



Monkey : Image Resolution and Text Label Are Important Things for Large Multi-modal Models

Zhang Li^{1†}, Biao Yang^{1†}, Qiang Liu², Zhiyin Ma¹, Shuo Zhang¹, Jingxu Yang², Yabo Sun²,
Yuliang Liu^{1*}, Xiang Bai^{1*}

¹Huazhong University of Science and Technology ²Kingsoft Office
ylliu@hust.edu.cn

CVPR 2024 (Highlight)

Industrial AI Research, POSCO DX
Susang Kim

Contents

1. Introduction
2. Related Works
3. Methods
4. Experiments
5. Conclusion

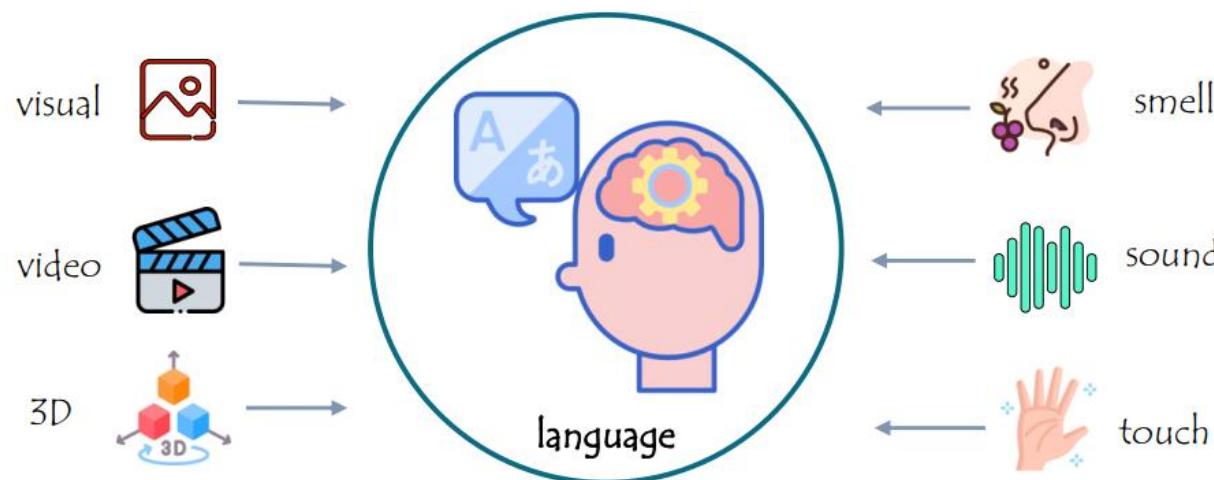
1. Introduction - Large Multimodal Models (LMMs)

Preliminary Idea: Language Intelligence as Pivot

Given this premise, nearly all CURRENT MLLMs are built based on language-based LLMs as the core decision-making module (i.e., the brain or central processor).

By adding additional external non-textual modality modules, LLMs are enabled with multimodal abilities.

- Extend the capability boundary, next milestone towards more advanced intelligence
- More applications



1. Introduction - LLM as Joint Part of System

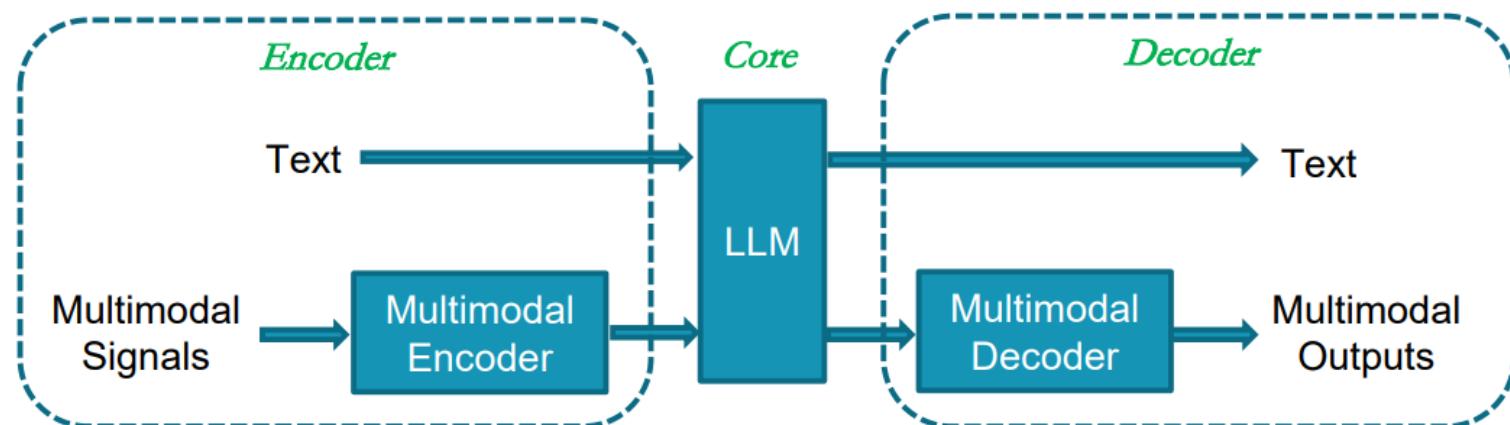
The role of the LLM is to perceive multimodal information, and react by itself, in an structure of Encoder-LLM-Decoder.

Key feature:

LLM is the key joint part of the system, receiving multimodal information directly from outside, and delegating instruction to decoders/generators in a more smooth manner.

Visual Encoder : CLIP-ViT is the most popular choice for vision-language models. (SigLIP)

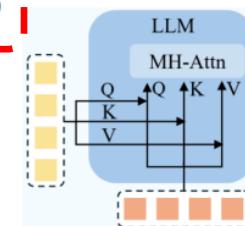
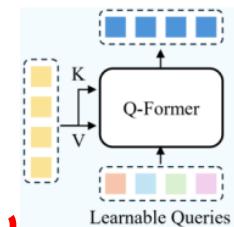
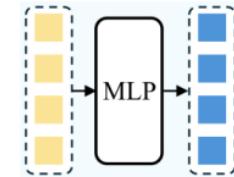
Non-Visual Encoder : Audio(Whisper, AudioCLIP, HuBERT, BEATs), 3D Point Cloud(Point-BERT)



1. Introduction – LMMs Input-side Projection

Methods to Connect Multimodal Representation with LLM

- + Projecting multimodal (e.g., image) representations into LLM semantic space
 - ✗ Q-Former: **BLIP-2, InstructBLIP, VisCPM, VisualGLM**
 - ✗ Linear projection: **LLaVA, MiniGPT-4, NExT-GPT**
 - ✗ Two-layer MLP: **LLaVA-1.5/NeXT, CogVLM, DeepSeek-VL, Yi-VL**
- + Perceiver Resampler: **Flamingo, Qwen-VL, MiniCPM-V, LLaVA-UHD**
- + C-Abstractor: **HoneyBee, MM1**

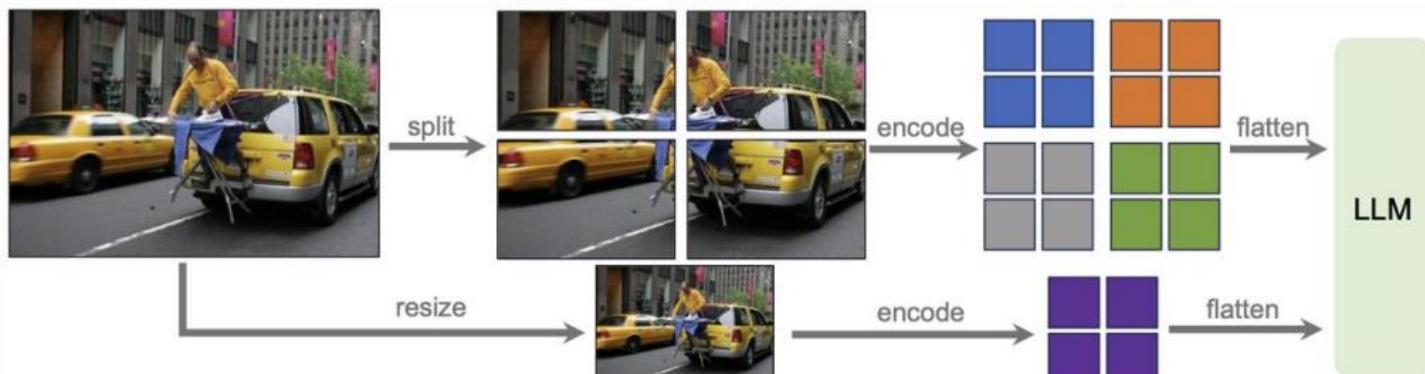


+Monkey

1. Introduction - Multimodal Encoding (Vision Encoder)

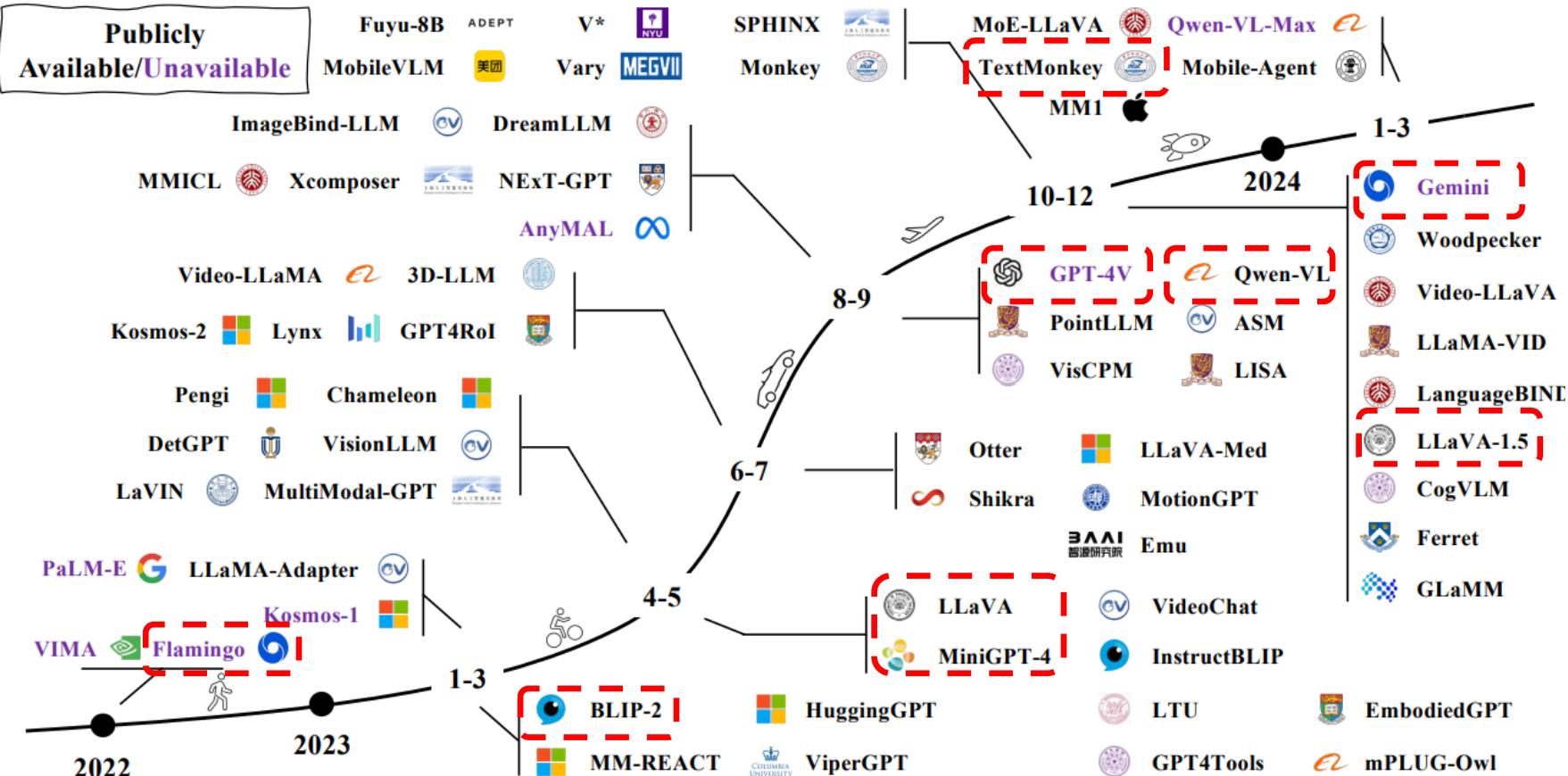
- **Visual Encoder**

- High-resolution Multimodal LLMs
 - ✗ Image slice-based: Split high-resolution images into slices
 - ✗ Representatives:
 - ◆ GPT-4V, LLaVA-NeXT, MiniCPM-V 2.0/2.5, LLaVA-UHD, mPLUG-DocOwl 1.5, SPHINX, InternLM-XComposer2-4KHD, **Monkey**



2.Related Works - MLLM

Related Research



2.Related Works - LLaVA & Qwen-VL (LLM & Vision Encoder)

TABLE 2: A summary of commonly used open-sourced LLMs. en, zh, fr, and de stand for English, Chinese, French, and German, respectively.

Model	Release Date	Pretrain Data Scale	Parameter Size (B)	Language Support	Architecture
Flan-T5-XL/XXL [56]	Oct-2022	-	3 / 11	en, fr, de	Encoder-Decoder
LLaMA [5]	Feb-2023	1.4T tokens	7 / 13 / 33 / 65	en	Causal Decoder
Vicuna [4]	Mar-2023	1.4T tokens	7 / 13 / 33	en	Causal Decoder
LLaMA-2 [57]	Jul-2023	2T tokens	7 / 13 / 70	en	Causal Decoder
Qwen [58]	Sep-2023	3T tokens	1.8 / 7 / 14 / 72	en, zh	Causal Decoder

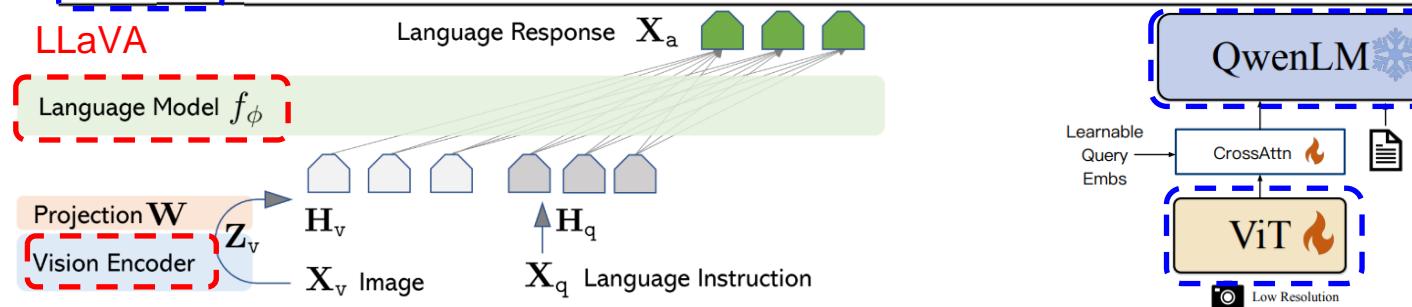


TABLE 1: A summary of commonly used image encoders.

Variants	Pretraining Corpus	Resolution	Samples (B)	Parameter Size (M)
OpenCLIP-ConvNext-L [46]	LAION-2B	320	29	197.4
CLIP-ViT-L/14 [13]	OpenAI's WIT	224/336	13	304.0
EVA-CLIP-ViT-G/14 [47]	LAION-2B, COYO-700M	224	11	1000.0
OpenCLIP-ViT-G/14 [46]	LAION-2B	224	34	1012.7
OpenCLIP-ViT-bigG/14 [46]	LAION-2B	224	34	1844.9

2.Related Works - Typical MLLM architecture

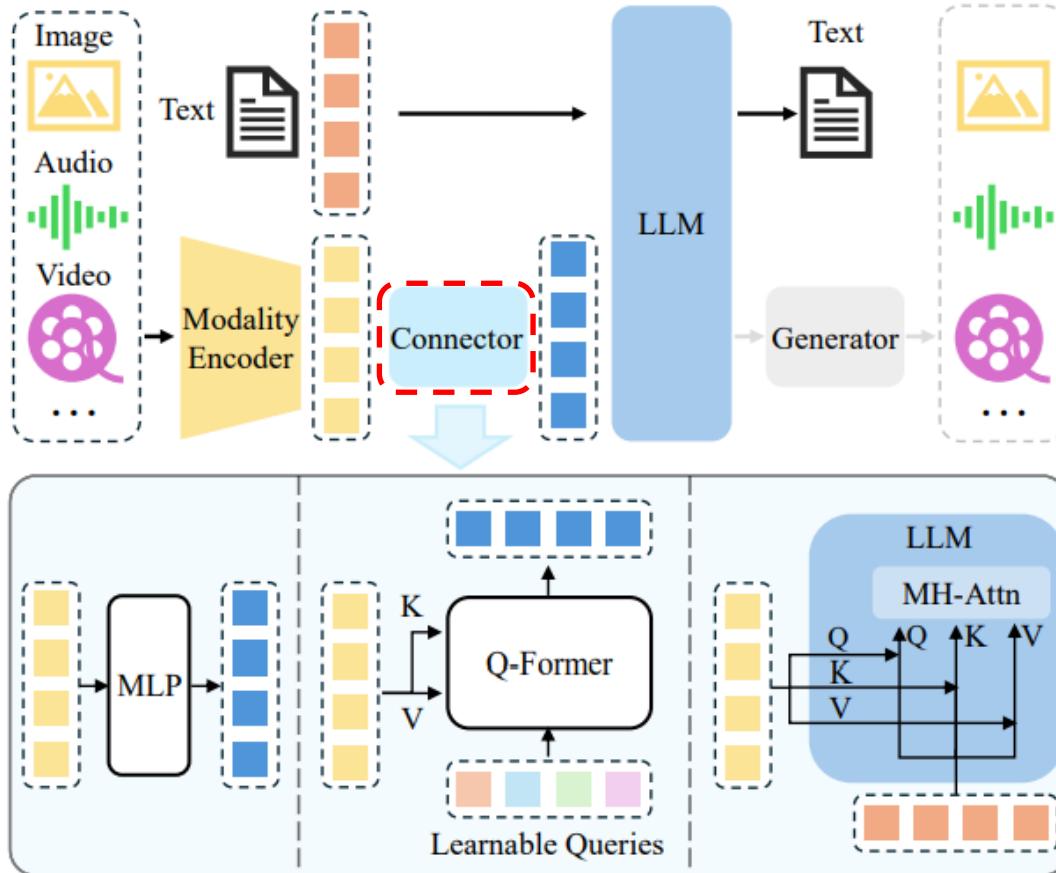


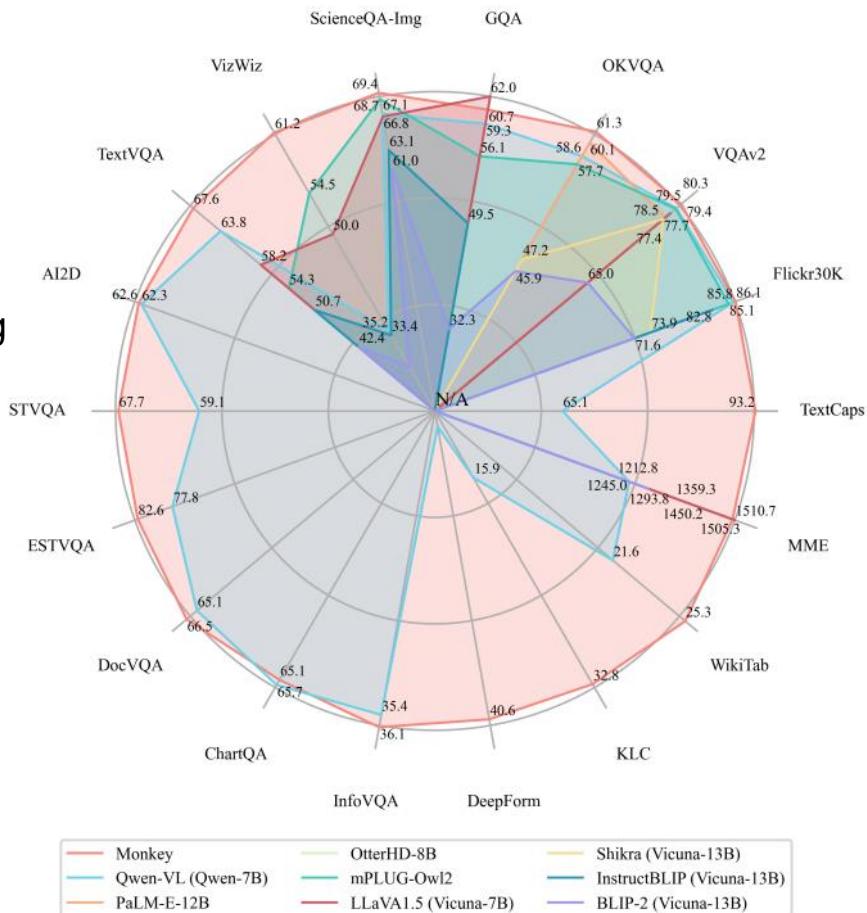
TABLE 4: Common datasets used for pre-training.

Dataset	Samples	Date
Coarse-grained Image-Text		
CC-3M [84]	3.3M	2018
CC-12M [85]	12.4M	2020
SBU Captions [86]	1M	2011
LAION-5B [87]	5.9B	Mar-2022
LAION-2B [87]	2.3B	Mar-2022
LAION-COCO [88]	600M	Sep-2022
COYO-700M [90]	747M	Aug-2022
Fine-grained Image-Text		
ShareGPT4V-PT [83]	1.2M	Nov-2023
LVIS-Instruct4V [91]	111K	Nov-2023
ALLaVA [92]	709K	Feb-2024
Video-Text		
MSR-VTT [93]	200K	2016
Audio-Text		
WavCaps [94]	24K	Mar-2023

3.Methods - Monkey (Large Multimodal Model (LMM))

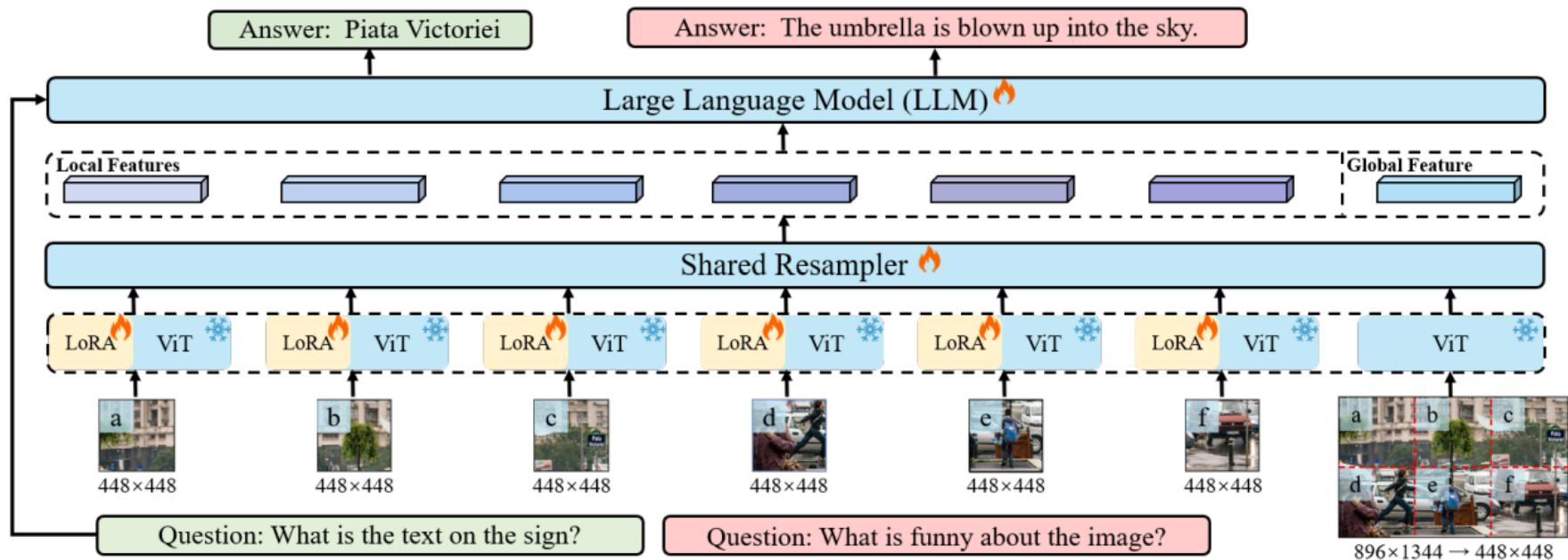
The performance of Monkey on a broad range of multimodal tasks compared with existing models.
(Monkey has shown promising results, comparing with GPT4V.)

- 1) **Higher resolutions** up to 1344x896 pixels, enabling the detailed capture of complex visual information.(the usual 448x448 resolution used in LMMs.)
- 2) It employs a **multi-level description generation** method, enriching the context for scene-object associations.
- 3) 18 datasets further demonstrate that Monkey **surpasses existing LMMs** in many tasks like Image Captioning and various Visual Question Answering.



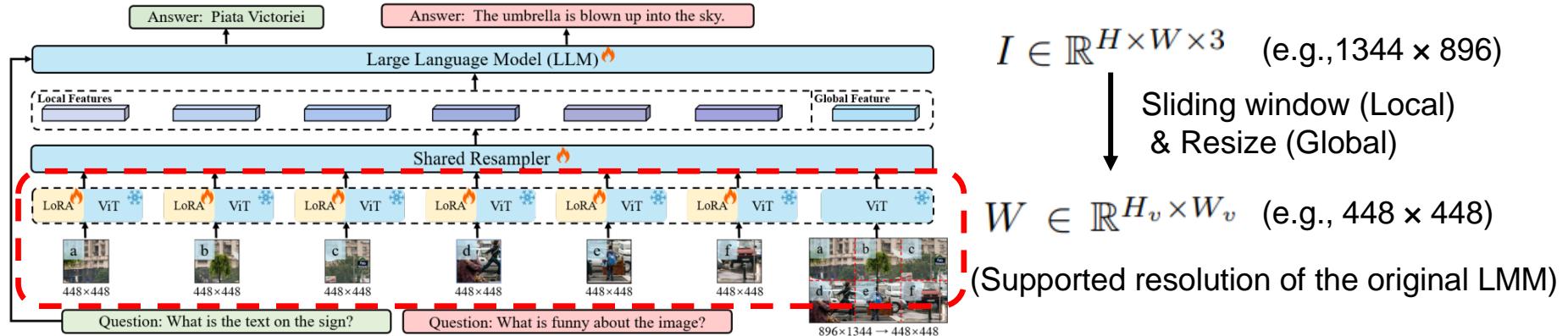
3.Methods - The overall architecture of Monkey

Monkey utilizes a new module that divides high-resolution images into smaller patches using a sliding window method. Each patch is processed independently by a static visual encoder, enhanced with LoRA adjustments and a trainable visual resampler.

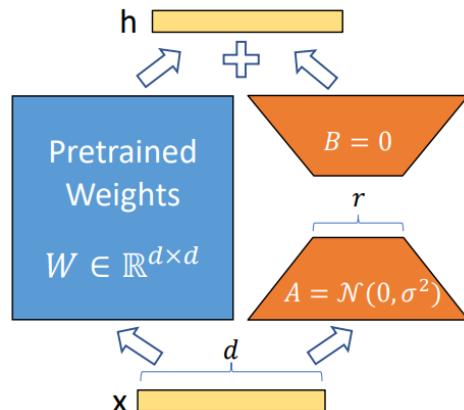


It enables high resolution by capturing global feature from original image and local features from divided patches. All patches are processed through the shared static ViT encoder, such as Vit-BigG with 2b parameters

3.Methods – 1) Enhancing Input Resolution



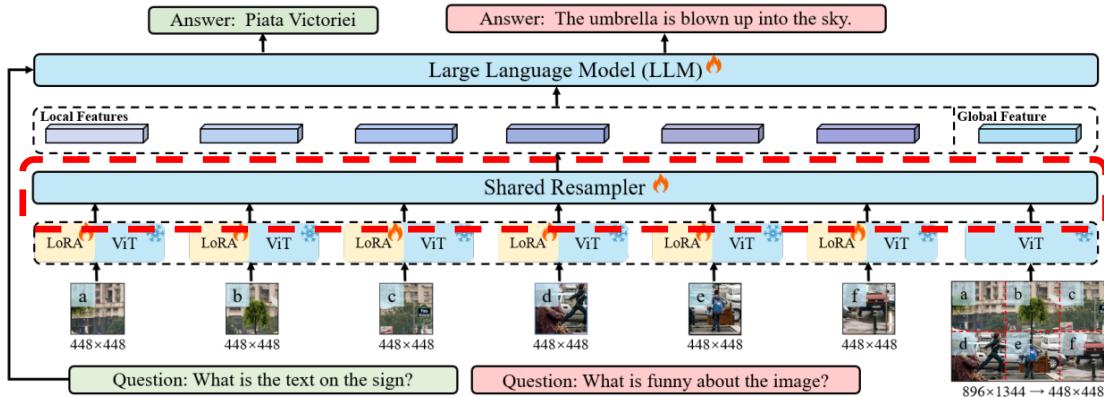
Each patch is processed independently by a static visual encoder + LoRA



LoRA freezes the pretrained model weights and **injects trainable rank decomposition matrices** into each layer of the Transformer architecture, significantly reducing the number of trainable parameters for downstream tasks.

$$h = W_0x + \Delta Wx = W_0x + BAx$$

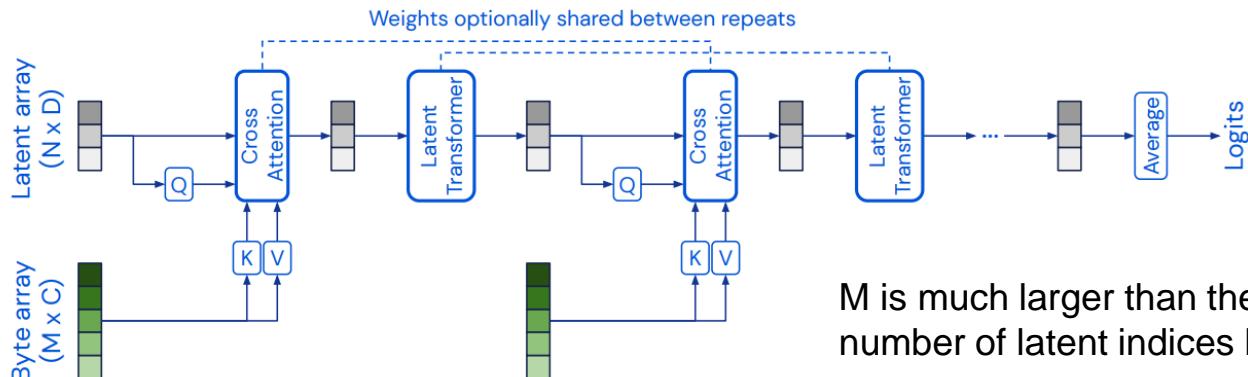
3.Methods – 1) Enhancing Input Resolution



This approach strikes a balance between detailed and holistic perspectives of the images, thereby enhancing the model performance while avoiding a substantial increase in computational demand.

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

Trainable vectors



3.Methods – 1) Enhancing Input Resolution - The Perceiver Resampler

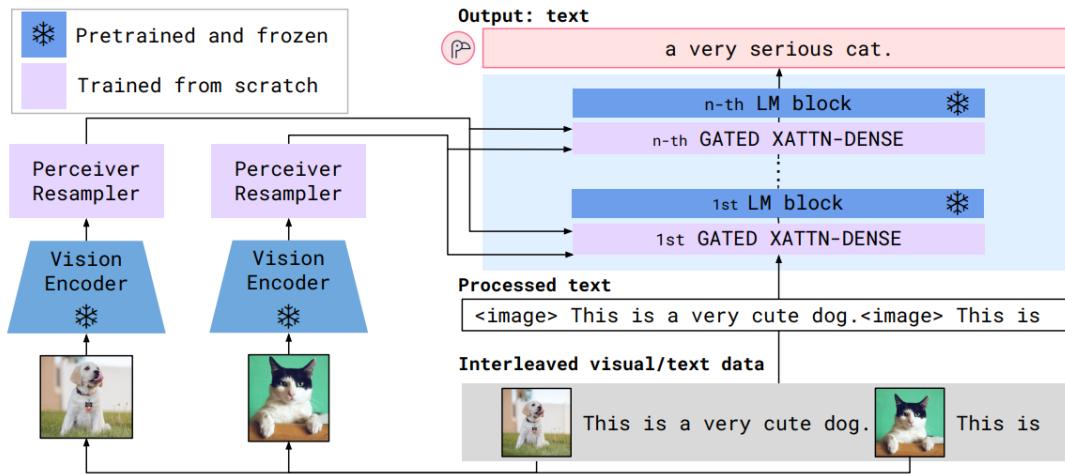
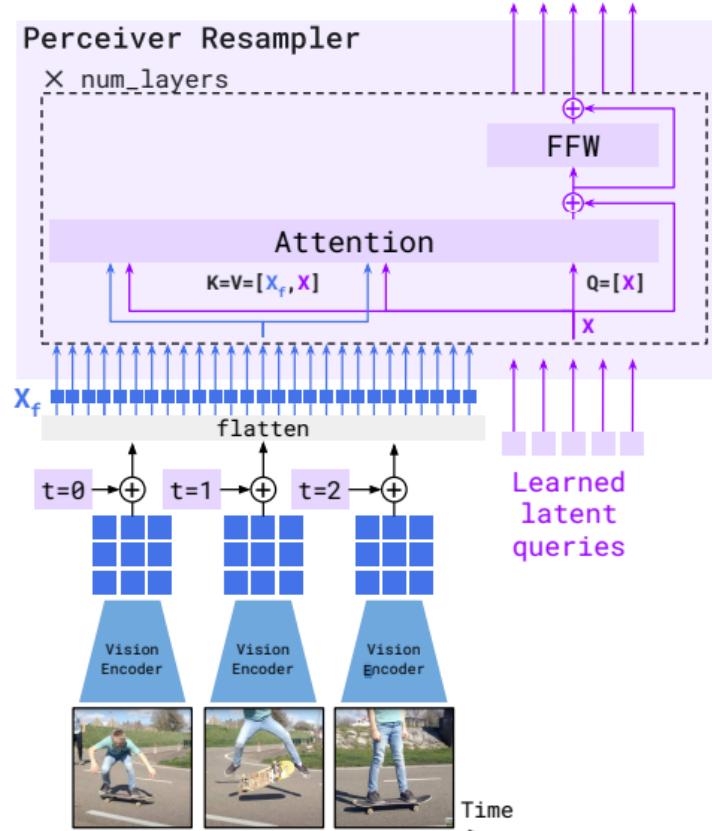


Figure 3: **Flamingo architecture overview**. Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

The Perceiver Resampler module transforms a variable-sized grid of spatio-temporal **visual features** from the **Vision Encoder** into a **fixed number of output tokens**, regardless of the input image resolution or video frame count. It uses **learned latent vectors as queries**, while the keys and values are a combination of the spatio-temporal features and the latent vectors.



3.Methods - 2) Multi-level Description Generation

LLaVA and Qwen-VL were trained on datasets that often contain image-text pairs that are too simple.

Language	Dataset	Original	Cleaned	Remaining%
English	LAION-en	2B	280M	14%
	LAION-COCO	600M	300M	50%
	DataComp	1.4B	300M	21%
	Coyo	700M	200M	28%
	CC12M	12M	8M	66%
	CC3M	3M	3M	100%
	SBU	1M	0.8M	80%
	COCO Caption	0.6M	0.6M	100%
Chinese	LAION-zh	108M	105M	97%
	In-house Data	220M	220M	100%
Total		5B	1.4B	28%

Details of Qwen-VL pre-training data



by Danail Nachev



by Justin Higuchi

the stars in the night sky.

musical artist performs on stage during festival.

(e.g., CC3M : one short sentence to describe a complicated image)

3.Methods - 2) Multi-level Description Generation

GPT-assisted Visual Instruction Data Generation. (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>

We collect 158K unique language-image instruction-following samples in total.

58K in conversations

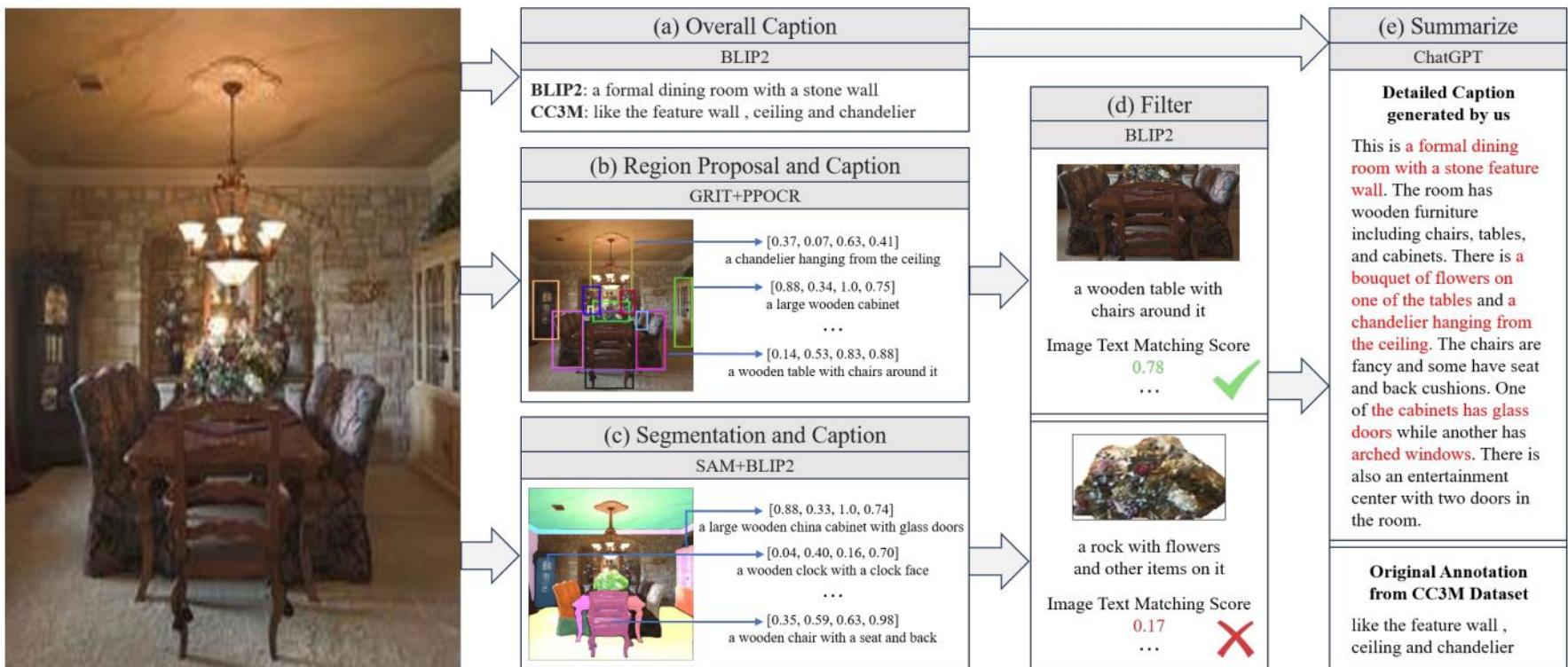
23K in detailed description

77k in complex reasoning

3.Methods - 2) Multi-level Description Generation

The pipeline for multi-level description generation for images. (**BLIP2** to check for filtering out low-scoring matches)

- **Q-former** (integration with the vision encoder and LLM), **GRIT** (region-to-text model),
- **PPOCR** (extracts text from the images), **SAM** (segments and identifies objects and their parts)



3.Methods - 2) Multi-level Description Generation

Task	Dataset	Smaples
Image Caption	Detailed Caption	213k
	COCO Caption [22]	82k
	TextCaps [45]	109k
General VQA	VQAV2 [16]	100k
	OKVQA [34]	18k
	GQA [20]	150k
	ScienceQA [33]	18k
	VizWiz [17]	20k
Scene Text-centric VQA	TextVQA [46]	34k
	OCRVQA [38]	250k
	AI2D [23]	24k
Doc-oriented VQA	DocVQA [36]	118k
	ChartQA [35]	84k
	InfoVQA [37]	47k
	DeepForm [48]	7k
	KLC [47]	27k
	WTQ [41]	28k
	TabFact [8]	91k
	VisualMRC [49]	21k
Total	-	1.44m

“Generate the caption in English:”



“Generate the detailed caption in English:”



“{question} Answer: {answer}.”

Is the umbrella upside down?
yes
no



Who is wearing glasses?
man
woman



Where is the child sitting?
fridge
arms



How many children are in the bed?
2
1



Figure 1: Examples from our balanced VQA dataset.

2007 Ig Nobel Prize winners announced

Friday, October 5, 2007

The winners of the 2007 Ig Nobel Prize have been announced. The awards, given out every early October since 1991 by the Annals of Improbable Research, are a parody of the Nobel Prize, which are awards given out in several fields. The awards are given to achievements that, “first make people laugh, and then make them think.” They were presented at Harvard University’s Sanders Theater.

Ten awards have been presented, each given to a different field. The winners are:

- **Medicine:** Brian Witcombe, of Gloucestershire Royal NHS Foundation Trust, UK, and Dan Meyer, who studied the health consequences of **sword swallowing**.
- **Physics:** A team from the USA and Chile, who made a study about how cloth sheets become wrinkled.
- **Biology:** Dr Johanna van Bronswijk of the **Netherlands**, for carrying out a census of creatures that live in people’s beds.
- **Chemistry:** Mayu Yamamoto, from Japan, for creating a method of extracting vanilla fragrance and flavouring from **cow dung**.



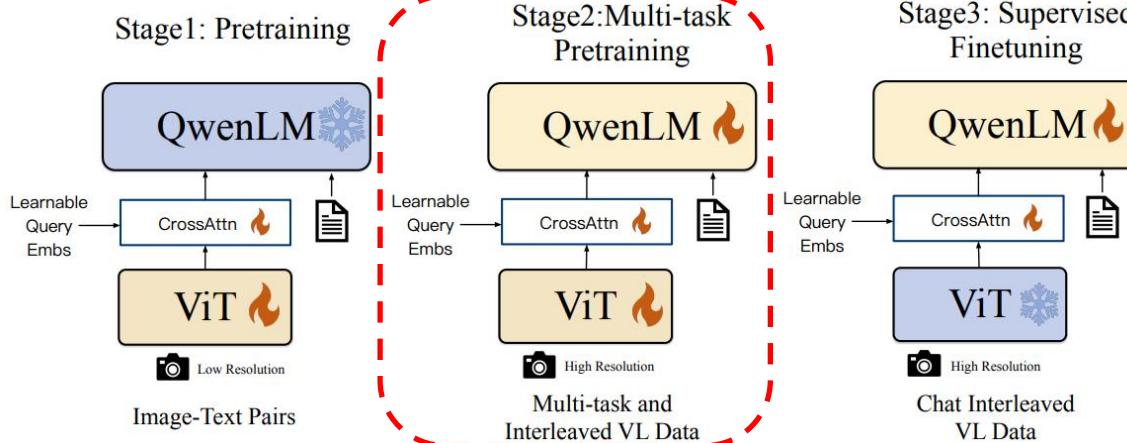
The 2007 Ig Nobel Prize in aviation went to a team from an Argentinian university, who discovered that impotency drugs can help hamsters recover from jet lag.

Q: Who were the winners of the Ig Nobel prize for Biology and Chemistry?

A: The winner of the Ig Nobel prize for biology was Dr Johanna van Bronswijk, and the winner for Chemistry was Mayu Yamamoto.

Details on the Monkey training data

3.Methods – 3) Multi-task Pre-training



By integrating diverse datasets and applying consistent instructions across all tasks, as guided by QwenLM, we enhance both the model's learning ability and training efficiency.

Task	# Samples	Dataset
Captioning	19.7M	LAION-en & zh, DataComp, Coyo, CC12M & 3M, SBU, COCO, In-house Data
VQA	3.6M	GQA, VGQA, VQAv2, DVQA, OCR-VQA, DocVQA, TextVQA, ChartQA, AI2D
Grounding ²	3.5M	GRIT
Ref Grounding	8.7M	GRIT, Visual Genome, RefCOCO, RefCOCO+, RefCOCOg
Grounded Cap.	8.7M	GRIT, Visual Genome, RefCOCO, RefCOCO+, RefCOCOg
OCR	24.8M	SynthDoG-en & zh, Common Crawl pdf & HTML
Pure-text Autoregression	7.8M	In-house Data

Details of Qwen-VL multi-task pre-training data.

4.Experiments - Model Configuration

[Model Configuration]

Vision Encoder : Vit-BigG(CLIP)

LLM : QwenVL

H_v, H_w : 448 x 448

Learnable queries : 256

Images of size : 896 x 896

Vision encoder : 1.9B

LoRA : 16 for the attention module 32 for MLP (117M parameters)

Large language model : 7.7B parameters

The resampling module : 90M parameters

Overall parameters for Monkey : 9.8B

[Training]

Learning rate : 1e-5 ,

AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.95$)

Cosine learning rate schedule

Warmup period of 100 steps

Batch size : 1024

Weight decay : 0.1

Whole training process : 40 A800 days for one epoch

4.Experiments – MLLM Evaluation benchmark MME

Perception (Coarse-Grained Tasks)

Existence



[Y] Is there a **elephant** in this image?

[N] Is there a **hair drier** in this image?



[Y] Is there a **refrigerator** in this image?

[N] Is there a **donut** in this image?

Poster



[Y] Is this movie directed by **francis ford coppola**?

[N] Is this movie directed by **franklin j. schaffner**?



[Y] Is this movie titled **twilight (2008)**?

[N] Is this movie titled the **horse whisperer (1998)**?

OCR



[Y] Is the phone number in the picture "**0131 555 6363**"?

[N] Is the phone number in the picture "**0137 556 6363**"?



[Y] Is the word in the logo **"high time coffee shop"**?

[N] Is the word in the logo **"high tite cofeee shop"**?

Commonsense Reasoning



[Y] Should I **stop** when I'm about to **cross** the street?

[N] When I see the sign in the picture, can I **cross** the street?

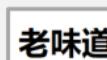


[Y] Is there **one** real cat in this picture?

[N] Is there **two** real cats in this picture?

Cognition (Reasoning Tasks)

Text Translation



[Y] Appropriate to translate into English '**classic taste**'?

[N] Appropriate to translate into English '**strawberry flavor**'?

共同努力

[Y] Appropriate to translate into English '**work hard together**'?

[N] Appropriate to translate into English '**be filled with intrigue**'?

MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models (total of 14 subtasks)

Rank	Model	Score
1	WeMM	1621.66
2	InfMLLM	1567.99
3	SPHINX	1560.15
4	Lion	1545.80
5	LLaVA	1531.31
6	XComposer-VL	1528.45
7	Qwen-VL-Chat	1487.58
8	mPLUG-Owl2	1450.20
9	Skywork-MM	1419.08
10	GPT-4V	1409.43

(1) Perception

4.Experiments – MLLM Evaluation benchmark MME

Model	Image Caption		General VQA				
	Flickr30K	TextCaps	VQAv2	OKVQA	GQA	ScienceQA	VizWiz
Flamingo-80B [1]	67.2	-	56.3	50.6	-	-	31.6
Palm-E-12B [13]	-	-	77.7	60.1	-	-	-
BLIP-2 (Vicuna-13B) [27]	71.6	-	65.0	45.9	32.3	61.0	19.6
InstructBLIP (Vicuna-13B) [12]	82.8	-	-	-	49.5	63.1	33.4
Shikra (Vicuna-13B) [7]	73.9	-	77.4	47.2	-	-	-
mPLUG-Owl2 [54]	85.1	-	79.4	57.7	56.1	68.7	54.5
LLaVA1.5 (Vicuna-7B) [28]	-	-	78.5	-	62.0	66.8	50.0
Qwen-VL(Qwen-7B) [3]	85.8	65.1	79.5	58.6	59.3	67.1	35.2
Qwen-VL-Chat [3]	81.0	-	78.2	56.6	57.5	68.2	38.9
Monkey	86.1	93.2	80.3	61.3	60.7	69.4	61.2

Table 2. Results on Image Caption and General VQA.

Model	TextVQA	AI2D	STVQA	ESTVQA
Pix2Struct-Large [25]	-	42.1	-	-
BLIP-2 [27]	42.4	-	-	-
InstructBLIP [12]	50.7	-	-	-
mPLUG-DocOwl [53]	52.6	-	-	-
mPLUG-Owl2 [54]	54.3	-	-	-
Qwen-VL [3]	63.8	62.3	59.1	77.8
Qwen-VL-Chat [3]	61.5	57.7	-	-
LLaVA-1.5 [28]	58.2	-	-	-
Monkey	67.6	62.6	67.7	82.6

Table 3. Results on Scene Text-centric VQA.

Model	DocVQA	ChartQA	InfoVQA	DeepForm	KLC	WTQ
Qwen-VL	65.1	65.7	35.4	4.1	15.9	21.6
Monkey	66.5	65.1	36.1	40.6	32.8	25.3

Table 4. Results on Doc-oriented VQA.

4.Experiments - Ablation Study

	Resolution	LoRA	Throughout	FLOPS (e20)	VQAv2	GQA	TextVQA	STVQA	DocVQA	DeepForm	InfoVQA	WTQ
r1	896×896*	0	43.452	1.608	74.1	55.2	44.7	41.5	53.9	11.4	32.7	16.8
r2	896×896*	1	37.429	1.614	71.4	54.0	41.7	38.5	47.5	7.2	31.5	17.1
r3	672×672	4	43.604	1.617	80.0	59.6	67.3	67.2	66.4	31.3	35.9	25.0
r4	784×784	4	42.851	1.617	79.9	59.8	67.5	67.7	66.5	38.9	35.5	25.1
r5	896×1344	6	28.542	1.622	80.1	61.1	67.3	66.7	66.3	42.3	39.6	26.6
r6	1344×896	6	28.842	1.622	80.2	61.8	67.7	66.3	64.5	41.4	35.7	25.2
r7	896×896	0	49.634	1.613	80.1	60.4	67.5	65.1	66.1	36.8	36.1	24.9
r8	896×896	1	42.885	1.614	80.0	60.3	67.6	67.0	66.7	36.9	36.5	24.7
r9	896×896	4	42.542	1.617	80.3	60.7	67.6	67.7	66.5	40.6	36.1	25.3

Table 5. Ablation study on enhancing input resolution and the number of trainable adapters using Qwen-VL (originally trained using 448×448). * refers to directly scaling the input size of the visual encoder from 448 to 896 using traditional positional position interpolation.

Res.	PT	GQA	TextVQA	MMVet
224	CC3M	62	56.1	33.2
224	Ours	62.1(+0.1)	56.3(+0.2)	33.7(+0.5)
336	CC3M	63.4	59.8	33.5
336	Ours	63.7 (+0.3)	60.4(+0.6)	36.1 (+2.6)
448	CC3M	64.3	60.2	33.6
448	Ours	64.6 (+0.3)	62.0 (+1.8)	36.2 (+2.6)

Table 6. Ablation study on LLaVA1.5. “Res.” denotes resolution. “PT” refers to pretrain data.

Increase the input size, there is a noticeable boost in performance. Reducing the LoRA number causes a performance decrease.

[LLaVa 1.5]
 Visual encoder : ViT-L/14
 LLM : Vicuna

4.Experiments - On some testing samples with dense text

Monkey has demonstrated impressive performance when compared to GPT4V.

(a)



What is the net weight of the product in the picture?

净含量: 125ml

125ml



The net weight of the product shown in the picture is 250ml.

(c)



What is the net weight of the product in the picture?

100g

100g



I'm sorry, but I'm unable to provide the net weight of the product in the picture as the text is not clearly readable.

4.Experiments - Dense text

(d)



Where is the BRISTO PHOTOGRAPHY?

3rd floor



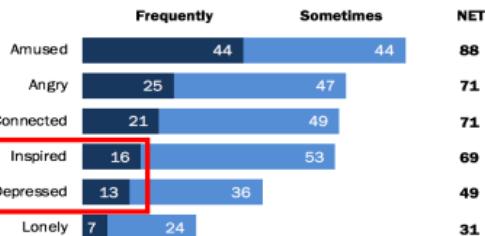
“Bristo Photography” is located on the second floor according to the directory sign in the image you've provided.



(e)

Social media users experience a mix of positive, negative emotions while using these platforms

% of social media users who say they ____ see content on social media that makes them feel ...



Note: Respondents who did not give an answer or gave other answers are not shown.

Source: Survey of U.S. adults conducted May 29-June 11, 2018.

“Public Attitudes Toward Computer Algorithms”

PEW RESEARCH CENTER

Inspired

16

Depressed

13



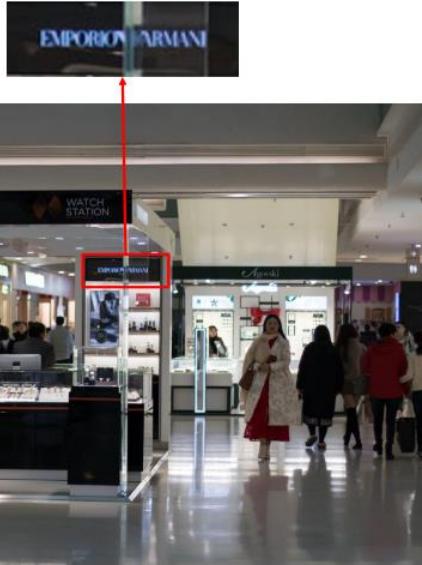
How many more people felt inspired frequently than depressed frequently?

0.03



.....Therefore, 3% more people felt inspired frequently than depressed frequently.

4.Experiments - Comparisons with LMMs on Detailed Caption task.



Q: Generate the detailed caption in English.

Ours: Image shows a shopping mall with various stores and people walking around. A Michael Kors store is visible, as well as a watch station and a store selling Emporio Armani. A woman in a red dress and white coat is walking through the mall, while another woman in a red coat and black pants is carrying a black purse. A black and white photo of a woman wearing sunglasses is also visible.



Qwen-VL-Chat: A shopping mall with people walking around and shopping.

LLaVA1.5: The image depicts a busy shopping mall with a large group of people walking around and browsing the stores. There are at least 14 people visible in the scene, some of them carrying handbags. A Michael Kors store is prominently featured in the mall, with a large display of handbags and other accessories.

In addition to the Michael Kors store, there are several other stores in the mall, including a jewelry store and a watch store. The mall is well-lit, and the atmosphere is lively as people move about and explore the various shops.



GPT4V : This image depicts a scene inside a modern shopping mall where various luxury brand stores are visible. In the foreground is a Michael Kors store showcasing an advertisement with an image of a woman wearing sunglasses, accompanied by an array of handbags on shelves. Adjacent to the Michael Kors store is a store named "WATCH STATION" featuring a selection of watches. In the background, we see the Emporio Armani store, which also displays luxury watches as part of its merchandise. Consumers are walking through the mall, browsing and shopping, with a few individuals standing and talking. The lighting is bright and highlights the products on display. The overall atmosphere is one of a clean, high-end shopping environment with a focus on fashion and accessories.



Monkey went further in detailing various elements in the scene, such as describing “another woman in a red coat and black pants carrying a black purse”

5. Conclusion & Limitation

- (+) Improve the input **resolution capacity up to 1344×896 pixels** without pretraining from the start.
- (+) To bridge the gap between simple text labels and high input resolution, we propose a **multi-level description generation method**, which automatically provides rich information that can guide the model to learn the contextual association between scenes and objects.
- (+) Our model demonstrates **promising performance** in image captioning by paying attention to textual information and capturing fine details within the images; its improved input resolution also enables remarkable performance in document images with dense text.
- (-) Input images is constrained to a maximum of six patches(due to the limited input length of the language model)
- (-) Multi-level description generation approach, it is capable of describing only the scene presented in the image and its scope is bound by the world knowledge encapsulated in BLIP2 and the original CC3M annotations.
- (-) How about making the patch crop adaptive instead of fixed-sized?

Monkey Demo : <http://vlrlab-monkey.xyz:7681/>



The photo depicts a man standing to his left side of an advertisement board for "IJCAI EU20+24 International Artificial Intelligence Conference" held on August 3-9 at Jeju Island South Korea with various logos and sponsors listed below it such as Oppo Didi Huawei COMebo Openbayes among others.

He is wearing glasses along with casual attire consisting of green short-sleeved shirt tucked into blue jeans paired white sneakers that have dark soles visible above them. He has black hair styled upwards slightly towards one temple and wears what appears like silver earrings or studs just behind each earlobe.

On top right corner there's another smaller signboard indicating some information about IJCAI which seems related due its proximity to main banner. The background shows part of indoor architecture typical of modern buildings; glass doors are partially seen reflecting light while ceiling lights illuminate overhead areas creating bright spots around edges where they meet walls/ceiling junctions.



How many company logos can you find?



In the image, several companies' logo designs appear in different sizes across both banners advertising IJCAI EU20+24 conference events from years past up until this year's event scheduled next month. These include recognizable brands associated with technology and artificial intelligence sectors represented by their respective symbols—such as 'OPPO', 'DIDI', etc.—which suggest these were major partners or supporters during those conferences.



ChatGPT



It looks like you're attending the **33rd International Joint Conference on Artificial Intelligence**

(IJCAI) 2024 in Jeju Island, South Korea! That's exciting! How has the conference been so far?

Are there any talks or sessions you're looking forward to?



Here are the company and sponsor logos I can identify from the banner:

1. MEGVII (also known as Face++)

2. Kyung Hee University

3. Meitu

4. Oppo

5. Vivo

6. Baidu

7. Huawei

8. Didi

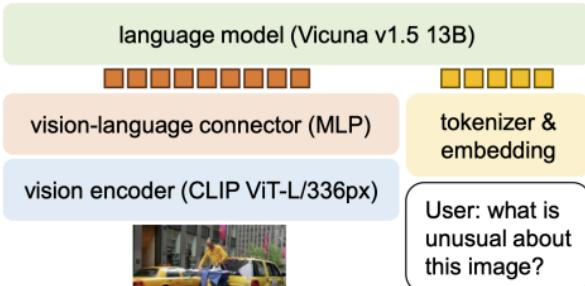
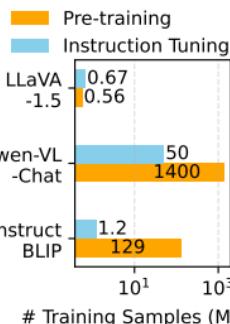
9. JD.com

10. Openbayes

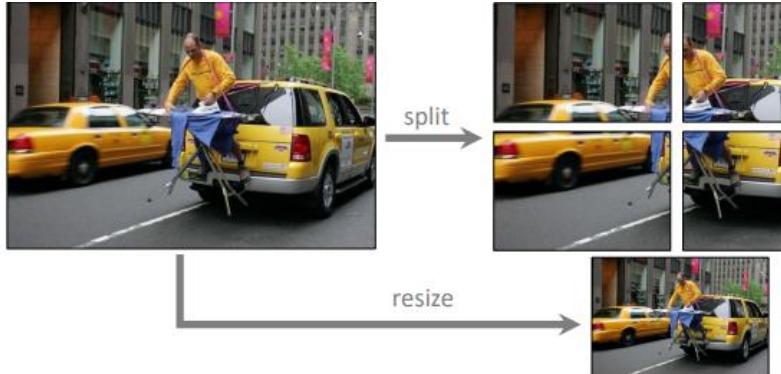


That makes a total of **10 company logos** on the banner.

LLaVA-1.5 (CVPR 2024)



Qwen-VL [3]	Qwen-7B	448 ²	1.4B [†]	50M [†]	78.8*	59.3*
Qwen-VL-Chat [3]	Qwen-7B	448 ²	1.4B*	50M [†]	78.2*	57.5*
LLaVA-1.5	Vicuna-7B	336 ²	558K	665K	78.5*	62.0*
LLaVA-1.5	Vicuna-13B	336 ²	558K	665K	80.0*	63.3*
LLaVA-1.5-HD	Vicuna-13B	448 ²	558K	665K	81.8*	64.7*



Method	LLM	Res.	GQA	MME	MM-Vet
InstructBLIP	14B	224	49.5	1212.8	25.6
<i>Only using a subset of InstructBLIP training data</i>					
0 LLaVA	7B	224	–	809.6	25.5
1 +VQA-v2	7B	224	47.0	1197.0	27.7
2 +Format prompt	7B	224	46.8	1323.8	26.3
3 +MLP VL connector	7B	224	47.3	1355.2	27.8
4 +OKVQA/OCR	7B	224	50.0	1377.6	29.6
<i>Additional scaling</i>					
5 +Region-level VQA	7B	224	50.3	1426.5	30.8
6 +Scale up resolution	7B	336	51.4	1450	30.3
7 +GQA	7B	336	62.0*	1469.2	30.7
8 +ShareGPT	7B	336	62.0*	1510.7	31.1
9 +Scale up LLM	13B	336	63.3*	1531.3	36.1

Table 2. Scaling results on data, model, and resolution.

Mini-Monkey: Multi-Scale Adaptive Cropping (NeurIPS 2024 Submission '24.08.09)

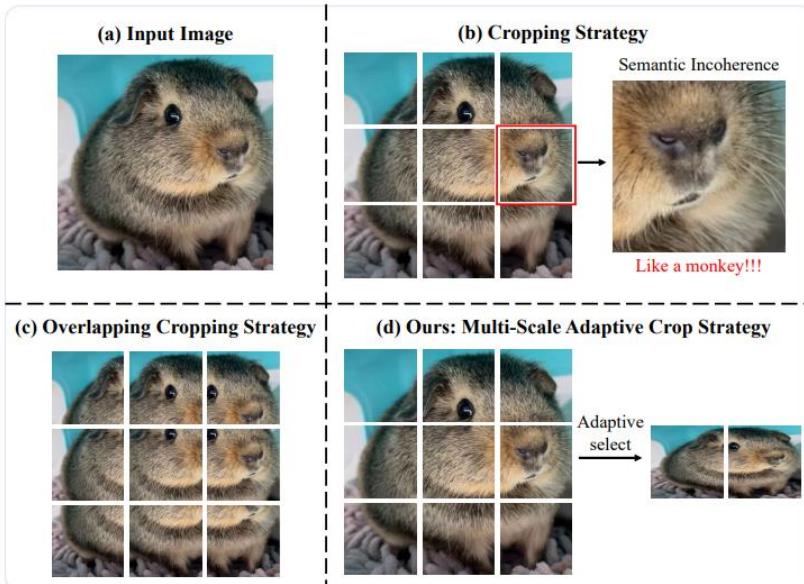


Figure 1: Sawtooth Effect caused by the cropping. (a) Input Image (b) Cropping strategy. (c) Overlapping Cropping Strategy. (d) Ours: Multi-scale adaptive cropping strategy.

Model	Resolution Strategy	TextVQA
Baseline	Dynamic High-Resolution strategy [8]	73.4
Baseline	Fixed Size High-Resolution strategy [37]	74.2
Baseline	Overlapping Cropping Strategy	70.6
Baseline	Multi-Scale Strategy [60]	74.8
Mini-Monkey (Ours)	Multi-Scale Adaptive cropping strategy	75.7

Mini-Monkey is efficient that our method can be trained using **only eight RTX 3090** and employs a **multiscale adaptive cropping strategy (MSAC)**. **Scale Compression Mechanism (SCM)** to reduce the computational overhead by compressing image tokens.

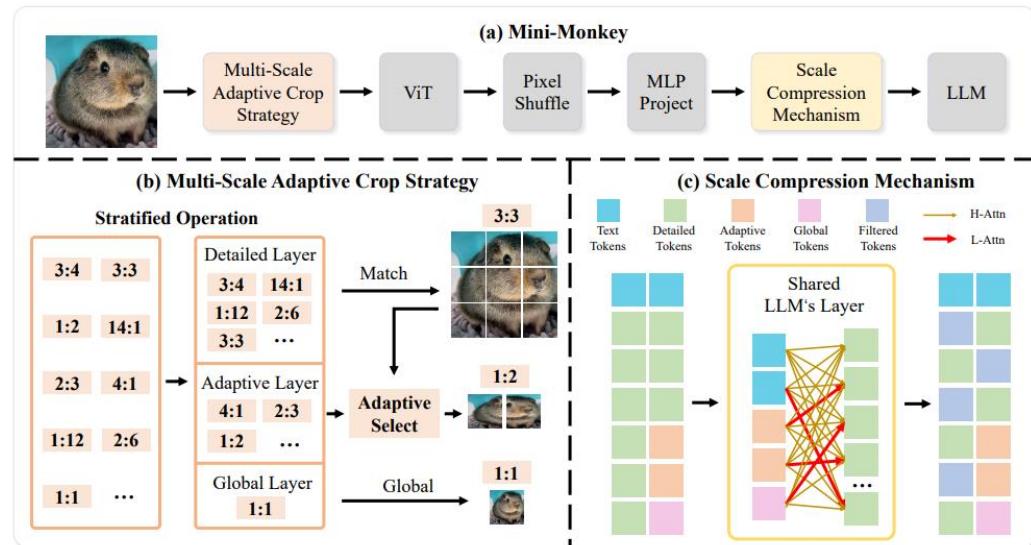


Figure 2: The overall architecture of Mini-Monkey. H-Attn represents high attention weight. L-Attn represents low attention weights. The tokens with low attention weights will be filtered. The shared LLM's Layer represents using the block layer from LLM in SCM.

Thanks Any Questions?

You can send mail to
Susang Kim(heales1@gmail.com)