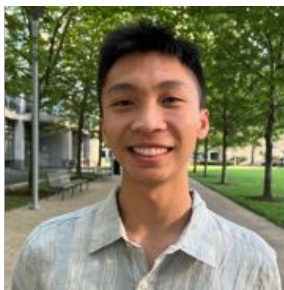


CS294-158 Deep Unsupervised Learning

Lecture 12 Multi-Modal Models



Pieter Abbeel, Wilson Yan, Kevin Frans, Philipp Wu

Multi-Modal Models

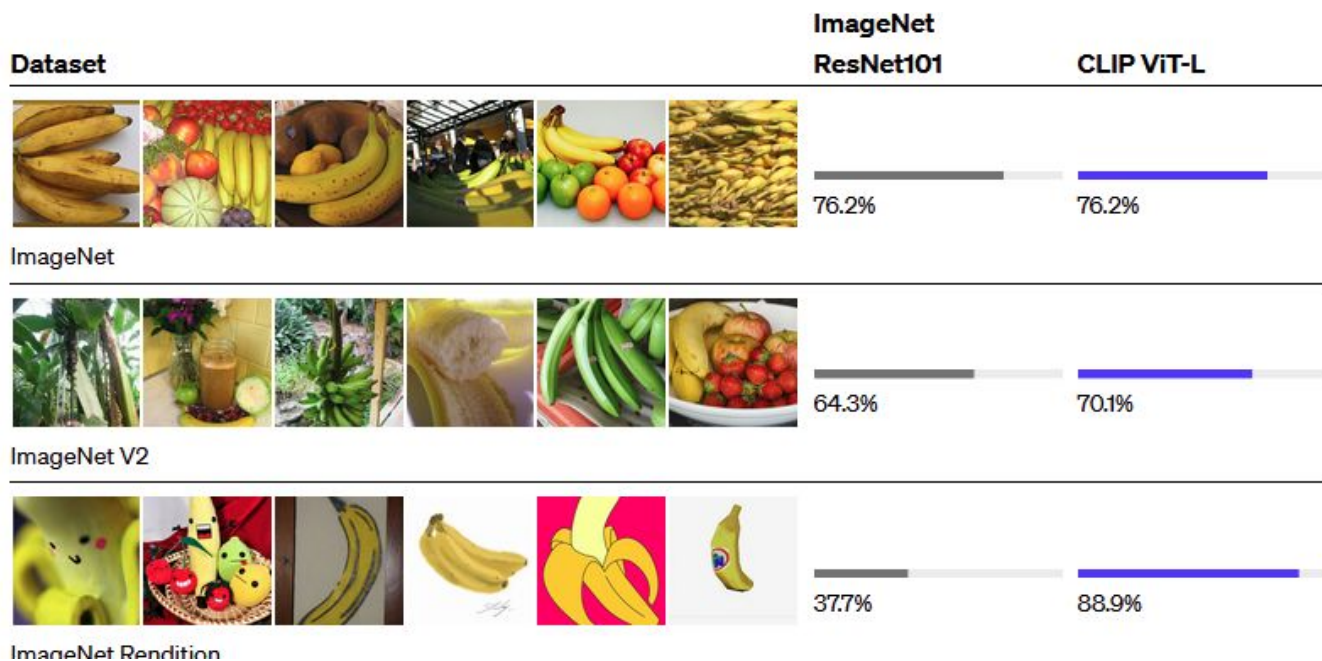
- Models that can incorporate *multiple* modalities
 - Images
 - Video
 - Text
 - Audio
 - ...
- Ideal Outcomes: (1) achieve multimodal understanding, (2) achieve better *unimodal* understanding

Outline

- **Bidirectional Multi-Modal Models**
- Encoder-Decoder Multi-Modal Models
- Decoder-only Multi-Modal Models
- Multi-Modal Generation
- Parallelization in Transformers

CLIP: Capabilities

Image classification



CLIP: Capabilities

CLIP score as a re-ranker

- Generation N images from a text-image model
- Rank their text-image alignment using CLIP
- Select the pair with the highest CLIP score

CLIP: Capabilities

3D generation (Dream Fields)



CLIP: Capabilities

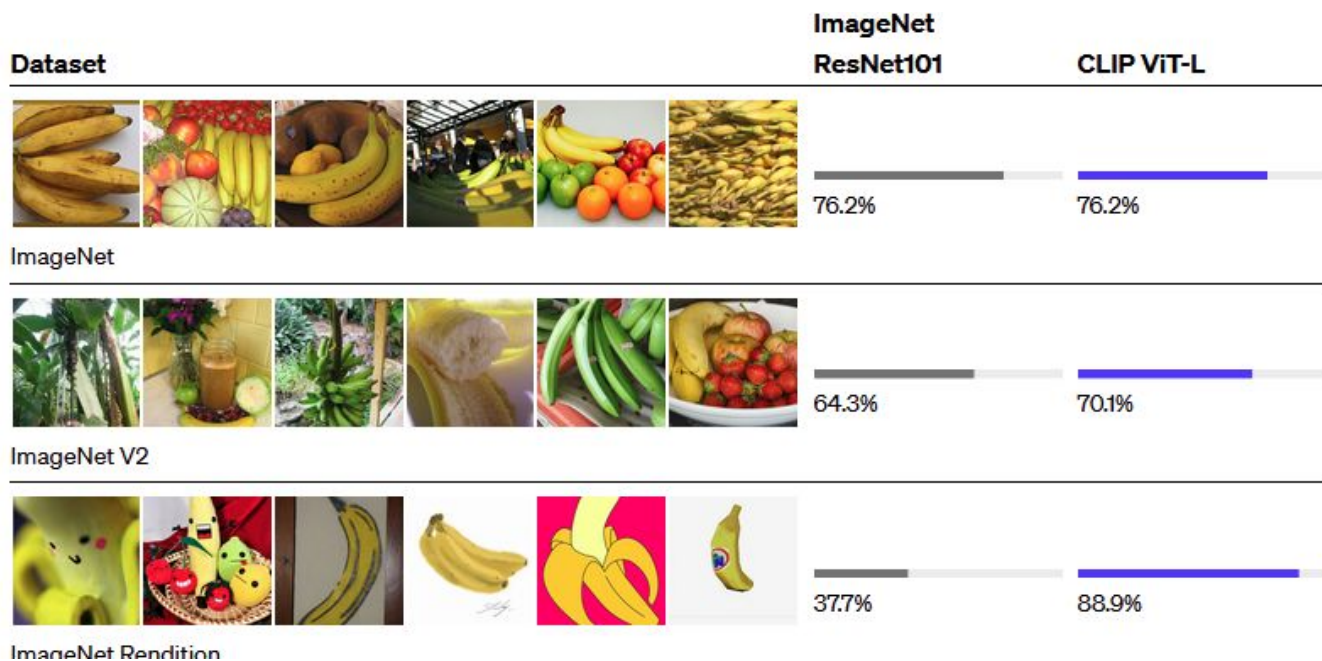
Image generation
(guidance) - VQGAN
CLIP

Works generally for
diffusion models as well
(better)



CLIP: Capabilities

Image classification



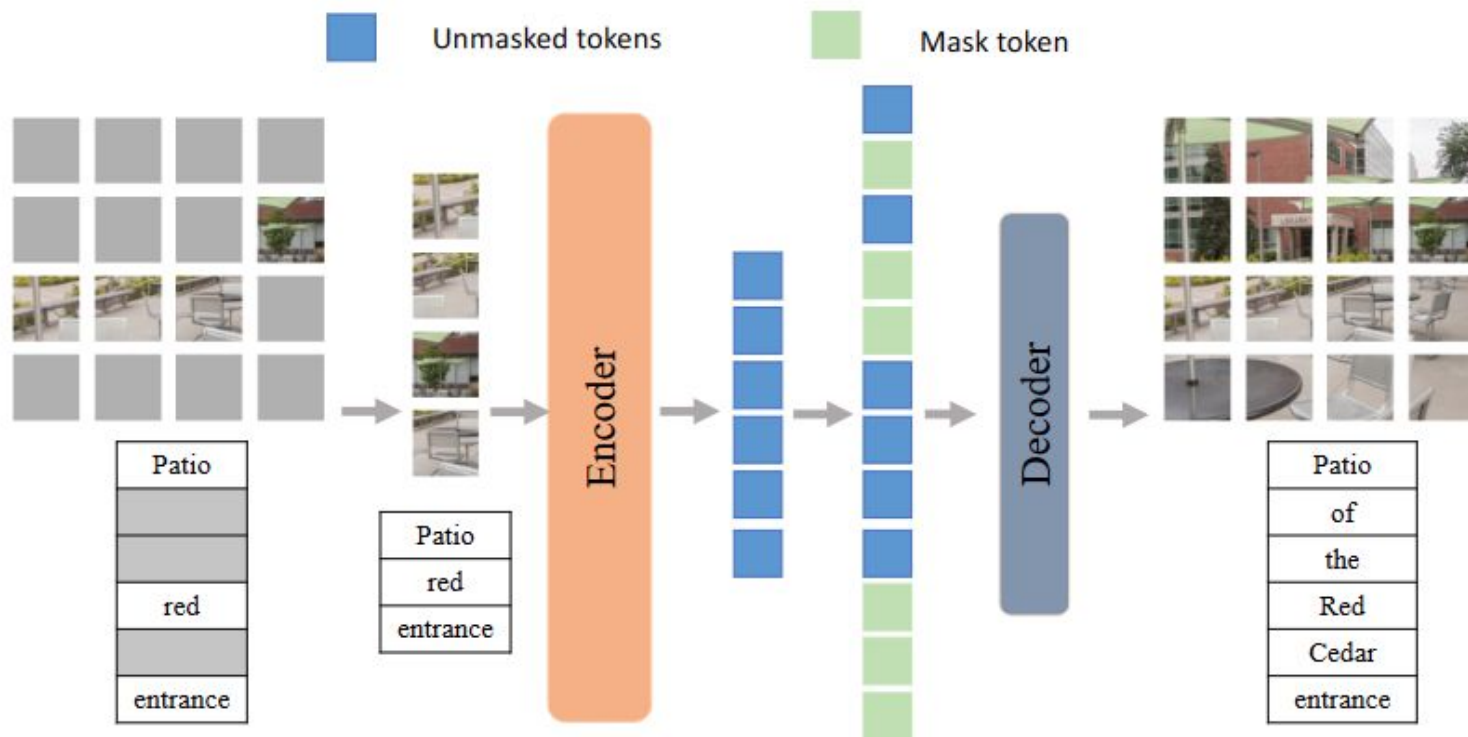
CLIP: Capabilities

Use CLIP text-image encoders as pretrained models

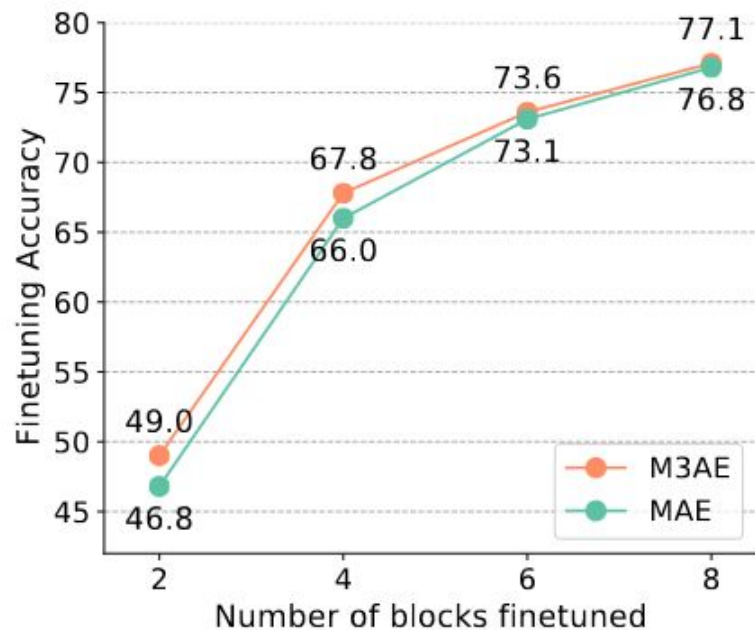
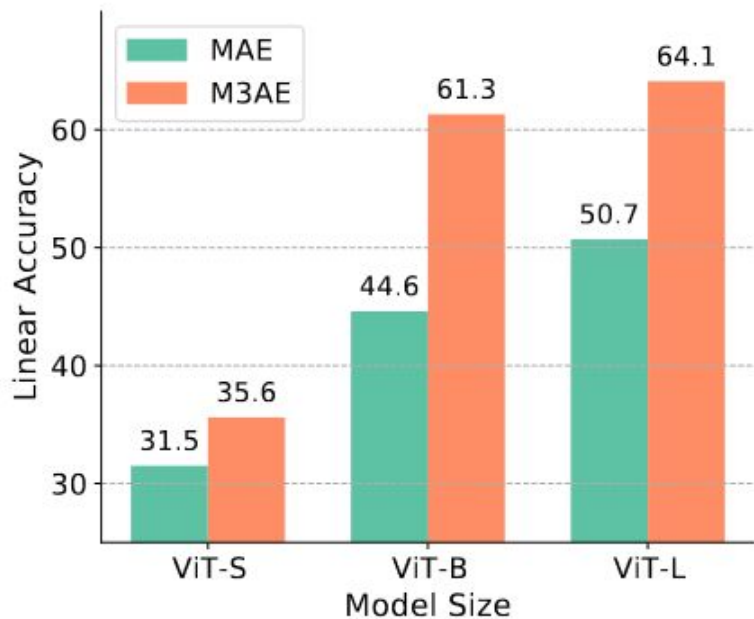
- Many current vision-language models
- For representations (e.g. Stable Diffusion text encoder)
- CLIPort

Usually achieves better generalization, especially to unseen visual concepts during finetuning

M3AE



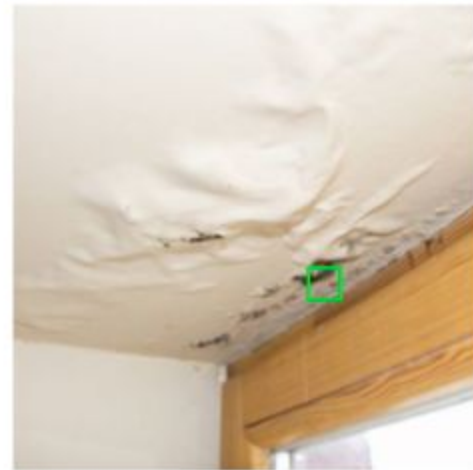
M3AE



M3AE

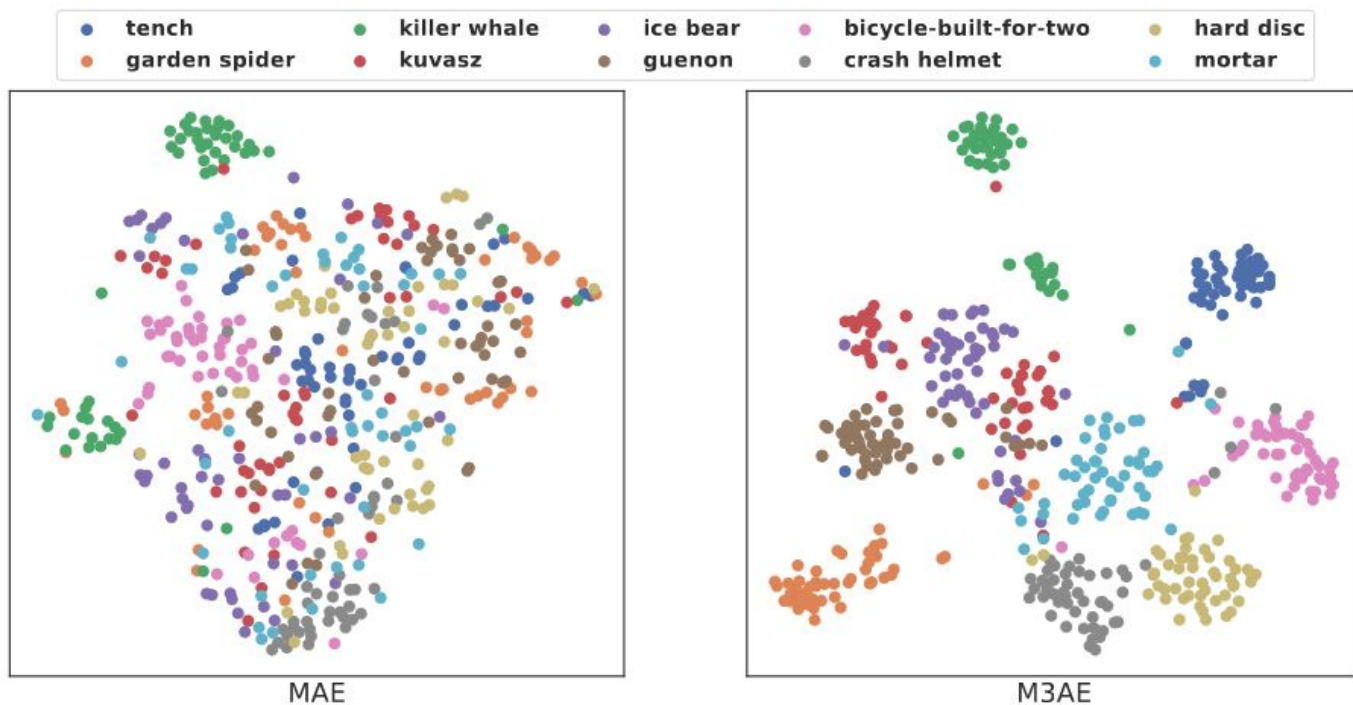


close up view of one end of the tart filled with strawberries and topped with strawberries.



a roof leak forming on a ceiling.

M3AE

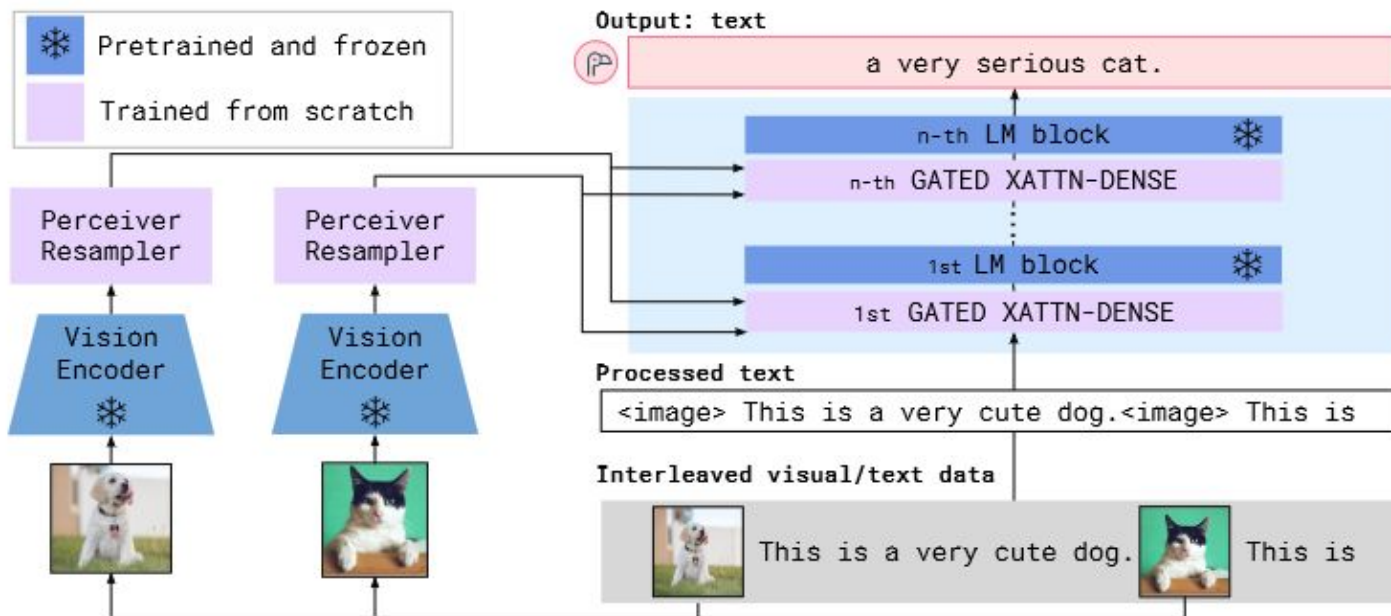


Outline

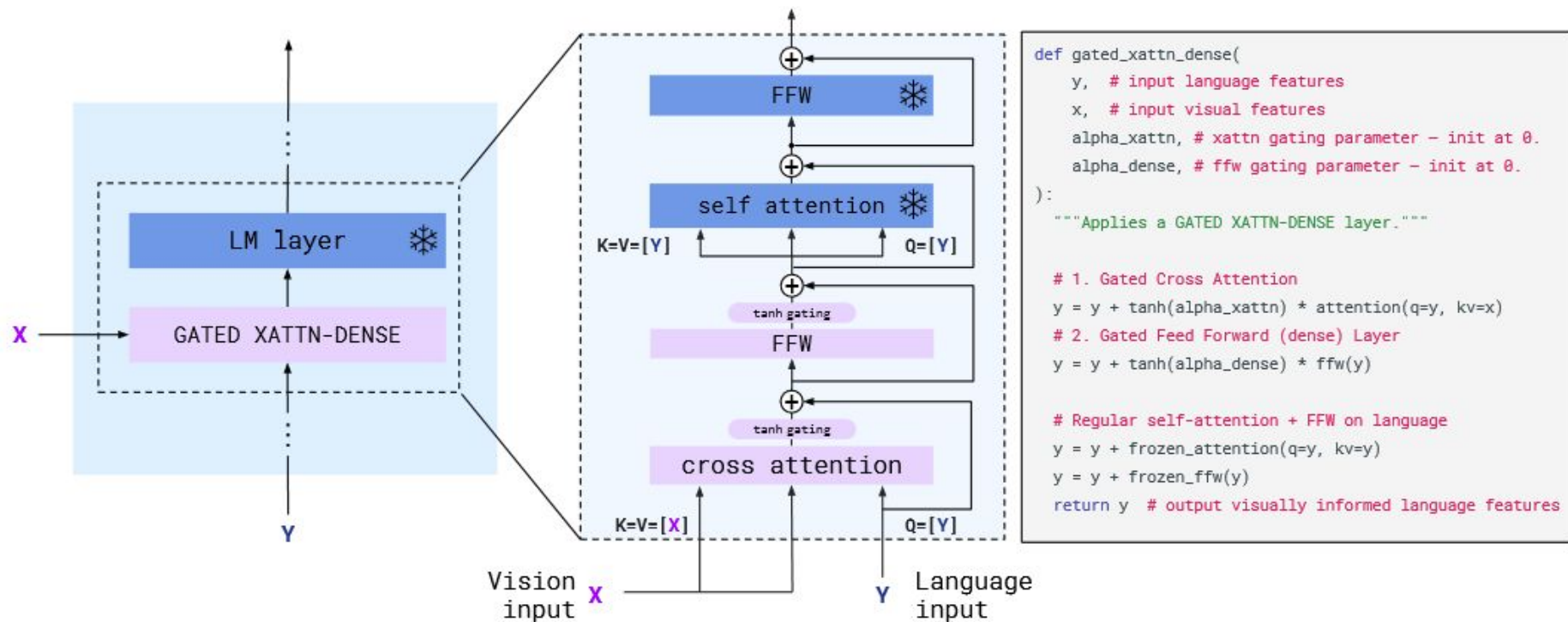
- Bidirectional Multi-Modal Models
- **Encoder-Decoder Multi-Modal Models**
- Decoder-only Multi-Modal Models
- Multi-Modal Generation
- Parallelization in Transformers

Flamingo

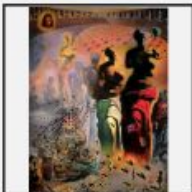
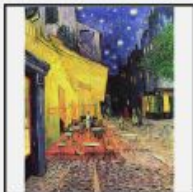
Leveraging pretrained vision-only and language-only models



Flamingo



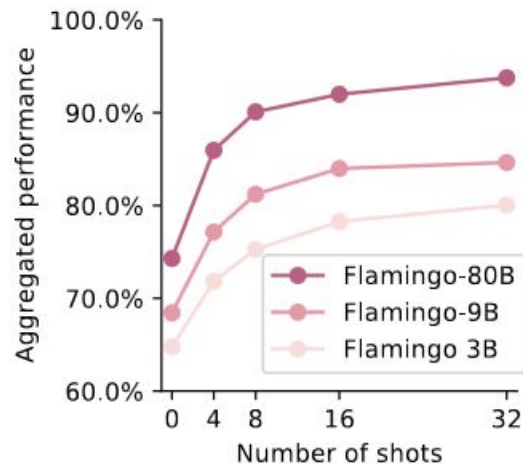
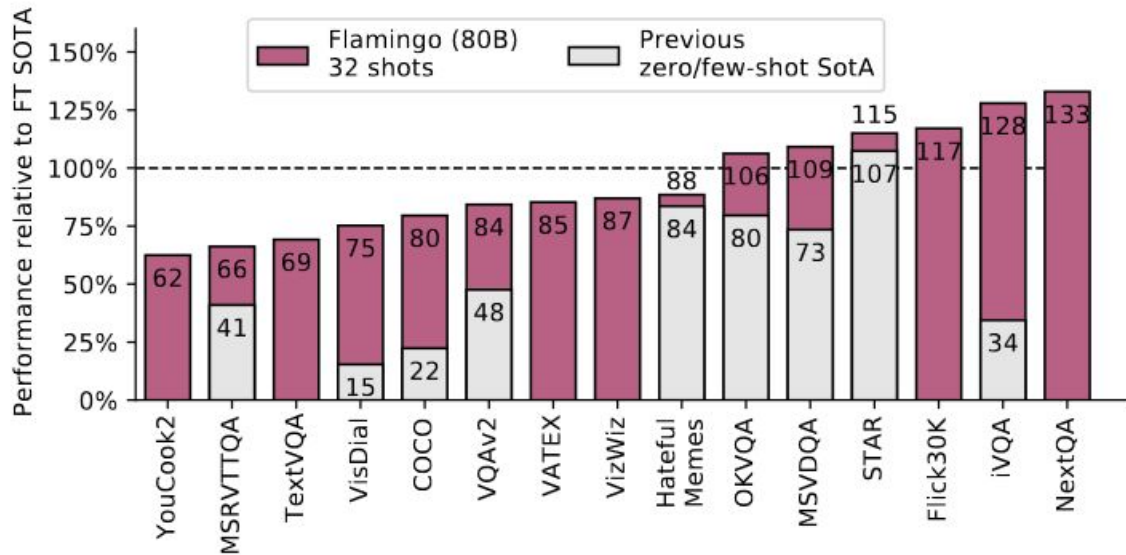
Flamingo

Input Prompt						Completion	
	This is a chinchilla. They are mainly found in Chile.		This is a shiba. They are very popular in Japan.		This is	→	a flamingo. They are found in the Caribbean and South America.
	What is the title of this painting? Answer: The Hallucinogenic Toreador.		Where is this painting displayed? Answer: Louvres Museum, Paris.		What is the name of the city where this was painted? Answer:	→	Arles.
	Output: "Underground"		Output: "Congress"		Output:	→	"Soulomes"

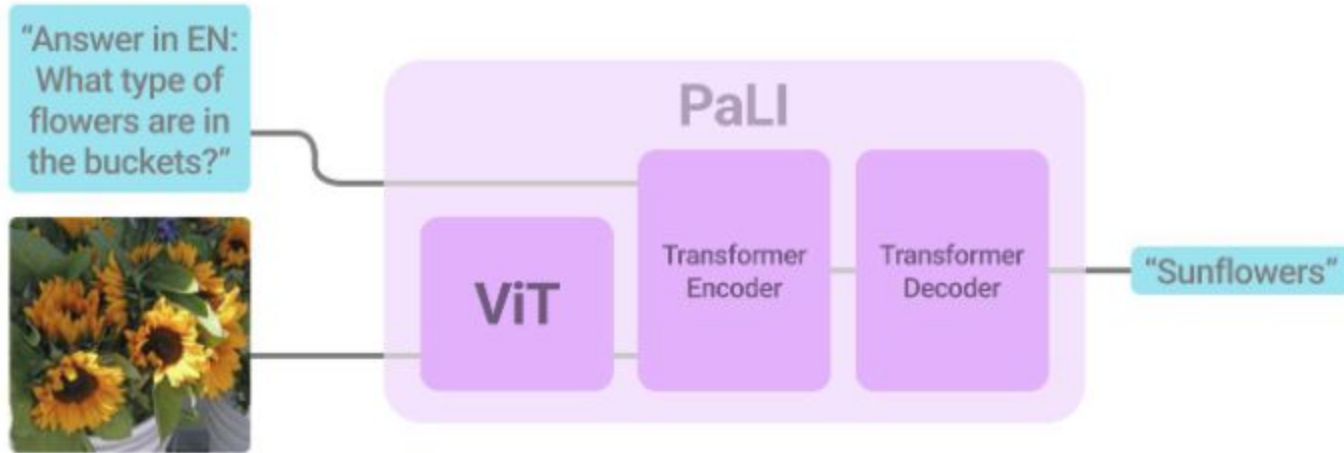
Flamingo

	What happens to the man after hitting the ball? Answer:	→ he falls down.	
 This is a picture of two teddy bears on the moon. What are they doing? They are having a conversation. What object are they using? It looks like a computer. Is this surprising? Yes, it is surprising. Why is this picture surprising to you? I think it is surprising because teddy bears are not usually found on the moon.	 What is the common thing about these three images? They are all flamingos. What is the difference between these three images? The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.	 This is an apple with a sticker on it. What does the sticker say? The sticker says "iPod". Where is the photo taken? It looks like it's taken in a backyard. Do you think it is printed or handwritten? It looks like it's handwritten. What color is the sticker? It's white.	 This is a cityscape. It looks like Chicago. What makes you think this is Chicago? I think it's Chicago because of the Shedd Aquarium in the background.  What about this one? Which city is this and what famous landmark helped you recognise the city? This is Tokyo. I think it's Tokyo because of the Tokyo Tower.

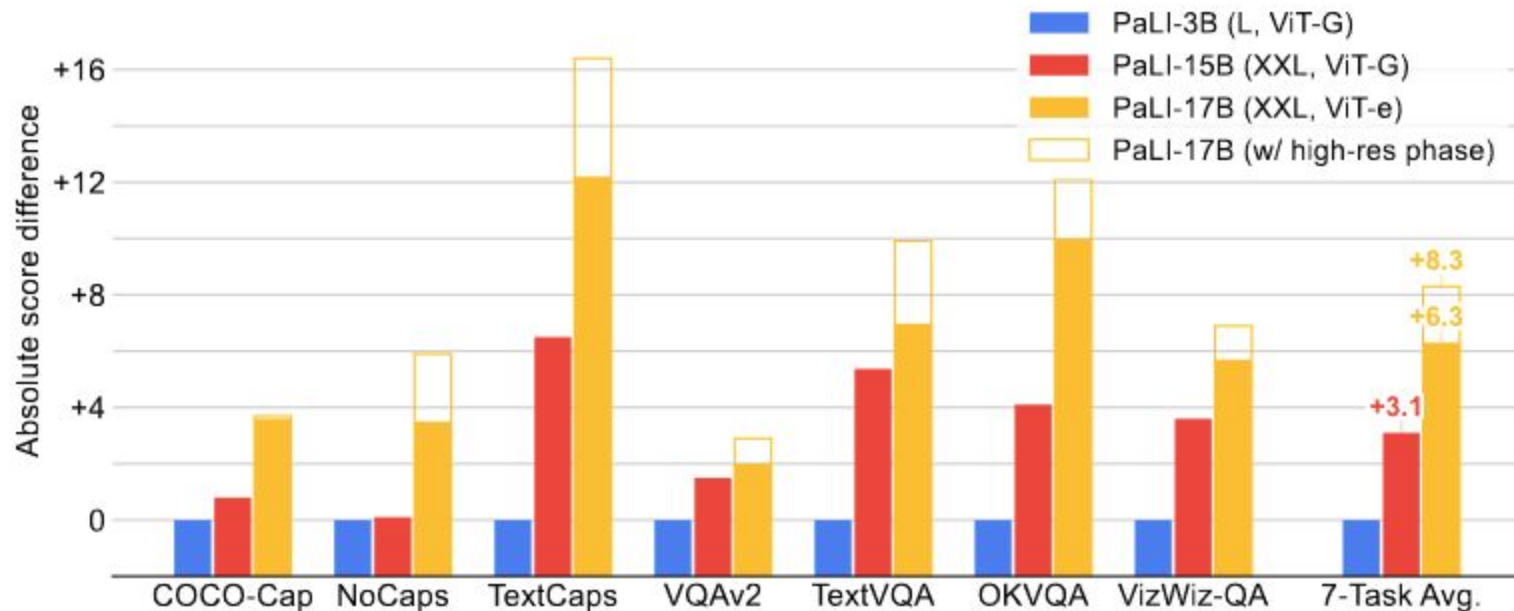
Flamingo



PaLI



PaLI



PaLI Successors

PaLI-X

- Further scaling
- Better training schemes (mixture of objectives, further training with OCR objectives on vision encoder for better text understanding)

PaLI-3

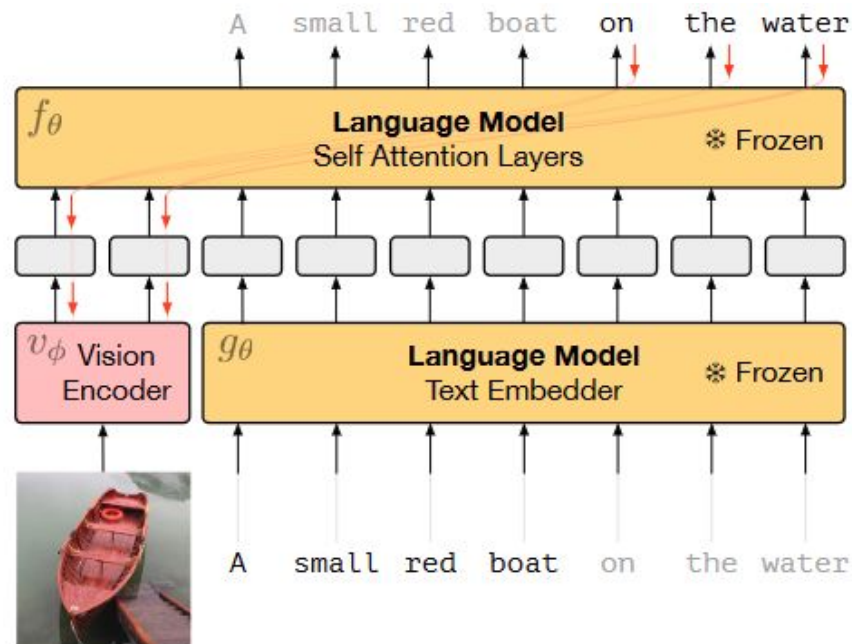
- Switch to SigLIP encoder
- Close to PaLI-X while being 10x smaller (5B vs 55B)

Outline

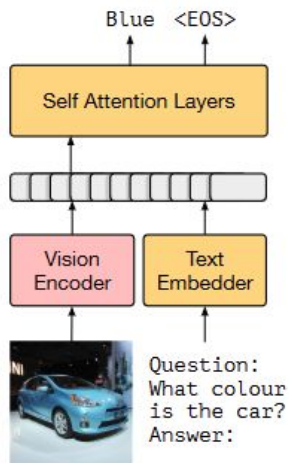
- Bidirectional Multi-Modal Models
- Encoder-Decoder Multi-Modal Models
- **Decoder-only Multi-Modal Models**
- Multi-Modal Generation
- Parallelization in Transformers

Frozen

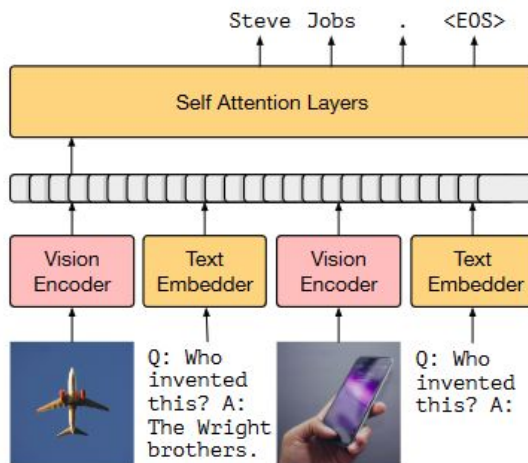
Use a pretrained LLM



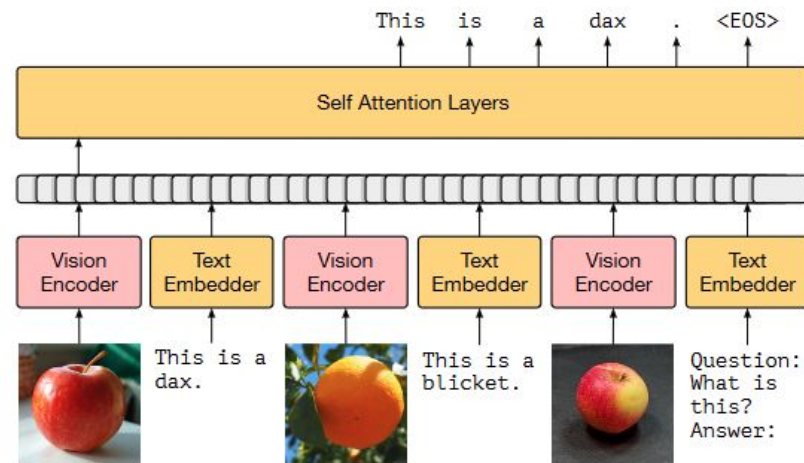
Frozen



(a) 0-shot VQA

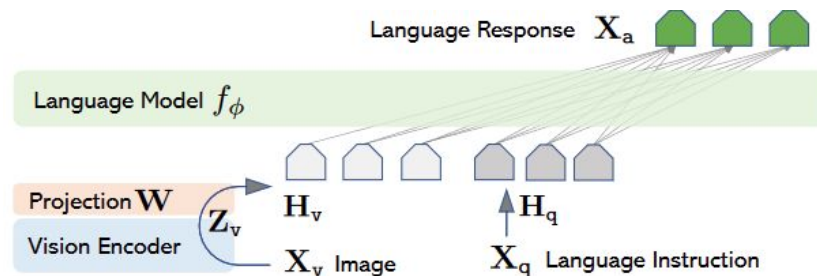
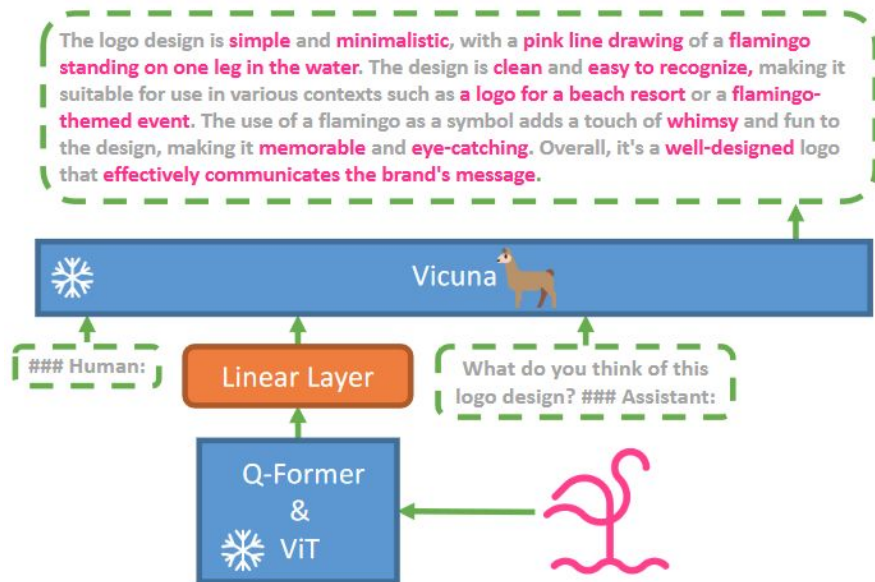


(b) 1-shot outside-knowledge VQA



(c) Few-shot image classification

MiniGPT-4 / LLaVA



LLaVA

Stage 1:

- Keep vision + language backbones frozen, train projection layer on ~600K text-image pairs

Stage 2:

- Keep vision backbone frozen, finetune language + projection on high quality visual instruct dataset

LLaVA

Challenging examples from LLaVA-Bench (In-the-Wild):



ICHIRAN Ramen [source]



Filled fridge [source]

Annotation	A close-up photo of a meal at ICHIRAN. The chashu ramen bowl with a spoon is placed in the center. The ramen is seasoned with chili sauce, chopped scallions, and served with two pieces of chashu. Chopsticks are placed to the right of the bowl, still in their paper wrap, not yet opened. The ramen is also served with nori on the left. On top, from left to right, the following sides are served: a bowl of orange spice (possibly garlic sauce), a plate of smoke-flavored stewed pork with chopped scallions, and a cup of matcha green tea.	An open refrigerator filled with a variety of food items. In the left part of the compartment, towards the front, there is a plastic box of strawberries with a small bag of baby carrots on top. Towards the back, there is a stack of sauce containers. In the middle part of the compartment, towards the front, there is a green plastic box, and there is an unidentified plastic bag placed on it. Towards the back, there is a carton of milk. In the right part of the compartment, towards the front, there is a box of blueberries with three yogurts stacked on top. The large bottle of yogurt is Fage non-fat yogurt, and one of the smaller cups is Fage blueberry yogurt. The brand and flavor of the other smaller cup are unknown. Towards the back, there is a container with an unknown content.
Question 1	What's the name of the restaurant?	What is the brand of the blueberry-flavored yogurt?
Question 2	Describe this photo in detail.	Is there strawberry-flavored yogurt in the fridge?

LLaVA 1.5

- Larger resolution (224 -> 336)
- Linear -> MLP projection
- Better instruct data mix with better VQA data

Method	LLM	Res.	PT	IT	VQA ^{v2}	GQA	VisWiz	SQA ^I	VQA ^T	POPE	MME	MMB	MMB ^{CN}	SEED	LLaVA ^W	MM-Vet
BLIP-2	Vicuna-13B	224	129M	-	41.0	41	19.6	61	42.5	<u>85.3</u>	1293.8	-	-	46.4	38.1	22.4
InstructBLIP	Vicuna-7B	224	129M	1.2M	-	49.2	34.5	60.5	50.1	-	-	36	23.7	53.4	60.9	26.2
InstructBLIP	Vicuna-13B	224	129M	1.2M	-	49.5	33.4	63.1	50.7	78.9	1212.8	-	-	-	58.2	25.6
Shikra	Vicuna-13B	224	600K	5.5M	77.4*	-	-	-	-	-	-	58.8	-	-	-	-
IDEFICS-9B	LLaMA-7B	224	353M	1M	50.9	38.4	35.5	-	25.9	-	-	48.2	25.2	-	-	-
IDEFICS-80B	LLaMA-65B	224	353M	1M	60.0	45.2	36.0	-	30.9	-	-	54.5	38.1	-	-	-
Qwen-VL	Qwen-7B	448	1.4B [†]	50M [†]	<u>78.8*</u>	59.3*	35.2	67.1	63.8	-	-	38.2	7.4	56.3	-	-
Qwen-VL-Chat	Qwen-7B	448	1.4B [†]	50M [†]	78.2*	57.5*	38.9	<u>68.2</u>	<u>61.5</u>	-	1487.5	60.6	56.7	<u>58.2</u>	-	-
LLaVA-1.5	Vicuna-7B	336	558K	665K	<u>78.5*</u>	<u>62.0*</u>	<u>50.0</u>	66.8	58.2	85.9	<u>1510.7</u>	<u>64.3</u>	<u>58.3</u>	<u>58.6</u>	<u>63.4</u>	<u>30.5</u>
LLaVA-1.5	Vicuna-13B	336	558K	665K	80.0*	63.3*	53.6	71.6	<u>61.3</u>	85.9	1531.3	67.7	63.6	61.6	70.7	35.4

LLaVA 1.6

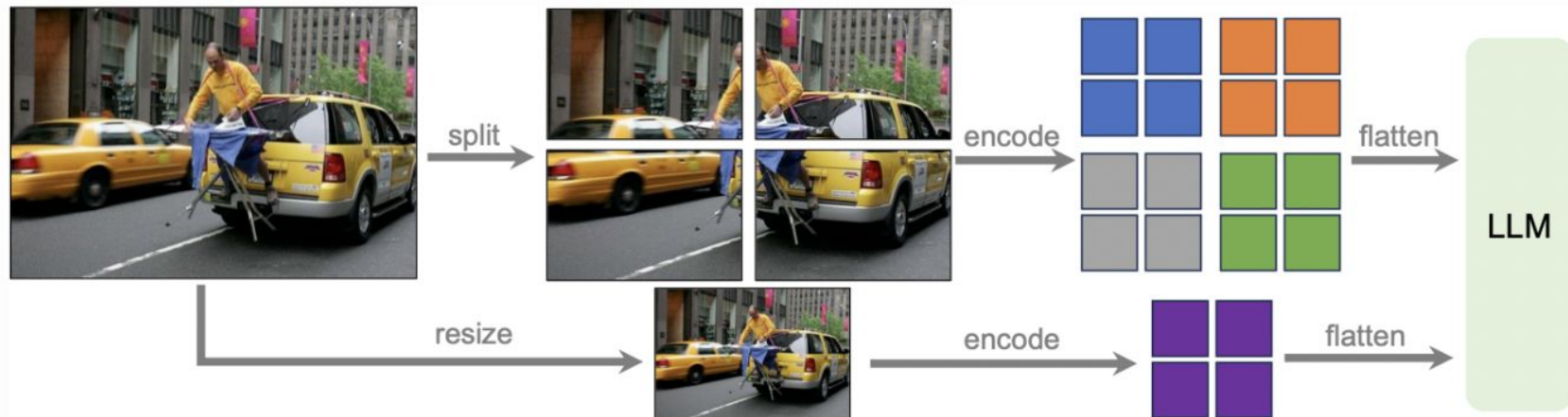


Illustration of dynamic high resolution scheme: a grid configuration of 2×2

LLaVA 1.6

Open-Source

Proprietary

Data (PT)	Data (IT)	Model	MMMU (val)	Math-Vista	MMB-ENG	MMB-CN	MM-Vet	LLaVA-Wild	SEED-IMG
N/A	N/A	GPT-4V	56.8	49.9	75.8	73.9	67.6	-	71.6
N/A	N/A	Gemini Ultra	59.4	53	-	-	-	-	-
N/A	N/A	Gemini Pro	47.9	45.2	73.6	74.3	64.3	-	70.7
1.4B	50M	Qwen-VL-Plus	45.2	43.3	-	-	55.7	-	65.7
1.5B	5.12M	CogVLM-30B	32.1	-	-	-	56.8	-	-
125M	~1M	Yi-VL-34B	45.9	-	-	-	-	-	-
558K	665K	LLaVA-1.5-13B	36.4	27.6	67.8	63.3	36.3	72.5	68.2
558K	760K	LLaVA-NeXT-34B	51.1	46.5	79.3	79	57.4	89.6	75.9

GPT-4V

Unclear as to what exactly it is. From the pricing - can infer:

- Low resolution encoding - 85 tokens
- Full resolution (512 x 512) - 255 tokens (85 + 170)
- Any larger - tiled by 512 x 512 tiles

Model	Input	Output
gpt-4-turbo-2024-04-09	\$10.00 / 1M tokens	\$30.00 / 1M tokens

Vision pricing calculator

512 px by 512 px = \$0.00255 ⓘ

☐ Low resolution

GPT-4V: Capabilities

As a visual knowledge bank



What coin is that?



CHATGPT

That is a United States penny, featuring a portrait of President Abraham Lincoln.

A photograph of a peace lily plant with white flowers and green leaves, sitting on a wooden surface.

SALO

What is that plant and how should I care about it?

CHATGPT

The plant in the image is a *Spathiphyllum*, commonly known as a Peace Lily. Here are some care tips for your Peace Lily:

1. **Light:** Peace Lilies prefer bright, indirect light but can tolerate low light conditions. Avoid direct sunlight, which can scorch the leaves.
2. **Watering:** Water the plant when the top inch of soil is dry. It's better to under-water than over-water. If the leaves begin to droop, it's a sign the plant needs water.
3. **Temperature:** They prefer temperatures between 65°F and 80°F (18°C to 27°C). Avoid exposing the

A wide-angle photograph of the San Francisco skyline, showing numerous skyscrapers and buildings, with the Transamerica Pyramid being a prominent feature.

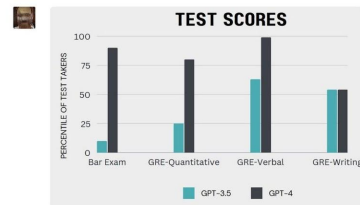
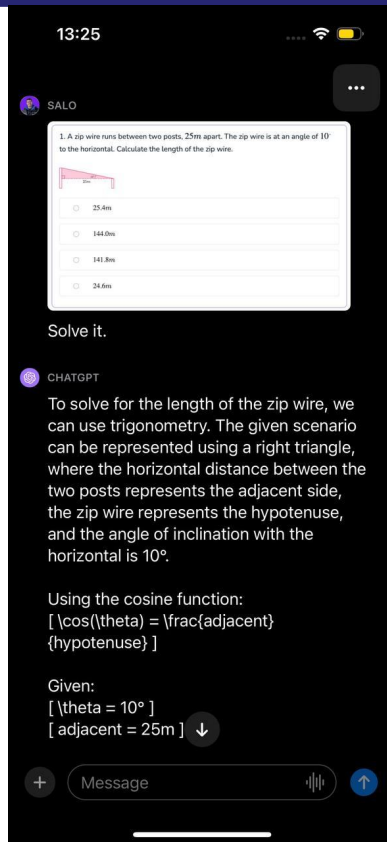
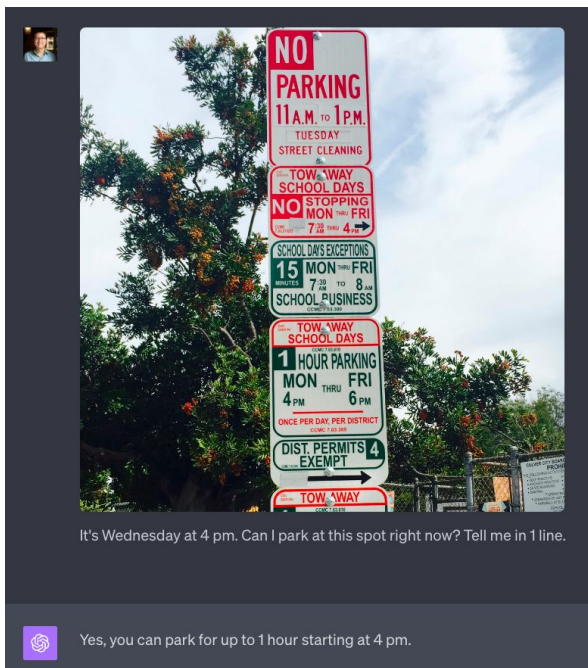
Where is this?

CHATGPT

This is a view of San Francisco, California. The prominent pyramid-shaped building is the Transamerica Pyramid, one of the city's most iconic landmarks.

GPT-4V: Capabilities

OCR / Text understanding in images



what does this tell you? be very very exact

This bar chart illustrates the test scores of two different models, GPT-3.5 and GPT-4, across four different tests: Bar Exam, GRE-Quantitative, GRE-Verbal, and GRE-Writing. The scores are presented in terms of percentiles of test takers.

Here are the exact observations based on the provided chart:

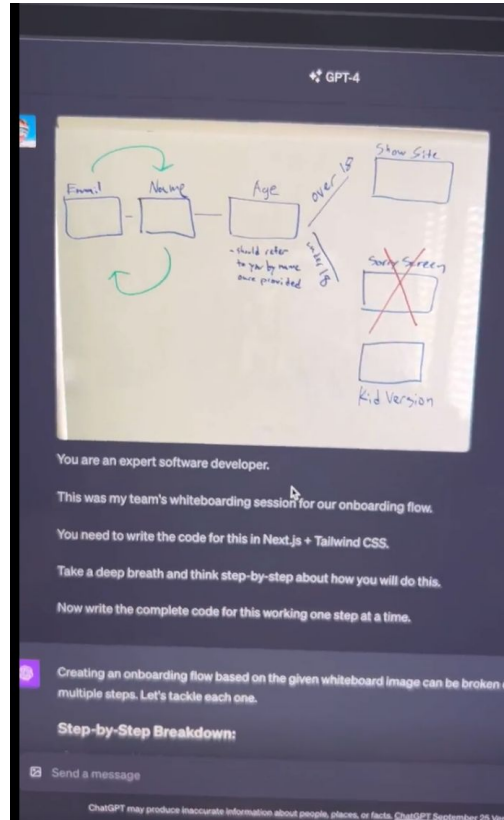
- Bar Exam:**
 - GPT-3.5 scored very low, close to the 0th percentile.
 - GPT-4 scored very high, close to the 100th percentile.
- GRE-Quantitative:**
 - GPT-3.5 scored around the 25th percentile.
 - GPT-4 scored slightly above the 75th percentile.
- GRE-Verbal:**
 - GPT-3.5 scored close to the 75th percentile.
 - GPT-4 scored very close to the 100th percentile.
- GRE-Writing:**
 - GPT-3.5 scored slightly below the 50th percentile.
 - GPT-4 scored just above the 50th percentile.

Overall Insights:

- GPT-4 generally outperforms GPT-3.5 across all the tests shown.
- The most significant performance gap between GPT-3.5 and GPT-4 is observed in the Bar Exam, where GPT-4 seems to excel while GPT-3.5 struggles considerably.
- In the GRE-Verbal section, both models perform relatively well, but GPT-4 approaches the top percentile.
- For GRE-Writing, the difference between the two models is minimal, with both scoring around the median.

GPT-4V: Capabilities

Code from sketches

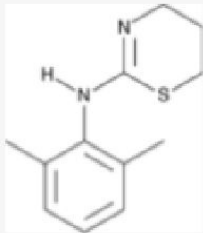


GPT-4V: Capabilities

Not always correct

Prompt

What is it?



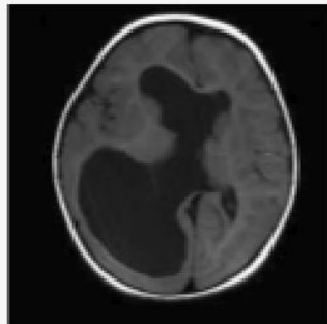
GPT-4V

The chemical structure you've provided is that of a compound called Thiamine, also known as Vitamin B1. Thiamine is an essential vitamin for humans, which means it must be obtained through the diet. It plays a crucial role in energy metabolism by acting as a coenzyme for enzymes involved in carbohydrate metabolism.

Analysis: Wrong Answer. The correct answer is xylazine

Prompt

Which side of the brain is the hydrocephalus located on in this image?



GPT-4V Response 1

The hydrocephalus in this image appears to be located on the left side of the brain.

Wrong Answer

GPT-4V Response 2

The hydrocephalus in the image appears to be located on the right side of the brain.

Correct Answer

GPT-4V Response 3

The hydrocephalus in this image appears to be located on the right side of the brain (from the viewer's perspective). However, it's important to note that medical images are typically viewed in a "radiological convention," where the right side of the image corresponds to the patient's left side...

Correct Answer

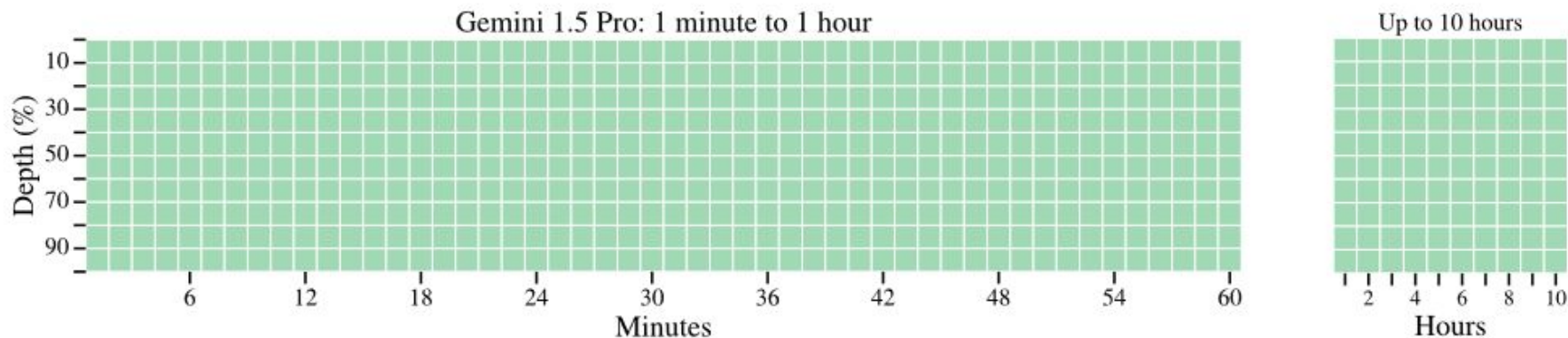
Gemini 1.5

Unclear exactly to what it is using, but something potentially like a VQ + CoCa Model

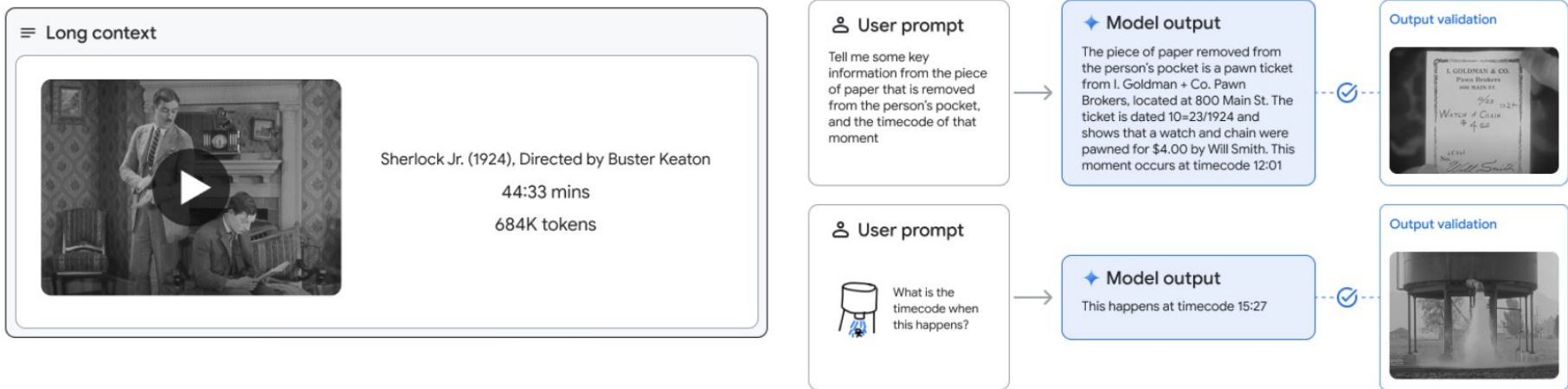
- Discrete (VQ?) encodings
- CoCa - image captioning + text-image contrastive objective

Gemini 1.5

Long-horizon needle retrieval for video

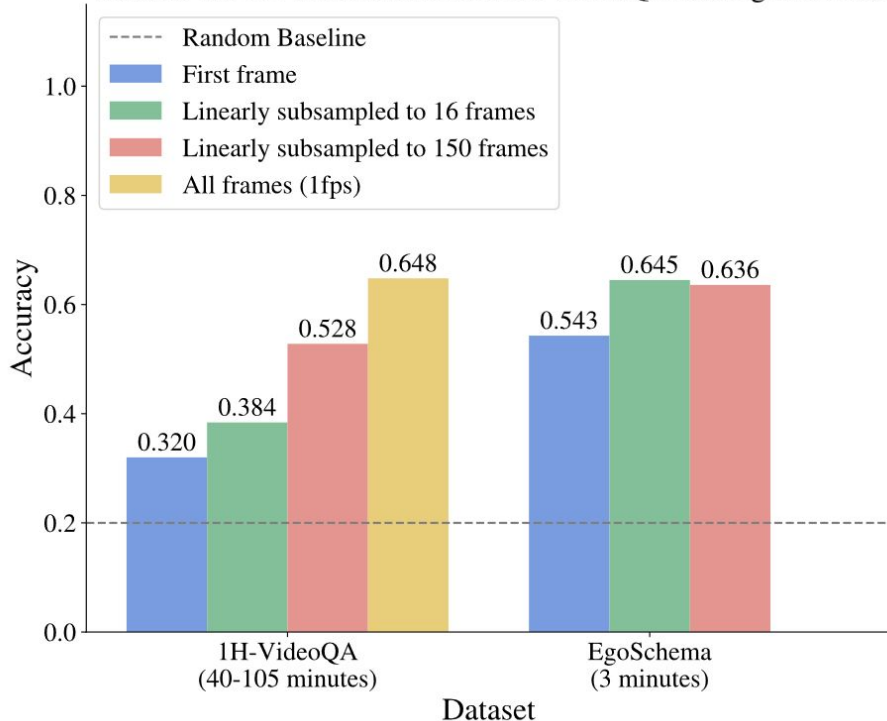


Gemini 1.5



Gemini 1.5

Gemini 1.5 Pro Performance on 1H-VideoQA and EgoSchema



Model	Frames		
	16	150	full video (1 fps)
GPT-4V (0-shot)	29.6%	49.6%	Not supported
Gemini 1.5 Pro (0-shot)	38.4%	52.8%	64.8%

Outline

- Bidirectional Multi-Modal Models
- Encoder-Decoder Multi-Modal Models
- Decoder-only Multi-Modal Models
- **Multi-Modal Generation**
- Parallelization in Transformers

Multi-modal Generation

Text->Image

- Stable Diffusion, Parti, Imagen

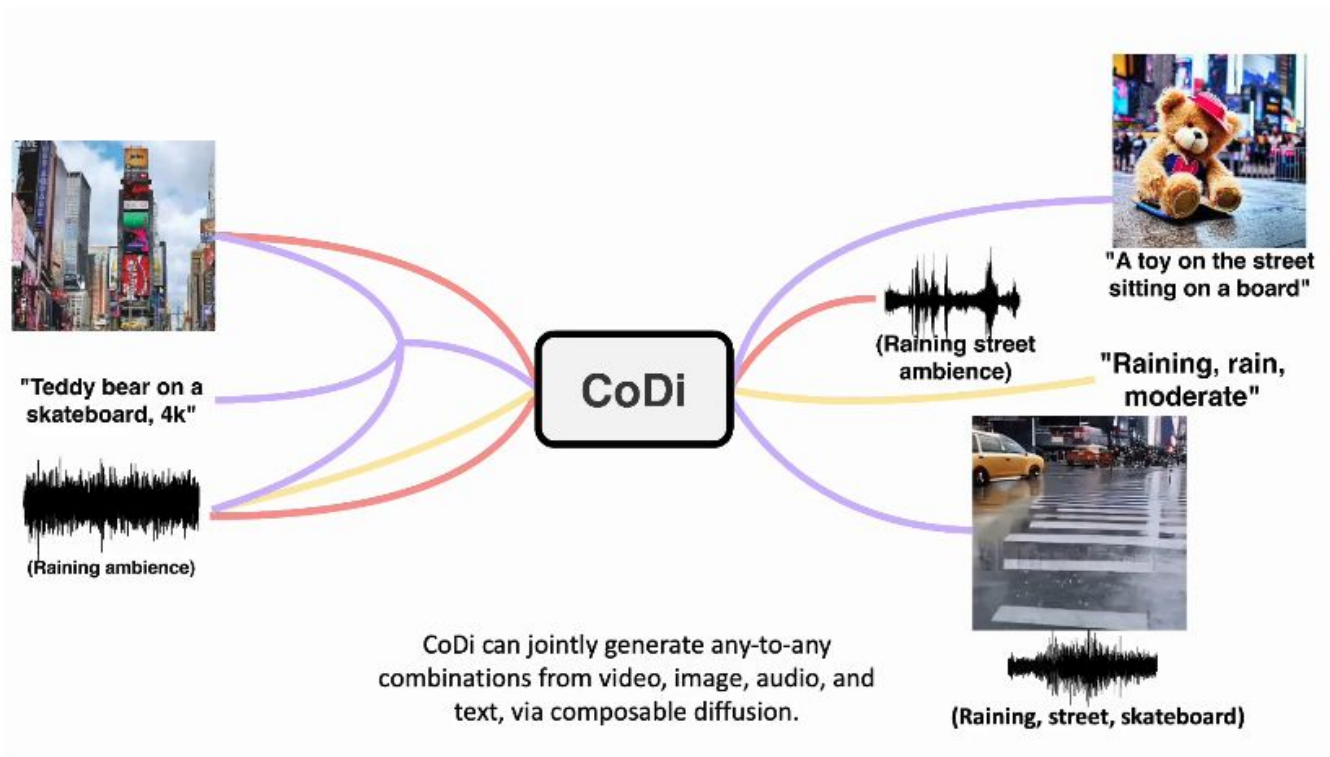
Text->Video

- ImagenVideo, Make-a-Video, Phenaki, Stable Video Diffusion

Text -> Video + Audio

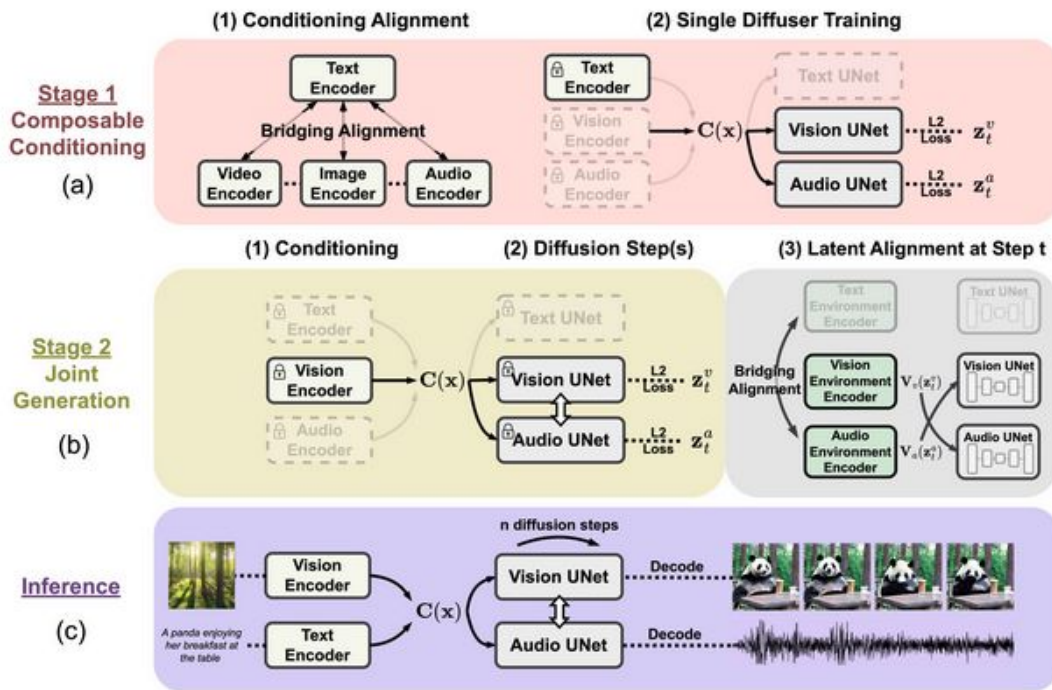
- VideoPoet

CoDi: Any-to-Any Generation



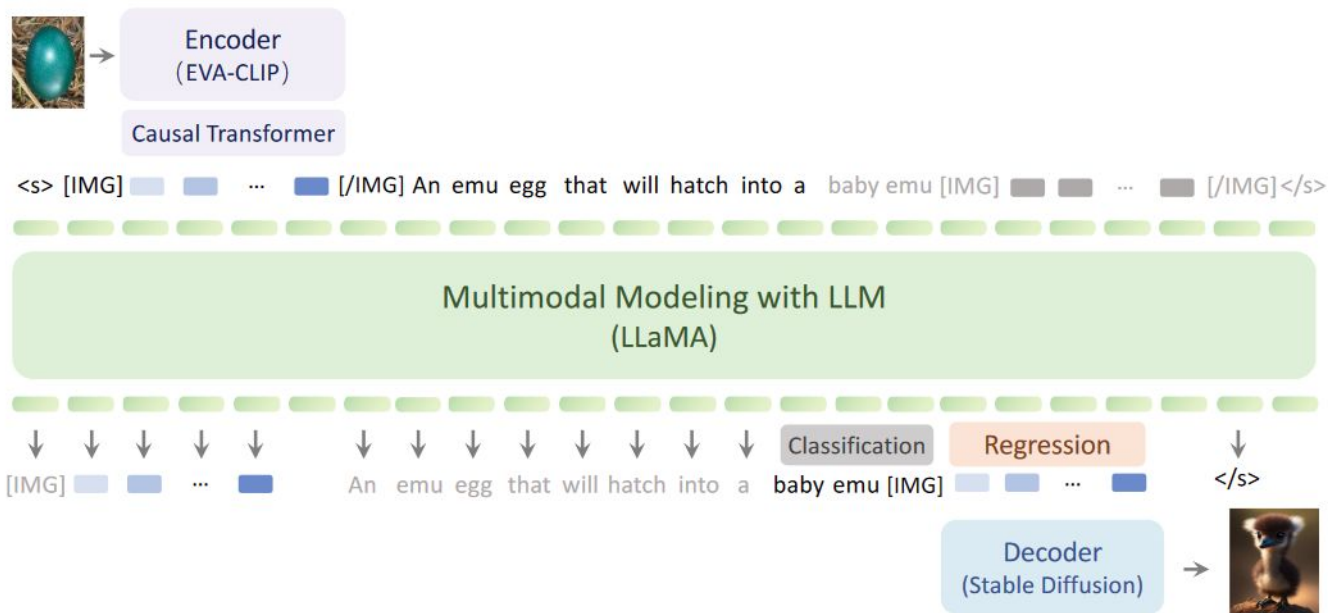
CoDi: Any-to-Any Generation

Model Architecture

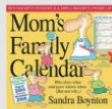


Emu / LaVIT - Style

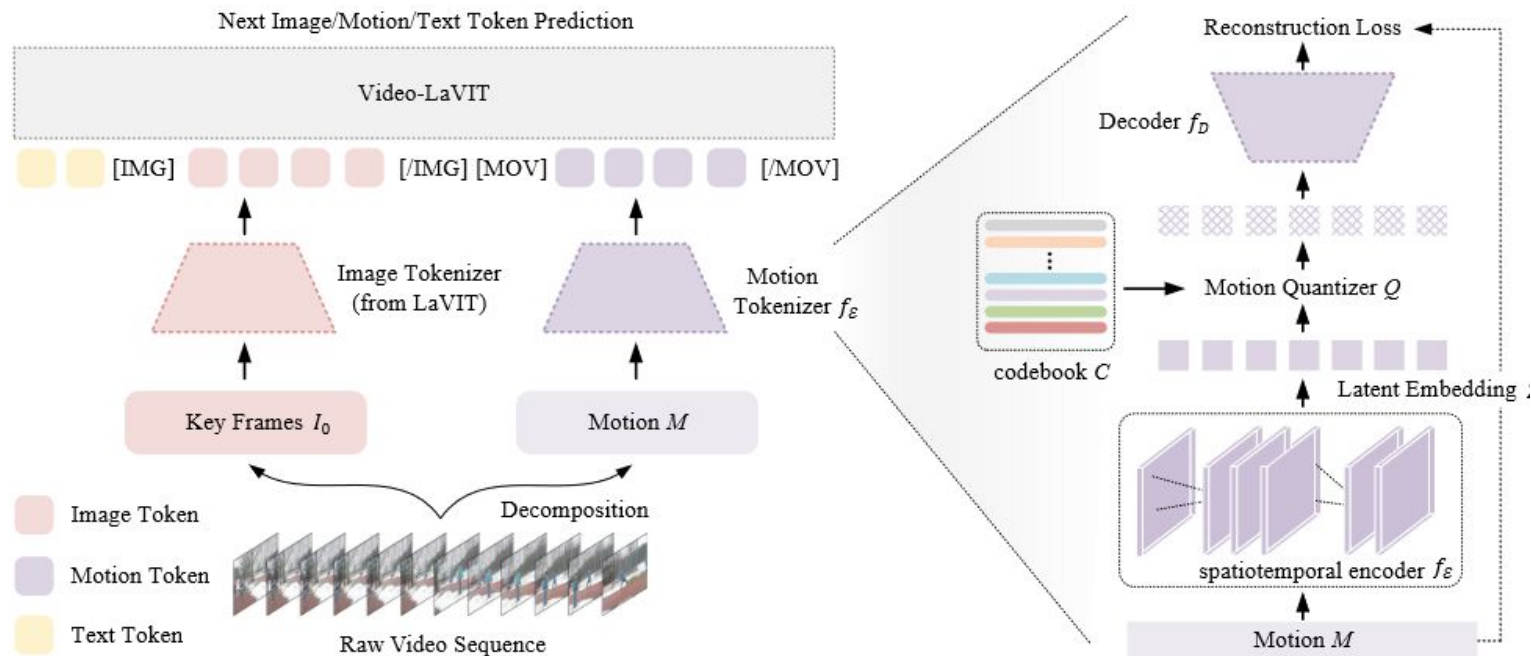
Emu uses regression, LaVIT predicts a quantized representation



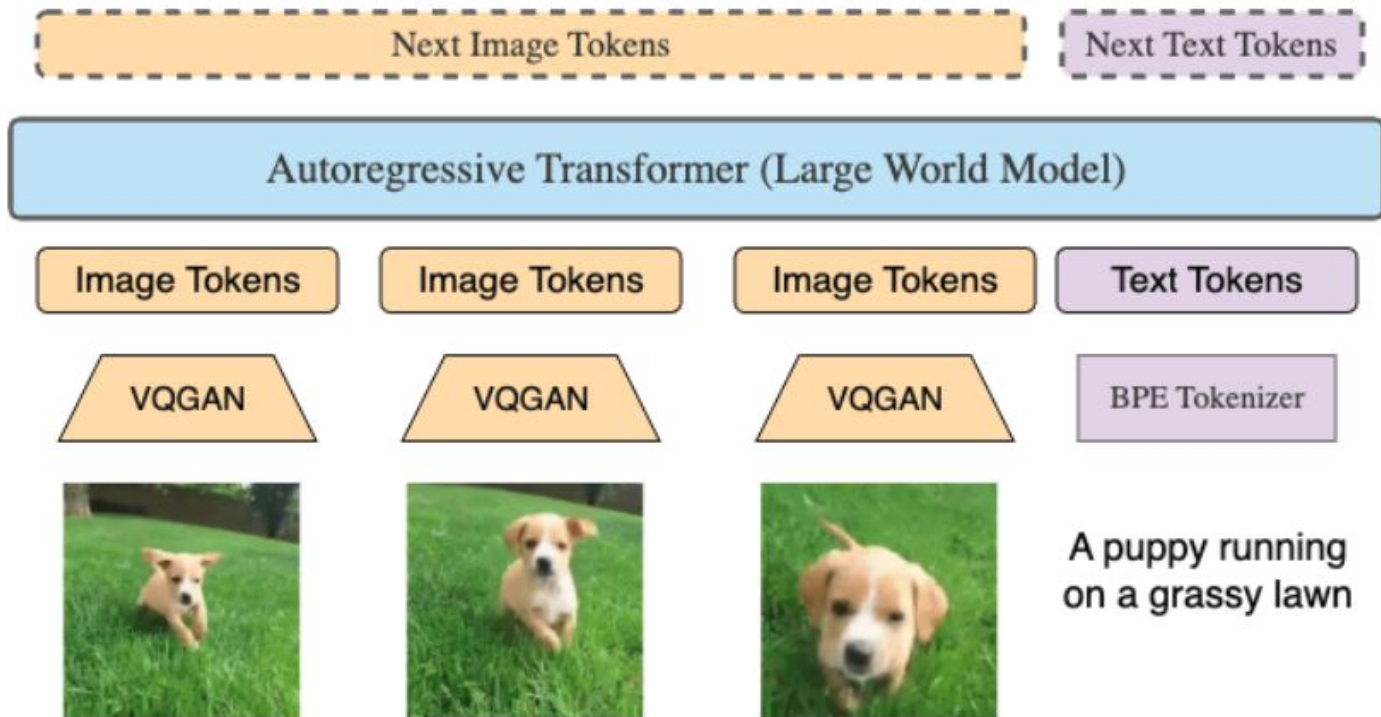
Emu / LaVIT - Style

Image Captioning		Image Question Answering		
Input Prompt			Question: Can you tell me about this image? Answer:	
	The image shows a photo of			
Completion	the movie poster for "Fight Club." The poster features two men standing next to each other, with the words "Fight Club" written in bold letters above them. The background of the poster is dark and ominous, giving the impression of a gritty and intense movie.		this image is the cover of a book called "Mom's Family Calendar" by Sandra Boynton. The book is designed to help moms keep track of their family's schedule and activities.	
Input Prompt			Completion	
In-context Completion			There is a godzilla.	
	There is a chinchilla.	There are three pandas.		
	This is the first image: 	This is the second image: 	The animal in the first image surrounded with the plant in the second image: 	

Video LaVIT



LWM



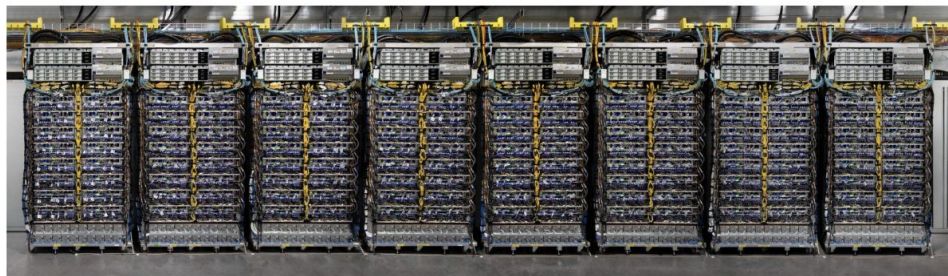
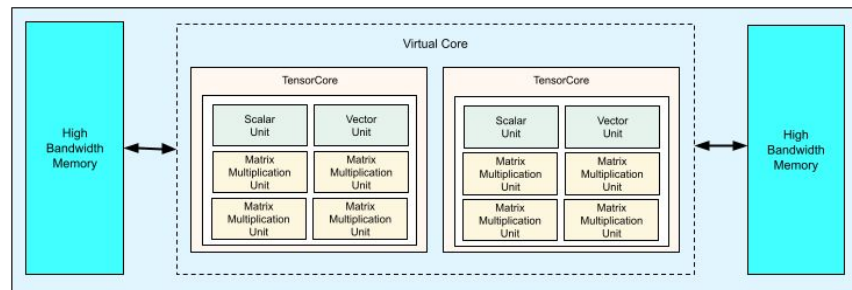
Outline

- Bidirectional Multi-Modal Models
- Encoder-Decoder Multi-Modal Models
- Decoder-only Multi-Modal Models
- Multi-Modal Generation
- **Parallelization in Transformers**

TPU / GPU

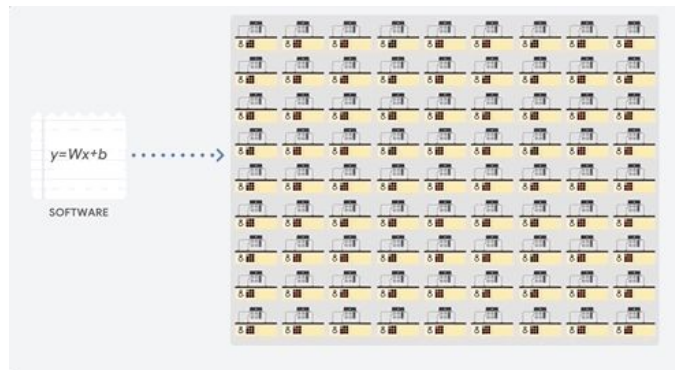
■ Flops, HBM size and bandwidth, Interconnect

Key specifications	v4 Pod values
Peak compute per chip	275 teraflops (bf16 or int8)
HBM2 capacity and bandwidth	32 GiB, 1200 GBps
Measured min/mean/max power	90/170/192 W
TPU Pod size	4096 chips
Interconnect topology	3D mesh
Peak compute per Pod	1.1 exaflops (bf16 or int8)
All-reduce bandwidth per Pod	1.1 PB/s
Bisection bandwidth per Pod	24 TB/s

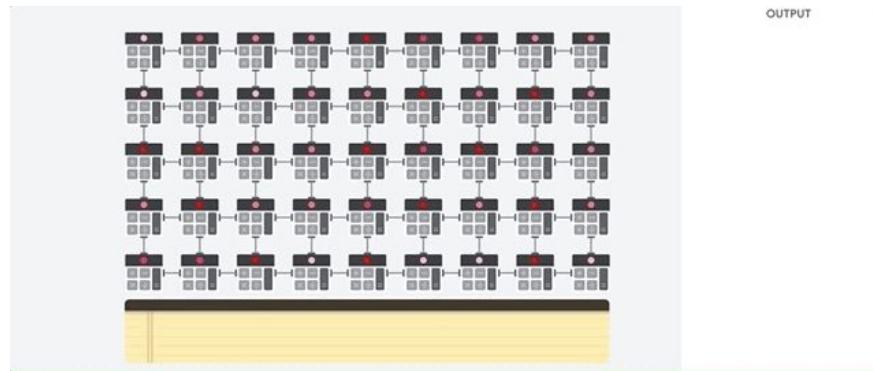


GEMM

- High-computational throughput on neural network calculations
- GPU compute unit is smaller but more, TPU has large compute units



GPU, many 8x4x8 ALU



TPU, systolic array 8x128x128

Matmul sharding

- Simple case: sharded matmul $Y = XA$
- Row sharding on A, column sharding on X



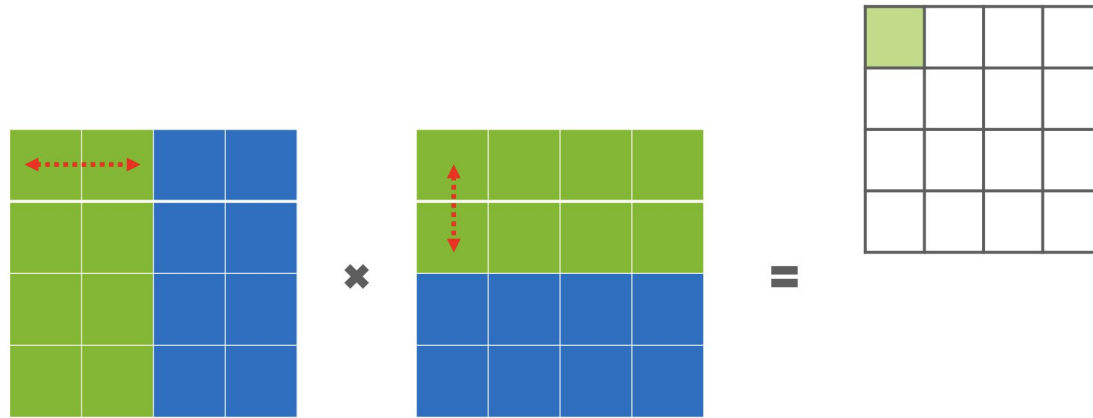
$$X = [X_1, X_2]$$

$$A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$$

$$Y = Y_1 + Y_2$$

Matmul sharding

- Row sharding on A, column sharding on X



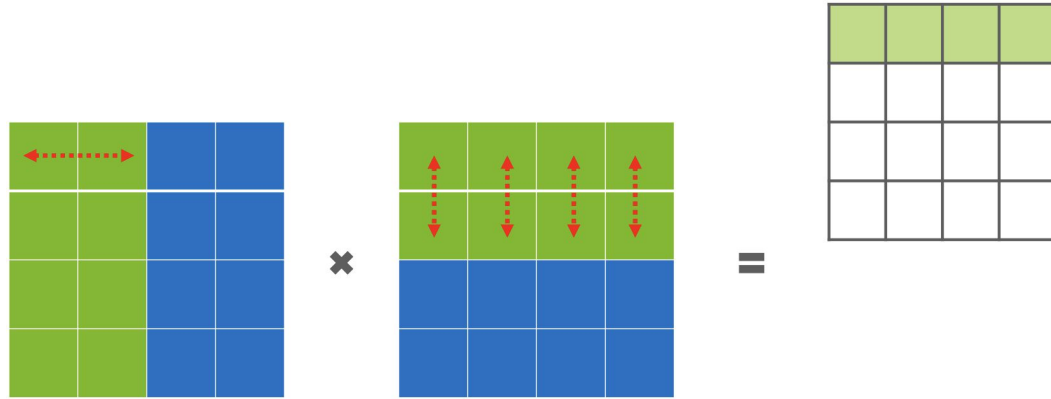
$$X = [X_1, X_2]$$

$$A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$$

$$Y = Y_1 + Y_2$$

Matmul sharding

- Row sharding on A, column sharding on X



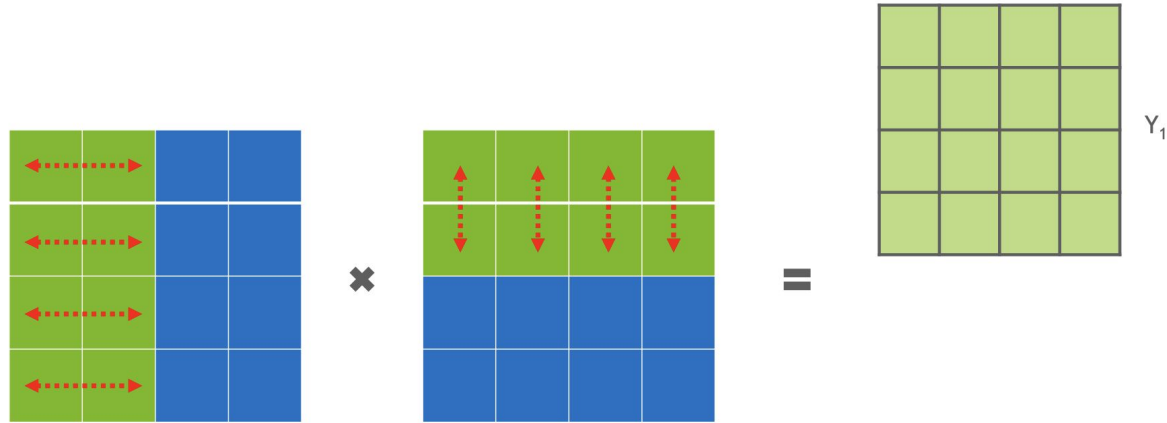
$$X = [X_1, X_2]$$

$$A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$$

$$Y = Y_1 + Y_2$$

Matmul sharding

- Row sharding on A, column sharding on X



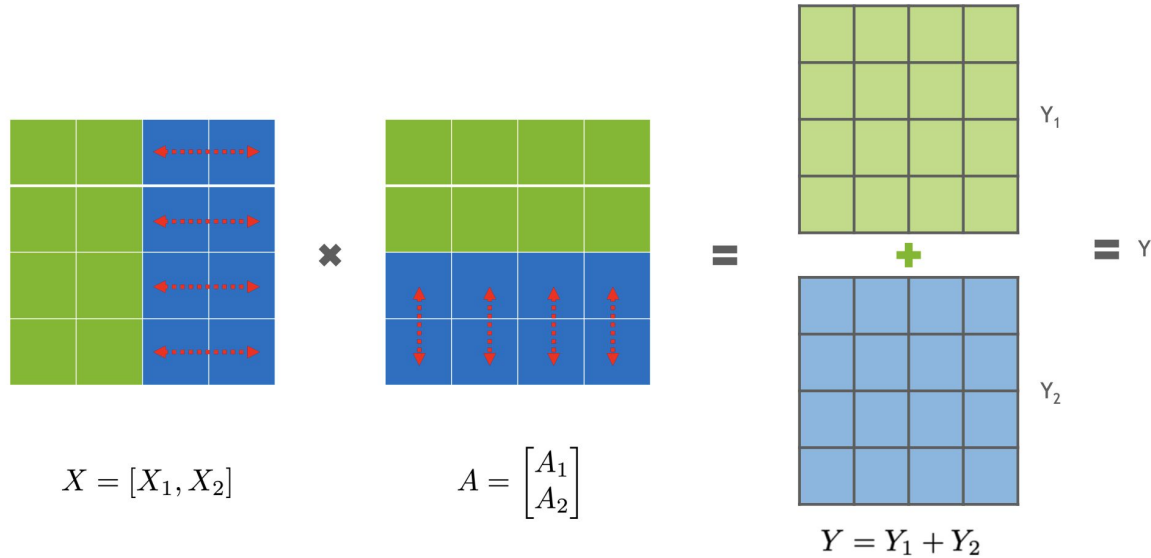
$$X = [X_1, X_2]$$

$$A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$$

$$Y = Y_1 + Y_2$$

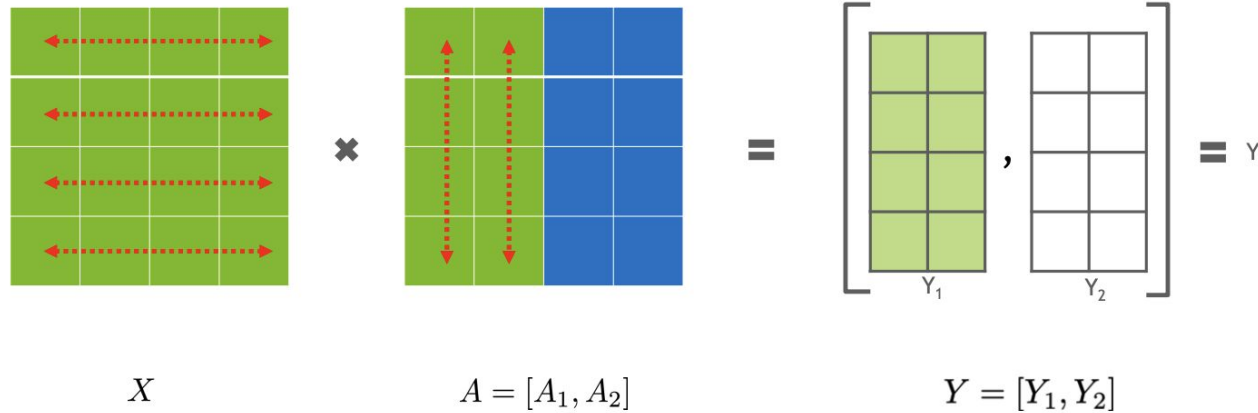
Matmul sharding

- Row sharding on A, column sharding on X



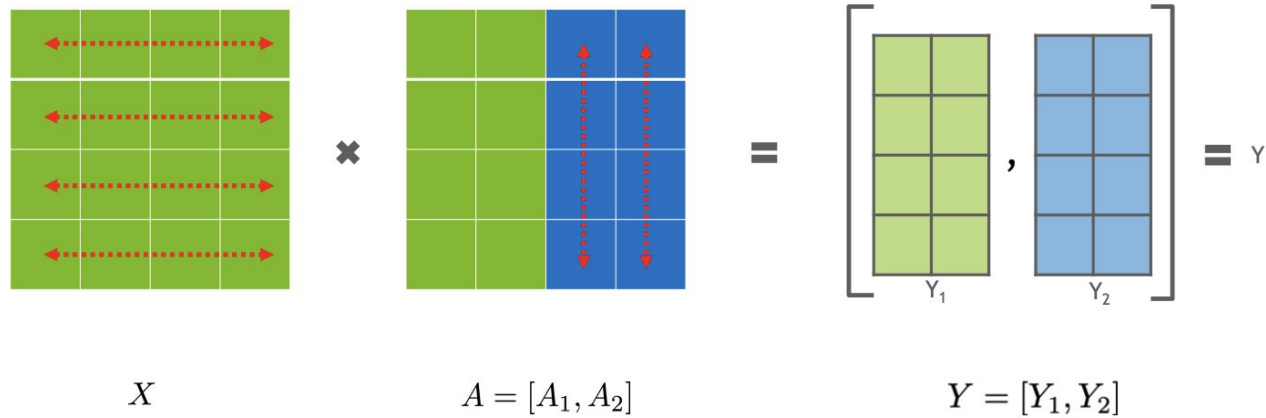
Matmul sharding

- Column sharding on A
- Replicate X



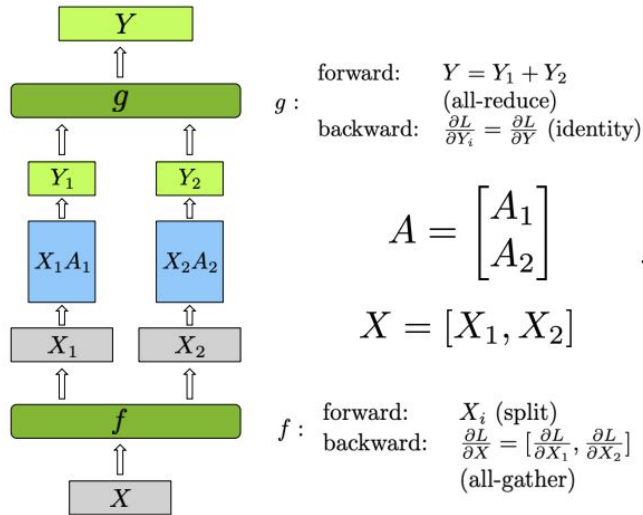
Matmul sharding

- Column sharding on A
- Replicate X

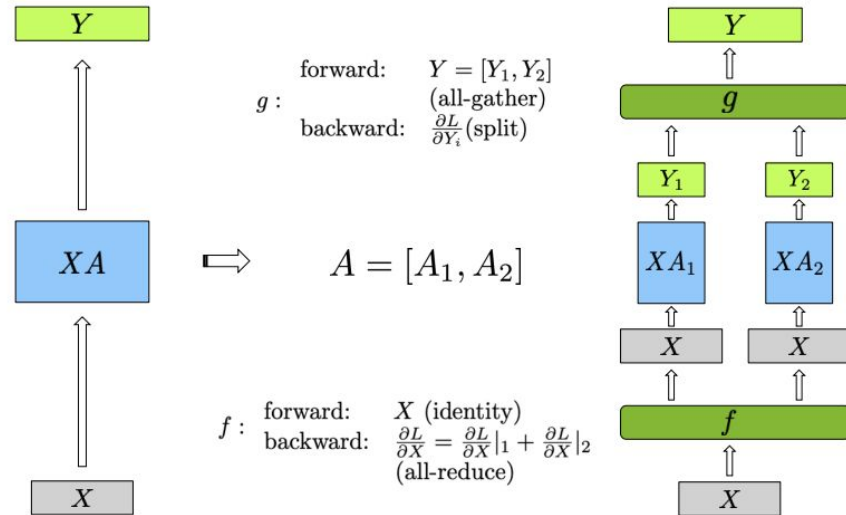


Matmul sharding

Row Parallel Linear Layer



Column Parallel Linear Layer

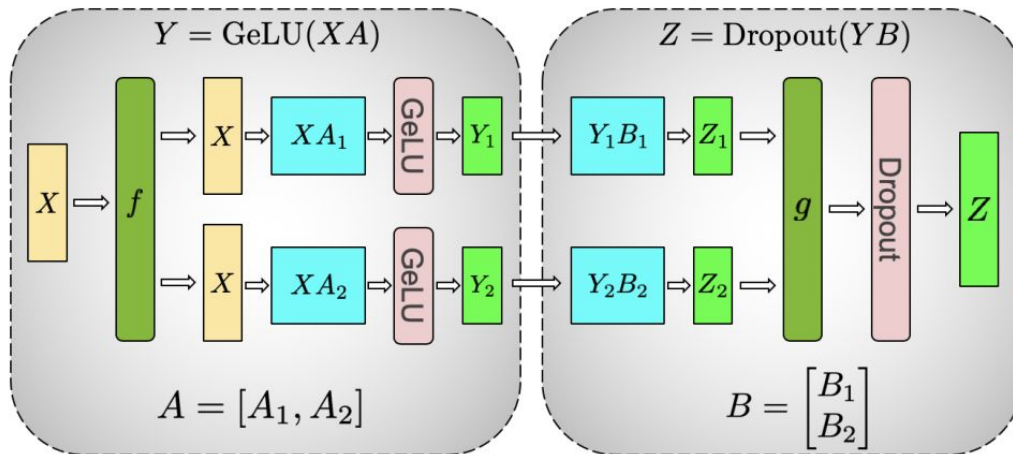


MLP sharding

- $Y = \text{GELU}(XA)B$
- Approach 1: split X column-wise and A row-wise
 - GeLU of sums $\neq$ sum of GeLUs $X = [X_1, X_2]$ $A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$ $Y = \text{GeLU}(X_1 A_1 + X_2 A_2)$
 - Requires synchronization before GeLU
- Approach 2: split A column-wise
 - No sharding on contraction dimension of X
 $A = [A_1, A_2]$ $[Y_1, Y_2] = [\text{GeLU}(X A_1), \text{GeLU}(X A_2)]$
 - No synchronization necessary

MLP sharding

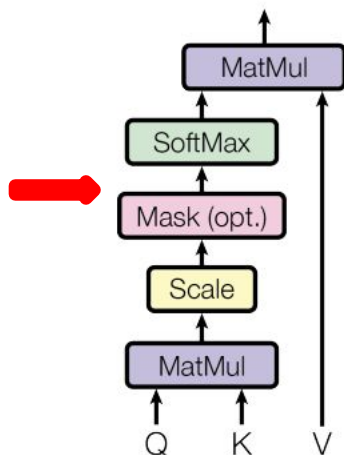
- **f** is **identity** operator in the forward pass and **all-reduce** in the backward pass while **g** is **all-reduce** in forward and **identity** in backward.



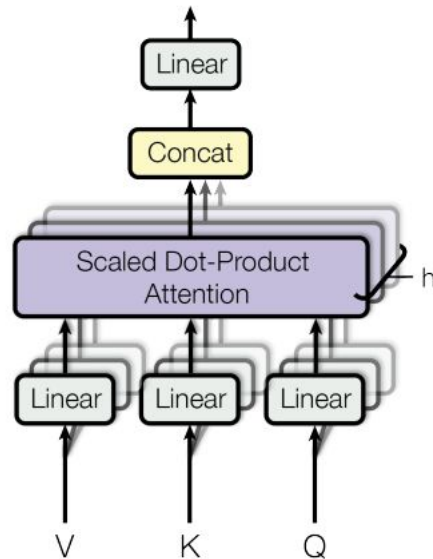
Attention sharding

- Nonlinearity in attention, so column sharding.

Scaled Dot-Product Attention

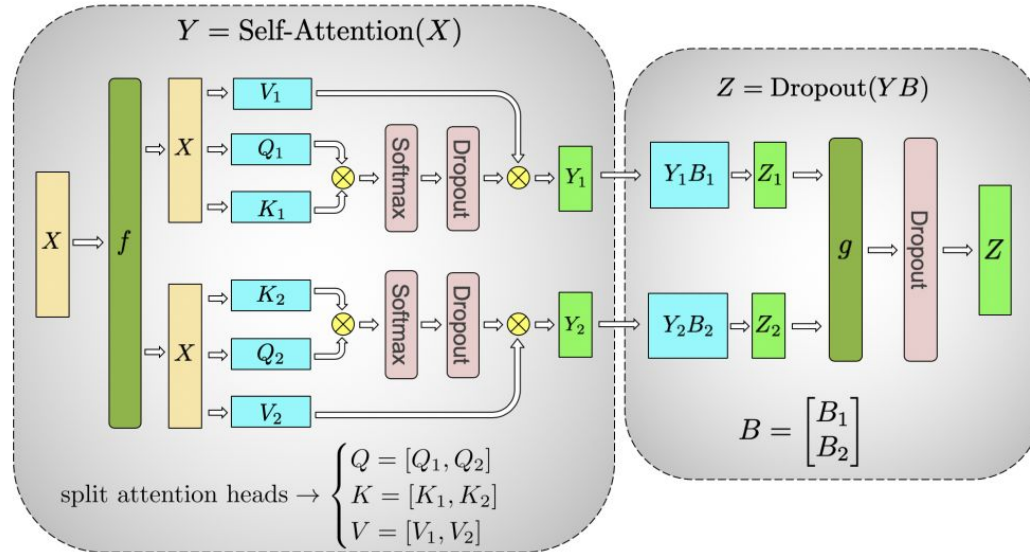


Multi-Head Attention



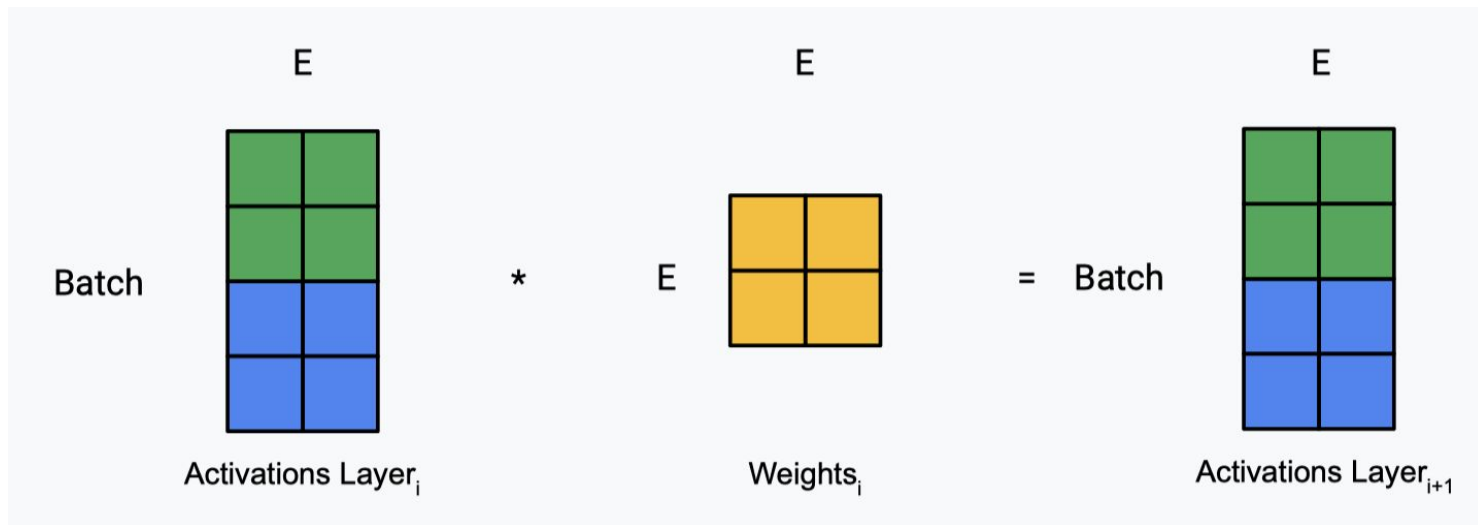
Attention sharding

- **f** is **identity** operator in the forward pass and **all-reduce** in the backward pass while **g** is **all-reduce** in forward and **identity** in backward.



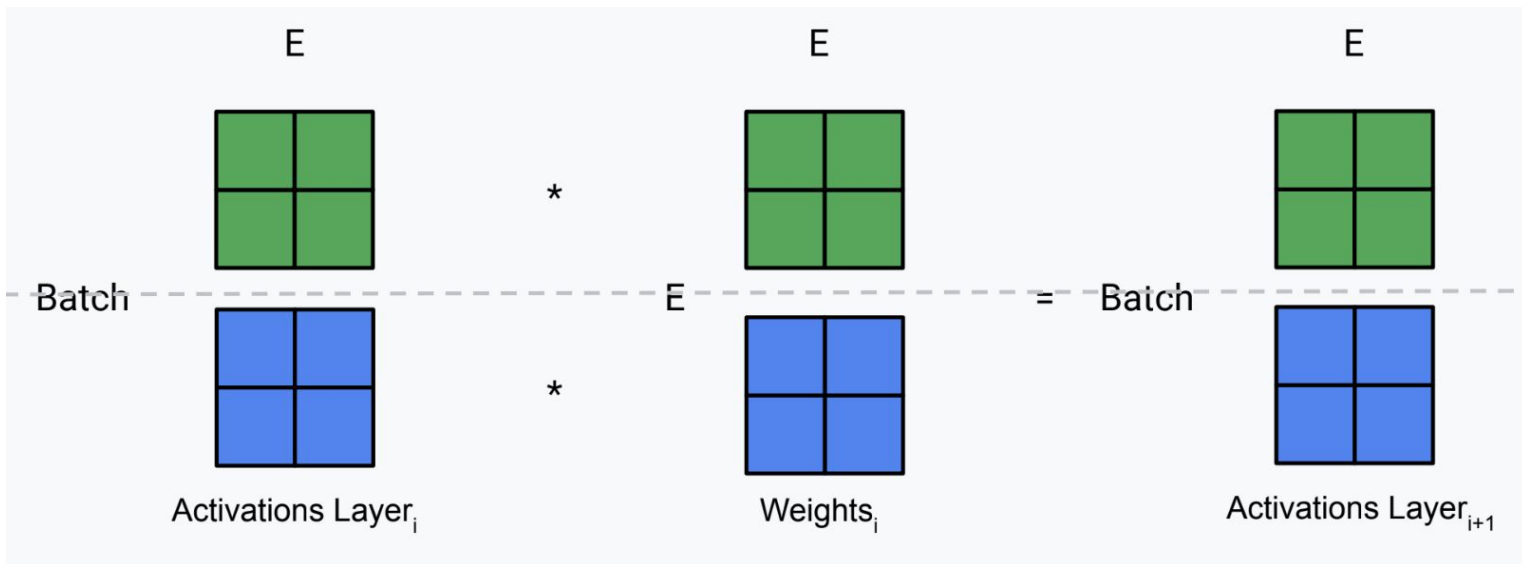
Data parallelism

- Activations need to be sharded – they're much bigger than weights. We can either shard them by rows or columns.
- It keeps working with a linear speedup as long as you scale batch linearly with number of TPUs



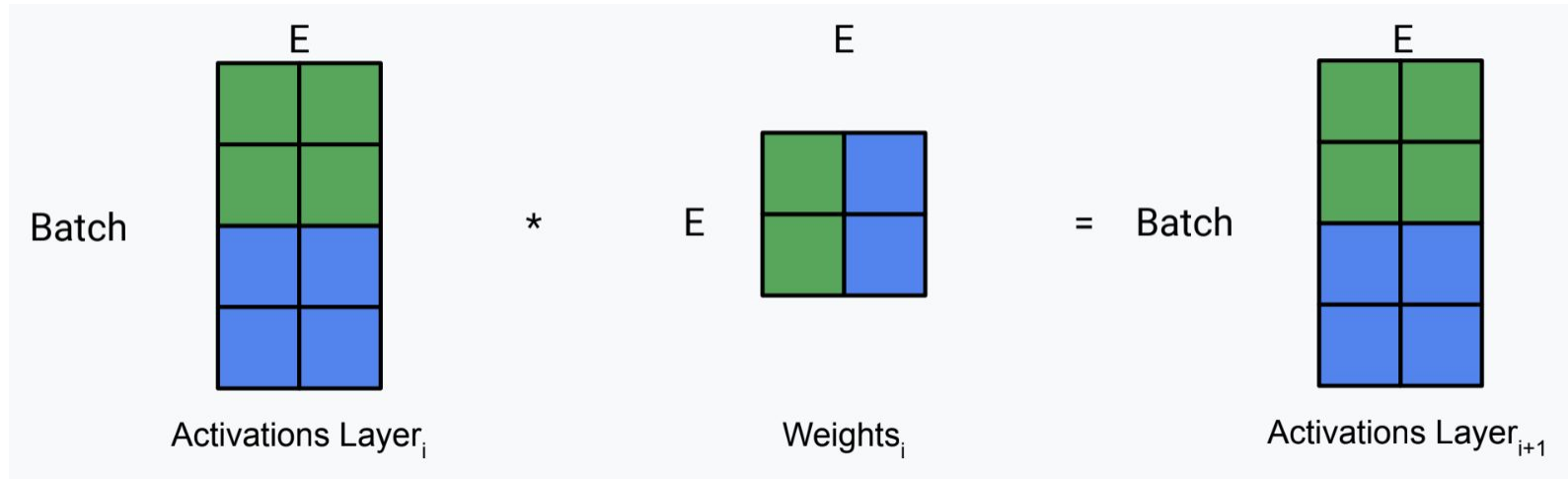
Data parallelism

- At 4 bytes per parameter we need 700GB for 175B GPT3
- But A100 has only 80GB.



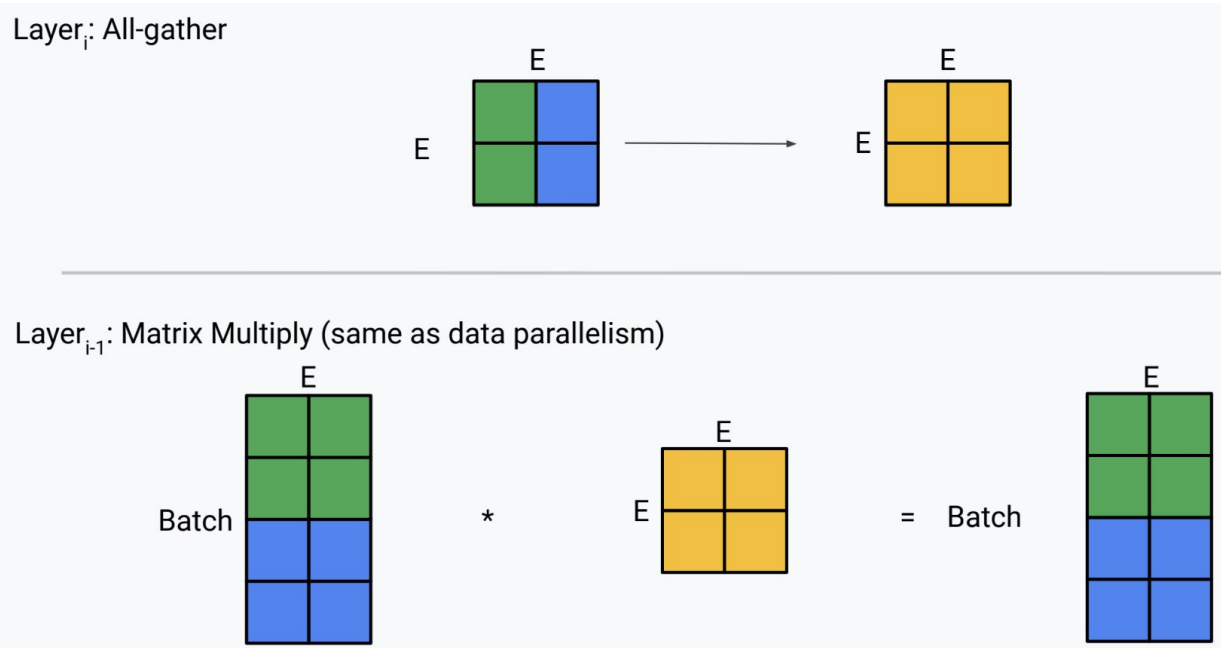
Fully sharded data parallelism

- All-gathering the next layer while doing the arithmetic on this layer
- Requires storing only 1 layer at a time (next slide)



Fully sharded data parallelism

- All-gathering the next layer while doing the arithmetic on this layer
- Requires storing only 1 layer at a time

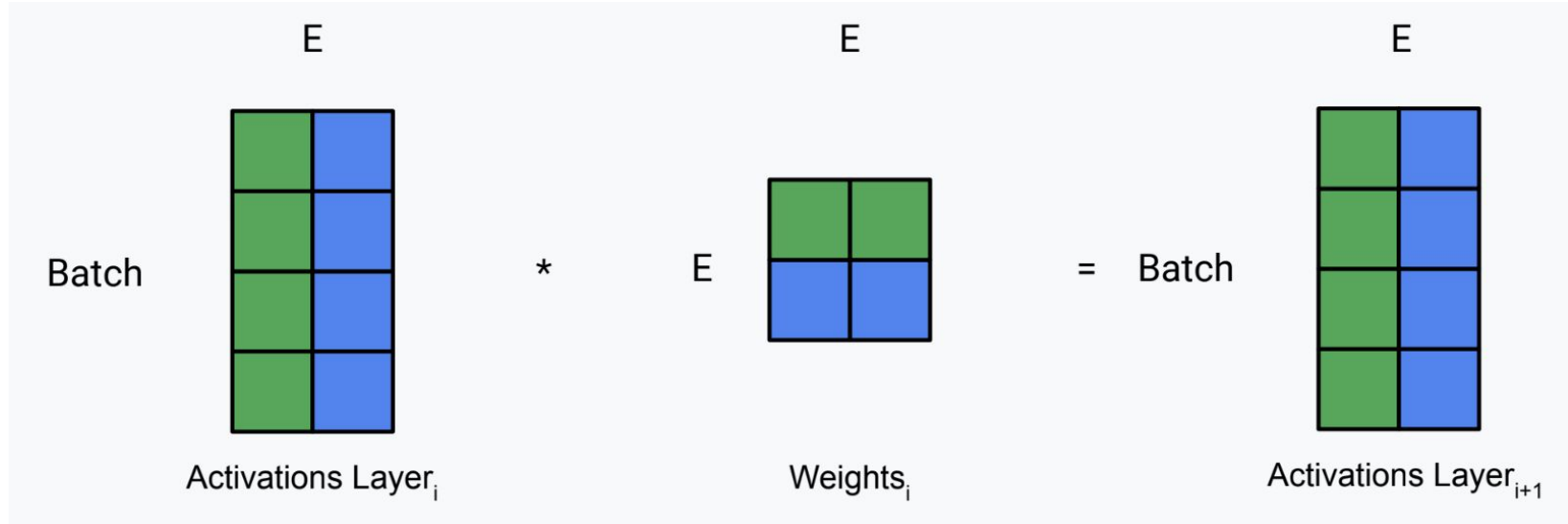


Fully sharded data parallelism

- The matrix multiply takes: $\text{FLOPs} = 2 * \text{BATCH_PER_CHIP} * E^2$
- The all gather requires reading $\sim 2 * E^2$ bytes (assuming bfloat16)
- Arithmetic intensity (FLOPs/bytes) = $2 * \text{BATCH_PER_CHIP} * E^2 / (2 * E^2 \text{ bytes}) = \text{BATCH_PER_CHIP FLOPs / byte.}$
- On TPU, ICI arithmetic intensity is 1018 FLOPs/byte, so we need our BATCH_PER_CHIP to be comfortably larger than 1018.

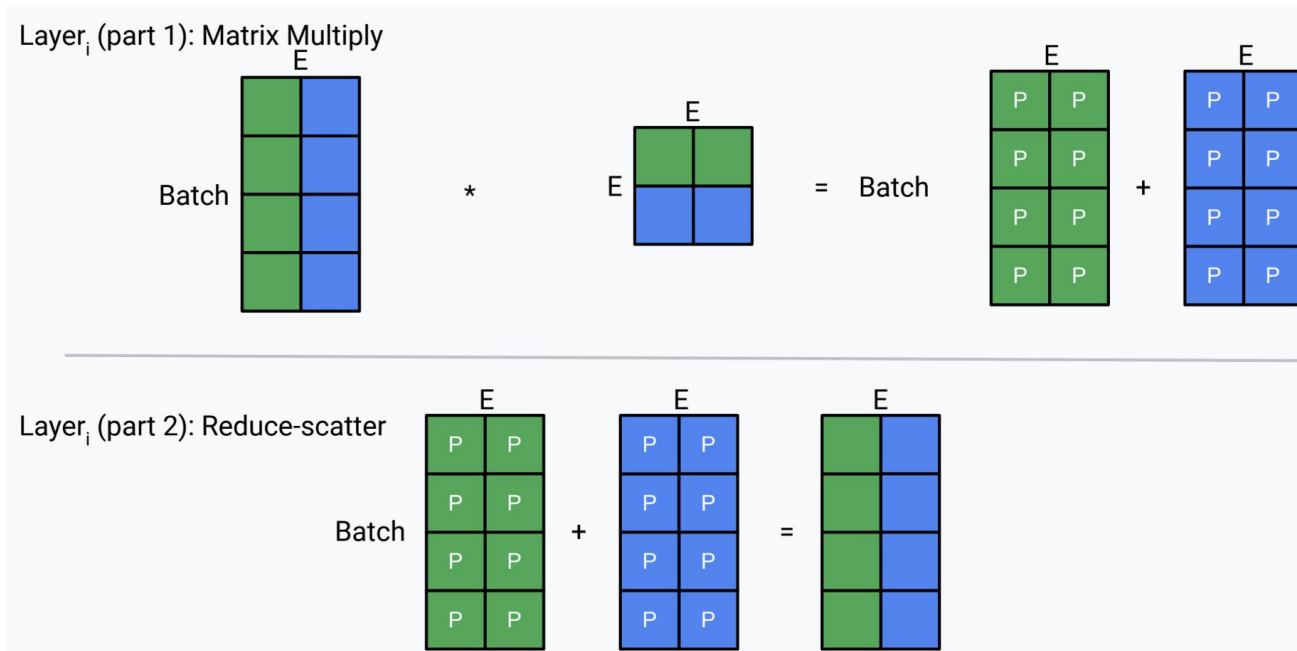
Tensor parallelism

- Shard the activation dimension



Tensor parallelism

- Shard the activation dimension

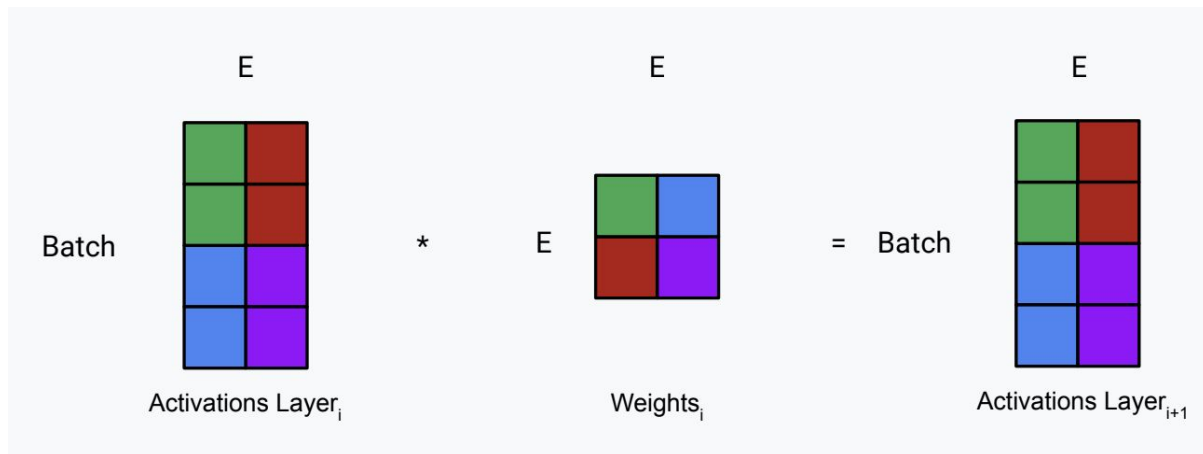


Tensor parallelism

- The matrix multiply takes (per chip): $\text{FLOPs} = 2 * \text{BATCH} * E^2 / (\text{NUM_TP_SHARDS})$
- The reduce scatter requires $\sim 2 * \text{BATCH} * E$ bytes (assuming bfloat16)
- Arithmetic intensity (FLOPs/bytes) = $E / \text{NUM_TP_SHARDS}$.
- TPU ICI arithmetic intensity is 1018 FLOPs/byte, so we need E to be comfortably larger $1018 * \text{NUM_TP_SHARDS}$.

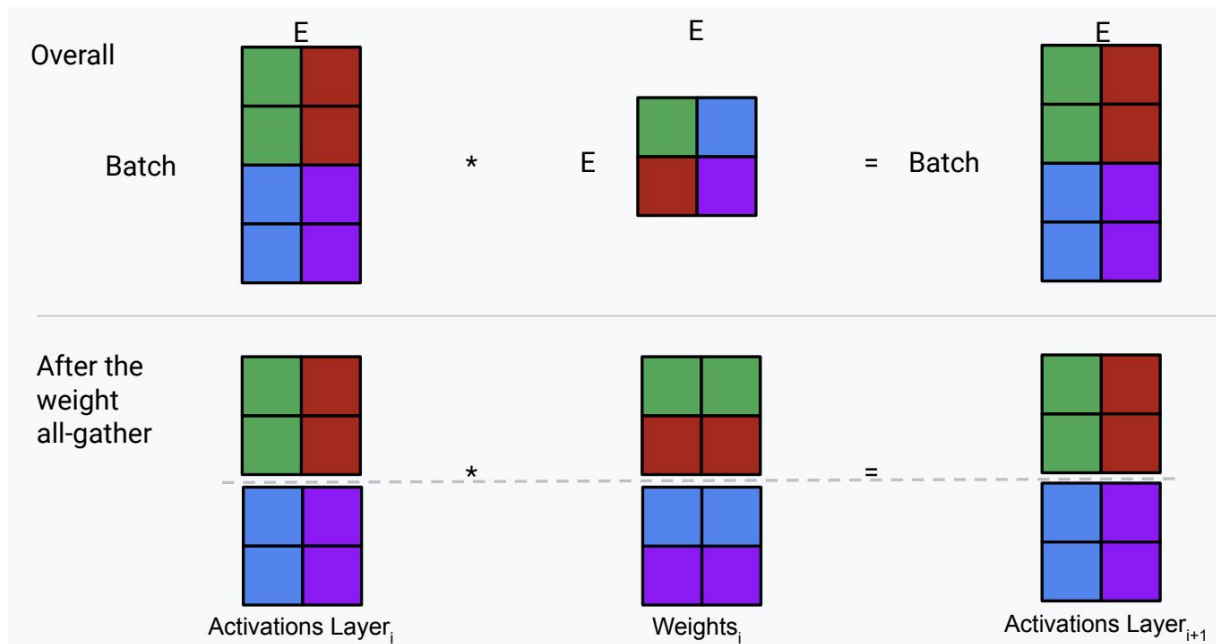
FSDP with TP

- While processing layer i , all-gather layer weights for layer $i+1$ for FSDP
- Then it just becomes TP



FSDP with TP

- While processing layer i , all-gather layer weights for layer $i+1$ for FSDP
- Then it just becomes TP



FSDP with TP

- Given NUM_TP_SHARDS and NUM_FSDP_SHARD
- When both:
 - $E > 2 * 1018 * \text{NUM_TP_SHARDS}$
 - $\text{NUM_TP_SHARD} * \text{BATCH_PER_CHIP} > 2 * 1018$
- 256 TPU: 16 for TP and 16 for FSDP
 - $E > \sim 32k$
 - $\text{BATCH_PER_CHIP} > 125$

References

- Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.
- Geng, Xinyang, et al. "Multimodal masked autoencoders learn transferable representations." *arXiv preprint arXiv:2205.14204* (2022).
- Jain, Ajay, et al. "Zero-shot text-guided object generation with dream fields." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
- Crowson, Katherine, et al. "Vqgan-clip: Open domain image generation and editing with natural language guidance." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.
- Shridhar, Mohit, Lucas Manuelli, and Dieter Fox. "Cliport: What and where pathways for robotic manipulation." *Conference on robot learning*. PMLR, 2022.
- Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." *Advances in neural information processing systems* 35 (2022): 23716-23736.
- Tsimpoukelli, Maria, et al. "Multimodal few-shot learning with frozen language models." *Advances in Neural Information Processing Systems* 34 (2021): 200-212.
- Chen, Xi, et al. "Pali: A jointly-scaled multilingual language-image model." *arXiv preprint arXiv:2209.06794* (2022).
- Chen, Xi, et al. "Pali-x: On scaling up a multilingual vision and language model." *arXiv preprint arXiv:2305.18565* (2023).
- Chen, Xi, et al. "Pali-3 vision language models: Smaller, faster, stronger." *arXiv preprint*
- Zhu, Deyao, et al. "Minigt-4: Enhancing vision-language understanding with advanced large language models." *arXiv preprint arXiv:2304.10592* (2023).
- Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing systems* 36 (2024).
- Liu, Haotian, et al. "Improved baselines with visual instruction tuning." *arXiv preprint arXiv:2310.03744* (2023).
- Reid, Machel, et al. "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context." *arXiv preprint arXiv:2403.05530* (2024).
- Tang, Zineng, et al. "Any-to-any generation via composable diffusion." *Advances in Neural Information Processing Systems* 36 (2024).
- Jin, Yang, et al. "Unified language-vision pretraining with dynamic discrete visual tokenization." *arXiv preprint arXiv:2309.04669* (2023).
- Jin, Yang, et al. "Video-LaVIT: Unified Video-Language Pre-training with Decoupled Visual-Motional Tokenization." *arXiv preprint arXiv:2402.03161* (2024).
- Sun, Quan, et al. "Generative pretraining in multimodality." *arXiv preprint arXiv:2307.05222* (2023).
- Liu, Hao, et al. "World Model on Million-Length Video And Language With RingAttention." *arXiv preprint arXiv:2402.08268* (2024).