Intro to MLLMs. Image modality

Vlad Shakhuro

20 March 2025

Outline

1. MLLMs as step towards AGI
2. Static benchmarks
3. Arena
4. Architectures
5. Overview of several popular models

Idea

- Make models multimodal
- Use world knowledge and reasoning from LLMs
- Explore various applications that emerge at the intersection of the modalities
- Boost quality of models using patterns in multimodal data that don't exist in unimodal data

Outline

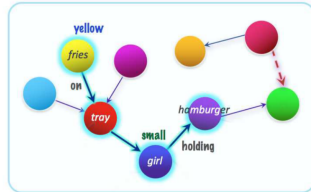
1. MLLMs as step towards AGI
2. Static benchmarks
3. Arena
4. Architectures
5. Overview of several popular models

Static benchmarks: GQA

Pattern: What|Which <type> [do you think] <is> <dobject>, <attr> or <decoy>?
Program: **Select:** <dobject> → **Choose** <type>: <attr>|<decoy>
Reference: The food on the red object left of the small girl that is holding a hamburger
Decoy: brown

What color is the food on the red object left of the small girl that is holding a hamburger, yellow or brown?

Select: hamburger → **Relate:** girl,holding → **Filter** size: small → **Relate:** object, left → **Filter** color: red → **Relate:** food,on → **Choose** color: yellow | brown



Graph Normalization

- Ontology construction
- Edge Pruning
- Object Augmentation
- Global Properties

Question Generation

- Pattern Collection
- Compositional References
- Decoy Selection
- Probabilistic Generation

Sampling and Balancing

- Distribution Balancing
- Type-Based Sampling
- Deduplication

Entailment Relations

- Functional Programs
- Entailment Relations
- Recursive Reachability

New Metrics

- Consistency
- Validity & Plausibility
- Distribution
- Grounding

Hudson et al. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. CVPR 2019

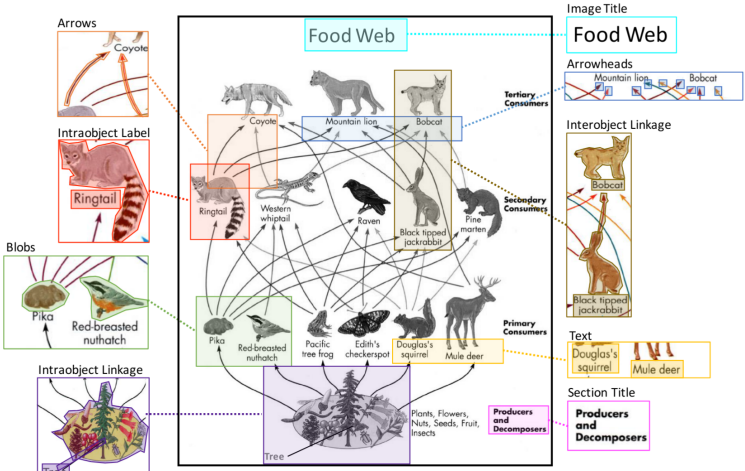
Static benchmarks: GQA

A1. Is the **tray** on top of the **table** black or light brown? light brown
A2. Are the **napkin** and the **cup** the same color? yes
A3. Is the small **table** both oval and wooden? yes
A4. Is there any **fruit** to the left of the **tray** the **cup** is on top of? yes
A5. Are there any **cups** to the left of the **tray** on top of the **table**? no
B1. What is the brown **animal** sitting inside of? **box**
B2. What is the large **container** made of? cardboard
B3. What **animal** is in the **box**? **bear**
B4. Is there a **bag** to the right of the green **door**? no
B5. Is there a **box** inside the plastic **bag**? no

- questions generated using scene graph of images
- 22.6M questions for 113k images
- evaluation metric: accuracy along with 5 more detailed metrics

Hudson et al. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. CVPR 2019

Static benchmarks: AI2D

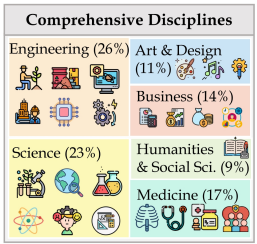


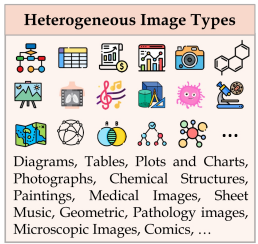
- 15k mutiple choice questions for 5k school grade diagrams
- diagram parse graphs are available
- 2 tasks: image parse graph; parse graph, question answer
- evaluation metric: accuracy

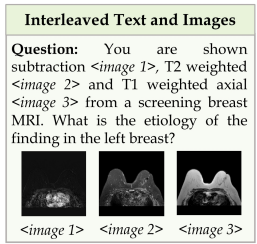
Kembhavi et al. A Diagram Is Worth A Dozen Images. ECCV 2016

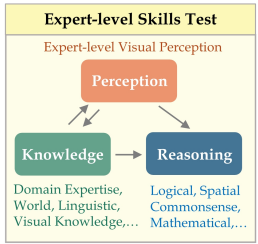
Static benchmarks: MMMU

- 11.5k questions from 6 university disciplines
- answers are extracted using regexps
- evaluation metric: accuracy









Yue et al. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. CVPR 2024

Static benchmarks: MMMU

Art & Design

Question: Among the following harmonic intervals, which one is constructed incorrectly?

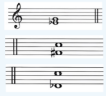
Options:

(A) Major third <*image 1*>

(B) Diminished fifth <*image 2*>

(C) Minor seventh <*image 3*>

(D) Diminished sixth <*image 4*>



Subject: Music; **Subfield:** Music;
Image Type: Sheet Music;
Difficulty: Medium

Business

Question: ...The graph shown is compiled from data collected by Gallup <*image 1*>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82?

Options:

(A) 0 (B) 0.2142

(C) 0.3571 (D) 0.5

Subject: Marketing; **Subfield:** Market Research; **Image Type:** Plots and Charts;
Difficulty: Medium

Science

Question: <*image 1*> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.

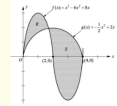
Options:

(A) $\int_0^{1.5}[f(x)-g(x)]dx$

(B) $\int_0^{1.5}[g(x)-f(x)]dx$

(C) $\int_0^{2}[f(x)-g(x)]dx$

(D) $\int_0^{2}[g(x)-x(x)]dx$



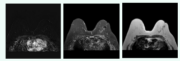
Subject: Math; **Subfield:** Calculus;
Image Type: Mathematical Notations;
Difficulty: Easy

Health & Medicine

Question: You are shown subtraction <*image 1*>, T2 weighted <*image 2*> and T1 weighted axial <*image 3*> from a screening breast MRI. What is the etiology of the finding in the left breast?

Options:

(A) Susceptibility artifact
(B) Hematoma
(C) Fat necrosis (D) Silicone granuloma



Subject: Clinical Medicine; **Subfield:** Clinical Radiology; **Image Type:** Body Scans: MRI, CT.;
Difficulty: Hard

Humanities & Social Science

Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? <*image* 1>

Option:
(A) Oppressor
(B) Imperialist
(C) Savior (D) Isolationist

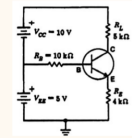
Subject: History; **Subfield:** Modern History; **Image Type:** Comics and Cartoons;
Difficulty: Easy

Tech & Engineering

Question: Find the VCE for the circuit shown in <*image 1*>. Neglect VBE

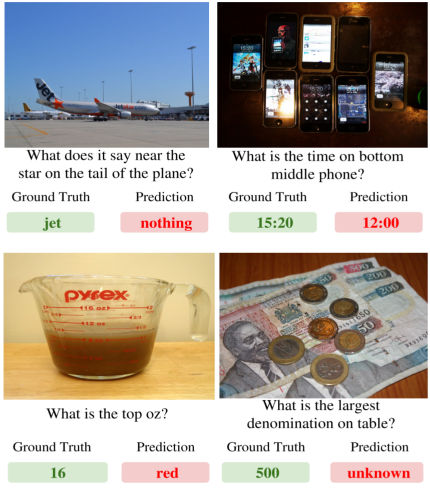
Answer: 3.75

Explanation: …IE = [(VEE) / (RE)] = [(5 V) / (4 k-ohm)] = 1.25 mA; VCE = VCC - IERL = 10 V - (1.25 mA) 5 k-ohm; VCE = 10 V - 6.25 V = 3.75 V



Subject: Electronics; **Subfield:** Analog electronics; **Image Type:** Diagrams;
Difficulty: Hard

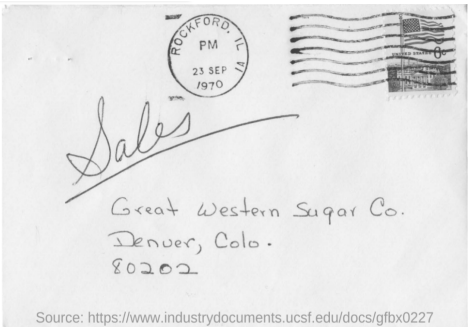
Static benchmarks: TextVQA



- 45k questions for 28k images, 10 answers per question
- evaluation metric: VQA accuracy (100% correct if 3 humans provided that answer)

Singh et al. Towards VQA Models That Can Read. CVPR 2019

Static benchmarks: DocVQA



Q: Mention the ZIP code written?
A: 80202
Q: What date is seen on the seal at the top of the letter?
A: 23 sep 1970
Q: Which company address is mentioned on the letter?
A: Great western sugar Co.

- 50k questions for 12k images
- documents mostly from 1960–2000, industries: tobacco, food, drug, chemical, fossil fuel
- evaluation metrics: Average Normalized Levenstein Similarity, Accuracy

Mathew et al. DocVQA: A Dataset for VQA on Document Images. WACV 2021

Static benchmarks: DocVQA


nitrogen

by Dr. Dwayne G. Westfall

Recommendations

Dr. Dwayne G. Westfall
Soils/Plant Nutritionist

A native of Herndon, Idaho, Dwayne received his Bachelor's Degree in Agronomy from the University of Idaho in 1961. He did his graduate work there and at Washington State University in 1967. After graduating from the University of Idaho, he worked for Simplot Soil Builders, Inc., a potato processing firm in American Falls, Idaho. Two years of Army service followed, including assignment to the Army Biological Laboratory. Dwayne joined Texas A & M University as assistant professor in 1969 and was promoted to associate professor in 1973. In September of 1975 he joined Great Western's agricultural research staff.

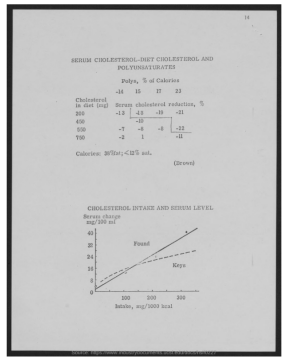
Effect of Nitrogen Level on Sucrose Percentage and Yield

N Level* lbs/A	District NCC Nebraska Sucrose %	Yield T/A	Sucrose %	Yield T/A	NEC&K Sucrose %	Yield T/A	Average Sucrose %	Yield T/A
100-200	16.4	22.7	17.4	20.4	16.0	20.8	16.6	21.3
201-300	16.4	22.8	17.1	21.9	16.4	20.8	16.6	21.8
301-400	16.2	22.5	16.5	22.0	15.8	19.7	16.2	21.4
401-500	15.9	22.5	15.9	22.8	15.3	22.4	15.7	22.6
501+			16.3	21.8	15.5	18.7	15.9	20.2

*Fertilizer nitrogen applied plus residual soil nitrogen

2

Source: https://www.industrydocuments.ucsf.edu/docs/klnn0226


SERUM CHOLESTEROL–DIET CHOLESTEROL AND POLYUNSATURATES

Polys, % of Calories

Cholesterol in diet (mg)	14	15	17	23
	Serum cholesterol reduction, %			
200	-13	-13	-13	-21
450		-12		
500	-7	-8	-8	-22
750	-2	1		-11

Calories: 36%fat; <12% sat.

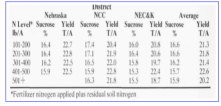
(Brown)

CHOLESTEROL INTAKE AND SERUM LEVEL

Serum change mg/100 ml

Source: https://www.industrydocuments.ucsf.edu/docs/klnn0226

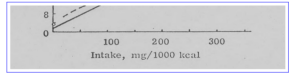
Q: Whose picture is given?
Question types: *photograph* and *layout*
A: Dr. Dwayne G. Westfall



N Level* lbs/A	District NCC Nebraska Sucrose %	Yield T/A	Sucrose %	Yield T/A	NEC&K Sucrose %	Yield T/A	Average Sucrose %	Yield T/A
100-200	16.4	22.7	17.4	20.4	16.0	20.8	16.6	21.3
201-300	16.4	22.8	17.1	21.9	16.4	20.8	16.6	21.8
301-400	16.2	22.5	16.5	22.0	15.8	19.7	16.2	21.4
401-500	15.9	22.5	15.9	22.8	15.3	22.4	15.7	22.6
501+			16.3	21.8	15.5	18.7	15.9	20.2

*Fertilizer nitrogen applied plus residual soil nitrogen

Q: What is the average sucrose % for N level 501+ ?
Question types: *table*
A: 15.9



Q: What is the highest value for "Intake, mg/1000kcal" plotted on the 'X' axis of the graph?
Question types: *figure*
A: 300

Mathew et al. DocVQA: A Dataset for VQA on Document Images. WACV 2021

Outline

1. MLLMs as step towards AGI
2. Static benchmarks
3. Arena
4. Architectures
5. Overview of several popular models

WildVision-Arena

Figure 1: WILDVISION-ARENA (WV-ARENA) supports multi-round multimodal chats with 20+ models, enabling the comparison of VLMs in real-world scenarios. We curate WILDVISION-BENCH (WV-BENCH) by selecting 500 samples from $20k+$ in-the-wild chats and $8k+$ user ratings. Automatic model scorings on WV-BENCH closely correlate with the Elo ratings on WV-ARENA.

UI

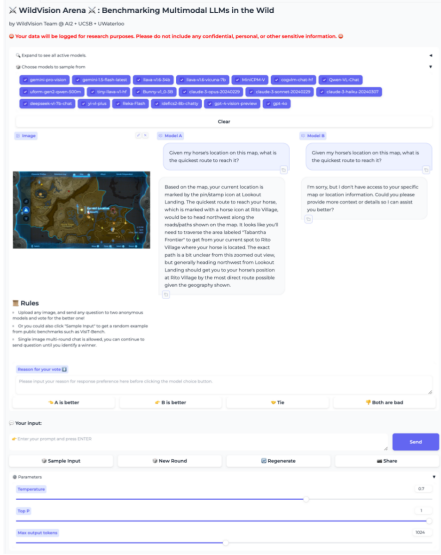


Figure 8: User Interface of WILDVISION-ARENA.

Question distribution

Statistic	Number
Total Votes	8,076
Anonymous	6,636
Non-anonymous	1,440
Left Vote	2,932
Right Vote	2,839
Tie Vote	979
Bad Vote	1,326
Days	102
Total Round	10,884
Avg Round	1.34
Avg Token Input	31.00
Avg Token Output	108.87

Table 1: Statistics of votings in WV-ARENA.

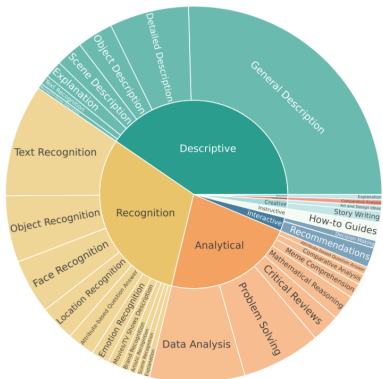


Figure 2: Question Category

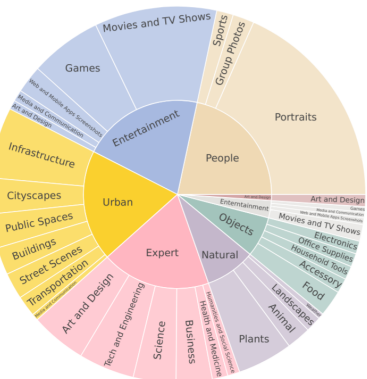


Figure 3: Image Domain

Battles

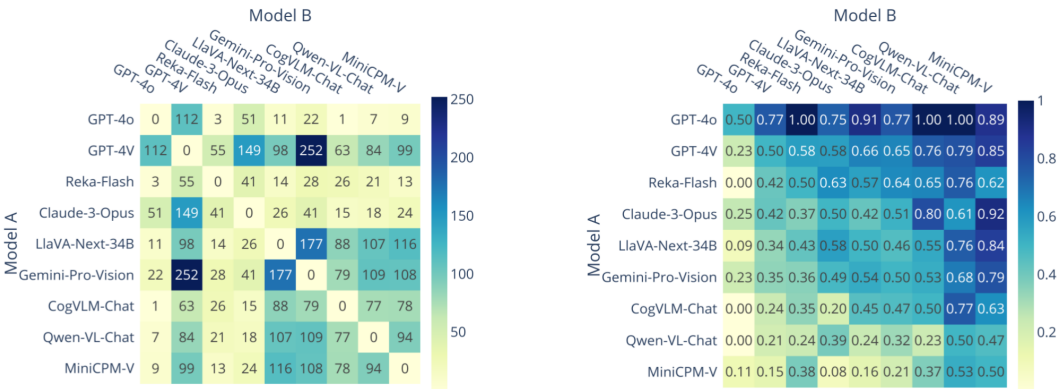


Figure 4: Battle Count Heatmap (Left): the number of voted comparisons between models. Win Fraction Heatmap (Right): the winning rate of Model A over Model B in voted comparisons.

Elo computation

Online Elo Rating Elo rating focuses on modeling the probability of player i winning against player j given their existing ratings R_i and R_j respectively, where $i, j \in N$. We define a binary outcome Y_{ij} for each comparison between player i and player j, where $Y_{ij} = 1$ if player i wins against player j, and $Y_{ij} = 0$ otherwise. Then the logistic probability is formulated as:

$$P(Y_{ij} = 1) = \frac{1}{1 + 10^{(R_j - R_i)/\alpha}}, \tag{1}$$

where $\alpha = 400$ for Elo rating computation. After a match, each player's rating is updated by the formula: $R'_i = R_i + K \times (S(i|j) - E(i|j))$, where $S(i|j)$ is the actual match outcome (1 for a win, 0.5 for a tie, and 0 for a loss), and $E(i|j) = P(Y_{ij} = 1)$. The higher-rated player will win fewer points if they win but lose more if they lose, while the lower-rated player will experience the opposite. The computation of the online Elo rating is correlated with the comparison order. Therefore, we follow Chatbot Arena to adopt the Bradley–Terry model [9] for a stable statistical estimation.

Statistical Estimation The Bradley–Terry model [9] estimates the Elo rating using a logistic regression model and maximum likelihood estimation (MLE). Let's say there are N players, and we have a series of pairwise comparisons, where W_{ij} is the number of times player i wins against player j. The log-likelihood function for all pairwise comparisons can be written as:

$$\mathcal{L}(\mathbf{R}) = \sum_{i,j \in N, i \neq j} (W_{ij} Y_{ij} \log P(Y_{ij} = 1)) \tag{2}$$

Leaderboard

Table 2: WILDVISION-ARENA Leaderboard. We show the full elo score and within three question categories (`Analytical`, `Descriptive`, `Recognition`) and three image domains (`Entertainment`, `Objects`, `Expert`) of 22 models with a time cutoff at May 29, 2024. **Best** Second Best Best among proprietary models Best among open-source models.

Models	Size	Elo	Battles	MMMU	Question Category			Image Domain		
					Analyt.	Descri.	Recogn.	Entert.	Objects	Expert
GPT-4O [69]	–	**1235**	434	**62.8**	**1290**	**1250**	**1236**	**1362**	**1203**	**1293**
GPT-4-Vision [68]	–	1132	2288	56.8	1154	1169	1099	1177	1109	1178
Reka-Flash [83]	–	1107	513	56.3	1093	1141	1067	1069	1101	1191
Claude-3-OPUS [2]	–	1100	908	59.4	1117	1096	1092	1111	1127	1128
Gemini-Pro-Vision [82]	–	1061	2229	47.9	1099	1041	1090	1088	1077	1041
Yi-VL-PLUS [1]	–	1061	283	–	1084	1040	1078	1001	1119	1101
LLaVA-NEXT [48]	$34B$	1059	1826	51.1	1068	1104	1021	1074	1015	1052
Gemini-1.5-Flash [81]	–	1055	132	–	1090	1018	1085	1190	990	1127
Claude-3-Sonnet [2]	–	1044	496	53.1	1063	1056	1041	1033	1023	1119
CogVLM-Chat-HF [89]	$13B$	1016	1024	32.1	950	947	1006	955	930	950
Claude-3-Haiku [2]	–	1002	419	50.2	964	1008	996	1033	1014	1005
LLaVA-NEXT [48]	$7B$	992	1367	35.1	963	1032	977	992	1023	1001
DeepSeek-VL [51]	$7B$	979	646	36.6	988	984	953	956	1026	962
Idefics2 [37]	$8B$	965	100	36.6	818	1003	1011	909	1071	1020
LLaVA-NEXT [48]	$13B$	956	201	35.9	965	974	1006	975	971	987
Qwen-VL-Chat [5]	$10B$	930	1328	35.9	898	937	940	923	942	902
Bunny-V1 [23]	$3B$	921	389	38.2	897	922	878	884	823	823
MiniCPM-V [26]	$3B$	910	1349	34.7	895	911	925	888	890	840
LLaVA-v1.5 [47]	$13B$	891	299	36.4	952	838	920	887	827	914
Tiny-LLaVA-v1-HF [111]	$3B$	879	288	33.1	901	828	821	808	853	894
InstructBLIP [14]	$7B$	862	807	30.6	834	856	891	840	902	763
UFORM-Gen2-Qwen [86]	$500M$	827	452	–	911	785	853	768	937	830

Per-domain quality

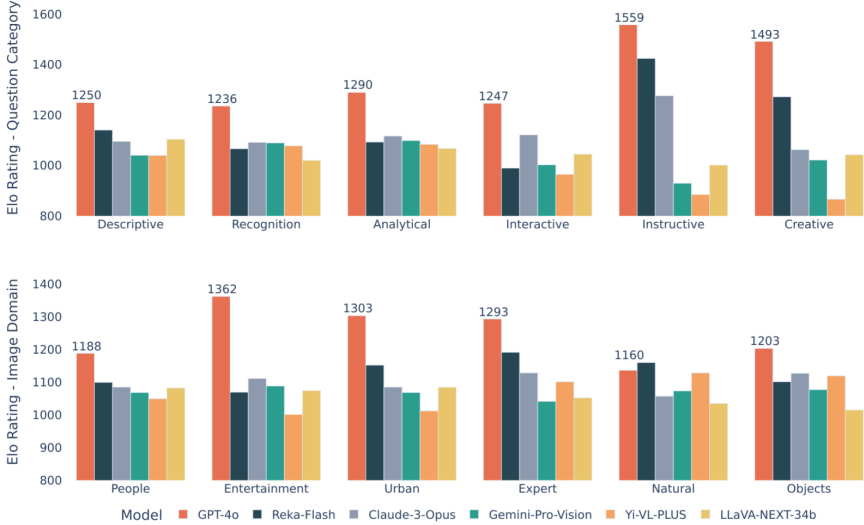


Figure 5: Elo ratings of six models across question categories (Top) and image domains (Bottom).

WV-Bench

500 data samples from Arena with expert annotation

Table 3: VLMs' responses on two cases from WILDVISION-BENCH expert annotated samples. The example #61 is a hard case that all models fall short at.

Image *#44* in WildVision Bench

Text Prompt: Answer shortly what is he doing? And what do we call him?

GPT-4V: He is performing a breakdance move. We call him a breakdancer or b-boy.
LLaVA-v1.6-34B: He is performing a dance move that involves spinning or flipping while balancing on one hand ..., he may be referred to as a dancer or acrobat.
Gemini Pro Vision: He is breakdancing. He is called a b-boy or breaker.

Image *#61* in WildVision Bench

	6	8	4		2	1		
			9			8		7
			6					4
	8		1		3	7	4	9
7	3	4		8		6		
	1	9	7			3	5	
8				9			7	6
		3		6		9	8	1
	2				7	4		

Text Prompt: Convert the first row of this sudoku matrix into the text format, each cell separated by a space. Use 0 to represent empty cells.

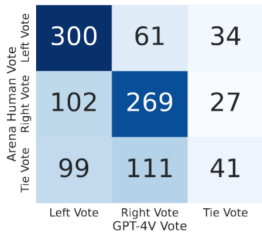
GPT-4V: 6 8 4 2 1 0 0 0 0

LLaVA-v1.6-34B: 6 8 4 0 2 1 0 7 3 4 0 6

Gemini Pro Vision: 6 8 4 2 1 9 8 7 7

WV-Bench

Evaluation is done using GPT-4o as judge and Claude-3-Sonnet as reference



Metric vs Human	GPT-4v		
	4-way	3-way	Binary
F1 Score (Macro)	0.4245	0.5143	0.7792
F1 Score (Micro)	0.5747	0.5842	0.7796
F1 Score (Weighted)	0.5407	0.5536	0.7798
Cohen's Kappa Score	0.3404	0.3442	0.5585
Pearson Correlation	0.2906	0.2880	0.5587

Figure 6: Left: GPT-4V vs. Arena Human Voting. Right: Agreement; 4-way: left/right/tie/bad vote. 3-way: left/right/other. Binary: left/right vote

WV-Bench

Evaluation is done using GPT-4o as judge and Claude-3-Sonnet as reference

Table 4: Estimated model scores of VLMs on WILDVISION-BENCHtest split of 500 samples.

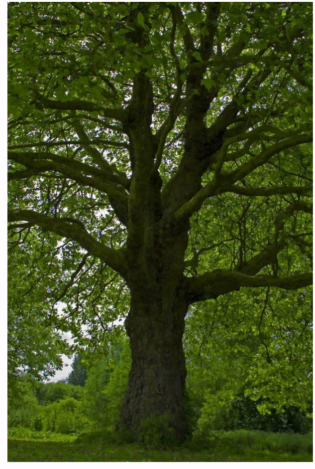
Model	Score	95% CI	Win Rate	Reward	Much Better	Better	Tie	Worse	Much Worse	Avg Tokens
GPT-4o [69]	89.41	(−1.7, 2.0)	80.6%	56.4	255.0	148.0	14.0	72.0	11.0	157
GPT-4-Vision [68]	80.01	(−1.9, 2.8)	71.8%	39.4	182.0	177.0	22.0	91.0	28.0	140
Reka-Flash [83]	64.79	(−2.9, 3.0)	58.8%	18.9	135.0	159.0	28.0	116.0	62.0	181
Claude-3-Opus [2]	62.15	(−2.8, 3.4)	53.0%	13.5	103.0	162.0	48.0	141.0	46.0	120
Yi-VL-PLUS [1]	55.09	(−2.9, 3.0)	52.8%	7.2	98.0	166.0	29.0	124.0	83.0	150
LLaVA-NEXT-34B [48]	51.91	(−3.1, 2.4)	49.2%	2.5	90.0	156.0	26.0	145.0	83.0	165
Claude-3-Sonnet [2]	50.00	–	–	–	–	–	–	–	–	120
Claude-3-Haiku [2]	37.70	(−3.2, 4.2)	30.6%	−16.5	54.0	99.0	47.0	228.0	72.0	97
Gemini-Pro-Vision [82]	35.45	(−2.6, 3.2)	32.6%	−21.0	80.0	83.0	27.0	167.0	143.0	66
LLaVA-NEXT-13B [48]	33.69	(−3.8, 2.7)	33.8%	−21.4	62.0	107.0	25.0	167.0	139.0	138
DeepSeek-VL-7B [51]	33.48	(−2.2, 3.0)	35.6%	−21.2	59.0	119.0	17.0	161.0	144.0	119
CogVLM-Chat-HF [89]	31.88	(−2.7, 2.4)	30.6%	−26.4	75.0	78.0	15.0	172.0	160.0	63
LLaVA-NEXT-7B [48]	26.15	(−2.7, 2.3)	27.0%	−31.4	45.0	90.0	36.0	164.0	165.0	139
Idefics2 [37]	23.71	(−2.4, 2.5)	26.4%	−35.8	44.0	88.0	19.0	164.0	185.0	128
Qwen-VL-Chat [5]	17.87	(−2.6, 2.2)	19.6%	−47.9	42.0	56.0	15.0	155.0	232.0	70
LLaVA-v1.5-13B [47]	14.15	(−2.2, 2.2)	16.8%	−52.5	28.0	56.0	19.0	157.0	240.0	87
Bunny-3B [23]	12.70	(−1.8, 1.9)	16.6%	−54.4	23.0	60.0	10.0	164.0	243.0	76
MiniCPM-V [26]	11.66	(−1.8, 2.1)	13.6%	−57.5	25.0	43.0	16.0	164.0	252.0	89
Tiny-LLaVA [111]	8.01	(−1.4, 1.4)	11.0%	−66.2	16.0	39.0	15.0	127.0	303.0	74
UFORM-Gen2-Qwen [86]	7.55	(−1.6, 1.1)	10.8%	−68.5	16.0	38.0	11.0	115.0	320.0	92
InstructBLIP-7B [14]	5.54	(−1.3, 1.5)	7.8%	−72.5	11.0	28.0	15.0	117.0	329.0	47

WV-Bench samples

Image [Entertainment-Movies/TV Shows]

[Descriptive-Movies/TV Shows] **Text Prompt:** What are the two giraffe characters on this movie poster doing?

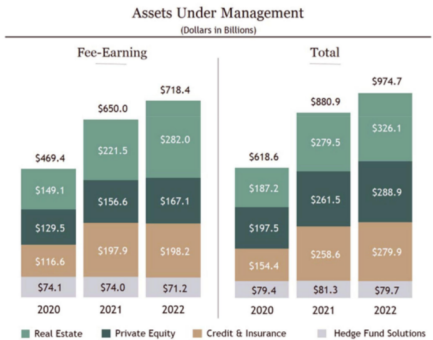
Image [Natural-Plants]



[Analytical-Problem Solving] **Text Prompt:** How likely is it to snow after this picture was taken? What would change with this type of tree before it's likely to snow?

WV-Bench samples

Image `[Expert-Business]`



`[Analytical-Data Analysis]` **Text Prompt:** Which of the companies featured in the dashboard are headquartered outside the US?

Image `[Urban-Infrastructure]`

`[Recognition-Text]` **Text Prompt:** Can you tell me the potential risks and the unreasonale parts in the image?

WV-Bench samples

Image [Urban-Buildings]

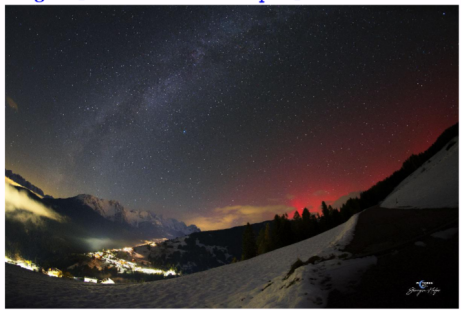
[Recognition-Location] **Text Prompt:** where is this?

Image [Expert-Science]

[Analytical-Safety Procedures] **Text Prompt:** Can you tell me the potential risks and the unreasonale parts in the image?

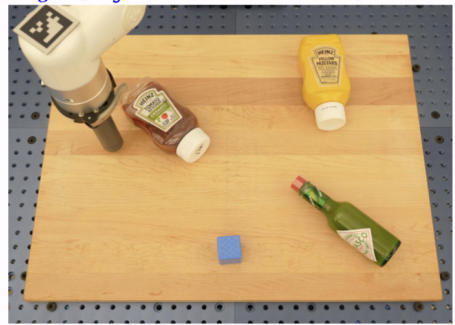
WV-Bench samples

Image `[Natural-Landscapes]`



`[Recognition-Location]` **Text Prompt:** where was this photo taken?

Image `[Objects-Household Tools]`



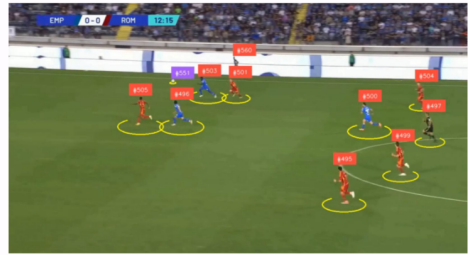
`[Descriptive-Object Description]` **Text Prompt:** describe the scene and objects

WV-Bench samples

Image [Entertainment-Web and Mobile Apps Screenshots]

[Interactive-Web Navigation] **Text Prompt:** I need to download flyer, you will be given screenshot from browser with elements marked with number. give next action to take on web page to download the flyersngive me response in below format example 1 action:[click,scroll,wait], box:1 format action:, box:

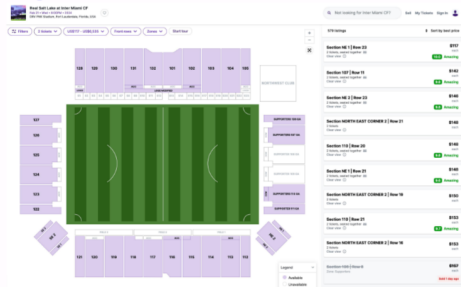
Image [Event-Sports]



[Descriptive-Scene Description] **Text Prompt:** this is a football match , every player has an identifier , describe every player action (example : player #501 is running)

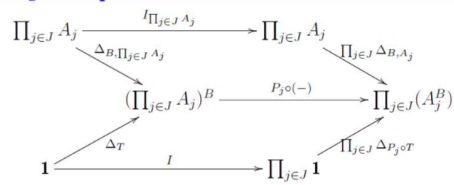
WV-Bench samples

Image `[Urban-Infrastructure]`



`[Interactive-Recommendations]` **Text Prompt:**
Which section's ticket would you recommend I purchase?

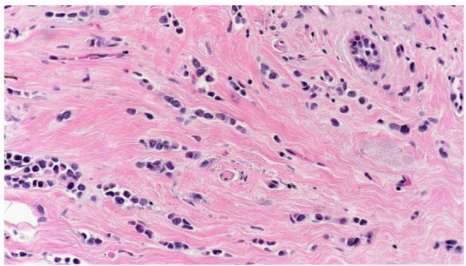
Image `[Expert-Science]`



`[Interactive-Code Generation]` **Text Prompt:**
Give me Latex code to create this diagram

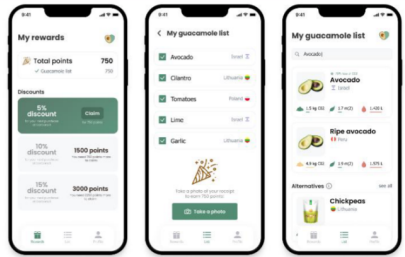
WV-Bench samples

Image [Expert-Health and Medcine]



[Recognition-Object] **Text Prompt:** what type of tumor is this?

Image [Entertainment-Web and Mobile Apps Screenshots]

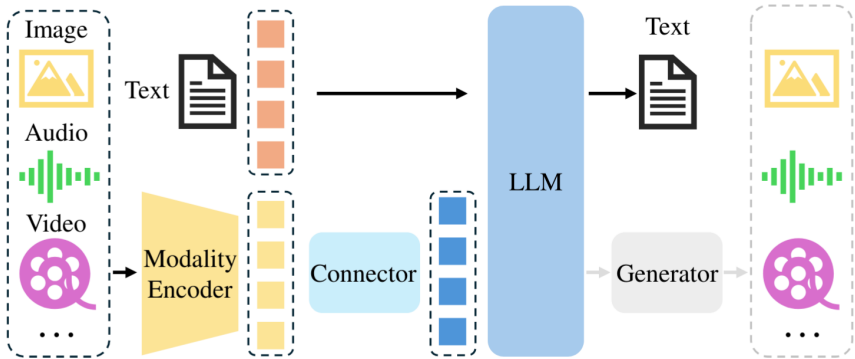


[Analytical-Critical Reviews] **Text Prompt:** Review each screenshot carefully, focusing on different aspects of usability...

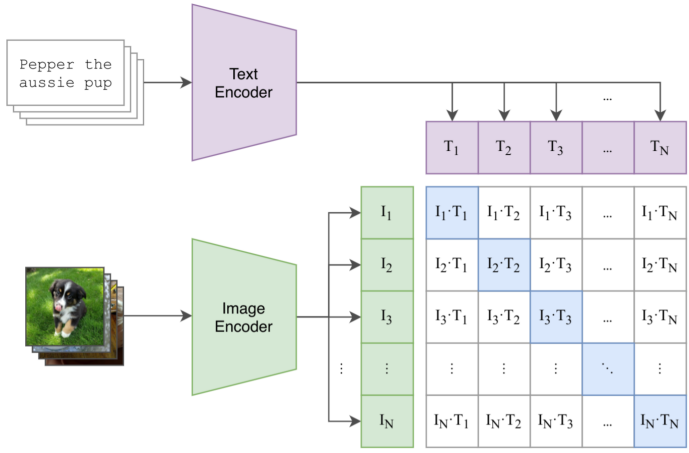
Outline

1. MLLMs as step towards AGI
2. Static benchmarks
3. Arena
4. Architectures
5. Overview of several popular models

General scheme



Encoder: CLIP



400M (image, text) pairs, 500×V100 GPUs for pretraining

Radford et al. Learning transferable visual models from natural language supervision. ICML 2021

Encoder: SigLIP

CLIP loss

$$-\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_i\cdot\mathbf{y}_j}}}^{\text{image}\to\text{text softmax}}+\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_j\cdot\mathbf{y}_i}}}^{\text{text}\to\text{image softmax}}\right)$$

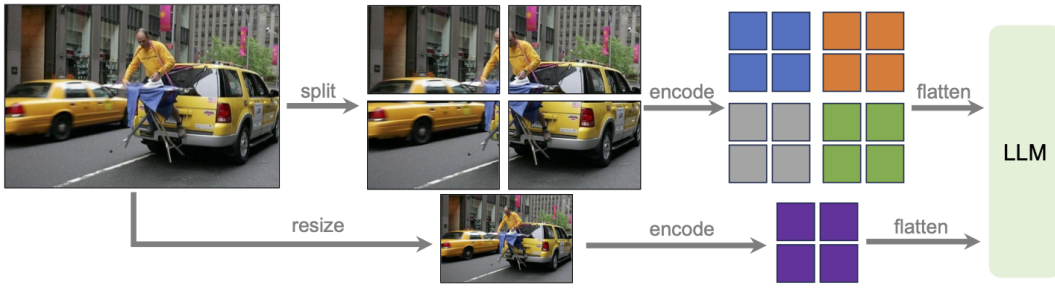
Sigmoid loss

$$-\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b)}}}_{\mathcal{L}_{ij}}$$

- turn multiclass classification into binary classification of all pair (image, text) combinations
- no global normalization, hence better scaling and memory efficiency

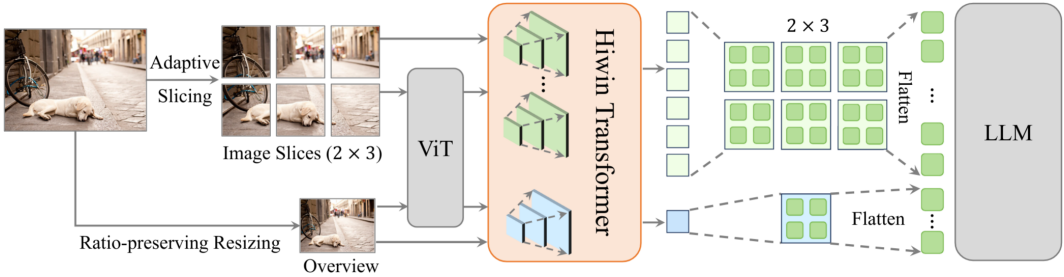
Zhai et al. Sigmoid Loss for Language Image Pre-Training. ICCV 2023

High-res: slicing and dual-branch



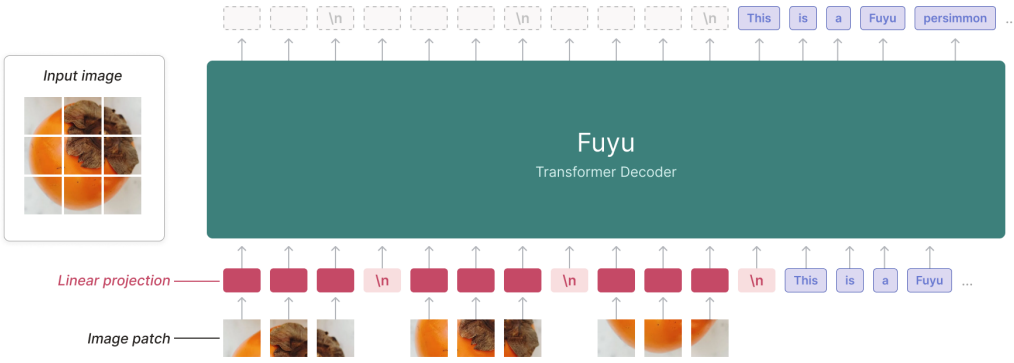
Liu et al. Improved Baselines with Visual Instruction Tuning. CVPR 2024

High-res: slicing and dual-branch

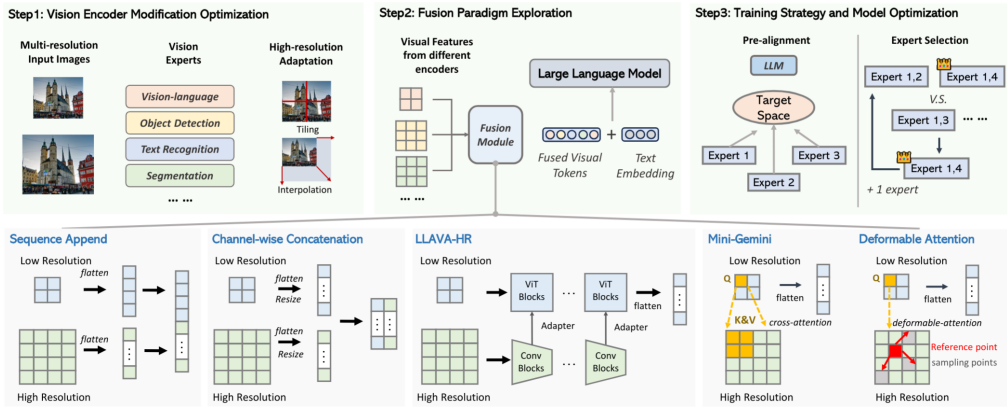


Guo et al. LLaVA-UHD v2: an MLLM Integrating High-Resolution Feature Pyramid via Hierarchical Window Transformer. arXiv:2412.13871

High-res: linear projection



Mixture of Encoders (MoE)



Shi et al. Eagle: Exploring The Design Space for Multimodal LLMs with Mixture of Encoders. ICLR 2025

Mixture of Encoders (MoE)

Table 4: **Comparison of different fusion methods for different vision experts.** "#Token(V)" denotes the number of visual tokens. "#Tokens/s" denotes the inference throughtput of the whole pipeline.

Vision Encoders	Fusion	#Token(V)	#Tokens/s	#Params	Avg.
CLIP + ConvNeXt	*Seq. Append*	2048	46.1	1200M	**690.5**
	Channel Concat.	1024	**47.3**	**1184M**	681.5
	LLaVA-HR	1024	47.0	1219M	678.7
	Mini-Gemini	1024	45.3	1201M	672.5
	Deformable Attn.	1024	**47.3**	1201M	674.3
CLIP + ConvNeXt + SAM	*Seq. Append*	3072	40.3	1529M	686.2
	Channel Concat.	1024	**46.3**	**1495M**	**690.4**

Shi et al. Eagle: Exploring The Design Space for Multimodal LLMs with Mixture of Encoders. ICLR 2025

MoVA

Table 1: **Comparison of CLIP *vs.* state-of-the-art task-specific vision encoders.** Our evaluation criteria encompass a variety of dimensions: comprehensive benchmarks [16], text-oriented Visual Question Answering (VQA) [17, 18], general VQA [19], object hallucination [20], Referring Expression Comprehension (REC) [21], Referring Expression Segmentation (RES) [21], and medical VQA benchmark SLAKE [22]. We use the same data for each model.

Vision Encoder	Task	MMB	DocVQA	ChartQA	GQA	POPE	REC	RES	SLAKE
CLIP [11]	Image-text Contrastive	**64.9**	35.6	35.3	62.5	85.7	81.5	43.3	63.7
DINOv2 [15]	Visual Grounding	57.5	14.7	15.9	**63.9**	86.7	**86.1**	47.5	59.4
Co-DETR [23]	Object Detection	48.4	14.2	14.8	58.6	**88.0**	82.1	48.6	55.3
SAM [24]	Image Segmentation	40.7	13.9	15.0	54.0	82.0	79.2	**49.3**	57.7
Pix2Struct [25]	Text Recognition	41.9	**57.3**	53.4	51.0	78.1	59.2	32.2	44.0
Deplot [26]	Chart Understanding	36.2	40.2	**55.8**	48.1	75.6	51.1	27.0	44.5
Vary [12]	Document Chart Parsing	28.1	47.8	41.8	42.6	69.1	21.6	16.0	40.9
BiomedCLIP [27]	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**

Zong et al. MoVA: Adapting Mixture of Vision Experts to Multimodal Context. NeurIPS 2024

MoVA

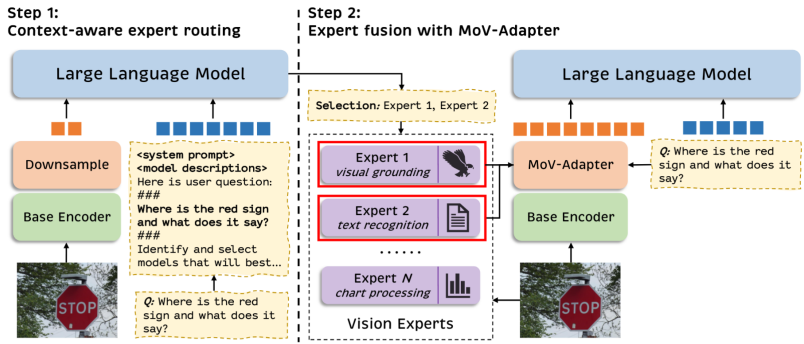


Figure 1: **The pipeline of MoVA.** MoVA performs coarse-to-fine routing to solve a given question. The coarse context-aware expert routing is performed in the first stage to select context-relevant experts. Next, we adopt the MoV-Adapter to extract and fuse the task-specific knowledge from these selected experts in a fine-grained manner.

Zong et al. MoVA: Adapting Mixture of Vision Experts to Multimodal Context. NeurIPS 2024

MoVA

Table 2: One example of the instruction-following data for context-aware expert routing. We present the multimodal inputs in the top block and the language response in the bottom block. The detailed model descriptions are released in the Appendix.

Routing Prompt Input

You are a helpful assistant router. Based on the visual content, questions, and model pool the user provides, you need to consider the expertise of these models to select the most 3 suitable models to help you answer the questions. Answer with the model's letter from the given choices directly. If no models are selected, just answer 'none'.

Model pool:

A. <DINOv2 model description>
B. <Co-DETR model description>
C. <SAM model description>
D. <Pix2Struct model description>
E. <Deplot model description>
F. <Vary model description>
G. <BiomedCLIP model description>

Question:

Where is the red sign and what does it say?

Routing Prompt Output

A, D

Context compression: MQT

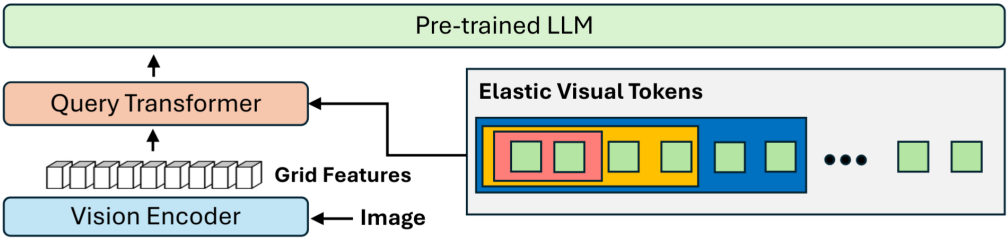
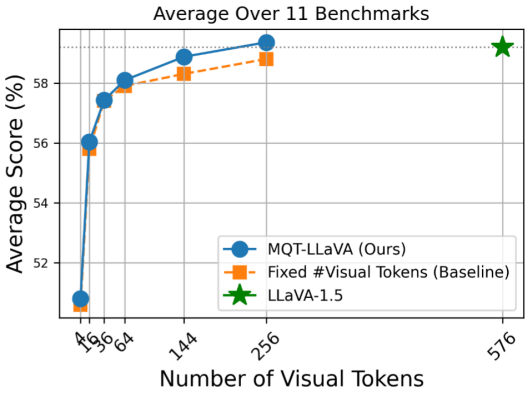
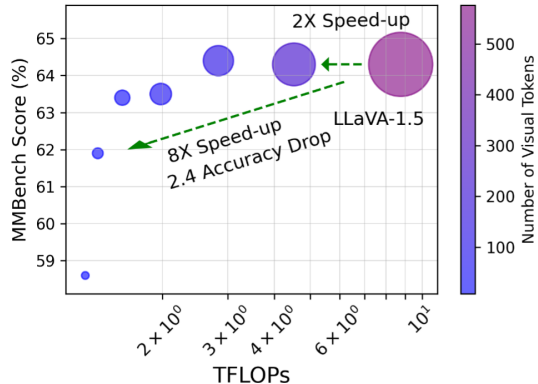


Figure 2: Our model employs a query transformer to encode images as visual tokens. We randomly select the first m tokens during training, and enable flexible choice of *any* m number under M during inference, where M is the maximum number of initialized tokens.

Hu et al. MQT-LLaVA: Matryoshka Query Transformer for Large Vision-Language Models. NeurIPS 2024

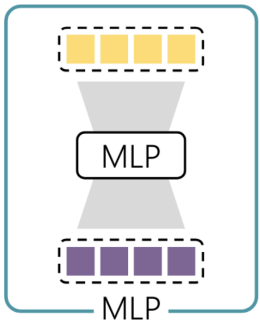
Context compression: MQT

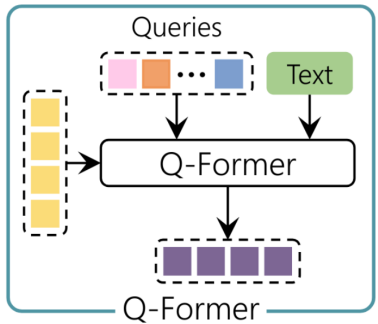


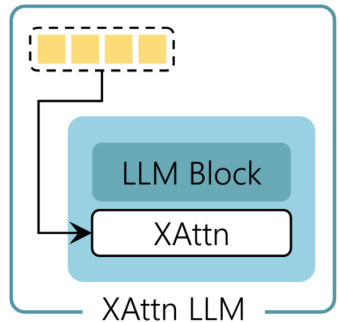


Hu et al. MQT-LLaVA: Matryoshka Query Transformer for Large Vision-Language Models. NeurIPS 2024

Connector


MLP
MLP

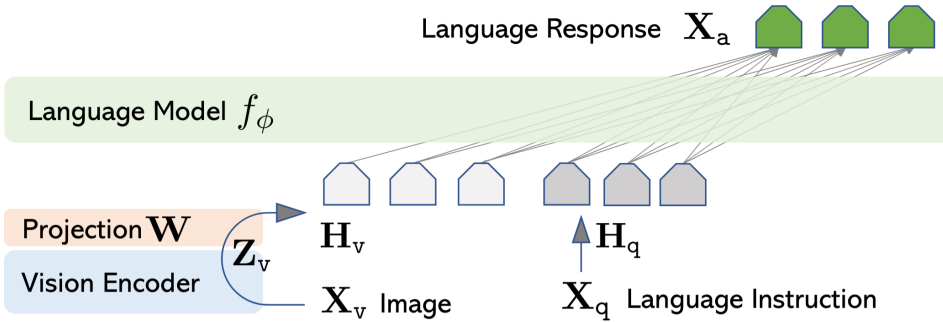

Queries
Text
Q-Former
Q-Former


LLM Block
XAttn
XAttn LLM

Outline

1. MLLMs as step towards AGI
2. Static benchmarks
3. Arena
4. Architectures
5. Overview of several popular models

LLaVA



- LLM — Vicuna-7B
- Vision Encoder — CLIP ViT-L/14
- Connector — Linear
- Train in 2 steps: 1) Projection 2) Projection & LLM

Liu et al. Visual Instruction Tuning. NeurIPS 2023

LLaVA

Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

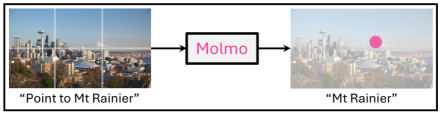
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

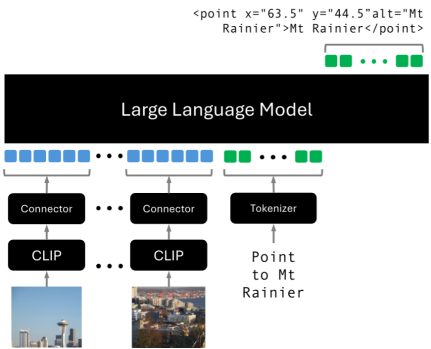
Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

Table 1: One example to illustrate the instruction-following data. The top block shows the contexts such as captions and boxes used to prompt GPT, and the bottom block shows the three types of responses. Note that the visual image is not used to prompt GPT, we only show it here as a reference.

Liu et al. Visual Instruction Tuning. NeurIPS 2023

Molmo





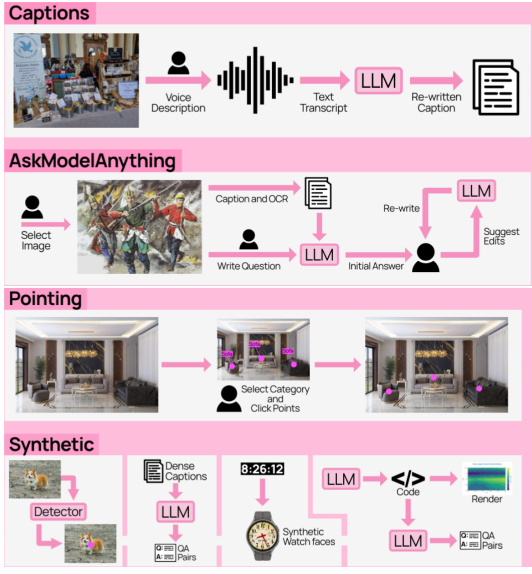
CLIP ViT-L/14 336px, high-res is processed using overlapped slicing Various LLMs

Training the whole model, no freezing:

1. Pretrain on PixMo
2. Finetune on PixMo and academic datasets

Deitke et al. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Multimodal Models. arXiv:2409.17146

PixMo (Pixels for Molmo)



PixMo (Pixels for Molmo)

1. **PixMo-Cap for pretraining:**
   3 labellers speak for 60 seconds → transcribe → improve with LLM → summarize with LLM; 712k images, 1.3M captions
2. **PixMo-AskModelAnything:**
   labellers use language-only LLMs to semi-automatically generate question; 73k images, 162k question-answer pairs
3. **PixMo-Points:**
   428k images, 2.3M question-point pairs
   Augment prev dataset with points, 29k images and 79k question-answer pairs
4. **PixMo-CapQA, PixMo-Docs, PixMo-Clocks:** generated using an LLM

Molmo openness

		VLM		LLM Backbone		Vision Encoder	
Category	Model	Open Weights	Open Data + Code	Open Weights	Open Data + Code	Open Weights	Open Data + Code
Molmo	Molmo-72B	Open	Open	Open	Closed	Open	Closed
	Molmo-7B-D	Open	Open	Open	Closed	Open	Closed
	Molmo-7B-O	Open	Open	Open	Open	Open	Closed
	MolmoE-1B	Open	Open	Open	Open	Open	Closed
API Models	GPT-4o	Closed	Closed	Closed	Closed	Closed	Closed
	GPT-4V	Closed	Closed	Closed	Closed	Closed	Closed
	Gemini 1.5 Pro	Closed	Closed	Closed	Closed	Closed	Closed
	Gemini 1.5 Flash	Closed	Closed	Closed	Closed	Closed	Closed
	Claude 3.5 Sonnet	Closed	Closed	Closed	Closed	Closed	Closed
	Claude 3 Opus	Closed	Closed	Closed	Closed	Closed	Closed
	Claude 3 Haiku	Closed	Closed	Closed	Closed	Closed	Closed
Open Weights	Qwen VL2 72B	Open	Closed	Open	Closed	Open	Closed
	Qwen VL2 7B	Open	Closed	Open	Closed	Open	Closed
	Intern VL2 LLAMA 76B	Open	Closed	Open	Closed	Open	Closed
	Intern VL2 8B	Open	Closed	Open	Closed	Open	Closed
	Pixtral 12B	Open	Closed	Open	Closed	Open	Closed
	Phi3.5-Vision 4B	Open	Closed	Open	Closed	Open	Closed
	PaliGemma 3B	Open	Closed	Open	Closed	Open	Closed
Open Weights & Data	LLAVA OneVision 72B	Open	Distilled	Open	Closed	Open	Closed
	LLAVA OneVision 7B	Open	Distilled	Open	Closed	Open	Closed
	Cambrian-1 34B	Open	Distilled	Open	Closed	Open	Closed
	Cambrian-1 8B	Open	Distilled	Open	Closed	Open	Closed
	xGen - MM - Interleave 4B	Open	Distilled	Open	Closed	Open	Closed
	LLAVA-1.5 13B	Open	Open	Open	Closed	Open	Closed
	LLAVA-1.5 7B	Open	Open	Open	Closed	Open	Closed

Molmo evaluation

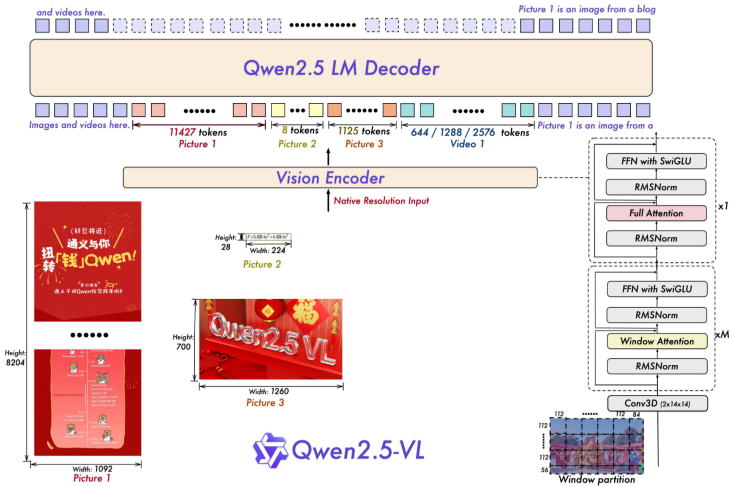
model	AI2D test [49]	ChartQA test [82]	VQA v2.0 testdev [36]	DocVQA test [83]	InfoQA test [84]	TextVQA val [100]	RealWorldQA [116]	MMMU val [129]	MathVista testmini [78]	CountBenchQA [10]	PixMo-Count test	Average	Elo score	Elo rank
API call only														
GPT-4V [88]	89.4	78.1	77.2	87.2	75.1	78.0	61.4	63.1	58.1	69.9	45.0	71.1	1041	10
GPT-4o-0513 [90]	94.2	85.7	78.7	92.8	79.2	77.4	75.4	**69.1**	63.8	87.9	59.6	78.5	**1079**	**1**
Gemini 1.5 Flash [103]	91.7	85.4	80.1	89.9	75.3	78.7	67.5	56.1	58.4	81.6	61.1	75.1	1054	7
Gemini 1.5 Pro [103]	94.4	87.2	80.2	93.1	81.0	78.7	70.4	62.2	63.9	85.8	64.3	78.3	1074	3
Claude-3 Haiku [7]	86.7	81.7	68.4	88.8	56.1	67.3	45.5	50.2	46.4	83.0	43.9	65.3	999	18
Claude-3 Opus [7]	88.1	80.8	66.3	89.3	55.6	67.5	49.8	59.4	50.5	83.6	43.3	66.7	971	21
Claude-3.5 Sonnet [7]	94.7	**90.8**	70.7	95.2	74.3	74.1	60.1	68.3	**67.7**	89.7	58.3	76.7	1069	4
Open weights only														
PaliGemma-mix-3B [10]	72.3	33.7	76.3	31.3	21.4	56.0	55.2	34.9	28.7	80.6	60.0	50.0	937	27
Phi3.5-Vision-4B [1]	78.1	81.8	75.7	69.3	36.6	72.0	53.6	43.0	43.9	64.6	38.3	59.7	982	19
Qwen2-VL-7B [111]	83.0	83.0	82.9	94.5	76.5	84.3	70.1	54.1	58.2	76.5	48.0	73.7	1025	14
Qwen2-VL-72B [111]	88.1	88.3	81.9	**96.5**	84.5	**85.5**	**77.8**	64.5	70.5	80.4	55.7	79.4	1037	12
InternVL2-8B [104]	83.8	83.3	76.7	91.6	74.8	77.4	64.2	51.2	58.3	57.8	43.9	69.4	953	23
InternVL2-Llama-3-76B [104]	87.6	88.4	85.6	94.1	**82.0**	84.4	72.7	58.2	65.5	74.7	54.6	77.1	1018	16
Pixtral-12B [3]	79.0	81.8	80.2	90.7	50.8	75.7	65.4	52.5	58.0	78.8	51.7	69.5	1016	17
Llama-3.2V-11B-Instruct [5]	91.1	83.4	75.2	88.4	63.6	79.7	64.1	50.7	51.5	73.1	47.4	69.8	1040	11
Llama-3.2V-90B-Instruct [5]	92.3	85.5	78.1	90.1	67.2	82.3	69.8	60.3	57.3	78.5	58.5	74.5	1063	5
Open weights + data († distilled)														
LLaVA-1.5-7B [69]	55.5	17.8	78.5	28.1	25.8	58.2	54.8	35.7	25.6	40.1	27.6	40.7	951	26
LLaVA-1.5-13B [69]	61.1	18.2	80.0	30.3	29.4	61.3	55.3	37.0	27.7	47.1	35.2	43.9	960	22
xGen-MM-interleave-4B† [119]	74.2	60.0	81.5	61.4	31.5	71.0	61.2	41.1	40.5	81.9	50.2	59.5	979	20
Cambrian-1-8B† [106]	73.0	73.3	81.2	77.8	41.6	71.7	64.2	42.7	49.0	76.4	46.6	63.4	952	25
Cambrian-1-34B† [106]	79.7	75.6	83.8	75.5	46.0	76.7	67.8	49.7	53.2	75.6	50.7	66.8	953	24
LLaVA OneVision-7B† [59]	81.4	80.0	84.0	87.5	68.8	78.3	66.3	48.8	63.2	78.8	54.4	72.0	1024	15
LLaVA OneVision-72B† [59]	85.6	83.7	85.2	91.3	74.9	80.5	71.9	56.8	67.5	84.3	60.7	76.6	1051	8
The Molmo family: Open weights, Open data, Open training code, Open evaluations														
MolmoE-1B	86.4	78.0	83.9	77.7	53.9	78.8	60.4	34.9	34.0	87.2	79.6	68.6	1032	13
Molmo-7B-O	90.7	80.4	85.3	90.8	70.0	80.4	67.5	39.3	44.5	89.0	83.3	74.6	1051	9
Molmo-7B-D	93.2	84.1	85.6	92.2	72.6	81.7	70.7	45.3	51.6	88.5	84.8	77.3	1056	6
Molmo-72B	**96.3**	87.3	**86.5**	93.5	81.9	83.1	75.2	54.1	58.6	**91.2**	**85.2**	**81.2**	1077	2

Molmo discussion

Benchmarks and human evaluation agree with exception of Qwen2-VL.
Key results:

1. MolmoE 1B nearly matches GPT-4V
2. Molmo 7B-D and Molmo 7B-O are between GPT-4V and GPT-4o
3. Molmo 72B is near to GPT-4o
4. 72B model outperforms Gemini 1.5 Pro and Claude 3.5 Sonnet

QWen2.5-VL



- similar arch, reworked ViT
- closed data, incl. PixMo
- 2-step training of full model

Bai et al. Qwen2.5-VL Technical Report. arXiv:2502. I3923

LISA

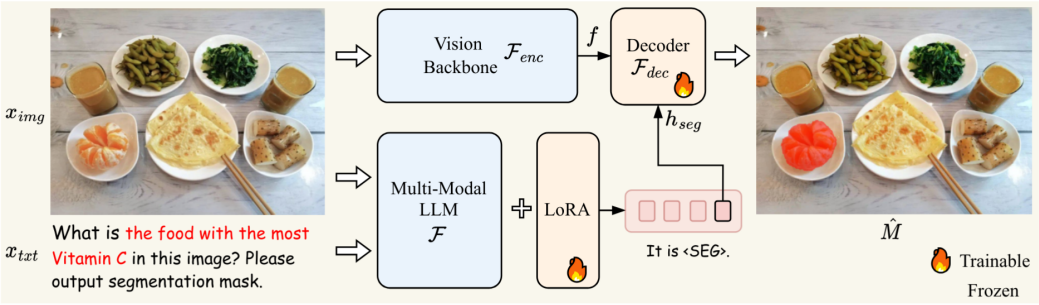


Figure 3. The pipeline of LISA. Given the input image and text query, the multimodal LLM (e.g., LLaVA [29]) generates text output. The last-layer embedding for the <SEG> token is then decoded into the segmentation mask via the decoder. We use LoRA [15] for efficient fine-tuning. The choice of vision backbone can be flexible (e.g., SAM [66], Mask2Former [9]).

Lai et al. LISA: Reasoning Segmentation via Large Language Model. CVPR 2024

OMG-LLaVA

Zhang et al. OMG-LLaVA: Bridging Image-level, Object-level, Pixel-level Reasoning and Understanding. NeurIPS 2024

OMG-LLaVA

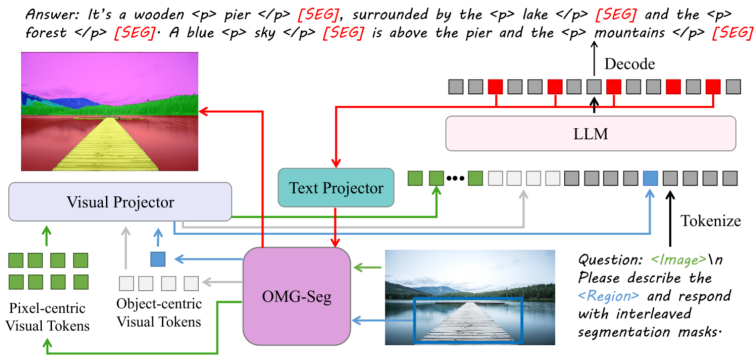


Figure 3: The Overview of OMG-LLaVA. OMG-LLaVA consists of OMG-Seg and LLM. OMG-Seg tokenizes the image into pixel-centric visual tokens, the detected objects, and inputs visual prompts into object-centric visual tokens. Additionally, the [SEG] token output by LLM is decoded by OMG-Seg into segmentation masks. OMG-Seg remains frozen at all stages.

Zhang et al. OMG-LLaVA: Bridging Image-level, Object-level, Pixel-level Reasoning and Understanding. NeurIPS 2024

Conclusion

We reviewed following topics:

- **Intro to image MLLMs and their applications.** Multimodal LLMs seems to be a step towards AGI with lots of interesting applications and challenges.
- **Benchmarking MLLMs.** It may be done using static benchmarks (as in CV or NLP) or using Arenas. Full-scale evaluation is very challenging, since models aim to solve a lot of useful tasks.
- **General architecture of MLLMs.** Typically models consist of vision encoder(s), connector, LLM and optional output modality decoder. There are a number of technical nuances that help to obtain best quality