

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

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

October 2023

Chunyuan Li  
Deep Learning Team  
Microsoft Research, Redmond

<https://chunyuan.li>

## □ Outline

### ① Computer Vision in the Wild (CVinW)

Definition and Current Status

### ② Text-to-Image Generation: GLIGEN (CVPR 2023)

A. Better Alignment with Human Intent

B. Much Lower Development Cost

### ③ Image-to-Text Generation

A. Visual Instruction Tuning with GPT-4 (LLaVA)

B. LLaVA Family

### ④ Towards Surpassing Multimodal GPT-4

# 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>



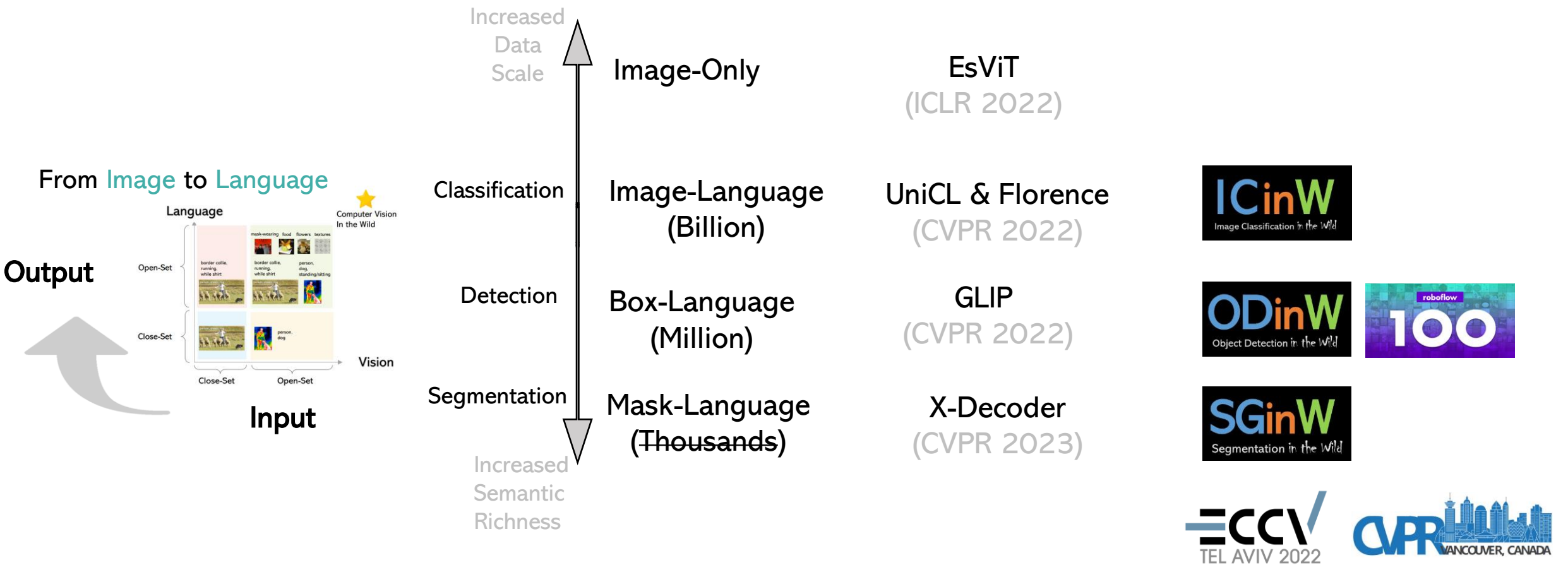
**YouTube** <https://www.youtube.com/@cvinw>

# Image Understanding

## A Data View

## A Modeling View

## A Benchmark View



## ② Text-to-Image Generation

# GLIGEN

Grounded Language-to-Image Generation

Acknowledgements to the V-Team Members

Yuheng Li<sup>1§</sup>, Haotian Liu<sup>1§</sup>, Qingyang Wu<sup>2</sup>, Fangzhou Mu<sup>1</sup>, Jianwei Yang<sup>3</sup>, Jianfeng Gao<sup>3</sup>,  
Chunyuan Li<sup>3¶</sup>, Yong Jae Lee<sup>1¶</sup>

<sup>1</sup>University of Wisconsin-Madison <sup>2</sup>Columbia University <sup>3</sup>Microsoft

§ Part of the work performed at Microsoft; ¶ Co-senior authors

CVPR 2023





Project: <https://gligen.github.io/>

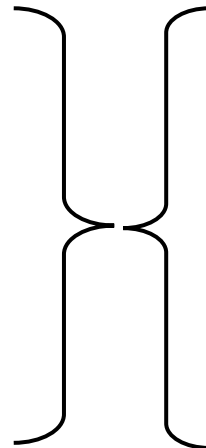
Demo: <https://aka.ms/gligen>

# The Space of Text-to-Image Generative AI

## Research

### Major Tech Companies

-  **OpenAI** DALLE | DALLE2
-  **Microsoft** NUWA
-  **Google** Imagen | Parti | Muse
-  **Meta** CM3 | Make-A-Scene




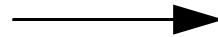
## Production

### Microsoft





### Open Source


-  **stability.ai** Latent Diffusion Models (LDM)  
Stable Diffusion (SD) 1 & 2



### Startup

 [huggingface / diffusers](#) ☆ Star 11.7k

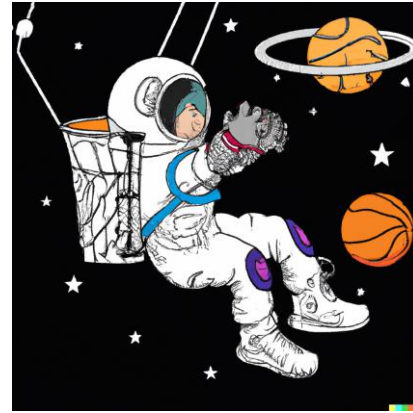
 **DreamStudio**

 **YOU.com**  
The AI Search Engine You Control

# Text-to-Image Generation Models



An astronaut playing basketball with cats in space



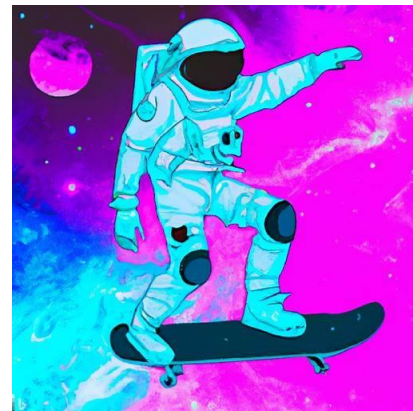
a women hugging a giant cat with a smile in the park, digital art



a pixar style character of a happy elderly man walking a dog



astronaut skate boarder in space, in the style of vaporwave



A castle in a fantasy world with a unicorn and a rainbow, painted in the style of Raphael



# Limitations with Language Prompt Alone

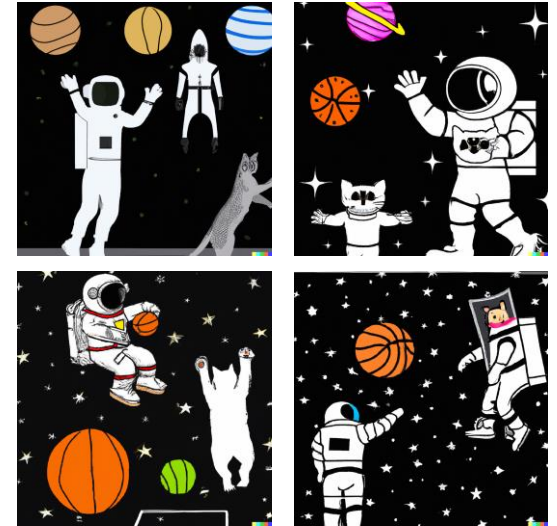
An astronaut playing basketball with cats in space



I'd like to put different objects in specific positions and sizes I want

Adding "blackball is on top, astronaut on left, cat on right"

DALLE2



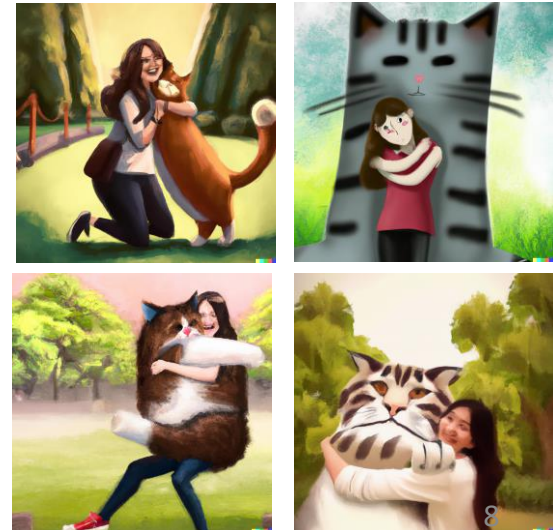
a women hugging a giant cat with a smile in the park, digital art



Emm, the cat is way too giant, I'd like to make the cat as giant as the girl

Adding "the cat as giant as the girl"

DALLE2



"Severely limited in their ability to generate multiple objects or the specified spatial relations"

*Benchmarking Spatial Relationships in Text-to-Image Generation*  
<https://arxiv.org/abs/2212.10015>



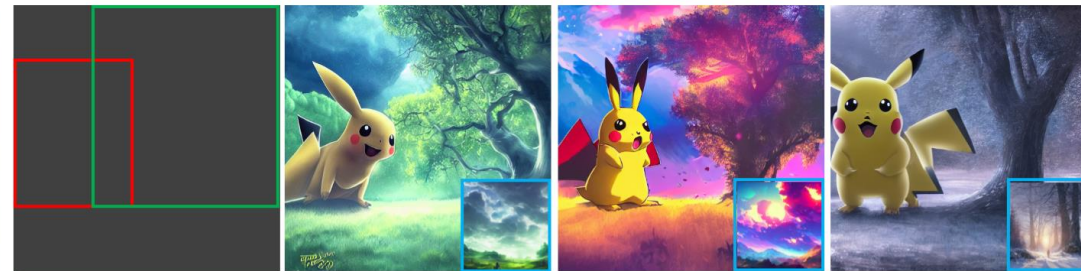
- A. Better Alignment with Human Intent**
- B. Much Lower Development Cost**

Disclaimer: The current GLIGEN is built with open-sourced Stable Diffusion, the technique is transferable to DALLE2





Caption: "a photo of a hybrid between a bee and a rabbit"  
Grounded text: hybrid between a bee and a rabbit, flower



Caption: "Pikachu is under a tree, digital art"  
Grounded text: Pikachu, tree; Grounded style image: blue inset



Caption: "A dog / bird / helmet / backpack is on the grass"  
Grounded image: red inset



Caption: "superman / monkey / Horner Simpson / is scratching its head"  
Grounded keypoints: plotted dots on the left image



Caption: "A vibrant colorful bird sitting on tree branch"  
Grounded depth map: the left image



Caption: "The beautiful scenery of a clam village near the sea"  
Grounded HED map: the left image

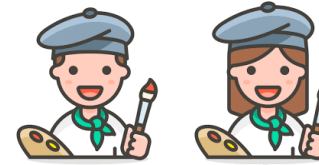


Caption: "Cars park on the snowy street"  
Grounded normal map: the left image



Caption: "A living room filled with lots of furniture and plants"  
Grounded semantic map: the left image

## Human Intents



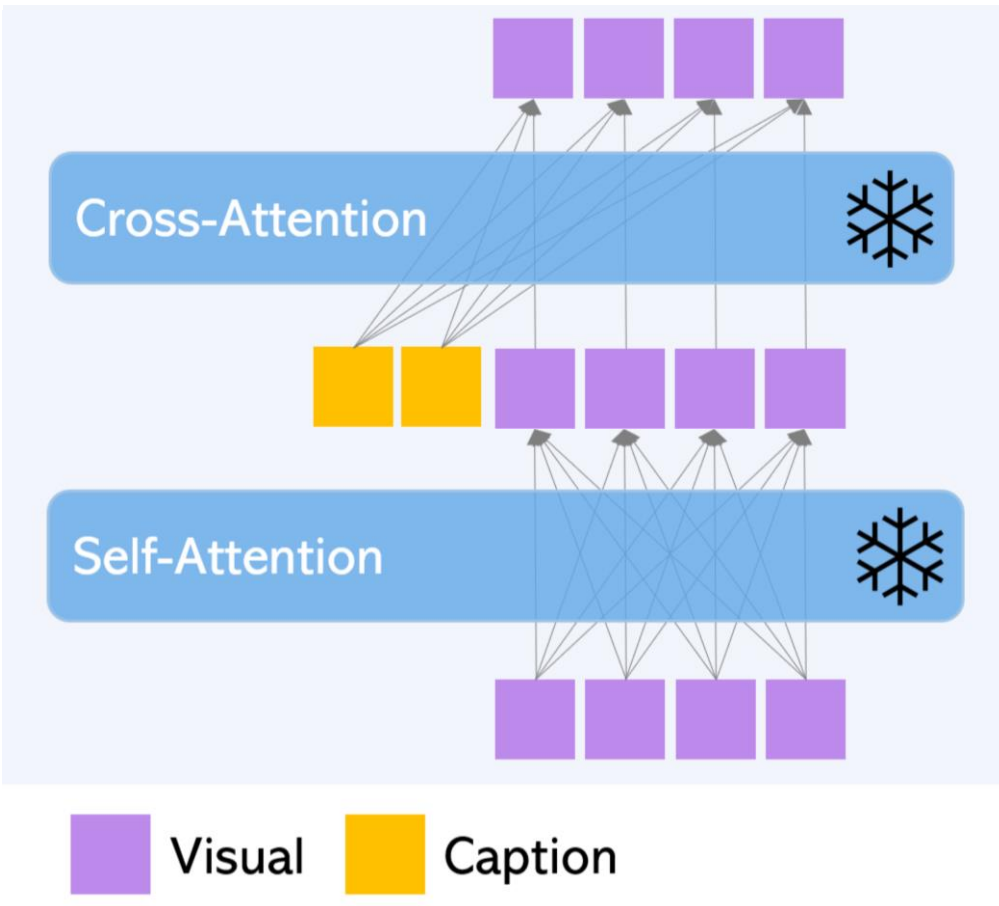
- Spatial: position, size, height/width ...
- Visual: artistic style, customer brand, personalization...

# GLIGEN

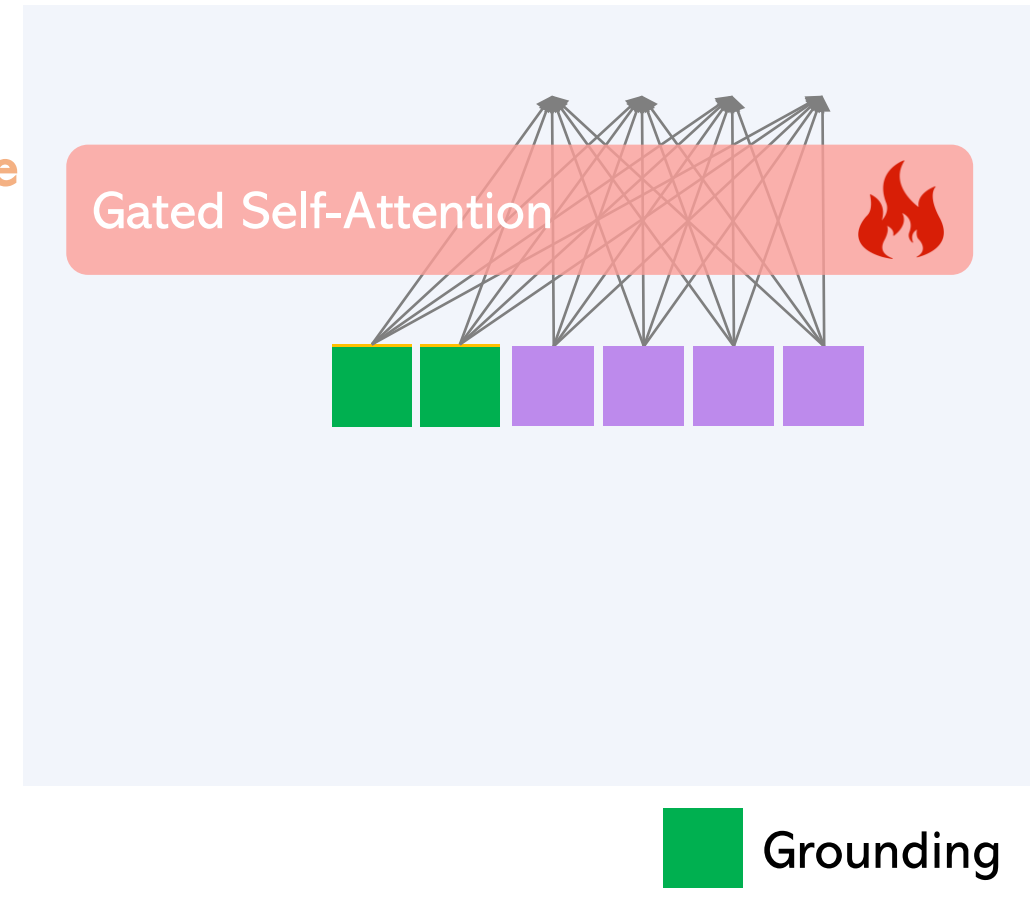


# Modulated Training

## Transformer in Diffusion Models

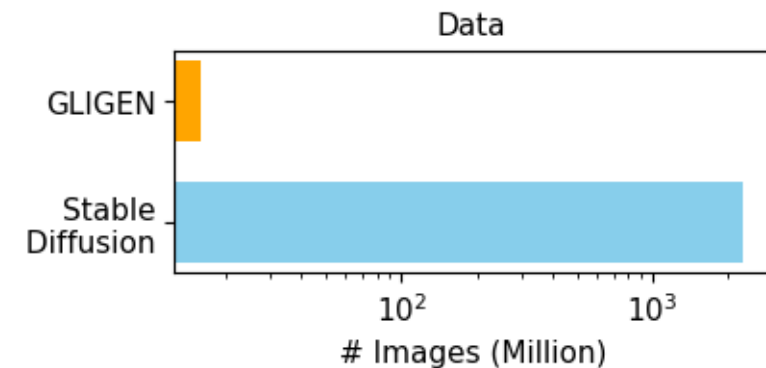
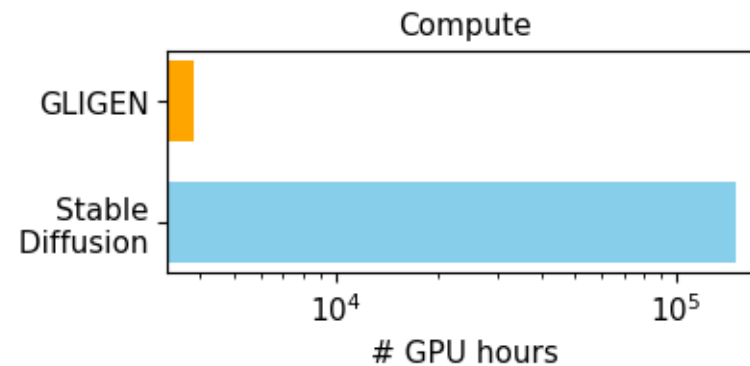


Plug-and-play  
trainable module  
»»»



# Training Cost

	<b>2.5%</b> GPU Hours	<b>0.7%</b> Training Data
<b>GLIGEN</b>	16 V100 GPUs for 10 days (Total: 3,840 GPU hours)	16 million images
<b>Stable Diffusion-v1 (from scratch)</b>	256 A100 GPUs for 24 days (150,000 GPU hours)	2.3 billion images





# The connections to the trends in NLP

*--The similar spirits, but in the image domain*

## A. Better Alignment with Human Intent

## B. Much Lower Development Cost

<p>Language Generation </p> <p><b>GPT3.5</b> → <b>ChatGPT</b></p>
<p>ChatGPT/InstructGPT are better aligned with users than GPT3/3.5 in following human intent and perform the language task that the user wants</p>
<p>Less than <b>2%</b> of the compute and data relative to model pretraining</p>

<p>Image Generation </p> <p><b>DALLE2</b> → <b>GLIGEN</b></p>
<p>GLIGEN is better aligned with users than DALLE2 in following human intent and perform the image generation/editing task that the user wants</p>
<p>Less than <b>3%</b> of the compute and <b>1%</b> of data relative to model pretraining</p>

<https://openai.com/blog/chatgpt>

<https://openai.com/blog/instruction-following/>

<https://openai.com/alignment/>

# ③ 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

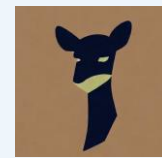
LLaMA



Alpaca



Vicuna



Our Contributions

**GPT-4-LLM**

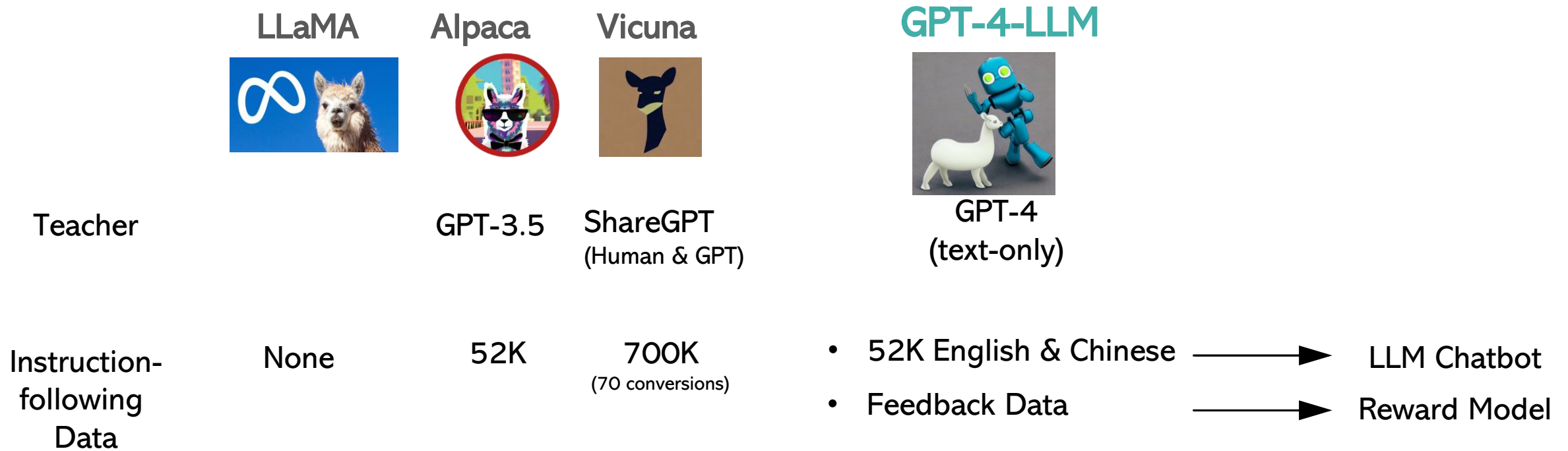
**LLaVA**

Data-Centric, NOT Model-Centric

# 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



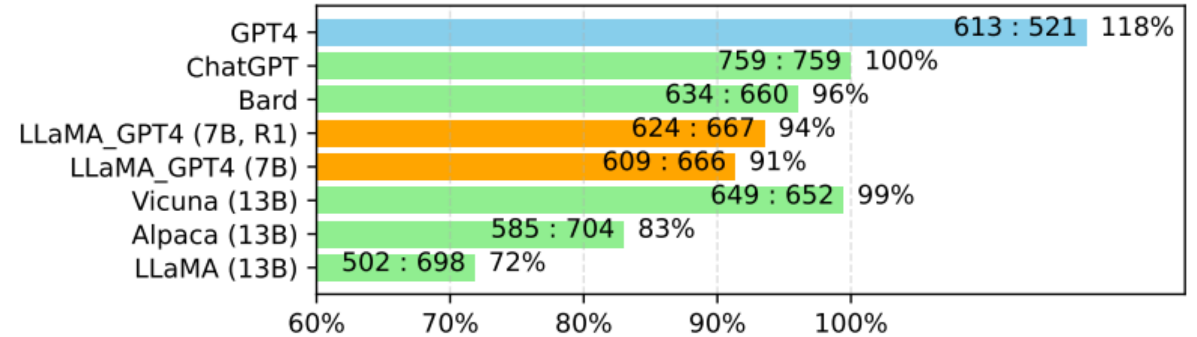
# Results on Chatbot

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

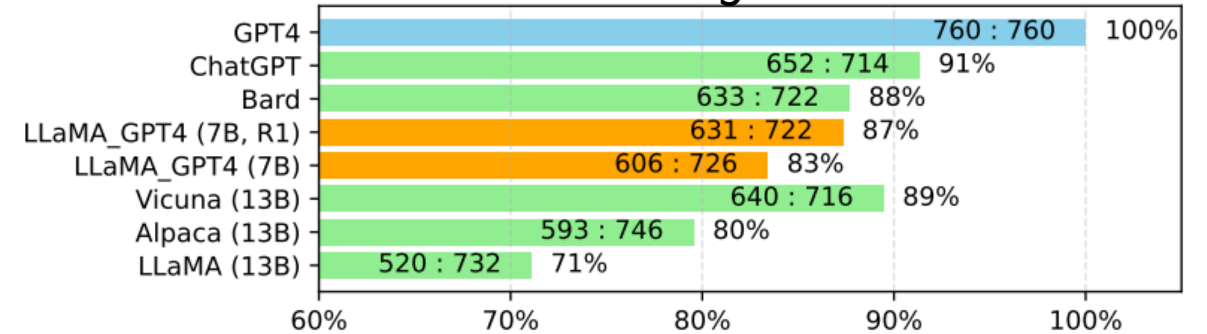
## Findings:

- 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://l1ava-vl.github.io/>

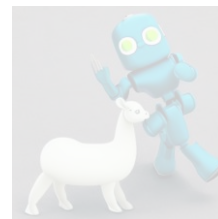
Haotian Liu\*, Chunyuan Li\*, Qingyang Wu, Yong Jae Lee (\* Equal contribution)

## Self-Instruct with Strong Teacher LLMs

## But No Teacher is available on multiGPT4?

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

## GPT-4-LLM



GPT-4  
(text-only)

## LLaVA



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

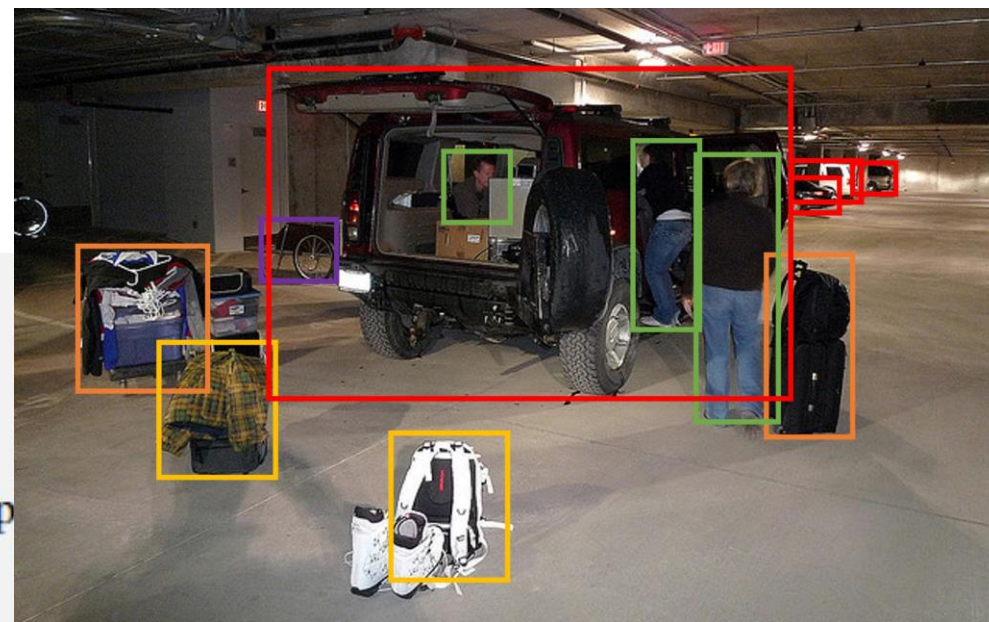
People try to fit all of their luggage in an SUV.

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.

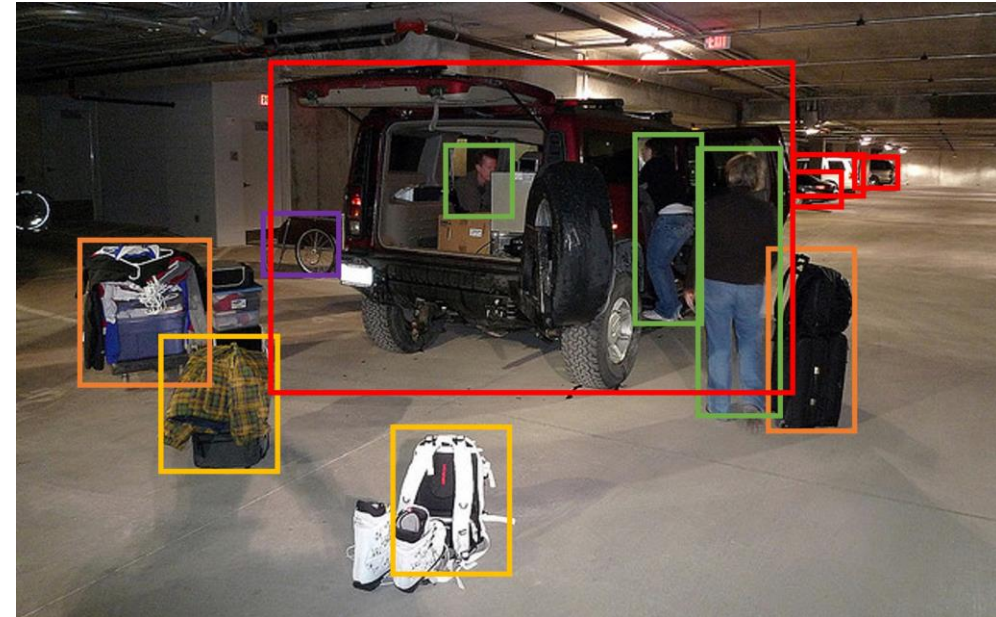
In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center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

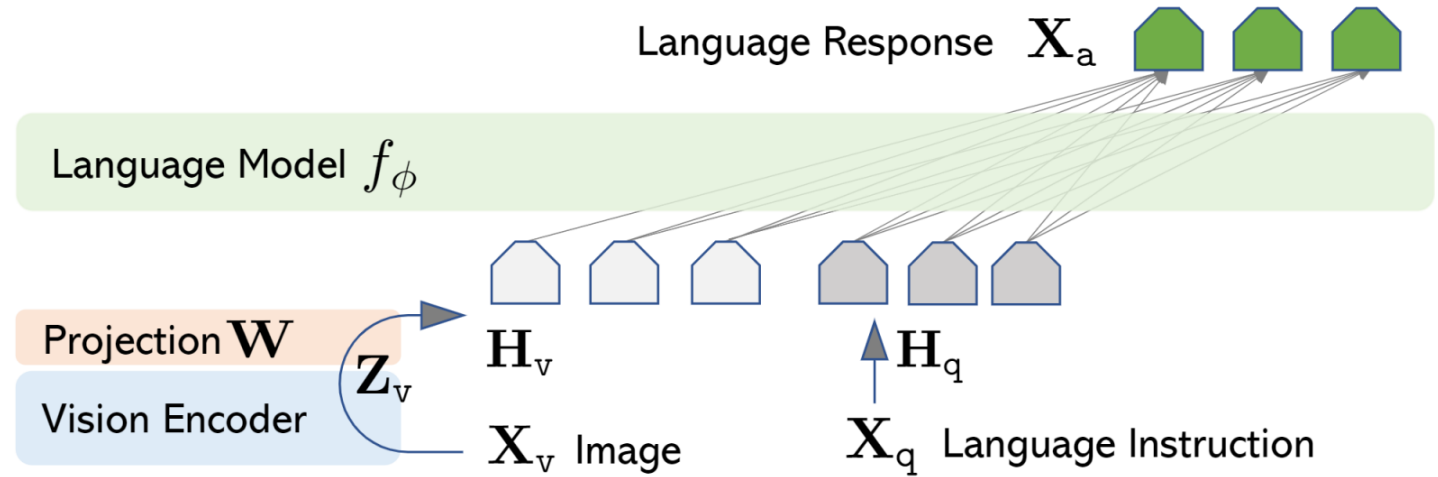
Question: What challenges do these people face?

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.

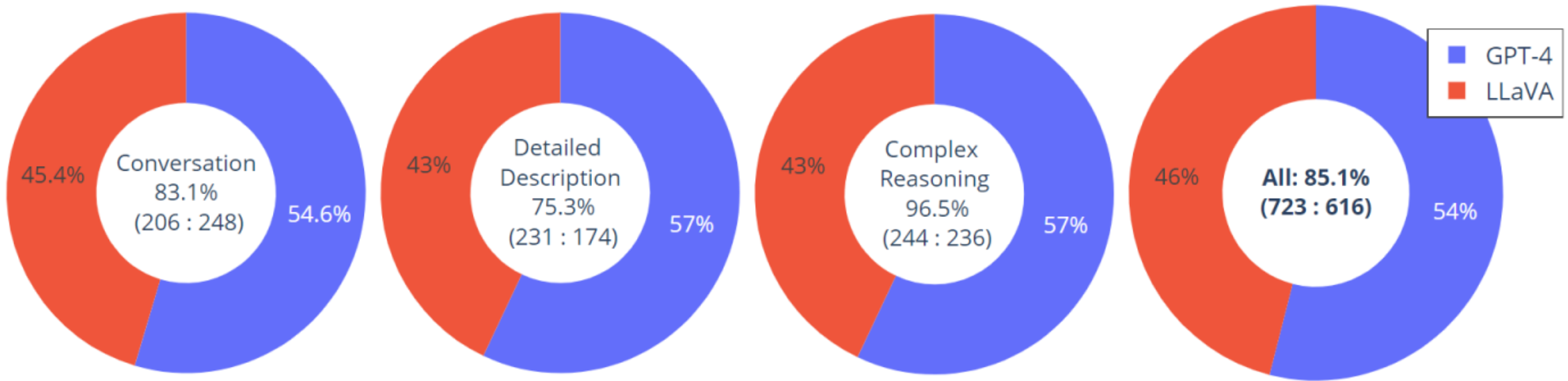
### •Stage 2: Fine-tuning End-to-End. Both the projection matrix and LLM are updated

•**Visual Chat:** Our generated multimodal instruction data for daily user-oriented applications.

•**Science QA:** Multimodal reasoning dataset for the science domain.

# Visual Chat: Towards building multimodal GPT-4 level chatbot

- LLaVA-Bench (COCO)



An evaluation dataset with 30 unseen images, 90 new language-image instructions

Overall, LLaVA achieves 85.1% relative score compared with GPT-4



# Visual Chat: Towards building multimodal GPT-4 level chatbot

- LLaVA-Bench (In-the-Wild) [LLaVA/docs/LLaVA\\_Bench.md](https://llava.lmsys.org/docs/LLaVA_Bench.md)

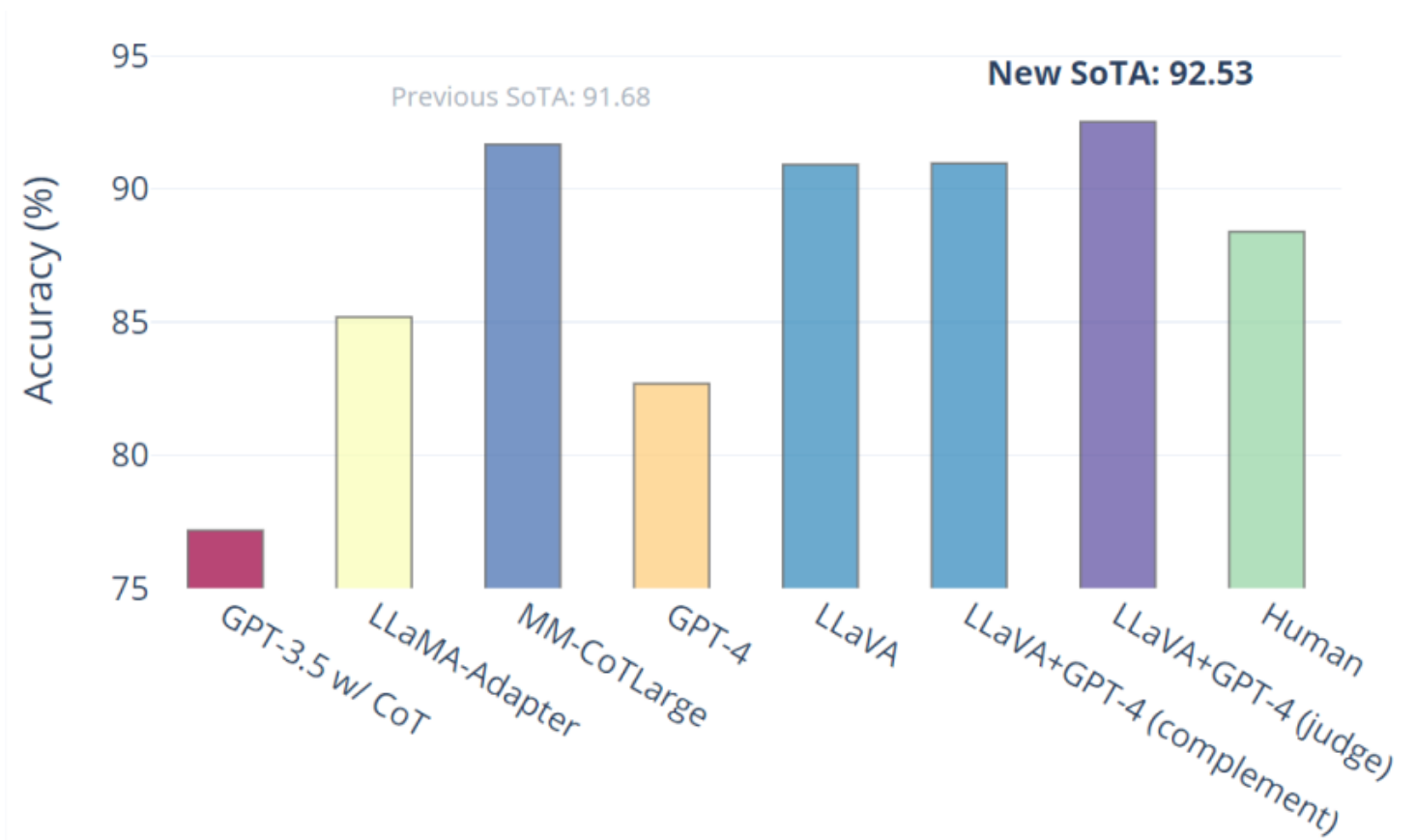
Approach	Conversation	Detail	Reasoning	Overall
Bard-0718	83.7	69.7	78.7	77.8
Bing-Chat-0629	59.6	52.2	90.1	71.5
LLaVA-13B-v1-336px-0719 (beam=1)	64.3	55.9	81.7	70.1
LLaVA-13B-v1-336px-0719 (beam=5)	68.4	59.9	84.3	73.5

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

# Science QA: New SoTA with the synergy of LLaVA with GPT-4

- LLaVA alones achieve 90.92%
- We use the text-only GPT-4 as the judge, to predict the final answer based on its own previous answers and the LLaVA answers.
- This ``GPT-4 as judge" scheme yields a new SOTA 92.53%
- GPT-4 is an effective model ensemble method





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.

User  
LLaVA

[Start a new conversation, and clear the history]  
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.

## Strong Visual Reasoning Ability

### 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.



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?

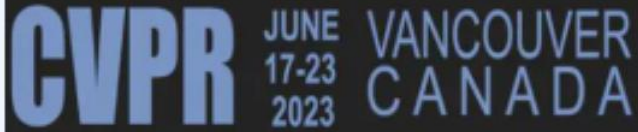
It's a picture of a chicken nugget on the International Space Station.

---

## Strong Visual Reasoning Ability

### Example 2: Chicken Nugget Map

Who will be interested in the content in this image?

The image shows a logo for the CVPR conference. It consists of a black horizontal bar with white text. On the left, 'CVPR' is written in a large, bold, sans-serif font. To its right, 'JUNE 17-23 2023' is written in a smaller font, stacked vertically. Further to the right, 'VANCOUVER CANADA' is written in a large, bold, sans-serif font, stacked vertically.

CVPR JUNE 17-23 2023 VANCOUVER CANADA

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

## Strong OCR Ability

Example 2: CVinW workshop

what is image about?

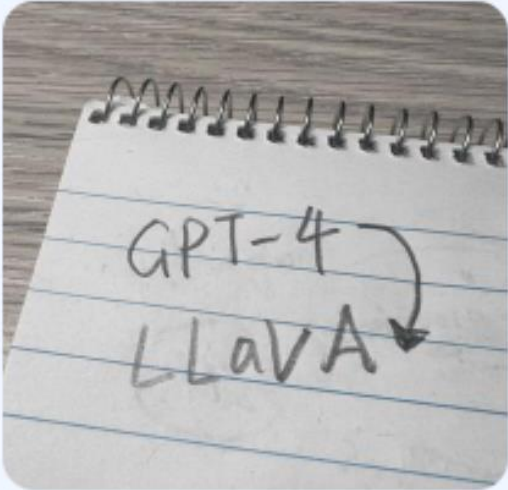


The image features a logo for "Computer Vision in the Wild," which is likely an educational or research program focused on computer vision and its applications in a wild setting. The logo consists of a large letter "C" with a smaller letter "v" on top of it, standing for "Computer Vision." The rest of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



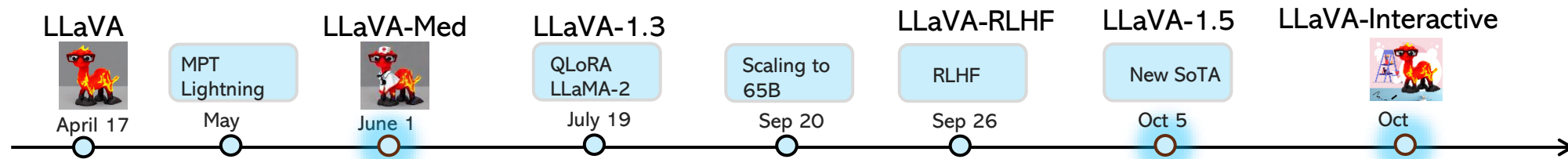
In the photo, there is written "Llava" which is presumably an abbreviation for a location or other identifier.

# Large Language and Vision Assistant

LLaVA ([llava-vl.github.io](https://llava-vl.github.io))

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

- 247 citations, and 9.4K GitHub stars, in ~6 months



**Stage 1**

Medical Concept Alignment

7 Hours


1 epoch on 600K samples

**Stage 2**


Medical Instruction Tuning

8 Hours

3 epochs on 60K samples



LLaVA-Med



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

**Performance: SoTA on 11 Benchmarks**

**LLaVA-1.5**

[https://llava-vl.github.io/](https://llava-vl.github.io)

- High Sample-Efficiency

- A Simple Architecture

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



# 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 Track (Spotlight)

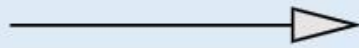
<https://aka.ms/llava-med>

LLaVA



Stage 1 (Optional)

Medical Concept Alignment



7 Hours

1 epoch on 600K samples

Stage 2

Medical Instruction Tuning



8 Hours

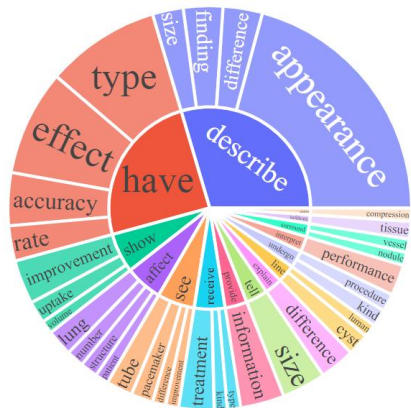
3 epochs on 60K samples

LLaVA-Med

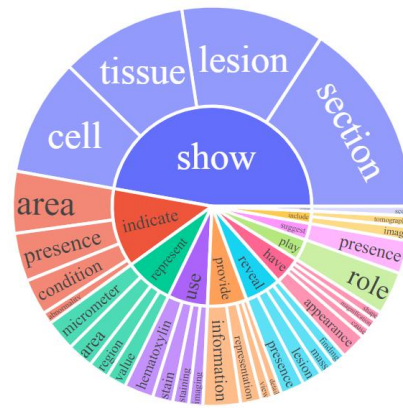


Downstream

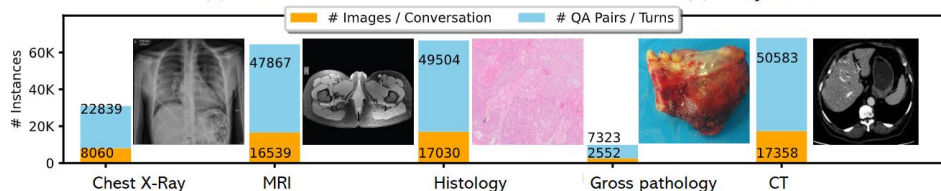
- Medical Visual Chat
- Medical VQA
  - VQA-Radiology
  - SLAKE
  - Pathology-VQA



(a) Instruction



(b) Responses



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

Visual input example, Biomedical image:



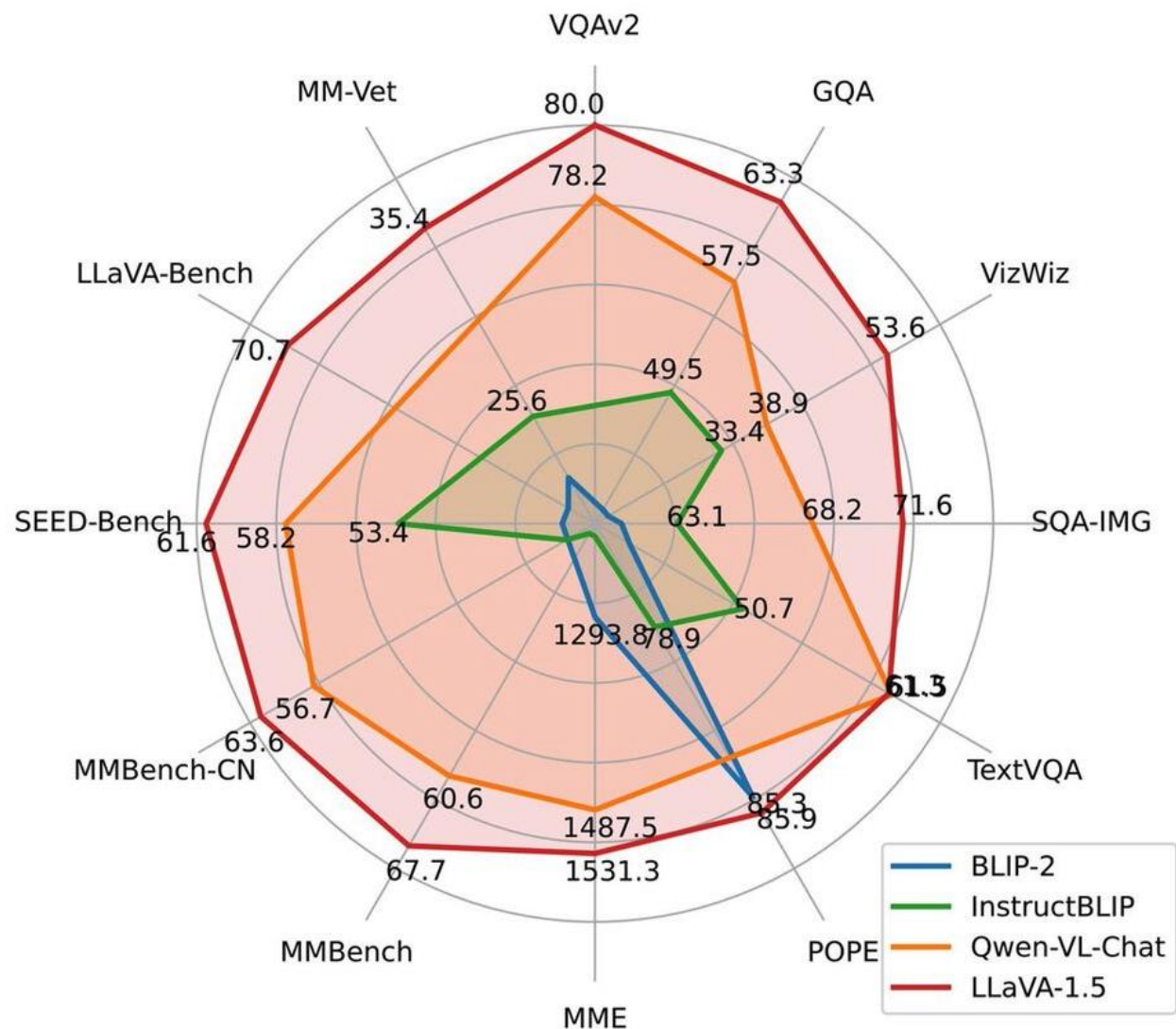
User: What is this image about?  
 LLaVA-Med: The image is a coronal view of a computed tomography (CT) scan of the abdomen and pelvis. It provides a detailed view of the internal organs and structures within the abdominal and pelvic cavities.

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...**

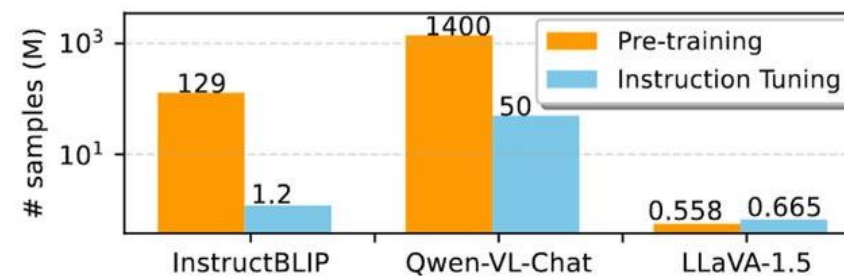
## Performance: SoTA on 11 Benchmarks



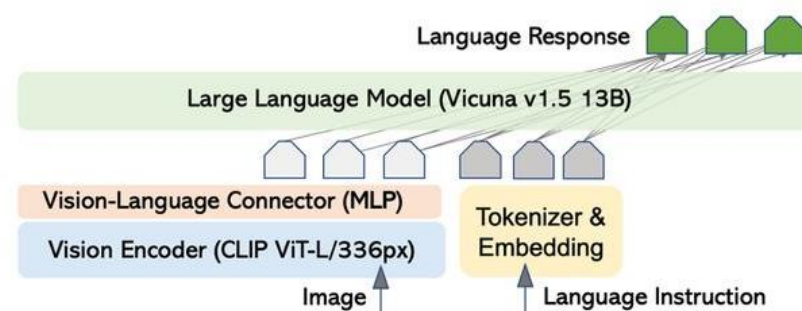
## LLaVA-1.5

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

### High Sample-Efficiency



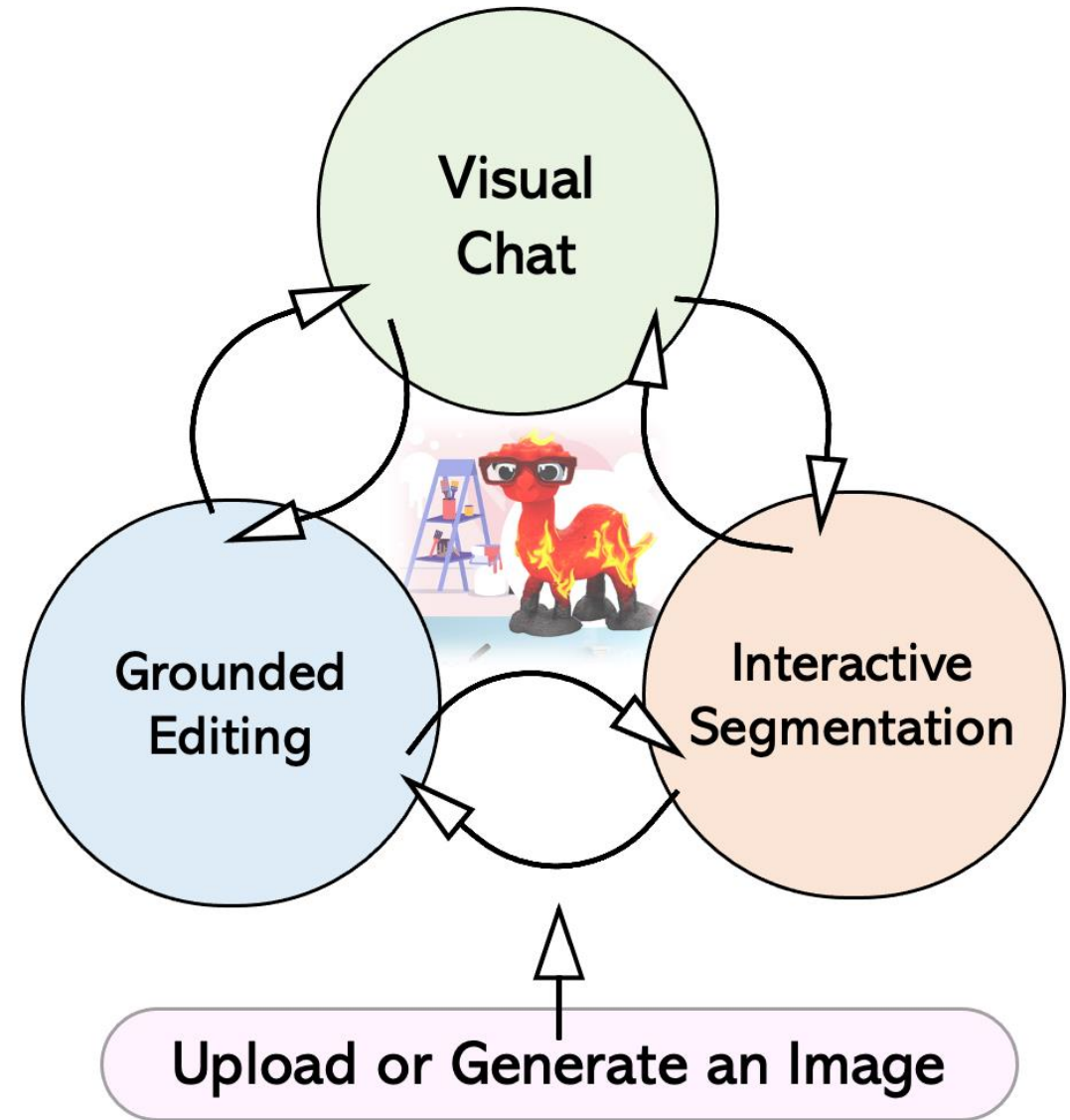
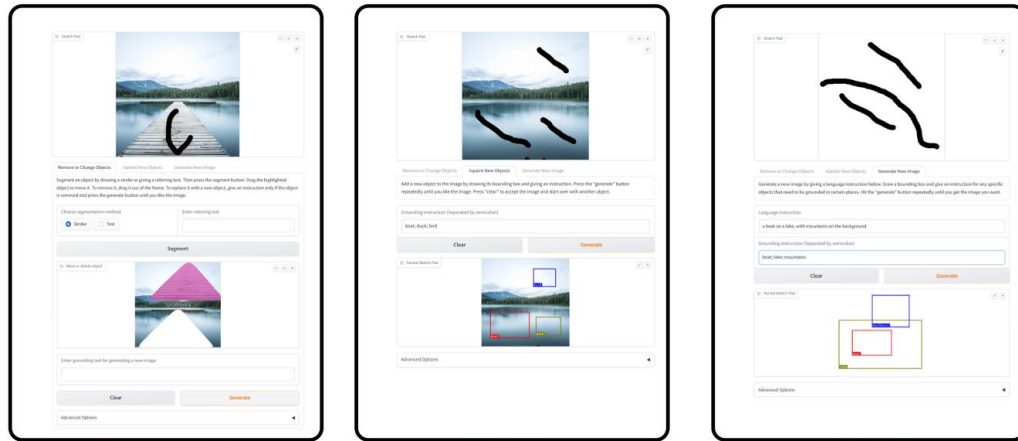
### A Simple Architecture



# LLaVA-Interactive

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

Go Beyond Language Prompt in Visual Interaction



# 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

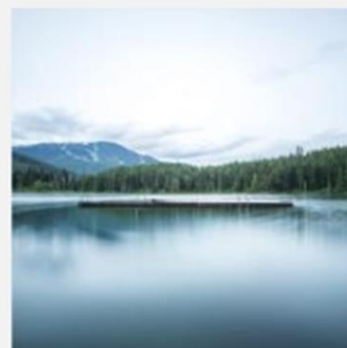
**User:** [Stroke on the pier]



**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?

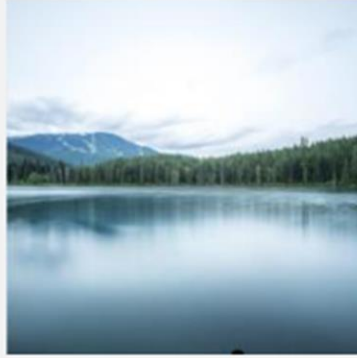
**Assistant:** The part of the picture that is not visually appealing to the viewer is the dock. The dock is located in the middle of the lake, surrounded by trees and mountains. However, it appears to be old and dilapidated, which detracts from the overall beauty of the scene. The presence of the dock 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: 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: how do you think about the current picture design for a kid cloth?  
User: 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

□ 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>

---

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

---

Chunyuan Li<sup>\*♣</sup>, Zhe Gan<sup>\*</sup>, Zhengyuan Yang<sup>\*</sup>, Jianwei Yang<sup>\*</sup>, Linjie Li<sup>\*</sup>,  
Lijuan Wang, Jianfeng Gao

Microsoft Corporation

{chunyl,zhgan,zhengyang,jianwyan,linjli,lijuanw,jfgao}@microsoft.com

<sup>\*</sup> Core Contribution <sup>♣</sup> 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.



# CVinW

Foundation Models

---

**GLIGEN**

Text-to-Image

**LLaVA**

Image-to-Text

Better Alignment with Human Intent

Q&A | Thanks

