

# Recent Advances in Vision Foundation Models

In conjunction with **CVPR 2024**

June 17<sup>th</sup> 2024 (9 a.m. PDT – 5 p.m. PDT)

Location: Summit 437- 439, Seattle Convention Center





## Morning Session

9:00 - 9:20	Opening Remarks	Lijuan Wang
9:20 - 10:10	Large Multimodal Models: Towards Building General-Purpose Multimodal Assistant	Chunyuan Li
10:10 - 11:00	Methods, Analysis & Insights from Multimodal LLM Pre-training	Zhe Gan
11:00 - 11:50	LMMs with Fine-Grained Grounding Capabilities	Haotian Zhang

## Afternoon Session

13:00 - 13:50	A Close Look at Vision in Large Multimodal Models	Jianwei Yang
13:50 - 14:40	Multimodal Agents	Linjie Li
14:40 - 15:00	Coffee Break & QA	
15:00 - 15:50	Recent Advances in Image Generative Foundation Models	Zhengyuan Yang
15:50 - 16:40	Video and 3D Generation	Kevin Lin

## 16:40 - 17:00 Closing Remarks & QA



Chunyuan Li

Tiktok



Zhe Gan

Apple



Haotian Zhang

Apple



Jianwei Yang

Microsoft



Linjie Li

Microsoft



Zhengyuan Yang

Microsoft



Kevin Lin

Microsoft



Jianfeng Gao

Microsoft



Lijuan Wang

Microsoft

# Previous Tutorials (2020-2023)

2020

## Agenda

- 1:15 - 1:25 **Opening Remarks** presented by JJ Liu and Xiaodong He ( [Slides](#), [YouTube](#), [Bilibili](#) )
- 1:25 - 2:15 **Visual QA and Reasoning** presented by Zhe Gan ( [Slides](#), [YouTube](#), [Bilibili](#) )
- 2:15 - 2:30 **Coffee Break**
- 2:30 - 3:10 **Visual Captioning** presented by Luowei Zhou ( [Slides](#), [YouTube](#), [Bilibili](#) )
- 3:10 - 3:40 **Text-to-image Synthesis** presented by Yu Cheng ( [Slides](#), [YouTube](#), [Bilibili](#) )
- 3:40 - 4:00 **Coffee Break**
- 4:00 - 5:00 **Self-supervised Learning** presented by Licheng Yu, Linjie Li and Yen-Chun Chen ( [Slides](#), [YouTube](#), [Bilibili](#) )

2022

### Morning Session

9:00 - 9:15	Opening Remarks ( <a href="#">Bilibili</a> , <a href="#">YouTube</a> )	Lijuan Wang
9:15 - 10:00	Overview of Image-Text Pre-training ( <a href="#">Slides</a> ) ( <a href="#">Bilibili</a> , <a href="#">YouTube</a> )	Jianfeng Wang
10:00 - 10:15	Coffee Break & QA	
10:15 - 11:00	Unified Image-Text Modeling ( <a href="#">Slides</a> ) ( <a href="#">Bilibili</a> , <a href="#">YouTube</a> )	Zhengyuan Yang
11:00 - 11:45	Advanced Topics in Image-Text Pre-training ( <a href="#">Slides</a> ) ( <a href="#">Bilibili</a> , <a href="#">YouTube</a> )	Zhe Gan
11:45 - 12:00	Q & A	

### Afternoon Session

13:00 - 13:30	Overview of Video-Text Pre-training ( <a href="#">Slides</a> ) ( <a href="#">Bilibili</a> , <a href="#">YouTube</a> )	Kevin Lin
13:30 - 14:00	Learning from Multi-channel Videos: Methods and Benchmarks ( <a href="#">Slides</a> ) ( <a href="#">Bilibili</a> , <a href="#">YouTube</a> )	Linjie Li
14:00 - 14:30	Advanced Topics in Video-Text Pre-training ( <a href="#">Slides</a> ) ( <a href="#">Bilibili</a> , <a href="#">YouTube</a> )	Chung-Ching Lin
14:30 - 14:45	Coffee Break & QA	
14:45 - 15:15	VLP for Image Classification ( <a href="#">Slides</a> ) ( <a href="#">Bilibili</a> , <a href="#">YouTube</a> )	Jianwei Yang
15:15 - 15:45	VLP for Object Detection ( <a href="#">Slides</a> ) ( <a href="#">Bilibili</a> , <a href="#">YouTube</a> )	Pengchuan Zhang
15:45 - 16:15	Benchmarks for Computer Vision in the Wild ( <a href="#">Slides</a> ) ( <a href="#">Bilibili</a> , <a href="#">YouTube</a> )	Chunyuan Li
16:15 - 17:00	VLP for Text-to-Image Synthesis ( <a href="#">Slides</a> ) ( <a href="#">Bilibili</a> , <a href="#">YouTube</a> )	Chenfei Wu
17:00 - 17:15	Q & A	

2021

### Prerecorded Sessions

4min	Opening Remarks ( <a href="#">Video</a> )	Jingjing Liu and Xiaodong He
50min	Representations and Training Strategies for VLP ( <a href="#">Video</a> ) ( <a href="#">Slides</a> )	Zhe Gan
40min	Robustness, Efficiency and Extensions for VLP ( <a href="#">Video</a> ) ( <a href="#">Slides</a> )	Linjie Li
40min	Video-and-Language Pre-training ( <a href="#">Video</a> ) ( <a href="#">Slides</a> )	Luowei Zhou
42min	Introduction to VLN ( <a href="#">Video</a> ) ( <a href="#">Slides</a> )	Qi Wu
55min	Generalizable VLN Methods ( <a href="#">Video</a> ) ( <a href="#">Slides</a> )	Xin Eric Wang
58min	Forward to Realistic VLN ( <a href="#">Video</a> ) ( <a href="#">Slides</a> )	Yoav Artzi and Peter Anderson
15min	VLN Summary ( <a href="#">Video</a> ) ( <a href="#">Slides</a> )	Qi Wu

### Live Session

16:00-17:00	Panel Discussion (🔊) LIVE on <a href="#">Zoom</a> ( <a href="#">Video</a> )	All speakers
-------------	---	--------------

2023

9:00 - 9:40	Opening Remarks & Visual and Vision-Language Pre-training ( <a href="#">Slides</a> ) ( <a href="#">Bilibili</a> , <a href="#">YouTube</a> )
9:40 - 10:20	From Representation to Interface: The Evolution of Foundation for Vision Understanding ( <a href="#">Slides</a> ) ( <a href="#">Bilibili</a> , <a href="#">YouTube</a> )
10:20 - 11:00	Alignments in Text-to-Image Generation ( <a href="#">Slides</a> ) ( <a href="#">Bilibili</a> , <a href="#">YouTube</a> )
11:00 - 11:40	Large Multimodal Models ( <a href="#">Slides</a> , <a href="#">Notes</a> ) ( <a href="#">Bilibili</a> , <a href="#">YouTube</a> )
11:40 - 12:10	Multimodal Agents: Chaining Multimodal Experts with LLMs ( <a href="#">Slides</a> ) ( <a href="#">Bilibili</a> , <a href="#">YouTube</a> )
12:10 - 12:30	Q & A

# Check out our survey paper

## Vision-Language Pre-training: Basics, Recent Advances, and Future Trends

2023

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao  
Microsoft Corporation  
{zhgan, linjli, chunyl, lijuanw, zliu, jfgao}@microsoft.com

### VQA & Visual Reasoning

Q: What is the dog holding with its paws?  
A: Frisbee.

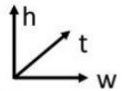
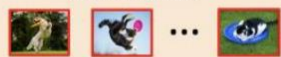
### Image Captioning

Caption: A dog is lying on the grass next to a frisbee.

### Text-to-Image Retrieval

Query: A dog is lying on the grass next to a frisbee.

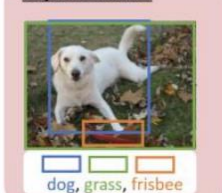
#### Negative Images



### Image Classification

Labels: [dog, grass, frisbee]

### Object Detection



### Segmentation



### Text-to-Video Retrieval

Query: A dog is lying on the grass next to a frisbee, while shaking its tail.

#### Negative Videos



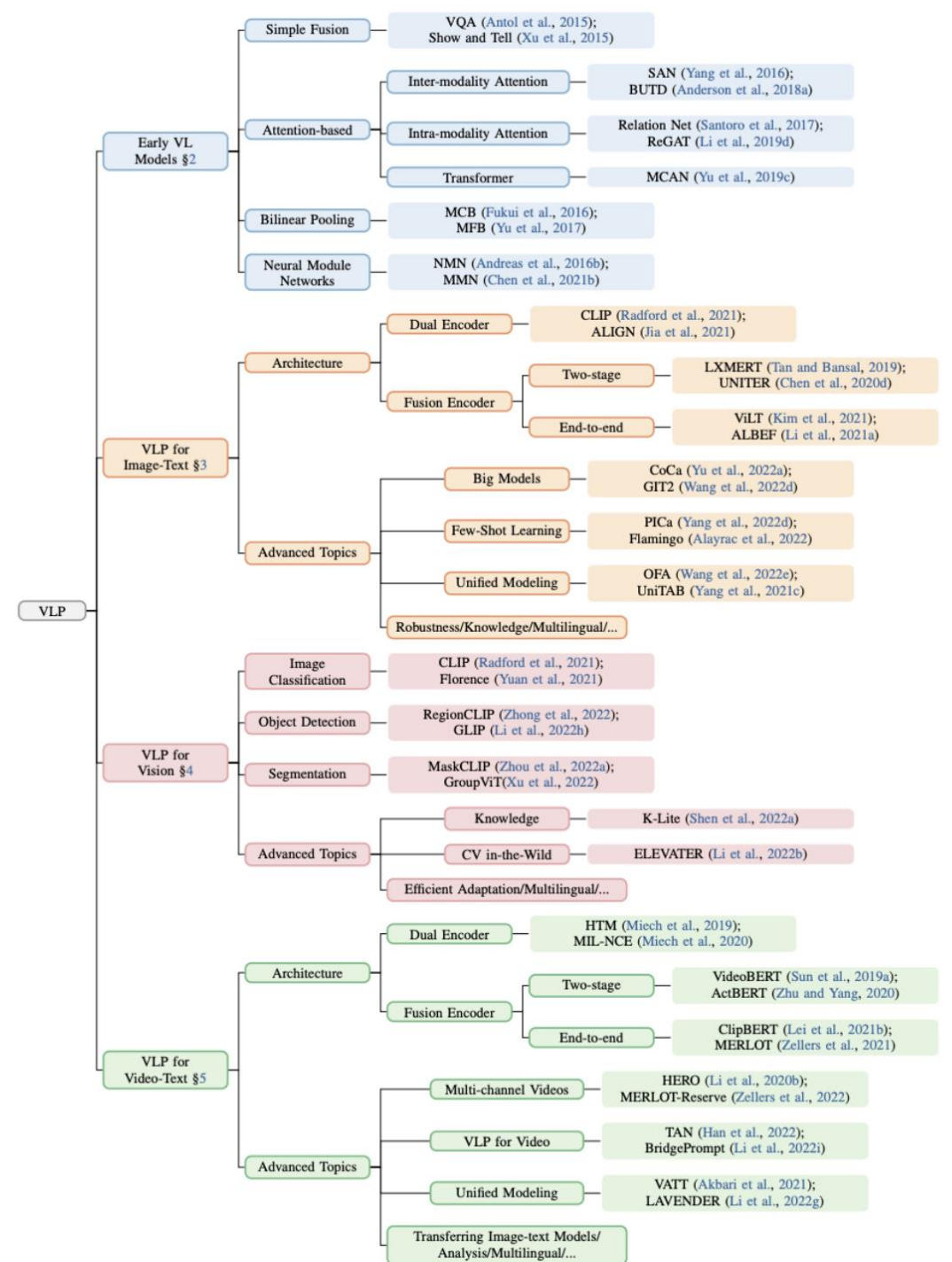
### Video Captioning

Caption: A dog is lying on the grass next to a frisbee, while shaking its tail.

### Video Question Answering

Q: Is the dog perfectly still?  
A: No.

Figure 1.2: Illustration of representative tasks from three categories of VL problems covered in this paper: image-text tasks, vision tasks as VL problems, and video-text tasks.

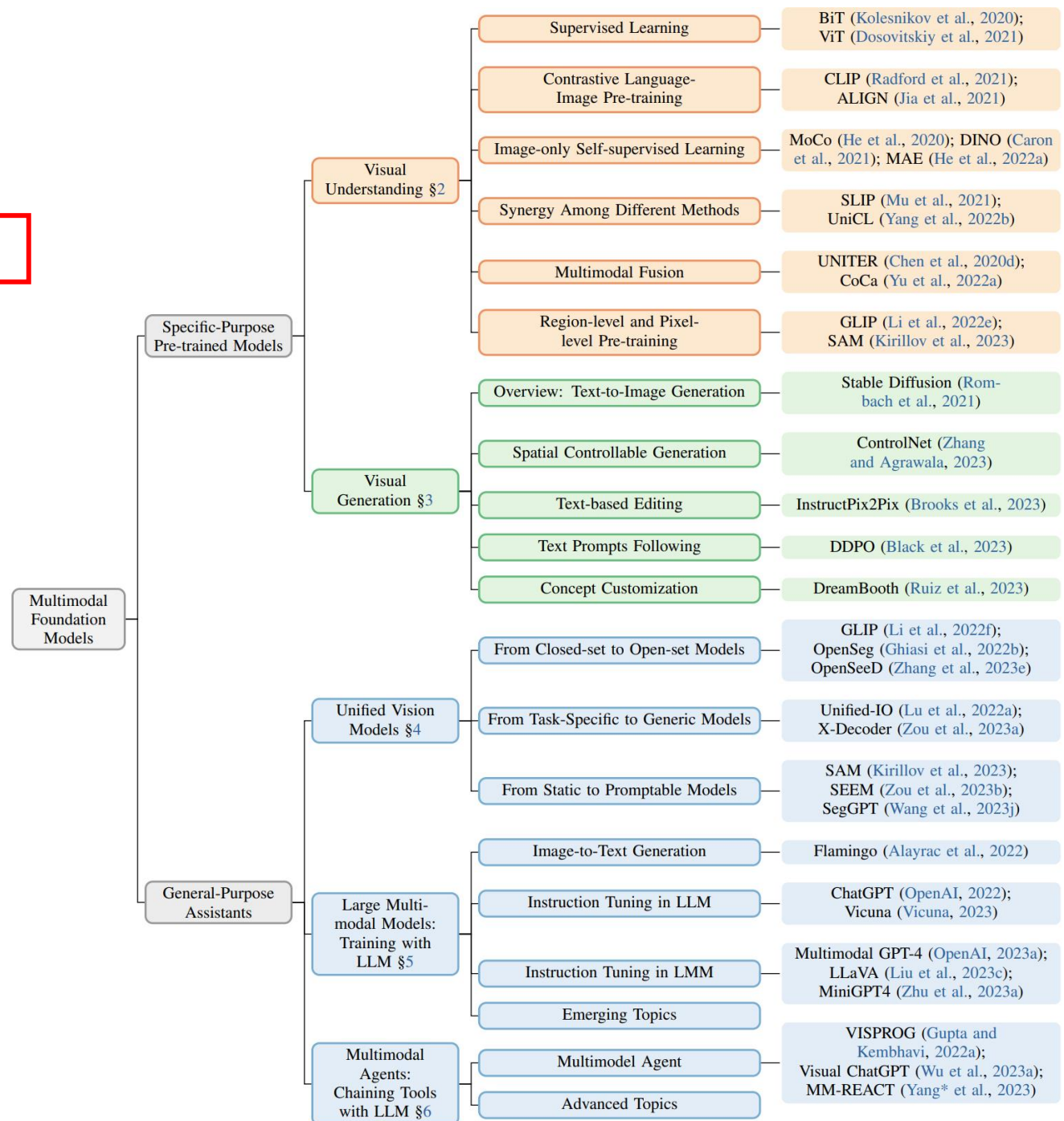




# Check out our survey paper

## Multimodal Foundation Models: 2024 From Specialists to General-Purpose Assistants

Chunyuan Li\*<sup>♣</sup>, Zhe Gan\*, Zhengyuan Yang\*, Jianwei Yang\*, Linjie Li\*,  
Lijuan Wang, Jianfeng Gao  
Microsoft Corporation  
{chunyl,zhgan,zhengyang,jianwyan,linjli,lijuanw,jfgao}@microsoft.com  
\* Core Contribution ♣ Project Lead



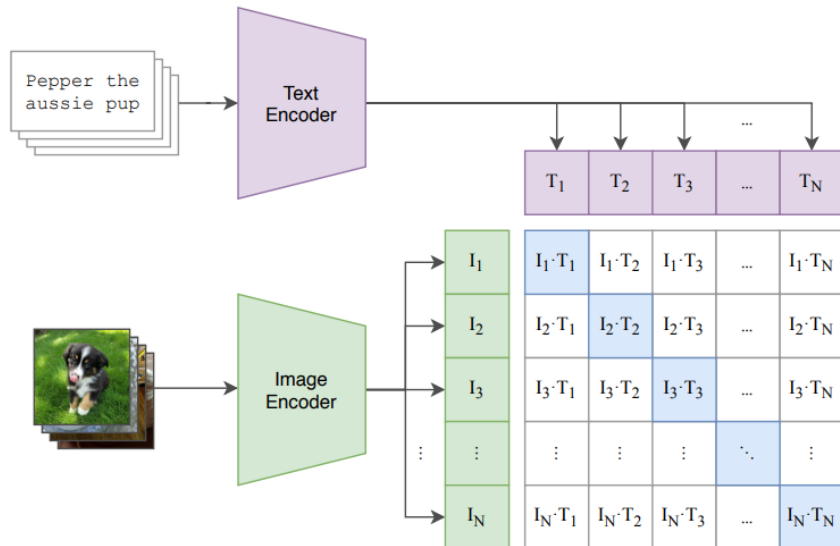
# Recent Advances in Vision Foundation Models

- Looking back, CLIP was a big paradigm shift
- LMMs extends LLMs with multi-sensory skill to achieve generic intelligence
  - Prelude of LMMs (Early Vision-Language Models, Flamingo, CoCa, GIT)
  - The era of LMMs starts from GPT-4V
  - Landscape of open-source and proprietary LMMs
  - New research areas: Grounding LMMs, Visual prompting, Multimodal Agent
- Diffusion model as a vision-centered representation learner
  - Your diffusion model is secretly a zero shot classifier
  - DALL-E 3: reconstruct image from ultra-descriptive caption
  - SORA: video generation models as world simulators

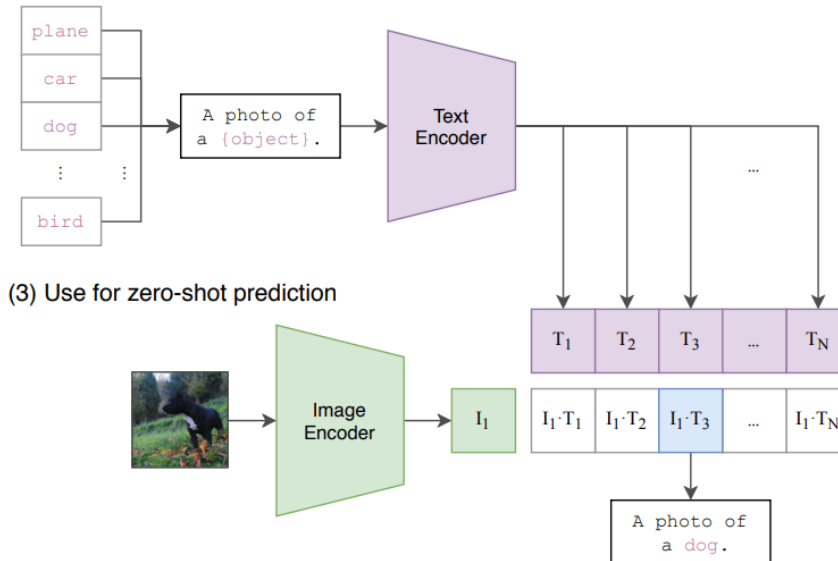
# CLIP was a Big Paradigm Shift

Rather than needing handcrafted labels to train a good classifier for a given domain, we can leverage free-form text from the internet to learn a model that is a good classifier for *all* domains

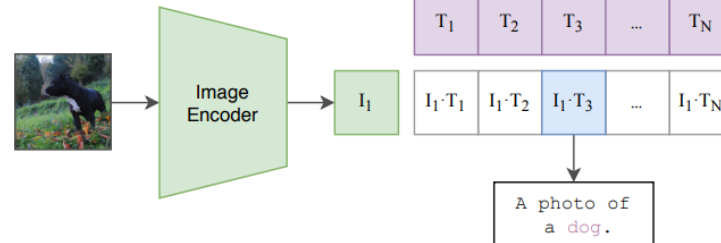
(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



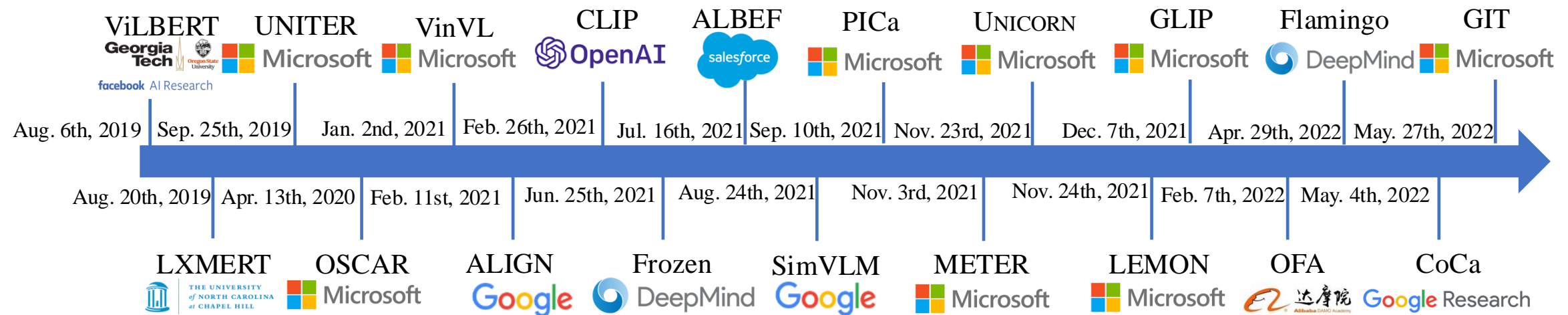
# Recent Advances in Vision Foundation Models

- Looking back, CLIP was a big paradigm shift
- LMMs extends LLMs with multi-sensory skill to achieve generic intelligence
  - Prelude of LMMs (Early Vision-Language Models, Flamingo, CoCa, GIT)
  - The dawn of LMMs starts from GPT-4V
  - Landscape of open-source and proprietary LMMs
  - New research areas: Grounding LMMs, Visual prompting, Multimodal Agent
- Diffusion model as a vision-centered representation learner
  - Your diffusion model is secretly a zero shot classifier
  - DALL-E 3: reconstruct image from ultra-descriptive caption
  - SORA: video generation models as world simulators



# Prelude of LMMs (Early Vision-Language Models)

- Early VLP models depend on pre-trained object detectors to extract visual features offline.
- Newer end-to-end VLP models achieve stronger performance with model and data scaling.
- Upscaled VLP models demonstrate new capabilities such as in-context learning and multimodal few shots.



# The Era of LMMs starts from GPT-4V

---

## The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)

---

Zhengyuan Yang\*, Linjie Li\*, Kevin Lin\*, Jianfeng Wang\*, Chung-Ching Lin\*,  
Zicheng Liu, Lijuan Wang\*♣  
Microsoft Corporation

\* Core Contributor ♣ Project Lead

“The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)” work from our colleagues at Microsoft covers a plethora of practical observations and strategies for using GPT-4V.

<https://openai.com/contributions/gpt-4v/>

<b>2</b>	<b>GPT-4V’s Input Modes</b>	<b>10</b>
2.1	Text-only Inputs . . . . .	10
2.2	Single Image-text Pair . . . . .	10
2.3	Interleaved Image-text Inputs . . . . .	11
<b>3</b>	<b>GPT-4V’s Working Modes and Prompting Techniques</b>	<b>12</b>
3.1	Following Text Instructions . . . . .	15
3.2	Visual Pointing and Visual Referring Prompting . . . . .	15
3.3	Visual + Text Prompting . . . . .	16
3.4	In-context Few-shot Learning . . . . .	19
<b>4</b>	<b>Vision-Language Capability</b>	<b>26</b>
4.1	Image Description on Diverse Domains . . . . .	26
4.2	Object Localization, Counting, and Dense Captioning . . . . .	38
4.3	Multimodal Knowledge and Commonsense . . . . .	43
4.4	Scene Text, Table, Chart, and Document Reasoning . . . . .	49
4.5	Multilingual Multimodal Understanding . . . . .	58
4.6	Coding Capability with Vision . . . . .	63
<b>5</b>	<b>Interaction with Humans: Visual Referring Prompting</b>	<b>66</b>
5.1	Understand Pointing Inputs . . . . .	66
5.2	Visual Referring Prompting . . . . .	69
5.3	Generate Pointing Outputs . . . . .	69
<b>6</b>	<b>Temporal and Video Understanding</b>	<b>73</b>
6.1	Multi-image Sequencing . . . . .	73
6.2	Video Understanding . . . . .	73
6.3	Visual Referring Prompting for Grounded Temporal Understanding . . . . .	79
<b>7</b>	<b>Abstract Visual Reasoning and Intelligence Quotient Test</b>	<b>80</b>
7.1	Abstract Visual Stimuli . . . . .	80
7.2	Discovery and Association of Parts and Objects . . . . .	80
7.3	Wechsler Adult Intelligence Scale . . . . .	80
7.4	Raven’s Progressive Matrices . . . . .	80
<b>8</b>	<b>Emotional Quotient Test</b>	<b>86</b>
8.1	Read Emotion from Facial Expressions . . . . .	86
8.2	Understand How Visual Content Arouses Emotions . . . . .	87
8.3	Emotion Conditioned Output . . . . .	89
<b>9</b>	<b>Emerging Application Highlights</b>	<b>90</b>
9.1	Spot the Difference . . . . .	90

# GPT-4V Emerging Capability Highlights

- LMM emergent capabilities

## The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)

Zhengyuan Yang\*, Linjie Li\*, Kevin Lin\*, Jianfeng Wang\*, Chung-Ching Lin\*,  
Zicheng Liu, Lijuan Wang\*  
Microsoft Corporation


\* Core Contributor \* Project Lead

### Abstract

Large multimodal models (LMMs) extend large language models (LLMs) with multi-sensory skills, such as visual understanding, to achieve stronger generic intelligence. In this paper, we analyze the latest model, GPT-4V(ision) [99–101, 1]<sup>1</sup>, to deepen the understanding of LMMs. The analysis focuses on the intriguing tasks that GPT-4V can perform, containing test samples to probe the quality and genericity of GPT-4V’s capabilities, its supported inputs and working modes, and the effective ways to prompt the model. In our approach to exploring GPT-4V, we curate and organize a collection of carefully designed qualitative samples spanning a variety of domains and tasks. Observations from these samples demonstrate that GPT-4V’s unprecedented ability in processing arbitrarily interleaved multimodal inputs and the genericity of its capabilities together make GPT-4V a powerful multimodal generalist system. Furthermore, GPT-4V’s unique capability of understanding visual markers drawn on input images can give rise to new human-computer interaction methods such as visual referring prompting. We conclude the report with in-depth discussions on the emerging application scenarios and the future research directions for GPT-4V-based systems. We hope that this preliminary exploration will inspire future research on the next-generation multimodal task formulation, new ways to exploit and enhance LMMs to solve real-world problems, and gaining better understanding of multimodal foundation models. Finally, we acknowledge that the model under our study is solely the product of OpenAI’s innovative work, and they should be fully credited for its development. Please see the GPT-4V contributions paper [101] for the authorship and credit attribution: <https://cdn.openai.com/contributions/gpt-4v.pdf>.

**USER** Please describe the image and answer the question shown in the image.

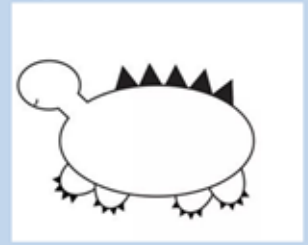
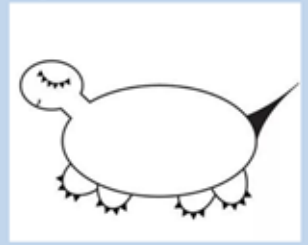
What does the arrow point to?



**GPT-V** The image shows a sock, a soccer ball, a tennis ball, a running shoe, and a skateboard. The arrow is pointing to the running shoe. So, the answer is that **the arrow points to a running shoe.**

## Visual Pointing

**USER** Compare Image 1 and image 2, tell me about the differences between image 1 and image 2.

Image 1:  Image 2: 

**GPT-V** 1. Image 1 has **spikes** on its back while image 2 does not.  
2. Image 2 has **a tail** while image 1 does not.  
3. Image 2 has **a mouth with teeth** while image 1 does not have a mouth

## Spot the Difference

**USER** Describe and summarize **what the circled person** doing?



**GPT-V** The circled person is **sitting on a bench and then stands up and adjusts his shirt.**

## Interleaved Image-text Sequence

[1] "The Dawn of LMMs: Preliminary Explorations with GPT-4V (ision)."



# GPT-4V Emerging Capability Highlights

- Genericity

---

## The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)

---

Zhengyuan Yang\*, Linjie Li\*, Kevin Lin\*, Jianfeng Wang\*, Chung-Ching Lin\*,  
Zicheng Liu, Lijuan Wang\*  
Microsoft Corporation

\* Core Contributor ♦ Project Lead

### Abstract

Large multimodal models (LMMs) extend large language models (LLMs) with multi-sensory skills, such as visual understanding, to achieve stronger generic intelligence. In this paper, we analyze the latest model, GPT-4V(ision) [99–101, 1]<sup>1</sup>, to deepen the understanding of LMMs. The analysis focuses on the intriguing tasks that GPT-4V can perform, containing test samples to probe the quality and genericity of GPT-4V’s capabilities, its supported inputs and working modes, and the effective ways to prompt the model. In our approach to exploring GPT-4V, we curate and organize a collection of carefully designed qualitative samples spanning a variety of domains and tasks. Observations from these samples demonstrate that GPT-4V’s unprecedented ability in processing arbitrarily interleaved multimodal inputs and the genericity of its capabilities together make GPT-4V a powerful multimodal generalist system. Furthermore, GPT-4V’s unique capability of understanding visual markers drawn on input images can give rise to new human-computer interaction methods such as visual referring prompting. We conclude the report with in-depth discussions on the emerging application scenarios and the future research directions for GPT-4V-based systems. We hope that this preliminary exploration will inspire future research on the next-generation multimodal task formulation, new ways to exploit and enhance LMMs to solve real-world problems, and gaining better understanding of multimodal foundation models. Finally, we acknowledge that the model under our study is solely the product of OpenAI’s innovative work, and they should be fully credited for its development. Please see the GPT-4V contributions paper [101] for the authorship and credit attribution: <https://cdn.openai.com/contributions/gpt-4v.pdf>.

**USER** Please read the table in this image and return a **markdown-style reconstructed table** in text.

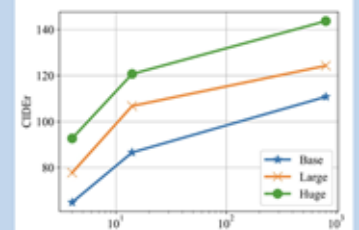
Country	Sales Volume	Revenue	Profit
USA	40.080	\$15,971,880	\$3,086,421
China	35.070	\$15,866,670	\$3,032,162
Australia	27.054	\$14,812,566	\$2,868,636
India	23.046	\$10,608,174	\$1,853,710
South Korea	16.032	\$10,494,948	\$1,975,844

**USER** Imagine that you are a robot operating a computer. Like how humans operate the computer, you can move the mouse, click an icon with the mouse, or type some texts with the keyboard.

Below is what you see on the computer screen. Describe the image and predict your next move. If the action involves moving the mouse, describe the location and the icon as detailed as possible.



**USER** Generate **python code** to draw similar curves.



million images in pre-training	Base CIDEr	Large CIDEr	Huge CIDEr
10 <sup>1</sup>	~75	~85	~95
10 <sup>2</sup>	~90	~110	~125
10 <sup>3</sup>	~110	~125	~140

**USER** Transcribe the video content, given the frames below.



**USER** Here is a list of reference images:



For below image, find what products are presented in the basket according to reference image above. Only tell me products that are exactly same type of reference images.



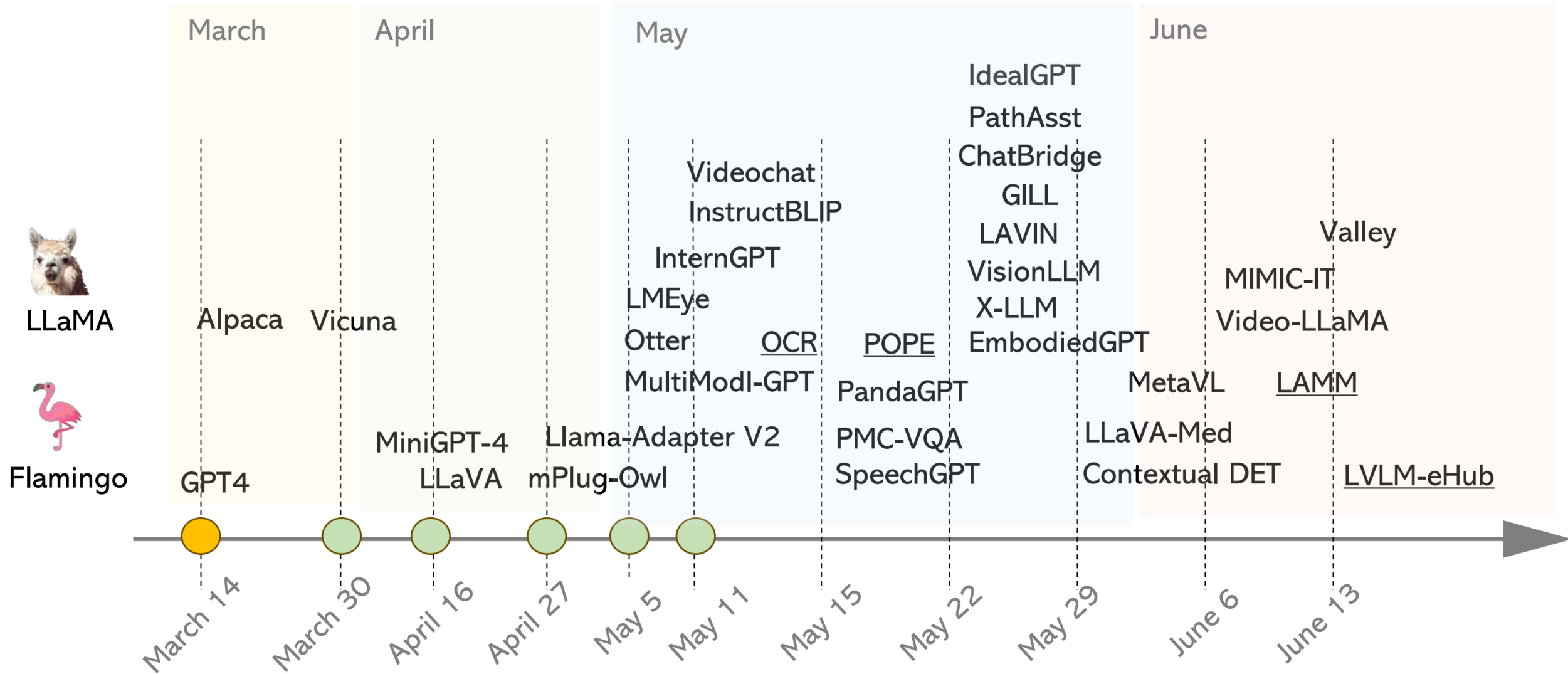
**USER** Imagine that you are a home robot, and is asked to go to the kitchen to fetch something from the fridge. The image below shows your current position. Please plan your next action.



Table, GUI, Coding, Video, Grocery, Embodied, etc.

[1] "The Dawn of LMMs: Preliminary Explorations with GPT-4V (ision)."

# Evolution of LMMs (a surge of Open Source LMMs since GPT-4V)



# Landscape of LMMs (Open-Source LMMs and Proprietary LMMs)

Human Expert   Open-Source   Proprietary

Reset	Size	Date	Overall	Art & Design	Business	Science	Health & Medicine	Human. & Social Sci.	Tech & Eng.
Human Expert (Best)	-	2024-01-31	88.6	89.2	90.7	90.0	87.3	89.2	86.2
Human Expert (Medium)	-	2024-01-31	82.6	84.2	86.0	84.7	78.8	85.0	79.1
Human Expert (Worst)	-	2024-01-31	76.2	80.8	78.0	78.0	73.3	74.2	74.3
GPT-4o*	-	2024-05-27	69.1	-	-	-	-	-	-
Gemini 1.5 Pro*	-	2024-05-31	<u>62.2</u>	-	-	-	-	-	-
Gemini 1.0 Ultra*	-	2023-12-11	59.4	70.0	56.7	48.0	67.3	78.3	47.1
Claude 3 Opus*	-	2024-03-05	59.4	67.5	67.2	48.9	61.1	70.0	50.6
GPT-4V(ision) (Playground)	-	2023-11-27	56.8	65.8	<u>59.3</u>	54.7	<u>64.7</u>	72.5	36.7
Reka Core*	-	2024-04-23	56.3	75.9	47.3	<u>49.3</u>	58.0	<u>75.0</u>	44.2
Gemini 1.5 Flash*	-	2024-05-31	56.1	-	-	-	-	-	-
SenseChat-Vision-0423-Preview*	-	2024-04-23	54.6	66.7	54.0	45.3	53.3	<u>75.0</u>	43.8
Reka Flash*	-	2024-04-23	53.3	61.7	42.7	47.3	59.3	74.2	44.3
Claude 3 Sonnet*	-	2024-03-05	53.1	61.7	58.2	37.1	57.1	68.7	45.0
HPT Pro*	-	2024-03-15	52.0	66.7	43.3	42.7	50.7	72.5	43.8
VILA1.5*	-	2024-05-04	51.9	60.8	43.3	36.0	57.3	73.3	<u>48.1</u>
InternVL-Chat-V1.2*	-	2024-02-23	51.6	62.5	40.7	39.3	58.7	70.0	46.2
Qwen-VL-MAX*	-	2024-01-27	51.4	<u>72.5</u>	43.3	40.0	58.0	69.2	38.6
Skywork-VL*	-	2024-05-30	51.4	66.7	41.3	38.7	55.3	68.3	46.7
LLaVA-1.6-34B *	34B	2024-02-01	51.1	67.5	46.0	39.3	52.0	67.5	43.8
Claude 3 Haiku*	-	2024-03-05	50.2	60.8	52.5	37.1	52.3	66.0	41.5
Adept Fuyu-Heavy*	-	2024-01-31	48.3	53.4	46.3	33.7	51.3	72.2	44.0
Gemini 1.0 Pro*	-	2023-12-11	47.9	-	-	-	-	-	-
Marco-VL-Plus*	-	2024-03-09	46.2	60.8	37.3	35.3	48.7	69.2	37.1

14

On MMMU leaderboard

- 30 Open-Source LMMs
- 21 Proprietary LMMs

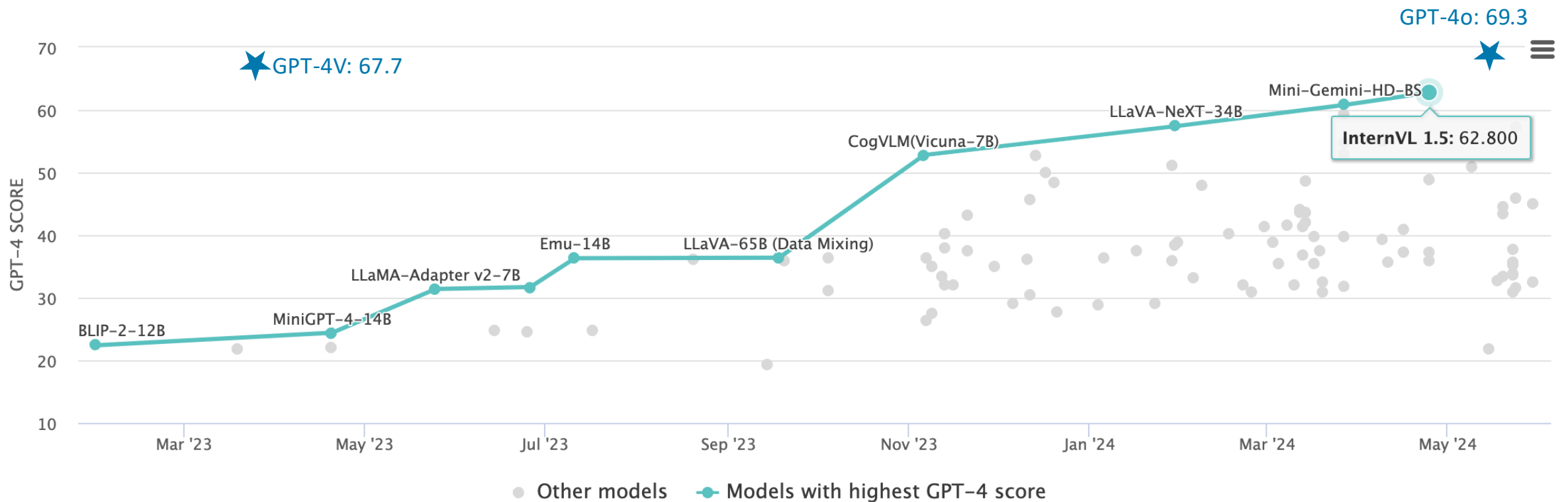
-- as of 6/13/2024



# Rapid Progress in LMMs



## MM-Vet : Evaluating integrated vision-language capabilities



GPT-4o (“o” for “omni”) is a step towards much more natural human-computer interaction—it accepts as input any combination of text, audio, image, and video and generates any combination of text, audio, and image outputs.

It would be interesting to investigate what additional capabilities a model combining all these modalities could achieve beyond current capabilities and how to make the native integration efficient so that different modalities enhance each other.

# LMM inspired new research area -- Visual Prompting

Input: Image



Input: Image + SoM



Conversation

**User** What is on the left side of the right laptop?

**GPT-4V** On the left side of the right laptop, there is a **cup or mug**. ❌

**User** I want to find a seat close to windows, where can I sit?

**GPT-4V** You can sit on **either of the two black chairs** in front of the white desks, as they are closest to the windowed wall. ❌

Conversation + 

**User** What is on the left side of the right laptop?

**GPT-4V** On the left side of the right laptop (numbered as 9), there is a **lamp, which is numbered as 12**. ✅

**User** I want to find a seat close to windows, where can I sit?

**GPT-4V** You can sit on the **chair numbered as 7**. It's the closest seat to the windows (represented by the area numbered as 1). ✅



# LMM inspired new research area – Multimodal Agent

## Agents with Multimodal Memory



**MM-Narrator**



**MM-Vid**



Audio Description

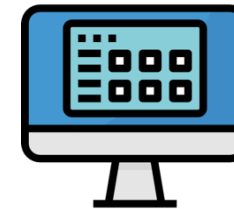
## Actionable Agents



**MM-Navigator**



**VideoGUI**



GUI Navigation

## Agent with Feedback



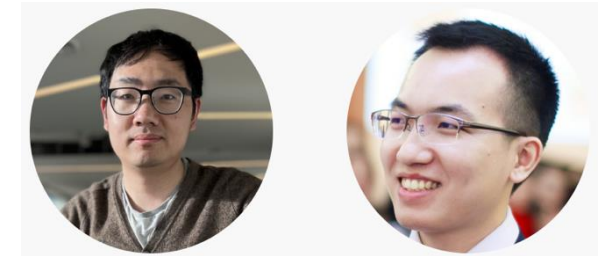
**Idea2Img**



Visual Design &  
Creation

# Recent Advances in Vision Foundation Models

- Looking back, CLIP is a big paradigm shift
- The past year was the year of LMMs
  - Prelude of LMMs (Early Vision-Language Models, Flamingo, CoCa, GIT)
  - The era of LMMs starts from GPT-4V
  - Landscape of open-source and proprietary LMMs
    - [Session 1: LLaVA and LMMs by Chunyuan Li](#)
    - [Session 2: LMMs pre-training by Zhe Gan](#)
- New research areas: Grounding LMMs, Visual prompting, Multimodal Agent
  - [Session 3: LMMs Grounding by Haotian Zhang](#)
  - [Session 4: Visual prompting by Jianwei Yang](#)
  - [Session 5: Multimodal agent by Linjie Li](#)



# Recent Advances in Vision Foundation Models

- Looking back, CLIP was a big paradigm shift
- The past year was the year of LMMs
  - Prelude of LMMs (Early Vision-Language Models, Flamingo, CoCa, GIT)
  - The era of LMMs starts from GPT-4V
  - Landscape of open-source and proprietary LMMs
  - New research areas: Grounding LMMs, Visual prompting, Multimodal Agent
- Diffusion model as a vision-centered representation learner
  - Your diffusion model is secretly a zero shot classifier
  - DALL-E 3: reconstruct image from ultra-descriptive caption
  - SORA: video generation models as world simulators

# Your Diffusion Model is Secretly a Zero-Shot Classifier

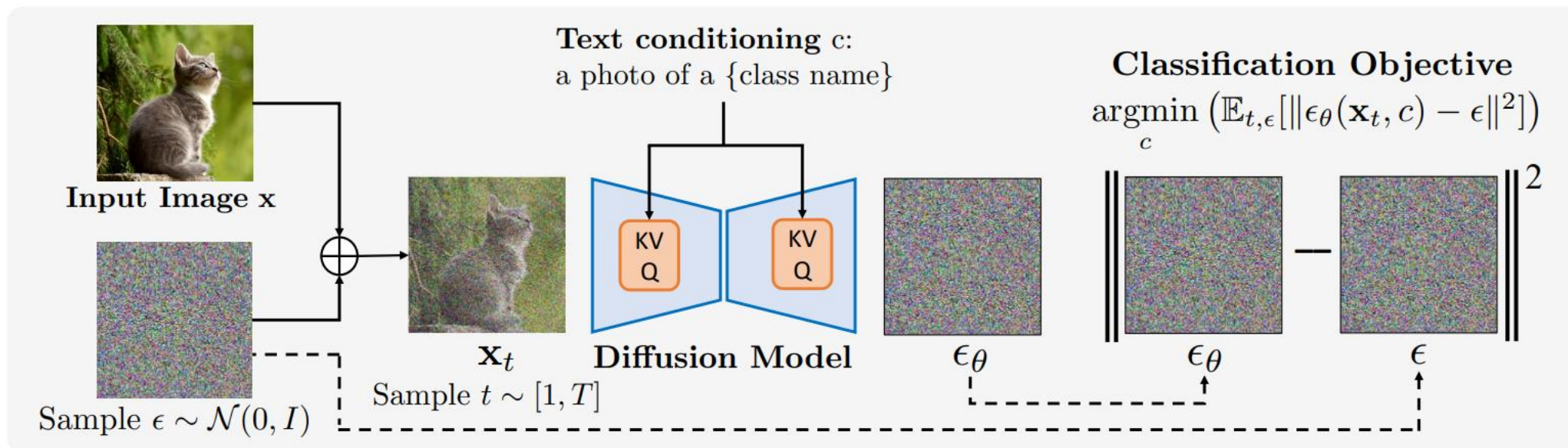
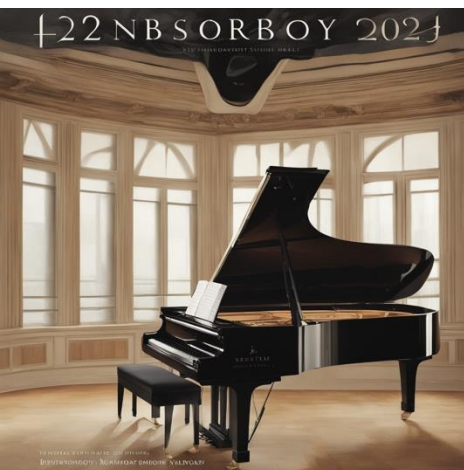
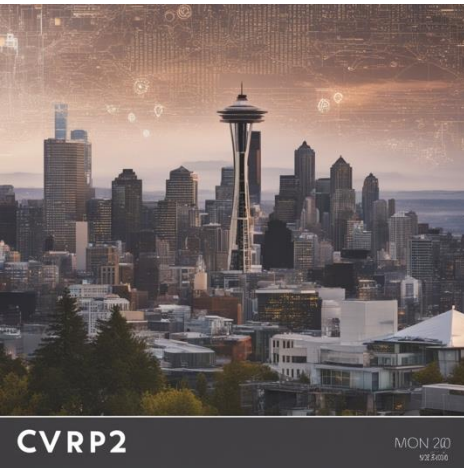


Figure 1. **Overview of our Diffusion Classifier approach:** Given an input image  $x$  and a set of possible conditioning inputs (e.g., text for Stable Diffusion or class index for DiT, an ImageNet class-conditional model), we use a diffusion model to choose the one that best fits this image. Diffusion Classifier is theoretically motivated through the variational view of diffusion models and uses the ELBO to approximate  $\log p_\theta(x | c)$ . Diffusion Classifier chooses the conditioning  $c$  that best predicts the noise added to the input image. *Diffusion Classifier can be used to extract a zero-shot classifier from Stable Diffusion and a standard classifier from DiT without any additional training.*





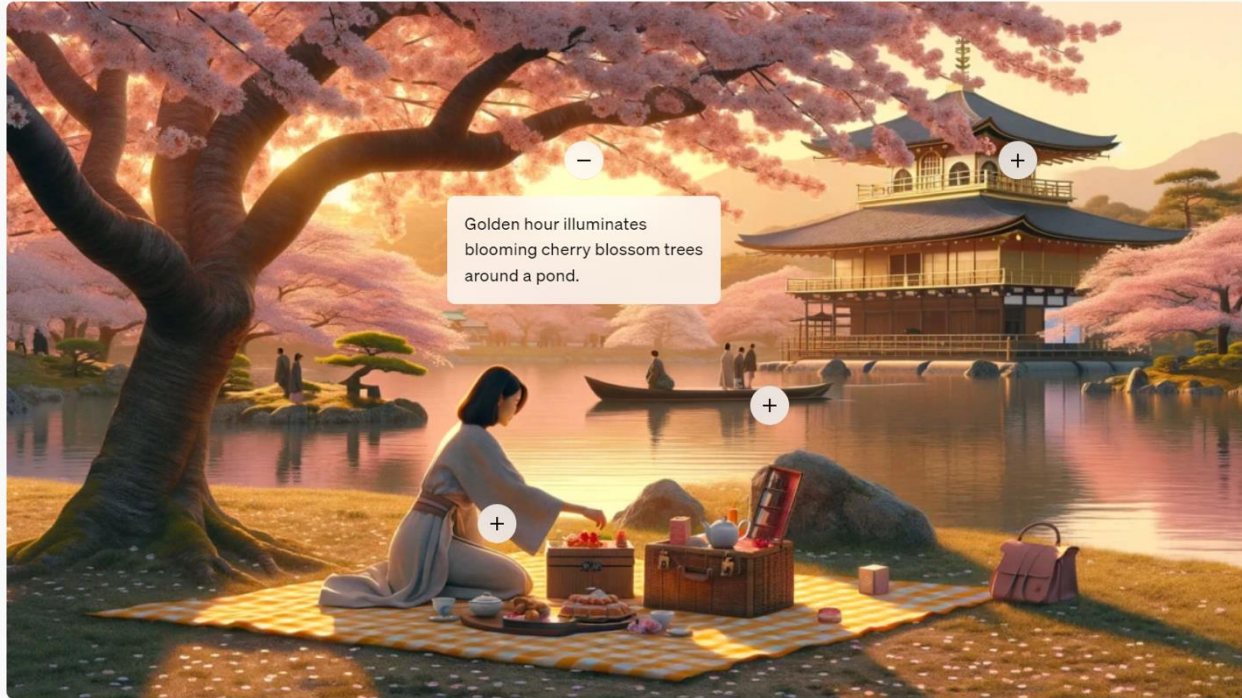
SDXL  
July 2023



Dalle-3  
Sept. 2023



# DALL-E 3: Reconstruct Image from Ultra-descriptive Caption



Core research and execution

Gabriel Goh, James Betker, Li Jing, Aditya Ramesh

Research contributors—primary

Tim Brooks, [Jianfeng Wang](#), [Lindsey Li](#), Long Ouyang, Juntang Zhuang, Joyce Lee, Prafulla Dhariwal, Casey Chu, Joy Jiao

Research contributors—secondary

Jong Wook Kim, Alex Nichol, Yang Song, [Lijuan Wang](#), Tao Xu



“Golden hour illuminates blooming cherry blossom trees around a pond.  
In the distance, a building with Japanese-inspired architecture is perched on the lake.  
In the pond, a group of people enjoying the serenity of the sunset in a rowboat.  
A woman underneath a cherry blossom tree is setting up a picnic on a yellow checkered blanket.”

- Takeaway from DALL-E 3: training on ultra-descriptive captions makes the model more compute-efficient, even when we measure the quality of samples produced with shorter captions
- Suggests that we can get better unconditional models by using UD captions as scaffolding, even if we don't use UD captions at inference

-- Aditya Ramesh

# Video Generation Models as World Simulators

Google

VDM 2022

Meta AI

Emu Video 2023

OpenAI

Sora 2024

Timeline



SORA: Our results suggest that scaling video generation models is a promising path towards building general purpose simulators of the physical world.

<https://openai.com/index/video-generation-models-as-world-simulators/>

# Recent Advances in Vision Foundation Models

- Looking back, CLIP was a big paradigm shift
- The past year was the year of LMMs
- Diffusion model as a vision-centered representation learner
  - Your diffusion model is secretly a zero shot classifier
  - DALL-E 3: reconstruct image from ultra-descriptive caption
  - SORA: video generation models as world simulators
    - [Session 6: Image Generation by Zhengyuan Yang](#)
    - [Session 7: Video and 3D Generation by Kevin Lin](#)





## Morning Session

9:00 - 9:20	Opening Remarks	Lijuan Wang
9:20 - 10:10	Large Multimodal Models: Towards Building General-Purpose Multimodal Assistant	Chunyuan Li
10:10 - 11:00	Methods, Analysis & Insights from Multimodal LLM Pre-training	Zhe Gan
11:00 - 11:50	LMMs with Fine-Grained Grounding Capabilities	Haotian Zhang

## Afternoon Session

13:00 - 13:50	A Close Look at Vision in Large Multimodal Models	Jianwei Yang
13:50 - 14:40	Multimodal Agents	Linjie Li
14:40 - 15:00	Coffee Break & QA	
15:00 - 15:50	Recent Advances in Image Generative Foundation Models	Zhengyuan Yang
15:50 - 16:40	Video and 3D Generation	Kevin Lin

## 16:40 - 17:00 Closing Remarks & QA



Chunyuan Li

Tiktok



Zhe Gan

Apple



Haotian Zhang

Apple



Jianwei Yang

Microsoft



Linjie Li

Microsoft



Zhengyuan Yang

Microsoft



Kevin Lin

Microsoft



Jianfeng Gao

Microsoft



Lijuan Wang

Microsoft



# Recent Advances in Vision Foundation Models

In conjunction with **CVPR 2024**

June 17<sup>th</sup> 2024 (9 a.m. PDT – 5 p.m. PDT)

Location: Summit 437- 439, Seattle Convention Center

