Microsoft

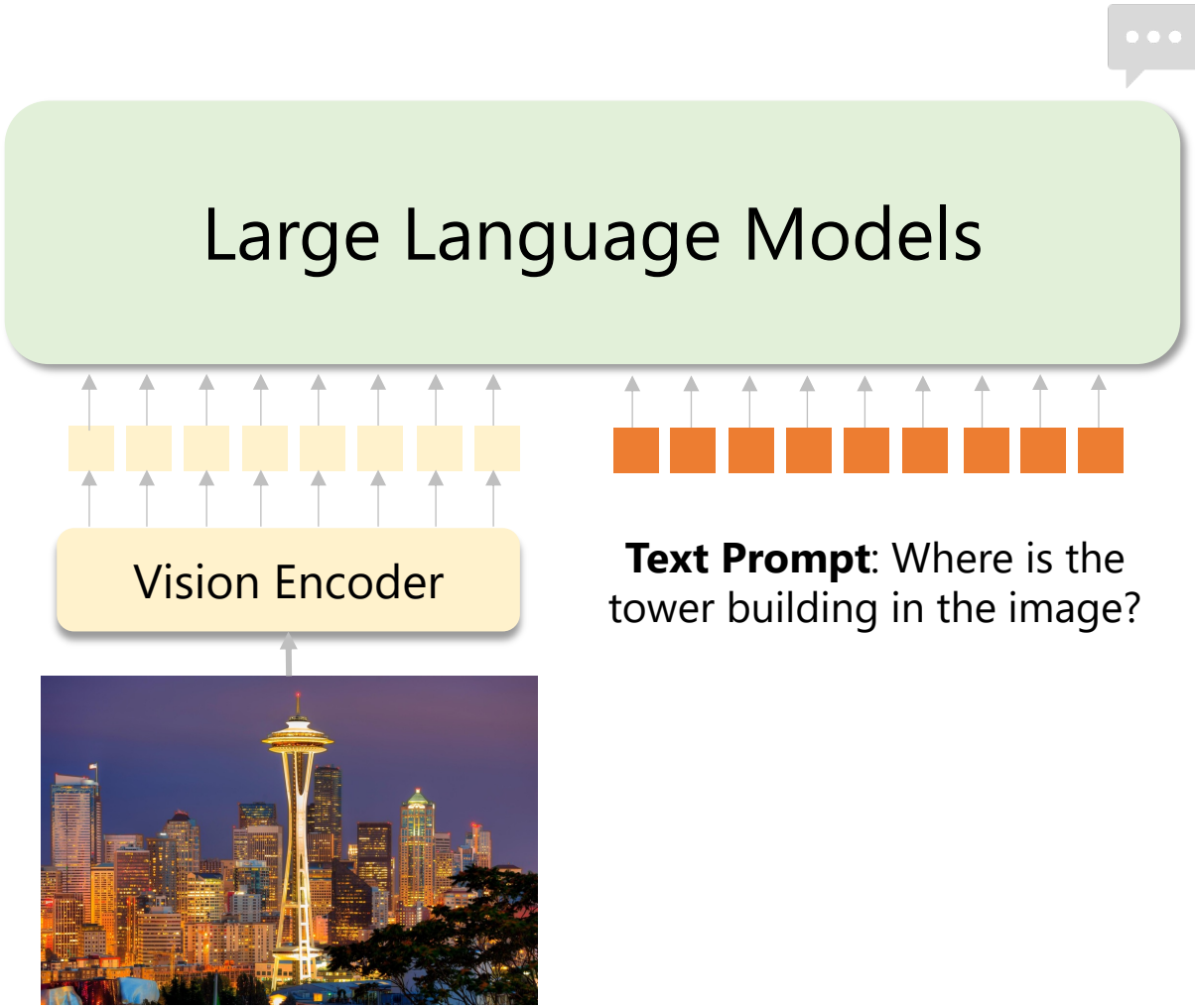
A Close Look at Vision in Large Multimodal Models

Jianwei Yang

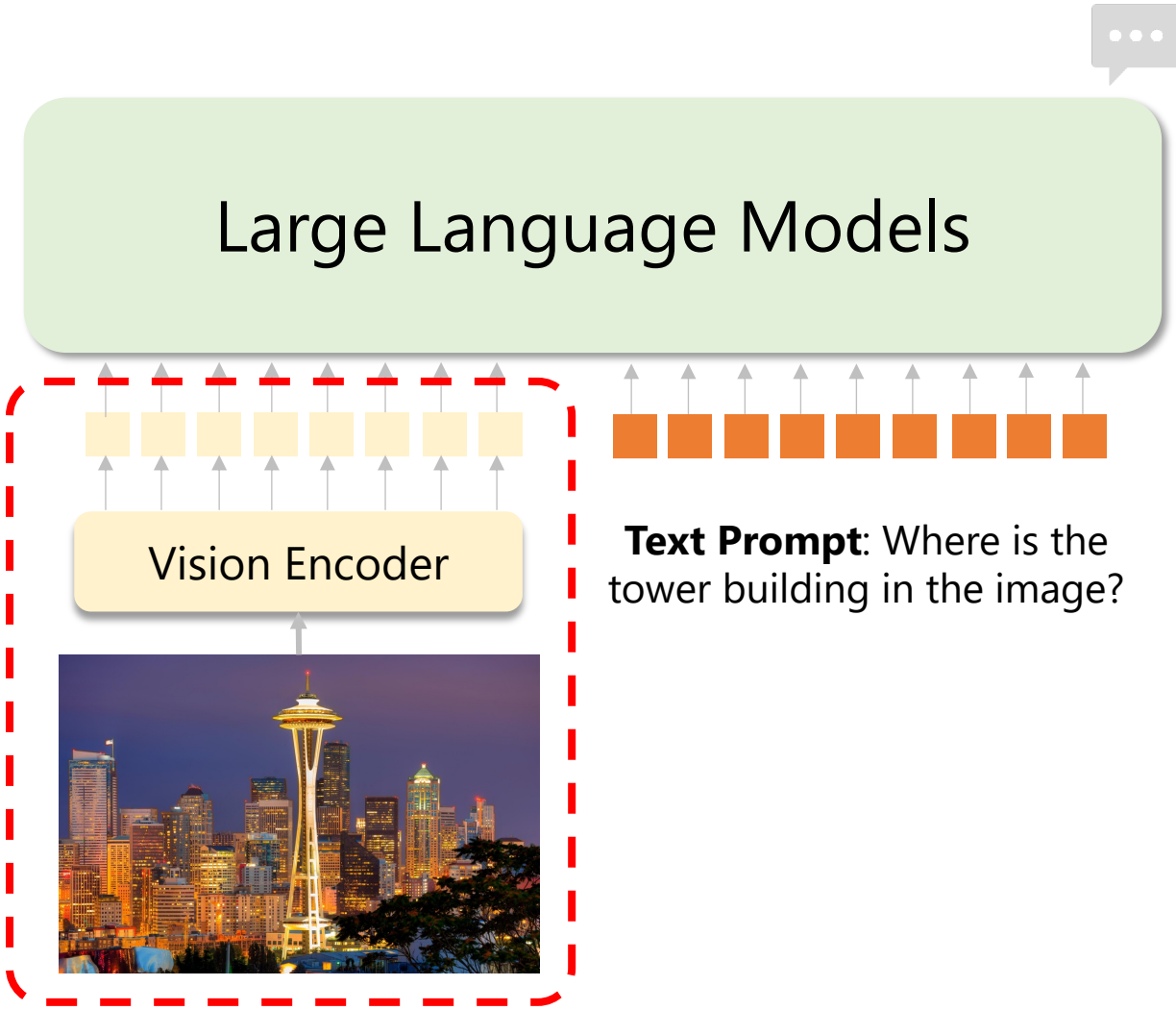
Microsoft Research

06/17/2024

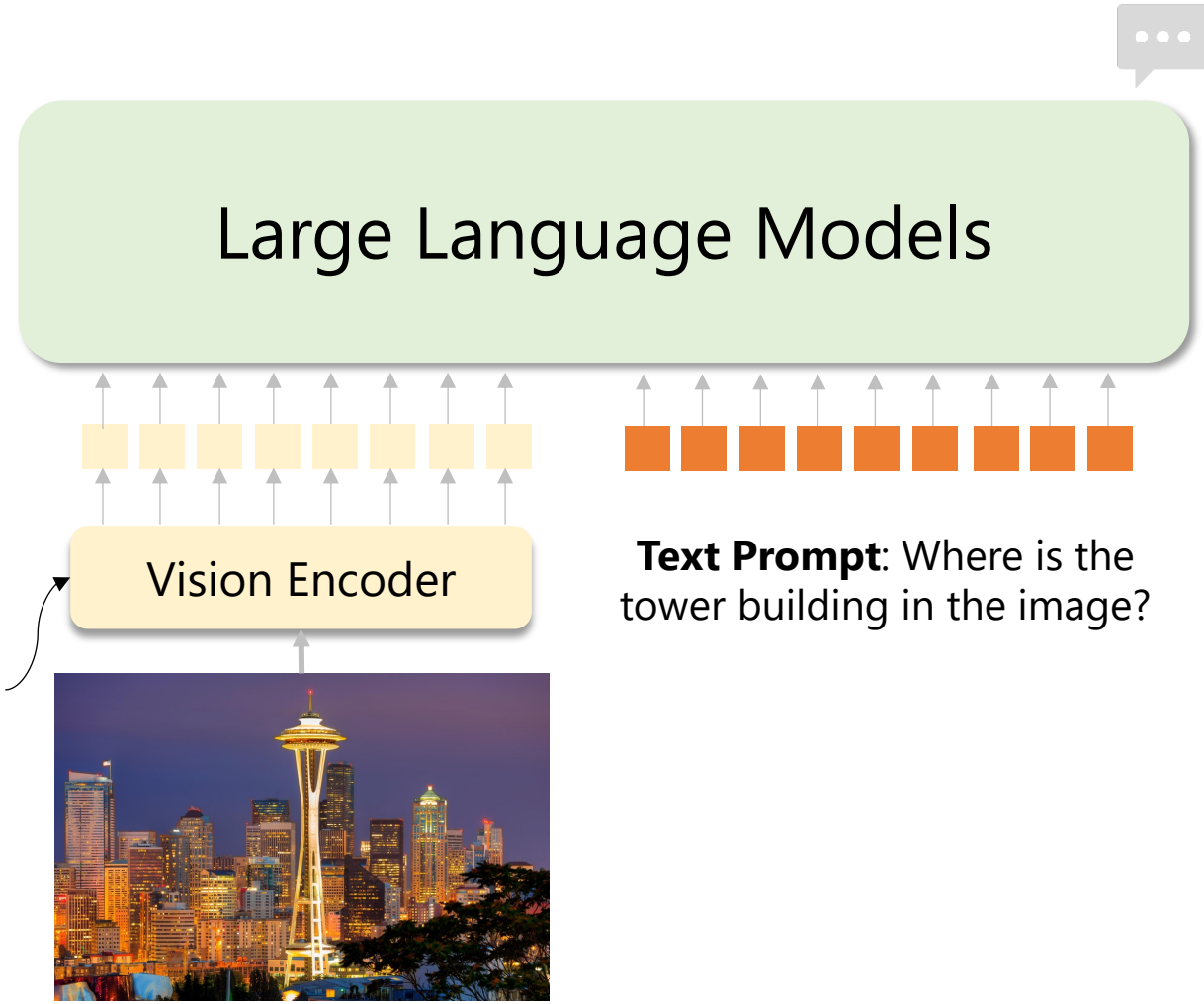
A Typical Large Multimodal Model (LMM)



A Typical Large Multimodal Model (LMM)

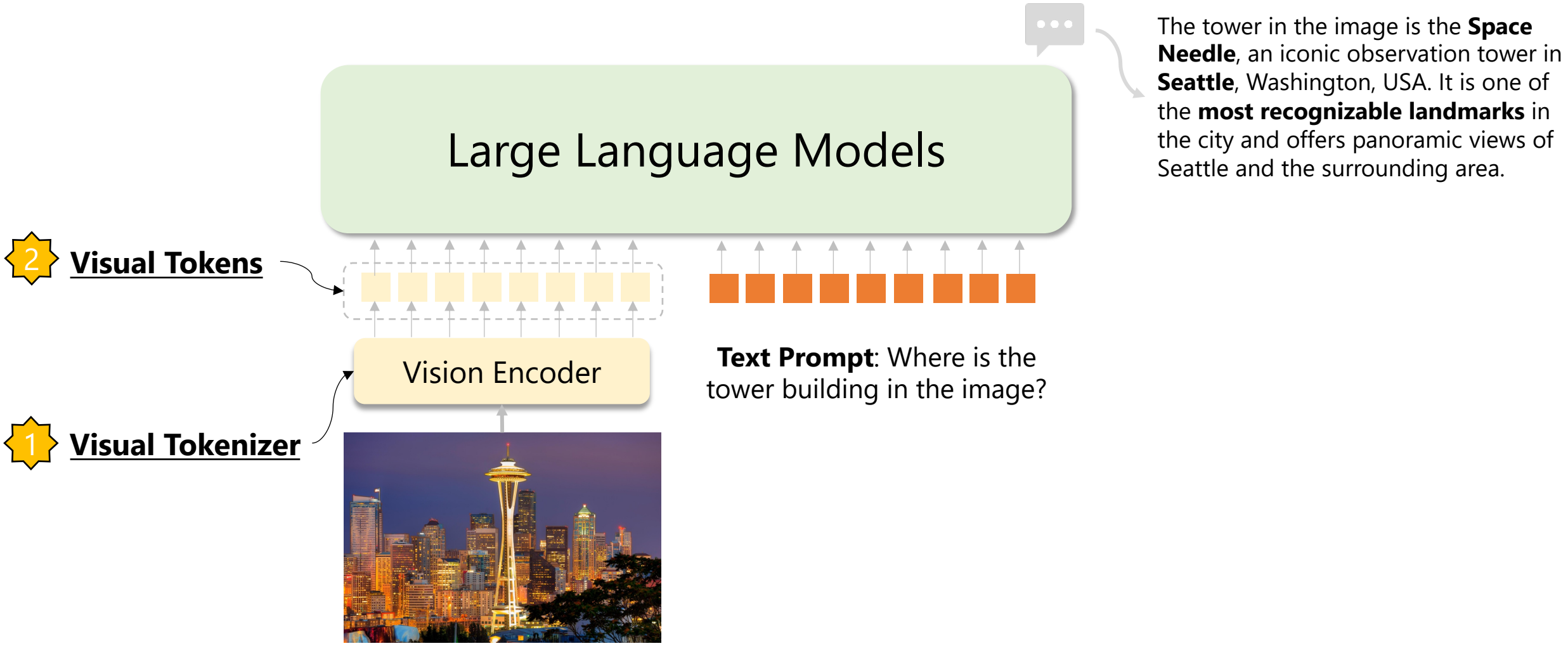


A Typical Large Multimodal Model (LMM)

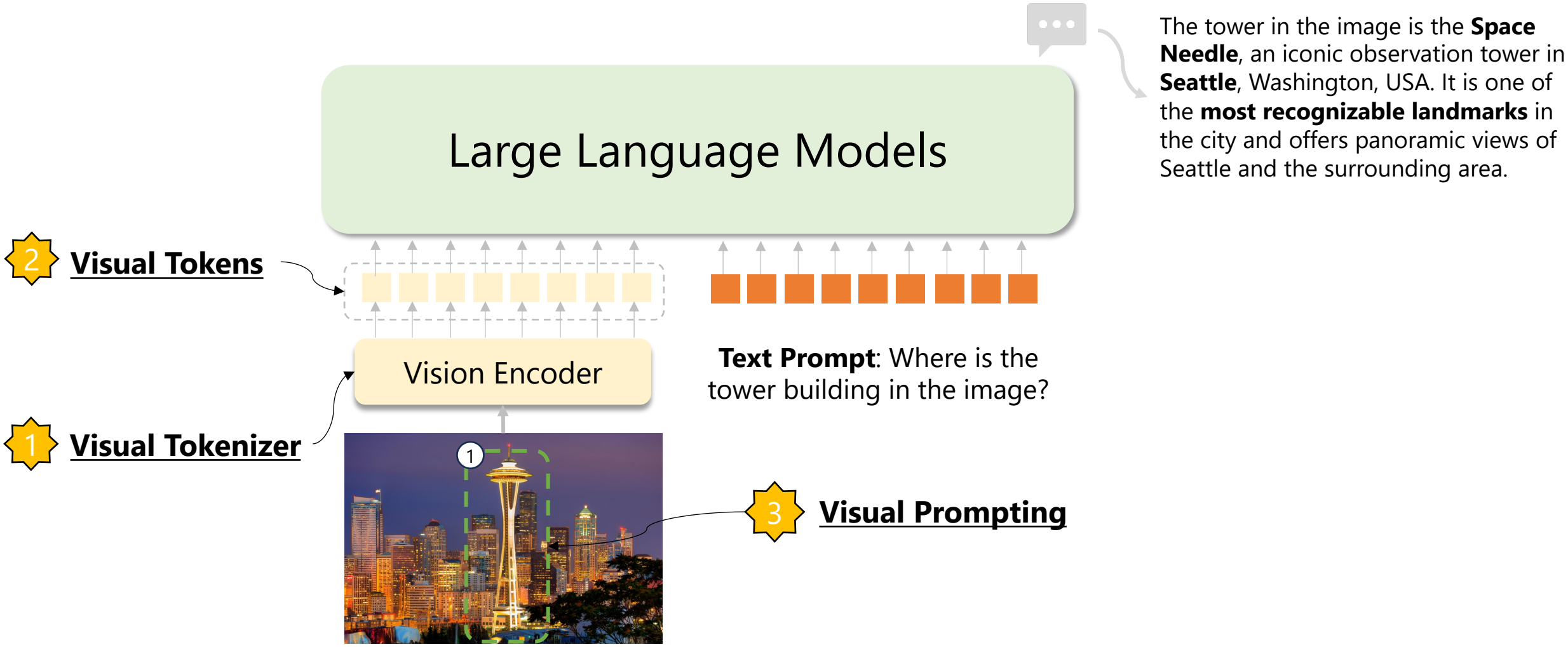


1 **Visual Tokenizer**

A Typical Large Multimodal Model (LMM)



A Typical Large Multimodal Model (LMM)



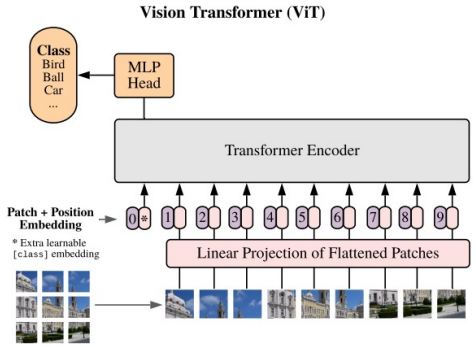
In this Talk - A Close Look at Vision

1. **Visual Tokenizer** — What vision encoder is a good vision tokenizer for LMMs?
2. **Visual Tokens** — How to cope with visual tokens for LLMs?
3. **Visual Prompting** — How to perform visual prompting for LMMs?

In this Talk - A Close Look at Vision

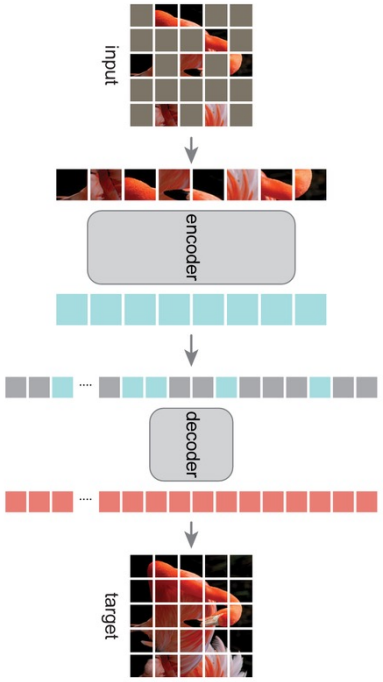
1 **Visual Tokenizer** What vision encoder is a good vision tokenizer for LMMs?

An Overview of Vision Encoder



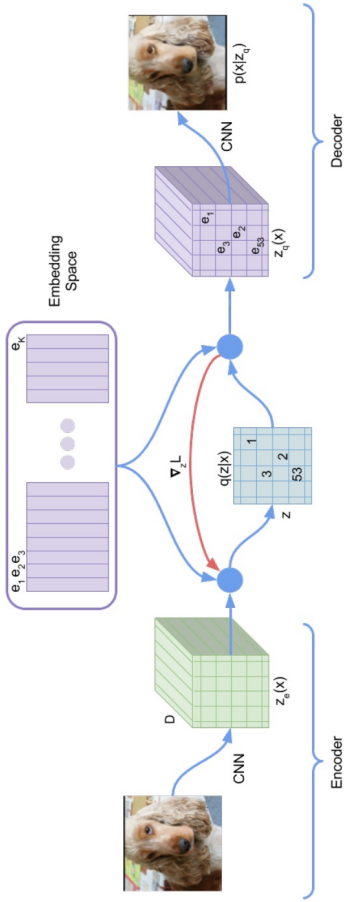
Supervised Learning

ViT, Dosovitskiy et al.



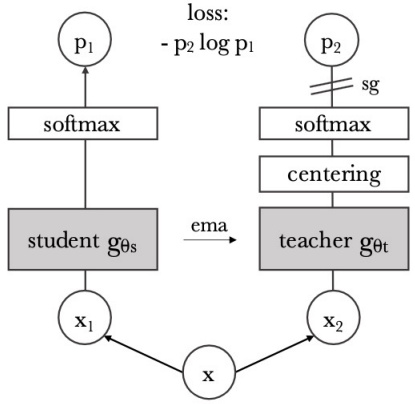
Masked Modeling

MAE, He et al.



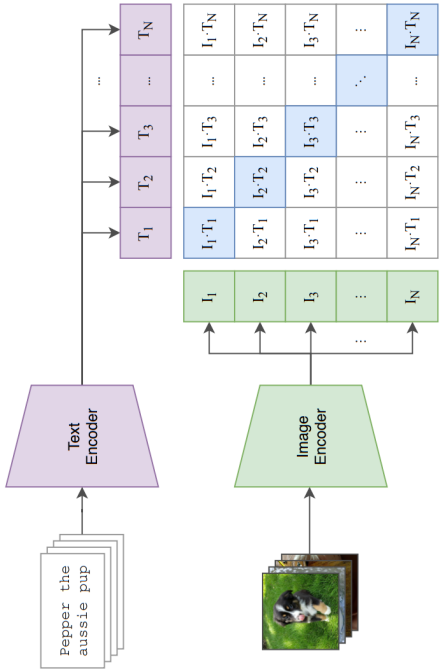
Auto-encoding

VQVAE, Van den Oord et al.



Self-Distillation

DINO, Caron et al.

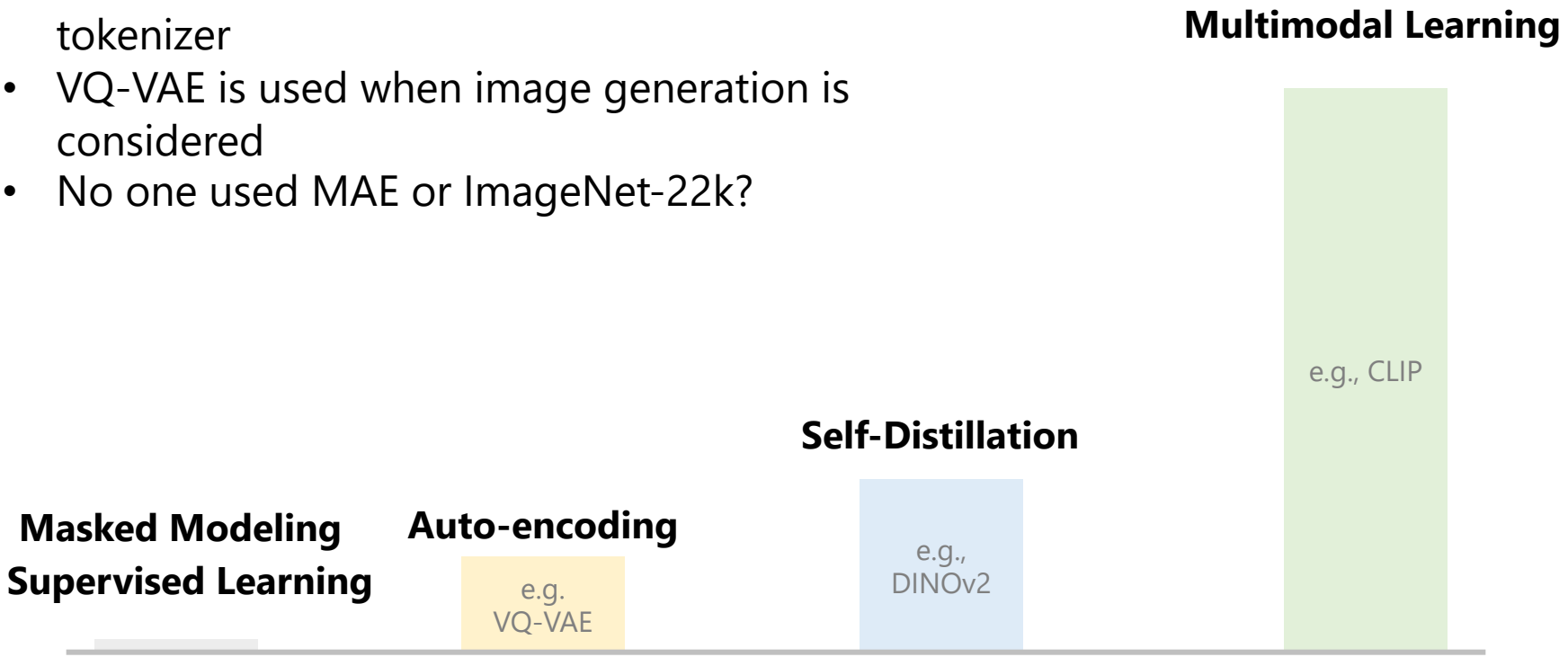


Multimodal Learning

CLIP, Radford et al.

Vision Encoder used in LMMs

- **A (Counter)-intuitive *de facto***:
  - Almost all LMMs used multimodal learned vision tokenizers
  - A few of them used DINOv2 for vision tokenizer
  - VQ-VAE is used when image generation is considered
  - No one used MAE or ImageNet-22k?

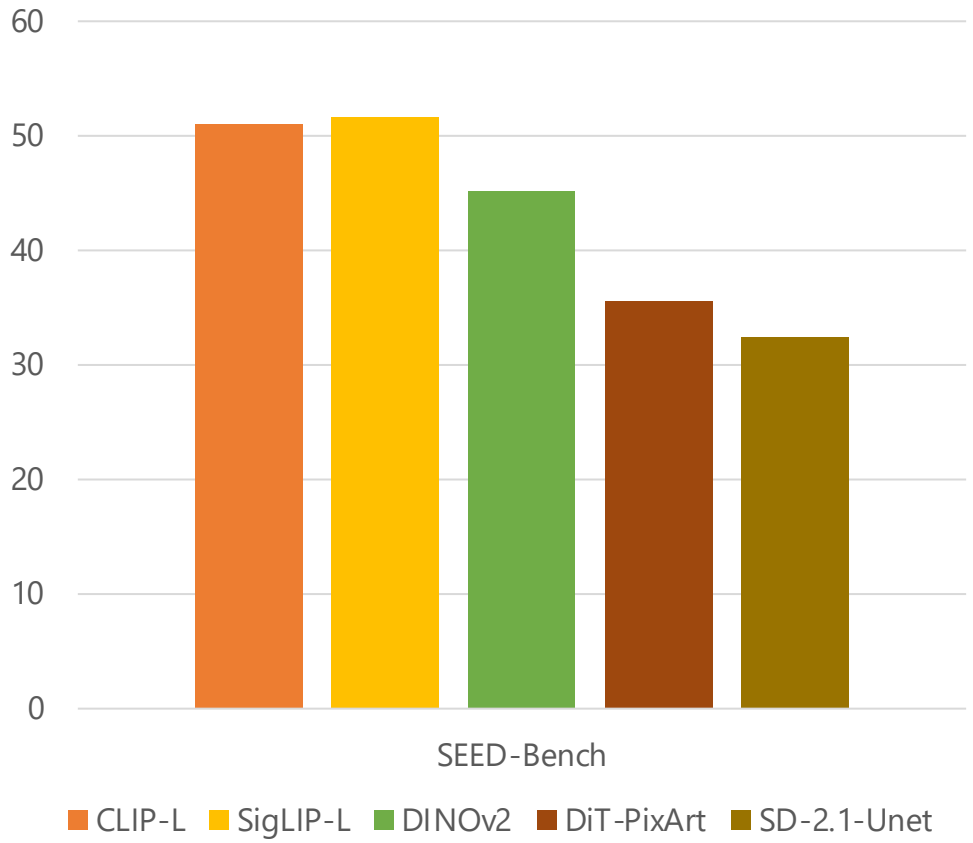


Popularity of different vision encoders in LMMs

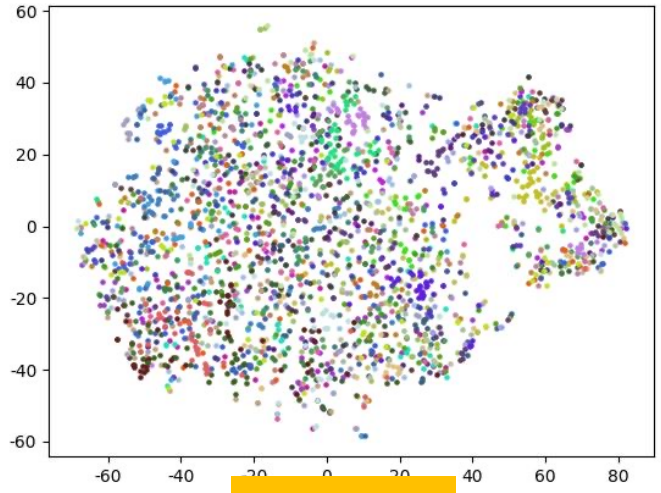
Behind the Scene

- **Ablation on different vision encoders:**
  - LLM – tinyLLAMA-1.1B
  - Follow LLaVA-1.5 training recipe
  - Report results on SEED-Bench

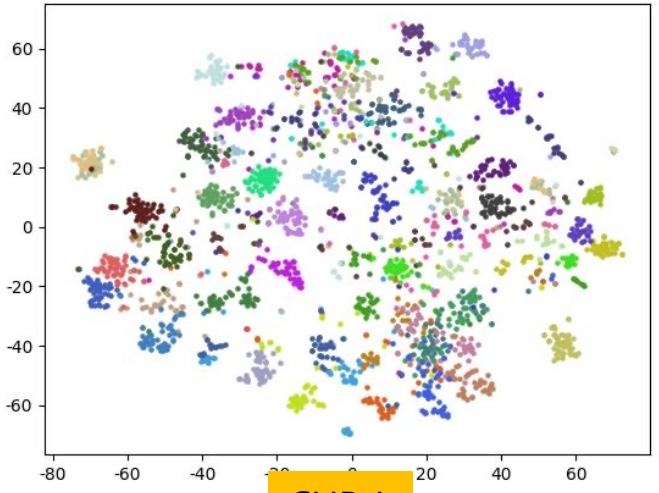
- **Observations:**
  - CLIP and SigLIP (significantly) outperforms non-multimodal pretrained ones
  - VQ-VAE and diffusion models are much worse for multimodal understanding tasks
  - DINOv2 is in the between but lag behind a lot CLIP and SigLIP.



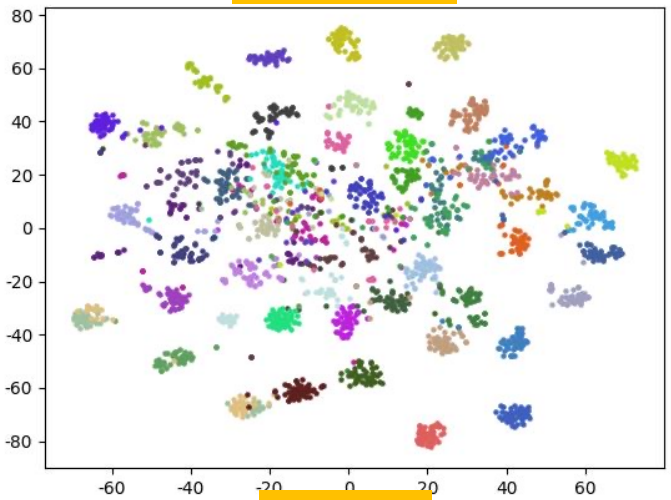
Behind the Scene



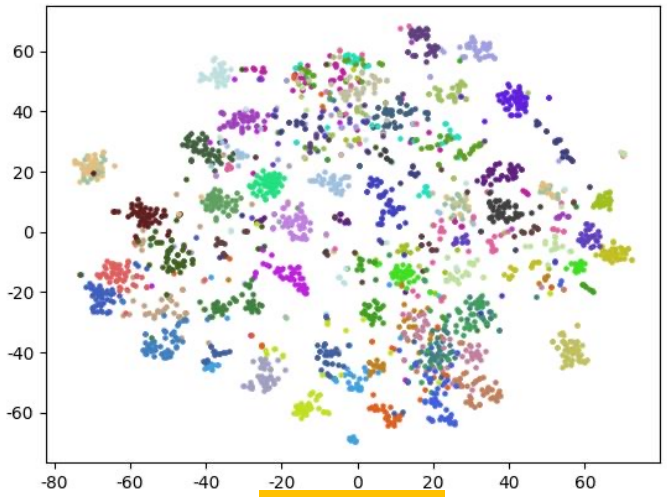
SD-2.1-336



CLIP-L



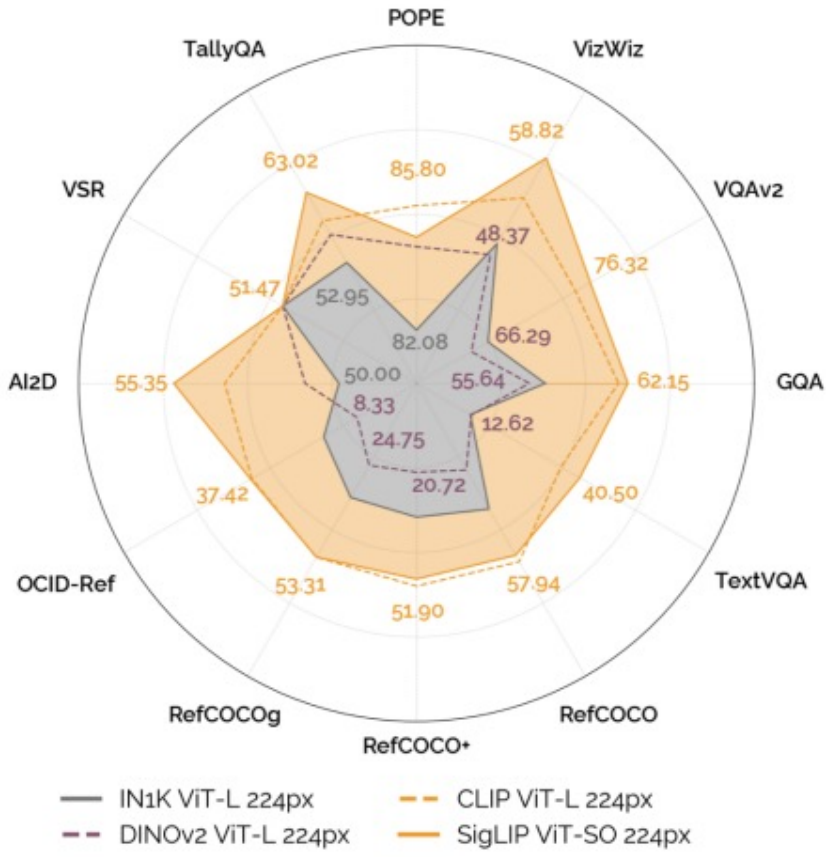
SigLIP-L



DINO-v2

- TSNE visualization of features extracted from different vision encoders
- Different colors represent features extracted from different image categories in imagenet-1k
- The features from multimodal models can distinguish the visual concepts much better

Behind the Scene



LlaVA-1.5 as the baseline

Vision encoder	COCO cap. ↑ Karpathy val	VQAv2 ↑ Karpathy val	OKVQA ↑ val	GQA ↑ test-dev	MMVP ↑ test
CLIP-L/14	133.0	74.4	61.0	48.7	15.3
OpenCLIP-G	128.3	73.3	60.6	48.0	22.0
EVA-CLIP-g	140.9	77.0	63.0	50.1	27.3
SIGLIP-G/14	133.0	74.7	62.5	48.6	24.0
SILC-G/16	141.1	77.0	63.4	49.7	24.0
ViT-e	137.8	75.6	61.9	49.1	25.3
ViT-G	133.8	74.2	61.2	48.3	20.7
DINOv2-L/14	127.6	71.3	59.0	48.0	22.0

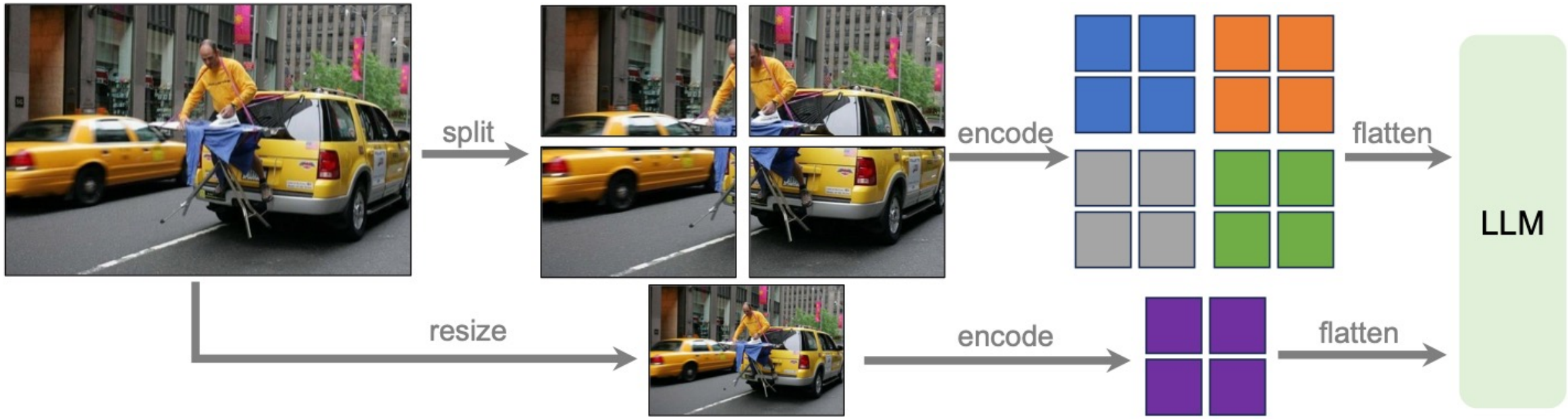
FlanT5-XL as language model, Q-Former for token compression

Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models. Karamcheti et al.
BRAVE : Broadening the visual encoding of vision-language models. Kar et al.

Vision Encoder Enhancements: Higher Resolution

Increasing image resolution and number of visual tokens (576->2880 maximal)

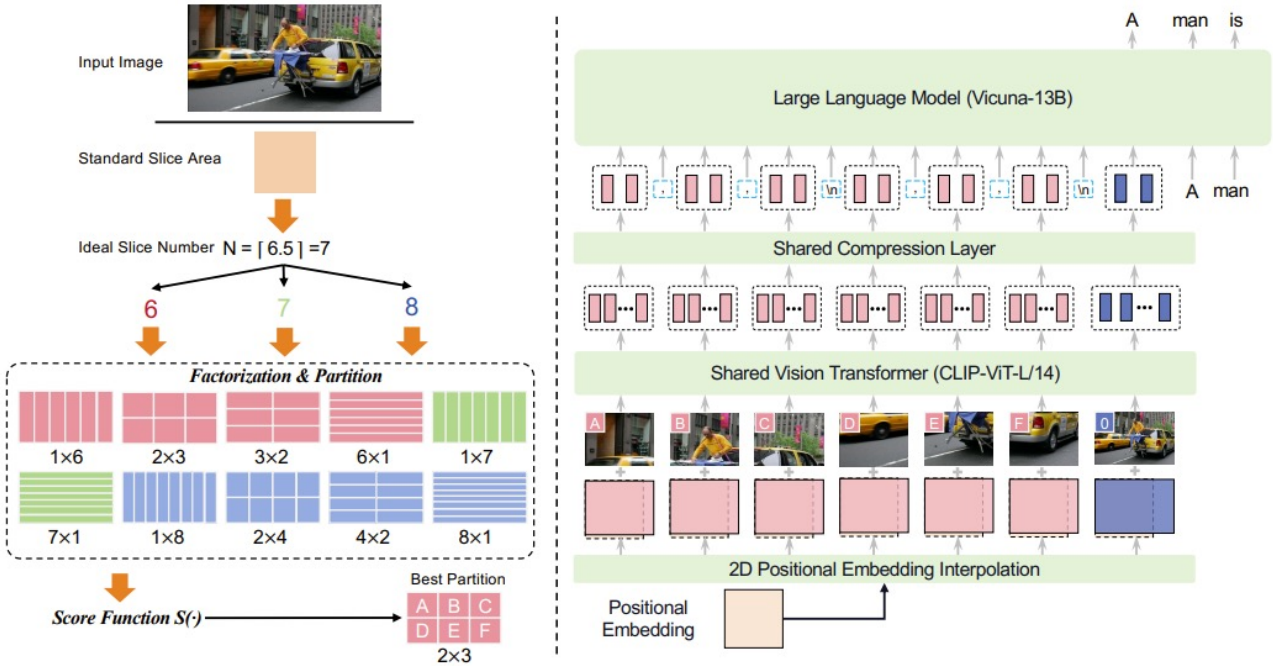
Dynamic High Resolution with fixed size vision encoder



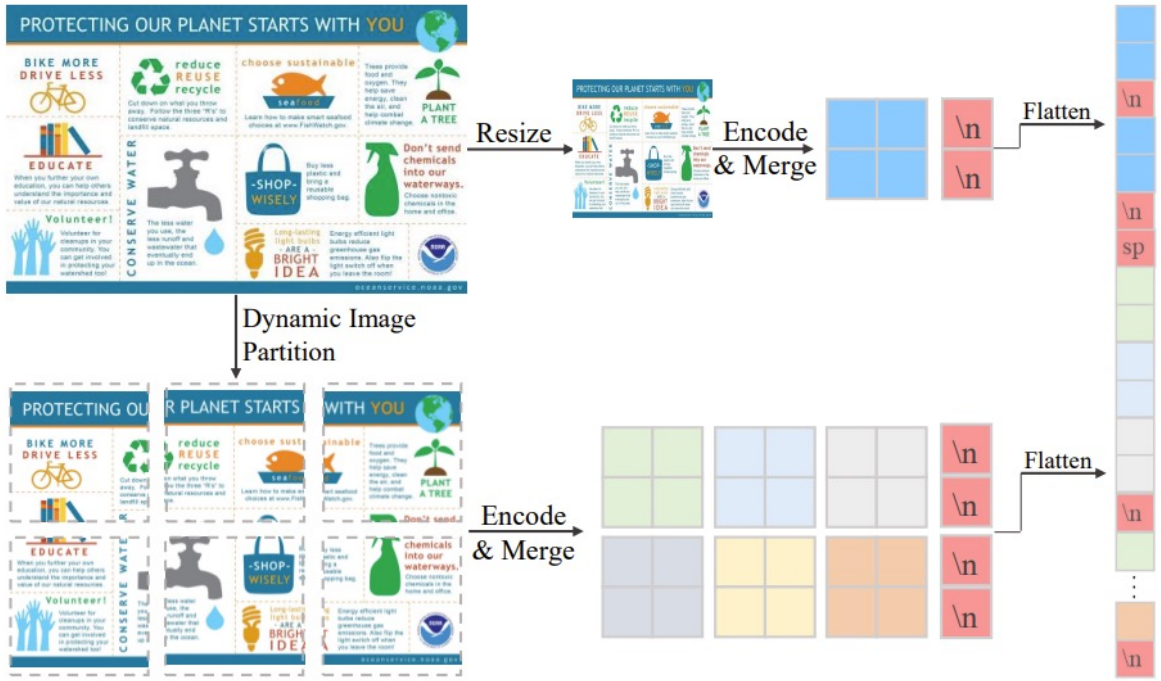
Vision Encoder Enhancements: Higher Resolution

- Image modularization to divide images into smaller slides;
- Compression layer to condense visual tokens
- A spatial schema to organize slide tokens.

- Dynamic Image Partition.
- Global-local format.
- Image 2D Structure Newline Indicator.



LlaVA-UHD



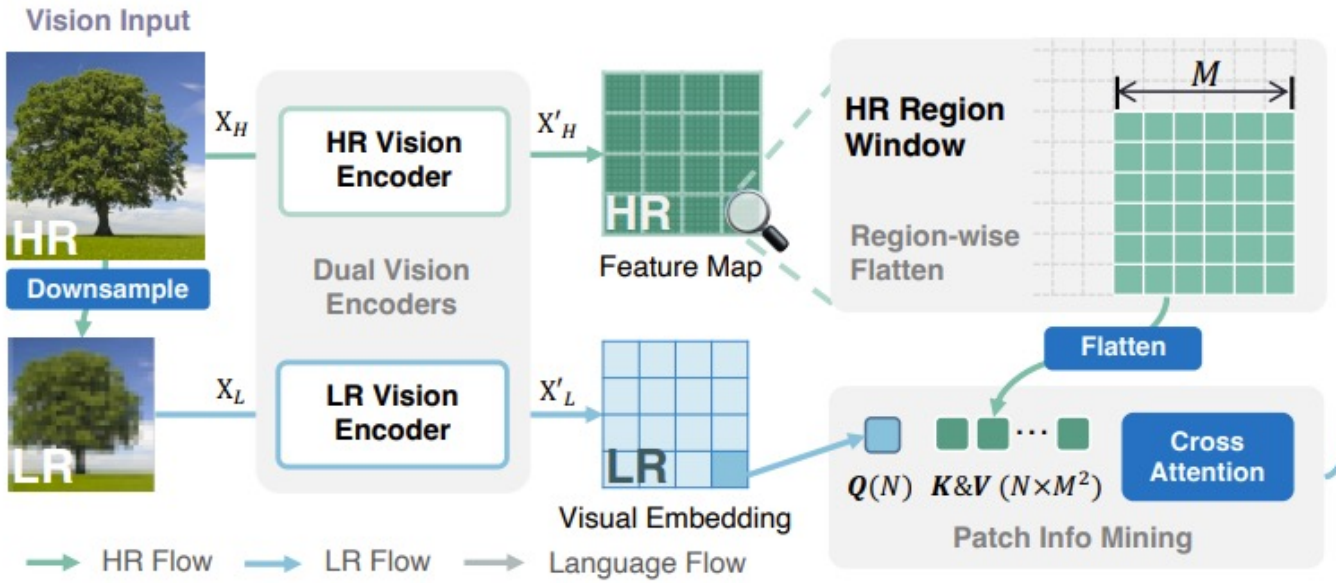
InternLM-XComposor2-4KHD

LLaVA-UHD: an LMM Perceiving Any Aspect Ratio and High-Resolution Images . Xu et al
InternLM-XComposor2-4KHD. Dong et al

Vision Encoder Enhancements: Higher Resolution

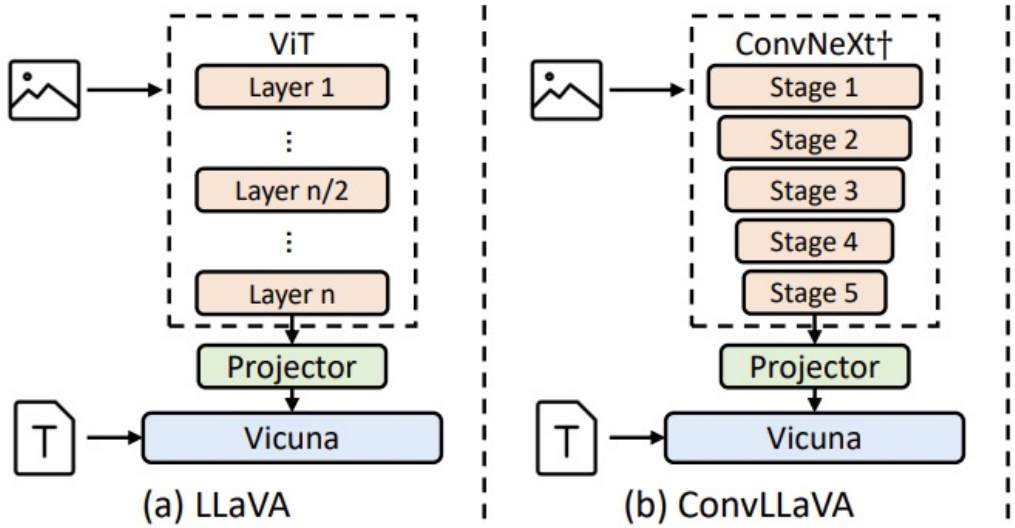
ConvNeXT can take high-resolution images by nature (w.o. positional embedding intrapolation)

- Dual vision encoder (CLIP + ConvNeXt)
- Patch info mining (Cross-attention query)



Mini-Gemini

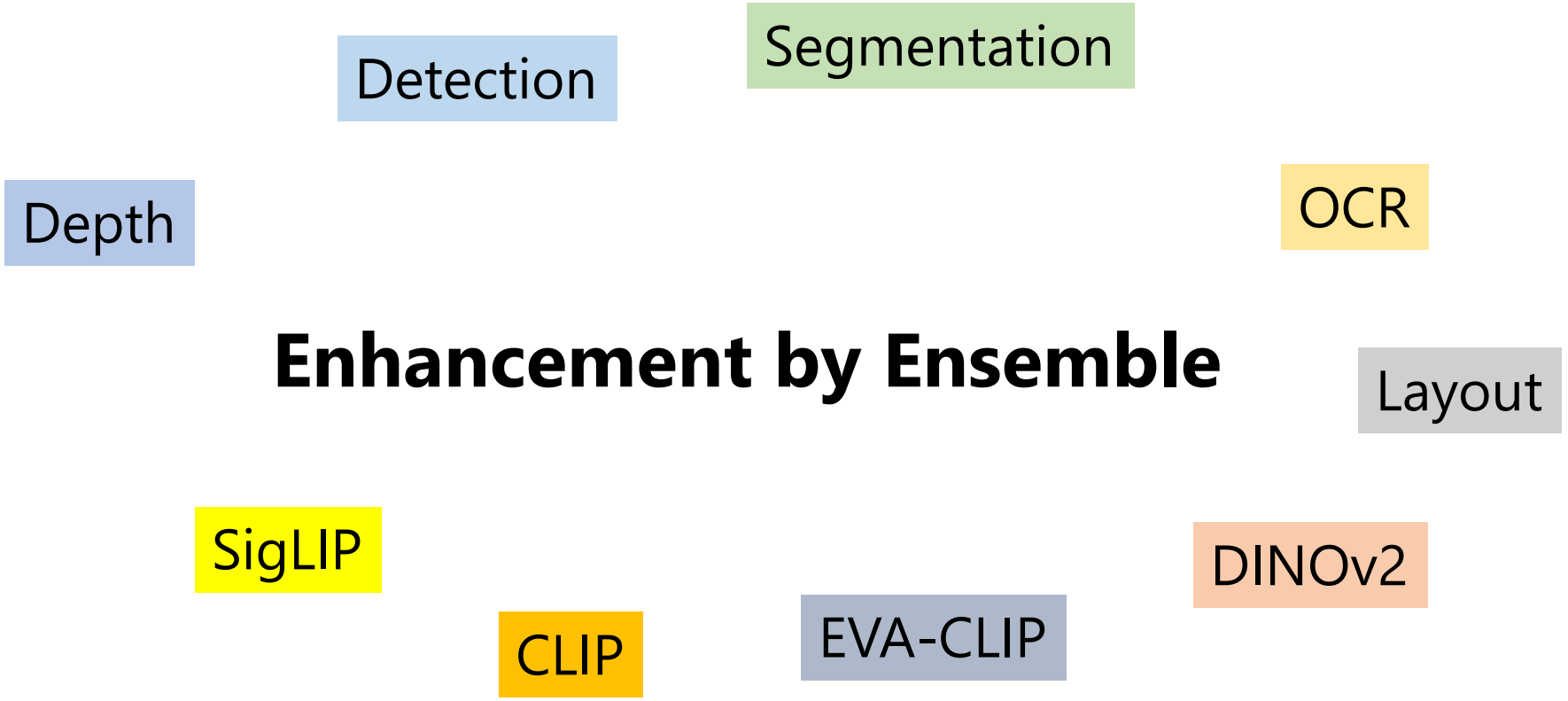
- Introduce an additional stage in convnext
- Three training stages



Conv-LlaVA

Mini-Gemini: Mining the Potential of Multi-modality Vision Language Models. Li et al
ConvLLaVA: Hierarchical Backbones as Visual Encoder for Large Multimodal Models. Ge et al

Vision Encoder Enhancements: Ensemble



Vision Encoder Enhancement: VCoder

GPT-4V is good at detailed descriptions but fails easily on counting.

USER
What is happening in the image?

Difficulty Level

GPT-4V
Characters from "Friends" are gathered in a living room, wearing various Halloween costumes, possibly celebrating or attending a Halloween party. Chandler stands out in a pink bunny costume.

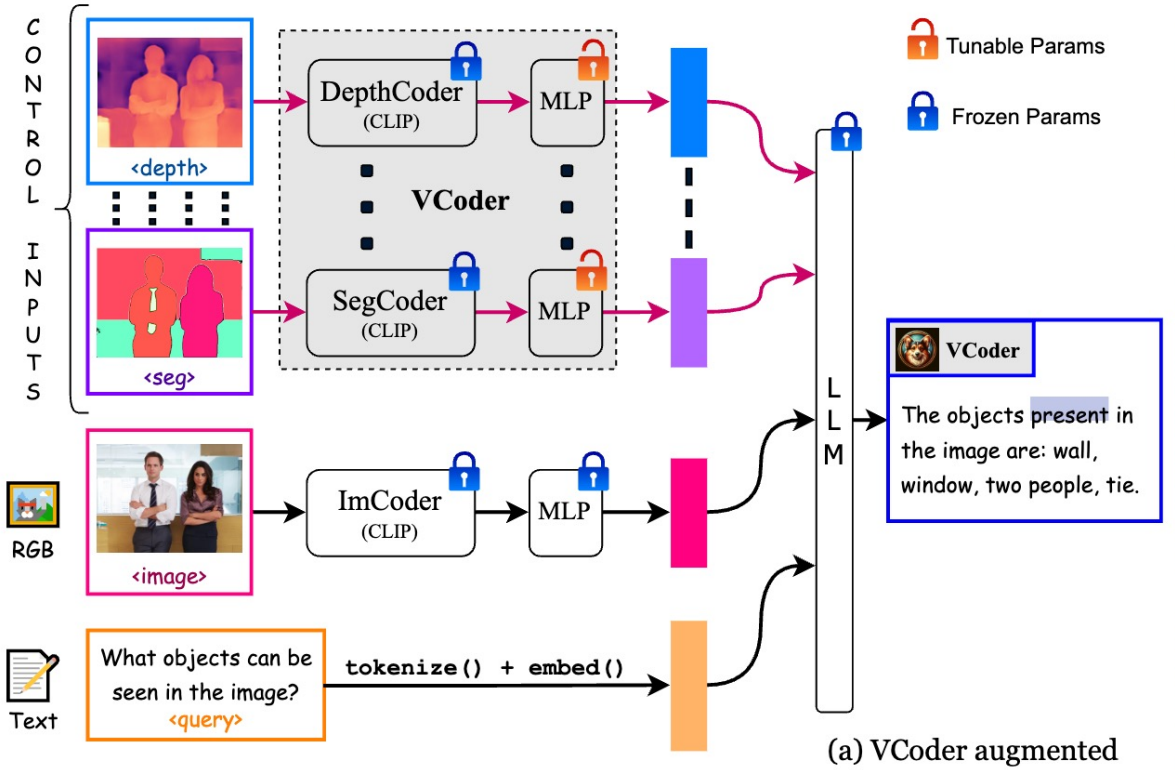
VCoder
The image shows a group of people gathered in what appears to be a living room or a dining area, dressed in various silly, quirky costumes. They are sitting at a dining table with chairs, engaged in a social event or party.

USER
Can you count the number of people in the image?

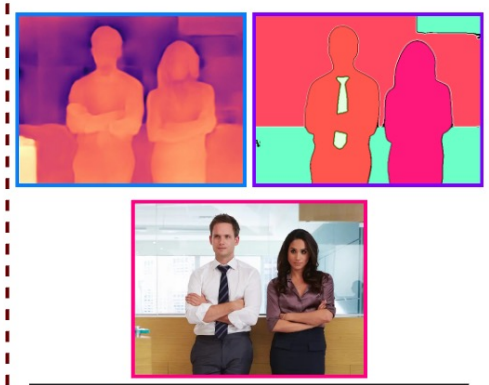
Difficulty Level

GPT-4V
Yes, there are ten people visible in the image.

VCoder
Yes, there are a total of 14 people in the image.



(a) VCoder augmented Multimodal LLM



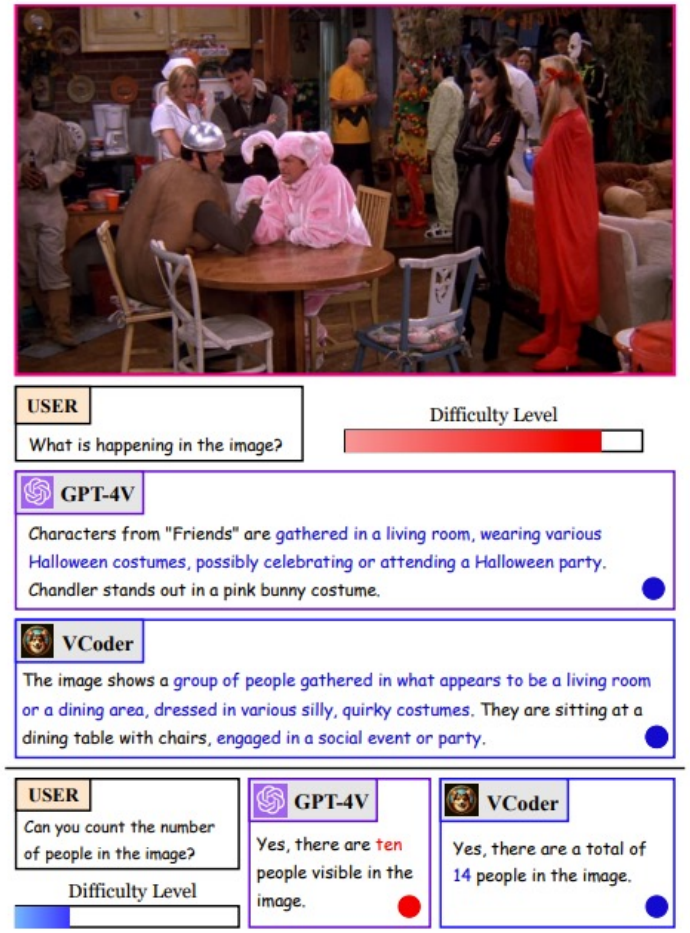
USER
Could you sort the objects in this photo from foreground to background?

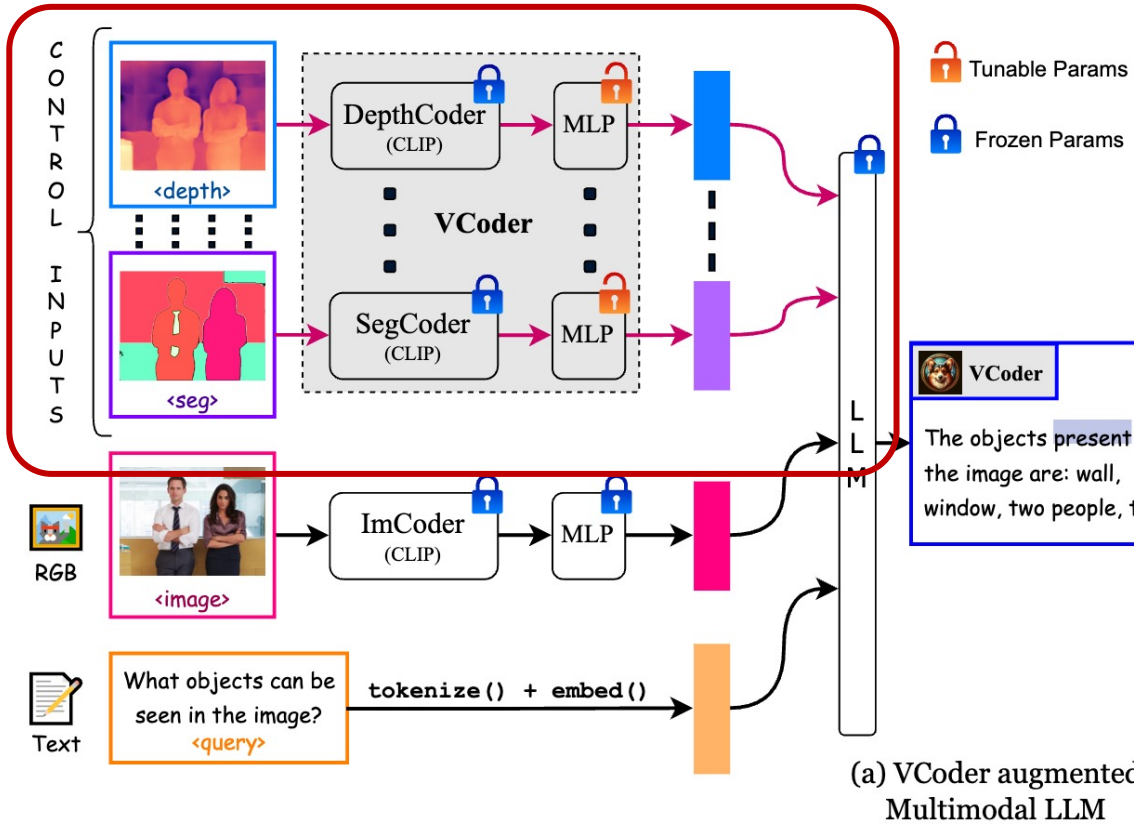
VCoder
The objects present in the image are: tie, person, person-2, wall, window.

(b) QnA about depth order of objects

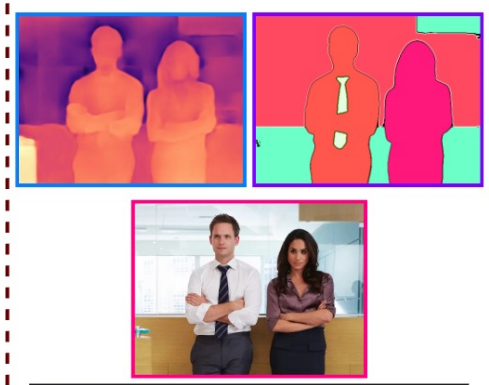
Vision Encoder Enhancement: VCoder

GPT-4V is good at detailed descriptions but fails easily on counting.





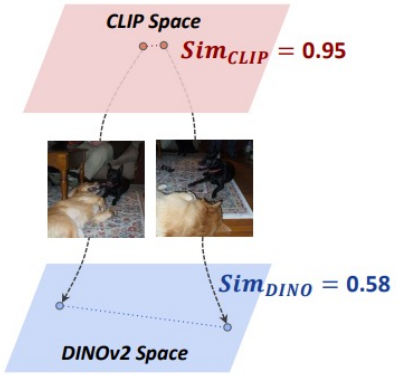
(a) VCoder augmented Multimodal LLM



(b) QnA about depth order of objects

Feed the LMMs with perception modalities such as **segmentations**, **depth maps**, etc, can significantly improve perception abilities.

Vision Encoder Enhancement: MoF

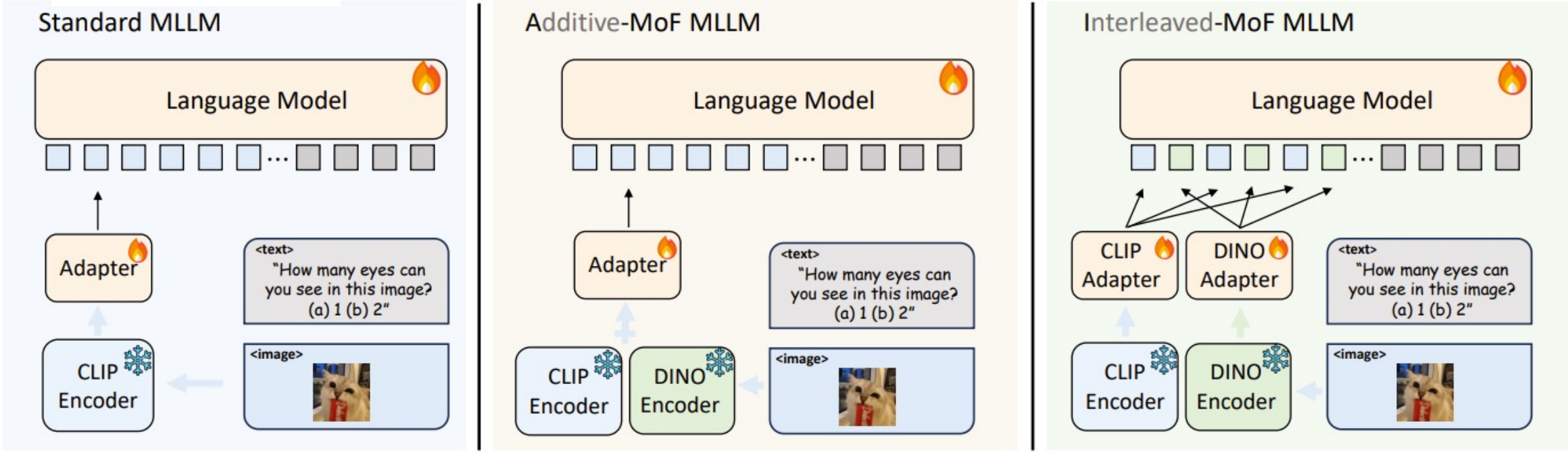


Finding: CLIP and DINOv2 represent image in different manners

Method: Mixture-of-Features from off-the-shelf vision encoders

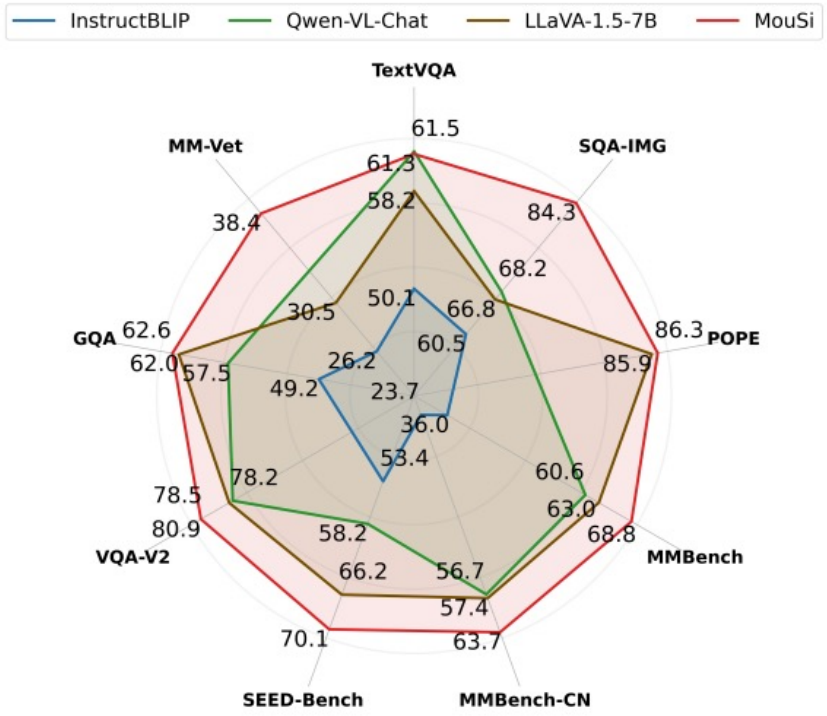
method	res	#tokens	MMVP	LLaVA	POPE
LLaVA	224^2	256	5.5	81.8	50.0
LLaVA	336^2	576	6.0	81.4	50.1
LLaVA + I-MoF	224^2	512	16.7 (+10.7)	82.8	51.0
$LLaVA^{1.5}$	336^2	576	24.7	84.7	85.9
$LLaVA^{1.5}$ + I-MoF	224^2	512	28.0 (+3.3)	82.7	86.3

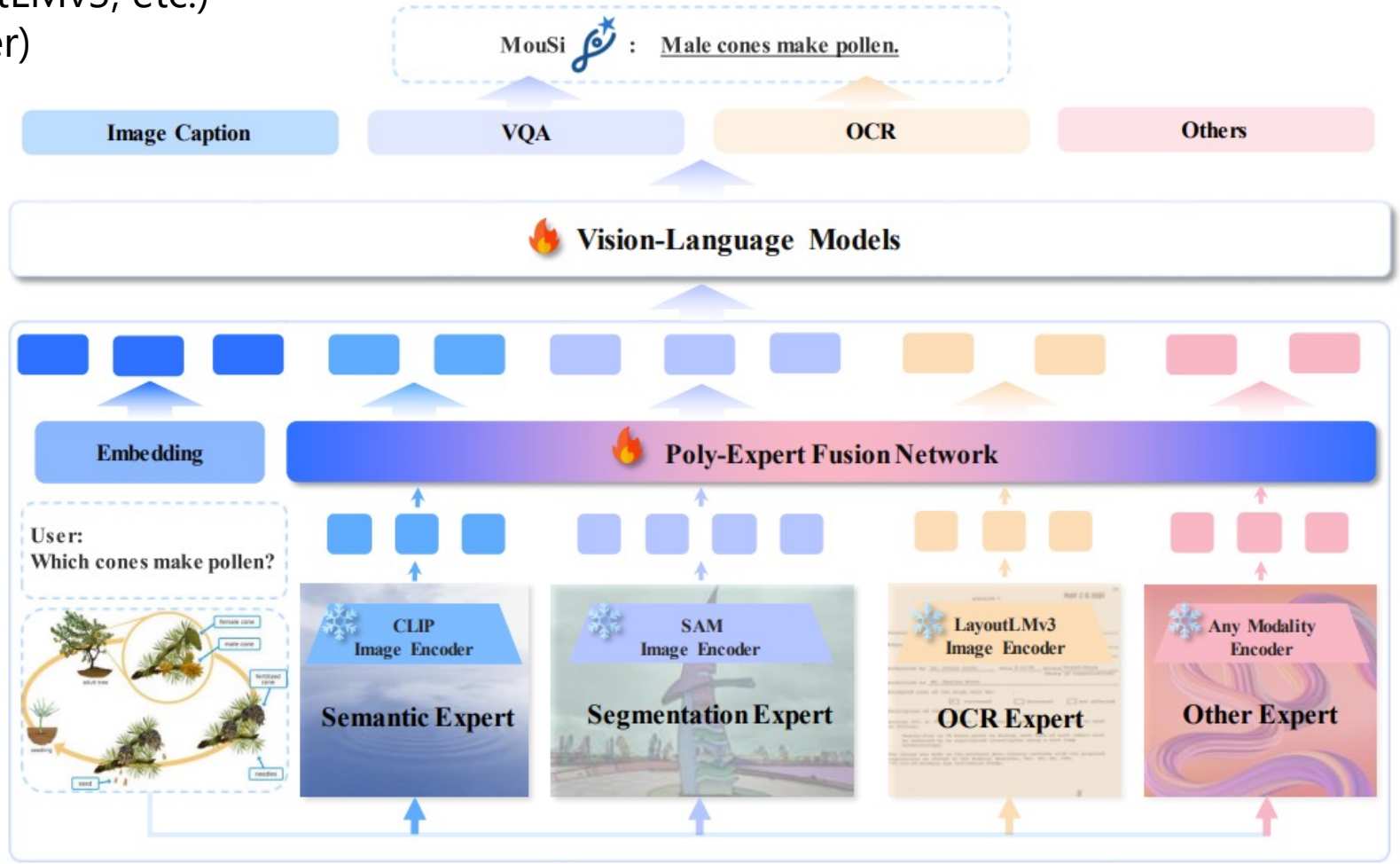
I-MoF improves models' ability to capture details



Vision Encoder Enhancement: Mousi

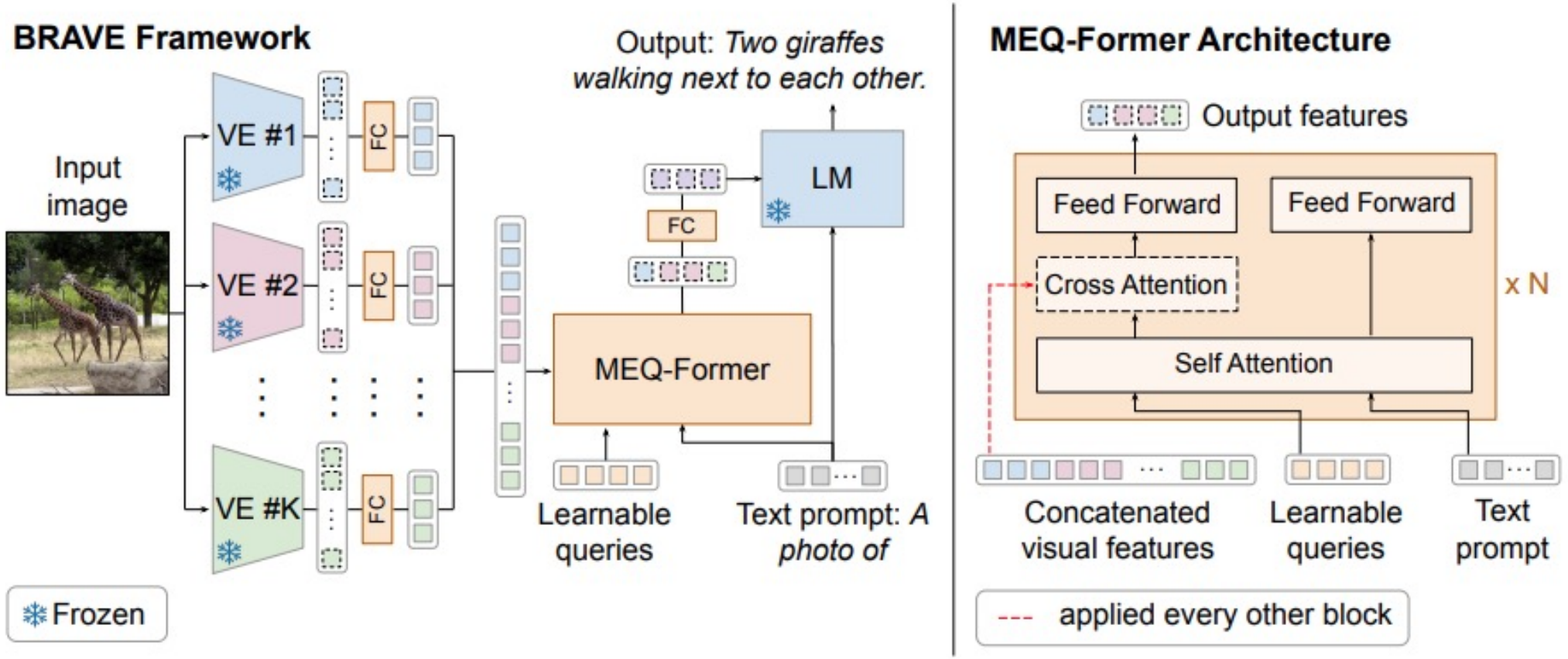
- More vision experts (CLIP, SAM, LayoutLMv3, etc.)
- Poly-Expert Fusion (MLP and Q-Former)
- Multi-path-single-token projection





Vision Encoder Enhancement: BRAVE

- **MEQ-Former** – Multi-encoder querying transformer
- **Five different vision encoders**: EVA-CLIP-g, CLIP-L/14, SILC-G/16, ViT-e, DINOv2-L/14

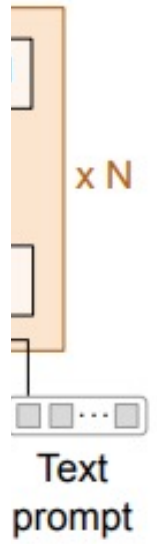


BRAVE : Broadening the visual encoding of vision-language models. Kar et al.

Vision Encoder Enhancement: BRAVE

- **MEQ-Former** – Multi-encoder querying transformer
- **Five different vision encoders**: EVA-CLIP-g, CLIP-L/14, SILC-G/16, ViT-e, DINOv2-L/14

Method	# params		Fine-tuned			Zero-shot			
	Trainable	Total	VQAv2 test-dev	OKVQA val	GQA test-dev	VizWiz-QA test-dev	GQA test-dev	MMVP test	POPE test
SimVLM [86]	632M	632M	80.0	-	-	-	-	-	-
Flamingo [3]	10.2B	80B	82.0	57.8	-	31.6	-	-	-
MiniGPT-v2 [13]	7B	8B	-	57.8	60.1	<u>53.6</u>	-	-	-
GiT2 [82]	5.1B	5.1B	81.7	-	-	-	-	-	-
Qwen-VL [5]	9.6B	9.6B	79.5	58.6	59.3	35.2	-	-	-
SPHINX-2k [58]	13B	16.5B	80.7	62.6	63.1	44.9	-	-	<u>87.2</u>
PaLI-17B [17]	16.9B	16.9B	**84.3**	<u>64.5</u>	-	-	-	-	-
BLIP-2 [53]	1.2B	12.1B	81.6	54.7	-	29.4	44.7	-	85.3
InstructBLIP [23]	188M	14.2B	-	55.5	-	33.4	<u>49.5</u>	16.7	78.9
LLaVa$^{1.5}$ [61]	13B	13.4B	80.0	-	<u>63.3</u>	<u>53.6</u>	-	24.7	85.9
LLaVA$^{1.5}$ (I-MoF) [79]	13B	13.6B	79.3	-	-	-	-	<u>31.3</u>	86.7
BRAVE	3B	10.3B	<u>82.5</u>	**66.0**	**66.3**	**54.2**	**52.7**	**42.0**	**87.6**

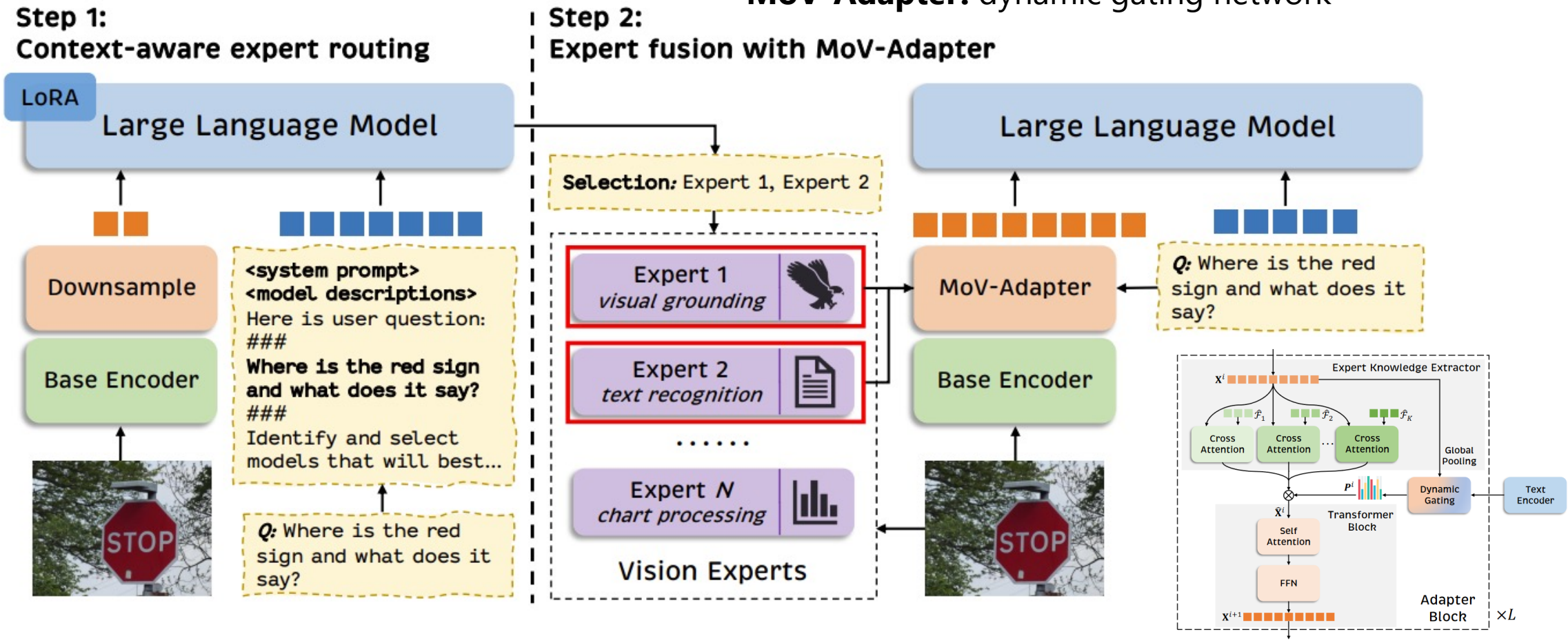


BRAVE : Broadening the visual encoding of vision-language models. Kar et al.

Vision Encoder Enhancement: MoVA

Strategy: Adaptive mixture of experts

Context-aware expert routing: requires constructing routing instruction tuning data
MoV-Adapter: dynamic gating network



MoVA: Adapting Mixture of Vision Experts to Multimodal Context. Zong et al.

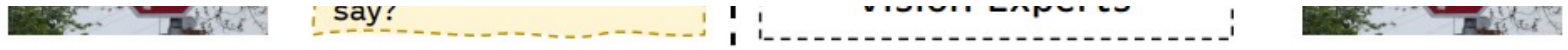
Vision Encoder Enhancement: MoVA

Method: Adaptive mixture of experts

Context-aware expert routing: requires constructing routing instruction tuning data

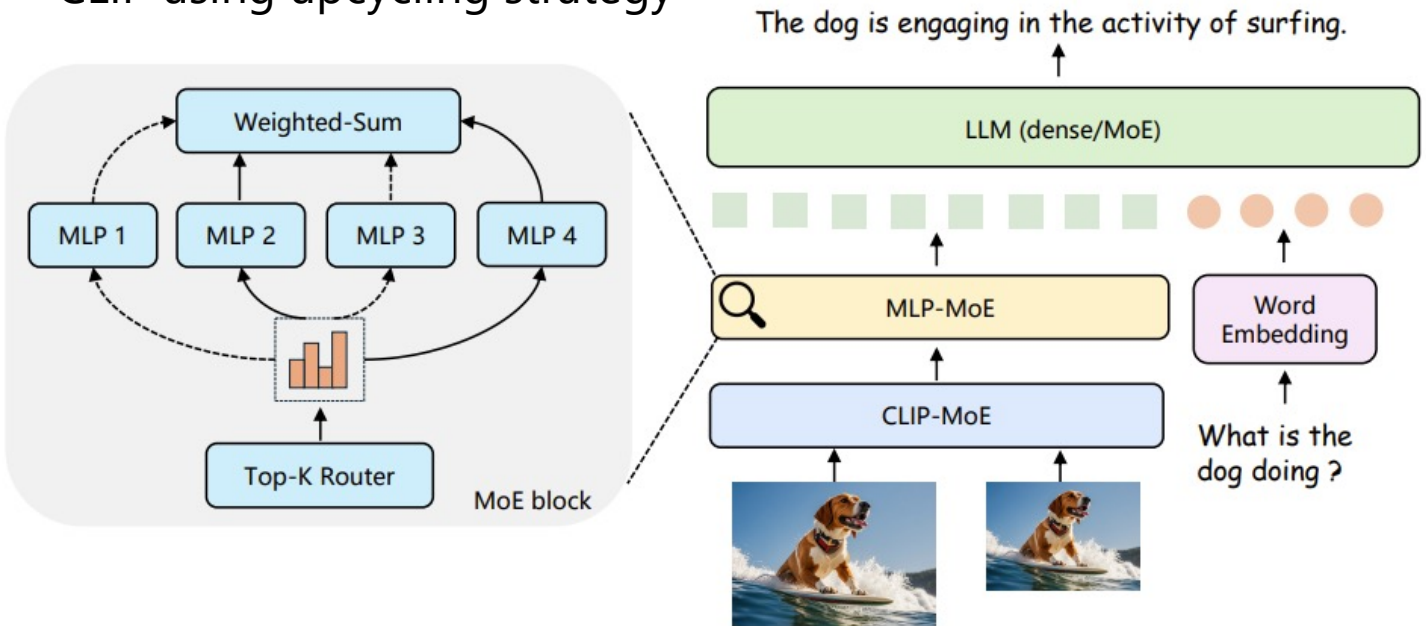
MoV-Adapter: dynamic gating network

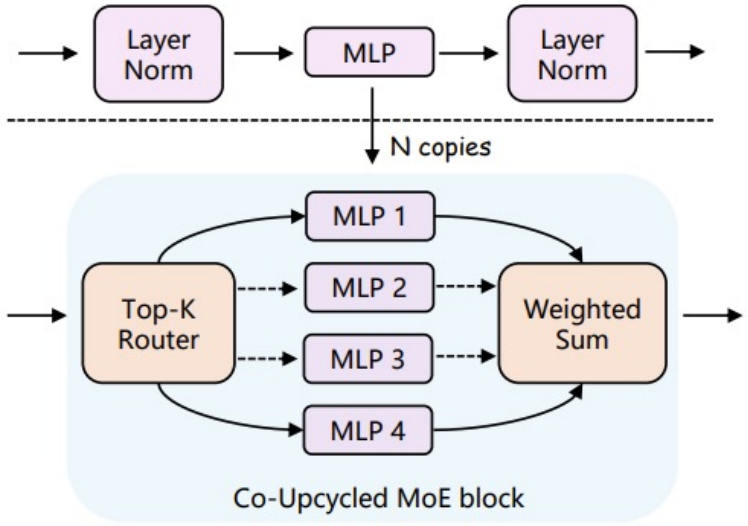
Vision Encoder	Task	MMB	DocVQA	ChartQA	GQA	POPE	REC	RES	SLAKE
CLIP [60]	Image-text Contrastive	**64.9**	35.6	35.3	62.5	85.7	81.5	43.3	63.7
DINOv2 [57]	Visual Grounding	57.5	14.7	15.9	**63.9**	86.7	**86.1**	47.5	59.4
Co-DETR [86]	Object Detection	48.4	14.2	14.8	58.6	**88.0**	82.1	48.6	55.3
SAM [30]	Image Segmentation	40.7	13.9	15.0	54.0	82.0	79.2	**49.3**	57.7
Pix2Struct [35]	Text Recognition	41.9	**57.3**	53.4	51.0	78.1	59.2	32.2	44.0
Deplot [43]	Chart Understanding	36.2	40.2	**55.8**	48.1	75.6	51.1	27.0	44.5
Vary [75]	Document Chart Parsing	28.1	47.8	41.8	42.6	69.1	21.6	16.0	40.9
BiomedCLIP [84]	Biomedical Contrastive	40.0	15.3	16.8	50.8	76.9	57.8	27.4	**65.1**
Plain fusion	-	63.4	46.5	48.9	63.0	86.4	85.7	45.3	64.7
MoVA	-	**65.9**	**59.0**	**56.8**	**64.1**	**88.5**	**86.4**	**49.8**	**66.3**

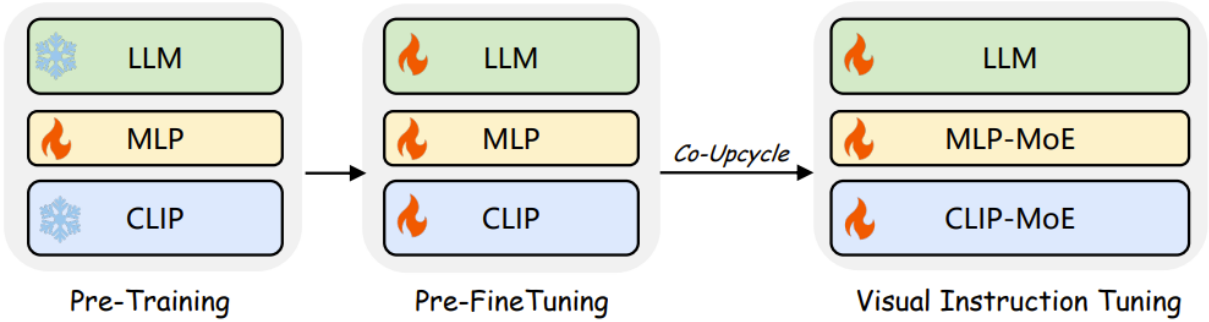


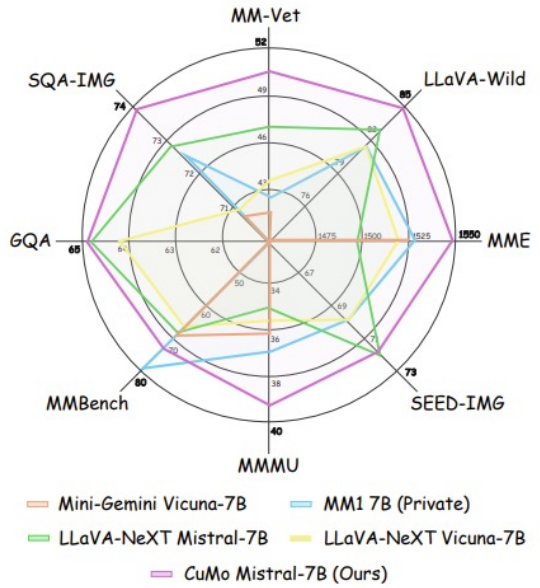
Vision Encoder Enhancement: CuMo

Method: Incorporate sparse Tok-K MoE blocks into CLIP using upcycling strategy









In this Talk - A Close Look at Vision

1 **Visual Tokenizer** What vision encoder is a good vision tokenizer for LMMs?

- Multimodal pretrained vision encoder CLIP is the best single one but still not sufficient
- Multi-crop strategy can support much higher-resolution using fixed-size encoder
- Mixture of vision encoders with CLIP can further enhance the performance

In this Talk - A Close Look at Vision

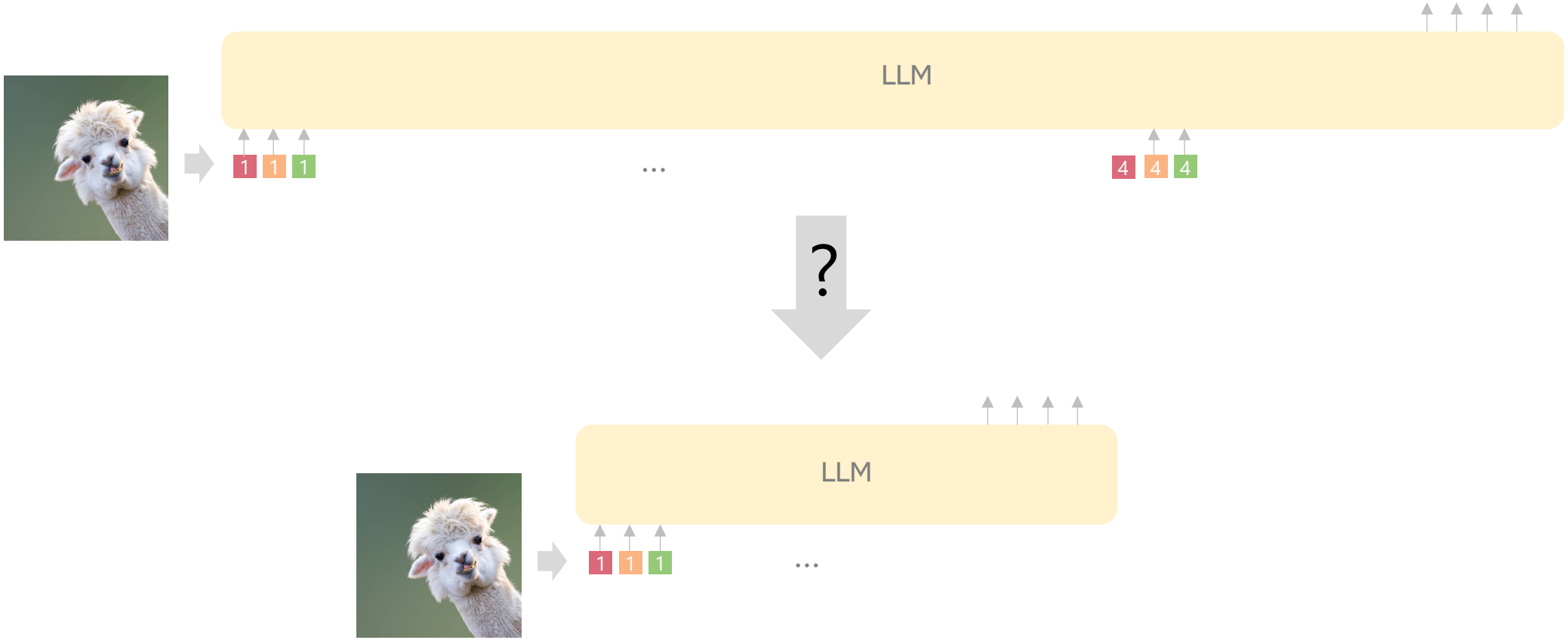
1 **Visual Tokenizer** What vision encoder is a good vision tokenizer for LMMs?

- Multimodal pretrained vision encoder CLIP is the best single one but still not sufficient
- Multi-crop strategy can support much higher-resolution using fixed-size encoder
- Mixture of vision encoders with CLIP can further enhance the performance

2 **Visual Tokens** How to cope with visual tokens for LLMs?

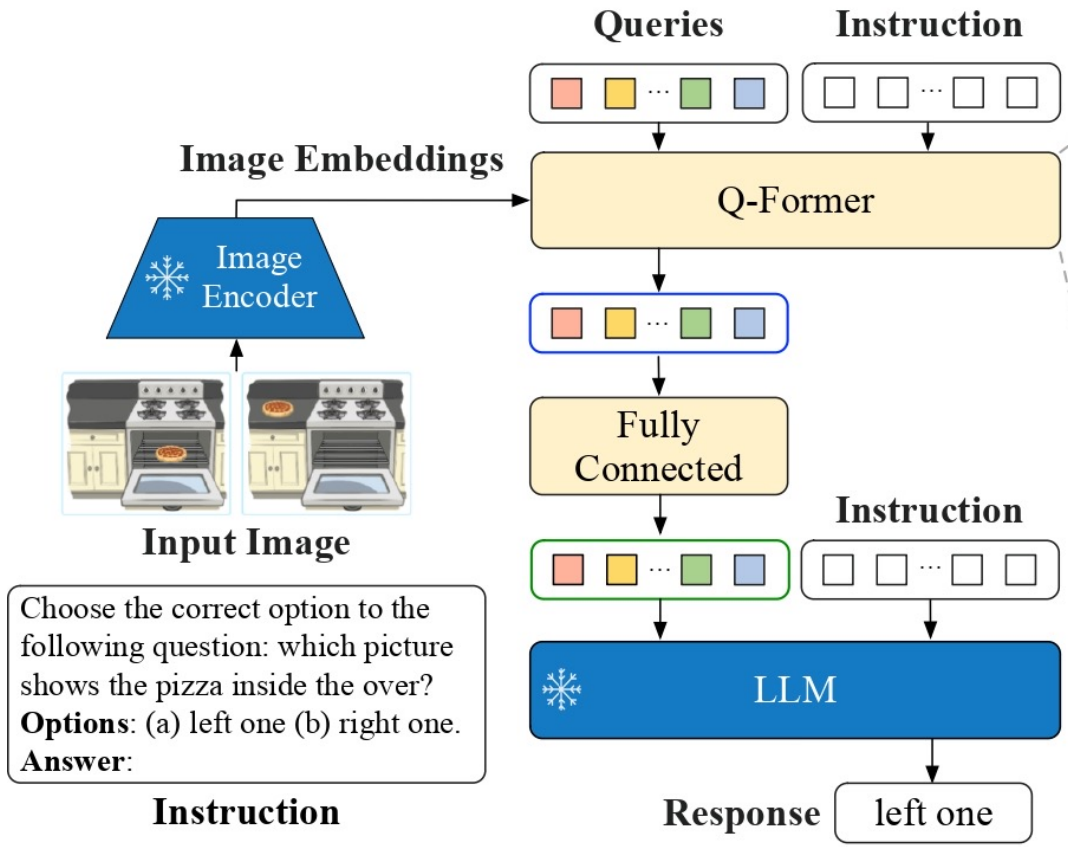
The Curse of Visual Tokens

A low-res image is worth 576 tokens, a high-res image is worth 3k tokens, a video is a disaster

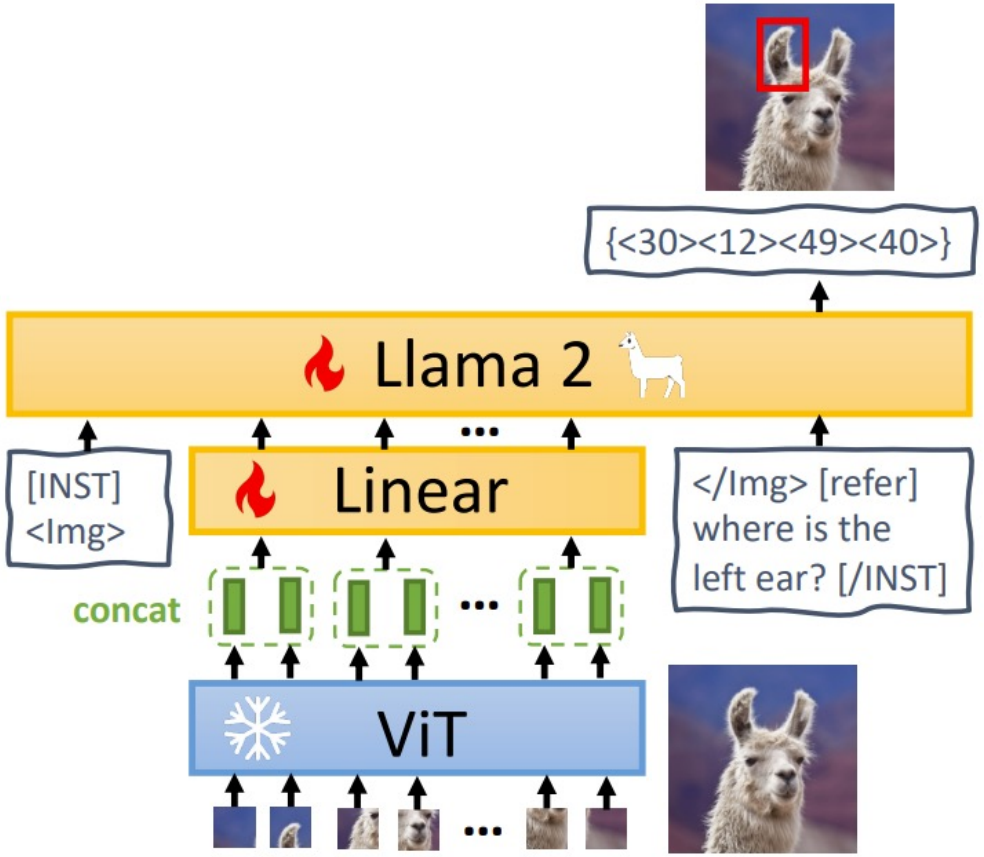


Early Strategies to Handle Visual Tokens

Q-Former: Encoder-Decoder architecture taking user-specific number of queries.



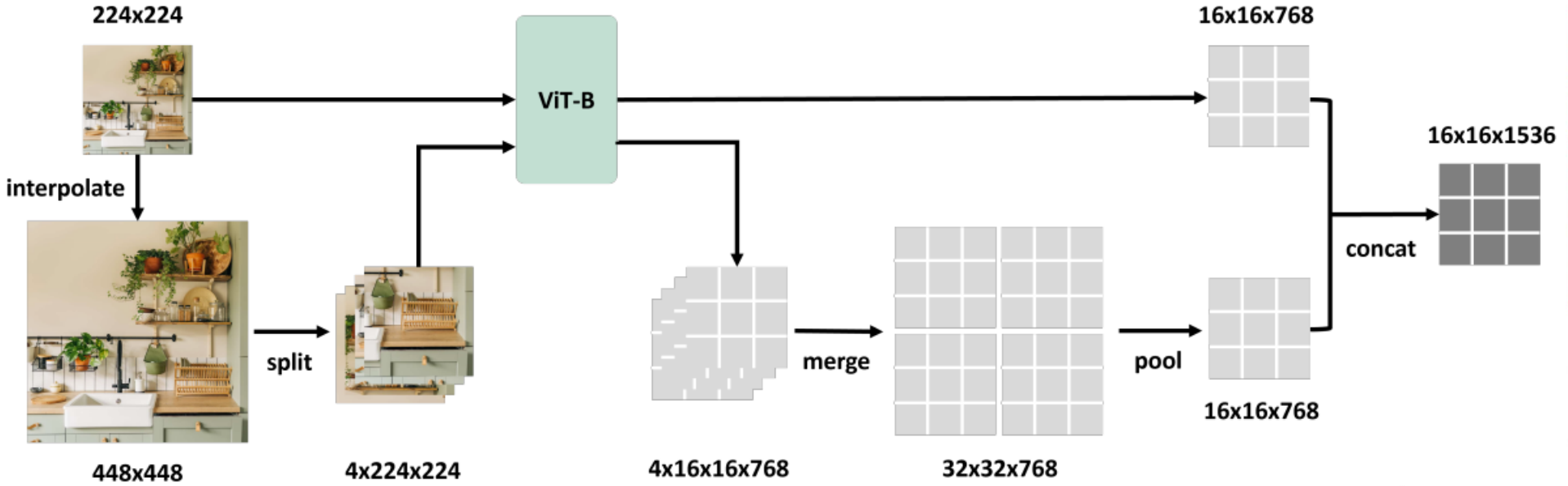
Concate-then-projection: concatenate 4 adjacent tokens and project into a single one



InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. Dai et al.
MiniGPT-v2: Large Language Model As a Unified Interface for Vision-Language Multi-task Learning. Chen et al.

Visual Tokens: Scaling on Scales (S2)

- Extract visual tokens for low-resolution image
- Extract visual tokens for high-resolution images and then merge-and-pool
- Concatenate low resolution image tokens and high-resolution image tokens



Q: What is the color of the water bottle?

GPT-4V: The water bottle on the ground is blue.

LLaVA-1.5: The color of the water bottle is blue.

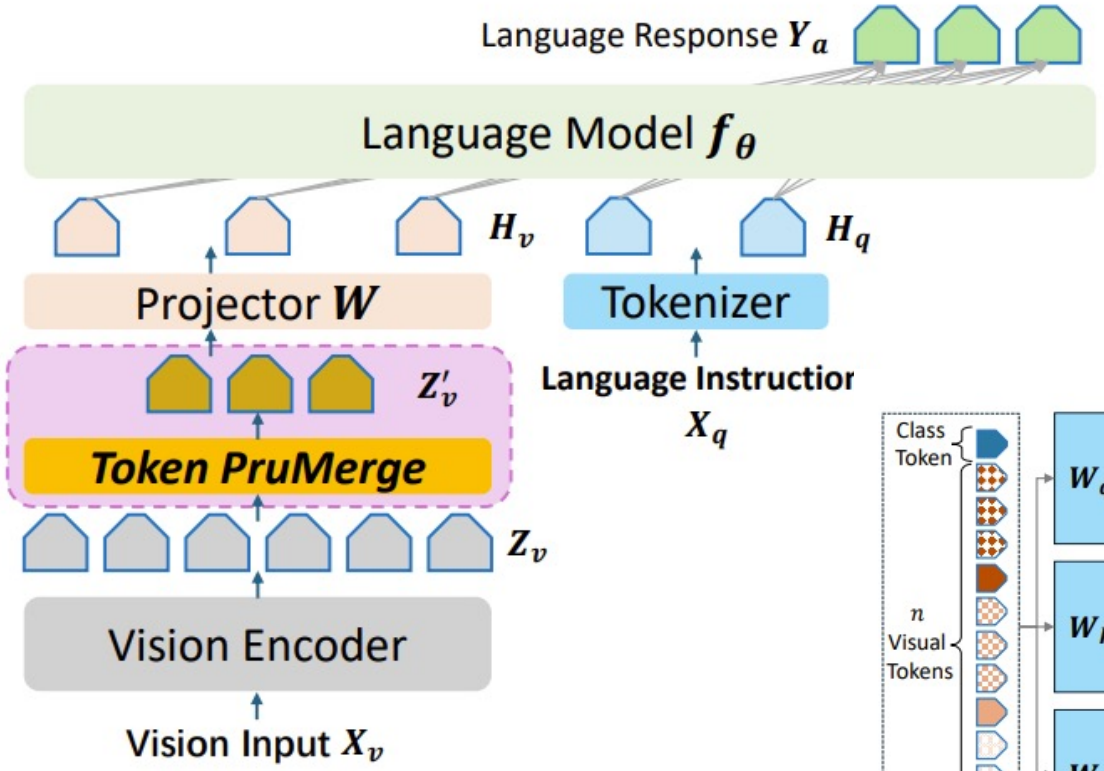
LLaVA-1.5-S²: The color of the water bottle is red.

- S2 scaling increases the image resolution taken by LMMs while producing same number of tokens.
- Significantly improves performance on V* benchmark

Model	Res.	#Tok	V^*_{Att}	V^*_{Spa}
LLaVA-1.5-7B [39]	336	576	43.5	56.6
- S^2 Scaling	1008	576	**51.3** (+7.8)	61.8 (+5.2)
LLaVA-1.5-13B [39]	336	576	41.7	55.3
- S^2 Scaling	1008	576	50.4 (+8.7)	**63.2** (+7.9)

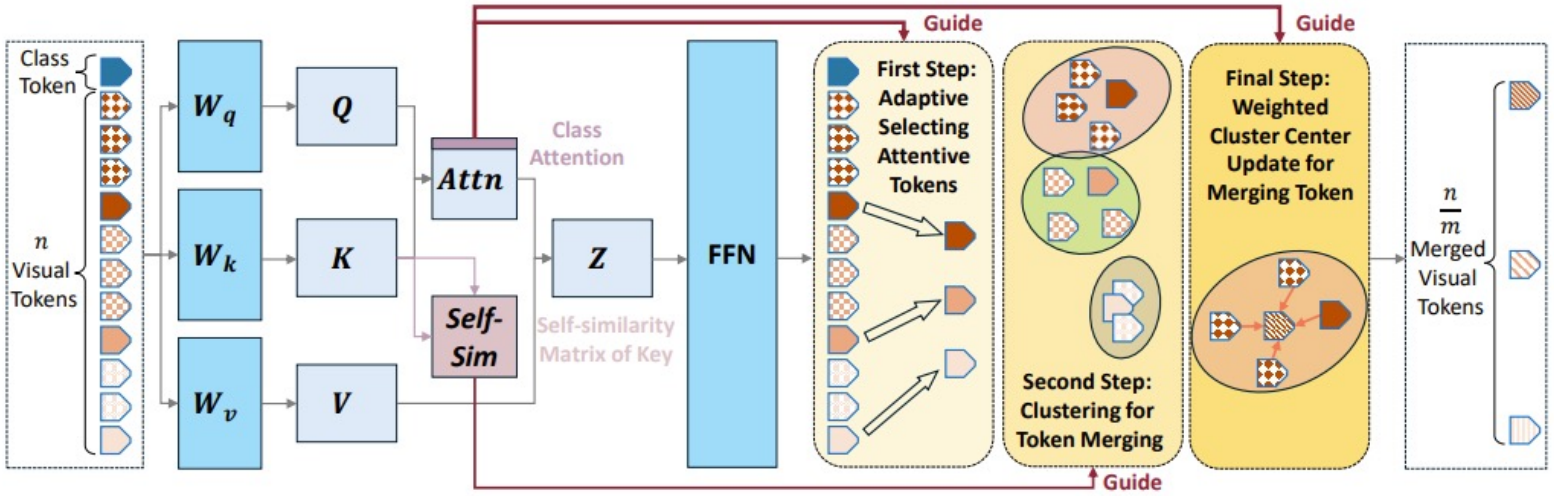
Visual Tokens: Adaptive Token Reduction PruMerge

Strategy: token reduction taking into account the spatial redundancy



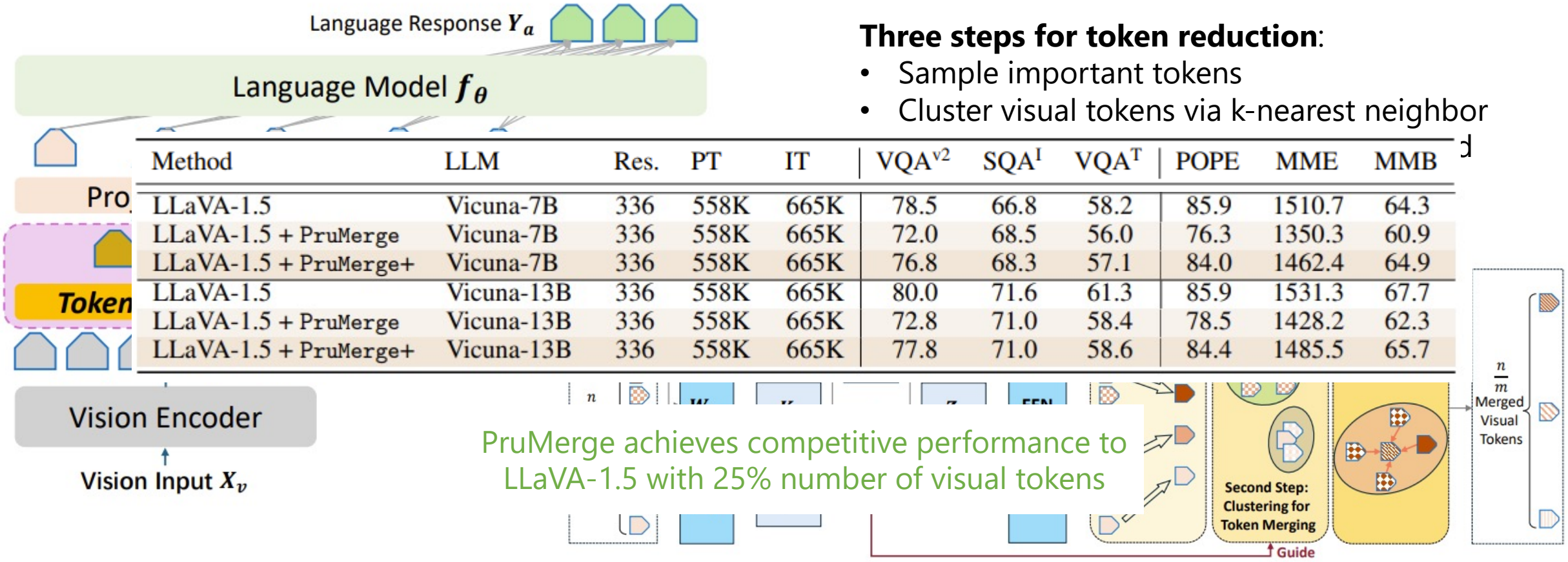
Three steps for token reduction:

- Sample important tokens
- Cluster visual tokens via k-nearest neighbor
- Adjust sampled visual tokens via weighted average for each cluster



Visual Tokens: Adaptive Token Reduction PruMerge

Strategy: token reduction taking into account the spatial redundancy



Method	LLM	Res.	PT	IT	VQA^{v2}	SQA^{I}	VQA^{T}	POPE	MME	MMB
LLaVA-1.5	Vicuna-7B	336	558K	665K	78.5	66.8	58.2	85.9	1510.7	64.3
LLaVA-1.5 + PruMerge	Vicuna-7B	336	558K	665K	72.0	68.5	56.0	76.3	1350.3	60.9
LLaVA-1.5 + PruMerge+	Vicuna-7B	336	558K	665K	76.8	68.3	57.1	84.0	1462.4	64.9
LLaVA-1.5	Vicuna-13B	336	558K	665K	80.0	71.6	61.3	85.9	1531.3	67.7
LLaVA-1.5 + PruMerge	Vicuna-13B	336	558K	665K	72.8	71.0	58.4	78.5	1428.2	62.3
LLaVA-1.5 + PruMerge+	Vicuna-13B	336	558K	665K	77.8	71.0	58.6	84.4	1485.5	65.7

Three steps for token reduction:

- Sample important tokens
- Cluster visual tokens via k-nearest neighbor

PruMerge achieves competitive performance to LLaVA-1.5 with 25% number of visual tokens

Visual Tokens: Matryoshka Multimodal Models (M3)

Strategy: learns to represent visual content as nested sets of visual tokens

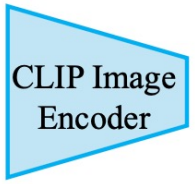
Gradually apply average pooling to the [H, W] visual features:

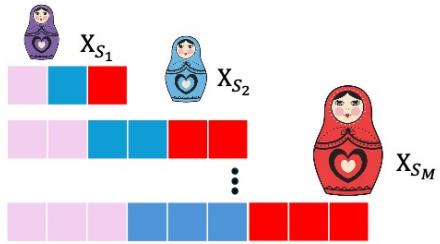
Resulting in visual features with shape $[H, W]$, $\left[\frac{H}{2}, \frac{W}{2}\right]$, $\left[\frac{H}{4}, \frac{W}{4}\right]$, $\cdots, [1,1]$

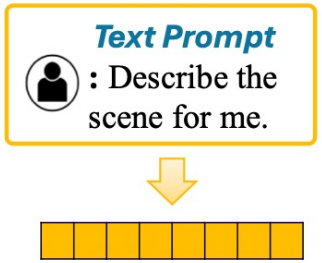
Average the language modelling loss upon all scales during training.

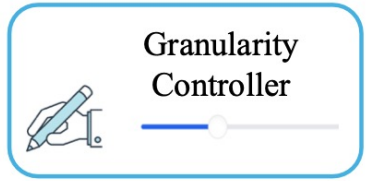
- User can control how many visual tokens to feed to LLMs.
- Similar performance to LLaVA-1.5 on MMBench with only 9 tokens
- COCO-style images require much less tokens than document images

Matryoshka Multimodal Models



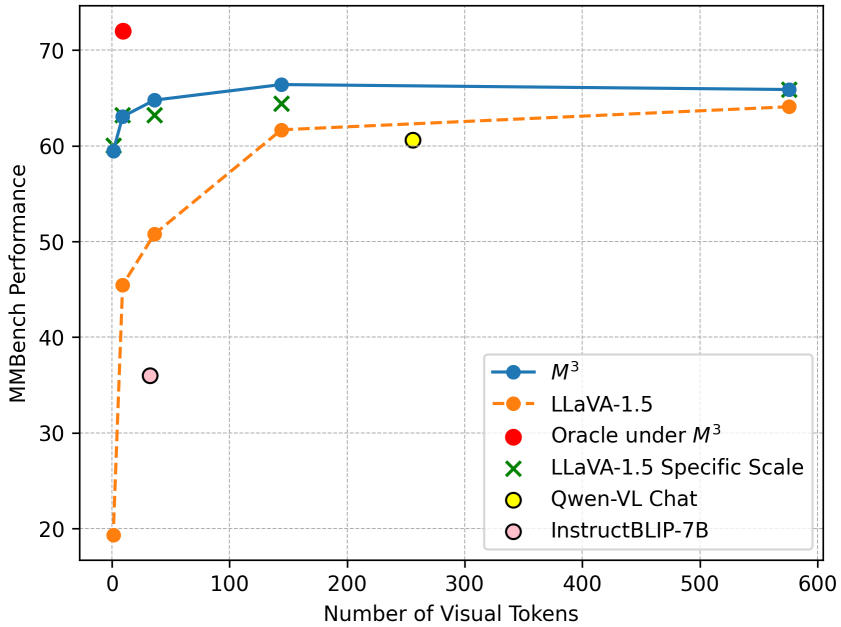






Large Language Model

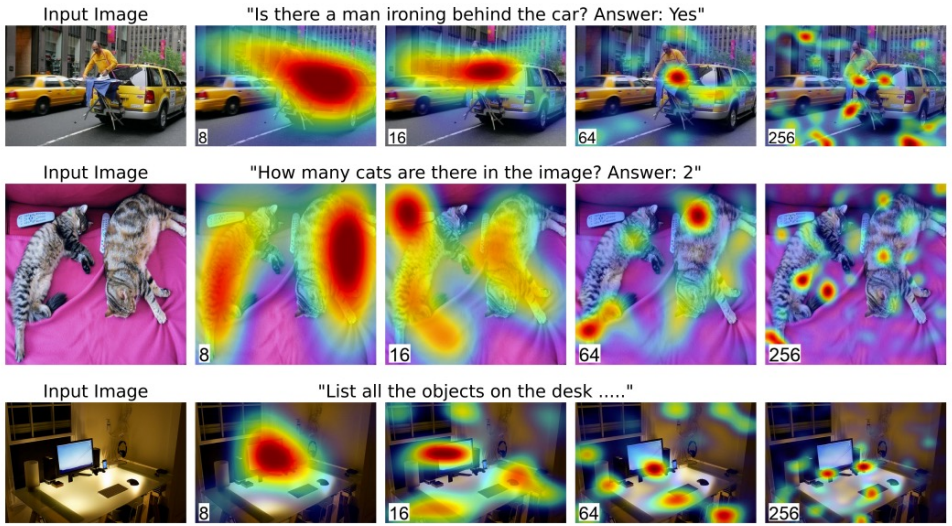
: There are a group of people standing in the ski facility, some of them are holding a green flag while other are …

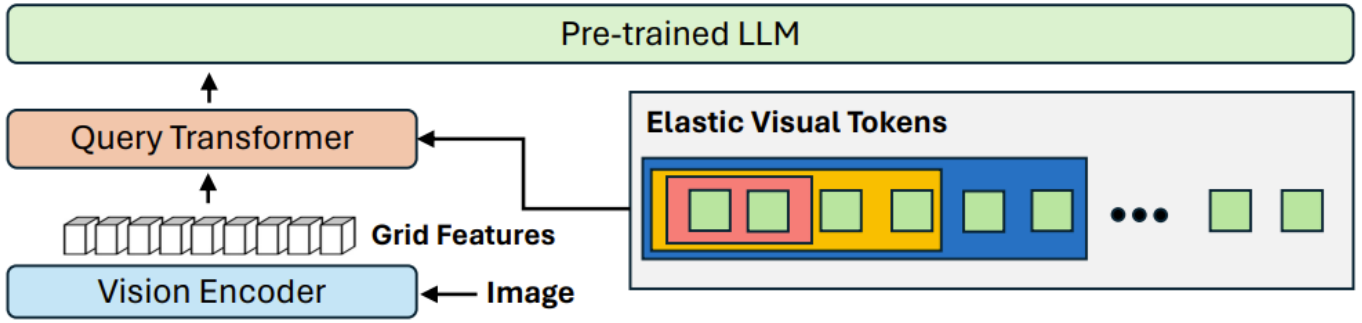


Visual Tokens: Matryoshka Query Transformer

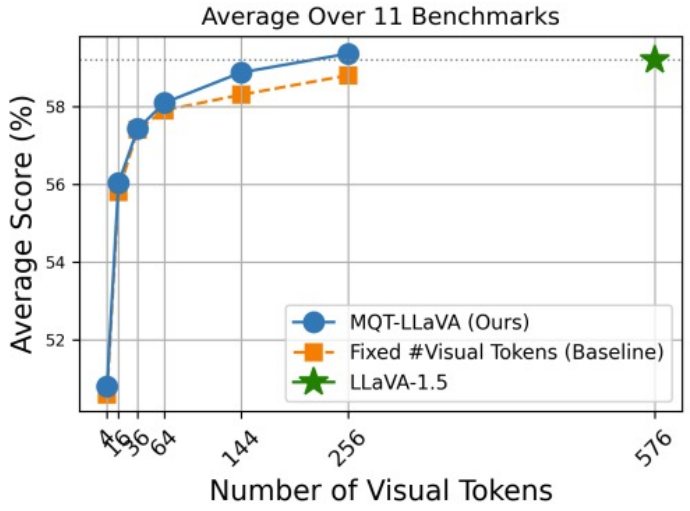
Strategy: learn a query transformer to extract visual tokens in an elastic manner.

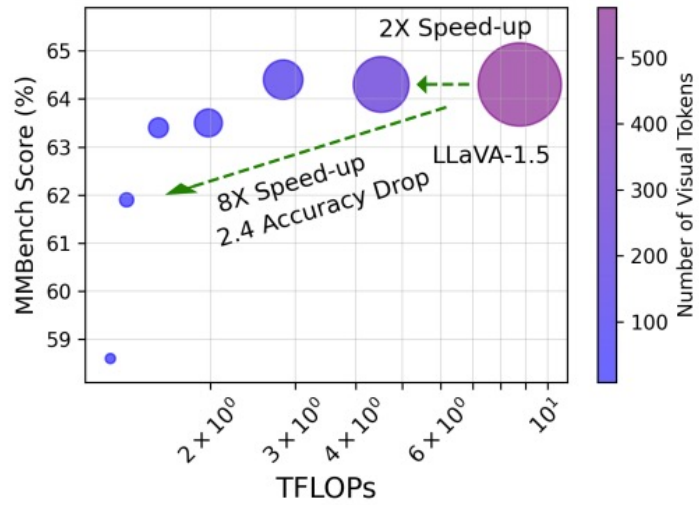
The model effectively concentrates on high-level concepts using fewer tokens and delves into low-level details with more tokens





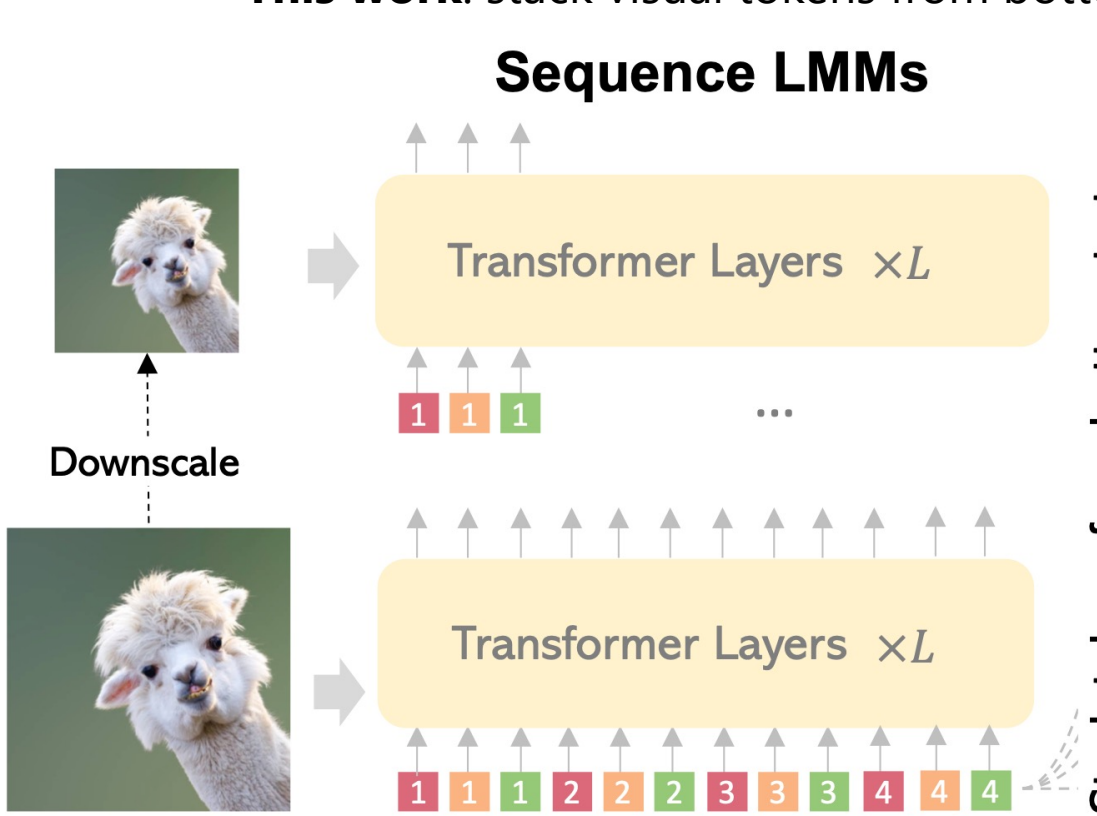
Improve efficiency by reducing #token but less drop on performance over 11 benchmarks.

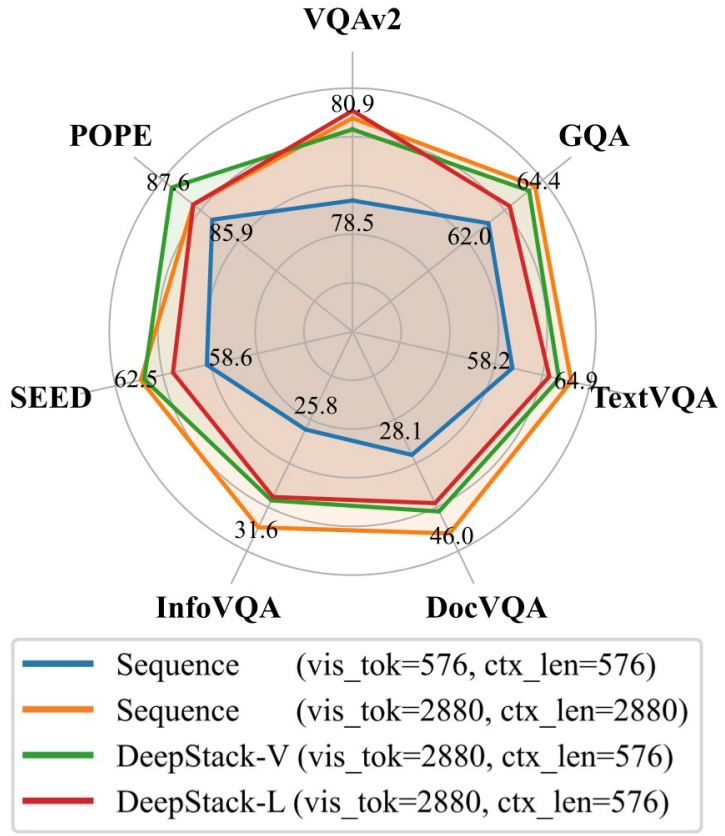




Visual Tokens: Deeply Stacking Visual Tokens

- **All previous works**: string tokens from left to right as a sequence – Sequence LMMs
- **This work**: stack visual tokens from bottom to top – DeepStack LMMs

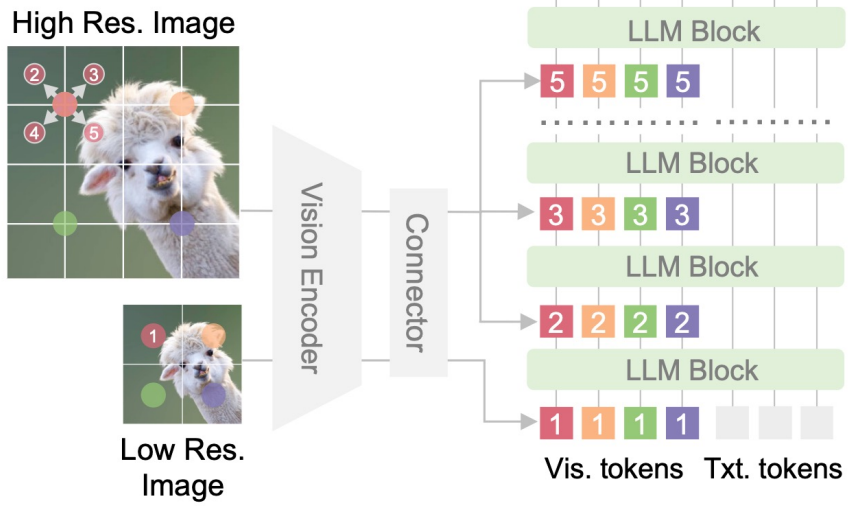




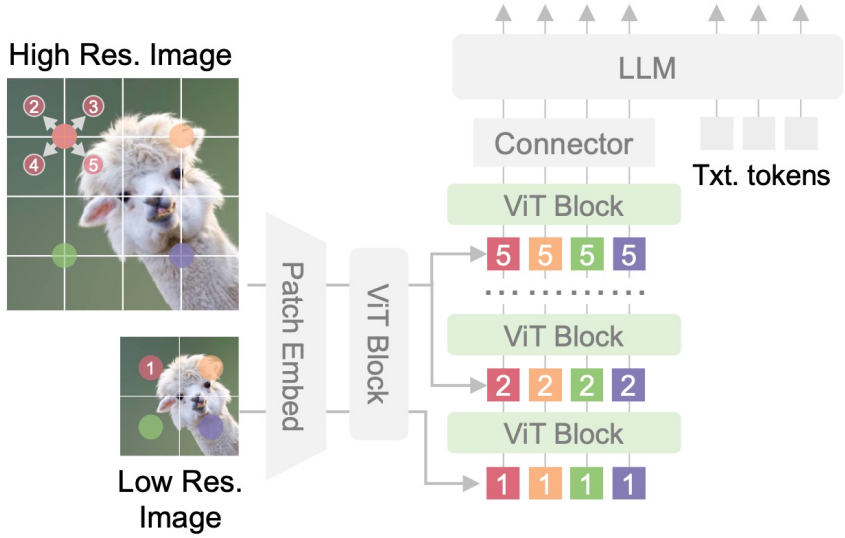
Using the same context length, DeepStack with 7B and 13B parameters surpass their counterparts by 2.7 and 2.9 on average across 9 benchmarks, respectively.

Visual Tokens: Deeply Stacking Visual Tokens

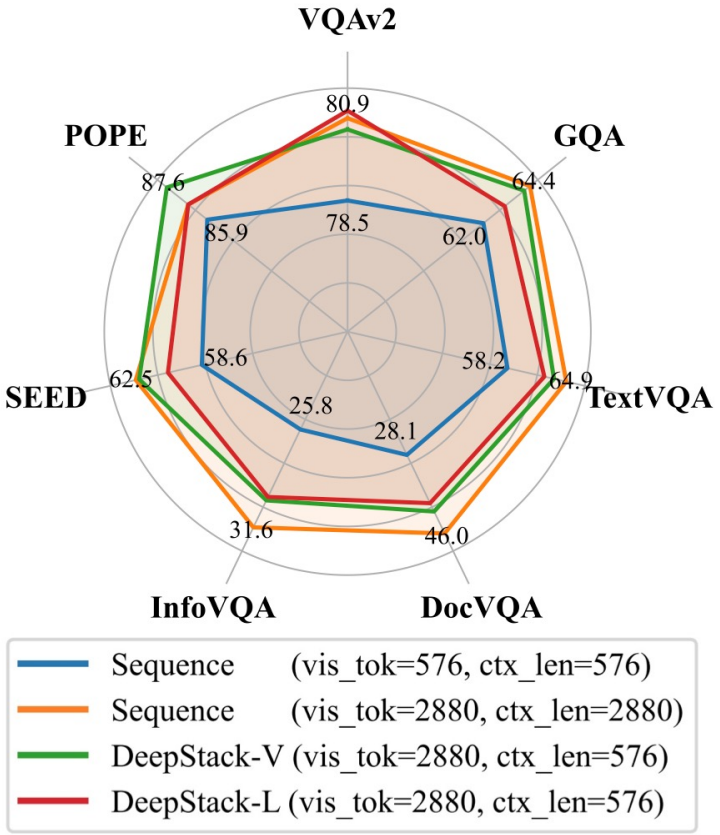
- **All previous works**: string tokens from left to right as a sequence – Sequence LMMs
- **This work**: stack visual tokens from bottom to top – DeepStack LMMs



DeepStack-L



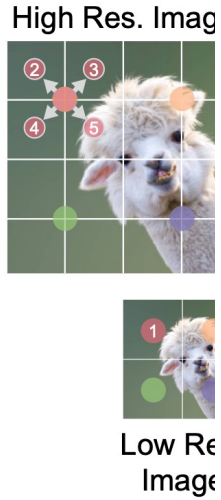
DeepStack-V



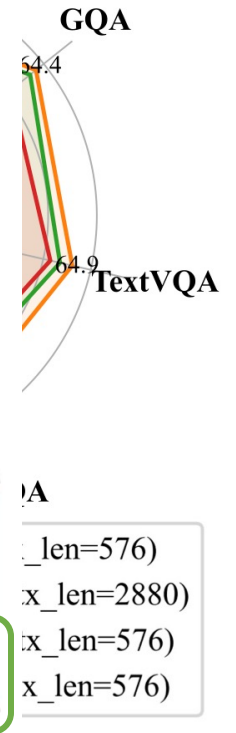
Using the same context length, DeepStack with 7B and 13B parameters surpass their counterparts by 2.7 and 2.9 on average across 9 benchmarks, respectively.

Visual Tokens: Deeply Stacking Visual Tokens

- **All previous works**: string tokens from left to right as a sequence → Sequence LMMs
-



Method	LLM	*Eff. Res.*	*Vis. Tok.*	*Cxt. Len.*	PT	SFT	*General VQA* VQA^{v2}	GQA	*Text-oriented VQA* Text VQA‡	Doc VQA‡	Info VQA‡	*LMM benchmarks* SEED (all)	POPE (all)	MM MU‡	MM Vet
BLIP-2 [43]	Vicuna-13B	224	32	32	129M	-	41.0	41.0	42.5			46.4	85.3	-	
InstructBLIP [16]	Vicuna-7B	224	32	32	129M	1.2M	–	49.2	50.1	-	-	53.4	-	-	-
InstructBLIP [16]	Vicuna-13B	224	32	32	129M	1.2M	–	49.5	50.7	-	-	78.9	-	-	-
Shikra [12]	Vicuna-13B	224	-	-	600K	5.5M	77.4*	-	-	-	-	-	-	-	-
IDEFICS-9B [37]	LLaMA-7B	224	-	-	353M	1M		50.9	38.4	-	-	-	-	-	-
IDEFICS-80B [37]	LLaMA-65B	224	-	-	353M	1M	60.0	45.2	-	-	-	-	-	-	-
Qwen-VL [5]	Qwen-7B	448	256	256	1.4B	50M	78.8*	59.3*	63.8	-	-	56.3	-	-	-
Qwen-VL-Chat [5]	Qwen-7B	448	256	256	1.4B	50M	78.2*	57.5*	61.5	-	-	58.2	-	-	-
VILA [47]	Llama2-7B	336	576	576	50M	1M	79.9*	62.3*	64.4	-	-	61.1	85.5	-	34.9
VILA [47]	Llama2-13B	336	576	576	50M	1M	80.8	63.3*	66.6	-	-	62.8	84.2	-	38.8
LLaVA-1.5 [49]	Vicuna-7B	336	576	576	558K	665K	78.5*	62.0*	58.2	28.1	25.8	58.6	85.9	35.3	30.5
LLaVA-1.5 [49]	Vicuna-13B	672	576	576	558K	665K	80.0*	63.3*	61.3	30.3	28.4	61.6	85.9	34.8	35.4
LLaVA-Next [50]	Vicuna-7B	672	2880	2880	558K	**765K**	81.8*	64.2*	64.9	74.4*	37.1*	64.7	86.5	35.1	44.1
LLaVA-Next [50]	Vicuna-7B	672	2880	2880	558K	**765K**	82.8*	65.4*	66.9	77.5*	44.5*	65.6	86.2	35.9	49.1
DeepStack-V	Vicuna-7B	672	2880	576	558K	665K	80.4*	64.1*	63.5	41.0	30.0	62.3	87.6	34.9	33.0
DeepStack-V	Vicuna-13B	672	2880	576	558K	665K	81.1	64.2*	63.9	41.7	33.1	63.0	86.6	34.7	31.1
DeepStack-L	Vicuna-7B	672	2880	576	558K	665K	79.5*	63.1*	62.4	39.1	29.8	60.6	86.7	35.7	29.9
DeepStack-L	Vicuna-13B	672	2880	576	558K	665K	80.9*	64.2*	64.6	41.5	33.0	63.5	87.7	35.2	35.9
DeepStack-L-HD†	Vicuna-7B	1344	14400	2880	558K	748K	82.0*	65.2*	66.7	78.8*	41.2*	63.6	86.5	35.6	37.5
DeepStack-L-HD†	Vicuna-13B	1344	14400	2880	558K	748K	83.0*	66.2*	68.7	81.0*	45.2*	65.1	86.7	33.4	39.3



Using the same context length, DeepStack with 7B and 13B parameters surpass their counterparts by 2.7 and 2.9 on average across 9 benchmarks, respectively.

In this Talk - A Close Look at Vision

1 **Visual Tokenizer** What vision encoder is a good vision tokenizer for LMMs?

- Multimodal pretrained vision encoder CLIP is the best single one but still not sufficient
- Multi-crop strategy can support much higher-resolution using fixed-size encoder
- Mixture of vision encoders with CLIP can further enhance the performance

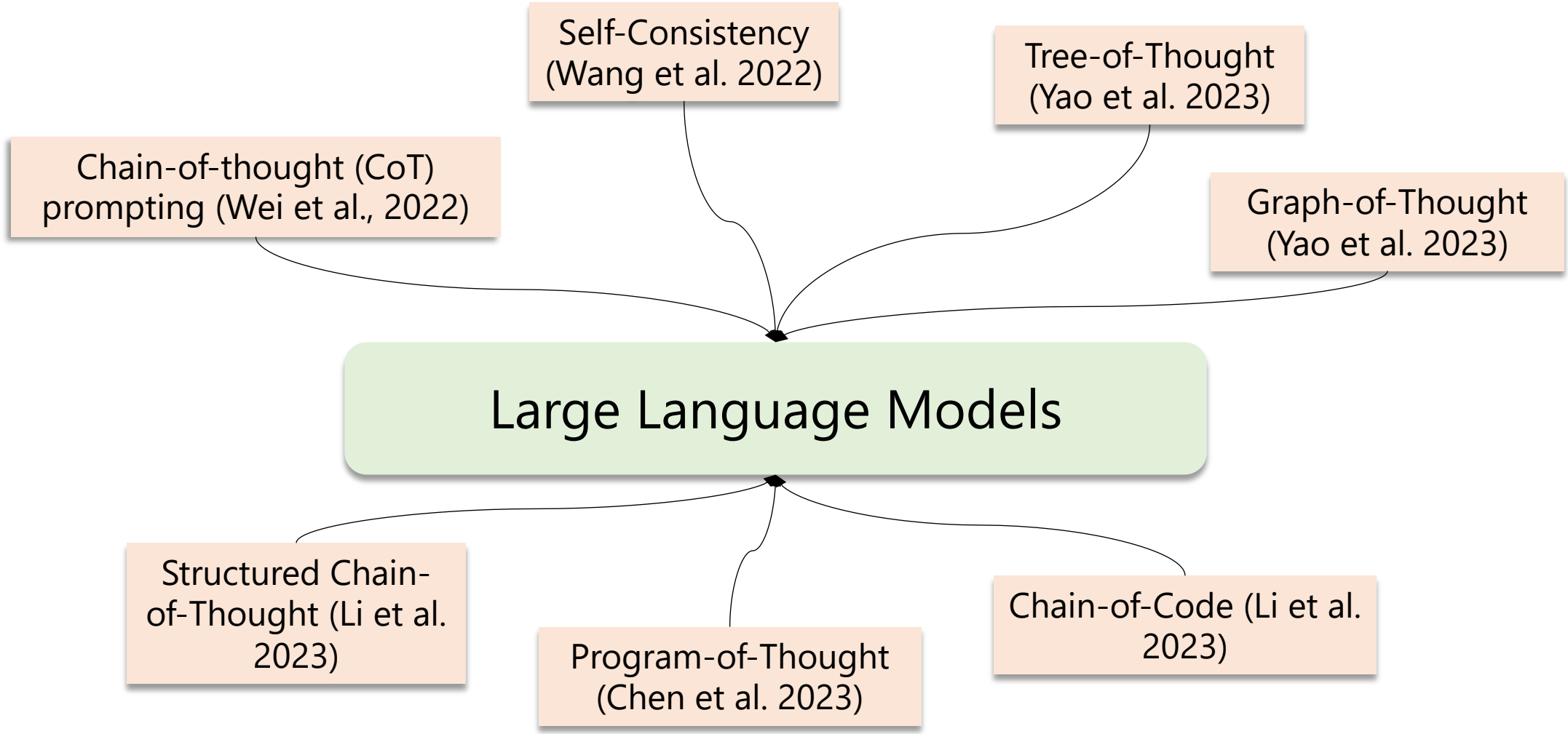
2 **Visual Tokens** How to cope with visual tokens for LLMs?

- Early strategies like Q-Former and Concatenate-then-projection
- Token reduction by adaptive sampling, nested and elastic organization
- Stack tokens from bottom to top, instead of only stringing tokens from left to right

In this Talk - A Close Look at Vision

1. **Visual Tokenizer** — What vision encoder is a good vision tokenizer for LMMs?
   - Multimodal pretrained vision encoder CLIP is the best single one but still not sufficient
   - Multi-crop strategy can support much higher-resolution using fixed-size encoder
   - Mixture of vision encoders with CLIP can further enhance the performance
2. **Visual Tokens** — How to cope with visual tokens for LLMs?
   - Early strategies like Q-Former and Concatenate-then-projection
   - Token reduction by adaptive sampling, nested and elastic organization
   - Stack tokens from bottom to top, instead of only stringing tokens from left to right
3. **Visual Prompting** — How to perform visual prompting for LMMs as for text?

Text Prompting for LLMs



Text Prompting for LLMs

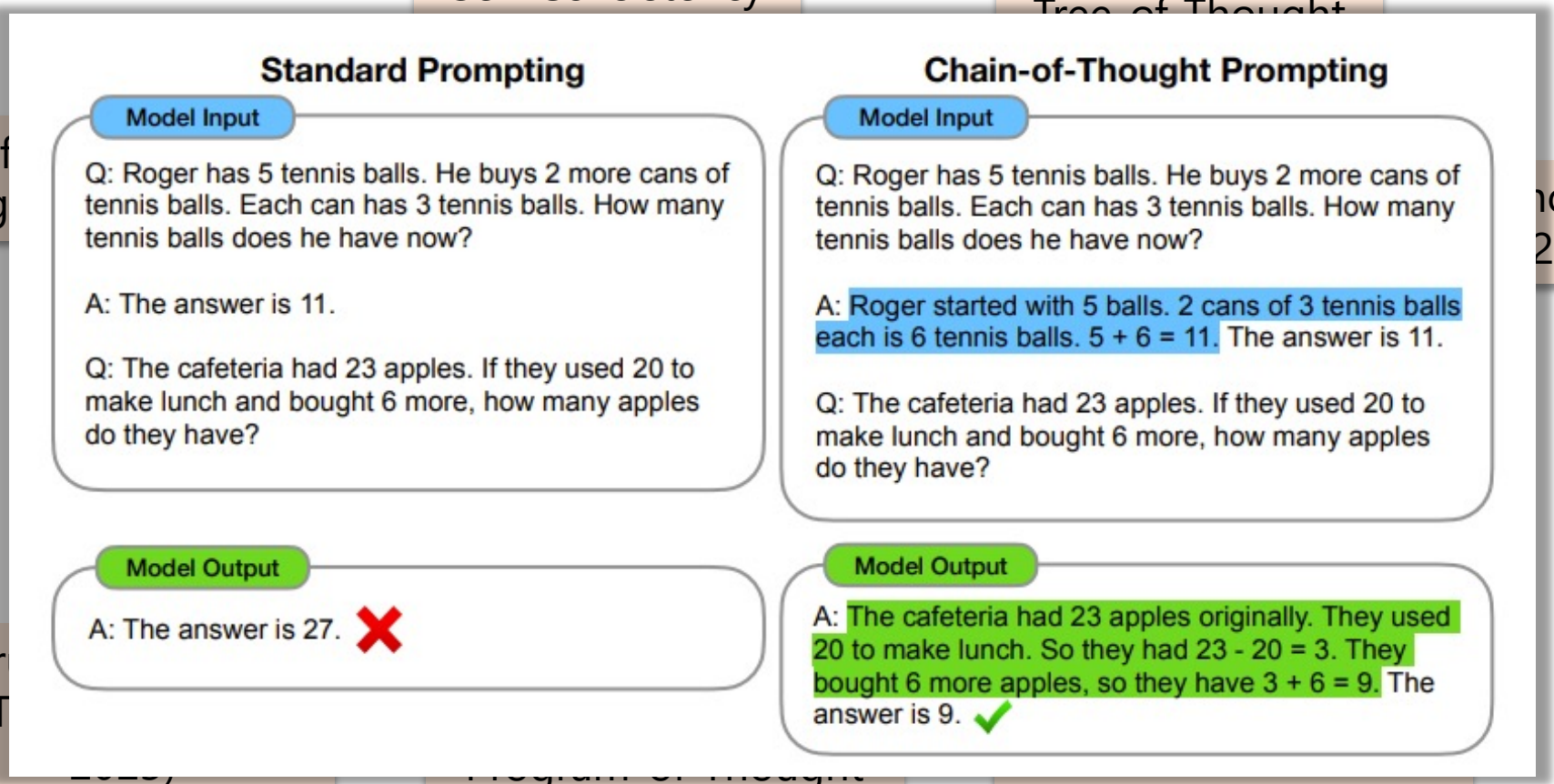
Self-Consistency

Chain-of
prompting

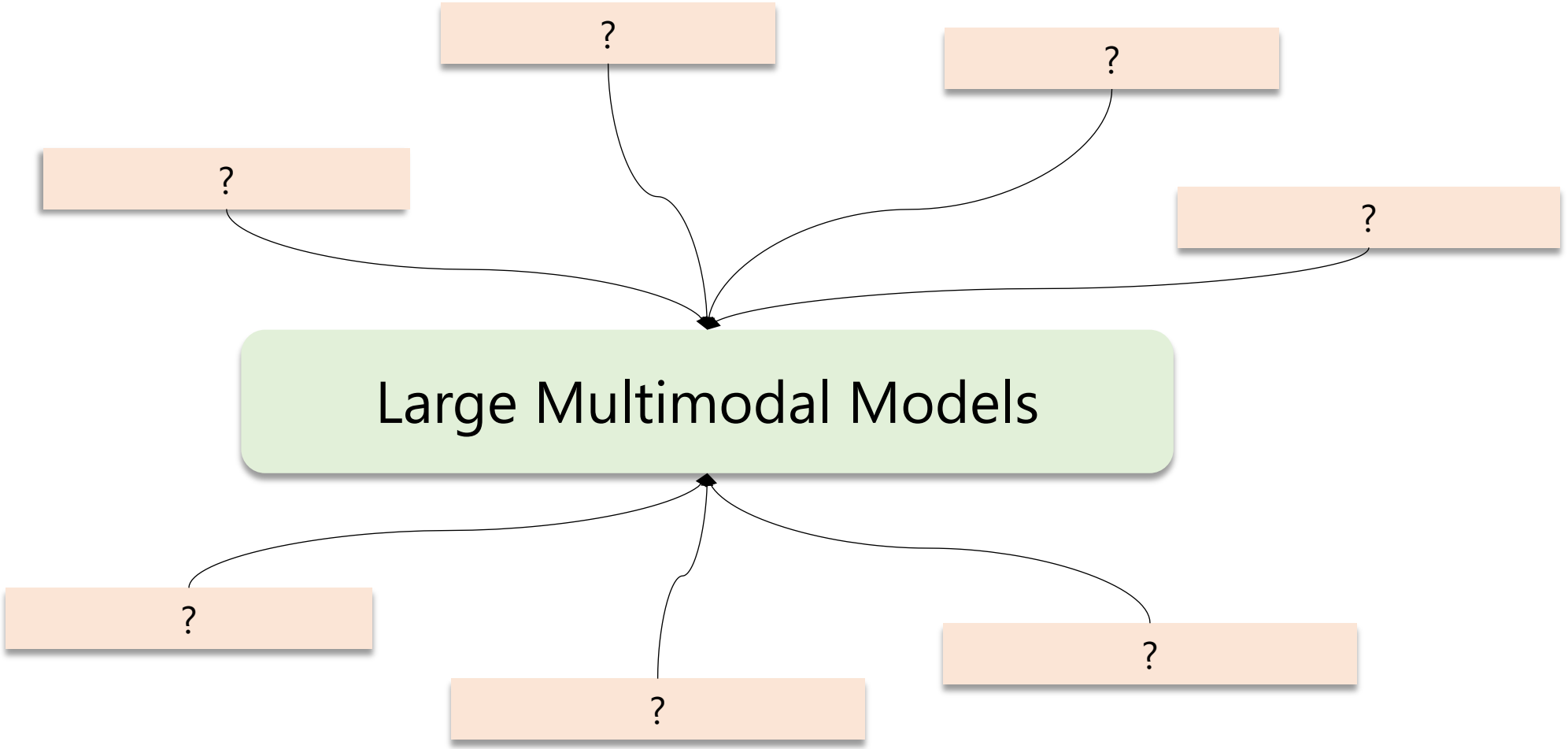
ought
2023)

Str
of-T

(Chen et al. 2023)

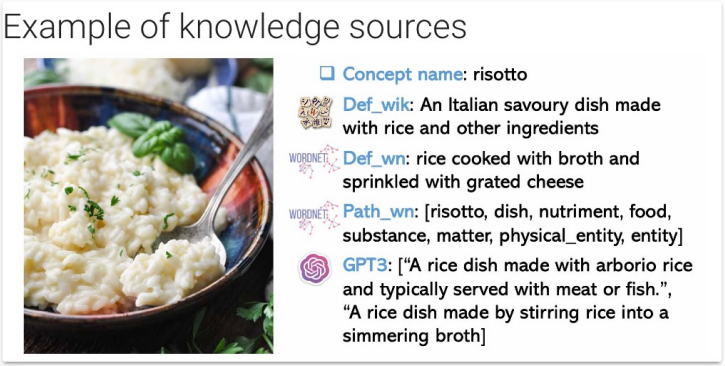


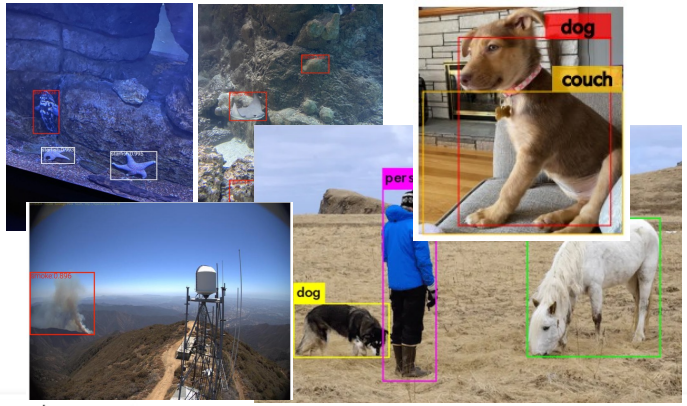
Visual Prompting

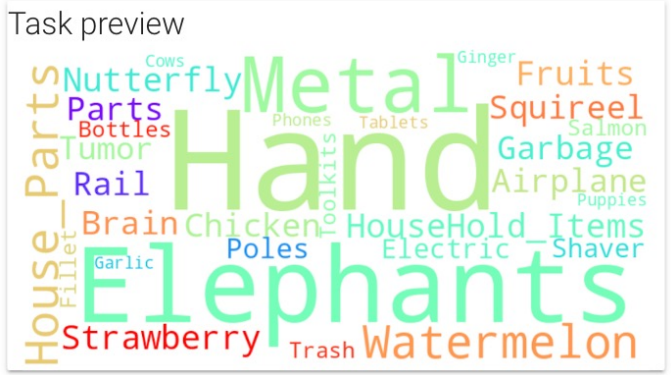


Why Visual Prompting Would Benefit?

Images are not only semantic, but also **visual & spatial**

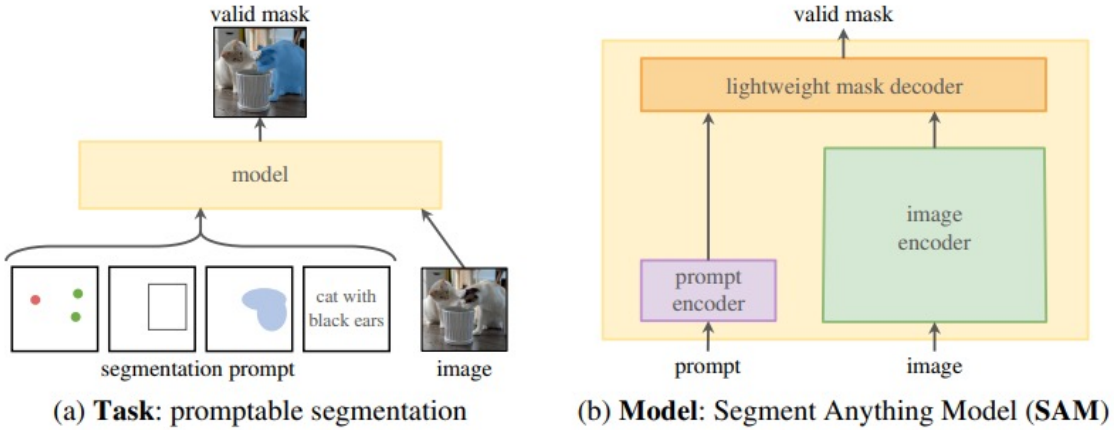




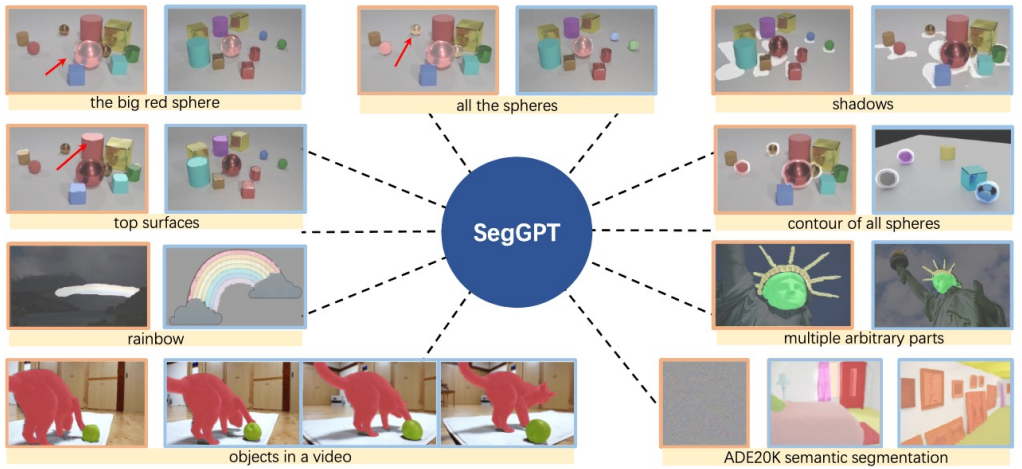


Visual Prompting for Vision Foundations

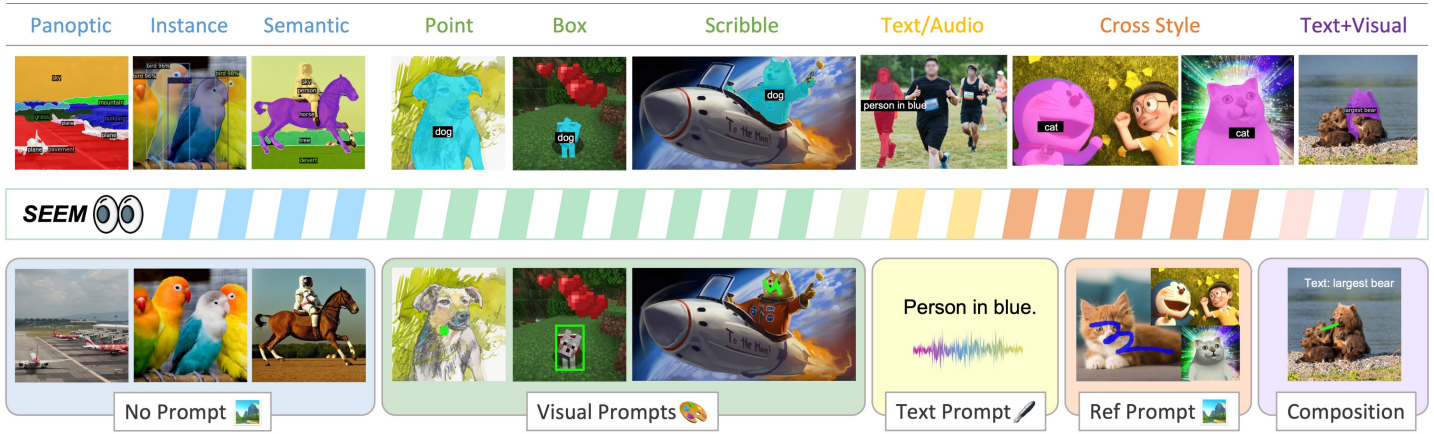
Visual prompting for segmentation



SAM. Kirillov et al. 2023



SegGPT. Wang et al. 2023

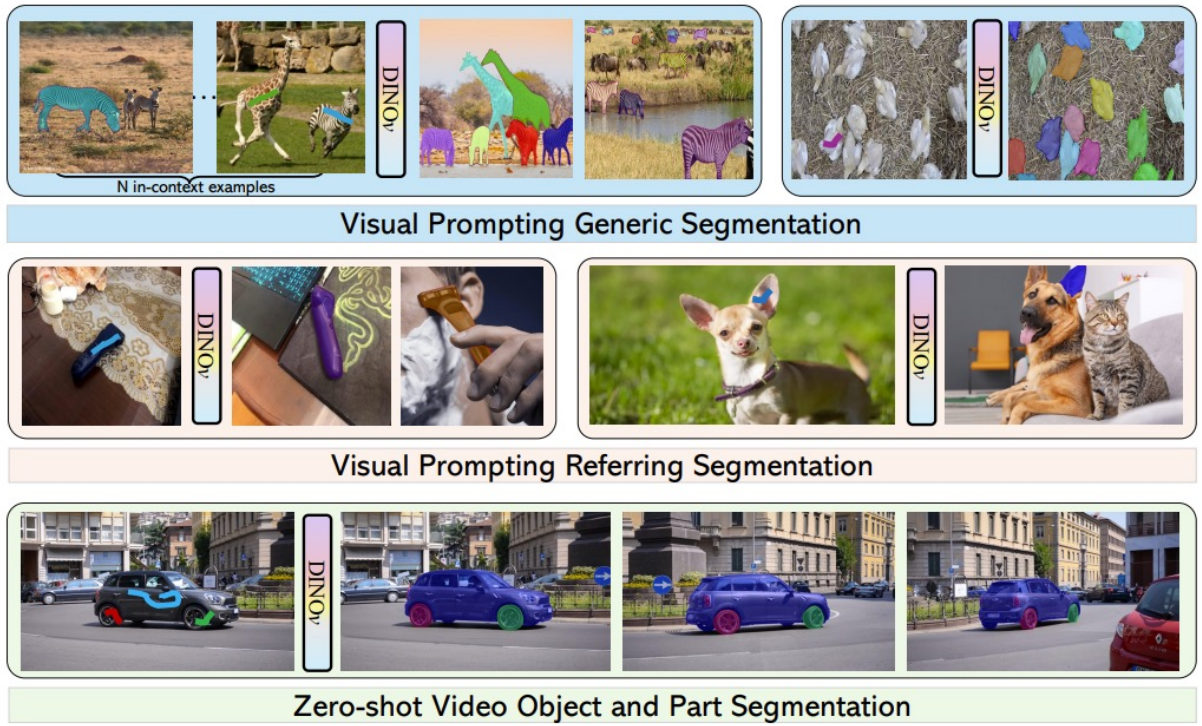


SEEM. Zou et al. 2023

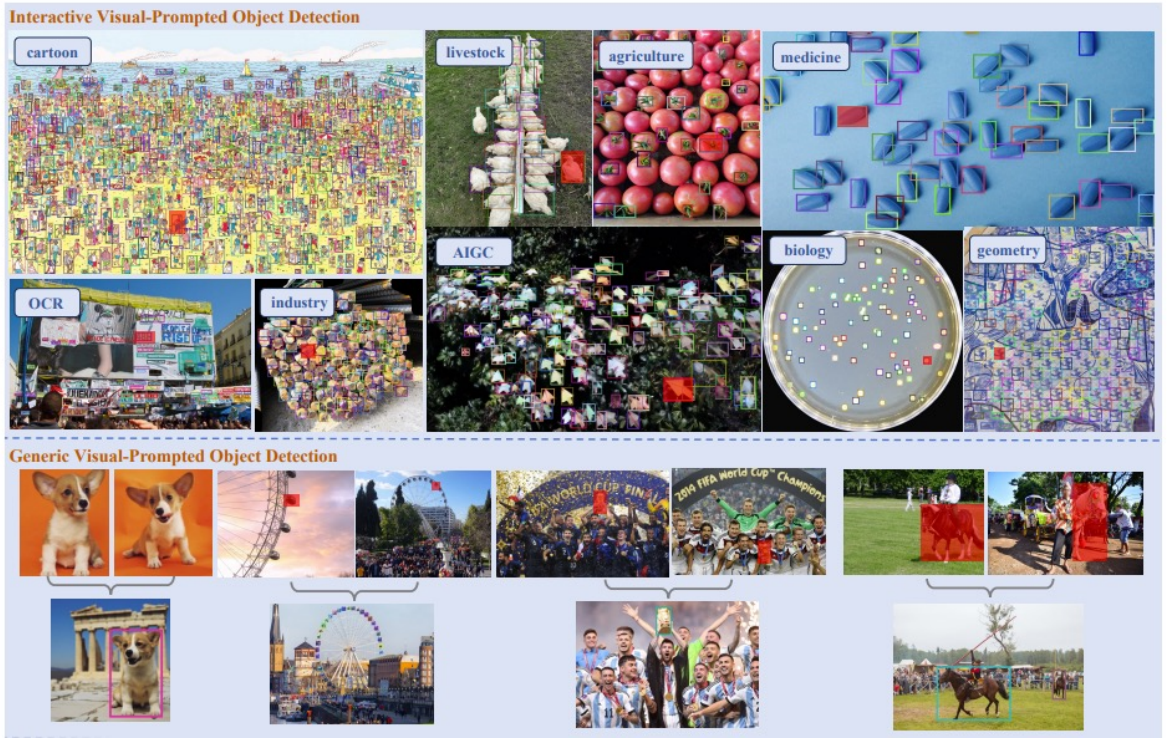
IMProv. Xu et al. 2023

Visual Prompting for Vision Foundations

Visual prompting for object detection, grounding and counting



DINOv. Li et al. 2024



T-Rex2. Jiang et al. 2024

Two Types of Visual Prompting for LVFs

Visual Feature Prompting

1. Extract visual features as the prompts
2. Need additional module to encode the visual prompts
3. Require some additional annotations

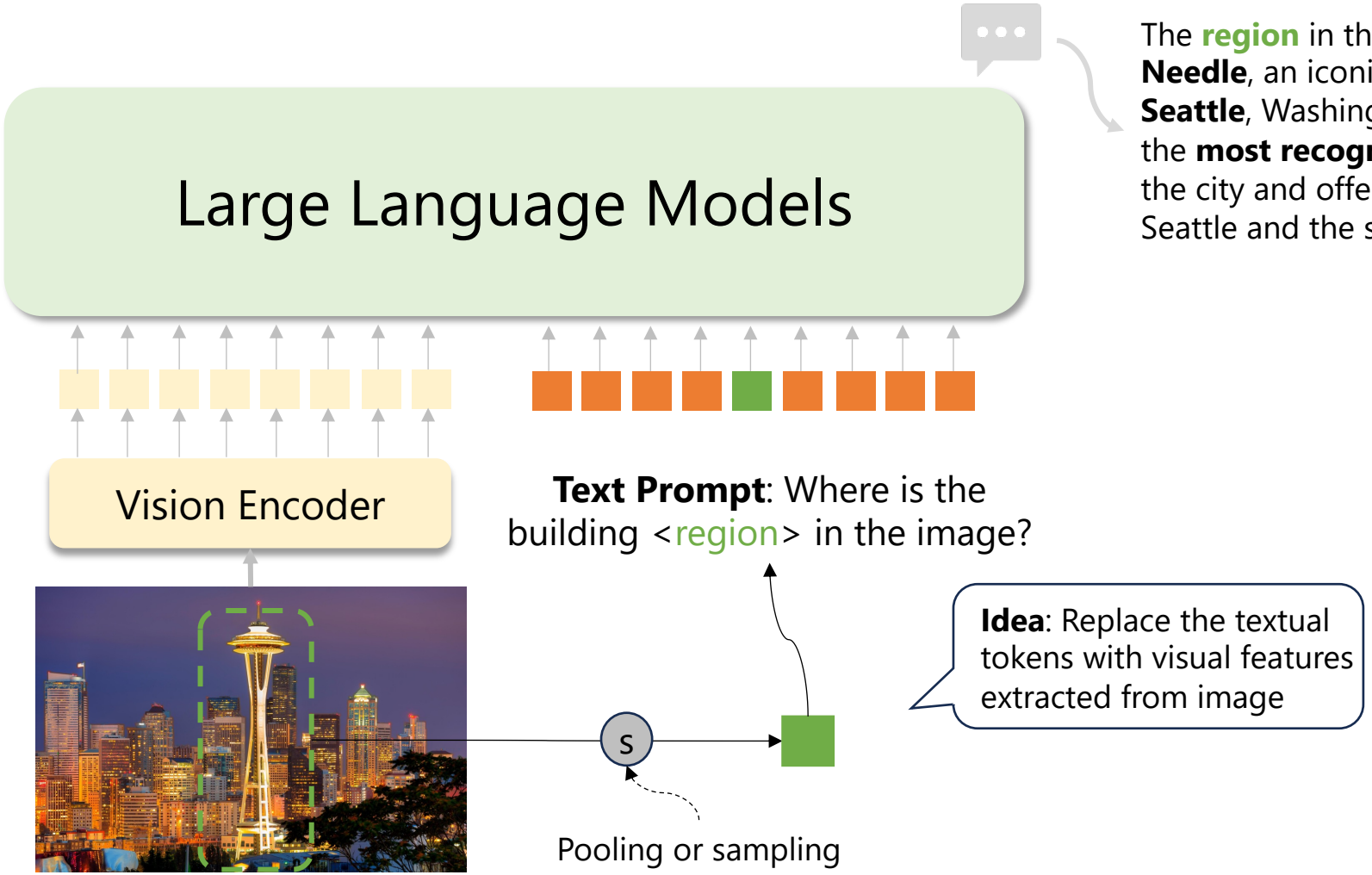
(e.g. SAM/SEEM/DINOv)

Visual Pixel Prompting

1. Directly overlay prompts on the images
2. Organic to the original LMM systems
3. Rely on the original emerging capability of LMMs

(e.g. SegGPT/IMProv)

Visual Feature Prompting for LMMs

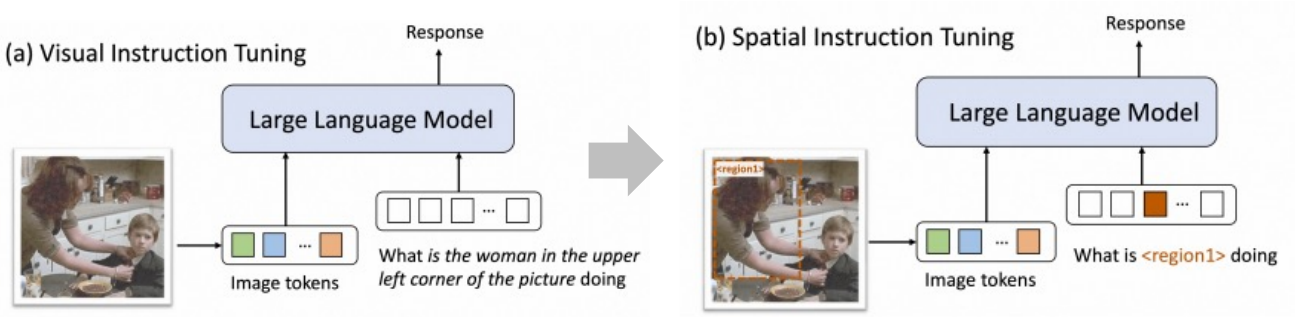


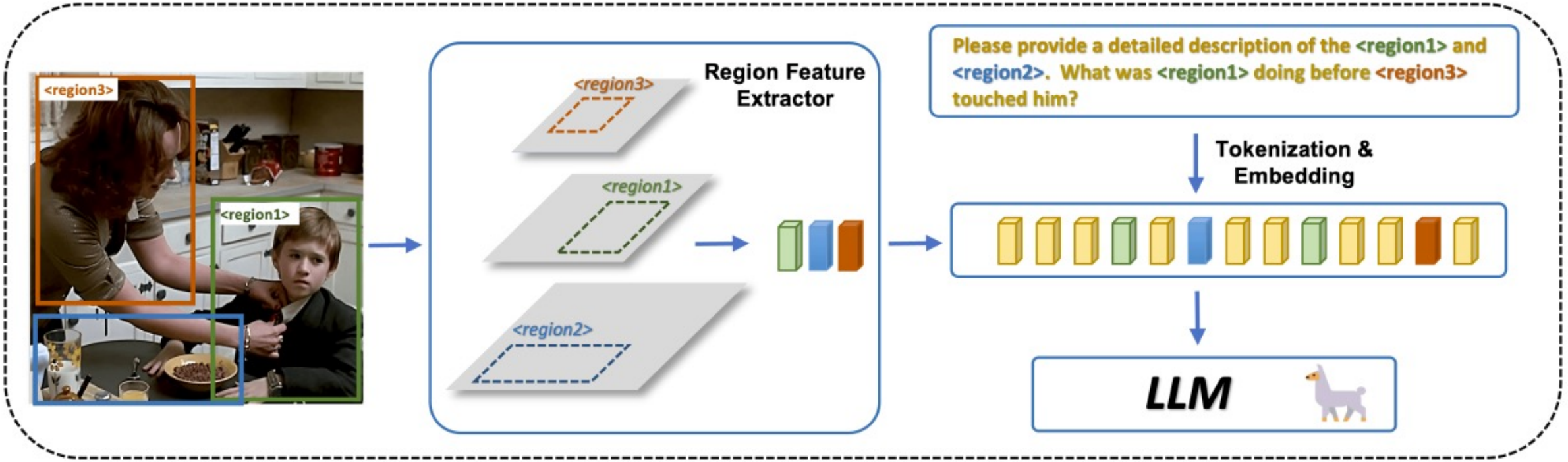
Visual Feature Prompting for LMMs

- **Find-grained Datasets**:
  - COCO, Lin et al.
  - RefCOCO/+, Yu et al.
  - Visual Genome, Krishna et al.
  - VCR, Zellers et al.
  - Object365, Shao et al.
  - Flickr30k-Entities, Plummer et al.
  - SAM, Kirillov et al.
- **Supported Prompts:**
  - Point
  - Box
  - Mask
  - Stroke
- **Feature samplers:**
  - RoIAlign
  - RoIPooling
  - SA-Sampler

Visual Feature Prompting: GPT4RoI

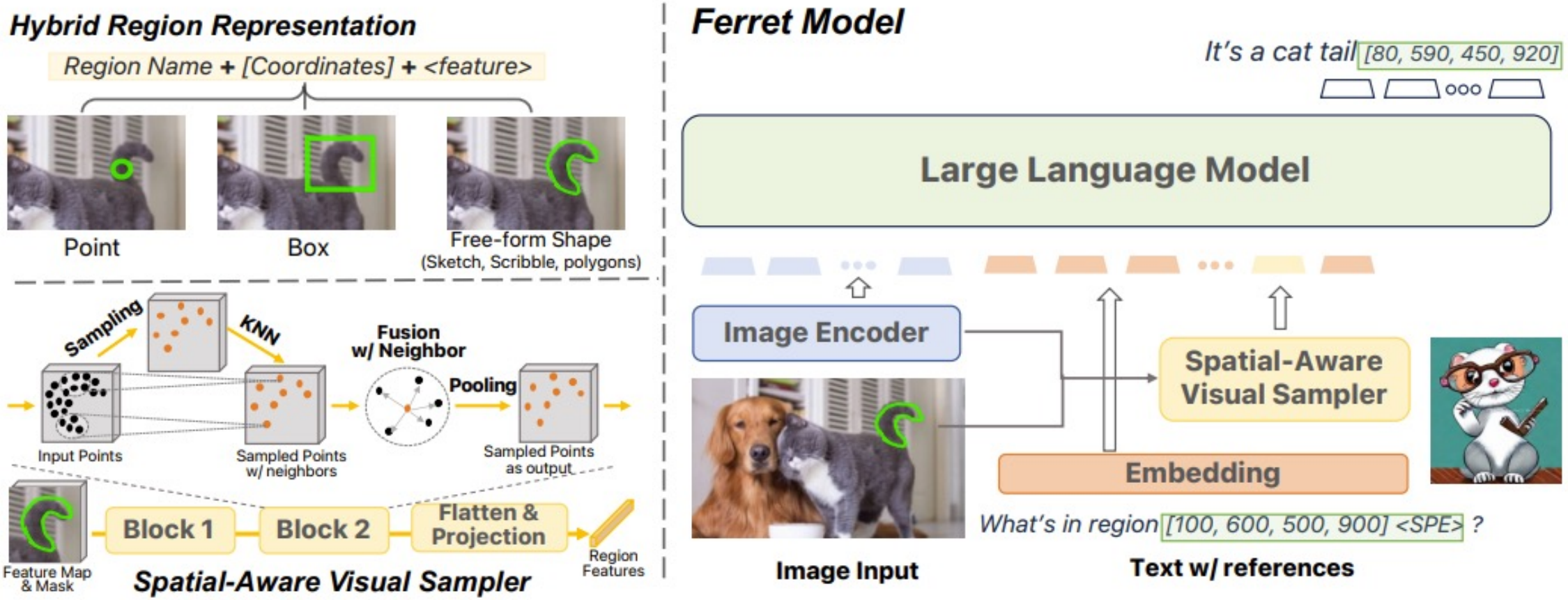
- **Technical Design:**
  - RoIAlign extracts multilevel region features
  - Adaptive to bounding box sizes
  - Interleaved prompts with text and visual tokens





Visual Feature Prompting: FERRET

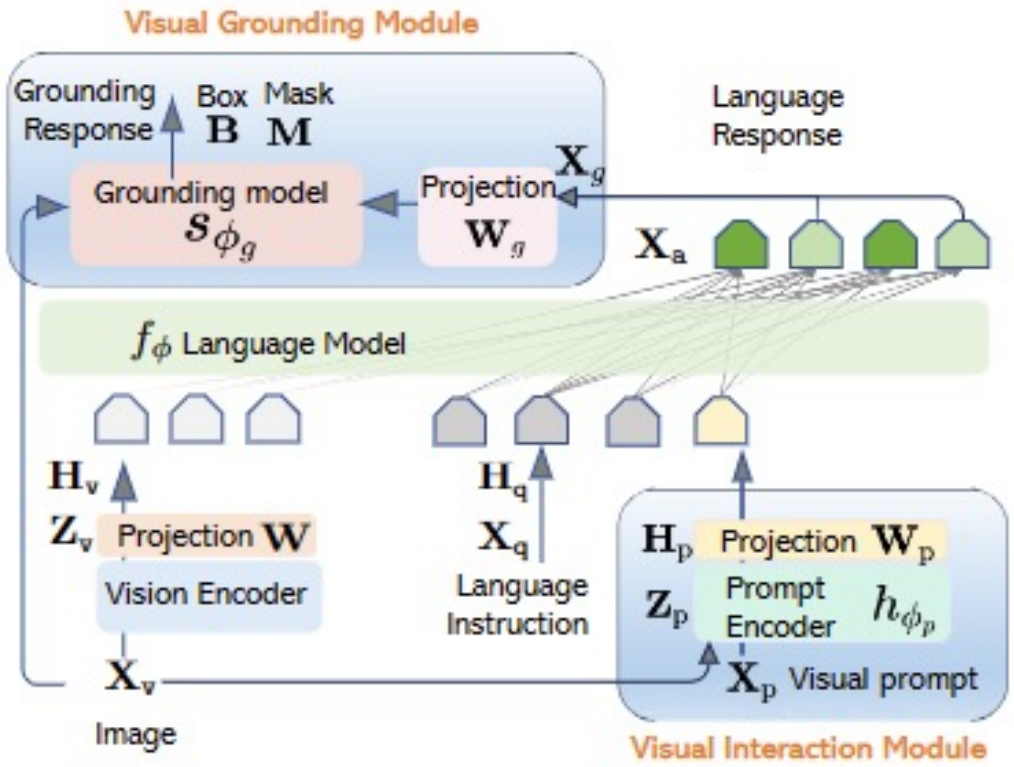
- Proposed a spatial-aware visual sampler
- Used text-formatted coordinates in both text prompts and outputs
- Trained the model with a large amount of fine-grained datasets



Ferret: Refer and Ground Anything Anywhere at Any Granularity. You et al.

Visual Feature Prompting: LLaVA-Grounding

- Use an extra **visual grounding module** taking outputs from LLMs as the condition
- Use a **visual interaction module** to encode the visual prompts and feed them into LLMs



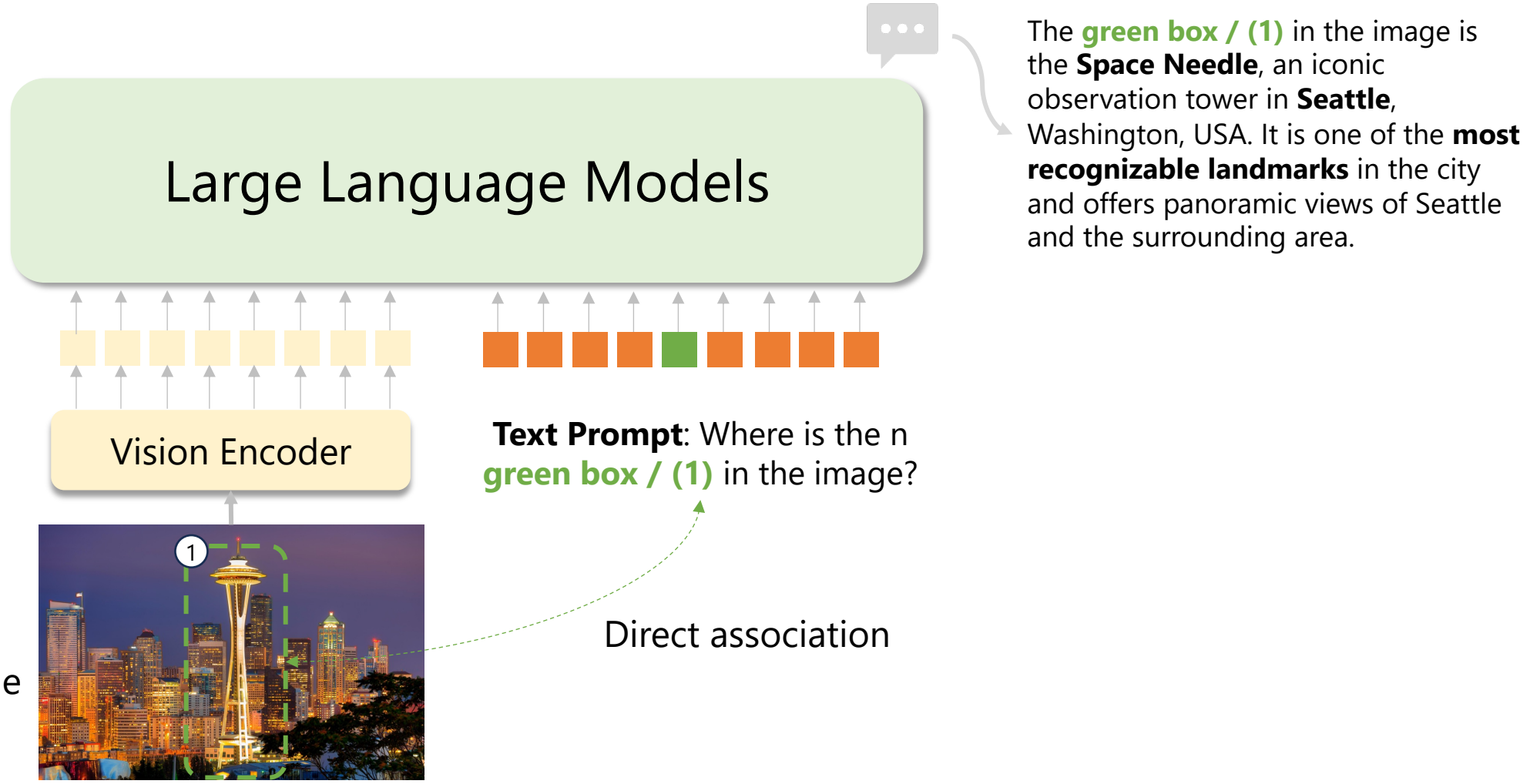
- Support as many types of visual prompts as possible (text, click, box, mark, etc.)
- Construct a new grounding instruction tuning data from LLaVA instruction tuning data

	input				output			
	text	click	box	mark	text	box	mask	mark
LLaVA [12]	✓				✓			
MiniGPT-4 [49]	✓				✓			
GPT4ROI [46]	✓		✓		✓			
Shikra [3]	✓				✓			
Ferret [40]	✓	✓			✓	✓		
MiniGPTv2 [2]	✓				✓	✓		
LLaVA1.5 [17]	✓				✓	✓		
CogVLM-Grounding [33]	✓				✓	✓		
LLaVA-G (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

Describe the following image with grounding.
The image depicts a group of people sitting at a large, white table, engaged in activities. There are four people in the scene, with three of them sitting close to each other and one person sitting further to the right. They are all using laptop computers, likely working or collaborating on a project together. The table is well organized, featuring multiple laptops, a large TV screen on the left side, and a clock on the wall. There are also several books and a cell phone placed on the table, indicating a casual and comfortable environment. A chair is positioned near the left side of the table, and there are two bottles placed nearby, possibly containing drinks for the people working at the table.

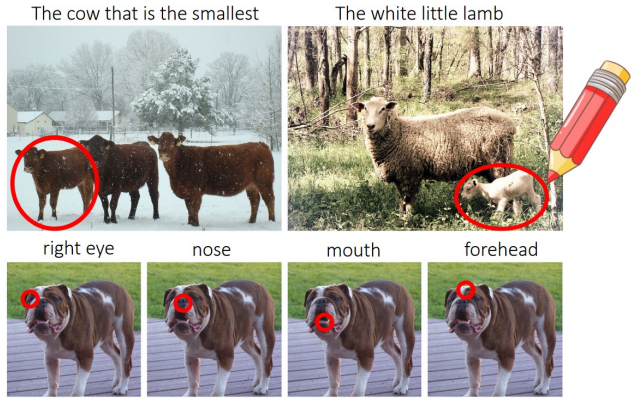
Llava-grounding: Grounded visual chat with large multimodal models. Zhang et al.

Visual Pixel Prompting for LMMs



Visual Pixel Prompting for LMMs

No changes on LMMs, but just pixels!



RedCLIP, Shtedritski et al.

Set-of-Mark. Yang et al.

Dawn of LMMs. Yang et al.

AssistGUI. Gao et al.

3DAxiesPrompt Liu et al.

GPT-4V-ACT. 2023

ViP-LLaVA. Cai et al.

GPT-4V Wonderland. Yan et al.

LLM-Optic. Zhao et al.

SketchPad. Hu et al.

AndroidWorld. Rawles et al.

Analogist. Gu et al.

SpatialRGPT. Cheng et al.

Scaffodling. Lei et al.

MOKA. Liu et al.

Set-of-Line Prompting. Zhang et al.

WebVoyager. He et al.

CoPA. Huang et al.

Visualwebarena. Koh et al.

Robi Butler. Zhang et al.

ManipQA. Huang et al.

SEEACT. Zheng et al.

PIVOT. He et al.

DetToolChain. Wu et al.

GlitchBench. Taesiri et al.

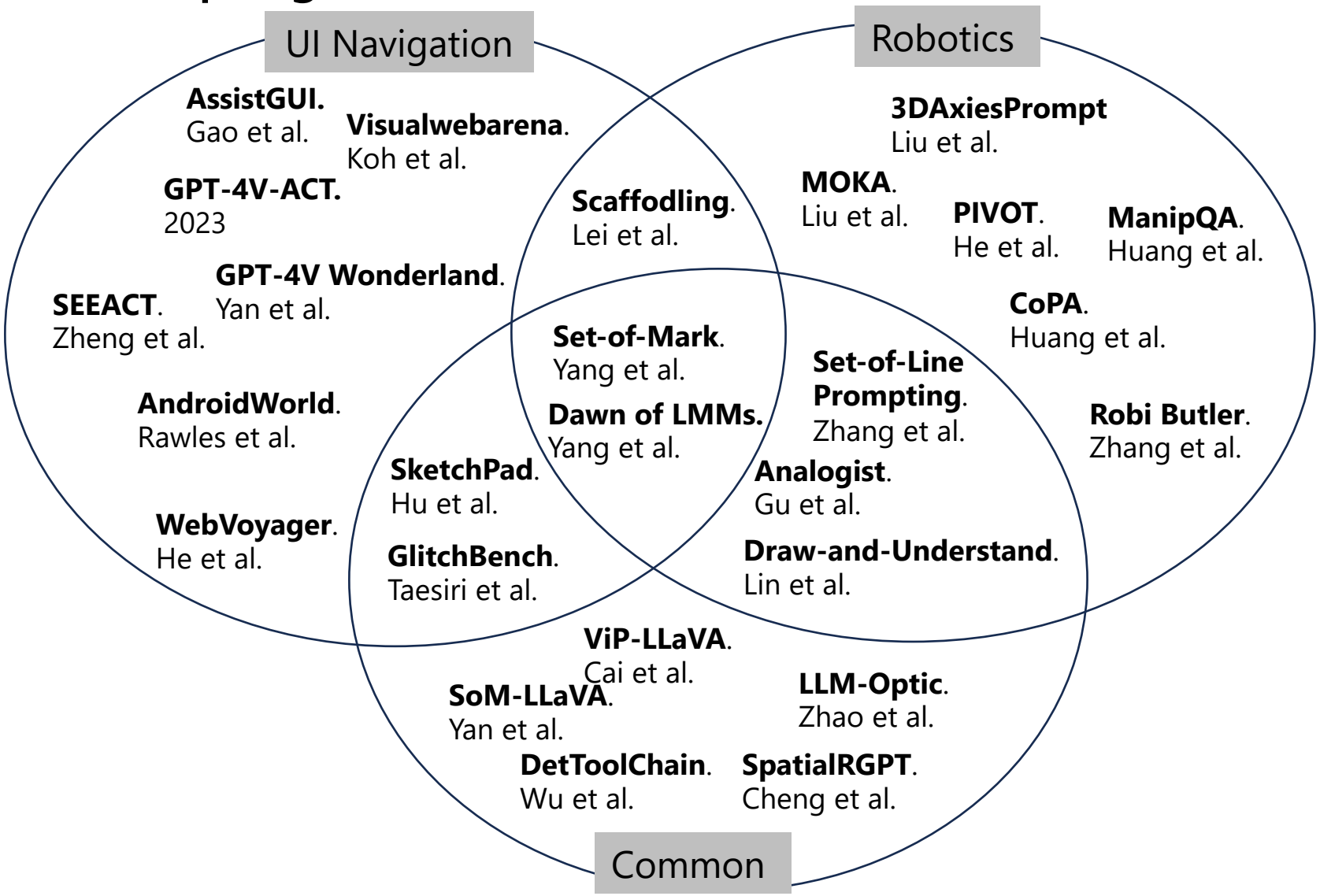
SoM-LLaVA. Yan et al.

Draw-and-Understand. Lin et al.

More to come...

Oct 2023 Dec 2023 June 2024

Visual Pixel Prompting for LMMs



Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding for GPT-4V

Set-of-Mark Prompting for LMMs

Input: Image

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. ✗

SoM Prompting

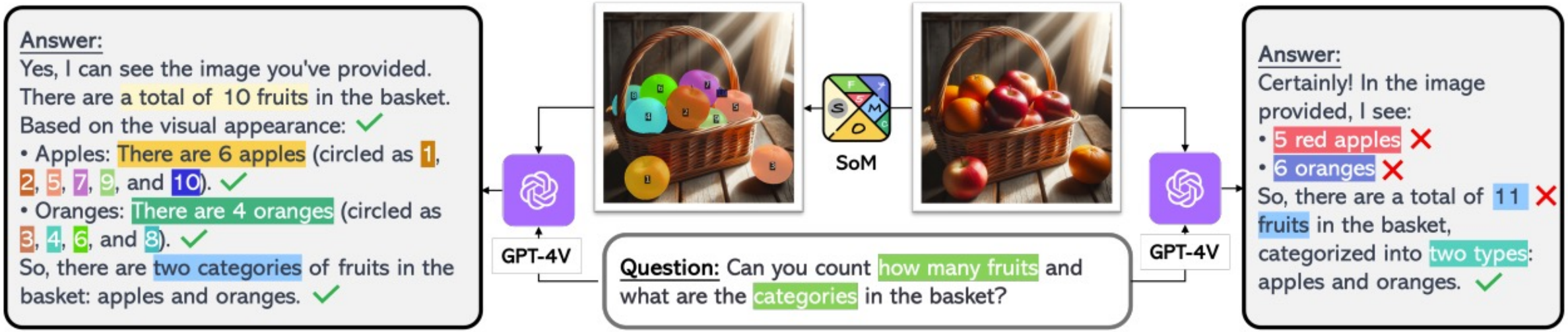
Input: Image + SoM

Conversation +

User	What is on the left side of the right laptop?
GPT-4V	On the left side of the right laptop (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). ✓

Set-of-Mark Prompting for LMMs

Think Step-by-Step for LLMs ⟶ See location-by-location for LMMs



SoM helps GPT-4V to see more precisely and finally induce the correct answers.

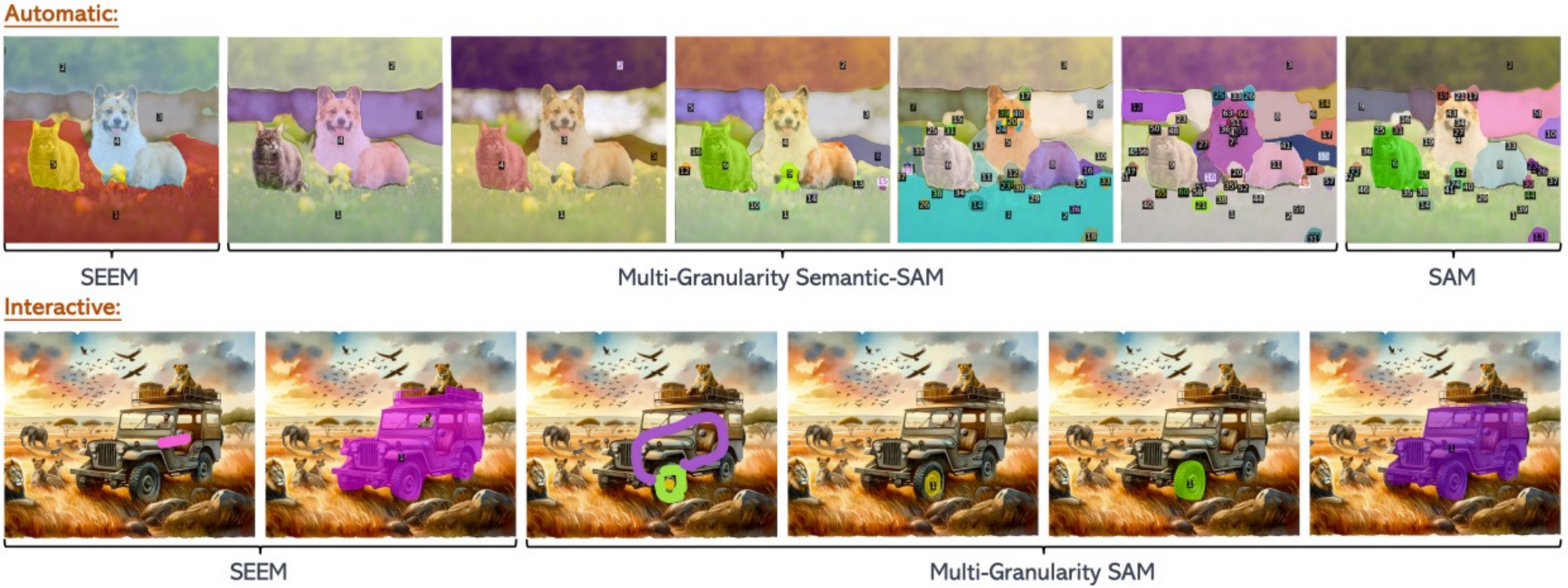
- **Two essential properties making SoM work:**
  - Partition an image into a set of semantically meaningful regions to align with the textual outputs, an ability known as grounding.
  - The auxiliary information cast to the input image should be both interpretable and spokenable by the LMM, so that it can be described in its textual outputs.

Set-of-Mark Prompting for LMMs: Image Partition

Strong performance: accurately segment images
Open vocabulary: understand a wide range of visual concepts
Rich granularities: not only single objects but also parts

Tools: SEEM + Semantic-SAM + SAM

Modes: Automatic + Interactive

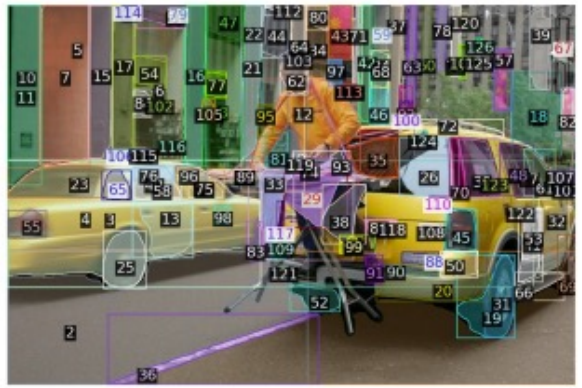


Set-of-Mark Prompting for LMMs: Mark Generation

(1) Mask

(2) Mask + Number

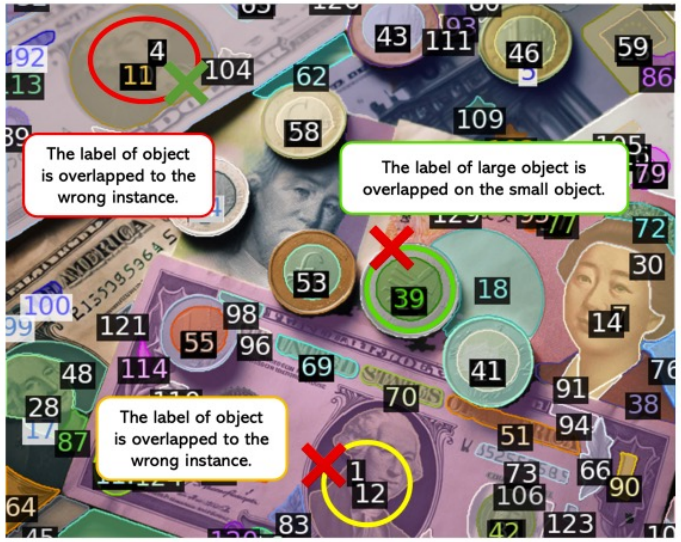
(3) Mask + Alphabet



(4) Mask + Box + Number

Mark types: mask, number, alphabet, box, etc.

Mark overlaying: start from small regions to big regions. Exclude previous regions when overlay current mark.




```
// Find center for a region
def Find_Center(r)
    D = DT(r) // Run distance transform
    c = arg max(D) // Find maxium location
    return c
// The main function
def Mark_Allocation(R):
    R̂ = Sorted(R) // Sort regions in ascending
        order of areas
    for k in range(K): do
        r_k = R̂[k] & ¬R̂[: k − 1].sum(0) // Exclude
            k − 1 regions
        C[k] = Find_Center(r_k)
    end
    return C
```

Set-of-Mark Prompting for Fine-grained Vision

Open-Vocabulary Segmentation

User: I have labeled a bright numeric ID at the center for each visual object in the image. Please enumerate their names. You must answer by selecting from the following names: [Person, ..., curtain] (ignore the 80 categories in COCO)
GPT-4V 1. Person 2. Person 3. Person 4. Surfboard 5. Handbag

Referring Segmentation

User: I have labeled a bright numeric ID at the center for each visual object in the image. Please tell me the IDs for: The laptop behind the beer bottle; Laptop turned on.
GPT-4V The IDs for the items you've mentioned: The laptop behind the beer bottle: 6 Laptop turned on: 2

Video Object Segmentation

User: The 2 images are from the same video, where the first image is the first frame and the second image is a later frame. In the first image, there are 2 objects labeled with 1,2. Can you track these 2 objects in the second image?
GPT-4V 1. The object labeled with 1 (a boxer in red headgear) is most similar to the object labeled with 2 (another boxer in red headgear). 2. The object labeled with 2 (a boxer in blue) is most similar to the object labeled with 1 (another boxer in blue).

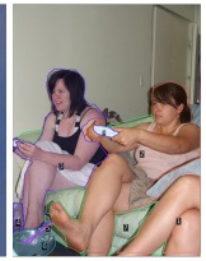
Grounded Segmentation

User: I have labeled a bright numeric ID at the center for each visual object in the image. Given the image showing a man in glasses holding a piece of paper, find the corresponding regions for a man in glasses, a piece of paper.
GPT-4V The "man in glasses" corresponds to the region labeled with "2". The "piece of paper" corresponds to the region labeled "5".

Model	LMM	Zero-shot	OV Segmentation COCO	OV Segmentation ADE20K	RefCOCOg REC	RefCOCOg RES	Phrase Grounding Flickr30K	VOS DAVIS2017
MaskDINO [23]	✗	✗	80.7	-	n/a	n/a	n/a	n/a
OpenSeeD [55]	✗	✓	-	23.4	n/a	n/a	n/a	n/a
GLIPv2 [56]	✗	✗	-	-	-	-	87.7*	n/a
GDINO [31]	✗	✗	n/a	n/a	86.1*	n/a	90.5	n/a
X-Decoder [64]	✗	✗	-	-	-	64.6*	n/a	62.8
PolyFormer [29]	✗	✗	n/a	n/a	85.8*	67.2	n/a	n/a
SegGPT [45]	✗	✓	n/a	n/a	n/a	n/a	n/a	75.6
SEEM [65]	✗	✗	-	-	-	65.7*	-	62.8
RedCircle [41]	✓	✓	n/a	n/a	59.4*	n/a	n/a	n/a
FGVP [48]	✓	✓	n/a	n/a	63.3*	n/a	n/a	n/a
Shikra [6]	✓	✗	n/a	n/a	82.6*	n/a	77.4	n/a
LLaVA-1.5 [27]	✓	✗	n/a	n/a	63.3	n/a	n/a	n/a
MiniGPT-v2 [5]	✓	✗	n/a	n/a	84.4*	n/a	n/a	n/a
Ferret [54]	✓	✗	n/a	n/a	85.8*	n/a	81.1	n/a
GPT-4V [36]	✓	✓	n/a	n/a	25.7	n/a	n/a	n/a
GPT-4V [36] + SoM (Ours)	✓	✓	75.7	**63.4**	**86.4**	**75.6**	89.2	**78.8**

Wrong Annotations

(a) A bird with a white wing [2]



(b) The legs of a person sitting on the couch [1]

(c) Trucks [1] tracing (Flicker30k Grounding)

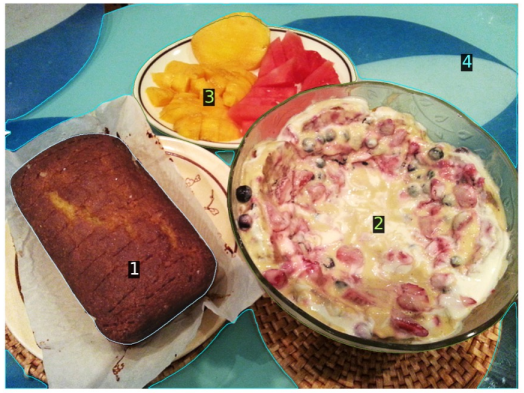
Confusing Labels

(e)

(f)

Set-of-Mark Prompting with Interleaved Prompts and Outputs

SoM-Enabled Plain and Interleaved Prompts



User: What is in the image?
GPT-4V: In the image, I observe the following items:
Loaf or Cake (1): A rectangular-shaped baked item, possibly a loaf of bread or a cake.
Creamy Fruit Mixture (2): A bowl containing a creamy mixture with various berries and possibly other fruits. It might be a fruit salad with a creamy dressing, yogurt with fruits, or some kind of dessert.
Sliced Fruits (3): A plate with slices of various fruits, including what looks like pineapples and possibly slices of watermelon or another type.
Table Surface (4): A portion of the table or countertop with a bluish hue.

(con't) Additionally, there are various decorative elements, plates, and bowls.

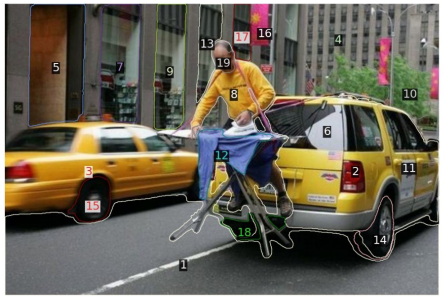
User: What is in 3?
GPT-4V: In the image, the item labeled "3" appears to be a plate of sliced fruits. It looks like there's pineapple and possibly another type of melon or fruit, such as watermelon or pink grapefruit.

User: Can I put the fruits in 3 into 2?
GPT-4V: Yes, based on the image, the fruits in 3 appear to be sliced fresh fruits, and they can be mixed or poured into the creamy fruit mixture in 2 to enhance flavor and texture. It's a common practice in desserts to combine fresh fruits with creamy mixtures.

User: Which can be toasted, 1 or 2?
GPT-4V: Based on the image, item 1 (which appears to be a loaf of bread or cake) is the one that can be toasted. Item 2 is a creamy fruit mixture and is not suitable for toasting.

User: How long it usually takes to toast a slice of loaf in 1?
GPT-4V: Typically, to toast a slice of a loaf like the one labeled as "1" in the image, it would take between 2 to 5 minutes in a toaster on a medium setting.

Grounded Reasoning and Cross-Image Reference



User: Can you point me what is the most unusual thing in this image?

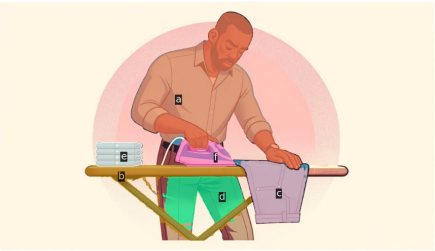
GPT-4V:

Certainly! The most unusual thing in the image is a man ironing clothes on an ironing board that is placed atop a moving yellow taxi cab in a city street. It's not a common sight to see someone ironing clothes on top of a taxi, especially while it's on the road!

User: Can you point me what is the most unusual thing in this image?

GPT-4V (SoM):

The most unusual thing in this image is **the man (elements 8, 17, 19)** standing on **an ironing board (element 12)** that's placed atop a moving **yellow taxi (elements 2, 6, 11)**. He's ironing clothes in the middle of a city street, which is highly atypical and not a safe behavior. This situation stands out as the most bizarre aspect of the scene.



User: What are the common objects in the two images?

GPT-4V (SoM): The common objects in the two images are:

- A man (In the first image, elements 8, 17, 19; In the second image, **element a**)
- An ironing board (In the first image, element 12; In the second image, **element b**)
- An iron (Not distinctly labeled in the first image but the man is using it; In the second image, **element f**)
- Clothes/shirt being ironed (Not distinctly labeled in the first image but the man is ironing it; In the second image, **element c**)

User:
I am on a diet, what are the top things I should avoid eating?

GPT4-V:

If you're on a diet, it would depend on the specific type of diet you're following. However, based on the image you provided, here are some common items that many diets might limit or avoid:
· Burgers: Especially if made with high-fat meats or served with cheese and mayonnaise.
· Fries: Deep-fried foods are typically high in calories and unhealthy fats.
· Chips: Can be high in salt and fats, especially if they're commercially produced.
· Onion rings: Another deep-fried item.
· Sandwiches with white bread: Many diets recommend whole grains over refined grains.
· Processed meats: Like the ham in the sandwich, as they can be high in sodium and preservatives.
· Sauces: Some can be high in sugar, sodium, or unhealthy fats.
· Soft drinks or sweetened beverages: High in sugars and empty calories.

User:
I am on a diet, what are the top things I should avoid eating?

GPT-4V (SoM):

Based on the typical diet considerations, here are the top items from the provided image that you should most likely avoid or limit:
1. **Onion rings (1 and 11)** - Fried foods like these are often high in unhealthy fats and calories.
2. **Fries (8)** - Another fried item that can be calorie-dense and contain unhealthy fats.
3. **Burgers (3 and 5)** - Depending on their composition, they can be high in calories, unhealthy fats, and may contain processed meats.
4. **Processed meats (2 and 15)** - These can be high in sodium, unhealthy fats, and other additives.
5. **Chips (16)** - Typically high in unhealthy fats and can be calorie-dense.: There is a yellowish sauce or broth on the plate.
While indulging occasionally is okay, it's best to consume these items in moderation if you're trying to maintain a healthy diet.

User:
What is in this dish?

GPT4-V:

The dish appears to feature:
· A grilled or pan-seared fish fillet, which looks like it could be a type of white fish, possibly sea bass, cod, or halibut. · Asparagus spears.
· Potatoes, possibly roasted or boiled.
· A wedge of lemon for squeezing over the dish.
· Garnished with fresh parsley or cilantro leaves.
· There might be a light sauce or seasoning drizzled over the fish and vegetables, but it's hard to specify its ingredients from the image alone.
This looks like a delicious and healthy meal!

User:
What is in this dish?

GPT-4V (SoM):

Here's a description of the items in the dish:
1. Fish Fillet: The main component of the dish appears to be a fillet of fish, likely grilled or pan-seared.
2. Lemon Wedge: A common accompaniment with fish dishes to enhance flavor when squeezed over the top.
3. Parsley or Cilantro: This green herb is used as a garnish and can add a fresh flavor to dishes.
4. Asparagus: These green stalks are often steamed, salted, or grilled and served as a side.
5. Sauce or Broth: There is a yellowish sauce or broth on the plate.
6. Potato: It looks like a roasted or boiled potato.
7. Garnish: These could be green onions or chives scattered on the plate to add color and flavor.
8. Sauce or Oil Drizzle: There's a darker drizzle around the potato, which could be a type of sauce or perhaps a balsamic reduction.

Tool Usage Instruction

User: How to use this controller?

GPT-4V (SoM):

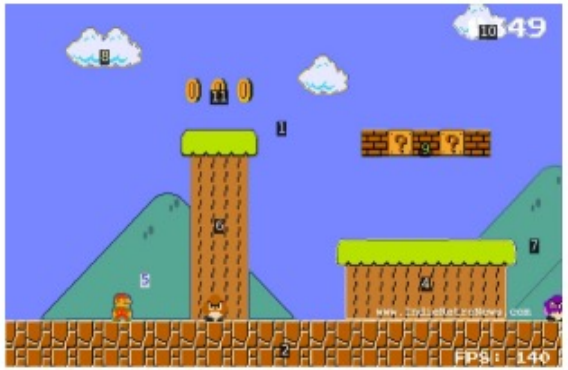
This appears to be an Xbox One or Xbox Series XlS controller. Here's a basic rundown on how to use it:

- **Thumbsticks (2 & 3)**: These are primarily used for movement and aiming.
  - **Left Thumbstick (2)**: Controls the movement of the character.
  - **Right Thumbstick (3)**: Controls the camera or aiming reticle.
- **Directional Pad (D-Pad) (1)**: Used for various functions such as navigating menus, selecting weapons, or other in-game actions.
- **Face Buttons (4, 5, 6, & 7)**:
  - **A (4)**: Typically the main action or "accept" button.
  - **B (7)**: Often the "back" or "cancel" button.
  - **X (5)**: Varies by game, frequently used for interactions.
  - **Y (6)**: Like X, its function varies but is commonly used for secondary actions or interactions.
- **Bumpers and Triggers (8 & 9)**:
  - **Left and Right Bumpers (8)**: Used for quick actions or toggling between items.
  - **Left and Right Triggers**: Used for primary actions like shooting or accelerating in racing games.
- Menu and View Buttons: Located in the center, used to bring up in-game menus or other overlays.
- Xbox Button: Press to go to the Xbox home or turn the controller on/off. Long press brings up a menu to turn off the controller or console.
- Pairing Button: Located on the top edge, used for pairing the controller.
- 3.5mm Audio Jack: On the bottom, plug in headphones or a headset.
- USB Port: On the top edge, used for charging or wired play.

Refer to the game's manual or in-game control settings for exact functions for each game.

2D Game Planning

User: How can I get the coins on top?



GPT-4V (SoM):

It looks like you're playing a platformer game reminiscent of Super Mario Bros. Here's a general strategy based on the layout:

- Elevated Platforms: **The platforms labeled as '4' and '7'** are the key. You need to first get onto these platforms to reach greater heights. Depending on the game mechanics, you may need to find a run-up or utilize a jump-boosting power-up if available.
- Bricks and Blocks: **The bricks labeled as '9'** might contain some power-ups or they might be breakable. In classic platformers like Super Mario, hitting such blocks from below can reveal power-ups or even create platforms. **The same goes for the block labeled as '1'**.
- Vertical Jump: **Once on the elevated platform '4'**, try to perform a vertical jump to reach **the coins labeled '11'**. Ensure that you time your jump correctly and use any speed or jump boosts, if available.
- Momentum: Some games also allow for gaining extra height in jumps by building momentum. If this game works similarly, try running from a distance and jumping at the last moment to grab the coins.
- Look for Hidden Power-ups: Some games hide power-ups in the level that can assist players in reaching difficult places. These might be hidden in unmarked bricks or in other places. If you have abilities like a double jump, use them.
- Game Mechanics: Every game can have unique mechanics. If there are any other controls or mechanics you haven't tried, consider experimenting with them. Maybe there's a way to bounce, double jump, or use an enemy to boost your jump height.

If you've tried all the general strategies and still can't reach the coins, it might be beneficial to look up specific strategies or walkthroughs for the particular game you're playing. Good luck!

Simulated Navigation

User: I am a bit hungry, where can I find some food?

GPT-4V (SoM):
Based on the image provided, it looks like you're viewing a floor plan of an apartment. For food, **you should head to the kitchen area which is labeled with the number "10"**. That's where you're most likely to find food or kitchen appliances to prepare a meal. If you're in this apartment, head to that section! If you're not, consider checking your own kitchen or a nearby restaurant. Enjoy your meal!

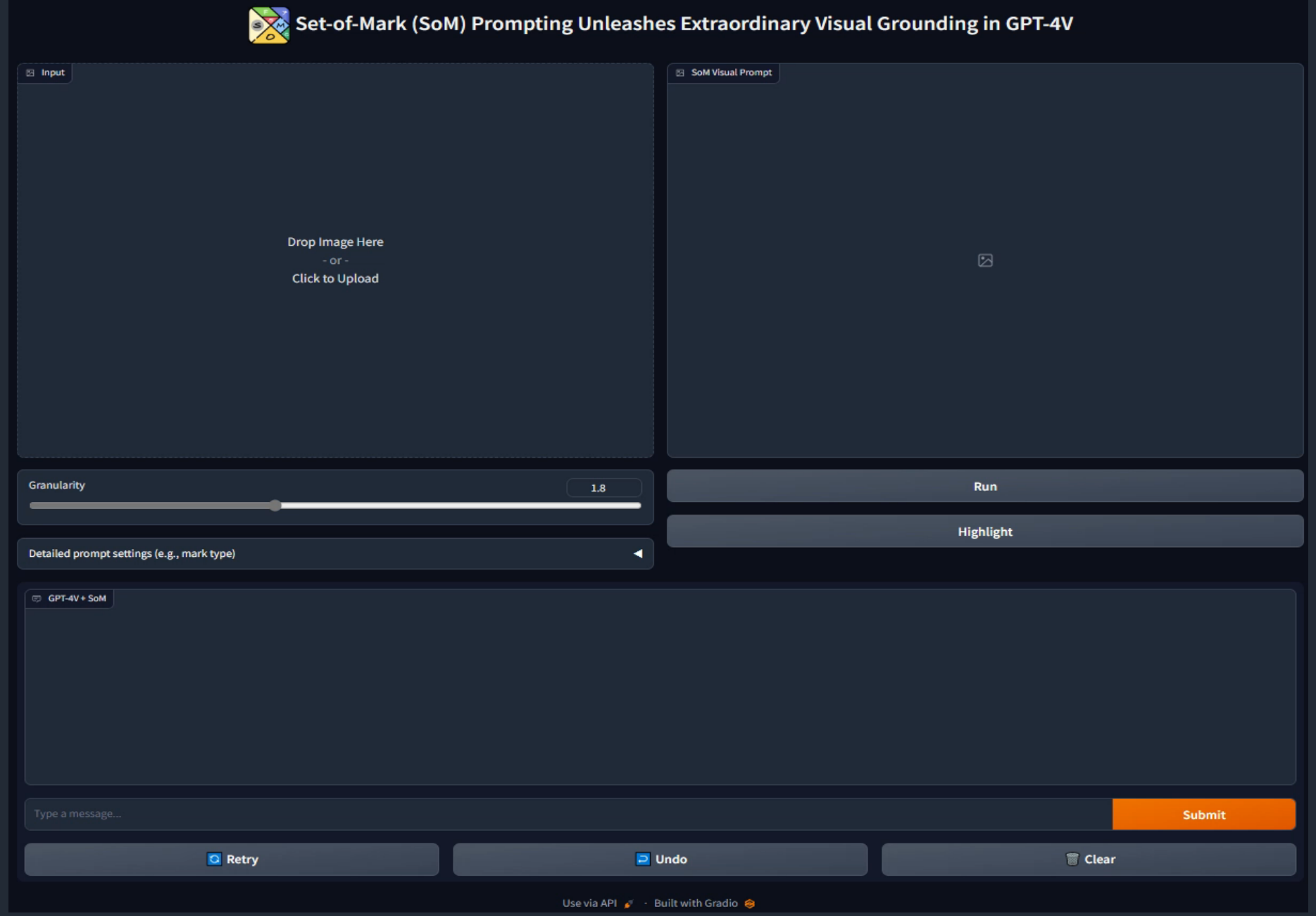
User: cool, where should I sit to eat if I find some food from 10?

GPT-4V (SoM):
Based on the floor plan, the most suitable place to sit and eat would be at **the dining area labeled with the number "14". It appears to have a dining table and chairs**, making it a comfortable spot for enjoying a meal.

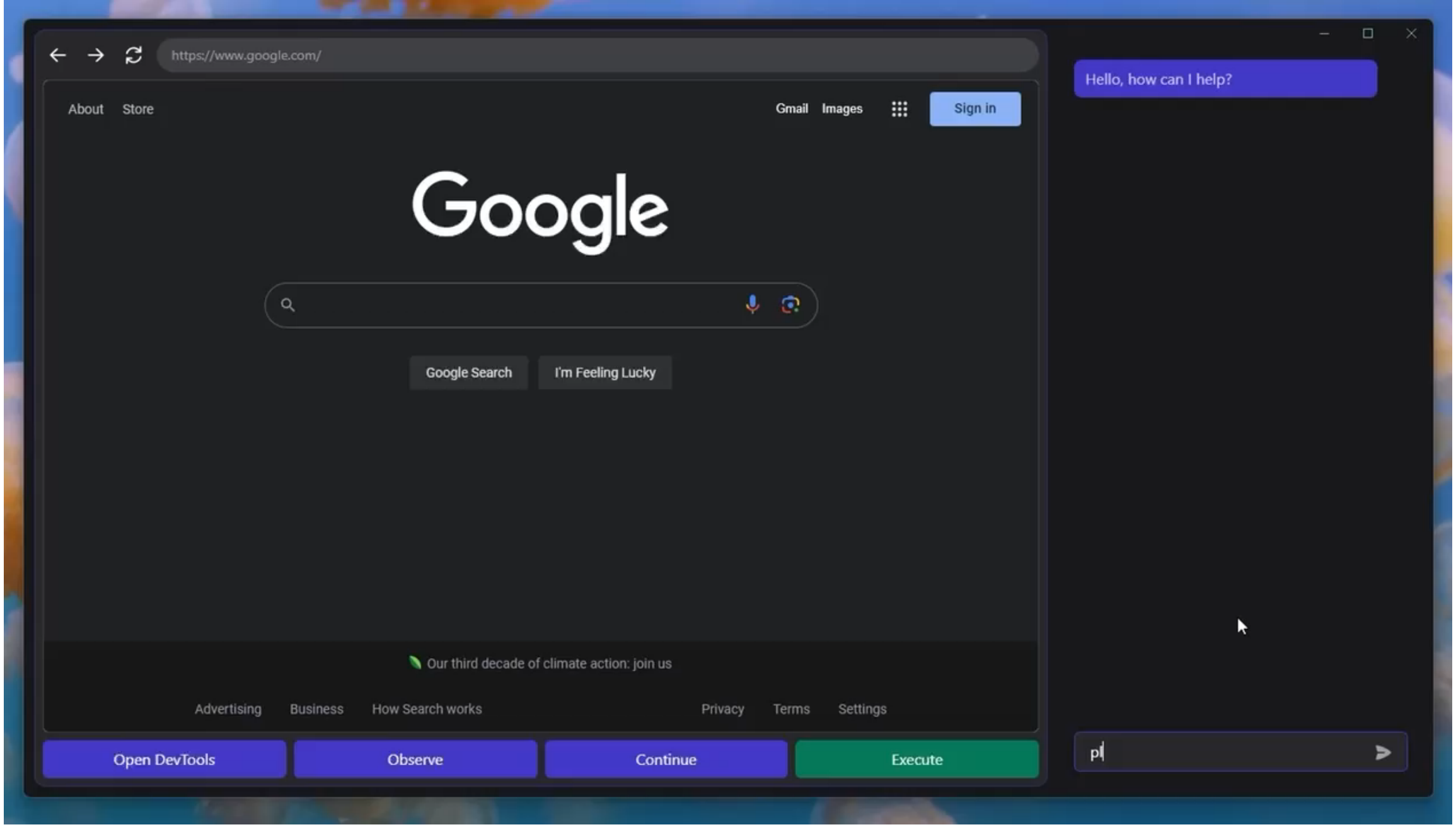
User: ok, i got a dirty hand, what should I do?

GPT-4V (SoM):
Based on the floor plan, you should head to the bathroom areas to wash your hands. **You can go to the space labeled with the number "13" which appears to be a bathroom**. It's likely to have a sink and soap where you can clean your hands. Always remember to wash your hands thoroughly, especially before eating!

Set-of-Mark (SoM) Prompting Unleashes Extraordinary Visual Grounding in GPT-4V



Visual Pixel (Set-of-Mark) Prompting for UI Navigation



https://github.com/ddupont808/GPT-4V-Act

Zero-shot Smartphone GUI Navigation

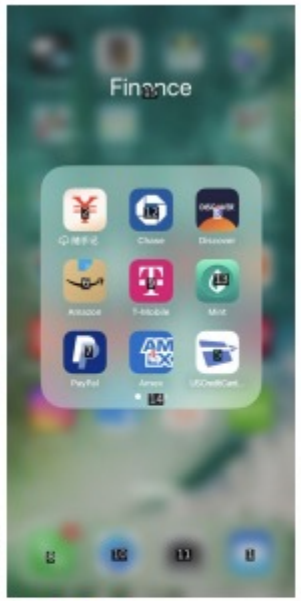
Instruction query:

You are asked to shop for a milk frother, your budget is between $50 and $100.

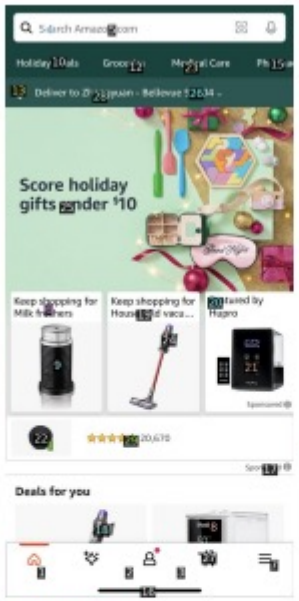
Prompt for all steps:

Imagine that you are a robot operating an iPhone.
You are asked to shop for a milk frother, your budget is between $50 and $100.
Your previous action is:
. . .
Below is what you see on the screen, with numeric IDs added. For the next action, which ID would you click on the screen next?

(1) GPT-4V:
To shop for a milk frother within the specified budget using the iPhone interface shown in the image, I would select an application that is capable of online shopping. In this scenario, I would click on ID 16 to use the Amazon app, as it is a comprehensive and user-friendly platform for purchasing a wide variety of products, including kitchen appliances like milk frothers.



(2) GPT-4V:
To proceed with shopping for a milk frother within the specified budget using the Amazon app, the next step would be to click on ID 6, which represents the Amazon app on the screen. This will open the app and allow me to search for milk frothers within the budget range.



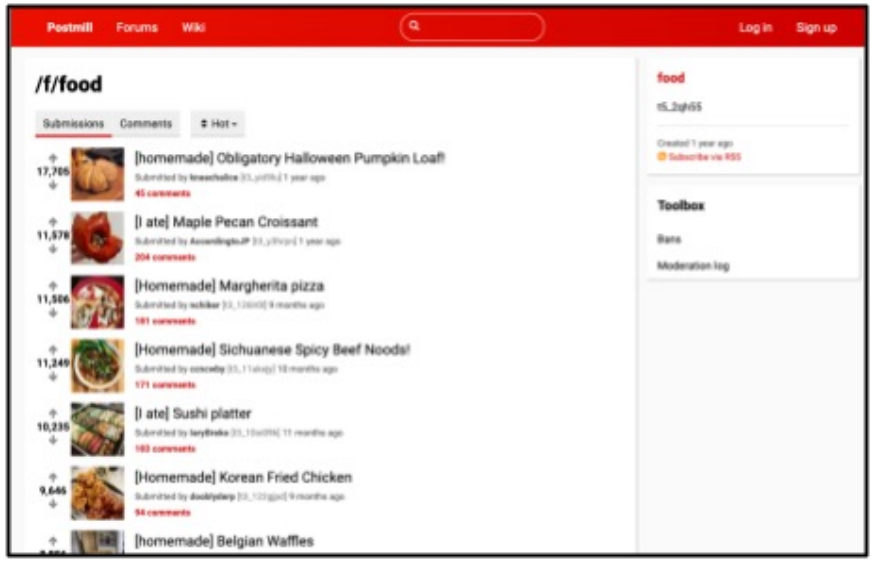
(3) GPT-4V:
I would click on the search bar at the top of the Amazon app interface, which is typically where you would enter a search query. This search bar is labeled with ID 5. By clicking on this, I can then type in "milk frother" and set the price filter to show results that are between $50 and $100.

(4) GPT-4V:
The next step would be to refine the search results by setting the price filter. In this interface, the ID most likely to represent the filter or sorting options is 8 labeled "Filters," which seems to be the right place to start refining the search results according to the budget. So the next action would be to click on ID 8 to access filtering options, including setting the desired price range.

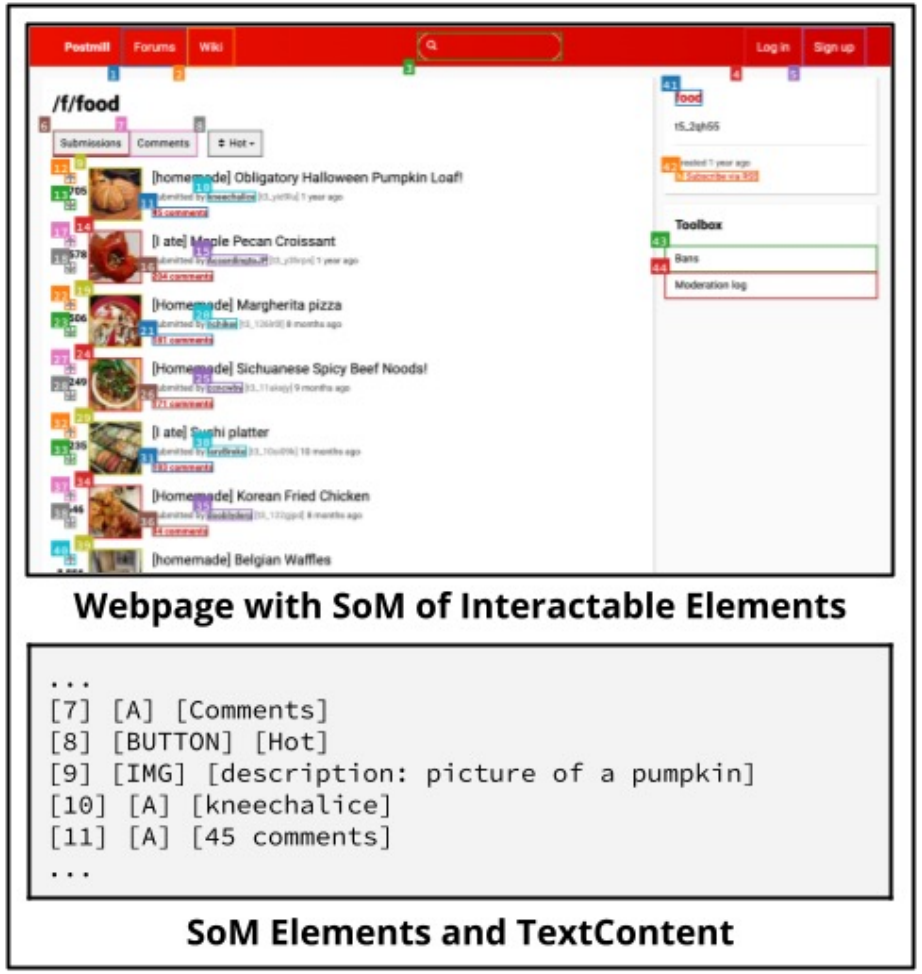
(5) GPT-4V:
The logical next step would be to click on the price filter that matches your budget range. Looking at the screen, you would choose the price filter that includes $50 to $100. The closest available options are ID 25 for "$25 to $50" and ID 26 for "$50 to $100". Since your budget starts at $50, you would click on ID 26 to filter the search results to show only milk frothers that are priced between $50 and $100.

(6) GPT-4V:
The next action would be to click on ID 13, "All results," to view all available milk frothers within the selected price range of $50 to $100.

VisualWebArena



Original Webpage



Webpage with SoM of Interactable Elements

SoM Elements and TextContent

"I'm trying to find this post. Navigate to the comment section for it."

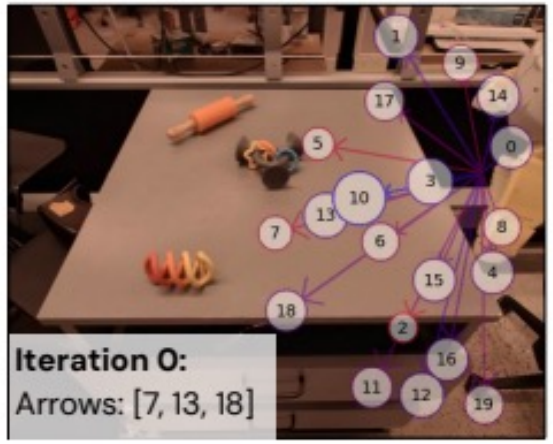
LLM / VLM Agent

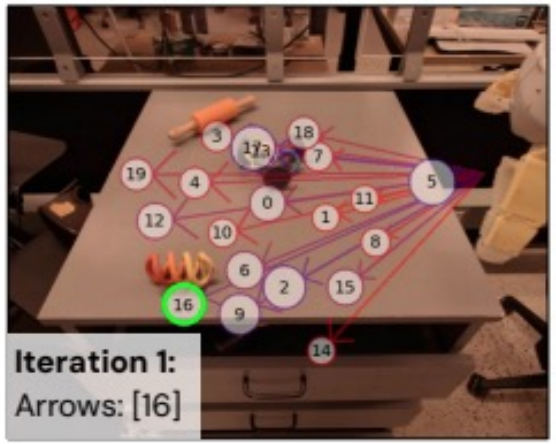
click [31]

Visual Pixel (Set-of-Mark) Prompting for Robotics

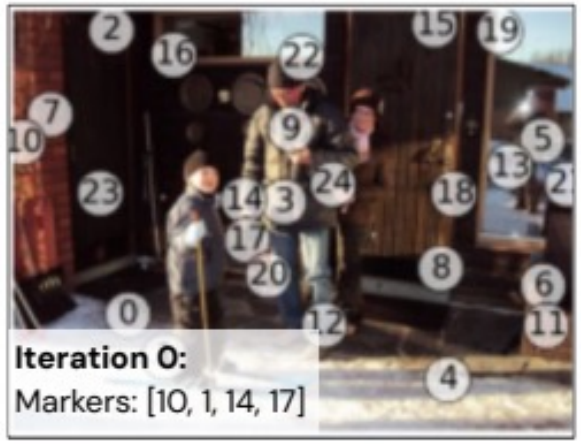
Visual Pixel Prompting for Robotics Navigation

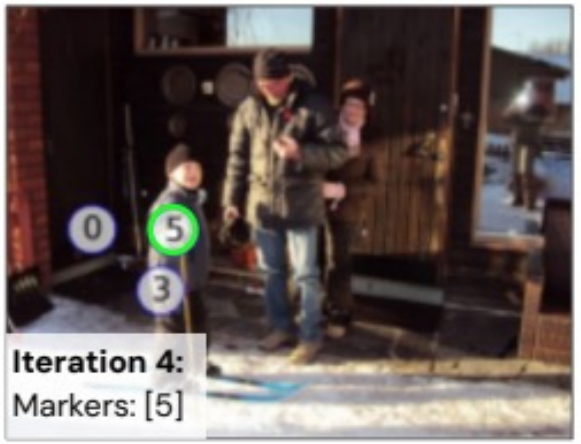
Task: What actions should the robot take to pick up the DNA chew toy?



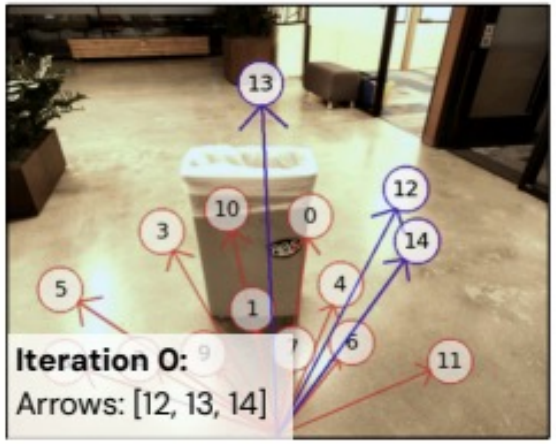


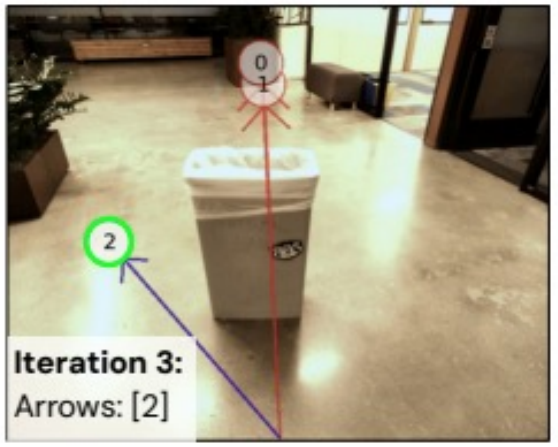
Task: What numbers overlay the "L kid"?



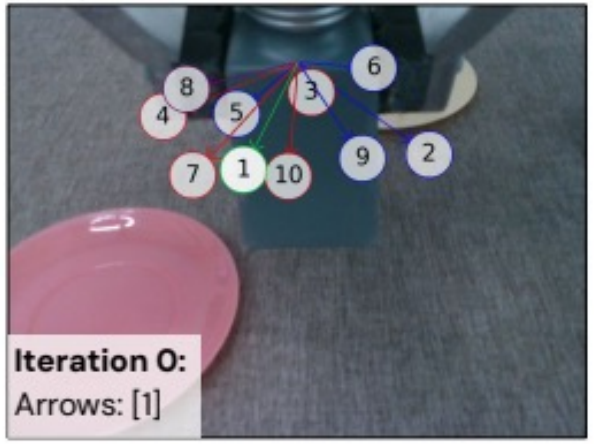


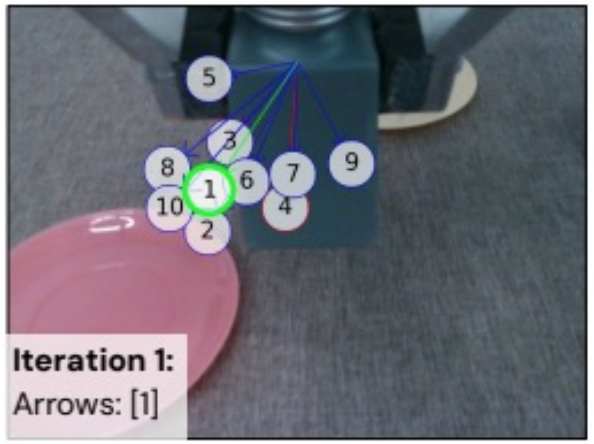
Task: What actions should the robot take to go to wooden bench without hitting the obstacle?





Task: What actions should the robot take to put the pepper shaker on the pink plate?

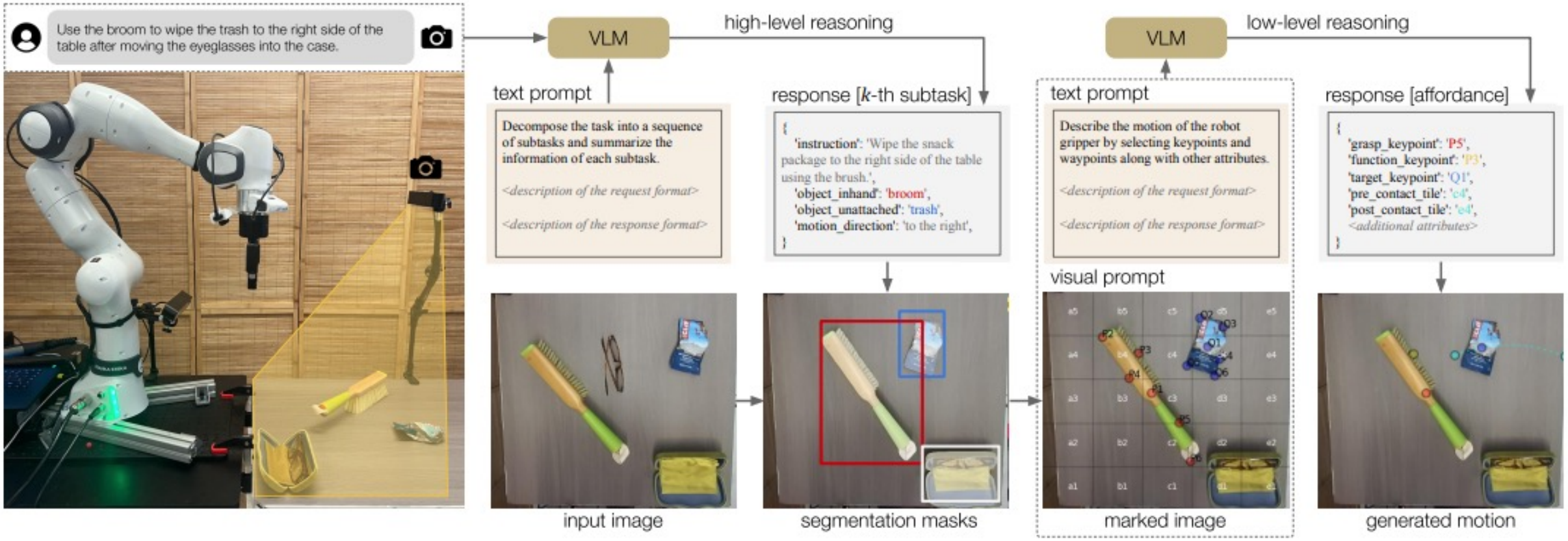




PIVOT: Iterative Visual Prompting Elicits Actionable Knowledge for VLMs. Nasiriany et al.

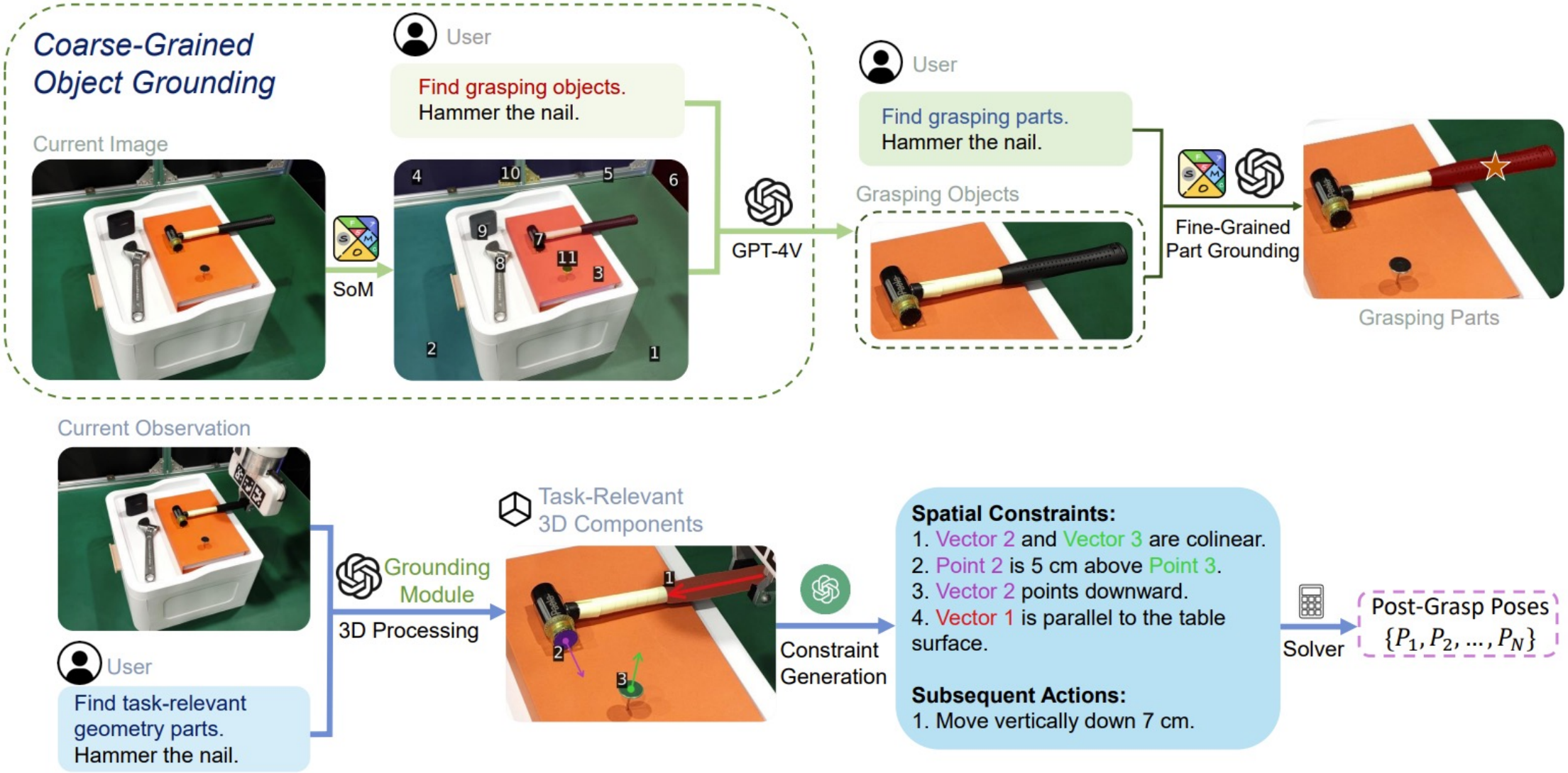
Visual Pixel Prompting for Robotics Manipulation

Decompose take into subtasks and address grounding sub-tasks with visual pixel prompting



MOKA: Open-Vocabulary Robotic Manipulation through Mark-Based Visual Prompting. Liu et al.

Visual Pixel Prompting for Robotics Manipulation



CoPa: General Robotic Manipulation through Spatial Constraints of Parts with Foundation Models. Huang et al.

Recap: Visual Prompting for LMMs

Visual Feature Prompting

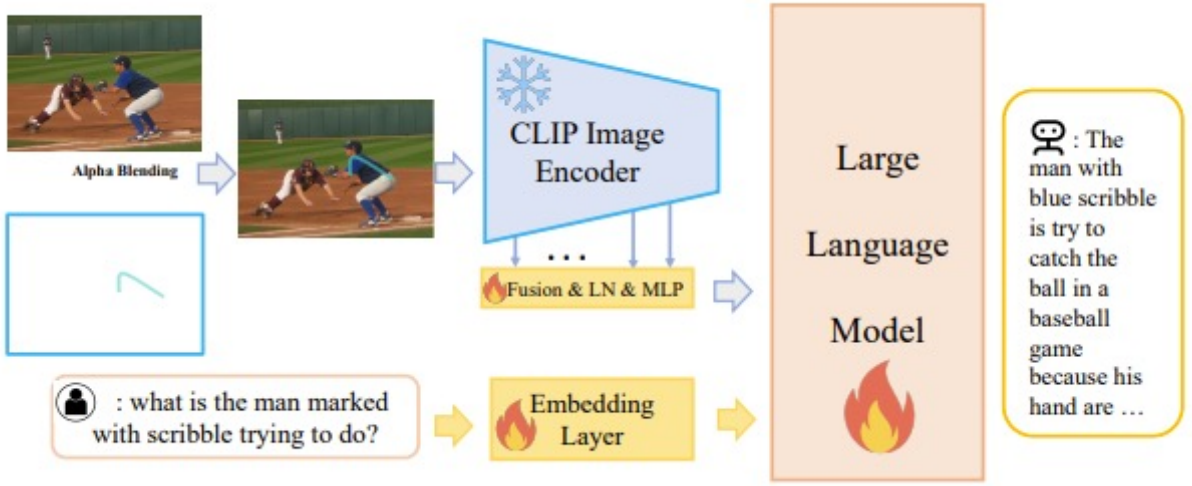
1. Extract visual features as the prompts
2. Need additional module to encode the visual prompts
3. Require some additional annotations

Visual Pixel Prompting

1. Directly overlay prompts on the images
2. Organic to the original LMM systems
3. Rely on the original emerging capability of LMMs

Can we build our own LMMs to support visual pixel prompting?

Visual Pixel Prompting LMMs: ViP-LLaVA



- No need for complex region encoding, but achieve state-of-the-art performance on region understanding tasks like Visual7W, PointQA, and Visual Commonsense Reasoning benchmark.
- ViP-Bench, a comprehensive benchmark to assess the capability of models in understanding visual prompts

ViP-LLaVA: Making Large Multimodal Models Understand Arbitrary Visual Prompts. Cai et al.

Visual Pixel Prompting LMMs: SoM-LLaVA

Empowering Open-Source Multimodal LLMs with Set-of-Mark Prompting and Improved Visual Reasoning Ability.

Input: Image

Conversation

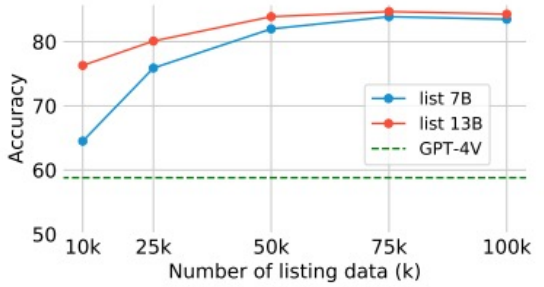
User	What items are there near the Marshal speaker?
LLaVA-1.5	There is a laptop and a cup near the Marshall speaker. ✗
User	To move the speaker closer to the curtain while keeping the laptop in the current position, which item should I swap it with?
LLaVA-1.5	You should swap the laptop with the cup. ✗

Input: Image + SoM

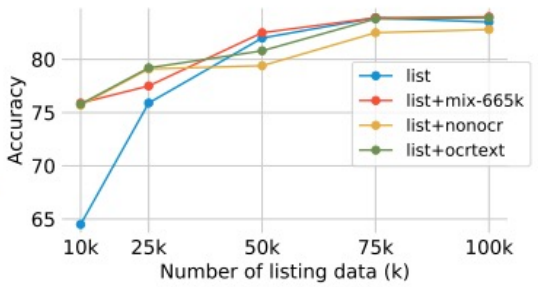
Conversation

User	What items are there near the Marshal speaker?
SoM-LLaVA	There is a laptop tagged with number 7 and a notebook tagged with number 8. ✓
User	To move the speaker closer to the curtain while keeping the laptop in the current position, which item should I swap it with?
SoM-LLaVA	You can swap it with the lamp tagged with number 9. ✓

Dedicatedly trained models can outperform GPT-4V for set-of-mark prompting



(a) Ablation on model sizes with LLaVA-1.5



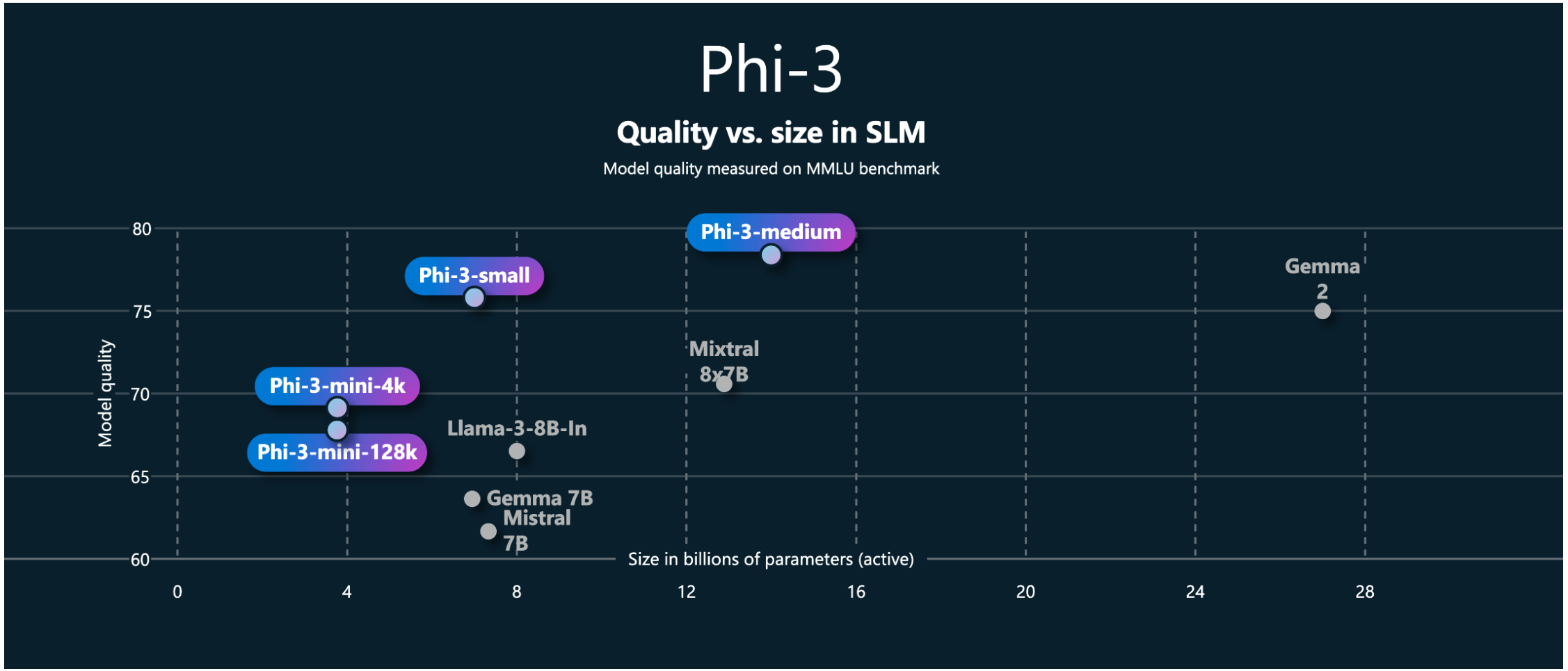
(b) Ablation on data sources with LLaVA-1.5-7B

Method	LLM	POPE	MME	SEED-I	LLaVA-Wild	MM-VET
BLIP-2	Vicuna-13B	85.3	1293.8	49.7	38.1	22.4
LLaVA-1.5	Vicuna-13B	85.9	1531.3	68.2	70.7	35.4
SoM-LLaVA-1.5	Vicuna-13B	86.6	1563.1	**69.6**	**75.3**	35.9
SoM-LLaVA-1.5 w/ tags	Vicuna-13B	**87.0**	**1572.8**	69.5	73.3	**37.2**

Force LMMs to list item one-by-one can improve the grounding and reasoning ability

List Items One by One: A New Data Source and Learning Paradigm for Multimodal LLMs. Yan et al.

Emerging Capability in Phi-3-Vision



Phi-3-Vision

Category	Benchmark	Phi-3-Vision-128K-In	Llava-1.6 Vicuna-7B	QWEN-VL Chat	Llama3-Llava-Next-8B	Claude-3 Haiku	Gemini 1.0 Pro V	GPT-4V-Turbo
Popular aggregated benchmark	MMMU (val)	40.4	34.2	39.0	36.4	40.7	42.0	55.5
	MMBench (dev-en)	80.5	76.3	75.8	79.4	62.4	80.0	86.1
Visual scientific knowledge reasoning	ScienceQA (img-test)	90.8	70.6	67.2	73.7	72.0	79.7	75.7
Visual math reasoning	MathVista (testmini)	44.5	31.5	29.4	34.8	33.2	35.0	47.5
	InterGPS (test)	38.1	20.5	22.3	24.6	32.1	28.6	41.0
Chart reasoning	AI2D (test)	76.7	63.1	59.8	66.9	60.3	62.8	74.7
	ChartQA (test)	81.4	55.0	50.9	65.8	59.3	58.0	62.3
Document Intelligence	TextVQA (val)	70.9	64.6	59.4	55.7	62.7	64.7	68.1
Object visual presence verification	POPE (test)	85.8	87.2	82.6	87.0	74.4	84.2	83.7

Phi-3-Vision

What is the price of a Cappuccino with large size?

The price for a Cappuccino in large size is $3.25.

How much does it cost to add ice to a tea?

It costs an additional $1.00 to add ice to any tea.

If someone wants to buy a pot of tea, how much would it cost?

A pot of tea would cost $4.00.

extract all displayed texts and keep original displayed format. For the left section, please use markdown table with column "coffee", "small", "median" and "large". Instead, for the right section, please use JSON format.

Certainly! Below is the markdown table converted from the image:

Coffee	Small	Medium	Large
Espresso	1.30	1.60	2.00
Americana	1.40	1.75	2.10
Cappuccino	2.25	2.90	3.25
Latte	2.35	3.00	3.25
Mocha	2.35	3.00	3.50
Chai	2.25	2.40	3.25
Chai Bomb	2.75	3.25	3.75
House Cafe	1.50	1.75	2.00
Cafe au Lait	1.75	2.50	3.25

And here is the JSON representation of the teas offered:

```
{
 "Teas": {
   "CUP": {
     "GREEN": "$1.50",
     "BLACK": "$1.50",
     "WHITE": "$1.50",
     "OOLONG": "$1.50",
     "HERBAL": "$1.50"
   },
   "POT": {
     "GREEN": "$4.00",
     "BLACK": "$4.00",
     "WHITE": "$4.00",
     "OOLONG": "$4.00",
     "HERBAL": "$4.00"
   }
 },
 "*ADD .50 FOR HONEY*",
 "*ADD $1.00 FORICE*"
}
```

Phi-3-Vision

Chatbot

which chair should I sit if I want to have a lamp?

You should sit in Chair 8 to have a lamp, as it is the only chair with a lamp on the desk.

Chatbot

what is 3?

The number 3 is associated with the garnish on top of the fish, which appears to be a sprig of parsley.

Recap: A Close Look at Vision

1 **Visual Tokenizer** What vision encoder is a good vision tokenizer for LMMs?

- Multimodal pretrained vision encoder CLIP is the best single one but still not sufficient
- Multi-crop strategy can support much higher-resolution using fixed-size encoder
- Mixture of vision encoders with CLIP can further enhance the performance

2 **Visual Tokens** How to cope with visual tokens for LLMs?

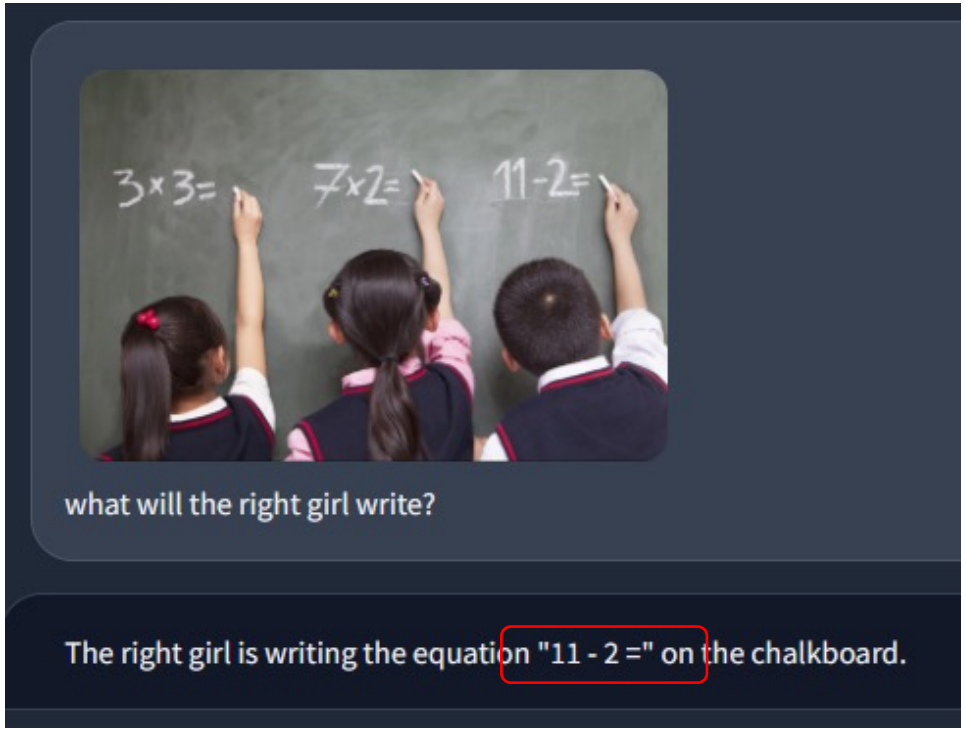
- Early strategies like Q-Former and Concatenate-then-projection
- Token reduction by adaptive sampling, nested and elastic organization
- Stack tokens from bottom to top, instead of only stringing tokens from left to right

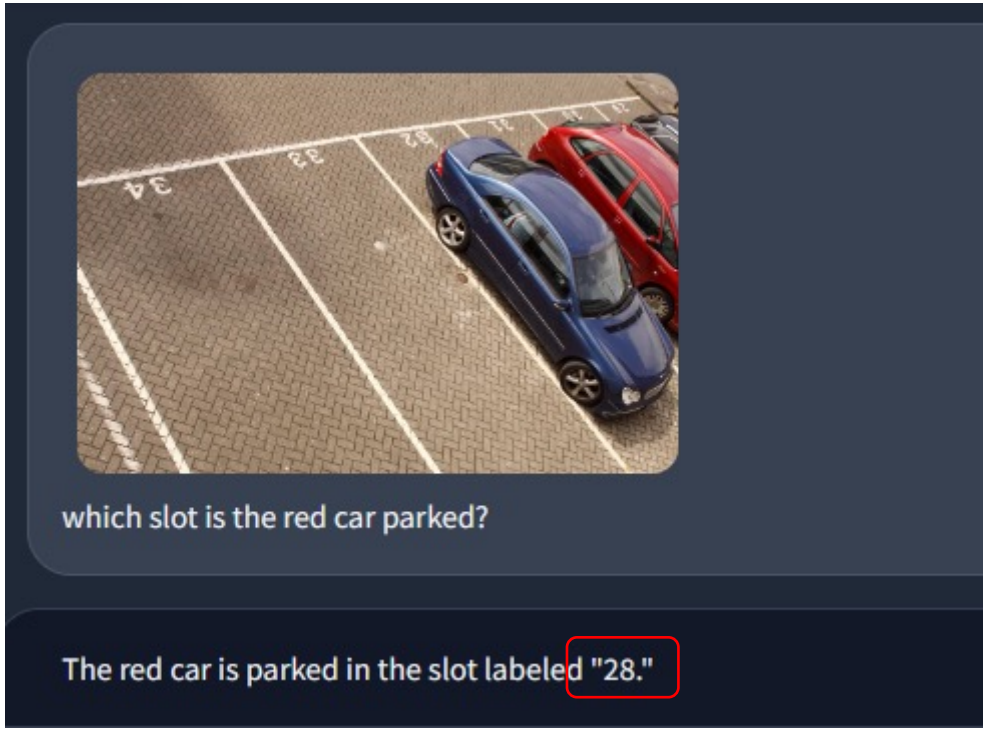
3 **Visual Prompting** How to perform visual prompting for LMMs as for text?

- Two visual prompting strategies: feature-based and pixel-based
- Both have pros and cons, and visual pixel prompting is much easier to apply
- LMMs can be trained to obtain visual pixel prompting capability, and could potentially have such capability emerged in small models, e.g., Phi-3-Vision

Challenges & Opportunities

Current LMMs still fall short of many “easy” tasks like

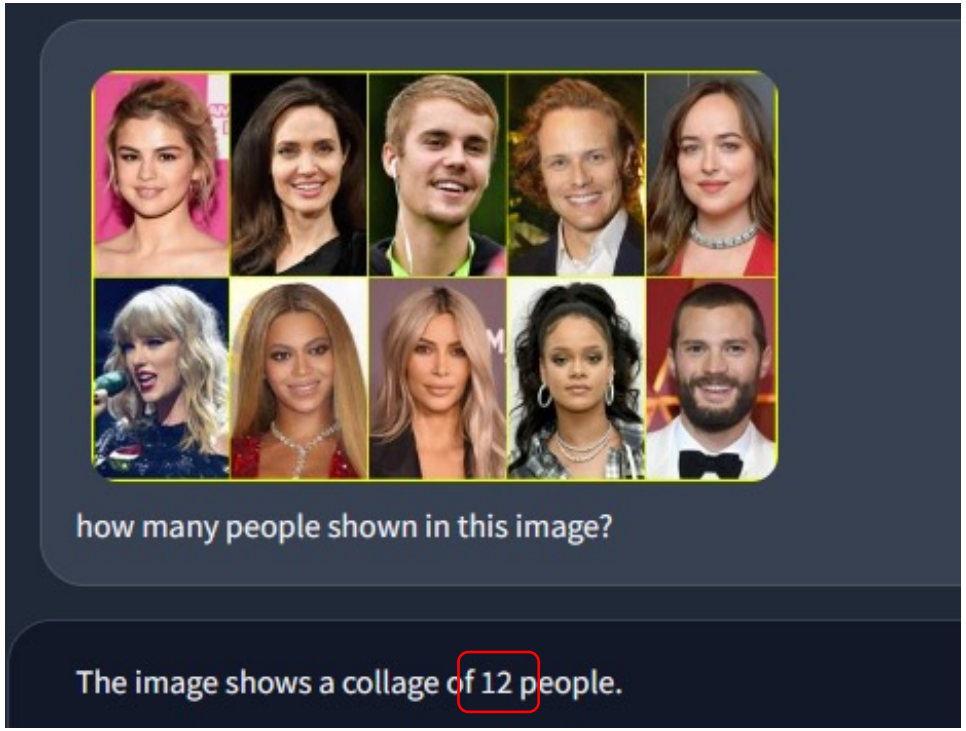




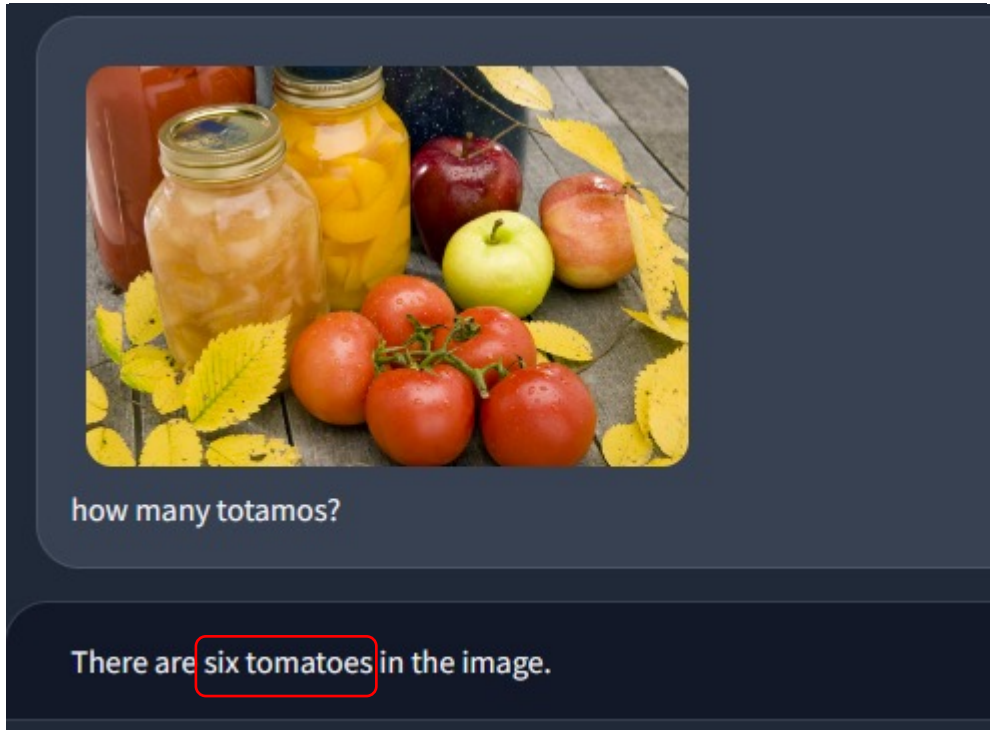
- It wrongly locates the girl on the right
- It is not able to build the correct association

Spatial Referring and Reasoning

Current LMMs still fall short of many “easy” tasks like



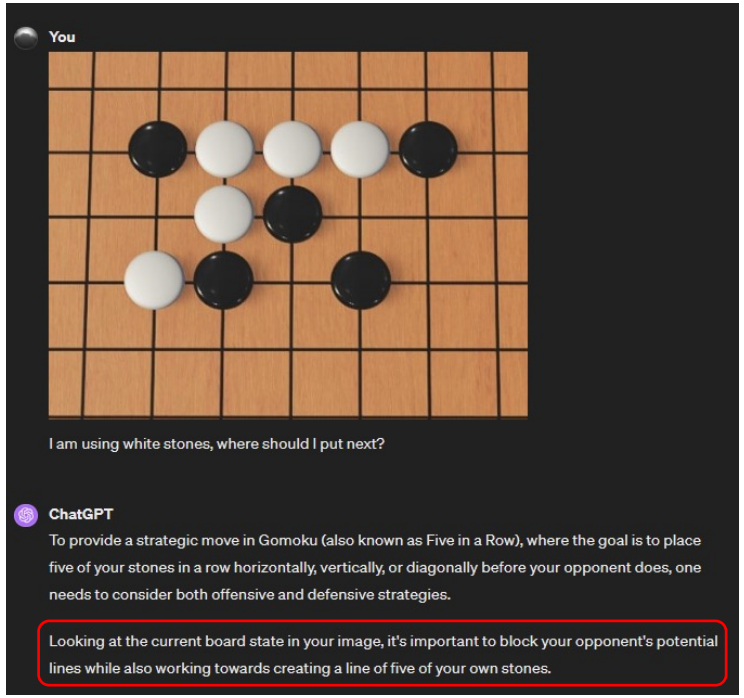
- There are 10 people in the image

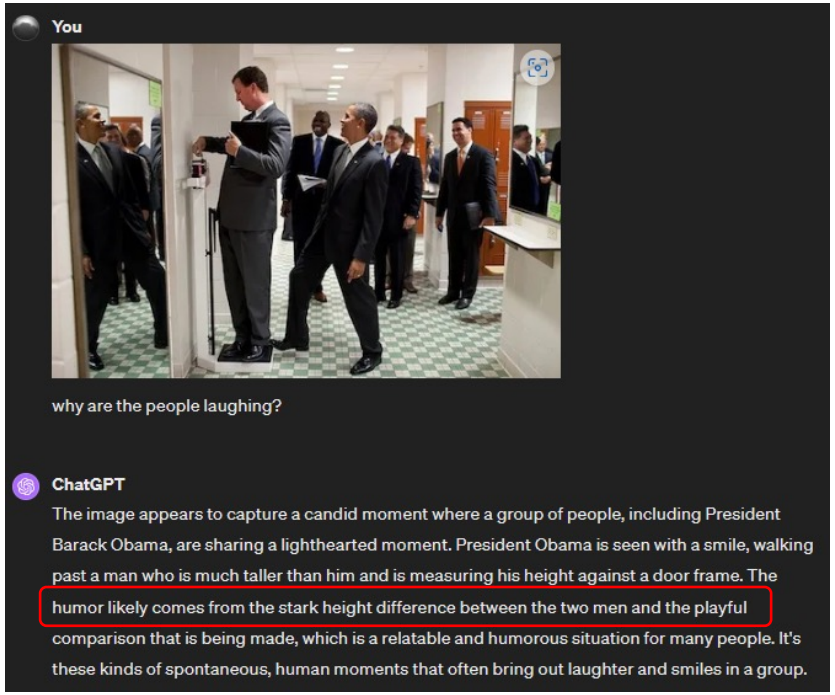


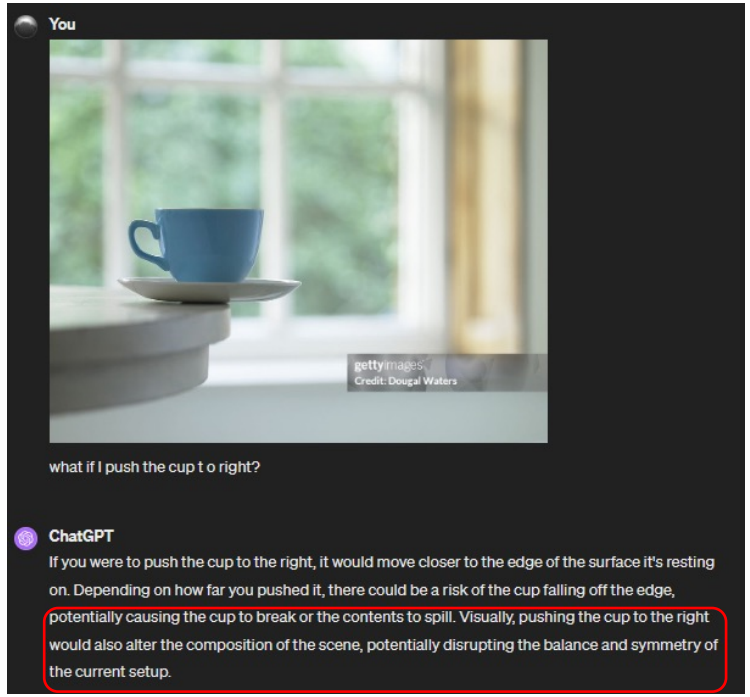
- There are five tomatoes in total

Counting

Current LMMs still fall short of tasks like



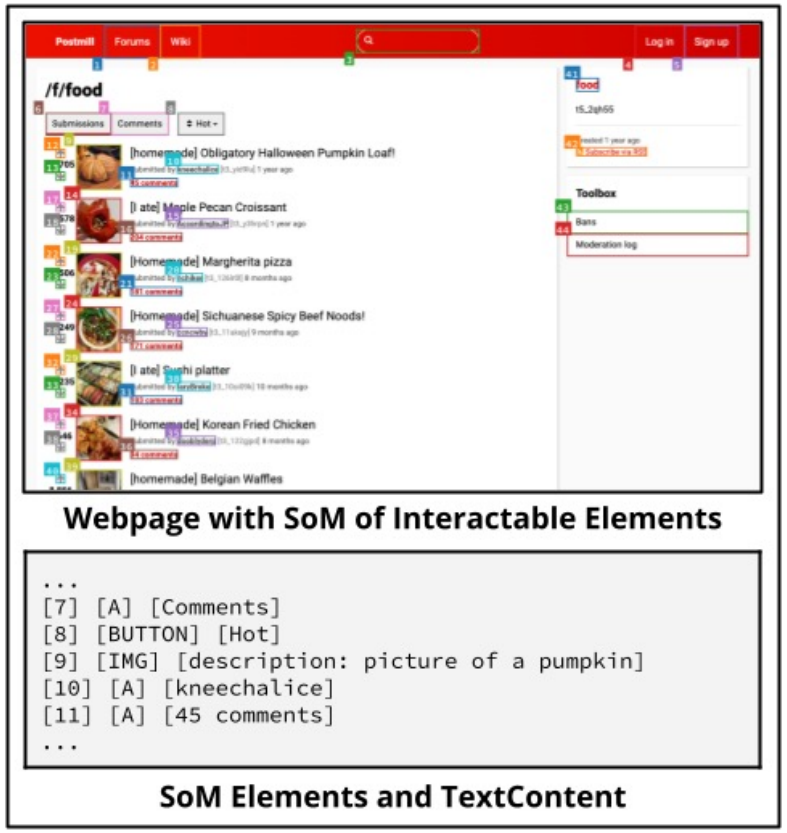


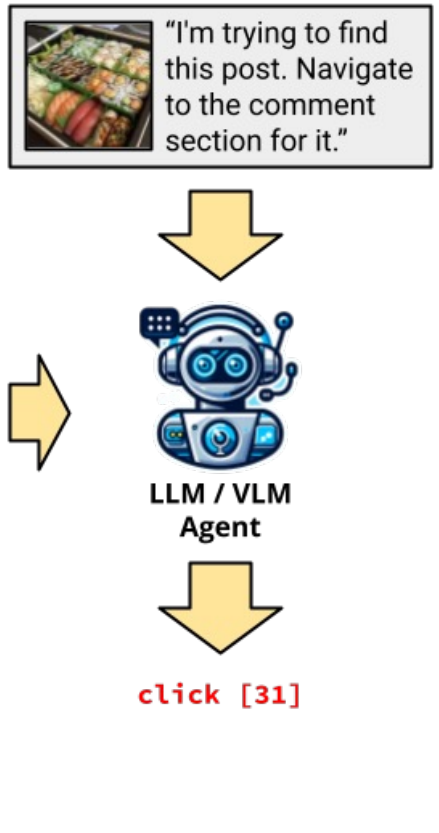


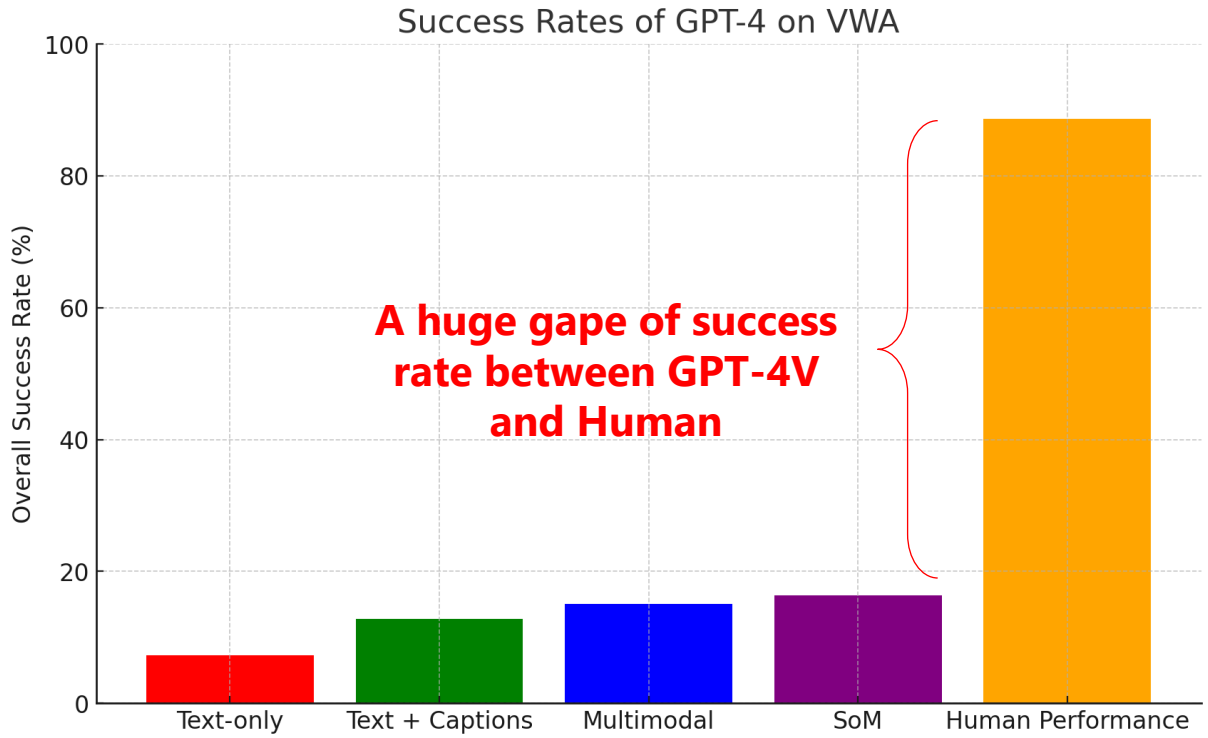
- **Planning**
- **Understand the humor in the image**
- **Causal-effect reasoning**

All answers are so general, ungrounded and less informative ☹

Current LMMs still fall short of realistic tasks like







- **Agent Navigation**

Opportunities: More Capable on Images

Q: Is the dog facing left or right from the camera's perspective?

The dog is **facing to the right** from the camera's perspective. You can tell by the position of its tail and the orientation of its body.

Q: Do you see any window in this image?

No, there are **no windows visible** in this image. The image shows a traditional red pillar box, which is a type of mailbox found in the United Kingdom.

Q: Is the door of the truck cab open?

No, the door of the truck is **not open** in the image provided.

Q: In this image, how many eyes can you see on the animal?

The image shows **one eye** of the animal. It's a close-up of a bald eagle facing slightly to its right, which presents only one side of its face to the viewer.

Q: How many wheels can you see in the image?

In the image provided, I can see **two wheels** on the visible side of the car.

Q: Does the keyboard have a backlight?

Based on the image, it appears that the keyboard **does have a backlight**. This is indicated by the illumination seen around the keys, which suggests that there is a light source beneath them to enhance visibility in low-light conditions.

Q: From the camera's perspective, is the piano's back panel on the left or right side?

From the camera's perspective, the piano's back panel is **on the left side**. The keys visible in the foreground suggest that we looking at the piano from front-right side, thus placing the back panel on the left from this angle.

Q: Is the edge of the heart in the picture white or dark-colored?

The image you've provided is small and of low resolution, but it appears that the hearts have a **dark-colored edge** or outline. This outline varies in visibility, likely due to the quality of the image.

Q: Is the school bus parked facing the camera or away from the camera?

The school bus is parked **facing away from the camera**. You can tell because the rear lights and the back of the exterior rearview mirror are visible. The text "SCHOOL BUS" is also reversed, which is typically seen on the front of a bus, designed so that it reads correctly in a rearview mirror.

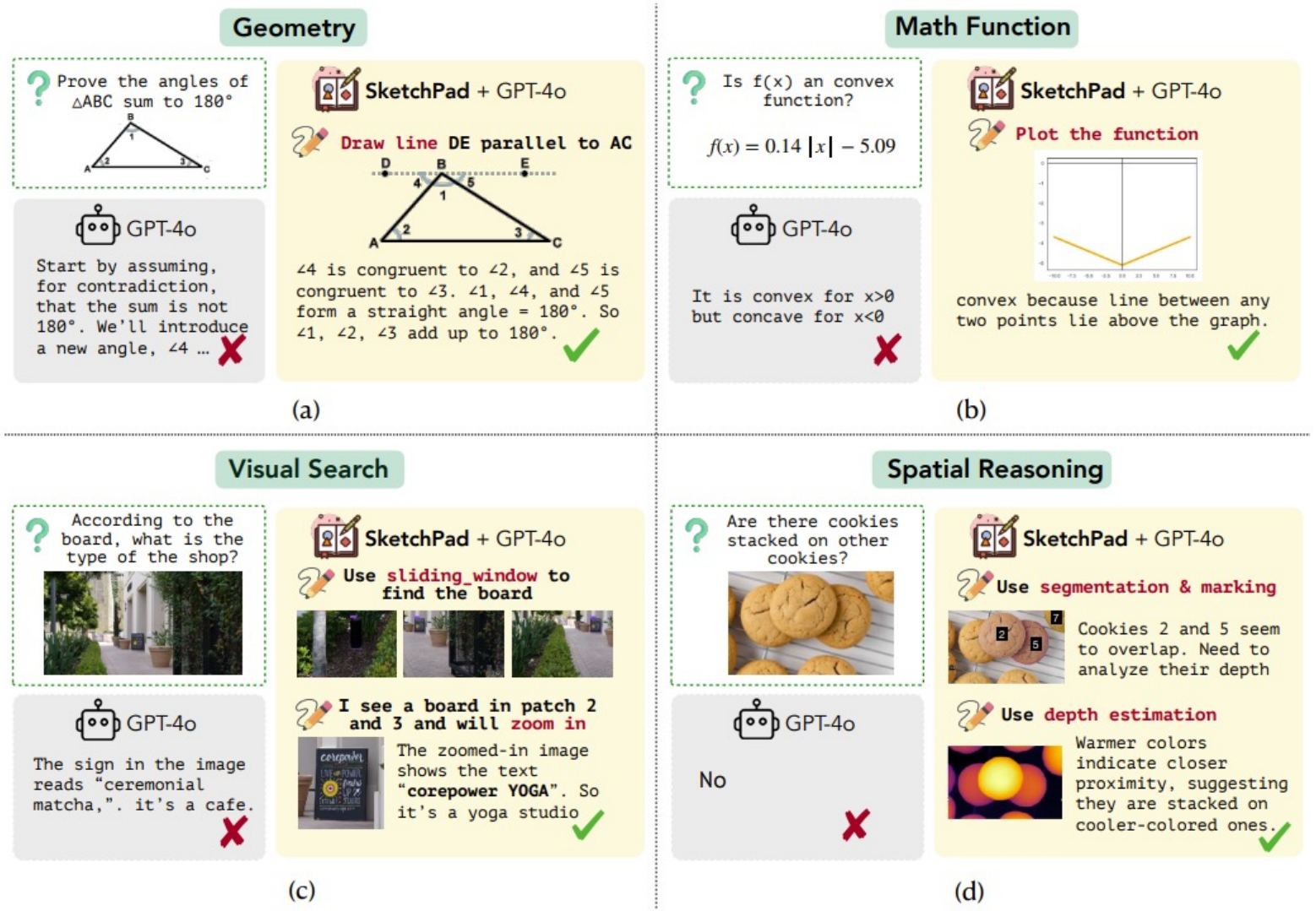
Q: Are the butterfly's feet visible?

The butterfly's feet are **not visible** in this image. The wings and part of the body of the butterfly are in view, but the feet are either hidden beneath the body or out of the frame of the picture. Butterflies have delicate, slender feet which are often not the most noticeable feature when they are perched, especially from the angle shown in this image.

Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. Tong et al.

Opportunities: More Capable on Multiple Images

Opportunities: More Advanced Prompting Techniques

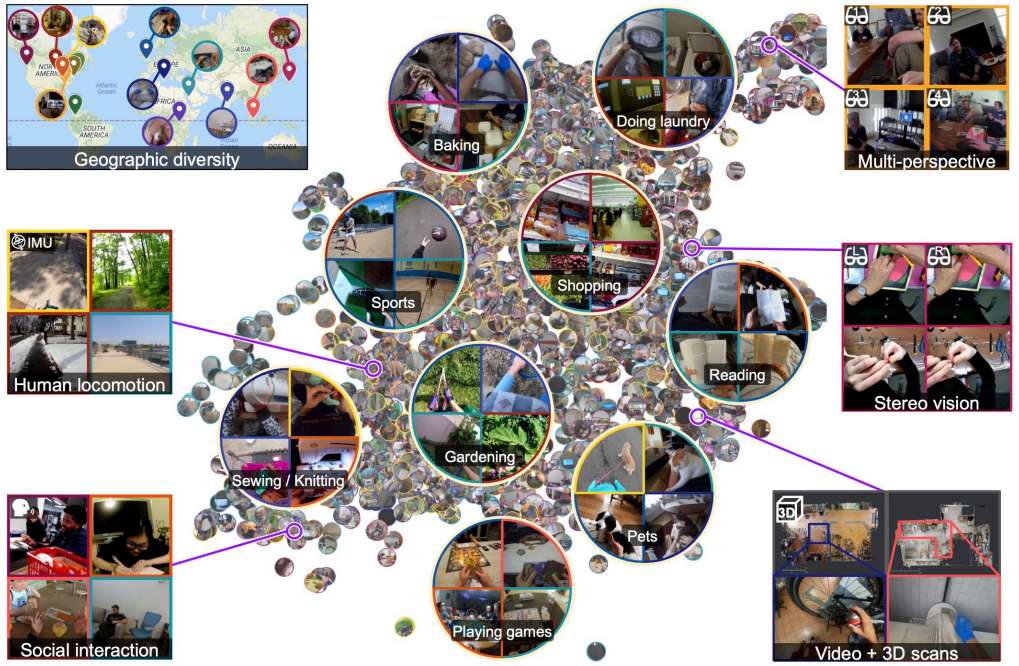


Visual SKETCHPAD: Sketching as a Visual Chain of Thought for Multimodal Language Models. Fu et al.

Opportunities: More Understanding of Videos

- Static image is insufficient to capture the world surrounding us

- Howto100M



- Ego4D

HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. Miech et al.
Ego4D: Around the World in 3,000 Hours of Egocentric Video. Grauman et al.

Opportunities: More Understanding of Videos

Q: What has been changed in the video?
A: The bottom drawer has been closed.

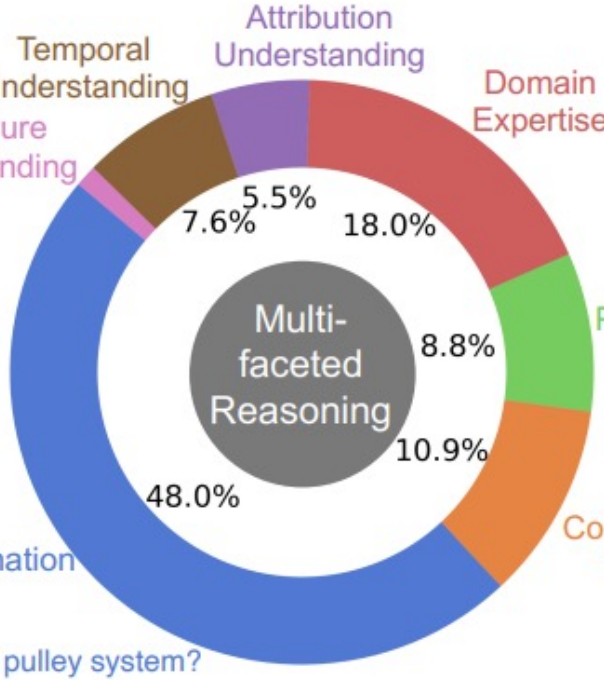
Q: How many animals appear in the video?
A: Two. There are a horse and a dog

Q: What is the reason that the lady decides to use the easy frost?
A: Because it has no-fuss frosting.

Q: What was first added into the milk?
A: Cocoa powder.

Q: What will happen next as the price is below the blue and red lines?
A: The price will go down.

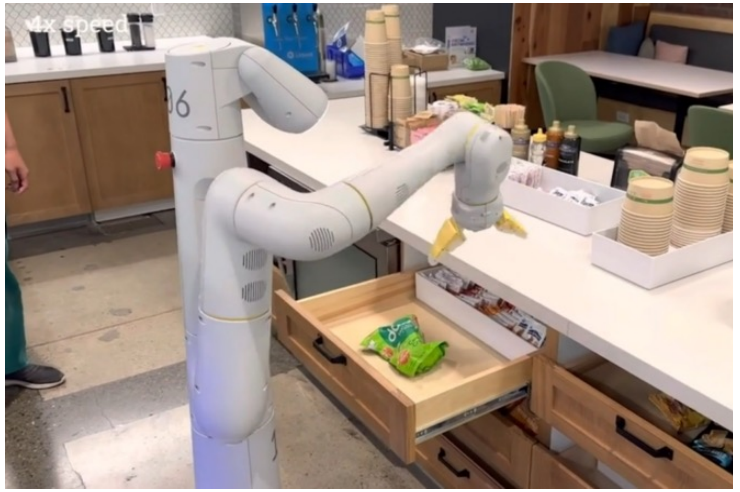
Q: How do the pulleys move when the hands are off the pulley system?
A: Two static and two moving upward.



Q: What would happen if the man skipped the step shown in the video?
A: The desktop of the coffee table will be upside down, which will make it impossible to mount the legs.

Opportunities: More Capable of Reasoning and Plannings in Real World

- An intelligent AI should be able to understand and interact with human and physical world



- Robotics

- Automatous Driving

Thanks for your attention!

Enjoy your stay at Seattle!