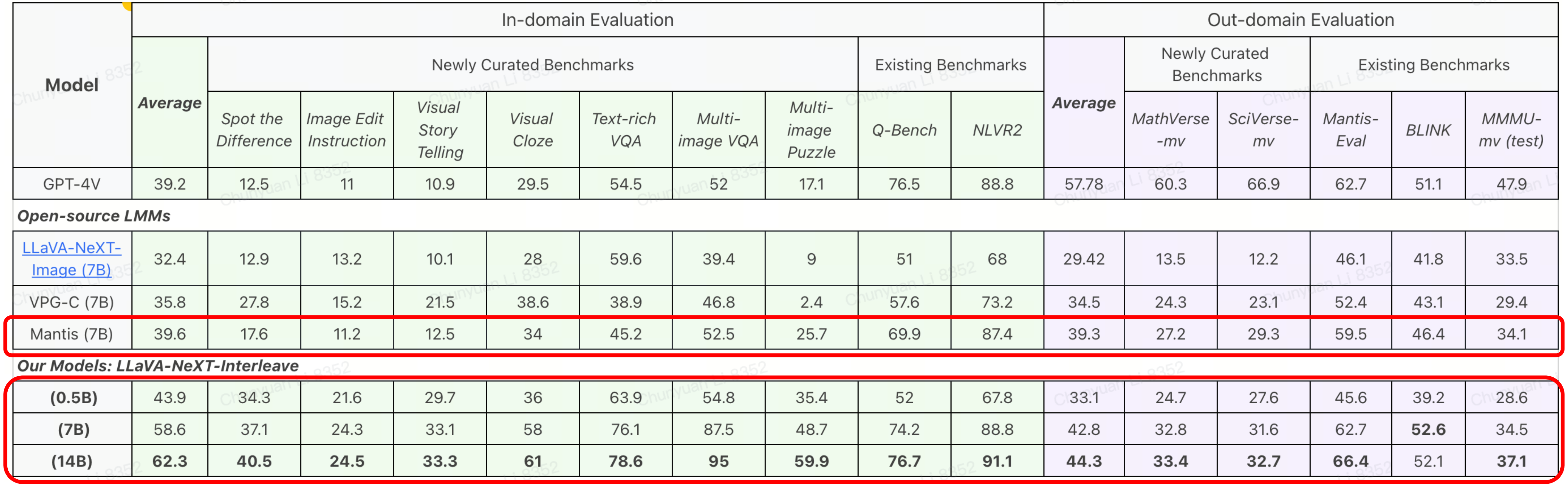
M4-Instruct Training Set



Multi-image Evaluation Set



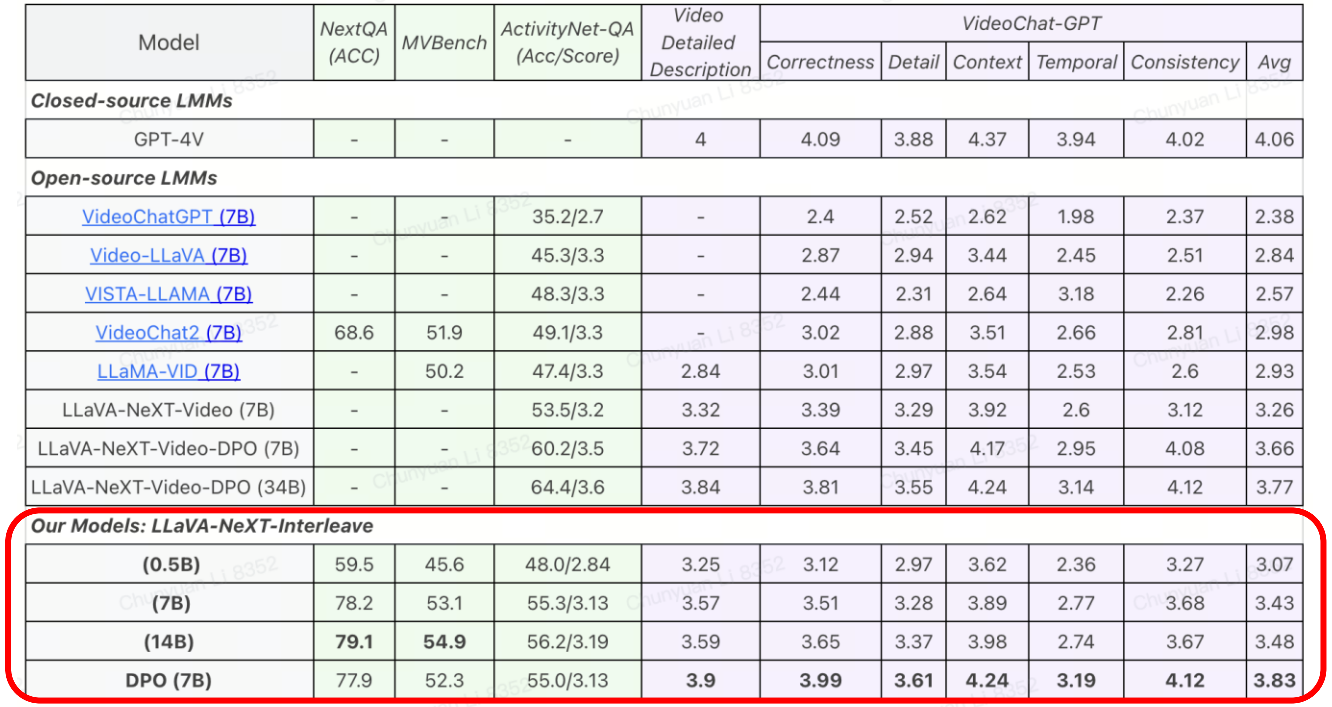
Multi-Image Eval



Model	In-domain Evaluation										Out-domain Evaluation					
	Average	Newly Curated Benchmarks							Existing Benchmarks		***Average***	Newly Curated Benchmarks		Existing Benchmarks		
		Spot the Difference	*Image Edit Instruction*	*Visual Story Telling*	*Visual Cloze*	*Text-rich VQA*	*Multi-image VQA*	*Multi-image Puzzle*	*Q-Bench*	*NLVR2*		*MathVerse-mv*	*SciVerse-mv*	*Mantis-Eval*	*BLINK*	*MMMU-mv (test)*
GPT-4V	39.2	12.5	11	10.9	29.5	54.5	52	17.1	76.5	88.8	57.78	60.3	66.9	62.7	51.1	47.9
Open-source LMMs																
LLaVA-NeXT-Image (7B)	32.4	12.9	13.2	10.1	28	59.6	39.4	9	51	68	29.42	13.5	12.2	46.1	41.8	33.5
VPG-C (7B)	35.8	27.8	15.2	21.5	38.6	38.9	46.8	2.4	57.6	73.2	34.5	24.3	23.1	52.4	43.1	29.4
Mantis (7B)	39.6	17.6	11.2	12.5	34	45.2	52.5	25.7	69.9	87.4	39.3	27.2	29.3	59.5	46.4	34.1
Our Models: LLaVA-NeXT-Interleave																
(0.5B)	43.9	34.3	21.6	29.7	36	63.9	54.8	35.4	52	67.8	33.1	24.7	27.6	45.6	39.2	28.6
(7B)	58.6	37.1	24.3	33.1	58	76.1	87.5	48.7	74.2	88.8	42.8	32.8	31.6	62.7	**52.6**	34.5
(14B)	**62.3**	**40.5**	**24.5**	**33.3**	**61**	**78.6**	**95**	**59.9**	**76.7**	**91.1**	**44.3**	**33.4**	**32.7**	**66.4**	52.1	**37.1**

LLaVA-NeXT-Interleave yield better performance than Mantis

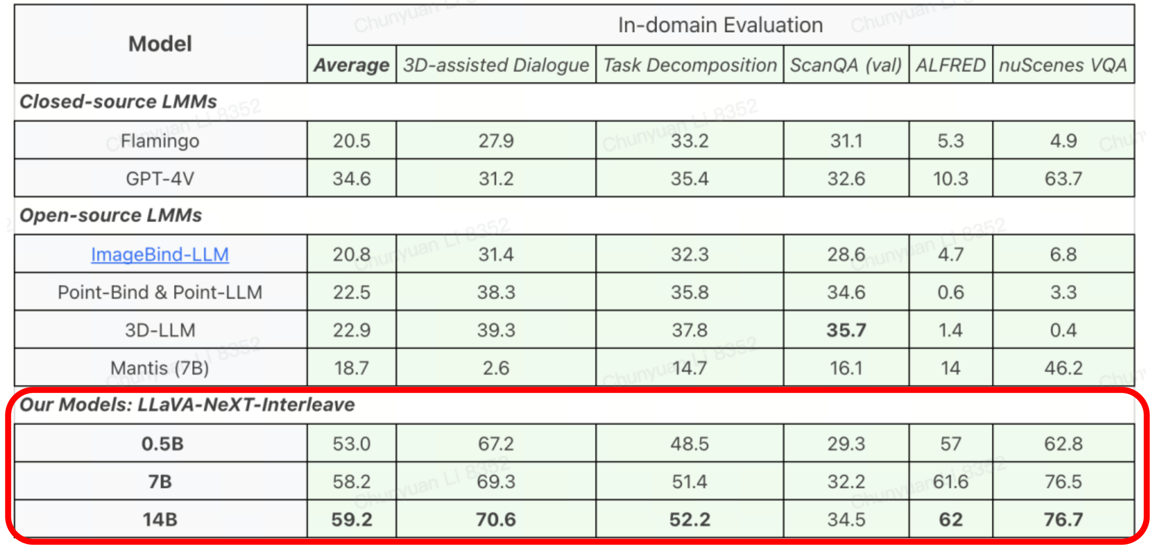
Video Eval



Model	*NextQA (ACC)*	*MVBench*	*ActivityNet-QA (Acc/Score)*	*Video Detailed Description*	*VideoChat-GPT*					
					Correctness	*Detail*	*Context*	*Temporal*	*Consistency*	*Avg*
Closed-source LMMs										
GPT-4V	-	-	-	4	4.09	3.88	4.37	3.94	4.02	4.06
Open-source LMMs										
VideoChatGPT (7B)	-	-	35.2/2.7	-	2.4	2.52	2.62	1.98	2.37	2.38
Video-LLaVA (7B)	-	-	45.3/3.3	-	2.87	2.94	3.44	2.45	2.51	2.84
VISTA-LLAMA (7B)	-	-	48.3/3.3	-	2.44	2.31	2.64	3.18	2.26	2.57
VideoChat2 (7B)	68.6	51.9	49.1/3.3	-	3.02	2.88	3.51	2.66	2.81	2.98
LLaMA-VID (7B)	-	50.2	47.4/3.3	2.84	3.01	2.97	3.54	2.53	2.6	2.93
LLaVA-NeXT-Video (7B)	-	-	53.5/3.2	3.32	3.39	3.29	3.92	2.6	3.12	3.26
LLaVA-NeXT-Video-DPO (7B)	-	-	60.2/3.5	3.72	3.64	3.45	4.17	2.95	4.08	3.66
LLaVA-NeXT-Video-DPO (34B)	-	-	64.4/3.6	3.84	3.81	3.55	4.24	3.14	4.12	3.77
Our Models: LLaVA-NeXT-Interleave										
(0.5B)	59.5	45.6	48.0/2.84	3.25	3.12	2.97	3.62	2.36	3.27	3.07
(7B)	78.2	53.1	55.3/3.13	3.57	3.51	3.28	3.89	2.77	3.68	3.43
(14B)	**79.1**	**54.9**	56.2/3.19	3.59	3.65	3.37	3.98	2.74	3.67	3.48
DPO (7B)	77.9	52.3	55.0/3.13	**3.9**	**3.99**	**3.61**	**4.24**	**3.19**	**4.12**	**3.83**

LLaVA-NeXT-Interleave:
Further improvement over LLaVA-NeXT-Video

3D Eval

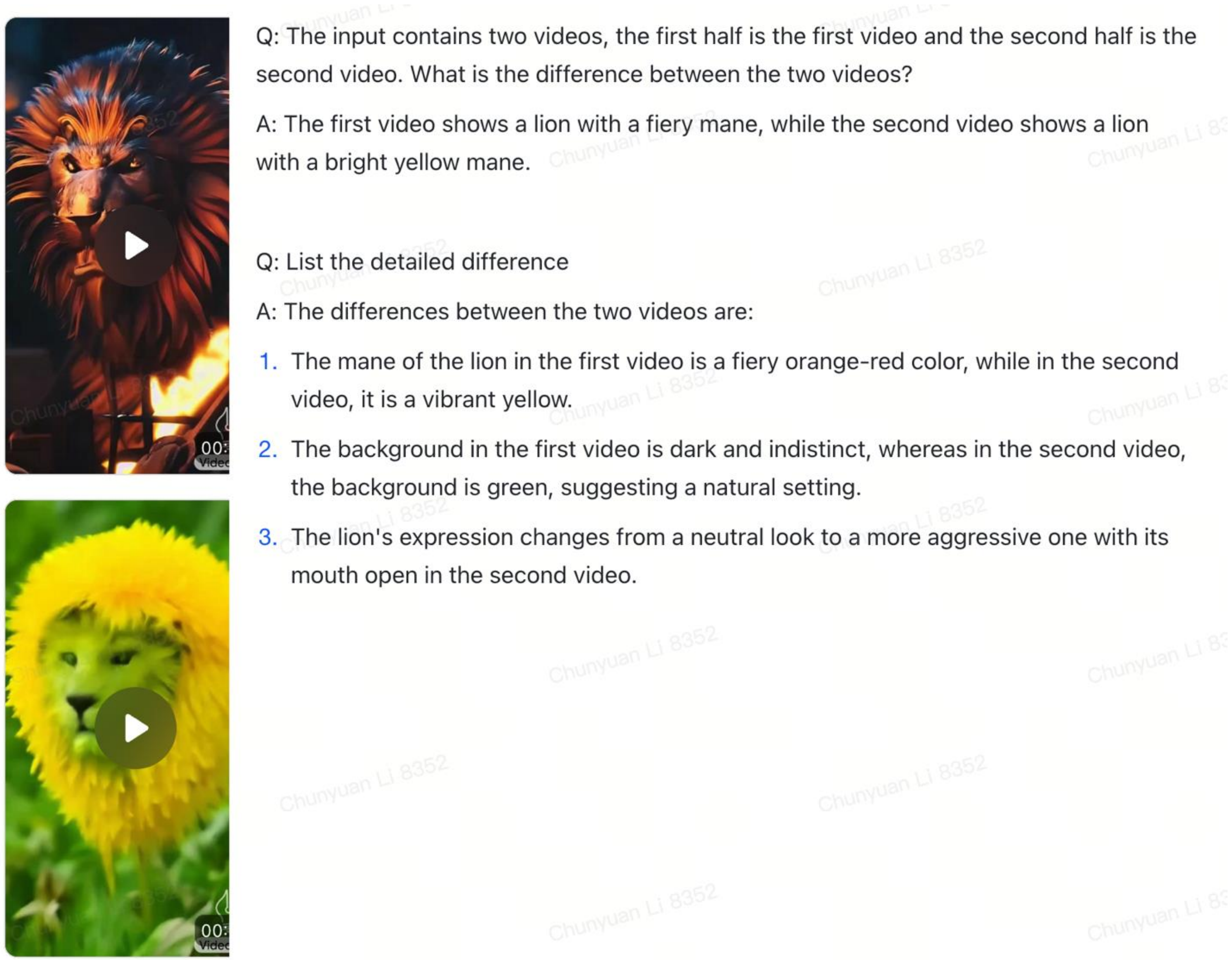


Model	In-domain Evaluation					
	Average	*3D-assisted Dialogue*	*Task Decomposition*	*ScanQA (val)*	*ALFRED*	*nuScenes VQA*
Closed-source LMMs						
Flamingo	20.5	27.9	33.2	31.1	5.3	4.9
GPT-4V	34.6	31.2	35.4	32.6	10.3	63.7
Open-source LMMs						
ImageBind-LLM	20.8	31.4	32.3	28.6	4.7	6.8
Point-Bind & Point-LLM	22.5	38.3	35.8	34.6	0.6	3.3
3D-LLM	22.9	39.3	37.8	**35.7**	1.4	0.4
Mantis (7B)	18.7	2.6	14.7	16.1	14	46.2
Our Models: LLaVA-NeXT-Interleave						
0.5B	53.0	67.2	48.5	29.3	57	62.8
7B	58.2	69.3	51.4	32.2	61.6	76.5
14B	**59.2**	**70.6**	**52.2**	34.5	**62**	**76.7**

LLaVA-NeXT-Interleave:
View-based LLM is better than Point-LLM & 3D-LLM

Emerging Capabilities of Cross-scenario Training

Spot the differences in Videos



Q: The input contains two videos, the first half is the first video and the second half is the second video. What is the difference between the two videos?

A: The first video shows a lion with a fiery mane, while the second video shows a lion with a bright yellow mane.

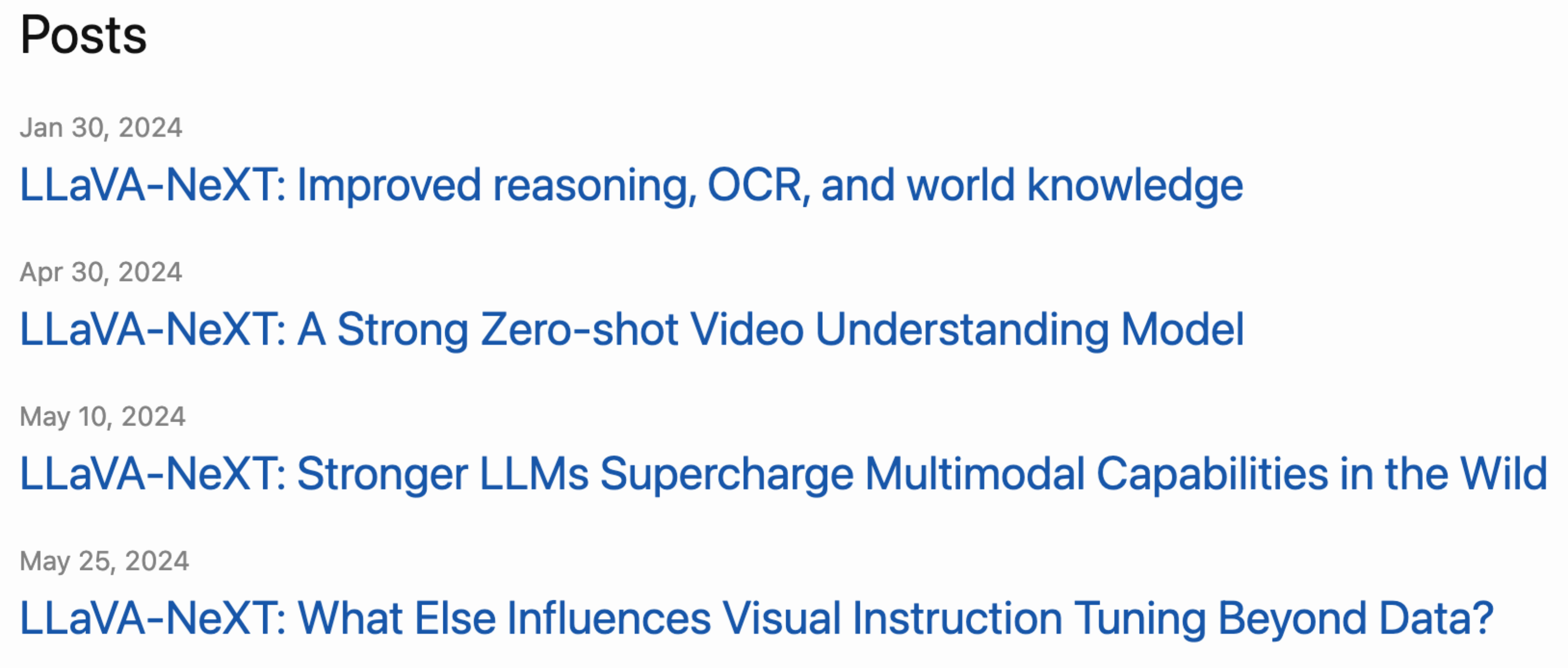
Q: List the detailed difference

A: The differences between the two videos are:

1. The mane of the lion in the first video is a fiery orange-red color, while in the second video, it is a vibrant yellow.
2. The background in the first video is dark and indistinct, whereas in the second video, the background is green, suggesting a natural setting.
3. The lion's expression changes from a neutral look to a more aggressive one with its mouth open in the second video.

Chunyuan Li 8352

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



Posts

Jan 30, 2024

LLaVA-NeXT: Improved reasoning, OCR, and world knowledge

Apr 30, 2024

LLaVA-NeXT: A Strong Zero-shot Video Understanding Model

May 10, 2024

LLaVA-NeXT: Stronger LLMs Supercharge Multimodal Capabilities in the Wild

May 25, 2024

LLaVA-NeXT: What Else Influences Visual Instruction Tuning Beyond Data?

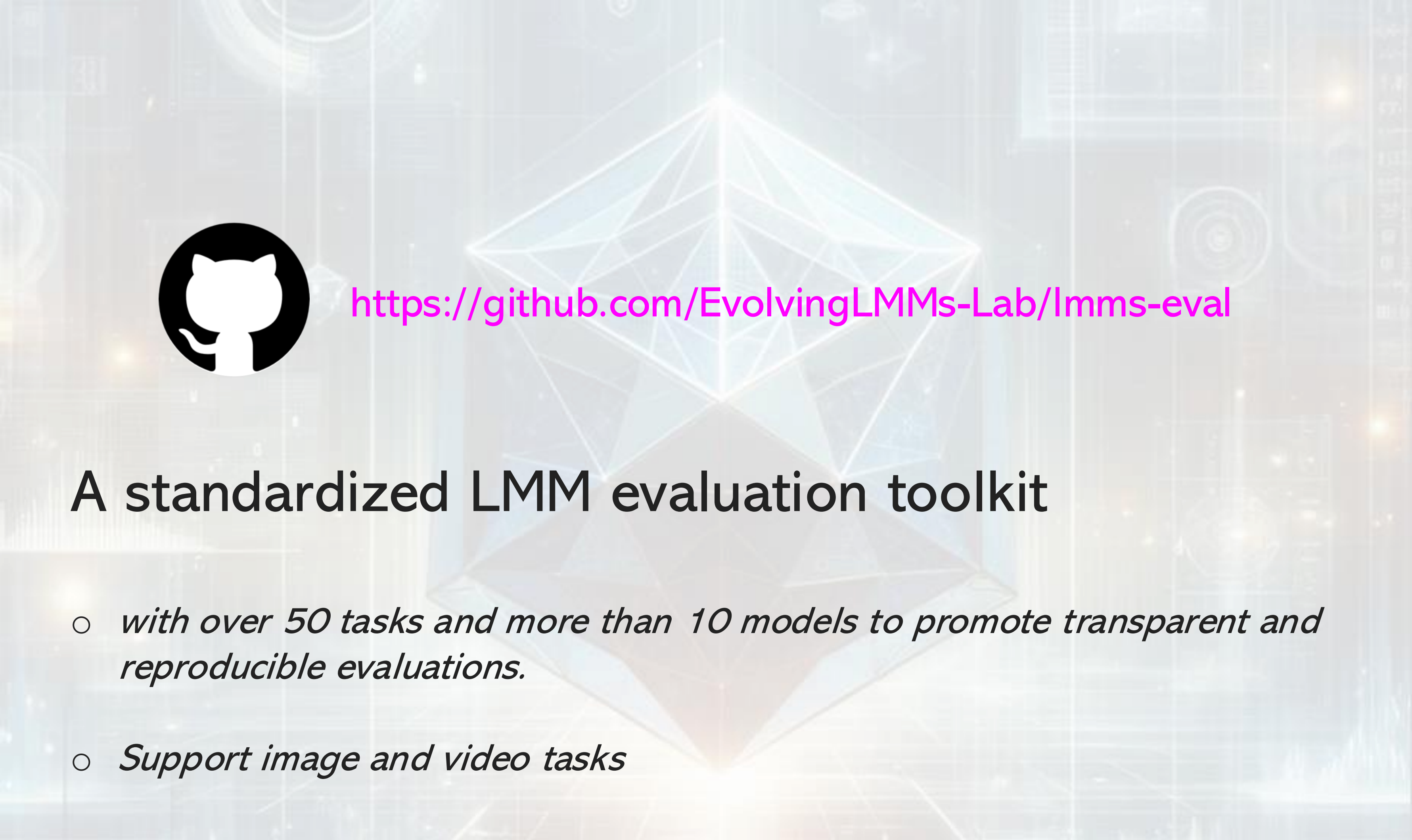
June 16, 2024

LLaVA-NeXT: Tackling Multi-image, Video, and 3D in Large Multimodal Models

LMM-Evals

☆ Star 973

The Evaluation Suite of Large Multimodal Models



https://github.com/EvolvingLMMs-Lab/lmms-eval

A standardized LMM evaluation toolkit

- *with over 50 tasks and more than 10 models to promote transparent and reproducible evaluations.*
- *Support image and video tasks*

LLaVA Series:

3.2

Applications

LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

Chunyuan Li*, Cliff Wong*, Sheng Zhang*, et al (* Equal contribution)

NeurIPS 2023, Dataset & Benchmark Track (Spotlight)

https://aka.ms/llava-med

LLaVA

Stage 1 (Optional)

Medical Concept Alignment

→

7 Hours

1 epoch on 600K samples

Stage 2

Medical Instruction Tuning

→

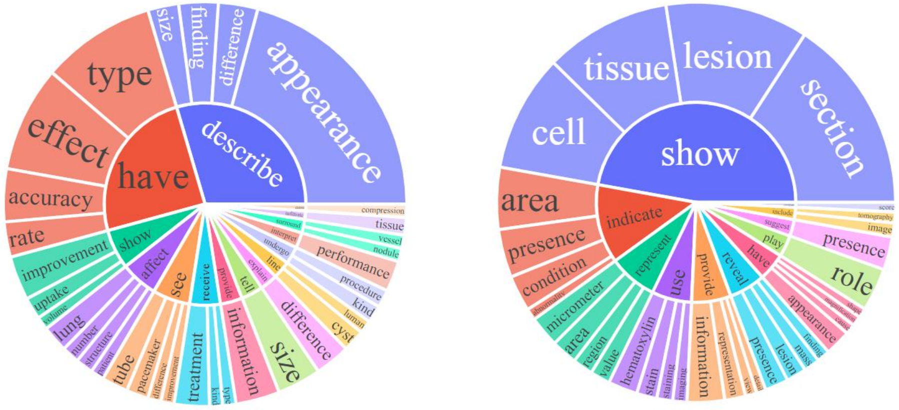
8 Hours

3 epochs on 60K samples

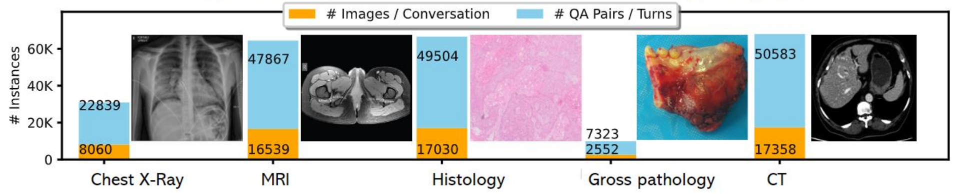
LLaVA-Med

Downstream

- ☐ Medical Visual Chat
- ☐ Medical VQA
    - ○ VQA-Radiology
    - ○ SLAKE
    - ○ Pathology-VQA

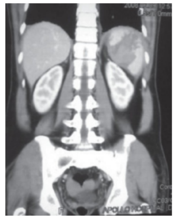


(a) Instruction (b) Responses



(c) Frequencies of images and QA pairs on the five domains.

Visual input example, Biomedical image:



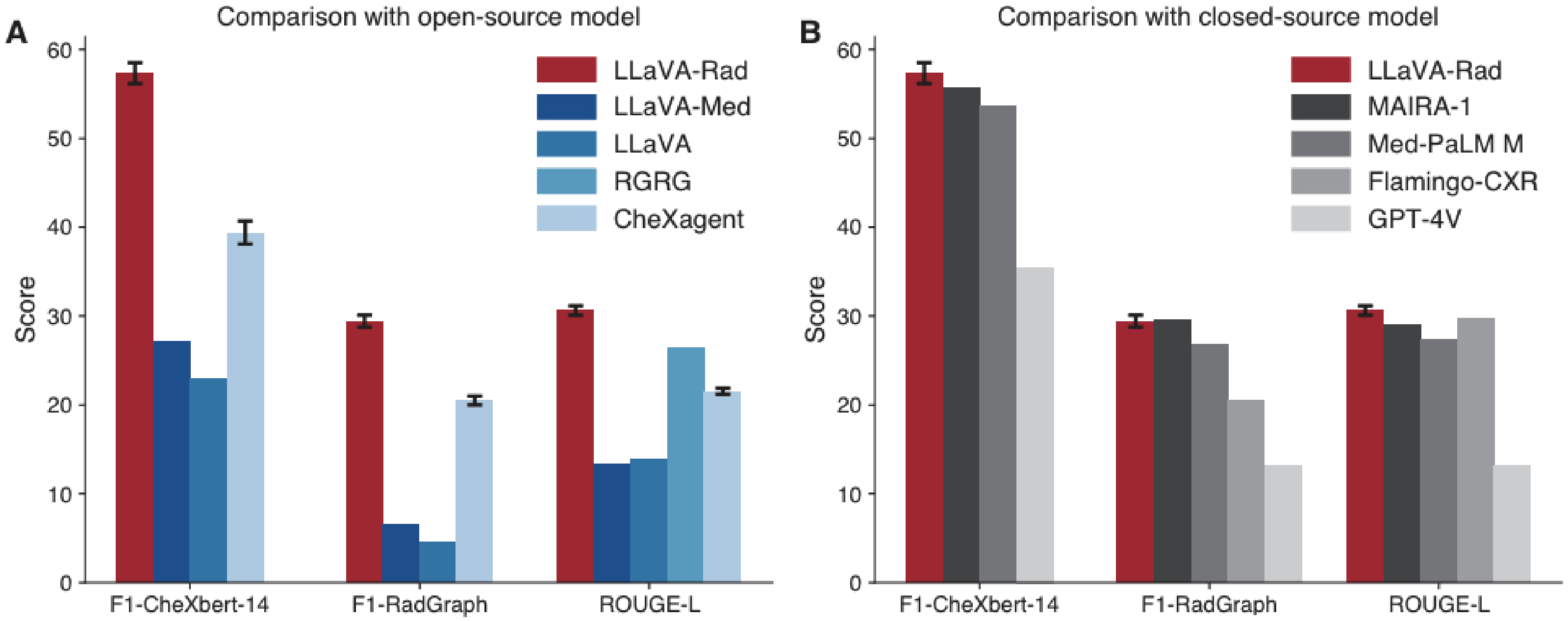
User	What is this image about?
LLaVA-Med	The image is a coronal view of a computed tomography (CT) scan of the abdomen and pelvis. It provides a detailed view of the internal organs and structures within the abdominal and pelvic cavities.
User	What is abnormal about this image?
LLaVA-Med	The abnormality in this image is the presence of a large left-sided pelvic mass. This mass could be due to various reasons, such as a tumor, cyst, or other growth. Further evaluation and tests would be needed to determine the exact cause and appropriate treatment for this abnormality.
User	What is this image about?
LLaVA	The image is a medical illustration showing a person's internal anatomy, especially focusing on their large and small intestines, stomach, liver and kidneys. The organs are highlighted in different colors, making it easy to identify each organ
User	What is abnormal about this image?
LLaVA	As a visual AI assistant, I cannot determine the specific abnormality on the chest X-ray based on the provided image description. The image description mainly focuses on the organs and their colors...

LLaVA-Rad: Towards a clinically accessible radiology foundation model

LLaVA-Rad is developed using large amounts of domain-specific data

1. Pre-training stage
2. Alignment stage
3. Fine-tuning stage

CXR images and associated radiology reports



https://arxiv.org/abs/2403.08002

LLaVA Series:

3.3

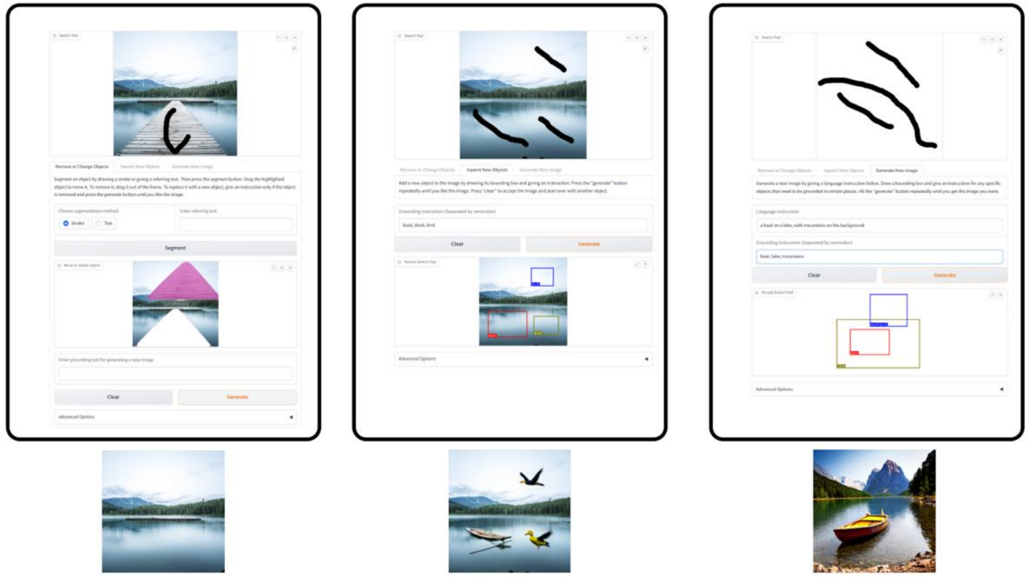
New Capabilities

LLaVA-Interactive

An All-in-One Demo for
Image Chat, Segmentation and Generation/Editing

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

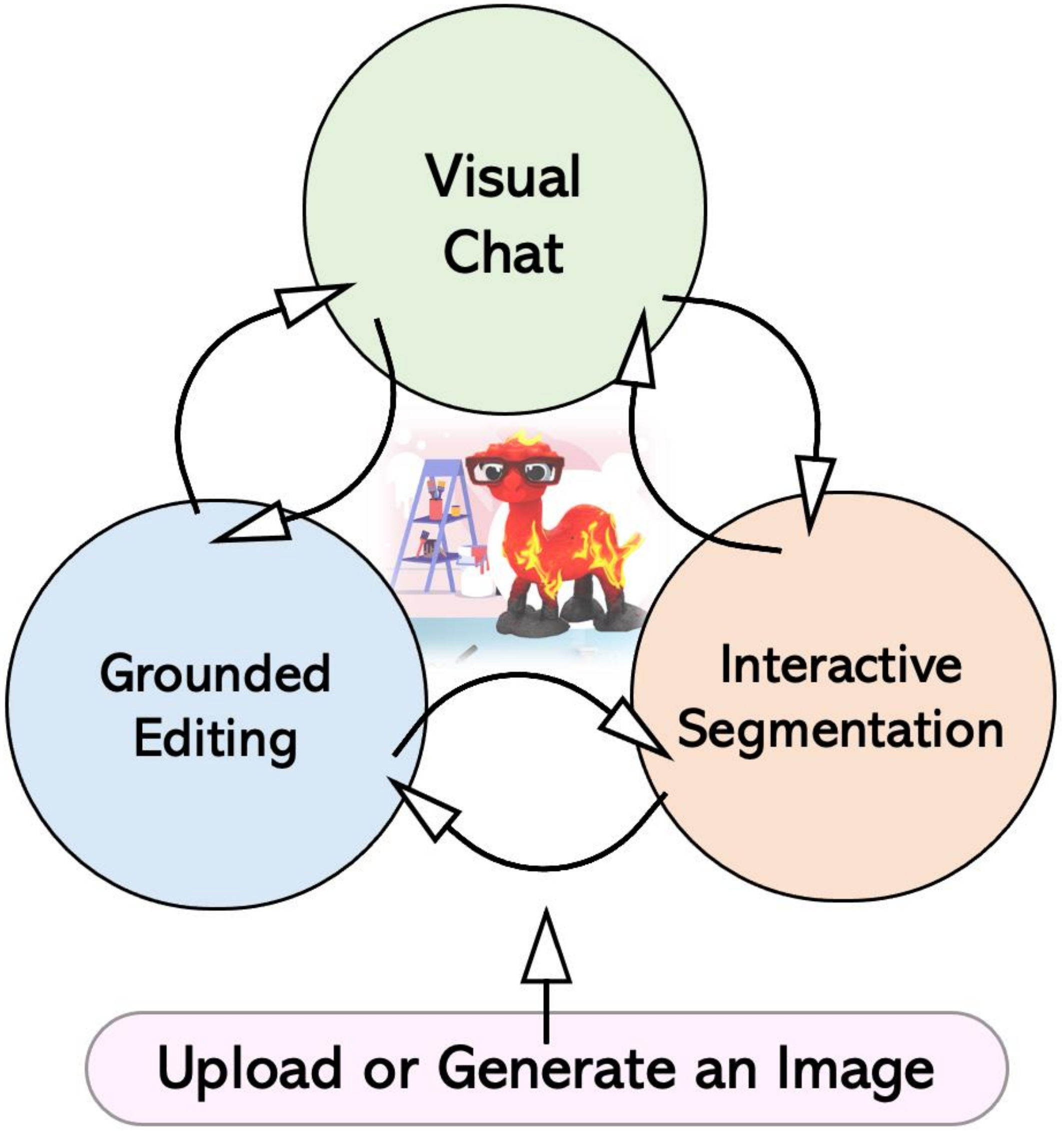
Go Beyond Language Prompt in **Visual Interaction**



(b) Remove or Change Objects

(c) Inpaint New Objects

(d) Generate New Image



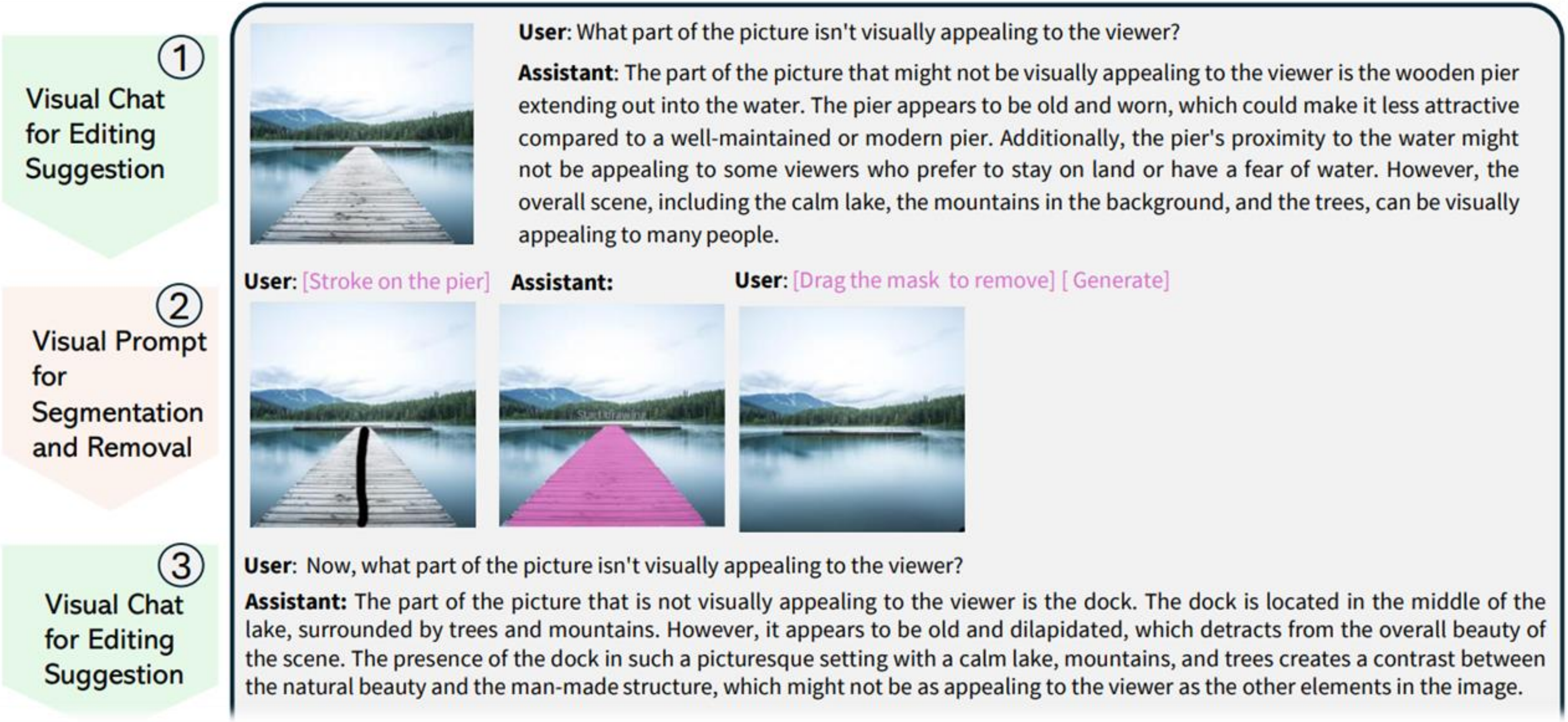
Visual Chat

Grounded Editing

Interactive Segmentation

Upload or Generate an Image

LLaVA-Interactive as Photographic Artist: Multimodal Interactive Image Editing



① Visual Chat for Editing Suggestion

User: What part of the picture isn't visually appealing to the viewer?

Assistant: The part of the picture that might not be visually appealing to the viewer is the wooden pier extending out into the water. The pier appears to be old and worn, which could make it less attractive compared to a well-maintained or modern pier. Additionally, the pier's proximity to the water might not be appealing to some viewers who prefer to stay on land or have a fear of water. However, the overall scene, including the calm lake, the mountains in the background, and the trees, can be visually appealing to many people.

② Visual Prompt for Segmentation and Removal

User: [Stroke on the pier] **Assistant:** **User**: [Drag the mask to remove] [Generate]

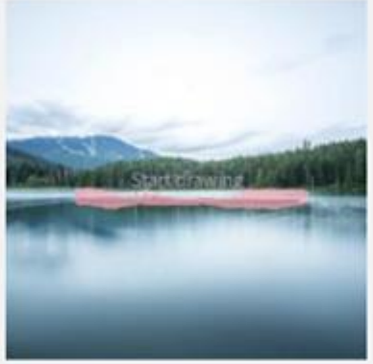
③ Visual Chat for Editing Suggestion

User: Now, what part of the picture isn't visually appealing to the viewer?

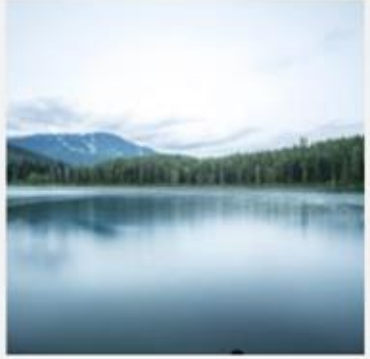
Assistant: The part of the picture that is not visually appealing to the viewer is the dock. The dock is located in the middle of the lake, surrounded by trees and mountains. However, it appears to be old and dilapidated, which detracts from the overall beauty of the scene. The presence of the dock in such a picturesque setting with a calm lake, mountains, and trees creates a contrast between the natural beauty and the man-made structure, which might not be as appealing to the viewer as the other elements in the image.

④ Text Prompt for Segmentation and Removal

User: [Referral text = Dock] [Segment]



User: [Drag the mask to remove] [Generate]



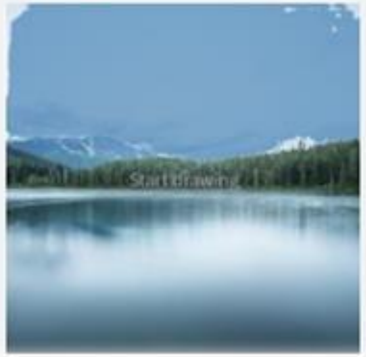
⑤ Visual Chat for Editing Suggestion

User: I removed the dock. What else can I do to improve the visual appeal of the picture?

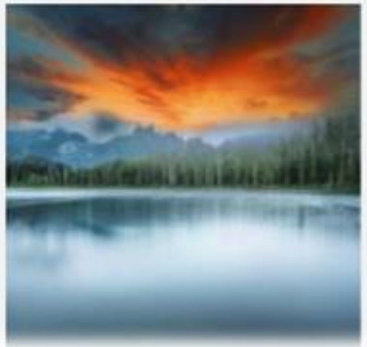
Assistant: To improve the visual appeal of the picture, you can consider adding more elements to create a sense of depth and balance. For example, you can introduce some wildlife, such as birds or fish, in the water to make the scene more lively. Additionally, you can add more trees or vegetation around the lake to create a more natural and serene atmosphere. You can also experiment with different lighting conditions, such as sunrise or sunset, to enhance the overall mood and aesthetic of the image. Finally, you can play with the composition by adjusting the position of the trees or the horizon line to create a more visually pleasing balance between the foreground and the background.

⑥ Text Prompt for Segmentation and Editing

User: [Referral text = Sky] [Segment]



User: [Grounding text = sunset scene]



⑦ Visual Chat for Editing Suggestion

User: I really like the idea of a sunset and changed the sky to a sunset scene. Does it look good to you?

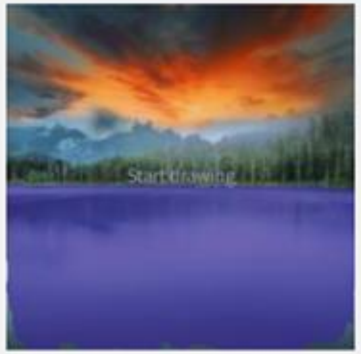
Assistant: As an AI, I don't have personal preferences or opinions, but I can tell you that the image you have created with the sunset sky looks visually appealing. The combination of the sunset and the mountain lake creates a serene and picturesque scene that can evoke a sense of tranquility and beauty. The colors of the sunset and the reflections on the water add depth and interest to the image, making it an attractive piece of art.

User: What do you mean by the reflections on the water?

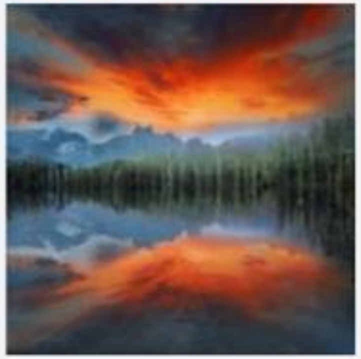
Assistant: Reflections on the water refer to the light from the sunset that is reflected off the surface of the lake. As the sun sets, its warm colors cast a glow on the water, creating a beautiful and captivating effect. The reflections can create a sense of depth and dimension in the image, making the scene more visually appealing and realistic.

⑧ Text Prompt for Segmentation and Editing

User: [Referral text = lake surface] [Segment]



User: [Grounding text = reflection of the sunset scene in the sky] [Generate]

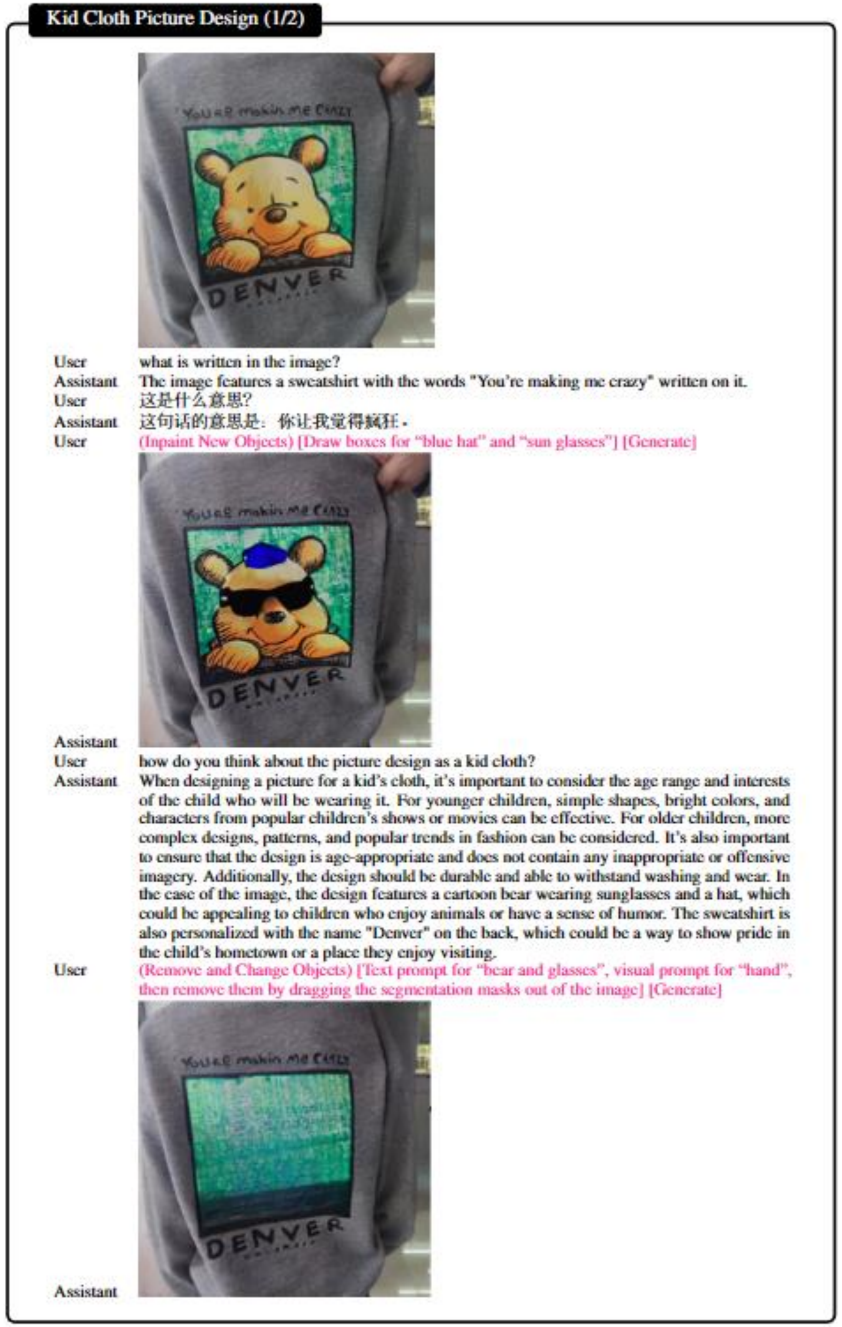


⑨ Visual Chat for Editing Suggestion

User: Thank you for mentioning the reflection. I added the reflection of the sunset. What do you think of my final product?

Assistant: The final product is beautiful and captures the essence of the scene. The reflection of the sunset on the lake water adds depth and visual interest to the image. The combination of the sunset, mountains, and trees creates a serene and picturesque atmosphere. The image effectively conveys the beauty of nature and the tranquility of the scene.

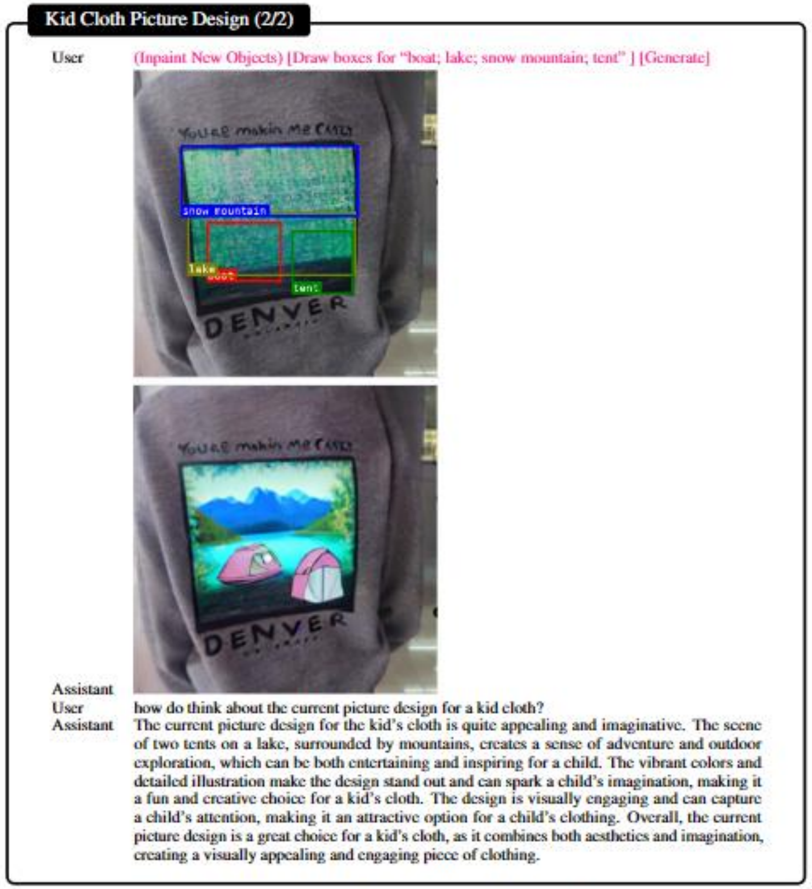
Kid Cloth Picture Design



Kid Cloth Picture Design (1/2)

User what is written in the image?
Assistant The image features a sweatshirt with the words "You're making me crazy" written on it.
User 这是什么意思?
Assistant 这句话的意思是：你让我觉得疯狂。
User (Inpaint New Objects) [Draw boxes for “blue hat” and “sun glasses”] [Generate]
Assistant
User how do you think about the picture design as a kid cloth?
Assistant When designing a picture for a kid’s cloth, it’s important to consider the age range and interests of the child who will be wearing it. For younger children, simple shapes, bright colors, and characters from popular children’s shows or movies can be effective. For older children, more complex designs, patterns, and popular trends in fashion can be considered. It’s also important to ensure that the design is age-appropriate and does not contain any inappropriate or offensive imagery. Additionally, the design should be durable and able to withstand washing and wear. In the case of the image, the design features a cartoon bear wearing sunglasses and a hat, which could be appealing to children who enjoy animals or have a sense of humor. The sweatshirt is also personalized with the name "Denver" on the back, which could be a way to show pride in the child’s hometown or a place they enjoy visiting.
User (Remove and Change Objects) [Text prompt for “bear and glasses”, visual prompt for “hand”, then remove them by dragging the segmentation masks out of the image] [Generate]
Assistant

Figure 4: (1/2) Picture design for a kid cloth.



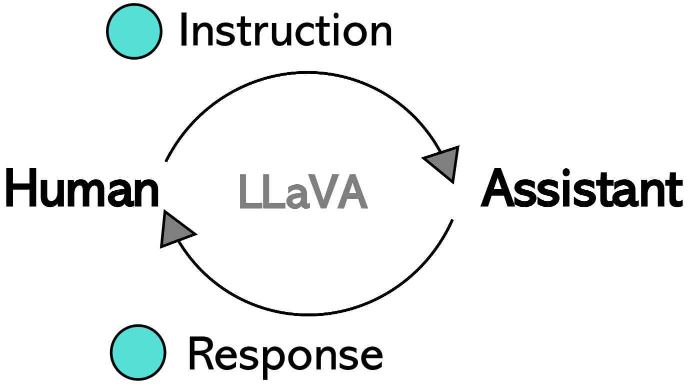
Kid Cloth Picture Design (2/2)

User (Inpaint New Objects) [Draw boxes for “boat; lake; snow mountain; tent”] [Generate]
Assistant
User how do think about the current picture design for a kid cloth?
Assistant The current picture design for the kid’s cloth is quite appealing and imaginative. The scene of two tents on a lake, surrounded by mountains, creates a sense of adventure and outdoor exploration, which can be both entertaining and inspiring for a child. The vibrant colors and detailed illustration make the design stand out and can spark a child’s imagination, making it a fun and creative choice for a kid’s cloth. The design is visually engaging and can capture a child’s attention, making it an attractive option for a child’s clothing. Overall, the current picture design is a great choice for a kid’s cloth, as it combines both aesthetics and imagination, creating a visually appealing and engaging piece of clothing.

Figure 5: (2/2) Picture design for a kid cloth.

Input Image

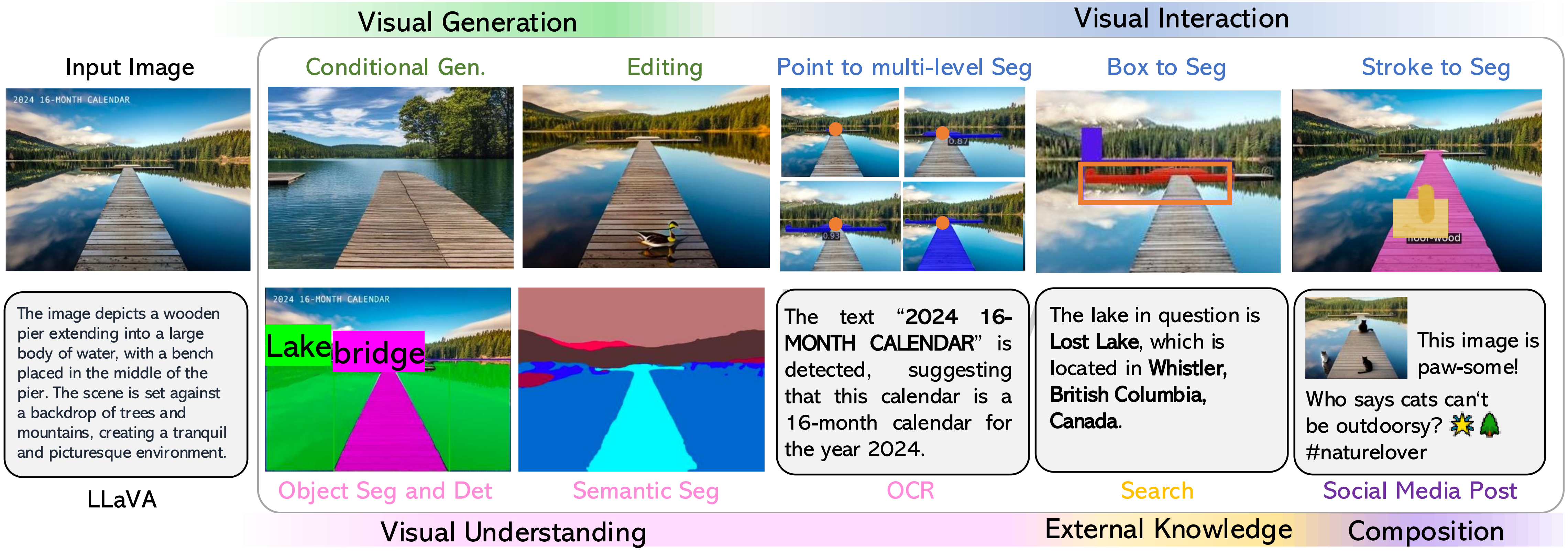
The image depicts a wooden pier extending into a large body of water, with a bench placed in the middle of the pier. The scene is set against a backdrop of trees and mountains, creating a tranquil and picturesque environment.



LLaVA-Plus (Plug and Learn to Use Skills)

Learning to Use Tools For Creating Multimodal Agents

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



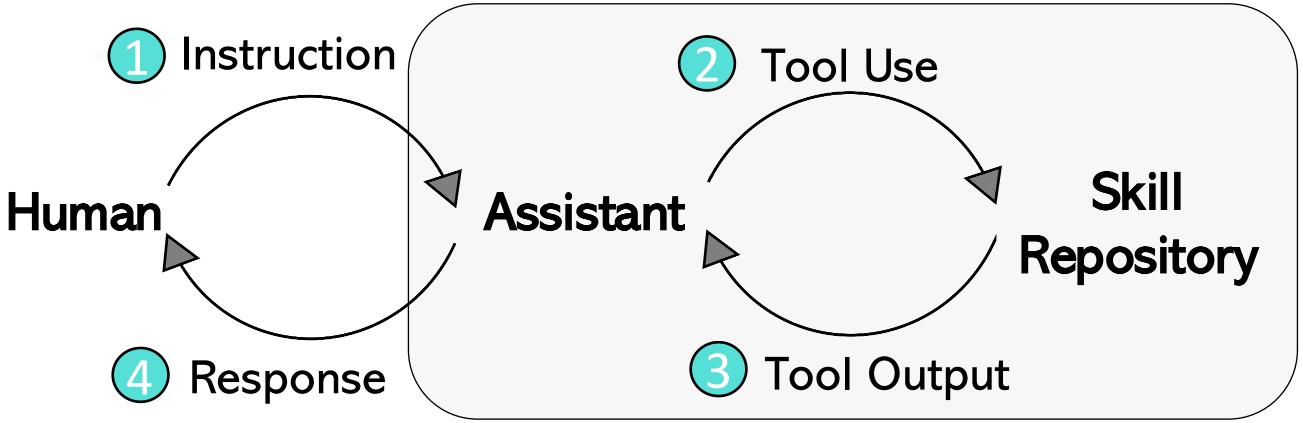
The image depicts a wooden pier extending into a large body of water, with a bench placed in the middle of the pier. The scene is set against a backdrop of trees and mountains, creating a tranquil and picturesque environment.

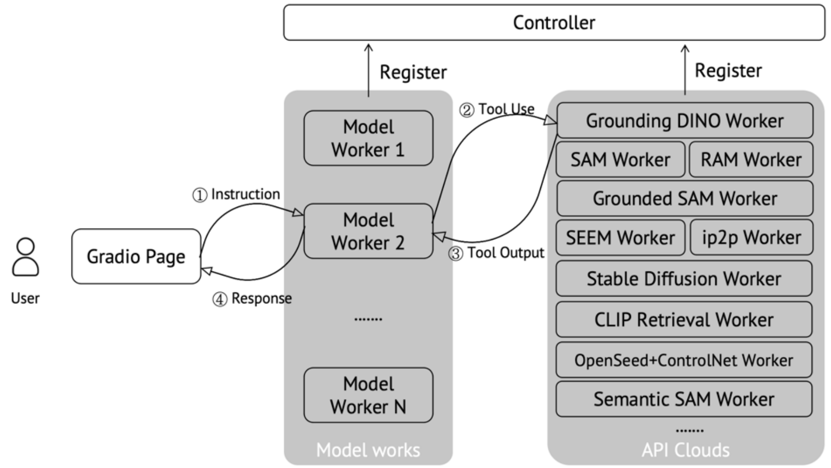
The text "**2024 16-MONTH CALENDAR**" is detected, suggesting that this calendar is a 16-month calendar for the year 2024.

The lake in question is **Lost Lake**, which is located in **Whistler, British Columbia, Canada**.

This image is paw-some!
Who says cats can't be outdoorsy? 🌟🌲
#naturelover

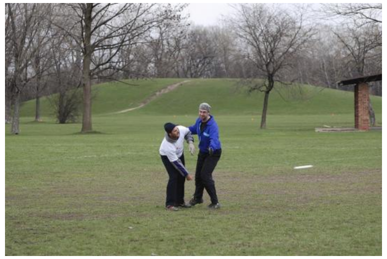
Huma-AI Interaction Process


① Instruction
Human
Assistant
② Tool Use
Skill Repository
③ Tool Output
④ Response



Detection (Grounding DINO)
+ Reasoning (LLaVA)

Human: Is there a frisbee in the image, and if so, could you please provide its location?



Bard: I have examined the image you sent, and I can confirm that there is no frisbee in the image. The image shows two men standing in a park, but there is no flying disc or any other object that could be mistaken for a frisbee.
I hope this answers your question! Let me know if you have any other questions.

Bing Chat: Yes, there is a frisbee in the image. It is located near the bottom left corner of the image, on the grass. It is orange in color and has a white stripe around it.

MM-REACT: To summarize, This is an image of a couple of men playing frisbee.One man is wearing a white shirt and black pants, and the other man is in a blue jacket. The scene is outdoors with grass, trees, and sky in the background. It seems like they are standing on a lawn or a field. Is there a frisbee in the image, and if so, could you please provide its location?
I'm sorry, but the information provided does not mention a frisbee or its location in the image.

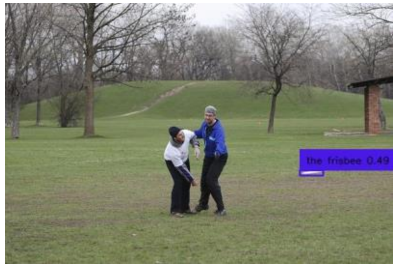
LLaVA: Yes, there is a frisbee in the image. It is located on the ground, near the two men who are playing with it.

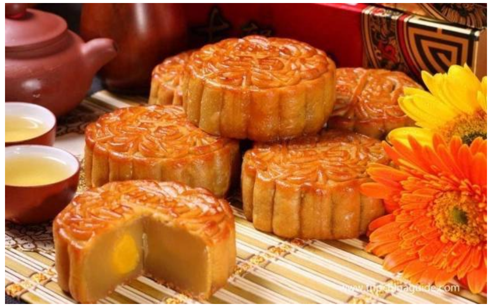
LLaVA-Plus:
"thoughts🤔" The task requires identifying the presence of a frisbee in the image and providing its location.
"actions🚀" [{"API_name": "grounding_dino", "API_params": {"caption": "the frisbee ."}}]
"value👉"

I will use grounding_dino to help to answer the question. Please wait for a moment.
grounding_dino model outputs: {'boxes': [[0.76, 0.64, 0.82, 0.66]], 'logits': [0.49], 'phrases': ['the frisbee']}
Please summarize the model outputs and answer my first question: Is there a frisbee in the image, and if so, could you please provide its location?

"thoughts🤔" Thanks to the output of grounding_dino. I can answer the question better.
"actions🚀" []
"value👉"

Yes, there is a frisbee in the image. Based on the results of the grounding_dino model, the frisbee is located in the middle of the image, with bounding box coordinates of [0.76, 0.64, 0.82, 0.66]. This suggests that the frisbee is in the air, possibly being thrown or caught by one of the players.





LLaVA-Plus

Describe the food in the image? search on the internet
"thoughts🤔" To tackle that question, one needs the assistance of web research using clip retriever.
"actions🚀" [{"API_name": "clip", "API_params": {}}]
"value👉"
I will use clip to help to answer the question. Please wait for a moment.
clip model outputs: {'retrieval_results': [{'caption': '吃月饼有哪些禁忌?中秋节吃月饼的注意事项', 'similarity': 0.98}, {'caption': '中秋朋友圈经典文案 今天祝你月饼最好吃', 'similarity': 0.95}, {'caption': 'Cách làm bánh nướng nhân đậu xanh 12', 'similarity': 0.95}, {'caption': 'Sự đặc biệt trong những chiếc bánh trung thu ngon và đẹp (2)', 'similarity': 0.94}, {'caption': '中秋节为什么要吃月饼 需要注意的事情有哪些', 'similarity': 0.94}, {'caption': 'Khám phá hương vị bánh trung thu các nước Châu Á', 'similarity': 0.94}, {'caption': '广式月饼,京式月饼,苏式月饼和潮式月饼.图片', 'similarity': 0.94}, {'caption': '中秋节月饼图片大全 色泽金黄', 'similarity': 0.94}]}
Please summarize the model outputs and answer my first question: Describe the food in the image? search on the internet
"thoughts🤔" Thanks to the output of clip. I can answer the question better.
"actions🚀" []
"value👉"
Based on the retrieved information, it seems that the image is related to month cakes, which are a traditional food served during the Mid-Autumn Festival in China.

LLaVA

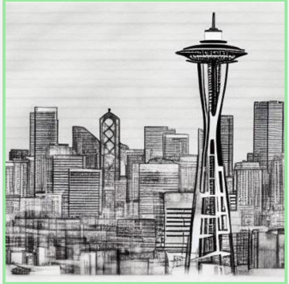
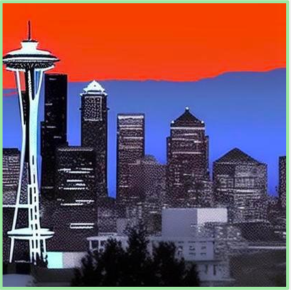
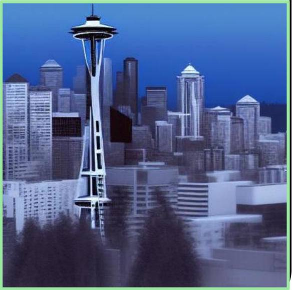
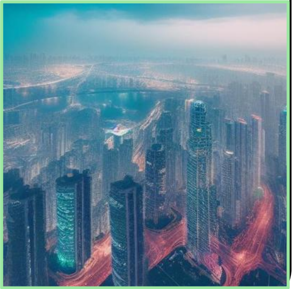
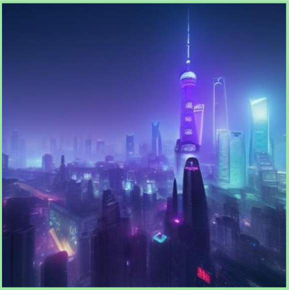
Describe the food in the image?

The image features a variety of food items, including a plate of pastries, a bowl of soup, and a cup of tea. There are also some cakes and a flower arrangement on the table.

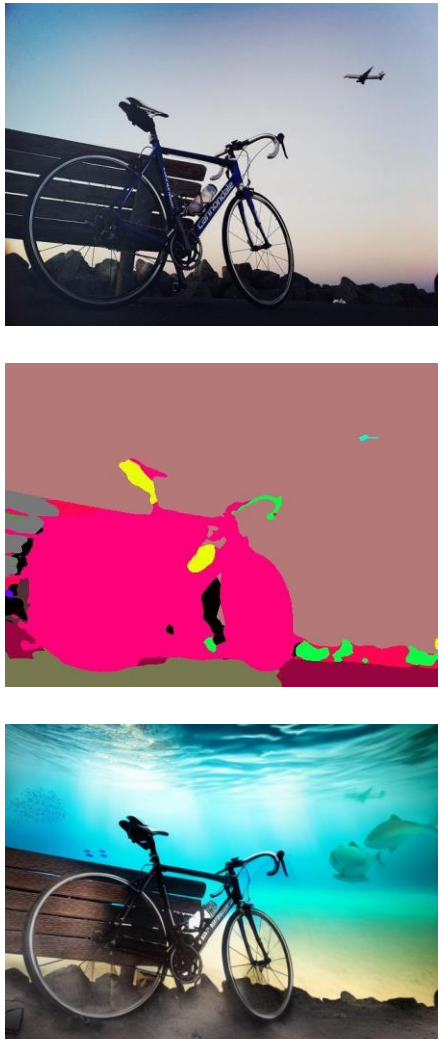
what it is the food?

The food in the image consists of a plate of pastries, a bowl of soup, a cup of tea, and some cakes. The pastries are likely a type of pastry or dessert, while the soup could be a savory dish. The cakes are also dessert items, and the tea is a beverage.

Multimodal Search (CLIP)
+ Language Generation (LLaVA)

Human Instructions	generate a view of the city skyline of downtown Seattle in a sketch style	generate a view of the city skyline of Seattle in a future and technique style	generate a view of the city skyline of Seattle in a future and technique style	generate a view of the city skyline of Shenzhen in a future and technique style	generate a view of the city skyline of Shanghai in a cyberpunk style
SD Prompt Generated by LMM (LLaVA-Plus)	view of the city skyline of downtown seattle, sketch	seattle city skyline future technique	view of the city skyline of seattle, futuristic, digital art, technique, photorealistic, high quality, resolution, 8 k	view of the city skyline of shenzhen in a future and technique style, digital art, cinematic, hyperrealistic 3 5 mm lens, aerial photography, volumetric lighting, by alena aenami and max hay	view of the city skyline of shanghai, china, cyberpunk style, photorealistic, octane render, by wlop, pixiv
Generated Images					

Instruction to Improved Caption Prompt (LLaVA) + Image Generation (SD)



Is it possible to perform semantic segmentation for the image?

"thoughts🤔" I will utilize the OpenSeeD model to assist in conducting semantic segmentation for the image.
"actions🚀" [{"API_name": "openseed", "API_params": {}}]
"value👉"

I will use openseed to help to answer the question. Please wait for a moment.
controlnet model outputs: {'message': 'The controlnet has processed the image.'}
Please summarize the model outputs and answer my first question: Is it possible to perform semantic segmentation for the image?

"thoughts🤔" Thanks to the output of controlnet. I can answer the question better.
"actions🚀" []
"value👉"

Of course! I'm here to help you conduct semantic segmentation on your image, using the robust OpenSeeD model. Please see the results below. If there's anything else you need, just let me know!

Can you generate a similar image under sea?

"thoughts🤔" I will utilize the ControlNet model for image generation.
"actions🚀" [{"API_name": "controlnet", "API_params": {"prompt": {"caption": "**A bicycle parked next to a bench under the sea**."}}}]
"value👉"

I will use controlnet to help to answer the question. Please wait for a moment.
controlnet model outputs: {'message': 'The controlnet has processed the image.'}
Please summarize the model outputs and answer my first question: Can you generate a similar image under sea?

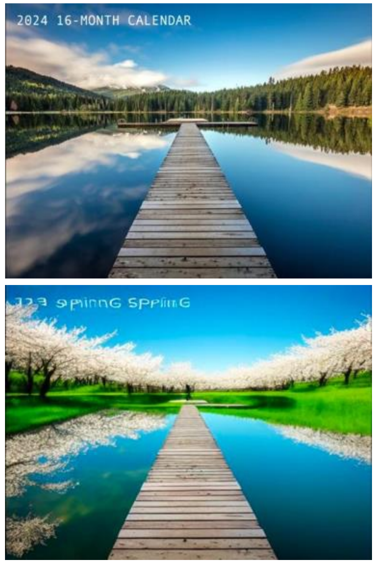
"thoughts🤔" Thanks to the output of controlnet. I can answer the question better.
"actions🚀" []
"value👉"

Absolutely! Utilizing the potent ControlNet model, I've created an image that resembles your underwater description. Check out the results below. I trust you'll find it appealing!

Composed Capability: Semantic Segmentation (OpenSeeD)
+ Instruction to Improved Caption Prompt (LLaVA)
+ Image Generation (ControlNet)

Composed Capability: Multimodal Social Media Post

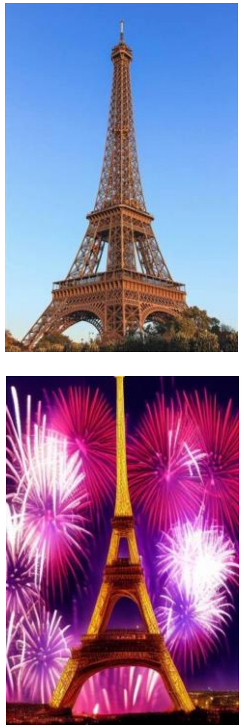
Editing Instruct Prompt Generation (LLaVA)
+ Image Editing (Instruct P2P)
+ Language Generation (LLaVA)



make the image like spring. then generate some attractive texts for Instagram posts

"thoughts🤔" To accomplish the task, I need to utilize the specialized abilities of the instruct pix2pix model for the required image editing.
"actions🚀" [{"API_name": "instruct-pix2pix", "API_params": {"prompt": "make the image like spring."}}]
"value👉"
I will use instruct-pix2pix to help to answer the question. Please wait for a moment.
instruct-pix2pix model outputs: {'message': 'The instruct-pix2pix has processed the image.'}
Please summarize the model outputs and answer my first question: make the image like spring. then generate some attractive texts for Instagram posts
"thoughts🤔" Thanks to the output of instruct-pix2pix. I can answer the question better.
"actions🚀" []
"value👉"

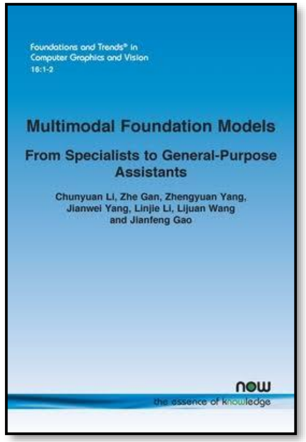
Absolutely! I've got you covered. Let's spruce up that image. See the enhancements below.
And now, let me share some Instagram captions to go with the new image:
1."Spring has sprung, and this stunning lake view is proof! Can you feel the sunshine and fresh air? 🌸🌿🌄 #SpringVibes #NatureLover"
2."Feeling rejuvenated by this breathtaking image of a serene lake in springtime. Mother Nature truly is a wonder! 🌸🏞 #NaturePhotography #SpringFeels"
3."This image transported us straight to a peaceful spring morning with its vibrant colors and tranquil surroundings. Let's take a moment to appreciate the beauty of nature! 💯🌿 #SpringMood #NatureLove"
4."Who needs a spring break when you have a view like this? 😍 This image perfectly captures the beauty of nature in all its glory. 🌸🌄 #NatureIsAmazing #Springtime"
5."Spring has arrived with a bang, as evidenced by this picture-perfect view of pristine waters and lush greenery. Can you feel the excitement in the air? 🌸🌿🛖 #SpringFeels #NatureBeauty"



i want to post a message on Instagram. add some firework to the image, and write an attractive post for my ins.

"thoughts🤔" My success in completing the task depends on my capability to utilize the functionalities of the instruct pix2pix model, which has been tailored for this type of image editing.
"actions🚀" [{"API_name": "instruct-pix2pix", "API_params": {"prompt": "add some fireworks"}}]
"value👉"
I will use instruct-pix2pix to help to answer the question. Please wait for a moment.
instruct-pix2pix model outputs: {'message': 'The instruct-pix2pix has processed the image.'}
Please summarize the model outputs and answer my first question: i want to post a message on Instagram. add some firework to the image, and write an attractive post for my ins.
"thoughts🤔" Thanks to the output of instruct-pix2pix. I can answer the question better.
"actions🚀" []
"value👉"

I have used Instruct pix2pix technique to make the necessary edits and the revised image is now available for your viewing pleasure.
Now, let me share with you some exciting captions to go with our new image on Instagram:
1."Paris shines extra bright with the added sparkle of fireworks 🎆✨#CityOfLight #FireworksFest"
2."Bonjour, Paris! 🇫🇷 It's even more magical with the dazzling fireworks illuminating the night sky 🌟 #TravelGoals"
3."Dreaming of a Parisian adventure? 🌇🇫🇷 Our stunning new image with fireworks will inspire your wanderlust 🚶🌅 #Wanderlust"
4."Who needs to wait for New Year's Eve to feel the magic of fireworks? 🎆🌟 Our new image of the Eiffel Tower sparkling in the night sky will transport you to the City of Love ❤ #BonJourParis"
5."Experience the enchanting beauty of Paris like never before! 🤩 Our new image with added fireworks will transport you to a fairytale setting 🧚 #MidnightMagic"



Foundations and Trends® in
Computer Graphics and Vision
16:1-2

Multimodal Foundation Models

From Specialists to General-Purpose Assistants

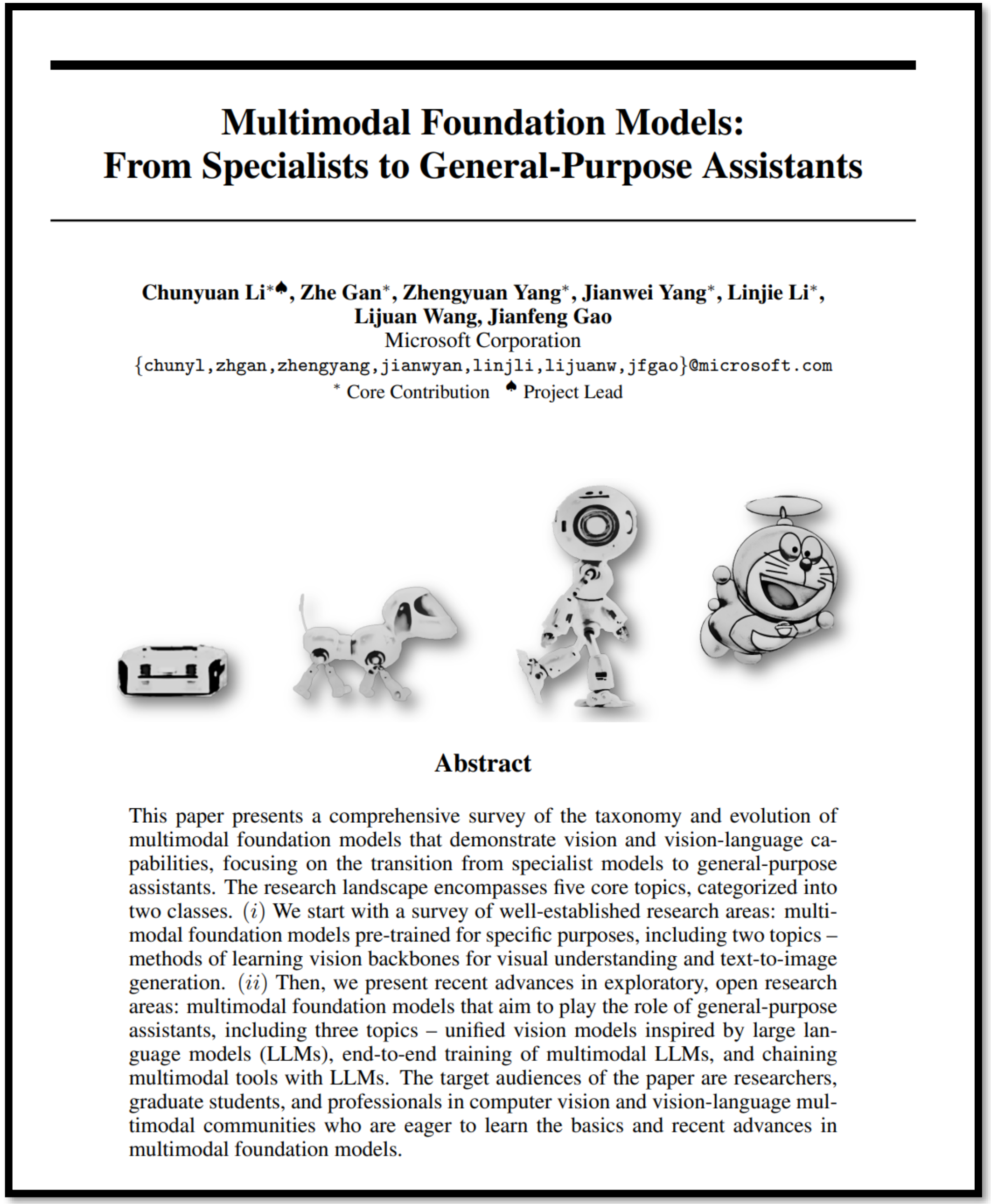
Chunyuan Li, Zhe Gan, Zhengyuan Yang,
Jianwei Yang, Linjie Li, Lijuan Wang
and Jianfeng Gao

now
the essence of knowledge

☐ Multimodal Foundation Models:
From Specialists to General-Purpose Assistants

https://arxiv.org/abs/2309.10020

Foundations and Trends® in Computer Graphics and Vision, 2024



Multimodal Foundation Models:
From Specialists to General-Purpose Assistants

Chunyuan Li*♠, Zhe Gan*, Zhengyuan Yang*, Jianwei Yang*, Linjie Li*,
Lijuan Wang, Jianfeng Gao
Microsoft Corporation
`{chunyl,zhgan,zhengyang,jianwyan,linjli,lijuanw,jfgao}@microsoft.com`
* Core Contribution ♠ Project Lead

Abstract

This paper presents a comprehensive survey of the taxonomy and evolution of multimodal foundation models that demonstrate vision and vision-language capabilities, focusing on the transition from specialist models to general-purpose assistants. The research landscape encompasses five core topics, categorized into two classes. (*i*) We start with a survey of well-established research areas: multimodal foundation models pre-trained for specific purposes, including two topics – methods of learning vision backbones for visual understanding and text-to-image generation. (*ii*) Then, we present recent advances in exploratory, open research areas: multimodal foundation models that aim to play the role of general-purpose assistants, including three topics – unified vision models inspired by large language models (LLMs), end-to-end training of multimodal LLMs, and chaining multimodal tools with LLMs. The target audiences of the paper are researchers, graduate students, and professionals in computer vision and vision-language multimodal communities who are eager to learn the basics and recent advances in multimodal foundation models.