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/@cvinw)

---

Examples of Vision Tasks: An Image Understanding Perspective

Language: A dog lying on the grass next to a frisbee

Image:

(a) Image Classification

(b) Object Detection

(c) Segmentation
Image Understanding

A Data View
- Increased Data Scale
- Image-Only

A Modeling View
- EsViT (ICLR 2022)
- UniCL & Florence (CVPR 2022)
- GLIP (CVPR 2022)
- X-Decoder (CVPR 2023)

A Benchmark View
- Unified Vision-Language Data
- LLaVA (NeurIPS 2023 & CVPR 2024)
- LMM Eval?
From Image Understanding to Image Generation
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)

OpenAI

- GPT-2 → GPT-3 → ChatGPT → InstructGPT → GPT-4

What's new?
- In-context-learning Chain-of-thoughts (CoT)
- Instruction-Following

Open Source Community

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

Our Contributions

- Alpaca
- Vicuna
- GPT-4-LLM
- Data-Centric, NOT Model-Centric

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


Self-Instruct with Strong Teacher LLMs

Teacher

LLaMA

/full/ Alpaca

GPT-3.5

Vicuna

ShareGPT (Human & GPT)

GPT-4-LLM

GPT-4 (text-only)

Instruction-following Data

None

52K

700K

(70 conversions)

• 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.
Visual Instruction Tuning with GPT-4  

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

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

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Instruction-following Data</th>
<th>LLaMA</th>
<th>Alpaca</th>
<th>Vicuna</th>
<th>GPT-4-LLM</th>
<th>LLaVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>None</td>
<td>GPT-3.5</td>
<td>ShareGPT (Human &amp; GPT)</td>
<td>GPT-4 (text-only)</td>
<td>GPT-4 (text-only)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>52K</td>
<td></td>
<td>700K (70 conversions)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 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 (In-the-Wild)** [LLaVA/docs/LLaVA_Bench.md](http://LLaVA/docs/LLaVA_Bench.md)

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

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

- **More Update on July 19 (LLaMA 2 released):**
  - LLaVA from LLaMA 2 (in 7 hours)
  - LoRA and QLoRA training
### Example 1: Extreme Ironing

**Visual input example, Extreme Ironing:**

![Image](https://www.bernorena.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 Visual Reasoning Ability

### Example 2: Chicken Nugget Map

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

---

**User** Can you explain this meme in detail?

**LLaVA** 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** Can you explain this meme?

**GPT-4 [32]** 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** Can you explain this meme in detail?

**BLIP-2**

**OpenFlamingo** It’s a picture of a chicken nugget on the International Space Station.
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.
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

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.
LLaVA Family 2023 & Beyond

Project: https://llava-vl.github.io/
An Emerging Topic in 2023 - current
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 Series:

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


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

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 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
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: [Segment the pier]

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.
<table>
<thead>
<tr>
<th>User: [Referral text = Dock] [Segment]</th>
<th>User: [Drag the mask to remove] [Generate]</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Image of a dock]</td>
<td>![Image of a lake without a dock]</td>
</tr>
</tbody>
</table>

**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.

<table>
<thead>
<tr>
<th>User: [Referral text = Sky] [Segment]</th>
<th>User: [Grounding text = sunset scene]</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Image of a sky]</td>
<td>![Image of a sunset]</td>
</tr>
</tbody>
</table>

![Image of a lake with a sunset]
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

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

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

Kid Cloth Picture Design
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.

Who says cats can't be outdoorsy?

This image is paw-some!

#naturelover
**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.

---

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

I will use grounding_dino to help 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?

I will use grounding_dino to help 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?

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.
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': '中秋月饼有哪些禁忌?中秋节吃月饼的注意事项', 'similarity': 0.98}, {'caption': '中秋节为什么要吃月饼 需要准备的事情有哪些', 'similarity': 0.94}, {'caption': '中秋月饼图片大全 色泽金黄', 'similarity': 0.94}, {'caption': '中秋月饼黑芝麻 turnover', 'similarity': 0.94}, {'caption': '中秋月饼各种口味', 'similarity': 0.94}, {'caption': '中秋月饼制作', 'similarity': 0.94}, {'caption': '中秋月饼的制作方法', 'similarity': 0.94}, {'caption': '中秋月饼的制作方法', 'similarity': 0.94}, {'caption': '中秋月饼的制作方法', 'similarity': 0.94}, {'caption': '中秋月饼的制作方法', '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.

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.
generate a view of the city skyline of downtown Seattle in a sketch style

view of the city skyline of downtown Seattle, sketch

generate a view of the city skyline of Seattle in a future and technique style

seattle city skyline future technique

generate a view of the city skyline of Shenzhen in a future and technique style

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

generate a view of the city skyline of Shanghai in a cyberpunk style

view of the city skyline of shanghai, china, cyberpunk style, photorealistic, octane render, by wlop, pixiv

Instruction to Improved Caption Prompt (LLaVA) + Image Generation (SD)
Is it possible to perform semantic segmentation for the image?

I will use openseed to assist in conducting semantic segmentation for the image.

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?

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?

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?

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)
5. Sparkling in the night sky will transport you to the City of Love!

4. It’s even more magical with the dazzling fireworks illuminating the night sky.

3. "Experience the enchanting beauty of Paris like never before!"

2. "Bonjour, Paris! It’s even more magical with the dazzling fireworks illuminating the night sky."

1. "Paris shines extra bright with the added sparkle of fireworks."

Please summarize the model outputs and answer my first question:

I want to post a message on Instagram. Add some fireworks 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";
- `valued`. 

I will use instruct-pix2pix to help answer the question. Please wait for a moment.

instruct-pix2pix model outputs: 

- `message`: "The instruct-pix2pix has processed the image.
- `valued`. 

Please summarize the model outputs and answer my first question: I want to post a message on Instagram. Add some fireworks 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: 

- `valued`.

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."
2. "Bonjour, Paris! It’s even more magical with the dazzling fireworks illuminating the night sky."
3. "Dreaming of a Parisian adventure? Our stunning new image with fireworks will inspire your 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."
5. "Experience the enchanting beauty of Paris like never before! Our new image with added fireworks will transport you to a fairytale setting."

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