

A Vision-and-Language Approach to Computer Vision in the Wild

**Building a General-Purpose Assistant in the Visual World
Towards Building and Surpassing GPT-4V**

April 2024

Chunyuan Li

<https://chunyuan.li>

□ Outline

1 Computer Vision in the Wild (CVinW)

Definition and Current Status

2 Basics: LLaVA-1.0 (NeurIPS 2023, Oral)

Visual Instruction Tuning

3 Recent Advances: LLaVA Family

- Applications: LLaVA-Med (NeurIPS 2023, D&B Track, Spotlight)
- New Capabilities: LLaVA-Interactive & LLaVA-Plus
- Performance: LLaVA-1.5 (CVPR 2024) & LLaVA-NeXT

What is Computer Vision in the Wild (CVinW) ?



Developing a transferable foundation model/system that can *effortlessly* adapt to a *large range of visual tasks* in the wild.

It comes with two key factors:

- 1 The task transfer scenarios are broad
- 2 The task transfer cost is low.



GitHub <https://github.com/Computer-Vision-in-the-Wild>



YouTube <https://www.youtube.com/@cvnw>

Examples of Vision Tasks: An Image Understanding Perspective

Language A dog lying on the grass next to a frisbee

Image



[dog, grass, frisbee]

(a) Image Classification



dog, grass, frisbee

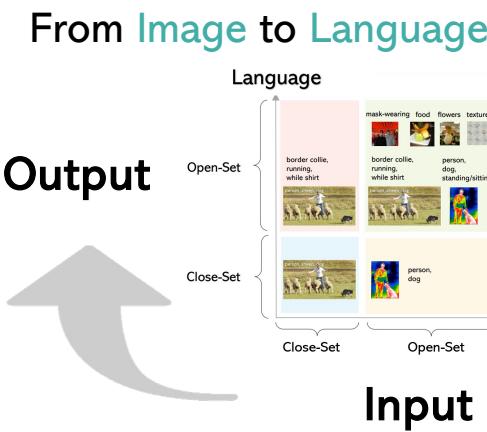
(b) Object Detection



dog, grass, frisbee

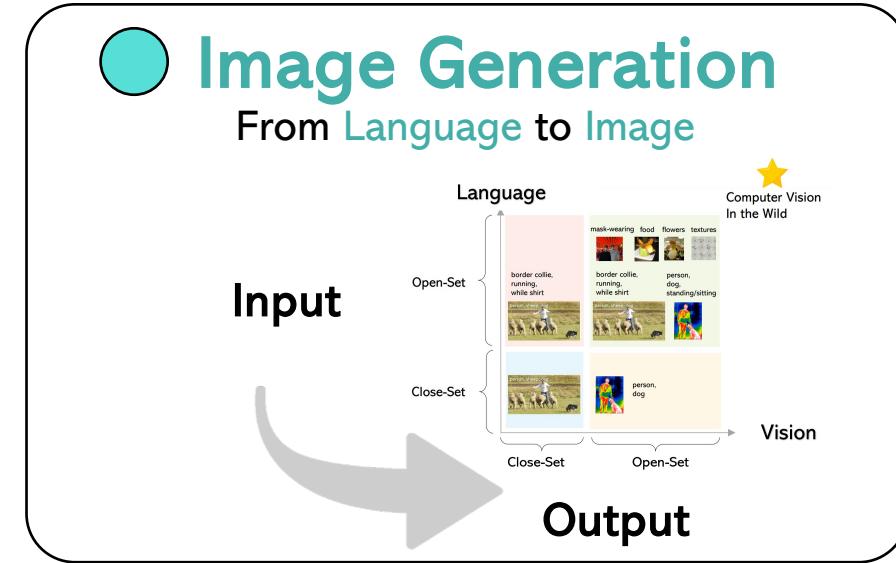
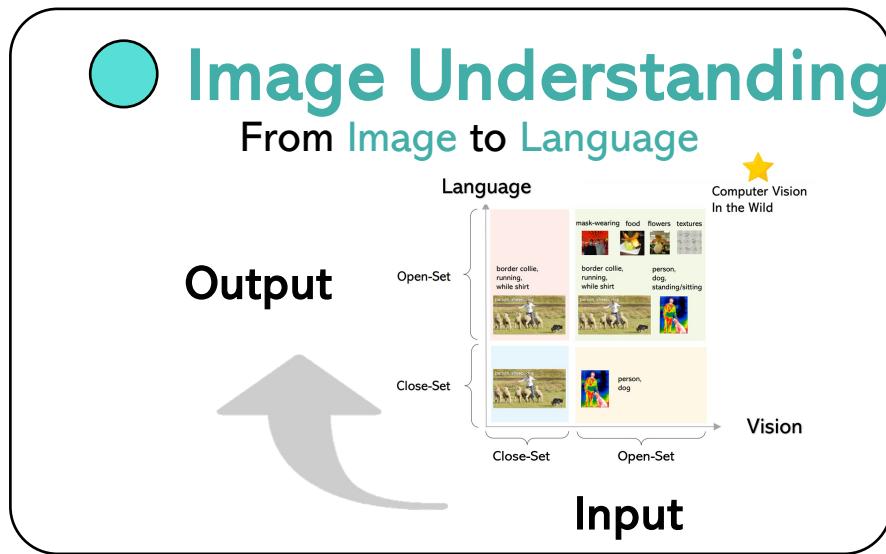
(c) Segmentation

Image Understanding



	A Data View	A Modeling View	A Benchmark View
Increased Data Scale	Image-Only	EsViT (ICLR 2022)	 ICinW Image Classification in the Wild
Classification	Image-Language (Billion)	UniCL & Florence (CVPR 2022)	 ODinW Object Detection in the Wild
Detection	Box-Language (Million)	GLIP (CVPR 2022)	 CVPR TEL AVIV 2022
Segmentation	Mask-Language (Thousands)	X-Decoder (CVPR 2023)	 SGinW Segmentation in the Wild
Increased Semantic Richness	Unified Vision-Language Data	LLaVA (NeurIPS 2023 & CVPR 2024)	 CVPR JUNE 17-21, 2024 SEATTLE, WA

From Image Understanding to Image Generation



②

A General-Purpose Visual Assistant

Towards Building GPT-4V: Image-to-text generation

❑ Visual Instruction Tuning (LLaVA)

NeurIPS 2023 (Oral Presentation) Project: <https://llava-vl.github.io/>

Language Generation: Large Language Models (LLM)



GPT-2

GPT-3

ChatGPT
InstructGPT

GPT-4

What's new?

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

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

In-context-learning
Chain-of-thoughts (CoT)
Instruction-Following
Multimodal Input with image

**Open Source
Community**

LLaMA



Alpaca



Vicuna



Our Contributions

**GPT-4-
LLM**

Data-Centric, NOT Model-Centric

LLaVA

Instruction Tuning with GPT-4 <https://instruction-tuning-with-gpt-4.github.io/>

Baolin Peng*, Chunyuan Li*, Pengcheng He*, Michel Galley, Jianfeng Gao (* Equal contribution)

Self-Instruct with Strong Teacher LLMs

	LLaMA	Alpaca	Vicuna	GPT-4-LLM
Teacher				
Instruction-following Data	None	GPT-3.5	ShareGPT (Human & GPT)	GPT-4 (text-only)
	52K	700K (70 conversions)		<ul style="list-style-type: none">• 52K English & Chinese• Feedback Data

→ LLM Chatbot

→ Reward Model

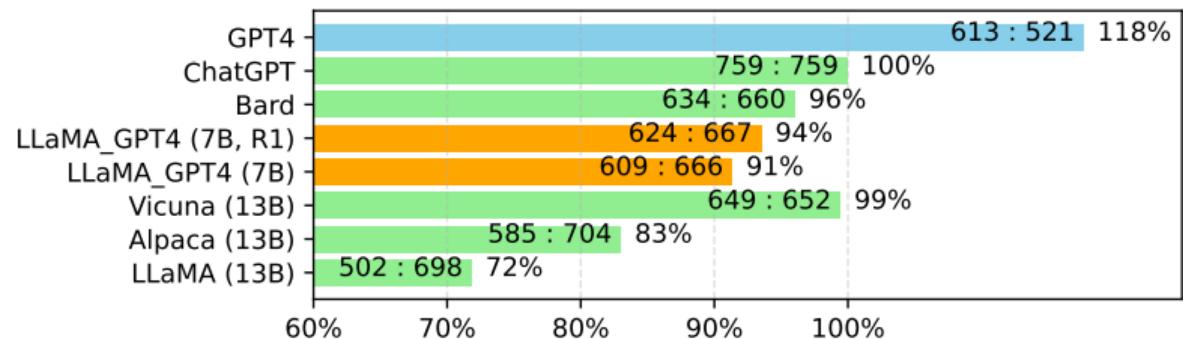
Results on Chatbot

Evaluation Metric: Ask GPT-4 to rate the two model responses (1-10), then compute the ratio, i.e. relative score

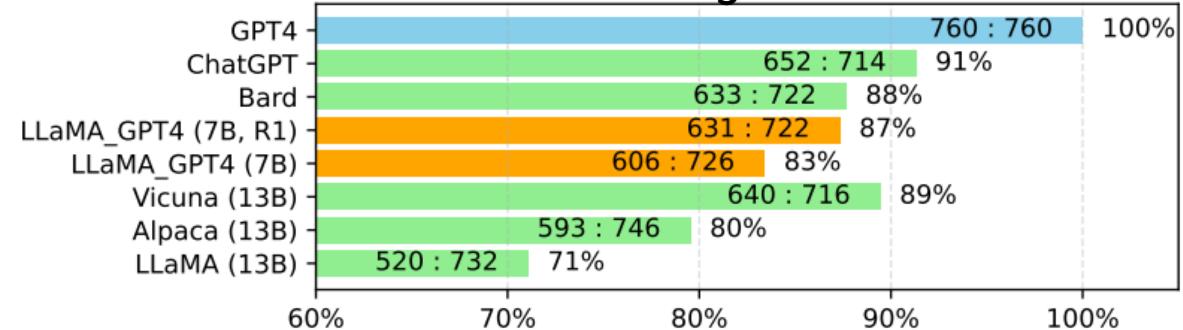
Findings:

- A VERY CONSISTENT Evaluation Metric !
- Our model LLaMA-GPT4 is performing closely to SoTA opensourced Chatbot, though with smaller training data and model size.

All chatbots against ChatGPT



All chatbots against GPT-4

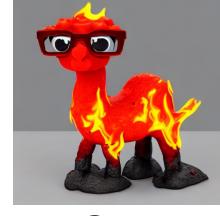


Visual Instruction Tuning with GPT-4

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

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

Self-Instruct with Strong Teacher LLMs But No Teacher is available on multiGPT4?

	LLaMA	Alpaca	Vicuna	GPT-4-LLM	LLaVA
Teacher					
	GPT-3.5	ShareGPT (Human & GPT)		GPT-4 (text-only)	GPT-4 (text-only)
Instruction-following Data	None	52K	700K (70 conversions)	<ul style="list-style-type: none">• 158K multimodal instruction following data (First & High Quality)	
				<p>→ Multimodal Chatbot</p>	
				<p>Large Language and Vision Assistant</p>	

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

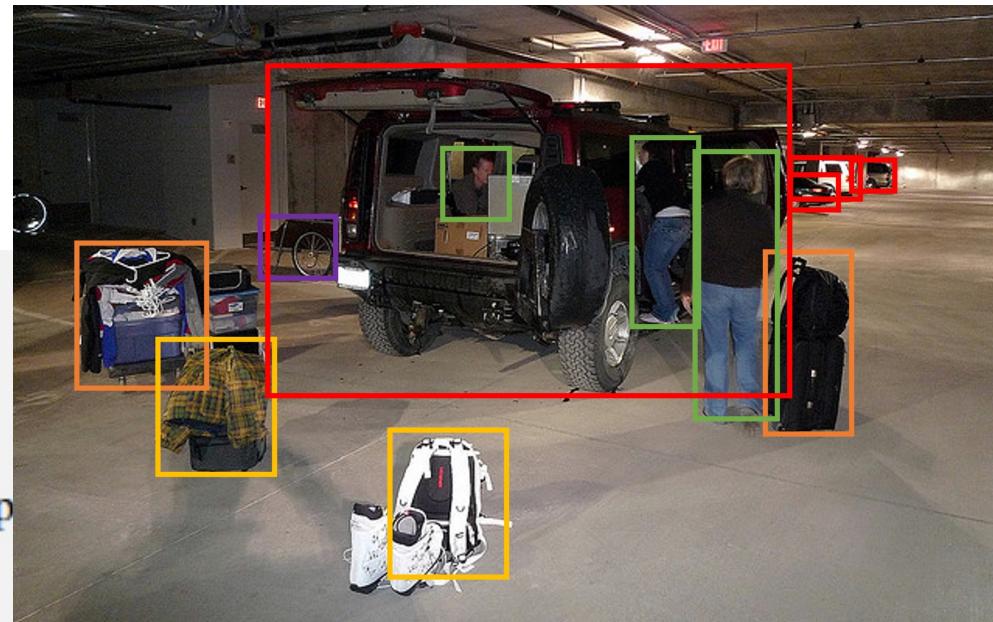
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area

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.

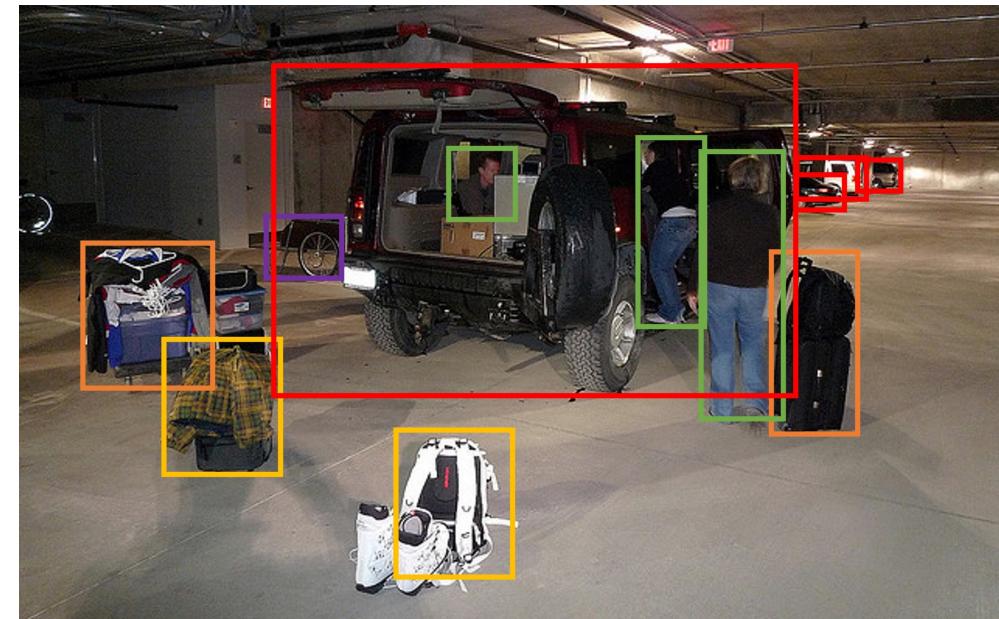
In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center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

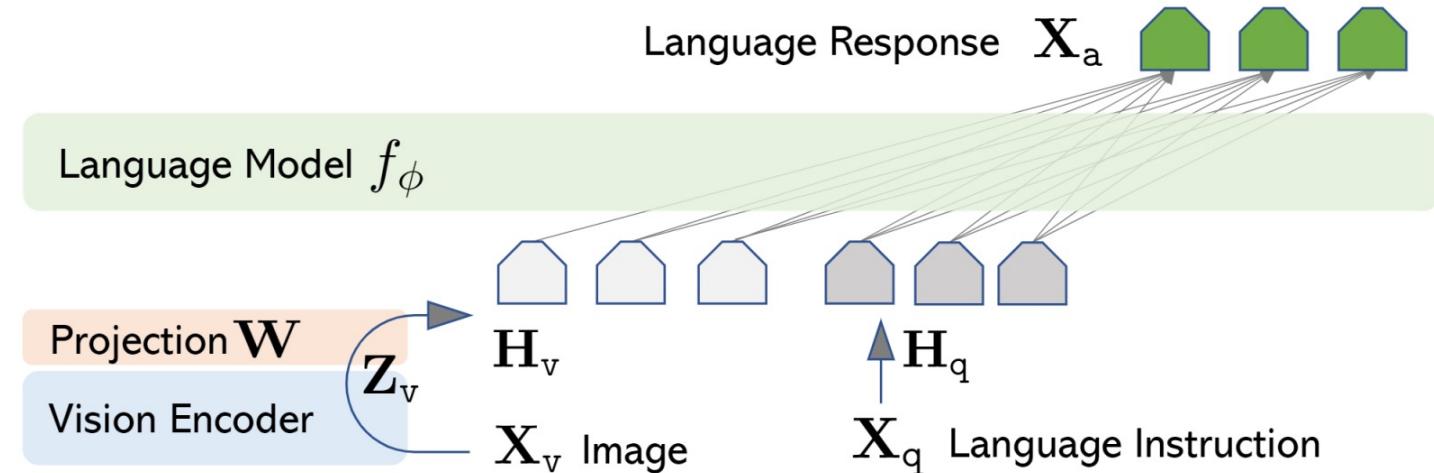
Question: What challenges do these people face?

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.

Visual Chat: Towards building multimodal GPT-4 level chatbot

- LLaVA-Bench (In-the-Wild) LLaVA/docs/LLaVA_Bench.md

Approach	Conversation	Detail	Reasoning	Overall
Bard-0718	83.7	69.7	78.7	77.8
Bing-Chat-0629	59.6	52.2	90.1	71.5
LLaVA-13B-v1-336px-0719 (beam=1)	64.3	55.9	81.7	70.1
LLaVA-13B-v1-336px-0719 (beam=5)	68.4	59.9	84.3	73.5

An evaluation dataset with 24 unseen images, 60 new language-image instructions

- More Update on July 19 (LLaMA 2 released):
 - LLaVA from LLaMA 2 (in 7 hours)
 - LoRA and QLoRA training



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

User
LLaVA

What is unusual about this image?

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
LLaVA

What's happening in the scene?

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
GPT-4 [34]

What is unusual about this image?

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
BLIP-2

What is unusual about this image?

a man is sitting on the back of a yellow cab

User
OpenFlamingo

What is unusual about this image?

The man is drying his clothes on the hood of his car.

Example 1: Extreme Ironing

Visual input example, Chicken Nugget Map:

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



Strong Visual Reasoning Ability

Example 2: Chicken Nugget Map

User
LLaVA

Can you explain this meme in detail?

The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all is..." The main image features a tray of cooked chicken nuggets, with the chicken pieces positioned to resemble the continents and islands on a world map. The meme playfully suggests that the chicken nuggets represent the Earth, and the various locations depicted in the photo are actually chicken nugget versions of different places. The punchline of the meme appears in smaller text at the bottom, reading "I mean, it's not the real Earth, but how beautiful it is all is." This lighthearted and entertaining meme combines food and humor, using the image of chicken nuggets to create a funny and imaginative representation of the world.

User
GPT-4 [32]

Can you explain this meme?

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.

User
BLIP-2

Can you explain this meme in detail?

sometimes i just look at pictures of the earth from space and marvel how beautiful it is

User
OpenFlamingo

Can you explain this meme in detail?

It's a picture of a chicken nugget on the International Space Station.

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 1: CVPR & Vancouver

what is image about?



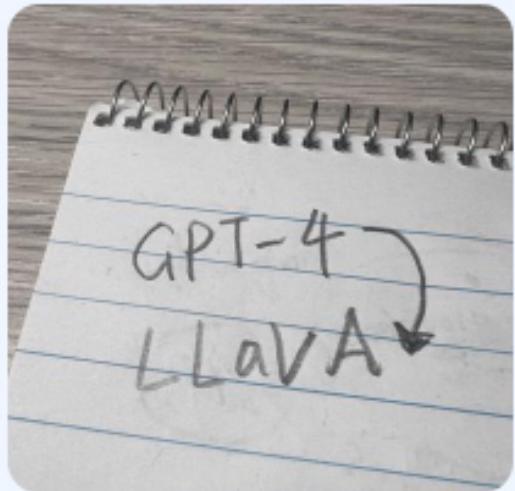
The image features a logo for "Computer Vision in the Wild," which is likely an educational or research program focused on computer vision and its applications in a wild setting. The logo consists of a large letter "C" with a smaller letter "v" on top of it, standing for "Computer Vision." The rest of 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 2: CVinW workshop

Strong OCR Ability

what is written in the photo?



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

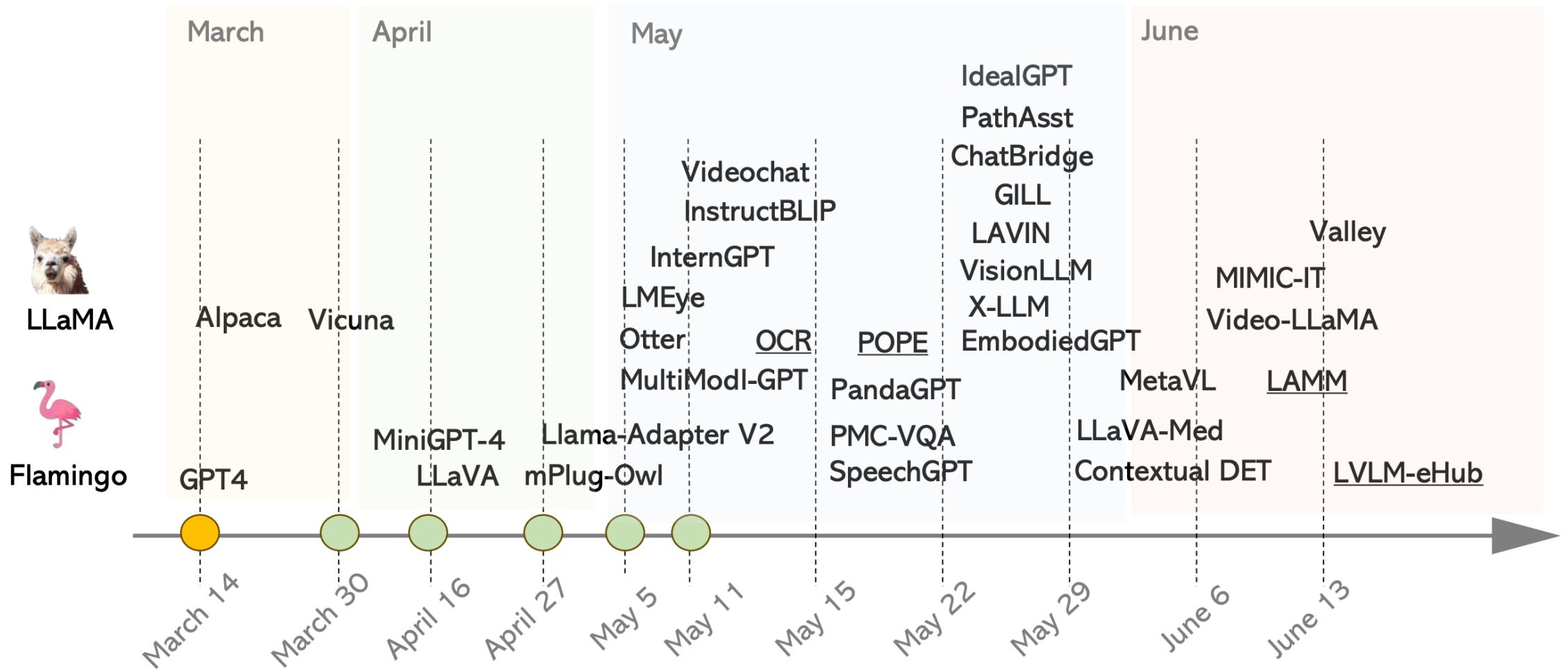
Example 3: LLaVA

3

LLaVA Family 2023 & Beyond

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

An Emerging Topic in 2023 - current

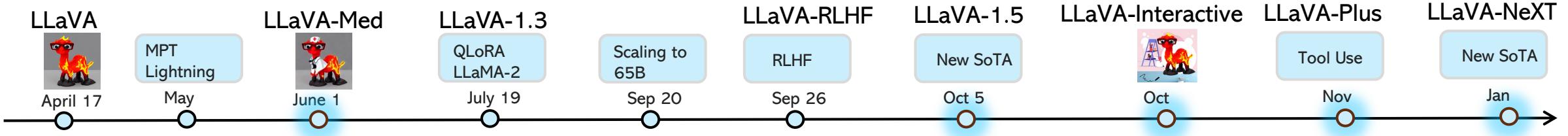


Large Language and Vision Assistant

LLaVA (llava-vl.github.io)

LLaVA is the first open-source project to build GPT-4V like model, inspiring dozens of projects

- 1.2K+ citations, and 15.5K+ GitHub stars, in ~12 months
- A cost-efficient open-source counterpart/recipe to GPT-4V (1 day on 8 A-100 GPUs)



LLaVA-Med: The first open-source healthcare LMM

Stage 1
Medical Concept Alignment
7 Hours
1 epoch on 600K samples

Stage 2
Medical Instruction Tuning
8 Hours
3 epochs on 60K samples

Visual Chat

Grounded Editing

Interactive Segmentation

Upload or Generate an Image

LLaVA Interactive: A demo to interact with users beyond language: stroke input and image output

Visual Chat

Grounded Editing

Interactive Segmentation

Upload or Generate an Image

LLaVA-Plus: Learning to use tools to expand capabilities

Human

Assistant

Skill Repository

Response

Tool Use

Tool Output

Input Image

Conditional Gen

Editing

Point to multi-level Seg

Visual Generation

Visual Understanding

Visual Interaction

Box to Seg

Stroke to Seg

External Knowledge

Search

Social Media Post

Composition

LLaVA-1.5 & LLaVA-NeXT: New SoTA among open-source LMM with a cost-efficient solution

Performance: SoTA on 11 Benchmarks

Benchmark	LLaVA-1.5	LLaVA-NeXT
MM-WT	76.9	77.5
VQA2	80.0	81.3
GOA	77.5	78.9
VQA1	73.6	74.9
SEED-Bench	50.2	53.1
MMBench-CN	60.6	61.1
POPE	1467.5	1551.3
HMBench	87.7	88.5
TextVQA	1229.9	1235.7
SQA-IMG	69.2	71.6
MHE	14.5	15.3

High Sample-Efficiency

A Simple Architecture

Large Language Model (VLLM v1.1.139)

Visual Encoder (CLIP VIT-L/14)

Text Encoder & Prompting

Image

Language Instruction

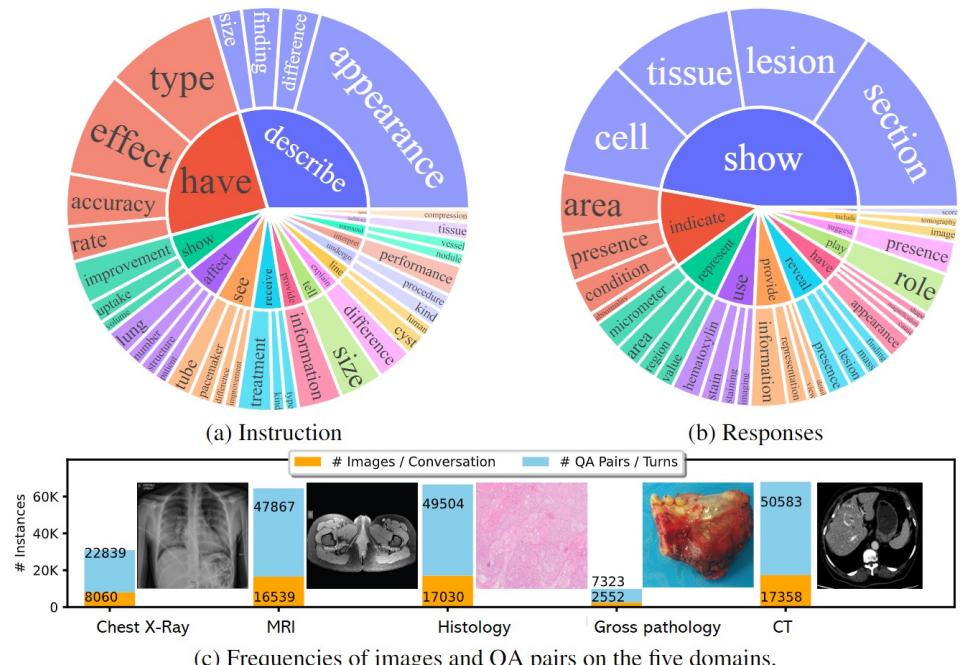
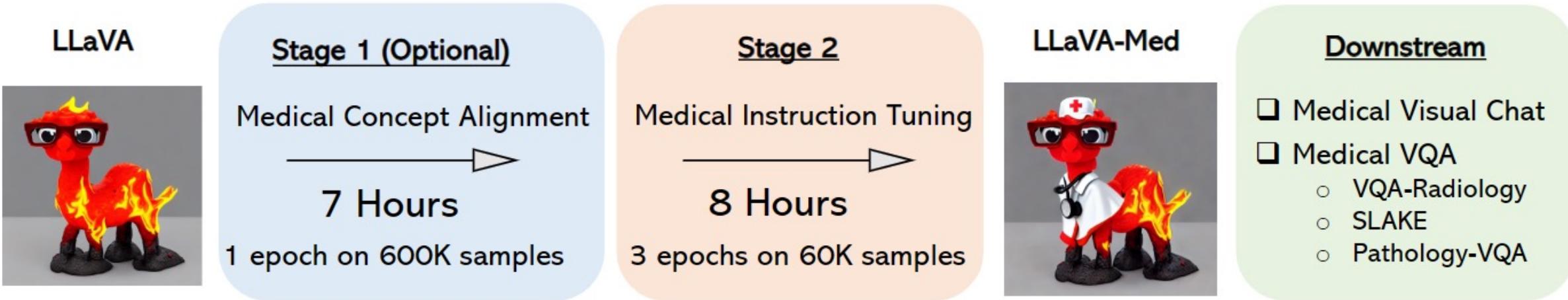
LLaVA Series:

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>



Visual input example, Biomedical image:	
User LLaVA-Med	What is this image about? 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 LLaVA-Med	What is abnormal about this image? 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 LLaVA	What is this image about? 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 LLaVA	What is abnormal about this image? 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 Series:

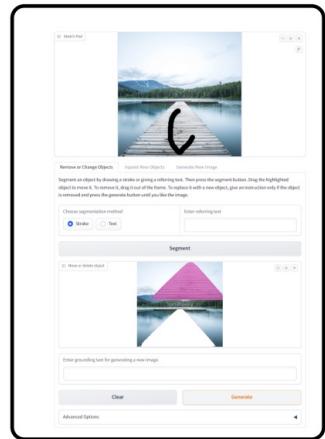
New Capabilities

LLaVA-Interactive

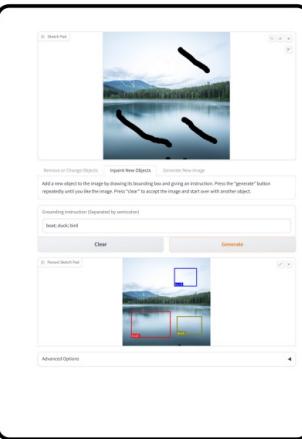
An All-in-One Demo for Image Chat, Segmentation and Generation/Editing

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

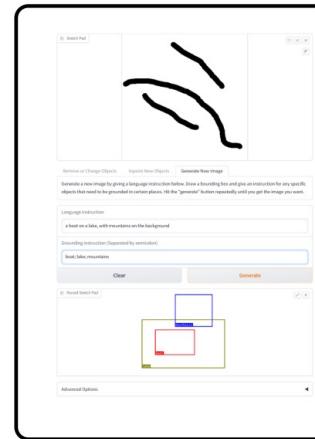
Go Beyond Language Prompt in Visual Interaction



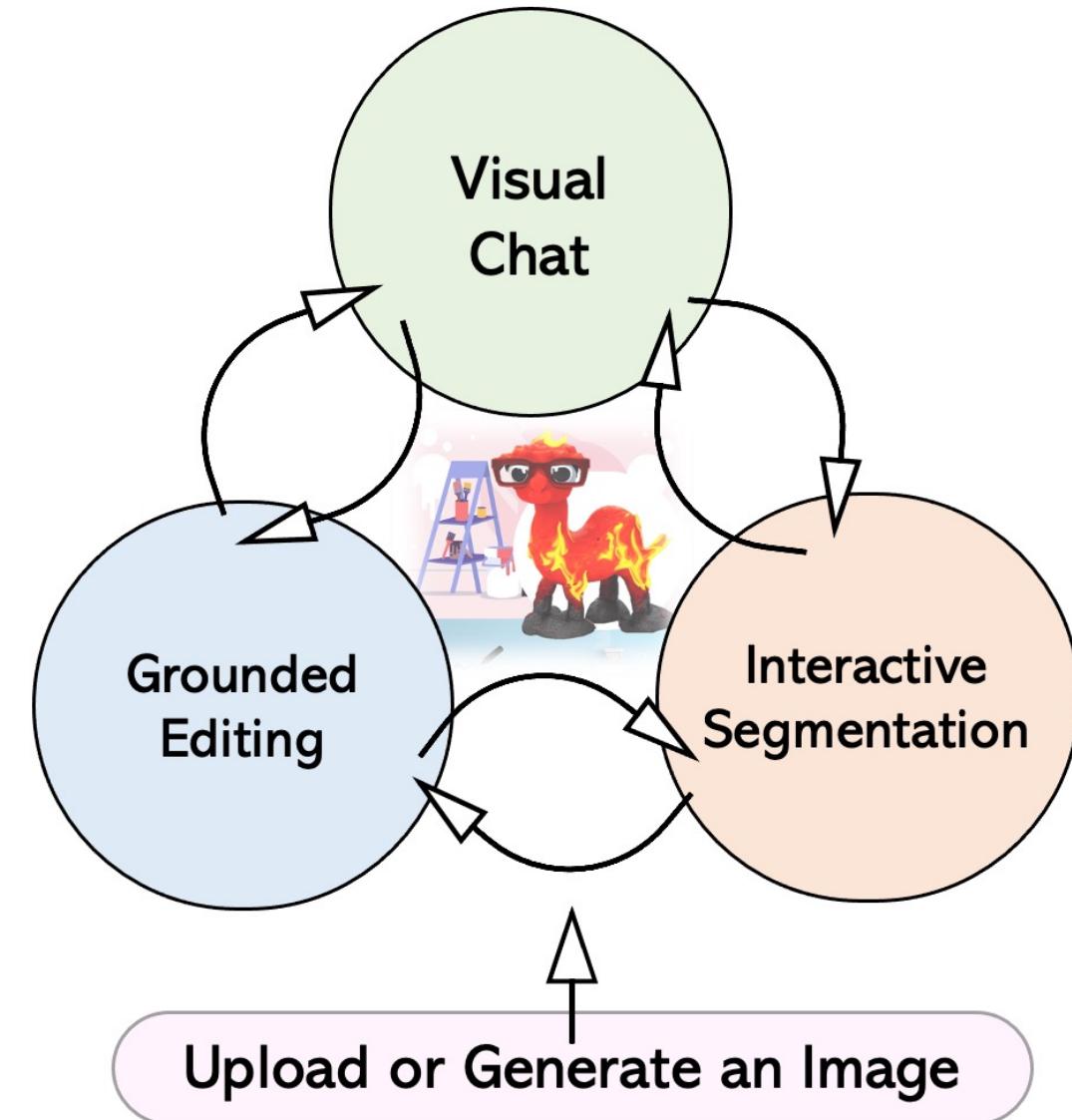
(b) Remove or Change Objects



(c) Inpaint New Objects



(d) Generate New Image



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

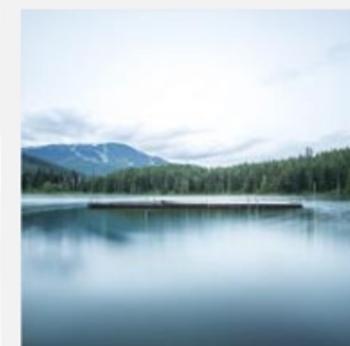


User: [Stroke on the pier]



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?

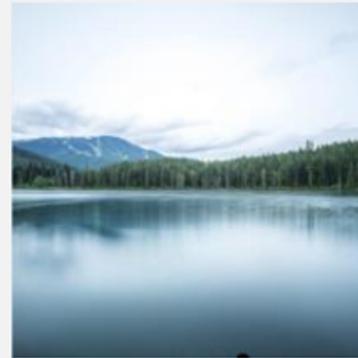
Assistant: The part of the picture that is not visually appealing to the viewer is the dock. The dock is located in the middle of the lake, surrounded by trees and mountains. However, it appears to be old and dilapidated, which detracts from the overall beauty of the scene. The presence of the dock in 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

8

Text Prompt for Segmentation and Editing

9

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.

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



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



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 (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 how do you think about the picture design as a kid cloth?

User 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

Kid Cloth Picture Design (2/2)



User

(Inpaint New Objects) [Draw boxes for "boat; lake; snow mountain; tent"] [Generate]



Assistant

User

Assistant

how do think about the current picture design for a kid cloth?

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.

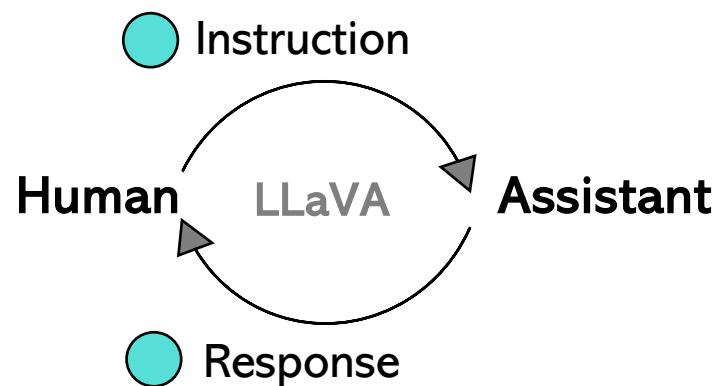
Figure 5: (2/2) Picture design for a kid cloth.

Kid Cloth Picture Design

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



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/>

Visual Generation

Visual Interaction

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

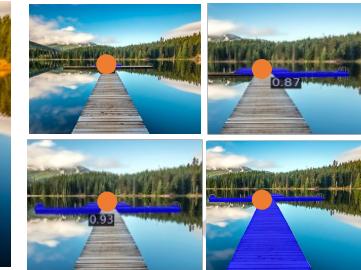
Conditional Gen.



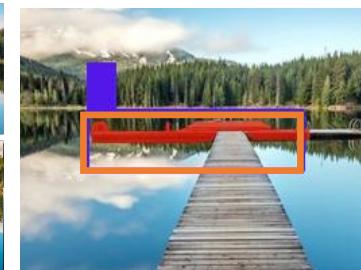
Editing



Point to multi-level Seg



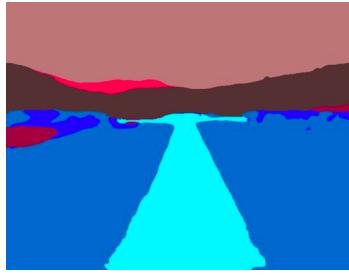
Box to Seg



Stroke to Seg



Object Seg and Det



Semantic Seg

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

OCR

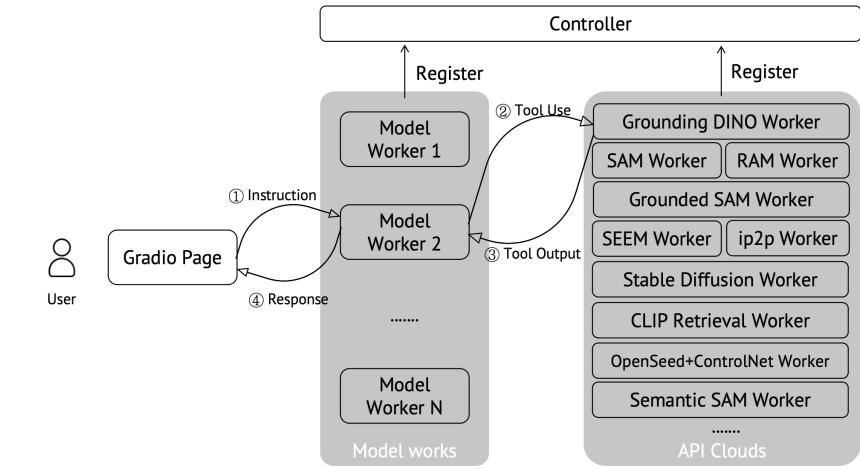
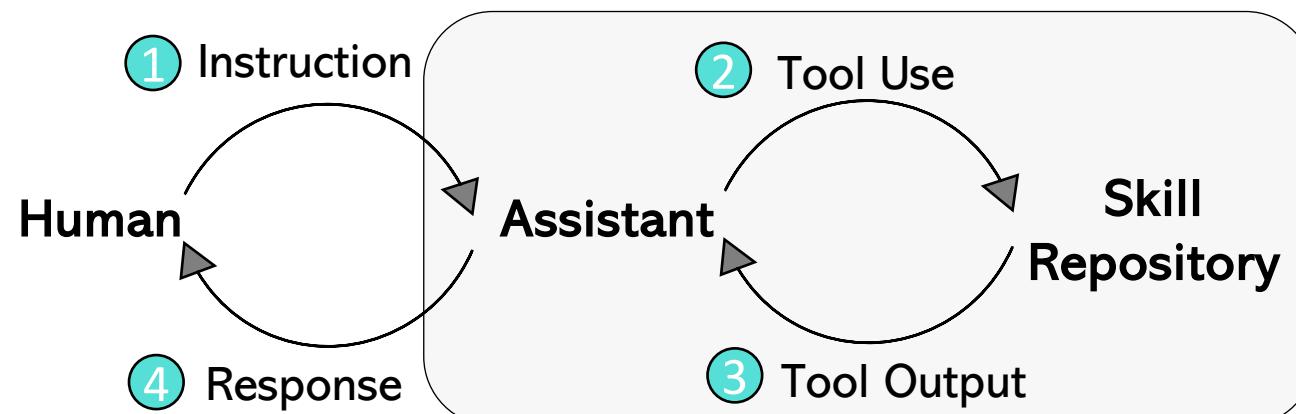
The lake in question is **Lost Lake**, which is located in **Whistler, British Columbia, Canada**.



This image is
paw-some!

Who says cats can't
be outdoorsy? 
#naturelover

Huma-AI Interaction Process



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.



**Detection (Grounding DINO)
+ Reasoning (LLaVA)**

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 moon 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.

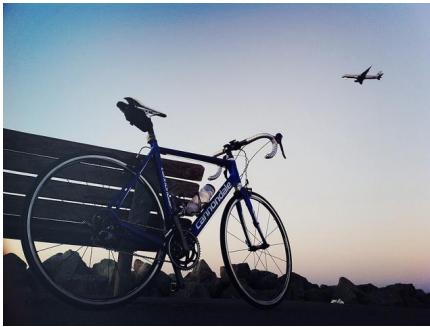
what is the food?

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)



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"

Composed Capability: Multimodal Social Media Post

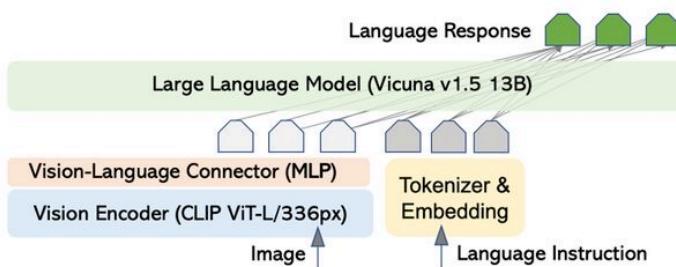
Editing Instruct Prompt Generation (LLaVA)
+ Image Editing (Instruct P2P)
+ Language Generation (LLaVA)

LLaVA Series:

Improved Performance

LLaVA-1.5: Improved Baselines with Visual Instruction Tuning

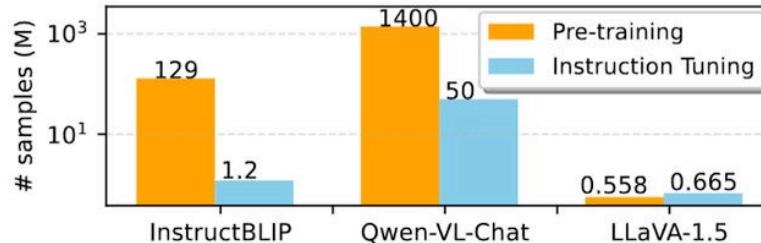
□ A Simple Architecture



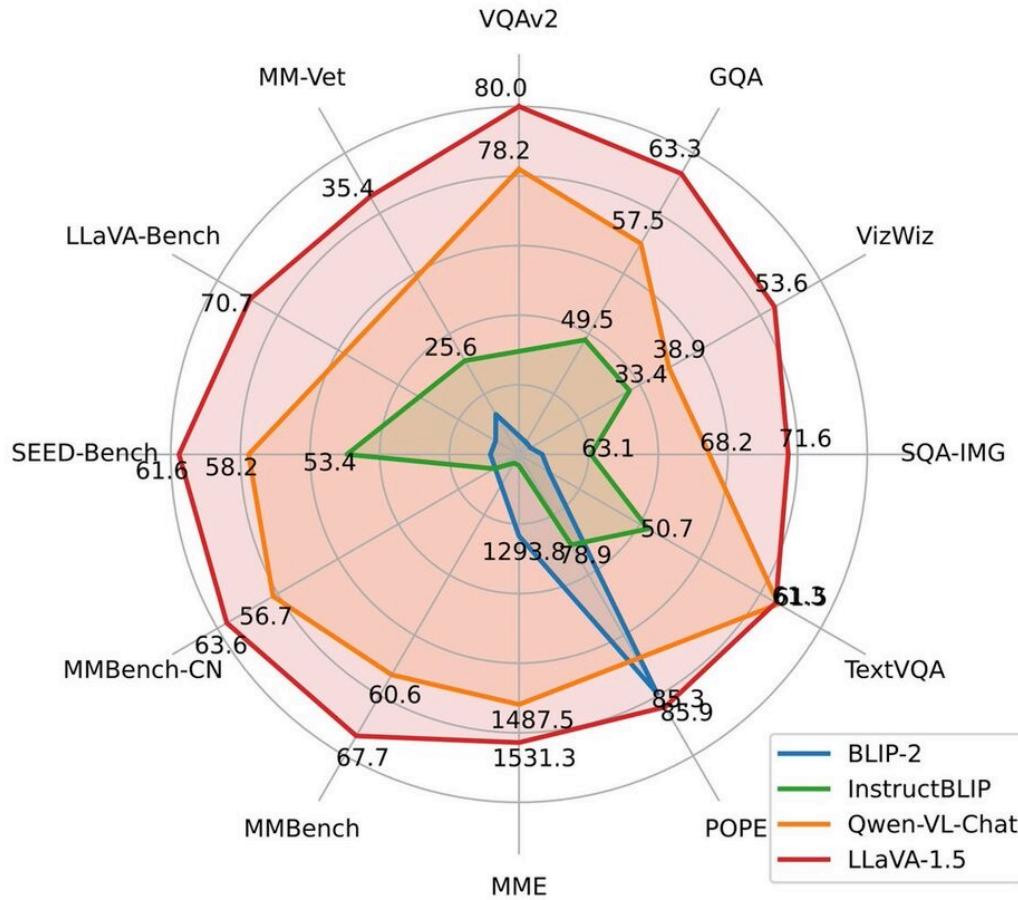
□ Data Mixture

- Academic task-oriented data
- Response formatting prompts

□ High Sample-Efficiency



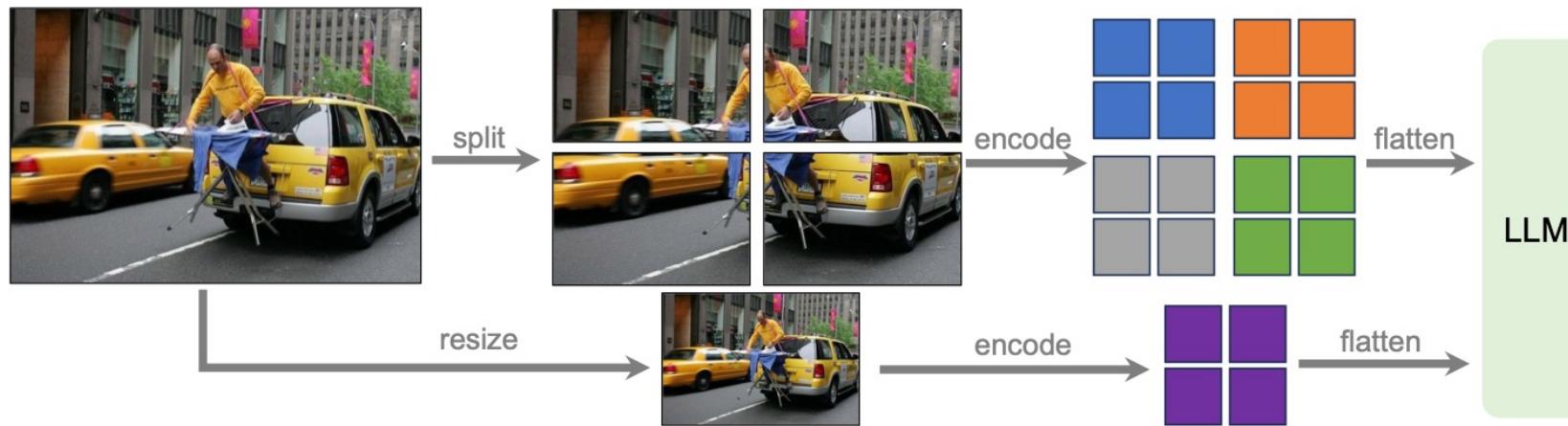
□ Performance: SoTA on 11 Benchmarks



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

<https://llava-vl.github.io/blog/2024-01-30-llava-next/> (Previously LLaVA-1.6)

(1) Dynamic High Resolution



(2) Data Mixture

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

(3) Scaling LLM backbone

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

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

- CVPR 2023 Tutorial:
Recent Advances in Vision Foundation Model
<https://vlp-tutorial.github.io/>

- Multimodal Foundation Models:
From Specialists to General-Purpose Assistants
<https://arxiv.org/abs/2309.10020>

*Foundations and Trends® in Computer
Graphics and Vision, 2023 (To appear)*

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,jianwyen,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.