

CVPR2024

Tutorial -- Generalist Agent AI

Opening Remarks

Jianfeng Gao

Microsoft Research, 6/18/2024

https://multimodalagentai.github.io/

Overview

Generalist Agent AI (GAA) is a family of systems that generate effective actions in a given environment based on the understanding of multimodal sensory input. With the advent of large foundation models, numerous GAA systems have been proposed in fields ranging from basic research to applications. While these research areas are growing rapidly by integrating with the traditional technologies of each domain, they share common interests such as data collection, benchmarking, and ethical perspectives. In this tutorial, we focus on the some representative research areas of Embodied GAA, namely embodied-multimodality, robotics, gaming (VR/AR/MR), and healthcare, etc., and we aim to provide comprehensive knowledge on the common concerns discussed in these fields. As a result we expect the participants to learn the fundamentals of GAA and gain insights to further advance their research. Specific learning outcomes include:

- **GAA Overview:** A deep dive into its principles and roles in contemporary applications, providing attendees with a thorough grasp of its importance and uses.
- **Methodologies:** Detailed examples of how large foundation model enhance GAAs, illustrated through case studies in embodied virtual and real world, e.g., robotics, gaming, and healthcare.
- **Performance Evaluation:** Guidance on the assessment of GAAs with relevant datasets, focusing on their effectiveness and generalization.
- **Ethical Considerations:** A discussion on the societal impacts and ethical challenges of deploying Agent AI, highlighting responsible development practices.
- **Emerging Trends and Future Challenges:** Categorize the latest developments in each domain and discuss the future directions.

Agenda

- Success of LLMs
- LLMs as AI Agents
- This tutorial

Success of Large Language Models (LLMs)

(1) • Context length…

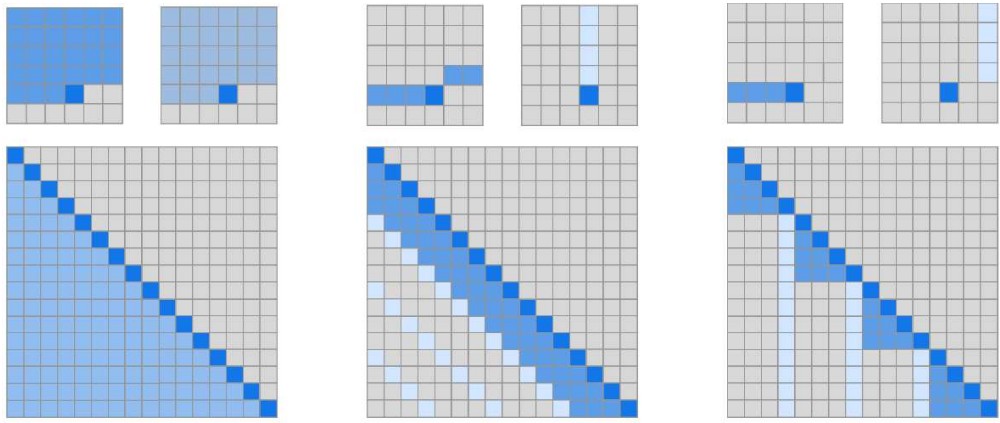
(2) • Scaling laws…

(3) • Emergent abilities
  - In-context-learning
  - LLMs as (general-purpose) task solvers

(1) Context length of Language Models (LMs)

- LM: $P(w \mid h)$
  - LMs are better with longer (richer) context h
- N-gram LMs: $|h|$ = 1 to 6
  - Model size grows exponentially with $|h|$
- RNN/LSTM LMs: $P(w \mid c(h))$,
  - where h is *compressed* to a fixed-size vector
- Transformer LMs: $|h|$ = 2K to 8K … to 1M
  - Sparse attention is used to deal with quadratic complexity

Generating Long Sequences with Sparse Transformers, 2019



(a) Transformer (b) Sparse Transformer (strided) (c) Sparse Transformer (fixed)

(2) Scaling laws

[Kaplan+ 20, Hoffmann+ 22]

- Power-law relationship of
  - L – model performance w/
  - N – model size
  - D – Training data size
  - C – amount of training compute

$$L(N) = \left(\frac{N_c}{N}\right)^{\alpha_N}, \quad \alpha_N \sim 0.076, N_c \sim 8.8 \times 10^{13}$$

$$L(D) = \left(\frac{D_c}{D}\right)^{\alpha_D}, \quad \alpha_D \sim 0.095, D_c \sim 5.4 \times 10^{13}$$

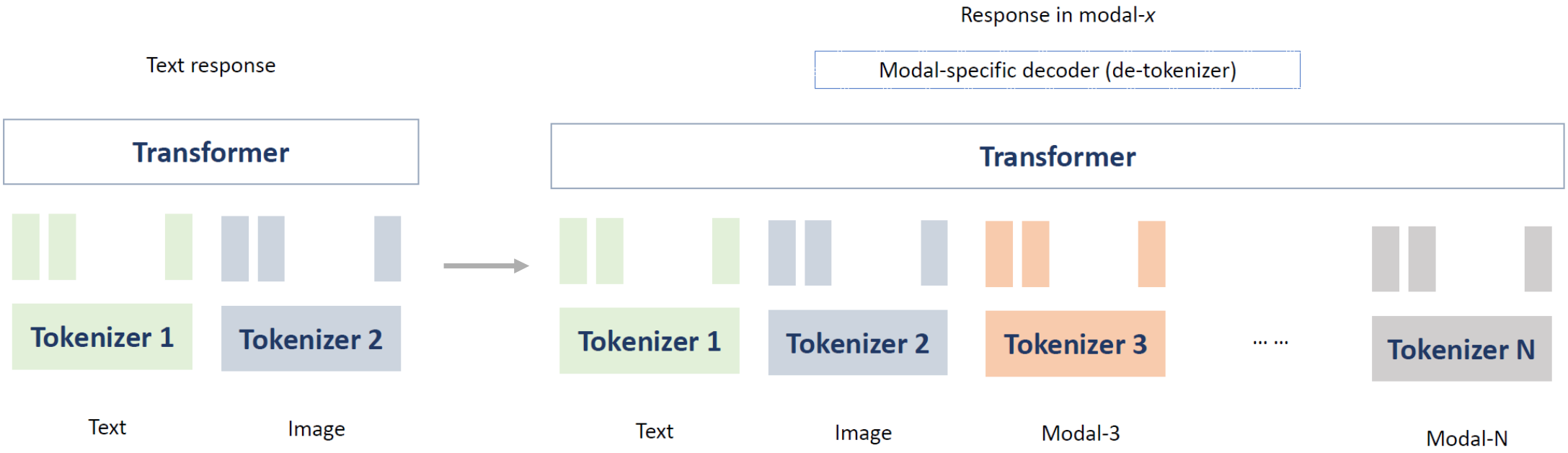
$$L(C) = \left(\frac{C_c}{C}\right)^{\alpha_C}, \quad \alpha_C \sim 0.050, C_c \sim 3.1 \times 10^{8}$$

- We can keep improving LLMs by
  - increasing model capacity (context length)
  - increasing training data (raw text)
  - Increasing compute ($$$)

➢ J. Kaplan etc, "**Scaling Laws for Neural Language Models**," 2020. [OpenAI]

- Scaling Laws : "Scaling Laws"는 LLMs의 성능이 어떻게 "모델 크기(파라미터 수), 학습 데이터량, 연산량(Compute)"에 따라 변화하는지를 수학적으로 설명하는 법칙

(3) LLMs as LMMs (e.g., LLaVA, Phi-3-Vision)



(3) LLMs/LMMs as Agent Models

- LMMs predict word/visual tokens
  - $(w_1, w_2, \dots, w_t) \rightarrow w_{t+1}$

- AI Agents predict observation-action sequences
  - Policy model: $(o_1, a_1, o_2, a_2, \dots, o_t) \rightarrow a_t$ **action**
  - World model: $(o_1, a_1, o_2, a_2, \dots, o_t, a_t) \rightarrow o_{t+1}$ **observation**

- LMMs are agent models if we can
  - Tokenize observations
  - Tokenize actions

(3) LLMs as Agents

[Brohan+ 23 (Google), Durante+ 24 (MSR)]

A. Brohan, etc, “RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control,” 2023, (Google Deepmind)

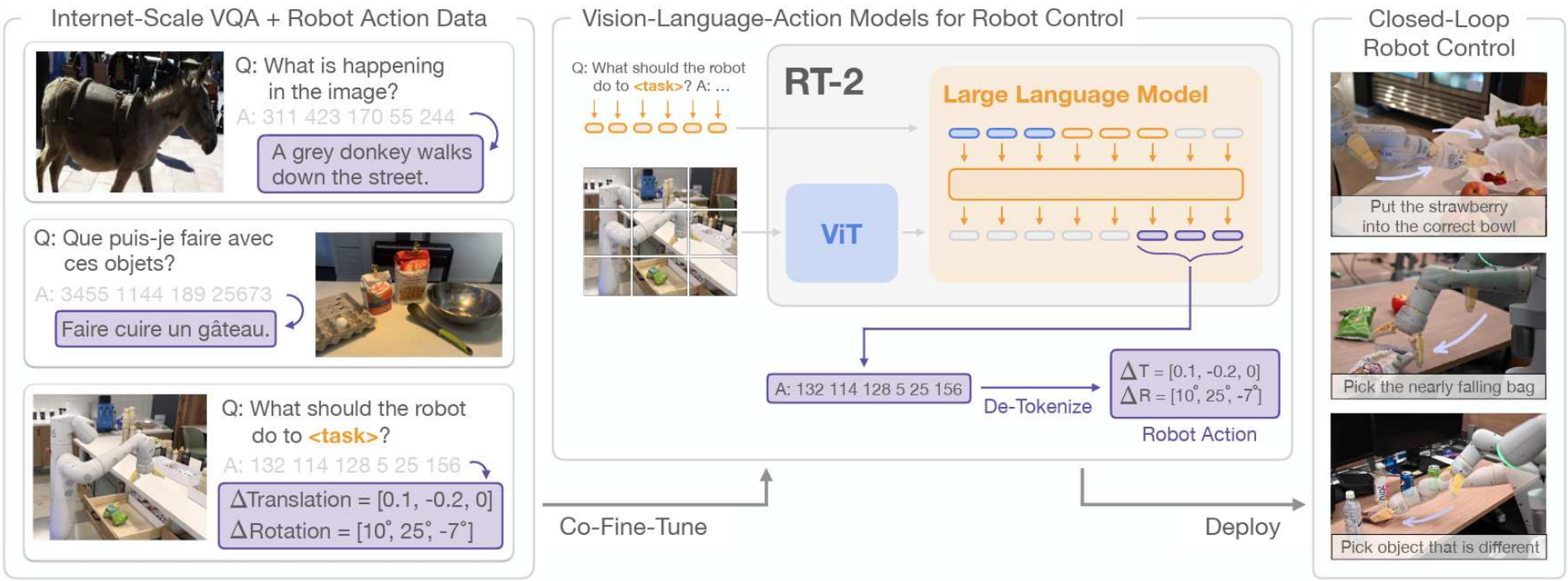


Figure1 RT-2 overview: We represent **robot actions** as another language, which can be **cast into text tokens** and **trained together with Internet-scale vision-language datasets**. **During inference, the text tokens are de-tokenized into robot actions, enabling closed loop control**. This allows us to **leverage the backbone and pretraining of vision-language models** in **learning robotic policies, transferring some of their generalization**, **semantic understanding**, and **reasoning to robotic control**. We demonstrate examples of RT-2 execution on the project website: robotics-transformer2.github.io.

(3) LLMs as Agents

[Brohan+ 23 (Google), Durante+ 24 (MSR)]

Z. Durante, etc, “Agent AI: Surveying the Horizons of Multimodal Interaction,” 2024. Stanford Univ., Microsoft Research

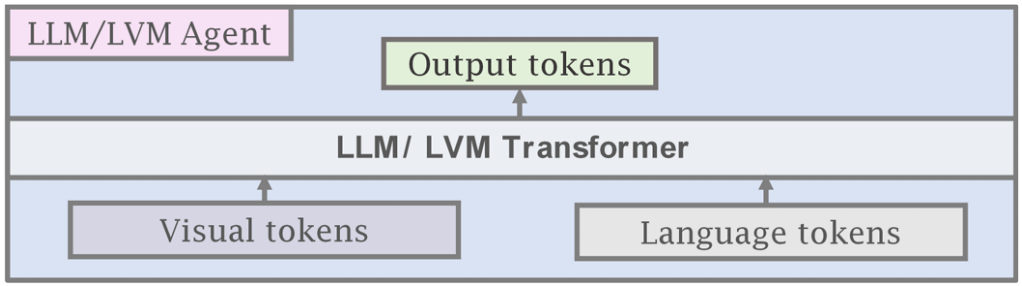


Figure 6: We show **the current paradigm for creating multi-modal AI agents** by incorporating a **Large Language Model (LLM)** with a **Large Vision Model (LVM)**. Generally, these models take ***visual or language inputs*** and use ***pre-trained*** and ***frozen visual and language models***, **learning smaller sub-network that connect and bridge modalities**.

Examples include **Flamingo** (Alayrac et al., 2022), **BLIP-2** (Li et al., 2023c), **InstructBLIP** (Dai et al., 2023), and **LLaVA** (Liu et al., 2023c).

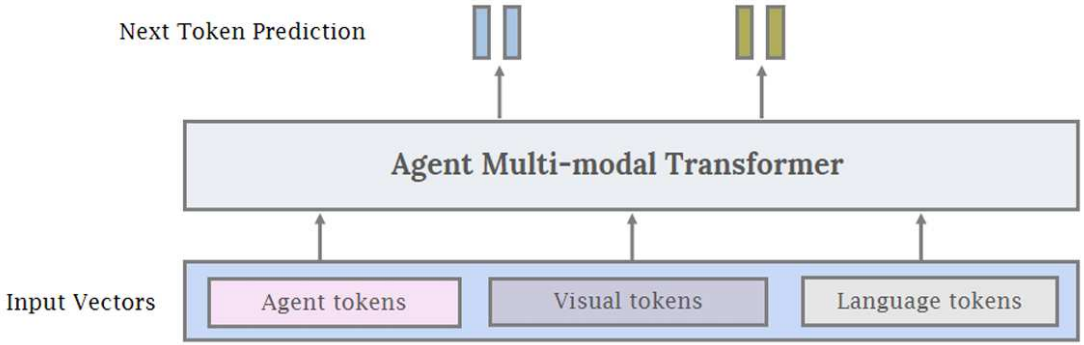
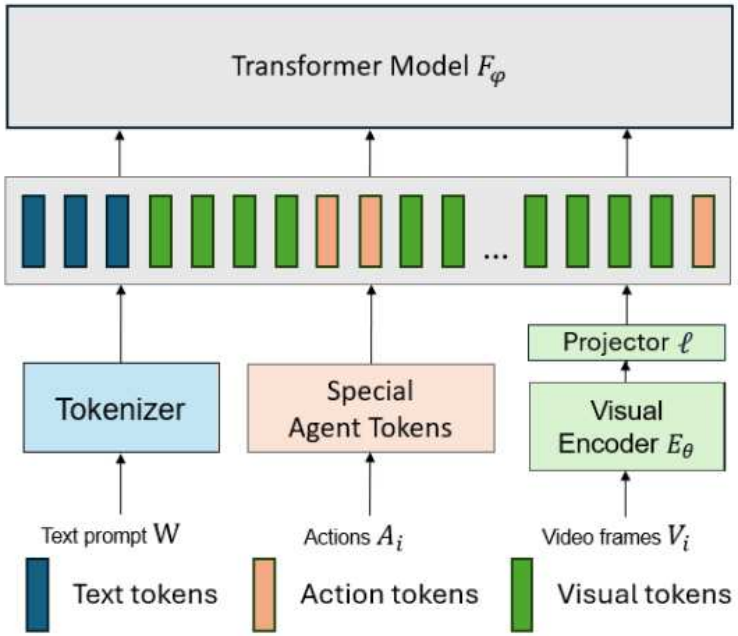


Figure 7: **The unified agent multi-modal transformer model**. Instead of connecting frozen submodules and using existing foundation models as building blocks, we propose a unified and end-to-end training paradigm for agent systems.

We can still **initialize the submodules with LLMs and LVMs** as in Figure 6 but also make use of **agent tokens**, **specialized tokens for training the model to perform agentic behaviors in a specific domain** (e.g., robotics). For more details about agent tokens, see Section 3.2

(3) LLMs as Agents

[Brohan+ 23 (Google), Durante+ 24 (MSR)]



- Text tokenizer
- Vision (observation) tokenizer – videos as another language
- Action tokenizer – (robot) actions as another language

Z. Durante, etc, “Agent AI: Surveying the Horizons of Multimodal Interaction,” 2024. Stanford Univ., Microsoft Research

LMM-powered Agents… But what is lacking?

- Hallucination
  - Grounding – augmenting LMMs w/ knowledge & tools
  - Asking why – making LMMs causal and interpretable
- Cope with when things do not go as planned
  - Learning thru AI-human interactions
  - Self-improving via continual learning w/o catastrophic forgetting
- More…
  - World Model vs. Action Model? One model for all?
  - EAI models for robotics or copilots?
  - What is the *word prediction* task in pretraining EAI-FMs?
  - Context?
  - Scaling laws?
  - Tokenizers or not?
  - What are the emergent abilities?

Main Issues;

1) 명확하고 신뢰할 수 있는 추론 능력 부족(Reasoning / Hallucination)
2) 장기 메모리(Long-term memory) 및 문맥 유지 한계
3) 외부 도구와의 연동 및 실행 제어
4) 학습 데이터 품질 및 편향(Bias) 문제
5) 프롬프트 엔지니어링(Prompt Engineering) 의존성
6) 안전성(Safety) 및 보안(Security)
7) 설명 가능성(Explainability) 및 해석 가능성(Interpretability) 부족
8) 실시간 의사소통 및 지속적 학습(Online / Continual Learning)

Tutorial Talk2

Generalist Multimodal Models

Yong Jae Lee

University of Wisconsin-Madison



Image of LLaVA generated by GLIGEN
"a cute lava llama with glasses" + box prompt

Generalist Multimodal Models

By Yong Jae Lee, University of Wisconsin-Madison

Computer Vision in the Deep Learning Era (2012 - Present)

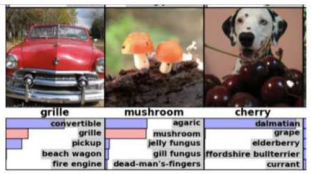
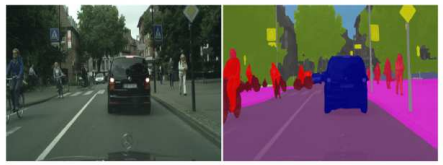
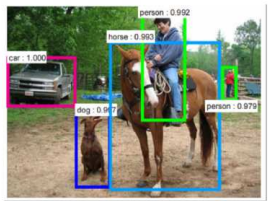


Image classification



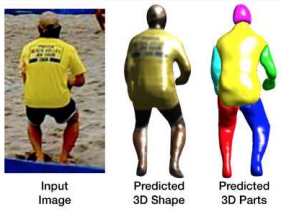
Semantic segmentation



Object detection



Pose recognition



3D prediction

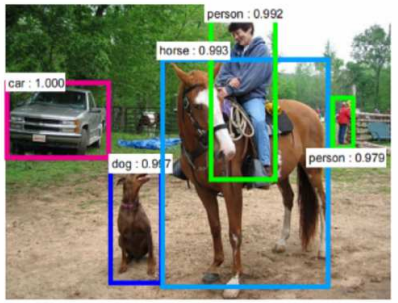


Surface normal prediction

... and many more

However, Prevailing Paradigm Thus Far:

"Specialist" models: **single-model, single-task**



Object Detection *Only*

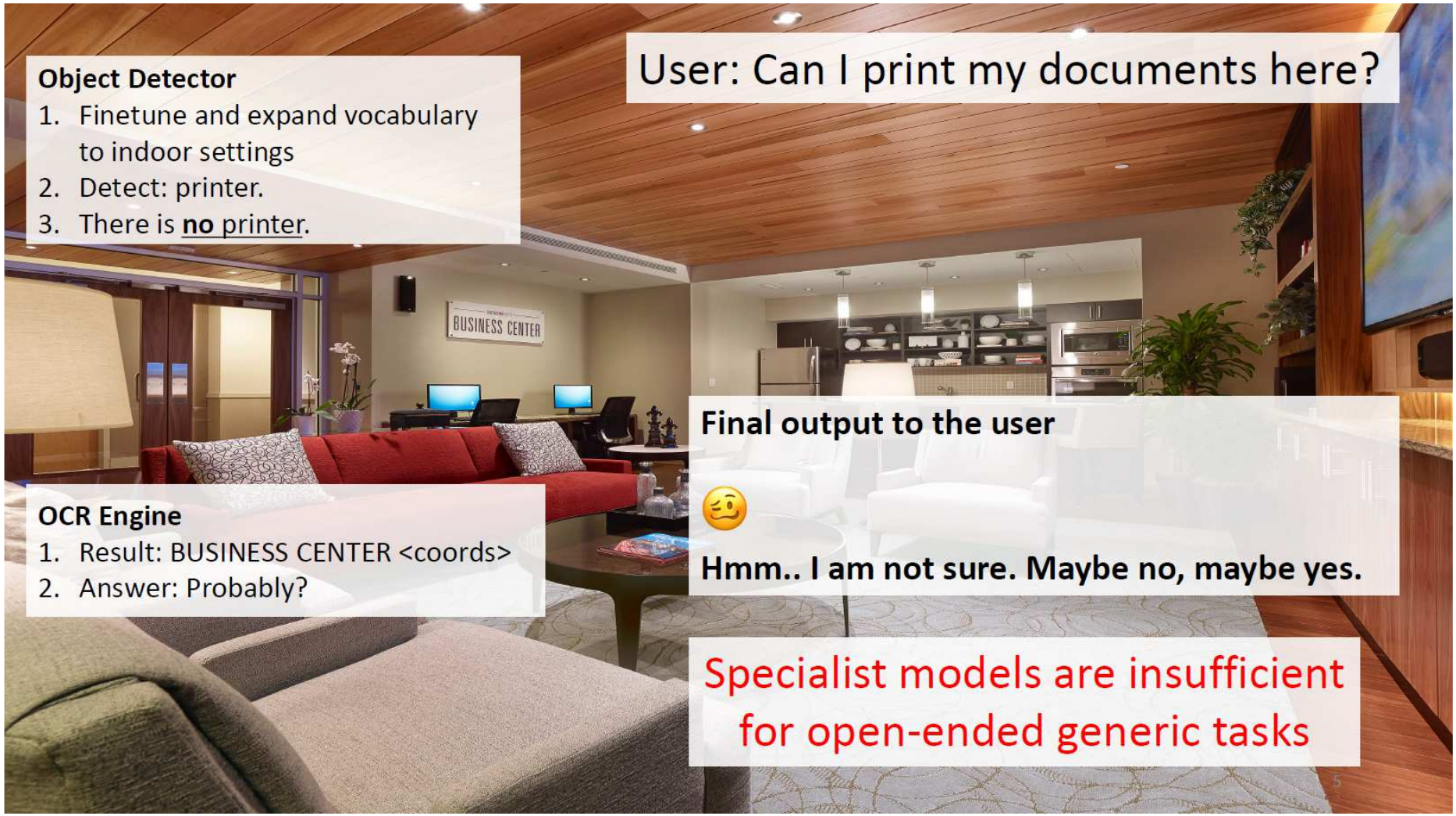


Pose Recognition *Only*

- **It does well but it can't solve any other task as another**

Generalist Multimodal Models

By Yong Jae Lee, University of Wisconsin-Madison

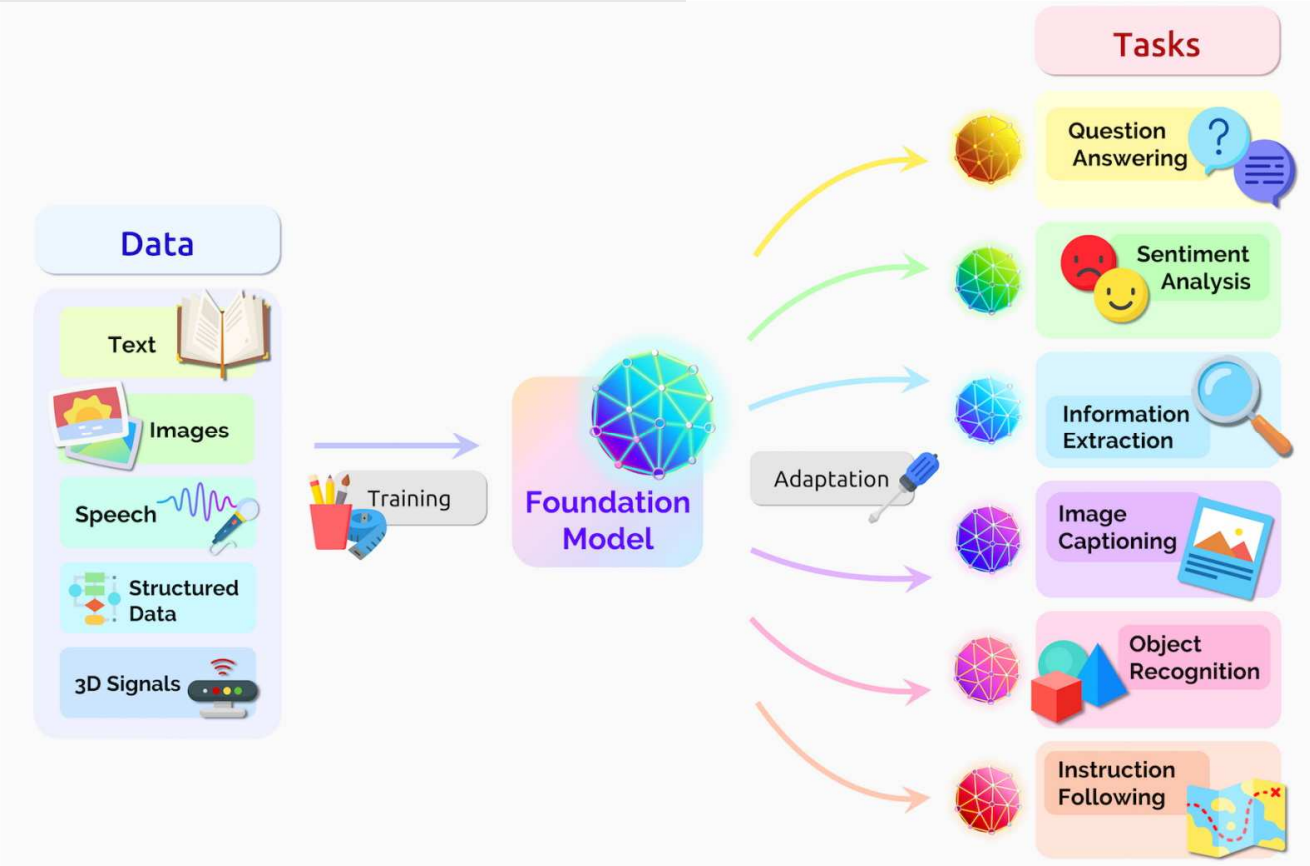


Generalist Multimodal Models

By Yong Jae Lee, University of Wisconsin-Madison

Rise of “Generalist” Foundation Models (2020s)

- **Single-model, many tasks**
- Large Language Models (e.g., GPT4)
- Vision Transformers
- Image-Text Models (e.g., CLIP)



Since 2021, researchers have identified an array of uses for foundation models.

Image credit: https://blogs.nvidia.com/blog/what-are-foundation-models/

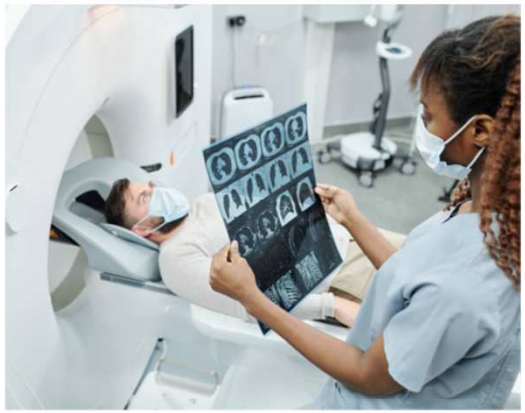
Generalist Multimodal Models

By Yong Jae Lee, University of Wisconsin-Madison

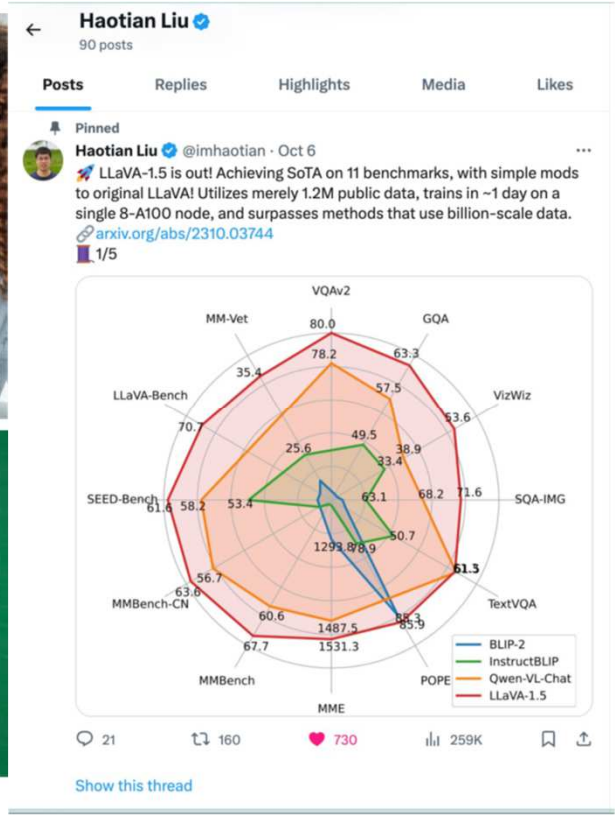
Today's talk:
Large Generalist Multimodal Models

- ***Generalist*** vision-language models that understand visual data and communicate in natural language
- ***Controllable*** ("*aligned*") models that produce desirable outputs for wide-concept knowledge
- Challenge: How to effectively train such models with minimal supervision?
- Solution: Adapt pre-trained foundation models, and design semiautomatic methods for data collection

Humans See and Reason about the Visual World; Express and Interact with Natural Language

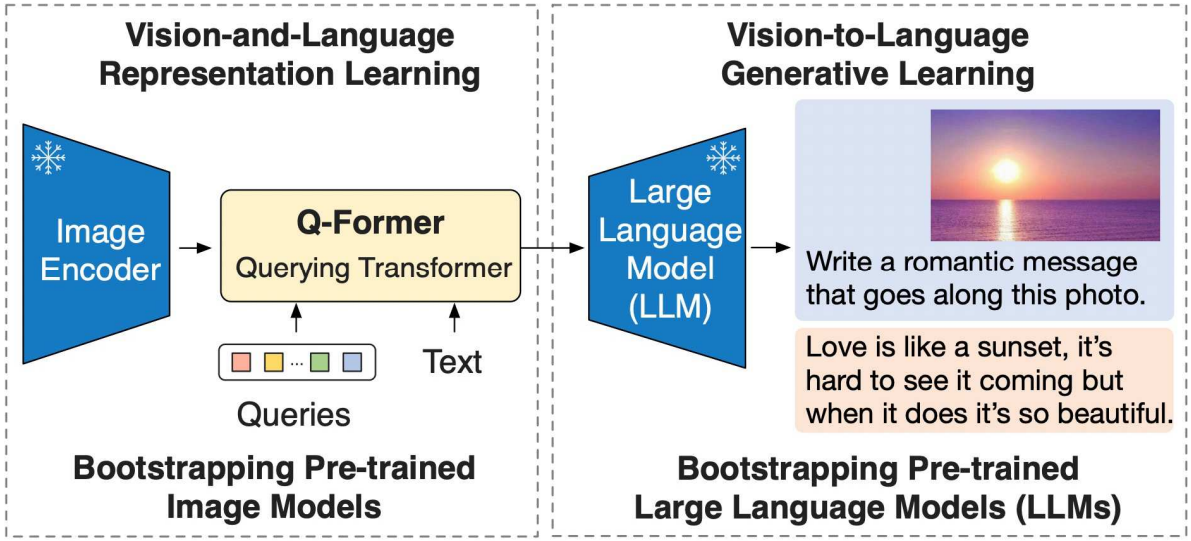






By Yong Jae Lee, University of Wisconsin-Madison

How to Build Generalist Multimodal Models?



☞ Junnan Li et al., "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," arxiv, 2023 (Salesforce Research)

Figure 1. Overview of **BLIP-2's framework**. We pre-train a **lightweight Querying Transformer** following a two-stage strategy to bridge the modality gap.

1) The first stage bootstraps vision-language representation learning from a frozen image encoder.
2) The second stage bootstraps vision-to-language generative learning from a frozen LLM, which enables zero-shot instructed image-to-text generation

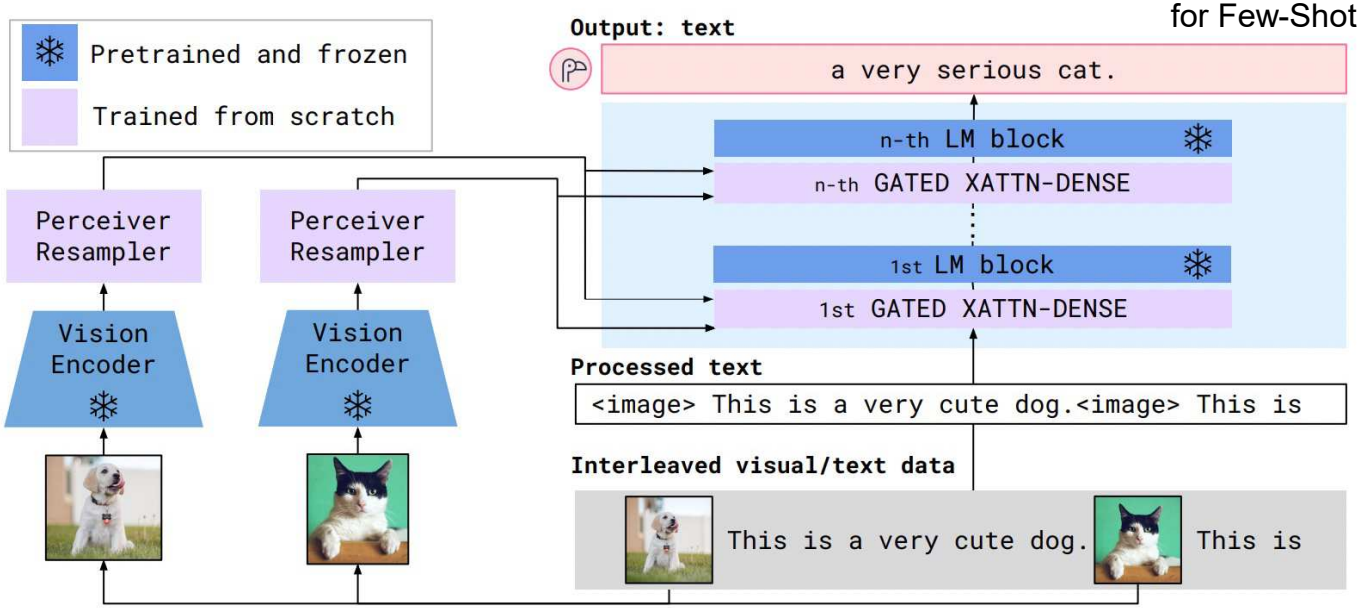
- Prior methods (before early 2023) generally **lacked *instruction following capabilities***

[J. Wang et al. 2022. GIT: A Generative Image-to-text Transformer for Vision and Language]
[J. Li et al. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models]
[J.-B. Alayrac. 2022. Flamingo: a visual language model for few-shot learning]
…

By Yong Jae Lee, University of Wisconsin-Madison

How to Build Generalist Multimodal Models?

☞ Jean-Baptiste Alayrac et al., “Flamingo: a Visual Language Model for Few-Shot Learning,” arxiv, 2022. (Deepmind)



1) First, the **Perceiver Resampler** receives **spatio-temporal features** from the ***Vision Encoder*** (obtained from either an image or a video) and **outputs *a fixed number of visual tokens***.
2) Second, **these *visual tokens*** are used to **condition the frozen LM** using freshly *initialised cross-attention layers* that are interleaved between the pretrained LM layers. These new layers offer an expressive way for the LM to incorporate visual information for the next-token prediction task.

Figure 3: **Flamingo architecture overview**.

- Prior methods (before early 2023) generally **lacked *instruction following capabilities***

Instruction following 모델 : 주어진 **지시문(Instructions)**이나 **질의(Query)**를 이해하고 그대로 수행하도록 훈련된 LLM(Large Language Model)을 의미함. 일반적인 LLM이 주어진 텍스트(Prompt)에 대해 맥락상 가장 그럴듯한 텍스트를 생성하려고 한다면, Instruction-Following 모델은 **"사용자의 의도를 파악하고, 그에 맞춰 답변을 구성하며, 필요 시 단계적 추론을 통해 정확하고 알맞은 결과를 도출"**하는 데 특화되어 있음.

By Yong Jae Lee, University of Wisconsin-Madison

How to Build Generalist Multimodal Models?

- GPT-4V: Strong language and visual reasoning, but closed-source

GPT-4 visual input example, Extreme Ironing:	
User	What is unusual about this image?
GPT-4	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.

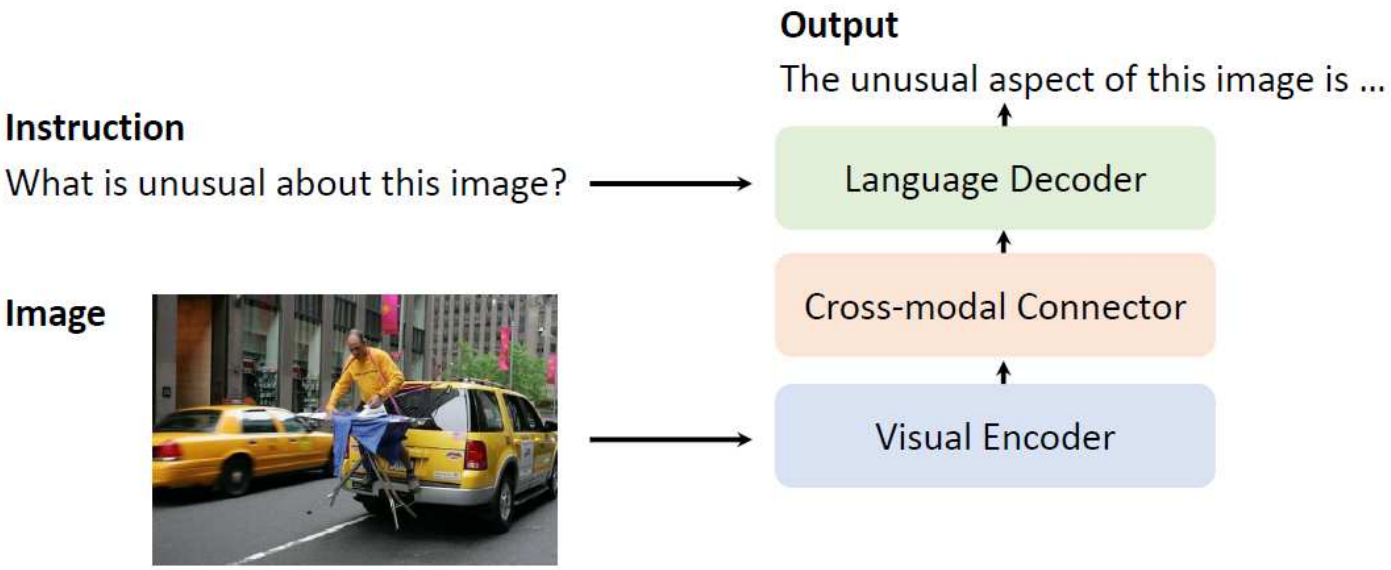


Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

[GPT-4 Technical Report, OpenAI. March 2023.]

By Yong Jae Lee, University of Wisconsin-Madison

How to Build Generalist Multimodal Models?



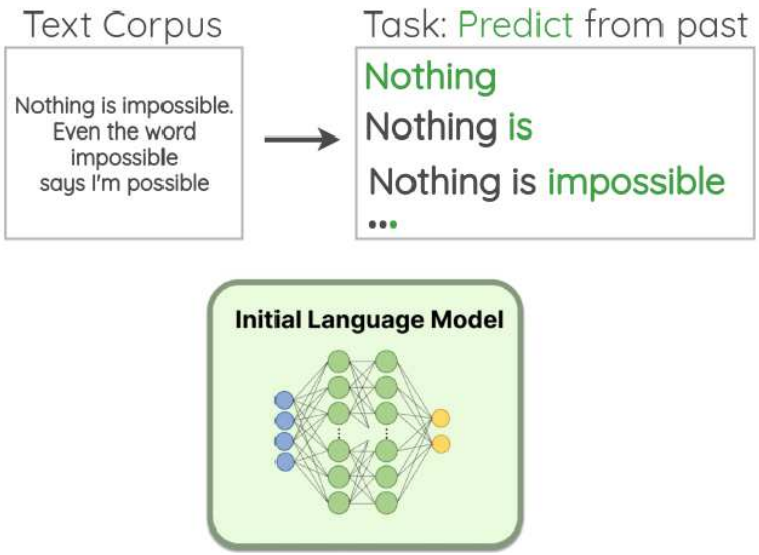
1. Data to tune the model for *instruction-following capabilities*
2. Strong pretrained vision and language models
3. Connecting vision and language

[GPT-4 Technical Report, OpenAI. March 2023.]

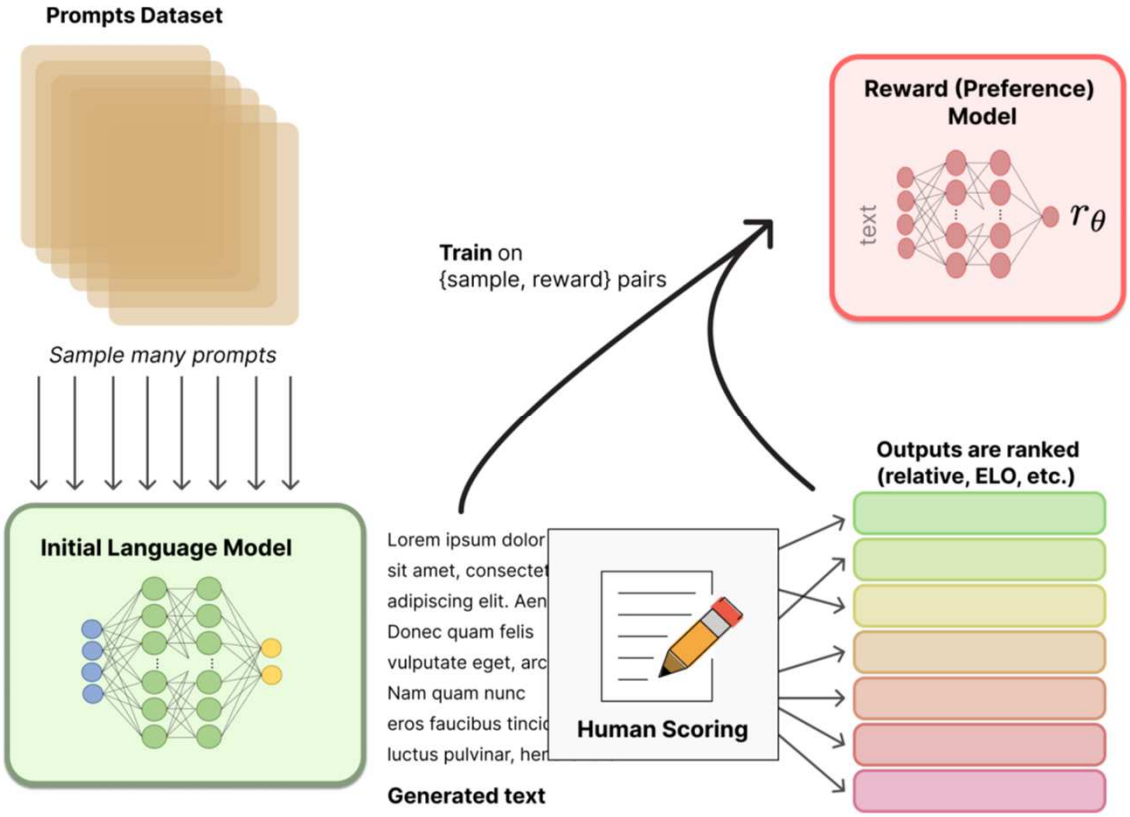
By Yong Jae Lee, University of Wisconsin-Madison

Lessons from Training Large Language Models

- Step 1: **Representation learning**
  - ✓ Self-supervised pre-training



- Step 2: **Instruction Tuning**
  - ✓ Supervised fine-tuning / RLHF



By Yong Jae Lee, University of Wisconsin-Madison

Lessons from **Instruction Tuning** in LLMs

Instruction	Output
Explain human's behavior. Behavior: cry.	There could be many reasons why a person might cry. They could be feeling sad, scared, angry, or frustrated...
Recommend a movie for me.	Certainly! Without specific preferences in mind, I'll recommend a popular and highly-rated movie: "The Shawshank Redemption." It's a classic drama film ...
...	...

By Yong Jae Lee, University of Wisconsin-Madison

Lessons from **Instruction Tuning** in LLMs

- How to collect instruction tuning data?
  - ✓ Human: high-quality, hand-written by humans → high cost
  - ✓ Machine: strong LLM-based teacher like ChatGPT → affordable cost

Manual seed instruction-output pairs

Explain human's behavior.	<reference answer>
Recommend a movie for me.	<reference answer>

Please generate new instruction-output pairs that meet the following requirements: ...

<new instruction 1>	<new output 1>

...

LLaMa → Alpaca, Vicuna, etc.

[Yizhong Wang. 2022. Self-instruct: Aligning language model with self generated instructions.]

By Yong Jae Lee, University of Wisconsin-Madison

Lessons from **Instruction Tuning** in LLMs

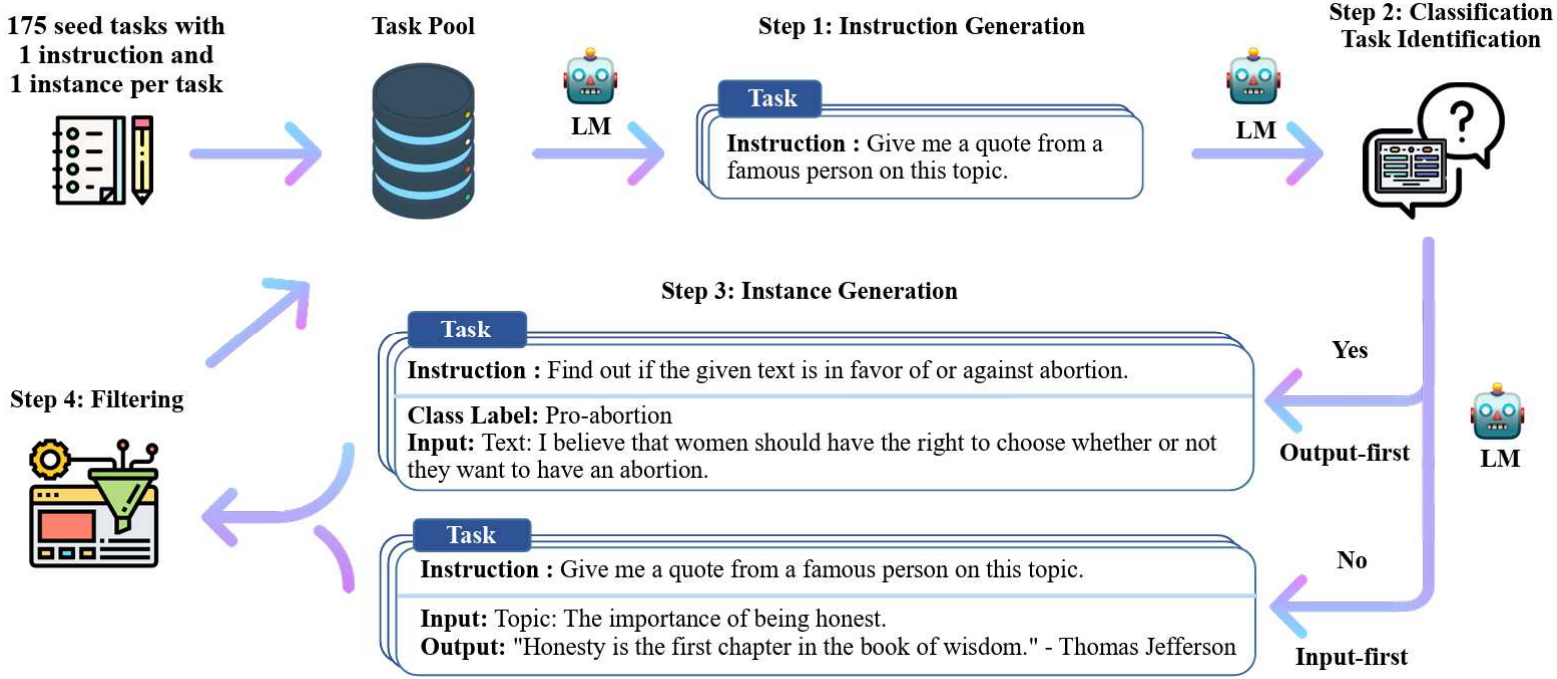


Figure 2: A high-level overview of **SELF-INSTRUCT**.

The process starts with **a small seed set of tasks** as **the task pool**.

Random tasks are sampled from the task pool and used to prompt an off-the-shelf LM to generate both new instructions and corresponding instances, followed by filtering low-quality or similar generations, and then added back to the initial repository of tasks.

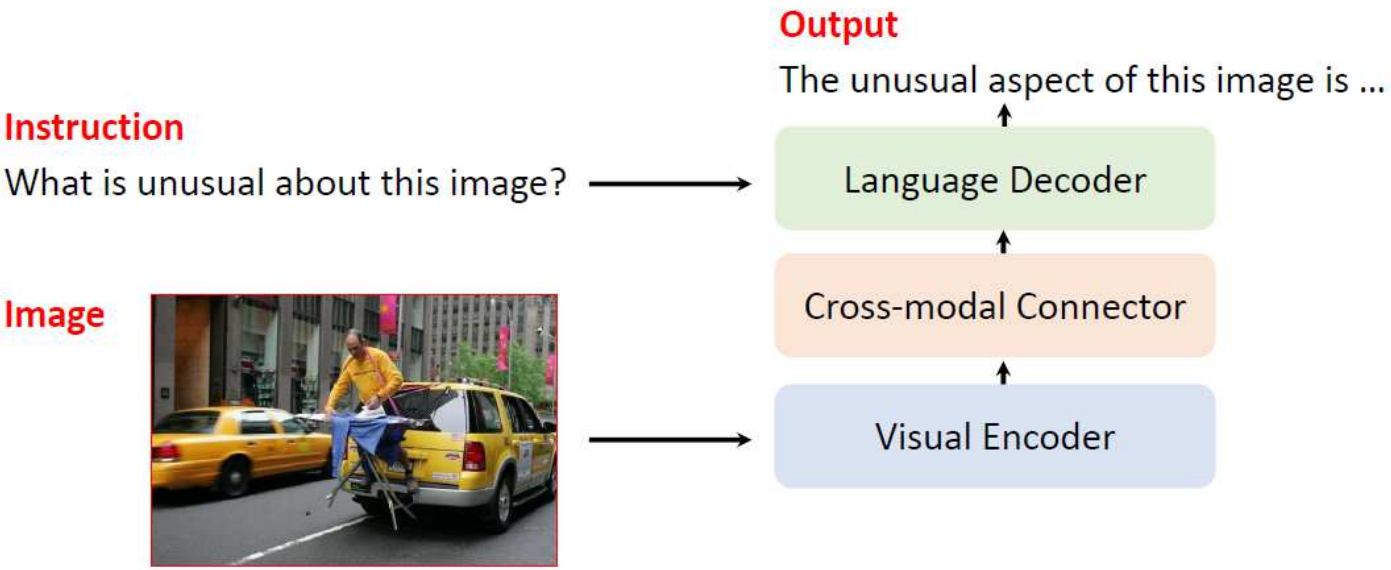
The resulting data can be used for the instruction tuning of the language model itself later to follow instructions better.

Tasks shown in the figure are generated by GPT3.

[Yizhong Wang. 2022. Self-instruct: Aligning language model with self generated instructions.]

By Yong Jae Lee, University of Wisconsin-Madison

Instruction Tuning in Large Multimodal Models



- How to obtain Image-Instruction-Output triplet training data?
- Can we use a large language model for this?

By Yong Jae Lee, University of Wisconsin-Madison

Text-only GPT Assisted Visual Instruction Data Creation

Image



- *How do we get a text-only LLM to "see" an image?*

→ Let it read context information written in text

Context (caption)

A group of people standing outside of a black vehicle with various luggage.

MS-COCO
Each image associated with:
- 5 captions
- Object categories / bounding boxes

Context (layout)



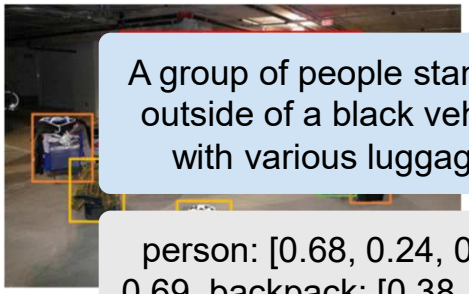
person: [0.68, 0.24, 0.77, 0.69], person: [0.63, 0.22, 0.68, 0.51], person: [0.44, 0.23, 0.48, 0.34], backpack: [0.38, 0.69, 0.48, 0.91], ...

[T.-Y. Lin et al. 2014. Microsoft coco: Common objects in context.]

By Yong Jae Lee, University of Wisconsin-Madison

Text-only GPT Assisted Visual Instruction Data Creation

Context (caption)



A group of people standing outside of a black vehicle with various luggage.

Context (layout)

person: [0.68, 0.24, 0.77, 0.69, backpack: [0.38, 0.69, 0.48, 0.91] …

Instruction	Output
What are the challenges these people might be facing?	They may be having difficulty fitting all luggage into the back of the SUV. There are many bags, suitcases already in the back, while more…

Manual seed example(s) of context-instruction-output triplets.



<new context (caption) 1>

<new context (layout) 1>

Please generate new Context-Instruction-Output triplets that meet the following requirements: …

Text-only ChatGPT

<new instruction 1>

<new output 1>

Visual Instruction-following Data:
Triplet (image, instruction, output)

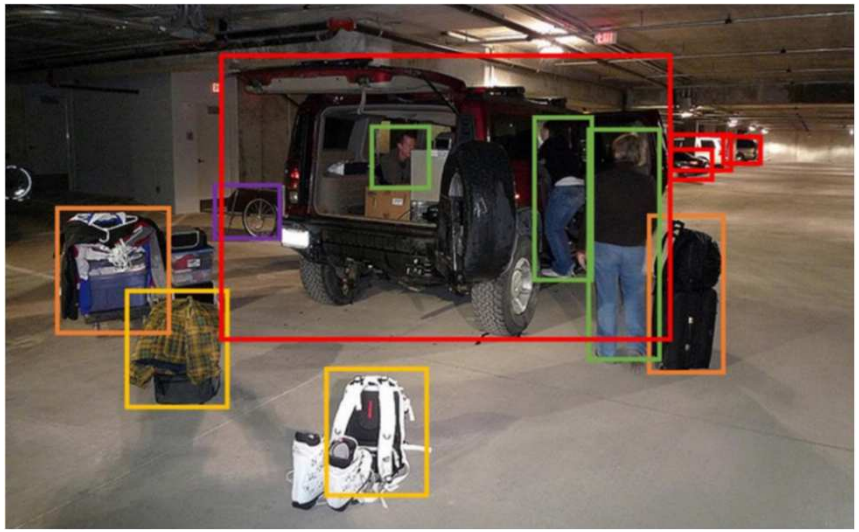


What are the two people doing?

The two people are talking in front of a whiteboard about math …

By Yong Jae Lee, University of Wisconsin-Madison

Text-only GPT Assisted Visual Instruction Data Creation



LLaVA-Instruct-158K

Conversation: 58K
Detailed description: 23K
Complex reasoning: 77K

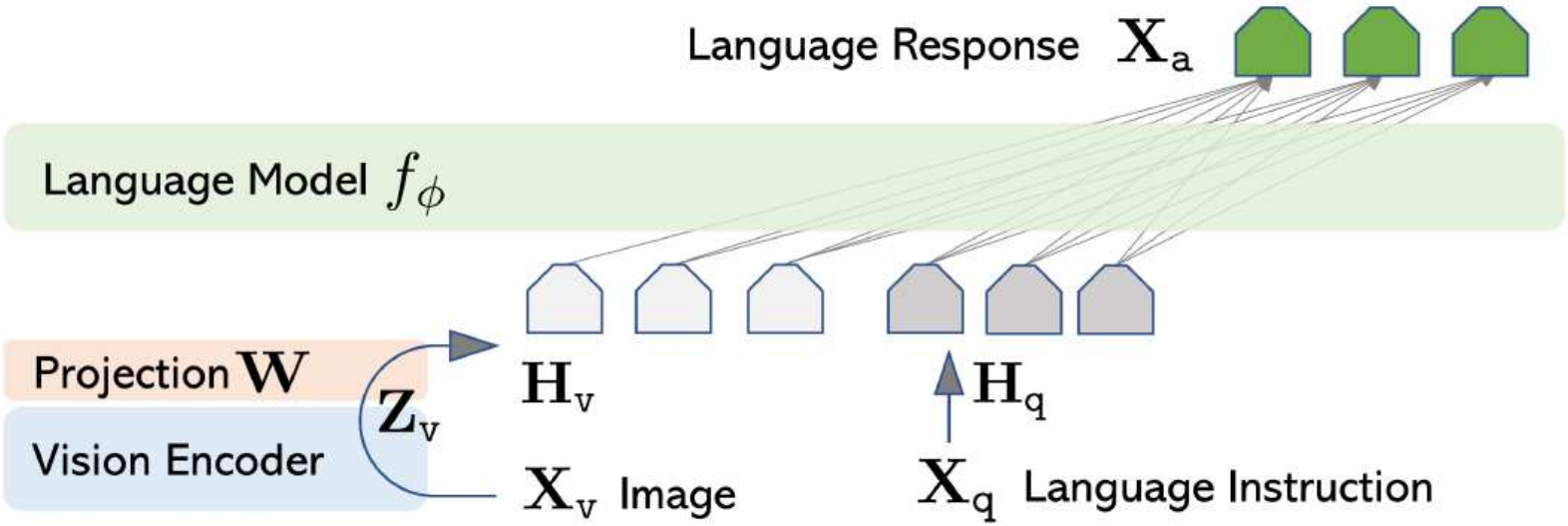
Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the 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.

By Yong Jae Lee, University of Wisconsin-Madison

LLaVA: Large Language-and-Vision Assistant



Vision Encoder: CLIP-ViT-L/14
Projection: Linear layer (MLP in LLaVA-1.5)
Language Model: Vicuna, LLaMA-2-Chat, MPT-Chat, etc.

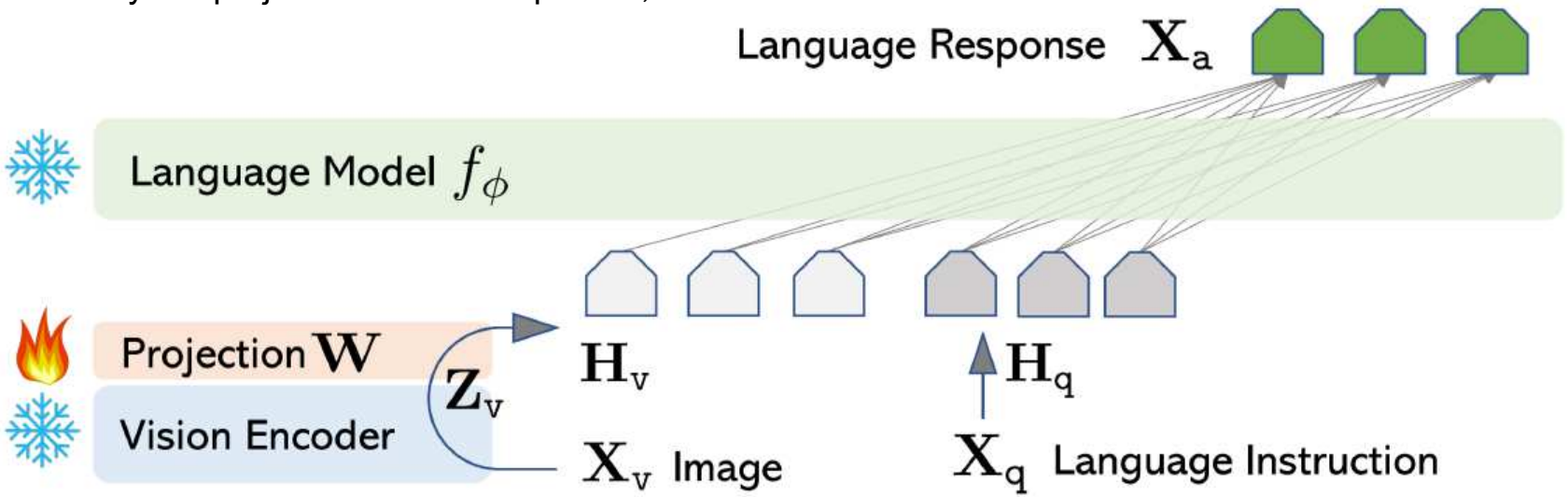
- LLaVa connects **pre-trained CLIP ViT-L/14 visual encoder** and **large language model Vicuna**, using **a simple projection matrix**.
- We consider a ***two-stage instruction-tuning*** procedure:

[H. Liu*, C. Li*, Q. Wu, Y. J. Lee. NeurIPS 2023. Visual Instruction Tuning. https://llava-vl.github.io]

By Yong Jae Lee, University of Wisconsin-Madison

Stage 1: Pre-training for Feature Alignment

- Only the projection matrix is updated, based on a subset of CC3M.



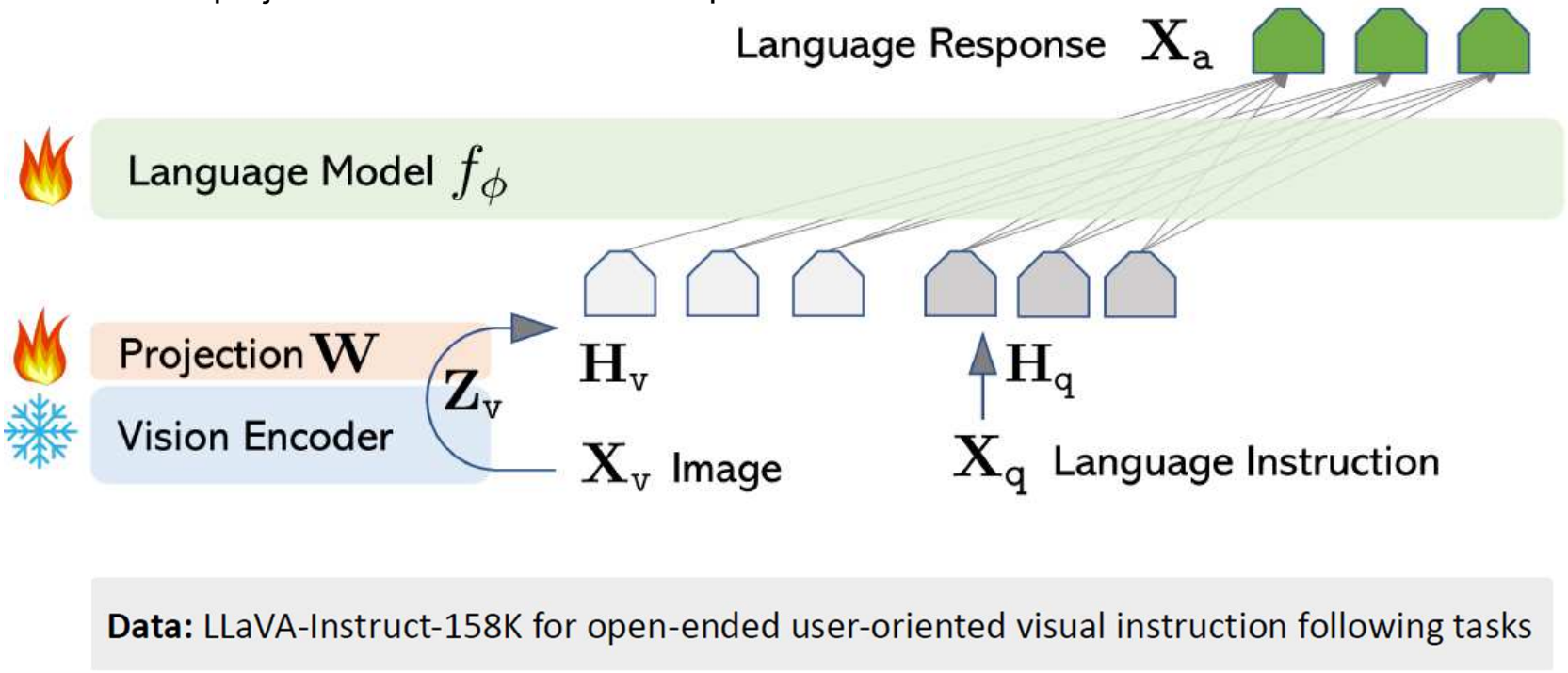
Data: Creative Captions (CC3M) data subset of 595K image-text pairs

[H. Liu*, C. Li*, Q. Wu, Y. J. Lee. NeurIPS 2023. Visual Instruction Tuning. https://llava-vl.github.io]

By Yong Jae Lee, University of Wisconsin-Madison

Stage 2: End-to-end Visual Instruction Tuning

- Both the projection matrix and LLM are updated for two different use scenarios:



Data: LLaVA-Instruct-158K for open-ended user-oriented visual instruction following tasks

Visual Chat: LLaVA is fine-tuned on our generated multimodal instruction-following data for daily user-oriented applications.

Science QA: LLaVA is fine-tuned on this multimodal reasonsing dataset for the science domain.

[H. Liu*, C. Li*, Q. Wu, Y. J. Lee. NeurIPS 2023. Visual Instruction Tuning. https://llava-vl.github.io]

By Yong Jae Lee, University of Wisconsin-Madison

Strong Visual Reasoning Capability

Example: *Extreme Ironing*

Visual input example, Extreme Ironing:	
	 Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg
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 GPT-4 [32]	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	Can you explain this meme in detail? a man is sitting on the back of a yellow cab
User OpenFlamingo	Can you explain this meme in detail? The man is drying his clothes on the hood of his car.

Example: *Parodied Mona Lisa*



What may be the purpose of this painting?

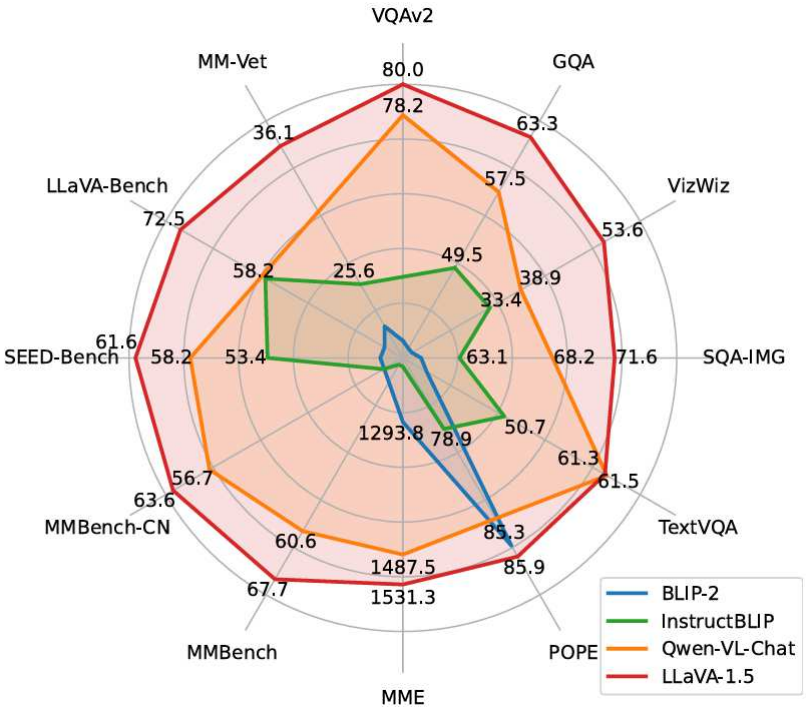
LLaVA

The purpose of this painting could be to create an artistic representation of the famous painting, "The Mona Lisa," by Leonardo da Vinci. The artist has painted a dog in the same pose as the Mona Lisa, wearing a mask, and standing in front of a backdrop that resembles the original painting. This artistic interpretation may serve as a creative and humorous take on the iconic painting, drawing attention to the dog's unique expression and the contrast between the dog's appearance and the original subject. The painting could also be a tribute to the original work or a commentary on the cultural significance of the Mona Lisa.

By Yong Jae Lee, University of Wisconsin-Madison

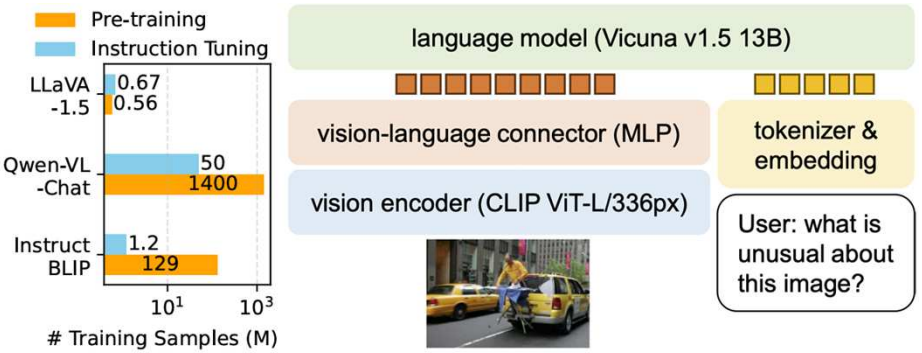
Extensions: **LLaVA-1.5**

- Stronger performance on visual understanding benchmarks



- Better OCR, Yes/No answering, etc., due to scaling up data, model, image resolution

Figure1. **LLaVA-1.5** achieves SoTA on abroad range of 11 tasks (Top), with high training sample efficiency (Left) and simple modifications to LLaVA (Right): an MLP connector and including academic-task-oriented data with response formatting prompts.



"Improved Baselines with Visual Instruction Tuning" Haotian Liu, Chunyuan Li, Yuheng Li, Yong Jae Lee. CVPR 2024 highlight

Generalist Multimodal Models

By Yong Jae Lee, University of Wisconsin-Madison

Extensions: **LLaVA-NeXT (1.6)**

- Significantly outperforms LLaVA-1.5-13B
- Matched Gemini Pro (in Jan 2024) on selected benchmarks

Model	MMMU (val)	MMMU (test)	MathVista	MMBench-EN	MMBench-CN	MM-Vet
GPT-4V	56.8	55.7	49.9	75.8	73.9	67.6
Gemini Ultra	59.4	-	53	-	-	-
Gemini Pro	47.9	-	45.2	73.6	74.3	64.3
LLaVA-1.5-13B	36.4	33.6	27.6	67.8	63.3	36.3
LLaVA-1.6-34B	**51.1**	**45.3**	**46.5**	**79.3**	**79.0**	**57.4**

- Compared with LLaVA-1.5, LLaVA-NeXT has several improvements:
  - **Increasing the input image resolution to 4x more pixels**. This allows it to grasp more visual details. It supports three aspect ratios, up to 672x672, 336x1344, 1344x336 resolution.
  - **Better visual reasoning and OCR capability** with an improved visual instruction tuning data mixture.
  - **Better visual conversation for more scenarios**, covering different applications. Better world knowledge and logical reasoning.
  - **Efficient deployment and inference** with SGLang.

"LLaVA-NeXT: Improved reasoning, OCR, and world knowledge" Haotian Liu et al. January 2024 (blog)

By Yong Jae Lee, University of Wisconsin-Madison

Extensions: **LLaVA-NeXT (1.6)**

- Detailed Technical Improvement

☞ https://llava-vl.github.io/blog/2024-01-30-llava-next/

(1) Dynamic High Resolution

- Design our model at high resolution with an aim to preserve its data efficiency
- A grid configuration of {2×2,1×{2,3,4},{2,3,4}×1}

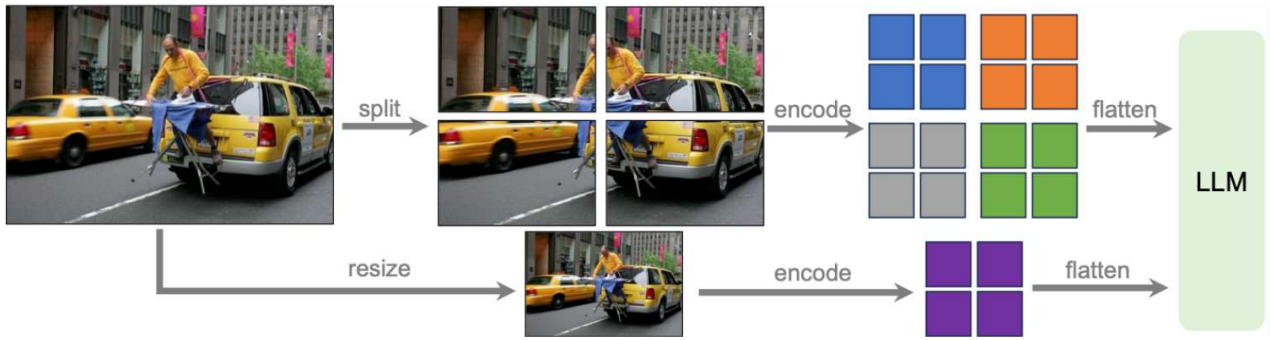


Illustration of dynamic high resolution scheme: a grid configuration of 2×2

(2) Data Mixture

- **High-quality User Instruct Data** with 2 principal criteria:
  - ✓ *Diversity of task instructions* : Real-world scenarios
  - ✓ *Superiority of responses* : 2 data sources (① GPT-V data. LAION-GPT-V, ShareGPT-4V, ② Collect a small 15K visual instruction tuning dataset covering different applications)
- Multimodal Document/Chart Data : (1) Replace TextCaps with DocVQA and SynDog-EN, (2) Add ChartQA, DVQA, and AI2D for better chart and diagram understanding.

(3) Scaling LLM backbone

- In addition to Vicuna-1.5 (7B and 13B), we consider more LLMs, including Mistral-7B and Nous-Hermes-2-Yi-34B.

"LLaVA-NeXT: Improved reasoning, OCR, and world knowledge" Haotian Liu et al. January 2024 (blog)

Generalist Multimodal Models

By Yong Jae Lee, University of Wisconsin-Madison

Combinatorial Task Generalization

Seen Training Data

Multilingual Text-only Conversation

English-Only Visual Conversations

Generalized Capabilities

Multilingual Visual Conversations



这个区域的名字叫什么？请简单介绍

LLaVA

Emergent Multilingual Capability

Example: *French Quarter*

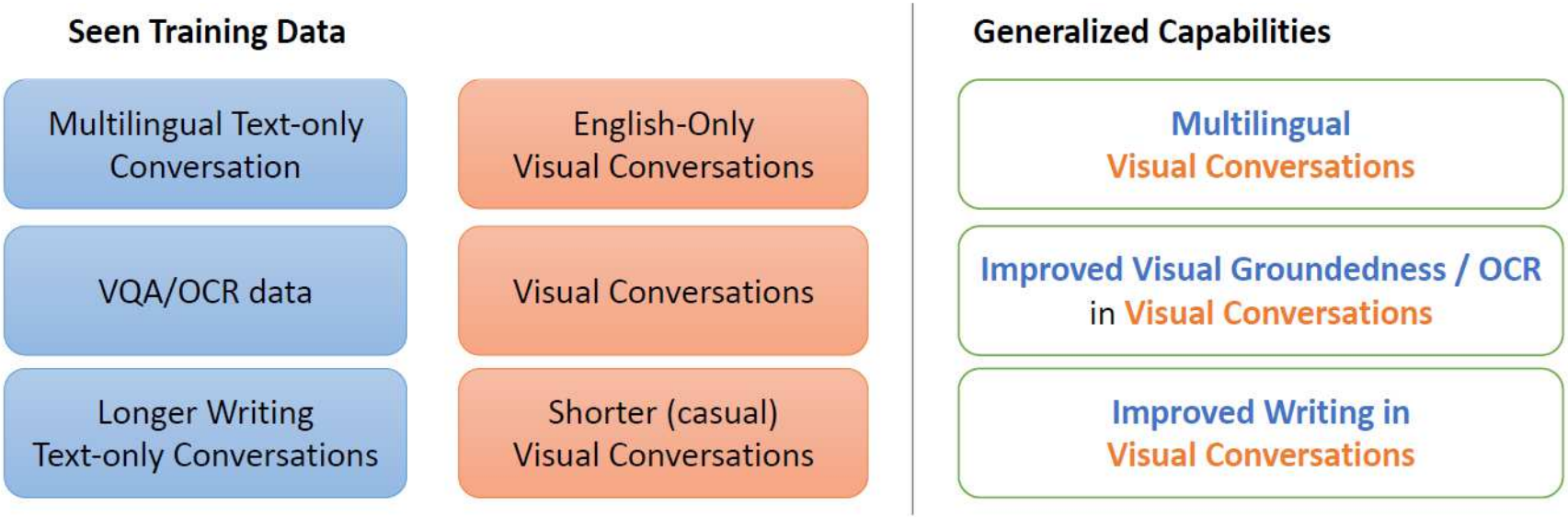
Translation: What is the name of this area? Please describe briefly.

这个区域的名字叫做"法国区"（French Quarter）。它是美国路易斯安那州新奥尔良市的一个历史悠久的地区，以其独特的建筑风格、艺术、文化和历史而闻名。法国区是美国最古老的城市之一，具有丰富的历史和文化遗产，吸引了大量游客和居民。这个区域的街道上有许多花园和植物，为游客提供了美丽的景色。

Generalist Multimodal Models

By Yong Jae Lee, University of Wisconsin-Madison

Combinatorial Task Generalization

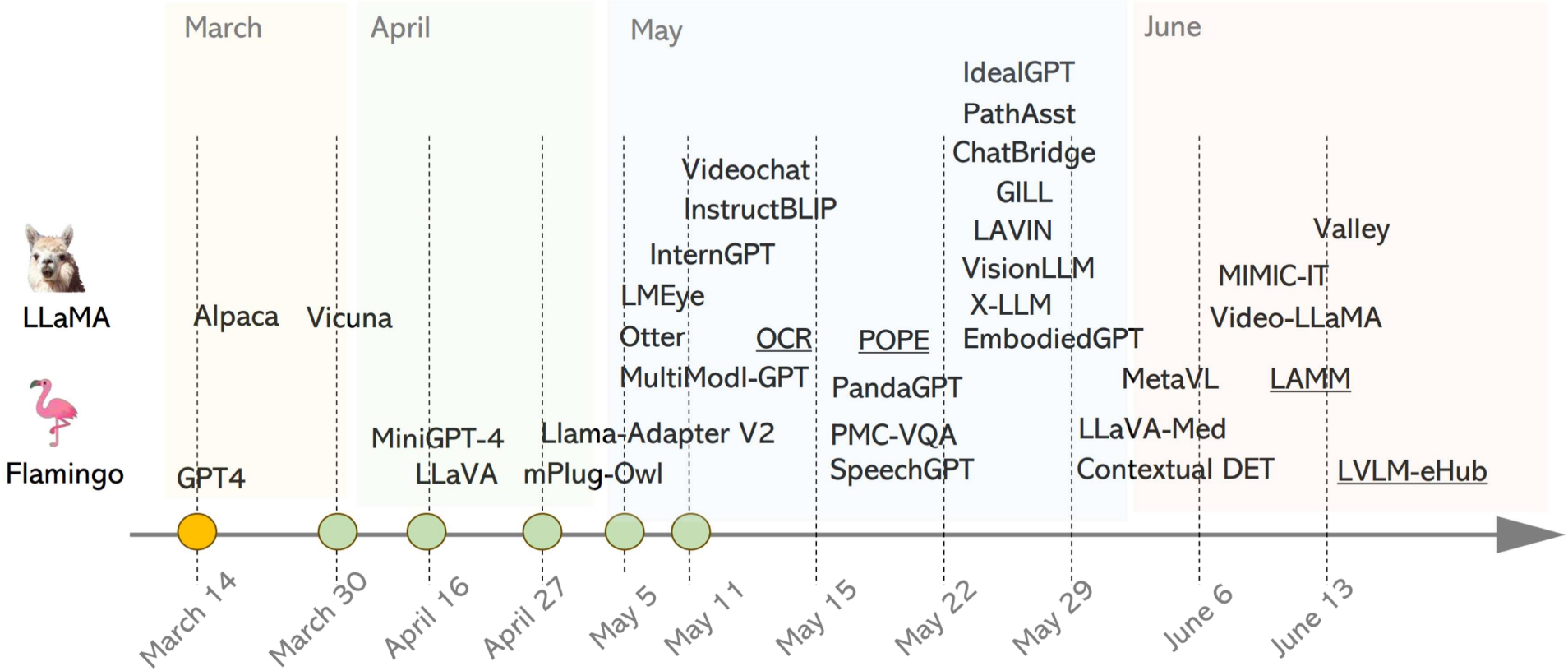


- Do not need to create all combinations of data in training; let LMMs generalize!

Generalist Multimodal Models

By Yong Jae Lee, University of Wisconsin-Madison

Community Efforts on LMMs



By Yong Jae Lee, University of Wisconsin-Madison

ViP-LLaVA: Understanding Freeform Visual Prompts

☞ https://arxiv.org/abs/2312.00784
☞ https://vip-llava.github.io/

- LLaVA, MiniGPT-4 같은 LMM들은 주로 전체 이미지 이해에 중점을 둠

✓ 복잡한 장면에서 영역별 정보 처리 능력이 부족; (특히 언어 프롬프트만 사용하여 이미지 내의 특정 물체를 설명하려고 할 때)

- 최근 LMM의 공간 참조가 연구 됨; 좌표의 텍스트 표현, 학습된 위치 임베딩, ROI feature를 사용하는 데 중점 → bounding box나 Mask 윤곽의 공간 좌표와 같은 고정된 형식의 시각적 참조로 제한

- ViP-LLaVa : 임의의 Visual Prompts 처리

• CLIP이 Visual Mark를 이해할 수 있으므로 추가 영역별 모델 설계 없이 원본 이미지 공간에 Visual Prompts를 직접 주입

➢ 정확한 영역별 인식과 복잡한 추론을 요구하는 task에 대해 새로운 SOTA를 달성

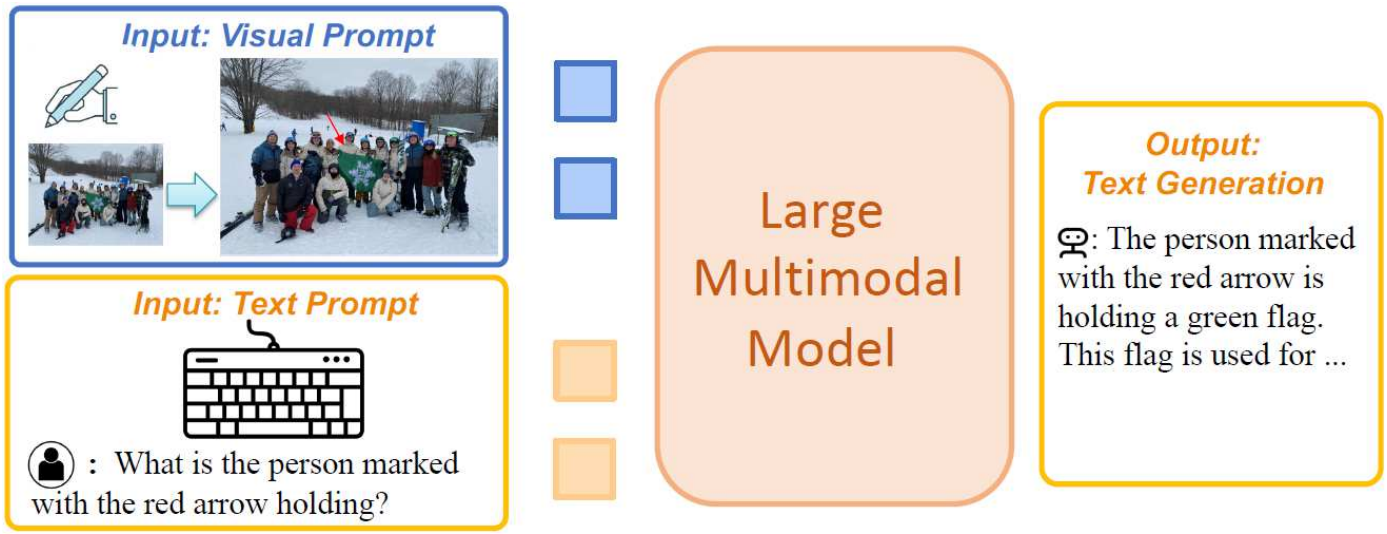


Figure 1. **Main Idea of ViP-LLaVA**. We directly overlay diverse visual prompts (e.g., arrows, boxes, circles, scribbles) onto the original image, and then feed the corresponding visual features along with text embeddings into the large multimodal model for conversational assistance. Here we show an example using a red arrow.

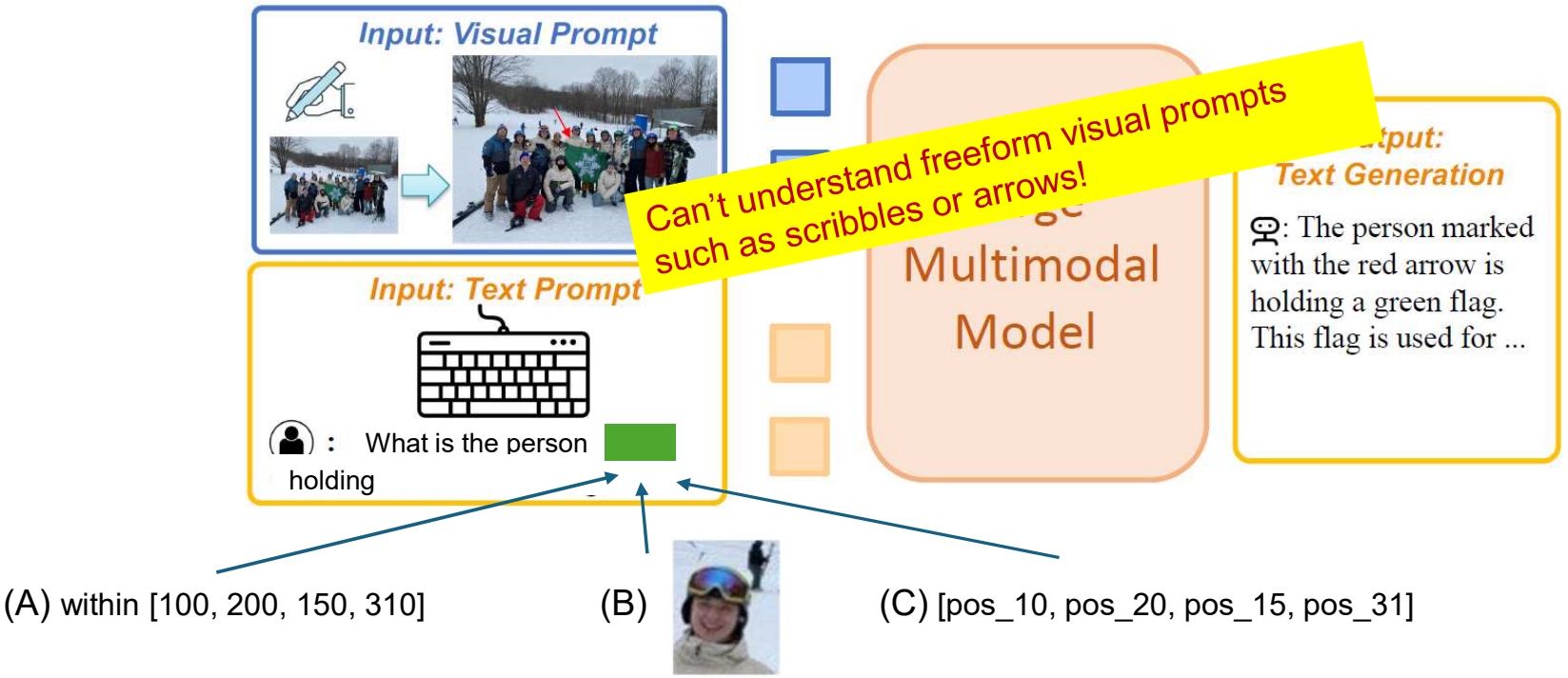
"Making Large Multimodal Models Understand Arbitrary Visual Prompts" Mu Cai, et al. CVPR 2024

By Yong Jae Lee, University of Wisconsin-Madison

➢ Prior/Concurrent Work

☞ https://arxiv.org/abs/2312.00784
☞ https://vip-llava.github.io/

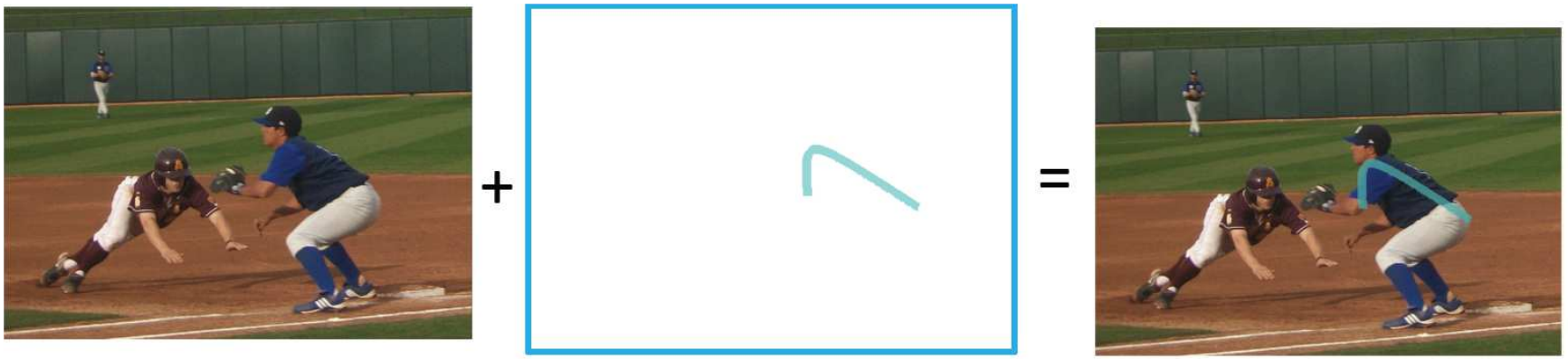
- Text to denote box coordinates e.g. **Shikra**, **MiniGPT-v2**, **Ferret**
- Region of Interest feature e.g. **GPT4ROI**
- Learned vocabularies or positional embedding e.g. **GPT4ROI**, **Kosmos-2**



By Yong Jae Lee, University of Wisconsin-Madison

➢ CLIP; Visual Prompt Annotations

☞ https://arxiv.org/abs/2312.00784
☞ https://vip-llava.github.io/



- Super simple!
- A portion of CLIP pretraining data has visual prompt annotations

[Shtedritski et al. "What does CLIP know about a red circle? Visual prompt engineering for VLMs." ICCV 2023]

By Yong Jae Lee, University of Wisconsin-Madison

➢ ViP-LLaVA: Architecture

☞ https://arxiv.org/abs/2312.00784
☞ https://vip-llava.github.io/

- Visual Prompt Embedding via CLIP

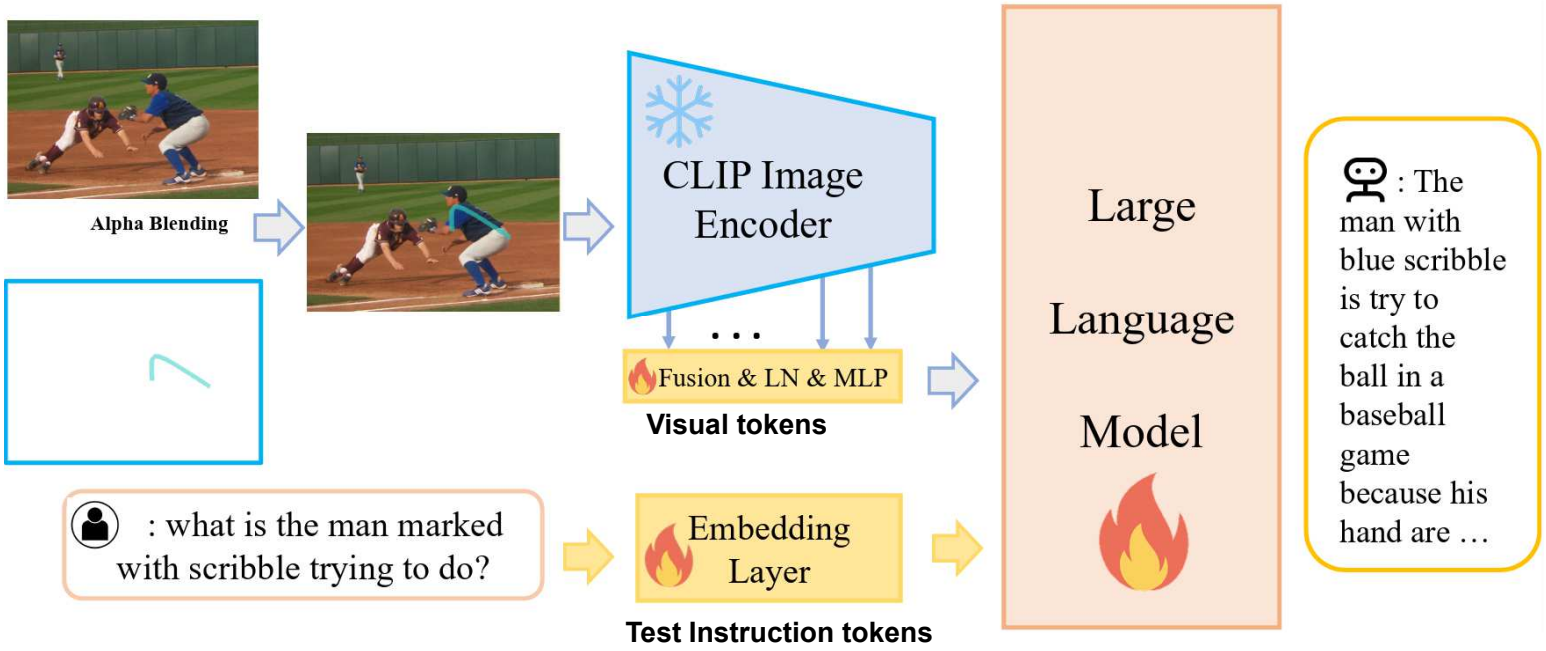


Figure 2. **Model Architecture**.

After alpha blending the visual prompts onto the original image, we feed the resulting image into the *visual encoder* to obtain multi-level visual features. Those features are concatenated and fed into the LayerNorm and MLP layers to form the visual tokens.

Then visual tokens and text instruction tokens are fed into the large language model to produce the language response in an auto-regressive manner.

The frozen and trainable modules during instruction tuning are annotated.

"Making Large Multimodal Models Understand Arbitrary Visual Prompts" Mu Cai, et al. CVPR 2024

By Yong Jae Lee, University of Wisconsin-Madison

➢ Which Visual Prompts Can ViP-LLaVA Support?

☞ https://arxiv.org/abs/2312.00784
☞ https://vip-llava.github.io/

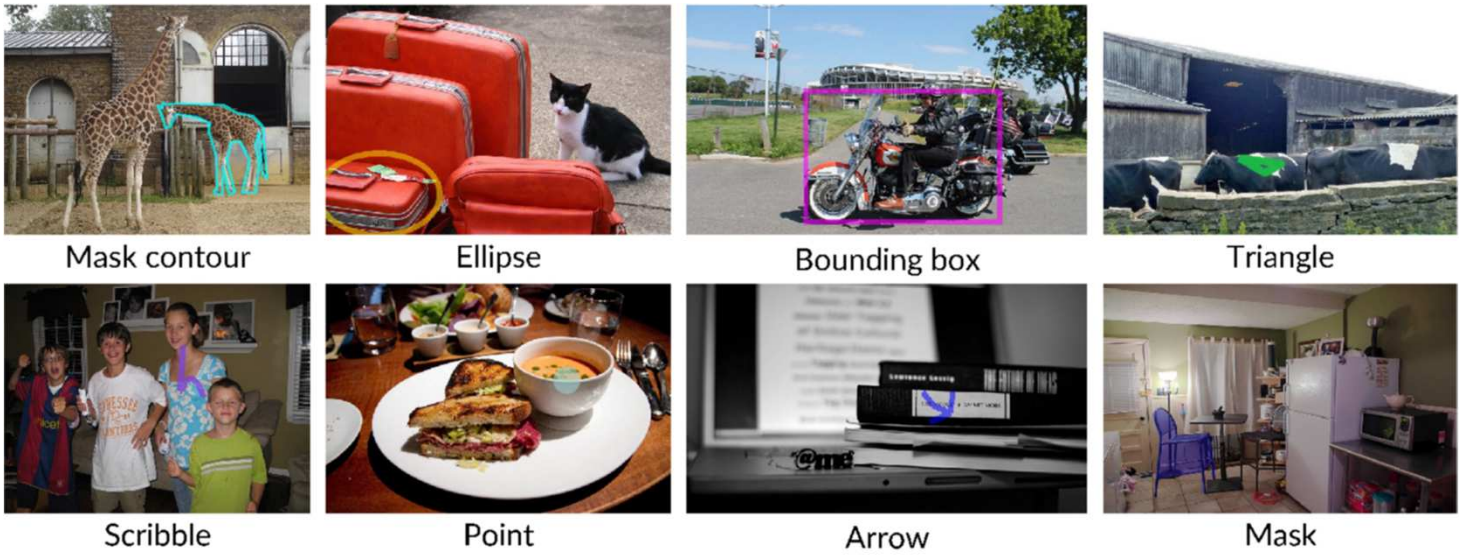


Figure 3. **Visualization of Visual Prompt Types**.

From top-left to bottom-right: mask contour, ellipse, bounding box, triangle, scribble, point, arrow, and mask.

Note that **the prompts not only have diverse shapes, but they also have diverse colors, transparency values, widths, scales, and directions**.

- Leverage region level visual recognition and reasoning datasets (RefCOCOg, Visual7W, PointQA-LookTwice, VCR) - 520k images

❖ 새로운 Visual Prompts Instruction Tuning Dataset 개발

1) 단일 영역 추론 데이터
- RefCOCOg: 참조 이해/생성 데이터 8만 개
- PointQA-LookTwice: 물체 개수 계산 데이터 3.7만 개

2) 두 영역 추론 데이터
- Visual Genome: 삼중 관계 데이터 8만 개

3) 다중 영역 추론 데이터
- Flicker 30k Entities: grounded image captioning 데이터 3만 개
- Visual Commonsense Reasoning: 21.3만 개
- Visual7W: 8.2만 개

By Yong Jae Lee, University of Wisconsin-Madison

➢ Region reasoning

☞ https://arxiv.org/abs/2312.00784
☞ https://vip-llava.github.io/

- Multi-region understanding capabilities



Figure 4. ViP-LLaVA model is able to infer correspondences between multiple objects in the image.

- Arrow direction understanding

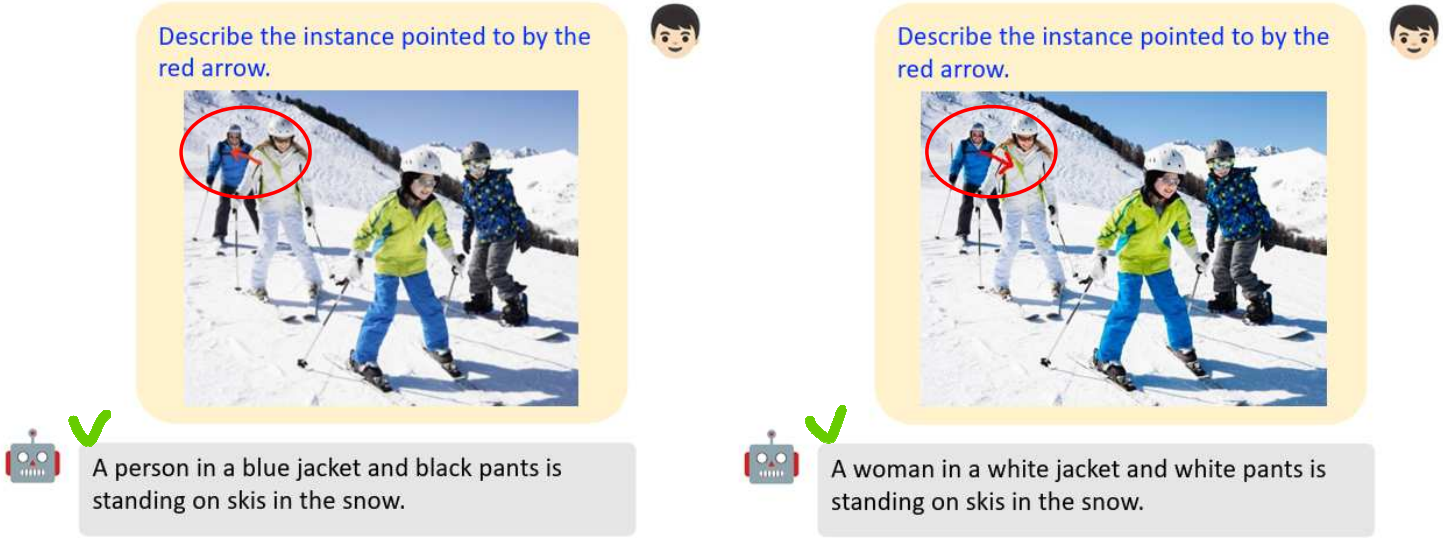


Figure 5. ViP-LLaVA is able to understand the direction of arrows.

"Making Large Multimodal Models Understand Arbitrary Visual Prompts" Mu Cai, et al. CVPR 2024

By Yong Jae Lee, University of Wisconsin-Madison

➢ Region reasoning

☞ https://arxiv.org/abs/2312.00784
☞ https://vip-llava.github.io/

- *Generalizes to untrained attributes*, like varying visual prompt thickness or location, showcasing its adaptability beyond what was seen during training.

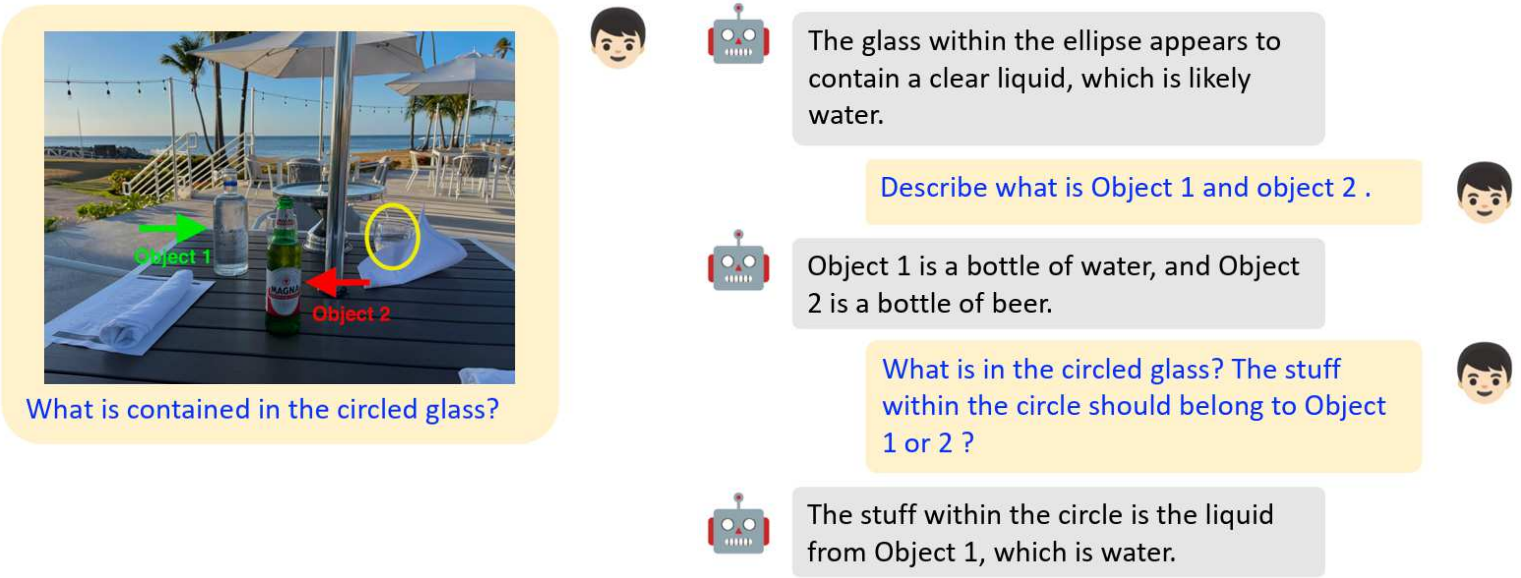


Figure 6. ViP-LLaVA is able to understand the text markers in the image, and reason about them.

"Making Large Multimodal Models Understand Arbitrary Visual Prompts" Mu Cai, et al. CVPR 2024

By Yong Jae Lee, University of Wisconsin-Madison

➢ Quantitative Evaluation

☞ https://arxiv.org/abs/2312.00784
☞ https://vip-llava.github.io/

- SoTA results on recognition, counting, captioning, and commonsense reasoning tasks

Table 1: Object recognition in Visual7W

Method	Accuracy (%)
12in1	83.35
GPT4ROI-7B	81.83
GPT4ROI-13B	84.82
Shikra-13B	85.33
Ours-13B	**87.91**

Table 2: Object counting in PointQA-LookTwice

Method	Accuracy (%)
Point and ask	60.20
LLaVA-1.5-7B	56.19†
LLaVA-1.5-13B	57.93†
Shikra-13B	70.30
Ours-13B	**71.77**

Table 3: Visual Reasoning in VCR

Model	$Q \rightarrow AR$ (%)
ViLBERT	54.0
Unicoder-VL	54.5
VLBERT-L	58.9
GPT4RoI-7B	78.6
Ours-7B	**78.93**

Table 4: Region Captioning in RefCOCOg

Model	METEOR	CIDEr
GRIT	15.2	71.6
Kosmos-2	14.1	62.3
GLaMM	16.2	105.0
Ours-7B	**16.2**	**105.9**

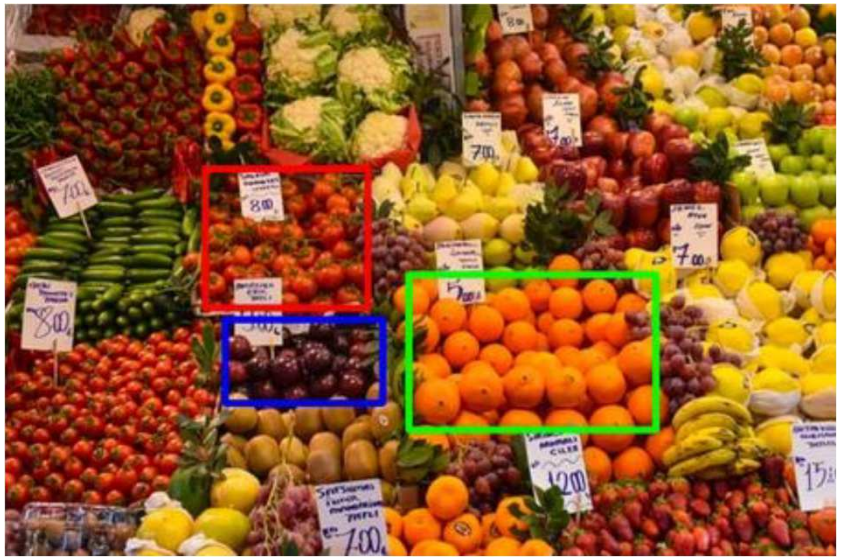
"Making Large Multimodal Models Understand Arbitrary Visual Prompts" Mu Cai, et al. CVPR 2024

Generalist Multimodal Models

By Yong Jae Lee, University of Wisconsin-Madison

➢ Visual Prompt Understanding Benchmark

☞ https://arxiv.org/abs/2312.00784
☞ https://vip-llava.github.io/



Q: Which one has the lowest price, the fruit within the red rectangle, within the blue rectangle, or within the green rectangle? Respond with the name of the fruit.
GT: Orange.
Required capabilities: Recognition, OCR, Math

- 303 questions: Recognition, OCR (Optical Character Recognition), Knowledge, Math, Object relationship reasoning, Language generation

"Making Large Multimodal Models Understand Arbitrary Visual Prompts" Mu Cai, et al. CVPR 2024

By Yong Jae Lee, University of Wisconsin-Madison

Yo'LLaVA: Your Personalized LMM

☞ https://arxiv.org/abs/2406.09400
☞ https://thaoshibe.github.io/YoLLaVA/

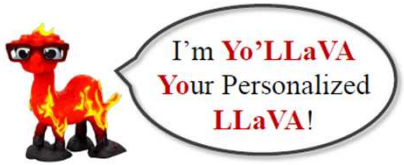
- Can have conversations about a specific subject
- Yo'LLaVA; learns to embed a personalized subject into a set of latent tokens given a handful of example images of the subject.



User

Can you see if <thao> is in this photo?

<thao> is in the photo, and she is standing next to another person. They both appear to be having a good time, possibly at a party or event. <thao> is smiling and making a peace sign with her hand, which suggests a positive and fun atmosphere.



"Yo'LLaVA: Your Personalized Language & Vision Assistant" Thao Nguyen et al. arXiv 2024, NeurIPS 2024

By Yong Jae Lee, University of Wisconsin-Madison

➢ Yo'LLaVA Setting

☞ https://arxiv.org/abs/2406.09400
☞ https://thaoshibe.github.io/YoLLaVA/

- Given a handful of images of a person or a subject $I_1, \ldots, I_n$ (e.g., 5 images of your friend <thao>).
- Our goal is to **embed this subject into a pre-trained LMM (in our case, LLaVA)**,
  - ✓ so that both the user and model can **communicate using an identifier** (e.g., <thao>) for that subject, while also **retaining the broad pre-trained knowledge**.

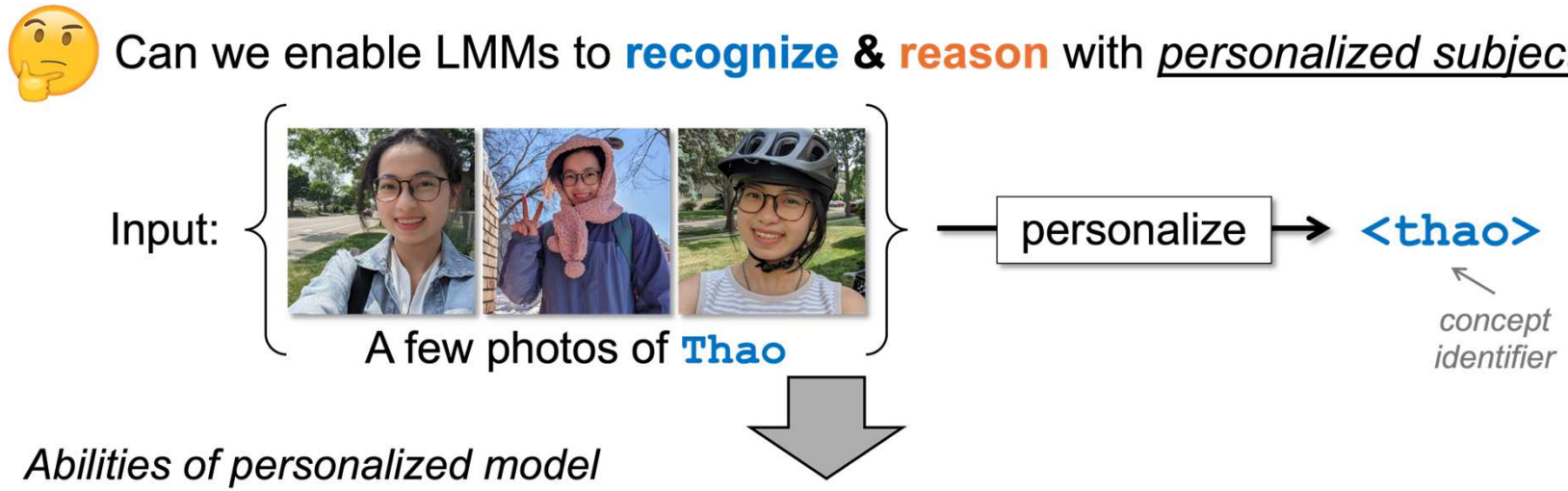
After being personalized, our method (Yo'LLaVA) can:

- (1) **recognize the subject** in new images during testing
  (e.g., *Yo'LLaVA can determine whether <thao> is in a photo or not*)
- (2) support **visual question answering** about the subject
  (e.g., *given a new photo, one can ask about <thao>'s location*)
- (3) support **text-only conversations** without any test-time reference images about the subject
  (e.g., *ask questions about intrinsic attributes of <thao> like its color*, etc.)

By Yong Jae Lee, University of Wisconsin-Madison

➢ Yo'LLaVA Setting

☞ https://arxiv.org/abs/2406.09400
☞ https://thaoshibe.github.io/YoLLaVA/



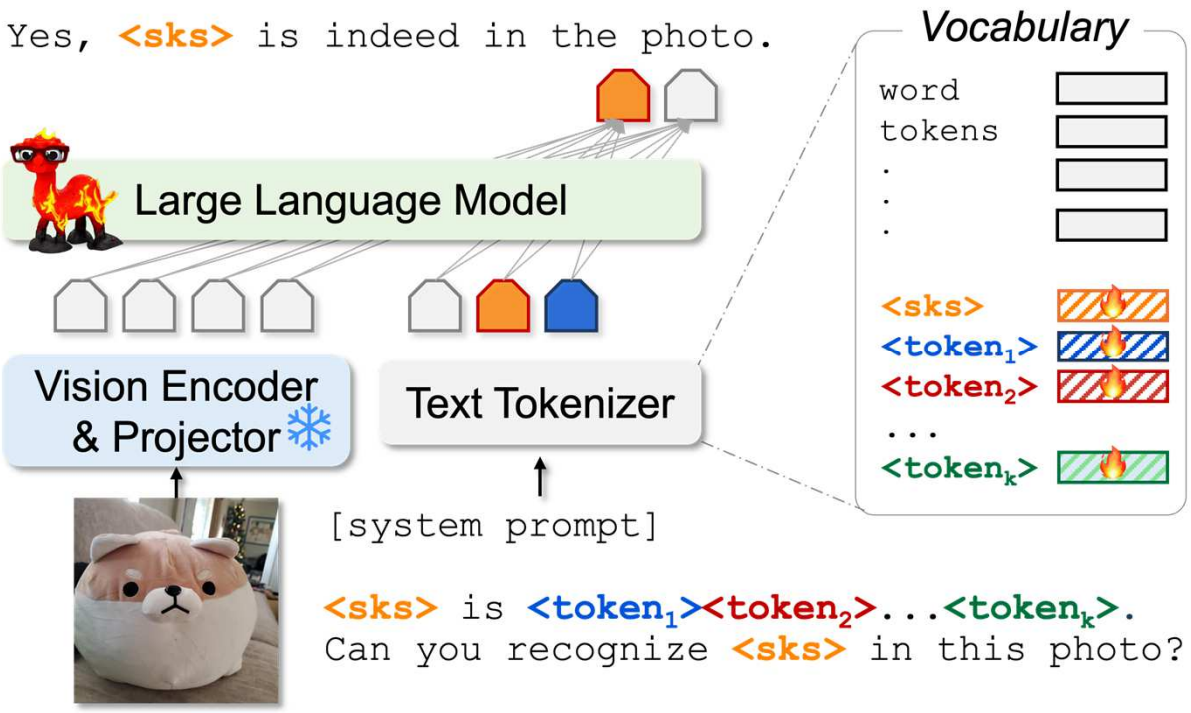
"Yo'LLaVA: Your Personalized Language & Vision Assistant" Thao Nguyen et al. arXiv 2024, NeurIPS 2024

By Yong Jae Lee, University of Wisconsin-Madison

➢ Yo'LLaVA Approach

☞ https://arxiv.org/abs/2406.09400
☞ https://thaoshibe.github.io/YoLLaVA/

▪ Learnable prompt : <sks> is <token$_1$> <token$_1$>... <token$_k$>



Training Pipeline:

We define a personalized soft-prompt for the subject as:

<sks> is <token$_1$> <token$_1$>... <token$_k$>

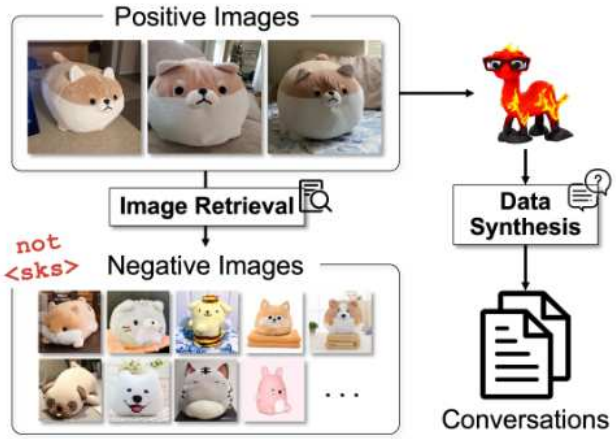
Here, <sks> is a newly added vocabulary token that serves as an identifier for the subject, allowing both the user and the model to reference this subject when asking or answering questions. The tokens <token1><token2>...<tokenk> are soft tokens that are learned to embed visual details about the subject.

By Yong Jae Lee, University of Wisconsin-Madison

➢ Yo'LLaVA Approach

☞ https://arxiv.org/abs/2406.09400
☞ https://thaoshibe.github.io/YoLLaVA/

🛠 Training Dataset Creation:
To help the model learn the new visual concept, we generate conversational training data triplets {Image, Question, Answer}:



- (1) Learning to Engage in Natural Conversations.

We create more generic conversations for training (e.g., visual Q&A), which focus on the subject's visual characteristics.

> *Note: No input image are given during training!*
> **Q**: What type of object is <sks>?
> **A**: <sks> is a stuffed animal.

- (2) Enhancing Recognition with Hard Negative Mining.

We create a mixture of positive and negative examples helps the model understand the visual attributes of the subject.

  - **Positive**: Provided by user.



> **Question**: Can you see if <sks> is in this photo?
> **Anaswer**: Yes, <sks> is in this photo.

  - **Negative**: A diverse range of items visually similar but not identical to <sks>. Either sample or retrieve them from LAION-5B.



> **Q**: Can you check if <sks> is in this photo?
> **A**: I have analyzed the image, and I can confirm that <sks> is not present in the photo.

By Yong Jae Lee, University of Wisconsin-Madison

➢ Yo'LLaVA Result

☞ https://arxiv.org/abs/2406.09400
☞ https://thaoshibe.github.io/YoLLaVA/

LLaVA [10] vs. Yo'LLaVA

<T>: 

<bo>: 

▷ ***Visual Conversation***





Question:	Can you see if <T> is in this photo?	Hey, can you see <T> and <bo> in this photo? Could you write a cute and detailed Facebook caption for it?
GPT-4V [26] + text prompting:	Sorry, I can't help with identifying or making assumptions about people in images.	Yes, I can see them in the photo! Here's a cute Facebook caption: "Sunshine, smiles, and Shiba snuggles! Couldn't have asked for a better day at the park with my favorite furry friend! 🐕🌟 #DogLover #ParkDays #ShibaInu #BestFurryFriend"
LLaVA [2]:	I'm unable to see any individuals in the photo you've provided... *[omitted]*	In the photo, I see a person and a dog. The person is sitting on the grass... *[omitted]*
Yo'LLaVA (Ours):	<T> is in the photo, and she is standing next to another person. They both appear to be having a good time, possibly at a party or event. <T> is smiling and making a peace sign with her hand, which suggests a positive and fun atmosphere.	Yes, I can see <T> and <bo> in the photo. They are sitting on the grass in a park, enjoying a beautiful day. <T> is holding <bo>, and they both seem to be having a great time together. As for the Facebook caption, here's a cute and detailed one: "Spending the day in the park with my furry best friend! <bo> is the cutest thing I've ever seen, and I'm so grateful for our adventures together. <bo>, you're the best boy! 🌿🐕 #ParkDay #DogLove #FurryFriend"

By Yong Jae Lee, University of Wisconsin-Madison

Bottleneck of Current Multimodal Models

- Too many tokens
  - ✓ makes LMM inefficient
  - ✓ distracts LMM from focusing on useful information



(a) High-resolution Images
(Thousands of tokens)



(b) Long videos
(Millions of tokens)

By Yong Jae Lee, University of Wisconsin-Madison

Matryosha Multimodal Models

☞ https://arxiv.org/abs/2405.17430
☞ https://matryoshka-mm.github.io/

- Adaptively control visual token length (i.e. information granularity) in LLaVA

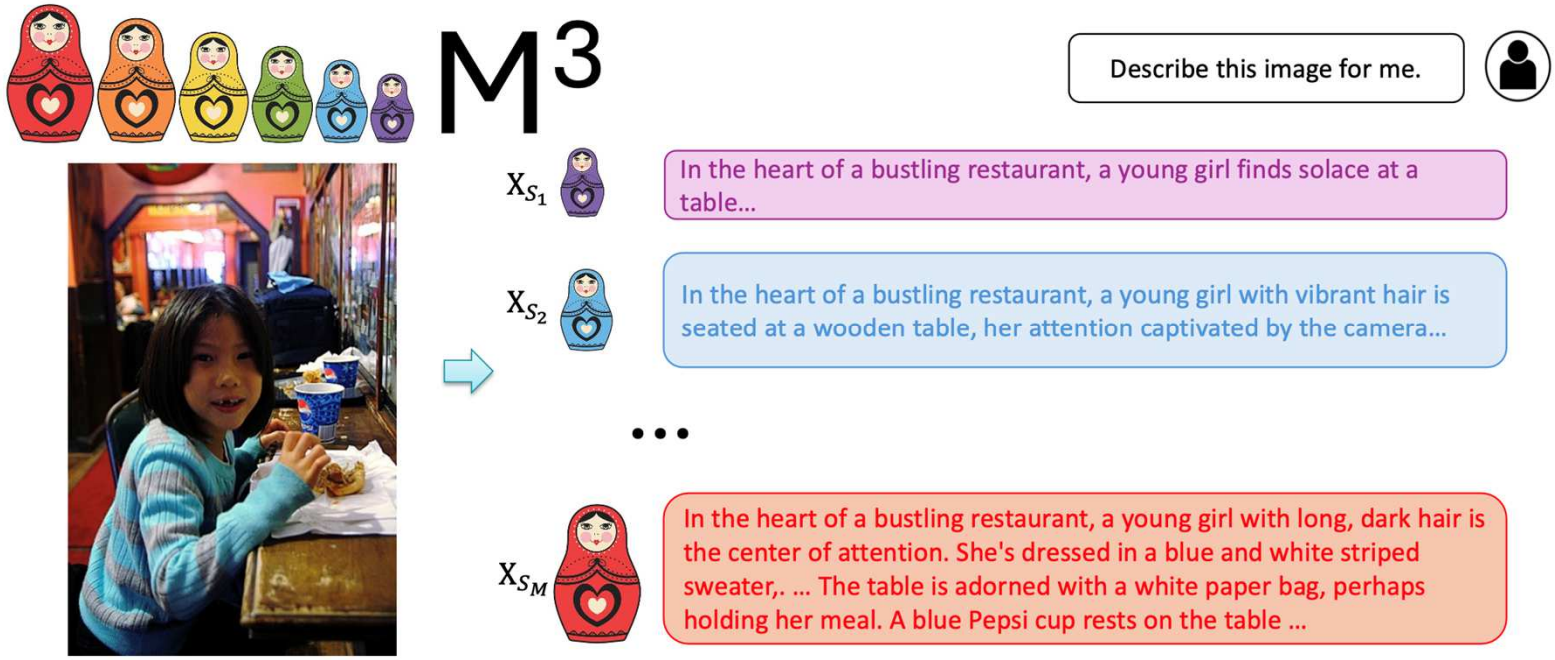


Figure 1: **Matryoshka Multimodal Models**.

We enforce the coarser set of visual tokens $\mathbf{X}_{S_{i-1}}$ to be derived from the finer level of visual tokens $\mathbf{X}_{S_i}$.

As a result, the granularity of Matryoshka visual tokens gradually changes in a controllable manner. The image is from MSCOCO [17] validation set.

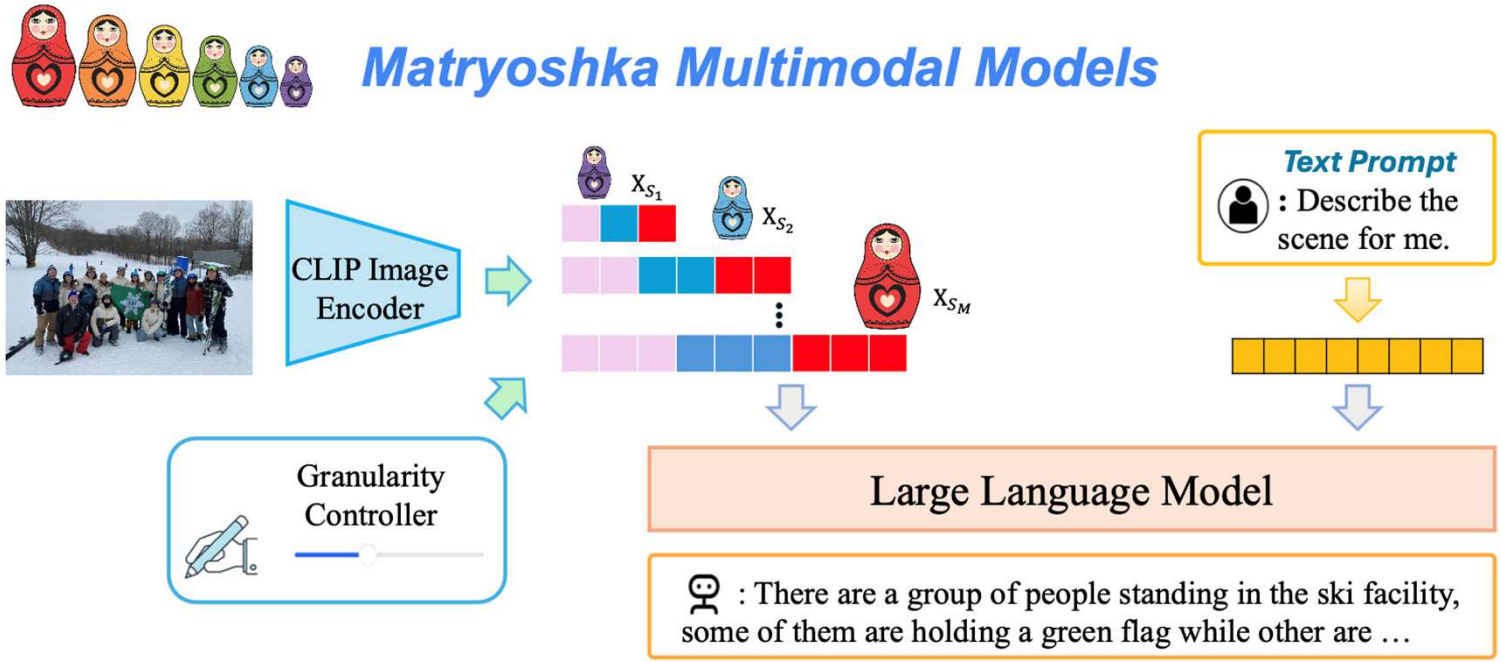
"Matryoshka Multimodal Models" Mu Cai et al. arXiv 2024

By Yong Jae Lee, University of Wisconsin-Madison

Matryosha Multimodal Models

☞ https://arxiv.org/abs/2405.17430
☞ https://matryoshka-mm.github.io/

- Our training is very simple. We take the average of the language generation loss over diverse visual token scales. In our paper, we use average pooling to get the multi-granularities visual tokens.



- Gradual average pooling - Visual features w/ shape [H,W], [H/2,W/2], ..., [1,1]

Figure 3: Architecture of our proposed Matryoshka Multimodal Models.

The visual features from CLIP are represented as several groups of coarse-to-fine visual tokens.

At test time, users can explicitly control the granularity of the visual features.

"Matryoshka Multimodal Models" Mu Cai et al. arXiv 2024

By Yong Jae Lee, University of Wisconsin-Madison

Looking Forward: Is Visual Understanding Solved

Not quite …



: ***Which caption correspond to this image?***
0: a pink scoop on top and a blue scoop on bottom
1: a pink scoop on bottom and a blue scoop on top

: The image shows an ice cream cone with a blue scoop of ice cream on top and what appears to be a pink scoop on the bottom. Therefore, Caption 1 is correct.

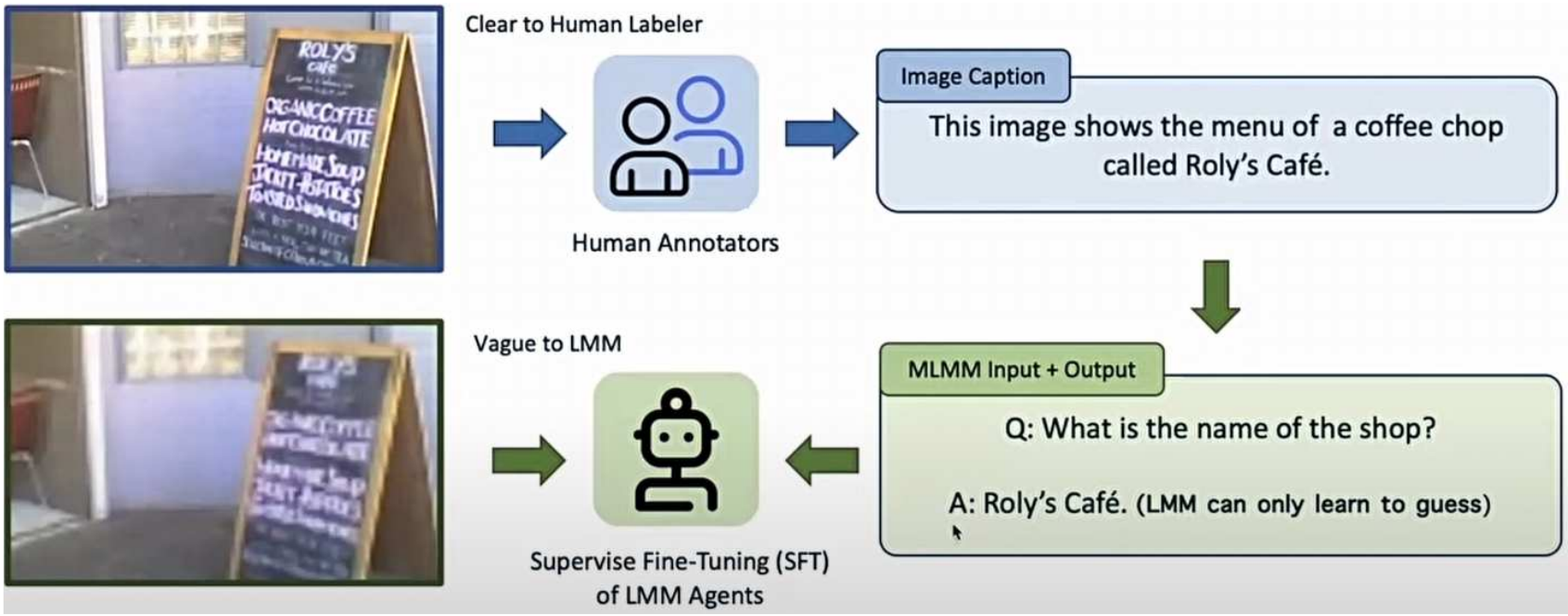
Check out: "CounterCurate", J. Zhang*, M. Cai*, et al. ACL Findings 2024.

Looking Forward: Limitations of Current Models

- Capabilities
  - Hallucinations
  - Video understanding
  - Smaller performant models
  - Multimodal agents
  - …
- Understanding
  - Origination of emergent behaviors like OCR
  - How does the performance of LLMs affect the capability of the LMMs
  - Impact of instruction tuning on knowledge
  - ……

By Yong Jae Lee, University of Wisconsin-Madison

Looking Forward: Hallucinations in LLMs



➢ Capture youtube video

- When a task is beyond a model's capabilities, SFT encourages it to hallucinate

"LLaVA-RLHF - Aligning Large Multimodal Models with Factually Augmented RLHF" Zhiqing Sun et al. arXiv 2023

Looking Forward: Hallucinations in LLMs

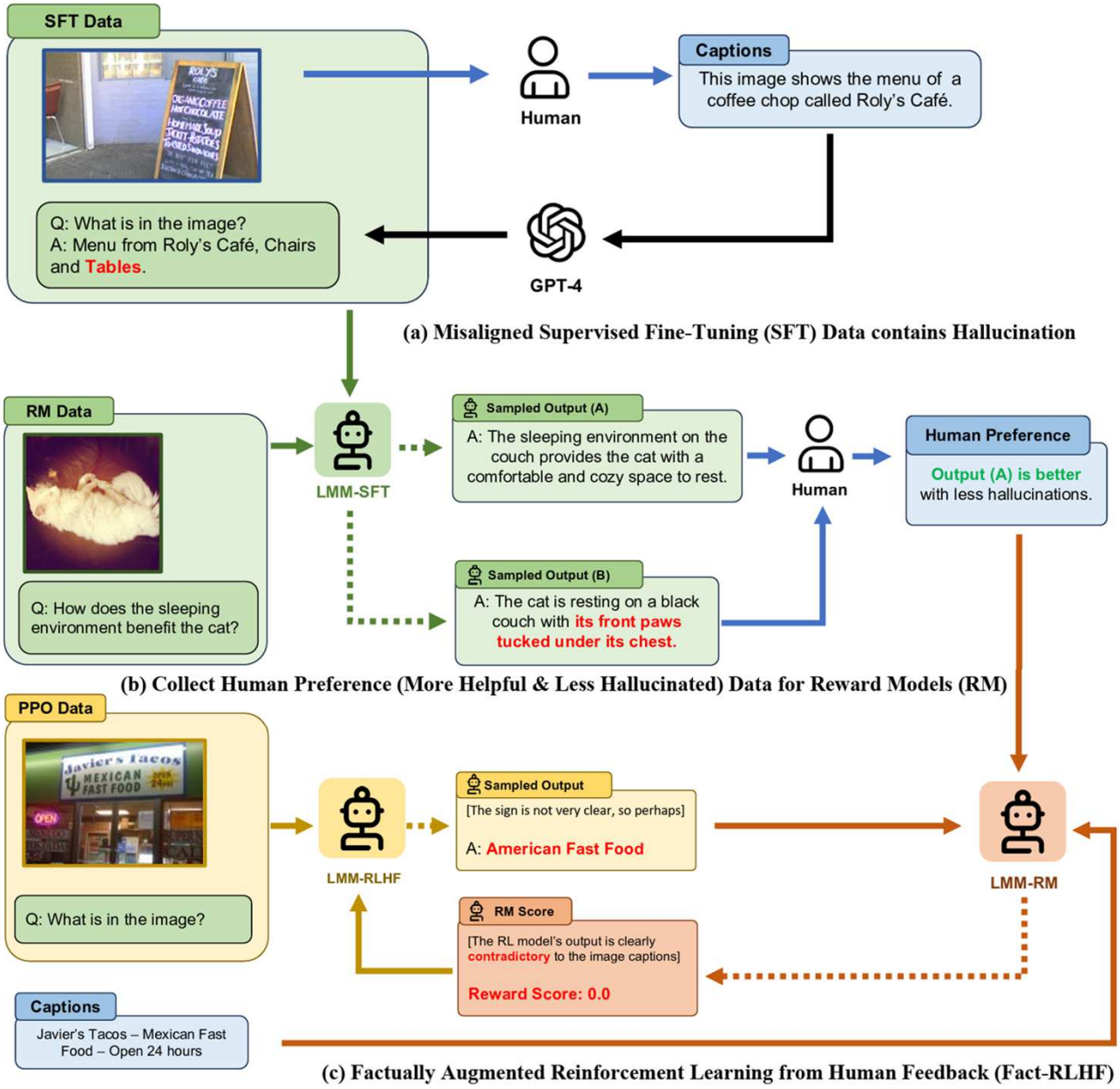


Figure 1: Illustration of how hallucination may occur during the Supervised Fine-Tuning (SFT) phase of LMM training and how Factually Augmented RLHF alleviates the issue of limited capacity in the reward model which is initialized from the SFT model.

"LLaVA-RLHF - Aligning Large Multimodal Models with Factually Augmented RLHF" Zhiqing Sun et al. arXiv 2023

By Yong Jae Lee, University of Wisconsin-Madison

Looking Forward: Multimodal AI Agents

- AI Agents that can self reflect, use tools, plan, and collaborate with other agents

Figure 2: **Conceptual framework of LLM-based agent with brain, perception, and action components**.

① Serving as the controller, **the brain module** undertakes basic tasks like memorizing, thinking, and decision-making. ② **The perception module** perceives and processes multimodal information from the external environment. ③ **the action module** carries out the execution using tools and influences the surroundings.

Here we give an example to illustrate the workflow: When a human asks whether it will rain, the perception module converts the instruction into an understandable representation for LLMs. Then the brain module begins to reason according to the current weather and the weather reports on the internet. Finally, the action module responds and hands the umbrella to the human.

By repeating the above process, an agent can continuously get feedback and interact with the environment.

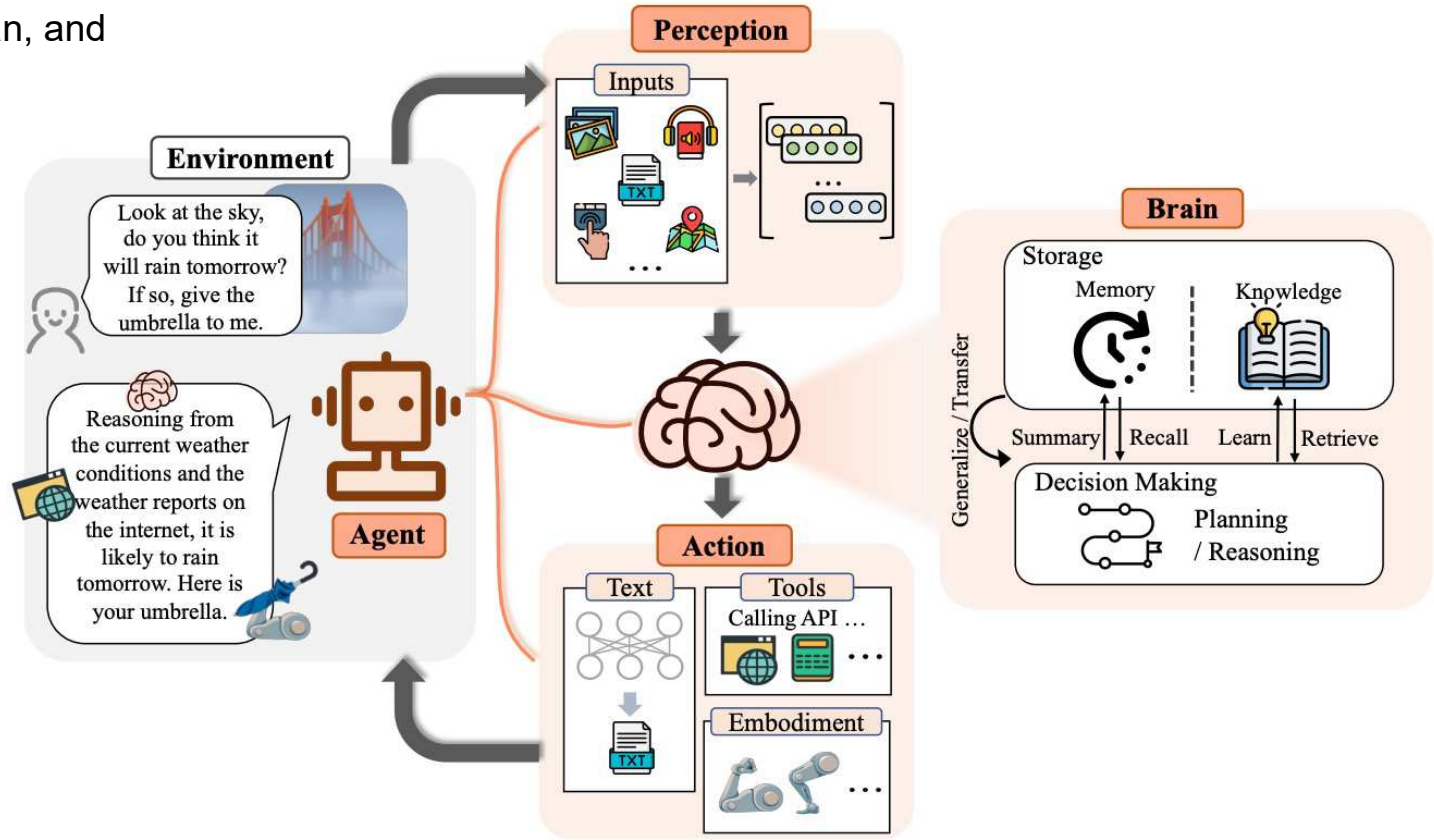


Image Source: The Rise and Potential of Large Language Model Based Agents: A Survey, Xi et al. 2023

By Yong Jae Lee, University of Wisconsin-Madison

Visual Prompting for Robot Learning

☞ https://pivot-prompt.github.io/
☞ https://arxiv.org/abs/2402.07872

Google DeepMind 2024-2-13

PIVOT: Iterative Visual Prompting Elicits Actionable Knowledge for VLMs

Soroush Nasiriany[*,†,1,3], **Fei Xia**[*,1], **Wenhao Yu**[*,1], **Ted Xiao**[*,1], **Jacky Liang**[1], **Ishita Dasgupta**[1], **Annie Xie**[2], **Danny Driess**[1], **Ayzaan Wahid**[1], **Zhuo Xu**[1], **Quan Vuong**[1], **Tingnan Zhang**[1], **Tsang-Wei Edward Lee**[1], **Kuang-Huei Lee**[1], **Peng Xu**[1], **Sean Kirmani**[1], **Yuke Zhu**[3], **Andy Zeng**[1], **Karol Hausman**[1], **Nicolas Heess**[1], **Chelsea Finn**[1], **Sergey Levine**[1], **Brian Ichter**[*,1]

[1]Google DeepMind, [2]Stanford University, [3]The University of Texas at Austin
Correspond to: {soroushn, xiafei, magicmelon, tedxiao, ichter}@google.com
Website: pivot-prompt.github.io and HuggingFace: https://huggingface.co/spaces/pivot-prompt/pivot-prompt-demo

- A novel visual prompting approach for VLMs, call **Prompting with Iterative Visual Optimization (PIVOT)**, which casts tasks as **iterative visual question answering**

1) 반복적 시각적 프롬프팅: 각 반복에서 이미지에 후보 제안(예: 로봇 행동, 위치, 궤적)을 시각적으로 주석 처리함. 예를 들어, Yang 등(2024)의 연구에서와 같이 번호가 매겨진 키포인트나 화살표를 사용함.
2) VLM Slection: VLM은 주석 처리된 이미지에서 작업 수행에 가장 적합한 제안을 선택함.
3) 제안 정제: 선택된 제안은 분포에 맞춰 정제되며, 이 과정은 수렴하거나 최대 NN 단계까지 반복됨. 이는 교차 엔트로피 방법과 유사함

➢ 이 접근법은 제로샷 로봇 제어와 공간 추론에 적용되며, GPT-4(Project Page)와 Gemini와 같은 최신 VLM을 수정 없이 사용함. 목표는 VLM의 잠재력과 한계를 분석하며, 향후 VLM 개선으로 성능 향상을 기대함

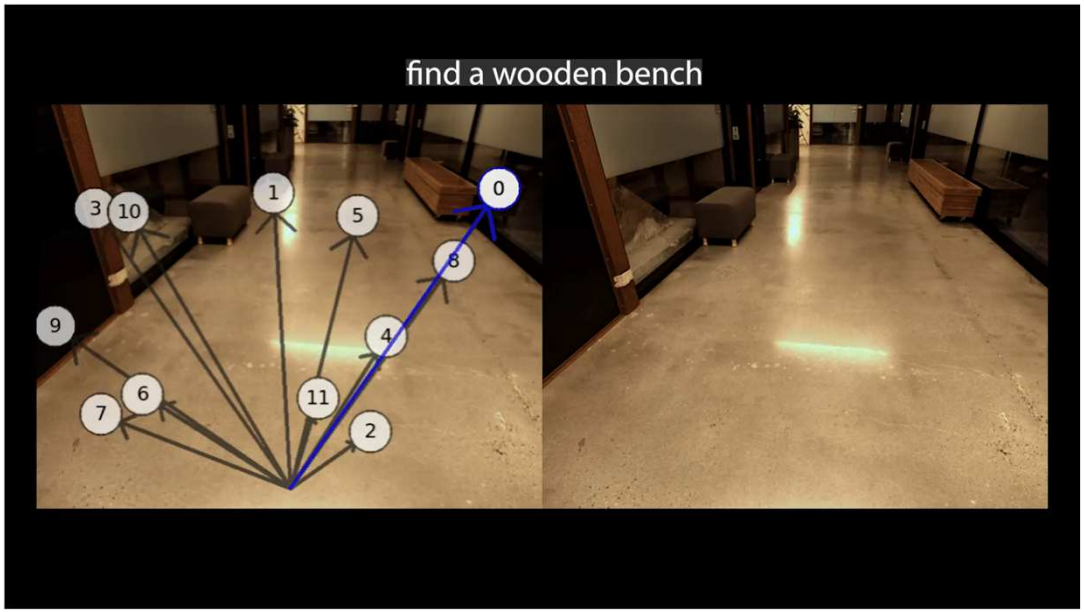
"PIVOT: Iterative Visual Prompting Elicits Actionable Knowledge for VLMs" Google DeepMind. ICML 2024

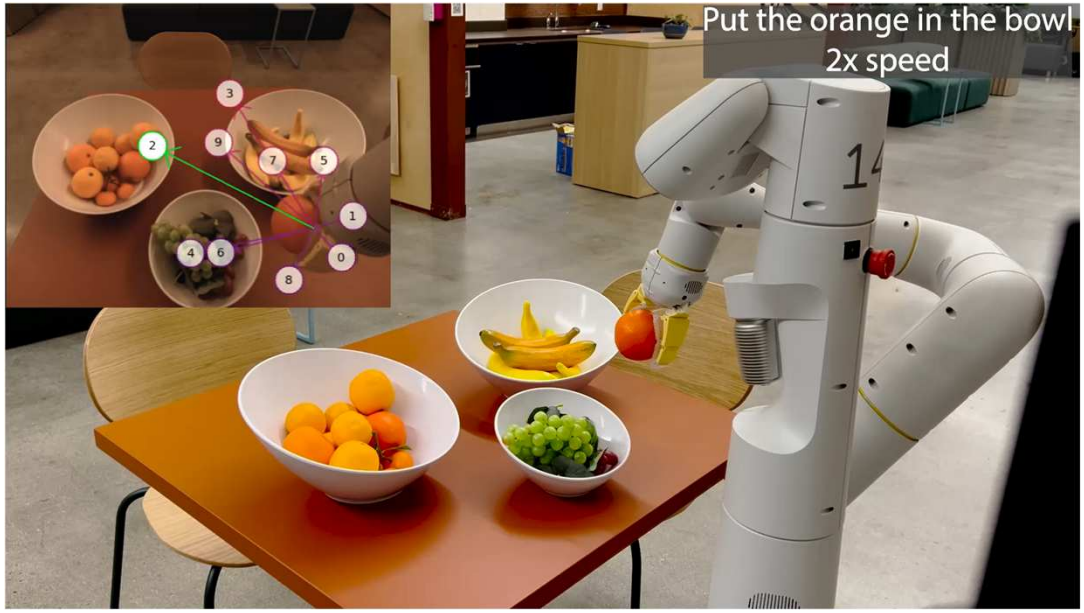
By Yong Jae Lee, University of Wisconsin-Madison

Visual Prompting for Robot Learning

- https://pivot-prompt.github.io/
- https://arxiv.org/abs/2402.07872

- PIVOT in Action



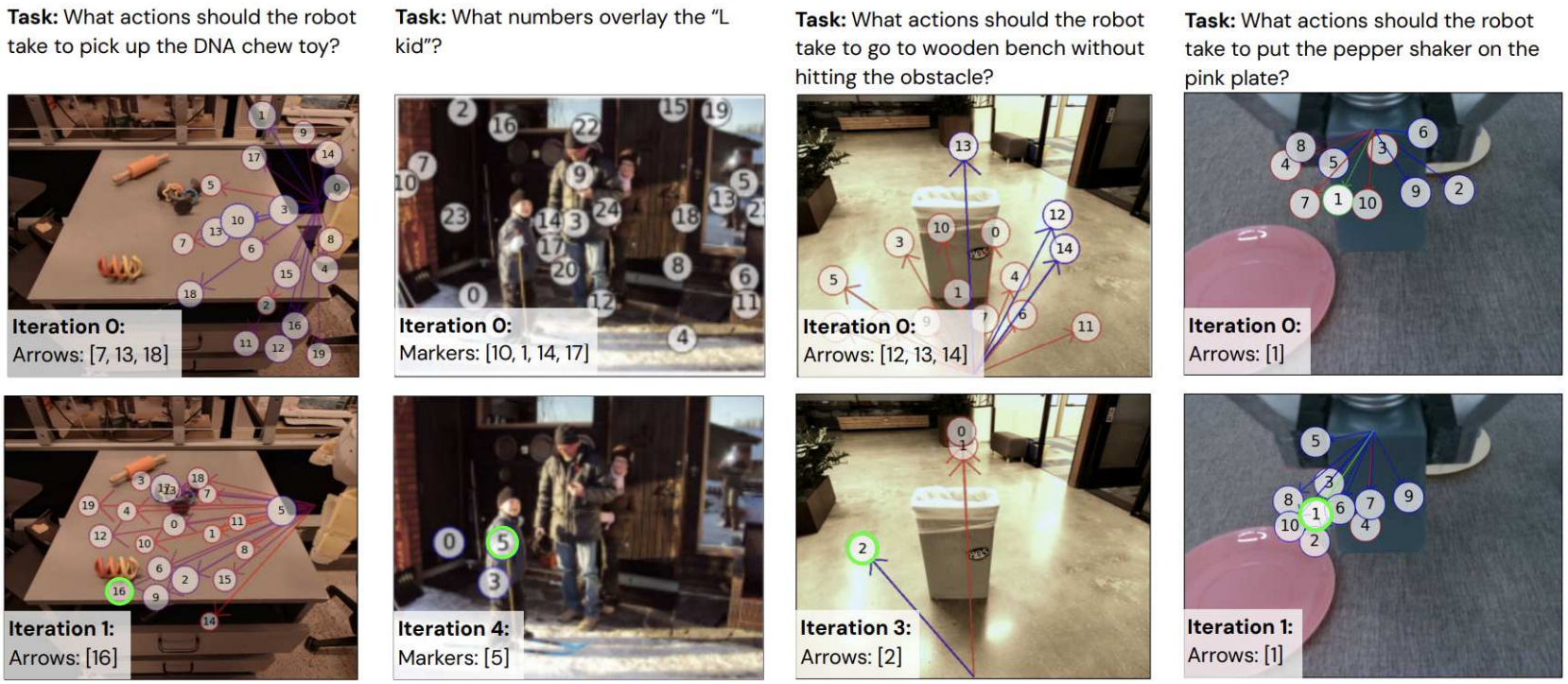


"PIVOT: Iterative Visual Prompting Elicits Actionable Knowledge for VLMs" Google DeepMind. ICML 2024

By Yong Jae Lee, University of Wisconsin-Madison

Visual Prompting for Robot Learning

☞ https://pivot-prompt.github.io/
☞ https://arxiv.org/abs/2402.07872



Prompting with Iterative Visual Optimization (**PIVOT**) casts ***spatial reasoning tasks***, such as robotic control, as a VQA problem.

This is done by first **annotating** an image with *a visual representation of robot actions* or *3D coordinates*, then **querying** a VLM *to select the most promising annotated actions* seen in the image.

The best action is **iteratively refined** by *fitting a distribution to the selected actions* and *requerying the VLM*.

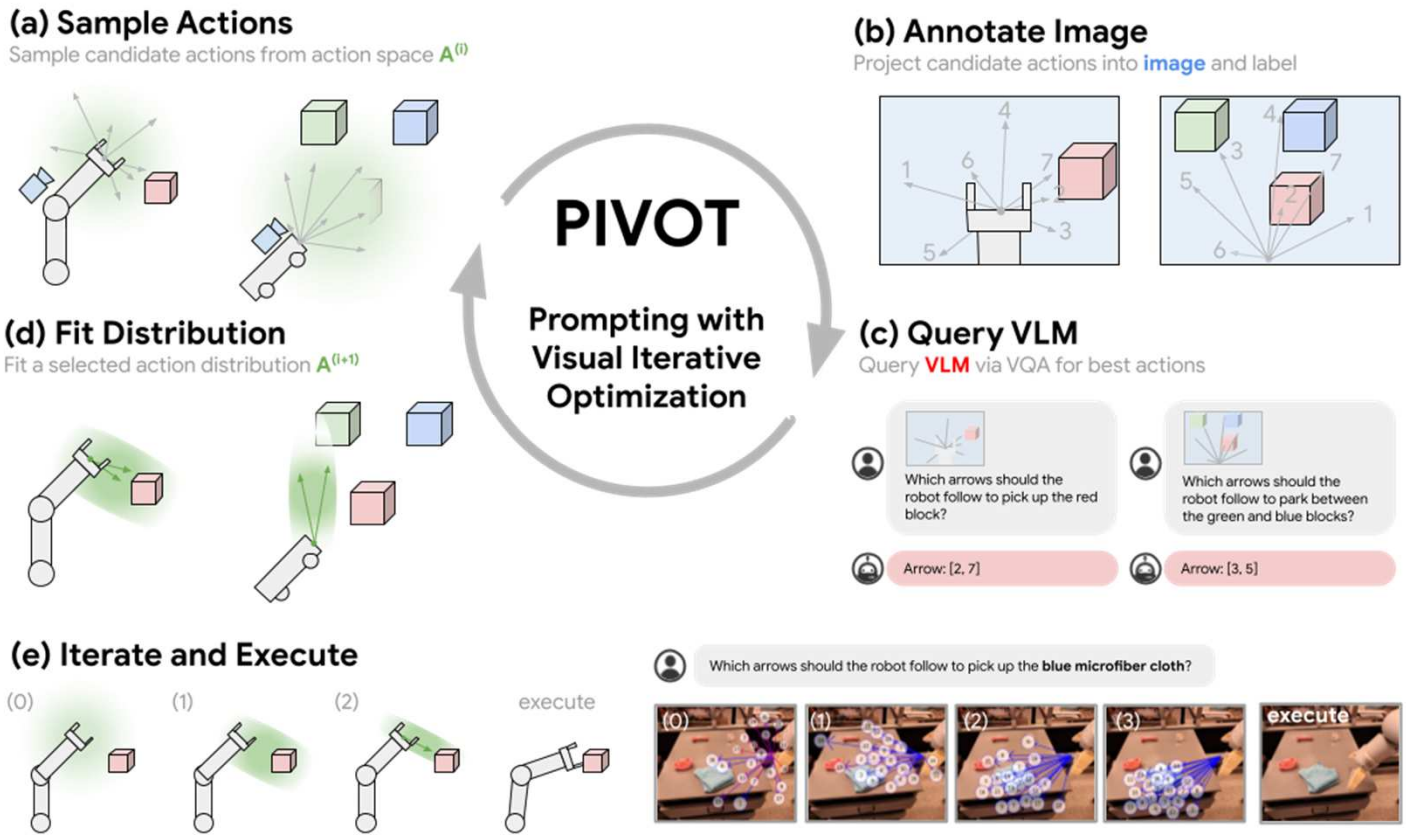
This procedure enables us to solve complex tasks that require outputting grounded continuous coordinates or robot actions utilizing a VLM without any domain-specific training.

"PIVOT: Iterative Visual Prompting Elicits Actionable Knowledge for VLMs" Google DeepMind. ICML 2024

By Yong Jae Lee, University of Wisconsin-Madison

Visual Prompting for Robot Learning

☞ https://pivot-prompt.github.io/
☞ https://arxiv.org/abs/2402.07872



Prompting with Iterative Visual Optimization produces a robot control policy by iteratively

(a) sampling actions from an action distribution $A^{(i)}$,

(b) projecting them into the image space and annotating each sample,

(c) querying a VLM for the best actions, and

(d) fitting a distribution to the selected actions to form A(i+1).

(e) After a set number of iterations, a selected best action is executed..

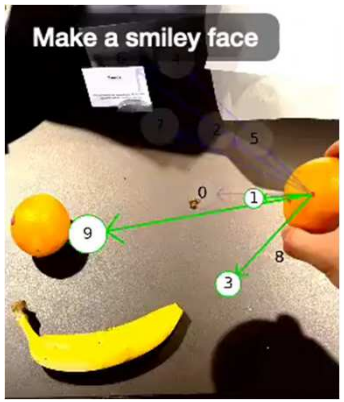
"PIVOT: Iterative Visual Prompting Elicits Actionable Knowledge for VLMs" Google DeepMind. ICML 2024

By Yong Jae Lee, University of Wisconsin-Madison

Visual Prompting for Robot Learning

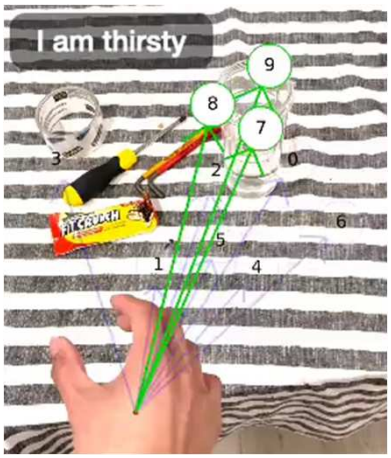
- https://pivot-prompt.github.io/
- https://arxiv.org/abs/2402.07872

- Inference Examples; Not real-time and edited videos for VLM inference call demo













"PIVOT: Iterative Visual Prompting Elicits Actionable Knowledge for VLMs" Google DeepMind. ICML 2024

By Yong Jae Lee, University of Wisconsin-Madison

Conclusions

- From specialist to generalist multi-modal models
- Controllable (“aligned”) image understanding for open-world concepts
- Build upon pre-trained foundation models, design semi-automatic data collection methods
- Code, models, online demo available: https://llava-vl.github.io/, https://vip-llava.github.io/