Methods, Analysis & Insights from Multimodal LLM Pre-training

Zhe Gan

CVPR 2024 I Apple I 2024.6.17

Multimodal LLM Has Been an Increasingly Hot Research Topic

Proprietary systems

OpenAI's GPT-4v and GPT-4o

Google's Gemini 1.5 Pro

Claude-3 Opus and Reka Core

xAI's Grok-1.5v, etc.

Open-source models

LLaVA-NeXT

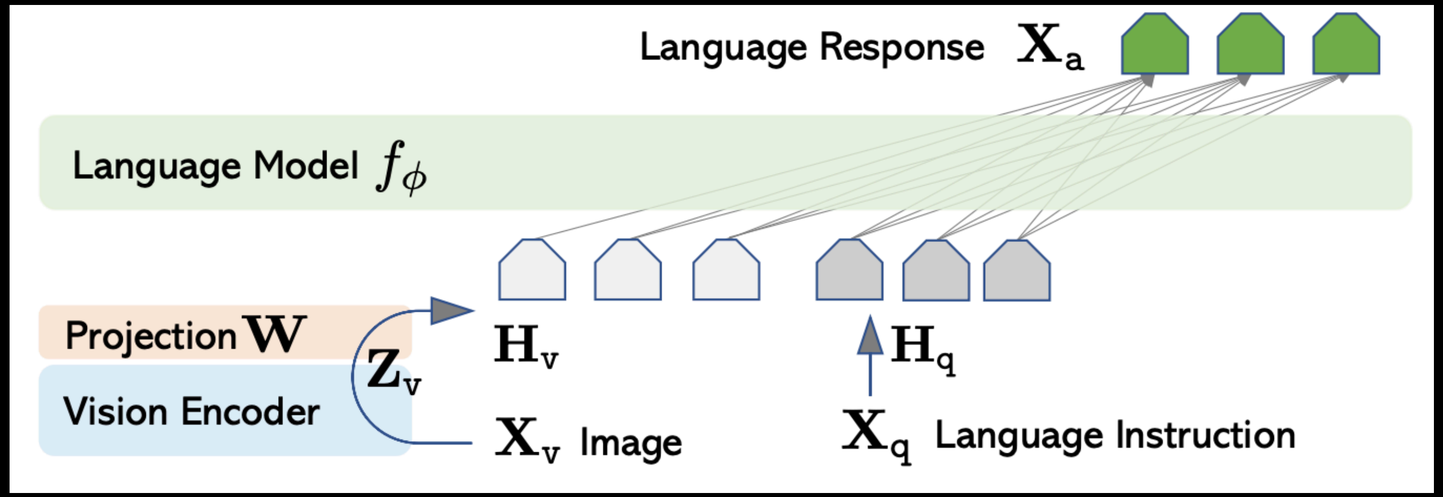
Qwen-VL-MAX

InternVL & InternLM-XComposer2-VL

VILA-1.5, Emu2, etc.

Visual Instruction Tuning

- Visual instruction tuning has been a key topic for open-source models
  - First, a lightweight alignment stage
  - Followed by SFT via using high-quality data, often distilled from GPT-4(v)
  - All the papers follow roughly the same pattern with recent focus on higher resolutions and stronger LLM backbones



[1] Visual Instruction Tuning, NeurIPS 2023

Indeed, the Performance is Very Strong

Results with LMMs-Eval				GPT4-V	LLaVA-NeXT (2024-05 Release)			LLaVA-NeXT (2024-01 Release)			
Datasets	Split	Metric	Instances		Qwen1.5-110B	Qwen1.5-72B	LLaMA3-8B	Yi-34B	Vicuna-1.5-13B	Vicuna-1.5-7B	Mistral-7B
AI2D*	test	Acc.	3088	78.2	**80.4**	77.4	71.6	74.9	70.0	66.6	60.8
ChartQA*	test	RelaxedAcc.	2500	78.5	**79.7**	77.0	69.5	68.7	62.2	54.8	38.8
DocVQA*	val	ANLS	5349	-	**85.7**	84.4	78.2	84.0	77.5	74.4	72.2
MathVista	test	Acc.	1000	**49.9**	49.0	46.6	37.5	46.0	35.1	34.4	37.4
MMBench	dev	Acc.	4377	75.0	**80.5**	**80.5**	72.1	79.3	-	-	-
MME-Cognition	test	Total Score	2374	**517.1**	453.9	459.6	367.8	397.1	316.8	322.5	323.9
MME-Perception	test			1409.4	**1746.5**	1699.3	1603.7	1633.2	1575.1	1519.3	1500.9
MMMU	val	Acc.	900	**56.8**	49.1	46.4	41.7	46.7	35.9	35.1	33.4
RealWorldQA	test	Acc.	765	61.4	63.1	**65.4**	60.0	61.0	-	-	54.4
LLaVA-W**	test	GPT4-Eval	60	**98.0**	90.4	89.2	80.1	88.8	72.3	72.3	71.7
LLaVA-Bench (Wilder)	Small	GPT4V-Eval	120	**71.5**	70.5	71.2	62.5	-	-	-	-
	Medium	GPT4V-Eval	1020	**78.5**	72.5	73.4	63.1	-	-	-	-

[1] https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/

So, do we still need large-scale multimodal pre-training if light-weight alignment is all you need?

Probably the Answer is Yes!

- Pre-training is crucial for the model to understand and digest the abundant multimodal knowledge and interactions in our visual world
- Pre-training learns deep modality fusion, rather than shallow alignment
- Image-text and text-only data used for CLIP and LLM pre-training inside the end-to-end LLaVA training lifecycle may not be sufficient
- Using interleaved image-text data mimics LLM pre-training, allowing us to incorporate multimodal data from the start, and even enhance LLM itself
- Multimodal in-context learning powered by pre-training is akin to instruction following, which can be lost via solely instruction tuning
- Pre-training provides the flexibility beyond LLaVA-style architectures

How to Perform Multimodal Pre-training? A Little Bit History

- The journey started roughly 5 years ago

LXMERT

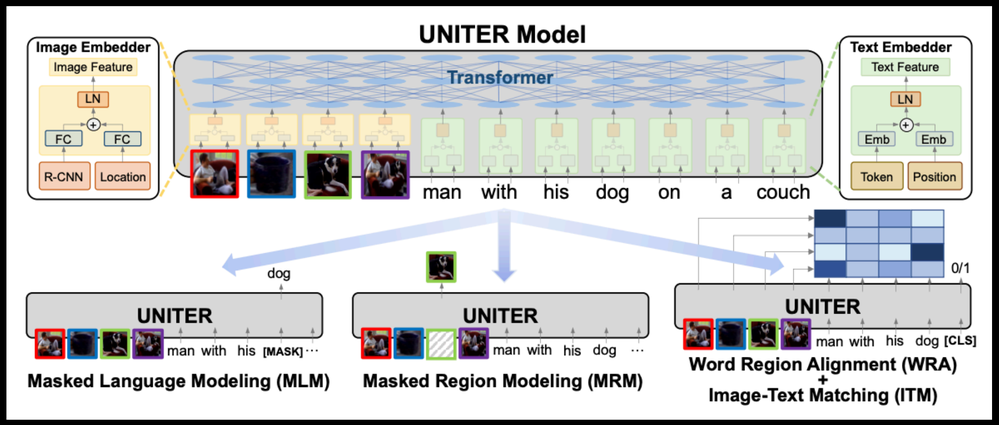
ViLBERT

UNITER

2019/8

BERT-style

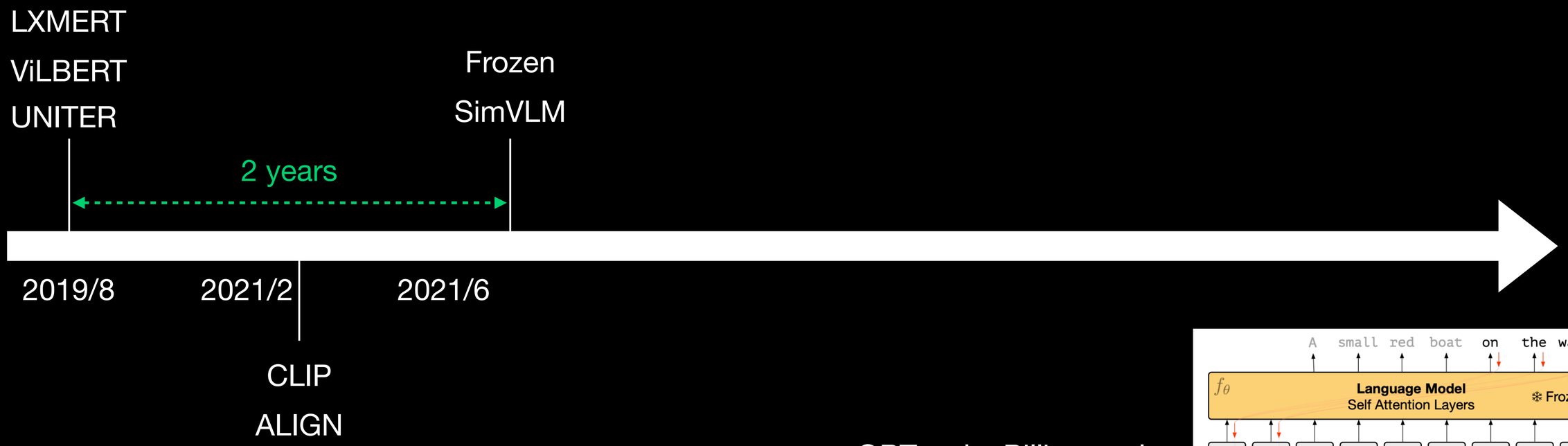
Million-scale



Note that there is a vast amount of literature on this topic, and we only select representative papers

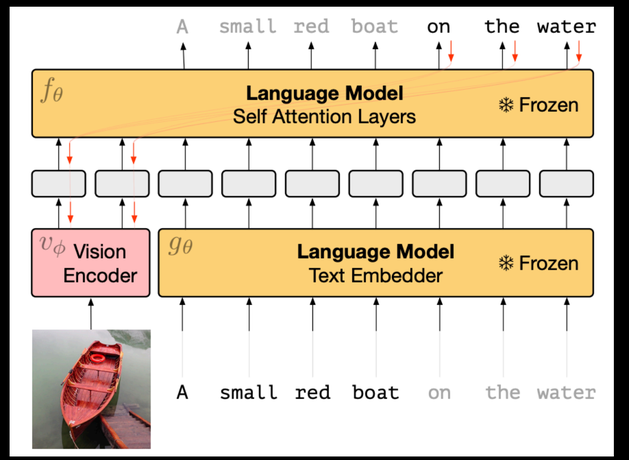
How to Perform Multimodal Pre-training? A Little Bit History

- It took 2 years from BERT-style pre-training to GPT-style pre-training



GPT-style, Billion-scale

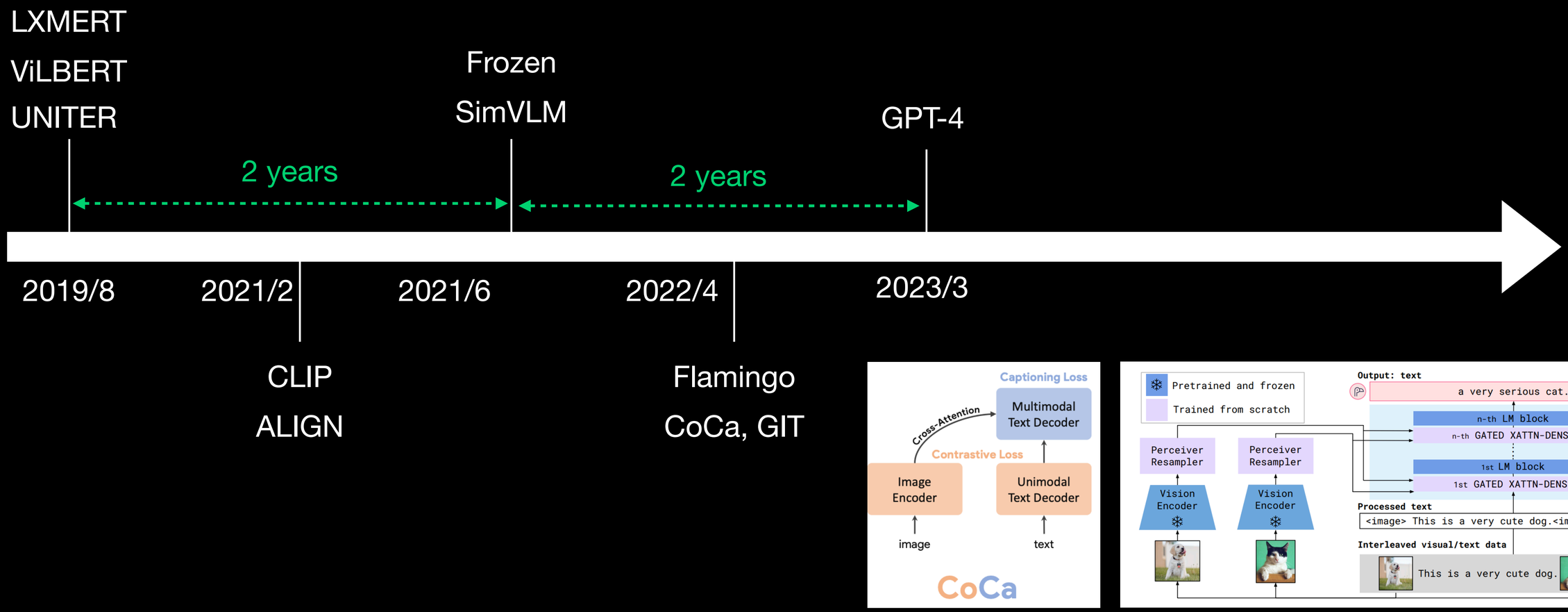
Simple pre-training loss



Note that there is a vast amount of literature on this topic, and we only select representative papers

How to Perform Multimodal Pre-training? A Little Bit History

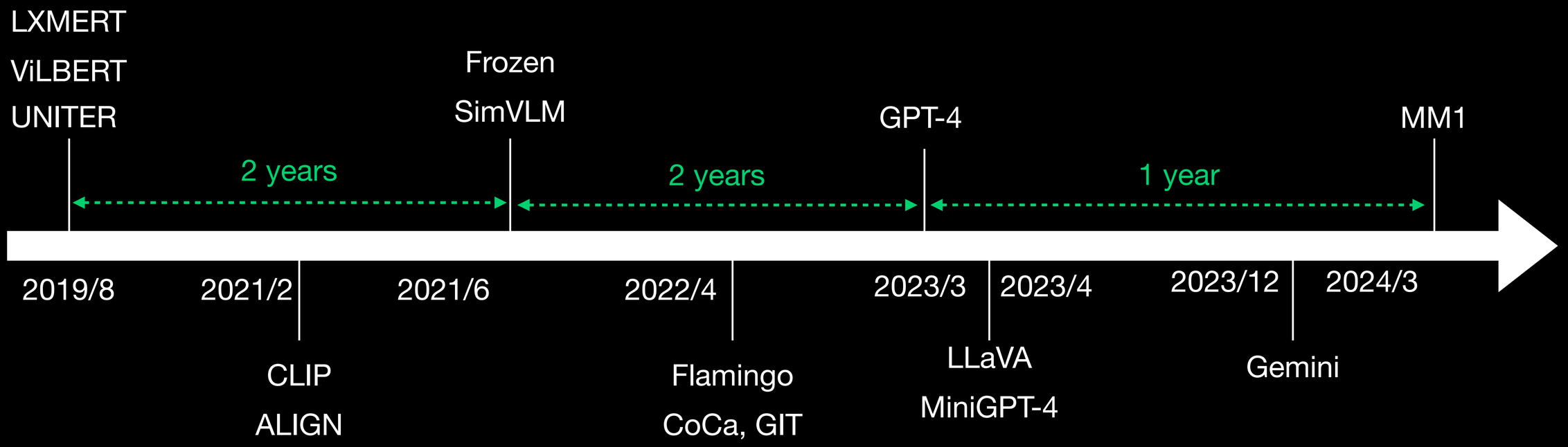
- It took another 2 years from Frozen to Flamingo, and to GPT-4



Note that there is a vast amount of literature on this topic, and we only select representative papers

How to Perform Multimodal Pre-training? A Little Bit History

- The past year has witnessed the boom of visual instruction tuning
- Public works that focus on pre-training include VILA, Emu2, MM1 etc.



Note that there is a vast amount of literature on this topic, and we only select representative papers

Agenda

- MM1 as a Case Study
  - Overview
  - Recipe for Building MM1
  - Final Model and Training Recipe
- Other Model Architecture Design
  - Fuyu, CM3 & SEED

MM1 Overview

Architecture Overview

Decoder-only Architecture

LLM Decoder

Embedding

"This Walnut and Blue Cheese Stuffed Mushrooms recipe is sponsored by Fisher Nuts."

Connector

Image Encoder (ViT)

Embedding

"Stuffed mushrooms are an appetizer that always grabs my attention at a party."

Connector

Image Encoder (ViT)

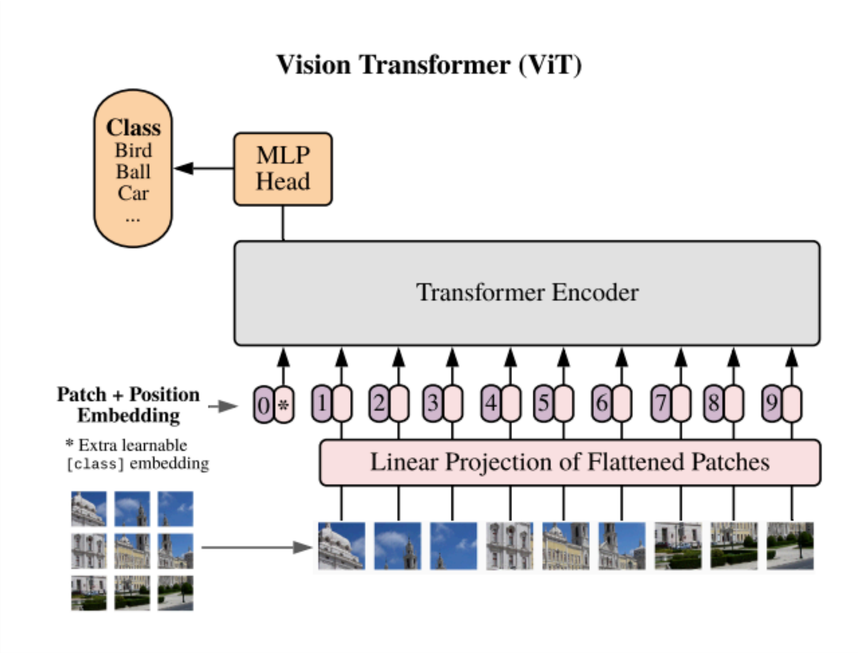
Key Components — LLM Decoder

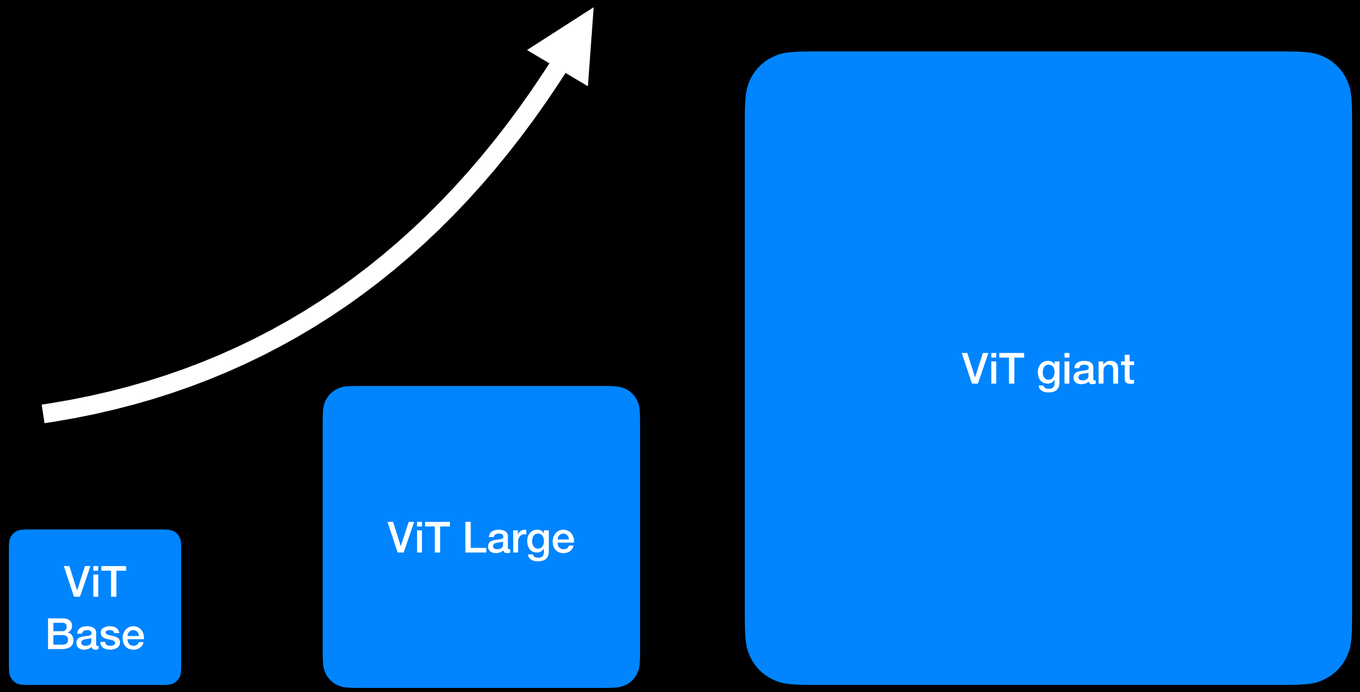
We explored both dense model (3B, 7B, 30B) and MoE variants

- We directly used the pre-trained LLM backbones, as multimodal pre-training will be performed later on
- Most open-source models use post-trained LLMs instead, such as Vicuna

Key Components — Image Encoder

Vision Transformer (ViT) trained using CLIP objective

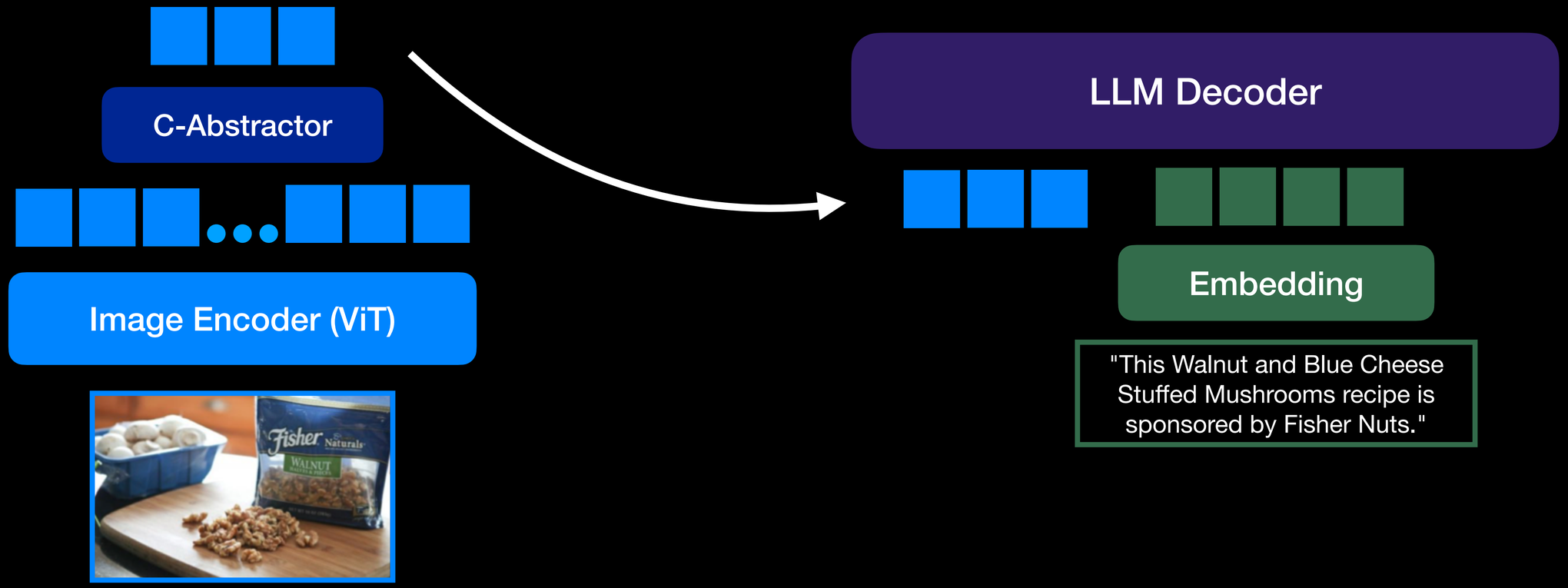




Key Components — Connector

1. Connector to aggregate patch representation to a fixed number of representations — > Current choice: **C-Abstractor**

2. The pooler representations are concatenated with text token embeddings



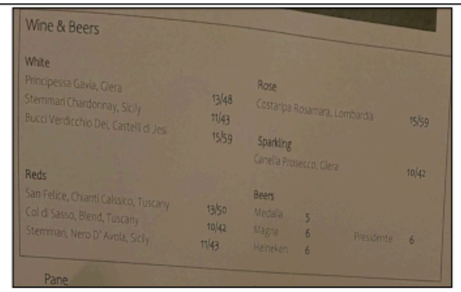
[1] Honeybee: Locality-enhanced Projector for Multimodal LLM, CVPR 2024

MM1 Exhibits Strong In-Context Learning Capability

Counting, visual prompting, and reasoning

MM1-Chat Can Follow Instructions and Reason across Images

User:
Photo: <img1> Menu: <img2> How much should I pay for all the beer on the table according to the price on the menu?



MM1-30B-Chat (Ours): 12

Emu-Chat-37B: According to the menu, the price for all the beer on the table is 15.99.

LLaVA-NeXT-34B: The image shows a menu with beer prices, but it's not entirely clear which specific beers are on the table. However, I can provide you with an estimate based on the visible prices. From the menu, it appears that the prices for beer are as follows: Magna: 5, Medalla: 5, Heineken: 6 Presidente: 6
Assuming that the beers on the table are Magna, Medalla, Heineken, and Presidente, and there are two of each, you would pay:
2 x Magna: 2 x 5 = 10, 2 x Medalla: 2 x 5 = 10, 2 x Heineken: 2 x 6 = 12, 2 x Presidente: 2 x 6 = 12 Total: 10 + 10 + 12 + 12 = 44
Please note that this is an estimate based on the visible prices and the assumption that there are two of each type of beer. The actual cost may vary if the beers on the table are not the ones listed on the menu or if there are different quantities.

User: Explain why.

MM1-30B-Chat (Ours): There are 2 beers on the table, and each beer costs 6 according to the price on the menu. So, 2*6=12.

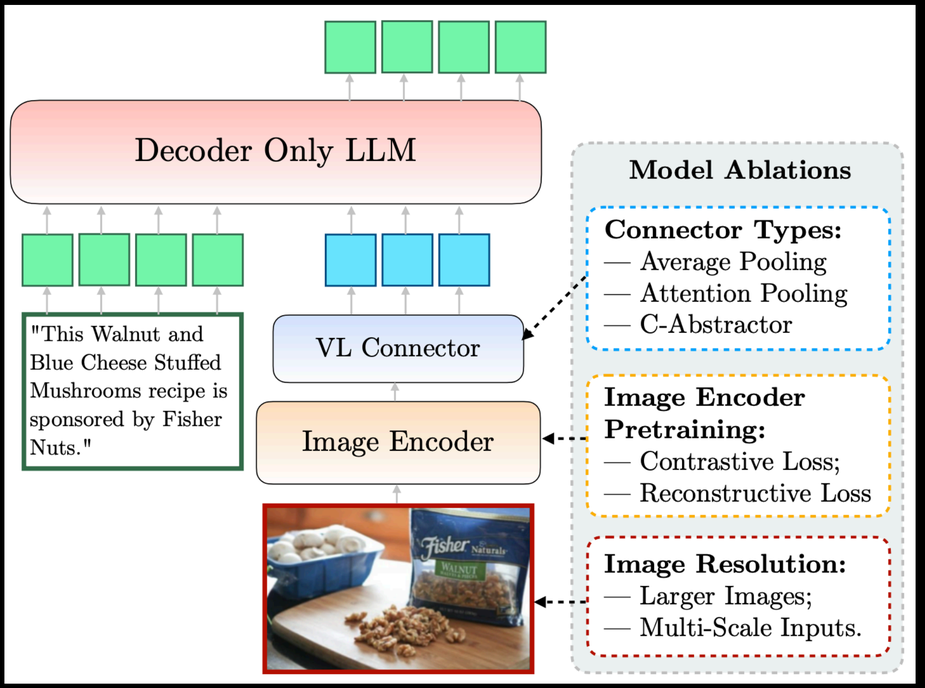
Emu-Chat-37B: The price for all the beer on the table is 15.99 because each beer has a price of 10.99, and there are a total of five beers on the table. When you multiply the number of beers (5) by the price per beer (10.99), you get a total cost of 54.95 (rounded up to 15.99).

LLaVA-NeXT-34B: The image shows a menu with beer prices, but it's not entirely clear which specific beers are on the table. However, I can provide you with an estimate based on the visible prices. From the menu, it appears that the prices for beer are as follows: (...)

Recipe for Building MM1

Model and Data Ablations

- For ablation, we use ViT-L/14, C-Abstractor, and a 1.2B LLM as the default setting
- A mix of captioned images (45%), interleaved image-text (45%) and text-only data (10%) as the pre-training data mixture
- 0/4/8-shot average on a variety of captioning and VQA datasets for evaluation



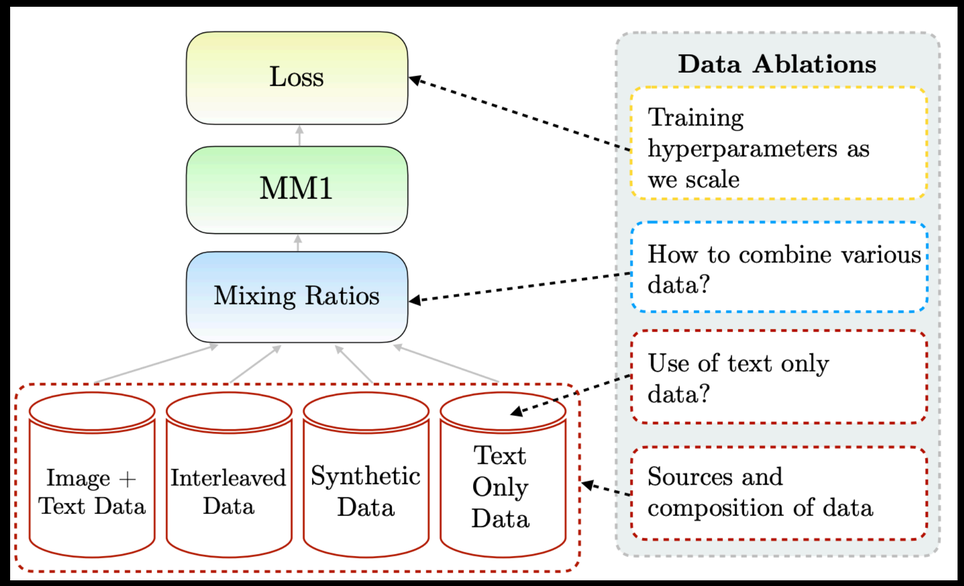
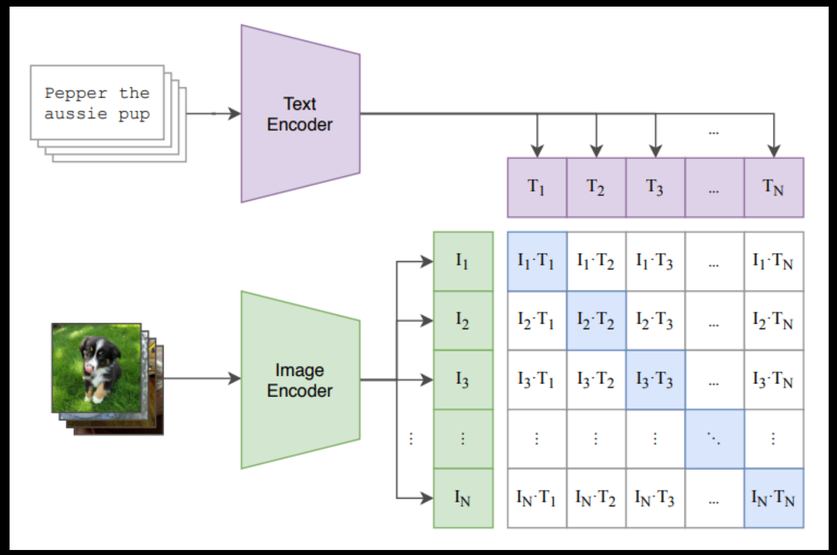
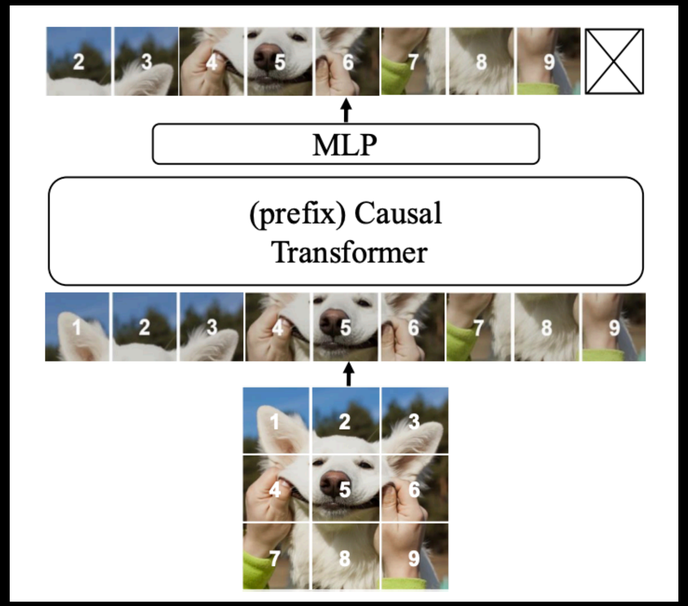


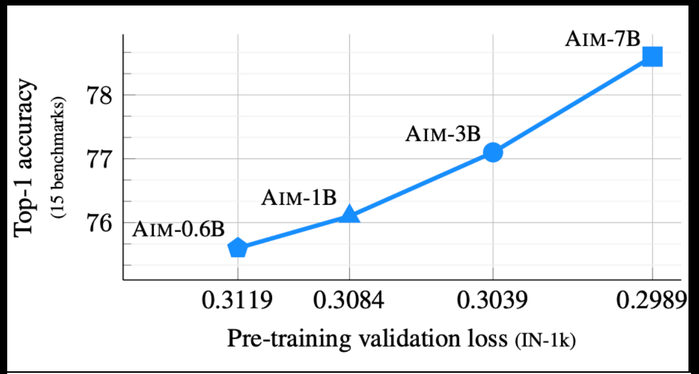
Image Encoder Pre-training

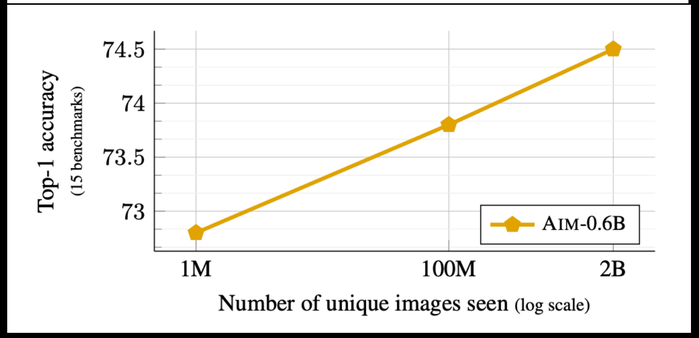
- We consider two types of pre-training methods to train image encoders: contrastive and reconstructive loss



CLIP







AIM

[1] Scalable Pre-training of Large Autoregressive Image Models, ICML 2024

Encoder Lesson

- We pre-train the image encoders (AIM & CLIP) with DFN-5B and VeCap-300M
- The highest native image resolution we can afford is 378px
- Perform further multimodal LLM pre-training with the pre-trained image encoders

	Setup				Results		
	Model	Arch.	Image Res.	Data	0-shot	4-shot	8-shot
Recon.	AIM_{600M}	ViT/600M	224	DFN-2B	36.6	56.6	60.7
Recon.	AIM_{1B}	ViT/1B	224	DFN-2B	37.9	59.5	63.3
Recon.	AIM_{3B}	ViT/3B	224	DFN-2B	38.9	60.9	64.9
Contrastive	$CLIP_{DFN+VeCap}$	ViT-L	224	DFN-5B+VeCap	36.9	58.7	62.2
Contrastive	$CLIP_{DFN}$	ViT-H	224	DFN-5B	37.5	57.0	61.4
Contrastive	$CLIP_{DFN+VeCap}$	ViT-H	224	DFN-5B+VeCap	37.5	60.0	63.6
Contrastive	$CLIP_{DFN+VeCap}$	ViT-L	336	DFN-5B+VeCap	39.9	62.4	66.0
Contrastive	$CLIP_{DFN+VeCap}$	ViT-H	336	DFN-5B+VeCap	40.5	**62.6**	66.3
Contrastive	$CLIP_{OpenAI}$	ViT-L	336	ImageText-400M	39.3	62.2	66.1
Contrastive	$CLIP_{DFN}$	ViT-H	378	DFN-5B	**40.9**	62.5	**66.4**

Table 1: MM1 pre-training ablation across different image encoders (with 2.9B LLM).

[1] Data Filtering Networks, 2023
[2] VeCLIP: Improving CLIP Training via Visual-enriched Captions, 2023

Encoder Lesson

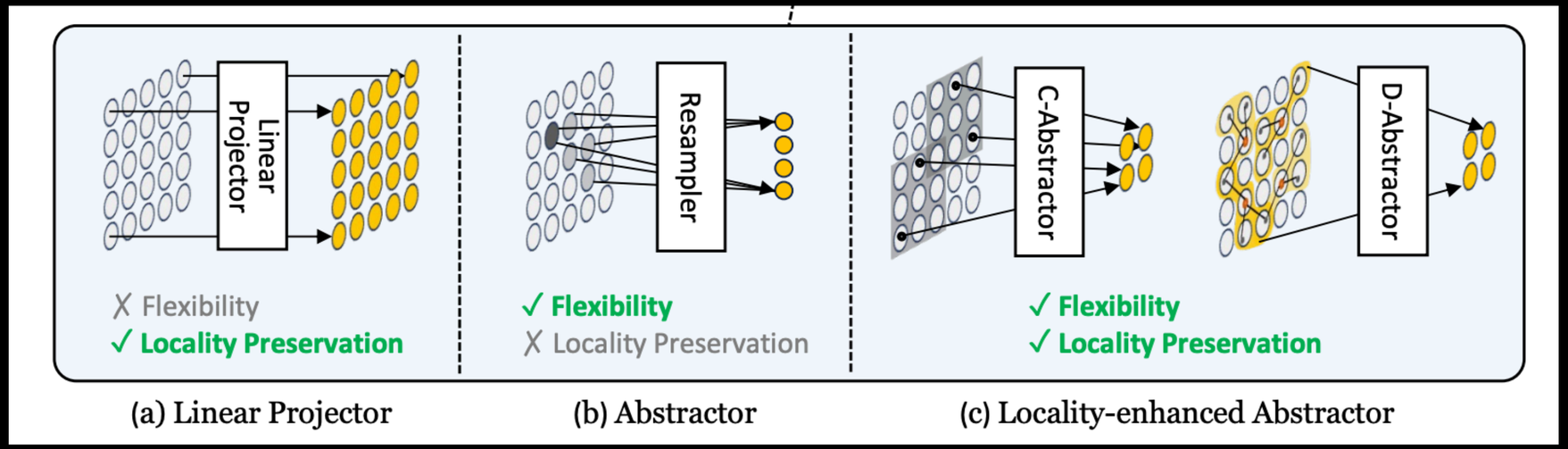
- AIM pre-trained image encoder shows great potential
- Image resolution has the highest impact, followed by model size

	Setup				Results		
	Model	Arch.	Image Res.	Data	0-shot	4-shot	8-shot
Recon.	AIM_{600M}	ViT/600M	224	DFN-2B	36.6	56.6	60.7
Recon.	AIM_{1B}	ViT/1B	224	DFN-2B	37.9	59.5	63.3
Recon.	AIM_{3B}	ViT/3B	224	DFN-2B	38.9	60.9	64.9
Contrastive	$CLIP_{DFN+VeCap}$	ViT-L	224	DFN-5B+VeCap	36.9	58.7	62.2
Contrastive	$CLIP_{DFN}$	ViT-H	224	DFN-5B	37.5	57.0	61.4
Contrastive	$CLIP_{DFN+VeCap}$	ViT-H	224	DFN-5B+VeCap	37.5	60.0	63.6
Contrastive	$CLIP_{DFN+VeCap}$	ViT-L	336	DFN-5B+VeCap	39.9	62.4	66.0
Contrastive	$CLIP_{DFN+VeCap}$	ViT-H	336	DFN-5B+VeCap	40.5	**62.6**	66.3
Contrastive	$CLIP_{OpenAI}$	ViT-L	336	ImageText-400M	39.3	62.2	66.1
Contrastive	$CLIP_{DFN}$	ViT-H	378	DFN-5B	**40.9**	62.5	**66.4**

Table 1: MM1 pre-training ablation across different image encoders (with 2.9B LLM).

VL Connector Lesson

- There are many VL connector (projection layers) choices
  - Linear layer, or simple MLP (LLaVA-1.5)
  - Average pooling (Emu2)
  - Q-former (InstructBLIP), Perceiver (Flamingo), Attention pooling (CoCa)
  - Convolutional mapping (Honeybee)



[1] Honeybee: Locality-enhanced Projector for Multimodal LLM, CVPR 2024

VL Connector Lesson

- Number of visual tokens and image resolution matters most, while the type of VL connector has little effect

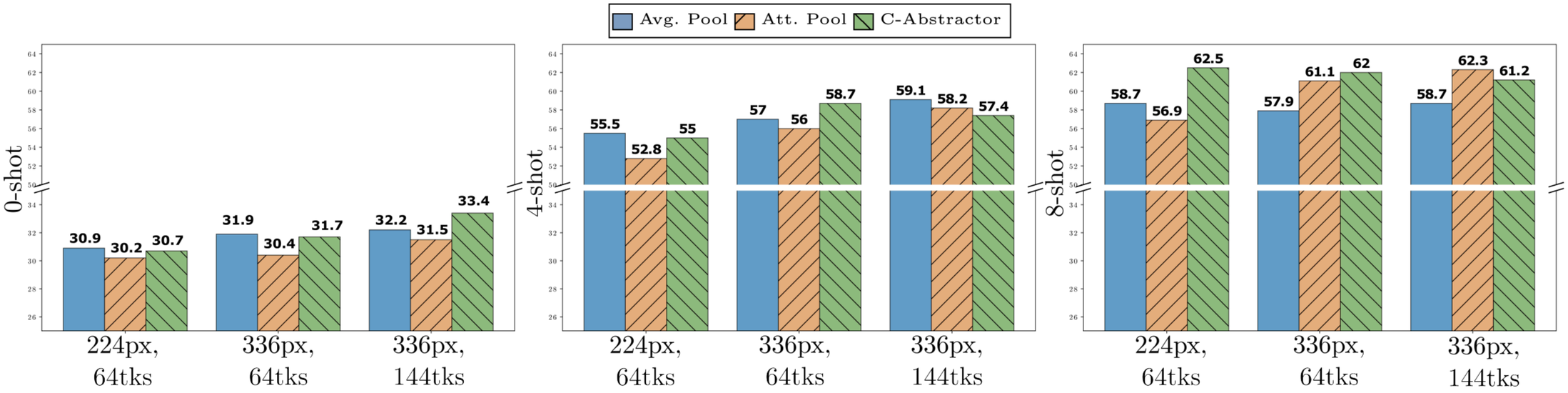


Fig. 4: 0-shot, 4-shot, and 8-shot ablations across different visual-language connectors for two image resolutions, and two image token sizes.

Pre-training Data Ablation

- Three types of data: Captioned Images, Interleaved Image-Text, Text-only Data

Data Type	Sources	Size
Captioned Images	CC3M [100], CC12M [13], HQIPT-204M [94], COYO [11], Web Image-Text-1B (Internal)	2B image-text pairs
Captioned Images (Synthetic)	VeCap [57]	300M image-text pairs
Interleaved Image-Text	OBELICS [58], Web Interleaved (Internal)	600M documents
Text-only	Webpages, Code, Social media, Books, Encyclopedic, Math	2T tokens

Table 2: List of datasets for pre-training multimodal large language models.

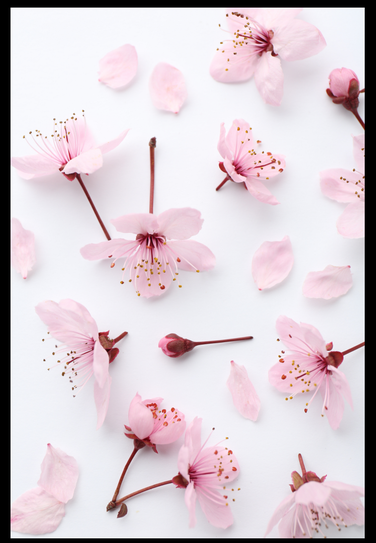
Captioned Images

- Image-text pairs
  - 2B web-mined image alt-text data
  - 300M model generated high-quality image captions (VeCap)

Large and Noisy !

Cute little black dachshund dog

Beautiful spring tree blossoms and petals on white background, flat lay



Paddle objects on blue turf

Fresh ripe red and green apples as background, top view

High-Quality VeCap Data

- Alt-text can be noisy, even after CLIP filtering
- Re-captioning: generating high-quality captions via an external model
- VeCap can be readily scaled to billion of images

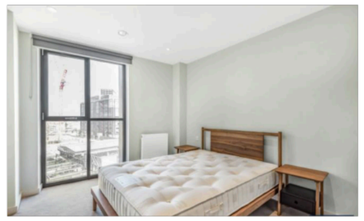
AltText:
112 Lafayette Dr.
LLM Rewrite:
112 Lafayette Drive.
VeCap:
A **red brick house** with a **white roof** and a **front porch**, surrounded by a **garden** at 112 Lafayette Dr.

AltText:
9 Misconceptions About Alcohol.
LLM Rewrite:
Debunking 9 Myths Surrounding Alcohol.
VeCap:
A **glass of beer** sits on a table with a **lit candle**, creating a cozy atmosphere.

AltText:
SURF IN, SURF OUT ~ LARGE condo, Steps to Beach + Parking.
LLM Rewrite:
Surf in, surf out – Spacious condo, just steps from the beach with parking.
VeCap:
A **small, modern home** with a **patio**, **outdoor furniture**, and a **stone wall**.

AltText:
Ring Car Cam.
LLM Rewrite:
Ring Automotive Camera.
VeCap:
A **hand** is holding a **smartphone** with Ring Car Cam, watching a car driving on a road.

AltText:
More than one million Australian households are already in mortgage stress.
LLM Rewrite:
Over a million Australian households are currently experiencing mortgage strain.
VeCap:
A **real estate sign** advertises a **house** in Australian for sale in a residential neighborhood, with a **tree** in the background.



AltText:
3 bedroom apartment to rent Croydon.
LLM Rewrite:
Croydon Rental: 3-Bedroom Apartment Available.
VeCap:
A bedroom with a **bed** with **white mattress**, a **window**, and a **wooden dresser**, providing a comfortable and well-organized space.

[1] VeCLIP: Improving CLIP Training via Visual-enriched Captions, 2023

Other High-Quality Image Caption Datasets

- Similar ideas have also been adopted in other works
  - DALL-E 3 for text-to-image generation
  - CapsFusion-120M
  - ShareGPT4V-PT, 1.2M image captions
    - First train a captioner on 100K GPT4v generated captions
  - LLaVA-ReCap via LLaVA-NeXT-34B
  - ReCap-DataComp-1B via LLaMA3-8B empowered LLaVA-1.5

[1] CapsFusion: Rethinking Image-Text Data at Scale, CVPR 2024
[2] ShareGPT4V: Improving Large Multi-Modal Models with Better Captions, 2023
[3] What IfWe Recaption Billions of Web Images with LLaMA-3?, 2024

Interleaved Image-Text Data

- 600M interleaved documents with 1B images and 500B text tokens
- Built from 3B HTML files with image filtering and de-duplication

Expand LLM-style data to images !

Example 1:

H2O2 (equation 2) in cloud droplets has been postulated to be the most important pathway for conversion of SO2 to NSS

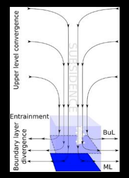
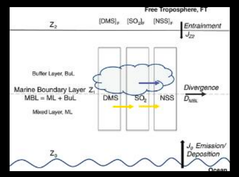


Figure 1 Open in figure view divergence er PowerPoint Model of subsidence, entrainment, and in the central Pacific MBL and FT system.



Example 2:

Christstollen is a traditional cake that is eaten at Christmas.

It takes the form of a rectangle or trapezium and is usually covered in icing sugar.

The main ingredients include raisins, sultanas, currants and citrus peel…

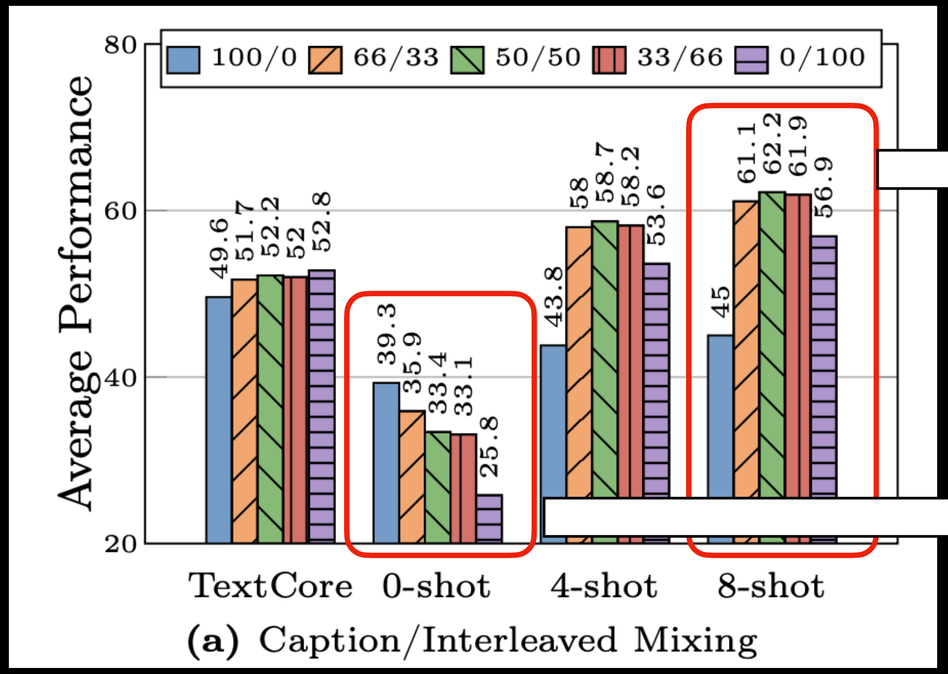
Interleaved Image-Text Data

- Similar datasets in the community
  - Early days: M3W (Flamingo) and Kosmos-1
  - Open-sourced: MMC4 and OBELICS, OmniCorpus
- MMC4 employed a retrieval strategy, inserting images into text sequence based on CLIP score
- OBELICS maintained the natural layout of the source webpage
- Model-based filtering has a great opportunity to further enhance the quality of the datasets

[1] Multimodal C4: An Open, Billion-scale Corpus of Images Interleaved with Text, 2023
[2] OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents, 2023
[3] OmniCorpus: A Unified Multimodal Corpus of 10 Billion-Level Images Interleaved with Text, 2024

Data Lesson 1

- Interleaved data is instrumental for few-shot and text-only performance, while captioning data lifts zero-shot performance

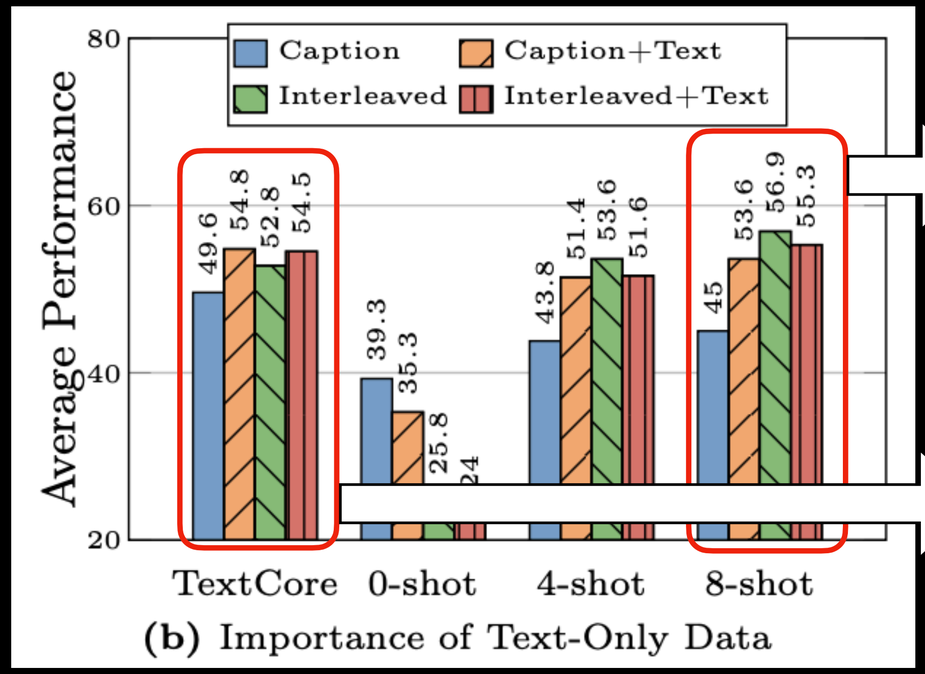


Blue (45) vs Green (62.2)

Blue (39.3) vs Green (33.4)

Data Lesson 2

- Text-only data helps with few-shot and text-only performance



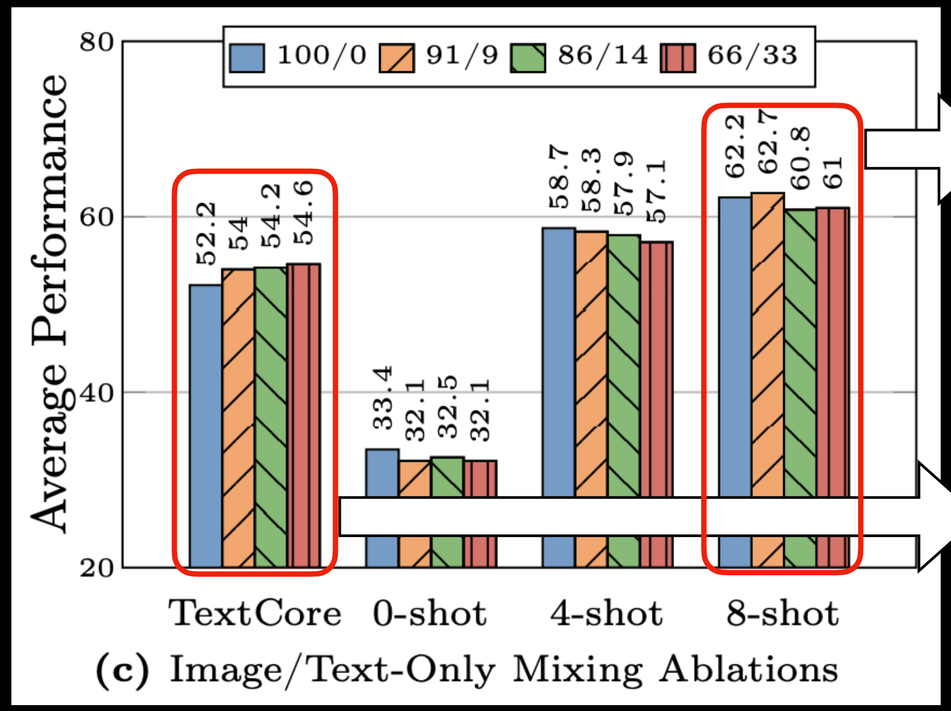
Blue (45) vs Orange (53.6) vs Green (56.9)

Blue (49.6) vs Orange (54.8)

Green (52.8) vs Red (54.5)

Data Lesson 3

- Careful mixture of image and text data can yield optimal multimodal performance and retrain strong text performance

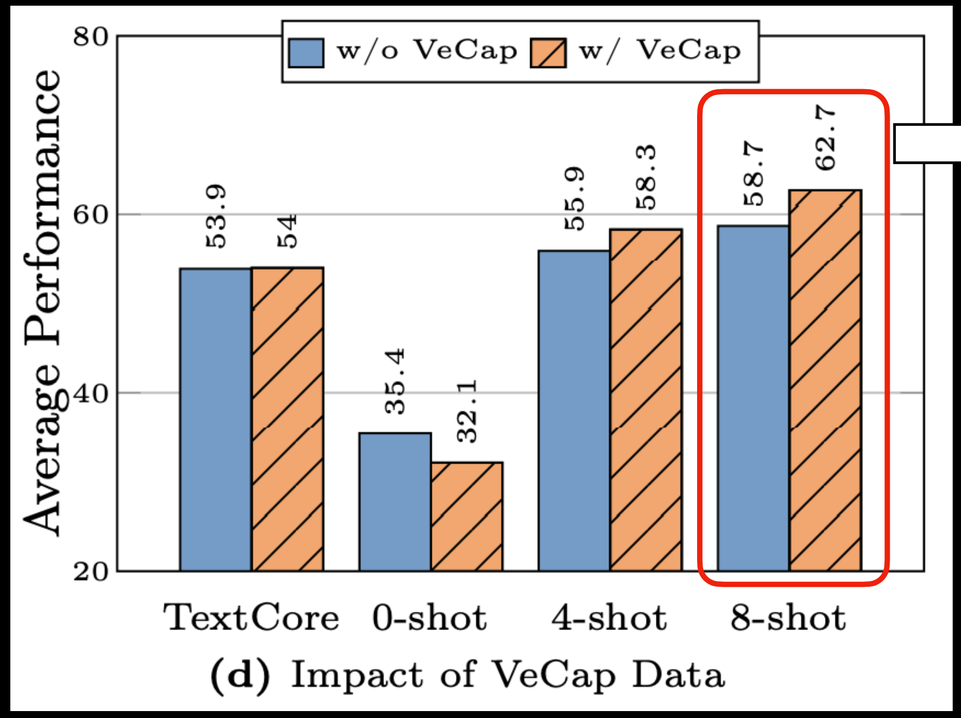


However, increasing text-only data can hurt multimodal performance

The more text-only data, the better for text-only performance

Data Lesson 4

- Synthetic data helps with few-shot learning



Blue (58.7) vs Orange (62.7)

Final Model and Training Recipe

Final Model and Training Recipe

- For this model family, we did not grow the image encoder size (600M) as LLM becomes larger
- Dense model scaling: 3B -> 7B -> 30B

– **Image Encoder**: Motivated by the importance of image resolution, we use a ViT-H [27] model with 378×378 resolution, pre-trained with a CLIP objective on DFN-5B [31].
– **Vision-Language Connector**: As the number of visual tokens is of highest importance, we use a VL connector with 144 tokens. The actual architecture seems to matter less, we opt for C-Abstractor [12].
– **Data**: In order to maintain both zero- and few-shot performance, we use the following careful mix of 45% interleaved image-text documents, 45% image-text pair documents, and 10% text-only documents.

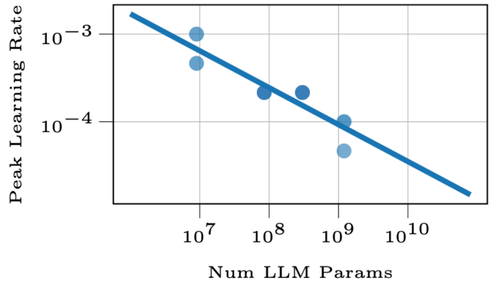
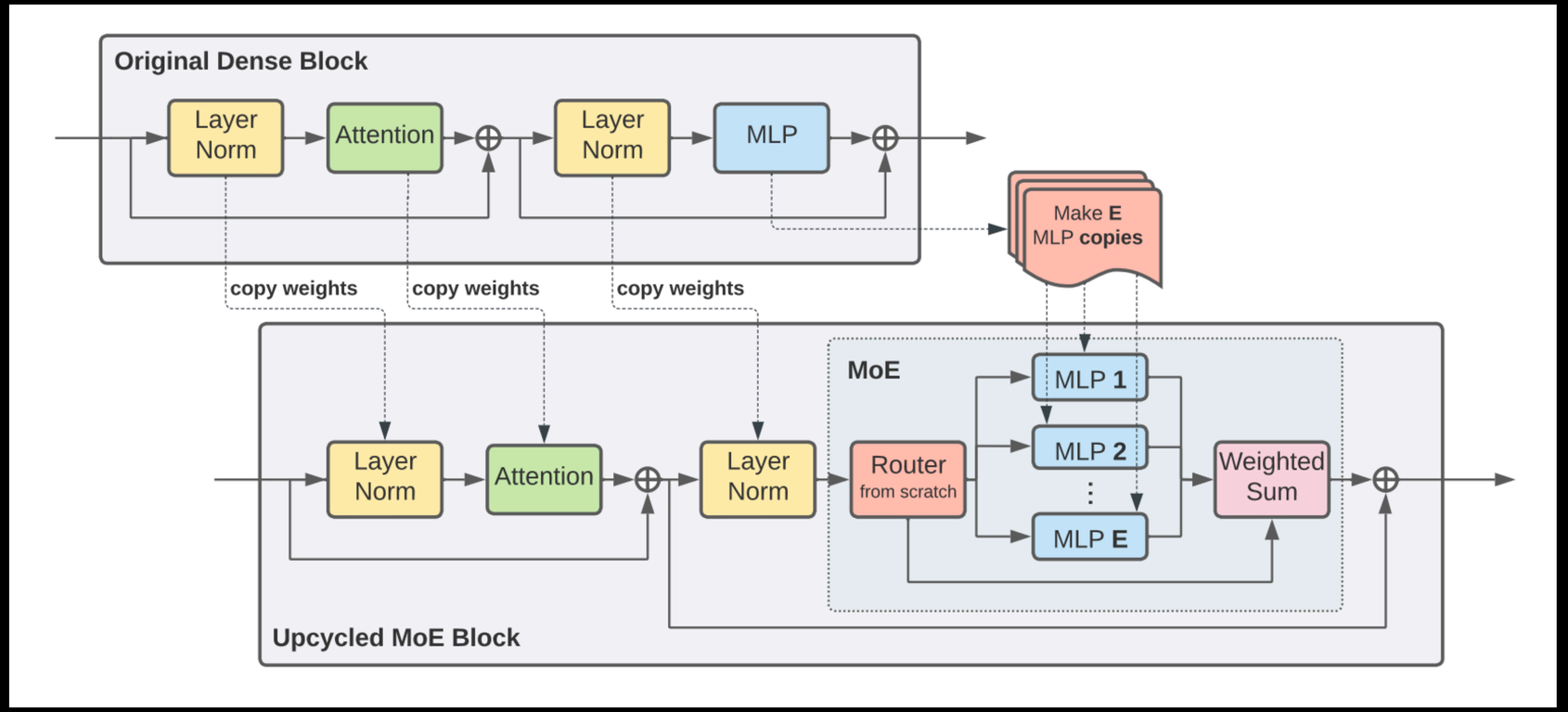


Fig. 6: Optimal peak learning rate as a function of model size. The data points represent experiments that achieved close-to-optimal 8-shot performance for their associated model size.

Scaling via Mixture-of-Experts (MoE)

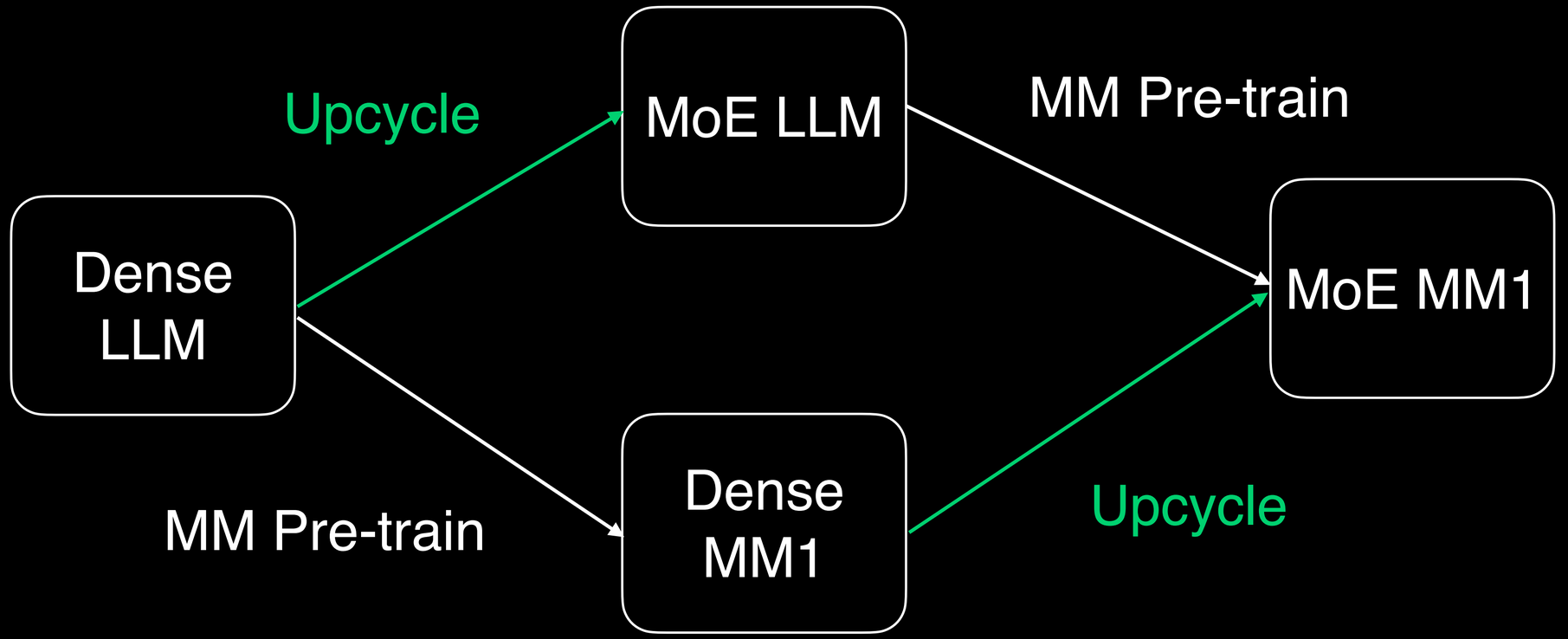
- MoE scales the total # of params while keeping the activated # of params constant
- We explore scaling the dense model by adding more experts in the FFN layers



[1] Sparse upcycling: Training mixture-ofexperts from dense checkpoints, ICLR 2023

Sparse Upcycling

- 3B-MoE with 64 experts, sparse layer in every-2 layers (64B in total)
- 7B-MoE with 32 experts, sparse layer in every-4 layers (47B in total)



Pre-training Results

- MM1 is mainly compared with Flamingo and Emu2
- SOTA few-shot results across model sizes
- Increasing number of shots keeps boosting performance
- Few-shot is more meaningful than zero-shot
- In-context learning is akin to instruction following, as the demonstrations can be considered as vivid instructions

Model	Shot	Captioning			Visual Question Answering			
		COCO	NoCaps	TextCaps	VQAv2	TextVQA	VizWiz	OKVQA
MM1-3B Model Comparisons								
Flamingo-3B [3]	0†	73.0	–	–	49.2	30.1	28.9	41.2
	8	90.6	–	–	55.4	32.4	38.4	44.6
MM1-3B	0	73.5	55.6	63.3	46.2	29.4	15.6	26.1
	8	**114.6**	**104.7**	**88.8**	**63.6**	**44.6**	**46.4**	**48.4**
MM1-7B Model Comparisons								
IDEFICS-9B [58]	0†	46.0*	36.8	25.4	50.9	25.9	35.5	38.4
	8	97.0*	86.8	63.2	56.4	27.5	40.4	47.7
Flamingo-9B [3]	0†	79.4	–	–	51.8	31.8	28.8	44.7
	8	99.0	–	–	58.0	33.6	39.4	50.0
Emu2-14B [105]	0†	–	–	–	52.9	–	34.4	42.8
	8	–	–	–	59.0	–	43.9	–
MM1-7B	0	76.3	61.0	64.2	47.8	28.8	15.6	22.6
	8	**116.3**	**106.6**	**88.2**	**63.6**	**46.3**	**45.3**	**51.4**
MM1-30B Model Comparisons								
IDEFICS-80B [58]	0†	91.8*	65.0	56.8	60.0	30.9	36.0	45.2
	8	114.3*	105.7	77.6	64.8	35.7	46.1	55.1
	16	116.6*	107.0	81.4	65.4	36.3	48.3	56.8
Flamingo-80B [3]	0†	84.3	–	–	56.3	35.0	31.6	50.6
	8	108.8	–	–	65.6	37.3	44.8	57.5
	16	110.5	–	–	66.8	37.6	48.4	57.8
Emu2-37B [105]	0	–	–	–	33.3	26.2	40.4	26.7
	8	–	–	–	67.8	49.3	54.7	54.1
	16	–	–	–	68.8	50.3	57.0	57.1
MM1-30B	0	70.3	54.6	64.9	48.9	28.2	14.5	24.1
	8	123.1	111.6	92.9	70.9	49.4	49.9	58.3
	16	**125.3**	**116.0**	**97.6**	**71.9**	**50.6**	**57.9**	**59.3**

Supervised Fine-Tuning

Instruction Tuning Data

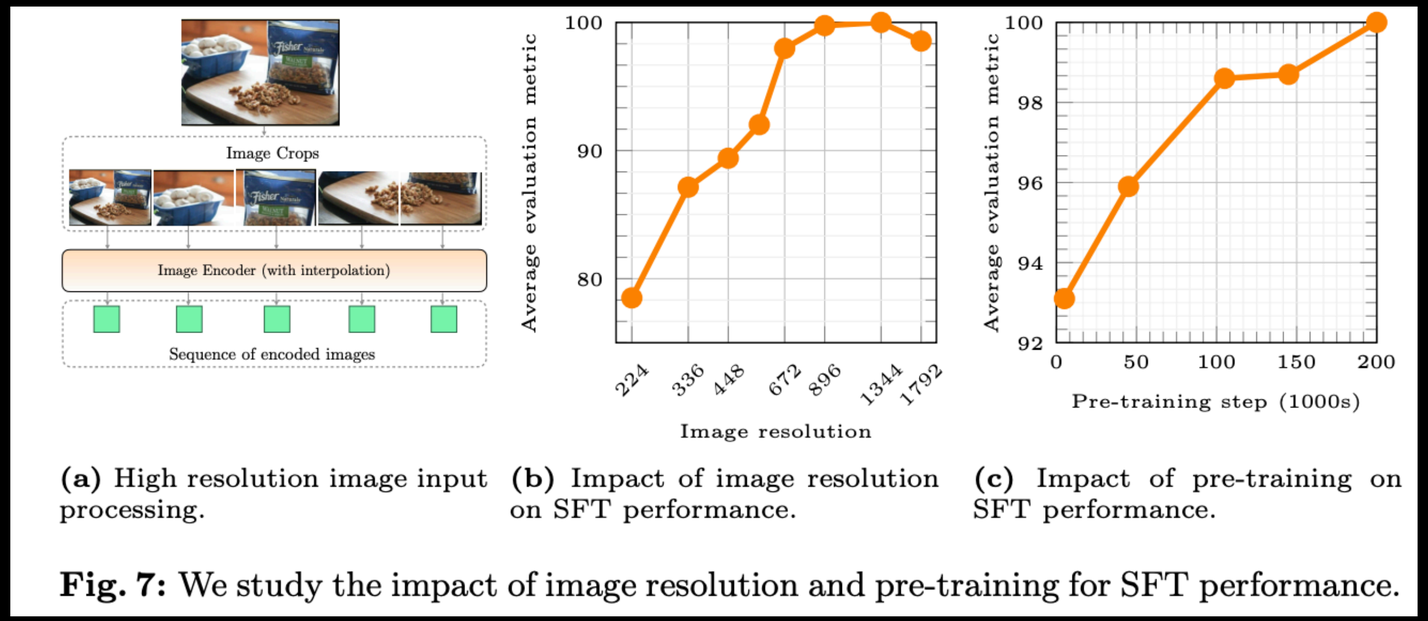
- We follow the setup in LLaVA-NeXT for SFT data mixture
- Order of ~1.45 M, diverse mixture for generalized instruction following
  - Long-form detailed image descriptions (e.g., ShareGPT4V)
  - Complex reasoning, conversational QA (e.g., LLaVA)
  - Academic datasets targeting general image understanding, OCR, knowledge, and more
  - Text-only SFT data (only 1.3%, need more)

Datasets	Size	Prompting Strategy
Text-only SFT	13k	–
LLaVA-Conv [76]	57k	–
LLaVA-Complex [76]	77k	
ShareGPT-4V [15]	102k	
VQAv2 [38]	83k	"Answer the question using a single word or phrase."
GQA [46]	72k	
OKVQA [82]	9k	
OCRVQA [86]	80k	
DVQA [51]	200k	
ChartQA [83]	18k	
AI2D [52]	3k	
DocVQA [85]	39k	
InfoVQA [84]	24k	
A-OKVQA [98]	66k	"Answer with the option's letter from the given choices directly."
COCO Captions [18]	83k	Sample from a pre-generated prompt list, *e.g.*, "Provide a brief description of the given image."
TextCaps [103]	22k	
SynthDog-EN [53]	500k	Sample from a pre-generated prompt list, *e.g.*, "Please transcribe all the text in the picture."
Total	**1.45M**	–

Table 5: List of datasets used for supervised fine-tuning.

Scaling to Higher-Resolution

- Cannot be affordable during pre-training, but can be done during SFT
- Positional embedding interpolation (from 378px to 672px) and Sub-image decomposition (from 672px to 1344px)
- Especially helpful for text-rich image understanding

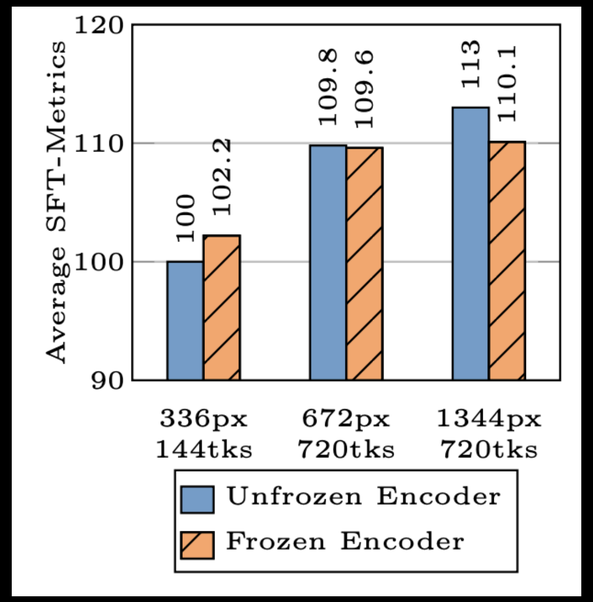


(a) High resolution image input processing.

(b) Impact of image resolution on SFT performance.

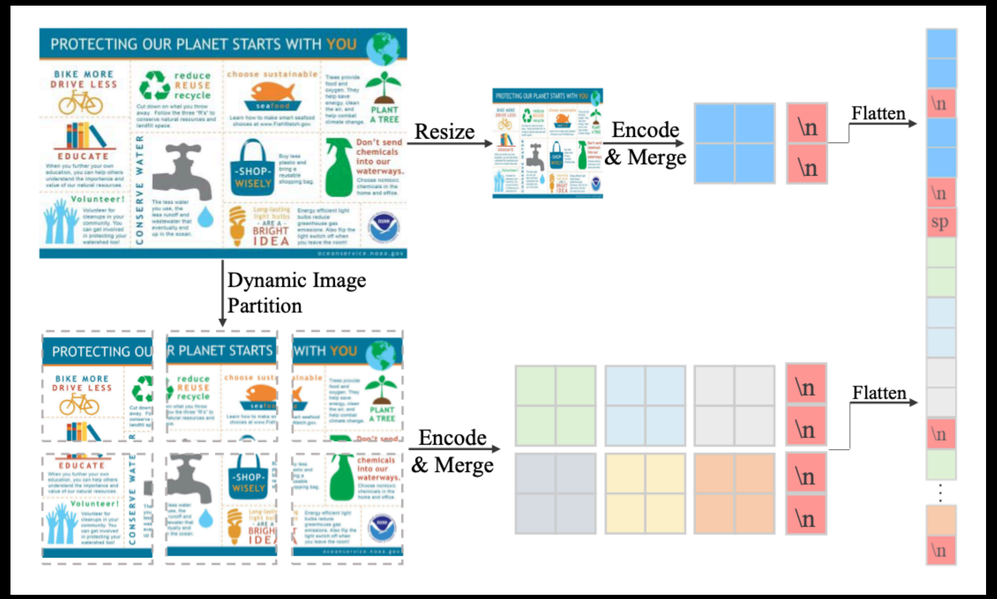
(c) Impact of pre-training on SFT performance.

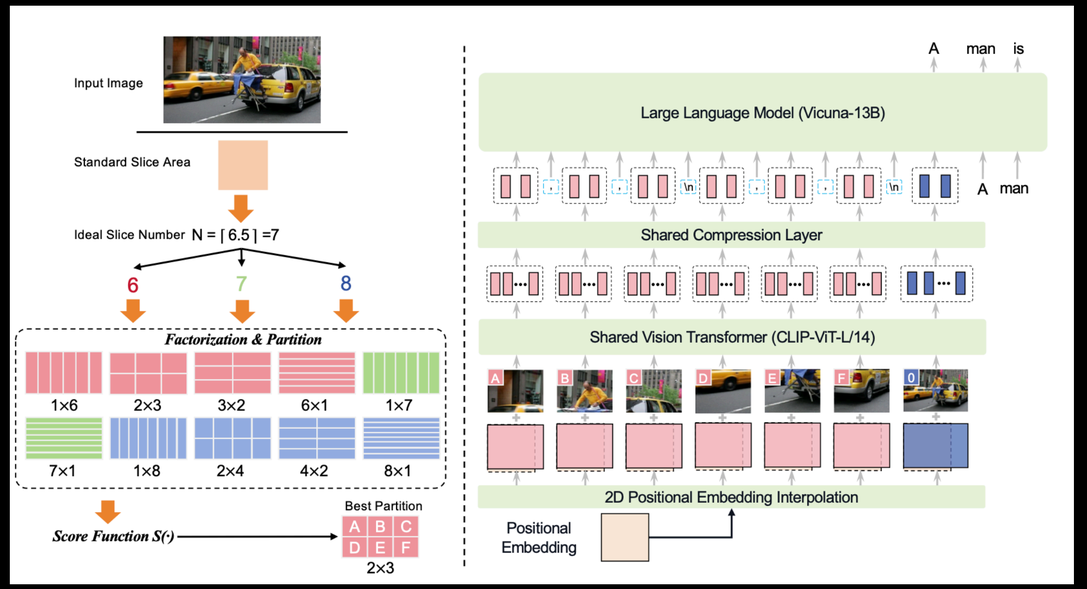
Fig. 7: We study the impact of image resolution and pre-training for SFT performance.



Scaling to Higher-Resolution

- The image split method is directly borrowed from other papers, and there is a growing body of literature on this topic
- One notable example is InternLM-XComposer2-4KHD, with dynamic resolution and automatic patch configuration, supporting resolution up to 4K





[1] InternLM-XComposer2-4KHD: A Pioneering Large Vision-Language Model Handling Resolutions from 336 Pixels to 4K HD, 2024
[2] LLaVA-UHD: an LMM Perceiving Any Aspect Ratio and High-Resolution Images, 2024

Comparison with SOTA

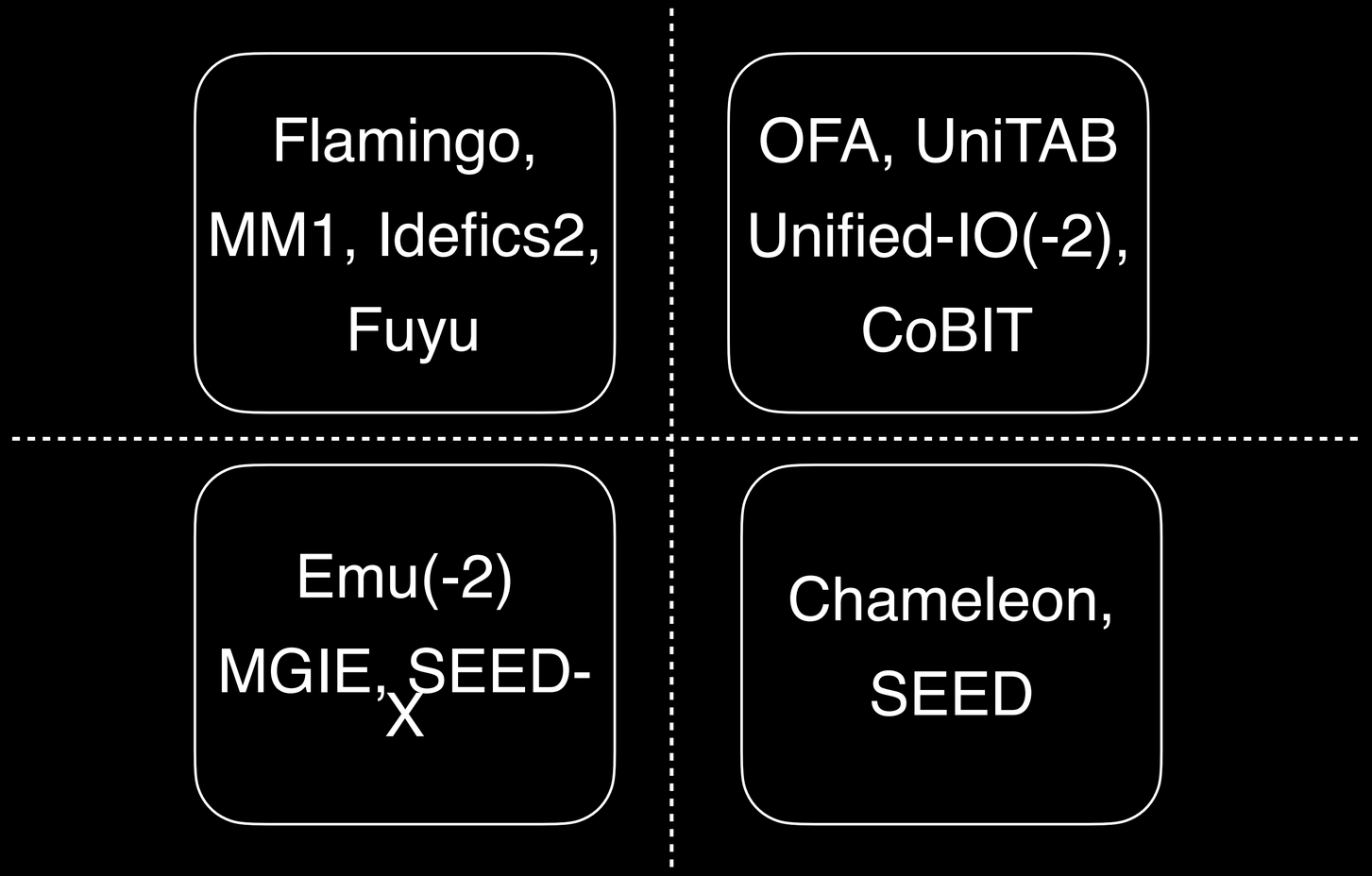
- SOTA on 3B back then, now lagging behind Phi-3-Vision
- MoE has great potential
- Competitive with LLaVA-NeXT-34B, and also support multi-image reasoning and few-shot-prompting
- 8-shot prompting on MathVista: 39.4 -> 44.4
- MMMU score needs enhancement

Model	VQA^{v2}	VQA^{T}	SQA^{I}	MMMU	MathV	MME^{P}	MME^{C}	MMB	SEED	POPE	$LLaVA^{W}$	MM-Vet
3B Model Comparison												
MobileVLM [20]	–	47.5	61.0	–/–	–	1288.9	–	59.6	–/–	84.9	–	–
LLaVA-Phi [135]	71.4	48.6	68.4	–/–	–	1335.1	–	59.8	–/–	85.0	–	28.9
Imp-v1 [99]	79.45	59.38	69.96	–/–	–	1434.0	–	66.49	–	88.02	–	33.1
TinyLLaVA [133]	79.9	59.1	69.1	–/–	–	1464.9	–	66.9	–/–	86.4	75.8	32.0
Bunny [42]	79.8	–	70.9	38.2/33.0	–	1488.8	289.3	68.6	62.5/–	86.8	–	–
Gemini Nano-2 [106]	67.5	65.9	–	32.6/–	30.6	–	–	–	–	–	–	–
MM1-3B-Chat	82.0	71.9	69.4	33.9/33.7	32.0	1482.5	279.3	67.8	63.0/68.8	87.4	72.1	43.7
MM1-3B-MoE-Chat	82.5	72.9	76.1	38.6/35.7	32.6	1469.4	303.1	70.8	63.9/69.4	87.6	76.8	42.2
7B Model Comparison												
InstructBLIP-7B [24]	–	50.1	60.5	–/–	25.3	–	–	36.0	53.4/–	–	60.9	26.2
Qwen-VL-Chat-7B [5]	78.2	61.5	68.2	35.9/32.9	–	1487.5	360.7	60.6	58.2/65.4	–	–	–
LLaVA-1.5-7B [74]	78.5	58.2	66.8	–/–	–	1510.7	316.1	64.3	58.6/66.1	85.9	63.4	31.1
ShareGPT4V-7B [15]	80.6	60.4	68.4	–/–	–	1567.4	376.4	68.8	–/–	–	72.6	–
LVIS-Ins4V-7B [113]	79.6	58.7	68.3	–/–	–	1528.2	–	66.2	60.6/–	86.0	67.0	31.5
VILA-7B [71]	79.9	64.4	68.2	–/–	–	1531.3	–	68.9	61.1/–	85.5	69.7	34.9
SPHINX-Intern2 [36]	75.5	–	70.4	–/–	35.5	1260.4	294.6	57.9	68.8/–	86.9	57.6	36.5
LLaVA-NeXT-7B [75]	81.8	64.9	70.1	35.8/–	34.6	1519	332	67.4	–/70.2	86.53	81.6	43.9
MM1-7B-Chat	82.8	72.8	72.6	37.0/35.6	35.9	1529.3	328.9	72.3	64.0/69.9	86.6	81.5	42.1
MM1-7B-MoE-Chat	83.4	73.8	74.4	40.9/37.9	40.9	1597.4	394.6	72.7	65.5/70.9	87.8	84.7	45.2
30B Model Comparison												
Emu2-Chat-37B [105]	84.9	66.6	–	36.3/34.1	–	–	–	–	62.8/–	–	–	48.5
CogVLM-30B [114]	83.4	68.1	–	32.1/30.1	–	–	–	–	–	–	–	56.8
LLaVA-NeXT-34B [75]	83.7	69.5	81.8	51.1/44.7	46.5	1631	397	79.3	–/75.9	87.73	89.6	57.4
MM1-30B-Chat	83.7	73.5	81.0	44.7/40.3	$39.4^{\dagger}$	1637.6	431.4	75.1	65.9/72.1	87.6	89.3	48.7
Gemini Pro [106]	71.2	74.6	–	47.9/–	45.2	–	436.79	73.6	–/70.7	–	–	64.3
Gemini Ultra [106]	77.8	82.3	–	59.4/–	53.0	–	–	–	–	–	–	–
GPT4V [1]	77.2	78.0	–	56.8/55.7	49.9	–	517.14	75.8	67.3/69.1	–	–	67.6

How about other model architecture designs?

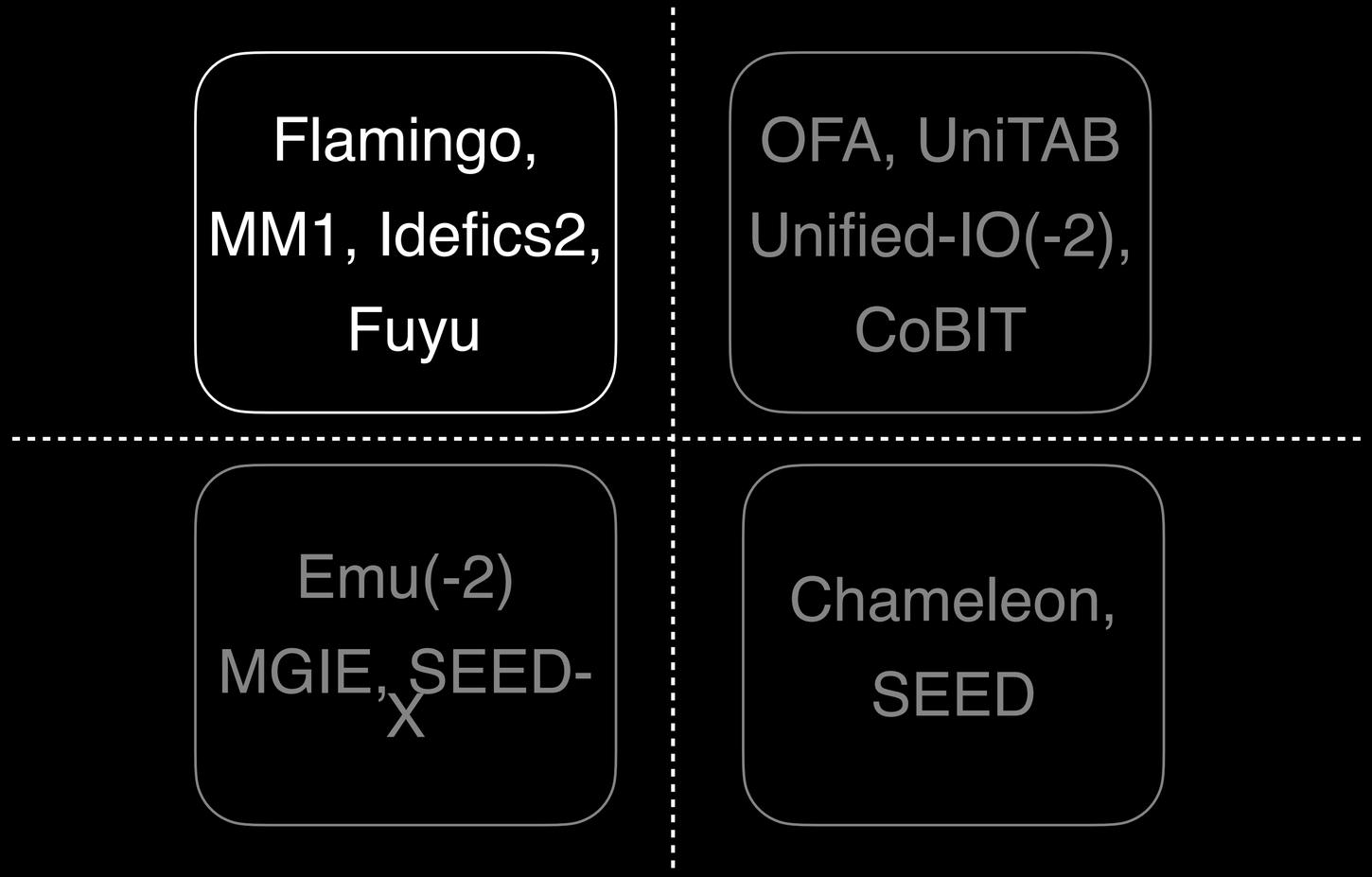
A Summary of Other Types of Model Architectures

- This also relates to the unification of image understanding and generation (and potentially also grounding)



A Summary of Other Types of Model Architectures

- This is what we have covered so far, continuous image features in, and text out, no image loss during training



Fuyu: No Image Encoder

- Pros: a vanilla decoder-only arch with no specialized image encoder
- Cons: model training can be challenging to reach SOTA performance

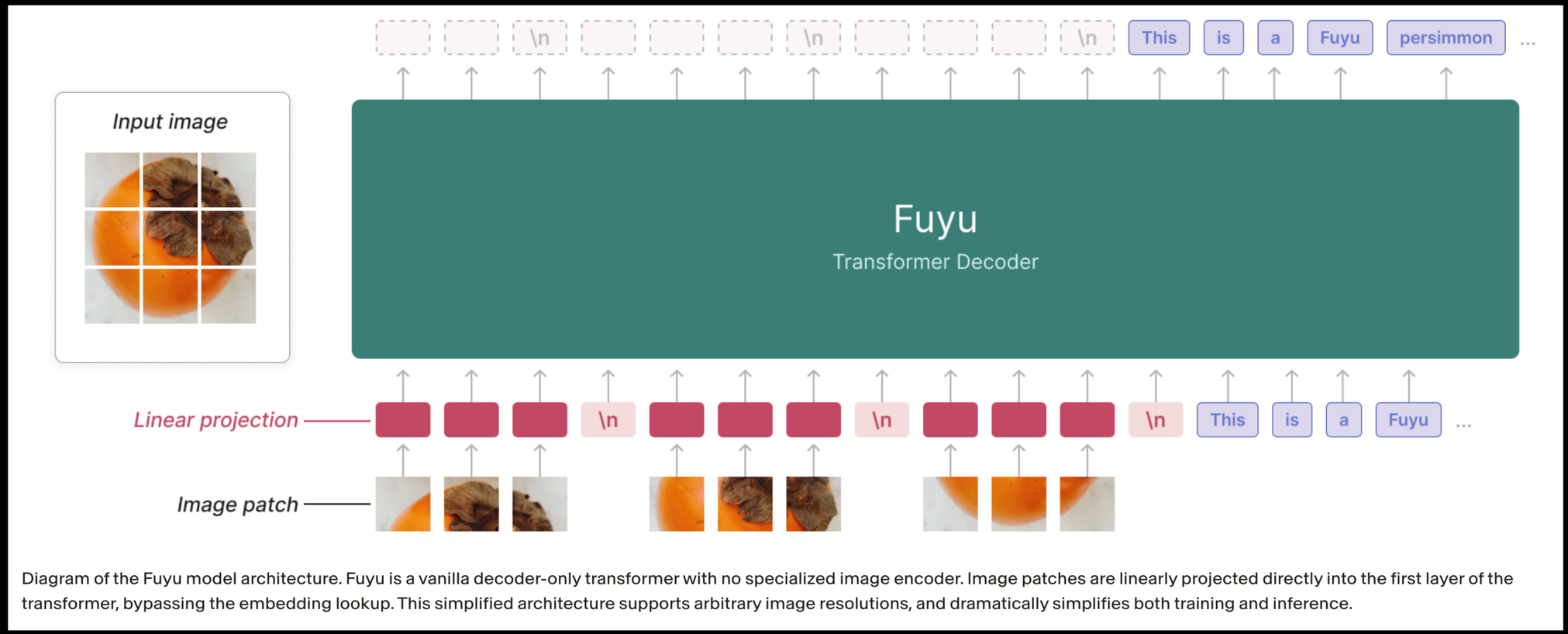
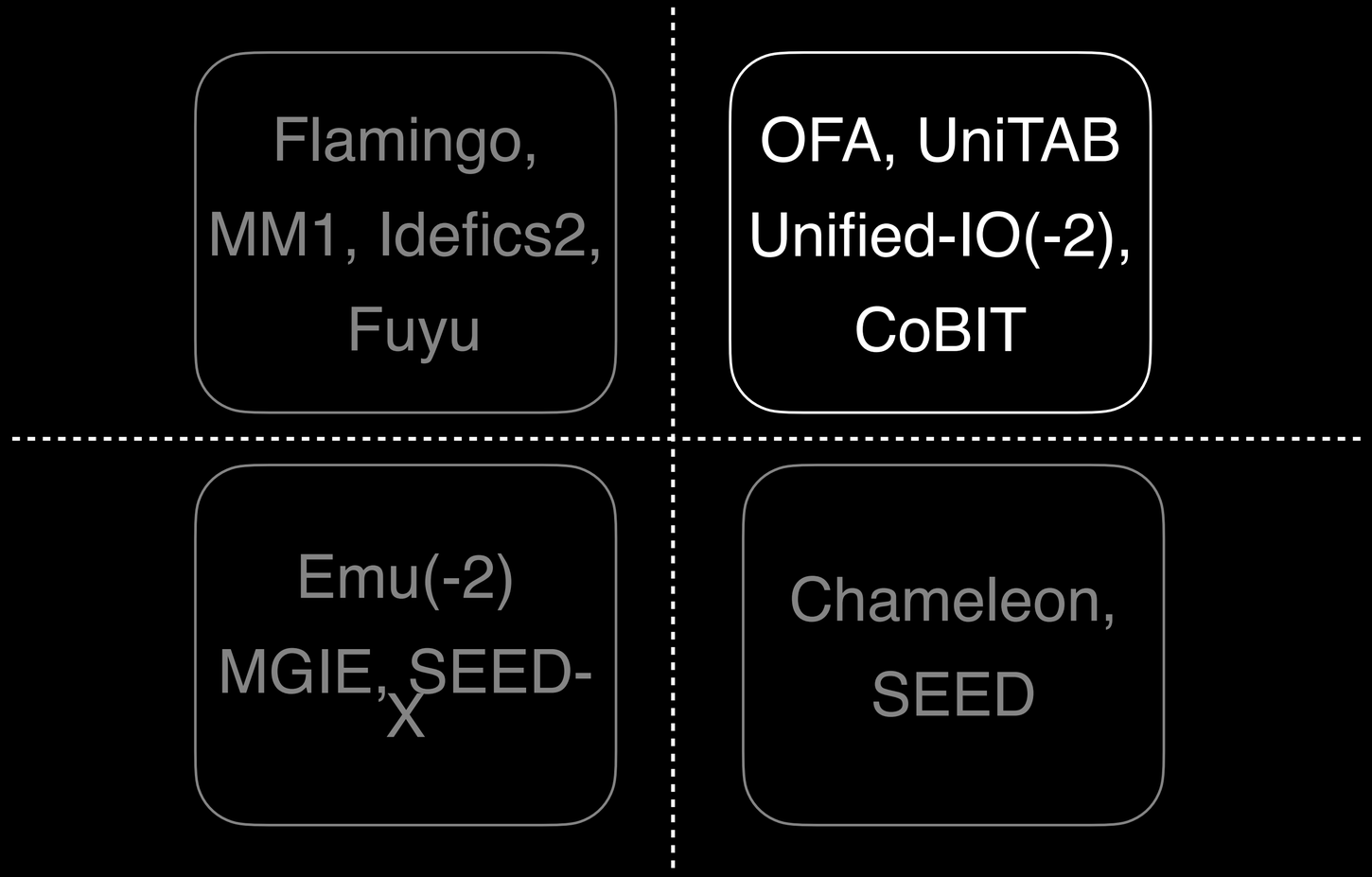


Diagram of the Fuyu model architecture. Fuyu is a vanilla decoder-only transformer with no specialized image encoder. Image patches are linearly projected directly into the first layer of the transformer, bypassing the embedding lookup. This simplified architecture supports arbitrary image resolutions, and dramatically simplifies both training and inference.

[1] https://www.adept.ai/blog/fuyu-8b/

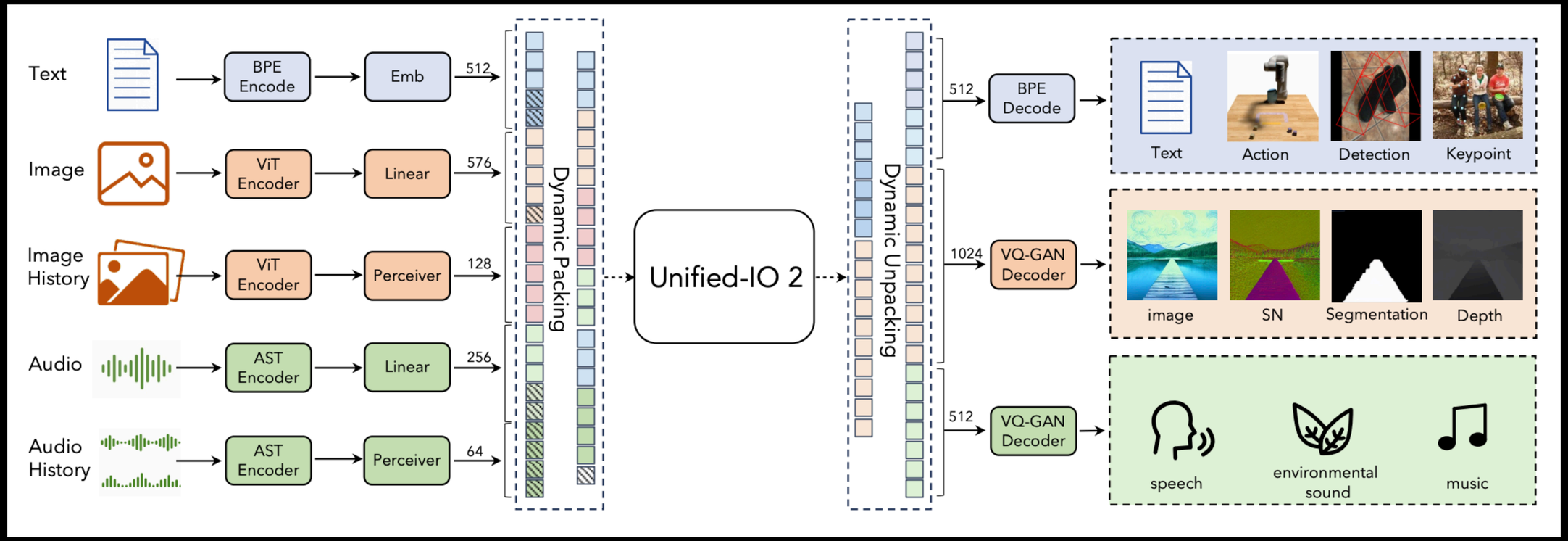
A Summary of Other Types of Model Architectures

- Continuous image features in, text and discrete image tokens out; there is cross-entropy loss on both image and text tokens



Using Unified-IO-2 as An Example

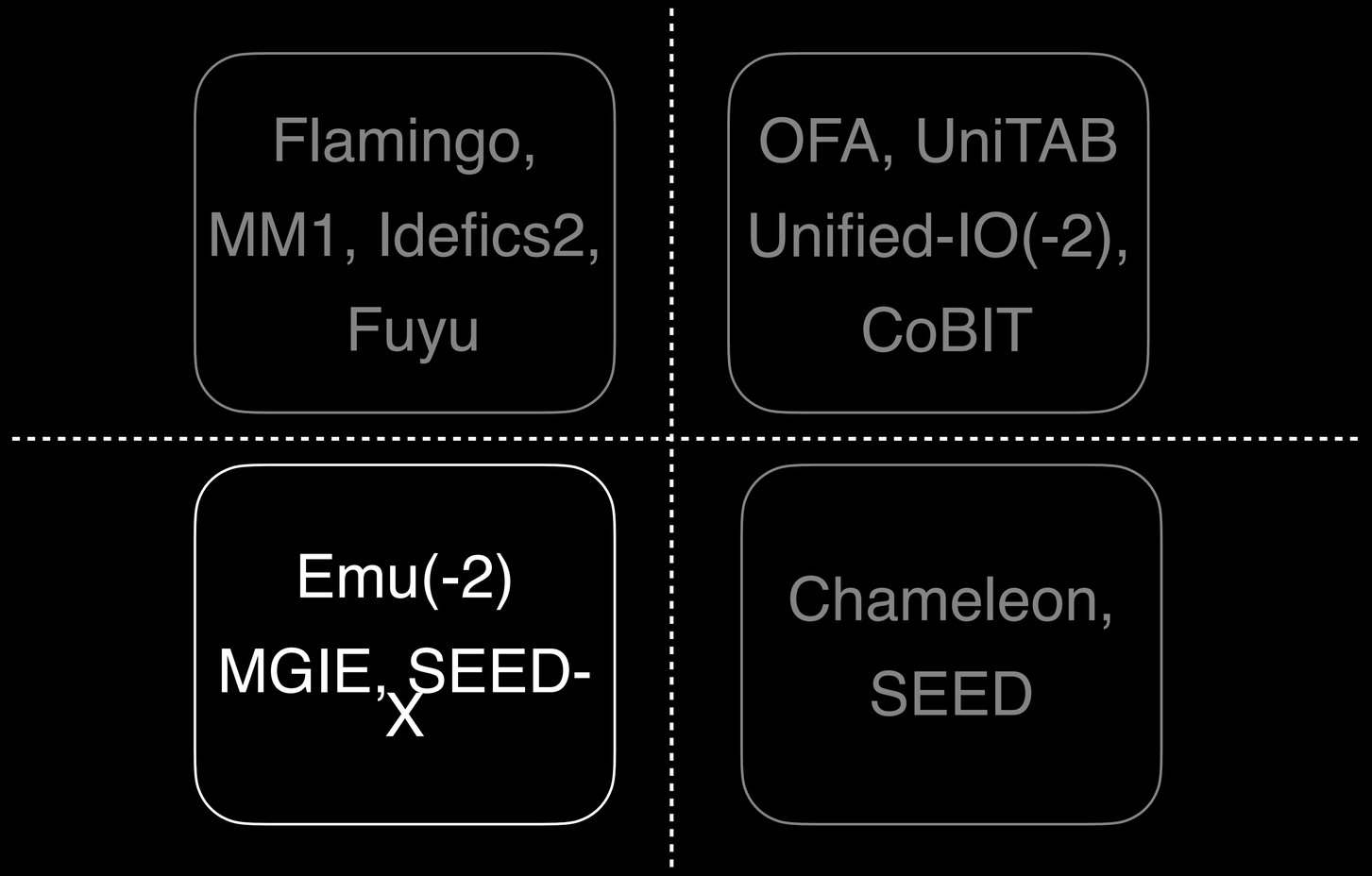
- Pros: unified model architecture across modalities and tasks
- Cons: how to leverage LLM is less clear, and the VQ-GAN tokenizer can be the bottleneck, and less popular nowadays compared with diffusion models



[1] Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action, 2023

A Summary of Other Types of Model Architectures

- Continuous image features in, text and continuous image features out, which is further connected to a diffusion model; L1 loss on image features



Using Emu2 as An Example

- Pros: maintains the strong image understanding capability
- Cons: The mixed use of X-entropy and L1 loss for training

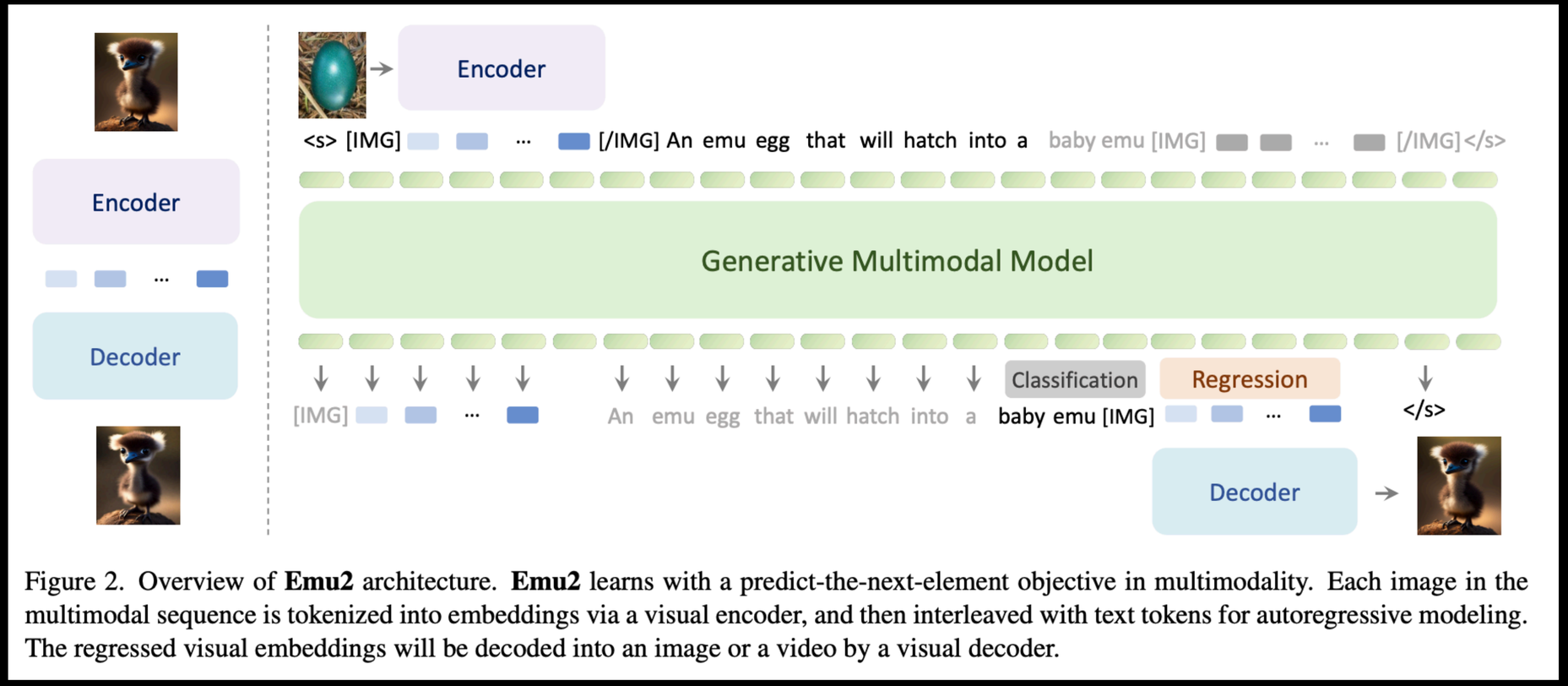
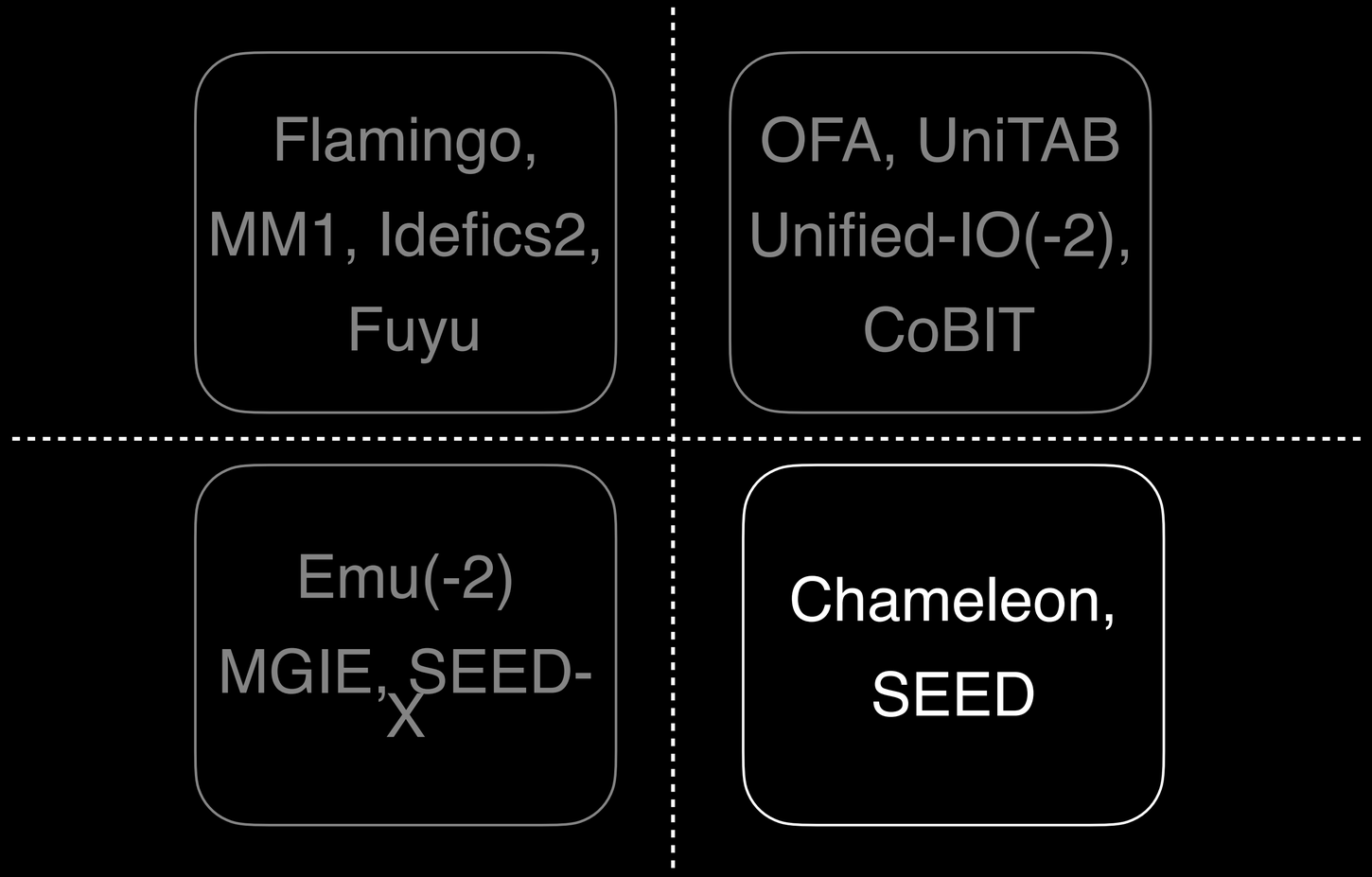


Figure 2. Overview of **Emu2** architecture. **Emu2** learns with a predict-the-next-element objective in multimodality. Each image in the multimodal sequence is tokenized into embeddings via a visual encoder, and then interleaved with text tokens for autoregressive modeling. The regressed visual embeddings will be decoded into an image or a video by a visual decoder.

[1] Generative Multimodal Models are In-Context Learners, CVPR 2024

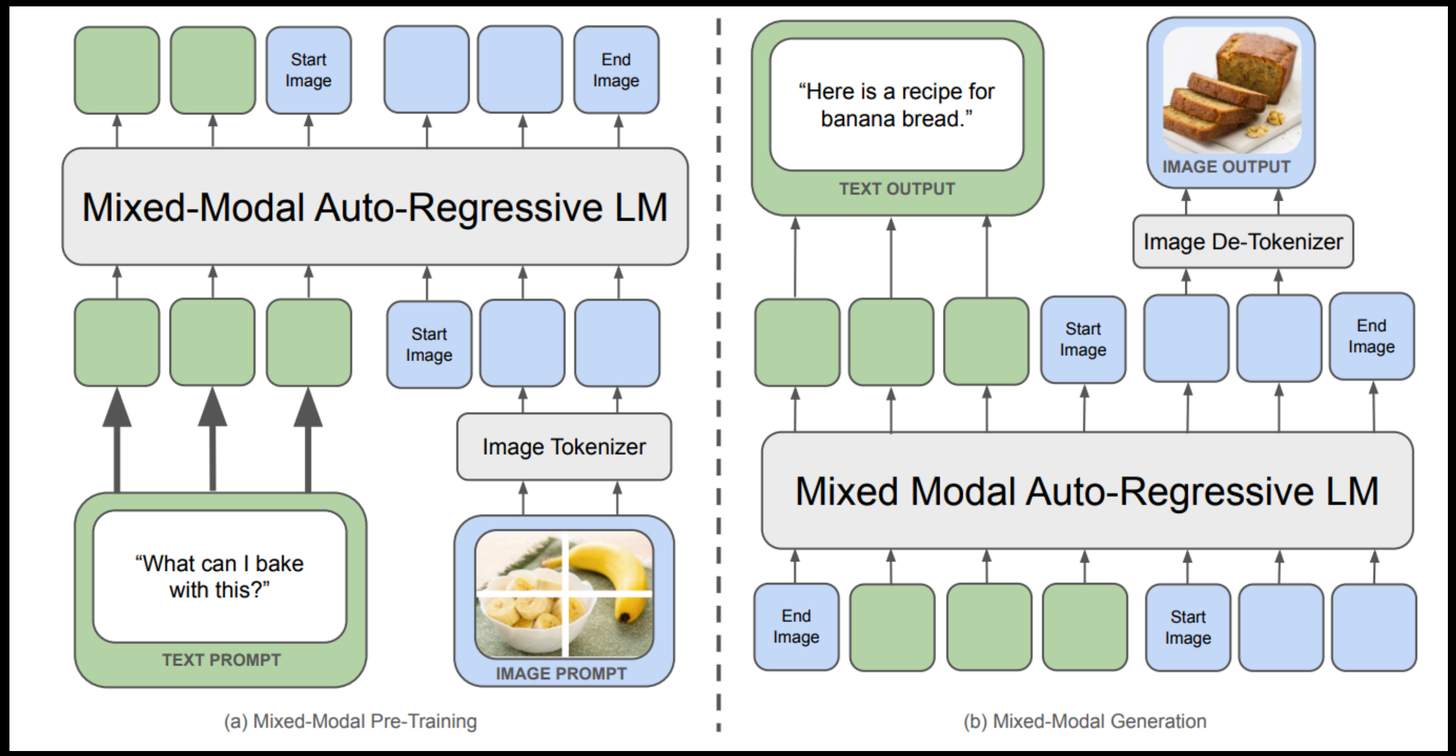
A Summary of Other Types of Model Architectures

- Discrete image and text tokens in and out



LLM for Unified Multimodal Understanding and Generation

- CM3 series of work: CM3, CM3Leon, and Chameleon
- However, the scores on multimodal understanding benchmarks are not strong
- Discrete image tokens trained via VQ-GAN like methods can be less informative, acting as an information bottleneck



[1] CM3: A Causal Masked Multimodal Model of the Internet, 2022
[2] Scaling Autoregressive Multi-Modal Models: Pretraining and Instruction Tuning, 2023
[3] Chameleon: Mixed-Modal Early-Fusion Foundation Models, 2024

Making Your Visual Tokenizers More Semantic

- Visual tokenizer is the key, making it more semantic can be helpful
- But still, compared with LLaVA-like models, the image understanding results are not ideal, and SEED-X chose back to use continuous image features

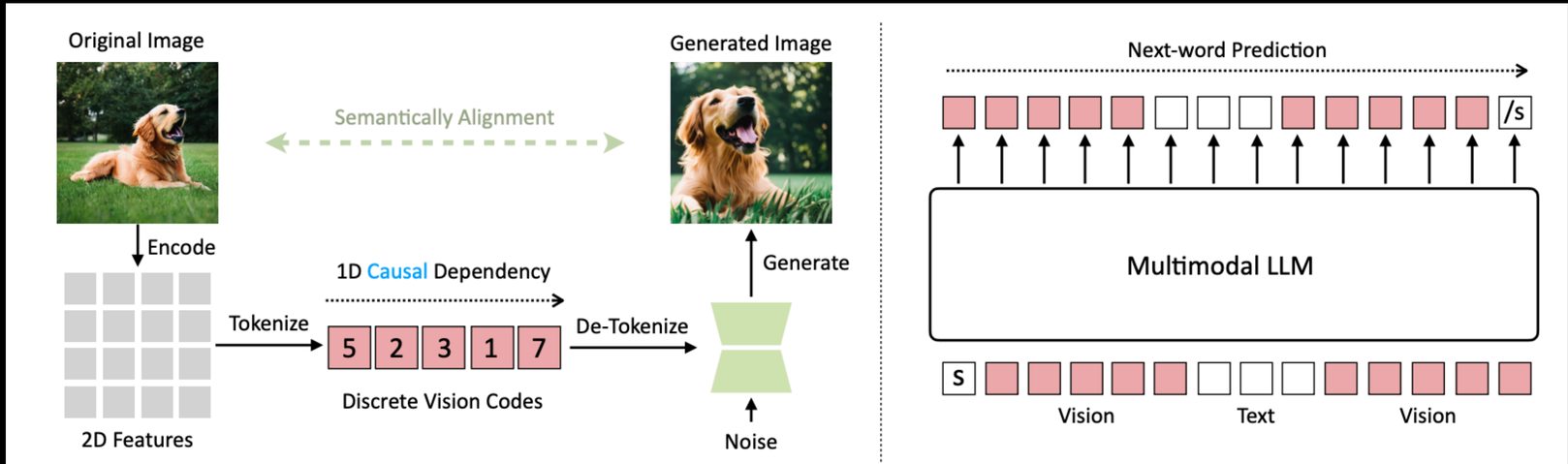


Figure 2: (a) SEED is a discrete image tokenizer, producing quantized visual codes with 1D causal dependency and high-level semantics. (b) With SEED tokenizer, LLM is able to perform scalable multimodal autoregression on interleaved visual and textual data with next-word-prediction objective.

[1] Making LLaMA SEE and Draw with SEED Tokenizer, ICLR 2024

Take-Away Messages

- We use MM1 to show how to pre-train a multimodal LLM
- We also discuss other types of architectures, such as Fuyu, Emu2, and SEED
- Reflections
  - How to obtain higher-quality pre-training data via model-based filtering and from more diverse data sources?
  - How to pre-train the next-gen visual encoder for multimodal LLM?
  - How to design better MoE architectures for multimodal LLM?
  - How to pre-train performant Fuyu and SEED-like models?
  - Can multimodal pre-training enhance LLM performance instead?
  - How to pre-train GPT-4o like models?