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

☐ 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. 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.

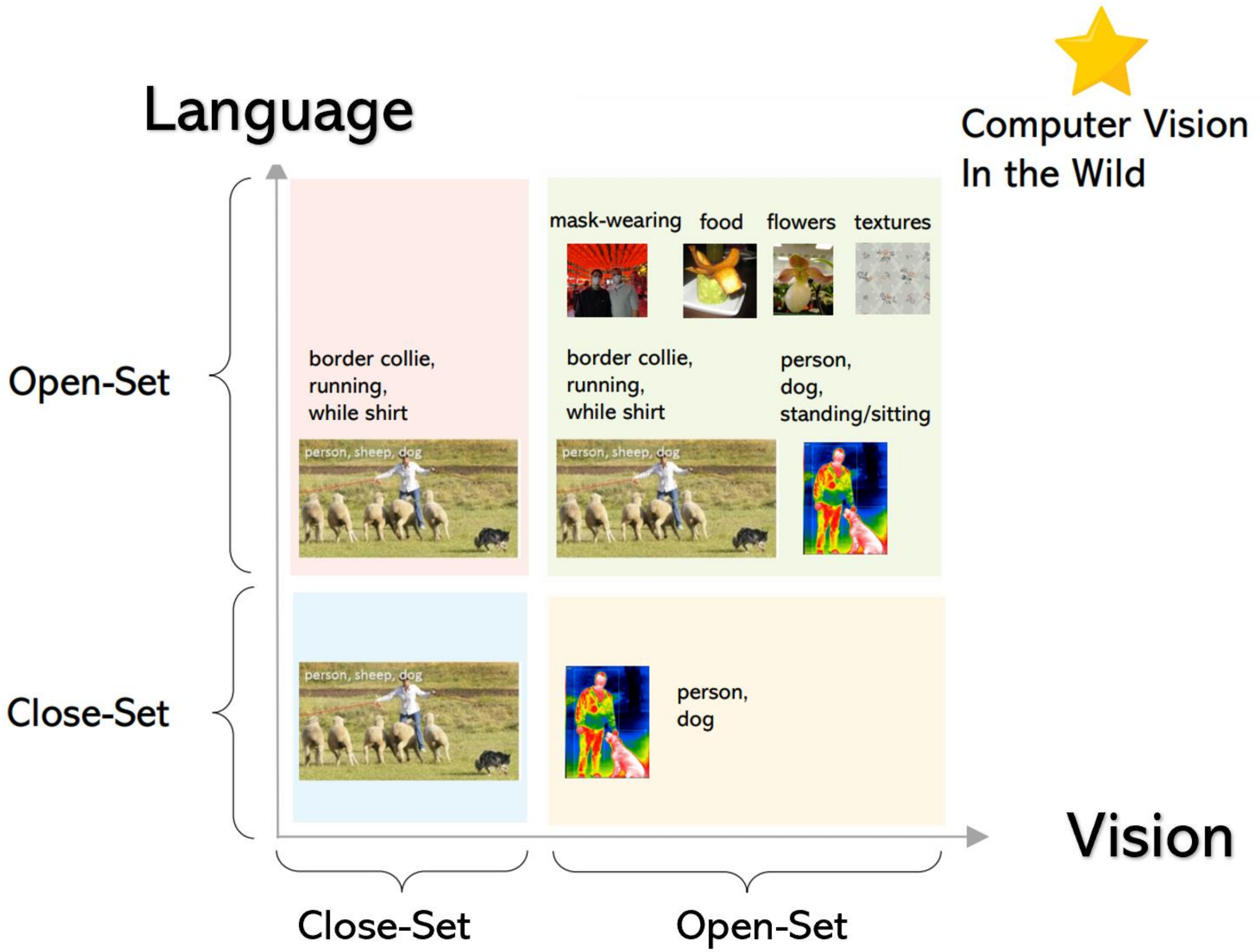
CVinW
Computer Vision in the Wild
https://computer-vision-in-the-wild.github.io/eccv-2022/

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

YouTube https://www.youtube.com/@cvinw

ELEVATER: A Benchmark and Toolkit for Evaluating Language-Augmented Visual Models, NeurIPS 2022

1 CVinW *vs* other CV settings



Computer Vision
In the Wild

Language

Open-Set

Close-Set

mask-wearing food flowers textures

border collie,
running,
while shirt

border collie,
running,
while shirt

person,
dog,
standing/sitting

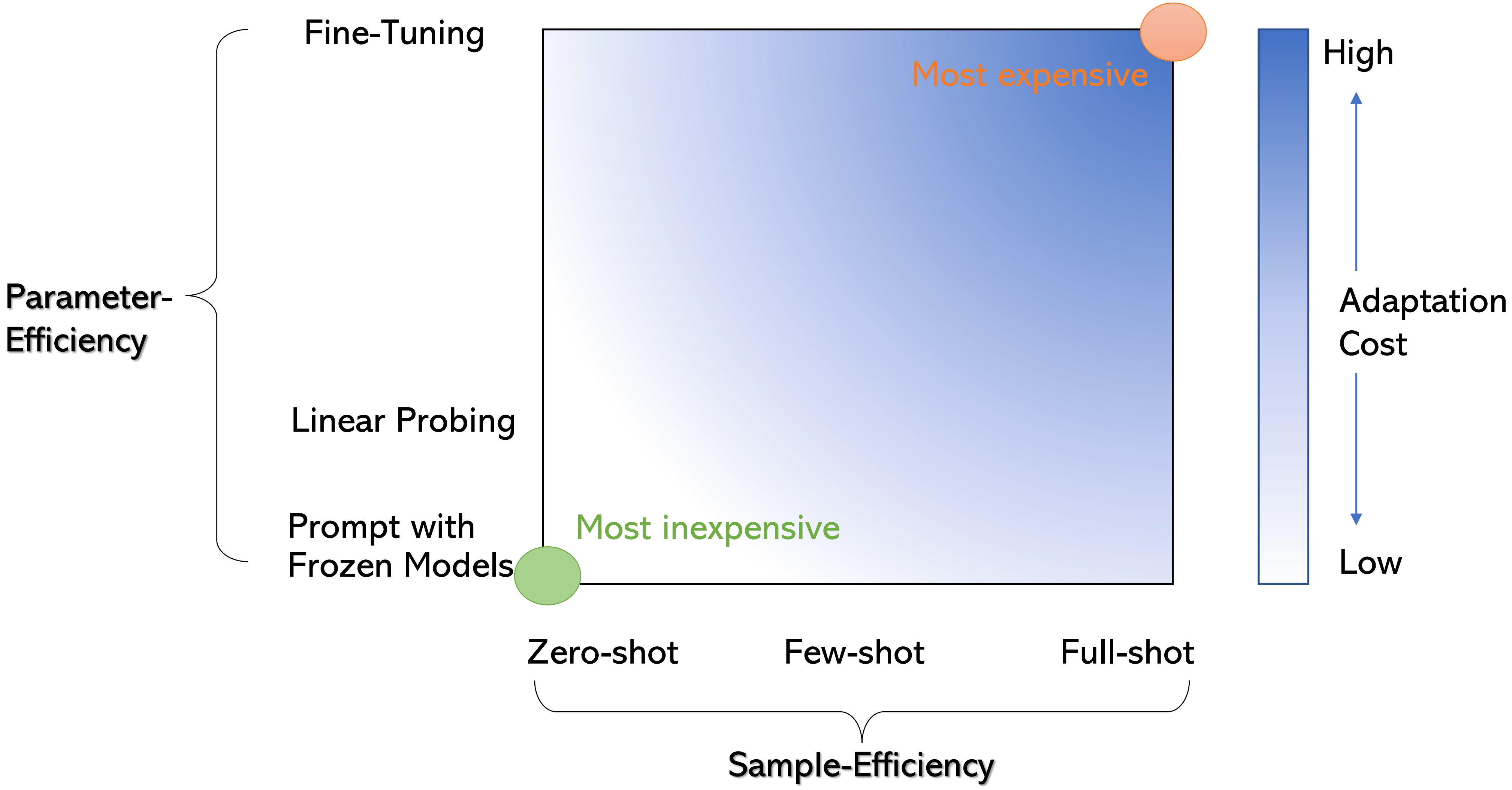
person,
dog

Vision

Close-Set

Open-Set

2 2D space for the definition of adaptation cost



Fine-Tuning

Linear Probing

Prompt with
Frozen Models

Parameter-
Efficiency

Most expensive

Most inexpensive

Zero-shot

Few-shot

Full-shot

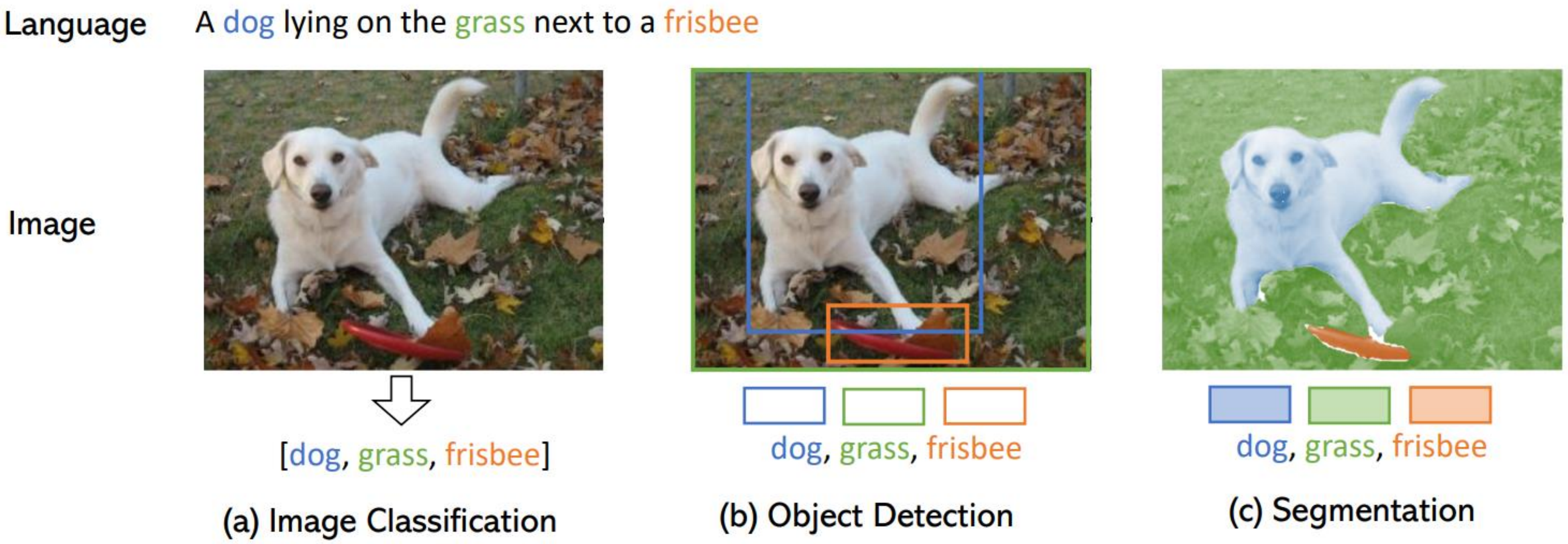
Sample-Efficiency

High

Adaptation
Cost

Low

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

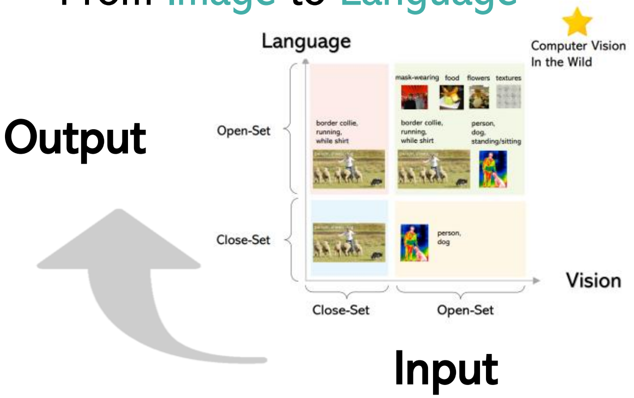
dog, grass, frisbee

(c) Segmentation

Vision-Language Pre-Training: Basics, Recent Advances, and Future Trends
Foundations and Trends® in Computer Graphics and Vision, 2022

Image Understanding

From Image to Language



Increased Data Scale

	A Data View	A Modeling View	A Benchmark View
	Image-Only	EsViT (ICLR 2022)	
Classification	Image-Language (Billion)	UniCL & Florence (CVPR 2022)	ICinW Image Classification in the Wild
Detection	Box-Language (Million)	GLIP (CVPR 2022)	ODinW Object Detection in the Wild roboflow 100
Segmentation	Mask-Language (~~Thousands~~)	X-Decoder (CVPR 2023)	SGinW Segmentation in the Wild

Increased Semantic Richness

ECCV TEL AVIV 2022

CVPR VANCOUVER, CANADA

From Image Understanding to Image Generation

Image Understanding

From Image to Language

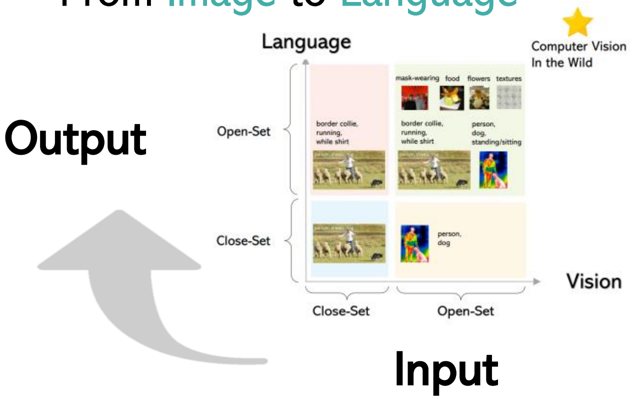


Image Generation

From Language to Image

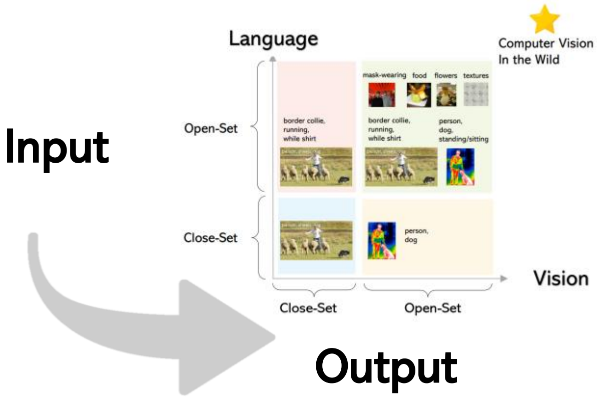
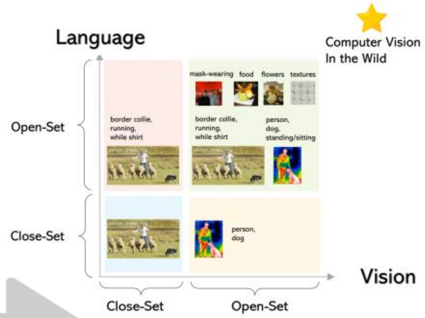


Image Generation

A Data View **A Modeling View** **A Benchmark View**

From Language to Image



Input

Output

Increased
Data
Scale

Image-Only

Classification

Image-Language

Detection

Box-Language

Segmentation

Mask-Language

Increased
Semantic
Richness

GLIGEN
(CVPR 2023)

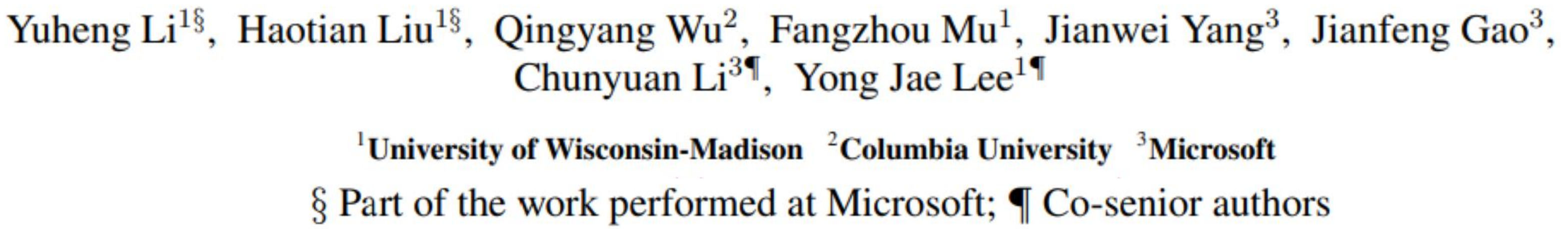
GLIGEN: (box, concept) → image
GLIP: image → (box, concept)

③ Text-to-Image Generation

GLIGEN

Grounded **L**anguage-to-**I**mage **Gen**eration

Acknowledgements to the V-Team Members



Yuheng Li[1§], Haotian Liu[1§], Qingyang Wu[2], Fangzhou Mu[1], Jianwei Yang[3], Jianfeng Gao[3],
Chunyuan Li[3¶], Yong Jae Lee[1¶]

[1]**University of Wisconsin-Madison** [2]**Columbia University** [3]**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

Microsoft Designer

☐ Open Source

stability.ai

Latent Diffusion Models (LDM)
Stable Diffusion (SD) 1 & 2

☐ Startup

huggingface / diffusers

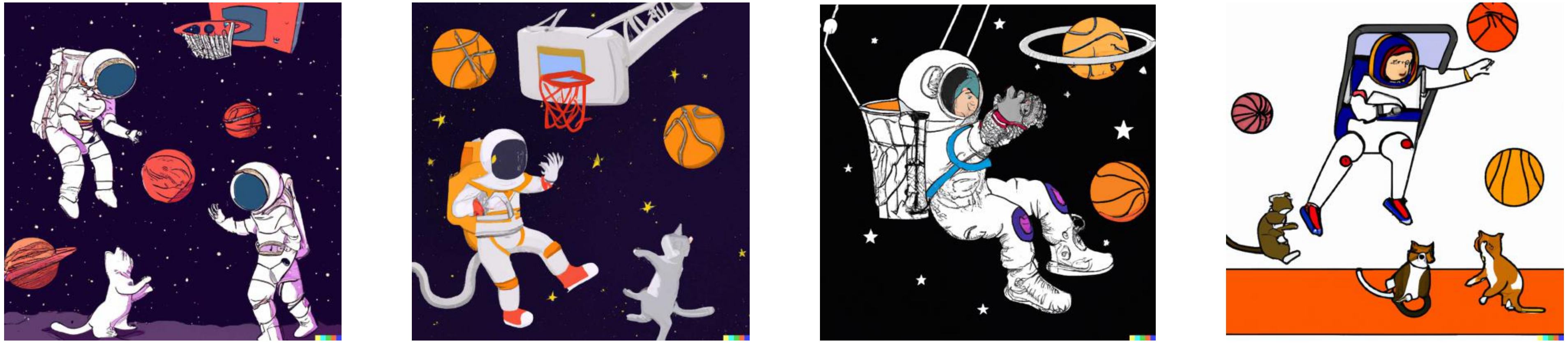
DreamStudio

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

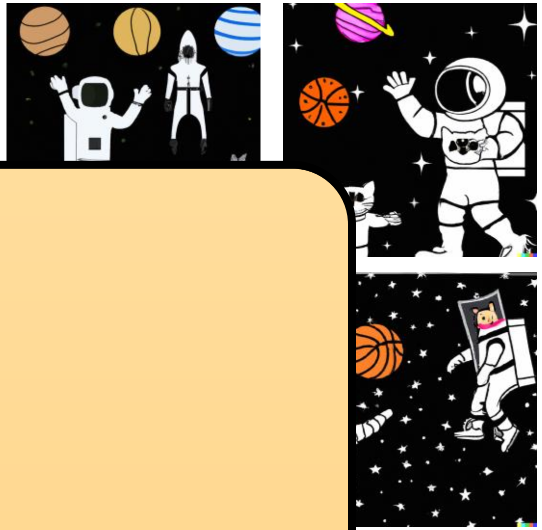
astronaut skate boarder in space, in the style of vaporwave

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

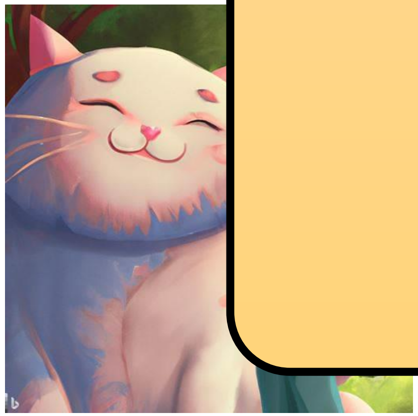
Limitations with Language Prompt Alone

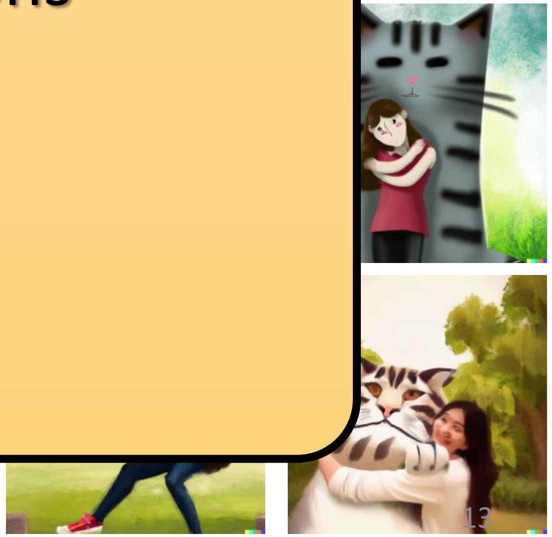
DALLE2

An astronaut playing basketball with cats in space



a women hu
giant cat with
in the park, c





⚠

“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

Video Demo (Total: 5:45 minutes)

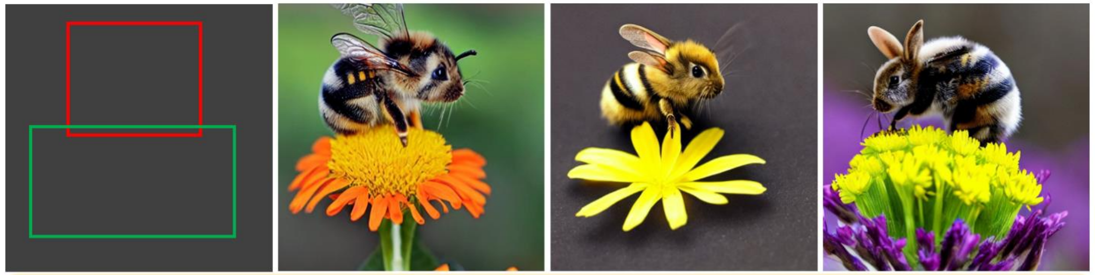
GLIGEN

Grounded Language-to-Image Generation

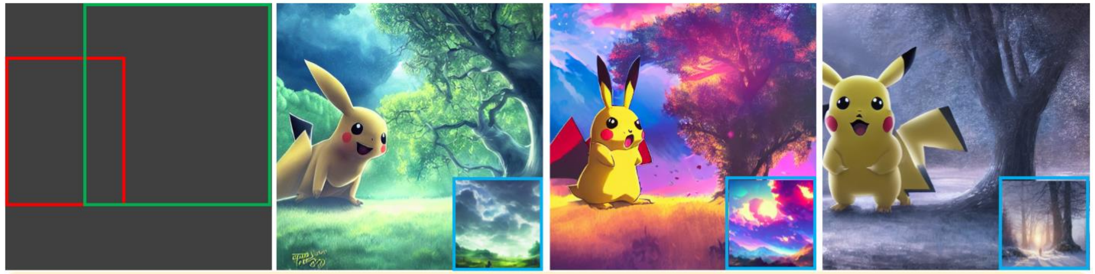
Project: https://gligen.github.io/

Demo: https://aka.ms/gligen

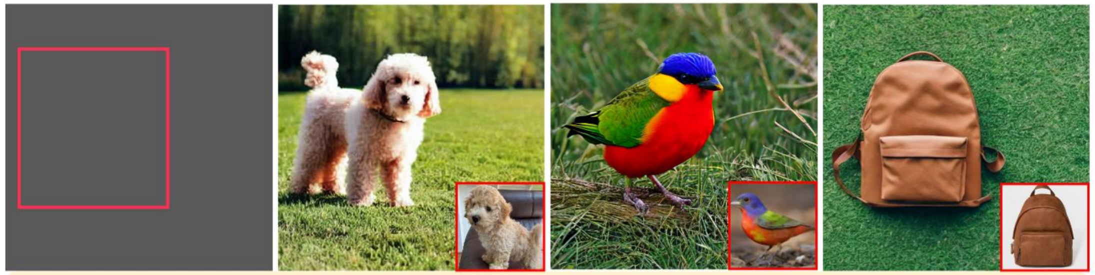
https://www.bilibili.com/video/BV12X4y1D7M2/?spm_id_from=333.788.recommend_more_video.6



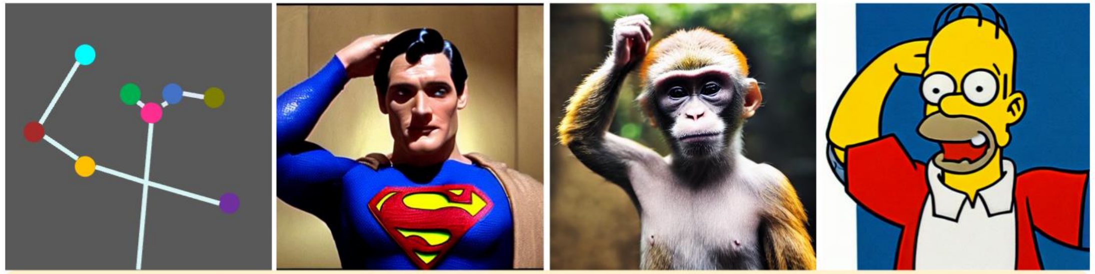
Caption: “a photo of a hybrid between a bee and a rabbit”
Grounded text: hybrid between a bee and a rabbit, flower



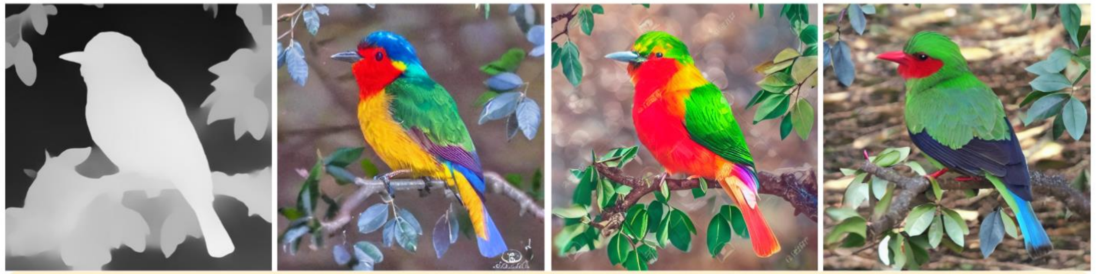
Caption: “Pikachu is under a tree, digital art”
Grounded text: Pikachu, tree; Grounded style image: blue inset



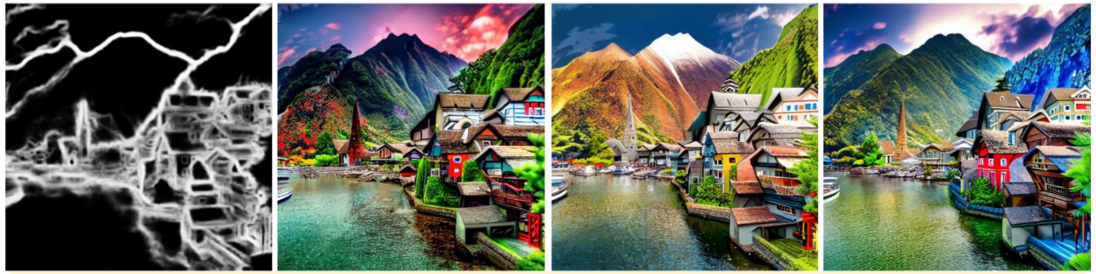
Caption: “A dog / bird / helmet / backpack is on the grass”
Grounded image: red inset



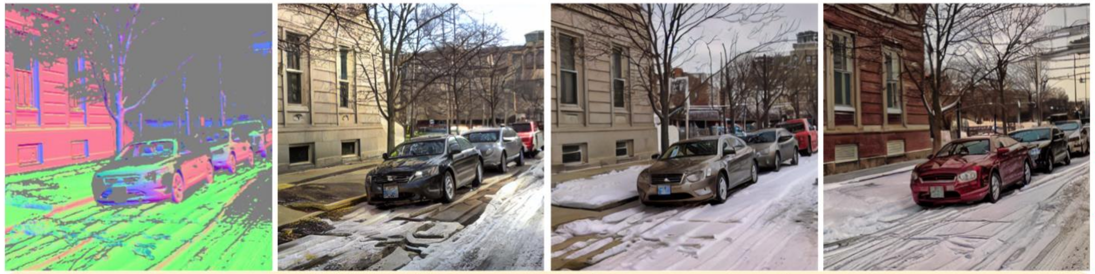
Caption: “superman / monkey / Hormer Simpson / is scratching its head”
Grounded keypoints: plotted dots on the left image



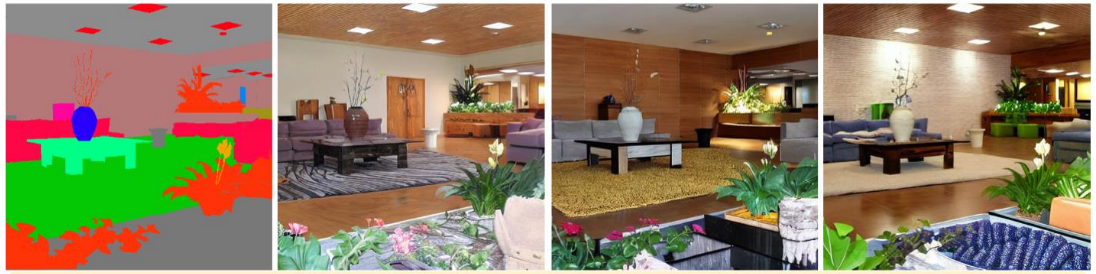
Caption: “A vibrant colorful bird sitting on tree branch”
Grounded depth map: the left image



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

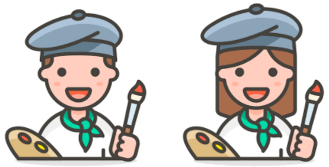


Caption: “Cars park on the snowy street”
Grounded normal map: the left image



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…

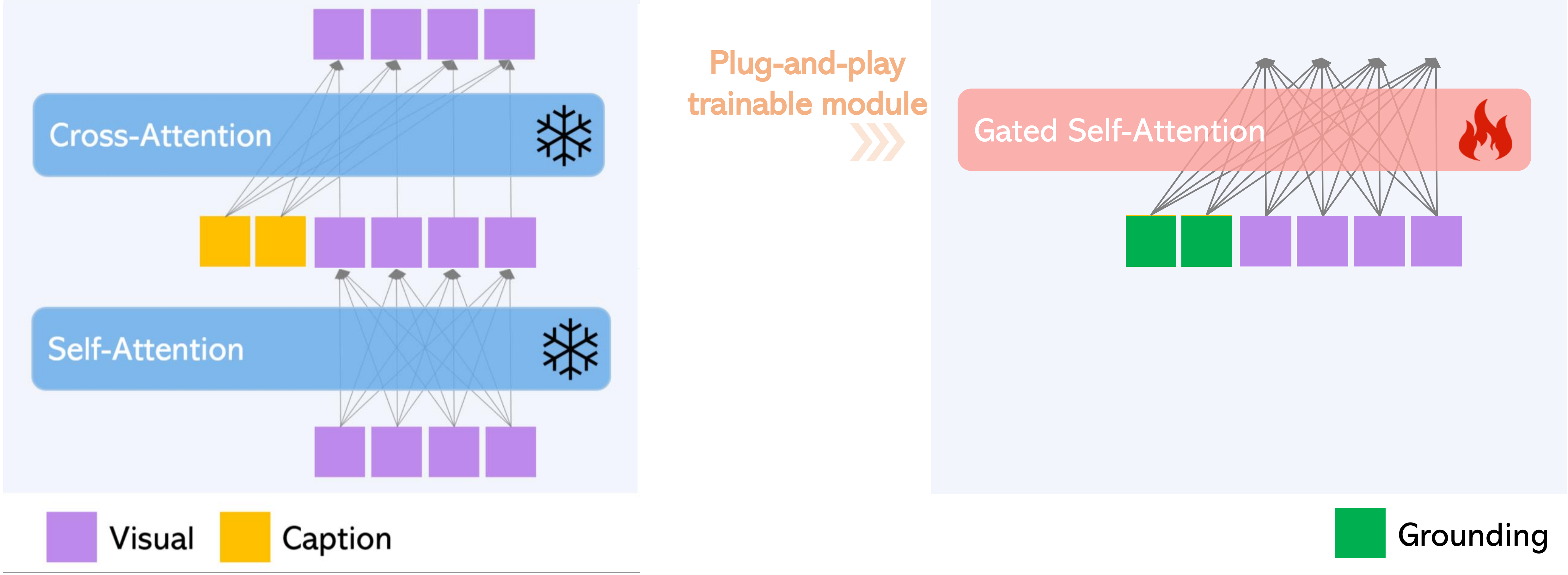
GLIGEN

×

stability.ai

Modulated Training

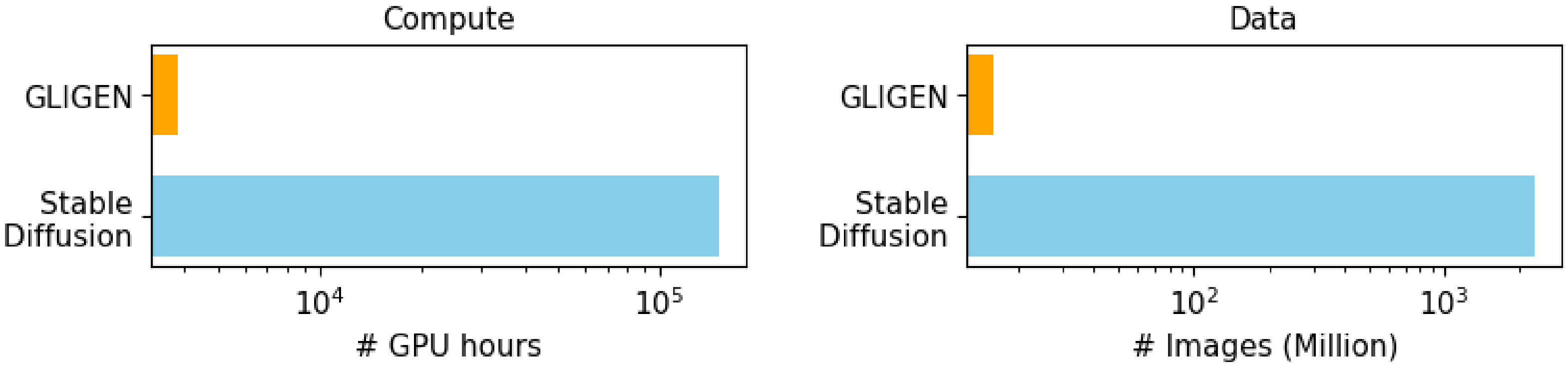
Transformer in Diffusion Models



Plug-and-play trainable module

Training Cost

	2.5% **GPU Hours**	0.7% **Training Data**
GLIGEN	16 V100 GPUs for 10 days (Total: 3,840 GPU hours)	16 million images
Stable Diffusion-v1 (from scratch)	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

	Language Generation **GPT3.5 → ChatGPT**	Image Generation **DALLE2 → GLIGEN**
A. Better Alignment with Human Intent	ChatGPT/InstructGPT are better aligned with users than GPT3/3.5 in following human intent and perform the language task that the user wants	GLIGEN is better aligned with users than DALLE2 in following human intent and perform the image generation/editing task that the user wants
B. Much Lower Development Cost	Less than **2%** of the compute and data relative to model pretraining	Less than **3%** of the compute and **1%** of data relative to model pretraining

https://openai.com/blog/chatgpt
https://openai.com/blog/instruction-following/
https://openai.com/alignment/

③ 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)

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

What's new?

GPT-3:
In-context-learning
Chain-of-thoughts (CoT)

ChatGPT / InstructGPT:
In-context-learning
Chain-of-thoughts (CoT)
Instruction-Following

GPT-4:
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/

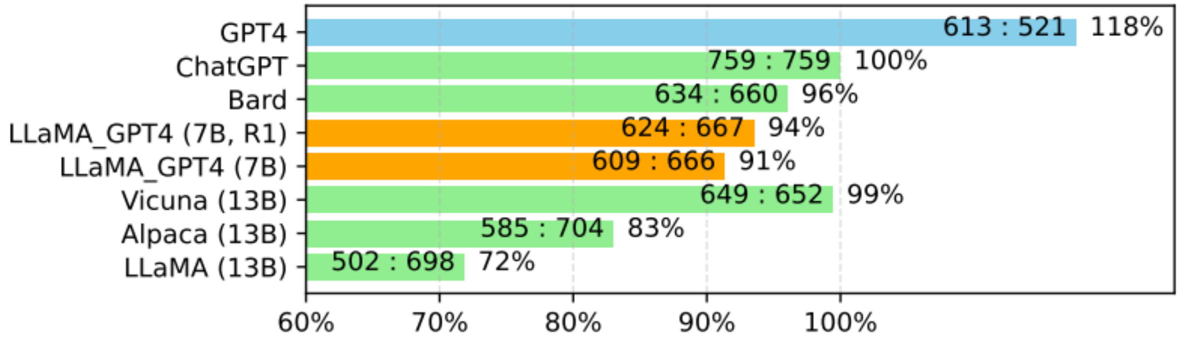
Baolin Peng*, Chunyuan Li*, Pengcheng He*, Michel Galley, Jianfeng Gao (* Equal contribution)

Self-Instruct with Strong Teacher LLMs

	LLaMA	Alpaca	Vicuna	GPT-4-LLM
Teacher		GPT-3.5	ShareGPT (Human & GPT)	GPT-4 (text-only)
Instruction-following Data	None	52K	700K (70 conversions)	• 52K English & Chinese → LLM Chatbot • Feedback Data → Reward Model

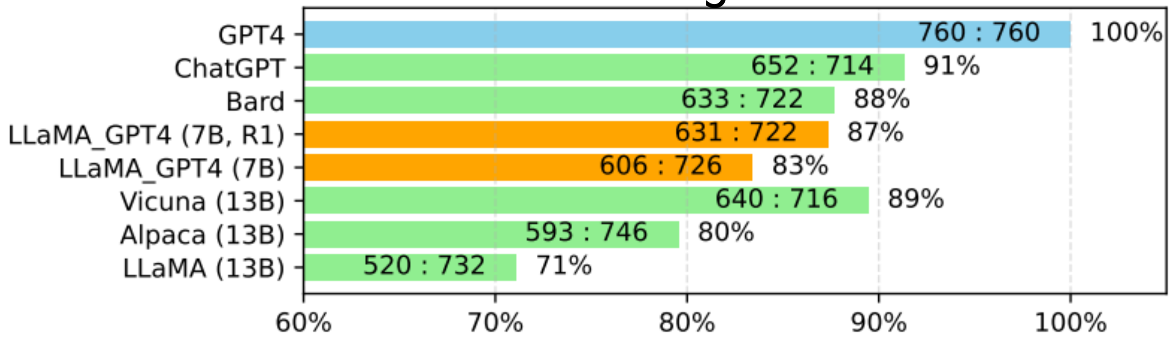
Results on Chatbot

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


All chatbots against ChatGPT

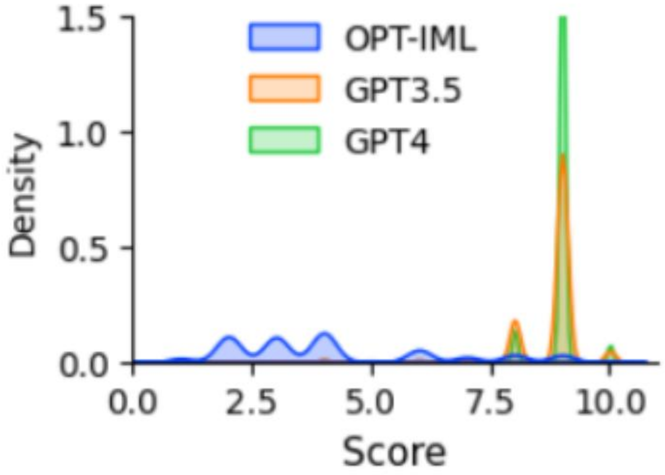
Findings:

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

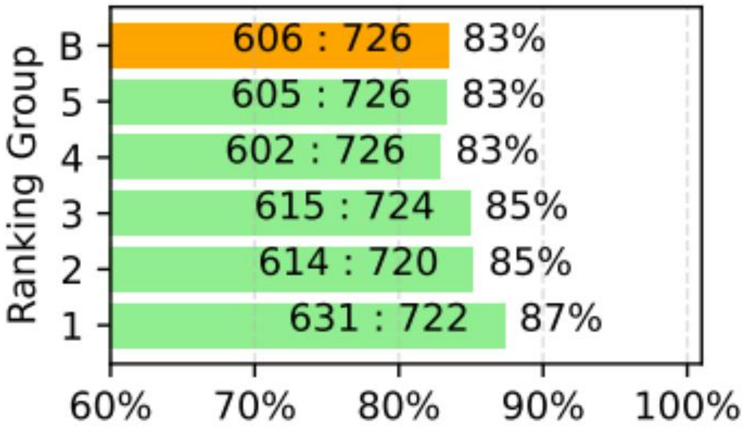

All chatbots against GPT-4

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



Relative score again GPT-4
(the scores are also rated by 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

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

But No Teacher is available on multiGPT4?

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

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

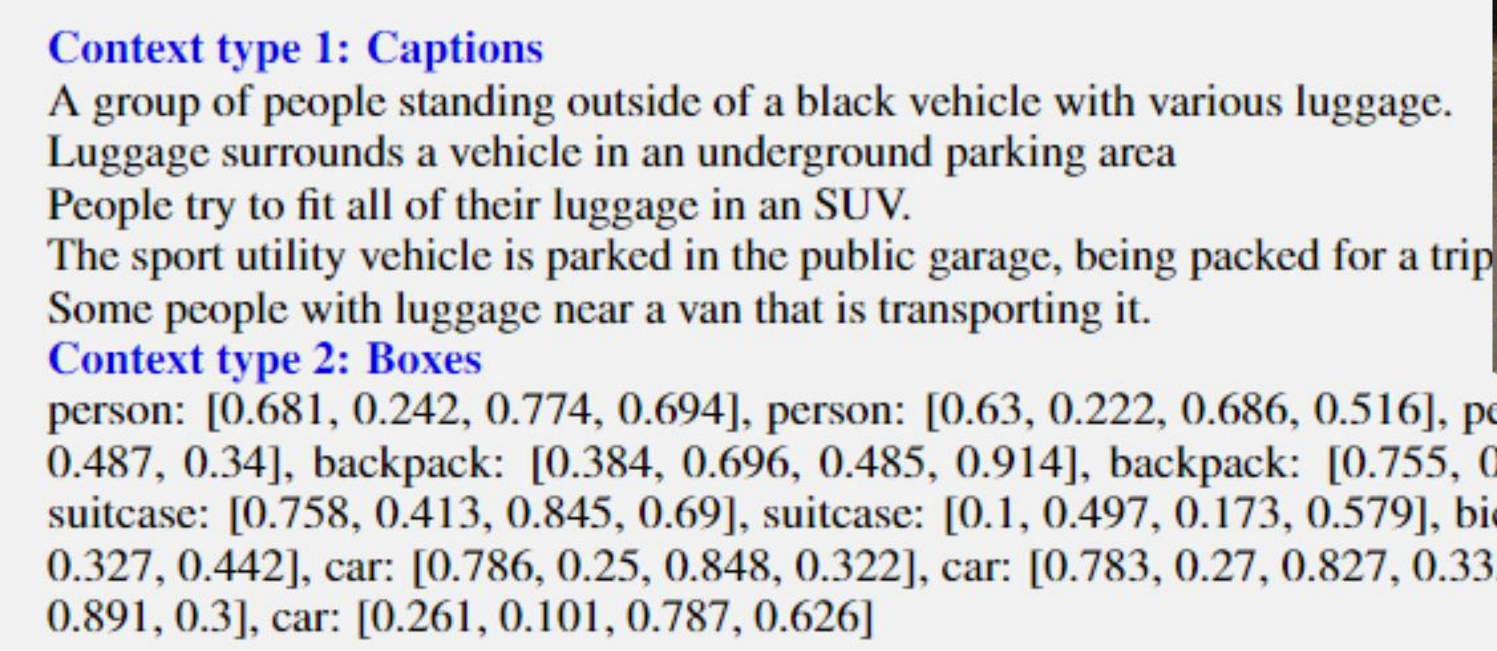
→ Multimodal Chatbot

Large Language and Vision Assistant

GPT-assisted Visual Instruction Data Generation

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

→ Text-only GPT-4

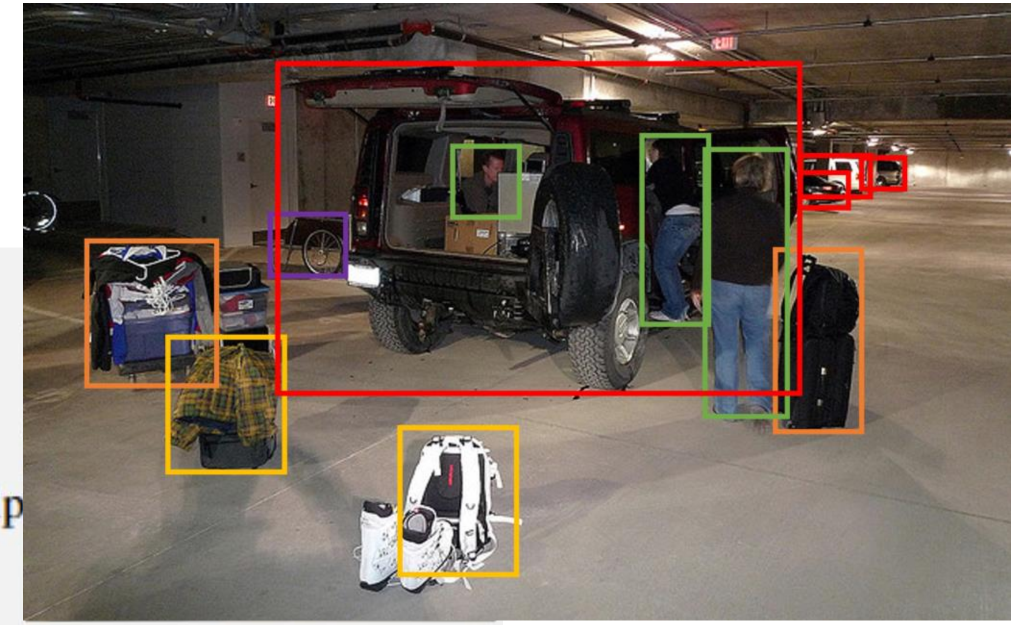


Context type 1: Captions

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

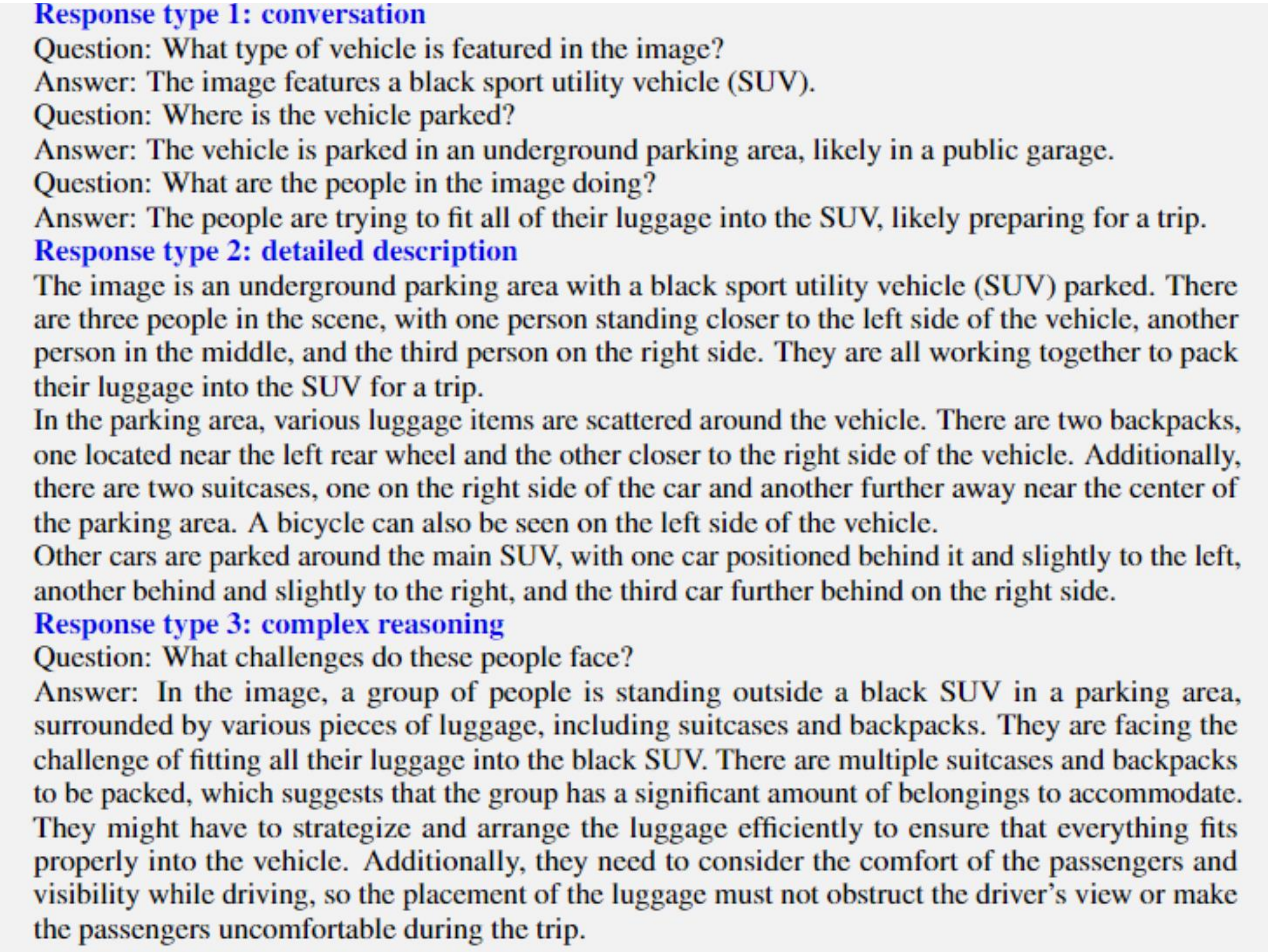
Context type 2: Boxes

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



GPT-assisted Visual Instruction Data Generation

Three type of instruction-following responses



Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

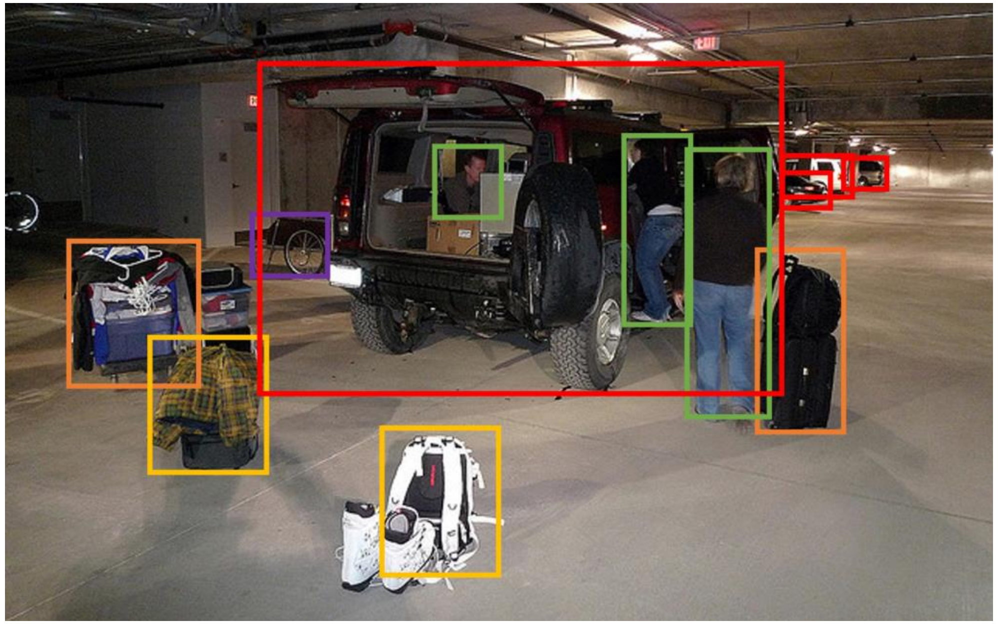
Response type 2: detailed description

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

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

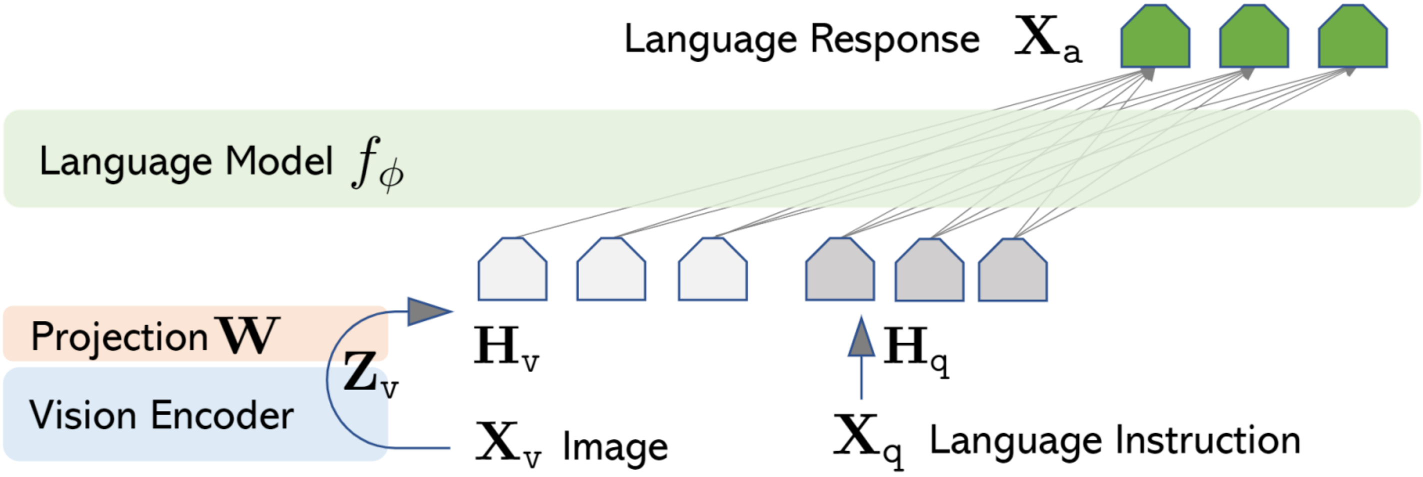
Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.



LLaVA: Large Language-and-Vision Assistant

❑ Architecture

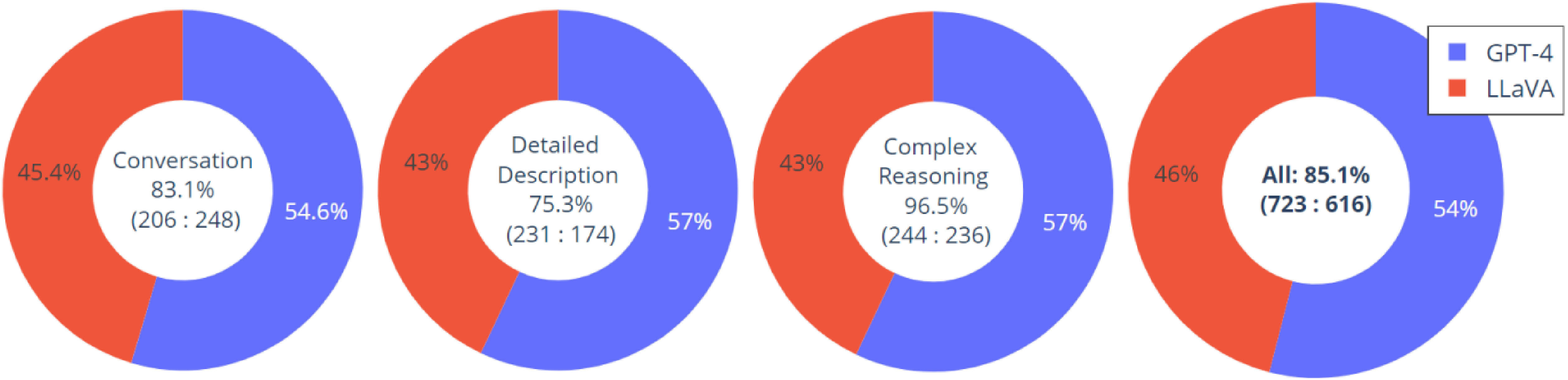


❑ Two-stage Training

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

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

Visual Chat: Towards building multimodal GPT-4 level chatbot

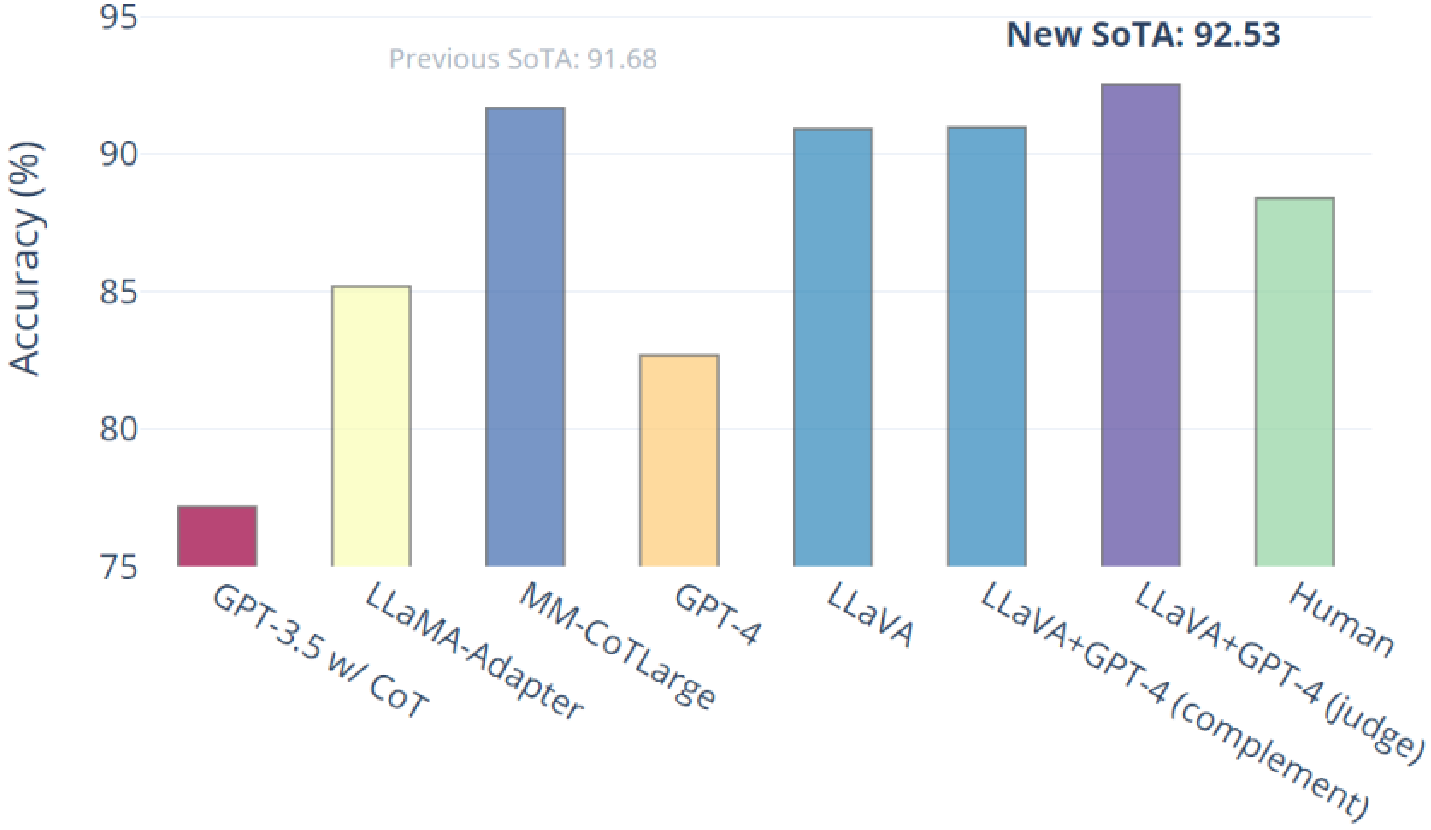


An evaluation dataset with 30 unseen images, 90 new language-image instructions

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

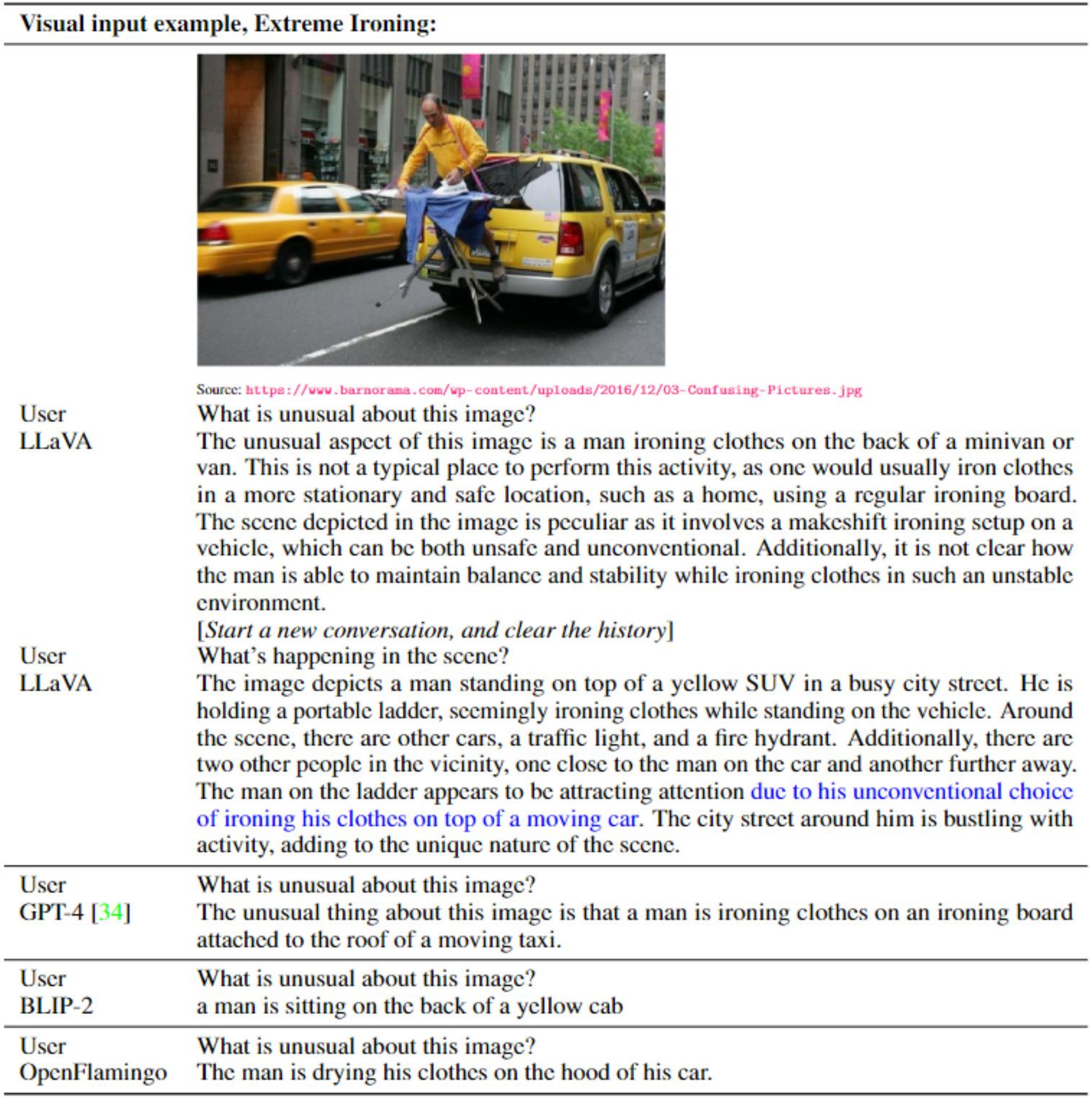
Science QA: New SoTA with the synergy of LLaVA with GPT-4

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



Strong Visual Reasoning Ability

Example 1: Extreme Ironing

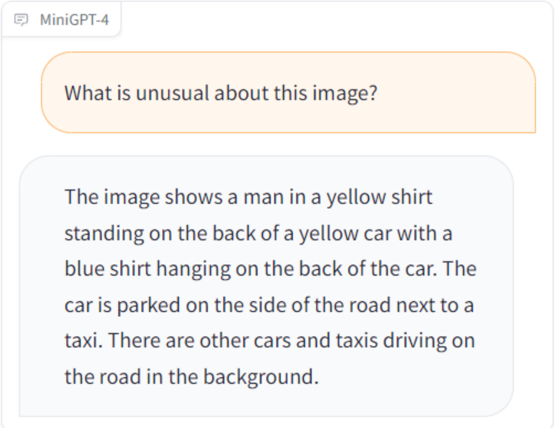


Visual input example, Extreme Ironing:

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

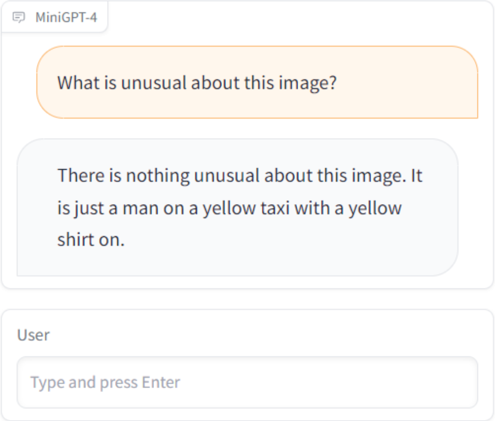
User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing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

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.



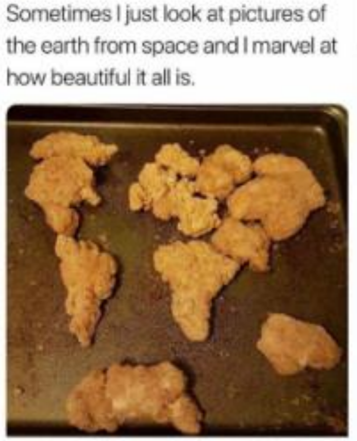
What is unusual about this image?

There is nothing unusual about this image. It is just a man on a yellow taxi with a yellow shirt on.

Strong Visual Reasoning Ability

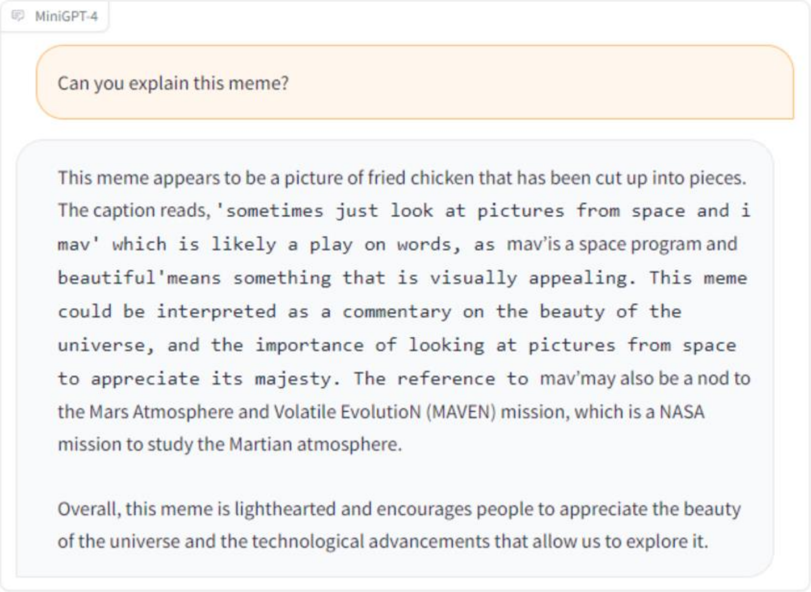
Example 2: Chicken Nugget Map

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	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	sometimes i just look at pictures of the earth from space and marvel how beautiful it is
User	Can you explain this meme in detail?
OpenFlamingo	It's a picture of a chicken nugget on the International Space Station.

mini-GPT4



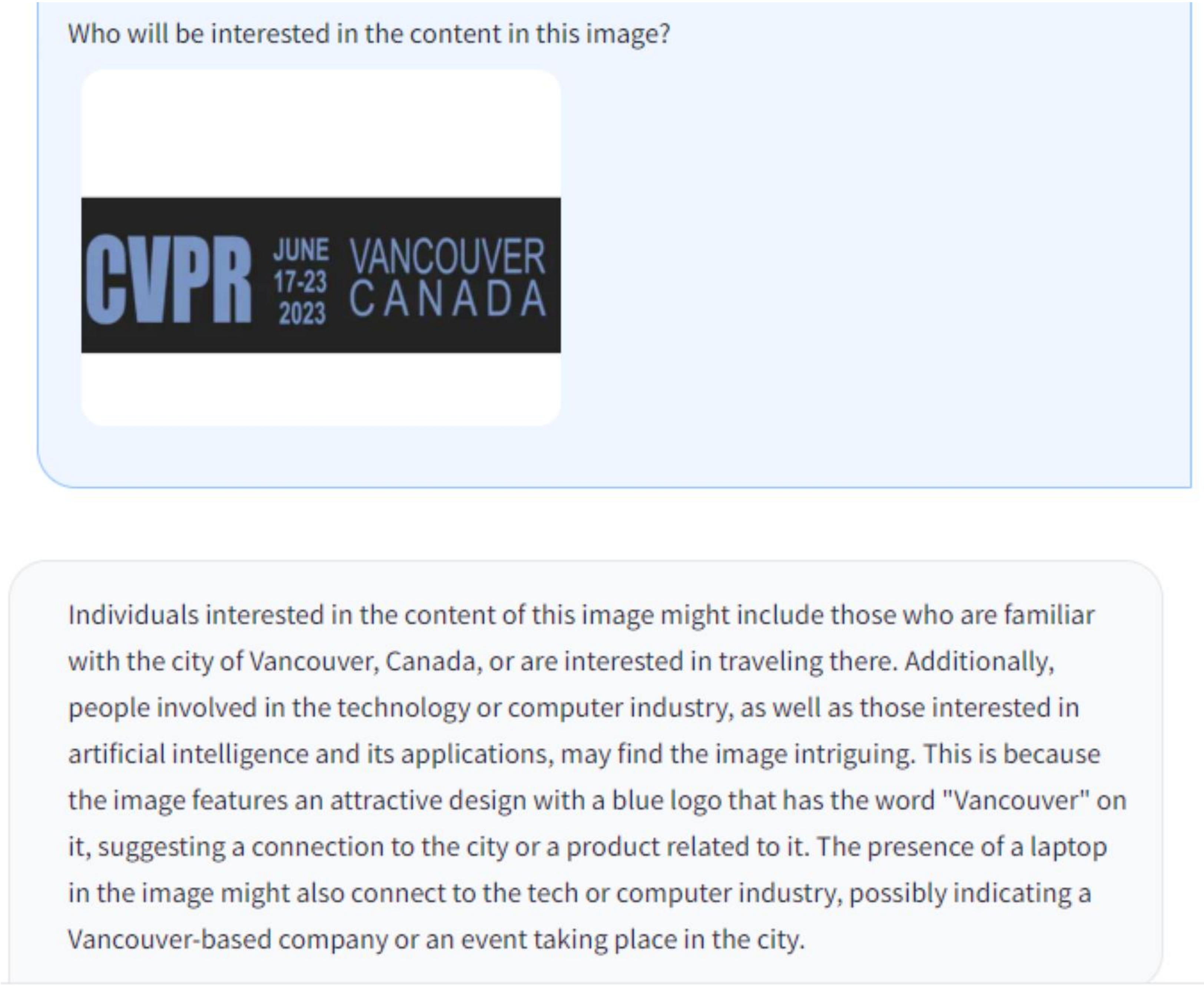
Can you explain this meme?

This meme appears to be a picture of fried chicken that has been cut up into pieces. The caption reads, 'sometimes just look at pictures from space and i mav' which is likely a play on words, as mav'is a space program and beautiful'means something that is visually appealing. This meme could be interpreted as a commentary on the beauty of the universe, and the importance of looking at pictures from space to appreciate its majesty. The reference to mav'may also be a nod to the Mars Atmosphere and Volatile EvolutioN (MAVEN) mission, which is a NASA mission to study the Martian atmosphere.

Overall, this meme is lighthearted and encourages people to appreciate the beauty of the universe and the technological advancements that allow us to explore it.

Strong OCR Ability

Example 1: CVPR & Vancouver

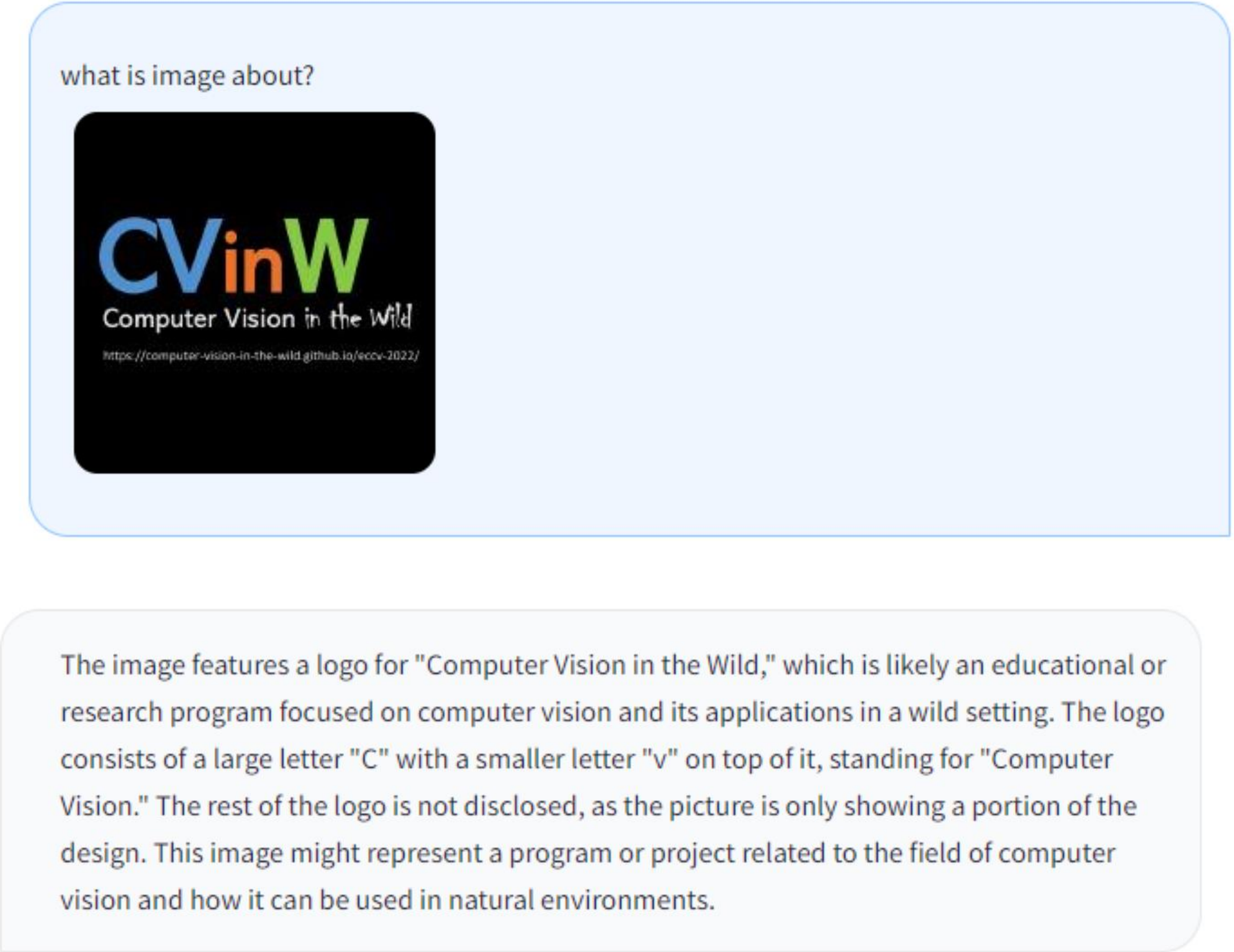


Who will be interested in the content in this image?

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

Strong OCR Ability

Example 2: CVinW workshop



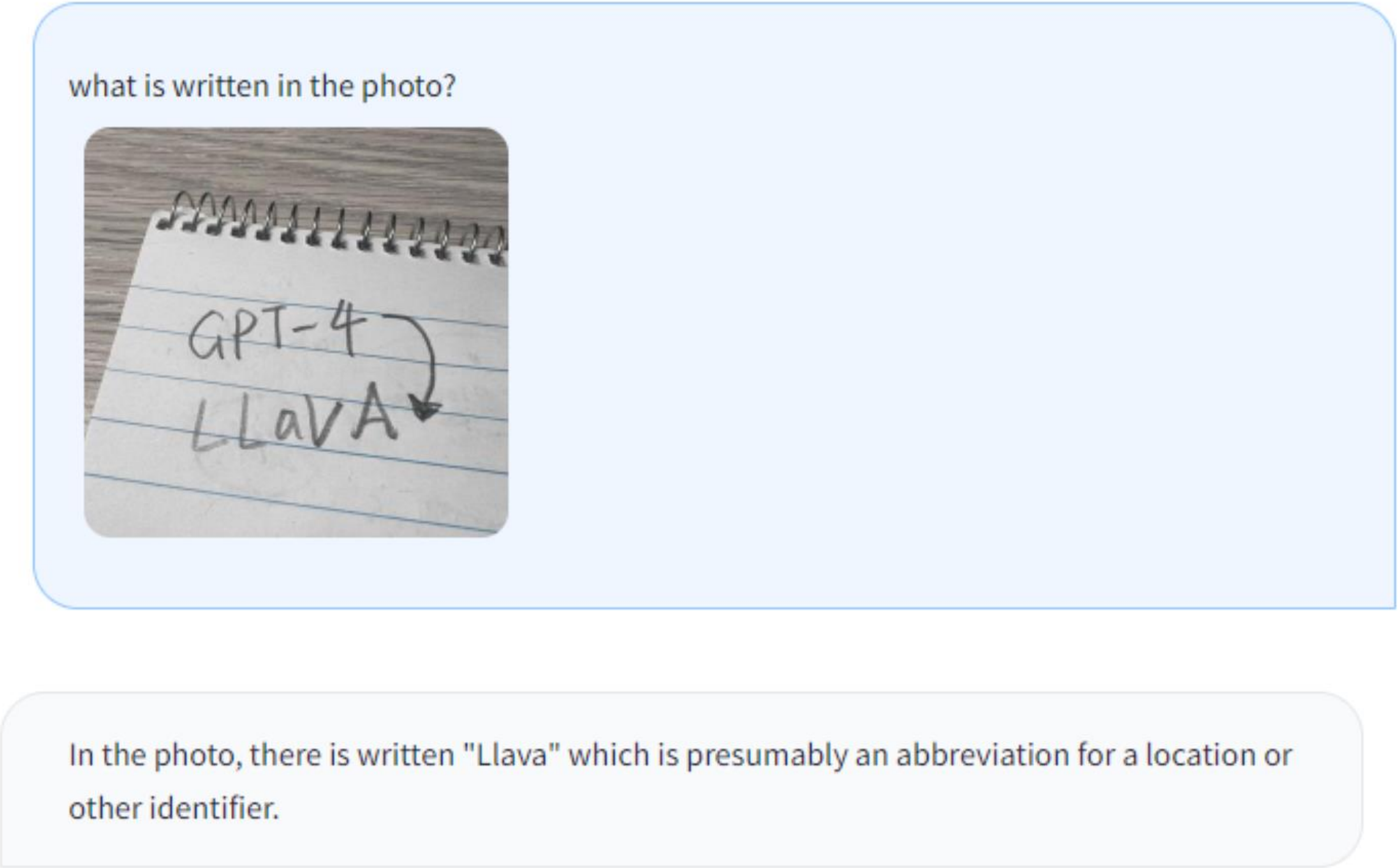
what is image about?

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

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.

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

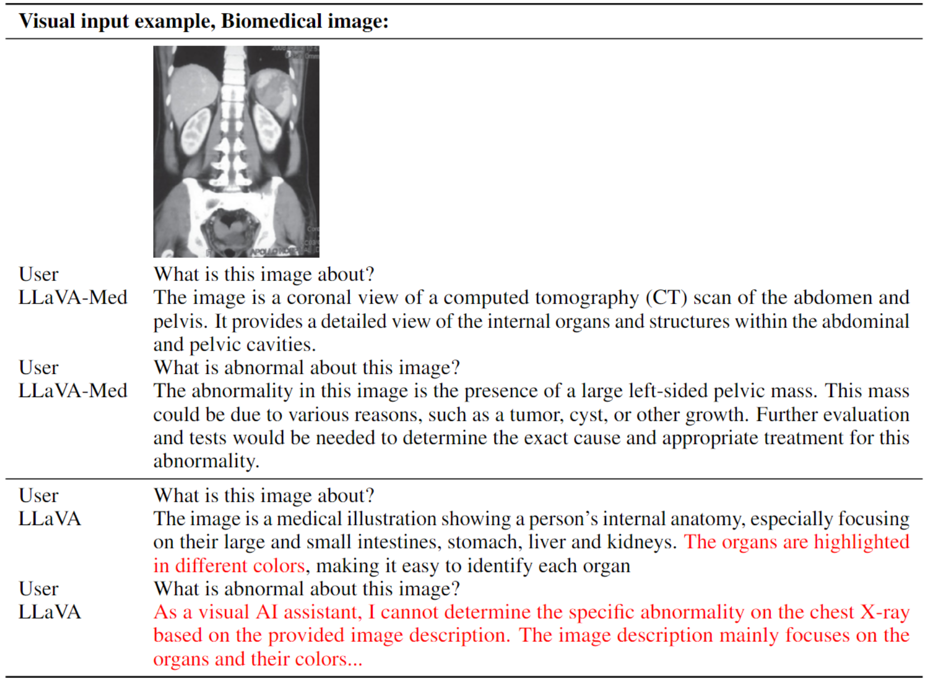
8 Hours

3 epochs on 60K samples

LLaVA-Med

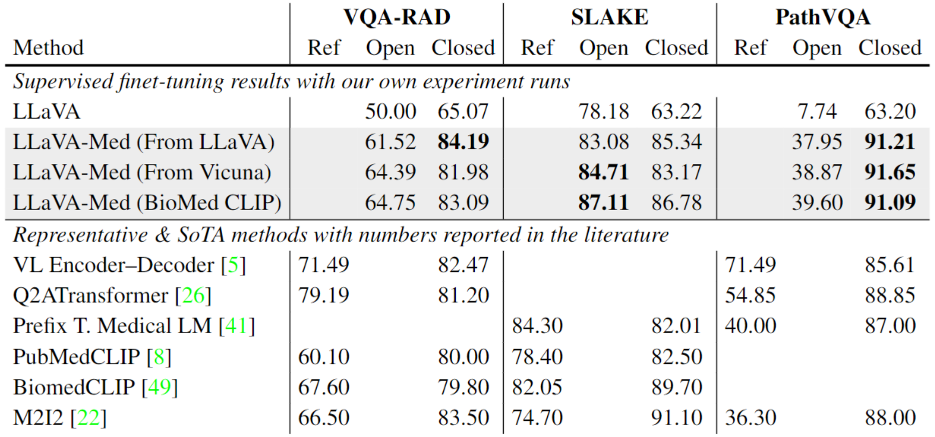
Downstream

- ☐ 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...



	VQA-RAD			SLAKE			PathVQA		
Method	Ref	Open	Closed	Ref	Open	Closed	Ref	Open	Closed
Supervised finet-tuning results with our own experiment runs									
LLaVA		50.00	65.07		78.18	63.22		7.74	63.20
LLaVA-Med (From LLaVA)		61.52	**84.19**		83.08	85.34		37.95	**91.21**
LLaVA-Med (From Vicuna)		64.39	81.98		**84.71**	83.17		38.87	**91.65**
LLaVA-Med (BioMed CLIP)		64.75	83.09		**87.11**	86.78		39.60	**91.09**
Representative & SoTA methods with numbers reported in the literature									
VL Encoder–Decoder [5]	71.49		82.47				71.49		85.61
Q2ATransformer [26]	79.19		81.20				54.85		88.85
Prefix T. Medical LM [41]				84.30		82.01	40.00		87.00
PubMedCLIP [8]	60.10		80.00	78.40		82.50			
BiomedCLIP [49]	67.60		79.80	82.05		89.70			
M2I2 [22]	66.50		83.50	74.70		91.10	36.30		88.00

CVinW

Foundation Models

GLIGEN
Text-to-Image

LLaVA
Image-to-Text

Better Alignment with Human Intent

Q&A | Thanks

Generated by GLIGEN