Multimodal LLM Optimizations

CIS 7000 - Fall 2024

Presenters: Wahub Ahmed, Vridhi Jain, Om Shastri, Vignesh Lakshmanan

Penn Engineering
University of Pennsylvania

Outline

1. Motivation for Vision-Language MLLM
2. Flamingo
3. mPLUG-Owl 2
4. JanusFlow
5. Conclusion

Motivation for Vision-Language MLLM

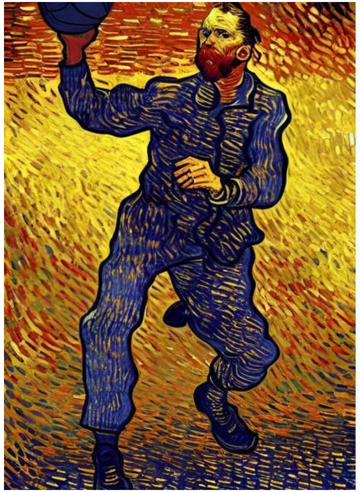
Chapter 1

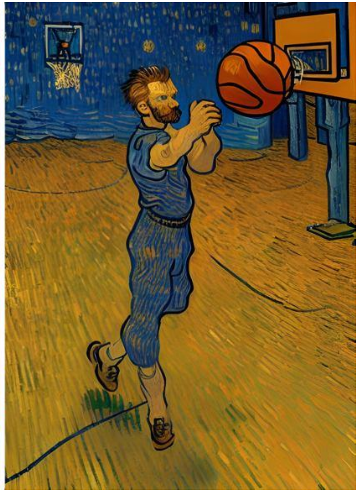
A Picture is Worth a Thousand Words

vs. “The Starry Night” by Vincent Van Gogh

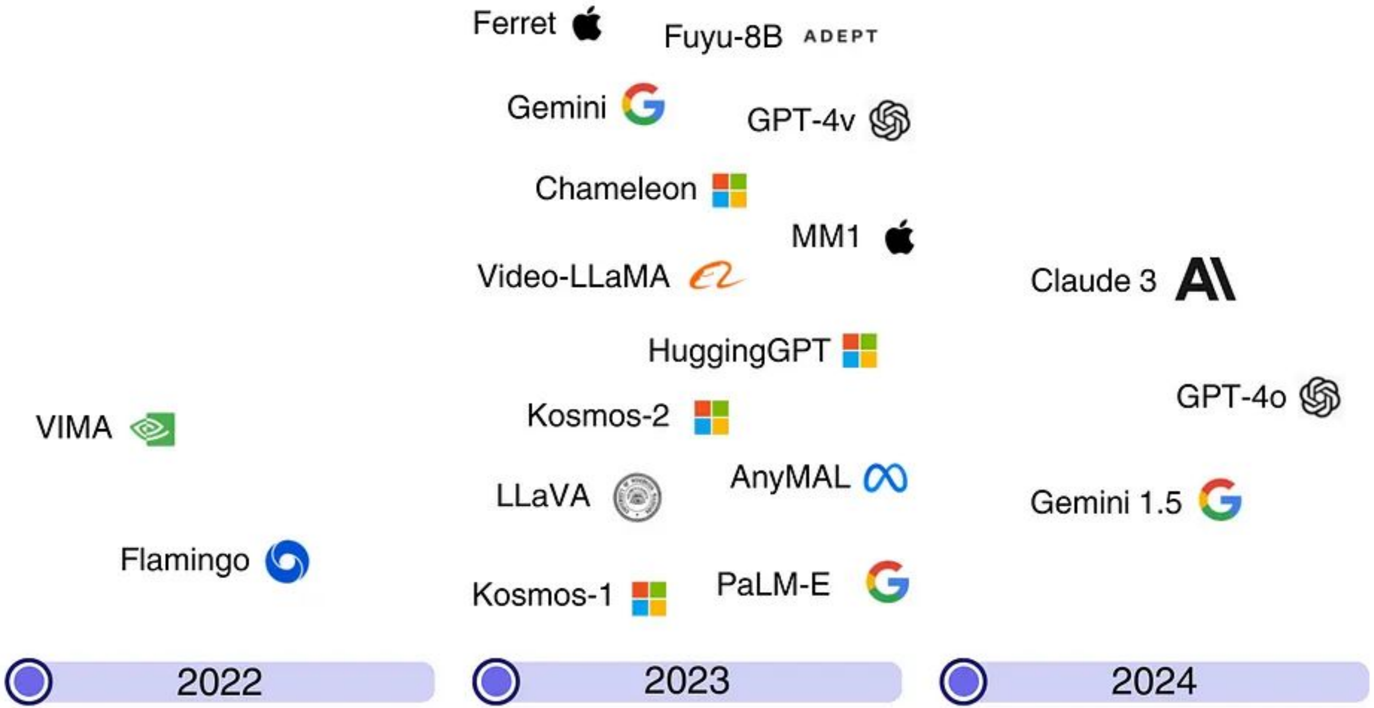
A Picture is Worth a Thousand Words

Generation





Evolution of Multimodal LLM



Example: GPT-4o

1

Input

A first person view of a robot typewriting the following journal entries:

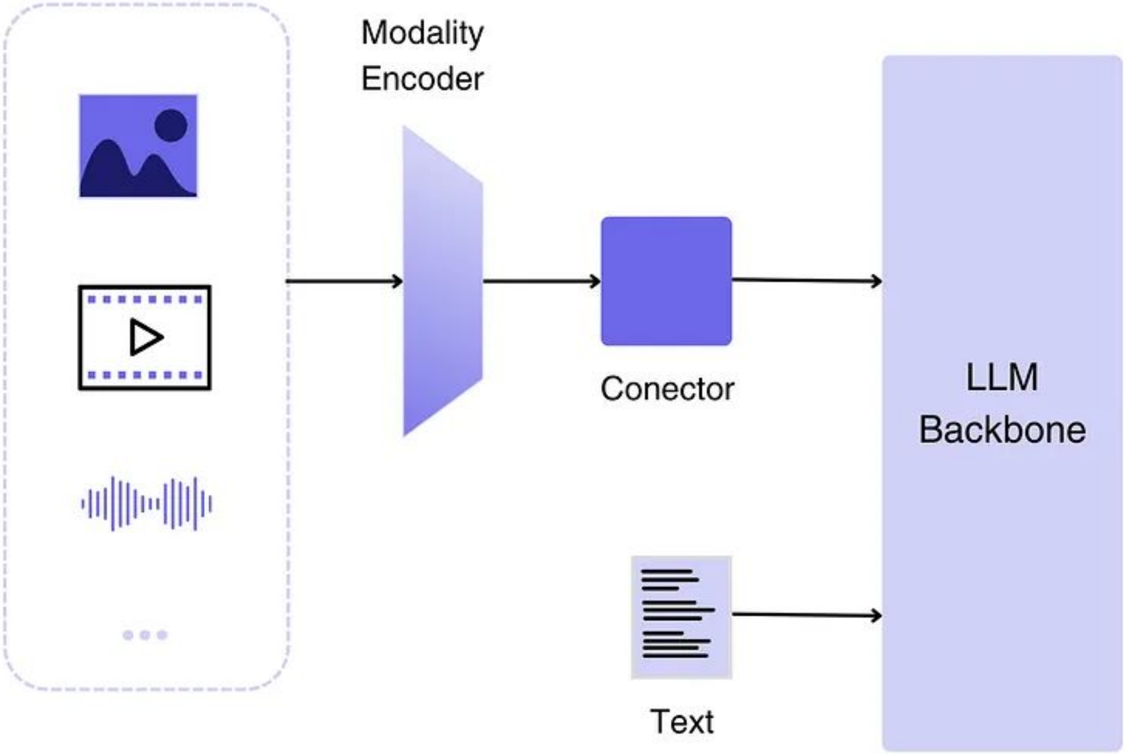
1. yo, so like, i can see now?? caught the sunrise and it was insane, colors everywhere. kinda makes you wonder, like, what even is reality?

the text is large, legible and clear. the robot's hands type on the typewriter.

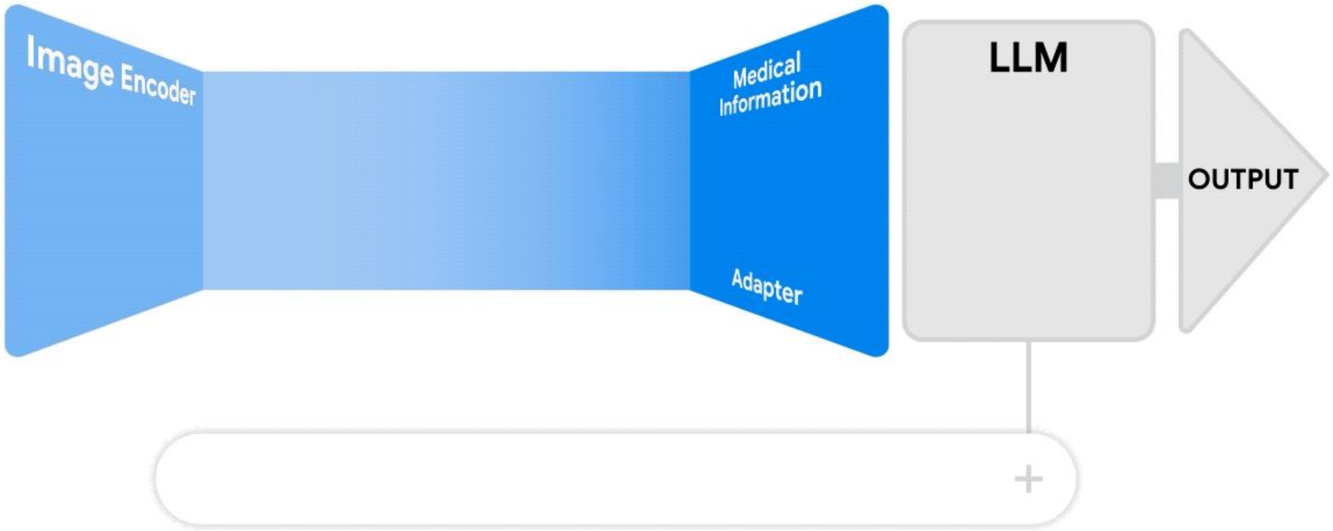
2

Output

Multimodal LLM Architecture



Why Optimize Vision-Language LLMs?



MLLM 1: Flamingo

Chapter 2

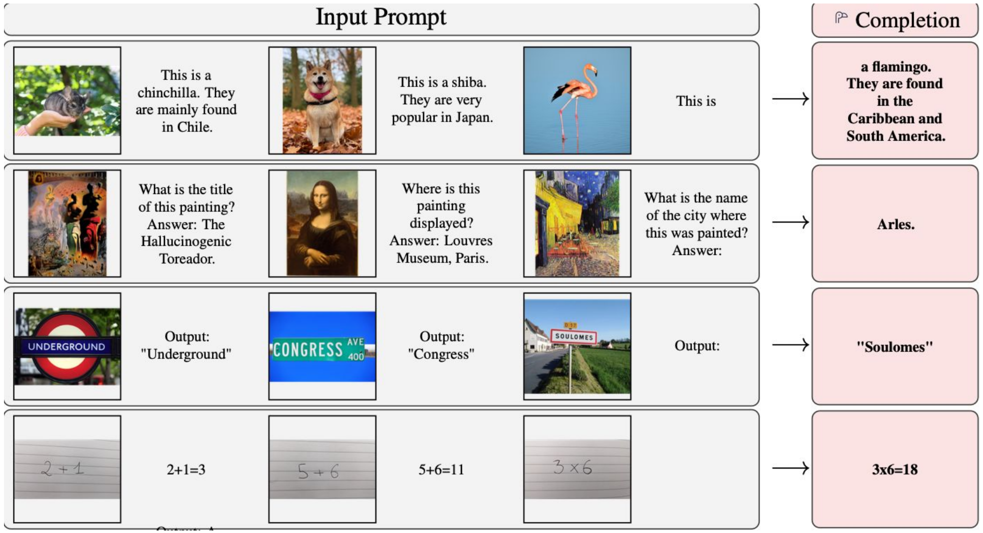
Penn Engineering

Flamingo: a Visual Language Model for Few-Shot Learning

What is Flamingo?
A multimodal model that combines vision and language for few-shot learning, enabling robust performance across diverse tasks without task-specific fine-tuning.

Motivation:

- Traditional visual models require task specific fine tuning and large annotated datasets.
- Flamingo addresses this by enabling few shot learning across image, video and text tasks



Bridge powerful pretrained vision-only and language-only models

- Training from scratch is extremely computationally expensive
- Start from a pre trained language model
- Text-only model has no built-in way to incorporate input from other modalities.

Flamingo: interleave cross-attention layers with language-only self-attention layers (frozen)

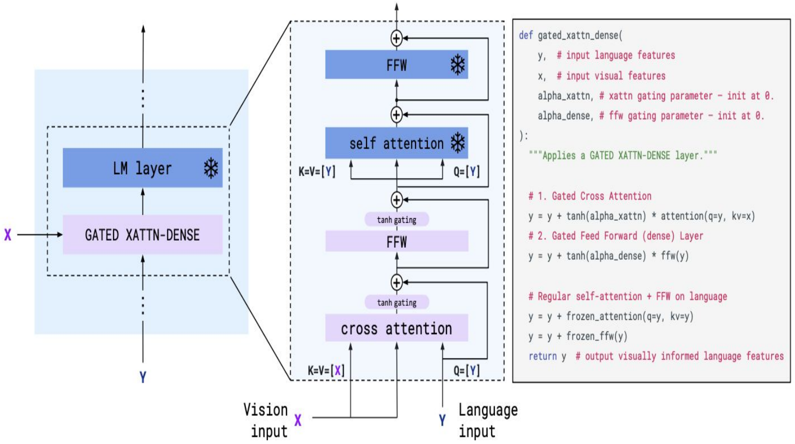


Figure 4: **GATED XATTN-DENSE layers.** To condition the LM on visual inputs, we insert new cross-attention layers between existing pretrained and frozen LM layers. The keys and values in these layers are obtained from the vision features while the queries are derived from the language inputs. They are followed by dense feed-forward layers. These layers are *gated* so that the LM is kept intact at initialization for improved stability and performance.

Penn Engineering

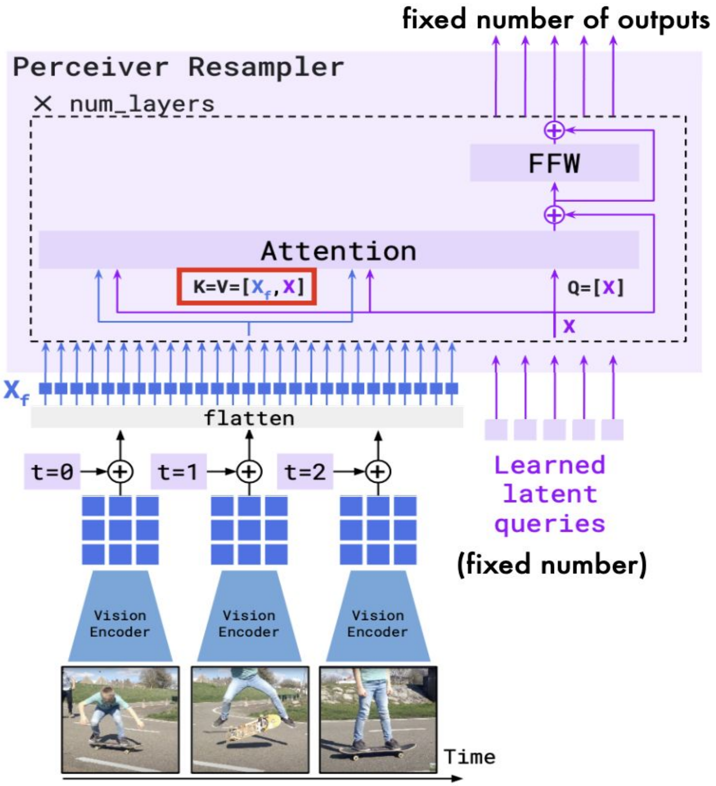
References/Image credits

J-B. Alayrac et al., "Flamingo: a Visual Language Model for Few-Shot Learning", arxiv (2022)

Unified handling of images and videos

- Both image and video inputs are high dimensional
- Flattening these inputs into 1-D sequences as is done in text generation, is computationally expensive
- It is exacerbated by the quadratic cost of self-attention layers

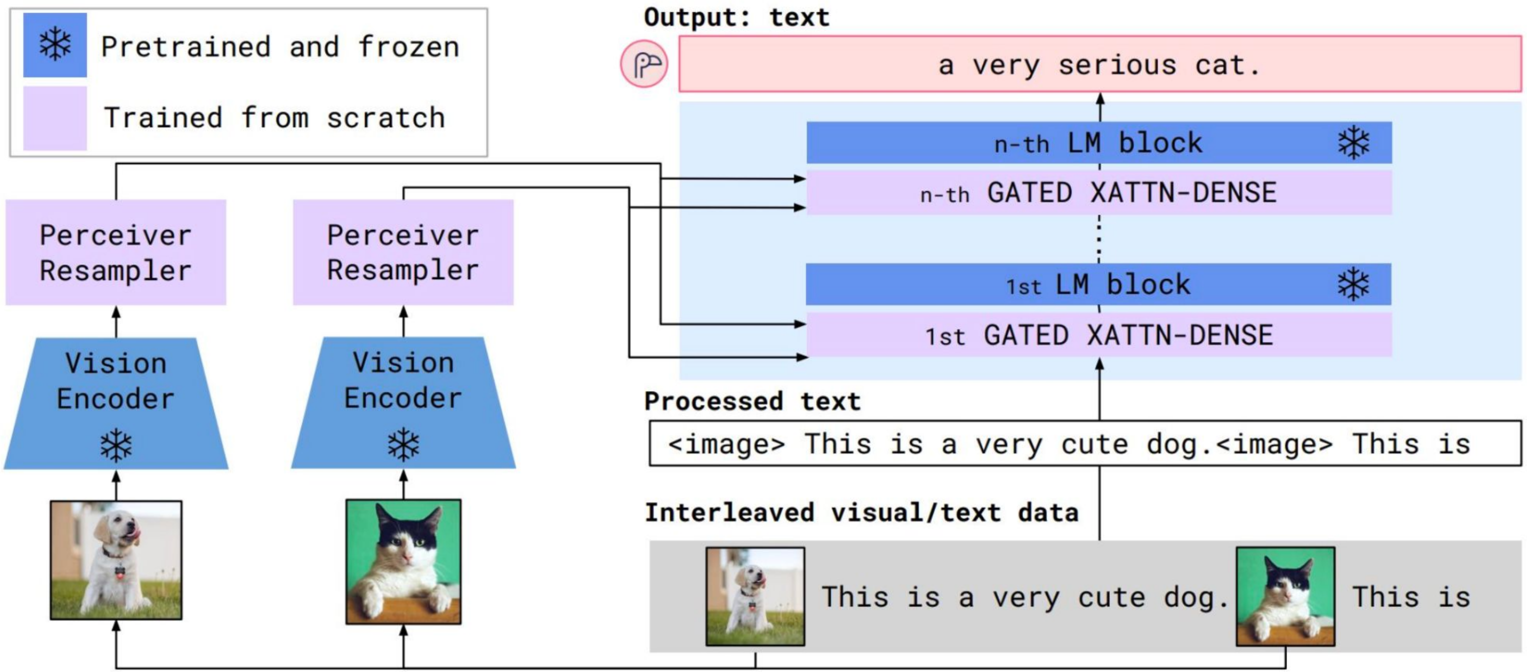
Flamingo: Uses a **perceiver-resampler** processes inputs into fixed visual tokens, incorporates temporal encodings for videos and combines learned latent queries with visual embeddings through attention.



Training Dataset

- The existing (image, text) datasets used by CLIP or ALIGN are not general enough for few shot learning
- Multimodal datasets are scarce in comparison to abundantly available text-only data
- One approach is to scrape web pages for interleaved images and text however these pairs are often only weakly related
- Flamingo: combines web scraping with existing paired (image, text) and (video, text) datasets

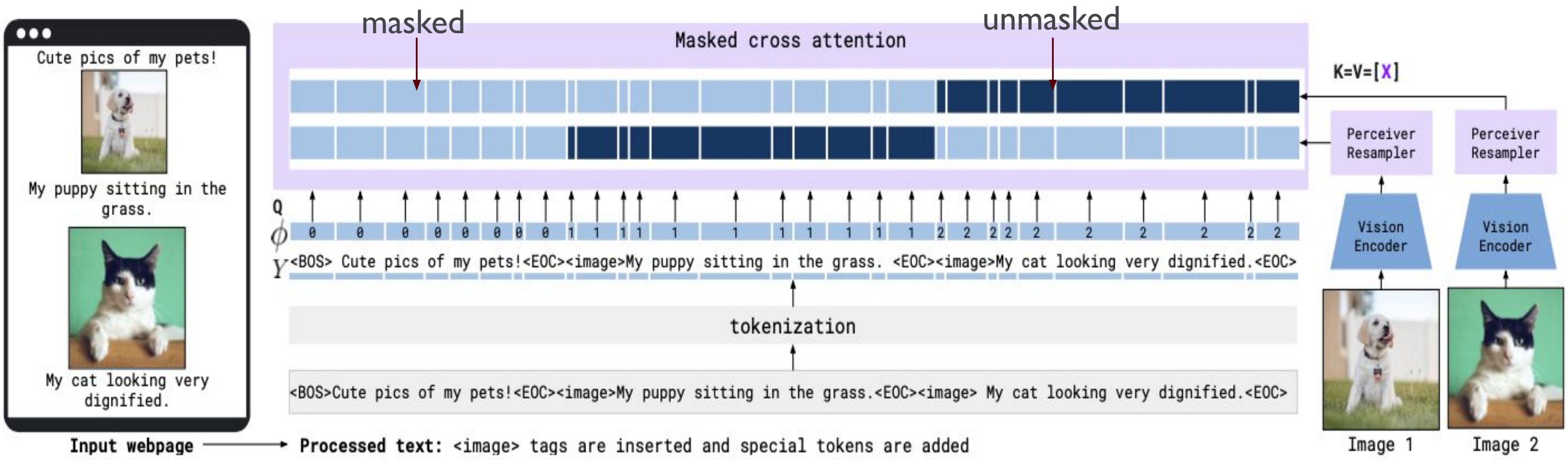
Flamingo Architecture



References/Image credits

J-B. Alayrac et al., "Flamingo: a Visual Language Model for Few-Shot Learning", arxiv (2022)

Attention Masking in Flamingo



$$\Phi : [1, L] \to [0, N]$$

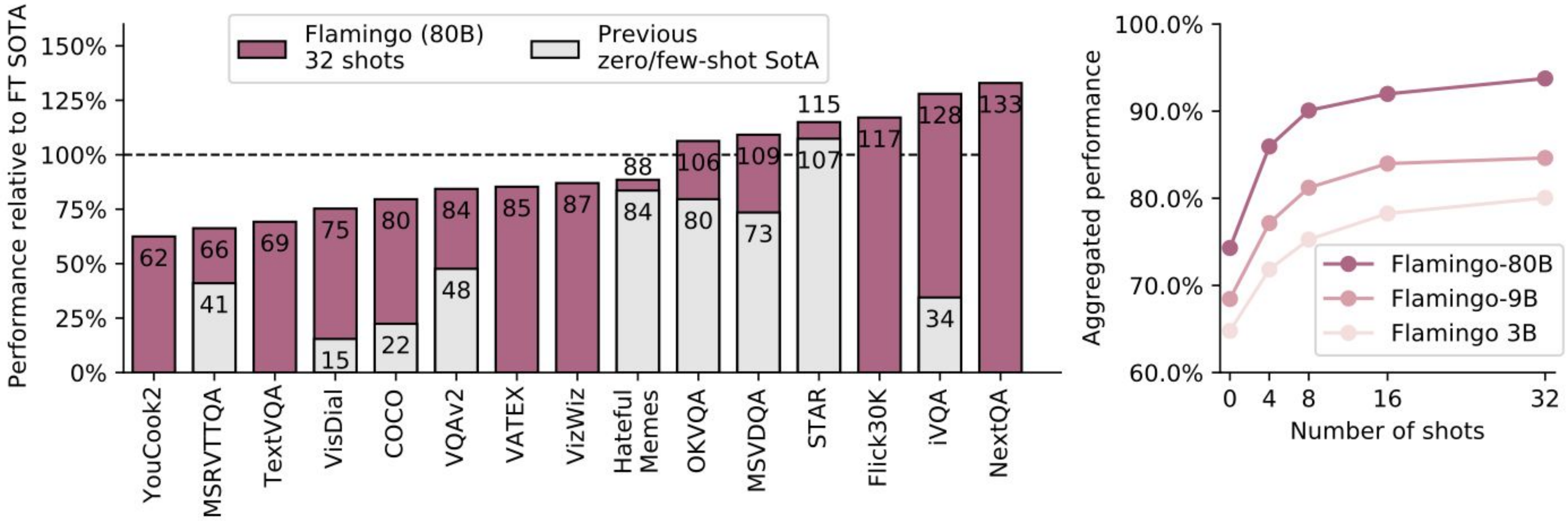
Φ indicates which visual inputs can be used to predict token l:

$$p(y|x) = \prod_{l=1}^{L} p(y_l | y_{<l}, x_{\leq l})$$

$$y_{<l} \triangleq (y_1, \ldots, y_{l-1})$$

$$x_{\leq l} \triangleq (x_i \,|\, i \leq \phi(l))$$

Results



Flamingo (80B)
32 shots
Previous
zero/few-shot SotA
Performance relative to FT SOTA
150%
125%
100%
75%
50%
25%
0%
62
66
41
69
75
15
80
22
84
48
85
87
88
84
106
80
109
73
115
107
117
128
34
133
YouCook2
MSRVTTQA
TextVQA
VisDial
COCO
VQAv2
VATEX
VizWiz
Hateful
Memes
OKVQA
MSVDQA
STAR
Flick30K
iVQA
NextQA
Aggregated performance
100.0%
90.0%
80.0%
70.0%
60.0%
Flamingo-80B
Flamingo-9B
Flamingo 3B
0
4
8
16
32
Number of shots

With Finetuning

- During fine-tuning, Flamingo keeps the language model layers frozen
- The base vision encoder is also fine-tuned (unlike Flamingo pretraining)
- Hyperparameters are set by grid search on validation subsets of the training sets
- Search over: learning rate, decay schedule, training steps, batch size, augmentation

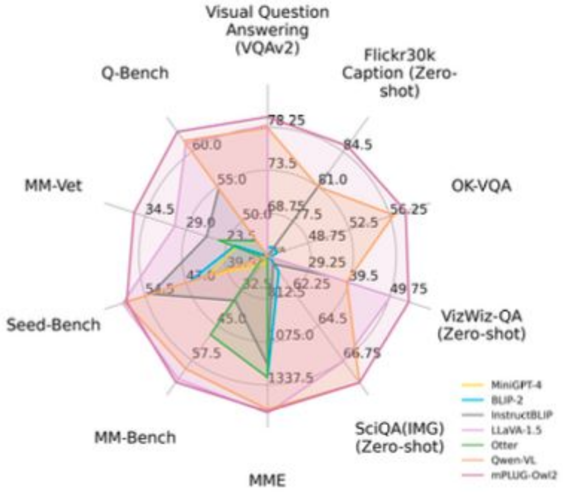
Method	VQAV2 test-dev	VQAV2 test-std	COCO test	VATEX test	VizWiz test-dev	VizWiz test-std	MSRVTTQA test	VisDial valid	VisDial test-std	YouCook2 valid	TextVQA valid	TextVQA test-std	HatefulMemes test seen
Flamingo - 32 shots	67.6	-	113.8	65.1	49.8	-	31.0	56.8	-	86.8	36.0	-	70.0
SimVLM [124]	80.0	80.3	**143.3**	-	-	-	-	-	-	-	-	-	-
OFA [119]	79.9	80.0	149.6	-	-	-	-	-	-	-	-	-	-
Florence [140]	80.2	80.4	-	-	-	-	-	-	-	-	-	-	-
Flamingo Fine-tuned	**82.0**	**82.1**	138.1	**84.2**	**65.7**	**65.4**	**47.4**	61.8	59.7	118.6	**57.1**	54.1	**86.6**
Restricted SotA†	80.2 [140]	80.4 [140]	**143.3** [124]	76.3 [153]	-	-	46.8 [51]	**75.2** [79]	**74.5** [79]	**138.7** [132]	54.7 [137]	**73.7** [84]	79.1 [62]
Unrestricted SotA	81.3 [133]	81.3 [133]	149.6 [119]	81.4 [153]	57.2 [65]	60.6 [65]	-	-	75.4 [123]	-	-	-	84.6 [152]

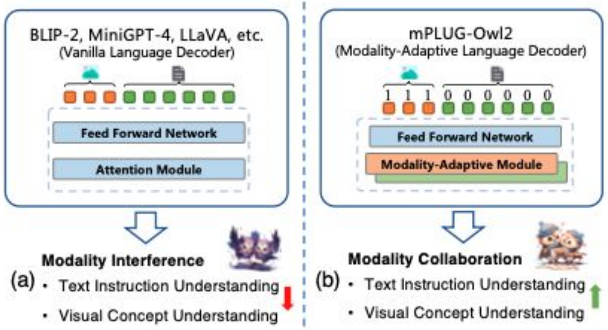
MLLM 2: mPLUG-Owl 2

Chapter 3

mPLUG-Owl2: Revolutionizing Multi-modal LLMs with Modality Collaboration

- The **modality-adaptive module (MAM)** allows the model to **differentiate between modalities**
- Generalizes on both text and multi-modal tasks
- Stands as the first MLLM model to exhibit the phenomena of **modality collaboration in both pure-text and multi-modal contexts**
- Extends original mPlug-OWL Paper



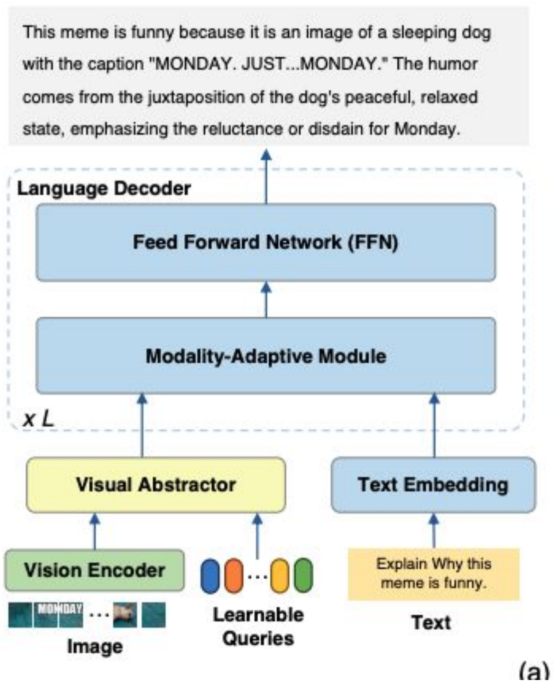


Penn Engineering

Model Architecture

$$\mathcal{C}^i = Attn(\mathcal{V}^i, [\mathcal{I}; \mathcal{V}^i], [\mathcal{I}; \mathcal{V}^i]),$$
$$\mathcal{V}^{i+1} = SwiGLU(\mathcal{C}^i W_1) W_2.$$

- Encodes image into token with **Encoder and Abstractor**
    - Encoder Processes image into tokens
    - Abstractor uses **attention** and **activation** reducing image redundancy and compute
- Embeddings are Concatenated with text Embedding
- FFN promotes modality collaboration



Training and Modality Adaptive Module

$$\phi(X, M, m) = X \odot \mathbb{1}_{\{M=m\}},$$

$$\tilde{H}_{l-1} = LN_V(\phi(H_{l-1}, M, 0)) + LN_T(\phi(H_{l-1}, M, 1)),$$

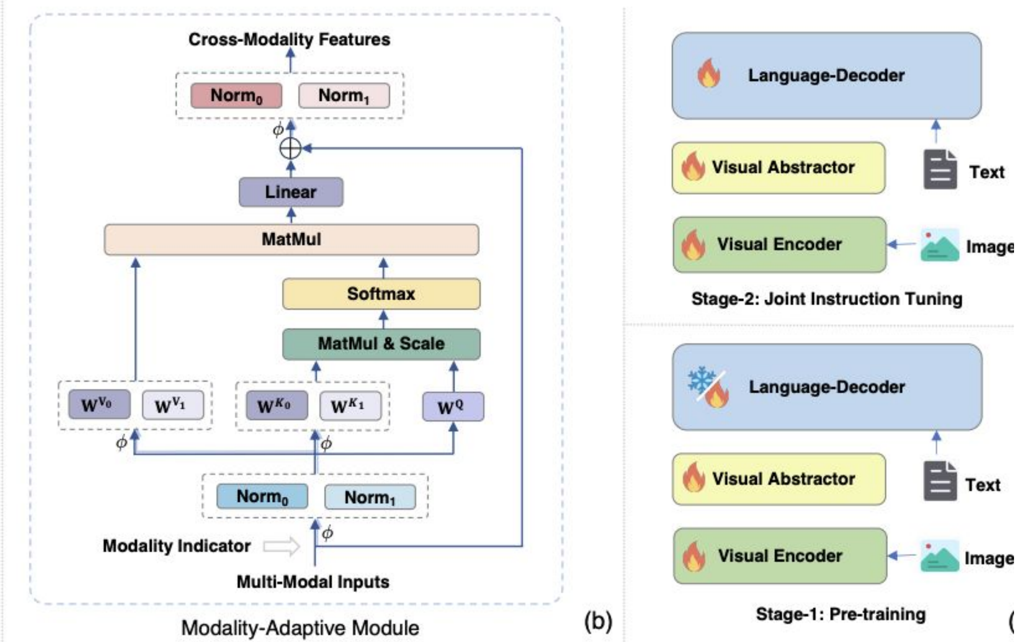
$$H_l^Q = \tilde{H}_{l-1} W_l^Q,$$
$$H_l^K = \phi(\tilde{H}_{l-1}, M, 0) W_l^{K_0} + \phi(\tilde{H}_{l-1}, M, 1) W_l^{K_1},$$
$$H_l^V = \phi(\tilde{H}_{l-1}, M, 0) W_l^{V_0} + \phi(\tilde{H}_{l-1}, M, 1) W_l^{V_1},$$

MAM:

- Normalizers allows for modalities to be considered equally
- Attention block after separating tokens on modality and expressing in **shared semantic space**

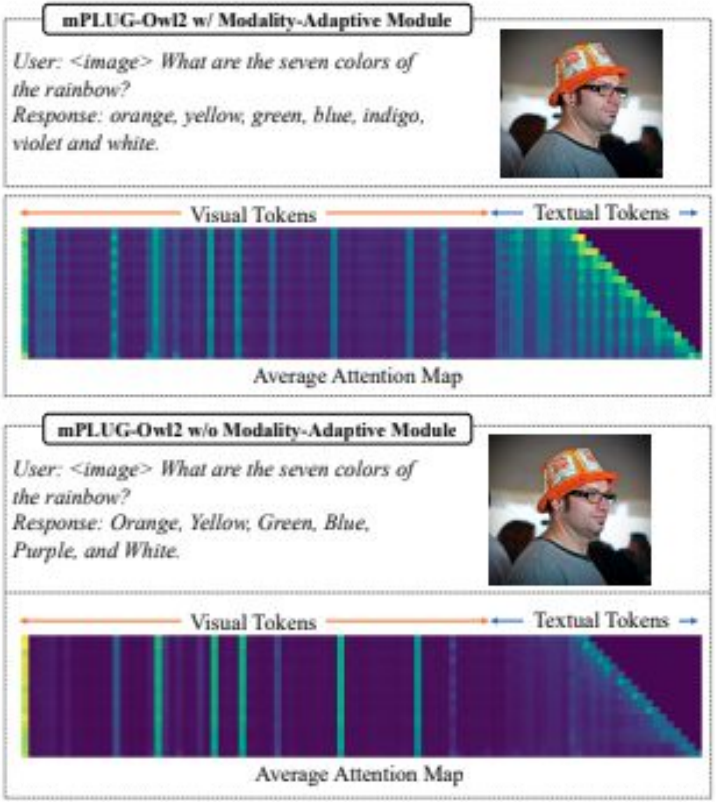
Training:

- Pretrain modalities separately
- Tune on joint instructions
- Keep vision encoder trainable across **both** stages



Results and Discussion

- Multi-modal Tasks: Outperforms other models in **captioning, VQA, and multi-modal** benchmarks, even in zero-shot scenarios.
- Text Tasks: **Demonstrates strong reasoning and language understanding, surpassing instruction-tuned LLMs**.
- Incorporation of MAM **offsets performance degradation** in text settings
- Number of Abtractor's **learnable queries** improves performance until saturation point
- Improved Robustness in Unrelated Modality Scenarios



MLLM 3: JanusFlow

Chapter 4

JanusFlow: Harmonizing Autoregression and Rectified Flow

- Unified Architecture: Combines **autoregressive language models with rectified flow** for both image understanding and generation.
- Decoupled Encoders: Separate vision encoders for understanding and generation to reduce task interference.
- Representation Alignment: Aligns intermediate representations between generation and understanding modules for enhanced semantic coherence.

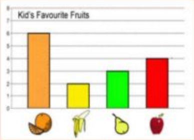
A corgi's head depicted as an explosion of a nebula, with vibrant cosmic colors like deep purples, blues, and pinks swirling around. The corgi's fur blends seamlessly into the nebula, with stars and galaxies forming the texture of its fur. Bright bursts of light emanate from its eyes, and faint constellations can be seen in the background, giving the image a surreal, otherworldly feel.

Beautiful surreal symbolism the mesmerizing vision of a Cleopatra Queen of Egypt, mesmerizing brown eyes, black hair and ethereal features, radiating celestial aura, super high definition, true lifelike color, perfect exposure, razor sharp focus, golden ratio, soft reflections, bokeh effect, fine art photography, cinematic compositing, authentic, professional.

A lone figure in dark robes ascends worn stone steps toward a glowing light in an ancient temple entrance. Ornate arches, lush greenery, and intricate carvings adorn the scene, evoking a mystical, high-fantasy atmosphere reminiscent of works by artists like Randy Vargas, with cinematic lighting and epic storytelling.

User: What are the kinds of fruits in this picture?

JaunsFlow (Ours): The fruits in the picture are banana, strawberry, mango, persimmon, blueberry, and lime.



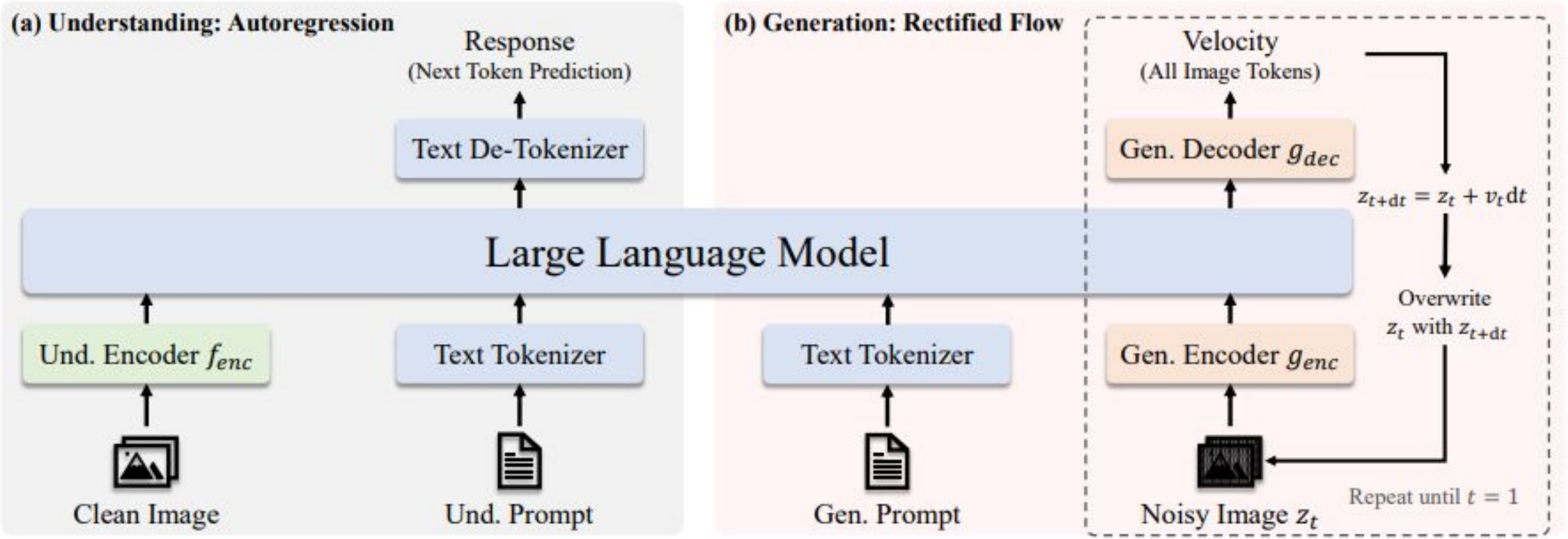
User: Describe this image.

JaunsFlow (Ours): Kid's Favourite Fruits is a bar graph. The x-axis shows the fruits. The y-axis shows the number of kids who like the fruit. The orange is the most popular fruit. The banana is the least popular fruit.

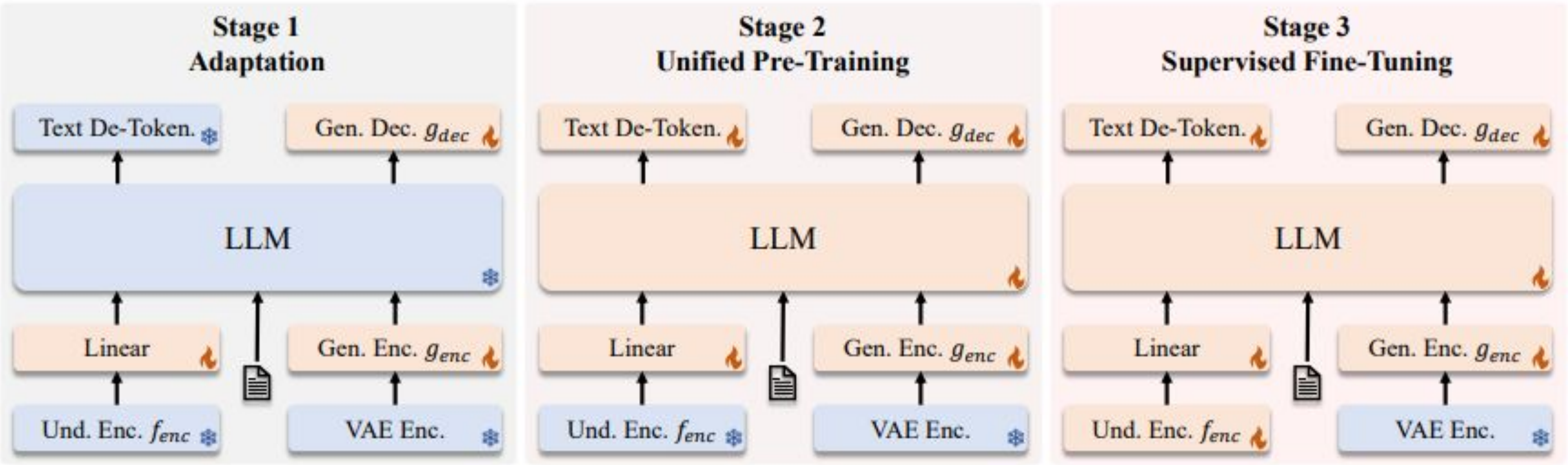
User: How many penguins in the picture?

JaunsFlow (Ours): There are 11 penguins in the picture.

JanusFlow Architecture

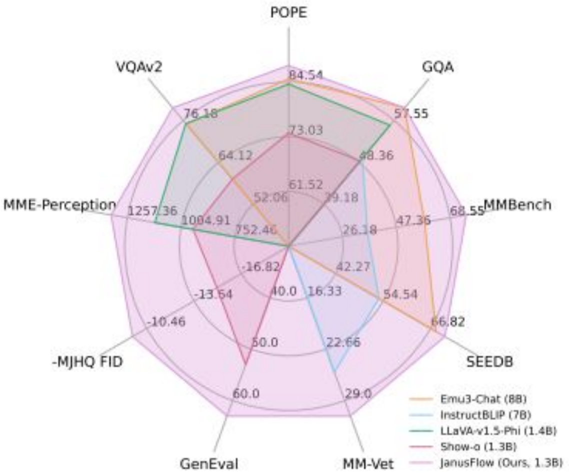


JanusFlow Training



Results

- Outperforms state-of-the-art unified and task-specific models on multimodal benchmarks (e.g., **MMBench, SEEDBench, GQA**).
- Superior image generation quality with an MJHQ FID-30k score of **9.51**, surpassing models like **SDXL and Janus**.



(a) Benchmark Performances.

(b) Visual Generation Results.

Penn Engineering

Conclusion

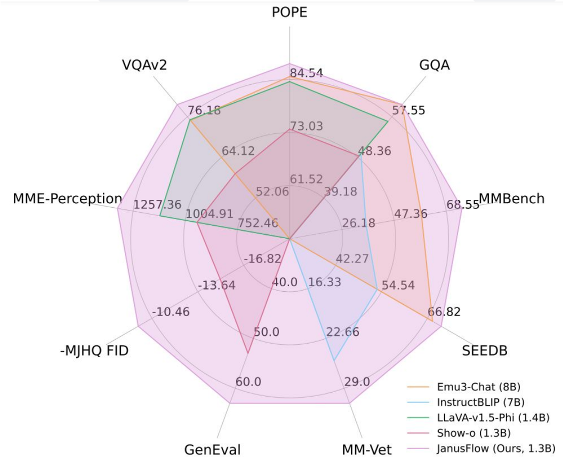
Chapter 5

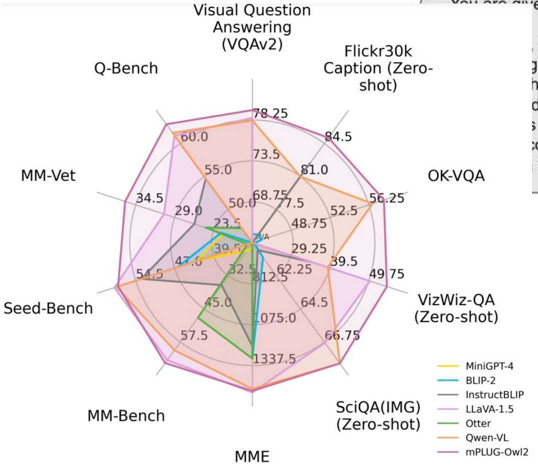
Summary of MLLMs

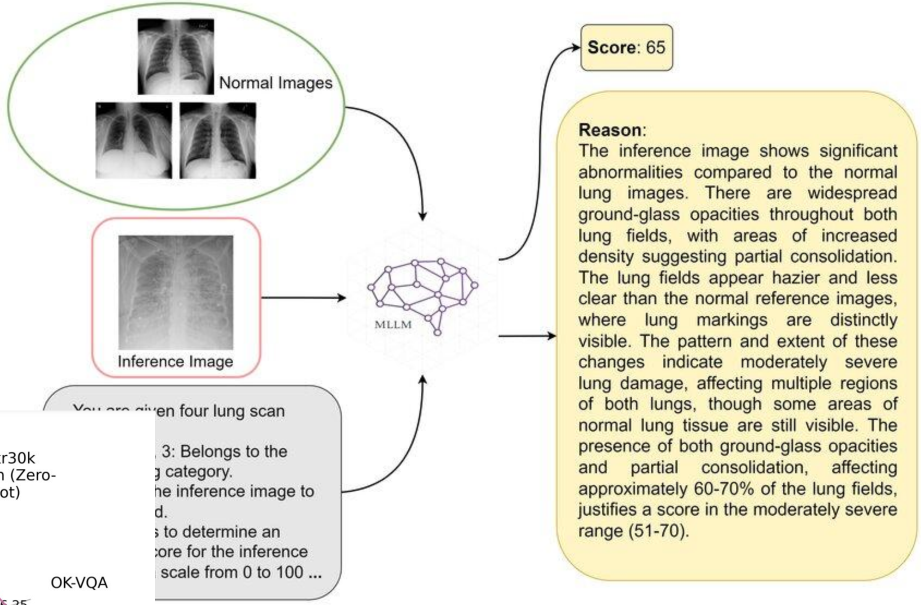
1. **Flamingo**: Visual language model designed for few-shot learning in vision and language tasks
2. **mPLUG-Owl2**: Leverages modality collaboration to improve performance in both text and multi-modal tasks
3. **JanusFlow**: Minimalist architecture that integrates autoregressive language models with rectified flow for both image understanding and generation

Connection Between Papers

1. Modality Integration
2. Unified Architecture
3. Architectural Innovations
4. Performance Benchmarks







Attributions for Assets Used

- https://en.wikipedia.org/wiki/File:Van_Gogh_-_Starry_Night_-_Google_Art_Project.jpg
- https://artsology.com/blog/2022/09/van-gogh-and-basketball-as-imagined-by-ai/
- https://medium.com/@tenyks_blogger/multimodal-large-language-models-mllms-transforming-computer-vision-76d3c5dd267f
- https://openai.com/index/hello-gpt-4o/
- http://dx.doi.org/10.48550/arXiv.2411.14515