Школа анализа данных

Seminar 1: Introduction to Multimodal Models

Supplementary slides to Google Colab notebook

Zinkovich Viktoriia

Introduction

Introduction to introduction :)

1 Focus on **VLM models**: Image + Text → Text

Introduction

Introduction to introduction :)

1. Focus on **VLM models**: Image + Text → Text
2. Classify multimodal models at a **high level**

Introduction

Introduction to introduction :)

1. Focus on **VLM models**: Image + Text → Text
2. Classify multimodal models at a **high level**
3. Explore code & architecture of the **most vivid exemplars**

Introduction

Introduction to introduction :)

1. Focus on **VLM models**: Image + Text → Text
2. Classify multimodal models at a **high level**
3. Explore code & architecture of the **most vivid exemplars**
4. Investigate **general approaches** applicable to other models

High-level Classification

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

1. Deep Fusion

deeply fuses multimodal inputs within internal layers

2. Early Fusion

multimodal inputs are fed to the model rather to its internals

Shakti N. Wadekar et al. The Evolution of Multimodal Model Architectures. 28 May 2024.

High-level Classification

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

1. Deep Fusion

deeply fuses multimodal inputs within internal layers

- 1.1. **Standard** Cross-Attention (SC-DF)
- 1.2. **Custom** Layers (CL-DF)

2. Early Fusion

multimodal inputs are fed to the model rather to its internals

Shakti N. Wadekar et al. The Evolution of Multimodal Model Architectures. 28 May 2024.

High-level Classification

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

1. Deep Fusion

deeply fuses multimodal inputs within internal layers

- 1.1. **Standard** Cross-Attention (SC-DF)
- 1.2. **Custom** Layers (CL-DF)

2. Early Fusion

multimodal inputs are fed to the model rather to its internals

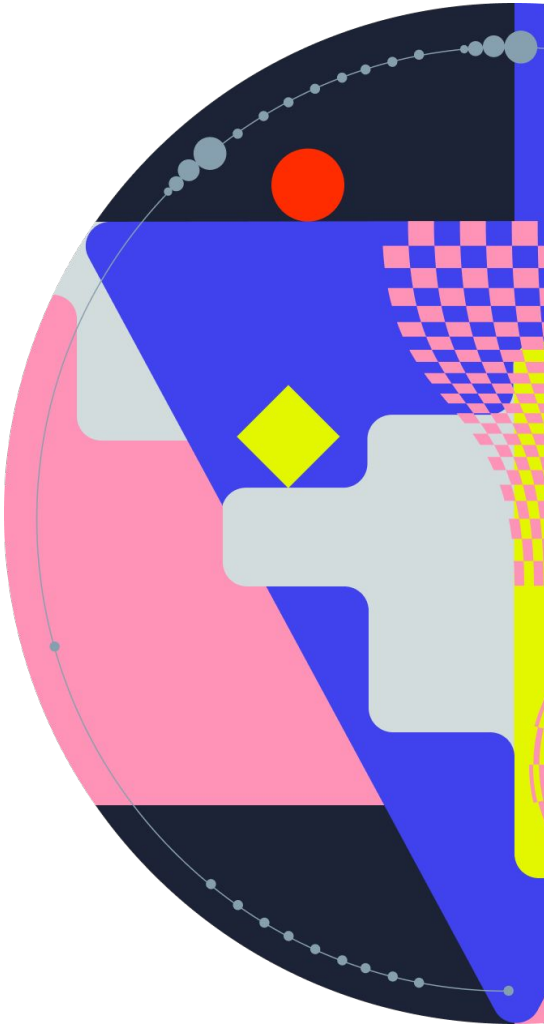
- 2.1. **Non-tokenized** (NT-EF)
- 2.2. **Tokenized** (T-EF)

Shakti N. Wadekar et al. The Evolution of Multimodal Model Architectures. 28 May 2024.

1.1

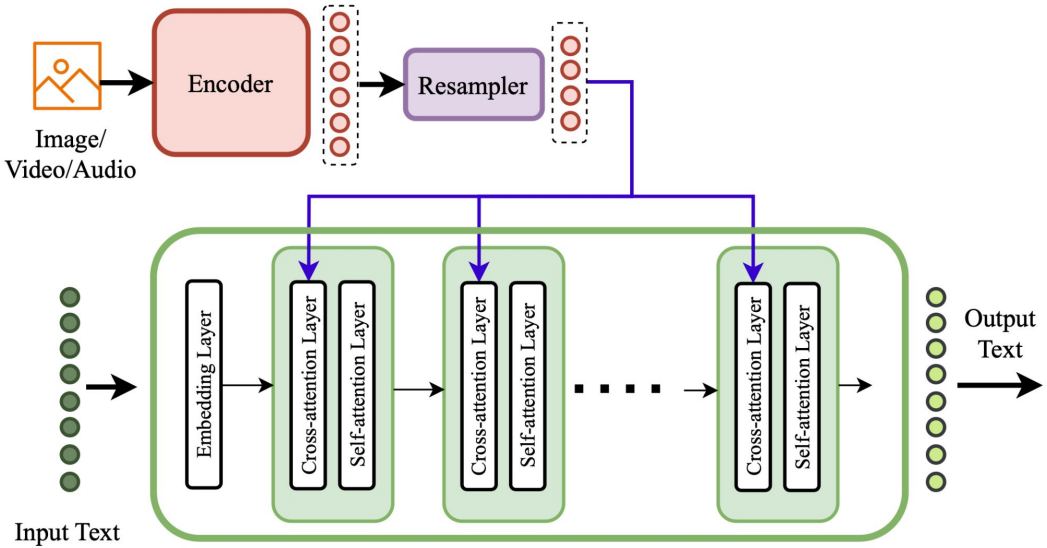
Deep Fusion:

Standard Cross-Attention Deep Fusion (SC-DF)



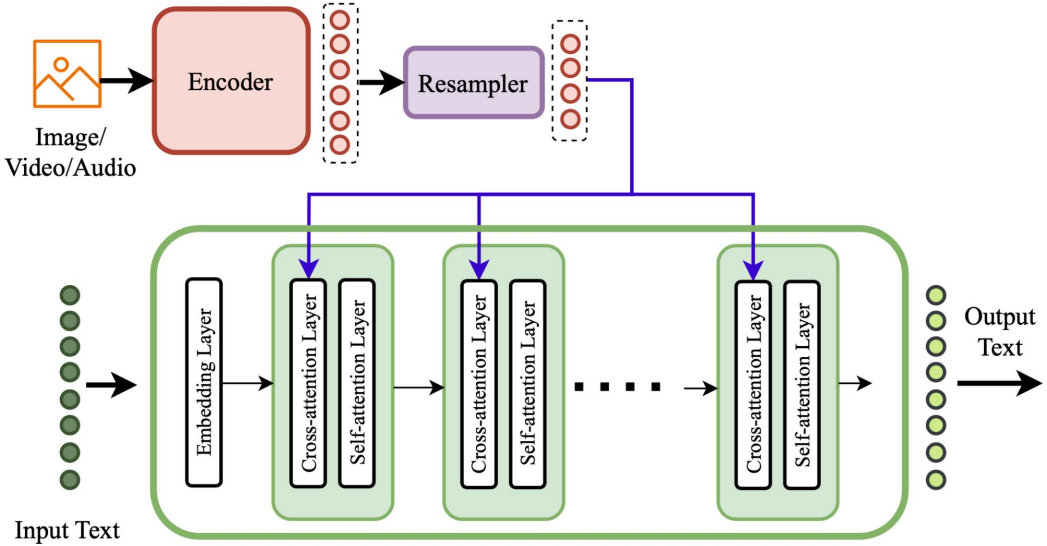
SC-DF: **Standard Cross-Attention**

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

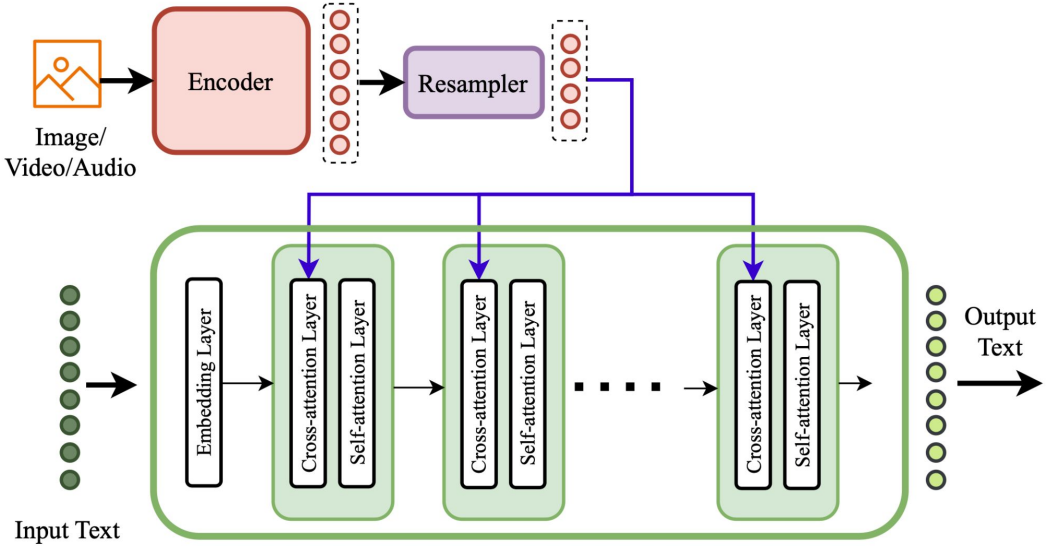
Flamingo
OpenFlamingo
Otter
Multimodal-GPT

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

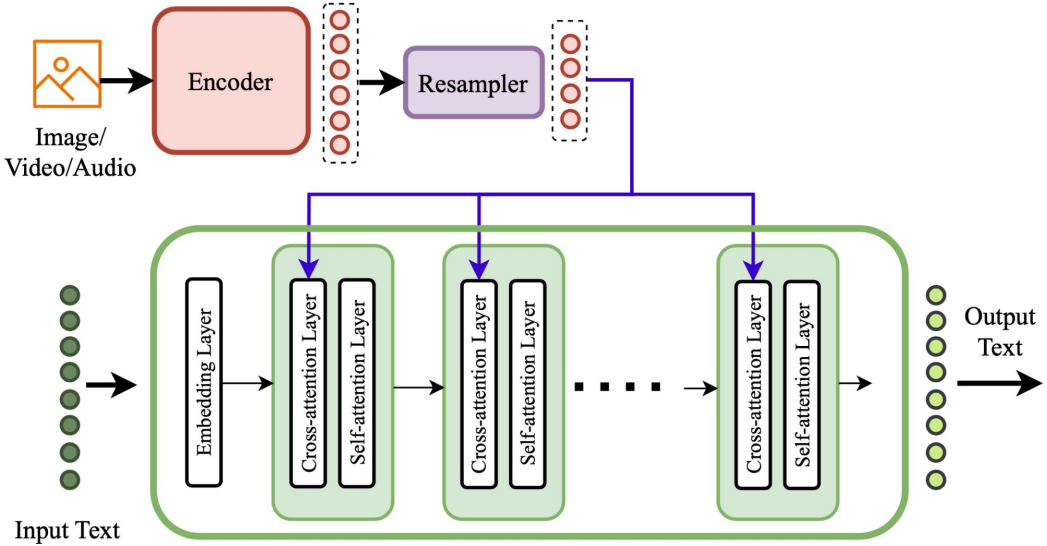
VL-BART
VL-T5

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



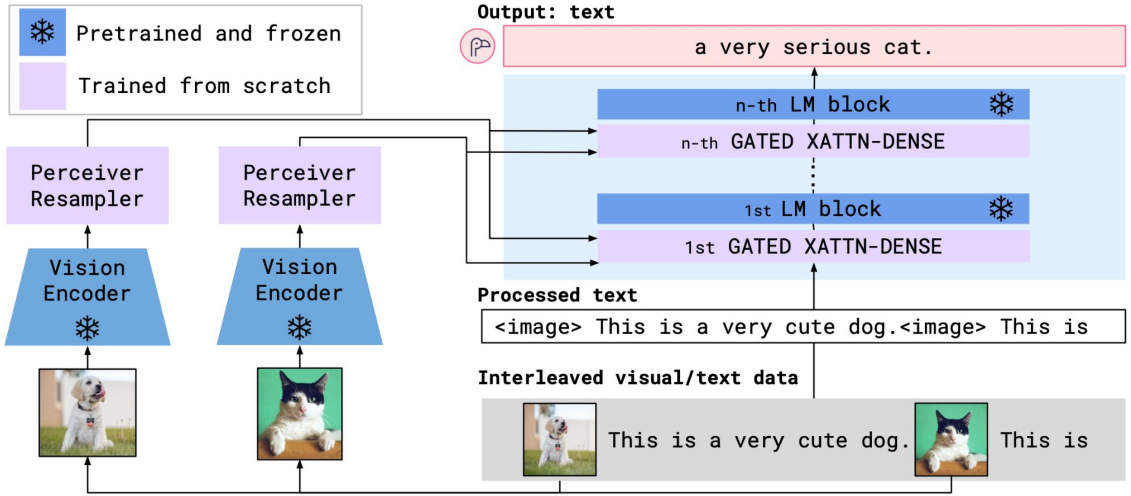
before

Flamingo
OpenFlamingo
Otter
Multimodal-GPT

after

VL-BART
VL-T5

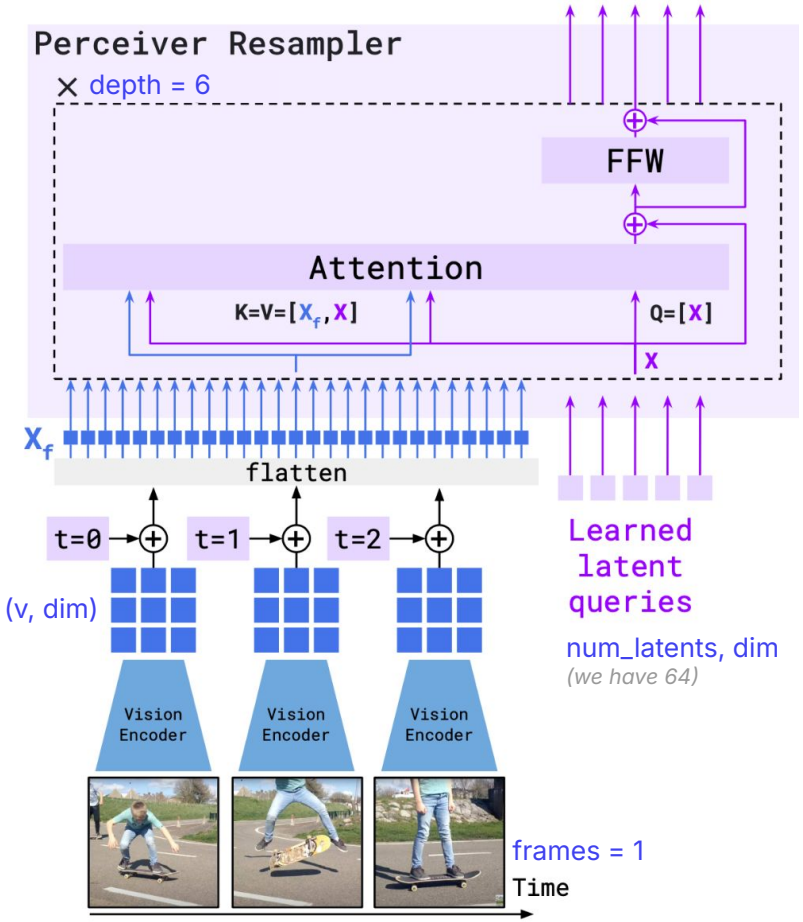
SC-DF: **OpenFlamingo** (Nov 2022)



OpenFlamingo follows Flamingo architecture

vision model
CLIP ViT-L/14
(NFNet)

language model
RedPajama / MPT
(Chinchilla)

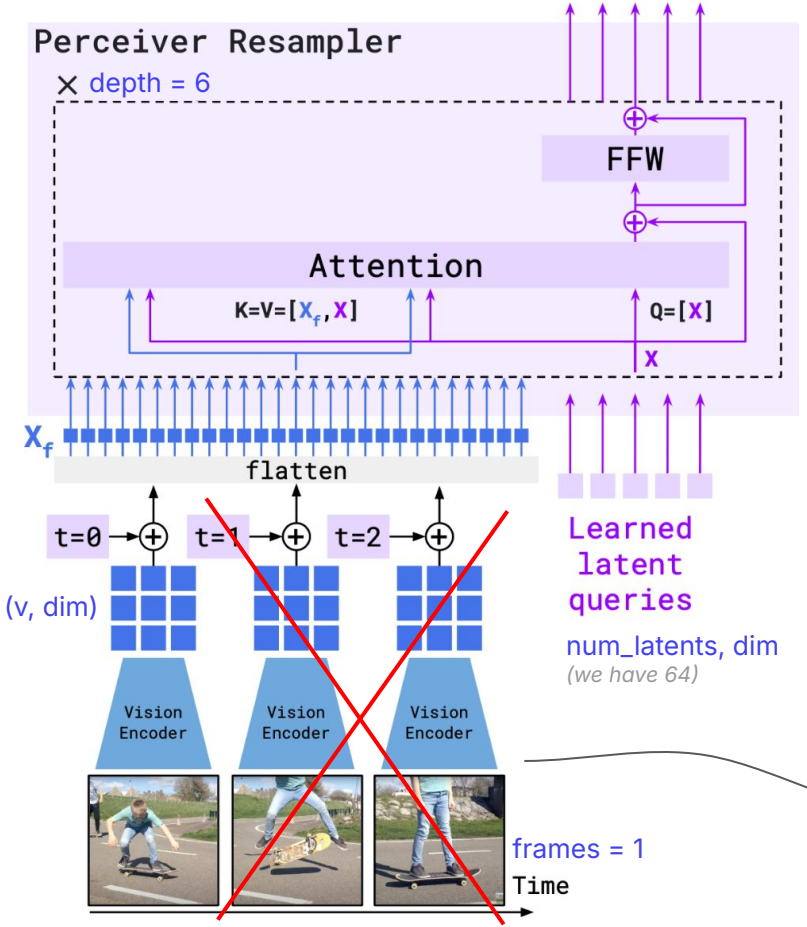


OpenFlamingo : Perceiver

b, t, f, v, d → b, t, (f * v), d → b, t, l, d

d = 1024 l = 64

OpenFlamingo : Perceiver

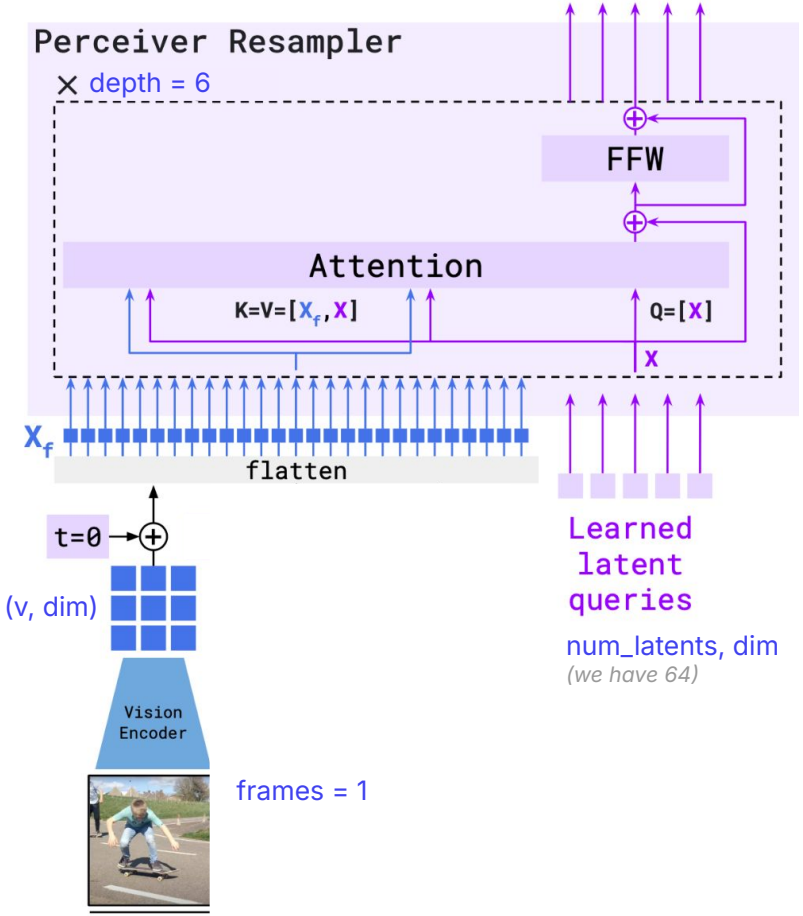


b, t, f, v, d → b, t, (f * v), d → b, t, l, d

d = 1024 l = 64

Do not have video modality in OpenFlamingo

OpenFlamingo : Perceiver



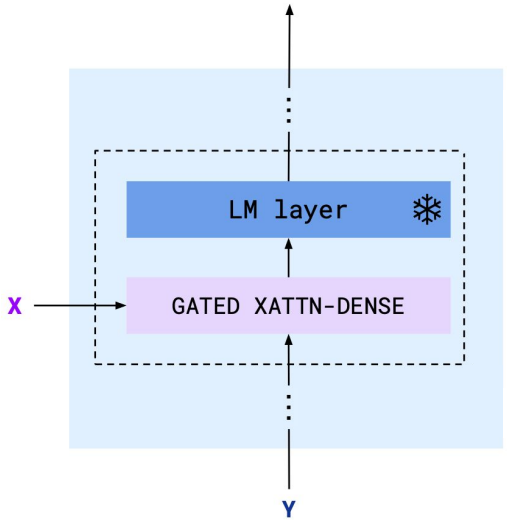
b, t, f, v, d → b, t, (f * v), d → b, t, l, d

d = 1024 l = 64

batch size	**b**
image examples	**t**
video frames	**f**
visual tokens	**v**
embed dim	**d**

OpenFlamingo: Feature Fusion

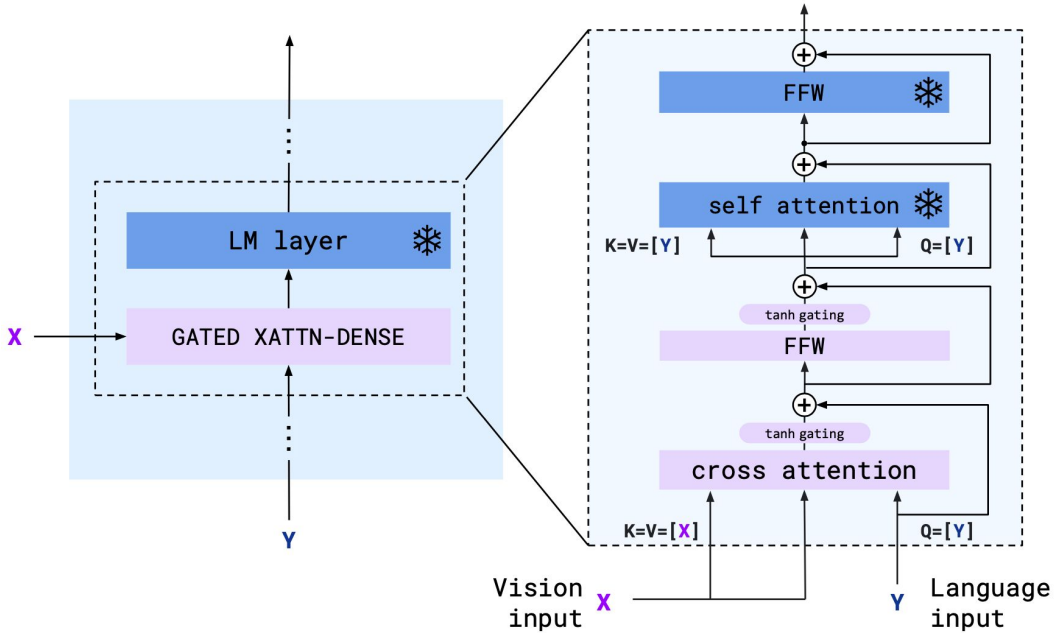
1. **freeze** the pretrained LM blocks
2. insert **gated cross-attention dense** blocks between the original layers
3. keep leyers gated to keep LM intact at initialization
4. **queries** = LM inputs



OpenFlamingo: Feature Fusion

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

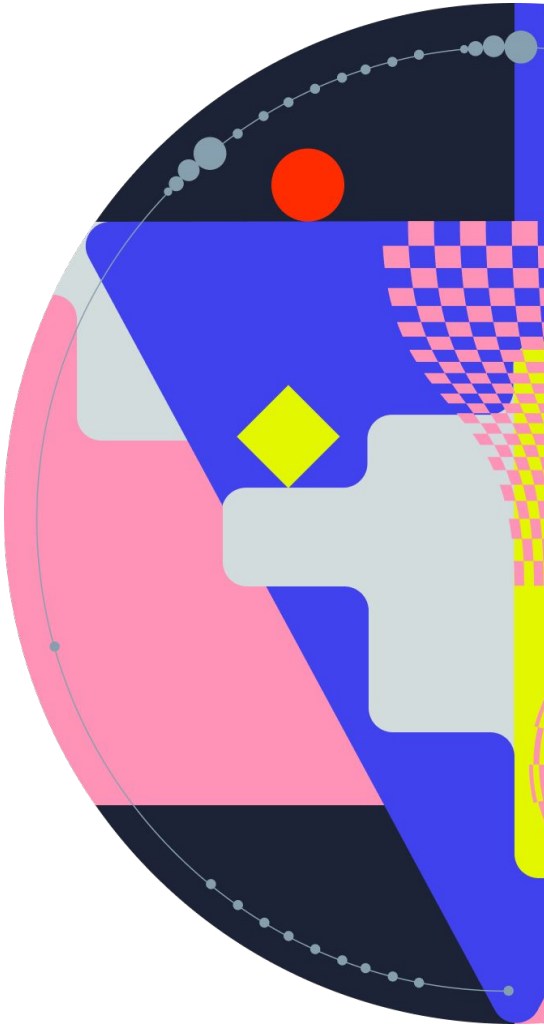
tanh-gating mechanism — multiples output of newly initialized layer by tanh(α)



1.2

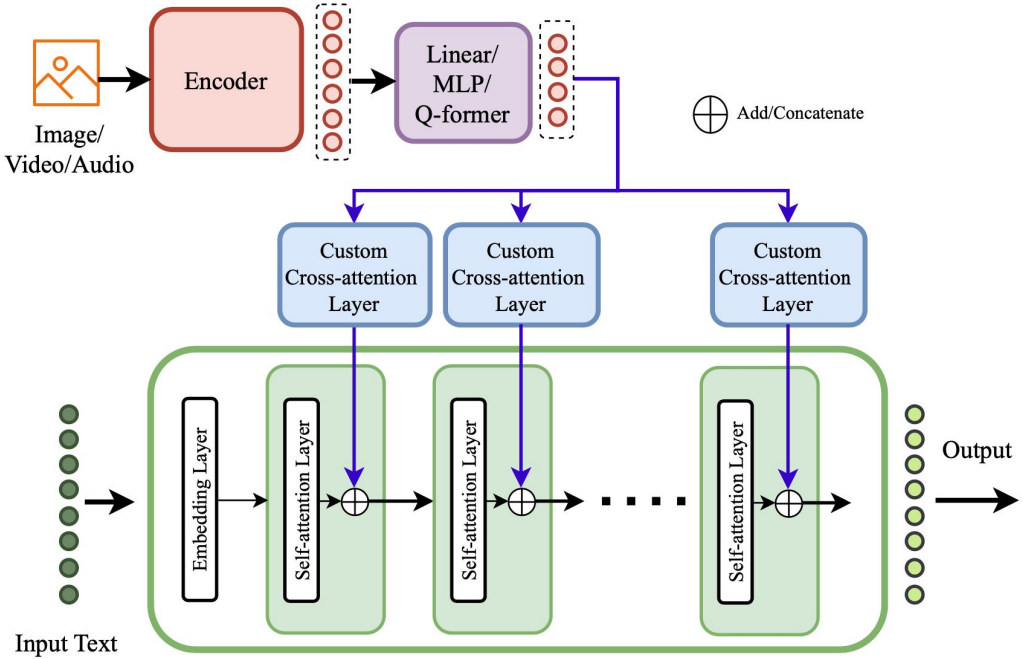
Deep Fusion:

Custom Layers Deep Fusion (CL-DF)



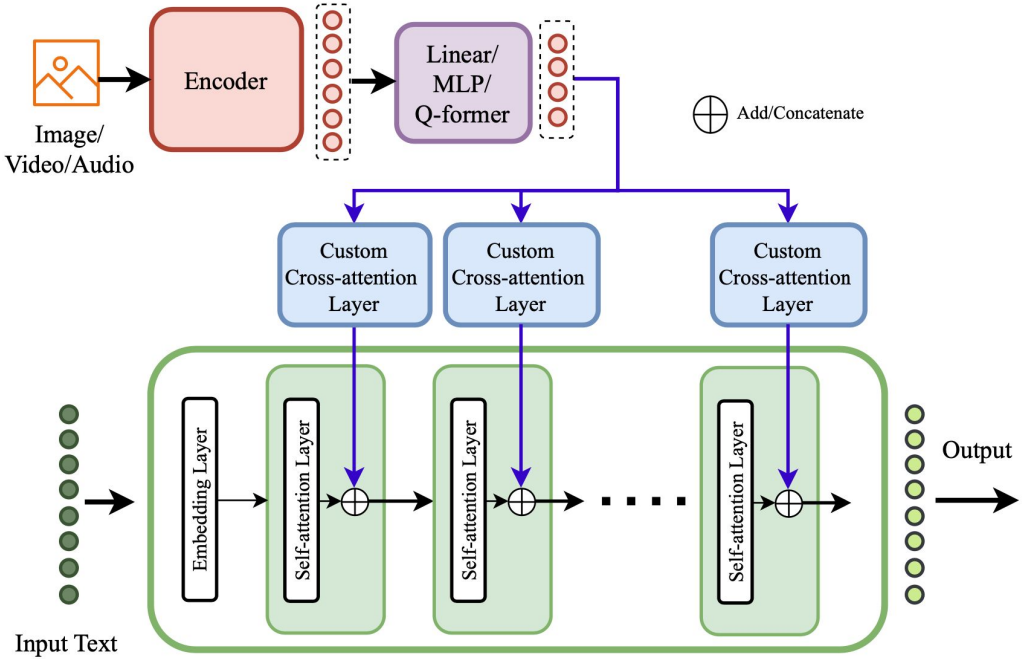
CL-DF: **Custom Layers**

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

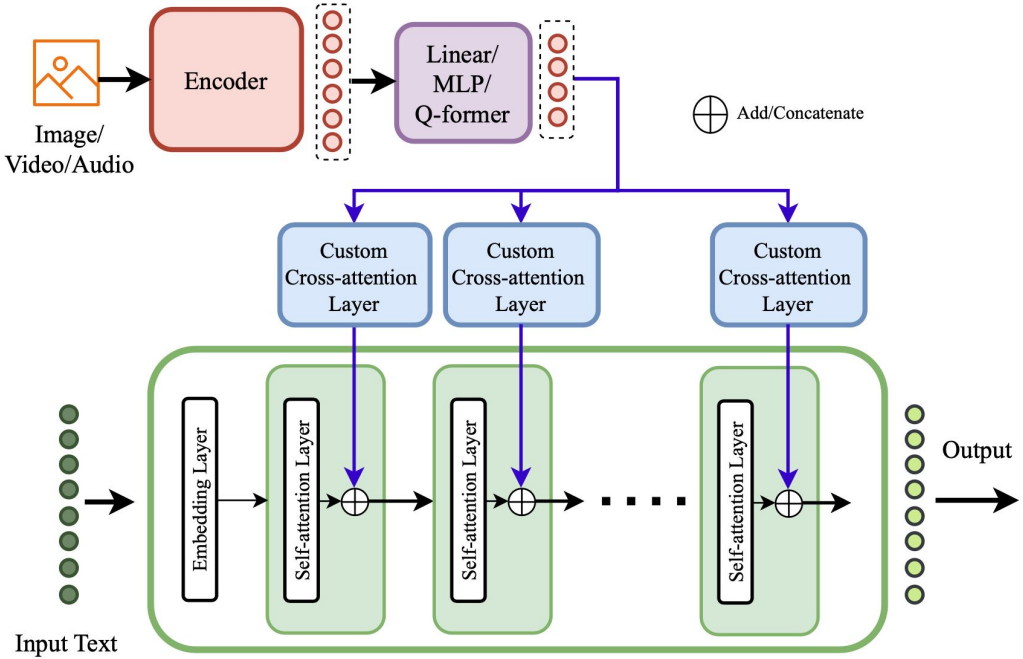
LLaMA-Adapter
CogVLM
InternVL
mPLUG-Owl2

other custom layers

MoE-LLaVA
LION

CL-DF: **Custom Layers**

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



custom cross-attention

LLaMA-Adapter
CogVLM
InternVL
mPLUG-Owl2

other custom layers

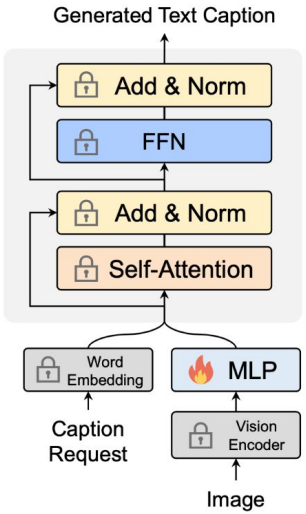
MoE-LLaVA
LION

CL-DF: **MoE-LLaVA** (Dec 2024)

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

Stage 1

adapt visual tokens

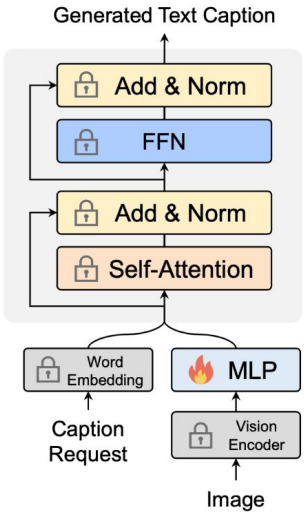


CL-DF: **MoE-LLaVA** (Dec 2024)

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

Stage 1

adapt visual tokens



vision model
CLIP-Large
(following LLaVA-1.5)

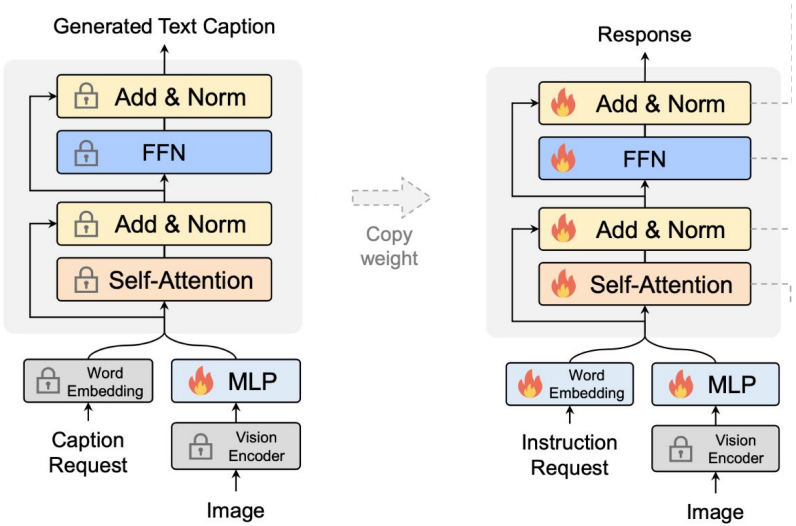
language model
LLaMA / Vicuna /Qwen…

CL-DF: **MoE-LLaVA** (Dec 2024)

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

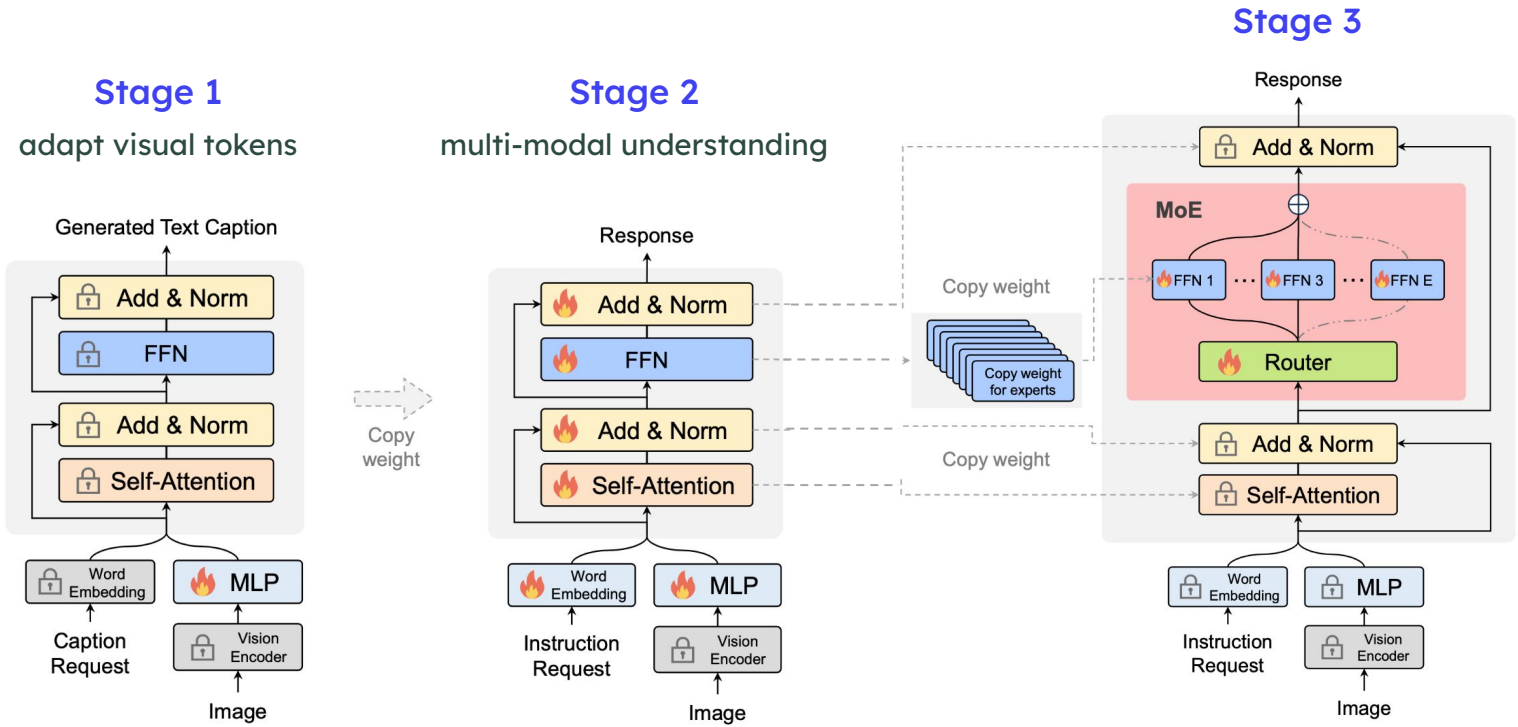
Stage 1
adapt visual tokens

Stage 2
multi-modal understanding



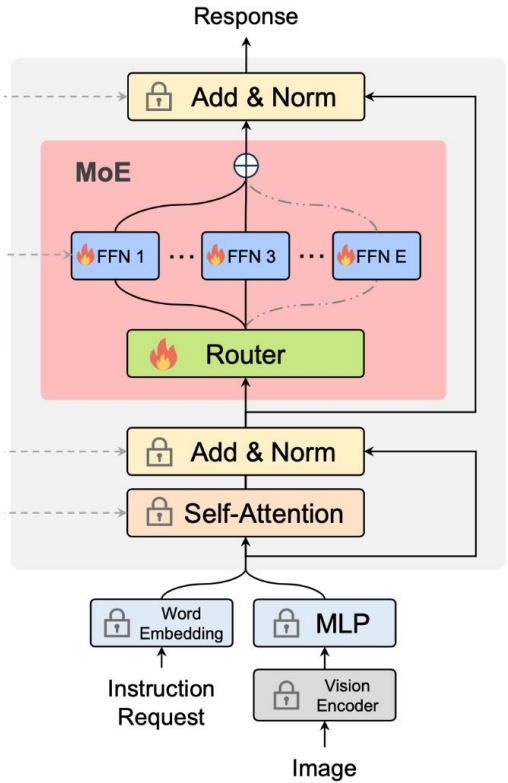
CL-DF: **MoE-LLaVA** (Dec 2024)

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



many LLaMAs!

MoE-LLaVA: Router



1 have E experts, each expert = FFN

$$\mathcal{E} = [e_1, e_2, \cdots, e_E]$$

2 router = linear layer that assigns probabilities to experts

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

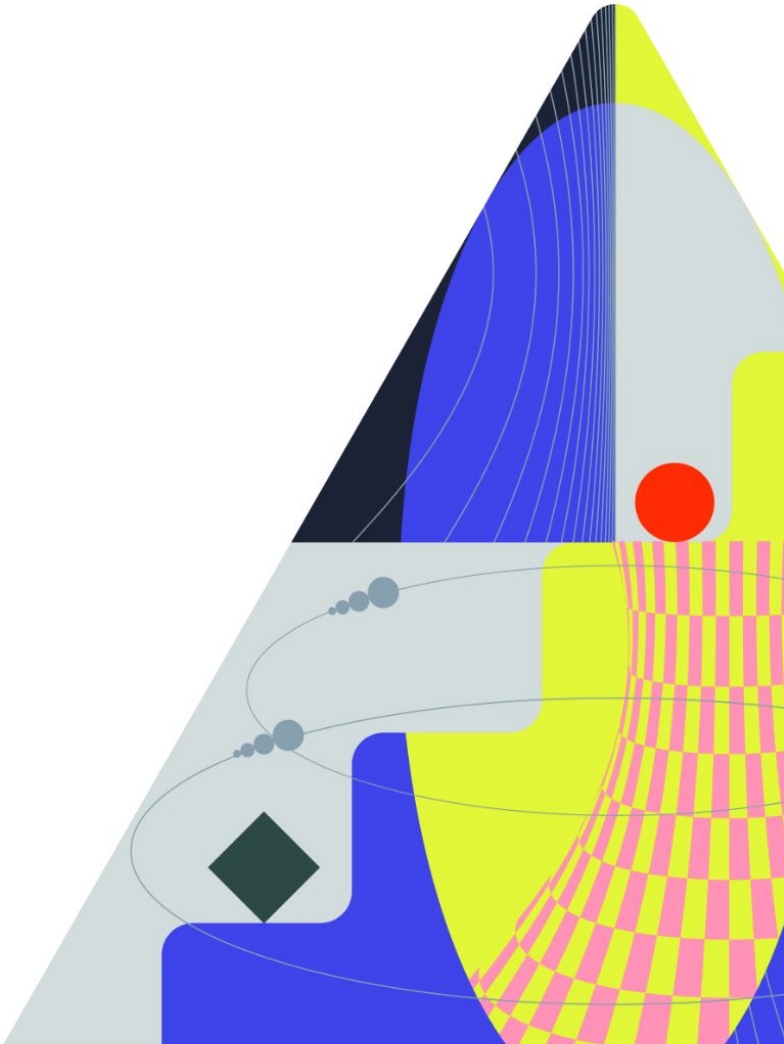
3 calculate weighted sum

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

2.1

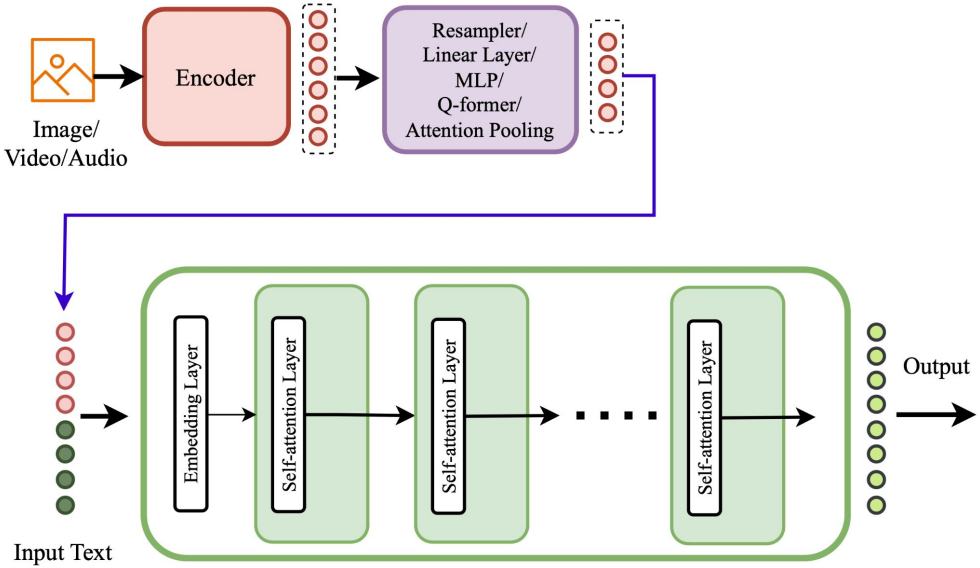
Early Fusion:

Non-Tokenized Early Fusion (NT-EF)



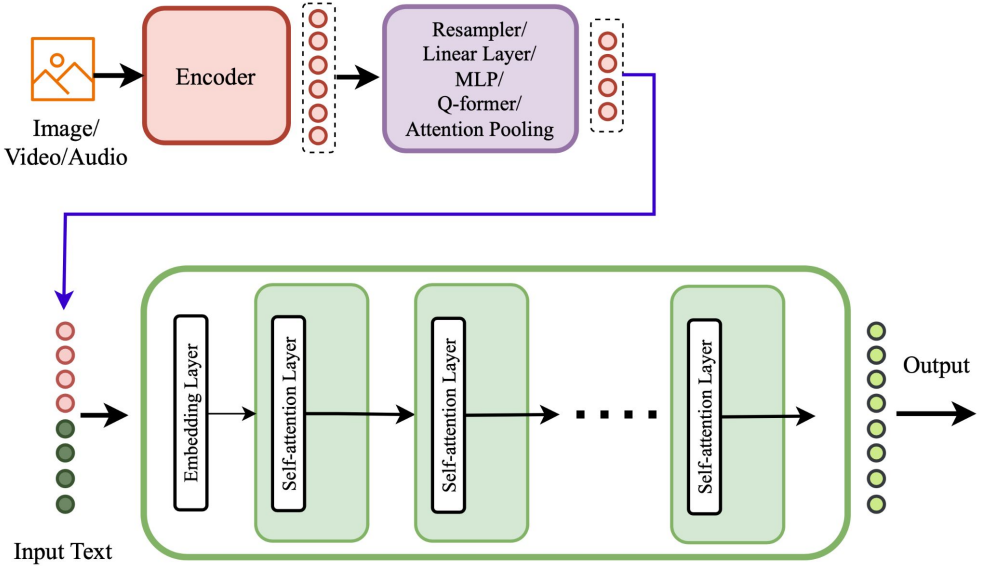
NT-EF: **Non-Tokenized**

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



NT-EF: **Non-Tokenized**

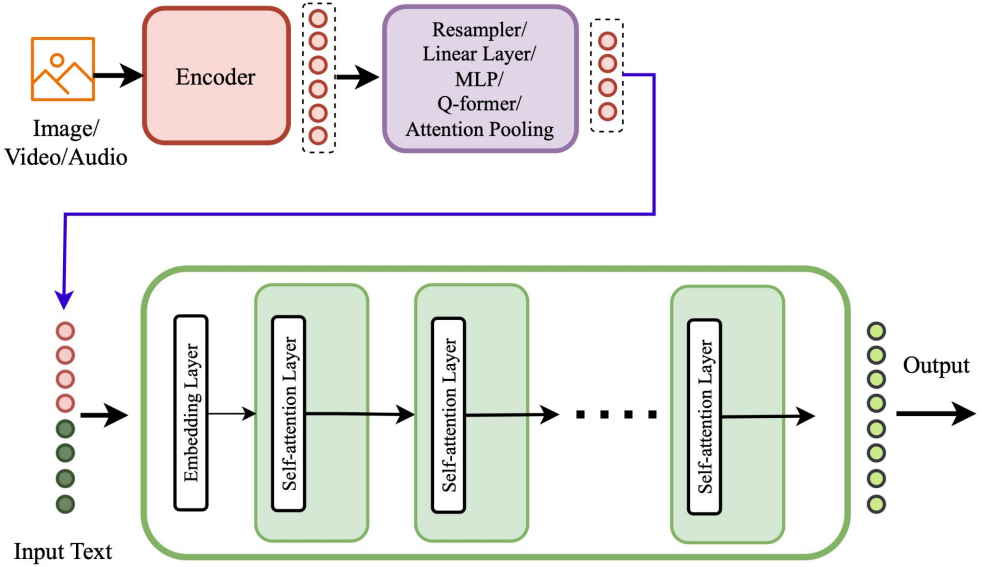
Non-tokenized input modalities are **directly fed to the model** rather than to internal layers



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

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)

Stage 1: Pretraining

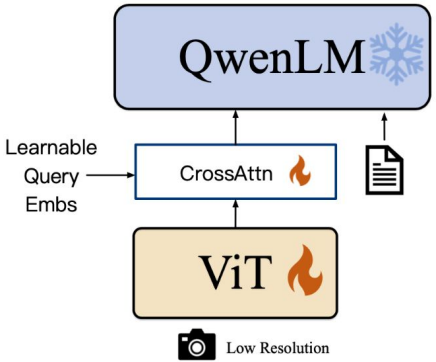
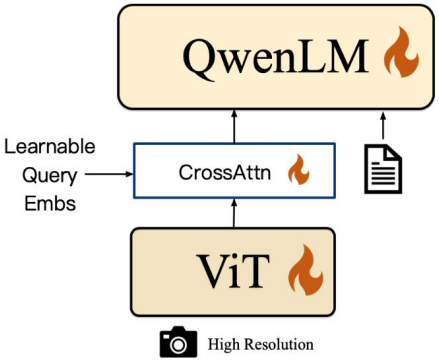


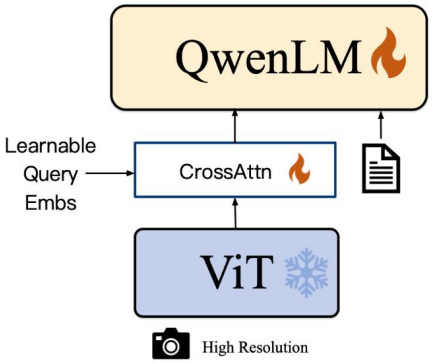
Image-Text pairs

Stage 2: Multi-task pretraining



Multi-task and Interleaved VL Data

Stage 3: Supervised Fine-tuning



Chat Interleaved VL Data

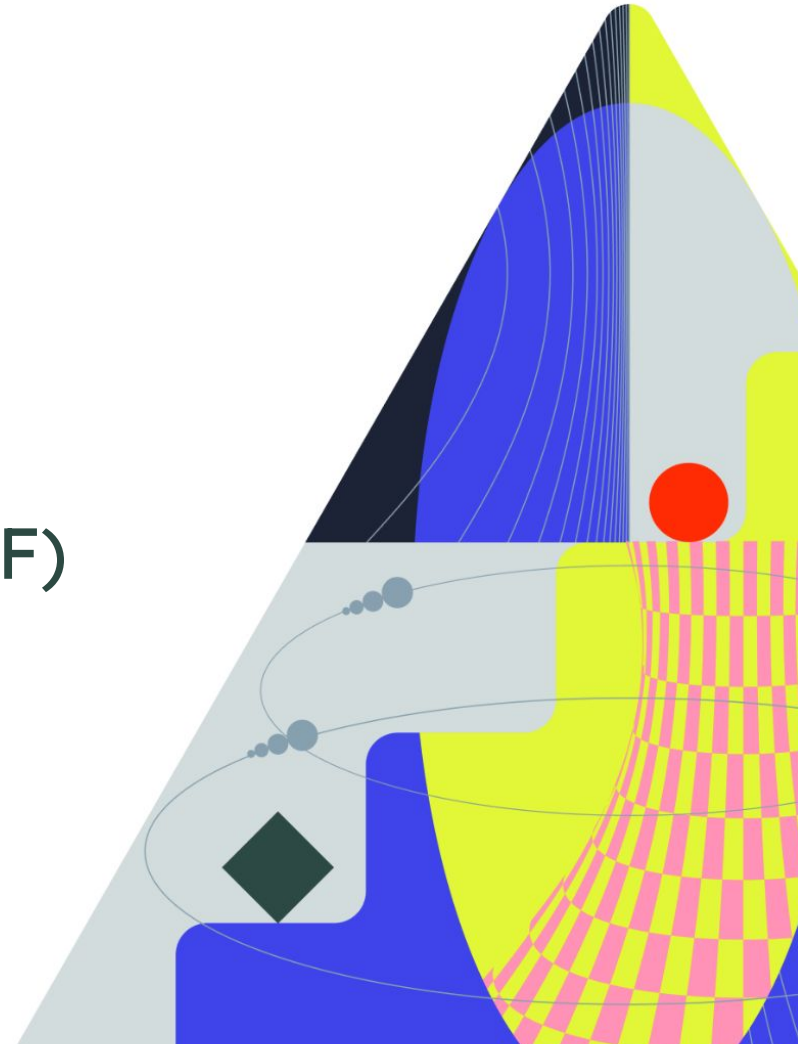
vision model
OpenClip ViT-bigG

language model
Qwen-7B

2.2

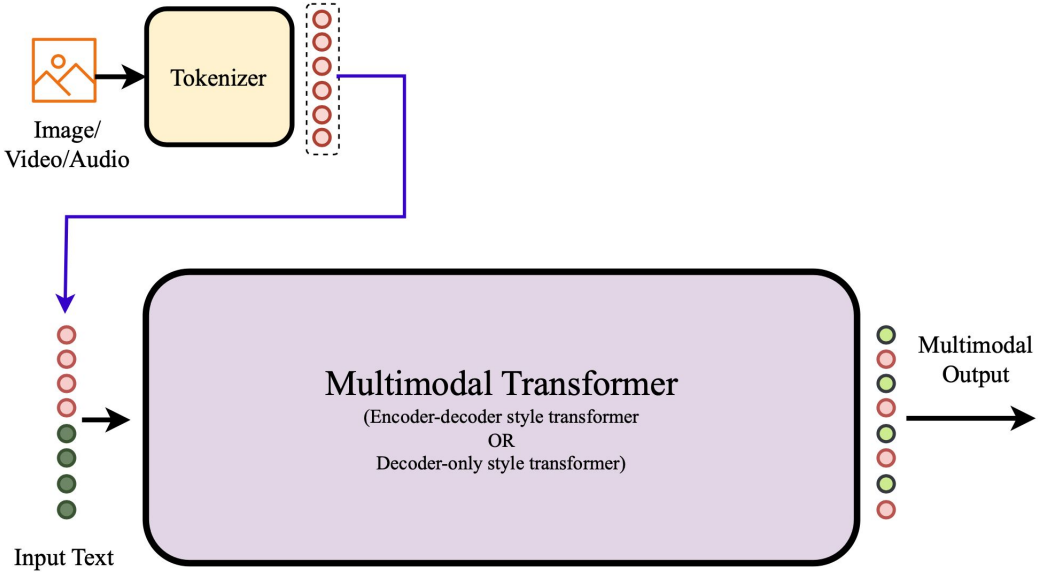
Early Fusion:

Tokenized Early Fusion (T-EF)



T-EF: **Tokenized**

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



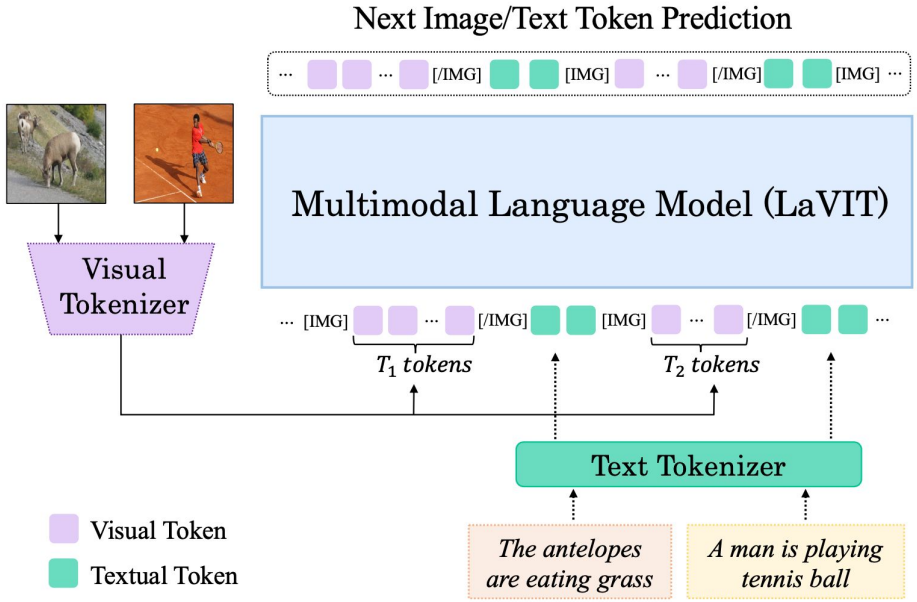
decoder-only transformer

- **LaVIT**
- TEAL
- CM3Leon
- VL-GPT

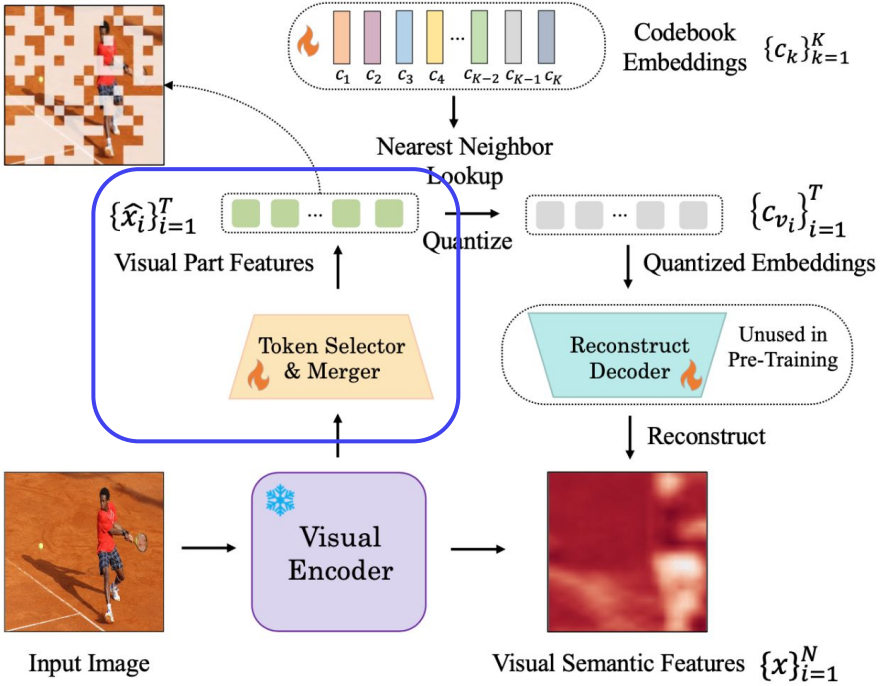
encoder-decoder transformer

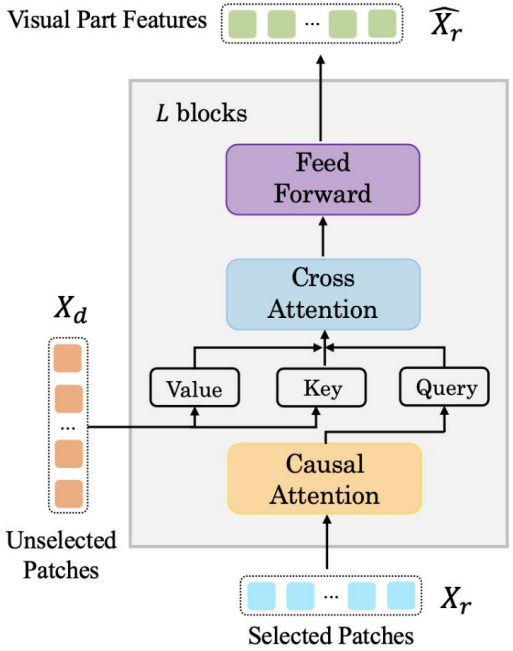
- Unified-IO
- Unified-IO
- 4M

T-EF: **LaVIT** (Mar 2024)



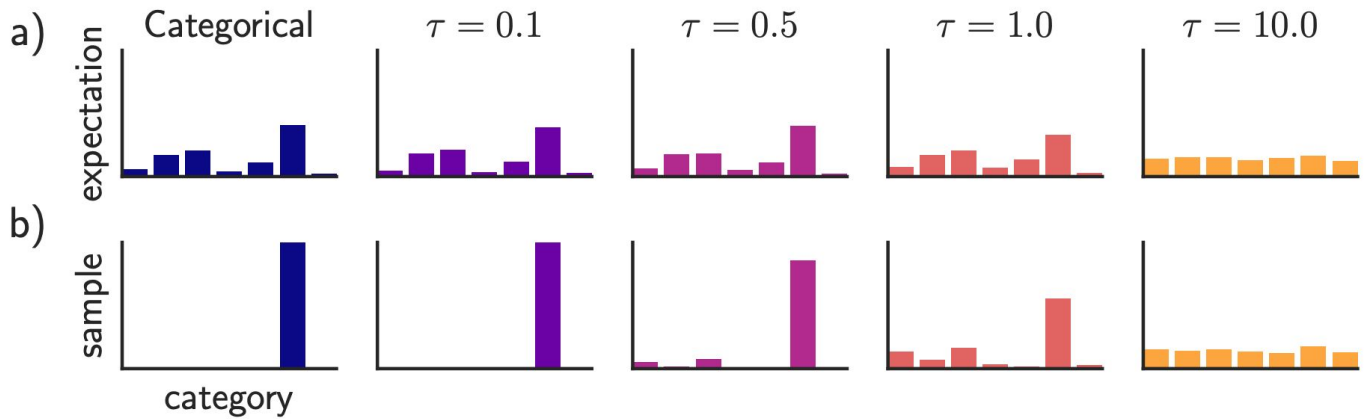
T-EF: **LaVIT** (Mar 2024)





T-EF: **LaVIT** (Mar 2024)

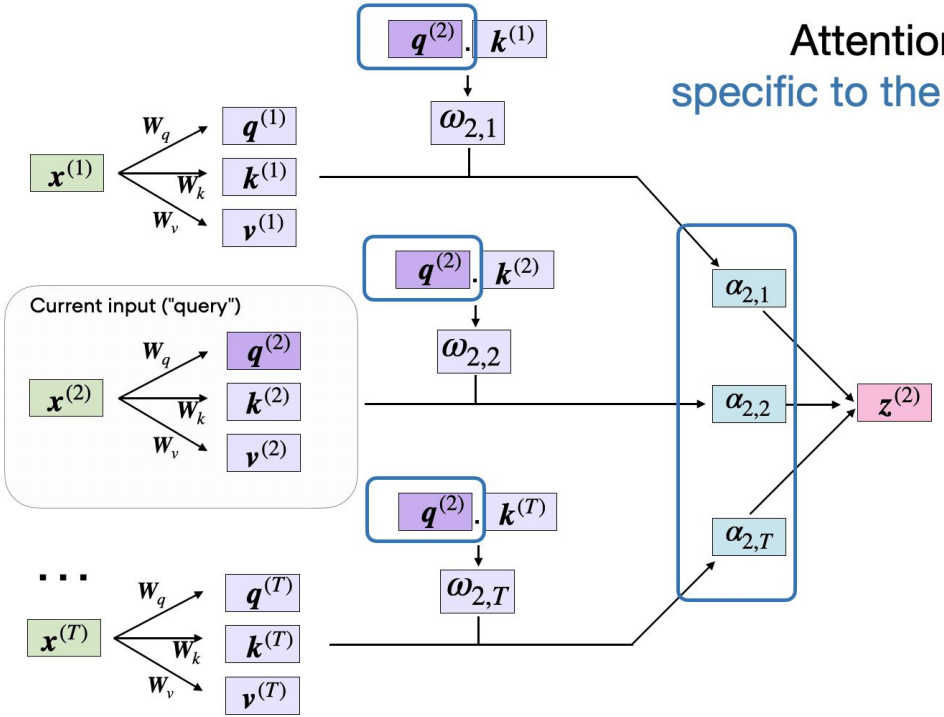
$$\hat{\pi_{i,j}} = \frac{\exp((\log \pi_{i,j} + G_{i,j})/\tau)}{\sum_{r=1}^{2} \exp((\log \pi_{i,r} + G_{i,r})/\tau)}$$



Conclusions

1. Classified multimodal models: **Deep** & **Early** Fusion
2. Focused on **VLM models**: Image + Text → Text
3. Investigated carefully **Flamingo**, **MoE-LLaVA**, **Qwen-VL**, and **LaVIT** models

Appendix: **Self-Attention**

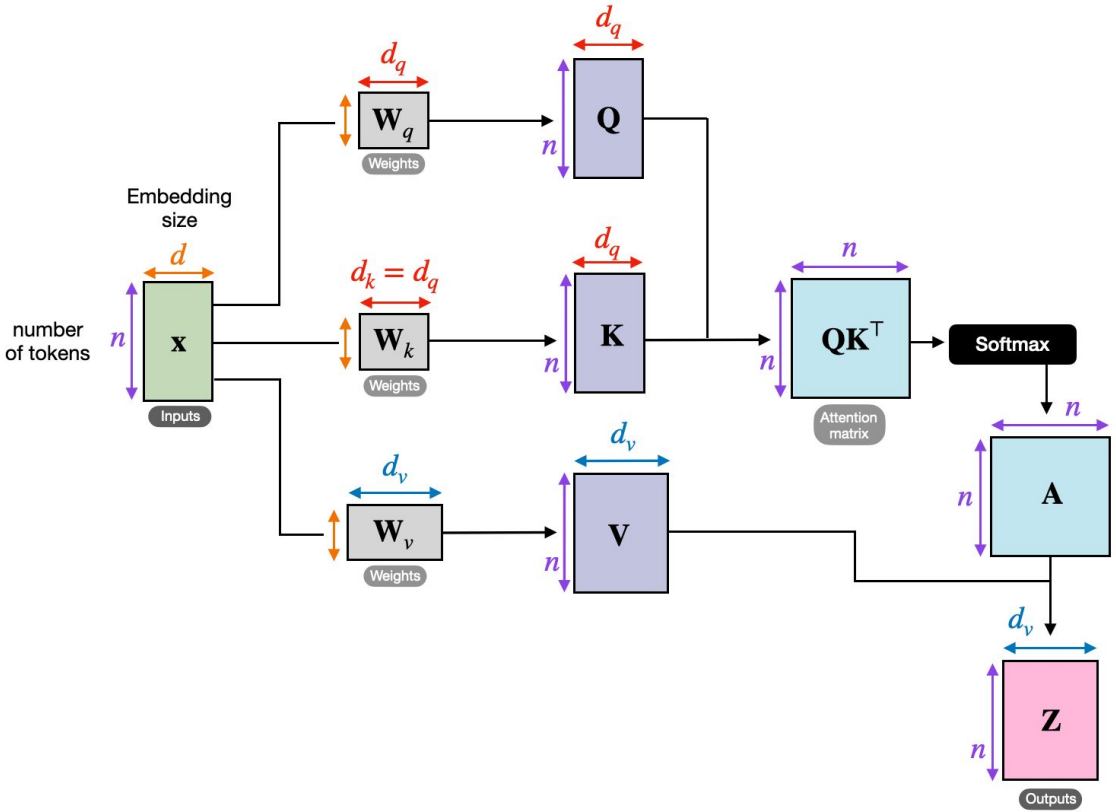


Attention weights are
specific to the current input token

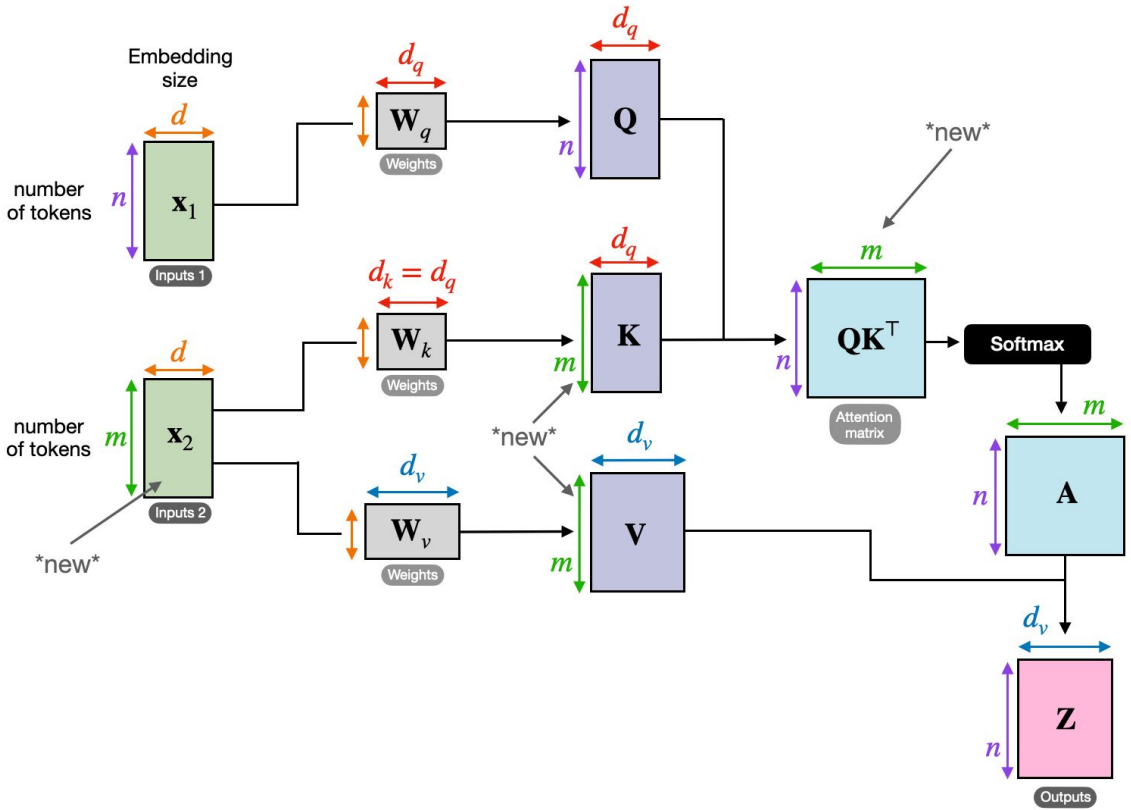
where $z^{(2)} = \sum_{j=1}^{T} \alpha_{2,j} \, v^{(j)}$

where $\alpha_2 = \text{softmax}\left(\frac{\omega_2}{\sqrt{d_k}}\right)$

Appendix: **Self-Attention**



Appendix: **Cross-Attention**



MoE-LLaVA: Router