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

May 2023

Chunyuan Li
Deep Learning Team
Microsoft Research, Redmond
https://chunyuan.li
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. Instruction Tuning with GPT-4 (GPT-4-LLM)
   B. Visual Instruction Tuning with GPT-4 (LLaVA)

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)
CVinW vs other CV settings
2D space for the definition of adaptation cost

- **Parameter-Efficiency**
  - Fine-Tuning
  - Linear Probing
  - Prompt with Frozen Models

- **Sample-Efficiency**
  - Zero-shot
  - Few-shot
  - Full-shot

- **Adaptation Cost**
  - Low
  - High

- Most inexpensive
- Most expensive

Prompt with Linear Probing
Frozen Models
Most inexpensive
Most expensive

2️⃣
Examples of Vision Tasks: An Image Understanding Perspective

Language

A dog lying on the grass next to a frisbee

Image

(a) Image Classification

[bbox color] dog, grass, frisbee

(b) Object Detection

[bbox color] dog, grass, frisbee

(c) Segmentation

[bbox color] dog, grass, frisbee
Image Understanding

A Data View

- Increased Data Scale
- Image-Only

A Modeling View

- EsViT (ICLR 2022)
- Image-Language (Billion)
- UniCL & Florence (CVPR 2022)
- Box-Language (Million)
- GLIP (CVPR 2022)
- Mask-Language (Thousands)
- X-Decoder (CVPR 2023)

A Benchmark View

- From Image to Language
- Classification
- Detection
- Segmentation
- Increased Semantic Richness

Output

Input

Classification

Detection

Segmentation

Increased Data Scale

Increased Semantic Richness
From Image Understanding to Image Generation
Image Generation

A Data View
- Image-Only
- Classification
- Detection
- Segmentation

A Modeling View
- Image-Language
- Box-Language
- Mask-Language

A Benchmark View
- Increased Data Scale
- Increased Semantic Richness

GLIGEN (CVPR 2023)
GLIGEN: (box, concept) → image
GLIP: image → (box, concept)
Text-to-Image Generation

GLIGEN

Grounded Language-to-Image Generation

Acknowledgements to the V-Team Members

Yuheng Li\textsuperscript{1}\textsuperscript{8}, Haotian Liu\textsuperscript{1}\textsuperscript{8}, Qingyang Wu\textsuperscript{2}, Fangzhou Mu\textsuperscript{1}, Jianwei Yang\textsuperscript{3}, Jianfeng Gao\textsuperscript{3}, Chunyuan Li\textsuperscript{3}, Yong Jae Lee\textsuperscript{3}\textsuperscript{8}

\textsuperscript{1}University of Wisconsin-Madison \textsuperscript{2}Columbia University \textsuperscript{3}Microsoft
\textsuperscript{8}Part of the work performed at Microsoft; \textsuperscript{7}Co-senior authors

CVPR 2023

Project: https://gligen.github.io/
Demo: https://aka.ms/gligen
The Space of Text-to-Image Generative AI

- Major Tech Companies
  - OpenAI: DALLE | DALLE2
  - Microsoft: NUWA
  - Google: Imagen | Parti | Muse
  - Meta: CM3 | Make-A-Scene

- Open Source
  - Latent Diffusion Models (LDM)
  - Stable Diffusion (SD) 1 & 2

- Microsoft

- Startup
  - huggingface / diffusers
  - 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 skateboarder 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

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

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

Adding "the cat as giant as the girl"

“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
A. Better Alignment with Human Intent

https://www.bilibili.com/video/BV12X4y1D7M2/?spm_id_from=333.788.recommend_more_video.6
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

Gated Self-Attention

Visual  Caption

Grounding
## Training Cost

<table>
<thead>
<tr>
<th></th>
<th>GLIGEN</th>
<th>Stable Diffusion-v1 (from scratch)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GPU Hours</strong></td>
<td>16 V100 GPUs for 10 days</td>
<td>256 A100 GPUs for 24 days</td>
</tr>
<tr>
<td>(Total: 3,840 GPU hours)</td>
<td></td>
<td>(150,000 GPU hours)</td>
</tr>
<tr>
<td><strong>Training Data</strong></td>
<td>16 million images</td>
<td>2.3 billion images</td>
</tr>
</tbody>
</table>

### Compute

- **GLIGEN**: $10^4$ GPU hours
- **Stable Diffusion**: $10^5$ GPU hours

### Data

- **GLIGEN**: $10^2$ images (Million)
- **Stable Diffusion**: $10^3$ images (Million)
The connections to the trends in NLP

--The similar spirits, but in the image domain

<table>
<thead>
<tr>
<th>Language Generation</th>
<th>Image Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GPT3.5</strong> $\rightarrow$ ChatGPT</td>
<td><strong>DALL-E2</strong> $\rightarrow$ GLIGEN</td>
</tr>
<tr>
<td>ChatGPT/InstructGPT are better aligned with users than GPT3/3.5 in following human intent and perform the language task that the user wants</td>
<td>GLIGEN is better aligned with users than DALL-E2 in following human intent and perform the image generation/editing task that the user wants</td>
</tr>
<tr>
<td>Less than 2% of the compute and data relative to model pretraining</td>
<td>Less than 3% of the compute and 1% of data relative to model pretraining</td>
</tr>
</tbody>
</table>

A. Better Alignment with Human Intent

B. Much Lower Development Cost

https://openai.com/blog/chatgpt
https://openai.com/blog/instruction-following/
https://openai.com/alignment/
A General-Purpose Visual Assistant

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

- Instruction Tuning with GPT-4 (GPT-4-LLM)
- Visual Instruction Tuning with GPT-4 (LLaVA)
Language Generation: Large Language Models (LLM)

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

Open Source Community
- LLaMA

Our Contributions
- Data-Centric, NOT Model-Centric

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

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

Alpaca
Vicuna

GPT-4-LLM

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

Self-Instruct with Strong Teacher LLMs

<table>
<thead>
<tr>
<th>Teacher</th>
<th>GPT-3.5</th>
<th>ShareGPT (Human &amp; GPT)</th>
<th>GPT-4 (text-only)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLaMA</td>
<td>Alpaca</td>
<td>Vicuna</td>
<td>GPT-4-LLM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instruction-following Data</td>
<td>None</td>
<td>52K (70K conversions)</td>
<td>52K English &amp; Chinese</td>
</tr>
<tr>
<td>Data</td>
<td></td>
<td></td>
<td>700K Feedback Data</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LLM Chatbot</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Reward Model</td>
</tr>
</tbody>
</table>
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.
Results on Reward Model

1. **Feedback Data**: Ask GPT-4 to rate the responses from three models from 1~20

2. **Reward Model**: Training OPT on the ranking data

3. **Evaluation**: Decode 5 responses for each unseen instruction, then rank them.

![Feedback Data Distribution](image1)

![Relative score again GPT-4](image2)

(the scores are also rated by GPT-4)
Visual Instruction Tuning with GPT-4

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

Self-Instruct with Strong Teacher LLMs

<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>52K</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>
</tbody>
</table>

But No Teacher is available on multiGPT4?

- 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
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**

- **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

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

Overall, LLaVA achieves 85.1% relative score compared with GPT-4
Science QA: New SoTA with the synergy of LLaVA with GPT-4

- LLaVA alone achieves 90.92% accuracy.
- 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% accuracy.
- GPT-4 is an effective model ensemble method.
Strong Visual Reasoning Ability

Example 1: Extreme Ironing

mini-GPT4

What is unusual about this image?

The image shows a man in a yellow shirt standing on the back of a yellow car with a blue shirt hanging on the back of the car. The car is parked on the side of the road next to a taxi. There are other cars and taxis driving on the road in the background.

What is unusual about this image?

There is nothing unusual about this image. It is just a man on a yellow taxi with a yellow shirt on.
### Visual input example, Chicken Nugget Map:

![Chicken Nugget Map](image)

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

<table>
<thead>
<tr>
<th>User</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLaVA</td>
<td>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 &quot;Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all is...&quot; 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 &quot;I mean, it’s not the real Earth, but how beautiful it is all is.&quot; 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.</td>
</tr>
<tr>
<td>GPT-4 [32]</td>
<td>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.</td>
</tr>
<tr>
<td>BLIP-2</td>
<td>Can you explain this meme in detail? Sometimes I just look at pictures of the Earth from space and marvel how beautiful it is</td>
</tr>
<tr>
<td>OpenFlamingo</td>
<td>It’s a picture of a chicken nugget on the International Space Station.</td>
</tr>
</tbody>
</table>
Strong OCR Ability

Example 1: CVPR & Vancouver

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.
In the photo, there is written "Llava" which is presumably an abbreviation for a location or other identifier.
LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

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

https://aka.ms/llava-med

**LLaVA**

**Stage 1**

Medical Concept Alignment

7 Hours

1 epoch on 600K samples

**Stage 2**

Medical Instruction Tuning

8 Hours

3 epochs on 60K samples

**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.

---

**Supervised fine-tuning results with our own experiment runs**

<table>
<thead>
<tr>
<th>Method</th>
<th>VQA-RAD</th>
<th>SLAKE</th>
<th>PathVQA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ref</td>
<td>Open</td>
<td>Closed</td>
</tr>
<tr>
<td>LLaVA</td>
<td>50.00</td>
<td>65.07</td>
<td>78.18</td>
</tr>
<tr>
<td>LLaVA-Med (From LLaVA)</td>
<td>61.52</td>
<td>84.19</td>
<td>83.08</td>
</tr>
<tr>
<td>LLaVA-Med (From Vicuna)</td>
<td>64.39</td>
<td>81.98</td>
<td>84.71</td>
</tr>
<tr>
<td>LLaVA-Med (BioMed CLIP)</td>
<td>64.75</td>
<td>83.09</td>
<td>87.11</td>
</tr>
</tbody>
</table>

**Representative & SoTA methods with numbers reported in the literature**

<table>
<thead>
<tr>
<th>Method</th>
<th>VQA-RAD</th>
<th>SLAKE</th>
<th>PathVQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>VL Encoder-Decoder [5]</td>
<td>71.49</td>
<td>82.47</td>
<td>71.49</td>
</tr>
<tr>
<td>Q2Transformer [26]</td>
<td>79.19</td>
<td>81.20</td>
<td>54.85</td>
</tr>
<tr>
<td>Prefix T Medical LM [41]</td>
<td>84.30</td>
<td>82.91</td>
<td>40.00</td>
</tr>
<tr>
<td>PubMedCLIP [6]</td>
<td>60.10</td>
<td>80.00</td>
<td>78.49</td>
</tr>
<tr>
<td>BiomedCLIP [49]</td>
<td>67.60</td>
<td>79.80</td>
<td>82.05</td>
</tr>
<tr>
<td>M212 [22]</td>
<td>66.50</td>
<td>83.50</td>
<td>74.70</td>
</tr>
</tbody>
</table>

---

37
CVinW

Foundation Models

GLIGEN  LLaVA
Text-to-Image  Image-to-Text

Better Alignment with  Human Intent

Q&A  |  Thanks