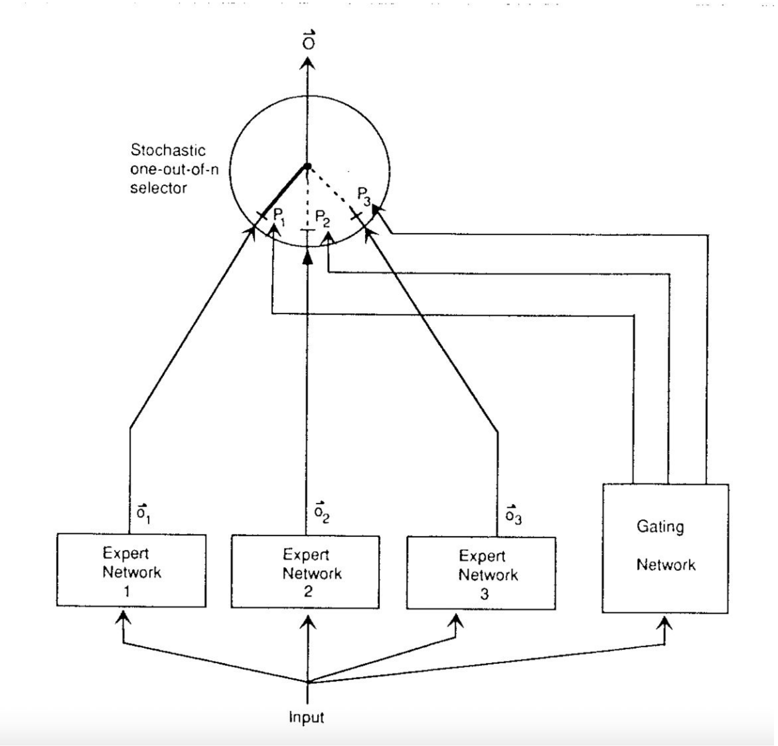
NEURAL INFORMATION
PROCESSING SYSTEMS

CuMo: Scaling Multimodal LLM with Co-Upcycled Mixture-of-Experts

Jiachen Li, Xinyao Wang, Sijie Wang, Chia-Wen Kuo, Lu Xu, Fan Chen,
Jitesh Jain, Humphrey Shi, Longyin Wen

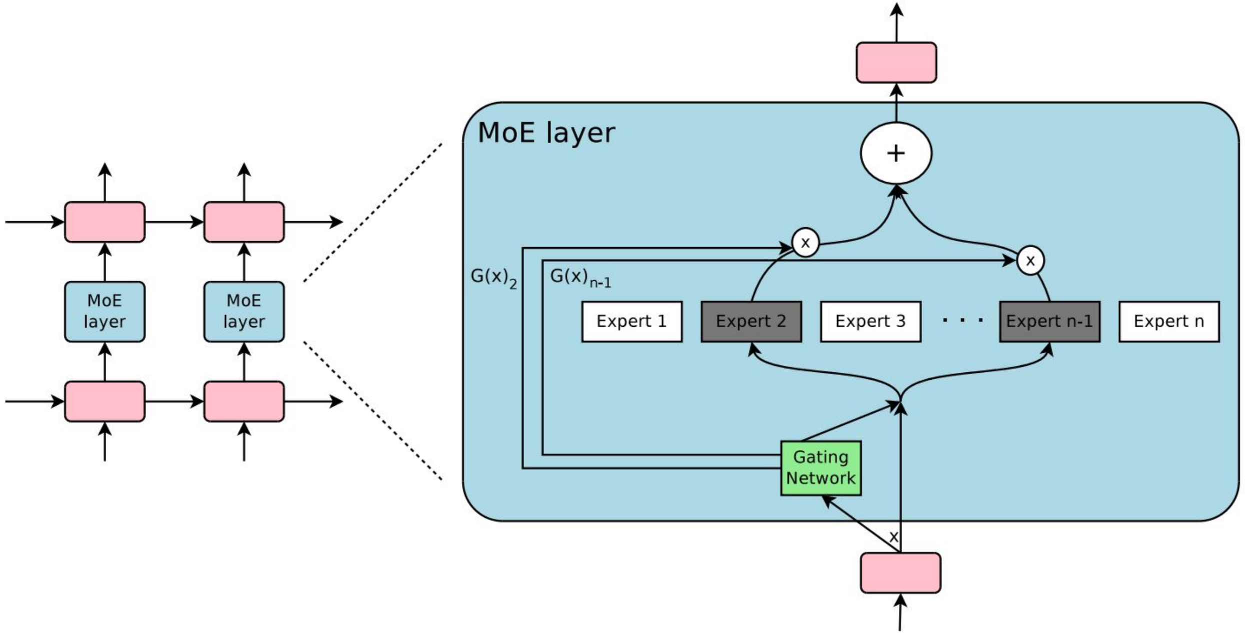
Background: Mixture-of-Expert



- Each expert is an independent network

Credits to *Robert A Jacobs, et al*

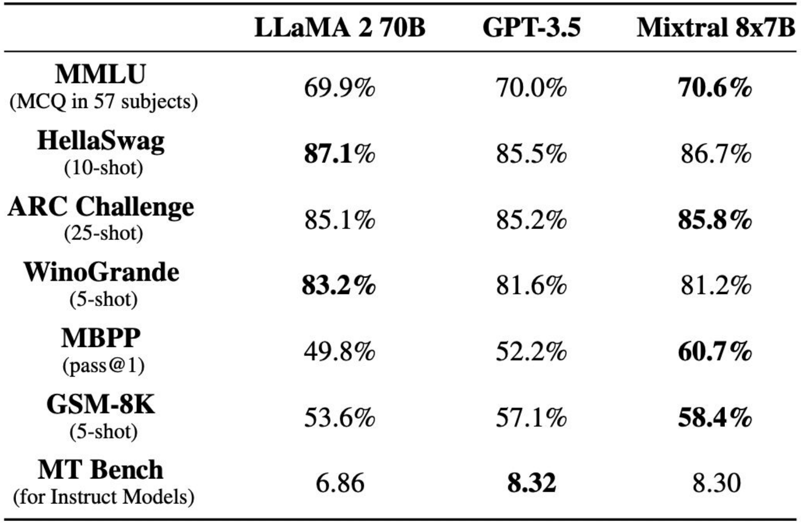
Background: Sparse MoE



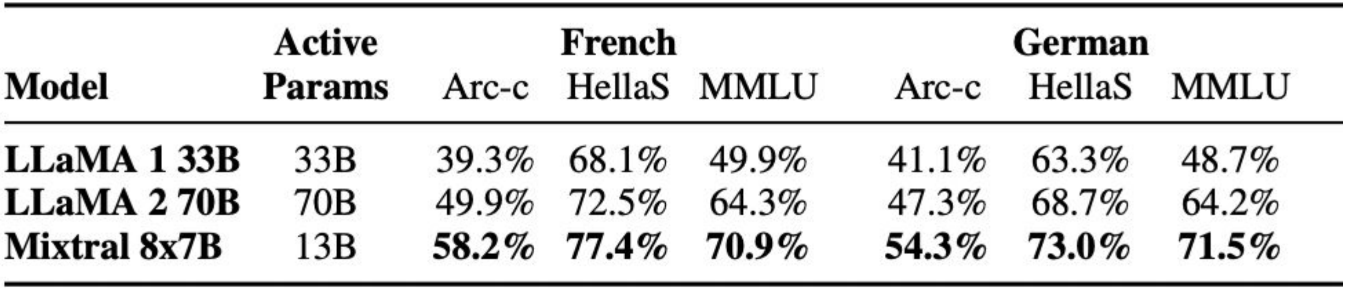
- Each expert is a MLP in the transformer architecture
- Gating network controls Top-K selection of experts

Credits to *Noam Shazeer, et al*

Background: Mixtral-8x7B

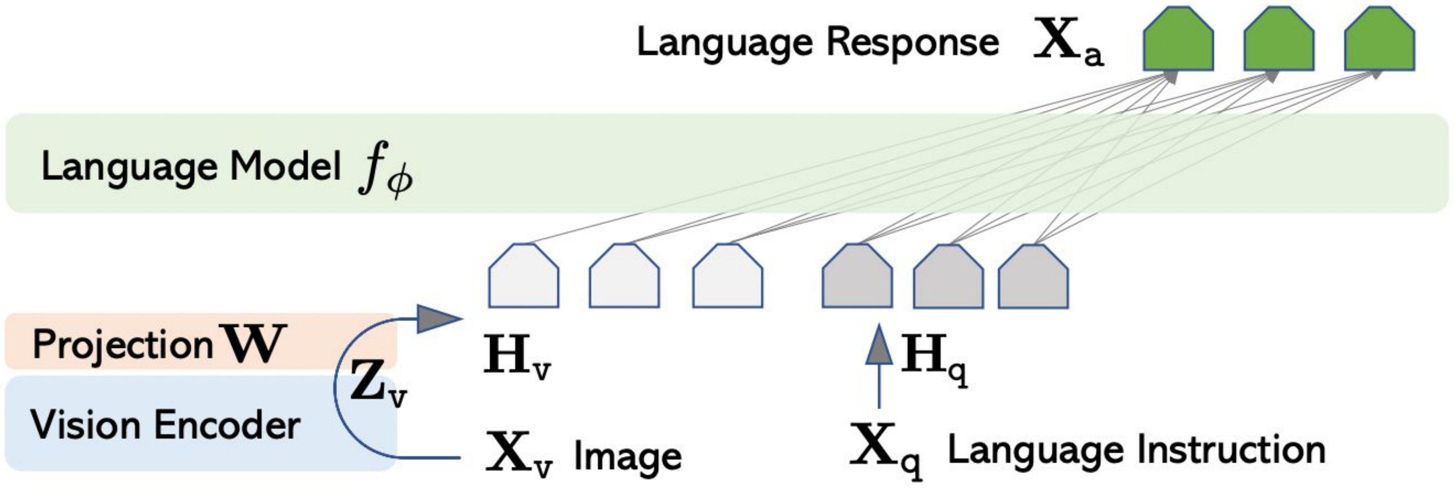


	LLaMA 2 70B	**GPT-3.5**	**Mixtral 8x7B**
MMLU (MCQ in 57 subjects)	69.9%	70.0%	**70.6%**
HellaSwag (10-shot)	**87.1%**	85.5%	86.7%
ARC Challenge (25-shot)	85.1%	85.2%	**85.8%**
WinoGrande (5-shot)	**83.2%**	81.6%	81.2%
MBPP (pass@1)	49.8%	52.2%	**60.7%**
GSM-8K (5-shot)	53.6%	57.1%	**58.4%**
MT Bench (for Instruct Models)	6.86	**8.32**	8.30



	Active		**French**			**German**	
Model	**Params**	Arc-c	HellaS	MMLU	Arc-c	HellaS	MMLU
LLaMA 1 33B	33B	39.3%	68.1%	49.9%	41.1%	63.3%	48.7%
LLaMA 2 70B	70B	49.9%	72.5%	64.3%	47.3%	68.7%	64.2%
Mixtral 8x7B	13B	**58.2%**	**77.4%**	**70.9%**	**54.3%**	**73.0%**	**71.5%**

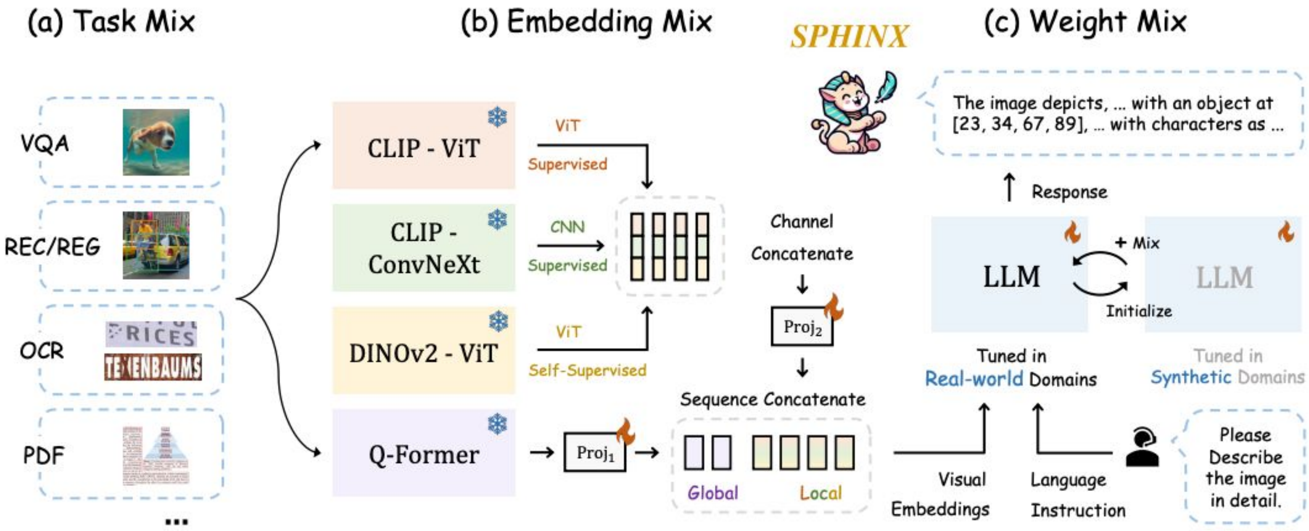
Background: Multimodal LLM



- LLaVA
  - CLIP + MLP Connector + LLM
  - Visual Instruction Tuning: Image Caption / OCR / VQA

Credits to *Liu Haotian, et al*

Background: Multimodal LLM

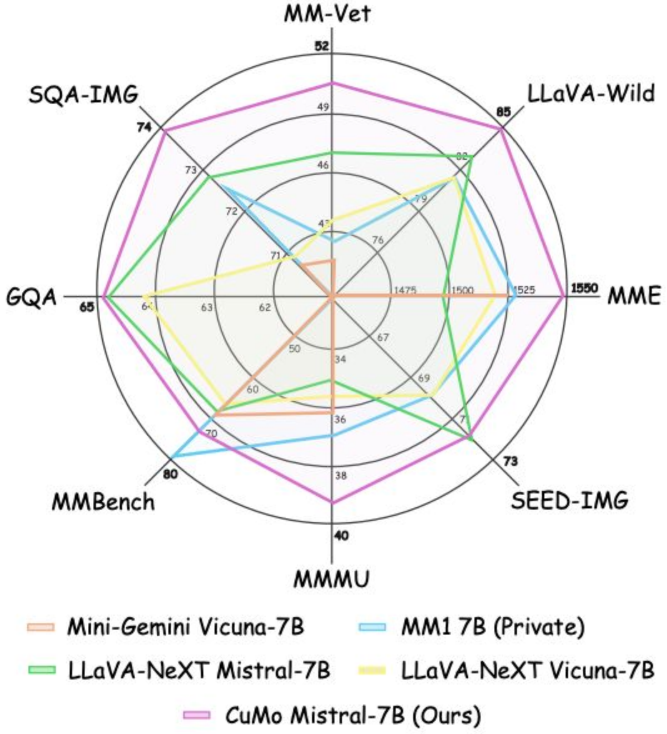


- Multiple vision encoders to improve the visual capacities of MLLM
- N times longer sequence of visual tokens

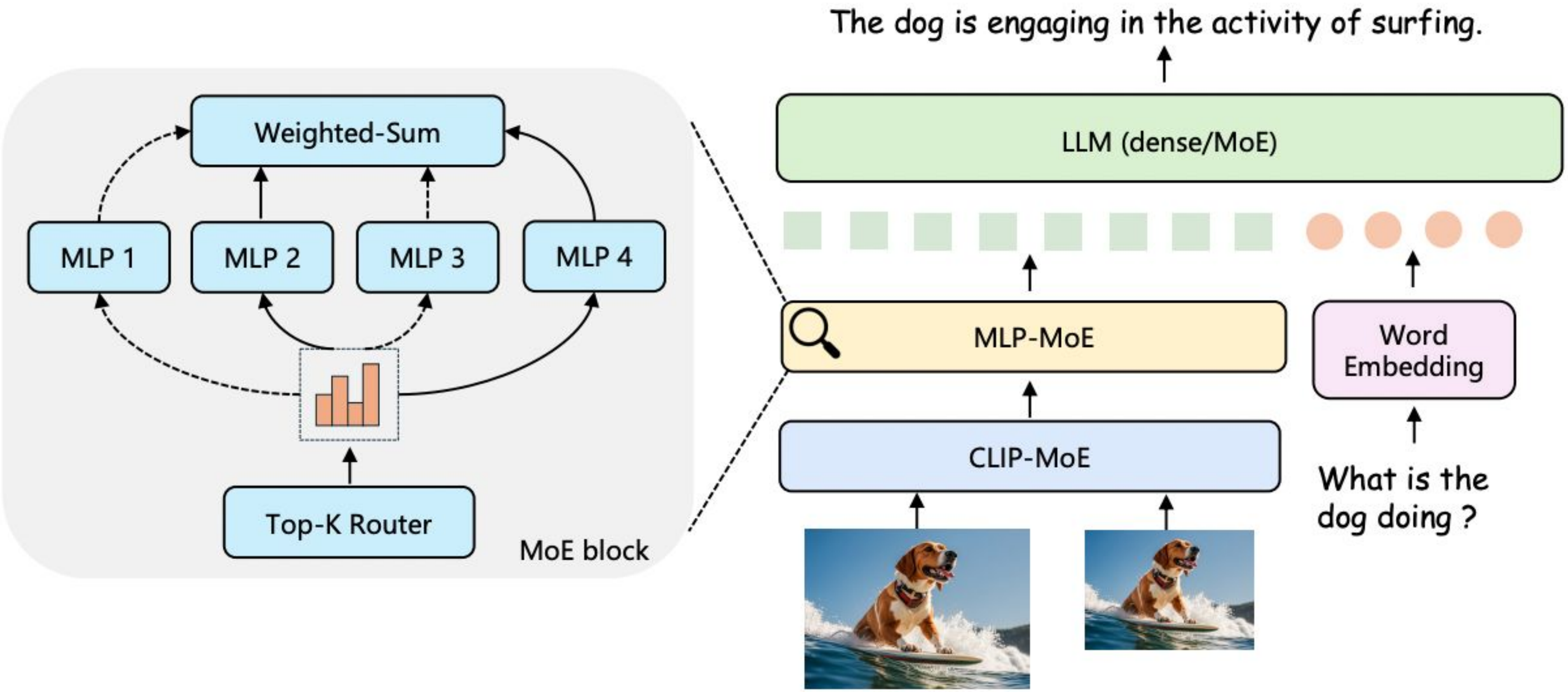
Credits to *Lin Ziyi, et al*

CuMo

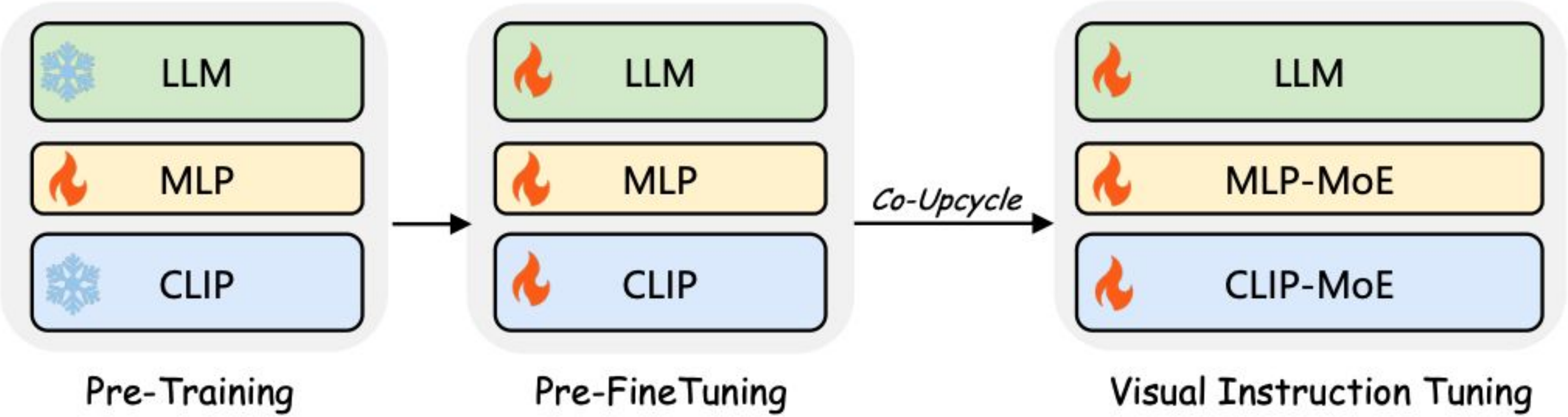
- MoE on CLIP-ViT & MLP
  - Improve visual capacities
  - Same length of input sequence to CLIP
- CuMo-7B
  - Mistral-7B + CLIP-MoE & MLP-MoE
- CuMo-8x7B
  - Mixtral-8x7B + CLIP-MoE & MLP-MoE



Architecture

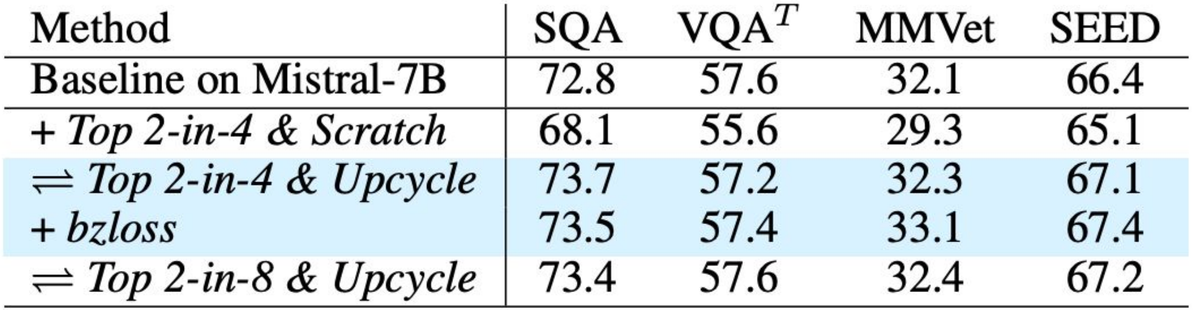


Training MoE



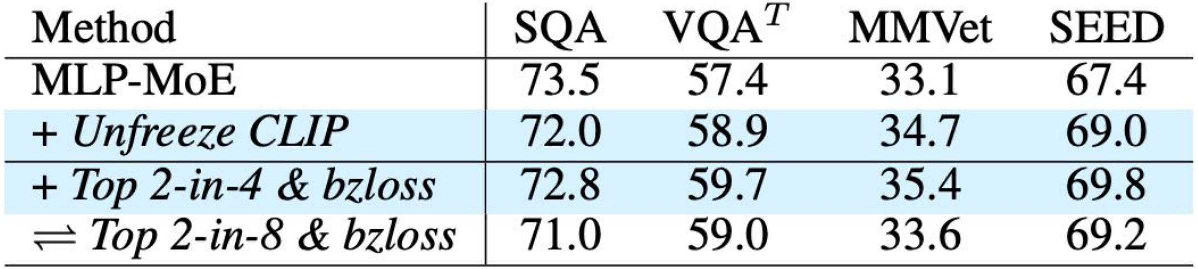
- Pre-Training : Simple coco caption data
- Pre-FineTuning : Complex long caption data
- Visual Instruction Tuning : Mixture of VQA / OCR / Caption / Text data

Ablation Study



Method	SQA	VQA^T	MMVet	SEED
Baseline on Mistral-7B	72.8	57.6	32.1	66.4
+ *Top 2-in-4 & Scratch*	68.1	55.6	29.3	65.1
$\rightleftharpoons$ *Top 2-in-4 & Upcycle*	73.7	57.2	32.3	67.1
+ *bzloss*	73.5	57.4	33.1	67.4
$\rightleftharpoons$ *Top 2-in-8 & Upcycle*	73.4	57.6	32.4	67.2

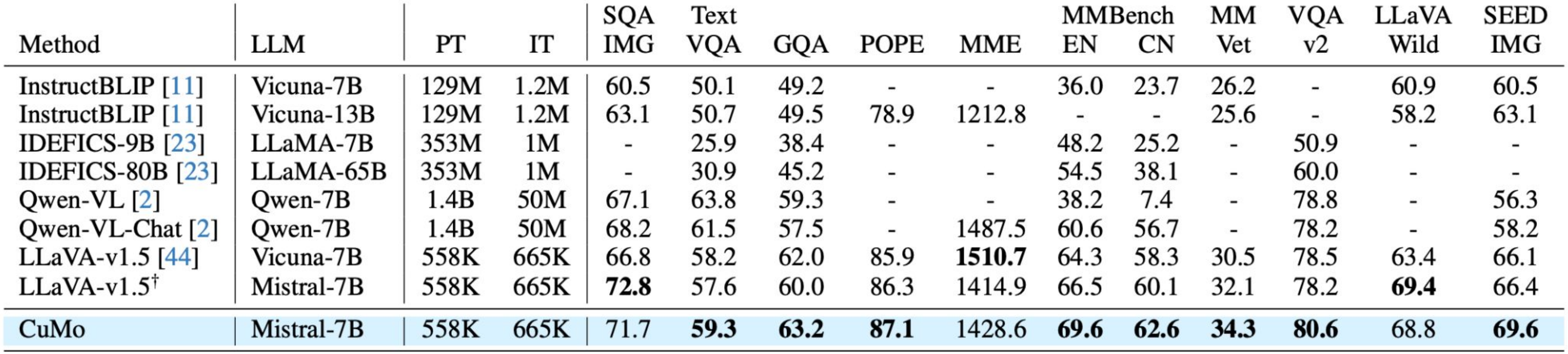
(a) MLP-MoE



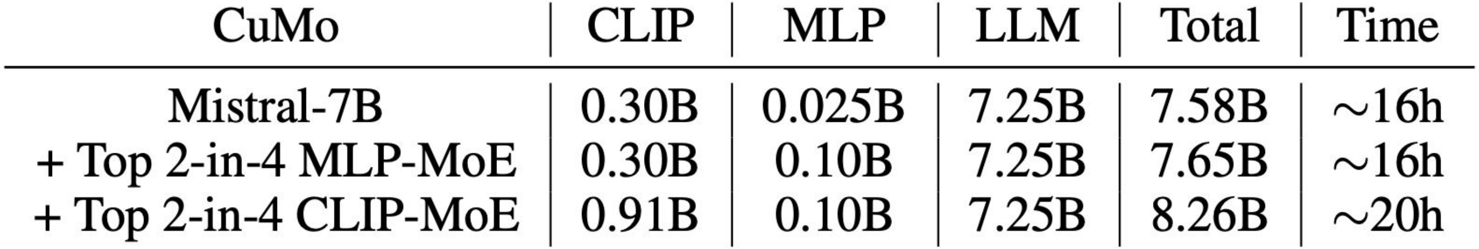
Method	SQA	VQA^T	MMVet	SEED
MLP-MoE	73.5	57.4	33.1	67.4
+ *Unfreeze CLIP*	72.0	58.9	34.7	69.0
+ *Top 2-in-4 & bzloss*	72.8	59.7	35.4	69.8
$\rightleftharpoons$ *Top 2-in-8 & bzloss*	71.0	59.0	33.6	69.2

(b) CLIP-MoE

Comparisons

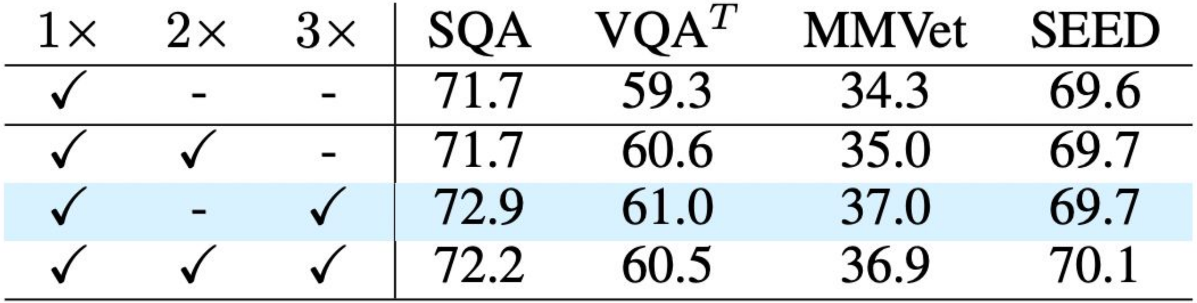


Method	LLM	PT	IT	SQA IMG	Text VQA	GQA	POPE	MME	MMBench EN	MMBench CN	MM Vet	VQA v2	LLaVA Wild	SEED IMG
InstructBLIP [11]	Vicuna-7B	129M	1.2M	60.5	50.1	49.2	-	-	36.0	23.7	26.2	-	60.9	60.5
InstructBLIP [11]	Vicuna-13B	129M	1.2M	63.1	50.7	49.5	78.9	1212.8	-	-	25.6	-	58.2	63.1
IDEFICS-9B [23]	LLaMA-7B	353M	1M	-	25.9	38.4	-	-	48.2	25.2	-	50.9	-	-
IDEFICS-80B [23]	LLaMA-65B	353M	1M	-	30.9	45.2	-	-	54.5	38.1	-	60.0	-	-
Qwen-VL [2]	Qwen-7B	1.4B	50M	67.1	63.8	59.3	-	-	38.2	7.4	-	78.8	-	56.3
Qwen-VL-Chat [2]	Qwen-7B	1.4B	50M	68.2	61.5	57.5	-	1487.5	60.6	56.7	-	78.2	-	58.2
LLaVA-v1.5 [44]	Vicuna-7B	558K	665K	66.8	58.2	62.0	85.9	**1510.7**	64.3	58.3	30.5	78.5	63.4	66.1
LLaVA-v1.5†	Mistral-7B	558K	665K	**72.8**	57.6	60.0	86.3	1414.9	66.5	60.1	32.1	78.2	**69.4**	66.4
CuMo	Mistral-7B	558K	665K	71.7	**59.3**	**63.2**	**87.1**	1428.6	**69.6**	**62.6**	**34.3**	**80.6**	68.8	**69.6**



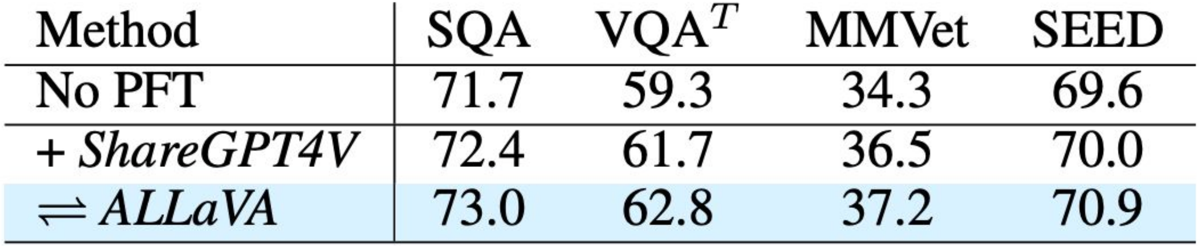
CuMo	CLIP	MLP	LLM	Total	Time
Mistral-7B	0.30B	0.025B	7.25B	7.58B	~16h
+ Top 2-in-4 MLP-MoE	0.30B	0.10B	7.25B	7.65B	~16h
+ Top 2-in-4 CLIP-MoE	0.91B	0.10B	7.25B	8.26B	~20h

Extra Ablation Study



1×	2×	3×	SQA	VQA^T	MMVet	SEED
✓	-	-	71.7	59.3	34.3	69.6
✓	✓	-	71.7	60.6	35.0	69.7
✓	-	✓	72.9	61.0	37.0	69.7
✓	✓	✓	72.2	60.5	36.9	70.1

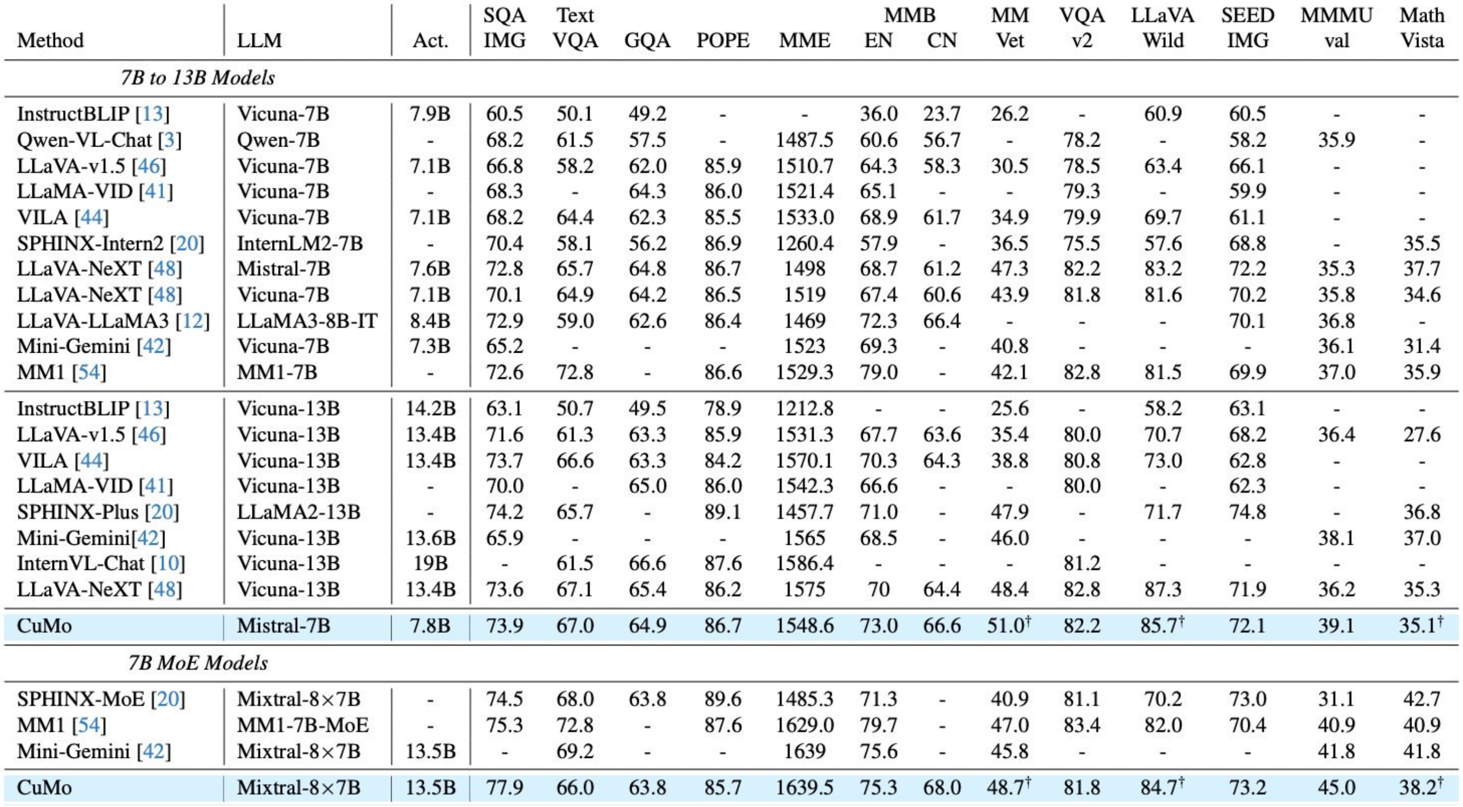
(d) Multi-resolution Feature



Method	SQA	VQA^T	MMVet	SEED
No PFT	71.7	59.3	34.3	69.6
+ *ShareGPT4V*	72.4	61.7	36.5	70.0
$\rightleftharpoons$ *ALLaVA*	73.0	62.8	37.2	70.9

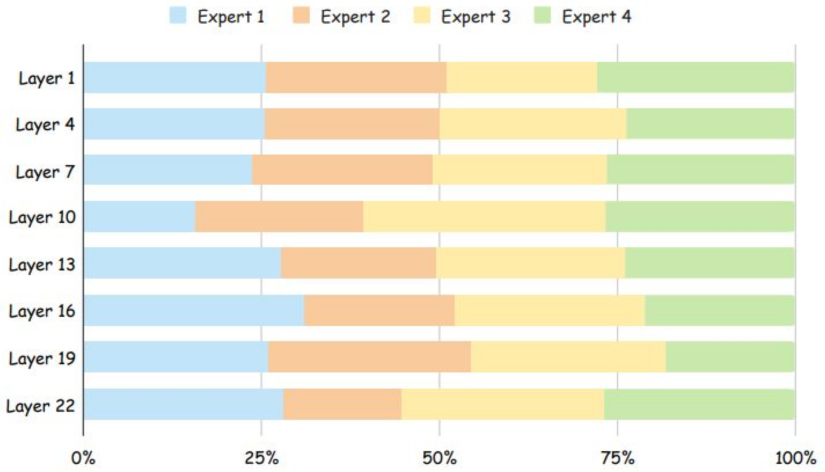
(e) Pre-FineTuning Stage

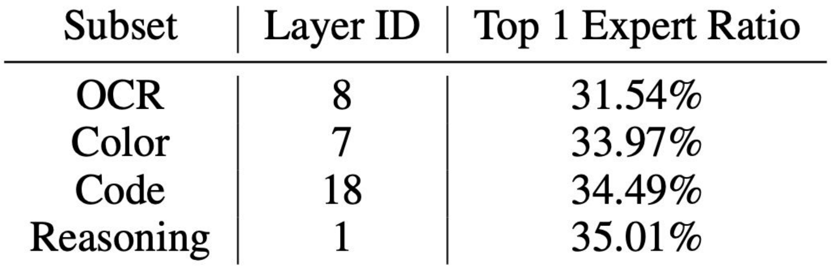
Main Table



Method	LLM	Act.	SQA IMG	Text VQA	GQA	POPE	MME	MMB EN	MMB CN	MM Vet	VQA v2	LLaVA Wild	SEED IMG	MMMU val	Math Vista
7B to 13B Models															
InstructBLIP [13]	Vicuna-7B	7.9B	60.5	50.1	49.2	-	-	36.0	23.7	26.2	-	60.9	60.5	-	-
Qwen-VL-Chat [3]	Qwen-7B	-	68.2	61.5	57.5	-	1487.5	60.6	56.7	-	78.2	-	58.2	35.9	-
LLaVA-v1.5 [46]	Vicuna-7B	7.1B	66.8	58.2	62.0	85.9	1510.7	64.3	58.3	30.5	78.5	63.4	66.1	-	-
LLaMA-VID [41]	Vicuna-7B	-	68.3	-	64.3	86.0	1521.4	65.1	-	-	79.3	-	59.9	-	-
VILA [44]	Vicuna-7B	7.1B	68.2	64.4	62.3	85.5	1533.0	68.9	61.7	34.9	79.9	69.7	61.1	-	-
SPHINX-Intern2 [20]	InternLM2-7B	-	70.4	58.1	56.2	86.9	1260.4	57.9	-	36.5	75.5	57.6	68.8	-	35.5
LLaVA-NeXT [48]	Mistral-7B	7.6B	72.8	65.7	64.8	86.7	1498	68.7	61.2	47.3	82.2	83.2	72.2	35.3	37.7
LLaVA-NeXT [48]	Vicuna-7B	7.1B	70.1	64.9	64.2	86.5	1519	67.4	60.6	43.9	81.8	81.6	70.2	35.8	34.6
LLaVA-LLaMA3 [12]	LLaMA3-8B-IT	8.4B	72.9	59.0	62.6	86.4	1469	72.3	66.4	-	-	-	70.1	36.8	-
Mini-Gemini [42]	Vicuna-7B	7.3B	65.2	-	-	-	1523	69.3	-	40.8	-	-	-	36.1	31.4
MM1 [54]	MM1-7B	-	72.6	72.8	-	86.6	1529.3	79.0	-	42.1	82.8	81.5	69.9	37.0	35.9
InstructBLIP [13]	Vicuna-13B	14.2B	63.1	50.7	49.5	78.9	1212.8	-	-	25.6	-	58.2	63.1	-	-
LLaVA-v1.5 [46]	Vicuna-13B	13.4B	71.6	61.3	63.3	85.9	1531.3	67.7	63.6	35.4	80.0	70.7	68.2	36.4	27.6
VILA [44]	Vicuna-13B	13.4B	73.7	66.6	63.3	84.2	1570.1	70.3	64.3	38.8	80.8	73.0	62.8	-	-
LLaMA-VID [41]	Vicuna-13B	-	70.0	-	65.0	86.0	1542.3	66.6	-	-	80.0	-	62.3	-	-
SPHINX-Plus [20]	LLaMA2-13B	-	74.2	65.7	-	89.1	1457.7	71.0	-	47.9	-	71.7	74.8	-	36.8
Mini-Gemini[42]	Vicuna-13B	13.6B	65.9	-	-	-	1565	68.5	-	46.0	-	-	-	38.1	37.0
InternVL-Chat [10]	Vicuna-13B	19B	-	61.5	66.6	87.6	1586.4	-	-	-	81.2	-	-	-	-
LLaVA-NeXT [48]	Vicuna-13B	13.4B	73.6	67.1	65.4	86.2	1575	70	64.4	48.4	82.8	87.3	71.9	36.2	35.3
CuMo	Mistral-7B	7.8B	73.9	67.0	64.9	86.7	1548.6	73.0	66.6	51.0†	82.2	85.7†	72.1	39.1	35.1†
7B MoE Models															
SPHINX-MoE [20]	Mixtral-8×7B	-	74.5	68.0	63.8	89.6	1485.3	71.3	-	40.9	81.1	70.2	73.0	31.1	42.7
MM1 [54]	MM1-7B-MoE	-	75.3	72.8	-	87.6	1629.0	79.7	-	47.0	83.4	82.0	70.4	40.9	40.9
Mini-Gemini [42]	Mixtral-8×7B	13.5B	-	69.2	-	-	1639	75.6	-	45.8	-	-	-	41.8	41.8
CuMo	Mixtral-8×7B	13.5B	77.9	66.0	63.8	85.7	1639.5	75.3	68.0	48.7†	81.8	84.7†	73.2	45.0	38.2†

Expert Analysis

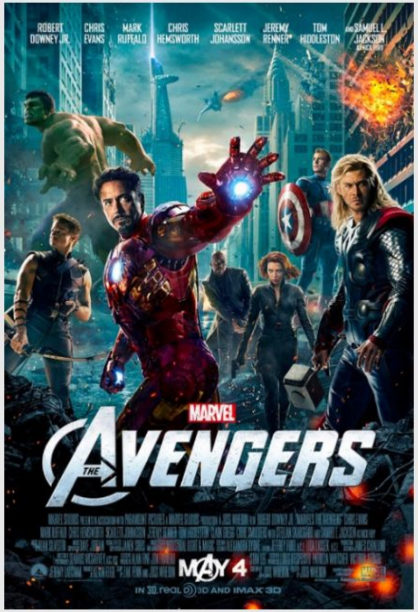




Subset	Layer ID	Top 1 Expert Ratio
OCR	8	31.54%
Color	7	33.97%
Code	18	34.49%
Reasoning	1	35.01%

- Expert loadings are overall equally distributed
- Expert selections show preference on certain subtasks

Examples



can you introduce this movie based on this poster

LLaVA-Yi-34B

This poster is for the movie "Avengers: Age of Ultron," which is a 2015 American superhero film based on the Marvel Comics superhero team the Avengers. The film is the sequel to "The Avengers" (2012) and the 11th film in the Marvel Cinematic Universe (MCU). It was directed by Joss Whedon and produced by Marvel Studios.

The poster features several of the main characters from the film, including Iron Man, Thor, Captain America, the Hulk, Black Widow, and Hawkeye, all in their superhero costumes. The setting appears to be a cityscape with a sense of destruction, suggesting a battle or conflict. The title "Avengers: Age of Ultron" is prominently displayed at the top of the poster, and the Marvel logo is visible at the bottom.

The film's plot revolves around the Avengers attempting to stop the villain Ultron from enacting his plans for human extinction. The movie is known for its action-packed sequences, special effects, and ensemble cast of superheroes. It was a critical and commercial success, contributing to the ongoing popularity of the Marvel Cinematic Universe.

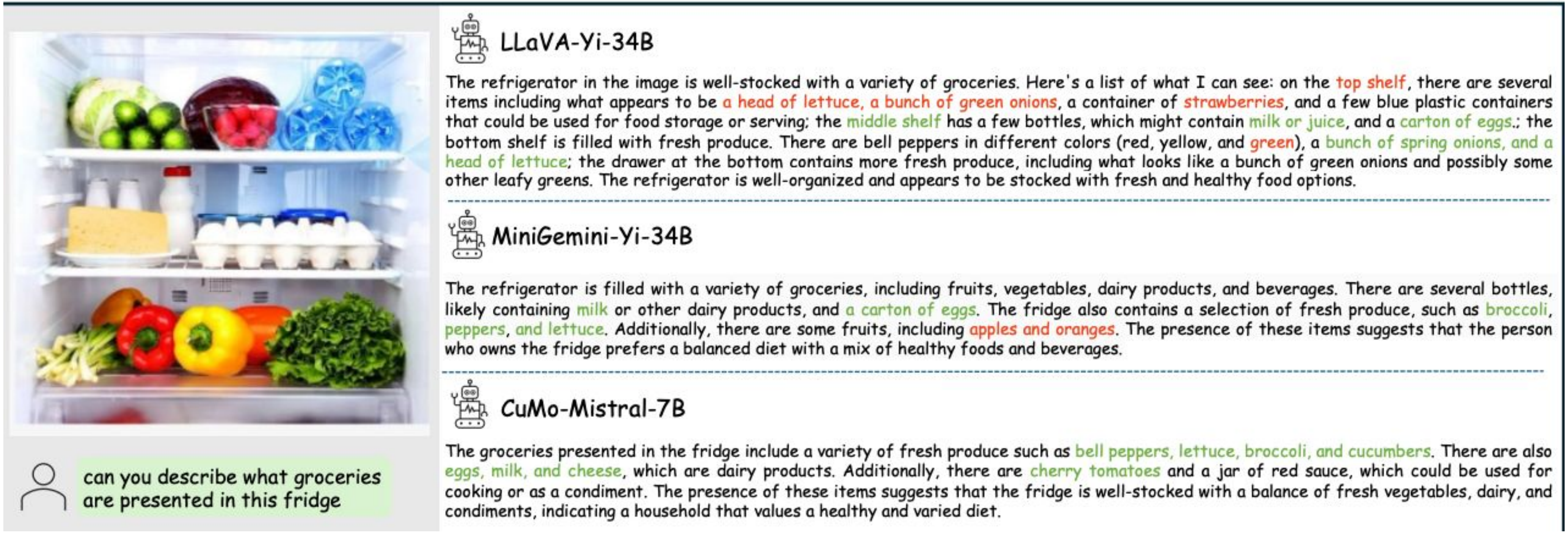
MiniGemini-Yi-34B

The movie poster is for "The Avengers," a 2012 American superhero film produced by Marvel Studios and distributed by Walt Disney Pictures. It is the sixth film in the Marvel Cinematic Universe (MCU). The film features an ensemble cast including Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth, Scarlett Johansson, Jeremy Renner, Tom Hiddleston, and Samuel L. Jackson. The plot revolves around Nick Fury, director of the peacekeeping organization S.H.I.E.L.D., recruiting a team of superheroes to help save the world from Loki, the adoptive brother of Thor, who is leading an alien invasion. The poster is designed to showcase the main characters and the action-packed nature of the film, with the characters in their iconic costumes and the cityscape in the background suggesting the scale of the conflict. The release date is prominently displayed, indicating the film's availability in theaters. The poster also highlights the film's availability in various formats, including 3D, which was a significant selling point at the time of its release.

CuMo-Mistral-7B

The poster is for the movie "The Avengers," which is part of the Marvel Cinematic Universe. It features a team of superheroes, each with unique abilities and costumes, coming together to face a significant threat. The central figure, with the iconic red and gold suit, is likely the leader of the group. The presence of the Hulk and Thor suggests that the film includes a mix of science fiction and fantasy elements. The cityscape and the explosion in the background imply that the movie's plot involves a large-scale conflict that impacts the urban environment. The release date and the mention of "3D and IMAX 3D" indicate that the movie was available in various formats, catering to different viewing preferences. The poster is designed to convey the epic scale and the high stakes of the film, as well as the diverse team of heroes that viewers can expect to see on screen.

Examples



can you describe what groceries are presented in this fridge

LLaVA-Yi-34B

The refrigerator in the image is well-stocked with a variety of groceries. Here's a list of what I can see: on the top shelf, there are several items including what appears to be a head of lettuce, a bunch of green onions, a container of strawberries, and a few blue plastic containers that could be used for food storage or serving; the middle shelf has a few bottles, which might contain milk or juice, and a carton of eggs.; the bottom shelf is filled with fresh produce. There are bell peppers in different colors (red, yellow, and green), a bunch of spring onions, and a head of lettuce; the drawer at the bottom contains more fresh produce, including what looks like a bunch of green onions and possibly some other leafy greens. The refrigerator is well-organized and appears to be stocked with fresh and healthy food options.

MiniGemini-Yi-34B

The refrigerator is filled with a variety of groceries, including fruits, vegetables, dairy products, and beverages. There are several bottles, likely containing milk or other dairy products, and a carton of eggs. The fridge also contains a selection of fresh produce, such as broccoli, peppers, and lettuce. Additionally, there are some fruits, including apples and oranges. The presence of these items suggests that the person who owns the fridge prefers a balanced diet with a mix of healthy foods and beverages.

CuMo-Mistral-7B

The groceries presented in the fridge include a variety of fresh produce such as bell peppers, lettuce, broccoli, and cucumbers. There are also eggs, milk, and cheese, which are dairy products. Additionally, there are cherry tomatoes and a jar of red sauce, which could be used for cooking or as a condiment. The presence of these items suggests that the fridge is well-stocked with a balance of fresh vegetables, dairy, and condiments, indicating a household that values a healthy and varied diet.

Future Works

- Alignment
  - Add RLHF to reduce hallucinations of CuMo's outputs
- Add Capacity & Modality
  - Capacity
    - Visual Grounding
    - Image Generation
  - Modality
    - Video
    - Audio

Thank You!

Code / Model

Demo

Project