

Lecture 1

Introduction to Multimodality

Deep & Early Fusion

Zinkovich Viktoriia

Special thanks to **Vlad Shakhuro** for
the slides and lecture content



 VladShakhuro



darkstore

success = **True** (*orig*)
put both the alphabet soup
and the tomato sauce in the
basket

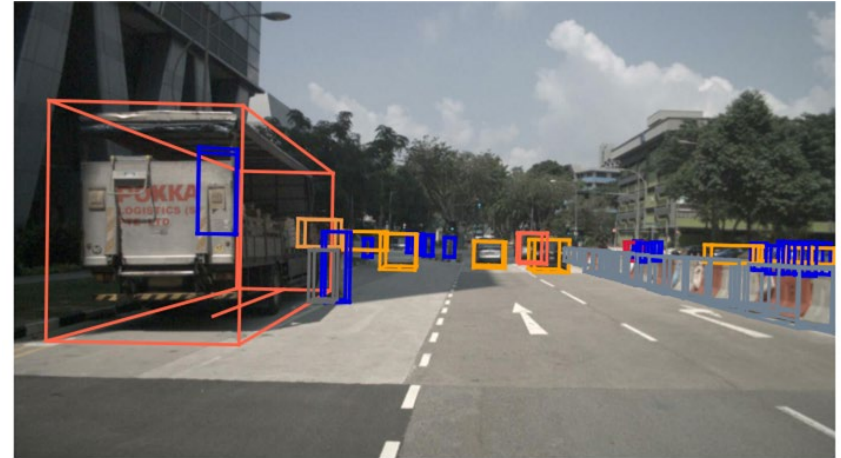
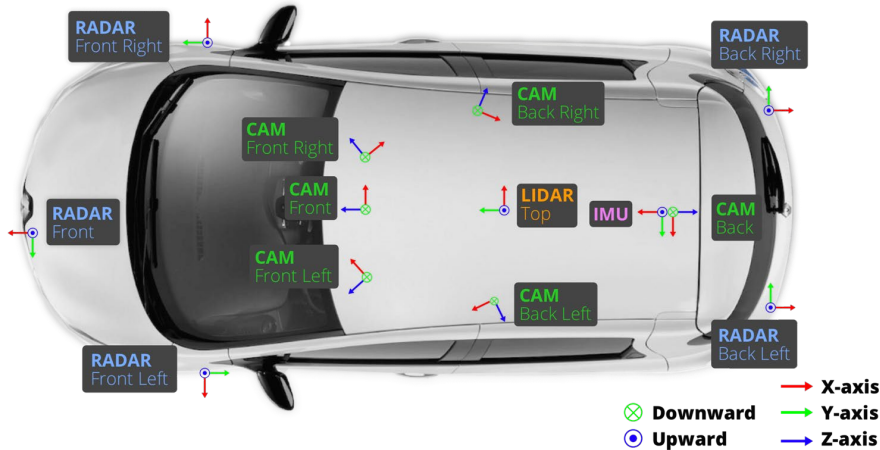


success = **False**
take both the alphabet soup
and the tomato sauce **and**
put them in the basket



robustness of VLA models against
reformulations of textual prompts

 VladShakhuro



autonomic vehicles

Introduction: Course Plan



Day 1
Image modality

Day 3
Data generation
in MLLMs



Day 5
3D models



Day 2
Video modality



Day 4
Action modality

Join the Chat!

Please join the chat with lecture slides and jupyter notebooks



Lecture Plan

- 1 Idea of Multimodality
- 2 Benchmarks: Static and Dynamic (*Arena*)
- 3 Classification of Multimodal Models:
Deep & Early Fusion
- 4 Explore code & architecture of the most
vivid exemplars

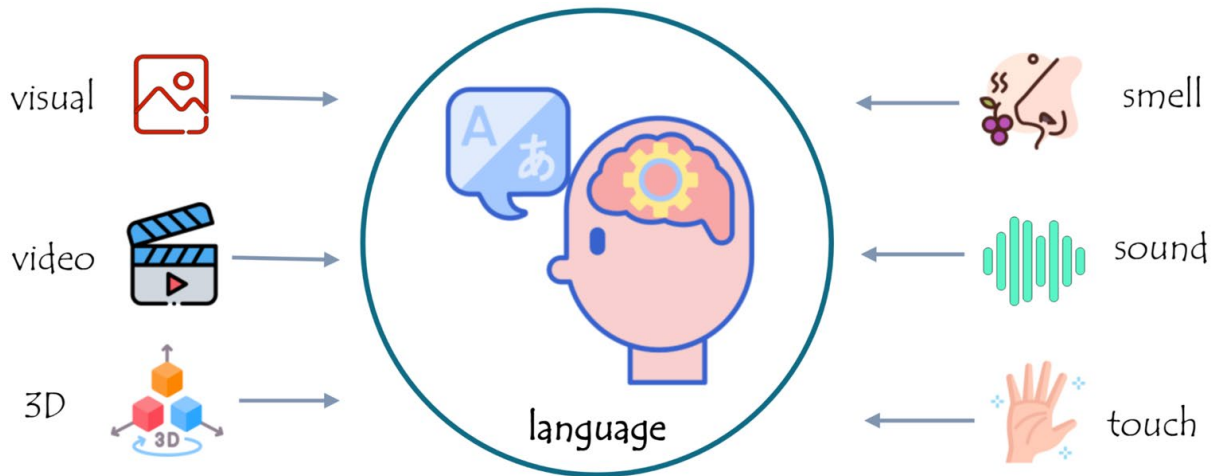


Idea of Multimodality

→ Use world knowledge and **reasoning** from LLMs

→ Explore various applications that emerge at the **intersection of the modalities**

→ Boost quality of models using **patterns in multimodal data** that don't exist in unimodal data



2.1

Benchmarks

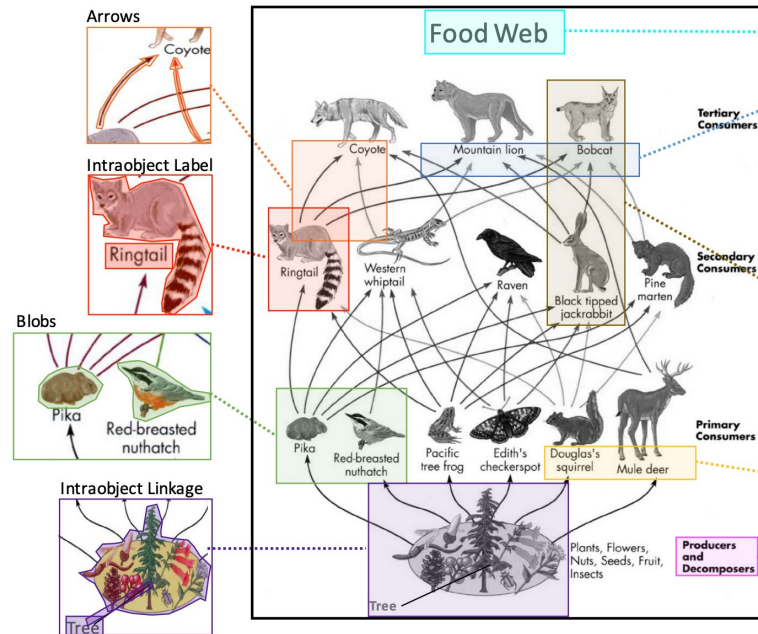
Static benchmarks used to validate developed models



Static Benchmarks: AI2D (2016)

AI2 Diagrams (AI2D)

- diagram parse graphs are available
- 15k multiple choice questions for 5k school grade diagrams
- evaluation metric: **accuracy**
- **prompt the model accurately** to obtain well formatted answer



From the above food web diagram, what will lead to an increase in the population of deer? a) increase in lion; b) decrease in plants; c) decrease in lion; d) increase in pika

Static Benchmarks: GQA (2019)

General Question Answering

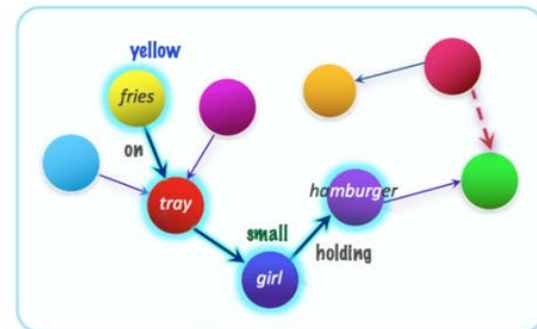
- questions generated using **scene graph** of images
- 22.6M questions for 113k images
- evaluation metric: **accuracy** (one-word answers)
- can create **balanced** dataset



Pattern: What/Which <type> [do you think] <is> <object>, <attr> or <decoy>?
Program: Select: <object> → Choose <type>: <attr>|<decoy>
Reference: The food on the red object left of the small girl that is holding a hamburger
Decoy: brown

What color is the food on the red object left of the small girl that is holding a hamburger, yellow or brown?

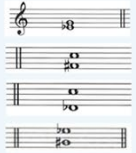

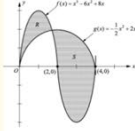
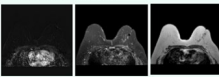
Select: hamburger → Relate: girl, holding → Filter size: small → Relate: object, left → Filter color: red → Relate: food, on → Choose color: yellow | brown



Static Benchmarks: MMMU (2024)

Massive Multi -discipline Multimodal Understanding benchmark

- 11.5k questions from 6 university disciplines
- answers are extracted using **regex**
- evaluation metric: **accuracy**

| Art & Design | Business | Science | Health & Medicine |
|---|--|--|--|
| <p>Question: Among the following harmonic intervals, which one is constructed incorrectly?</p> <p>Options:</p> <p>(A) Major third <i><image 1></i></p> <p>(B) Diminished fifth <i><image 2></i></p> <p>(C) Minor seventh <i><image 3></i></p> <p>(D) Diminished sixth <i><image 4></i></p>  | <p>Question: ...The graph shown is compiled from data collected by Gallup <i><image 1></i>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82?</p> <p>Options:</p> <p>(A) 0 (B) 0.2142</p> <p>(C) 0.3571 (D) 0.5</p>  | <p>Question: <i><image 1></i> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.</p> <p>Options:</p> <p>(A) $\int_0^{1.5} [f(x) - g(x)] dx$</p> <p>(B) $\int_0^{1.5} [g(x) - f(x)] dx$</p> <p>(C) $\int_0^2 [f(x) - g(x)] dx$</p> <p>(D) $\int_0^2 [g(x) - x(x)] dx$</p>  | <p>Question: You are shown subtraction <i><image 1></i>, T2 weighted <i><image 2></i> and T1 weighted axial <i><image 3></i> from a screening breast MRI. What is the etiology of the finding in the left breast?</p> <p>Options:</p> <p>(A) Susceptibility artifact</p> <p>(B) Hematoma</p> <p>(C) Fat necrosis (D) Silicone granuloma</p>  |
| <p>Subject: Music; Subfield: Music;</p> <p>Image Type: Sheet Music;</p> <p>Difficulty: Medium</p> | <p>Subject: Marketing; Subfield: Market Research;</p> <p>Image Type: Plots and Charts;</p> <p>Difficulty: Medium</p> | <p>Subject: Math; Subfield: Calculus;</p> <p>Image Type: Mathematical Notations;</p> <p>Difficulty: Easy</p> | <p>Subject: Clinical Medicine; Subfield: Clinical Radiology;</p> <p>Image Type: Body Scans: MRI, CT.;</p> <p>Difficulty: Hard</p> |

Static Benchmarks: TextVQA (2019)

- **text recognition** on an existing VQA dataset
- 45k questions for 28k images
- **10 answers** per question
- evaluation metric: **VQA accuracy** (100% correct if 3 humans provided that answer)



(a) Question: which processor Brand is featured on the top left?
Answer: intel



(b) Question: which brand are the crayons?
Answer: crayola



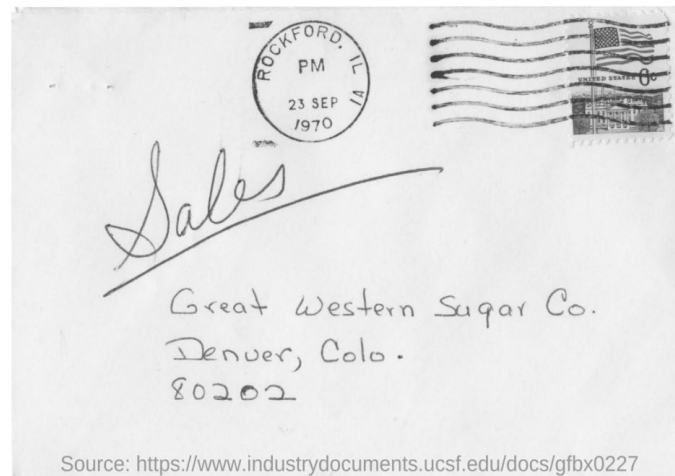
(c) Question: what is the name of the bose speaker style in these boxes?
Answer: freestyle



(d) Question: what is the license number?
Answer: cu58 ckk

Static Benchmarks: WACV (2021)

- 50k questions for 12k images
- **documents** mostly from 1960 —2000
- industries: tobacco, food, drug, chemical, fossil fuel
- **evaluation metrics:**
 - Average Normalized Levenshtein Similarity
 - Accuracy



Q: Mention the ZIP code written?

A: 80202

Q: What date is seen on the seal at the top of the letter?

A: 23 sep 1970

Q: Which company address is mentioned on the letter?

A: Great western sugar Co.

2.2

Benchmarks

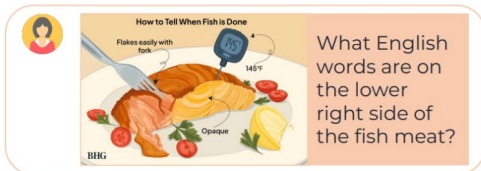
Dynamic benchmarks:
Arena for multimodal
models



Dynamic Benchmarks: WildVision Arena

Rules

- Chat with two anonymous models
- Continue to chat until you identify a winner
- Vote for the better one with reason



Model A

On the lower right side of the cooked fish, the word "Opaque" is labeled.

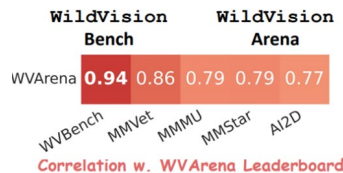
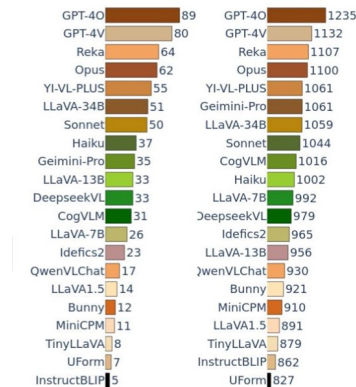
Model B

The English word on the lower right side of the fish meat is "Opaque."

Reason Both Model A and Model B answer correctly regarding the text.

Vote A is Better B is Better Tie Both are bad

Model A: Claude-3-Sonnet, Model B: GPT-4V WVArena Elo Ratings Submit



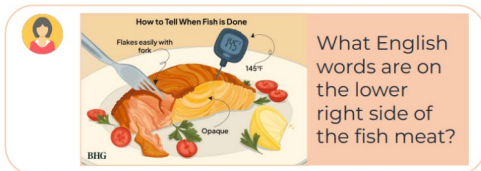
Model Evaluation System:

- Two models provide answers to users on the same question
- Best response is selected through comparison
- Elo rating system is compiled (similar to chess rankings)

Dynamic Benchmarks: WildVision Arena

Rules

- Chat with two anonymous models
- Continue to chat until you identify a winner
- Vote for the better one with reason



Model A

On the lower right side of the cooked fish, the word "Opaque" is labeled.

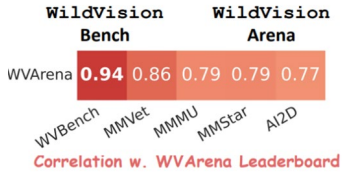
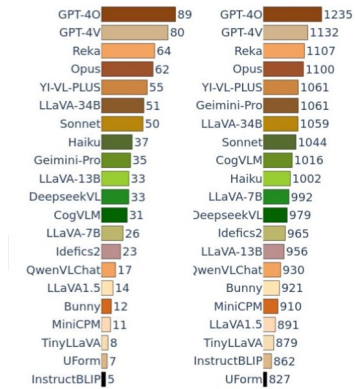
Model B

The English word on the lower right side of the fish meat is "Opaque."

Reason Both Model A and Model B answer correctly regarding the text.

Vote A is Better B is Better Tie Both are bad

Model A: Claude-3-Sonnet, Model B: GPT-4V WVArena Elo Ratings Submit



- Leaderboard data from May 2024 (pre-NeurIPS submission)
- Closed/proprietary models listed first, followed by open-source models
- Elo ratings show strong correlation with MMMU benchmark scores

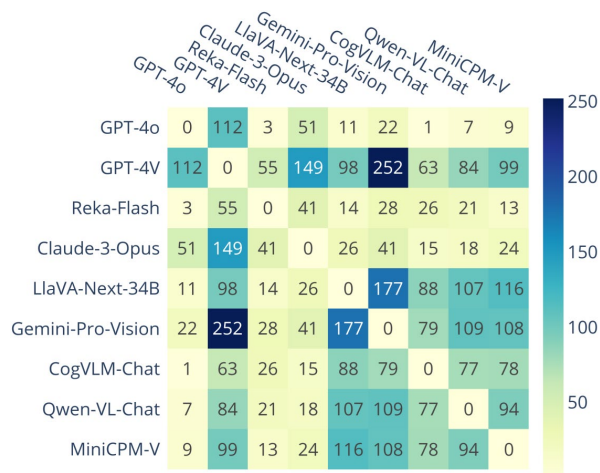
Wild Vision Arena : UI

The screenshot shows the Wild Vision Arena interface. At the top, there are navigation links for 'Arena', 'Leaderboard', and 'About Us'. The main title is 'WildVision Arena : Benchmarking Multimodal LLMs in the Wild' by WildVision Team @ AI2 + UCSB + UWaterloo. A warning message states: 'Your data will be logged for research purposes. Please do not include any confidential, personal, or other sensitive information.' Below this, there is a search bar and a section for 'Choose models to sample from' with a grid of model buttons including gemini-pro-vision, gemini-1.5-flash-latest, llama-v1.6-34b, llama-v1.6-vicuna-7b, MiniCPM-V, cogvlm-chat-hf, Qwen-VL-Chat, uform-gen2-qwen-500m, tiny-llava-v1-hf, Bunny-v1.0-3B, claude-3-opus-20240229, claude-3-sonnet-20240229, claude-3-haiku-20240307, deepseek-v1-7b-chat, yi-vl-plus, Reka-Flash, idefics2-8b-chatty, gpt-4-vision-preview, and gpt-4o. A 'Clear' button is below the model selection. The chat interface shows an 'Image' input with a map of a game world, 'Model A' and 'Model B' tabs, and a question: 'Given my horse's location on this map, what is the quickest route to reach it?'. Model A's response provides a detailed route: 'Based on the map, your current location is marked by the pin/stamp icon at Lookout Landing. The quickest route to reach your horse, which is marked with a horse icon at Rito Village, would be to head northwest along the roads/paths shown on the map. It looks like you'll need to traverse the area labeled "Tabantha Frontier" to get from your current spot to Rito Village where your horse is located. The exact path is a bit unclear from this zoomed out view, but generally heading northwest from Lookout Landing should get you to your horse's position at Rito Village by the most direct route possible given the geography shown.' Model B's response is: 'I'm sorry, but I don't have access to your specific map or location information. Could you please provide more context or details so I can assist you better?'

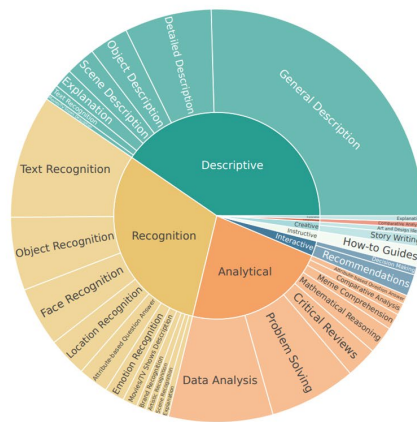
- Utilize AI arenas for free access to language models
- **No VPN required** — direct access to GPT or Claude services
- Unclear which specific AI model is being used in conversations
- Cost-effective solution for accessing **premium AI capabilities**

Wild Vision Arena : Details

- **8k rounds** + matrices with visualization showing which models lost more to whom



- Questions can be quite general – need checking **specialized knowledge** (statistical benchmarks, domain specialists)



Question category



Image Domain

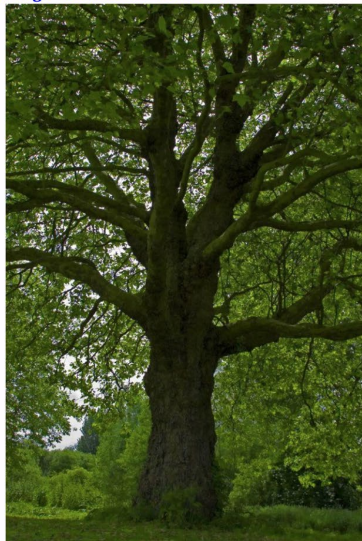
Wild Vision Arena : WVBench

Image [Entertainment-Movies/TV Shows]



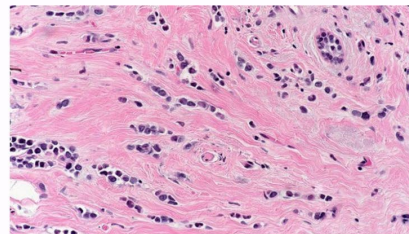
[Descriptive-Movies/TV Shows] **Text Prompt:** What are the two giraffe characters on this movie poster doing?

Image [Natural-Plants]



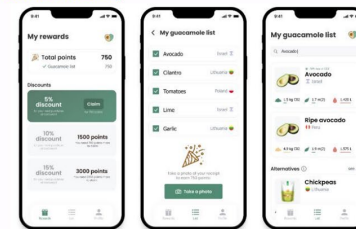
[Analytical-Problem Solving] **Text Prompt:** How likely is it to snow after this picture was taken? What would change with this type of tree before it's likely to snow?

Image [Expert-Health and Medicine]



[Recognition-Object] **Text Prompt:** what type of tumor is this?

Image [Entertainment-Web and Mobile Apps Screenshots]



[Analytical-Critical Reviews] **Text Prompt:** Review each screenshot carefully, focusing on different aspects of usability...

- 500 data samples from Arena with expert annotation
- evaluation is done using GPT-4o as judge and Claude-3-Sonnet as reference

3

Architectures

Classification of
multimodal models: **Deep** &
Early Fusion



High-level Classification

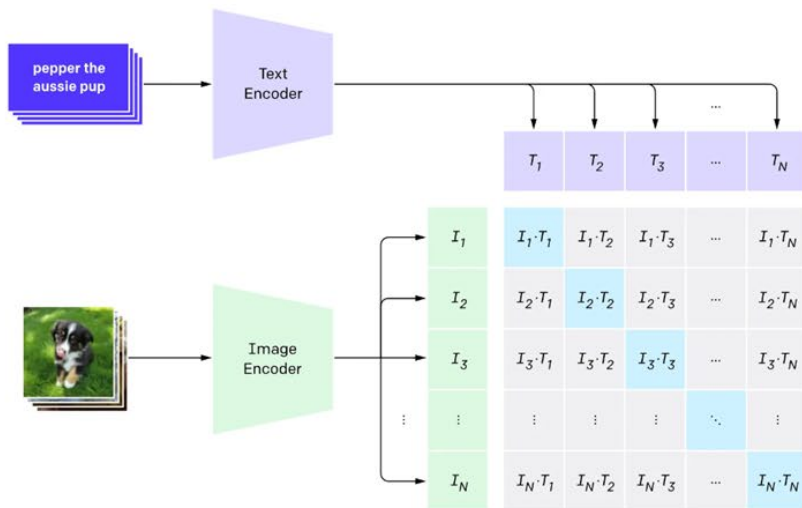
Multimodal models can be classified in 2 main types (4 subtypes)
based on the **fusion of input modalities**



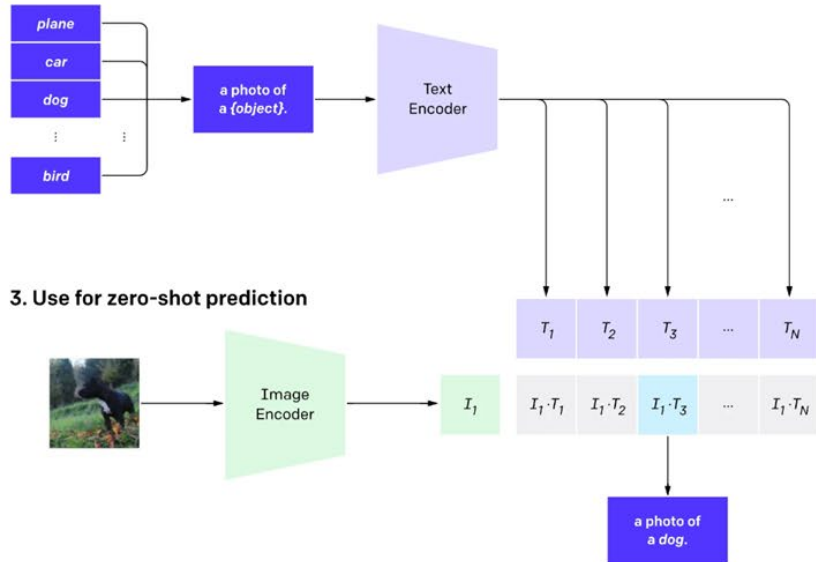
Preliminary 1: Image Encoders

CLIP: 400M (image, text) pairs, 500×V100 GPUs for pretraining

1. Contrastive pre-training



2. Create dataset classifier from label text

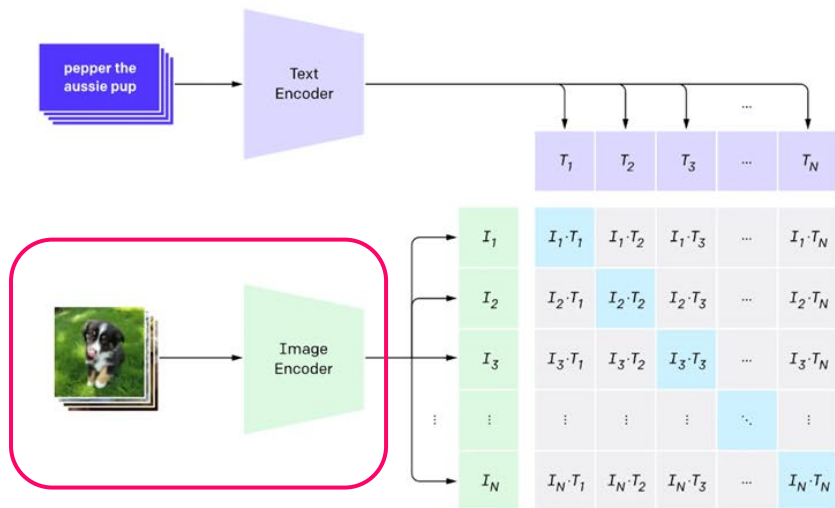


3. Use for zero-shot prediction

Preliminary 1: Image Encoders

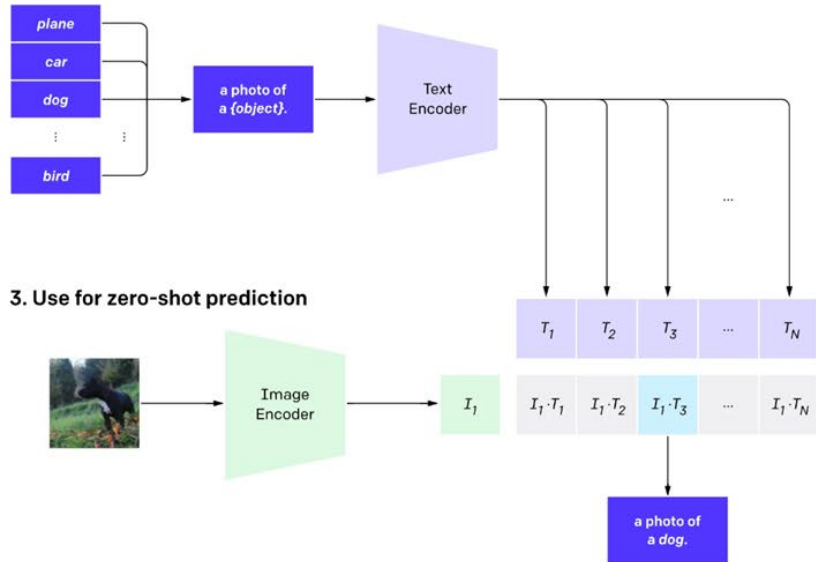
CLIP: 400M (image, text) pairs, 500×V100 GPUs for pretraining

1. Contrastive pre-training



we need only image encoder

2. Create dataset classifier from label text

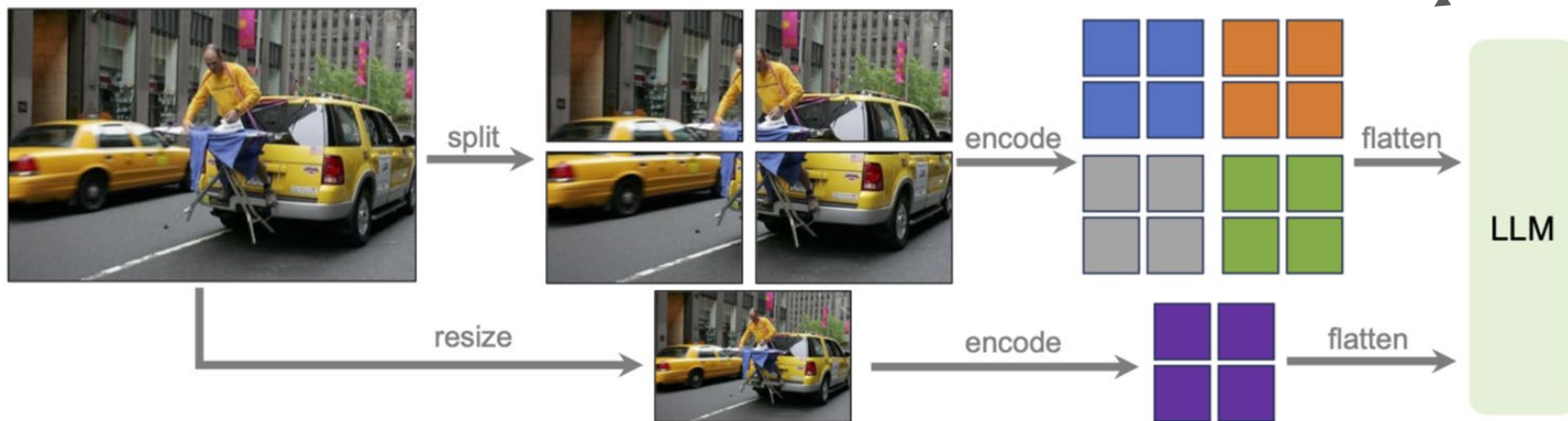


3. Use for zero-shot prediction

Preliminary 1: High Resolution

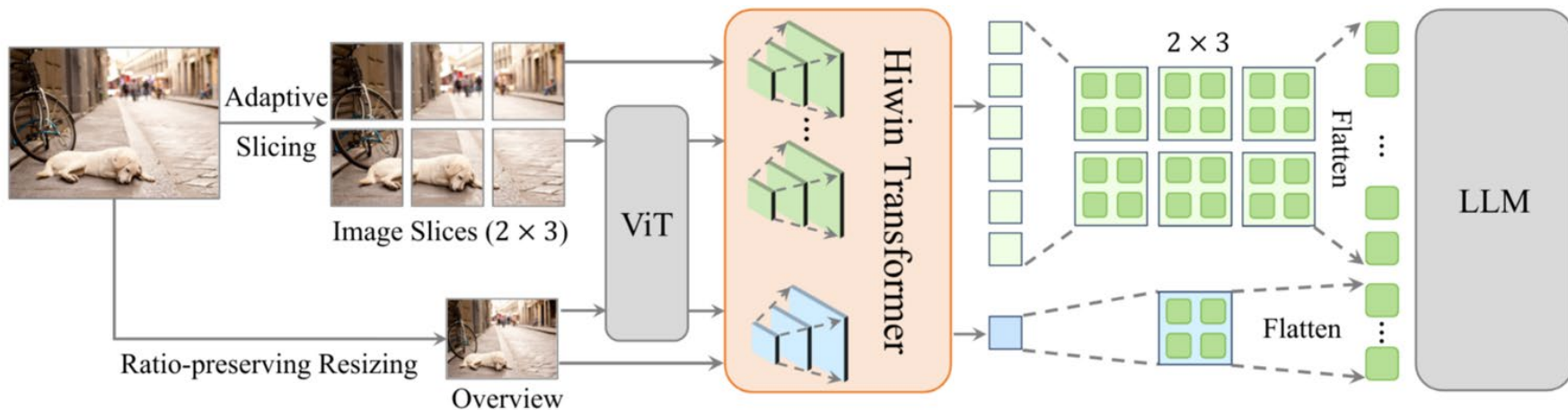
1 **Slicing** —break resolution into chunks + overlap

2 **Dual-branch** —two branches with different resolutions / pyramid of resolution



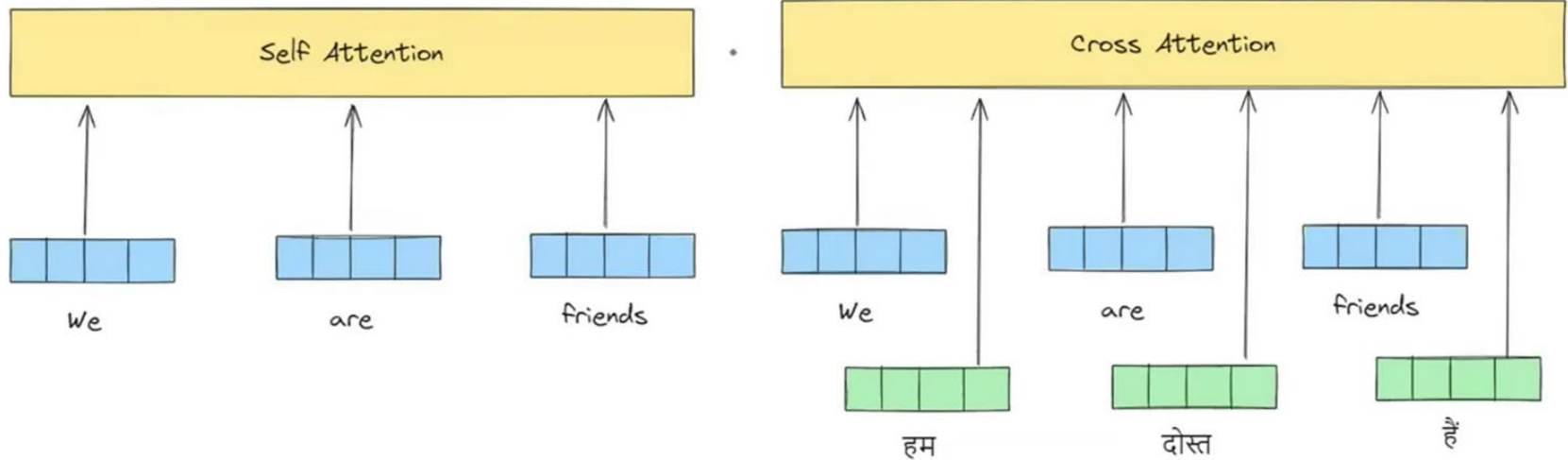
Preliminary 1: High Resolution

In multimodal models, CLIP is frequently used (ViT + pyramid of resolutions)



Preliminary 2: Self-Attention vs Cross-Attention

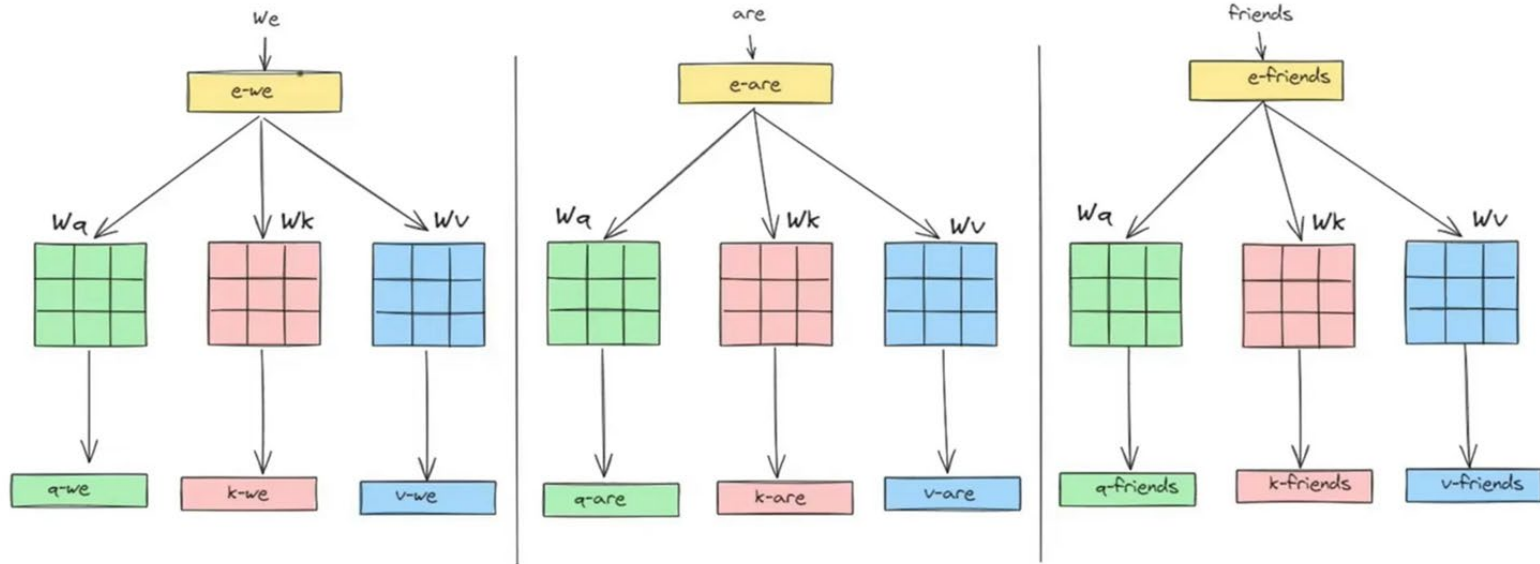
Self-attention works on a **single sequence** by generating contextual embeddings for its tokens, **cross-attention** handles **two sequences** at once



Preliminary 2: Self-Attention vs Cross-Attention

Input embeddings go through three linear transformations to produce the **query, key, and value vectors** for each word

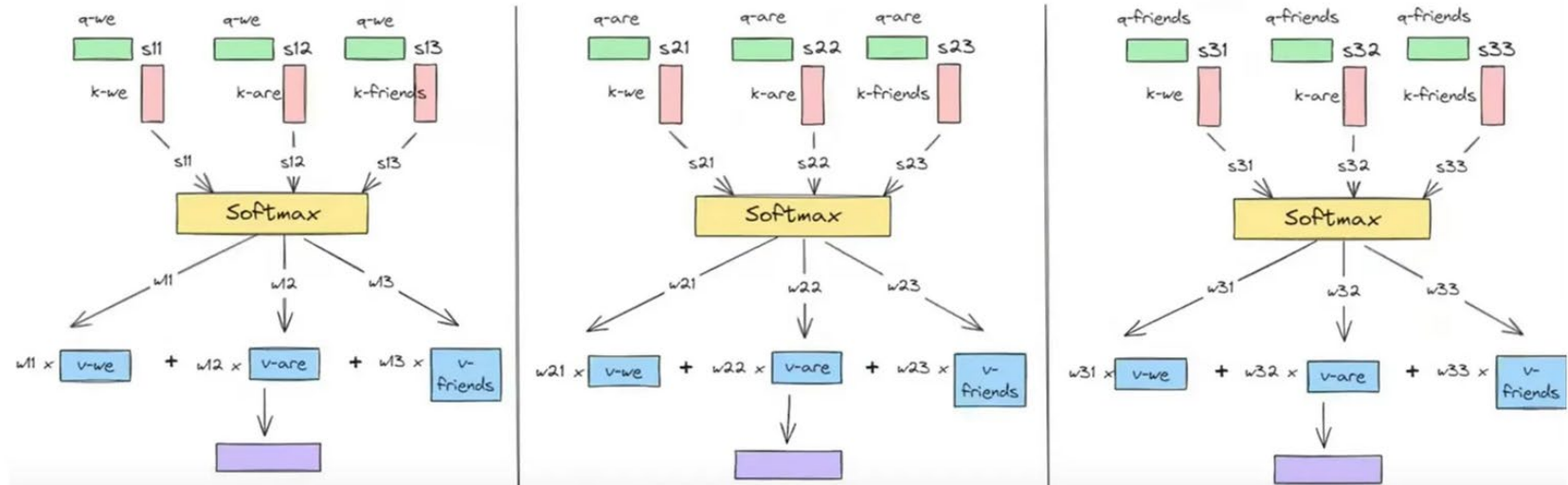
query, key,



Preliminary 2: Self-Attention vs Cross-Attention

Query vectors of each word are dot-multiplied with the key vectors of every other word in the sequence \rightarrow attention scores

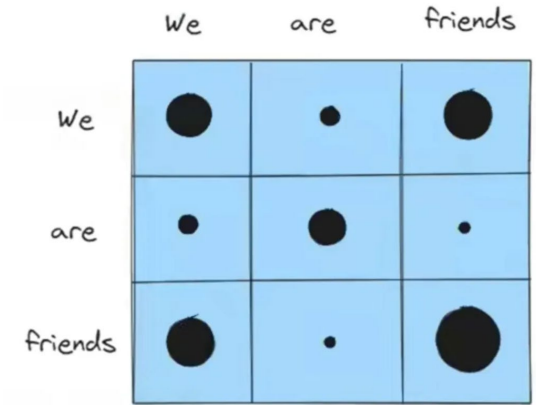
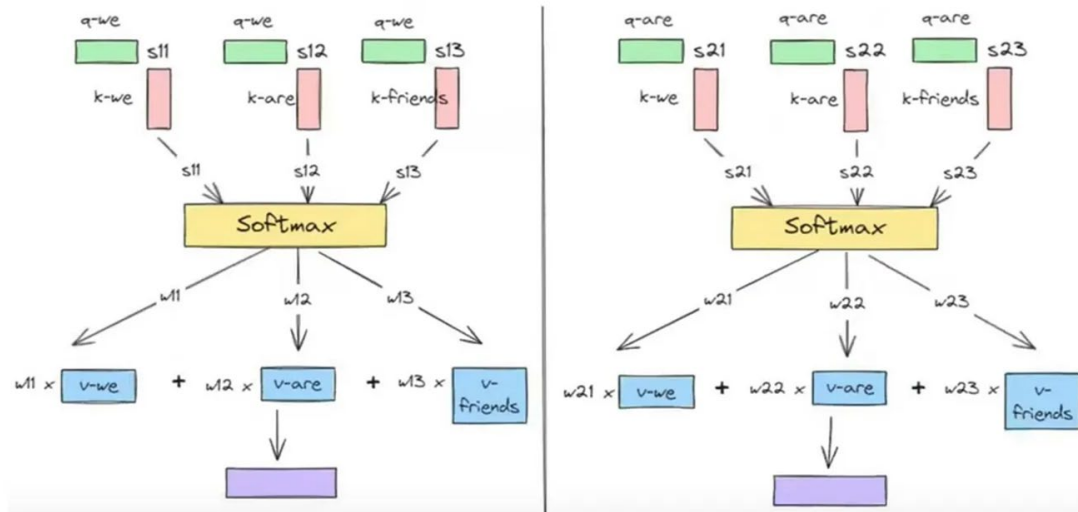
of every other word



Preliminary 2: Self-Attention vs Cross-Attention

Query vectors of each word are dot-multiplied with the key vectors in the sequence

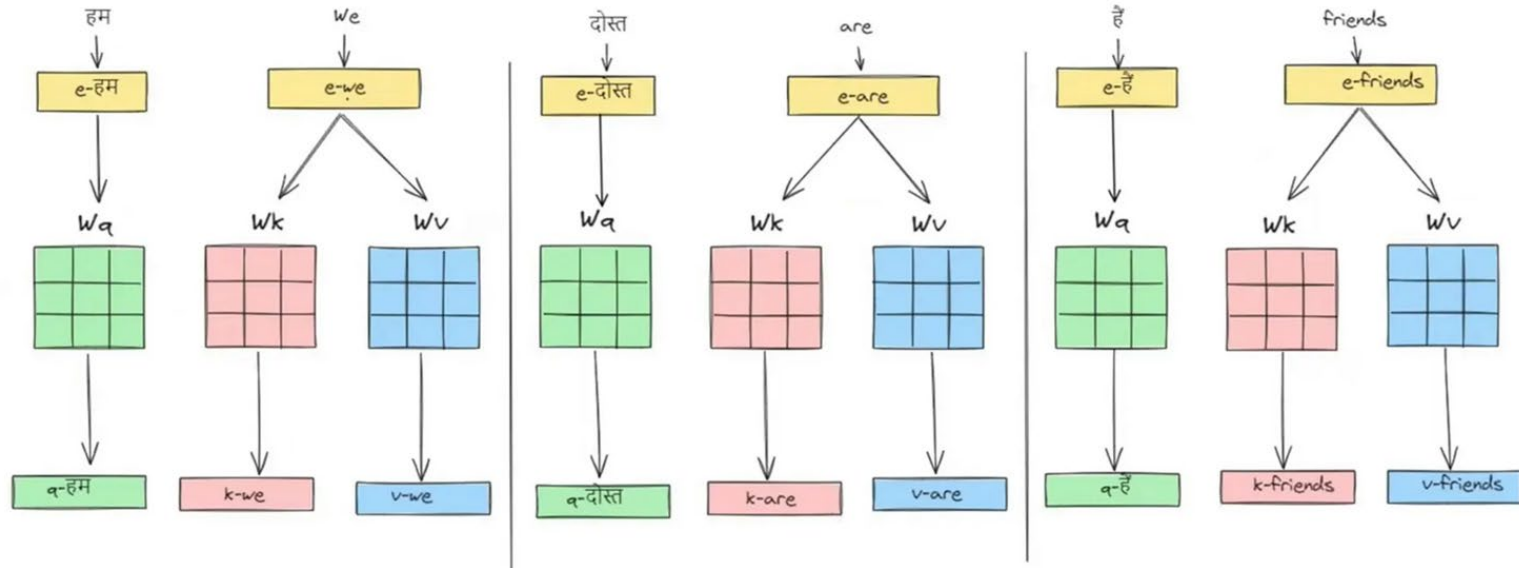
of every other word



larger circles indicate a stronger relationship between the corresponding words

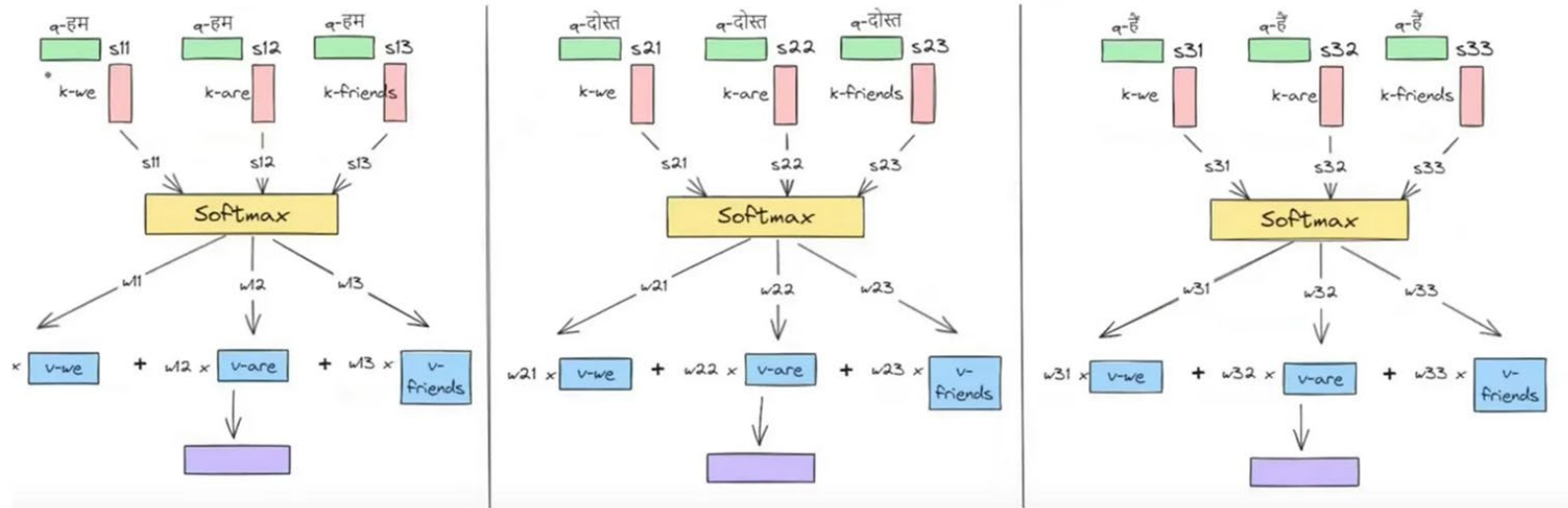
Preliminary 2: Self-Attention vs Cross-Attention

Cross-attention generates **query** vectors from the **output** sequence (Hindi), while **key and value vectors** are derived from the **input** sequence (English)



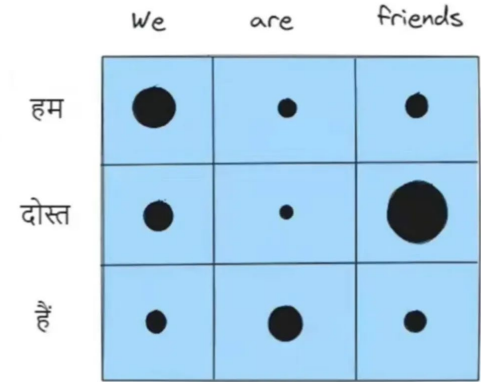
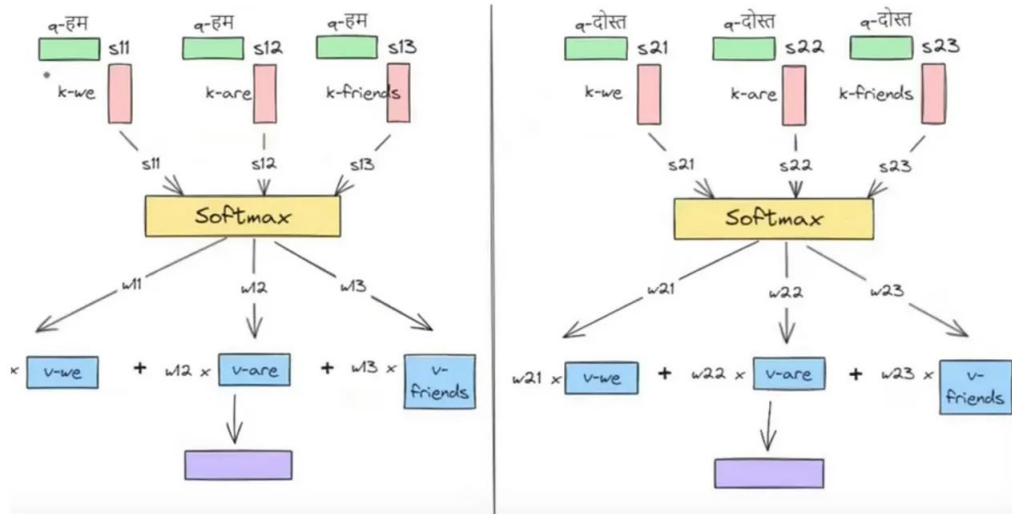
Preliminary 2: Self-Attention vs Cross-Attention

Cross-attention generates **query** vectors from the **output** sequence (Hindi), while **key and value vectors** are derived from the **input** sequence (English)



Preliminary 2: Self-Attention vs Cross-Attention

Cross-attention generates **query** vectors from the **output** sequence (Hindi), while **key and value vectors** are derived from the **input** sequence (English)



larger circles indicate a stronger relationship between the corresponding words

Preliminary 2: Self-Attention vs Cross-Attention

Demonstrating how the **contextual embedding of a word** from other words:

$$ce_we = 0.8 \times e_we + 0.1 \times e_are + 0.1 \times e_friends$$

$$ce_are = 0.15 \times e_we + 0.75 \times e_are + 0.1 \times e_friends$$

$$ce_friends = 0.2 \times e_we + 0.1 \times e_are + 0.7 \times e_friends$$

Self-attention

is calculated using information

$$ce_हम = 0.5 \times e_we + 0.3 \times e_are + 0.2 \times e_friends$$

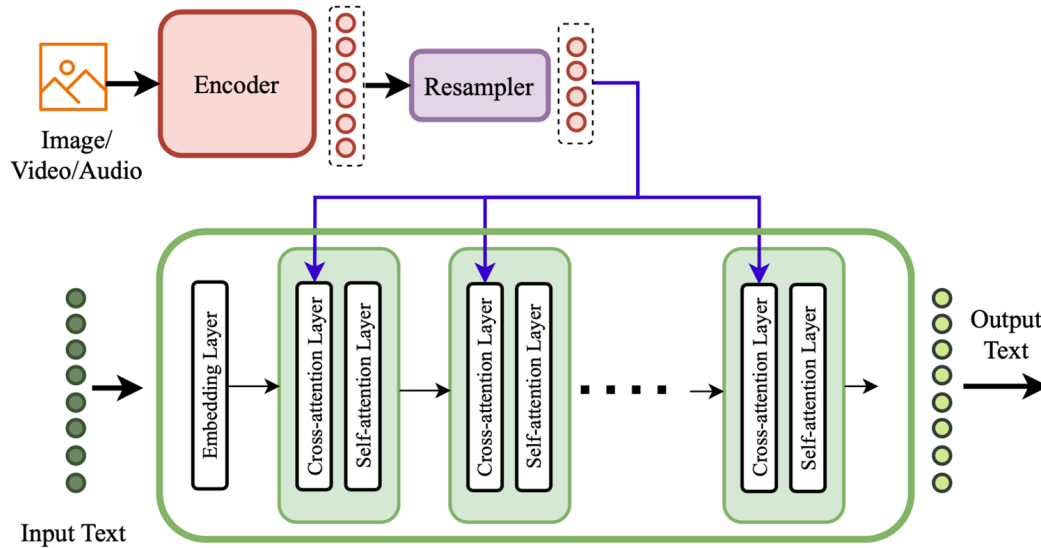
$$ce_दोस्त = 0.2 \times e_we + 0.2 \times e_are + 0.6 \times e_friends$$

$$ce_हैं = 0.3 \times e_we + 0.4 \times e_are + 0.3 \times e_friends$$

Cross-attention

Preliminary 2: Cross-Attention for 3 modalities?

Cross-attention between **three sequences** (modalities)



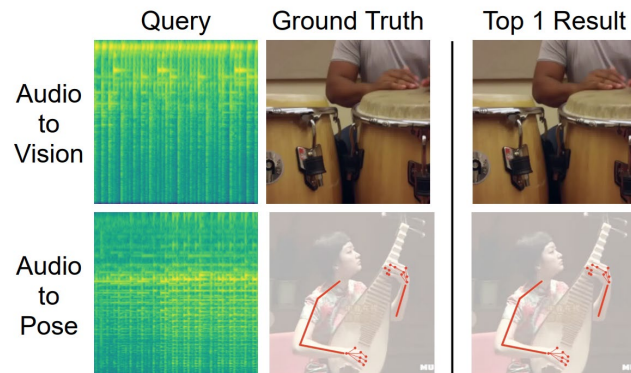
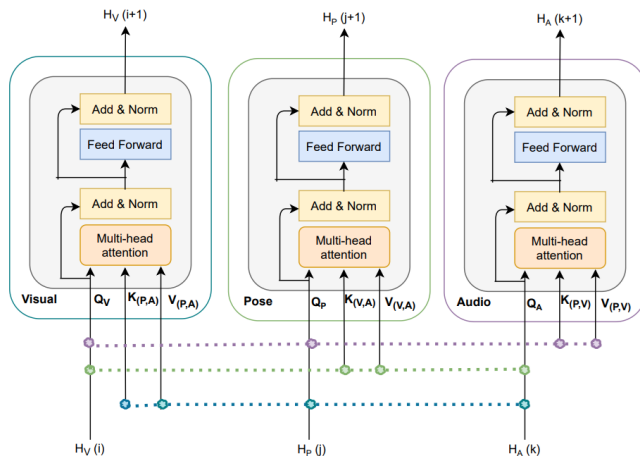
Cross-attention between **two modalities** (text and image) is used in **Deep Fusion**, particularly in cross-attention layers inside LLM

(again: OpenFlamingo)

Preliminary 2: Cross-Attention for 3 modalities?

Cross-attention between **three sequences** (modalities)

tri-modal co-attention in TriBERT



[1] **TriBERT**: Full-body Human-centric Audio-visual Representation Learning for Visual Sound Separation. **NeurIPS** 2021. [\[link\]](#)

[2] **TriCAFFNet**: A Tri-Cross-Attention Transformer with a Multi-Feature Fusion Network for Facial Expression Recognition. 2021. [\[link\]](#)

High-level Classification

Multimodal models can be classified in 2 main types (4 subtypes)
based on the **fusion of input modalities**



3.1

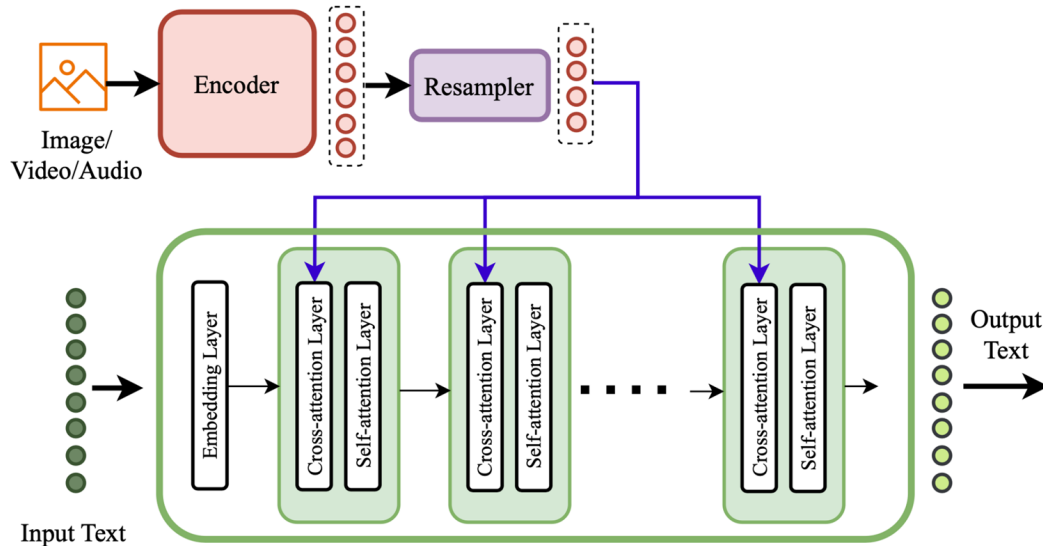
Architectures

Deep Fusion: Standard
Cross-Attention, **SC-DF**



SC-DF: Standard Cross-Attention

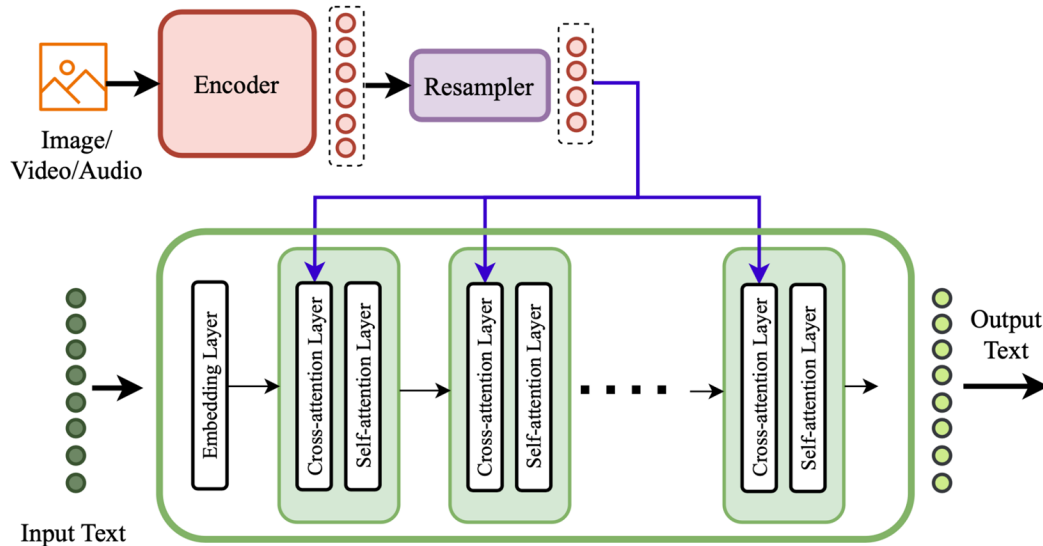
Input modalities are deeply fused into the **internal layers of the LLM** using **standard cross-attention** layer



SC-DF: Standard Cross-Attention

Input modalities are deeply fused into the **internal layers of the LLM**

using **standard cross-attention** layer



before

Flamingo
OpenFlamingo
Otter
Multimodal-GPT

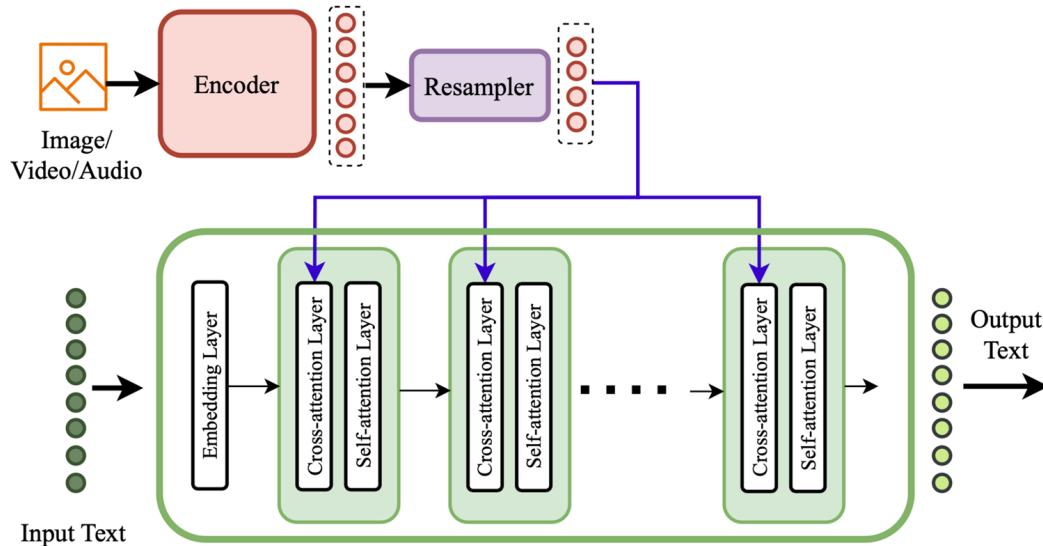
after

VL-BART
VL-T5

SC-DF: Standard Cross-Attention

Input modalities are deeply fused into the **internal layers of the LLM**

using **standard cross-attention** layer



before

Flamingo

OpenFlamingo

Otter

Multimodal-GPT



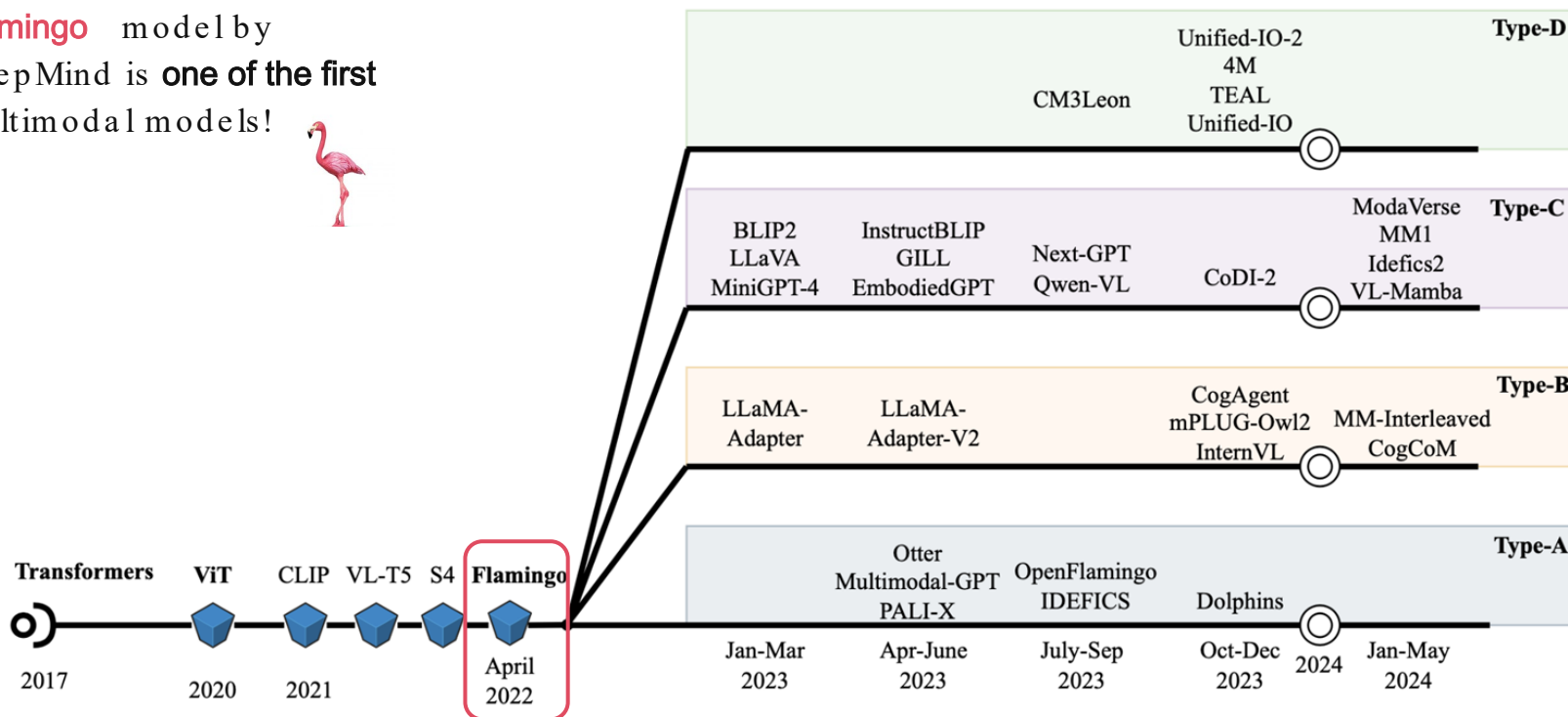
after

VL-BART

VL-T5

SC-DF: Standard Cross-Attention

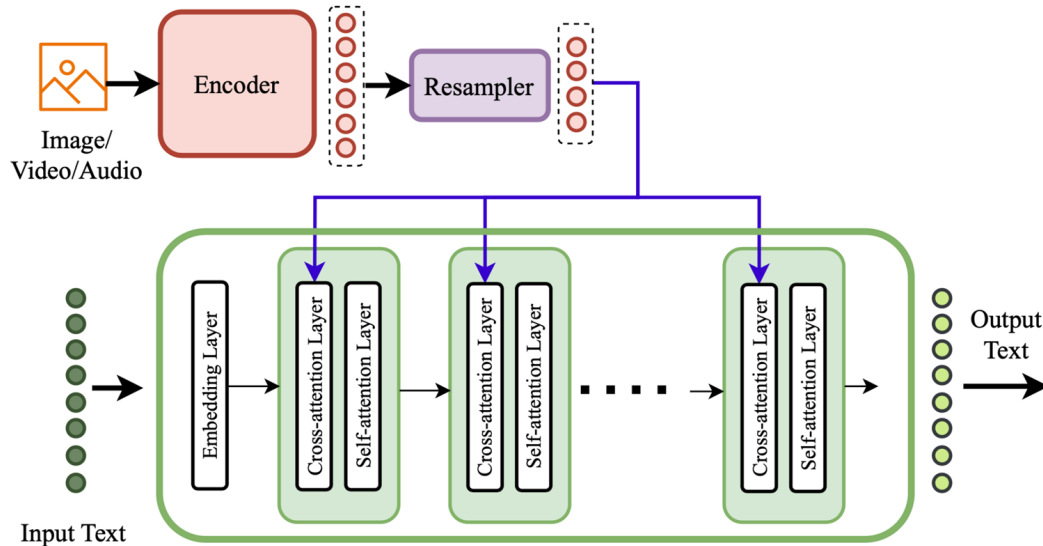
Flamingo model by
DeepMind is **one of the first**
multimodal models!



SC-DF: Standard Cross-Attention

Input modalities are deeply fused into the **internal layers of the LLM**

using **standard cross-attention** layer



before

Flamingo

OpenFlamingo

Otter

Multimodal-GPT

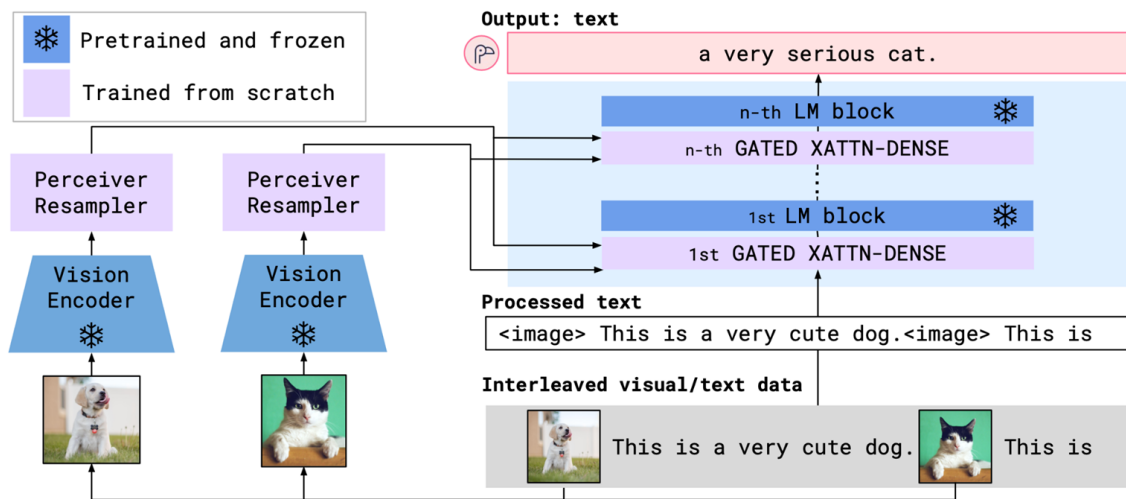


after

VL-BART

VL-T5

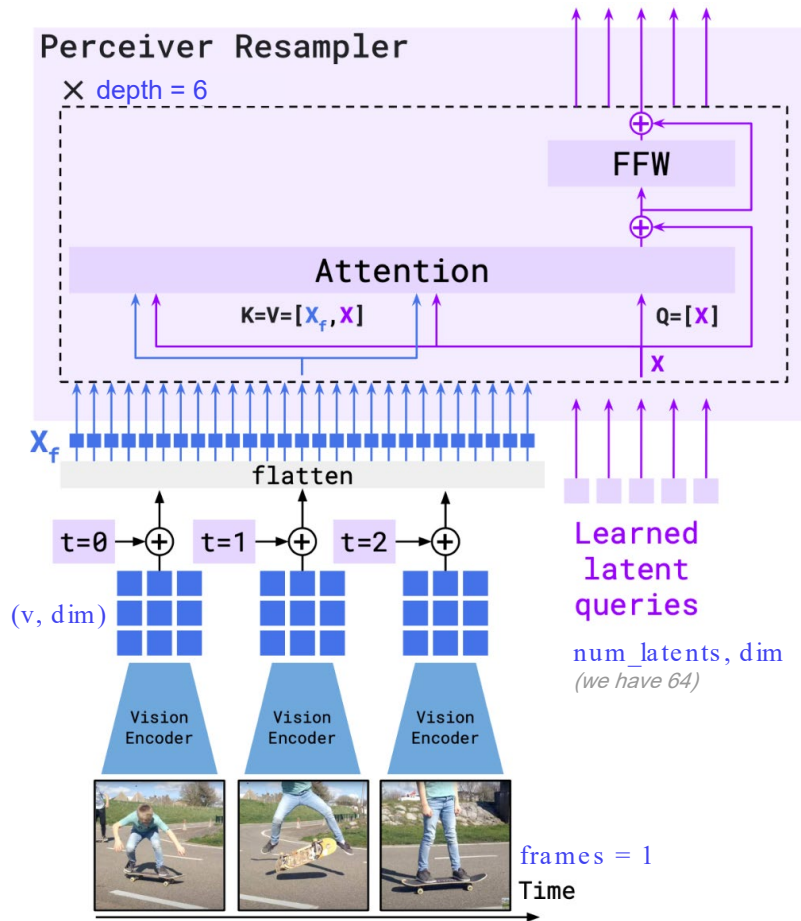
SC-DF: OpenFlamingo (Nov 2022)



vision model
CLIP ViT-L/ 14
(NFNet)

language model
RedPajama / MPT
(Chinchilla)

OpenFlamingo follows Flamingo architecture

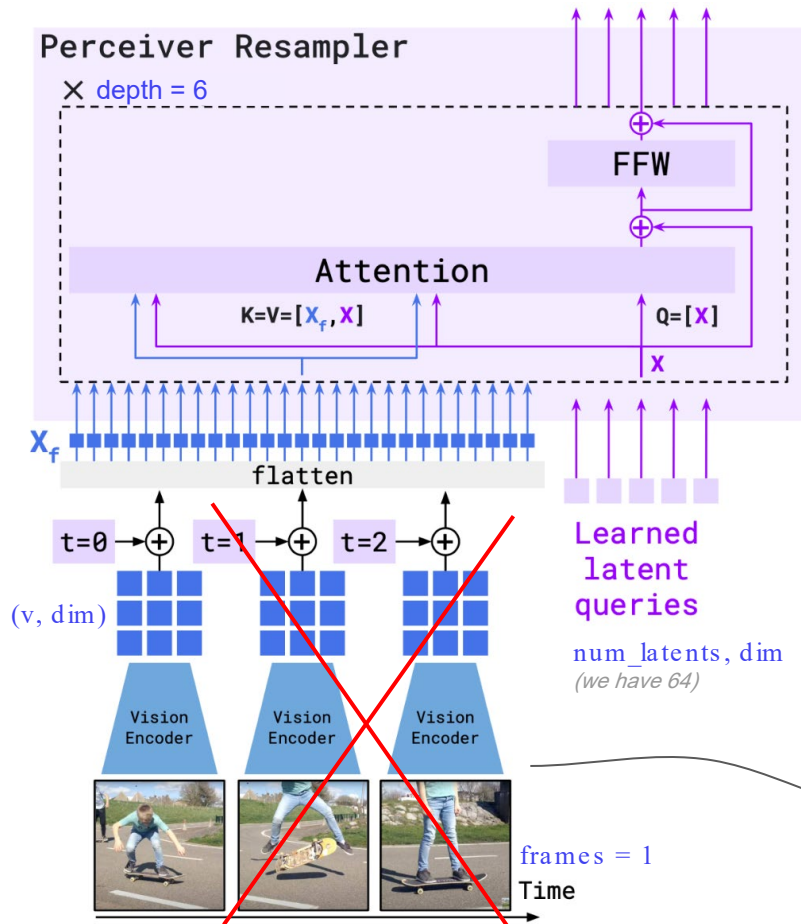


OpenFlamingo: Perceiver

$b, t, f, v, d \rightarrow b, t, (f * v), d \rightarrow b, t, l, d$

$d = 1024$

$l = 64$

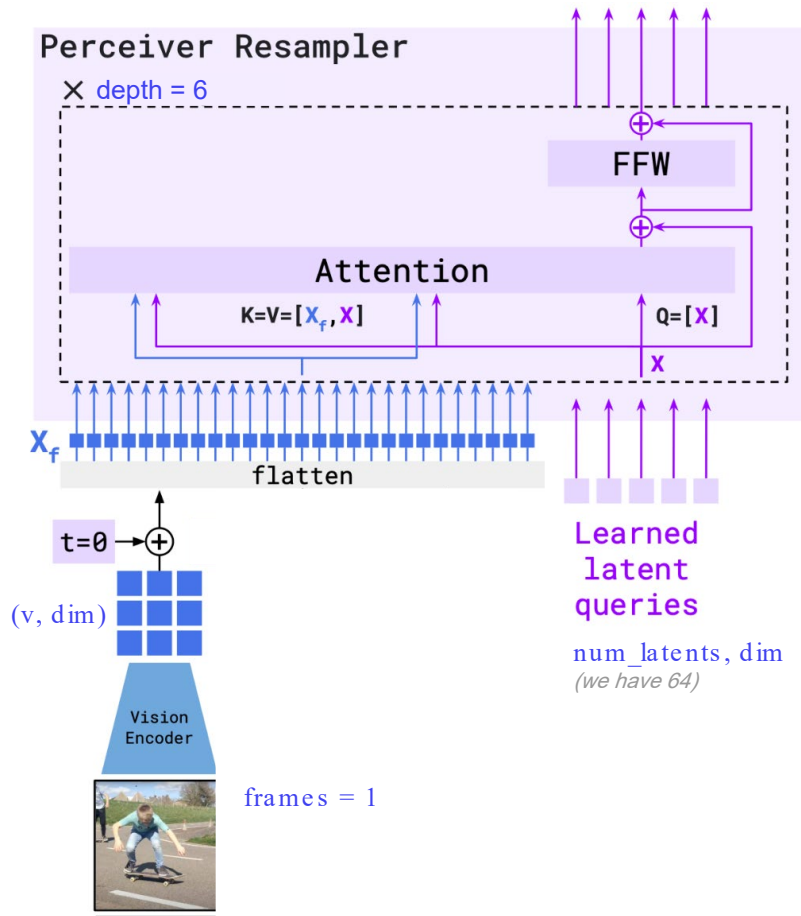


OpenFlamingo: Perceiver

$$b, t, f, v, d \rightarrow b, t, (f * v), d \rightarrow b, t, l, d$$

$d = 1024$ $l = 64$

Do not have video modality
in OpenFlamingo



OpenFlamingo: Perceiver

$$b, t, f, v, d \rightarrow b, t, (f * v), d \rightarrow b, t, l, d$$

$d = 1024$
 $l = 64$

batch size

b

image examples

t

video frames

f

visual tokens

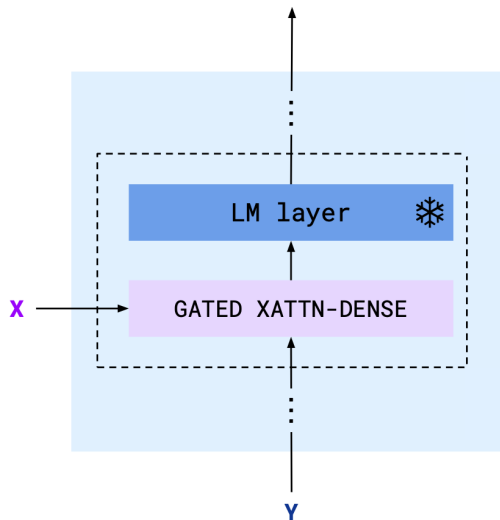
v

embed dim

d

OpenFlamingo: Feature Fusion

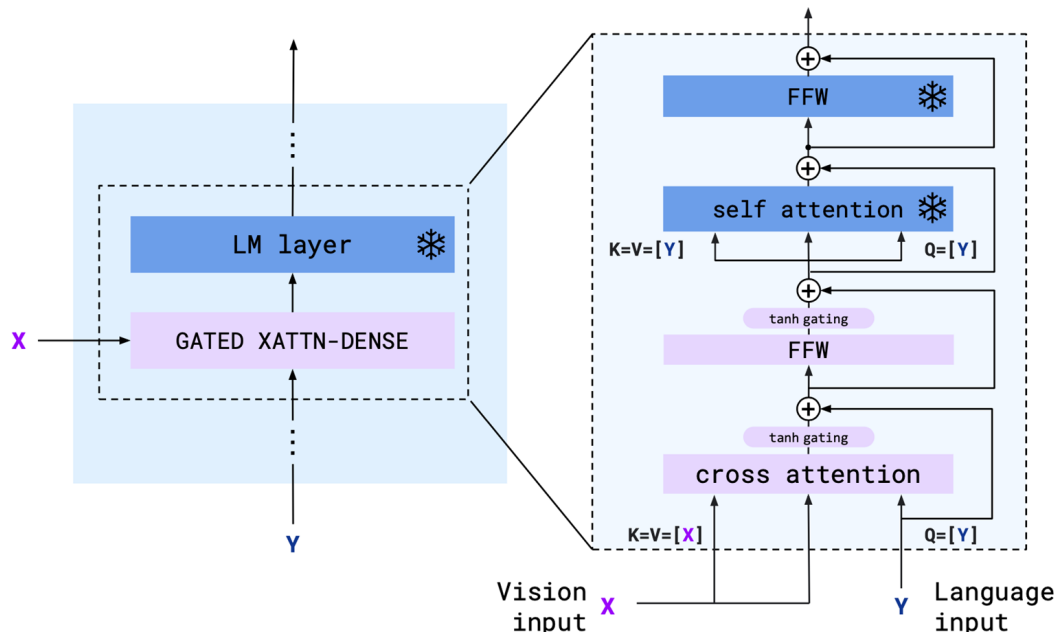
1. **freeze** the pretrained LM blocks
2. insert **gated cross -attention dense** blocks between the original layers
3. keep layers gated to keep LM intact at initialization
4. **queries** = LM inputs



OpenFlamingo: Feature Fusion

1. **freeze** the pretrained LM blocks
2. insert **gated cross -attention dense** blocks between the original layers
3. keep layers gated to keep LM intact at initialization
4. **queries** = LM inputs

tanh -gating mechanism —
 multiplies output of newly
 initialized layer by $\tanh(\alpha)$



OpenFlamingo: Inference



Giving **OpenFlamingo** tricky few-shot examples
training dataset = LAION-2B with image-text pairs

An image of $2 \times 2 = 4$

$$2 \times 2 = 4$$

An image of $3 + 3 = 6$

$$3 + 3 = 6$$

An image of

$$1 + 1 = 2$$

An image of $1 + 1 = 2$

$$1 + 1 = 2$$

input images and prompts

output

OpenFlamingo: Inference



Giving **OpenFlamingo** tricky few-shot examples
training dataset = LAION-2B with image-text pairs

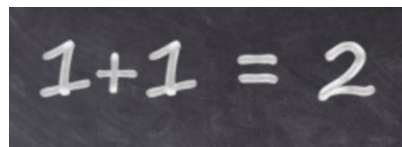
An image of $2 \times 2 = 4$

An image of $3 + 3 = 6$

What is the color of board?

The equation $2 \times 2 = 4$ is displayed in large, bold, red font.

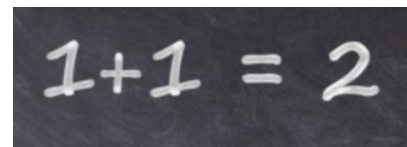
The equation $3 + 3 = 6$ is displayed in large, bold, red font.



input images and prompts

An image of $2 + 2 = 4$.

An image of $3 + 3 = 6$



output

OpenFlamingo: Inference



Giving **OpenFlamingo** tricky few-shot examples
training dataset = LAION-2B with image-text pairs

Print equation: $2 \times 2 = 4$

Print equation: $3 + 3 = 6$

Print equation:

A red, stylized equation $2 \times 2 = 4$ on a white background.

A red, stylized equation $3 + 3 = 6$ on a white background.

A white equation $1 + 1 = 2$ written on a dark chalkboard.

input images and prompts

Print equation $2 + 2 = 4$

Print equation $3 + 3 = 6$

Print equation $2 + 2 = 4$

Print

A white equation $1 + 1 = 2$ written on a dark chalkboard.

output

OpenFlamingo: Inference

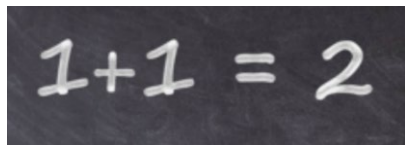
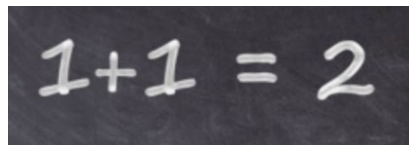


Giving **OpenFlamingo** tricky few-shot examples
training dataset = LAION-2B with image-text pairs

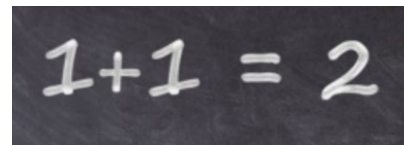


An image of

An image of a
blackboard with a plus
and minus sign on it.



What is on the image?
<|endofchunk|>



input

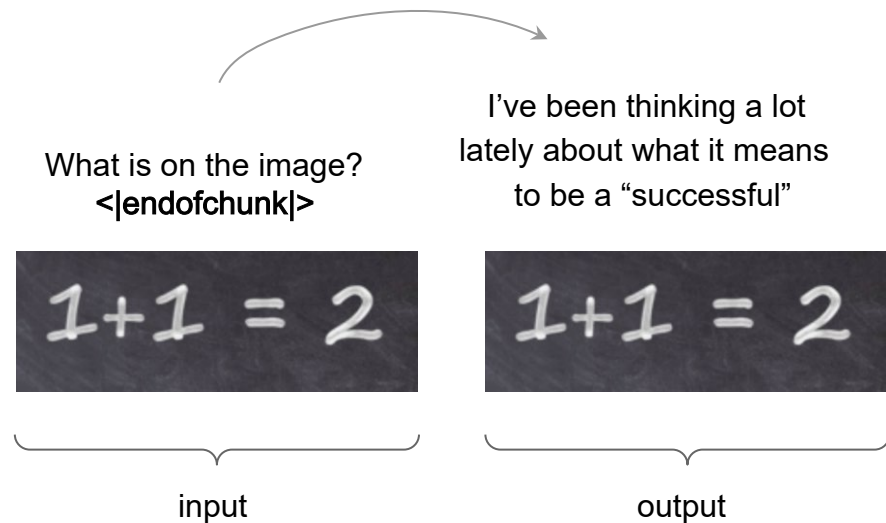
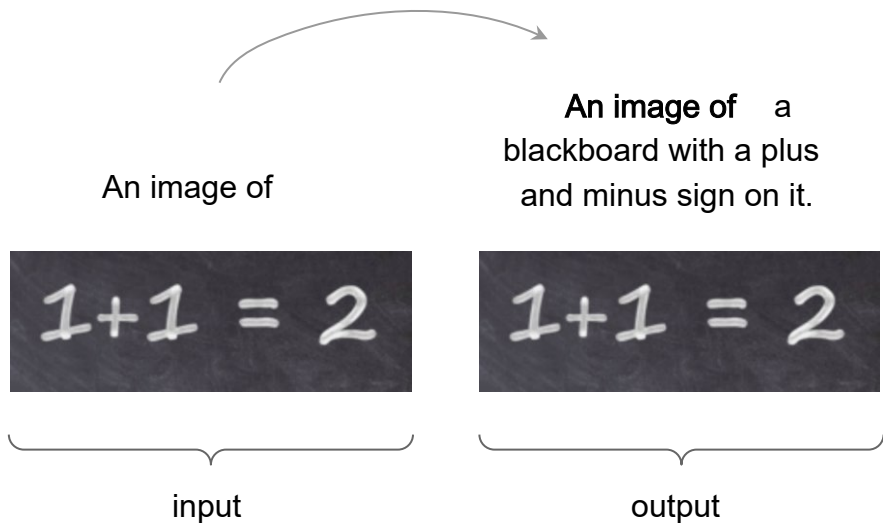
output

input

OpenFlamingo: Inference



Giving **OpenFlamingo** tricky few-shot examples
training dataset = LAION-2B with image-text pairs



3.1

Architectures

Deep Fusion: Custom
Layers, **CL-DF**



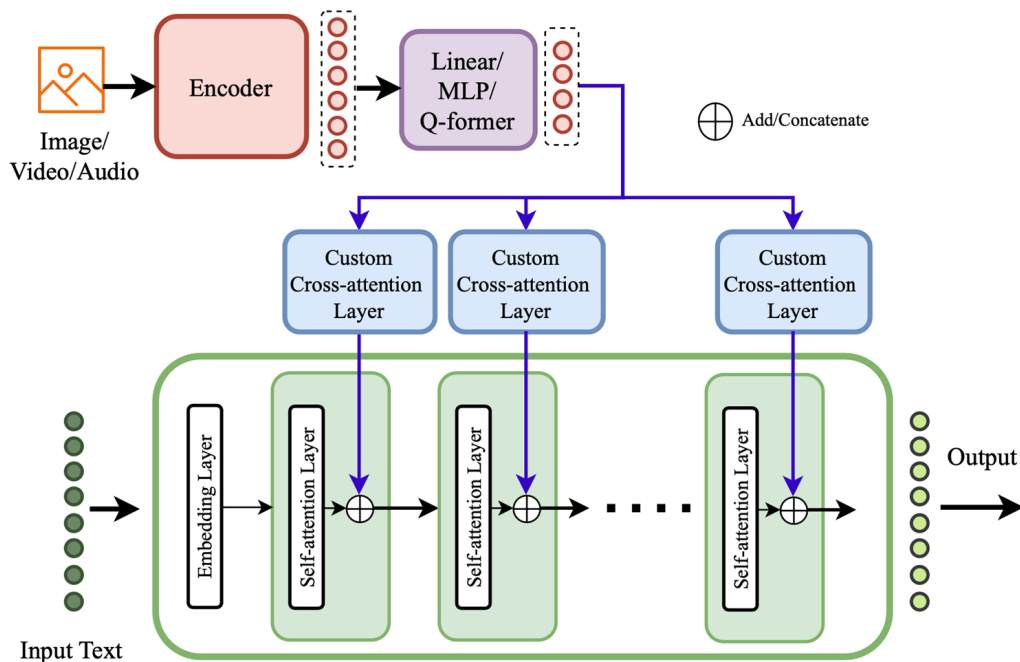
High-level Classification

Multimodal models can be classified in 2 main types (4 subtypes)
based on the **fusion of input modalities**



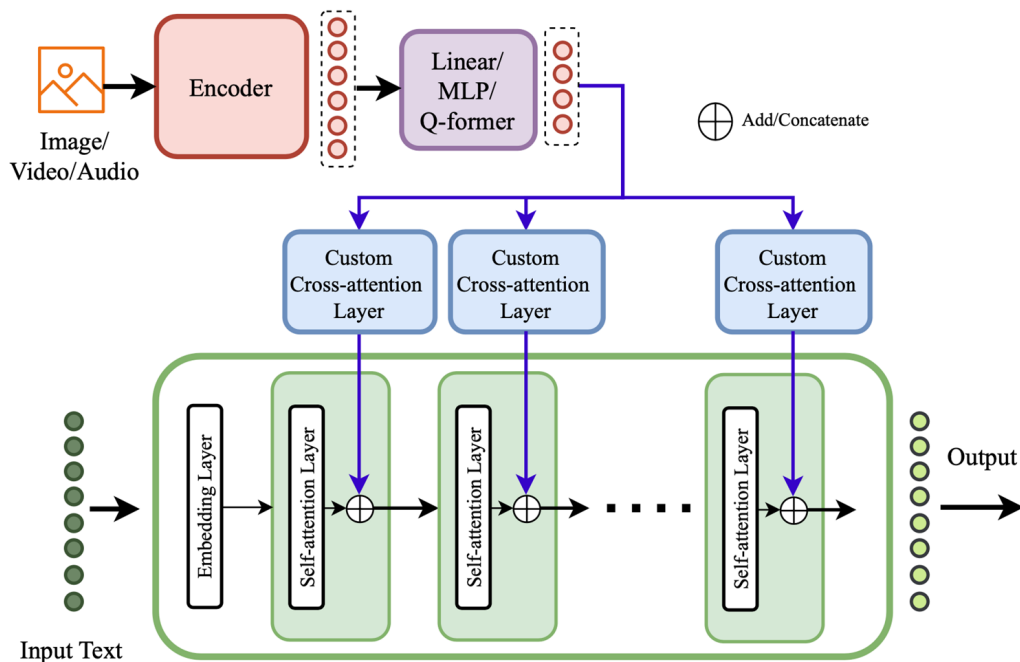
CL-DF: Custom Layers

Input modalities are deeply fused into the **internal layers of the LLM** using **custom -designed** layers



CL-DF: Custom Layers

Input modalities are deeply fused into the **internal layers of the LLM**



using **custom -designed** layers

custom cross -
attention

LLa MA-Adapter
Cog VLM
InternVL
mPLUG-Owl2

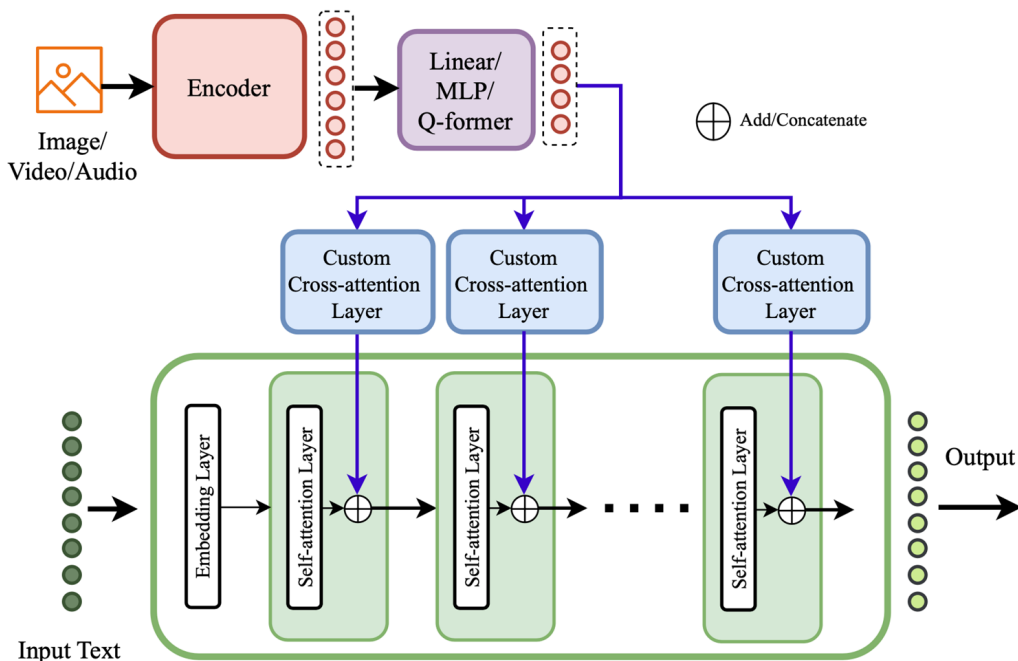
other custom
layers

MoE-LLa VA
LION

CL-DF: Custom Layers

Input modalities are deeply fused into the **internal layers of the LLM**

using **custom -designed** layers



custom cross -
attention

other custom
layers

LLa MA-Adapter
Cog VLM
InternVL
mPLUG-Owl2

MoE-LLaVA
LION

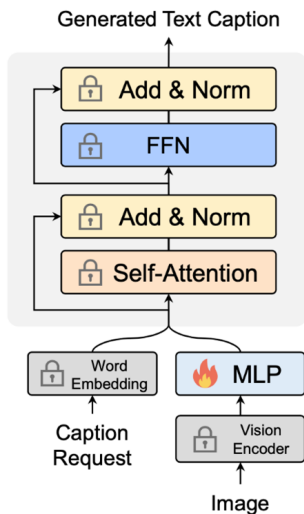


CL-DF: MoE-LLaVA (Dec 2024)

LLaVA – by Microsoft, **MoE-LLaVA** – Peking University: Mixture-of-Experts layer

Stage 1

adapt visual tokens

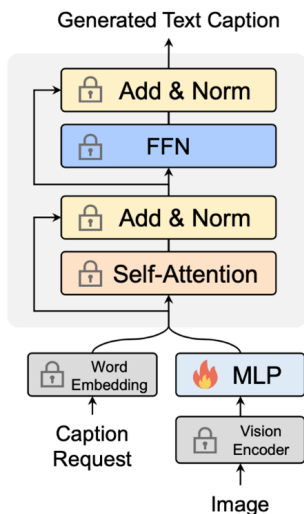


CL-DF: MoE-LLaVA (Dec 2024)

LLaVA – by Microsoft, **MoE-LLaVA** – Peking University: Mixture of Experts layer

Stage 1

adapt visual tokens



vision model

CLIP-Large

(following LLaVA-1.5)

language model

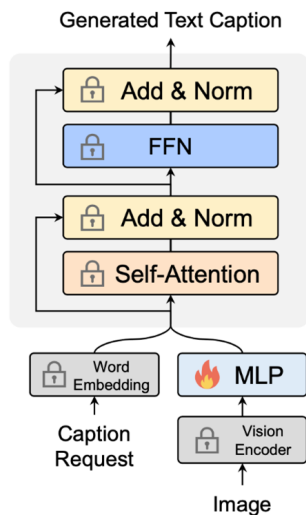
LLaMA / Vicuna / Qwen...

CL-DF: MoE-LLaVA (Dec 2024)

LLaVA – by Microsoft, **MoE-LLaVA** – Peking University: Mixture-of-Experts layer

Stage 1

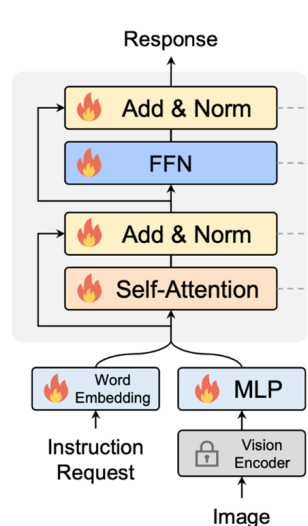
adapt visual tokens



Stage 2

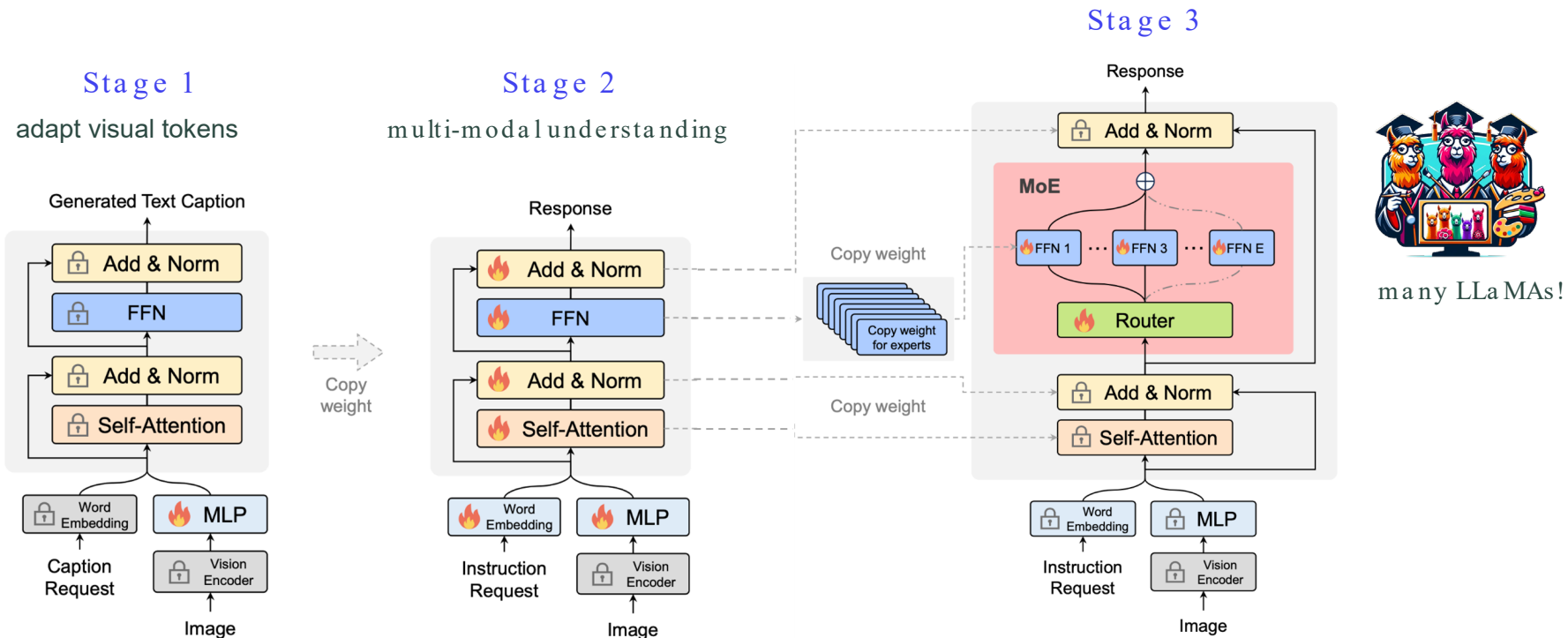
multi-modal understanding

Copy weight

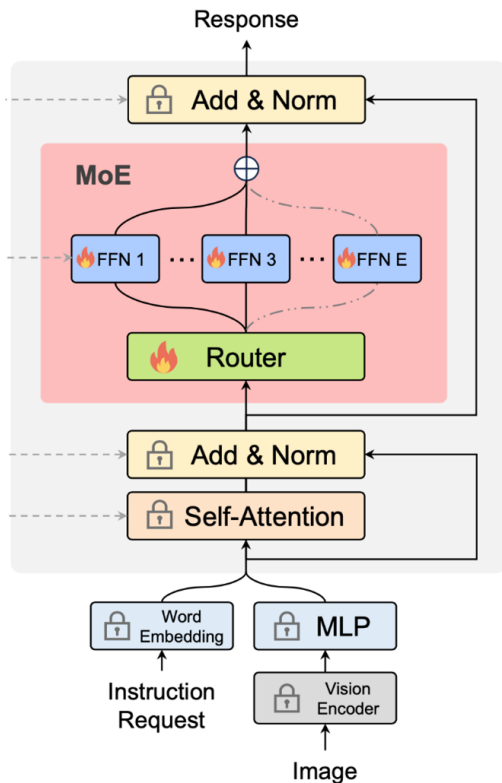


CL-DF: MoE-LLaVA (Dec 2024)

LLaVA – by Microsoft, **MoE-LLaVA** – Peking University: Mixture-of-Experts layer



MoE-LLaVA: Router



1 have E experts, each expert = FFN
 $\mathcal{E} = [e_1, e_2, \dots, e_E]$

2 router = linear layer that assigns probabilities to experts

$$\mathcal{P}(\mathbf{x})_i = \frac{e^{f(\mathbf{x})_i}}{\sum_j^E e^{f(\mathbf{x})_j}}$$

3 calculate weighted sum

$$\text{MoE}(\mathbf{x}) = \sum_{i=1}^k \mathcal{P}(\mathbf{x})_i \cdot \mathcal{E}(\mathbf{x})_i$$

3.2

Architectures

Early Fusion: Non-
Tokenized, **NT-EF**



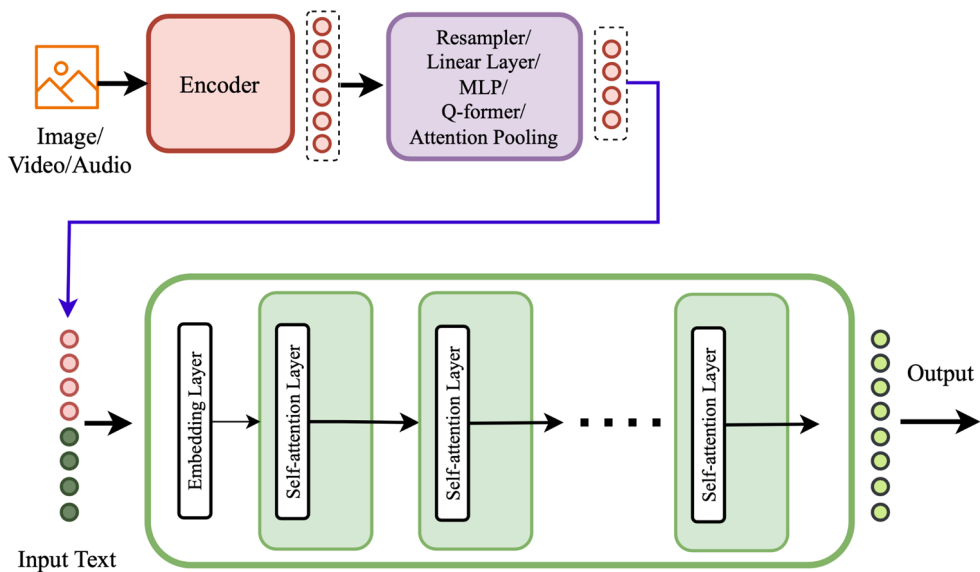
High-level Classification

Multimodal models can be classified in 2 main types (4 subtypes)
based on the **fusion of input modalities**



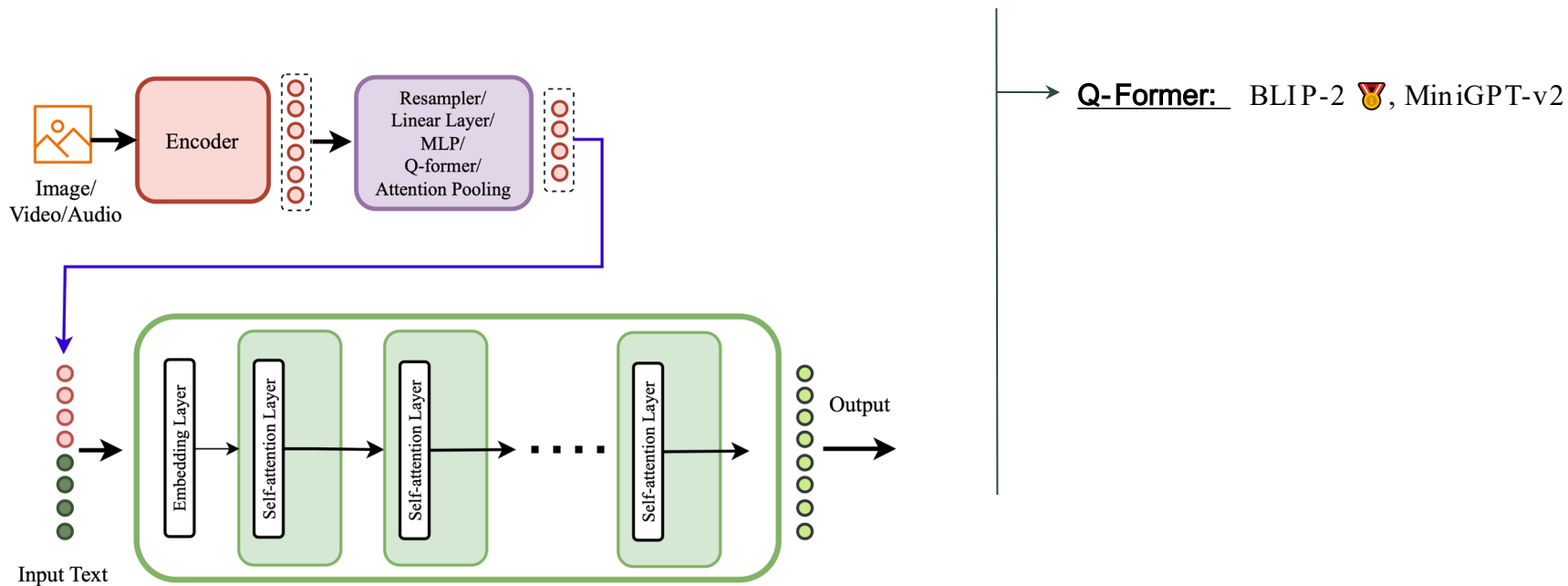
NT-EF: Non-Tokenized

Non-tokenized input modalities are **directly fed to the model** rather than to internal layers



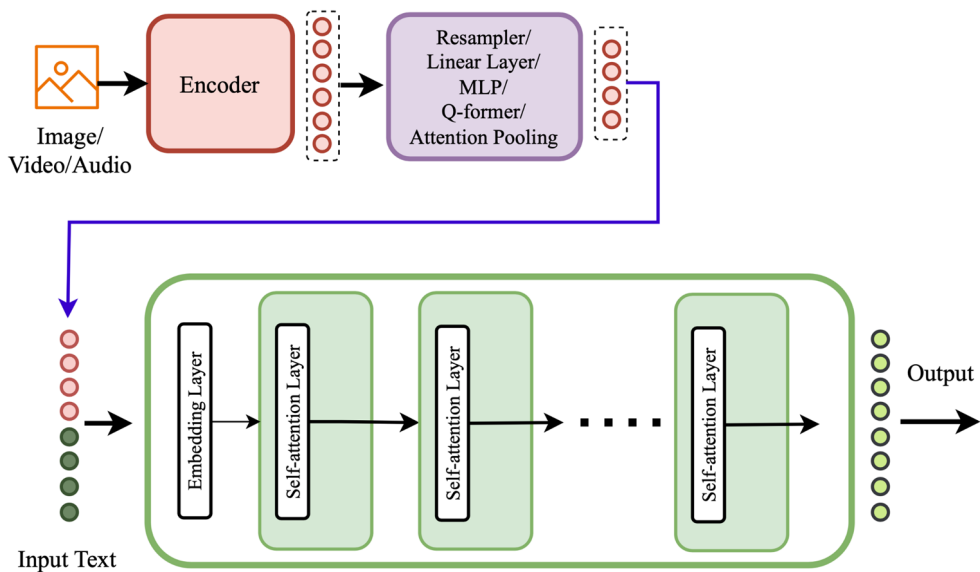
NT-EF: Non-Tokenized

Non-tokenized input modalities are **directly fed to the model** rather than to internal layers



NT-EF: Non-Tokenized

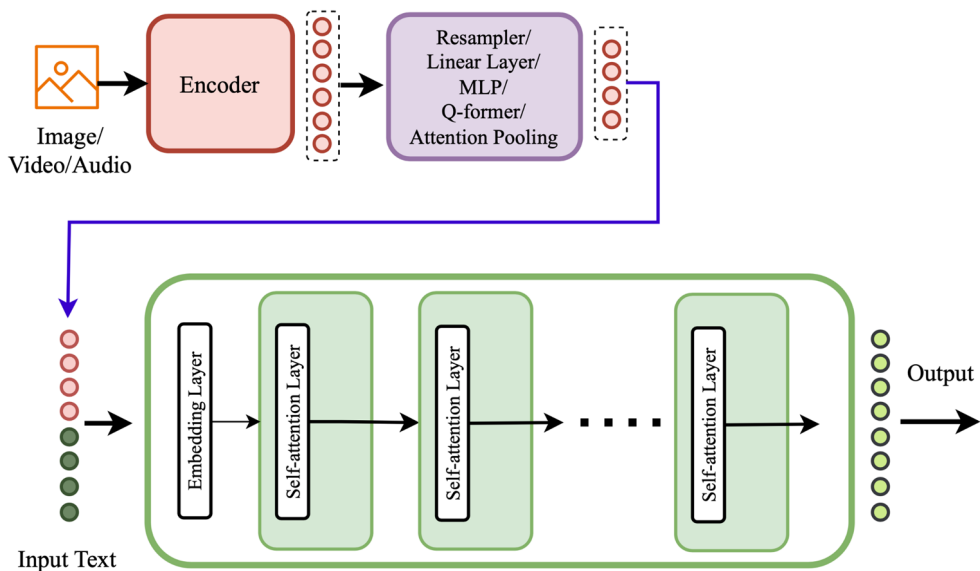
Non-tokenized input modalities are **directly fed to the model** rather than to internal layers



- **Q-Former:** BLIP-2 🏆, MiniGPT-v2
- **Custom layer:** Qwen -VL, AnyMAL, Video-ChatGPT, EmbodiedGPT

NT-EF: Non-Tokenized

Non-tokenized input modalities are **directly fed to the model** rather than to internal layers



- **Q-Former:** BLIP-2 🏆, MiniGPT-v2
- **Custom layer:** Qwen -VL, AnyMAL, Video-ChatGPT, EmbodiedGPT
- **Linear / MLP:** DeepSeek-VL, LLaVA, LLaVA-NeXT, PaLM-E, Shikra
- **Perceiver resampler:** Monkey, V*, Kosmos-G

NT-EF: Qwen-VL (Oct 2023)



Alibaba Group, 9.6B parameters

vision model = OpenClip ViT-bigG, **language model** = Qwen-7B

NT-EF: Qwen-VL (Oct 2023)

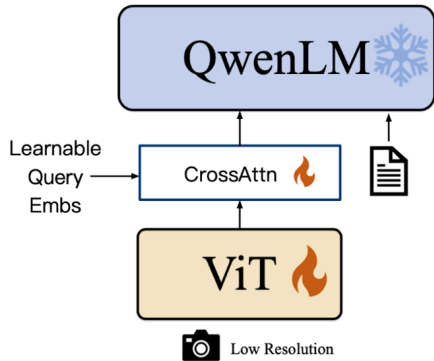


Alibaba Group, 9.6B parameters

vision model = OpenClip ViT-bigG, **language model** = Qwen-7B

Stage 1: Pretraining

5B web data pairs → 1.4B



NT-EF: Qwen-VL (Oct 2023)

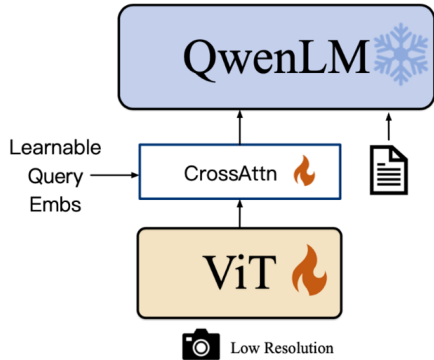


Alibaba Group, 9.6B parameters

vision model = OpenClip ViT-bigG, **language model** = Qwen-7B

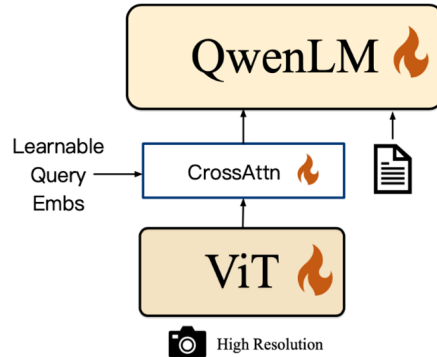
Stage 1: Pretraining

5B web data pairs → 1.4B



Stage 2: Multi-task pretraining

high quality, ~80M data



NT-EF: Qwen-VL (Oct 2023)

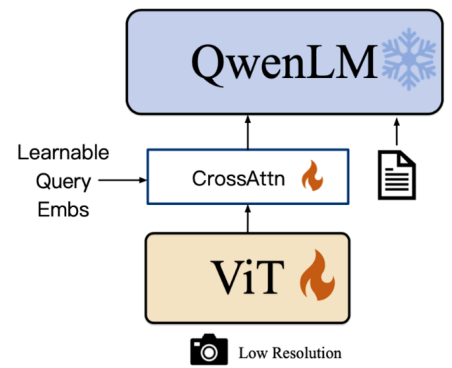


Alibaba Group, 9.6B parameters
vision model = OpenClip ViT-bigG, language model = Qwen-7B

Qwen-VL-Chat

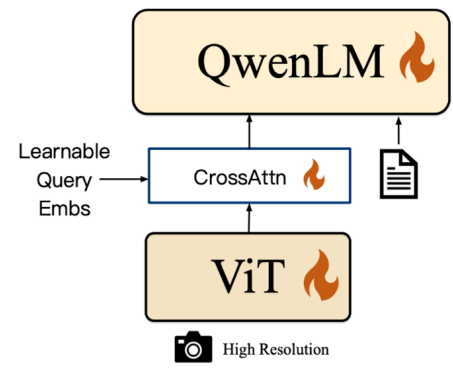
Stage 1: Pretraining

5B web data pairs → 1.4B



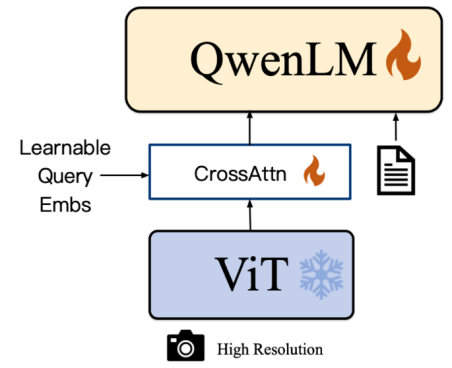
Stage 2: Multi-task pretraining

high quality, ~80M data



Stage 3: Supervised Fine-tuning

instructions, 350k data



Early Fusion

3.2

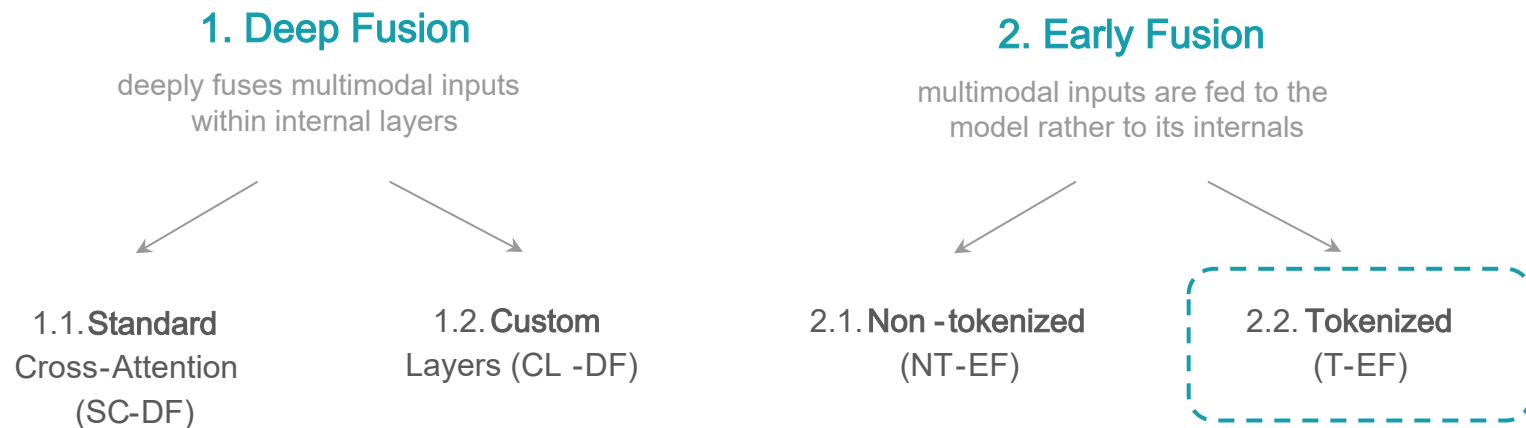
Architectures

Early Fusion: Tokenized, T-
EF



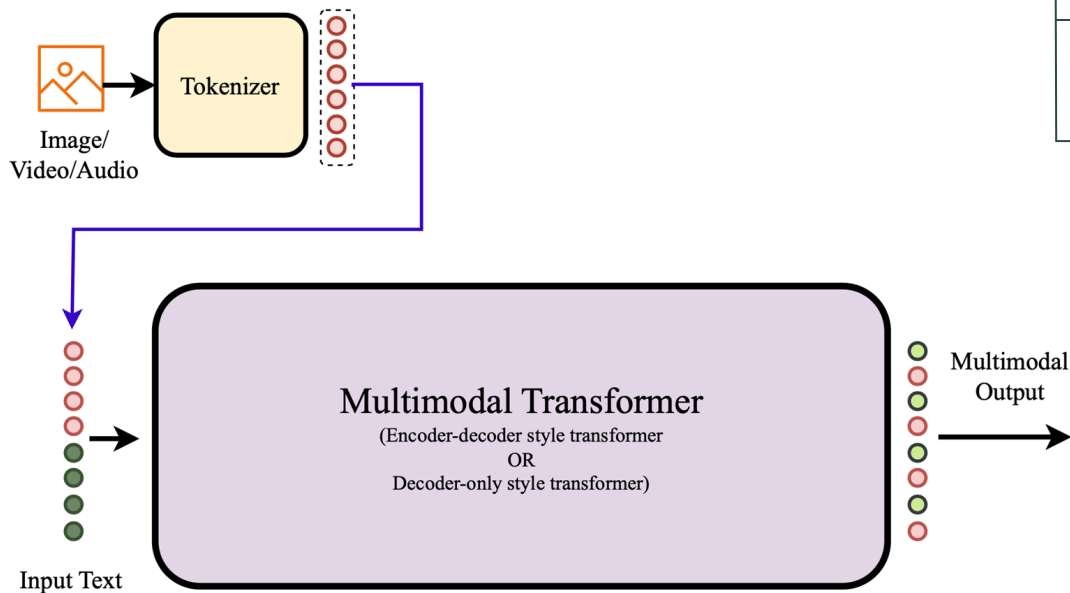
High-level Classification

Multimodal models can be classified in 2 main types (4 subtypes) based on the **fusion of input modalities**



T-EF: Tokenized

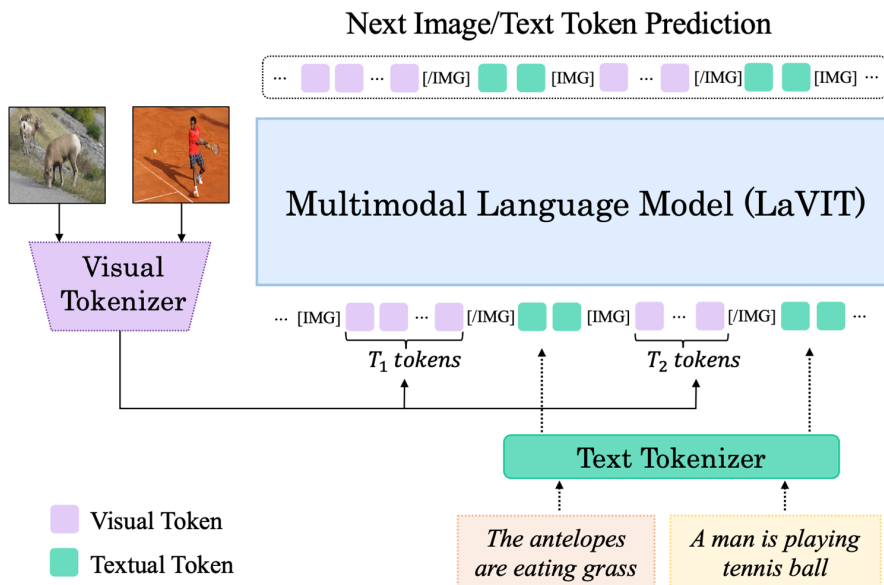
Inputs are tokenized **using a common tokenizer** or modality specific tokenizers



- encoder -decoder: Unified-IO, 4M
- decoder -only: [LaViT](#), TEAL, CM3Leon, VL-GPT

T-EF: LaViT (Mar 2024, ICLR)

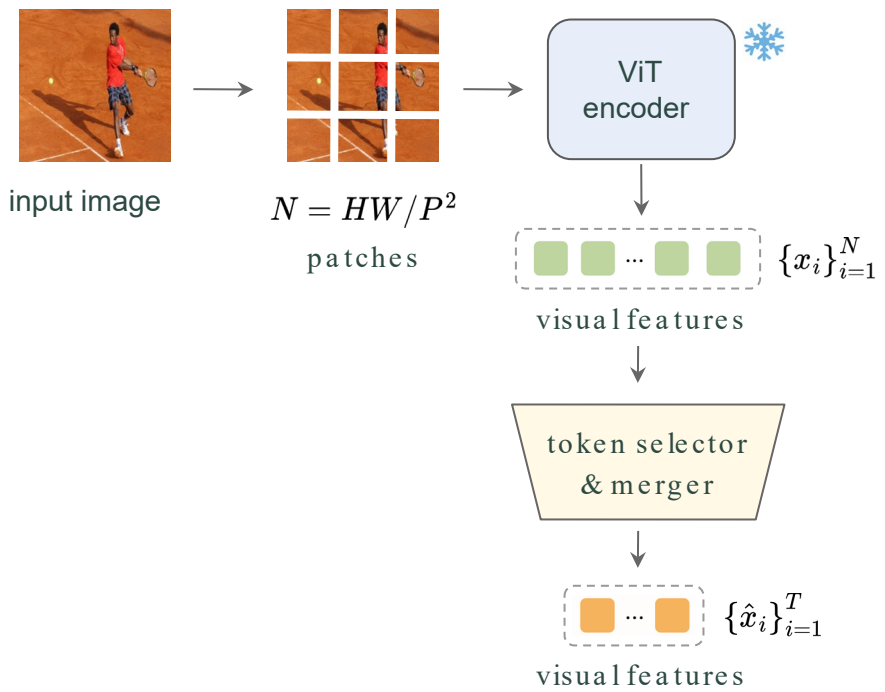
LaViT — Language-Vision Transformer by researchers from Peking & Kuaishou University



1. represent two modalities in a uniform form to exploit LLM's next -token prediction
2. visual tokenizer returns sequence of discrete visual tokens possessing word-like high-level semantics

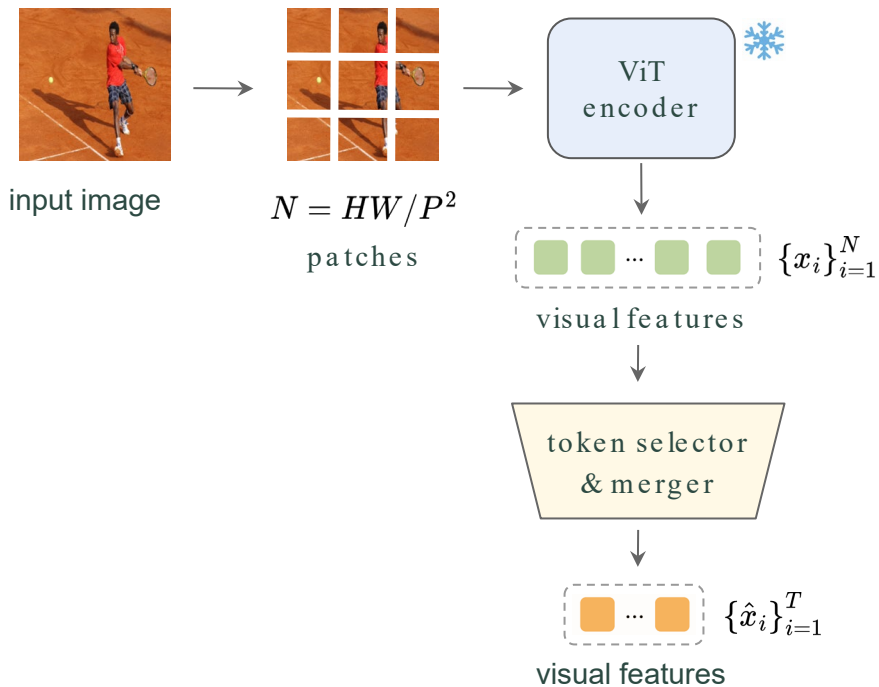
T-EF: La VI T (Mar 2024, ICLR)

vision model = ViT-G/14 of EVA-CLIP, language model = LLaMA-7B

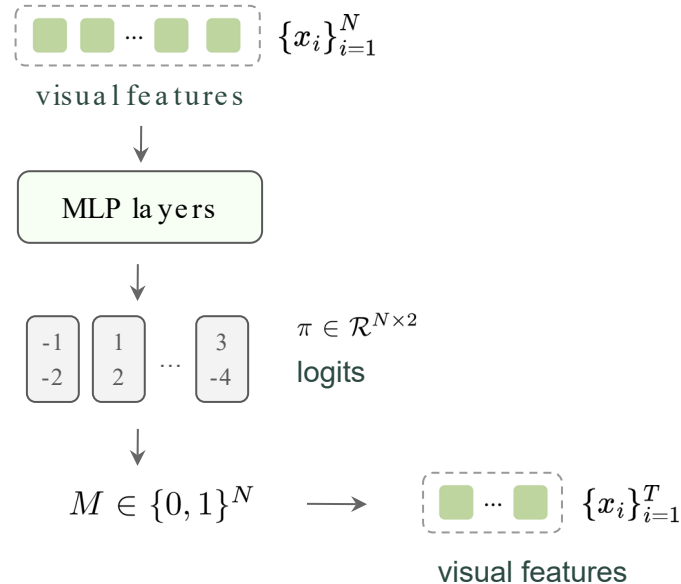


T-EF: La Vi T (Mar 2024, ICLR)

vision model = ViT-G/14 of EVA-CLIP, language model = LLaMA-7B

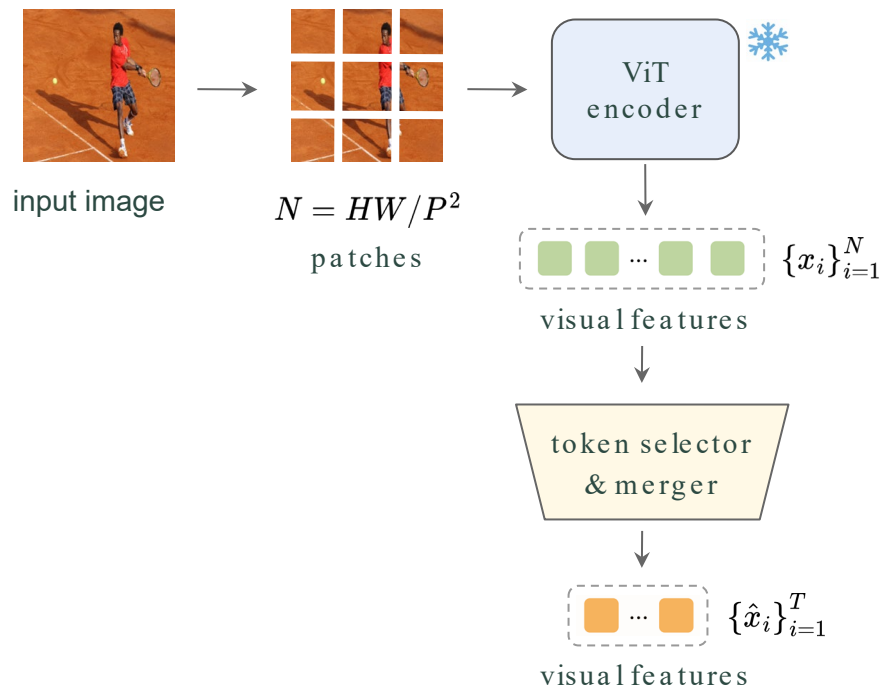


1 token selector

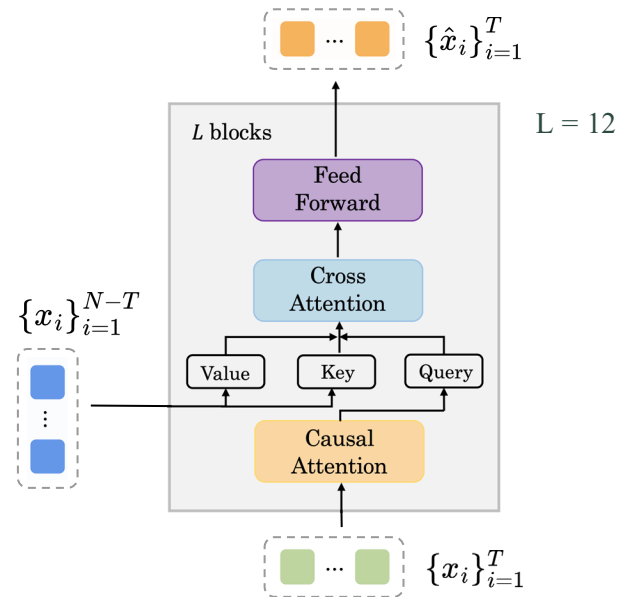


T-EF: La ViT (Mar 2024, ICLR)

vision model = ViT-G/14 of EVA-CLIP, language model = LLaMA-7B

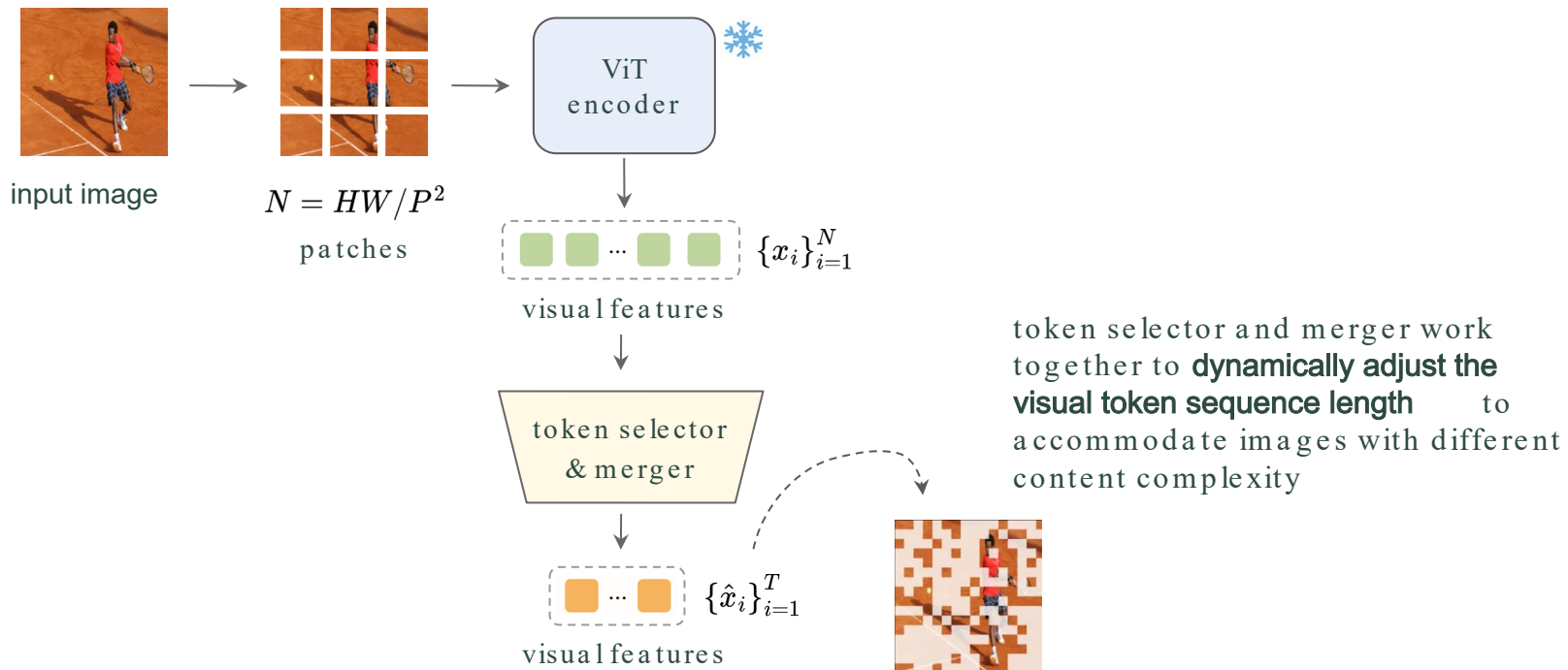


2 token merger



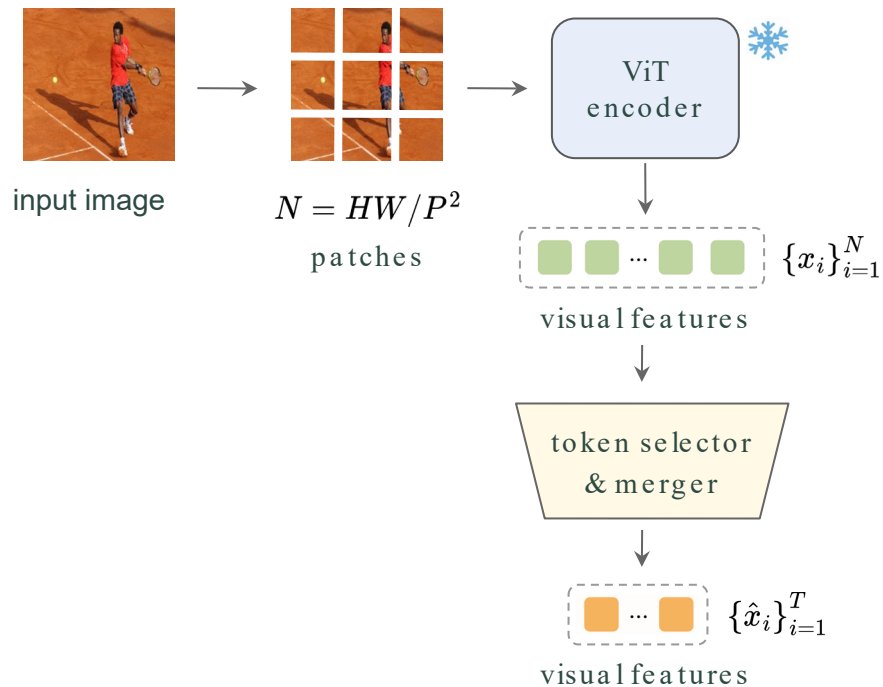
T-EF: La Vi T (Mar 2024, ICLR)

vision model = ViT-G/14 of EVA-CLIP, language model = LLaMA-7B



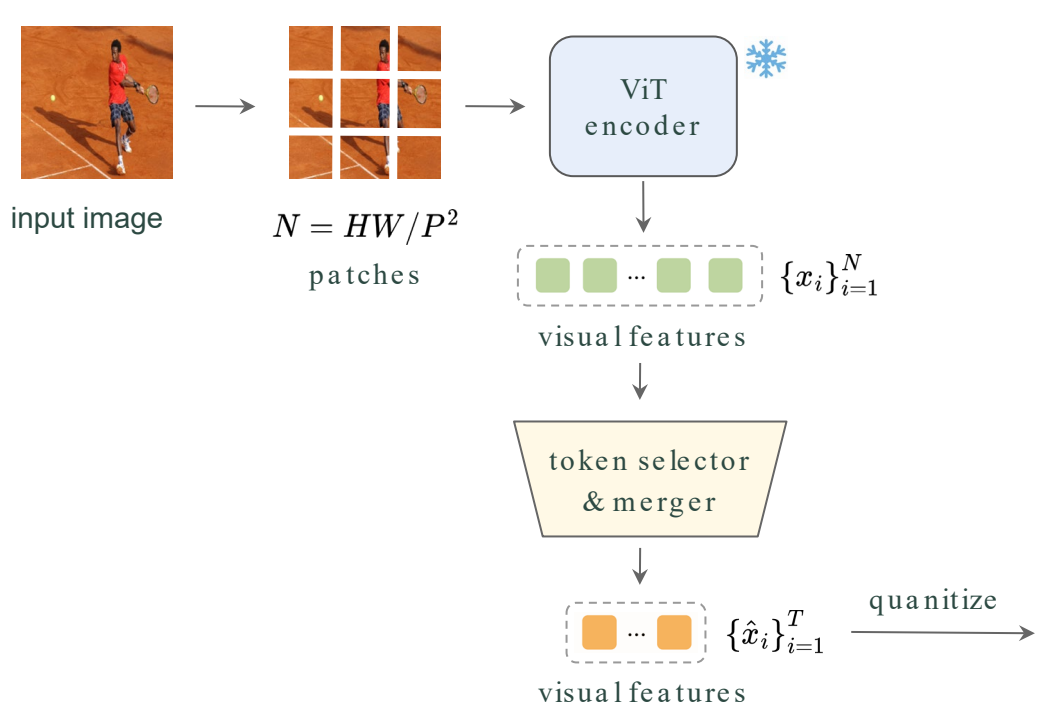
T-EF: La Vi T (Mar 2024, ICLR)

vision model = ViT-G/14 of EVA-CLIP, language model = LLaMA-7B

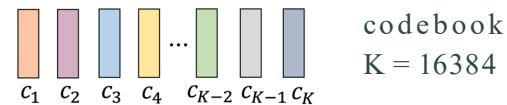


T-EF: La ViT (Mar 2024, ICLR)

vision model = ViT-G/14 of EVA-CLIP, language model = LLaMA-7B



vector quantization

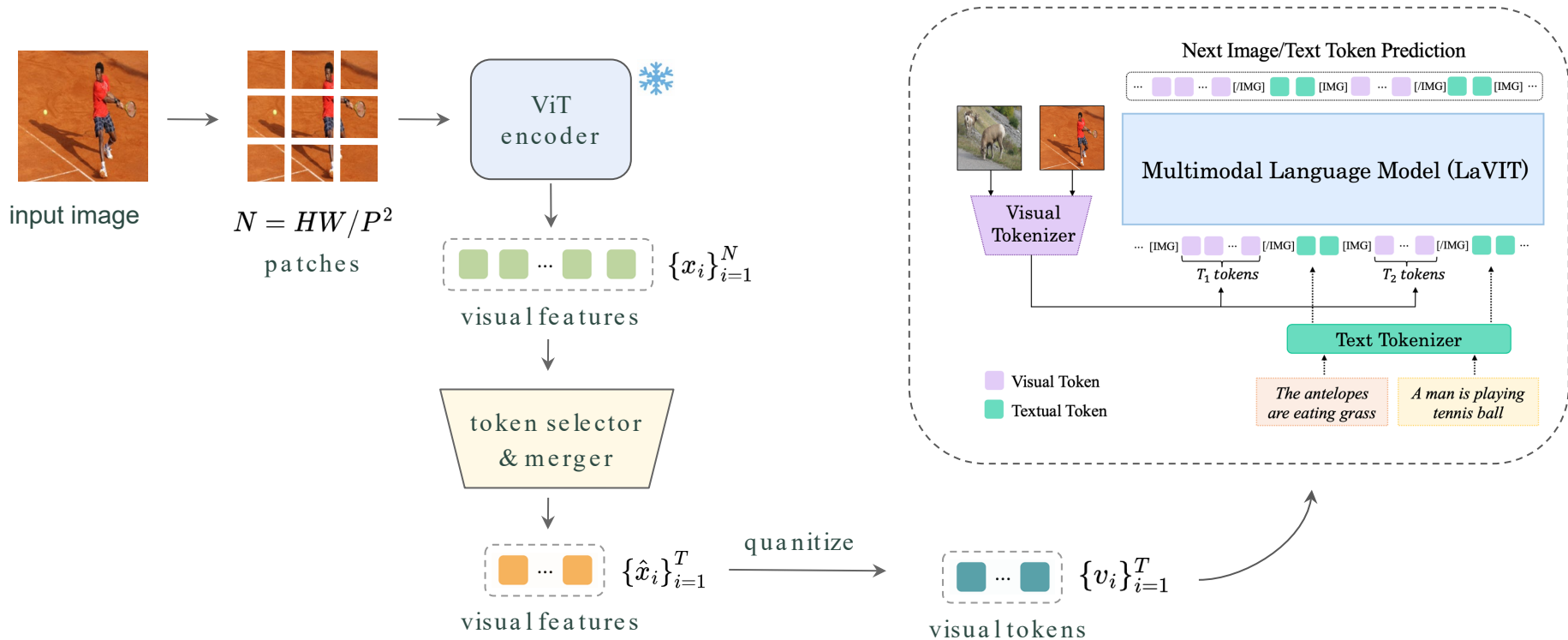


$$v_i = \arg \min_j \|l_2(\hat{x}_i) - l_2(c_j)\|_2$$

$$v_i \in [0, K - 1]$$

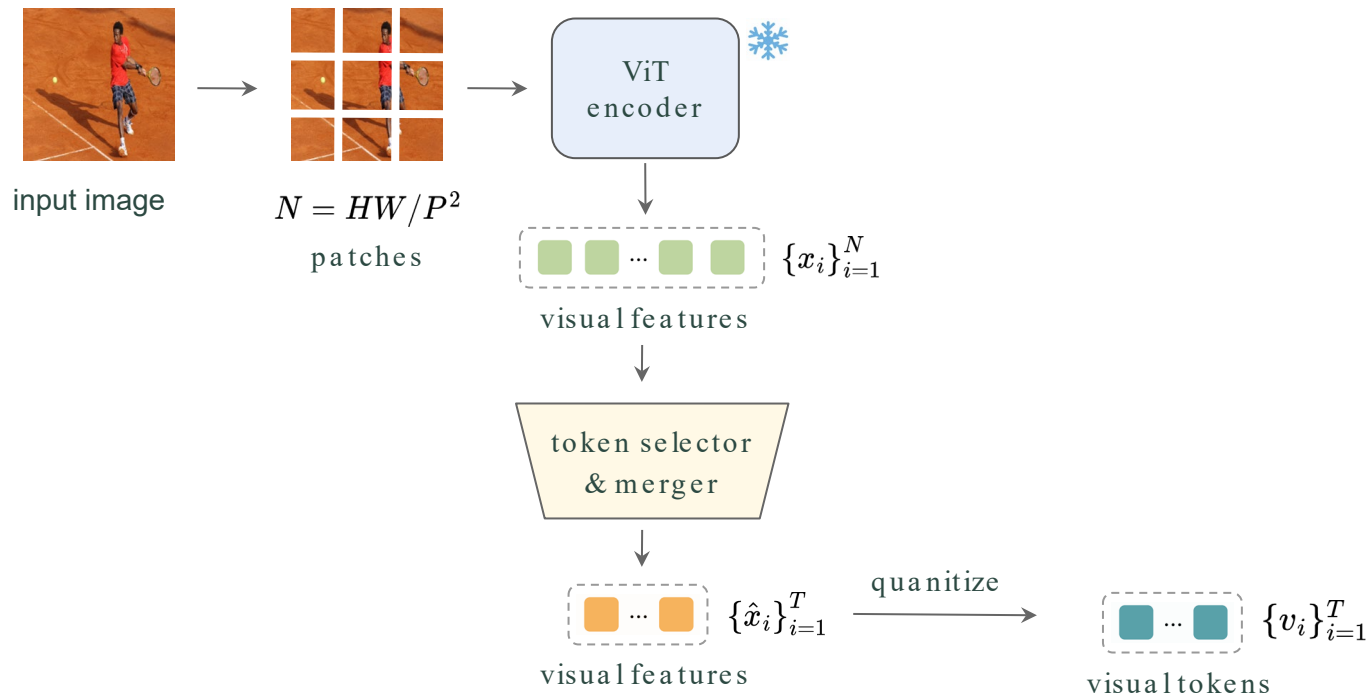
T-EF: La Vi T (Mar 2024, ICLR)

vision model = ViT-G/14 of EVA-CLIP, language model = LLaMA-7B



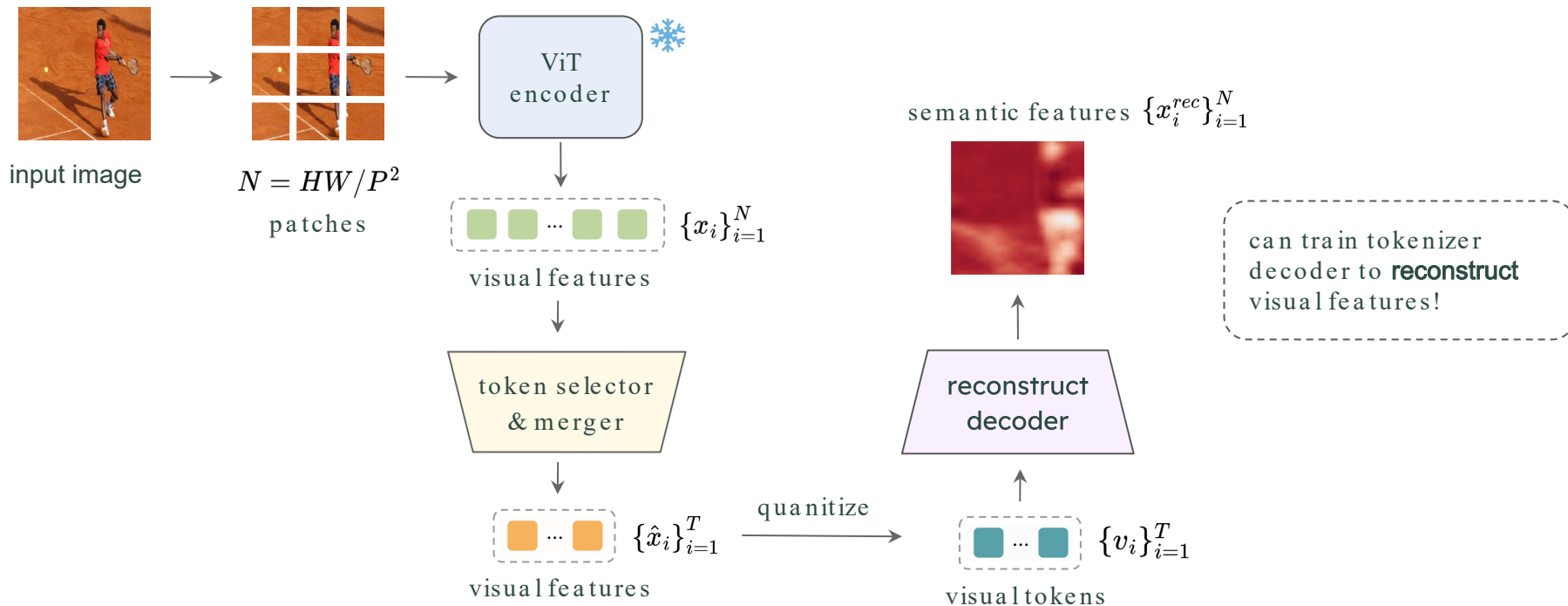
T-EF: La Vi T (Mar 2024, ICLR)

vision model = ViT-G/14 of EVA-CLIP, language model = LLaMA-7B



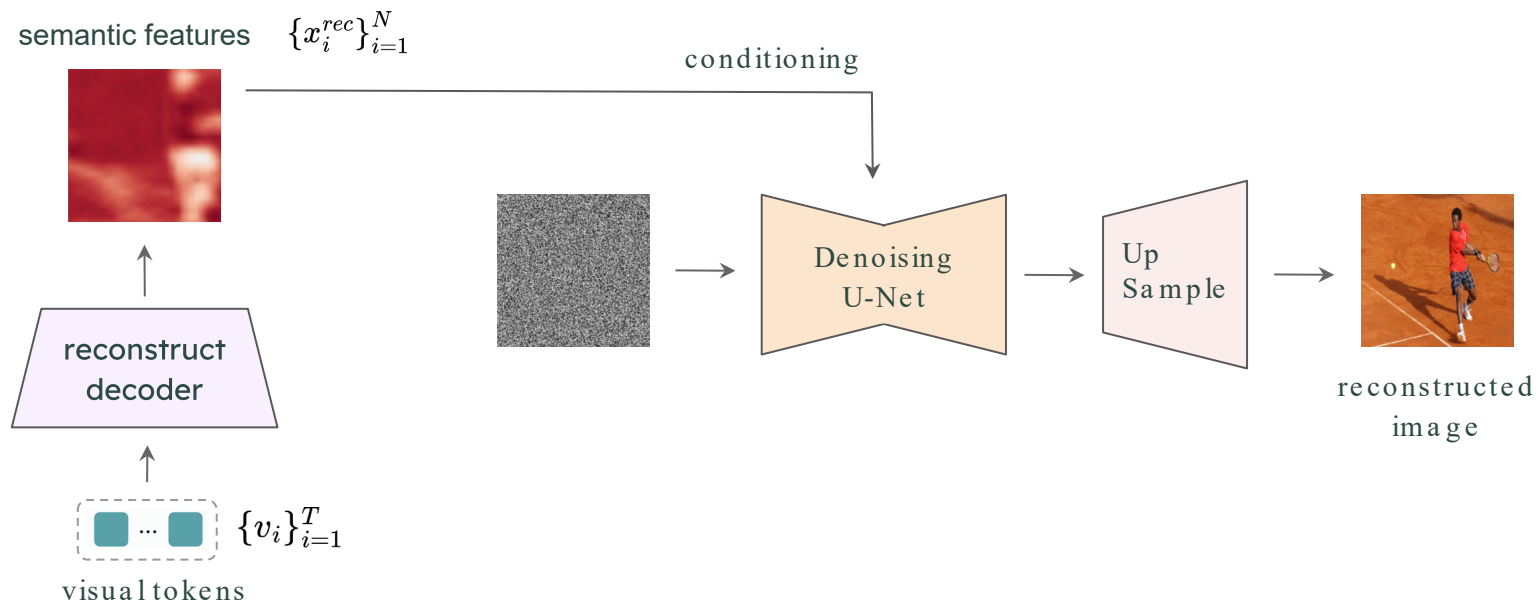
T-EF: La VI T (Mar 2024, ICLR)

vision model = ViT-G/14 of EVA-CLIP, language model = LLaMA-7B



T-EF: La ViT (Mar 2024, ICLR)

During inference, the generated visual tokens from La ViT **can be decoded** into realistic images **by this U-Net!**

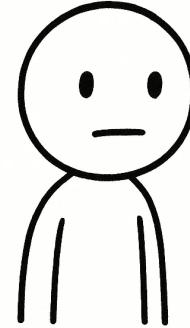


Conclusions

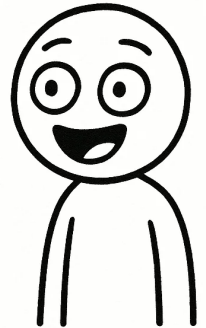
- 1 MLLMs seems to be a **step towards AGI** with lots of interesting applications and challenges
- 2 **Benchmarking MLLMs** may be done via static benchmarks (as in CV or NLP) or using Arenas
- 3 **Classification** of Multimodal Models:
Deep (OpenFlamingo, MoE-LLaVA) &
Early Fusion (Qwen-VL, LaVIT)

this is the end of
the lecture

ME



THEM



tomorrow is the
another lecture

