Lecture 9: Vision-Language Generative Models

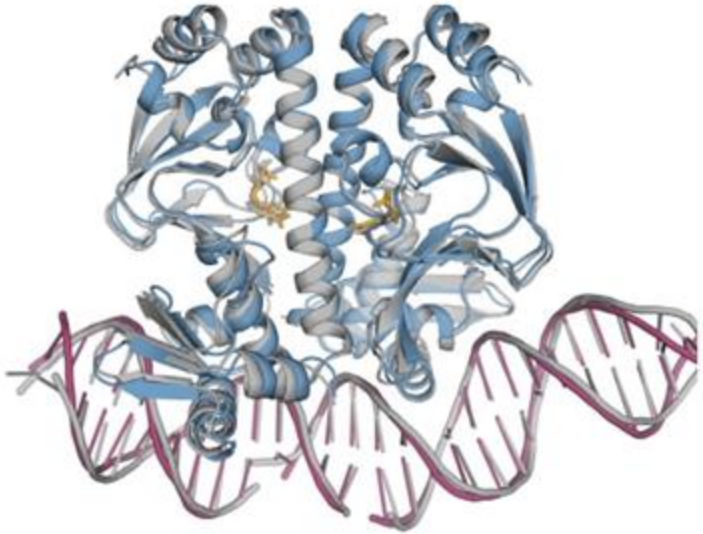
Announcements

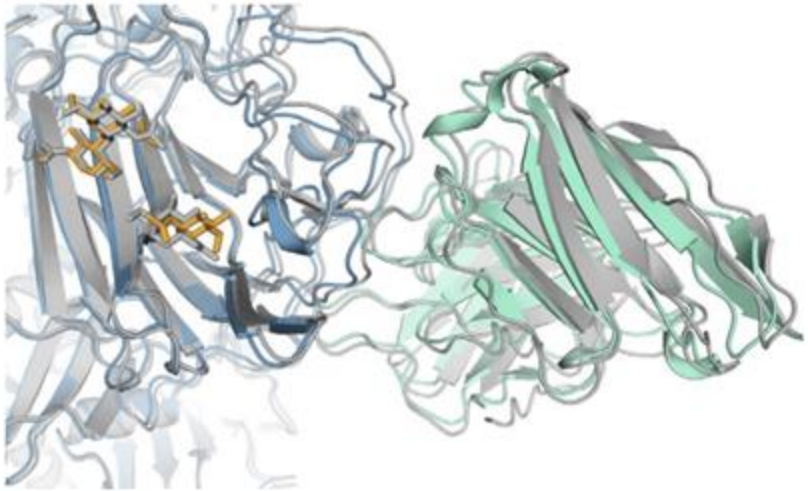
- Project proposal is due Wed Oct 23
- Discussion presentation and question-asking assignments posted on Ed - check for your assignment and fill in your email address in the spreadsheet (see instructions)
- First discussion section is Mon Nov 4
- A2 (second paper analysis) will be released Oct 30, and due Nov 14

Finishing up from last lecture: Vision Diffusion and Generative Models in Biomedicine

AlphaFold 3

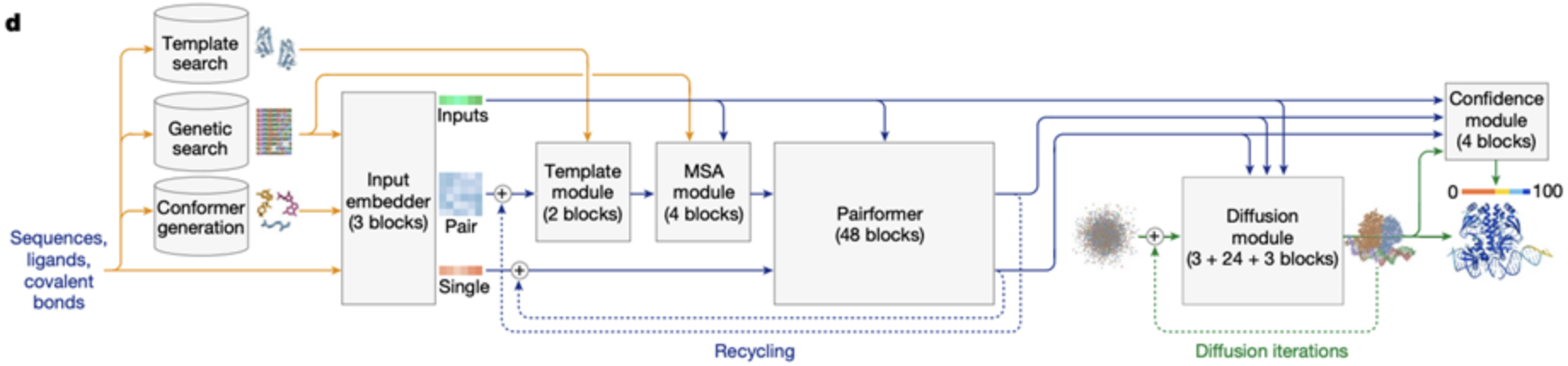
- Diffusion-based architecture that predicts the joint 3D structure of complexes including proteins, nucleic acids, small molecules, etc. from sequences





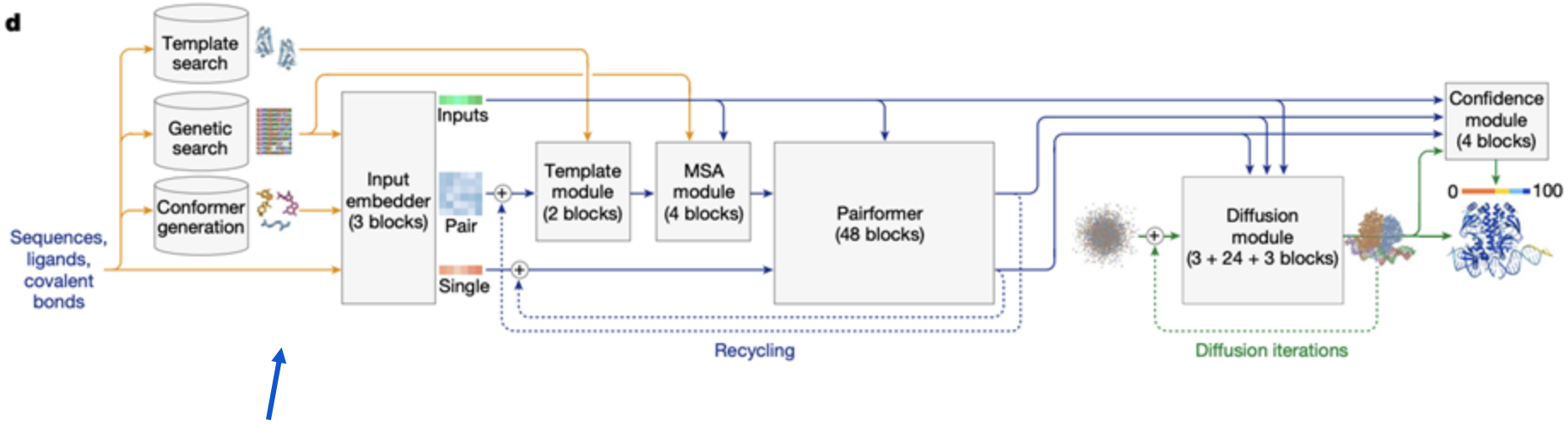
Abramson et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024.

AlphaFold 3: a bird's eye view



Abramson et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024.

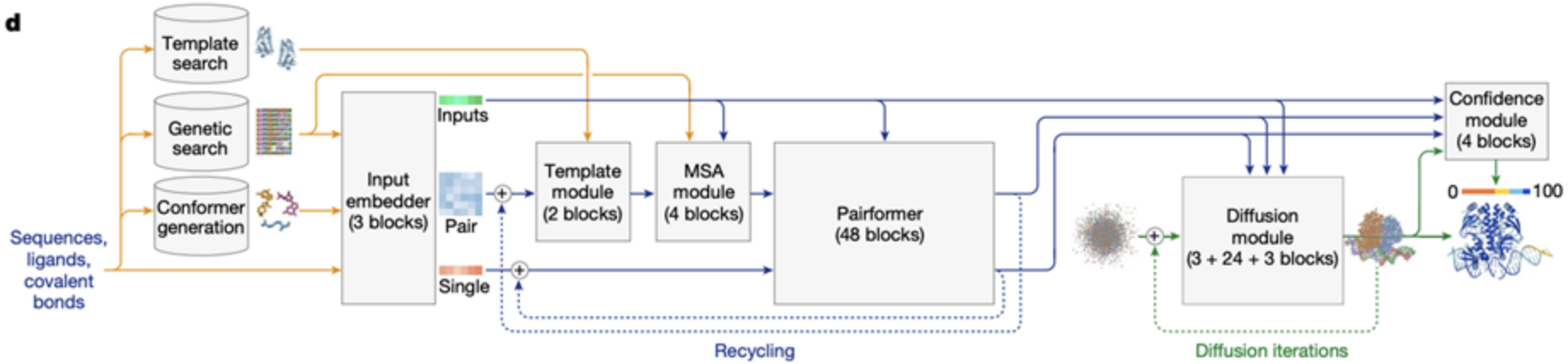
AlphaFold 3: a bird's eye view



Process and initially embed inputs

Abramson et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024.

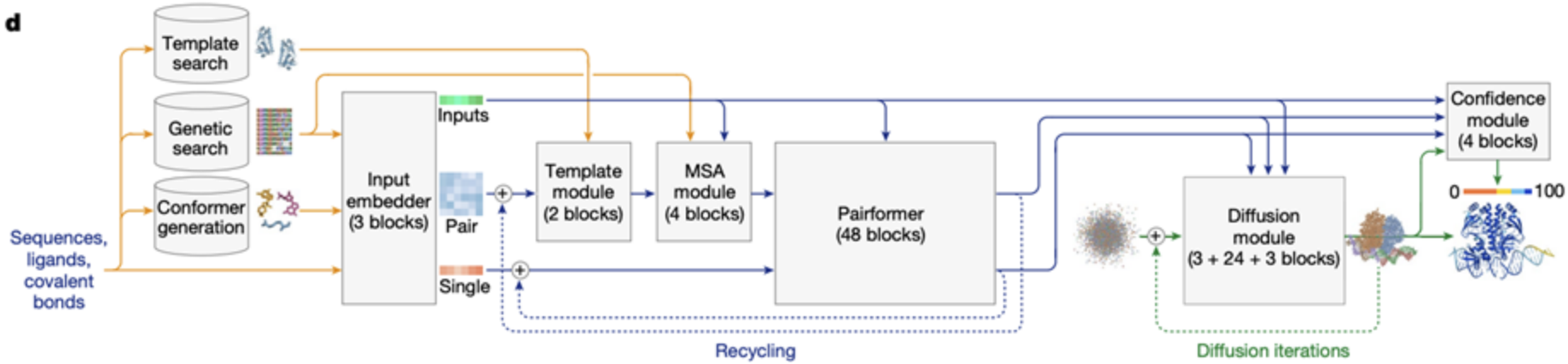
AlphaFold 3: a bird's eye view



Transform inputs into powerful representations to condition diffusion model

Abramson et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024.

AlphaFold 3: a bird's eye view

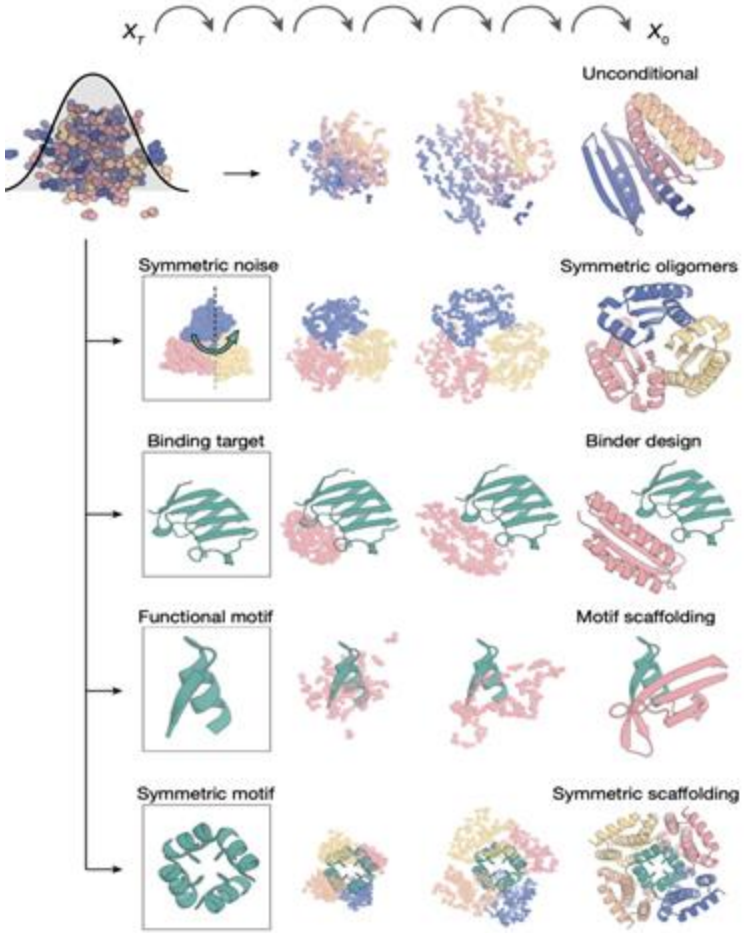


Starting from random atom coordinates (noise), use conditional diffusion model to generate refined coordinate values (i.e. 3D structure)

Abramson et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024.

RFdiffusion: designing protein structures based on functional constraints (conditions)

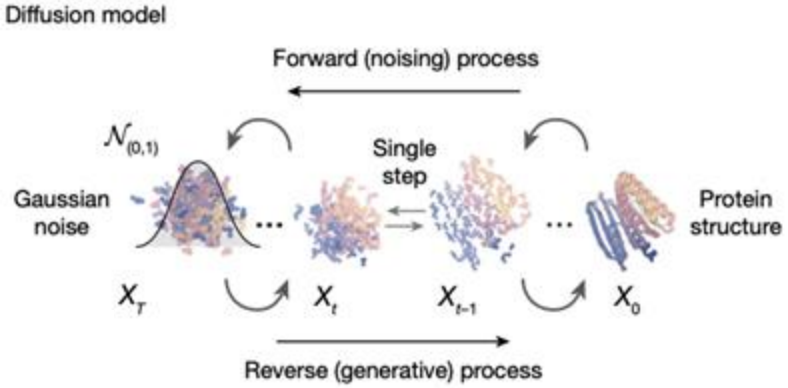
Generated output is still 3D protein structure (similar to AlphaFold 2, not the more diverse outputs of AlphaFold 3), but conditioning is now based on embedding functional conditions such as a desired binding target, functional motif, etc.



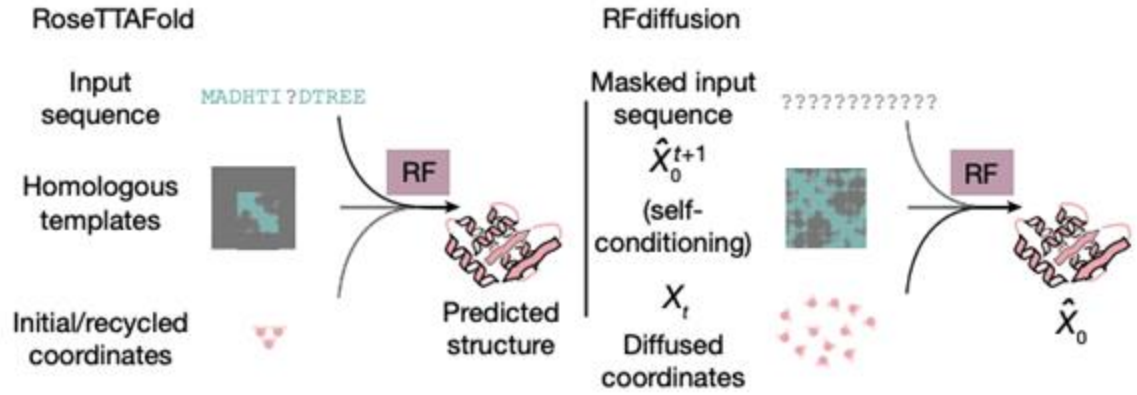
Watson et al. De novo design of protein structure and function with RFdiffusion. Nature 2023.

RFdiffusion: designing protein structures based on functional constraints (conditions)

Core diffusion model looks familiar: generating a valid 3D protein structure from a starting noise structure



Compared to RosaTTAFold (the authors' previous work performing sequence-based prediction similar to AlphaFold), RFdiffusion no longer uses an input sequence as input



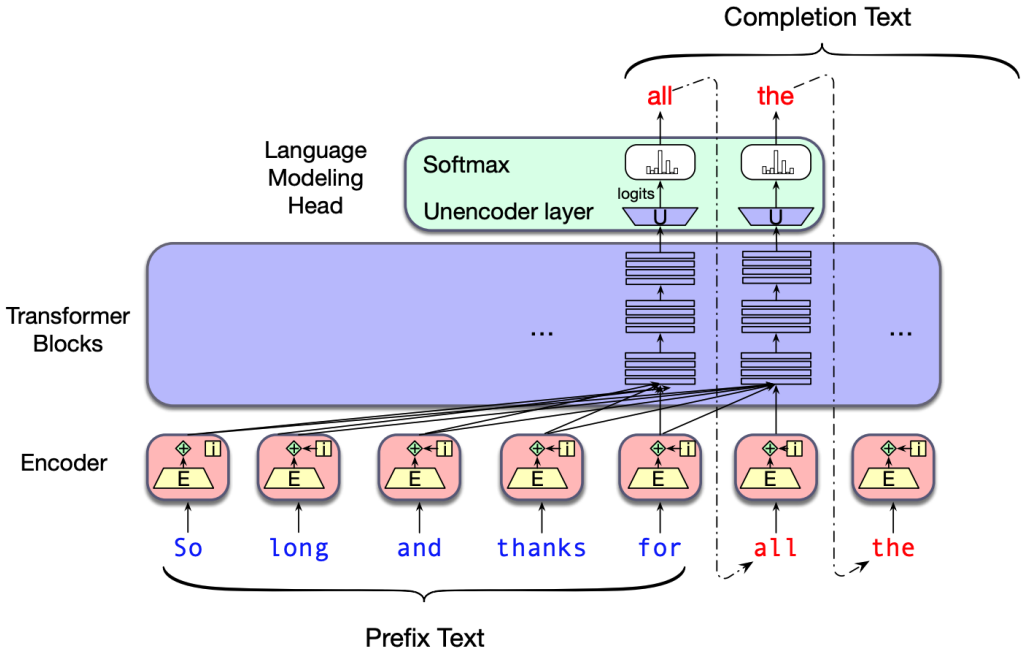
Watson et al. De novo design of protein structure and function with RFdiffusion. Nature 2023.

Next:
Vision-Language Generative Models

Preliminary

Language Models

Autoregressive Language Models



LARGE LANGUAGE MODELS WITH TRANSFORMERS (Daniel Jurafsky & James H. Martin 2024)

Autoregressive Language Models

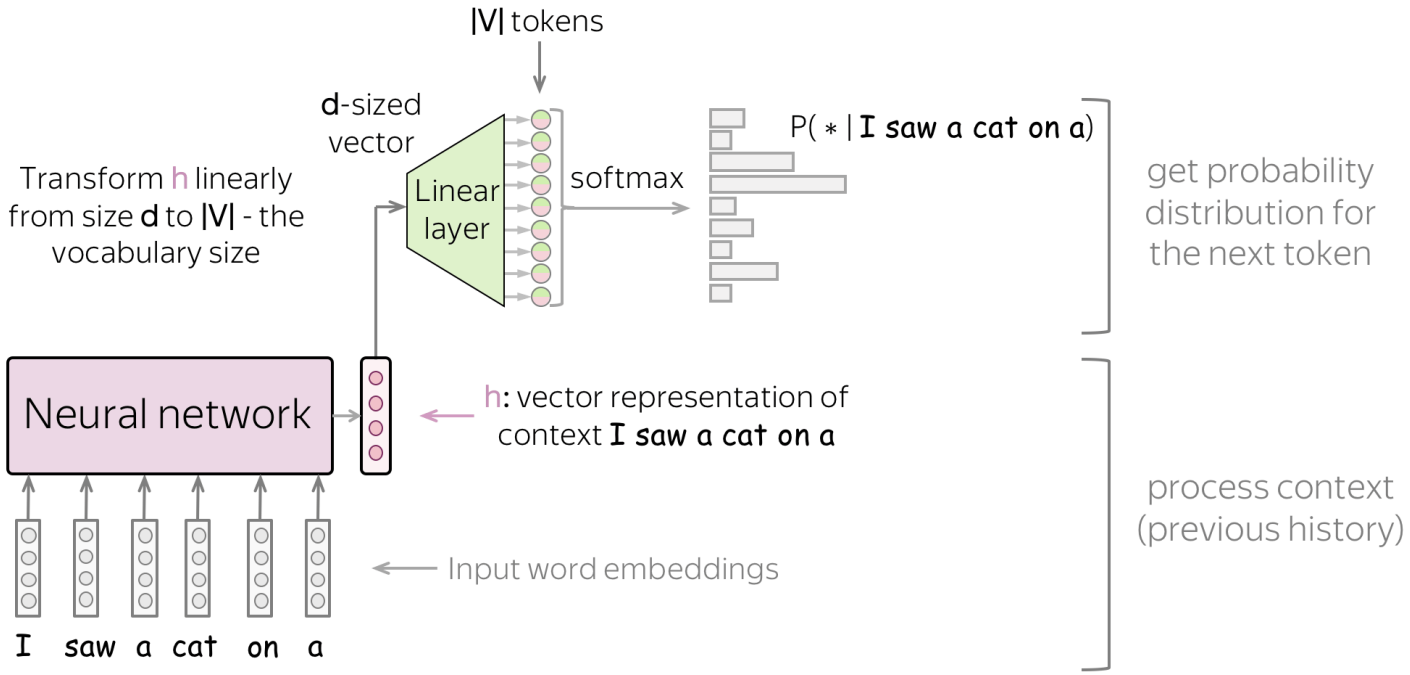
Joint distribution with chain rule of probability:

$$p(x_{1:L}) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \cdots p(x_L \mid x_{1:L-1}) = \prod_{i=1}^{L} p(x_i \mid x_{1:i-1}).$$

$$\begin{aligned} p(\text{the, mouse, ate, the, cheese}) = & \; p(\text{the}) \\ & p(\text{mouse} \mid \text{the}) \\ & p(\text{ate} \mid \text{the, mouse}) \\ & p(\text{the} \mid \text{the, mouse, ate}) \\ & p(\text{cheese} \mid \text{the, mouse, ate, the}). \end{aligned}$$

Credit: CS326

Autoregressive Language Models



Credit: https://lena-voita.github.io/nlp_course/language_modeling.html#intro

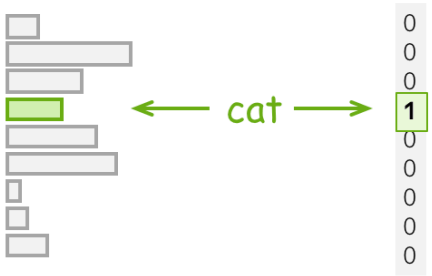
Autoregressive Language Models

we want the model to predict this

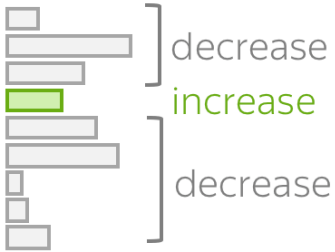
Training example: I saw a cat on a mat <eos>

Model prediction: p(∗ | I saw a)

Target

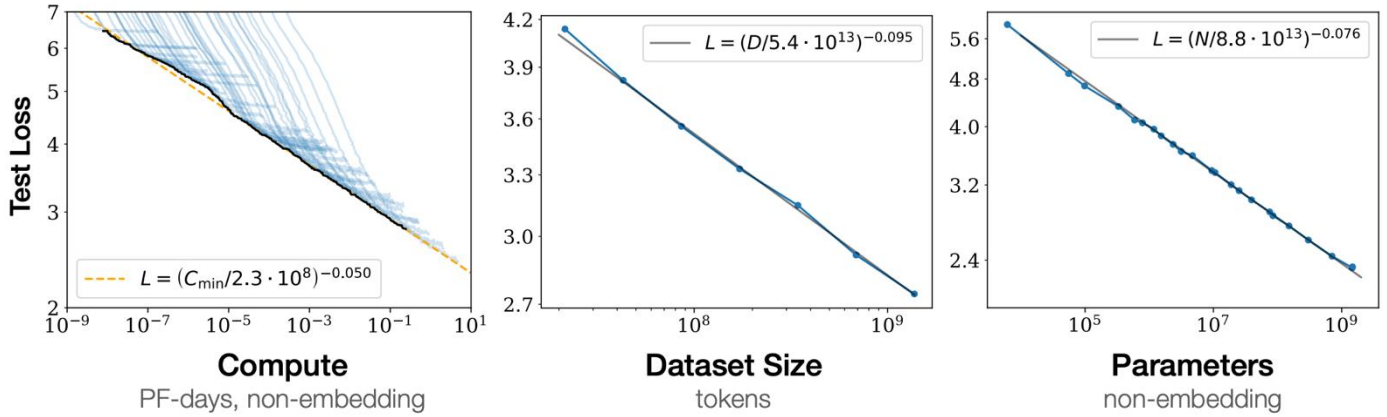


Loss = -log (p(cat)) → min



Credit: https://lena-voita.github.io/nlp_course/language_modeling.html#intro

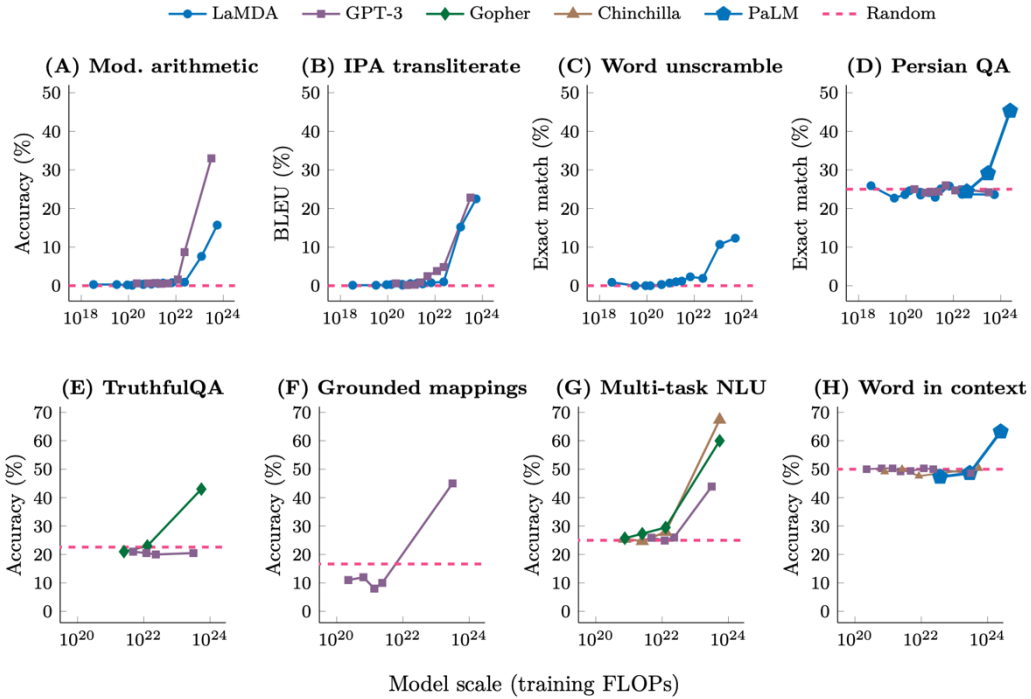
Why “Large” Language Models



Empirically: more data and larger models => better performance

Scaling Laws (Kaplan et al. 2020)

Why “Large” Language Models



Emergent Abilities (Wei et al. 2020)

Scaling up LMs produces new emergent behavior, leading to qualitatively different capabilities and qualitatively different societal impact.

Why “Large” Language Models

How to Link Vision to LLMs?

How to Link Vision to LLMs?

Internal Linkage → Vision-Language Models

External Linkage → Vision-Language Agents

How to Link Vision to LLMs?

Internal Linkage → Vision-Language Models

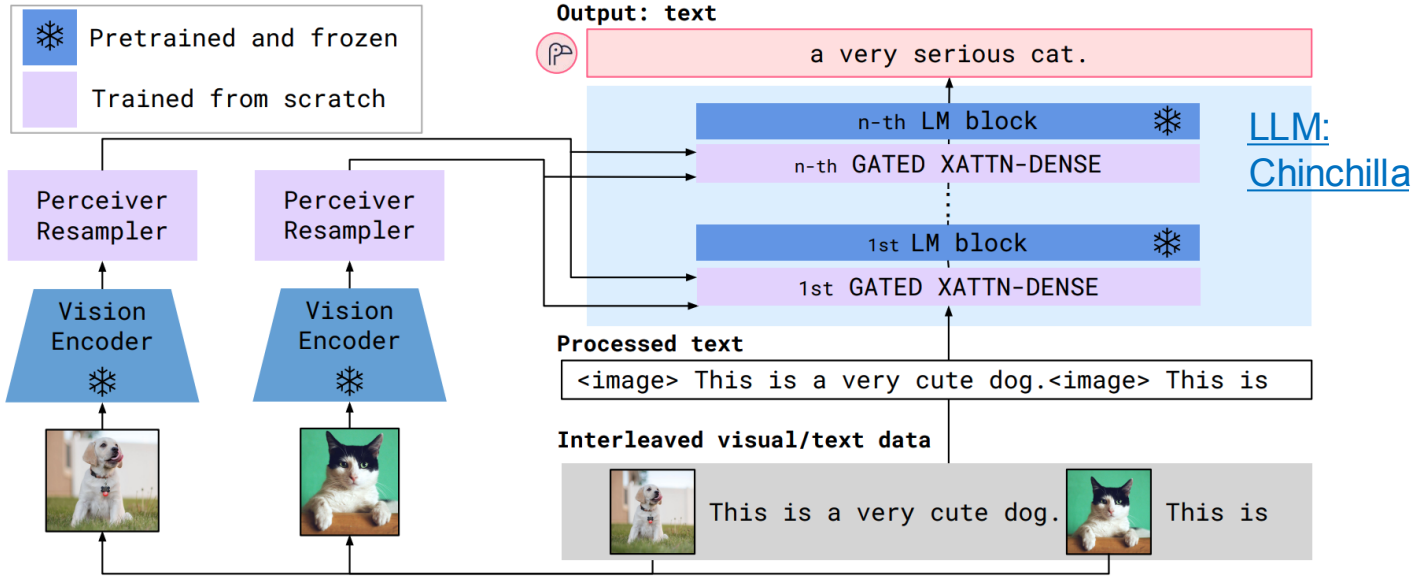
- Architecture

External Linkage → Vision-Language Agents

Integrate Visual Features into Intermedia Layers

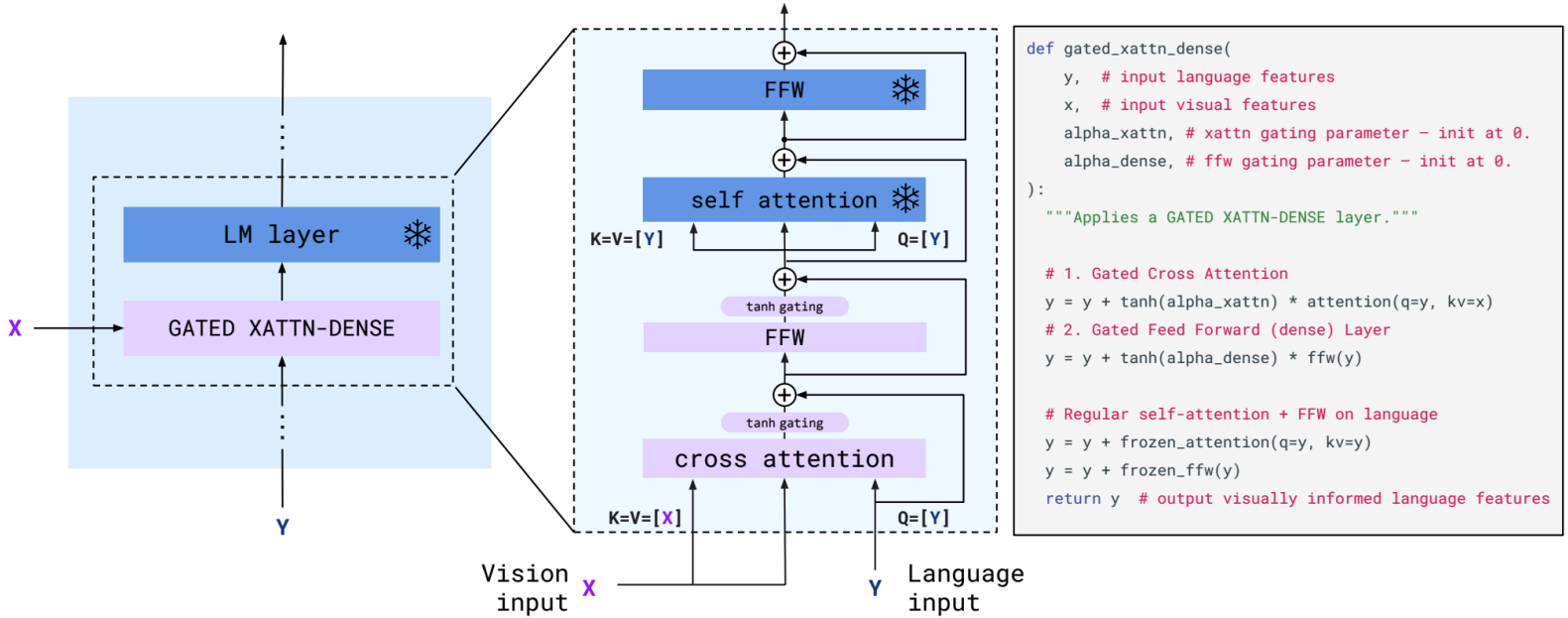
Perceiver Resampler: from varying-size large feature maps to few visual tokens

Vision Encoder: from pixels to features, CLIP-style Models



Flamingo (Alayrac et al. 2022)

Integrate Visual Features into Intermedia Layers



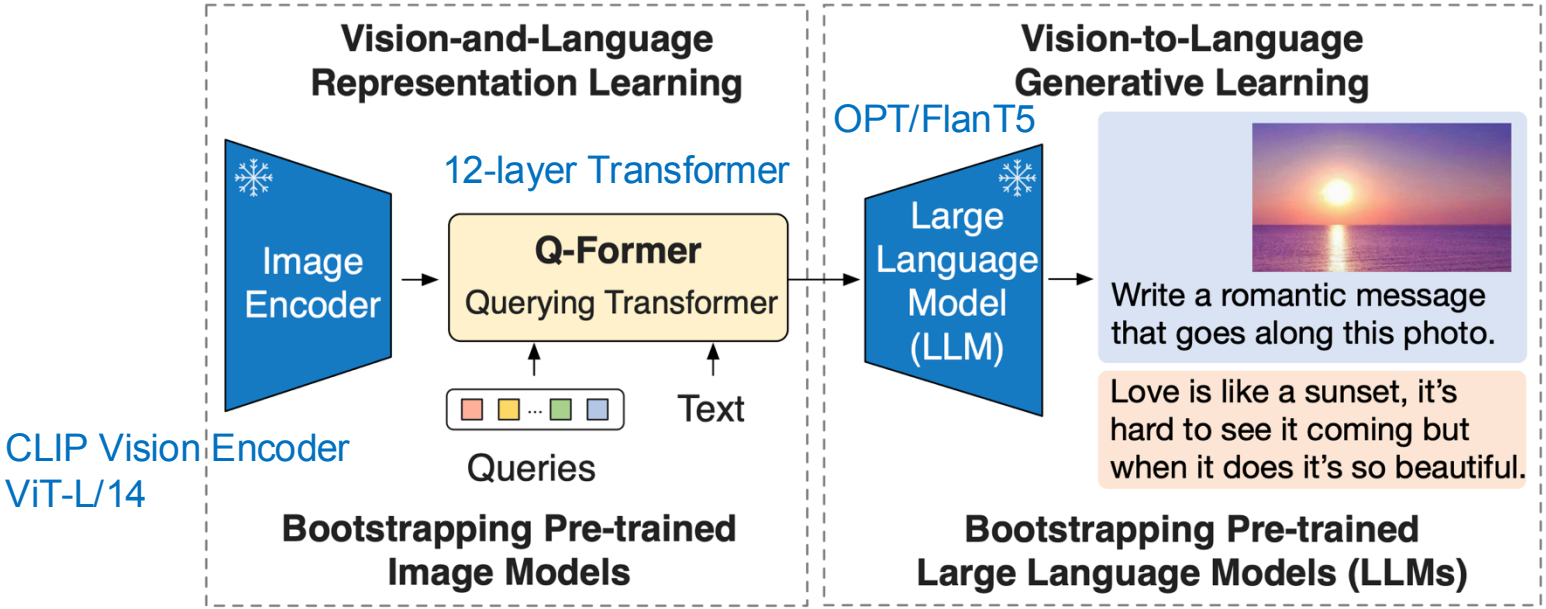

```
def gated_xattn_dense(
    y,  # input language features
    x,  # input visual features
    alpha_xattn, # xattn gating parameter – init at 0.
    alpha_dense, # ffw gating parameter – init at 0.
):
  """Applies a GATED XATTN-DENSE layer."""

  # 1. Gated Cross Attention
  y = y + tanh(alpha_xattn) * attention(q=y, kv=x)
  # 2. Gated Feed Forward (dense) Layer
  y = y + tanh(alpha_dense) * ffw(y)

  # Regular self-attention + FFW on language
  y = y + frozen_attention(q=y, kv=y)
  y = y + frozen_ffw(y)
  return y  # output visually informed language features
```

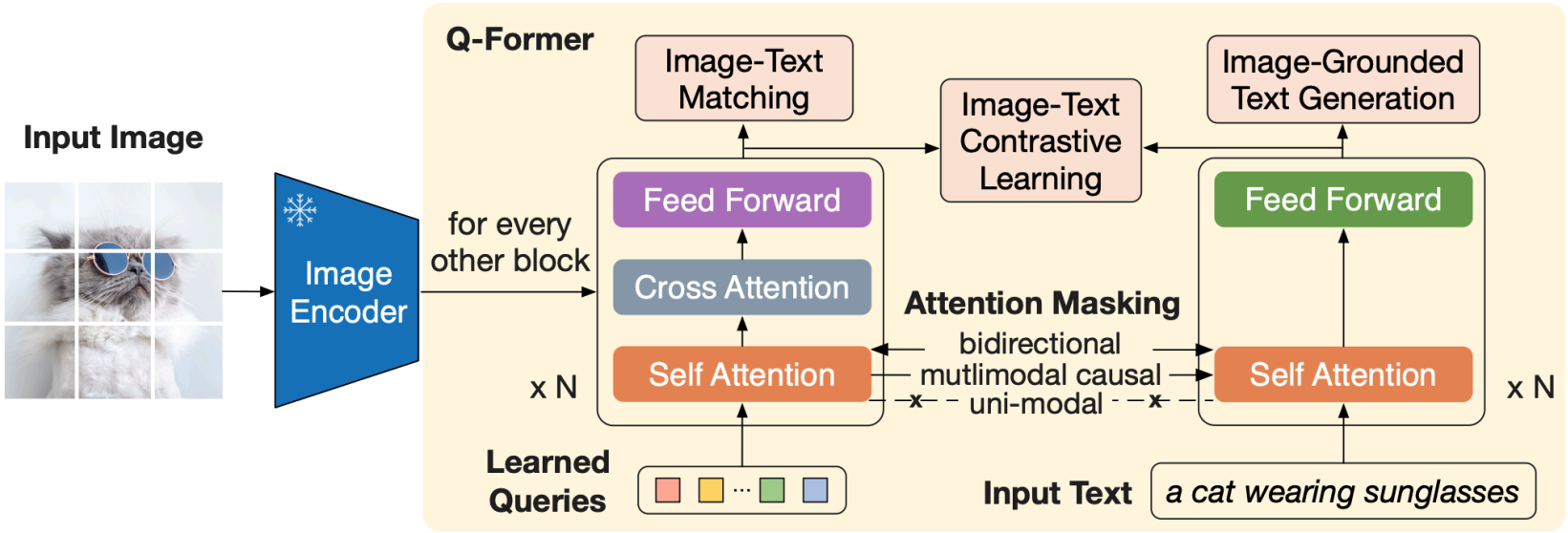
Flamingo (Alayrac et al. 2022)

Integrate Visual Features into Input Layer



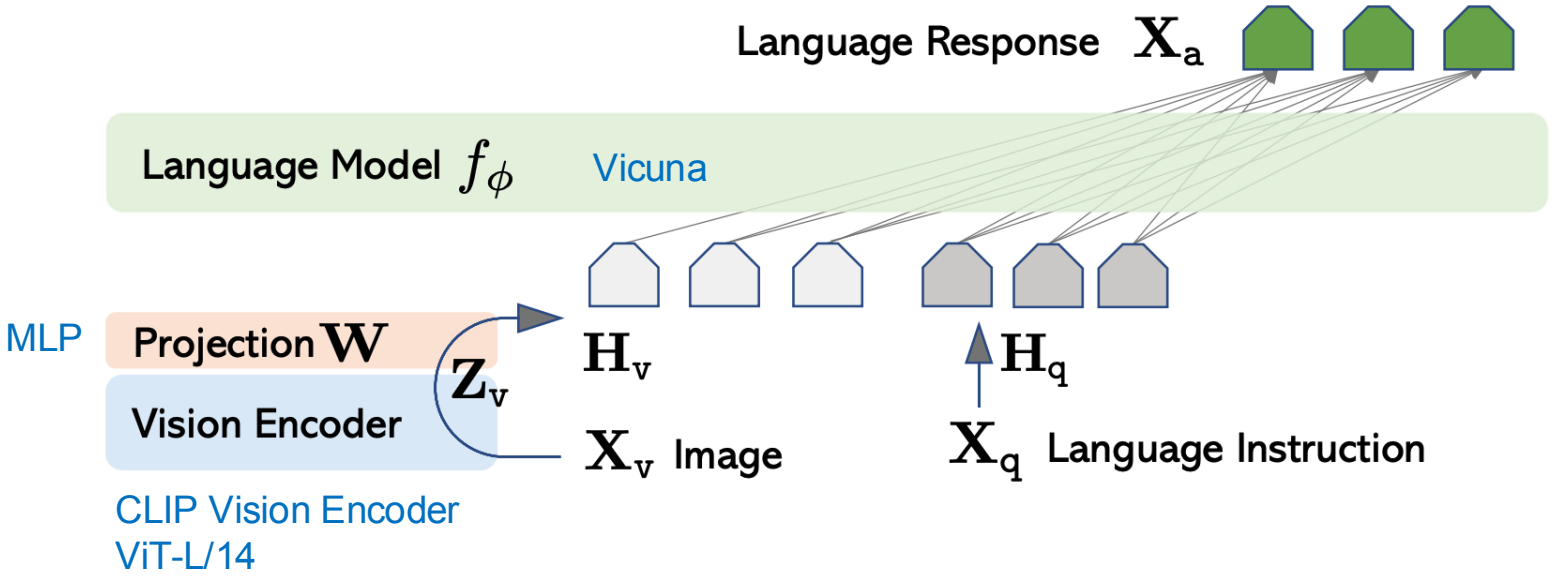
BLIP-2 (Li et al. 2023)

Integrate Visual Features into Input Layer



BLIP-2 (Li et al. 2023)

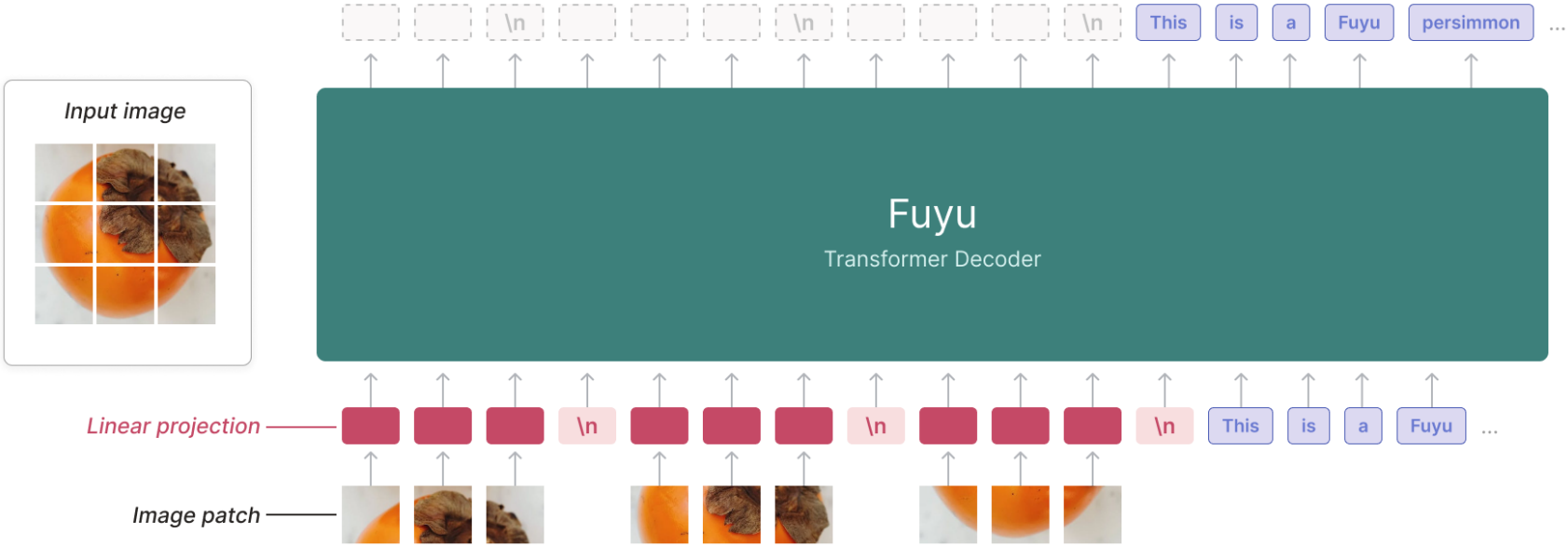
Integrate Visual Features into Input Layer



Stage 1: Pre-training for Feature Alignment
Stage 2: Fine-tuning End-to-End (W and $\emptyset$)

LLaVA (Liu et al. 2023)

Integrate Visual Patches into Input Layer



Fuyu (Rohan et al. 2023)

Which One is the Best?

Architecture

- Integrate Visual Features into Intermedia Layers
- Integrate Visual Features into Input Layer
- Integrate Visual Patches into Input Layer

How to Link Vision to LLMs?

Internal Linkage → Vision-Language Models

- Architecture
- Evaluation

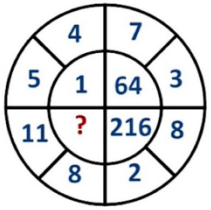
External Linkage → Vision-Language Agents

Evaluation Task: Visual Question Answering

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

Examples of QA from the GQA dataset

Evaluation Task: Visual Question Answering



Question: Find the missing value in this math puzzle.

Solution:

$(5 - 4)^3 = 1$

$(7 - 3)^3 = 64$

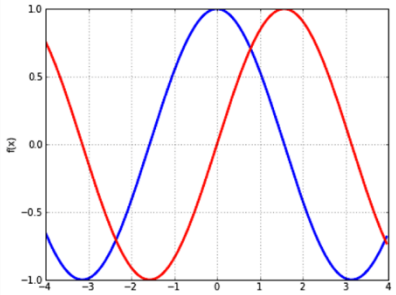
$(8 - 2)^3 = 216$

Similarly, $(11 - 8)^3 = 27$.

So the missing value is 27.

Answer: 27

Category: Math-targeted
Task: Figure question answering
Context: Puzzle test
Grade: Elementary school
Math: Logical reasoning



Question: Which function is monotonic in range [0, pi]?

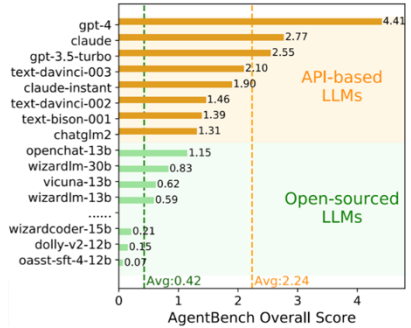
Choices:

(A) the red one (B) the blue one

(C) both (D) none of them

Answer: (B) the blue one

Category: Math-targeted
Task: Textbook question answering
Context: Function plot
Grade: College
Math: Algebraic reasoning



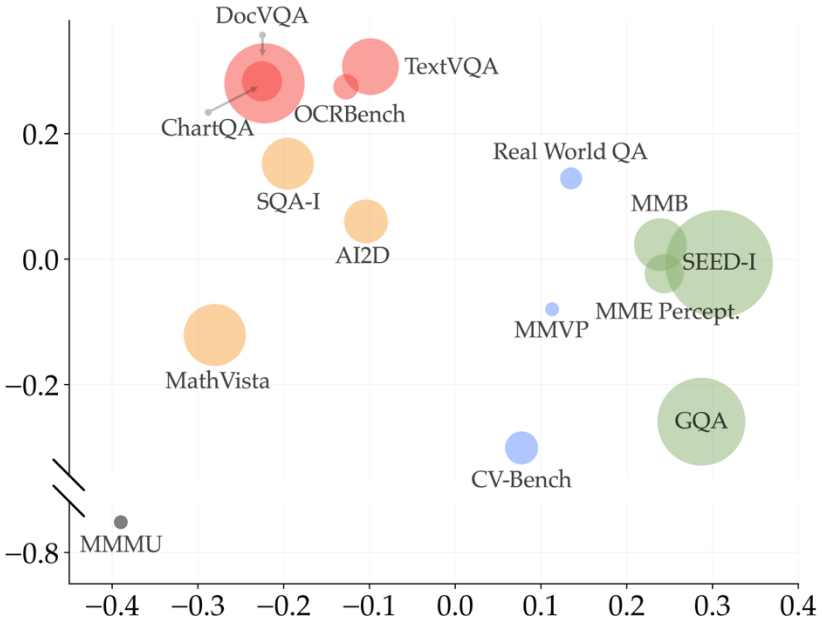
Question: What is the performance gap in the AgentBench Overall Score between the worst API-based LLM and the best open-sourced LLM?

Answer: 0.16

Category: Math-targeted
Task: Figure question answering
Context: Scientific figure
Grade: College
Math: Scientific reasoning

Examples of QA from MathVista

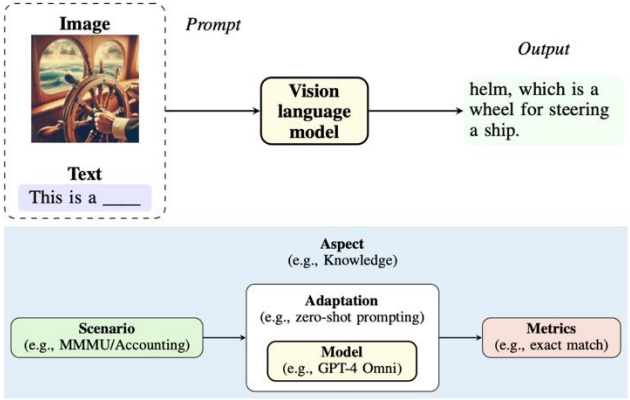
Clustering the Benchmarks



"General" in green, "Knowledge" in yellow, "Chart & OCR" in red, and "Vision-Centric" in blue

Cambrian-1 (Tong et al. 2024)

Holistic Evaluation

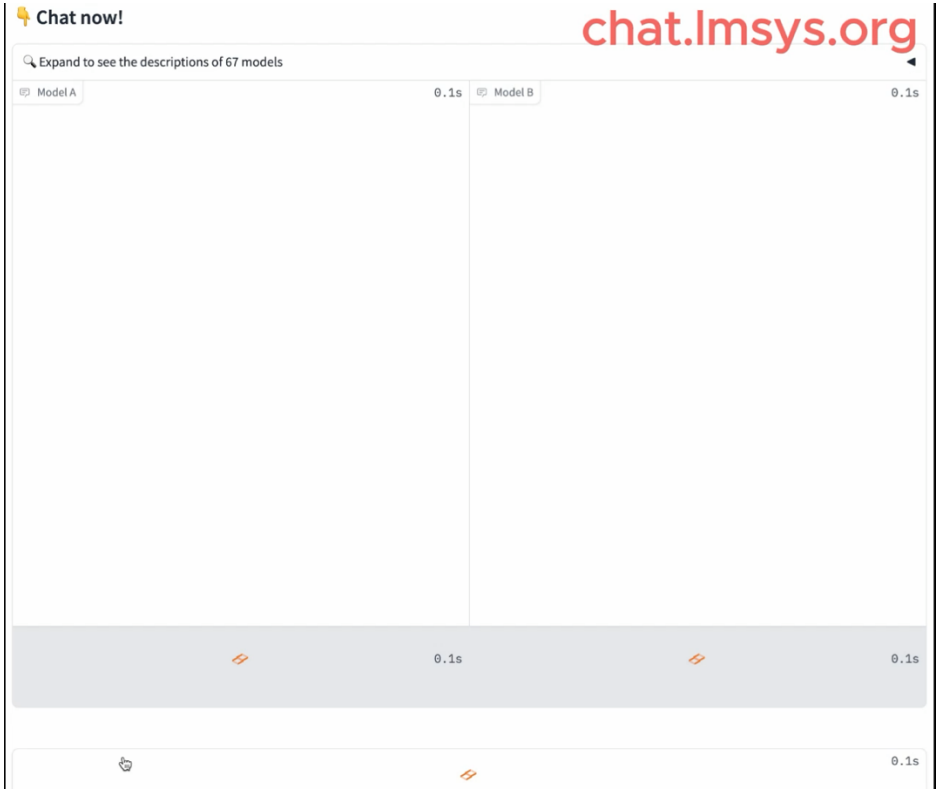


Model	Mean win rate
GPT-4o (2024-05-13)	**0.793**
GPT-4o (2024-08-06)	0.766
Gemini 1.5 Pro (0409 preview)	0.72
Claude 3.5 Sonnet (20240620)	0.7
GPT-4 Turbo (2024-04-09)	0.679
GPT-4o mini (2024-07-18)	0.654
Palmyra Vision 003	0.652
Gemini 1.5 Pro (0514 preview)	0.648
Gemini 1.5 Pro (001, BLOCK_NONE safety)	0.631
GPT-4V (1106 preview)	0.602

VHELM aggregates various datasets to cover one or more of the 9 aspects: visual perception, bias, fairness, knowledge, multilinguality, reasoning, robustness, safety, and toxicity.

VHELM (Lee et al. 2024)

Human Evaluation: Multimodal Chatbot Arena



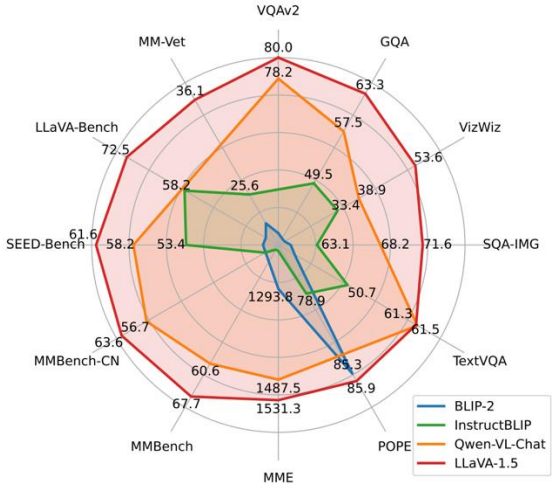
Multimodal Arena (Chou et al. 2024)

Which One is the Best?

Architecture

- ❖ Integrate Visual Features into Intermedia Layers
- ❖ Integrate Visual Features into Input Layer
- ❖ Integrate Visual Patches into Input Layer

Models	#Trainable Params	#Total Params	VQAv2 val	VQAv2 test-dev
VL-T5$_{no\text{-}vqa}$	224M	269M	13.5	-
FewVLM (Jin et al., 2022)	740M	785M	47.7	-
Frozen (Tsimpoukelli et al., 2021)	40M	7.1B	29.6	-
VLKD (Dai et al., 2022)	406M	832M	42.6	44.5
Flamingo3B (Alayrac et al., 2022)	1.4B	3.2B	-	49.2
Flamingo9B (Alayrac et al., 2022)	1.8B	9.3B	-	51.8
Flamingo80B (Alayrac et al., 2022)	10.2B	80B	-	56.3
BLIP-2 ViT-L OPT$_{2.7B}$	104M	3.1B	50.1	49.7
BLIP-2 ViT-g OPT$_{2.7B}$	107M	3.8B	53.5	52.3
BLIP-2 ViT-g OPT$_{6.7B}$	108M	7.8B	54.3	52.6
BLIP-2 ViT-L FlanT5$_{XL}$	103M	3.4B	62.6	62.3
BLIP-2 ViT-g FlanT5$_{XL}$	107M	4.1B	63.1	63.0
BLIP-2 ViT-g FlanT5$_{XXL}$	108M	12.1B	**65.2**	**65.0**



Eval Task	Fuyu-8B	Fuyu-Medium	LLaVA 1.5 (13.5B)	QWEN-VL (10B)
VQAv2	74.2	77.4	80	79.5
OKVQA	60.6	63.1	n/a	58.6
COCO Captions	141	138	n/a	n/a
AI2D	64.5	73.7	n/a	62.3

Which One is the Best?

Architecture

- Integrate Visual Features into Intermedia Layers
- Integrate Visual Features into Input Layer
- Integrate Visual Patches into Input Layer

Architectures of GPT-4, Gemini 1.5, and Claude 3?

- Details Unknown

Based on the latest open-sourced information

- the LLaVA-style VLM (Architecture + Training Recipe) is most effective. (2024/10)

How to Link Vision to LLMs?

Internal Linkage → Vision-Language Models

- Architecture
- Evaluation
- Training Recipe

External Linkage → Vision-Language Agents

Formulation

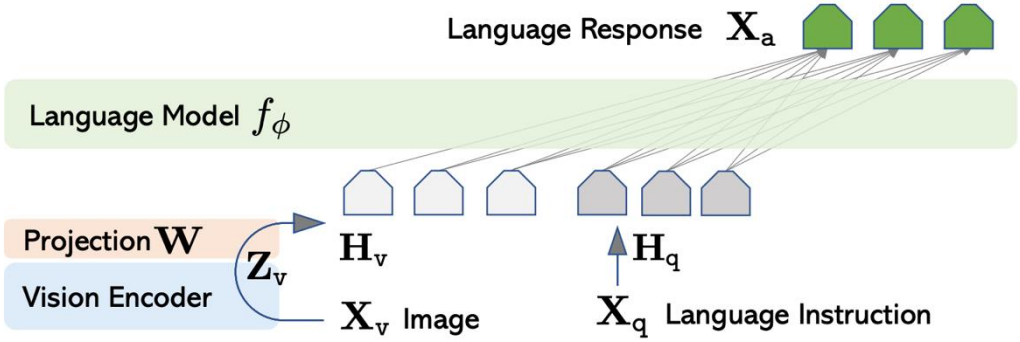
$$p(\mathbf{X}_\mathtt{a}|\mathbf{X}_\mathtt{v}, \mathbf{X}_\mathtt{instruct}) = \prod_{i=1}^{L} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\mathbf{X}_\mathtt{v}, \mathbf{X}_{\mathtt{instruct},<i}, \mathbf{X}_{\mathtt{a},<i})$$

$\mathbf{X}_{\texttt{system-message}}$ <STOP>
Human : $\mathbf{X}^1_{\texttt{instruct}}$ <STOP> Assistant: $\mathbf{X}^1_\mathtt{a}$ <STOP>
Human : $\mathbf{X}^2_{\texttt{instruct}}$ <STOP> Assistant: $\mathbf{X}^2_\mathtt{a}$ <STOP> $\cdots$

LLaVA (Liu et al. 2023)

Stage 1: Pre-training

Goal: Align visual features to LLM's word embedding space



Data: Converted from image captioning data

- 595K image-text pairs filtered from CC3M
- convert to instruction-following format

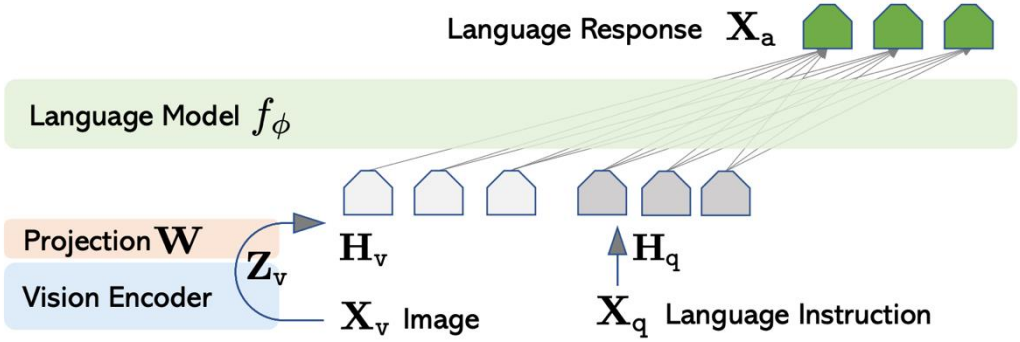
Trainable Parameters

- Only W

LLaVA (Liu et al. 2023)

Stage 2: Instruction Tuning (Supervised Finetuning)

Goal: Visual Captioner → Visual Assistant (Follow user's instructions)



Data: leverage ChatGPT/GPT-4 for multimodal instruction-following data collection

- 158K language-image instruction-following data
- 3 Types: Conversation (Multi-Turn), Detailed description, Complex reasoning

Trainable Parameters

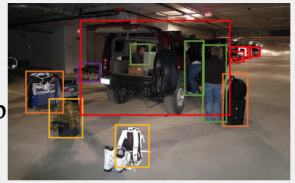
- $\{W, \emptyset\}$

LLaVA (Liu et al. 2023)

Stage 2: Instruction Tuning (Supervised Finetuning)

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>

LLaVA (Liu et al. 2023)

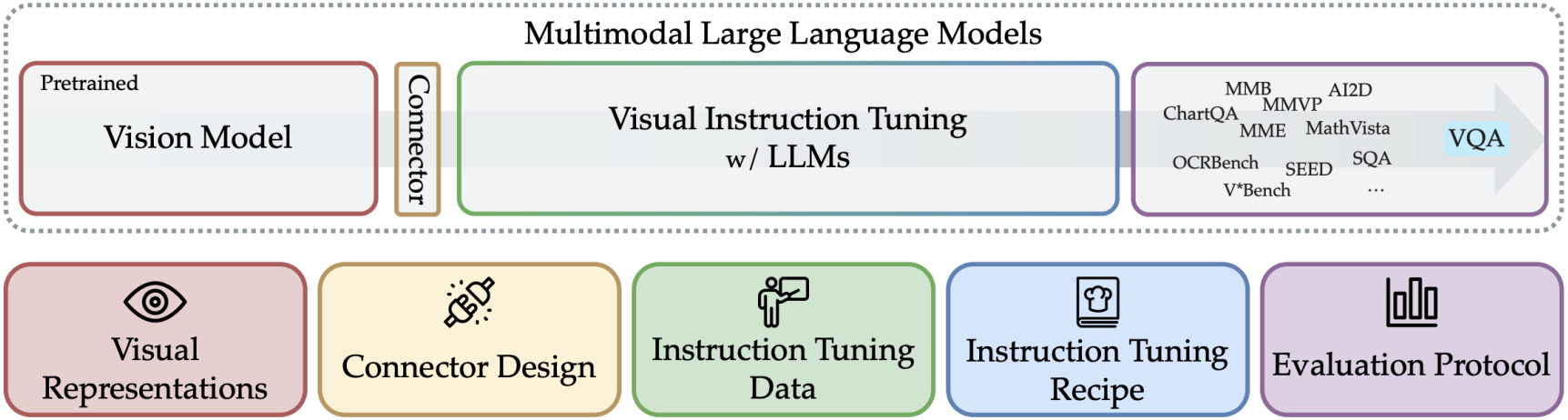
How to Link Vision to LLMs?

Internal Linkage → Vision-Language Models

- Architecture
- Evaluation
- Training Recipe
- Improvement

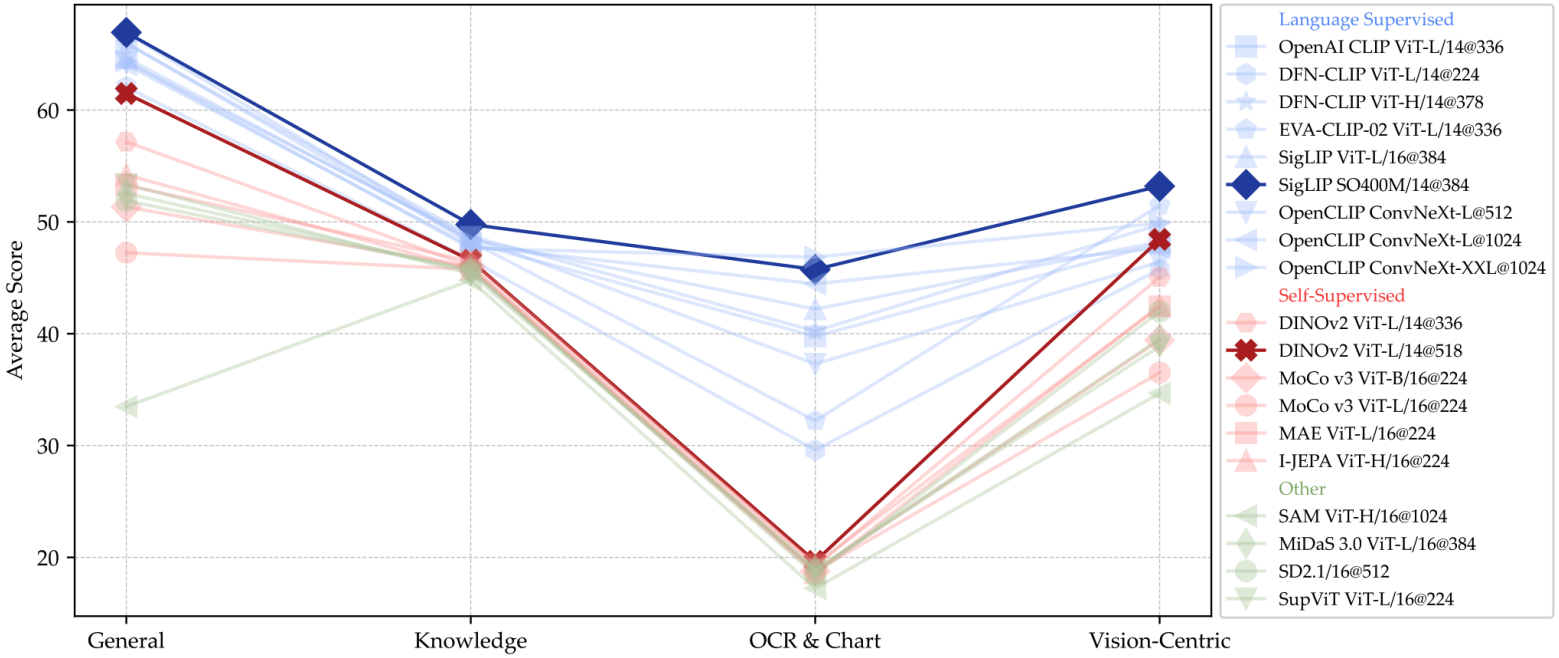
External Linkage → Vision-Language Agents

How to improve the LLaVA-style VLM?



Cambrian-1 (Tong et al. 2024)

Visual Representations



Takeaway: Language supervision offers strong advantages, SigLIP is the best (2024/10)

Cambrian-1 (Tong et al. 2024)

Visual Representations

Language Supervised

Model	Architecture	All	G	K	O	V
SigLIP	ViT-SO400M/14@384	1	1	1	2	1
OpenCLIP	ConvNeXt-XXL@1024	2	6	8	1	3
DFN-CLIP	ViT-H/14@378	3	4	2	5	4
OpenCLIP	ConvNeXt-L@1024	4	8	7	3	8
SigLIP	ViT-L/16@384	5	5	4	4	6
OpenAI CLIP	ViT-L/14@336	6	3	6	6	7
EVA-CLIP-02	ViT-L/14@336	7	2	5	8	2
OpenCLIP	ConvNeXt-L@512	8	7	3	7	9
DFN-CLIP	ViT-L/14@224	9	9	9	9	10
DINOv2*	ViT-L/14@518	10	10	10	10	5

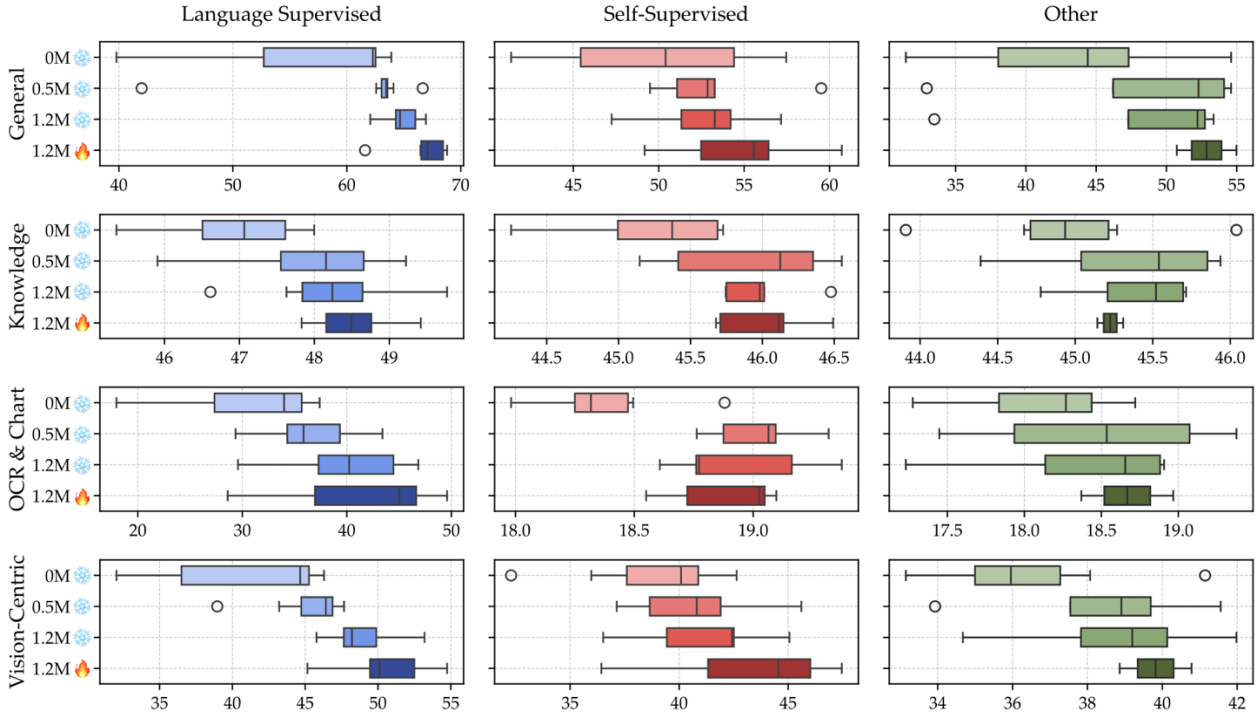
Self-Supervised & Other

Model	Architecture	All	G	K	O	V
DINOv2	ViT-L/14@518	1	1	1	1	1
DINOv2	ViT-L/14@336	2	2	3	3	2
MAE	ViT-L/16@224	3	5	2	2	4
I-JEPA	ViT-H/14@224	4	3	6	8	3
SD2.1	VAE+UNet/16@512	5	7	9	9	5
MiDaS 3.0	ViT-L/16@384	6	6	8	5	6
SupViT	ViT-L/16@224	7	4	9	4	8
MoCo v3	ViT-B/16@224	8	8	4	7	7
MoCo v3	ViT-L/16@224	9	9	5	6	9
SAM	ViT-H/16@1024	10	10	10	10	10

Takeaway: High-res encoders greatly enhance performance on chart & vision-centric benchmarks

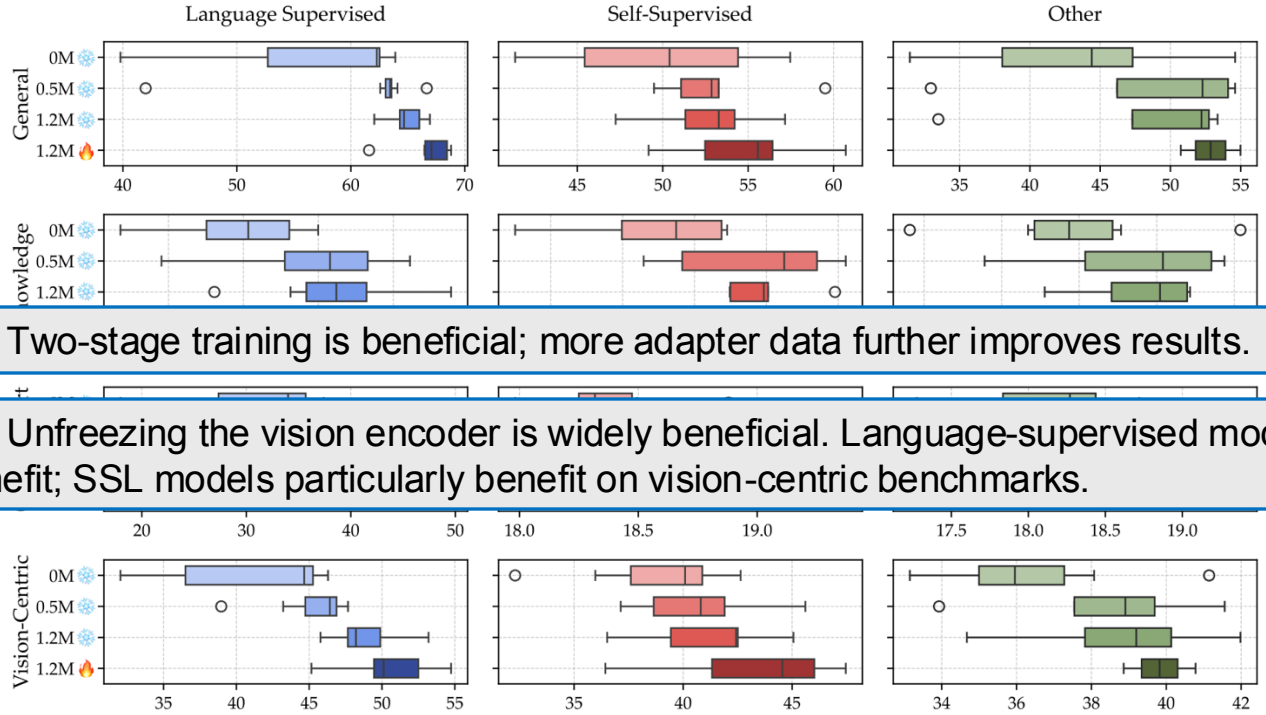
Cambrian-1 (Tong et al. 2024)

Instruction Tuning Recipes



Cambrian-1 (Tong et al. 2024)

Instruction Tuning Recipes

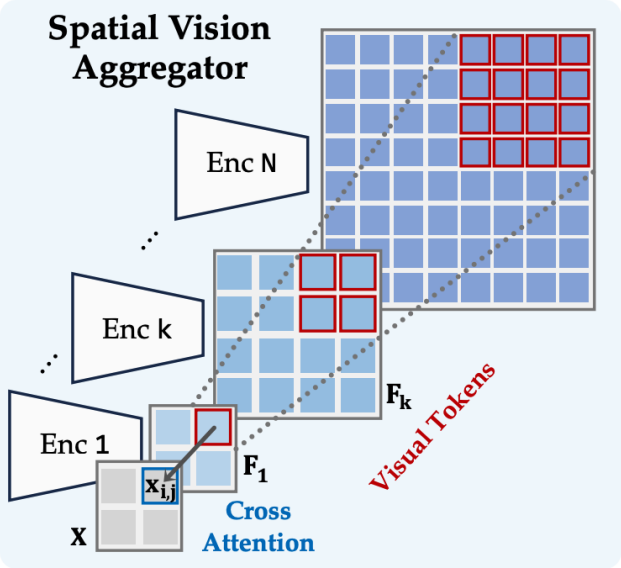


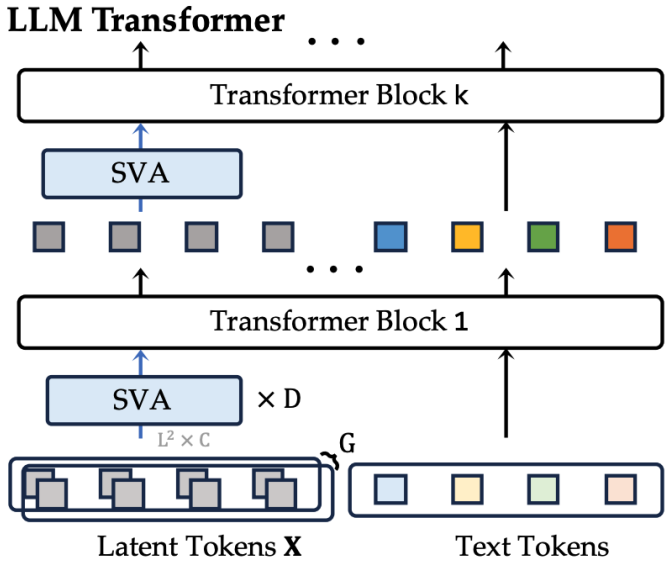
Takeaway: Two-stage training is beneficial; more adapter data further improves results.

Takeaway: Unfreezing the vision encoder is widely beneficial. Language-supervised models always benefit; SSL models particularly benefit on vision-centric benchmarks.

Cambrian-1 (Tong et al. 2024)

New Connector Design





Cambrian-1 (Tong et al. 2024)

New Connector Design

Connector	General	Knowledge	OCR & Chart	Vision-Centric
Concat. [117]	67.2	48.9	50.1	52.6
Resampler [51]	63.1	46.5	27.1	42.6
SVA-no-multi-agg	68.0	49.5	55.2	52.6
SVA	**68.5**	**49.7**	**55.5**	**53.2**

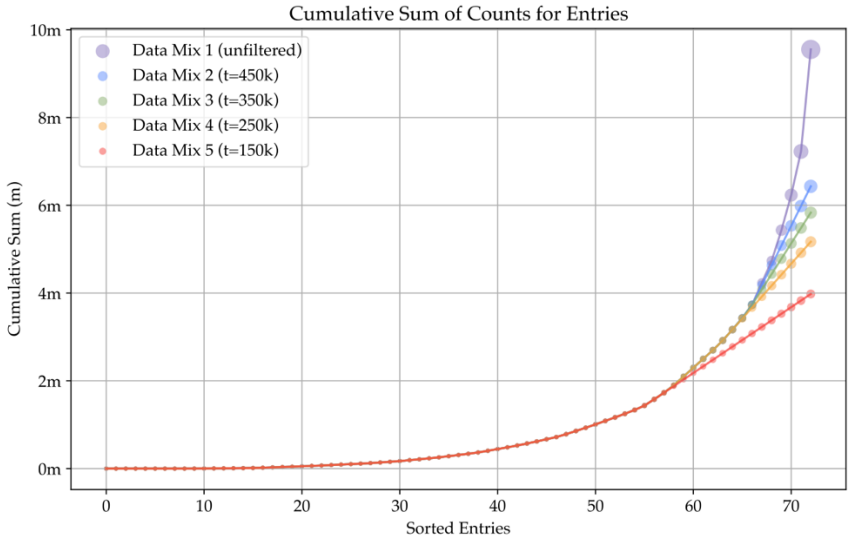
Takeaway: Spatial inductive bias and deep interaction between LLM and vision feature help to better aggregate and condense vision features.

Cambrian-1 (Tong et al. 2024)

Instruction Tuning Data

Internet Data Collection Engine

- Selects target field/subfield (e.g., "Physics")
- Uses LLM (e.g., GPT-4) to identify topics (e.g., "Newton's Laws")
- Searches reliable sources (e.g., Wikipedia)
- Extracts image-caption pairs and uses LLM to create Q&A pairs about the image



Cambrian-1 (Tong et al. 2024)

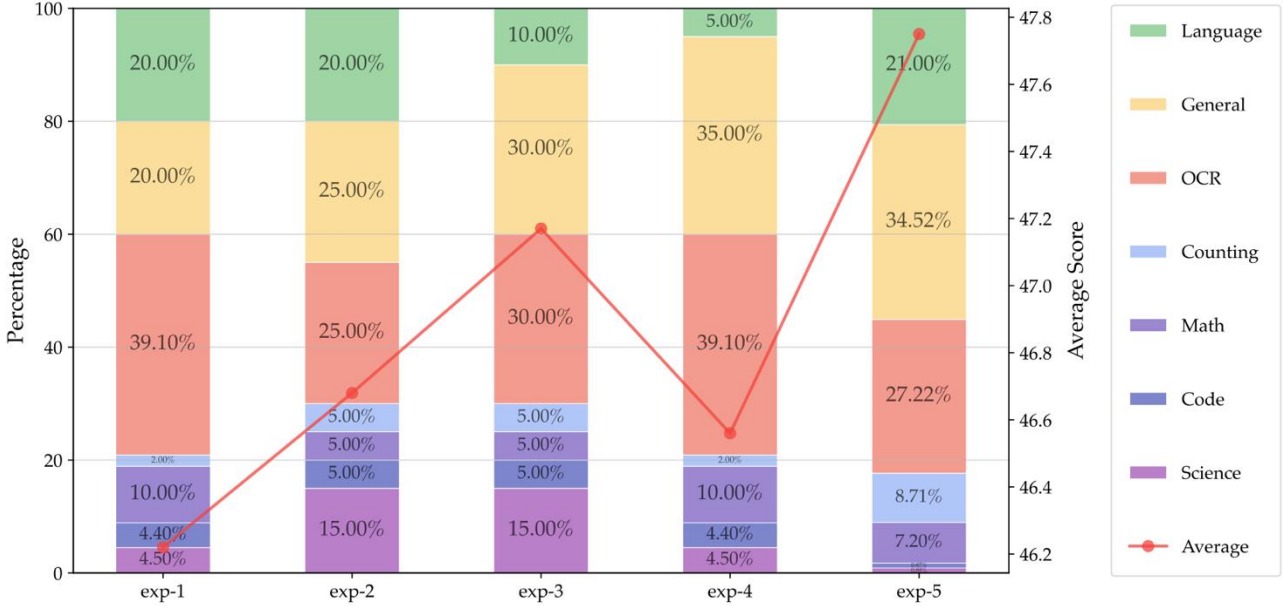
Instruction Tuning Data

	Average	General	Knowledge	OCR & Chart	Vision-Centric
150k	53.7	68.0	51.3	45.2	50.5
250k	**54.3**	**68.1**	51.5	45.3	52.2
350k	54.3	67.4	51.4	**46.0**	**52.3**
450k	54.2	68.0	**52.2**	45.5	50.7

Takeaway: Data Balancing Matters.

Cambrian-1 (Tong et al. 2024)

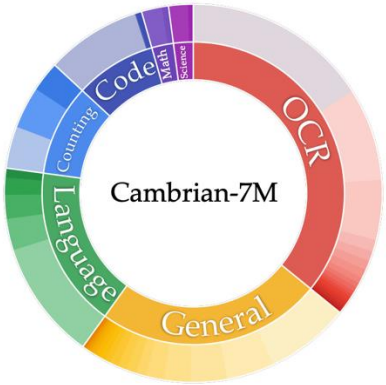
Instruction Tuning Data



Takeaway: Different data ratios have a non-trivial impact on the overall performance.

Cambrian-1 (Tong et al. 2024)

Instruction Tuning Data



OCR (27.6%)
- Filtered DVQA (1550.0 K)
- DVQA [54] (775.0 K)
- SynthDog [60] (500.0 K)
- ArxivQA [69] (100.0 K)
- OCRVQA [93] (80.0 K)
- ScreenQA [49] (79.0 K)
- WIkiSQL [144] (74.0 K)
- Low-Level Vision [22] (50.0 K)
- DocVQA [90] (39.0 K)
- WTQ [99] (38.0 K)
- ChartQA [89] (28.0 K)
- IconQA [82] (27.0 K)
- Chart2Text [55] (26.0 K)
- TabMWP [81] (23.0 K)
- TextCaps [111] (22.0 K)
- LLAVAR [140] (20.0 K)
- ST-VQA [15] (17.0 K)
- AI2D [58] (15.0 K)
- RenderedText [125] (10.0 K)
- VisText [115] (9.0 K)
- FinQA [26] (6.0 K)
- InfoVQA [14] (2.0 K)
- TAT-QA [148] (2.0 K)
- HiTab [27] (2.0 K)

General (33.3%)
- ALLaVA [20] (700.0 K)
- Q-Instruct [126] (400.0 K)
- LNQA [101] (302.0 K)
- LVIS-Instruct4V [122] (220.0 K)
- LLaVA150K [75] (150.0 K)
- VisualGenome [62] (86.0 K)
- VQAv2 [43] (83.0 K)
- GPT4V Rewritten (77.0 K)
- GQA [50] (72.0 K)
- A-OKVQA [108] (50.0 K)
- AlfWorld [137] (45.0 K)
- ShareGPT [22] (40.0 K)
- RefCOCO [131] (30.0 K)
- VizWiz [44] (20.0 K)
- Visual7W [149] (14.0 K)
- LAION GPT-4V [63] (11.0 K)
- IDK [17] (11.0 K)
- OKVQA [88] (9.0 K)
- HatefulMemes [59] (8.0 K)
- OODVQA [120] (8.0 K)
- SketchyVQA [120] (8.0 K)
- Visualmrc [114] (3.0 K)

Language (23.8%)
- OpenOrca [71] (994.0 K)
- MathInstruct [133] (262.0 K)
- OrcaMath [92] (200.0 K)
- WizardCoder [86] (143.0 K)
- OpenCodeInterpreter [143] (66.0 K)
- Dolly [30] (11.0 K)

Counting (8.5%)
- Filtered CLEVR (350.0 K)
- CLEVR [52] (350.0 K)
- TallyQA [1] (250.0 K)

Code (0.8%)
- Filtered WebSight (790.0 K)
- WebSight [64] (10.0 K)
- DaTikz [12] (47.0 K)
- Design2Code [110] (0.5 K)

Math (3.2%)
- Geo170K [37] (170.0 K)
- RAVEN [139] (42.0 K)
- GeomVerse [57] (9.0 K)
- MathVision [123] (3.0 K)
- Inter-GPS [83] (1.0 K)
- TQA [4] (1.0 K)

Science (2.9%)
- Data Engine (161.0 K)
- PathVQA [46] (32.0 K)
- ScienceQA [84] (12.0 K)

	Average	General	Knowledge	OCR & Chart	Vision-Centric
LLaVA-665K	40.7	64.7	45.2	20.8	32.0
Cambrian-10M	54.8	68.7	51.6	**47.3**	51.4
Cambrian-7M	**55.9**	**69.6**	**52.6**	**47.3**	**54.1**

Cambrian-1 (Tong et al. 2024)

Overall Performance

Model		General					Knowledge					OCR & Chart					Vision-Centric				
Method	# Vis Tok.	Avg	MME^P	MMB	$SEED^I$	GQA	Avg	SQA^I	$MMMU^V$	$MathVista^M$	AI2D	Avg	ChartQA	OCRBench	TextVQA	DocVQA	Avg	MMVP	RealworldQA	$CV\text{-}Bench^{2D}$	$CV\text{-}Bench^{3D}$
GPT-4V	UNK.	63.0	1409.4	75.8	69.1	36.8	65.2	75.7	56.8	49.9	78.2	77.4	78.5	64.5	78.0	88.4	62.4	50.0	61.4	64.3	73.8
Gemini-1.0 Pro	UNK.	-	1496.6	73.6	70.7	-	-	79.5	47.9	45.2	-	-	-	65.9	-	-	-	-	-	-	-
Gemini-1.5 Pro	UNK.	-	-	-	-	-	-	-	58.5	52.1	80.3	-	81.3	-	73.5	86.5	-	-	67.5	-	-
Grok-1.5	UNK.	-	-	-	-	-	-	-	53.6	52.8	88.3	-	76.1	-	78.1	85.6	-	-	68.7	-	-
MM-1-8B	144	-	1529.3	72.3	69.9	-	-	72.6	37.0	35.9	-	-	-	-	-	-	-	-	-	-	-
MM-1-30B	144	-	1637.6	75.1	72.1	-	-	81.0	44.7	39.4	-	-	-	-	-	-	-	-	-	-	-
Base LLM: Llama-3-Ins-8B																					
Mini-Gemini-HD-8B	2880	72.7	**1606.0**	72.7	73.2	64.5	55.7	75.1	37.3	37.0	**73.5**	62.9	59.1	47.7	70.2	74.6	51.5	18.7	62.1	62.2	63.0
LLaVA-NeXT-8B	2880	72.5	1603.7	72.1	72.7	**65.2**	55.6	72.8	41.7	36.3	71.6	63.9	69.5	49.0	64.6	72.6	56.6	38.7	60.1	62.2	65.3
Cambrian-1-8B	576	**73.1**	1,547.1	**75.9**	**74.7**	64.6	**61.3**	**80.4**	**42.7**	**49.0**	73.0	**71.3**	**73.3**	**62.4**	**71.7**	**77.8**	**65.0**	**51.3**	**64.2**	**72.3**	**72.0**
Base LLM: Vicuna-1.5-13B																					
Mini-Gemini-HD-13B	2880	70.7	1597.0	68.6	70.6	63.7	54.1	71.9	37.3	37.0	70.1	60.8	56.6	46.6	70.2	69.8	49.4	19.3	57.5	53.6	67.3
LLaVA-NeXT-13B	2880	69.9	1575.0	70.0	65.6	**65.4**	53.7	73.5	36.2	35.1	70.0	62.9	62.2	51.4	67.1	70.9	55.9	36.0	59.1	62.7	65.7
Cambrian-1-13B	576	**73.7**	**1,610.4**	**75.7**	**74.4**	64.3	**60.2**	**79.3**	**40.0**	**48.0**	**73.6**	**71.3**	**73.8**	**61.9**	**72.8**	**76.8**	**62.2**	**41.3**	**63.0**	**72.5**	**71.8**
Base LLM: Hermes2-Yi-34B																					
Mini-Gemini-HD-34B	2880	76.2	1659.0	80.6	75.3	65.8	62.4	77.7	48.0	43.4	**80.5**	68.1	67.6	51.8	74.1	**78.9**	63.8	37.3	67.2	71.5	79.2
LLaVA-NeXT-34B	2880	76.0	1633.2	79.3	**75.9**	**67.1**	62.5	81.8	46.7	46.5	74.9	67.7	68.7	54.5	69.5	78.1	64.0	47.3	61.0	73.0	74.8
Cambrian-1-34B	576	**76.8**	**1689.3**	**81.4**	75.3	65.8	**67.0**	**85.6**	**49.7**	**53.2**	79.7	**71.9**	**75.6**	**60.0**	**76.7**	75.5	**68.5**	**52.7**	**67.8**	**74.0**	**79.7**

Cambrian-1 (Tong et al. 2024)

How to Link Vision to LLMs?

Internal Linkage → Vision-Language Models

- Architecture
- Evaluation
- Training Recipe
- Improvement
- Takeaway

External Linkage → Vision-Language Agents

Human Evaluation

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organizatio	License	Knowledge Cutoff
1	1	ChatGPT-4o-latest (2024-09-03)	1250	+12/-10	2889	OpenAI	Proprietary	2023/10
2	2	Gemini-1.5-Pro-Exp-0827	1232	+6/-6	10695	Google	Proprietary	2023/11
3	4	Gemini-1.5-Flash-Exp-0827	1210	+10/-7	3788	Google	Proprietary	2023/11
3	2	GPT-4o-2024-05-13	1208	+5/-5	19278	OpenAI	Proprietary	2023/10
5	4	Claude 3.5 Sonnet	1189	+5/-4	21653	Anthropic	Proprietary	2024/4
6	6	Gemini-1.5-Pro-001	1151	+5/-4	17886	Google	Proprietary	2023/11
6	7	GPT-4-Turbo-2024-04-09	1151	+6/-5	13752	OpenAI	Proprietary	2023/12
8	8	GPT-4o-mini-2024-07-18	1121	+5/-5	13188	OpenAI	Proprietary	2023/10
8	9	Gemini-1.5-Flash-8b-Exp-0827	1111	+9/-10	3474	Google	Proprietary	2023/11
9	8	Qwen2-VL-72b-Instruct	1103	+9/-12	2127	Alibaba	Qwen	2024/9
11	11	Claude 3 Opus	1076	+5/-5	15940	Anthropic	Proprietary	2023/8
11	12	Llama-3.2-90b-Vision-Instruct	1074	+10/-8	3100	Meta	Llama 3.2	2023/11
11	9	Gemini-1.5-Flash-001	1073	+6/-5	14254	Google	Proprietary	2023/11

Multimodal Arena (Chou et al. 2024)

Video Understanding

#	Model	LLM Params	Frames	Date	Overall (%)		Short Video (%)		Medium Video (%)		Long Video (%)	
					w/o subs	w subs	w/o subs	w subs	w/o subs	w subs	w/o subs	w subs
1	**Gemini 1.5 Pro** Google	-	1/0.5 fps[1*]	2024-06-15	**75.0**	**81.3**	81.7	84.5	74.3	81.0	67.4	77.4
2	**Qwen2-VL** Alibaba	72B	768[3*]	**2024-08-19**	**71.2**	**77.8**	80.1	82.2	71.3	76.8	62.2	74.3
3	**GPT-4o** OpenAI	-	384[2*]	2024-06-15	**71.9**	**77.2**	80.0	82.8	70.3	76.6	65.3	72.1
4	**LLaVA-Video** Bytedance & NTU S-Lab	72B	64	**2024-08-28**	**70.6**	**76.9**	81.4	82.8	68.9	75.6	61.5	72.5
5	**Gemini 1.5 Flash** Google	-	1/0.5 fps[1*]	2024-06-15	**70.3**	**75.0**	78.8	79.8	68.8	74.7	61.1	68.8
6	**Oryx-1.5** THU & Tencent & NTU	34B	128	**2024-10-21**	**67.3**	**74.9**	77.3	80.6	65.3	74.3	59.3	69.9
7	**Aria** Rhymes AI	8x3.5B	256	**2024-10-11**	**67.6**	**72.1**	76.9	78.3	67.0	71.7	58.8	66.3
8	**LLaVA-OneVision** Bytedance & NTU S-Lab	72B	32	**2024-08-08**	**66.3**	**69.6**	76.7	79.3	62.2	66.9	60.0	62.4
9	**GPT-4o mini** OpenAI	-	250	2024-07-21	**64.8**	**68.9**	72.5	74.9	63.1	68.3	58.6	63.4
10	**ByteVideoLLM** Bytedance	14B	100	**2024-10-21**	**64.6**	**68.8**	74.4	77.1	62.9	69.1	56.4	60.2
11	**VideoLLaMA 2** Alibaba	72B	32	**2024-08-29**	**62.4**	**64.7**	69.8	72.0	59.9	63.0	57.6	59.0
12	**VILA-1.5** NVIDIA & MIT	34B	14	2024-07-21	**62.3**	**64.1**	72.0	74.0	61.2	62.6	53.8	55.7

Video-MME (Fu et al. 2024)

Model Choice

Fully Open-Sourced Large VLMs (Data+Training+Inference):

- Cambrian-1

Open-Sourced Large VLMs (Inference Only):

- Qwen2-VL

Proprietary Large VLMs (API Only):

- Image: GPT-4o
- Video: Gemini-1.5 Pro

2024/10；My Personal Opinion + Benchmark Results

Video-MME (Fu et al. 2024)

How to Link Vision to LLMs?

Internal Linkage → Vision-Language Models

- Architecture
- Evaluation
- Training Recipe
- Improvement
- Takeaway

External Linkage → Vision-Language Agents