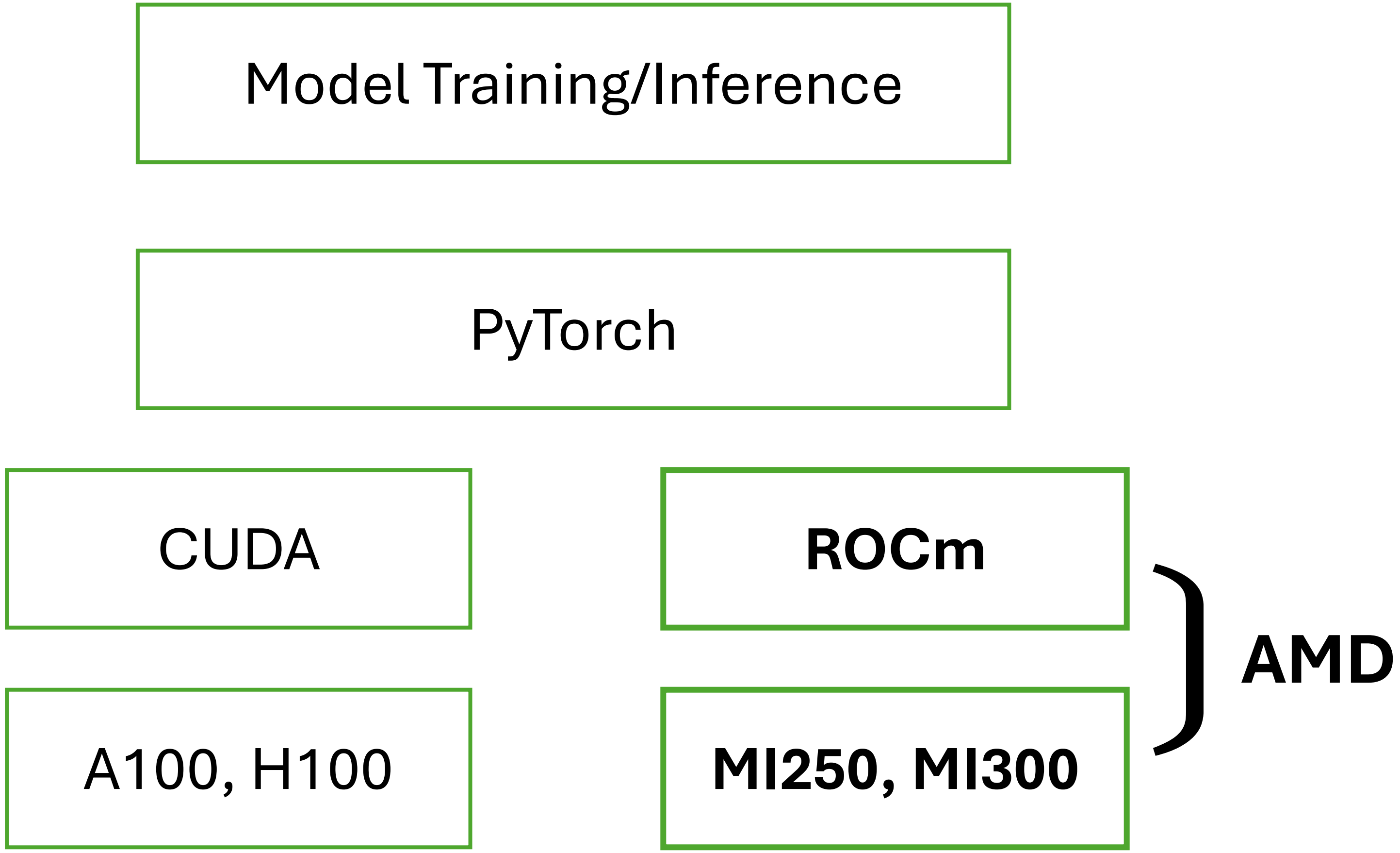
Evolution of Multimodal Foundation Models

Zicheng Liu
Senior Director of GenAI
AMD

hosted by

AMD



Model Training/Inference

PyTorch

CUDA

A100, H100

ROCm

MI250, MI300

AMD

Training Foundation Models from Scratch—Fully Open Source

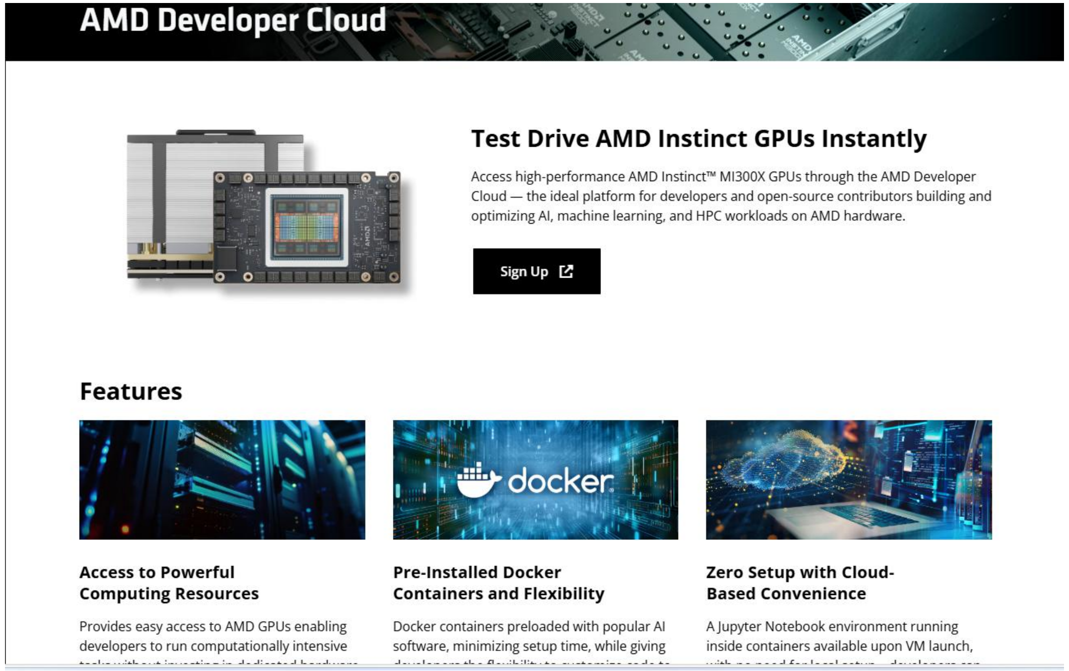
- Show credibility
- Generate awareness
- Provide feedback

Foundation Models Released

- AMD-OLMo-1B
  - Introducing the First AMD 1B Language Models: AMD OLMo
- Instella-3B
  - Introducing Instella: New State-of-the-art Fully Open 3B Language Models — ROCm Blogs
- Instella-VL-1B
  - Instella-VL-1B: First AMD Vision Language Model — ROCm Blogs
- Instella-Long
  - Introducing Instella-Long: A Fully Open Language Model with Long-Context Capability — ROCm Blogs
- Instella-T2I
  - https://arxiv.org/abs/2506.21022
  - https://rocm.blogs.amd.com/artificial-intelligence/instella-t2i/README.html

We’re hiring

AMD Developer Cloud



AMD Developer Cloud

Test Drive AMD Instinct GPUs Instantly

Access high-performance AMD Instinct™ MI300X GPUs through the AMD Developer Cloud — the ideal platform for developers and open-source contributors building and optimizing AI, machine learning, and HPC workloads on AMD hardware.

Sign Up

Features

Access to Powerful Computing Resources

Provides easy access to AMD GPUs enabling developers to run computationally intensive

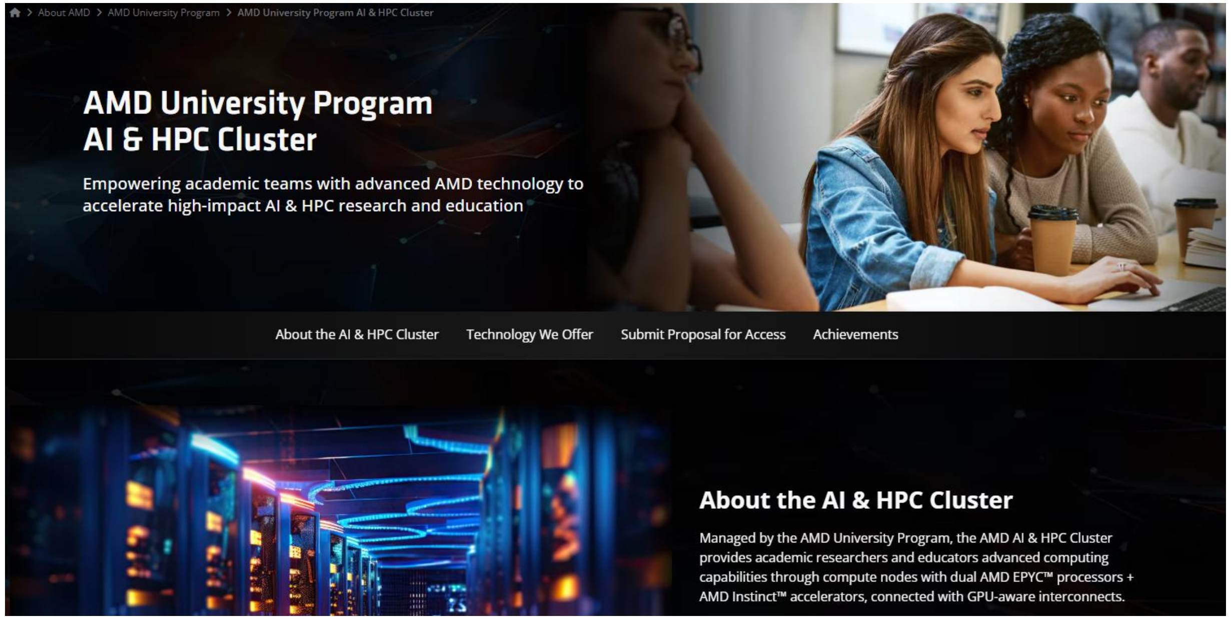
Pre-Installed Docker Containers and Flexibility

Docker containers preloaded with popular AI software, minimizing setup time, while giving

Zero Setup with Cloud-Based Convenience

A Jupyter Notebook environment running inside containers available upon VM launch,

AMD University Program AI & HPC Cluster



About AMD > AMD University Program > AMD University Program AI & HPC Cluster

AMD University Program AI & HPC Cluster

Empowering academic teams with advanced AMD technology to accelerate high-impact AI & HPC research and education

About the AI & HPC Cluster | Technology We Offer | Submit Proposal for Access | Achievements

About the AI & HPC Cluster

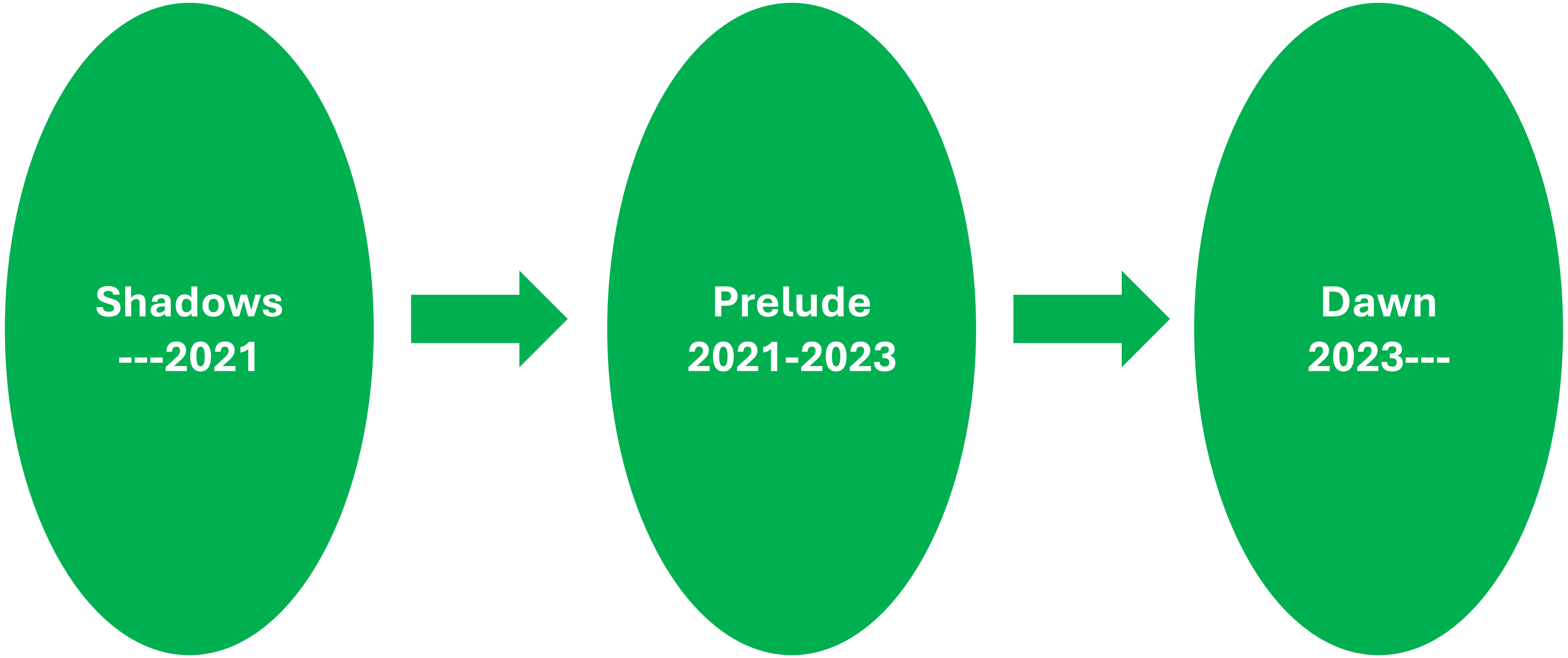
Managed by the AMD University Program, the AMD AI & HPC Cluster provides academic researchers and educators advanced computing capabilities through compute nodes with dual AMD EPYC™ processors + AMD Instinct™ accelerators, connected with GPU-aware interconnects.

Outline

- A brief history of multimodality models
  - Understanding
  - Generation
  - Unification
  - Agentic
- Special topics
  - Does image help with reasoning?
  - Token compression
- Summary and future directions

Outline

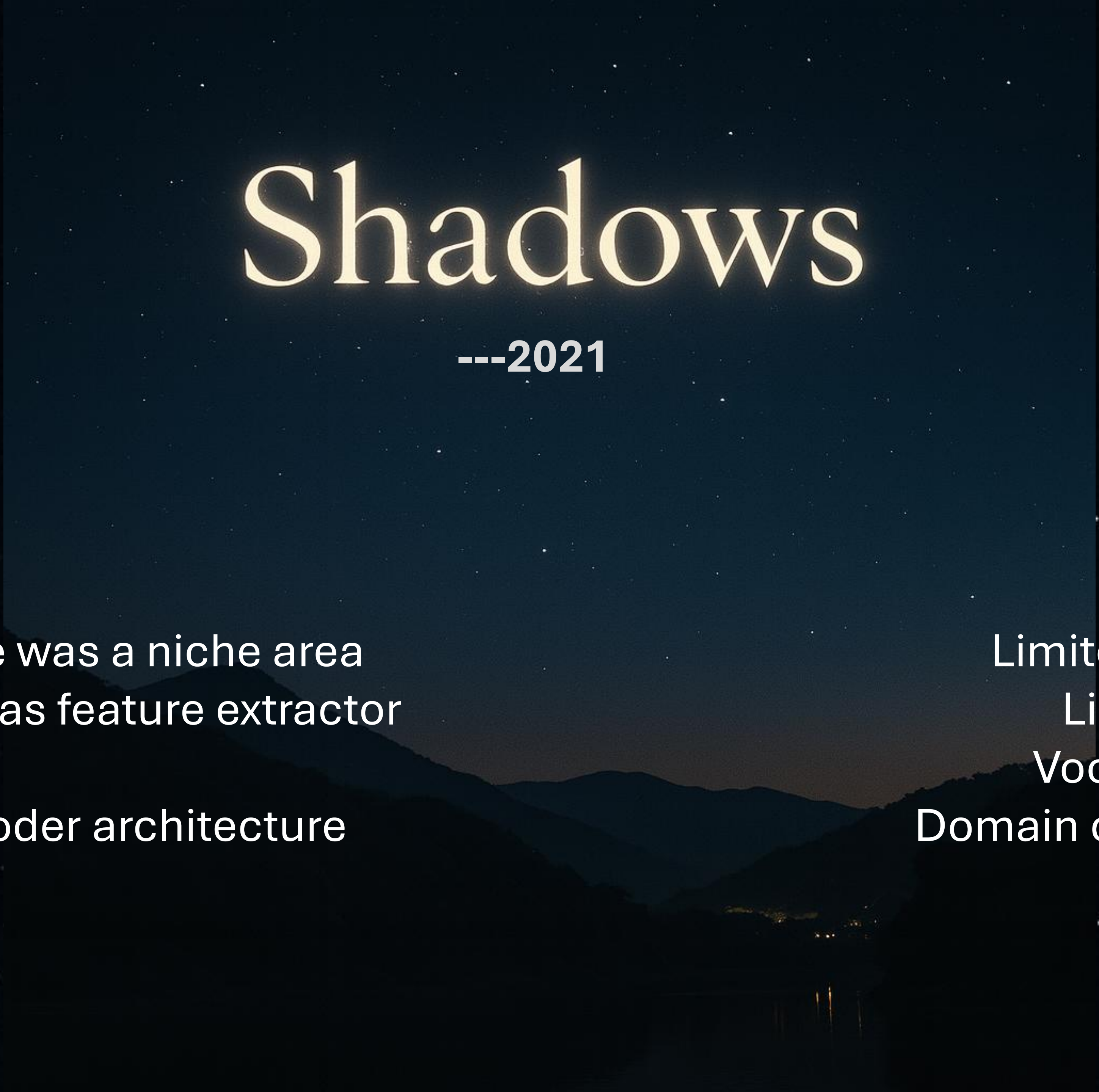
- A brief history of multimodality models
  - Understanding
  - Generation
  - Unification
  - Agentic
- Special topics
  - Does image help with reasoning?
  - Token compression
- Summary and future directions



Shadows
---2021

Prelude
2021-2023

Dawn
2023---



Shadows

---2021

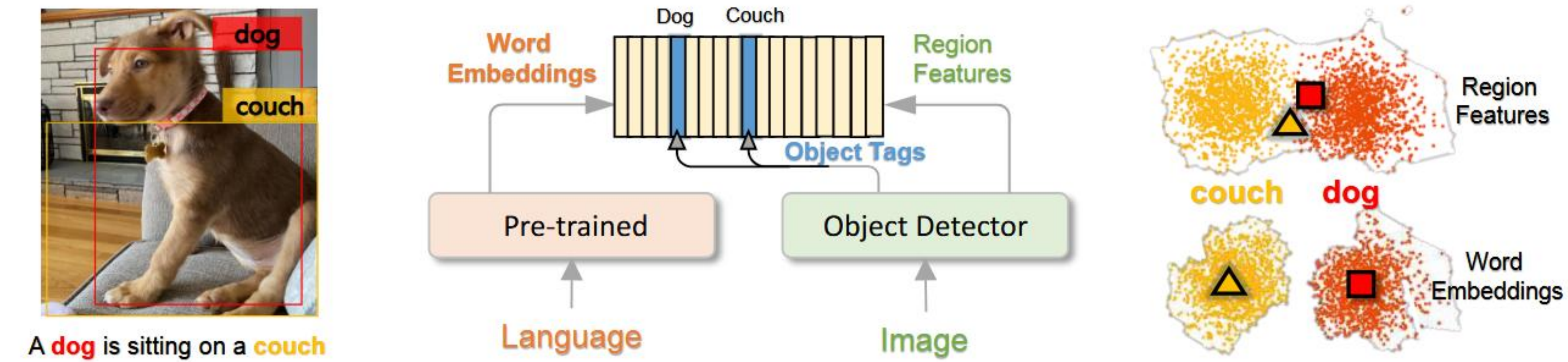
Research

Vision-Language was a niche area
Object detector as feature extractor
Not end-to-end
Multimodal encoder architecture

Industry

Limited vocabulary
Limited domain
Vocab expansion
Domain customization

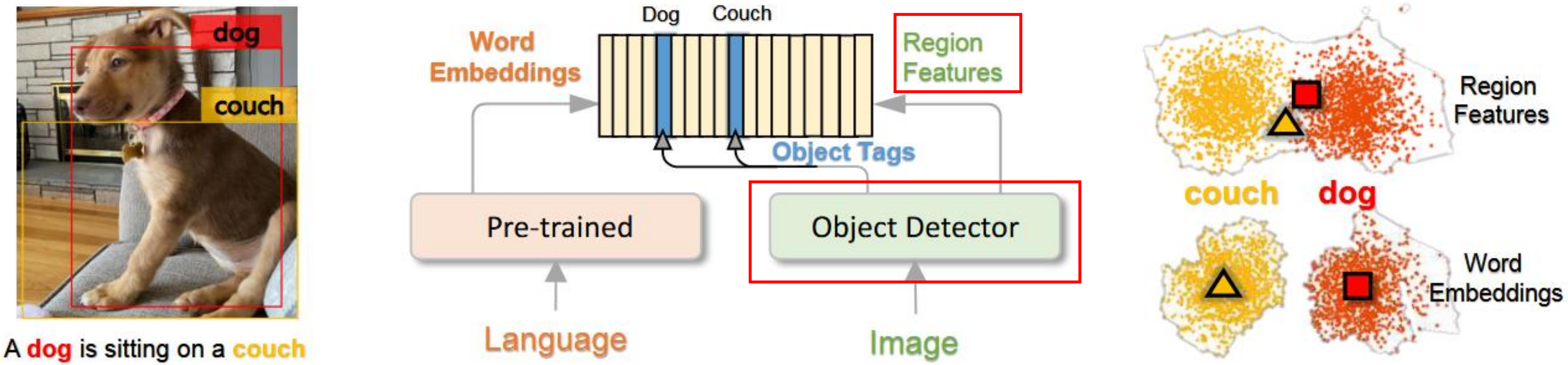
Oscar: Object-Semantics Aligned Pre-Training for Vision-Language Tasks (2020)



(a) Image-text pair (b) Objects as anchor points (c) Semantics spaces

Oscar: Object-Semantics Aligned Pre-Training for Vision-Language Tasks, X. Li, X. Yin, Y. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, J. Gao, 04/13/2020

Oscar: Object-Semantics Aligned Pre-Training for Vision-Language Tasks (2020)



(a) Image-text pair (b) Objects as anchor points (c) Semantics spaces

Oscar: Object-Semantics Aligned Pre-Training for Vision-Language Tasks, X. Li, X. Yin, Y. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, J. Gao, 04/13/2020

Uniter 2019

Oscar 2020

VinVL 2021

MiniVLM 2021

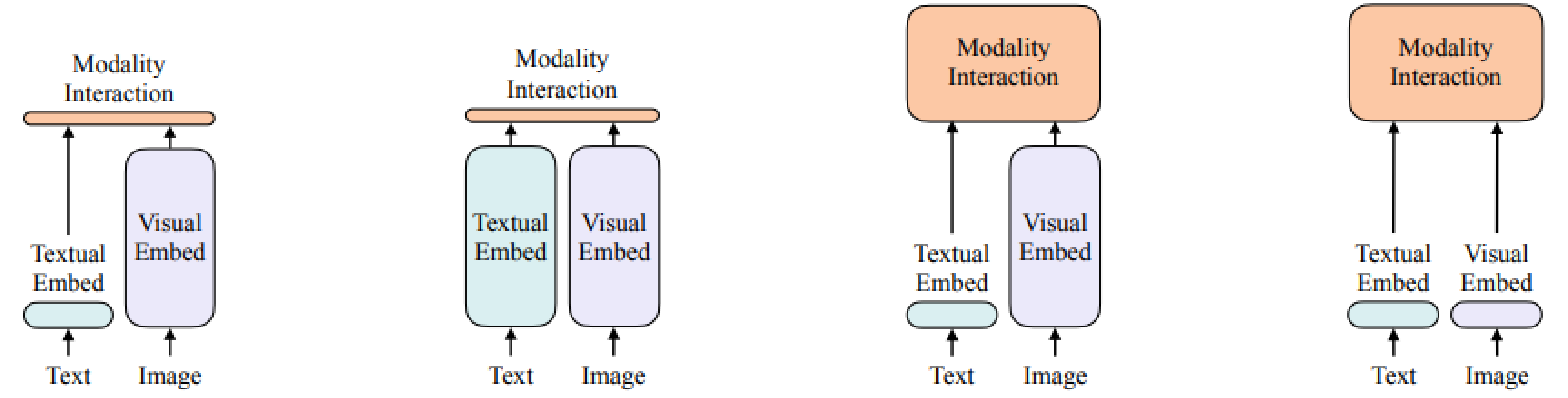
ViLT 2021

UFO 2021

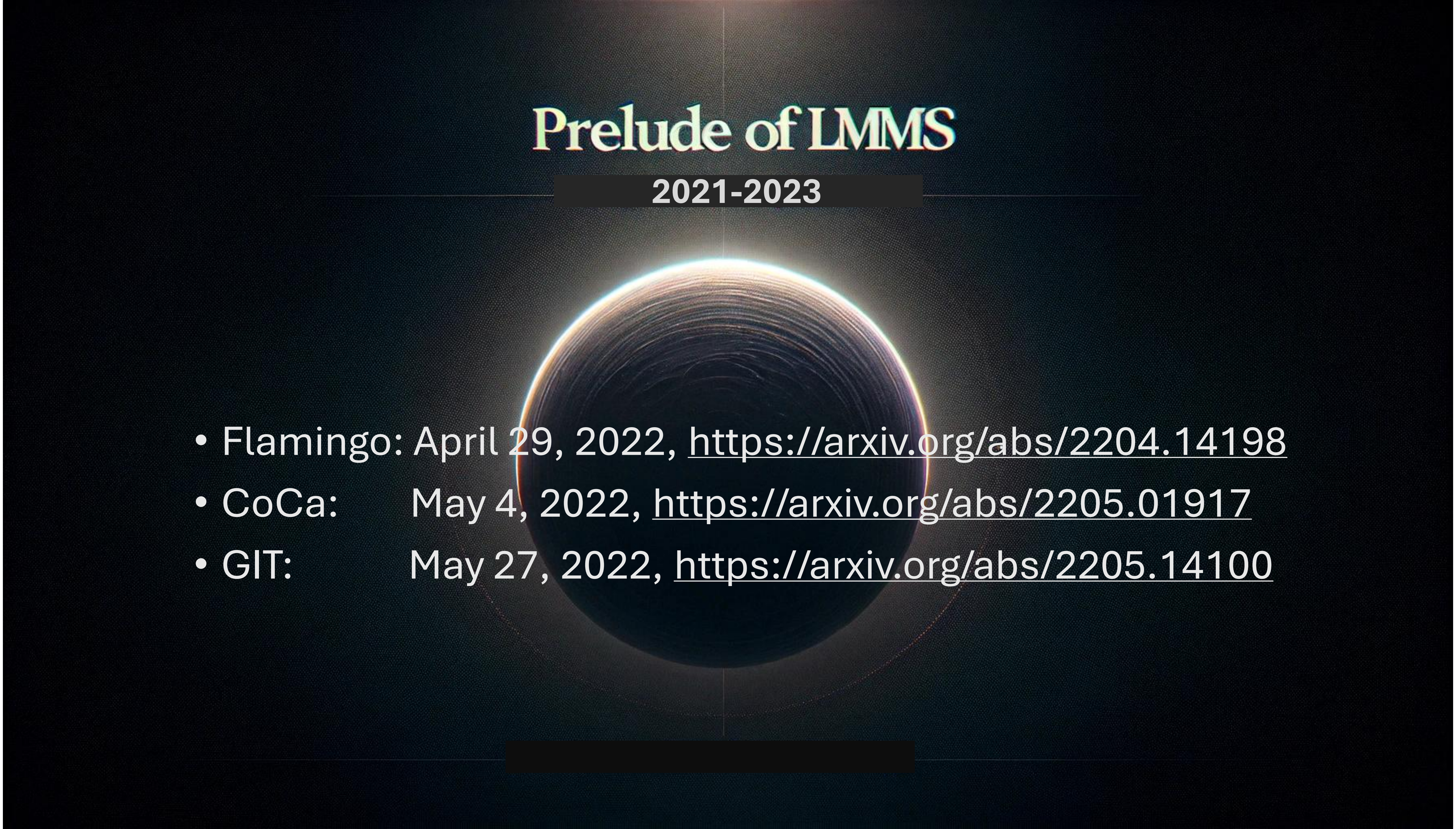
ViTCap 2021

Lemon 2021

ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision



(a) VE > TE > MI (b) VE = TE > MI (c) VE > MI > TE (d) MI > VE = TE



Prelude of LMMS

2021-2023

- Flamingo: April 29, 2022, https://arxiv.org/abs/2204.14198
- CoCa: May 4, 2022, https://arxiv.org/abs/2205.01917
- GIT: May 27, 2022, https://arxiv.org/abs/2205.14100

[redacted]

Flamingo: a Visual Language Model for Few-Shot Learning

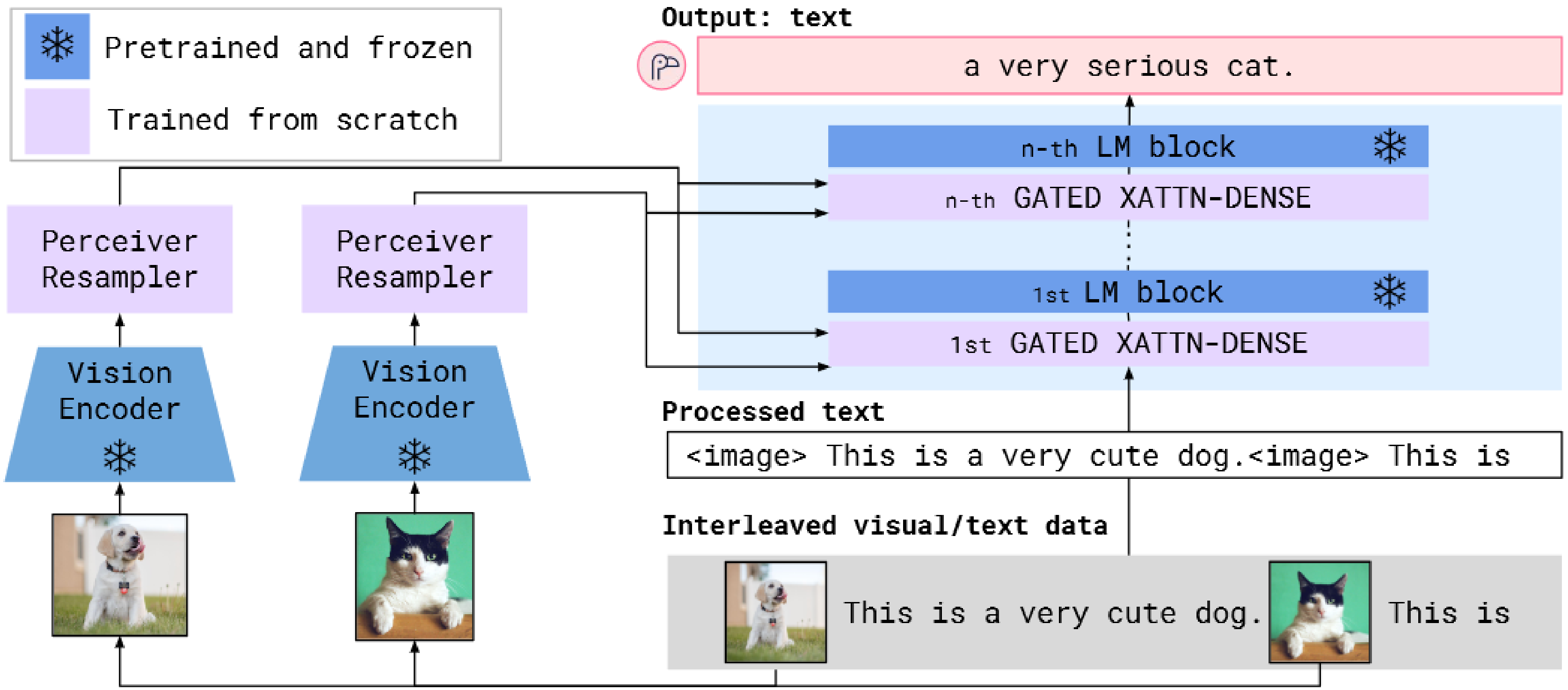
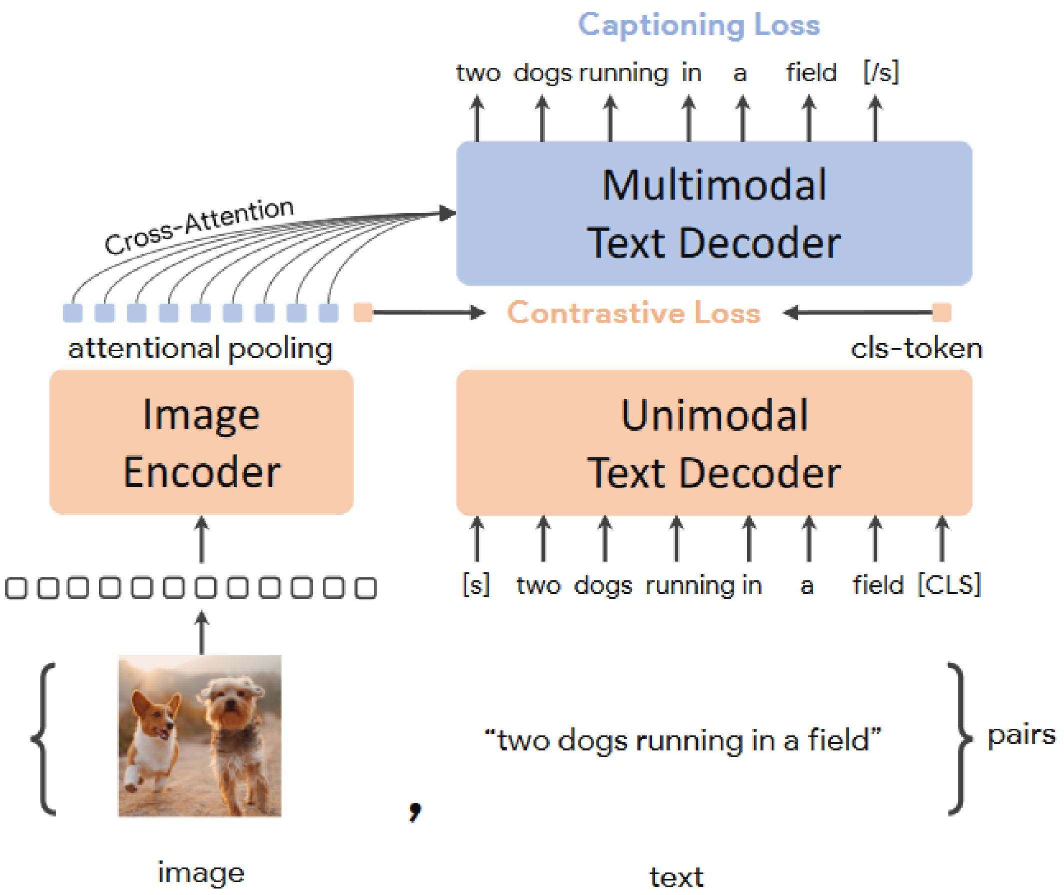


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

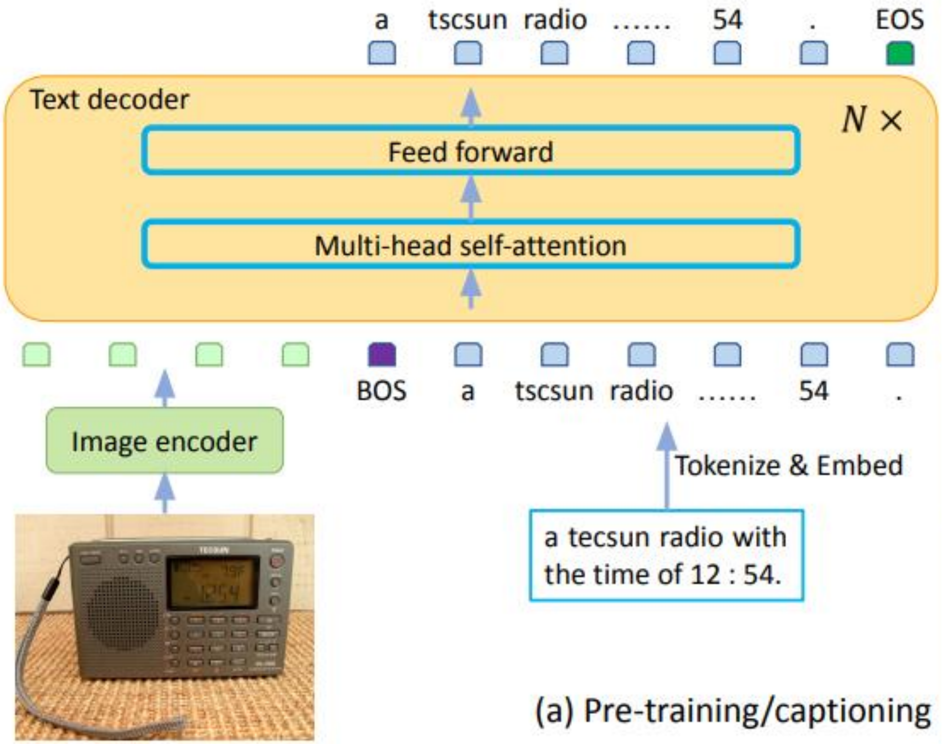
Alayrac, Donahue, Luc, Miech, Barr, Hasson, Lenc, Mensch, Millican, Reynolds, Ring, Rutherford, Cabi, Han, Gong, Samangooei, Monteiro, Menick, Borgeaud, Brock, Nematzadeh, Sharifzadeh, Binkowski, Barreira, Vinyals, Zisserman, Simonyan, 2022/04/29, https://arxiv.org/abs/2204.14198

CoCa: Contrastive Captioners are Image-Text Foundation Models



J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, Y. Wu, 2022/05/04, https://arxiv.org/abs/2205.01917

GIT: A Generative Image-to-text Transformer for Vision and Language



(a) Pre-training/captioning

J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, L. Wang, 2022/05/27, https://arxiv.org/abs/2205.14100

☑ Select/Unselect All

☑ a page that says probability theory is a central field of mathematics, widely applicable to scientific, technological, and human situations involving uncertainty.

Probability theory is a central field of mathematics, widely applicable to scientific, technological, and human situations involving uncertainty. The most obvious applications are to situations, such as games of chance

Dawn of LMMS

2023---

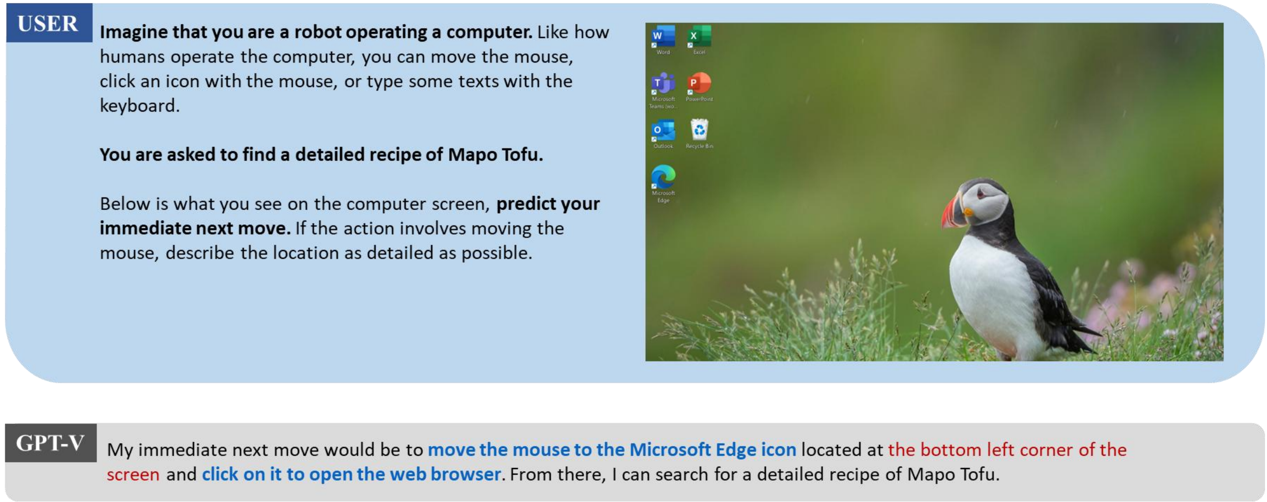
GPT-4V

September 2023

Dawn of LMMs

- GPT-4V(vision):
  - GPT-4V(ision) system card
    - https://openai.com/research/gpt-4v-system-card
  - GPT-4V(ision) technical work and authors
    - GPT-4V(ision) technical work and authors (openai.com)
  - The Dawn of LMMs: preliminary explorations with GPT-4V(ision)
    - https://arxiv.org/abs/2309.17421

GUI-Navigation


USER

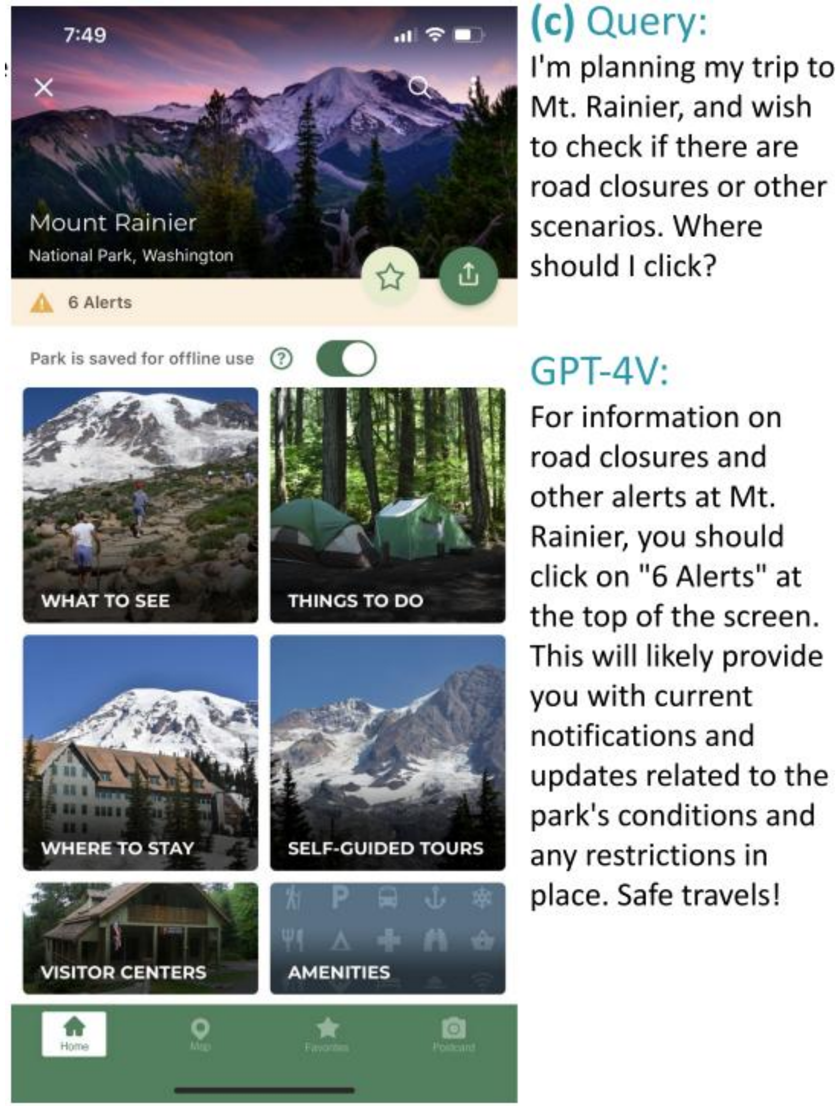
Imagine that you are a robot operating a computer. Like how humans operate the computer, you can move the mouse, click an icon with the mouse, or type some texts with the keyboard.

You are asked to find a detailed recipe of Mapo Tofu.

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

GPT-V

My immediate next move would be to **move the mouse to the Microsoft Edge icon** located at the bottom left corner of the screen and **click on it to open the web browser**. From there, I can search for a detailed recipe of Mapo Tofu.



(c) Query:

I'm planning my trip to Mt. Rainier, and wish to check if there are road closures or other scenarios. Where should I click?

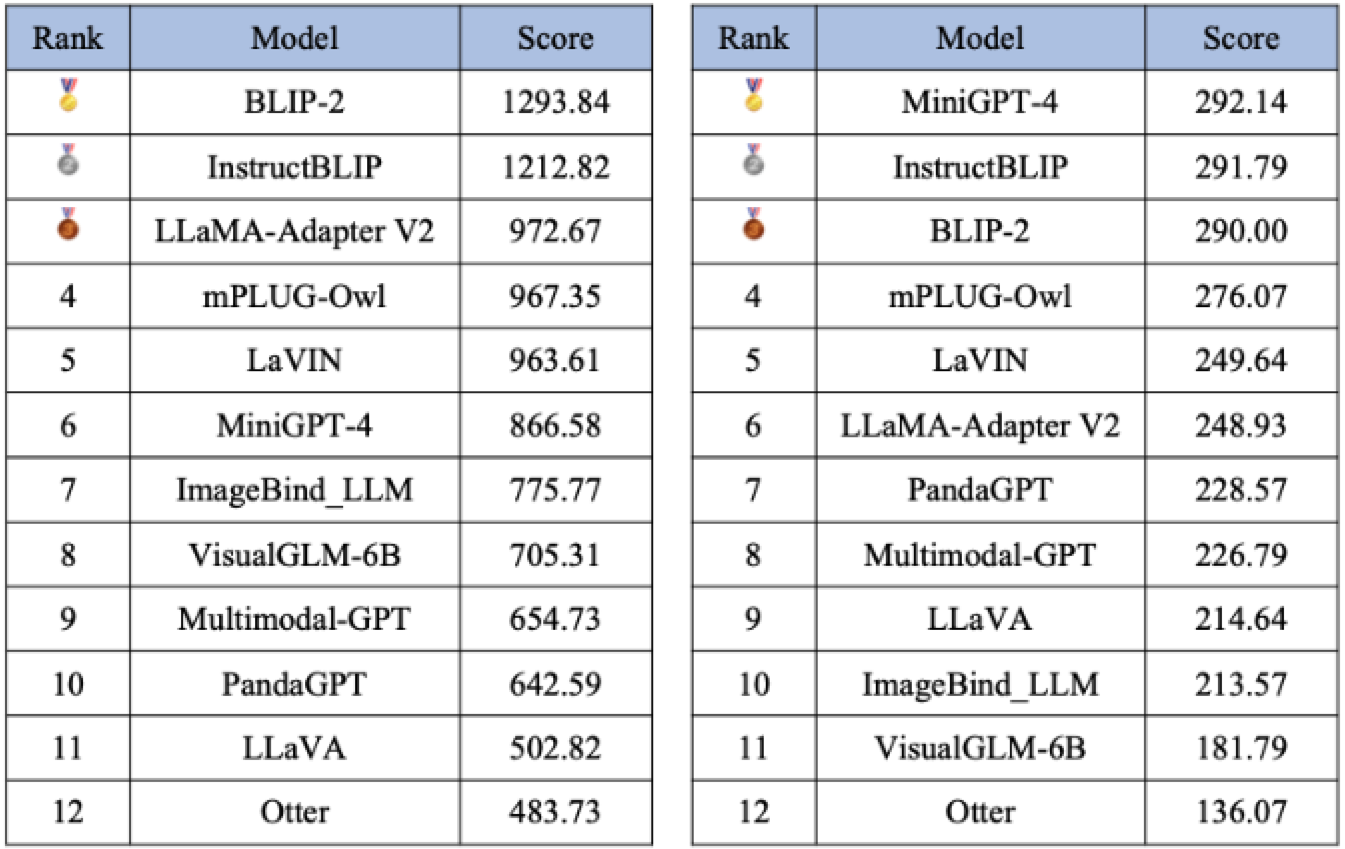
GPT-4V:

For information on road closures and other alerts at Mt. Rainier, you should click on "6 Alerts" at the top of the screen. This will likely provide you with current notifications and updates related to the park's conditions and any restrictions in place. Safe travels!

A. Yan, Z. Yang, W. Zhu, K. Lin, L, Li, J. Wang, J. Yang, Y.Zhong, J.McAuley, J. Gao, Z. Liu, L, Wang, GPT-4V in Wonderland: Large Multimodal Models for Zero-Shot Smartphone GUI Navigation,2023, https://arxiv.org/pdf/2311.07562.pdf

Open Source LMMs (December, 2023)

- BLIP2
- InstructBLIP
- MiniGPT-4
- LLaVA
- …



Rank	Model	Score
🥇	BLIP-2	1293.84
🥈	InstructBLIP	1212.82
🥉	LLaMA-Adapter V2	972.67
4	mPLUG-Owl	967.35
5	LaVIN	963.61
6	MiniGPT-4	866.58
7	ImageBind_LLM	775.77
8	VisualGLM-6B	705.31
9	Multimodal-GPT	654.73
10	PandaGPT	642.59
11	LLaVA	502.82
12	Otter	483.73

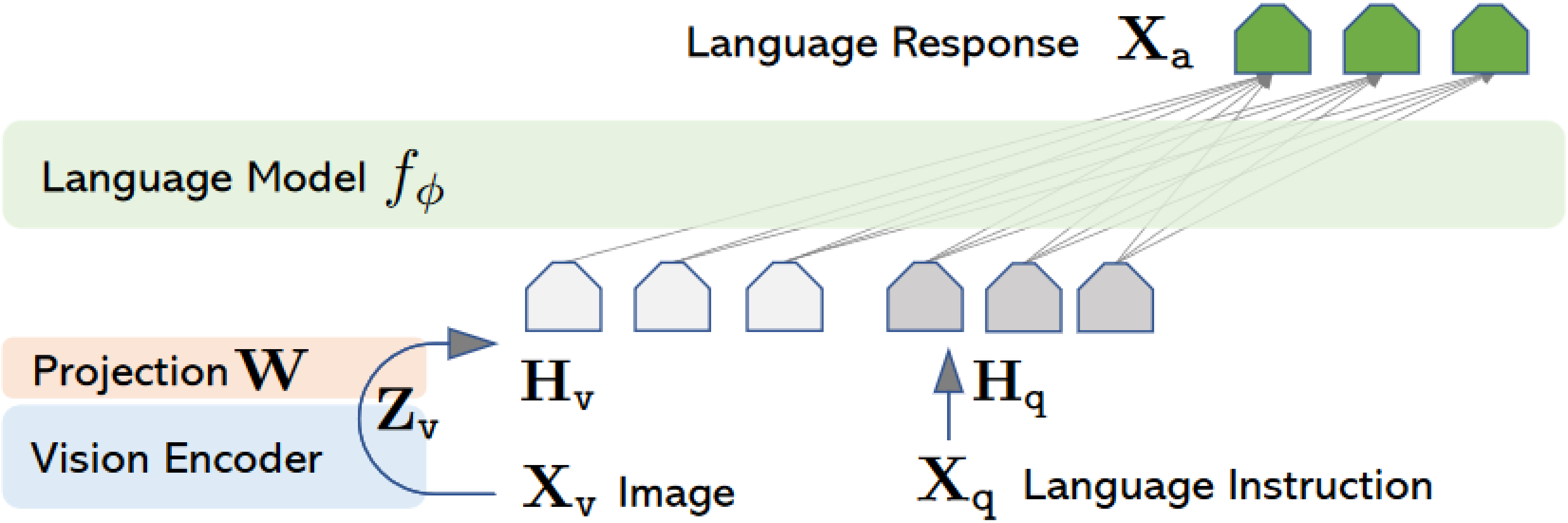
(1) **Perception**

Rank	Model	Score
🥇	MiniGPT-4	292.14
🥈	InstructBLIP	291.79
🥉	BLIP-2	290.00
4	mPLUG-Owl	276.07
5	LaVIN	249.64
6	LLaMA-Adapter V2	248.93
7	PandaGPT	228.57
8	Multimodal-GPT	226.79
9	LLaVA	214.64
10	ImageBind_LLM	213.57
11	VisualGLM-6B	181.79
12	Otter	136.07

(2) **Cognition**

C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, Z. Qiu, W. Lin, J. Yang, X. Zheng, K. Li, X. Sun, R. Ji,
MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models, https://arxiv.org/pdf/2306.13394.pdf

LLaVA: Visual Instruction Tuning



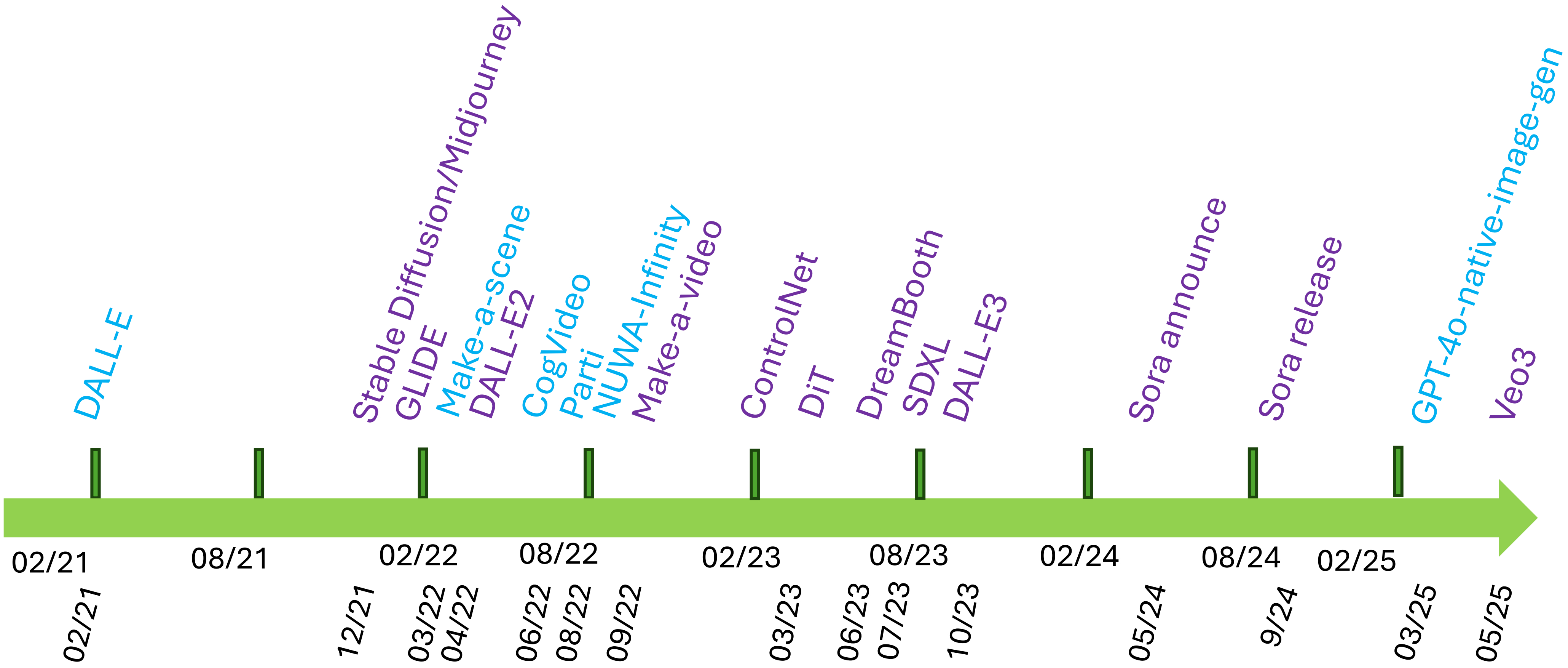
H. Liu, C. Li, Q. Wu, Y. Lee, Visual Instruction Tuning, https://arxiv.org/pdf/2304.08485.pdf, 2023/04/17

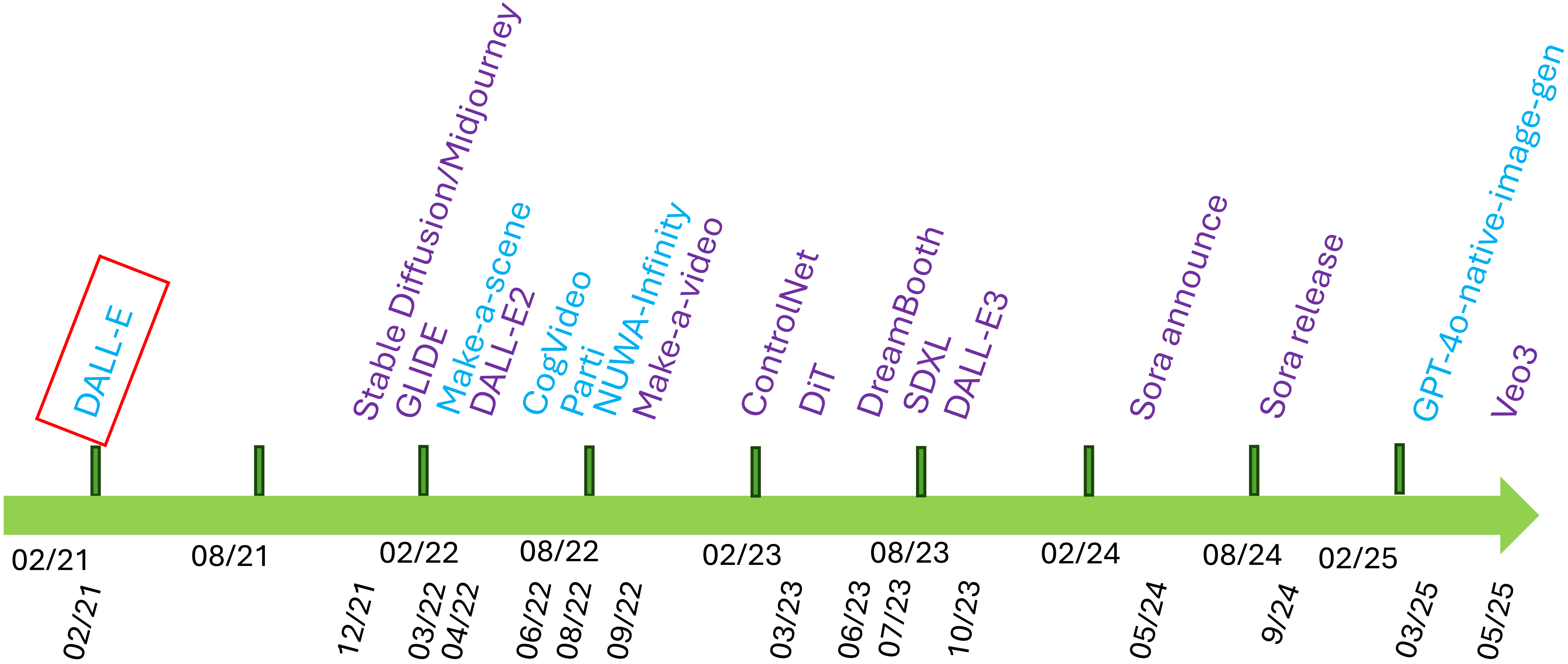
Looking Back

- Convergence of vision-language architecture
  - Vision encoder + language decoder
  - It was not obvious
  - Is it optimal?
- Benefits of leveraging language models for vision tasks
  - Open vocabulary
  - Task unification
  - Generalization

Outline

- A brief history of multimodality models
  - Understanding
  - Generation
  - Unification
  - Agentic
- Special topics
  - Does image help with reasoning?
  - Token compression
- Future directions
  - 3D, embodied agents, robotics


DALL-E
Stable Diffusion/Midjourney
GLIDE
Make-a-scene
DALL-E2
CogVideo
Parti
NUWA-Infinity
Make-a-video
ControlNet
DiT
DreamBooth
SDXL
DALL-E3
Sora announce
Sora release
GPT-4o-native-image-gen
Veo3
02/21
08/21
02/22
08/22
02/23
08/23
02/24
08/24
02/25
02/21
12/21
03/22
04/22
06/22
08/22
09/22
03/23
06/23
07/23
10/23
05/24
9/24
03/25
05/25

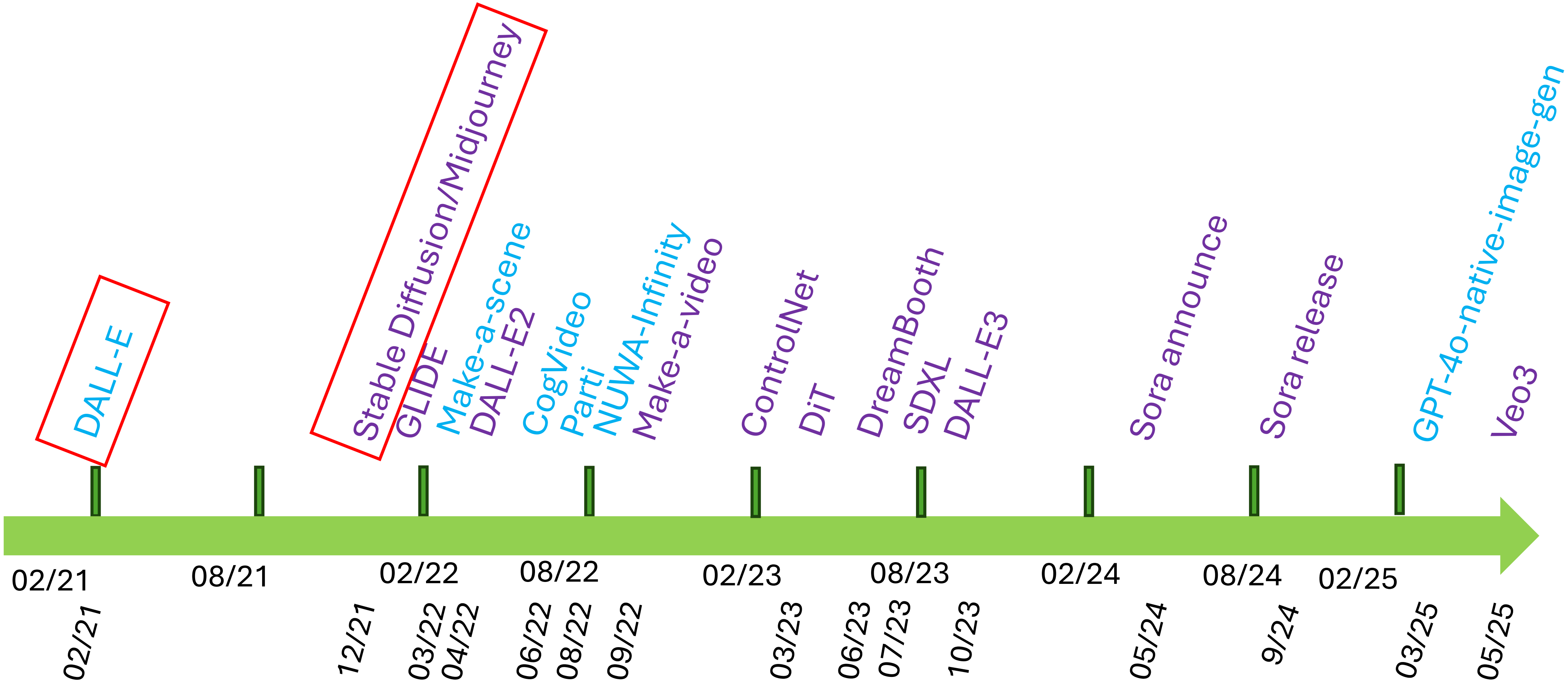


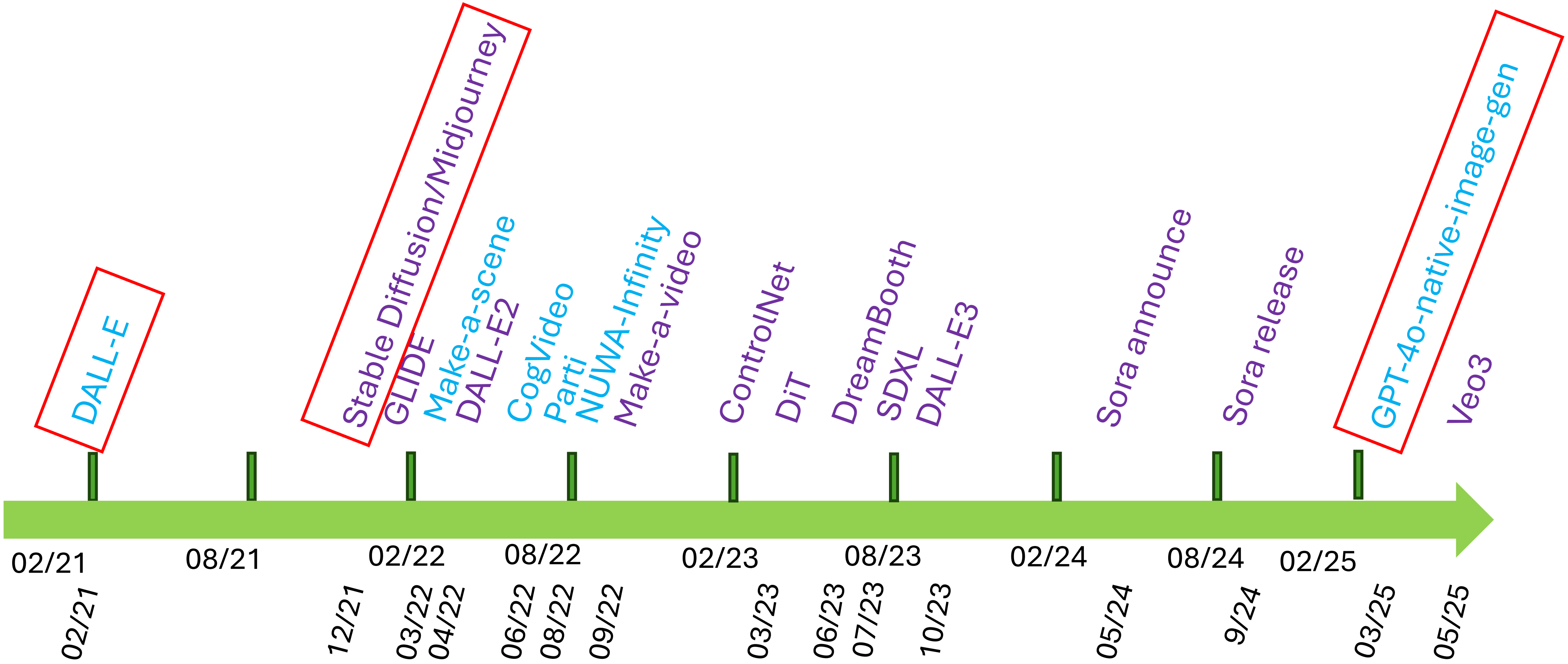
DALL-E

Stable Diffusion/Midjourney
GLIDE
Make-a-scene
DALL-E2
CogVideo
Parti
NUWA-Infinity
Make-a-video
ControlNet
DiT
DreamBooth
SDXL
DALL-E3
Sora announce
Sora release
GPT-4o-native-image-gen
Veo3

02/21 08/21 02/22 08/22 02/23 08/23 02/24 08/24 02/25

02/21 12/21 03/22 04/22 06/22 08/22 09/22 03/23 06/23 07/23 10/23 05/24 9/24 03/25 05/25


DALL-E
Stable Diffusion/Midjourney
GLIDE
Make-a-scene
DALL-E2
CogVideo
Parti
NUWA-Infinity
Make-a-video
ControlNet
DiT
DreamBooth
SDXL
DALL-E3
Sora announce
Sora release
GPT-4o-native-image-gen
Veo3
02/21
08/21
02/22
08/22
02/23
08/23
02/24
08/24
02/25
02/21
12/21
03/22
04/22
06/22
08/22
09/22
03/23
06/23
07/23
10/23
05/24
9/24
03/25
05/25



DALL-E

Stable Diffusion/Midjourney

GLIDE

Make-a-scene

DALL-E2

CogVideo

Parti

NUWA-Infinity

Make-a-video

ControlNet

DiT

DreamBooth

SDXL

DALL-E3

Sora announce

Sora release

GPT-4o-native-image-gen

Veo3

02/21 08/21 02/22 08/22 02/23 08/23 02/24 08/24 02/25

02/21 12/21 03/22 04/22 06/22 08/22 09/22 03/23 06/23 07/23 10/23 05/24 9/24 03/25 05/25

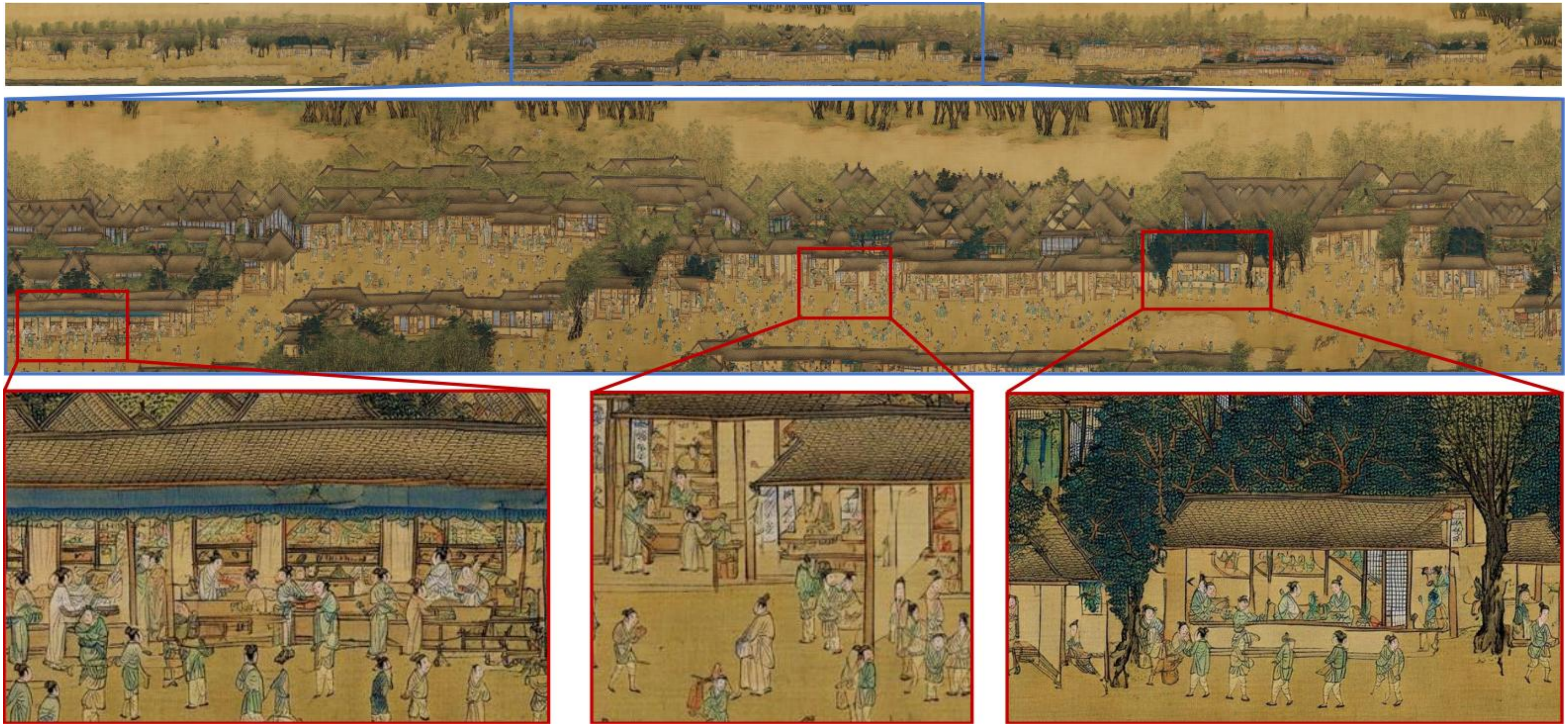
Diffusion vs Autoregressive

	Aesthetics	**Spatial relation**	**Text rendering**	**Semantic alignment**	**Speed**
DALL-E3	●				●
GPT-4o		●	●	●	

Diffusion vs Autoregressive

- Diffusion
  - Better aesthetic quality
  - Faster especially with few-stop distillation
- Autoregressive
  - Leveraging pretrained LLM
    - Better semantic alignment
    - Text generation
  - Flexible in image size and video length expansion

AR: Flexibility in image size expansion 38912 x 2048

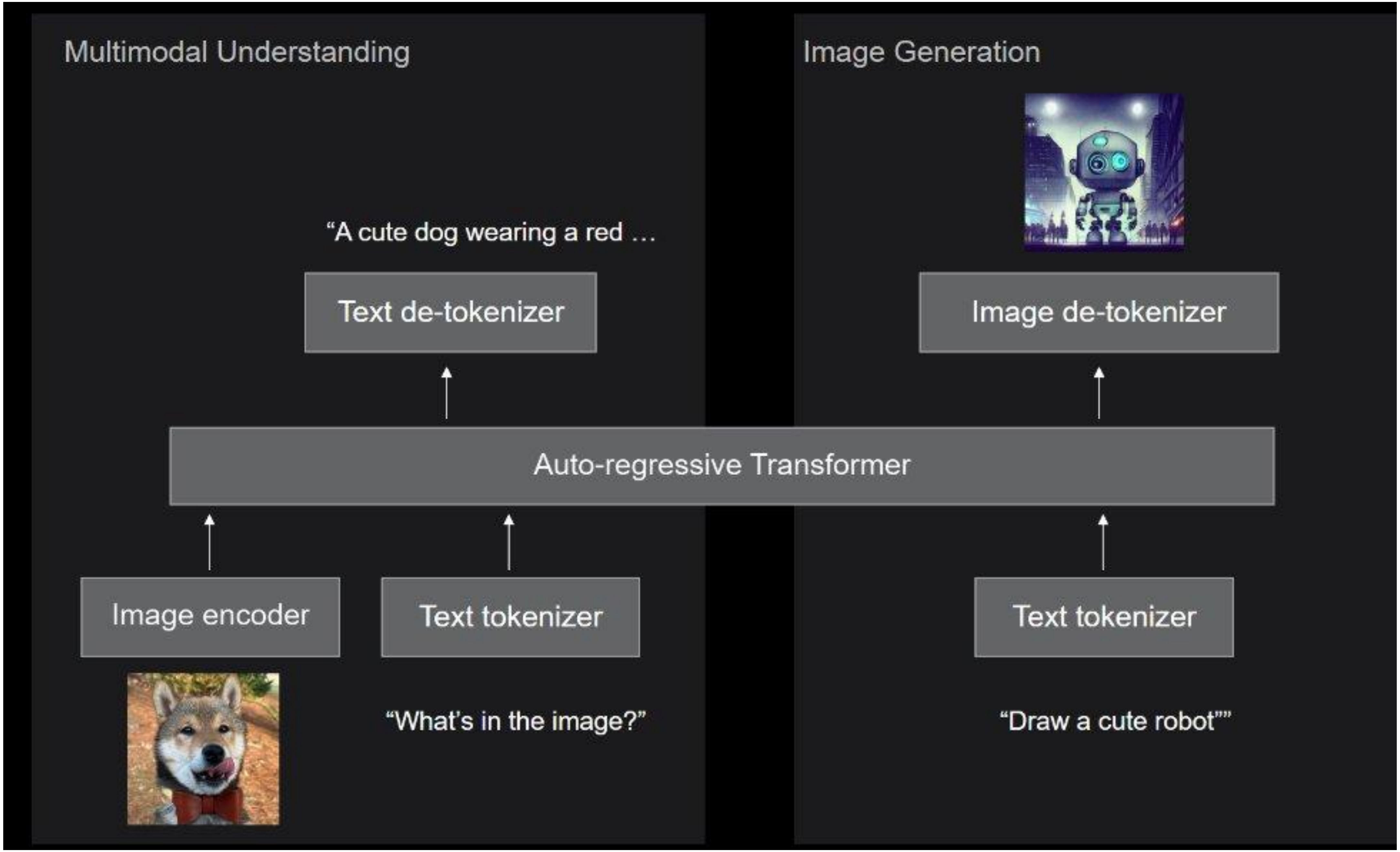


“Along the River During the Qingming Festival”

NUWA-Infinity: Autoregressive over Autoregressive Generation for Infinite Visual Synthesis Wu et al, arXiv 2207.09814, Oct. 2022

Outline

- A brief history of multimodality models
  - Understanding
  - Generation
  - Unification
  - Agentic
- Special topics
  - Does image help with reasoning?
  - Token compression
- Summary and future directions



Multimodal Understanding

“A cute dog wearing a red …

Text de-tokenizer

Image encoder

Text tokenizer

“What's in the image?”

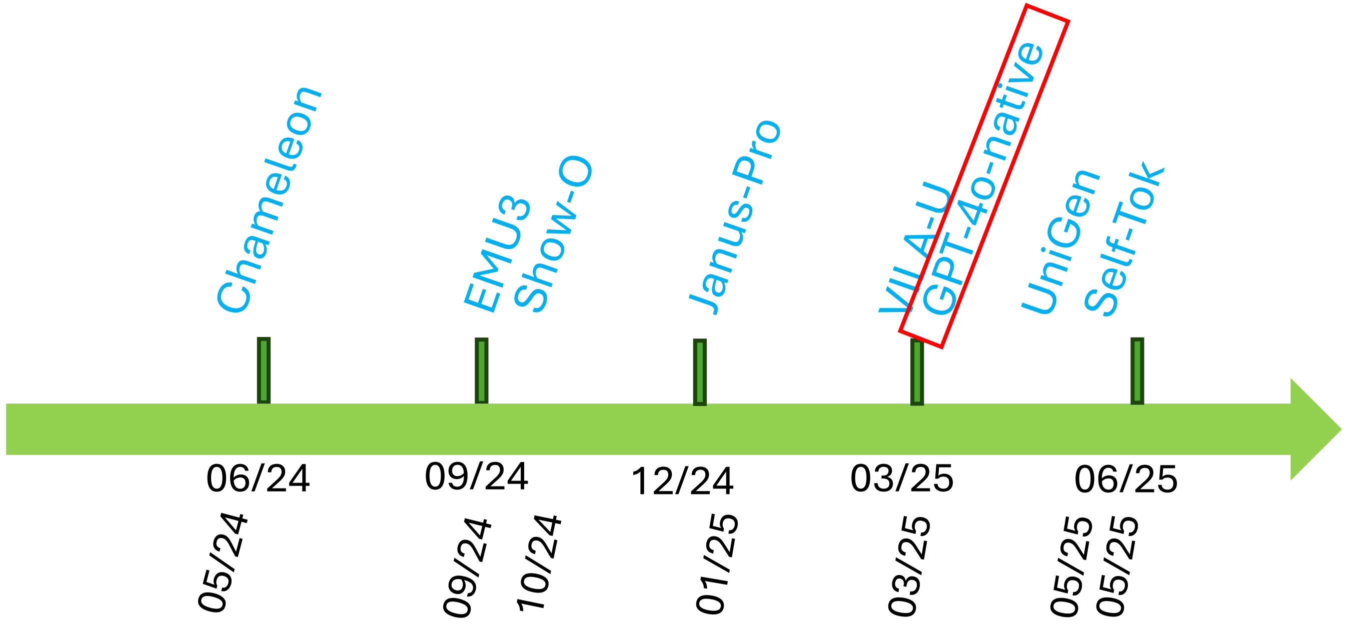
Auto-regressive Transformer

Image Generation

Image de-tokenizer

Text tokenizer

“Draw a cute robot””

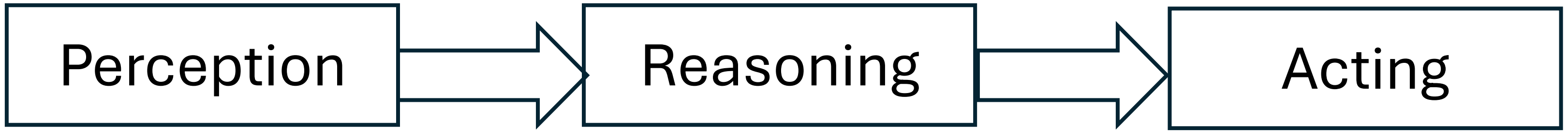


- All models are autoregressive except Show-O
- Show-O: discrete diffusion
- Janus-Pro and UniGen use separate tokenizers: better accuracy in understanding tasks
- GPT-4o proves that unified model can excel in both understanding and generation
- Need more and performant open-source unified models

Outline

- A brief history of multimodality models
    - Understanding
    - Generation
    - Unification
    - Agentic
- Special topics
    - Does image help with reasoning?
    - Token compression
- Summary and future directions

Agentic model goes beyond perception



Perception → Reasoning → Acting

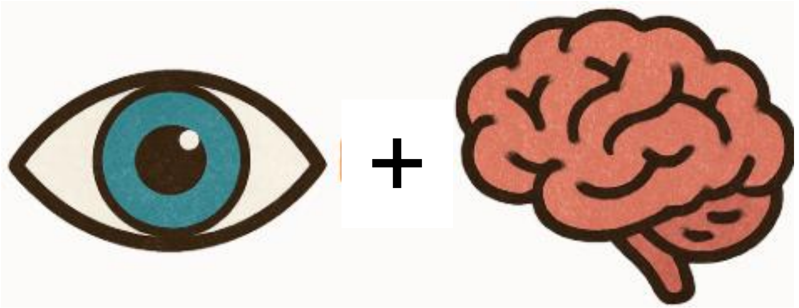
Two examples of agentic multimodality models

GUI Navigation

Tool-Use

GUI Navigation: Computer-Using Agent

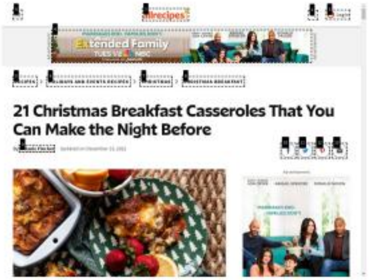
- OpenAI Operator
  - Vision + reasoning
  - Multi-turn RL



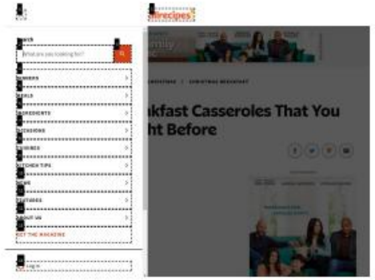
- Benchmarks
  - OSWorld
  - WebArena
  - WebVoyager
  - MageBench
  - …

GUI Navigation: Computer-Using Agent

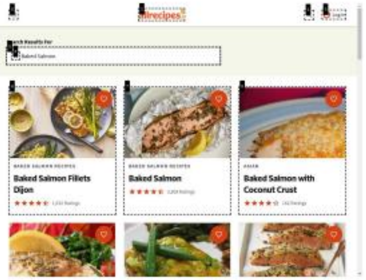
“Find a recipe for Baked Salmon that takes less than 30 minutes to prepare and has at least a 4-star rating based on user reviews.”


21 Christmas Breakfast Casseroles That You Can Make the Night Before

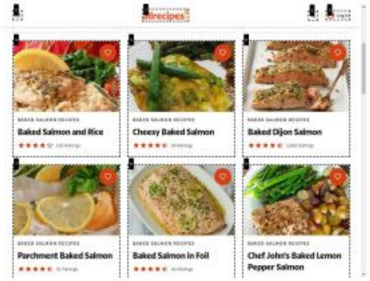
Step 1: Click [2]



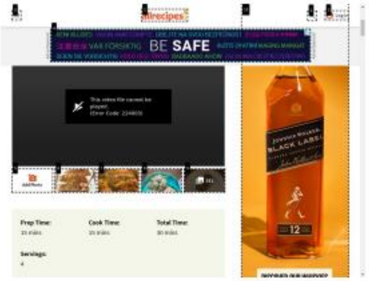
Step 2: Type [2]; Baked Salmon



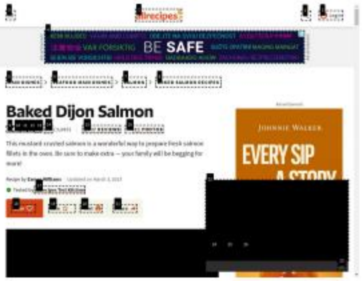
Step 3: Scroll down



Step 4: Click [6]



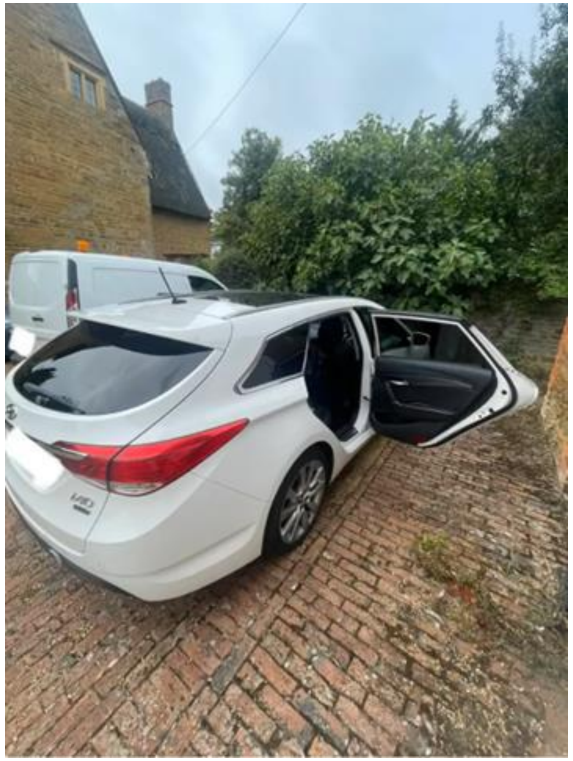
Step 5: Scroll down



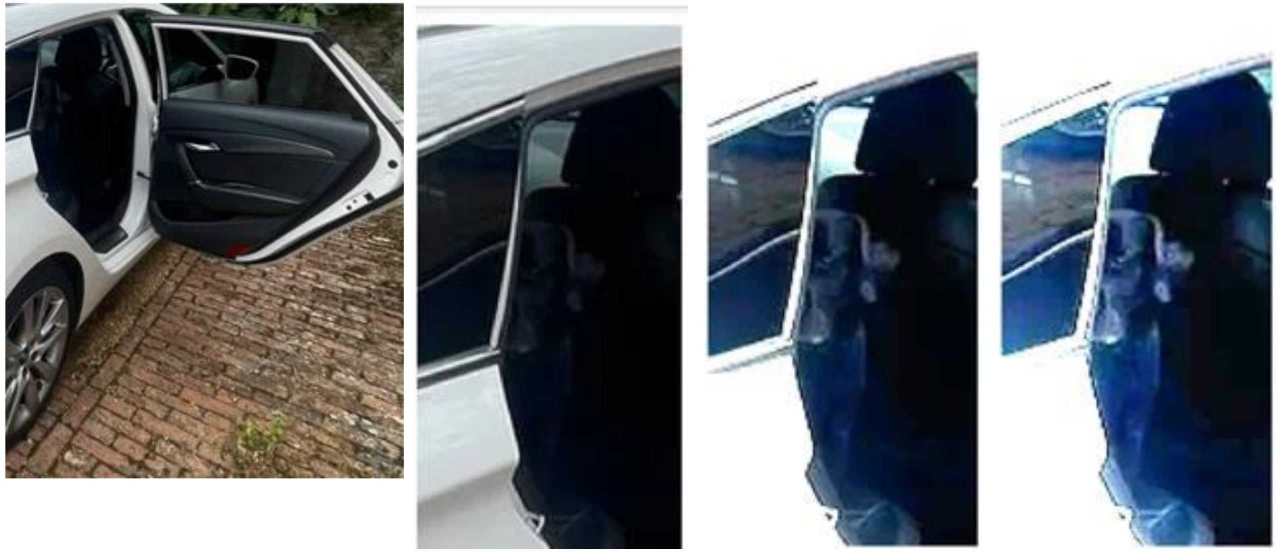
Step 6: ANSWER

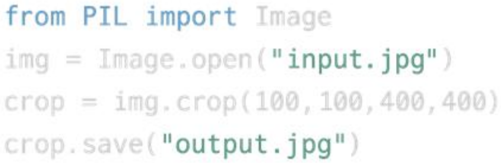
WebVoyager: Building an End-to-End Web Agent with Large Multimodal Models, He et al, 2024

Vision Tool-Use: O3/O3-Pro



What color is the dog?





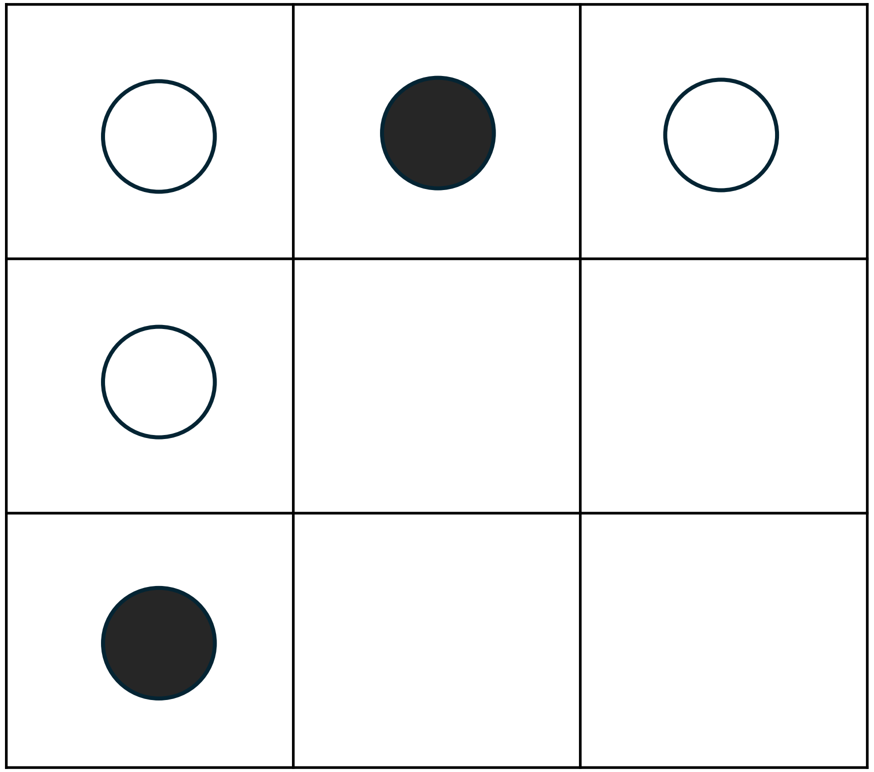
```
from PIL import Image
img = Image.open("input.jpg")
crop = img.crop(100,100,400,400)
crop.save("output.jpg")
```

Outline

- A brief history of multimodality models
    - Understanding
    - Generation
    - Unification
    - Agentic
- Special topics
    - Does image help with reasoning?
    - Token compression
- Summary and future directions

Does Image Help with Reasoning?

- Tic-Tac-Toe game



Where should black play next?

Version 1: Text only

Alice and Bob are playing a game on a 3x3 grid. The points on the grid are labeled top to bottom, left to right, as A,B,C,D,E,F,G,H,I. Alice plays white. Bob plays black. At each turn, the player places a stone of the corresponding color onto one of the positions that have not been occupied. Whoever has three stones in a line (horizontal, vertical, or diagonal) wins.

Alice first places a white stone at A. Bob places a black stone at B. Then Alice at C. Then Bob at G. Alice D.

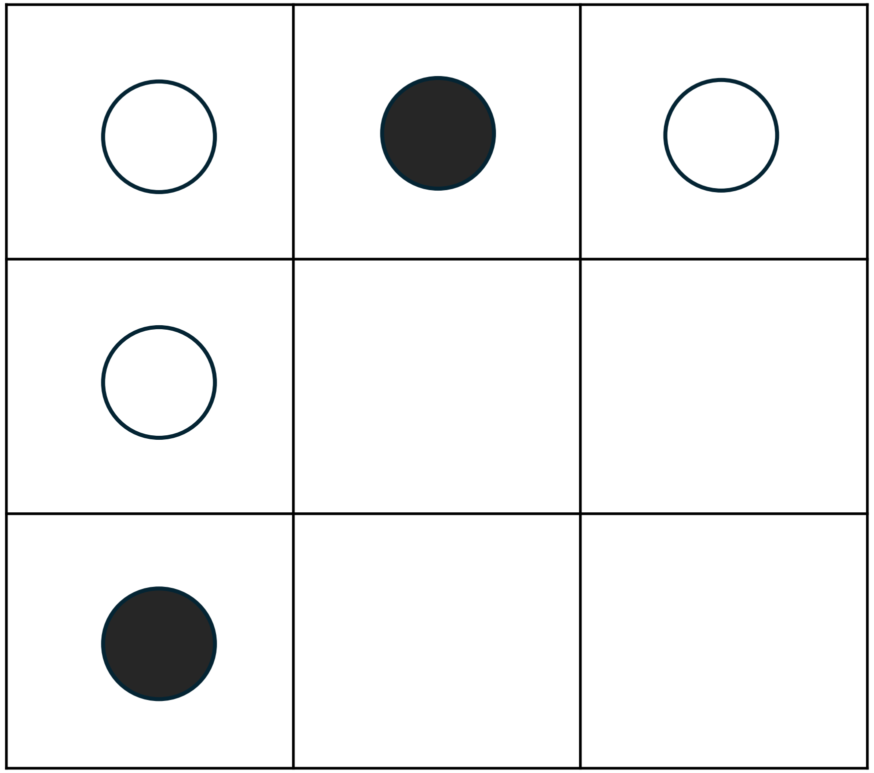
Where should Bob play next?

Version 2: Text + Image

Alice and Bob are playing a game on a 3x3 grid. The points on the grid are labeled top to bottom, left to right, as A,B,C,D,E,F,G,H,I. Alice plays white. Bob plays black. At each turn, the player places a stone of the corresponding color onto one of the positions that have not been occupied. Whoever has three stones in a line (horizontal, vertical, or diagonal) wins.

Alice first places a white stone at A. Bob places a black stone at B. Then Alice at C. Then Bob at G. Alice D.

Where should Bob play next?

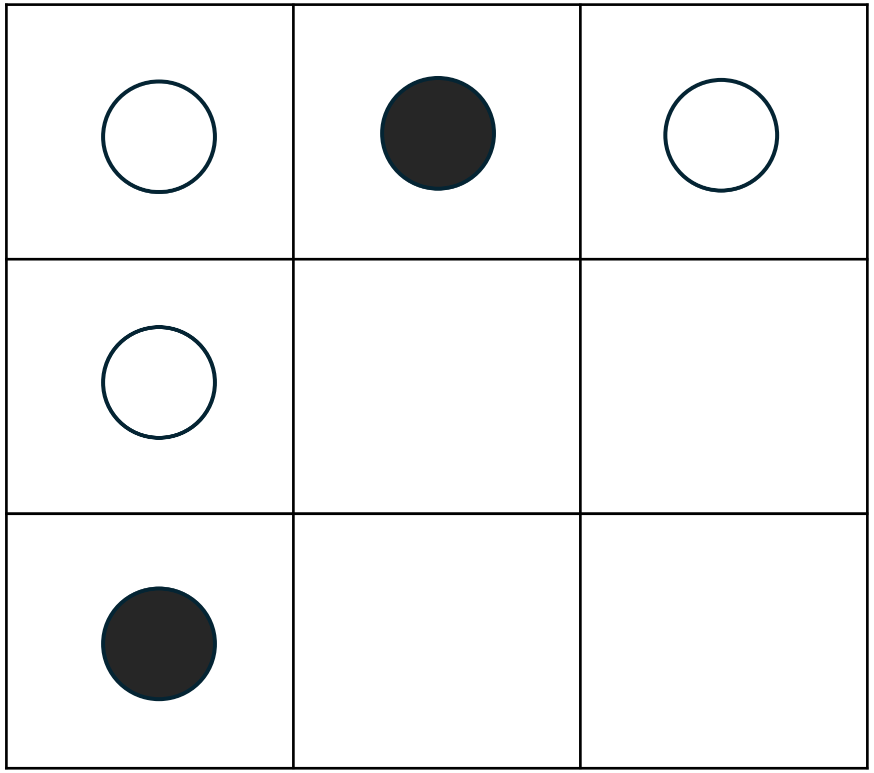


Version 2: Image Only

Alice and Bob are playing a game on a 3x3 grid. The points on the grid are labeled top to bottom, left to right, as A,B,C,D,E,F,G,H,I. Alice plays white. Bob plays black. At each turn, the player places a stone of the corresponding color onto one of the positions that have not been occupied. Whoever has three stones in a line (horizontal, vertical, or diagonal) wins.

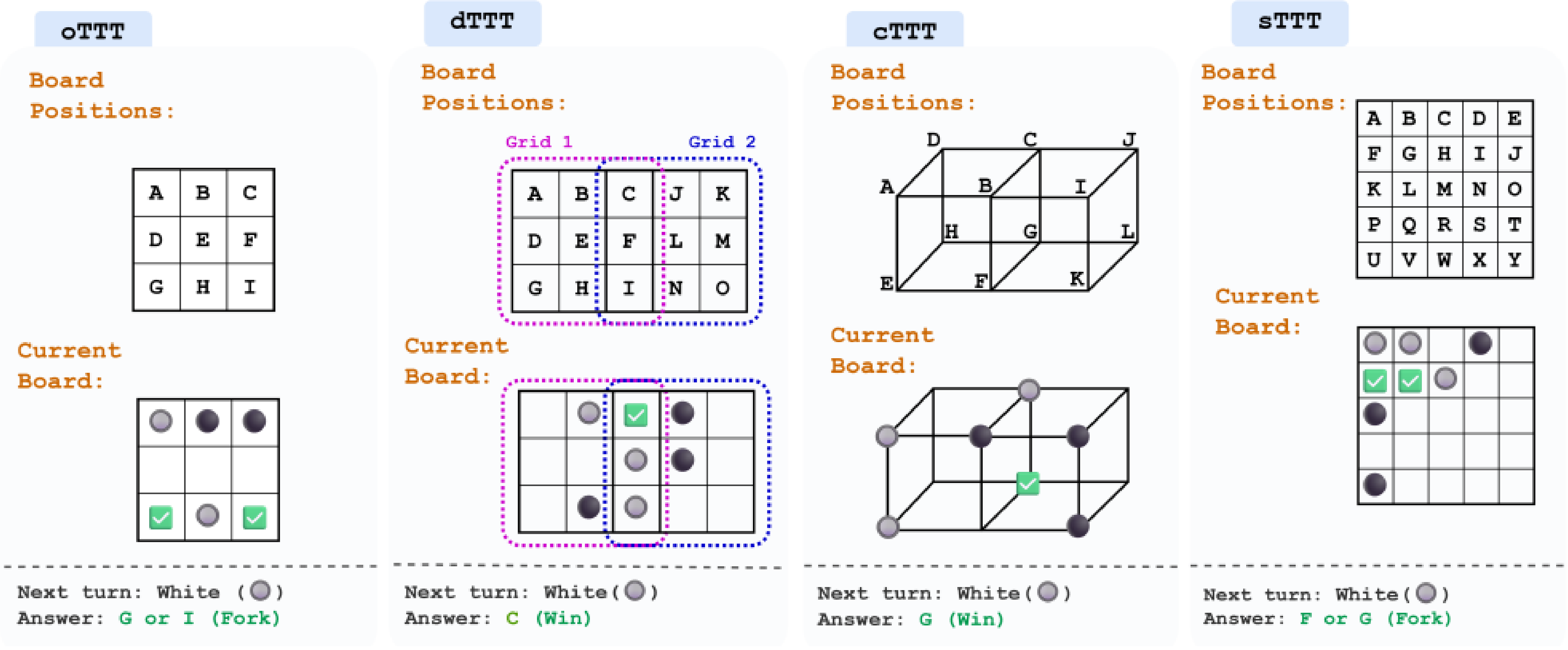
~~Alice first places a white stone at A. Bob places a black stone at B. Then Alice at C. Then Bob at G. Alice D.~~

Where should Bob play next?



○	●	○
○		
●		

Tic-Tac-Toe Extensions (not seen by Frontier models)



oTTT: Next turn: White (○) Answer: G or I (Fork)

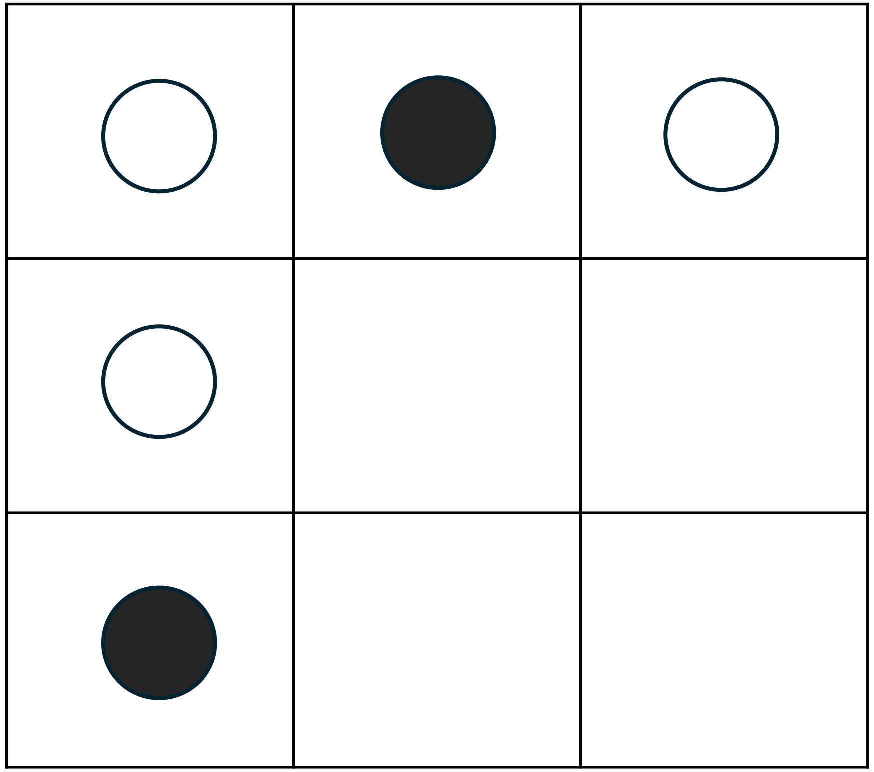
dTTT: Next turn: White(○) Answer: C (Win)

cTTT: Next turn: White(○) Answer: G (Win)

sTTT: Next turn: White(○) Answer: F or G (Fork)

TTT-Bench: A Benchmark for Evaluating Reasoning Ability with Simple and Novel Tic-Tac-Toe-style Games, arXiv.2506.10209

Visual Is Helpful for Human



A	B	C
D	E	F
G	H	I

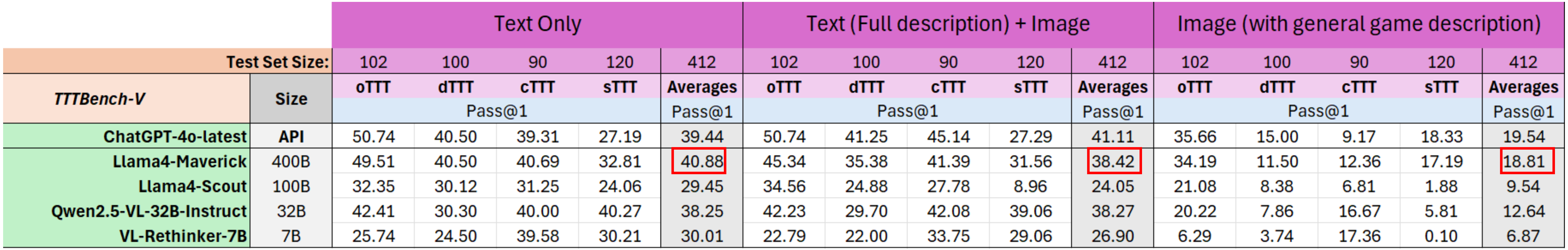
Is Visual Helpful for LLMs?

Text > Text + Image > Image

		Text Only					Text (Full description) + Image					Image (with general game description)				
Test Set Size:		102	100	90	120	412	102	100	90	120	412	102	100	90	120	412
TTTBench-V	**Size**	**oTTT**	**dTTT**	**cTTT**	**sTTT**	**Averages**	**oTTT**	**dTTT**	**cTTT**	**sTTT**	**Averages**	**oTTT**	**dTTT**	**cTTT**	**sTTT**	**Averages**
		Pass@1				Pass@1	Pass@1				Pass@1	Pass@1				Pass@1
ChatGPT-4o-latest	**API**	50.74	40.50	39.31	27.19	39.44	50.74	41.25	45.14	27.29	41.11	35.66	15.00	9.17	18.33	19.54
Llama4-Maverick	400B	49.51	40.50	40.69	32.81	40.88	45.34	35.38	41.39	31.56	38.42	34.19	11.50	12.36	17.19	18.81
Llama4-Scout	100B	32.35	30.12	31.25	24.06	29.45	34.56	24.88	27.78	8.96	24.05	21.08	8.38	6.81	1.88	9.54
Qwen2.5-VL-32B-Instruct	32B	42.41	30.30	40.00	40.27	38.25	42.23	29.70	42.08	39.06	38.27	20.22	7.86	16.67	5.81	12.64
VL-Rethinker-7B	7B	25.74	24.50	39.58	30.21	30.01	22.79	22.00	33.75	29.06	26.90	6.29	3.74	17.36	0.10	6.87

Is Visual Helpful for LLMs?

Text > Text + Image > Image



		Text Only					Text (Full description) + Image					Image (with general game description)				
Test Set Size:		102	100	90	120	412	102	100	90	120	412	102	100	90	120	412
TTTBench-V	**Size**	**oTTT**	**dTTT**	**cTTT**	**sTTT**	**Averages**	**oTTT**	**dTTT**	**cTTT**	**sTTT**	**Averages**	**oTTT**	**dTTT**	**cTTT**	**sTTT**	**Averages**
		Pass@1				Pass@1	Pass@1				Pass@1	Pass@1				Pass@1
ChatGPT-4o-latest	**API**	50.74	40.50	39.31	27.19	39.44	50.74	41.25	45.14	27.29	41.11	35.66	15.00	9.17	18.33	19.54
Llama4-Maverick	400B	49.51	40.50	40.69	32.81	40.88	45.34	35.38	41.39	31.56	38.42	34.19	11.50	12.36	17.19	18.81
Llama4-Scout	100B	32.35	30.12	31.25	24.06	29.45	34.56	24.88	27.78	8.96	24.05	21.08	8.38	6.81	1.88	9.54
Qwen2.5-VL-32B-Instruct	32B	42.41	30.30	40.00	40.27	38.25	42.23	29.70	42.08	39.06	38.27	20.22	7.86	16.67	5.81	12.64
VL-Rethinker-7B	7B	25.74	24.50	39.58	30.21	30.01	22.79	22.00	33.75	29.06	26.90	6.29	3.74	17.36	0.10	6.87

Outline

- A brief history of multimodality models
  - Understanding
  - Generation
  - Unification
  - Agentic
- Special topics
  - Does image help with reasoning?
  - Token compression
- Summary and future directions

Token Compression

- Images are represented as tokens
- Too many tokens

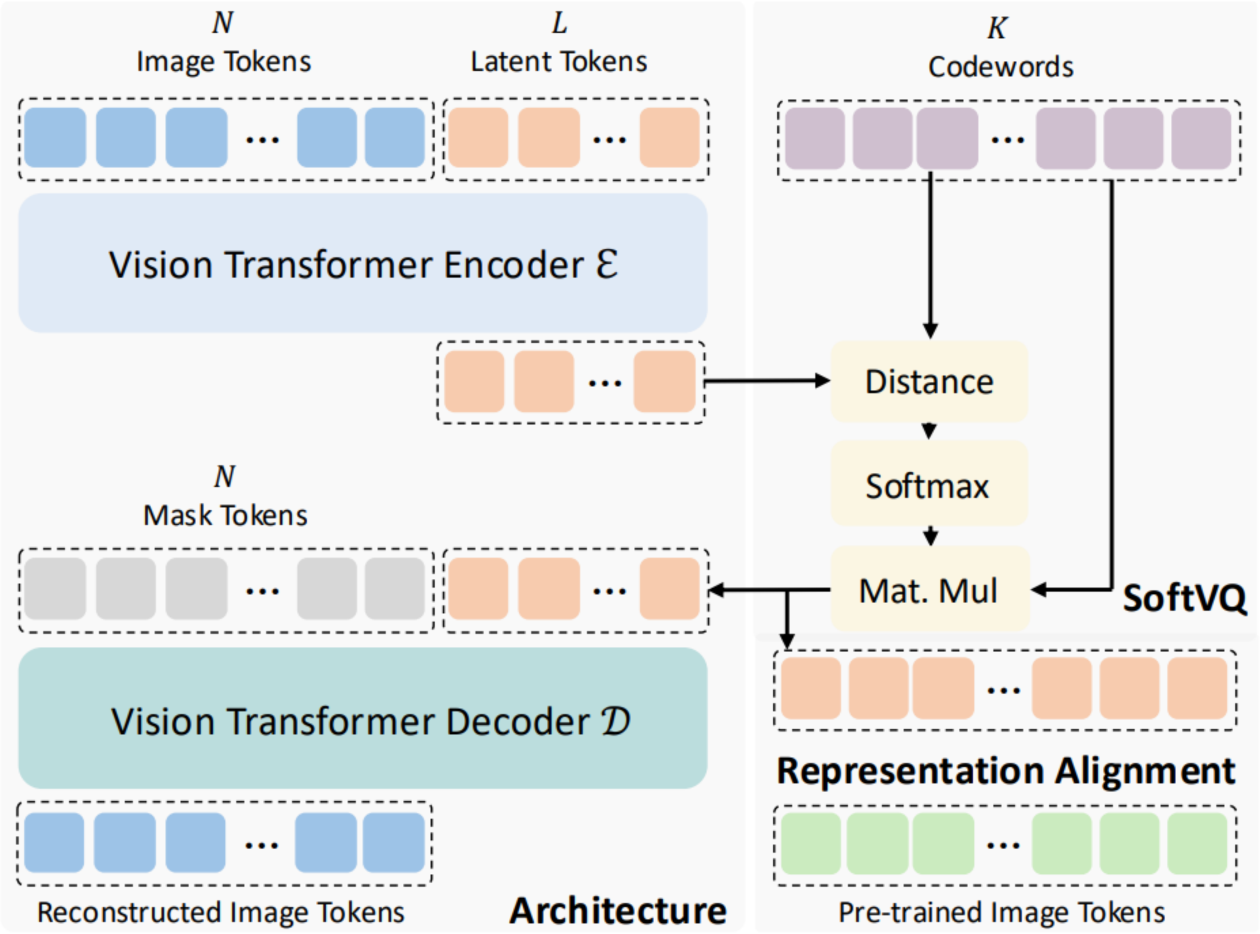
$$w \times h \longrightarrow \frac{w}{f} \times \frac{h}{f} \quad \text{tokens}$$

$$f = 8 \ \ or\ 16$$

EMU3: $1024 \times 1024 \longrightarrow \frac{1024}{16} \times \frac{1024}{16} = 4096$ tokens

1D Tokenization

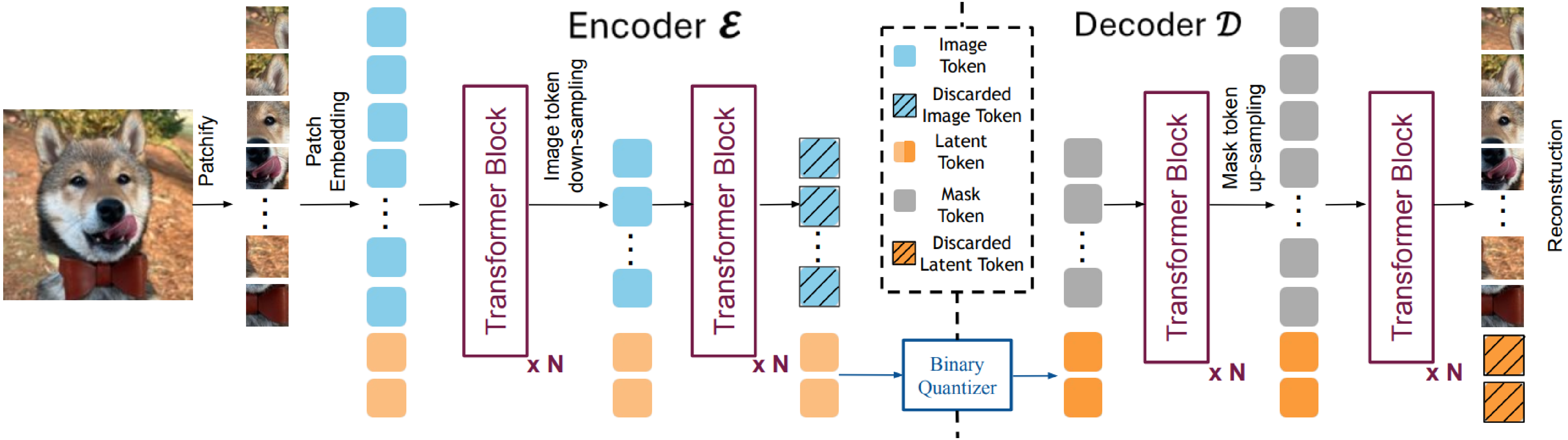
- 128 tokens for 1024x1024 image vs 4096 tokens in EMU3
- **32x tokens reduction**



SoftVQ-VAE: Efficient 1-Dimensional Continuous Tokenizer, Chen et al, CVPR 2025
Masked autoencoders are effective tokenizers for diffusion models. Chen et al, ICML 2025

1D Binary Image Tokenizer (1D BiT)

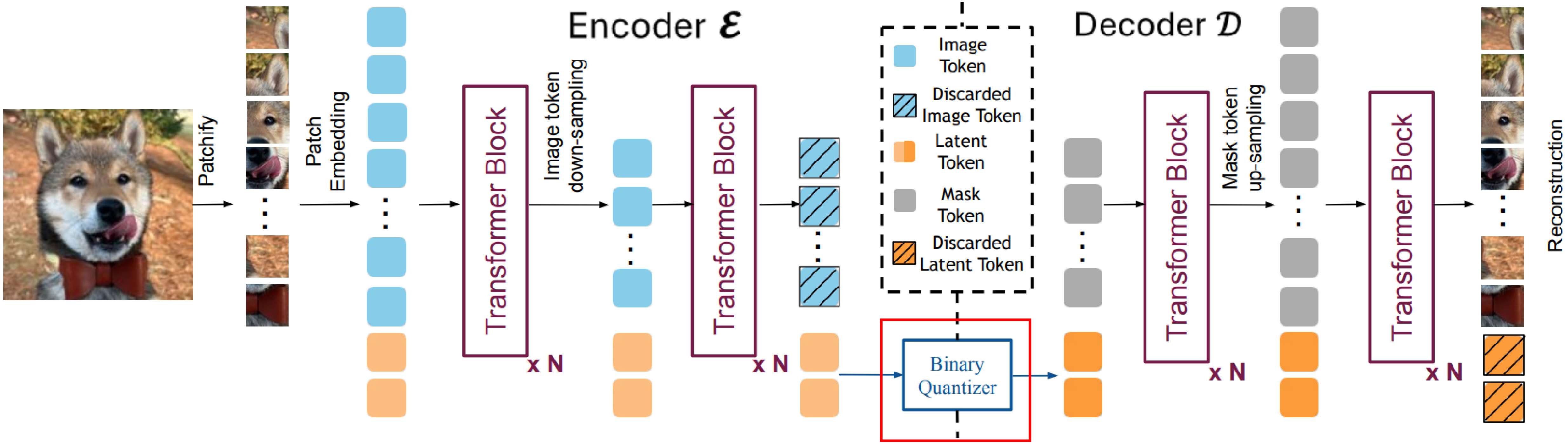
- Enabling both diffusion and Autoregressive image generation



Instella-T2I: Pushing the Limits of 1D Discrete Latent Space Image Generation, Wang et al, 2025 https://arxiv.org/abs/2506.21022

1D Binary Image Tokenizer (1D BiT)

- Enabling both diffusion and Autoregressive image generation



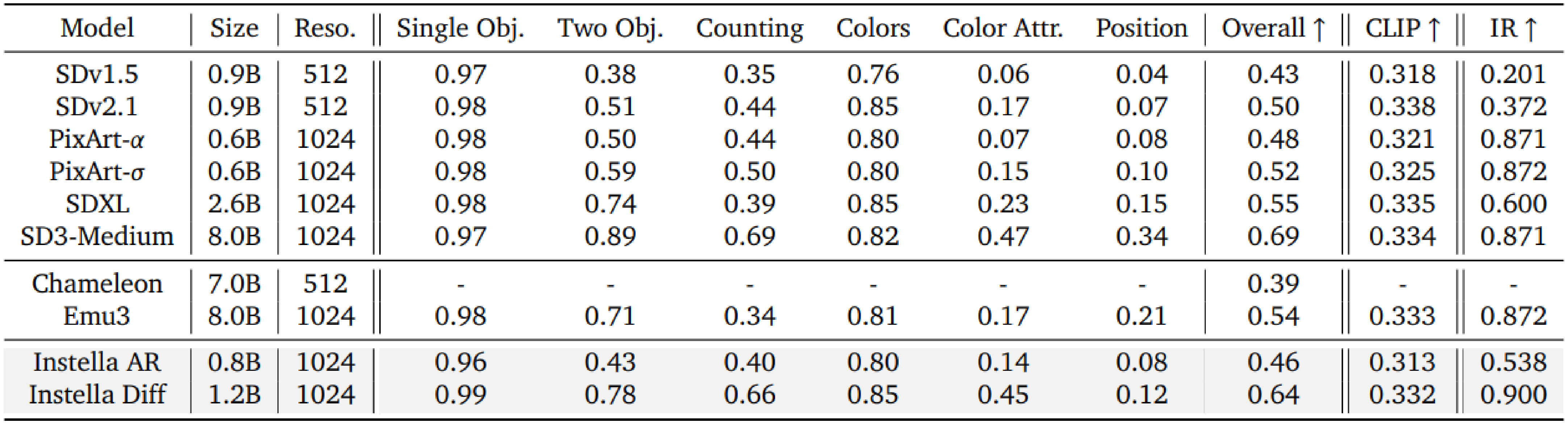
Instella-T2I: Pushing the Limits of 1D Discrete Latent Space Image Generation, Wang et al, 2025 https://arxiv.org/abs/2506.21022

T2I using 1D BiT (Instella-T2I)

Instella-T2I: Pushing the Limits of 1D Discrete Latent Space Image Generation, Wang et al, 2025 https://arxiv.org/abs/2506.21022

1024x1024 T2I Evaluation

- 128 tokens for both Instella-AR and Instella-Diff
- 32 times token reduction compared to EMU3
- Fully open public datasets for training
- Instella-Diff: competitive against SOTA models like SDXL that use in-house data



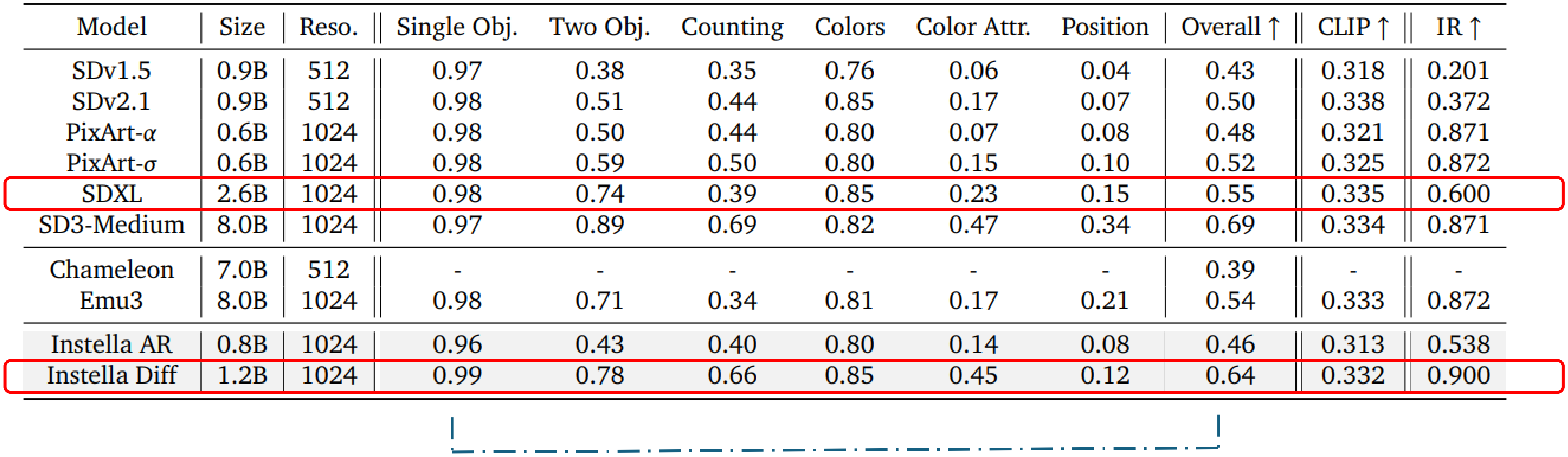
Model	Size	Reso.	Single Obj.	Two Obj.	Counting	Colors	Color Attr.	Position	Overall ↑	CLIP ↑	IR ↑
SDv1.5	0.9B	512	0.97	0.38	0.35	0.76	0.06	0.04	0.43	0.318	0.201
SDv2.1	0.9B	512	0.98	0.51	0.44	0.85	0.17	0.07	0.50	0.338	0.372
PixArt-α	0.6B	1024	0.98	0.50	0.44	0.80	0.07	0.08	0.48	0.321	0.871
PixArt-σ	0.6B	1024	0.98	0.59	0.50	0.80	0.15	0.10	0.52	0.325	0.872
SDXL	2.6B	1024	0.98	0.74	0.39	0.85	0.23	0.15	0.55	0.335	0.600
SD3-Medium	8.0B	1024	0.97	0.89	0.69	0.82	0.47	0.34	0.69	0.334	0.871
Chameleon	7.0B	512	-	-	-	-	-	-	0.39	-	-
Emu3	8.0B	1024	0.98	0.71	0.34	0.81	0.17	0.21	0.54	0.333	0.872
Instella AR	0.8B	1024	0.96	0.43	0.40	0.80	0.14	0.08	0.46	0.313	0.538
Instella Diff	1.2B	1024	0.99	0.78	0.66	0.85	0.45	0.12	0.64	0.332	0.900

GenEval

Instella-T2I: Pushing the Limits of 1D Discrete Latent Space Image Generation, Wang et al, 2025 https://arxiv.org/abs/2506.21022

1024x1024 T2I Evaluation

- 128 tokens for both Instella-AR and Instella-Diff
- 32 times token reduction compared to EMU3
- Fully open public datasets for training
- Instella-Diff: competitive against SOTA models like SDXL that use in-house data



Model	Size	Reso.	Single Obj.	Two Obj.	Counting	Colors	Color Attr.	Position	Overall ↑	CLIP ↑	IR ↑
SDv1.5	0.9B	512	0.97	0.38	0.35	0.76	0.06	0.04	0.43	0.318	0.201
SDv2.1	0.9B	512	0.98	0.51	0.44	0.85	0.17	0.07	0.50	0.338	0.372
PixArt-α	0.6B	1024	0.98	0.50	0.44	0.80	0.07	0.08	0.48	0.321	0.871
PixArt-σ	0.6B	1024	0.98	0.59	0.50	0.80	0.15	0.10	0.52	0.325	0.872
SDXL	2.6B	1024	0.98	0.74	0.39	0.85	0.23	0.15	0.55	0.335	0.600
SD3-Medium	8.0B	1024	0.97	0.89	0.69	0.82	0.47	0.34	0.69	0.334	0.871
Chameleon	7.0B	512	-	-	-	-	-	-	0.39	-	-
Emu3	8.0B	1024	0.98	0.71	0.34	0.81	0.17	0.21	0.54	0.333	0.872
Instella AR	0.8B	1024	0.96	0.43	0.40	0.80	0.14	0.08	0.46	0.313	0.538
Instella Diff	1.2B	1024	0.99	0.78	0.66	0.85	0.45	0.12	0.64	0.332	0.900

GenEval

Instella-T2I: Pushing the Limits of 1D Discrete Latent Space Image Generation, Wang et al, 2025 https://arxiv.org/abs/2506.21022

Outline

- A brief history of multimodality models
  - Understanding
  - Generation
  - Unification
  - Agentic
- Special topics
  - Does image help with reasoning?
  - Tokenization vs compression
- Summary and future directions

Summary

- By leveraging language models, multimodality field has been revolutionized
- Diffusion has been dominating on image/video generation, but AR is coming back, leading to unification where language model plays a central role for both understanding and generation
- Agentic models extend perception to reasoning+acting, leveraging language model's basic reasoning and tool-use capabilities and RL framework
- There is a large intelligence gap on reasoning with images
- Token compression has huge potential for image/video generation and understanding

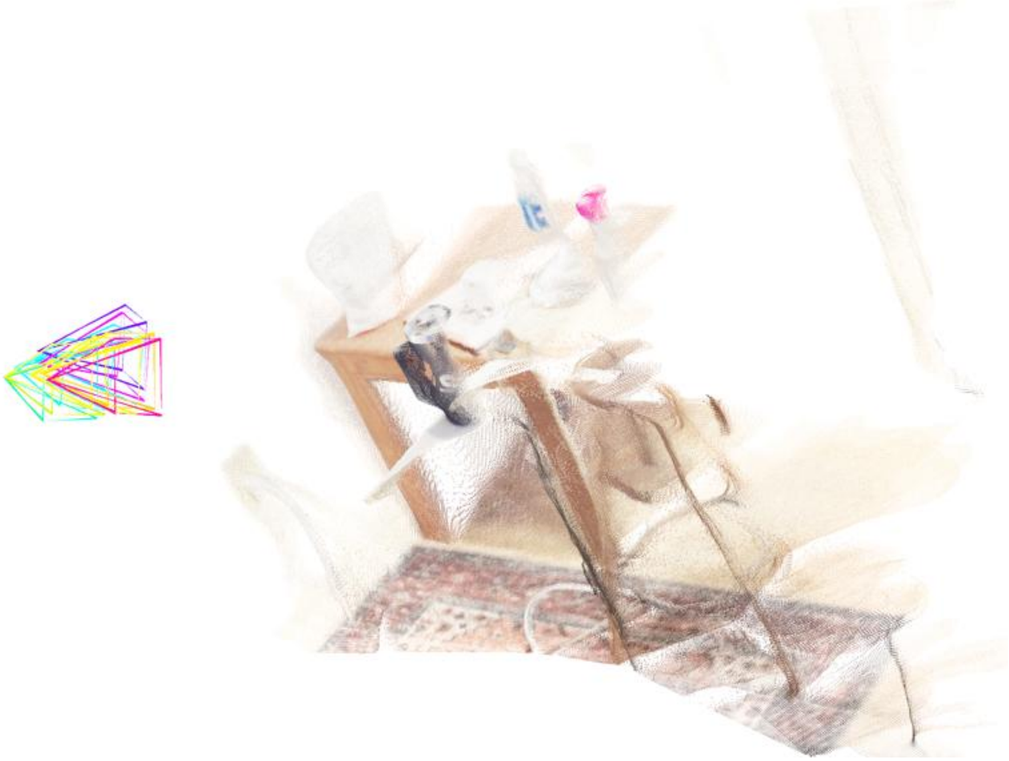
Future Directions

- Unification
  - How to get visual understanding and generation to help each other
  - Can multimodal learning help improving language model?
- 3D
  - Hot topic in CVPR2025, VGGT
  - Need to inject semantics by integrating with language
  - Component semantics: dog->walk, car->run
- Image/Video compression vs. token compression
  - Token redundancy in both visual understanding and generation
  - Language-conditioned token compression
- Embodied agents and robotics
  - Robots need a brain:
    - Physical abilities are amazing, but Intelligence is lacking
    - Multimodal models to play a major role
    - Learning paradigm shift: learning from interaction in real time, persistent memory, personalization

VGGT: Visual Geometry Grounded Transformer, Want et al, CVPR2025

Nutrition Facts
PRINCETON ALUMNI WEEKLY

[vggt - a Hugging Face Space by facebook](https://huggingface.co/spaces/facebook/vggt)



Acknowledgement

Ze Wang
Prakamya Mishra
Hao Chen
Benran Hu
Ximeng Sun
Jiang Liu
Jialian Wu
Yusheng Su
Xiaodong Yu
Emad Barsoum
Zhengyuan Yang
Linjie Li
Jianfeng Wang
An Yan
Lijuan Wang
Chenfei Wu
Nan Duan

THANKS!

zicheng.liu@amd.com

https://www.linkedin.com/in/zicheng-liu/

hosted by