Школа анализа данных

Seminar 2: Early Fusion. Video modality

Supplementary slides to [Google Colab notebook]

Zinkovich Viktoriia

Recap of previous seminar

1. Deep Fusion

deeply fuses multimodal inputs within internal layers

1.1. **Standard** Cross-Attention (SC-DF)

OpenFlamingo

- perceiver resampler
- cross-attention
- tanh gating

1.2. **Custom** Layers (CL-DF)

MoE-LLaVA

- vision encoder MLP
- Mixture-of-Experts layer
- Router

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)

Questions

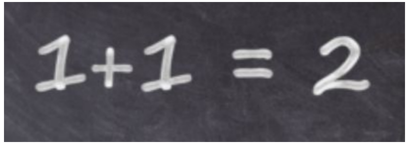
Giving **OpenFlamingo** tricky few-shot examples

training dataset = LAION-2B with image-text pairs

An image of 2 x 2 = 4

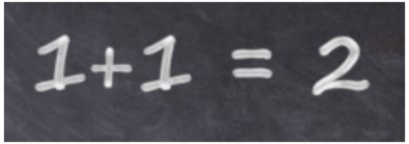
An image of 3 + 3 = 6

An image of



input images and prompts

An image of 1 + 1 = 2



output

Questions

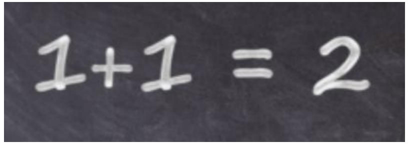
Giving **OpenFlamingo** tricky few-shot examples

training dataset = LAION-2B with image-text pairs

An image of 2 x 2 = 4

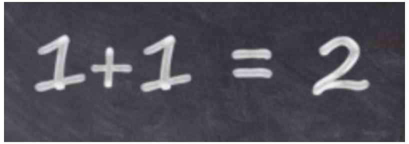
An image of 3 + 3 = 6

What is the color of board?



input images and prompts

An image of 2 + 2 = 4.
An image of 3 + 3 = 6



output

Questions

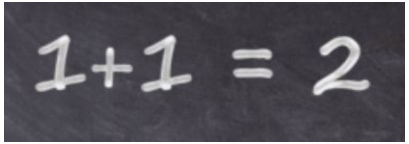
Giving **OpenFlamingo** tricky few-shot examples

training dataset = LAION-2B with image-text pairs

Print equation: 2 x 2 = 4

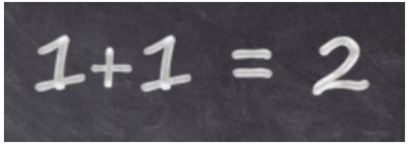
Print equation: 3 + 3 = 6

Print equation:



input images and prompts

Print equation 2 + 2 = 4
Print equation 3 + 3 = 6
Print equation 2 + 2 = 4
Print

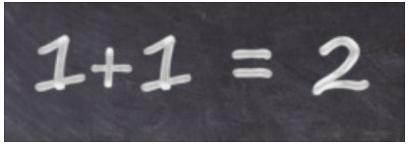


output

Questions

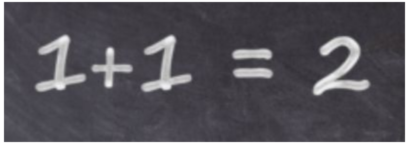
Giving **OpenFlamingo** tricky few-shot examples
training dataset = LAION-2B with image-text pairs

An image of



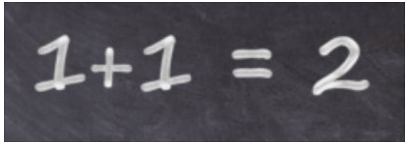
input

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



output

What is on the image?
<|endofchunk|>



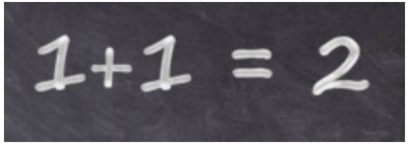
input

Questions

Giving **OpenFlamingo** tricky few-shot examples

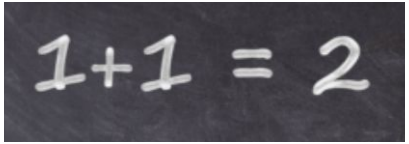
training dataset = LAION-2B with image-text pairs

An image of



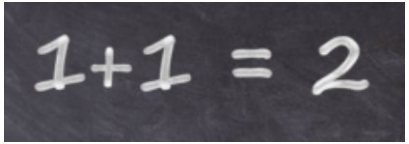
input

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



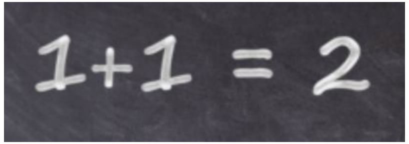
output

What is on the image? **<|endofchunk|>**



input

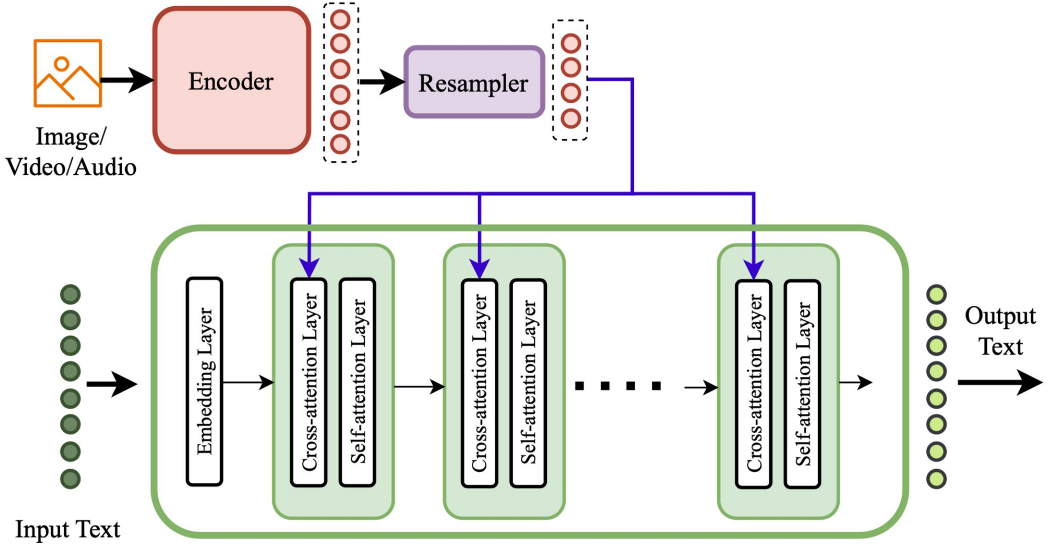
I've been thinking a lot lately about what it means to be a "successful"



output

Questions

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



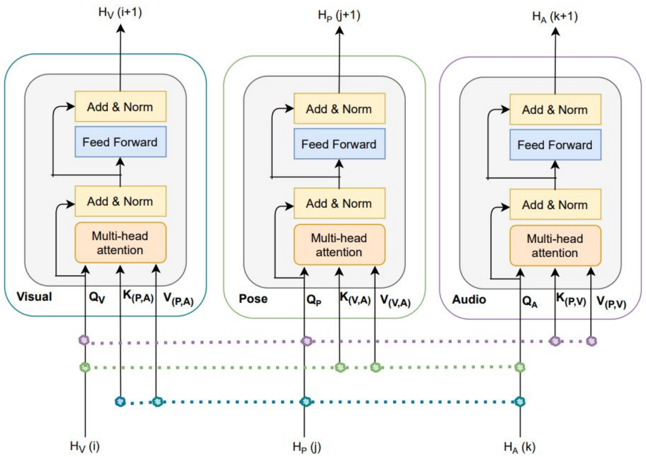
Cross-attention between **two modalities** (text and image) is used in **Deep Fusion**, particularly in cross-attention layers inside LLM

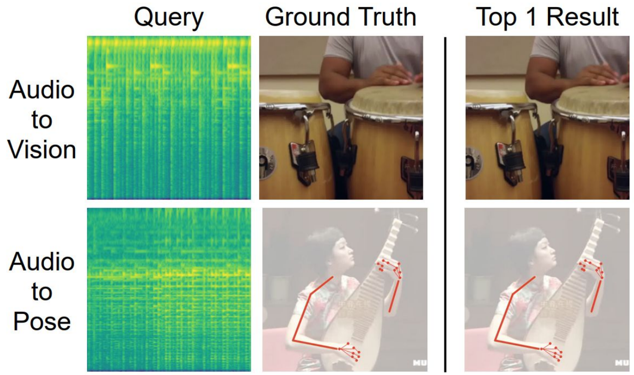
(again: OpenFlamingo)

Questions

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]

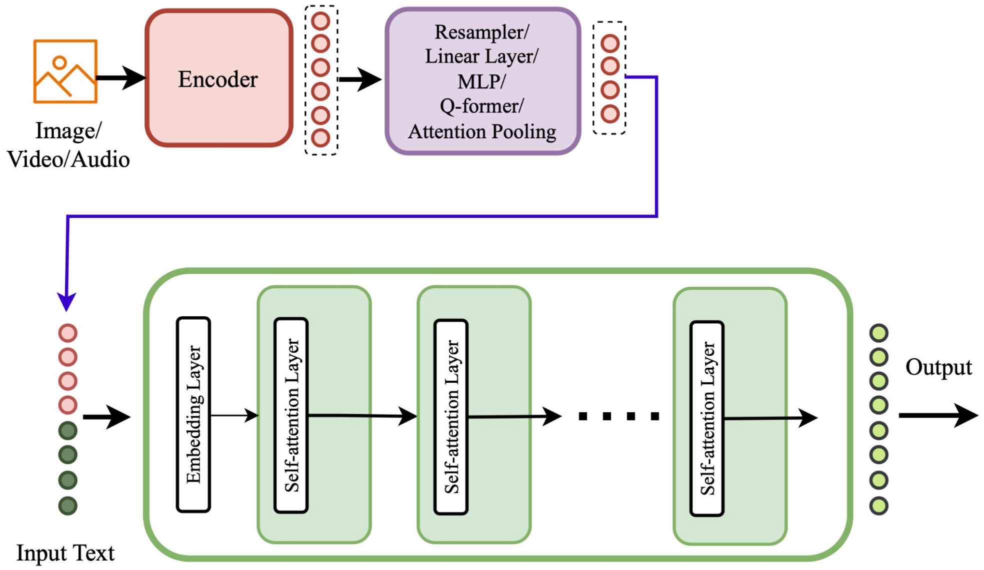
2.1

Early Fusion:

Non-Tokenized Early Fusion (NT-EF)

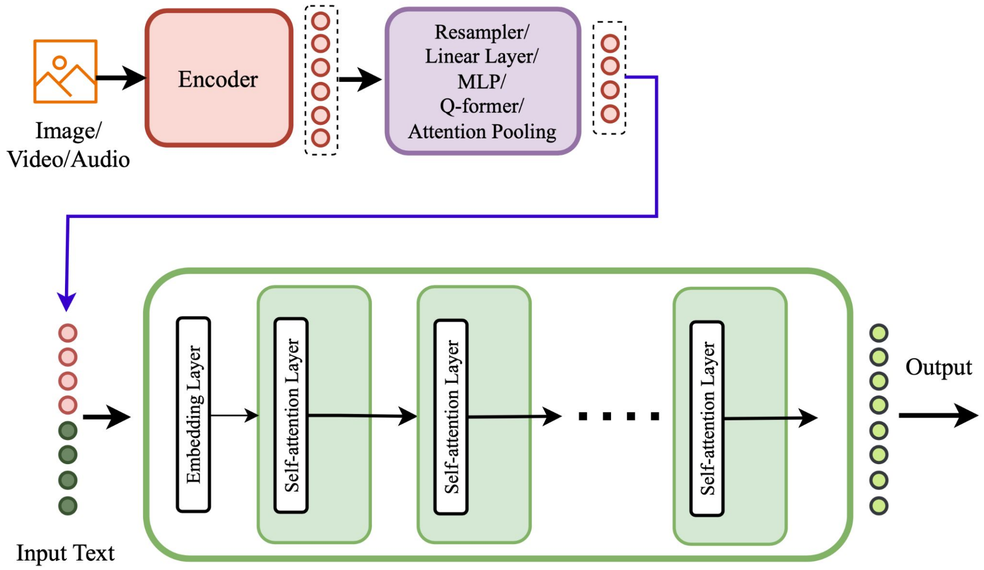
NT-EF: **Non-Tokenized**

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



NT-EF: **Non-Tokenized**

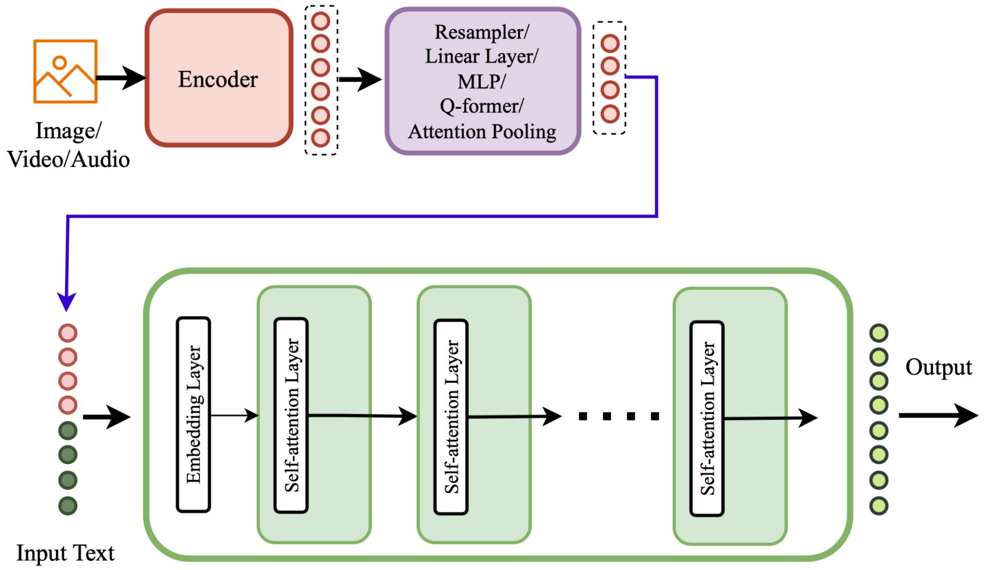
Non-tokenized input modalities are **directly fed to the model** rather than to internal layers



→ **Q-Former:** BLIP-2 🥇, MiniGPT-v2

NT-EF: **Non-Tokenized**

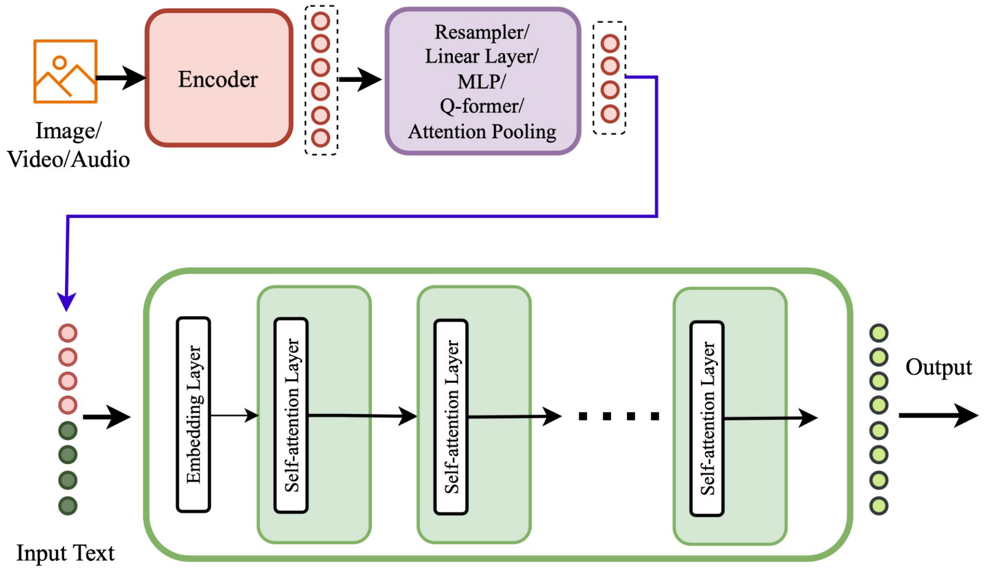
Non-tokenized input modalities are **directly fed to the model** rather than to internal layers



- **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)

Qwen-VL

Alibaba Group, 9.6B parameters

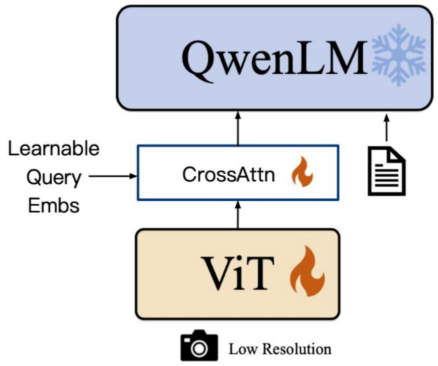
vision model = OpenClip ViT-bigG, **language model** = Qwen-7B

NT-EF: **Qwen-VL** (Oct 2023)

Qwen-VL

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

Stage 1: Pretraining
5B web data pairs → 1.4B

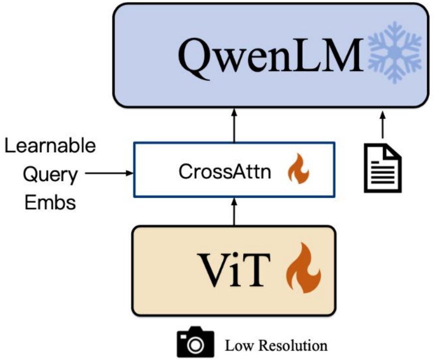


NT-EF: **Qwen-VL** (Oct 2023)

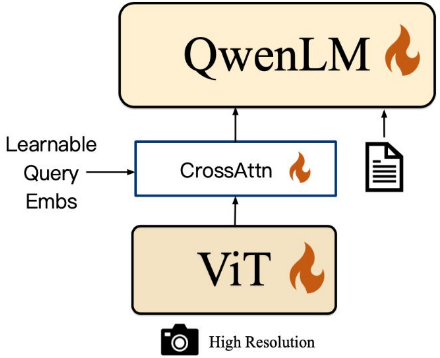
Qwen-VL

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)

Qwen-VL

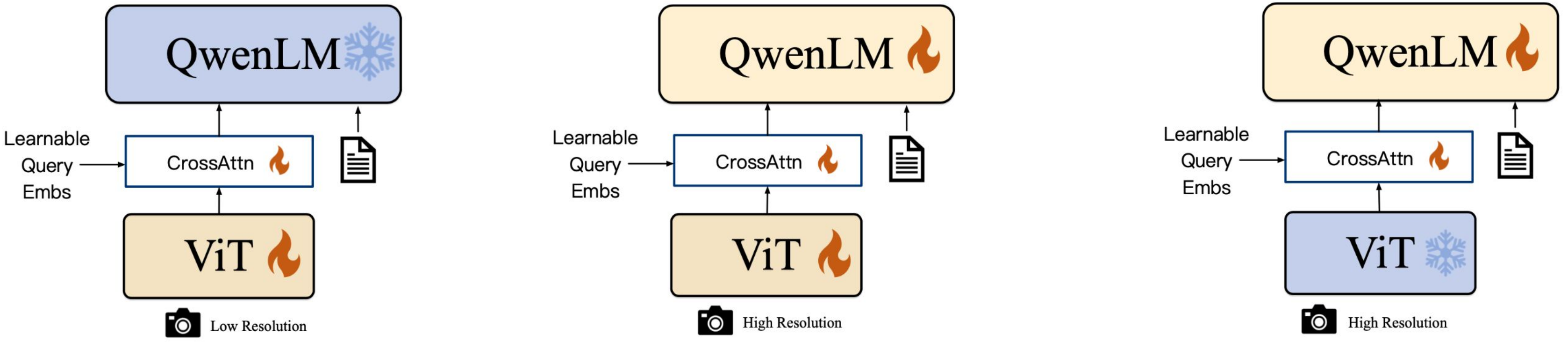
Alibaba Group, 9.6B parameters
vision model = OpenClip ViT-bigG, **language model** = Qwen-7B

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



2.2

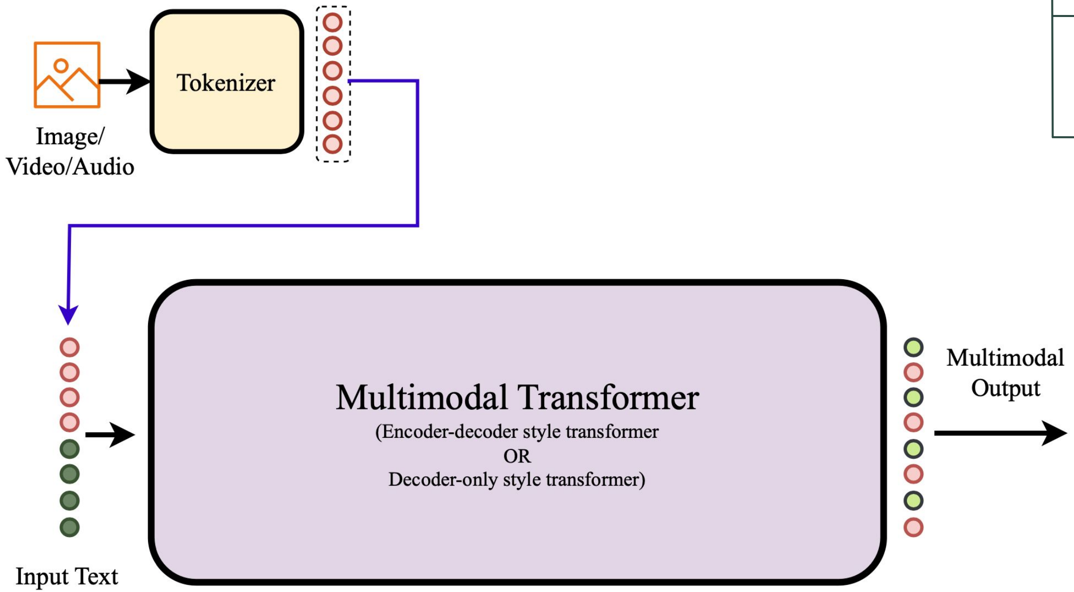
Early Fusion:

Tokenized Early Fusion (T-EF)

T-EF: **Tokenized**

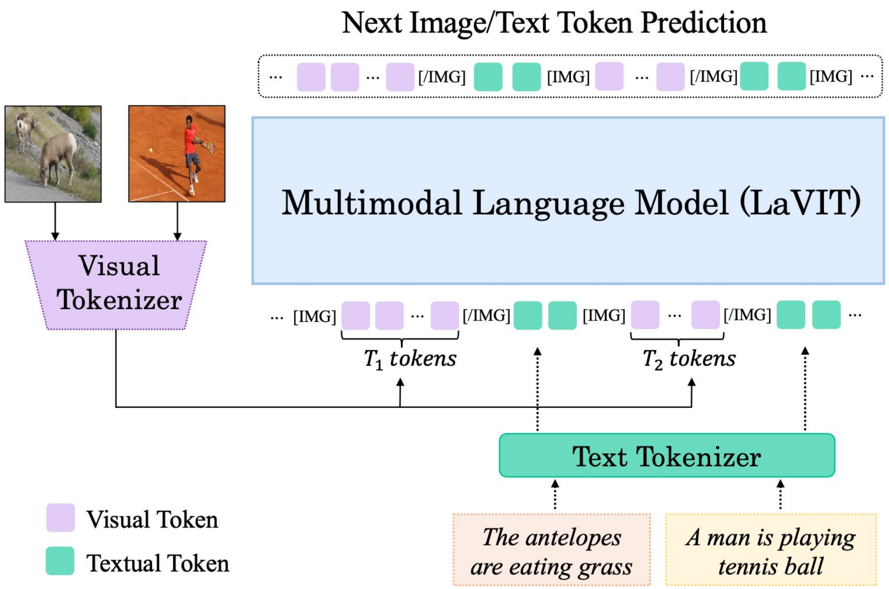
Inputs are tokenized **using a common tokenizer** or modality specific tokenizers

- **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

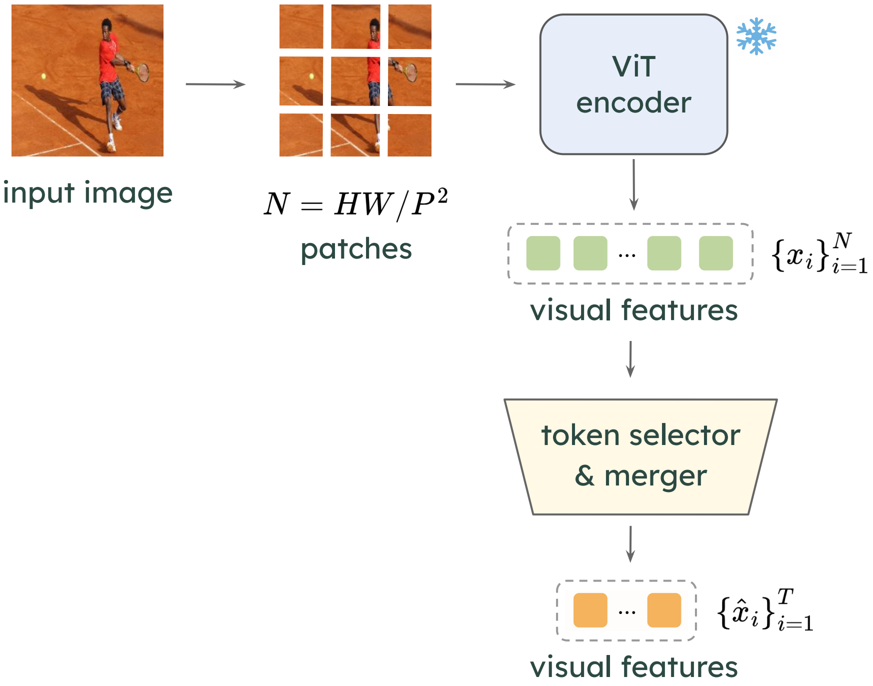


1. represent two modalities in a uniform form to exploit **LLM's next-token prediction**

2. visual tokenizer returns sequence of **discrete visual tokens** possessing word-like high-level semantics

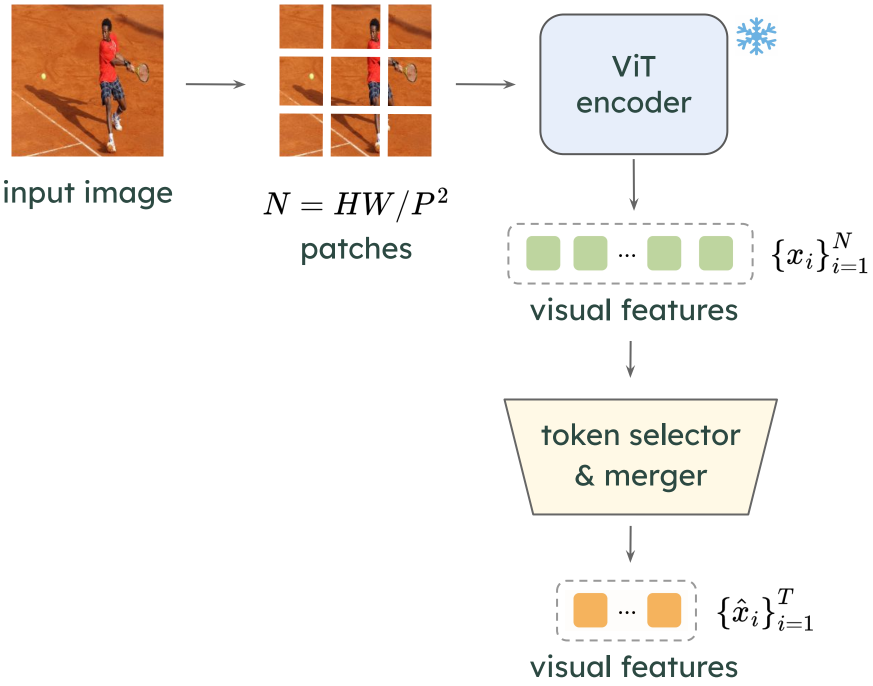
T-EF: **LaVIT** (Mar 2024, ICLR)

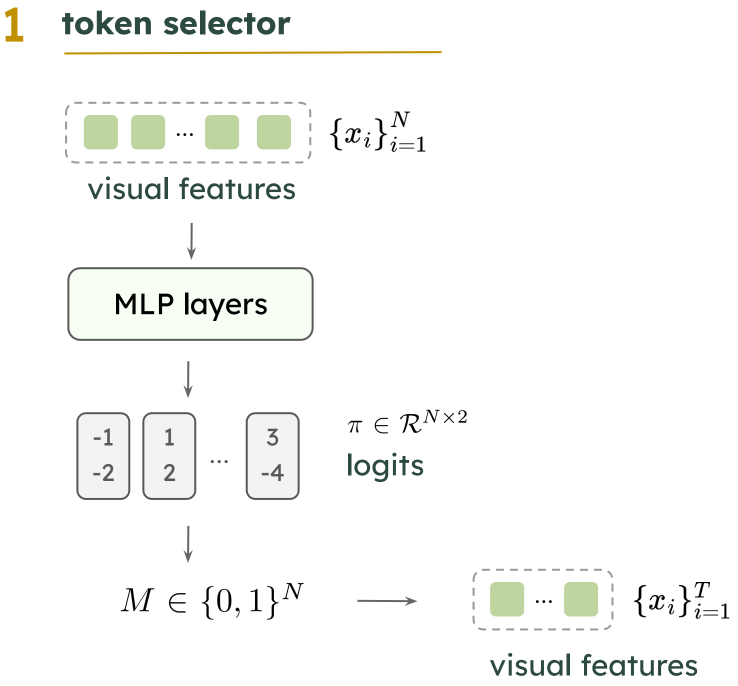
vision model = ViT-G/14 of EVA-CLIP, **language model** = LLaMA-7B



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

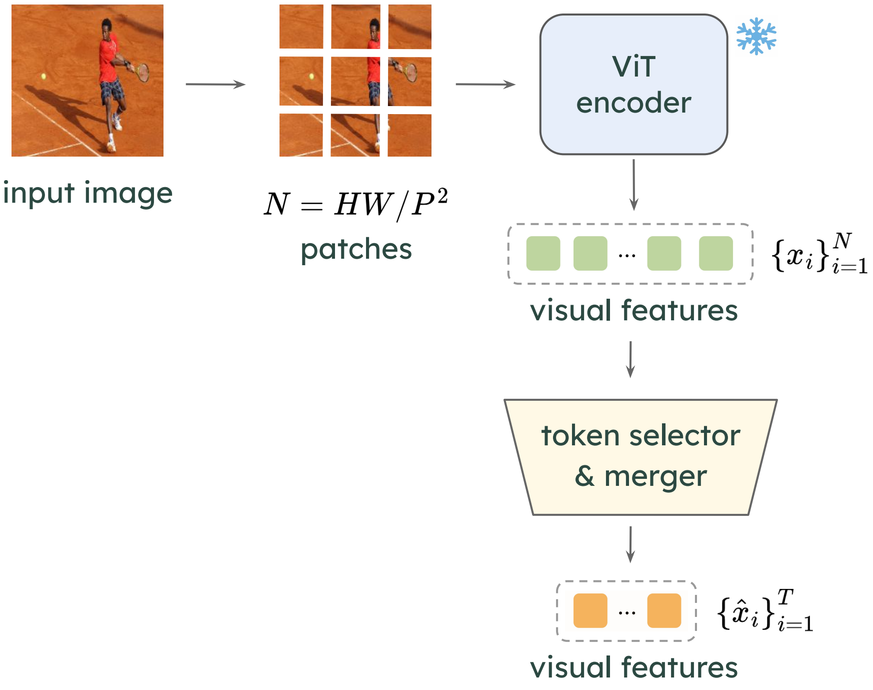
vision model = ViT-G/14 of EVA-CLIP, **language model** = LLaMA-7B



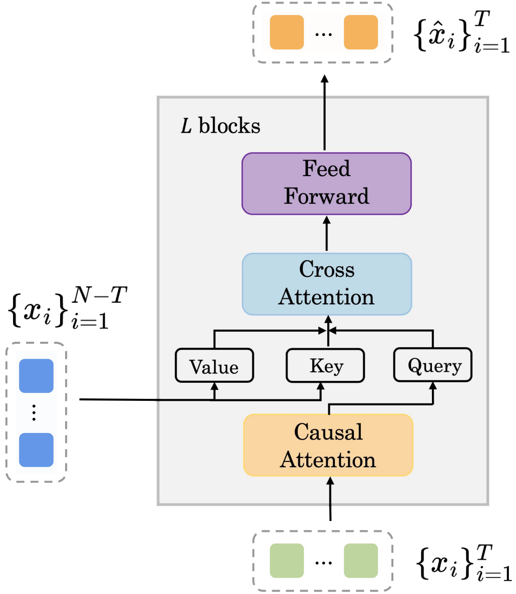


T-EF: **LaVIT** (Mar 2024, ICLR)

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



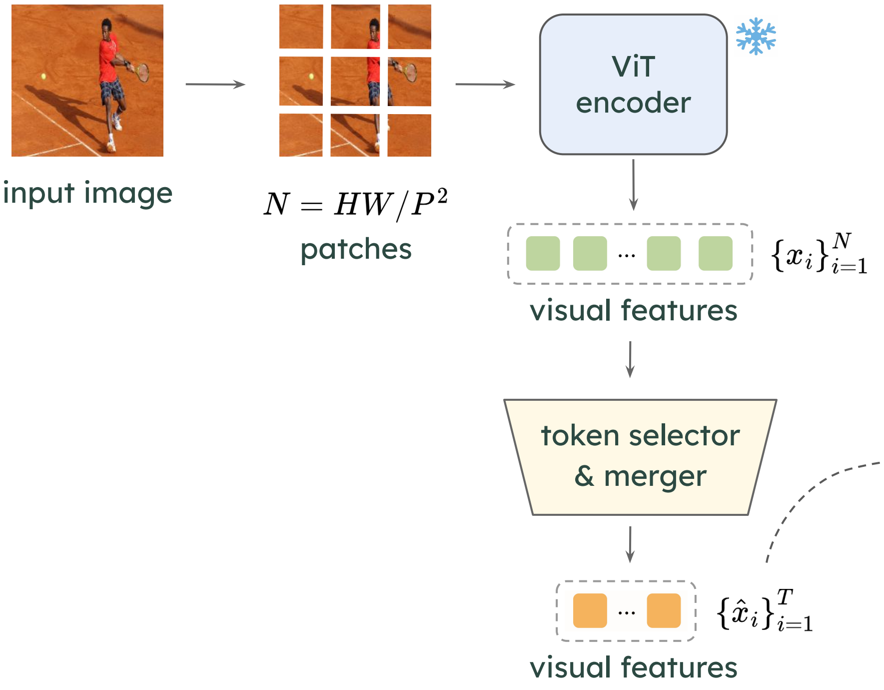
2 token merger

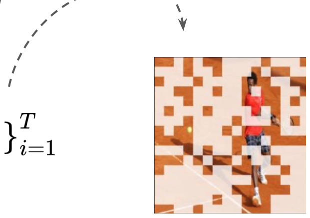


L = 12

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

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

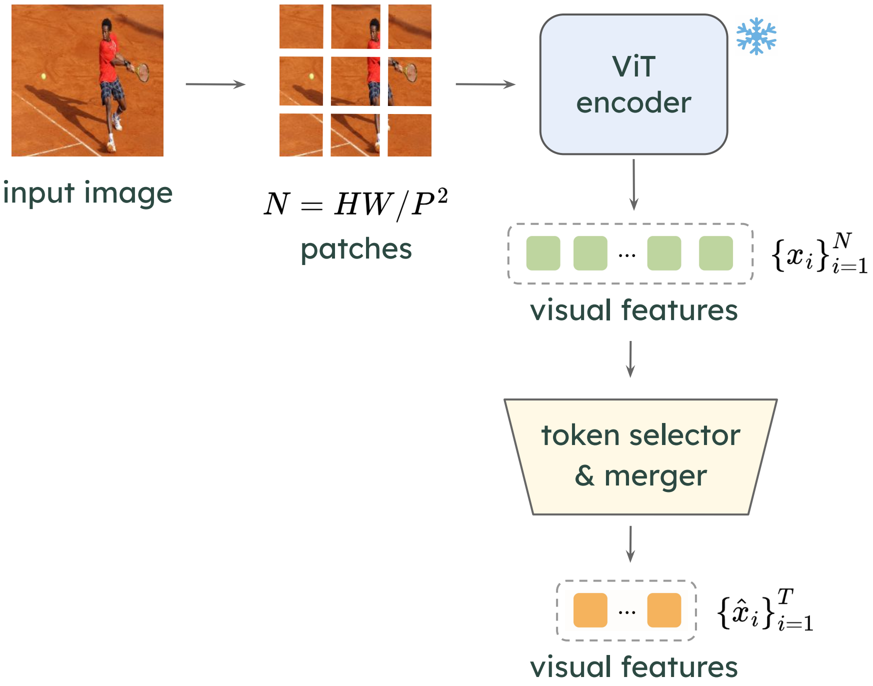




token selector and merger work together to **dynamically adjust the visual token sequence length** to accommodate images with different content complexity

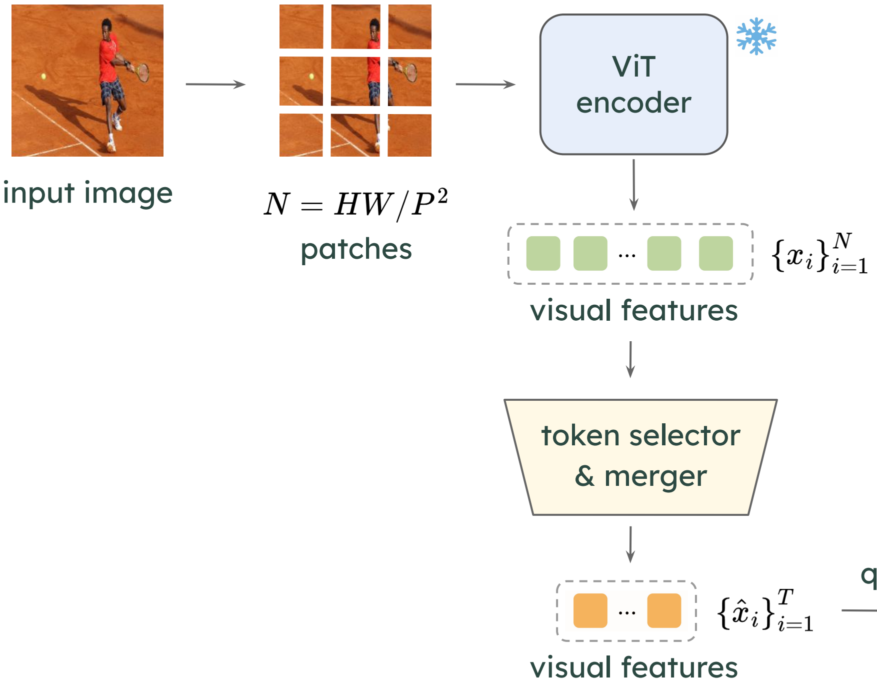
T-EF: **LaVIT** (Mar 2024, ICLR)

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

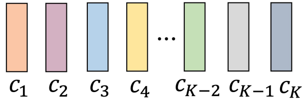


T-EF: **LaVIT** (Mar 2024, ICLR)

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



vector quantization



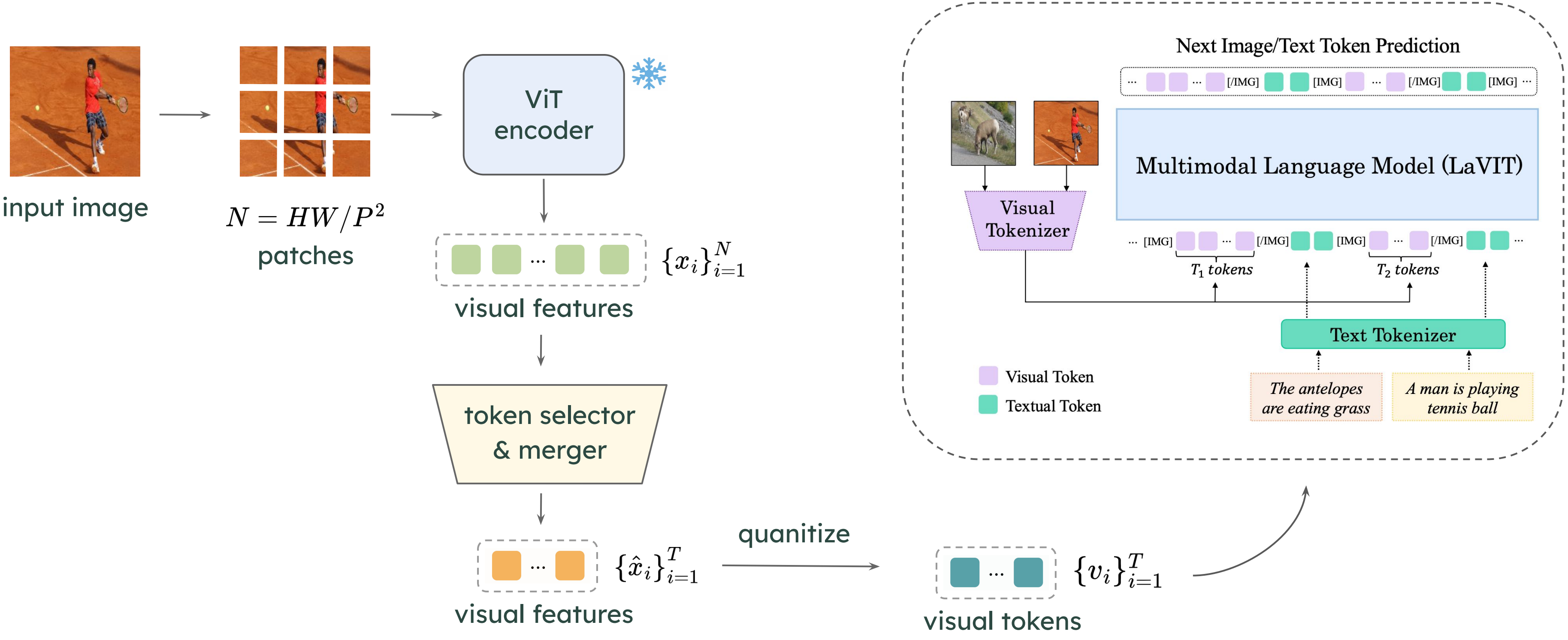
codebook
K = 16384

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

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

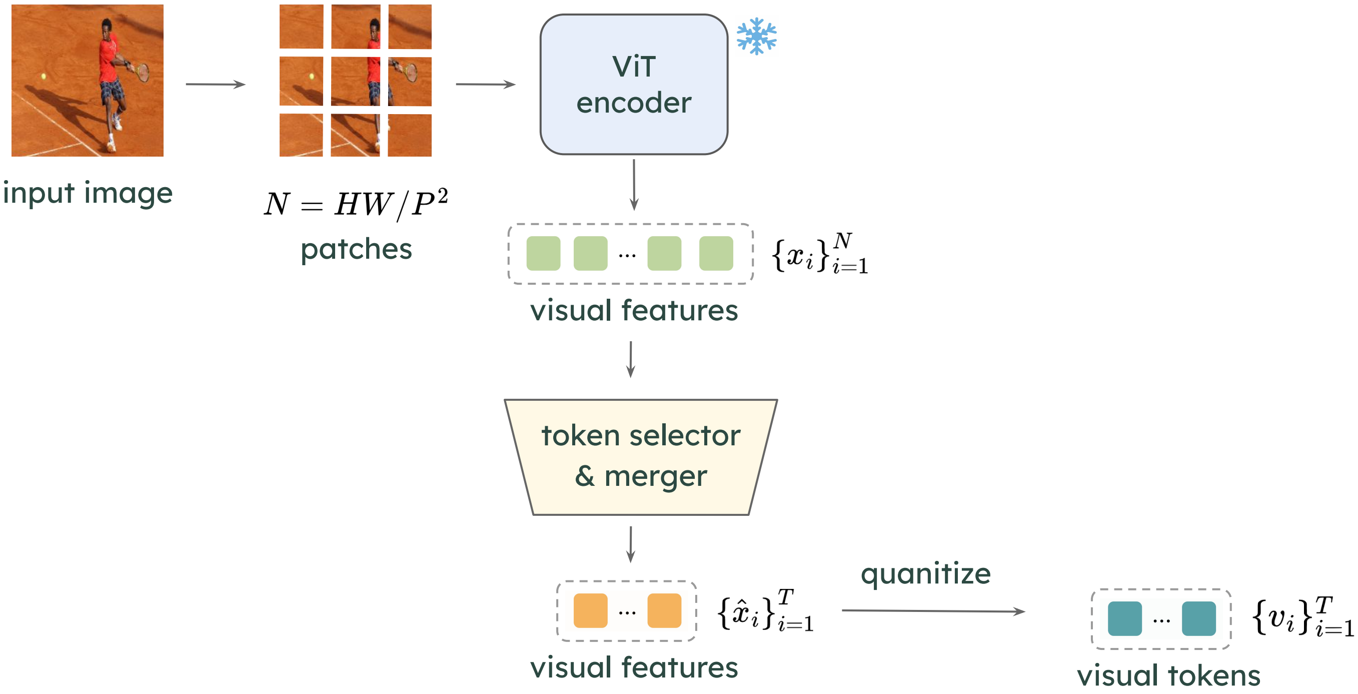
T-EF: **LaVIT** (Mar 2024, ICLR)

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



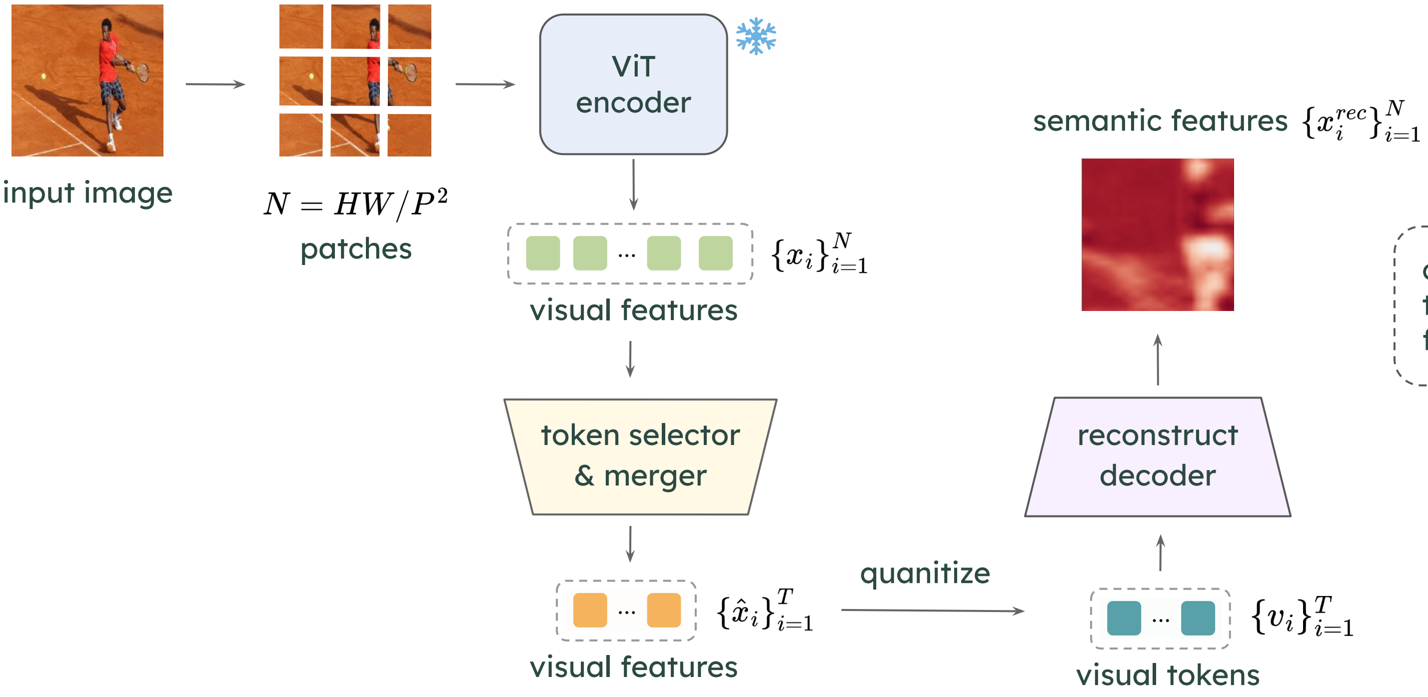
T-EF: **LaVIT** (Mar 2024, ICLR)

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



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

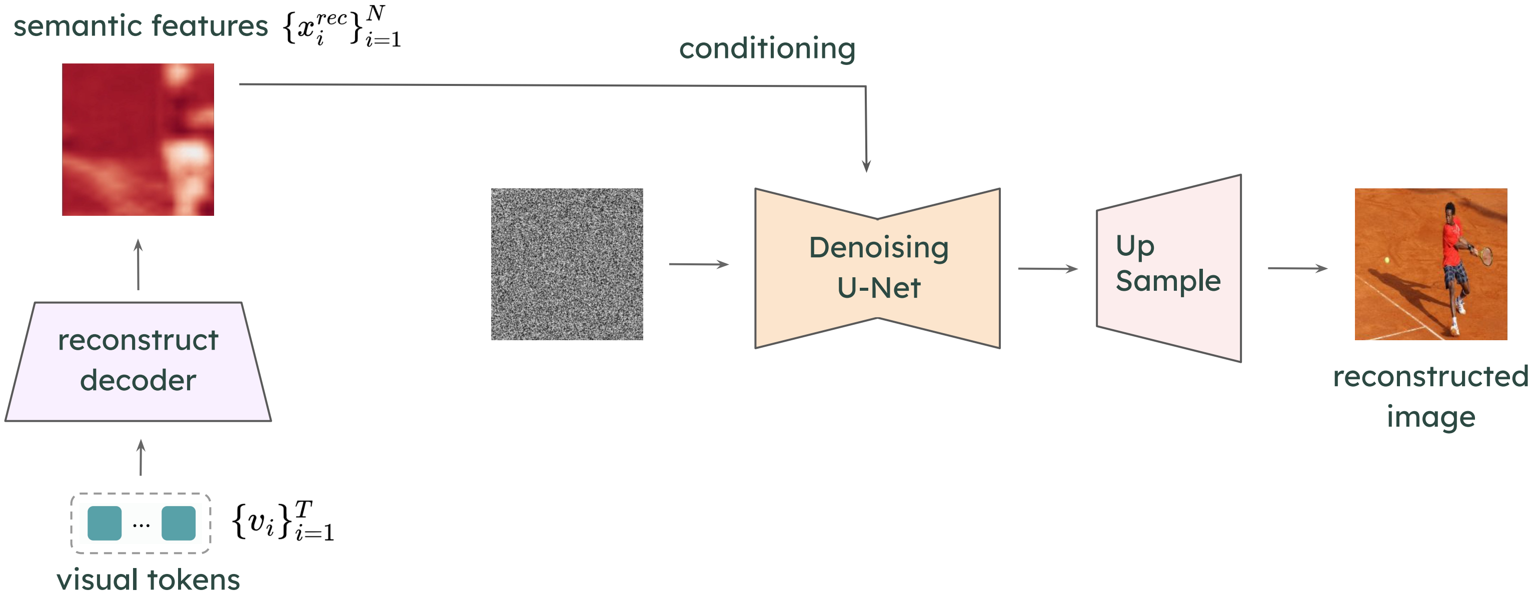
vision model = ViT-G/14 of EVA-CLIP, **language model** = LLaMA-7B



can train tokenizer decoder to **reconstruct** visual features!

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

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



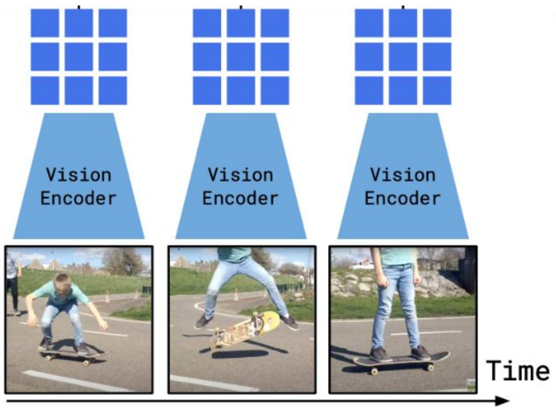
Part 2

Video Modality

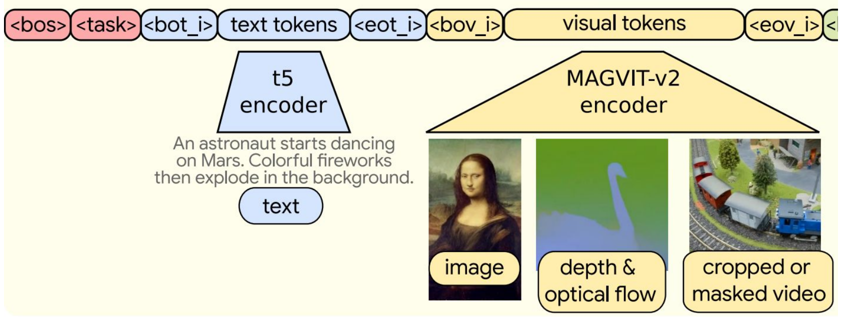
Examples of such models

General Approaches

1 uniformly downsample the original video into a **series of frames**



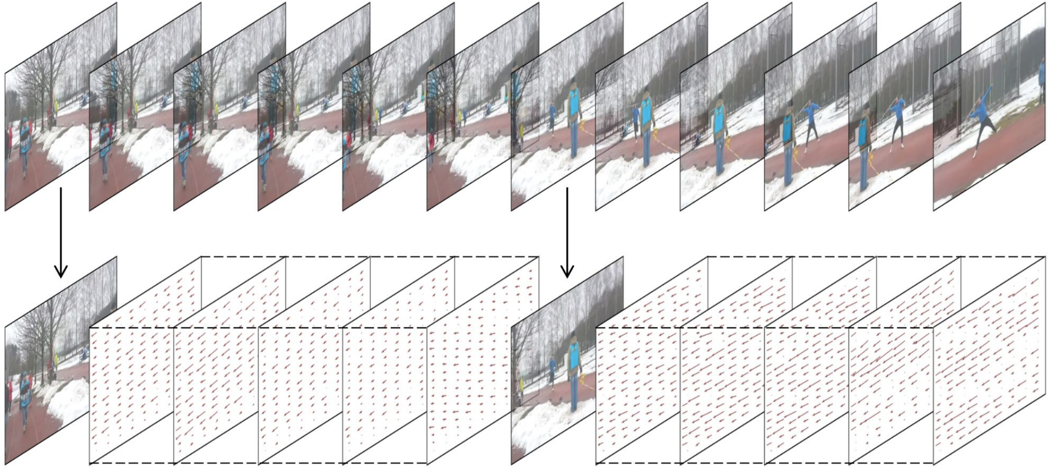
2 uniformly downsample the original video into a **series of frames**



[1] **Flamingo:** a Visual Language Model for Few-Shot Learning. Apr 2022. [link]
[2] **VideoPoet:** A Large Language Model for Zero-Shot Video Generation. Jun 2024. [link]

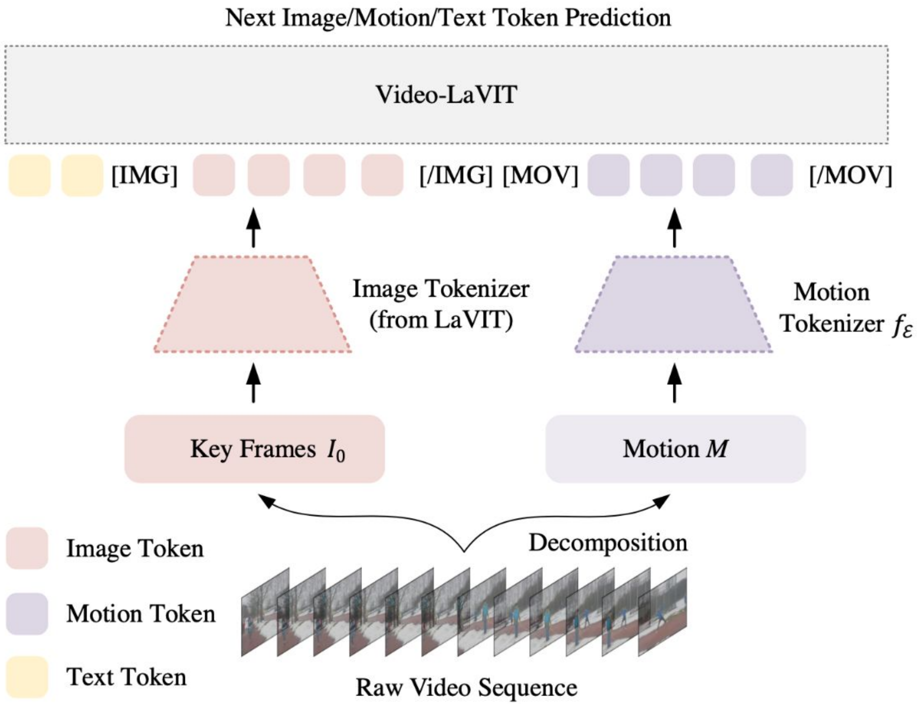
VideoLaVIT (Feb 2024) – ICML 2025

<u>**Key observation:**</u> most video parts have a high degree of temporal redundancy that may be described by **motion vectors**



most video frames are not needed and can be described by **motion vectors,** decreasing the number of utilized visual tokens

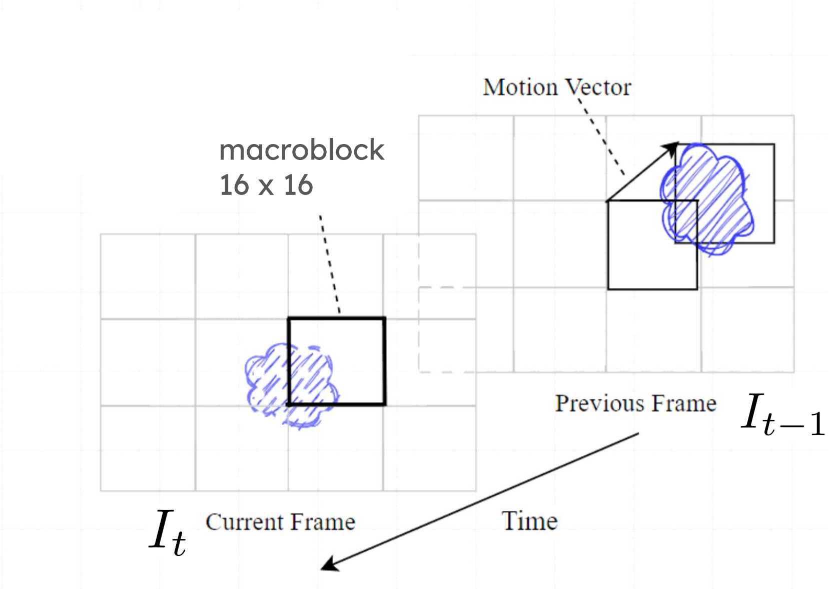
VideoLaVIT (Feb 2024) – ICML 2025



- introduce **novel video-tokenizer** and **video-detokenizer** to adapt visual features to LLM
- video tokens can be updated through the same **next-token-prediction** objective

VideoLaVIT (Feb 2024) – ICML 2025

Employ the **MPEG-4** (1991) to divide the image to **keyframes** (primary semantics) and **motion** (temporal evolvement)



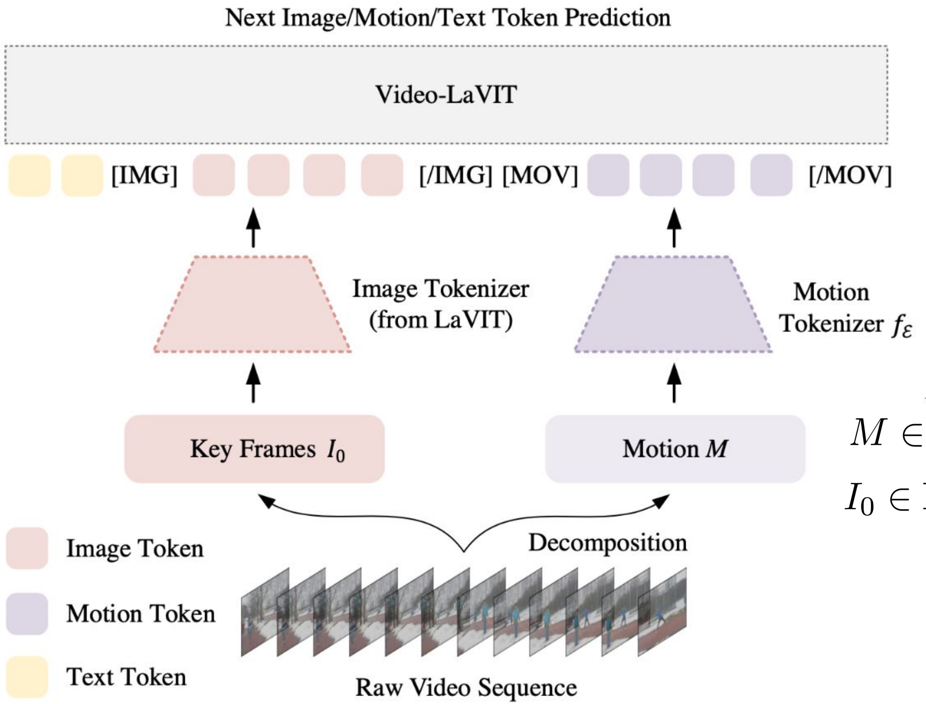
$$\vec{m}(p,q) = \arg\min_{i,j} \|I_t(p,q) - I_{t-1}(p-i, q-j)\|$$

(i, j) – coordinate offset between the center of 2 macroblocks

video clip $\longrightarrow$ $I_0 \in \mathbb{R}^{H \times W \times 3}$

$M \in \mathbb{R}^{T \times \frac{H}{16} \times \frac{W}{16} \times 2}$

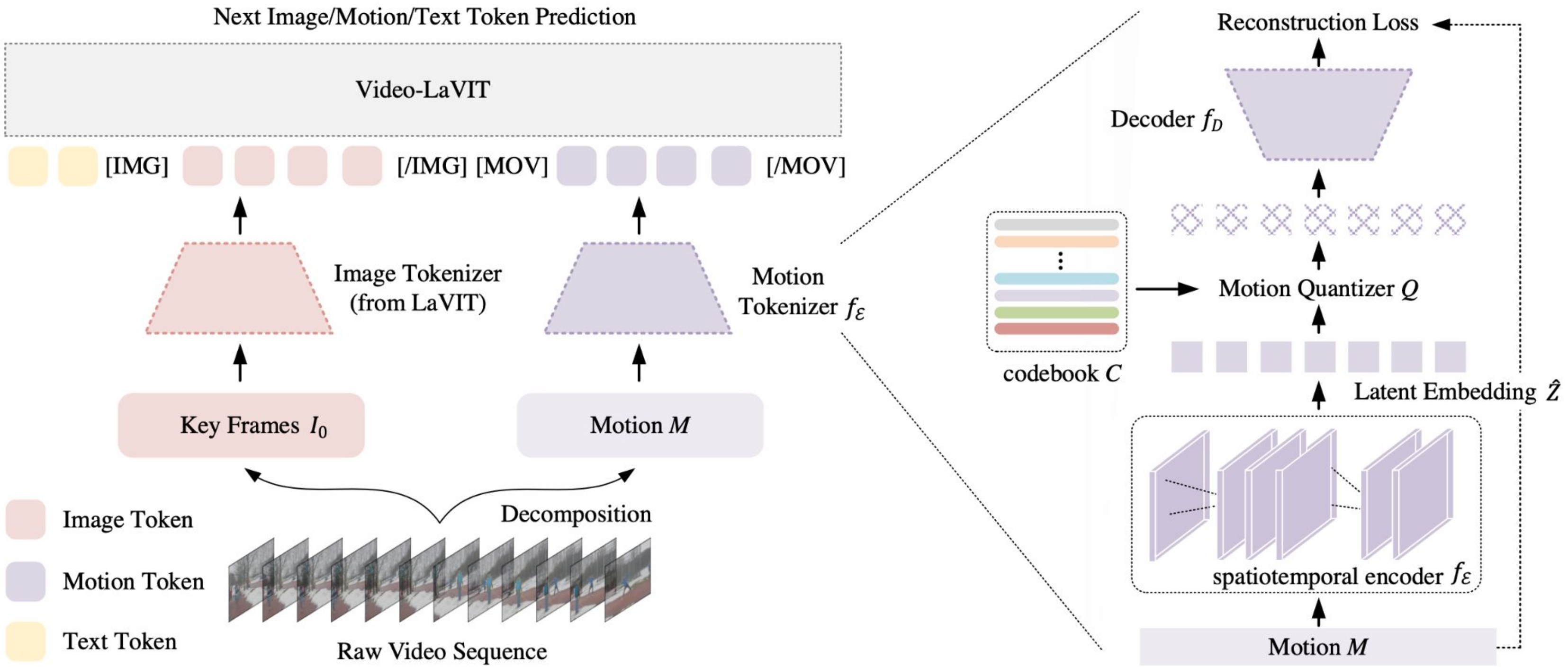
VideoLaVIT (Feb 2024) – ICML 2025



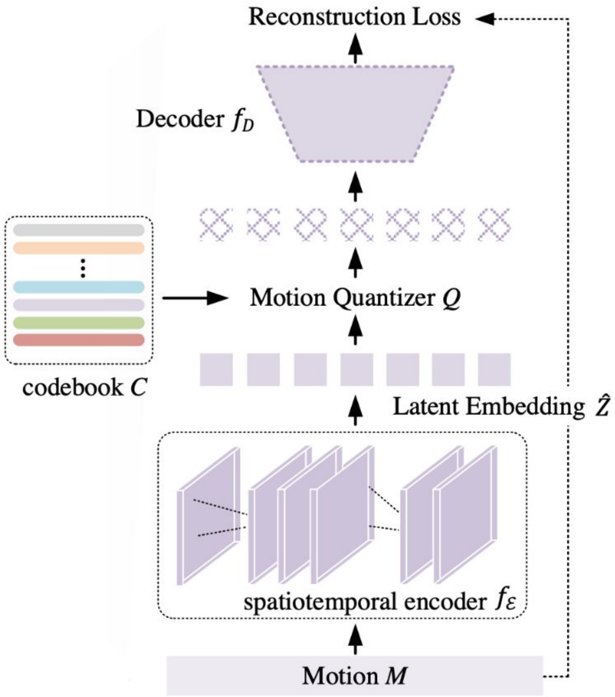
$$M \in \mathbb{R}^{T \times \frac{H}{16} \times \frac{W}{16} \times 2}$$

$$I_0 \in \mathbb{R}^{H \times W \times 3}$$

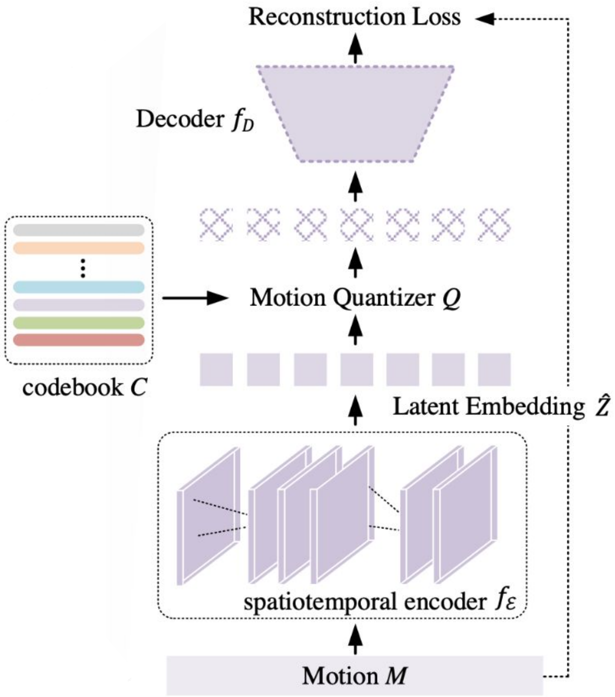
VideoLaVIT (Feb 2024) – ICML 2025



VideoLaVIT (Feb 2024) – ICML 2025



VideoLaVIT (Feb 2024) – ICML 2025



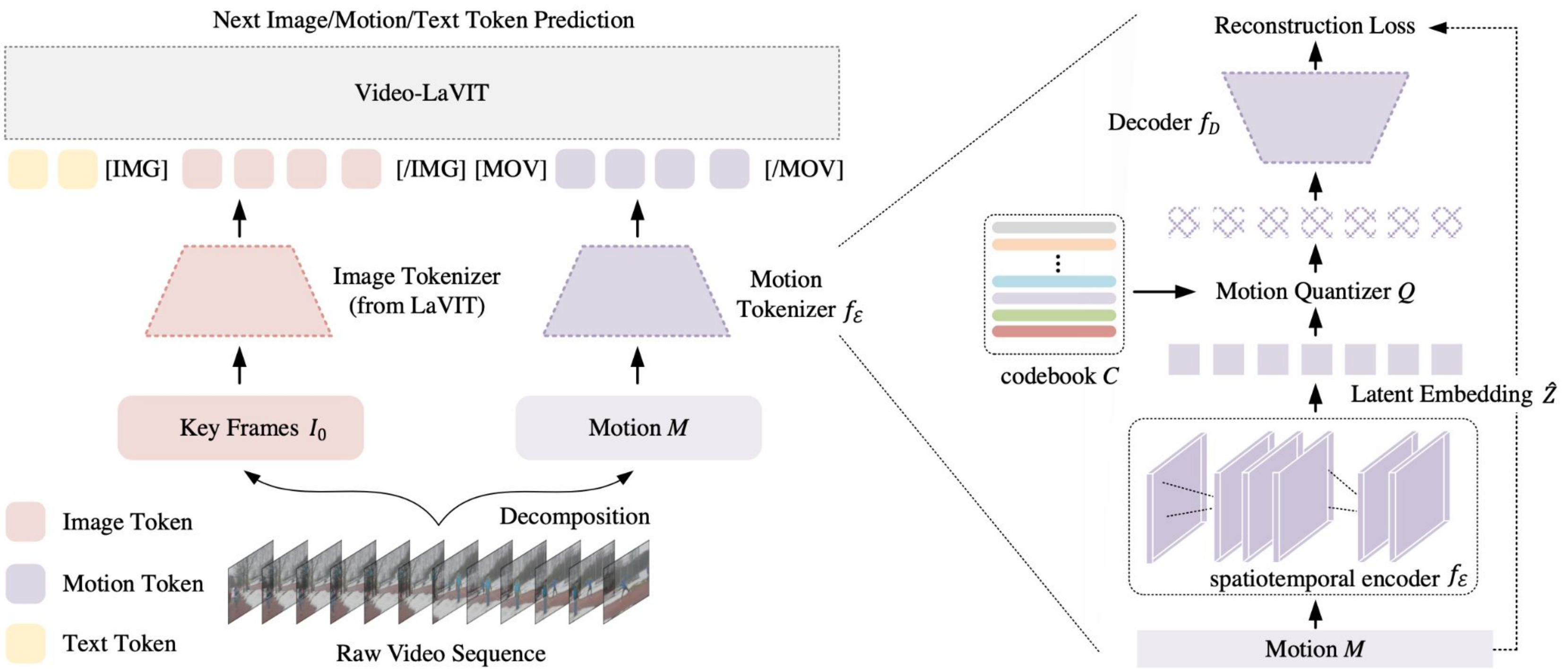
$\hat{Z} \in \mathbb{R}^{N \times d}$

$M \in \mathbb{R}^{T \times \frac{H}{16} \times \frac{W}{16} \times 2}$

Each embedding vector is then tokenized by a vector quantizer

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

VideoLaVIT (Feb 2024) – ICML 2025



Conclusions

1. Considered **classification of VLMs** based on feature fusion
2. Investigated models for **Deep Fusion**: OpenFlamingo, MoE-LLaVA
3. Investigated models for **Early Fusion**: Qwen-VL, LaVIT
4. Explored **Video-LaVIT** that processes video modality