



Language
Technologies
Institute

Carnegie
Mellon
University

Multimodal Machine Learning

Lecture 6.1: Visual-Language Models

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis.
Major revision co-developed with Paul Liang. Co-lecturers
included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Administrative Stuff

Upcoming Course Assignments

Second project assignment (due Sunday 10/4 at 8pm ET)

- Look at your data, before modeling it! 😊

Week 7 Reading assignment (due Friday 10/9 at 8pm ET)

- Two papers (with a focus on visual-language and omni models)

Fall break (Oct 12-16): No classes

Midterm project assignment (due Sunday 10/25 at 8pm ET)

- For each N-1 research idea, implement a first (baseline) model
- In-class presentations: Tuesday 10/27 and Thursday 10/29

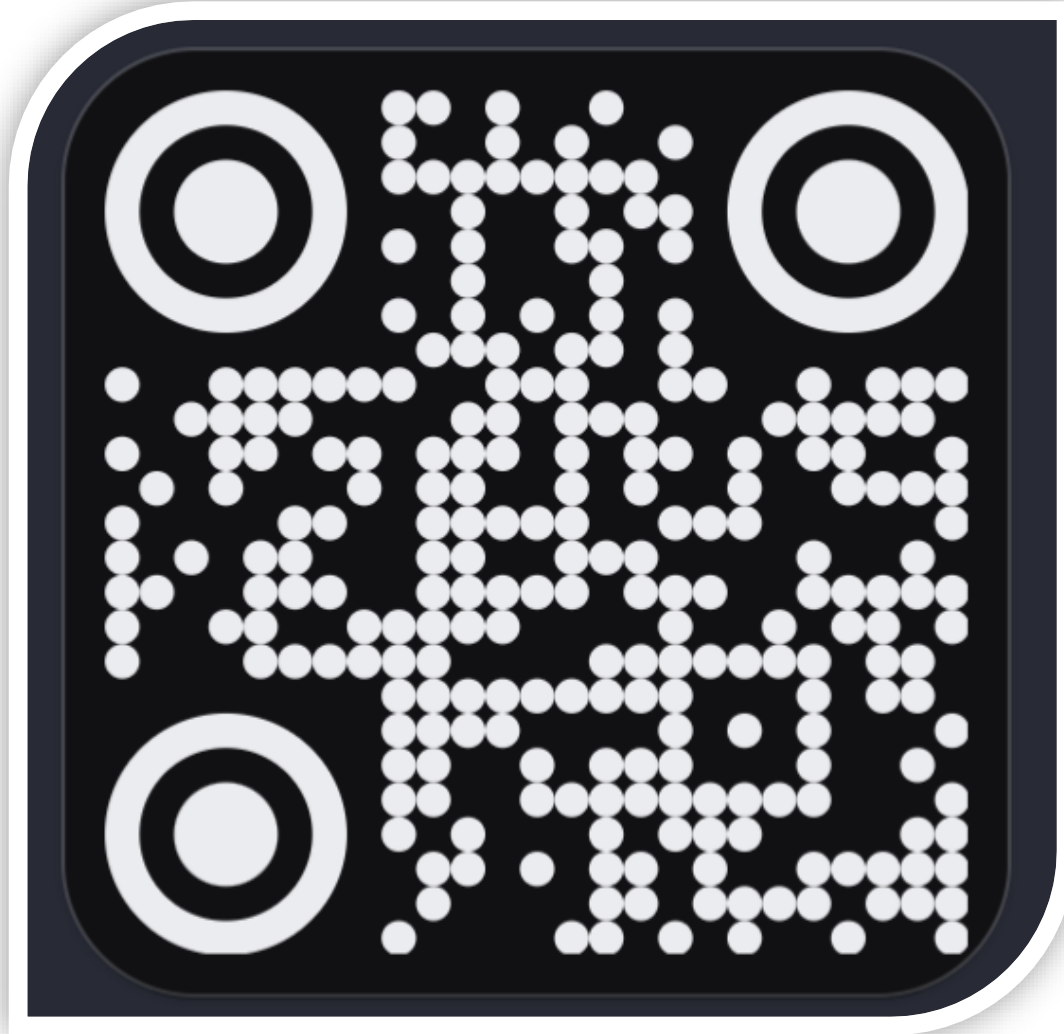
Guidelines for Course Projects

- Each research idea could be a workshop paper, if successful
 - Research ideas should not be simple variants of each others
- Second assignment: look at your data
 - Replace AI with “human intelligence” 😊
- Midterm report: Implementation of your first “baseline” model
 - Implement simple version of your idea with SOTA components.
 - Not just running a previous model on your data.
 - Analyze results. Run some ablations. Understand failure cases.
 - This is Version 1.0 of your research idea.

Guidelines for Course Projects

- Final report: Improve your initial model
 - You have 4 weeks to do research and try some alternatives
 - This is Version 1.1 (and maybe V1.2) of your research idea
 - Reserve the last week before the deadline for results analyses
 - Grades are primarily based on analyses and discussion (not the improved accuracy!)
 - Do not train models until the last minute
 - Research ideas must be multimodal (not just the input data!)
 - You need to show that you are putting in practice concepts learned in class
 - The novel component of your research idea should be multimodal
 - Final report with more than one unimodal/generic research idea will be penalized

Attendance using Poll Everywhere



- 1 Scan QR Code
or
<https://pe.app/mmm1>
- 2 Enter your AndrewID
(as your “name tag”)
- 3 Click “Share location”

✗ *Do not try to sign-in*



Language
Technologies
Institute

Carnegie
Mellon
University

Multimodal Machine Learning

Lecture 6.1: Visual-Language Models

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis.
Major revision co-developed with Paul Liang. Co-lecturers
included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Objectives of today's class

Three generations of visual-language models:

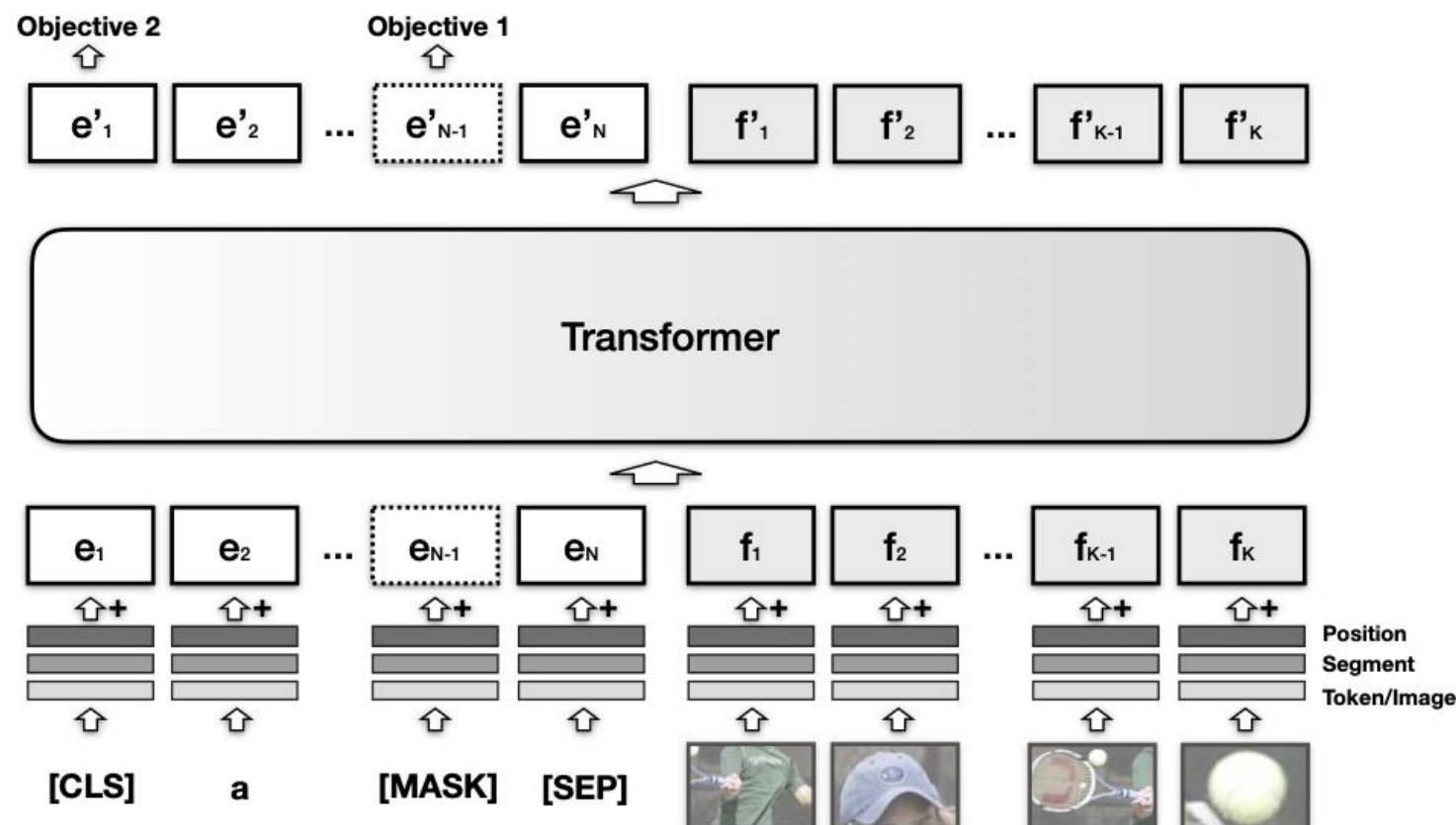
1. Object-based visual encoders
 - Fully-connected transformer (e.g., VisualBERT) – *from last week*
 - Cross-model attention (e.g., ViLBERT, LXMERT)
2. Patch-based visual encoders
 - Based on Visual Transformer (e.g., ViLT)
 - Based on Masked Autoencoder (e.g., VLC)
3. Auto-regressive multimodal LLMs
 - Prefix tuning (e.g., Frozen)
 - Cross-attention layers (e.g., Flamingo)
 - Linear projection (e.g., LiMBeR, Fromage)
 - Visual instruction tuning (e.g., LLaVA, LLaVA-NeXT)

1st VLM Generation: Object-based Visual Encoders

Option 1: Fully-connected Transformer (e.g., VisualBERT)

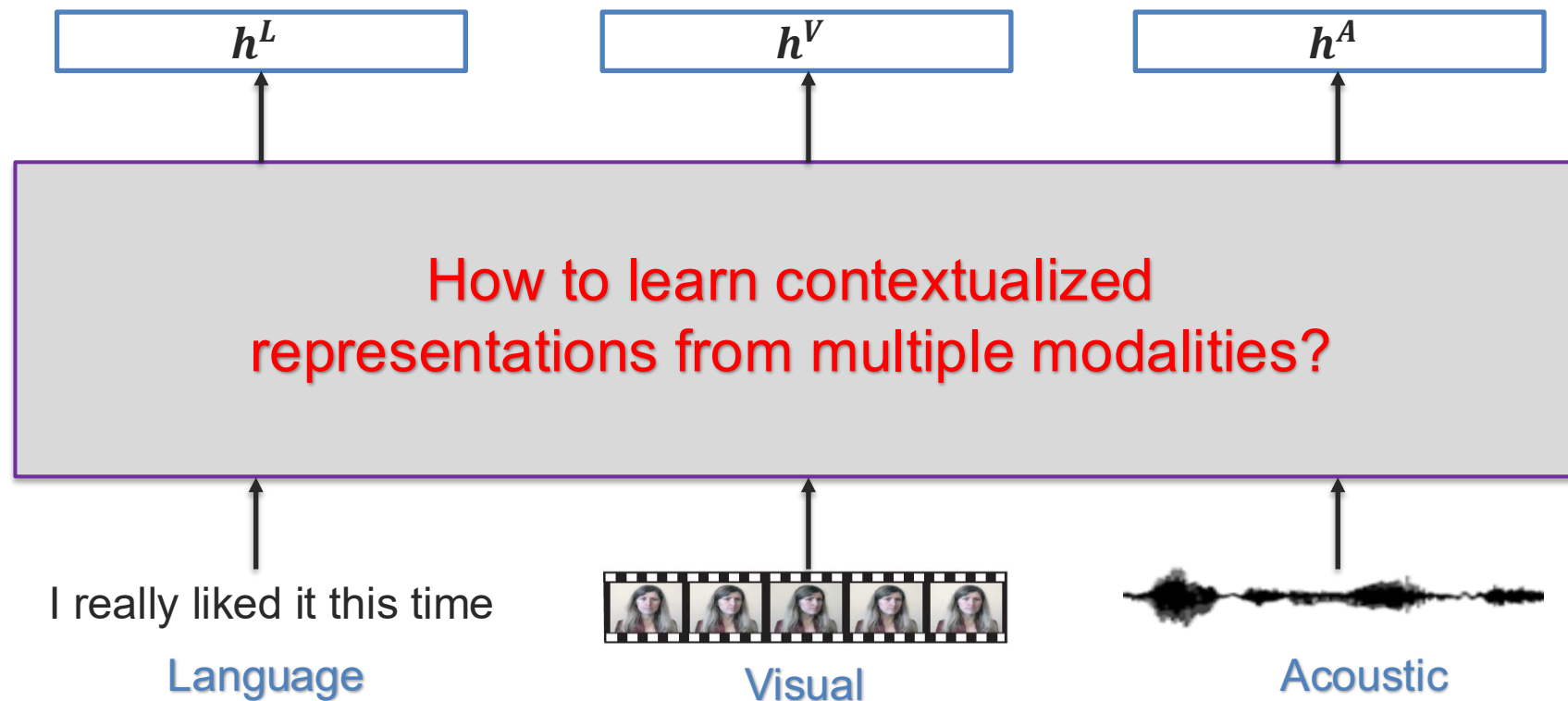


A person hits a ball with a tennis racket



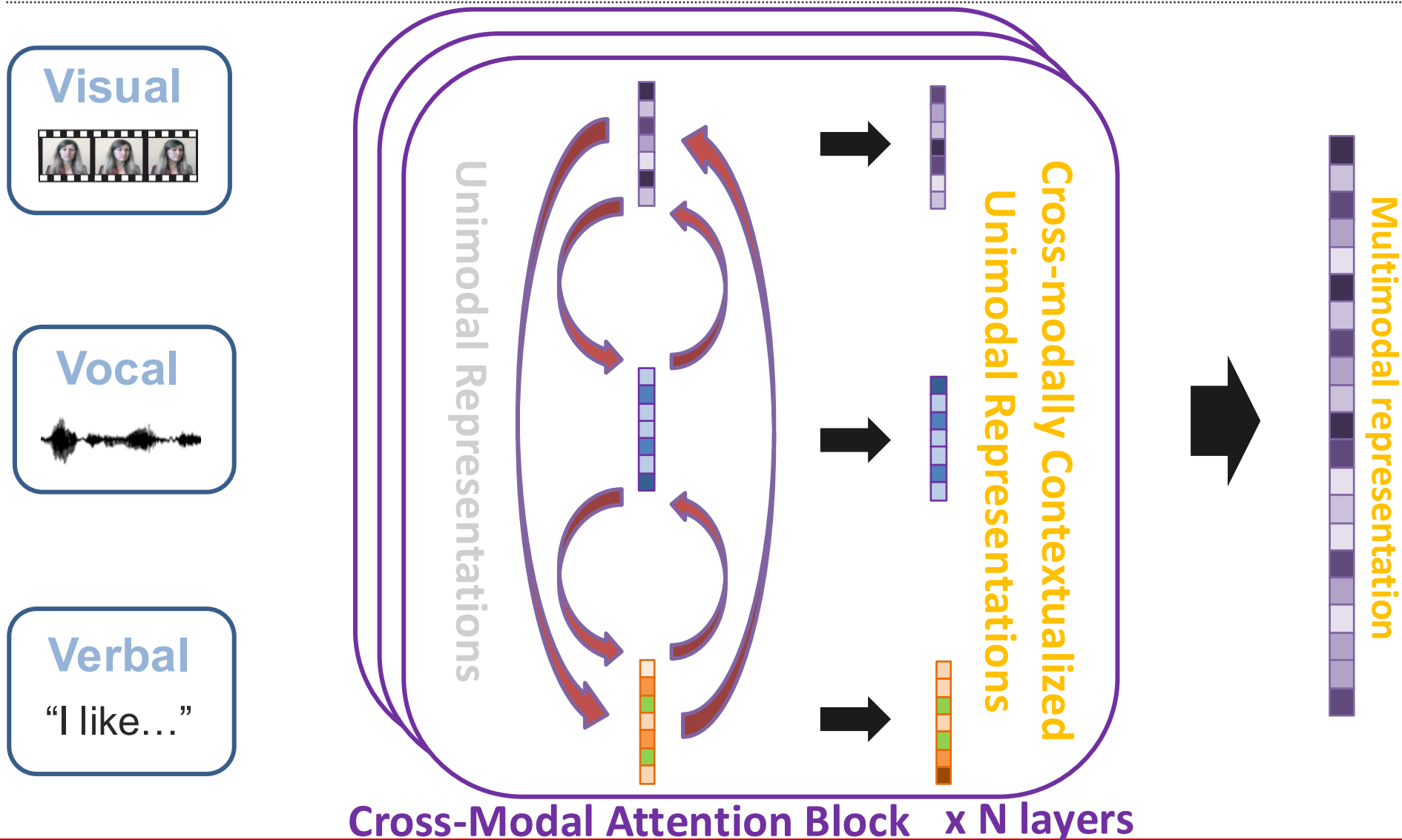
Li, Liunian Harold, et al. "Visualbert: A simple and performant baseline for vision and language." *ACL* (2020).

Option 2: Focusing on Cross-modal Interactions

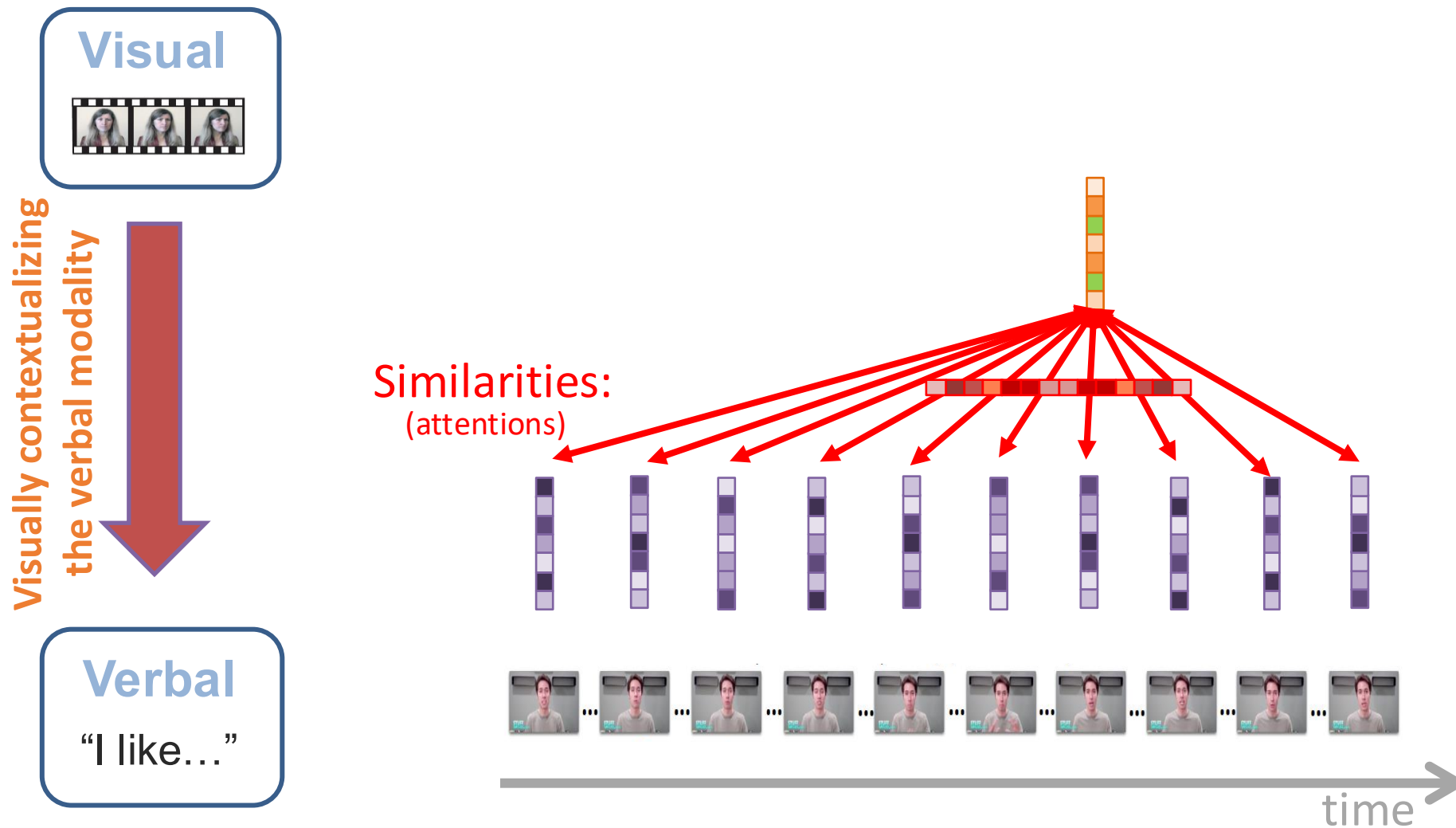


If more than 2 modalities: Look at pairwise interactions between modalities

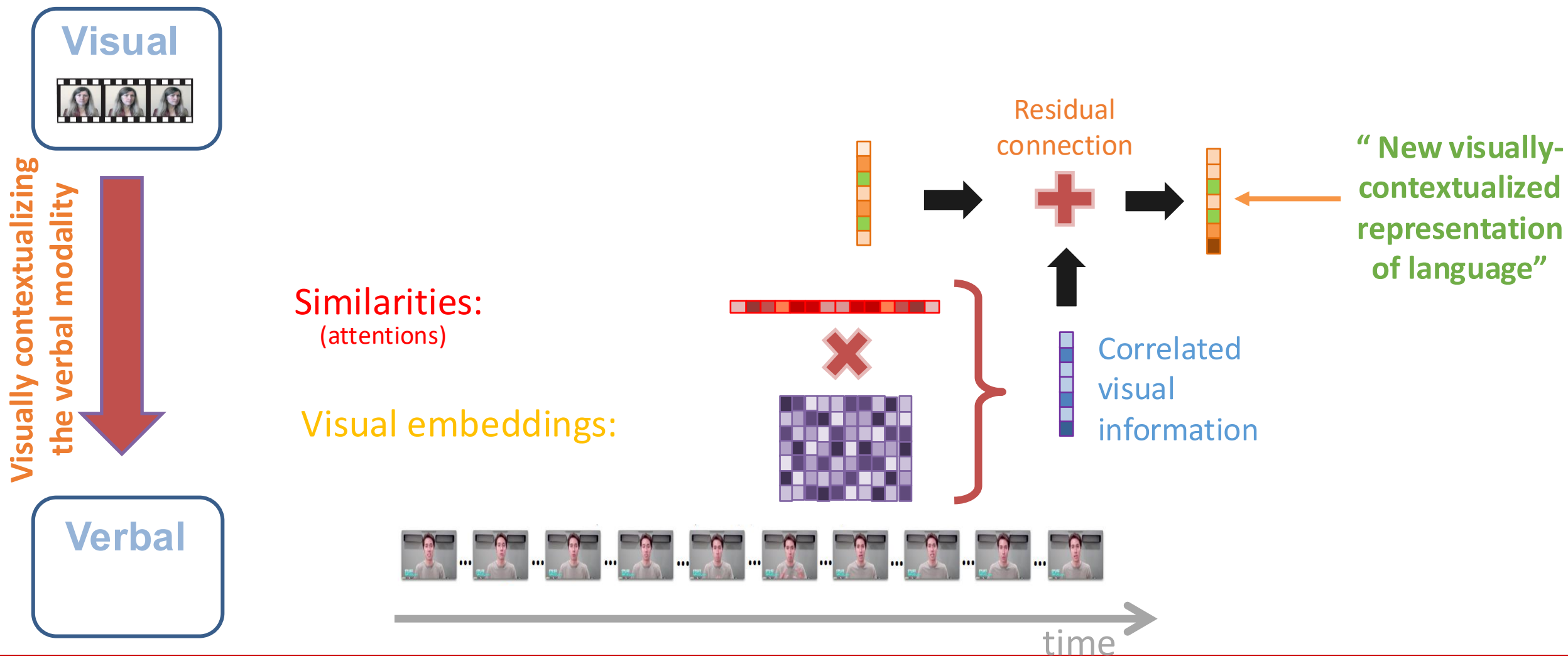
Cross-Modal Transformer – with Pairwise Cross-Modal



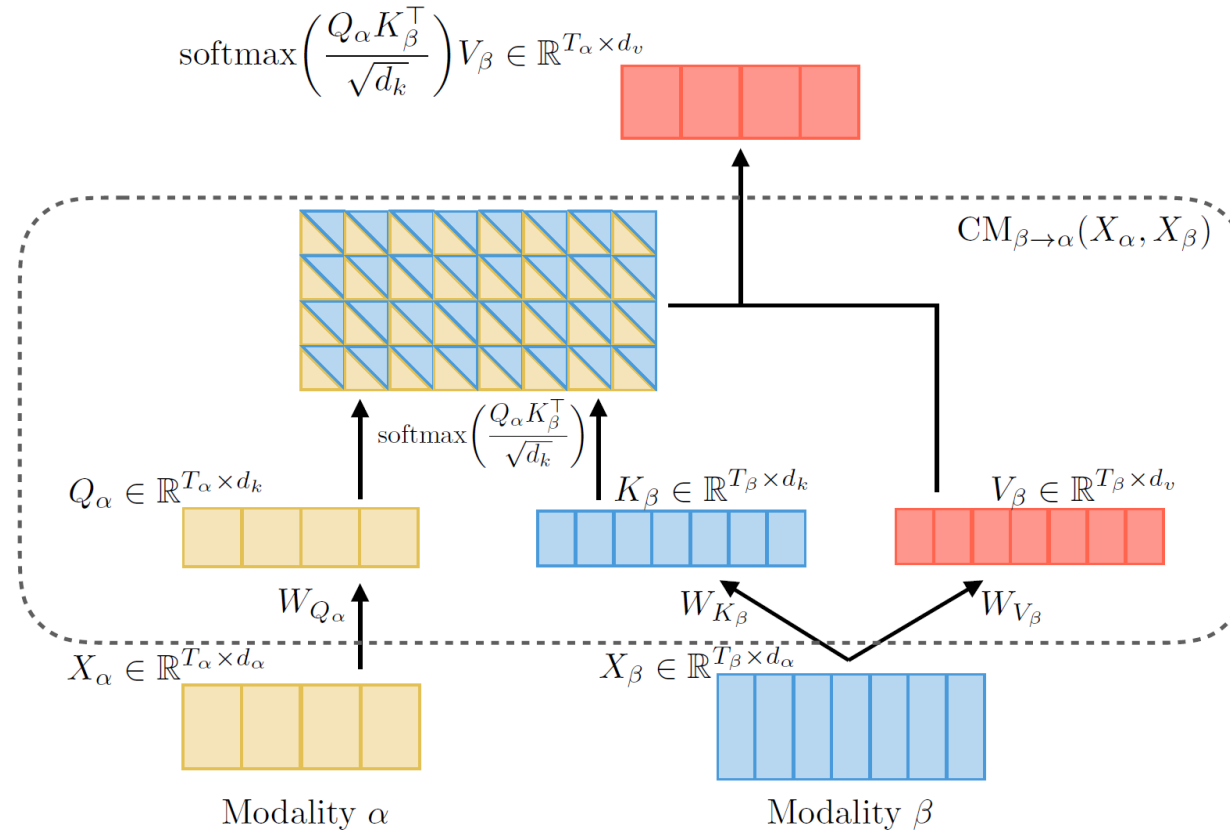
Cross-Modal Transformer Module ($V \rightarrow L$)



Cross-Modal Transformer Module ($V \rightarrow L$)

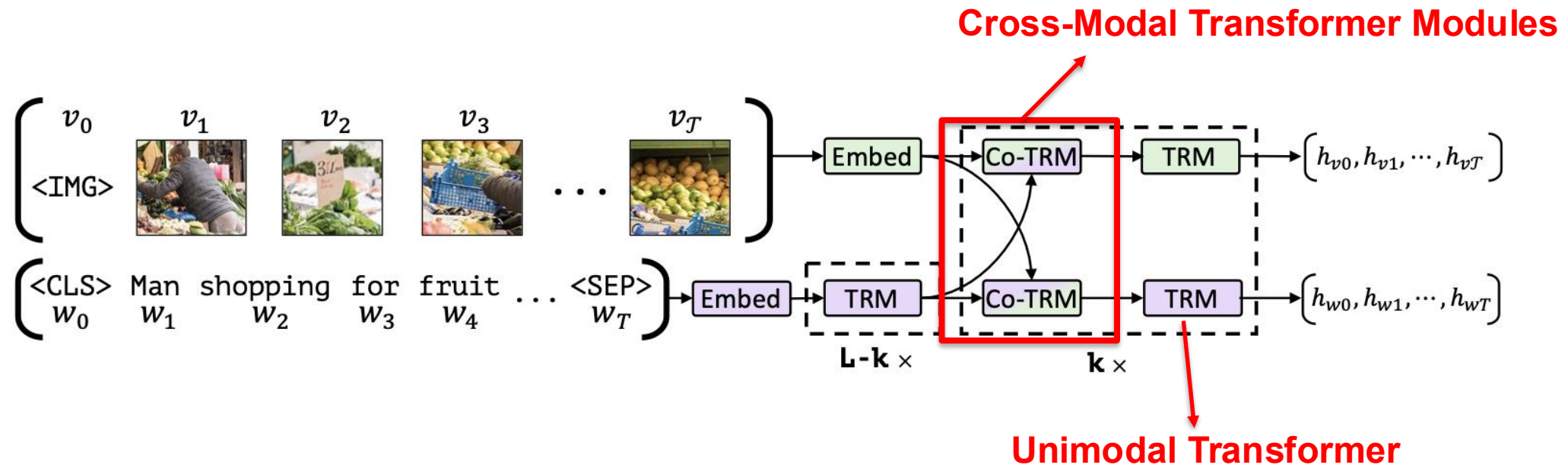


Cross-Modal Transformer Module ($\beta \rightarrow \alpha$)



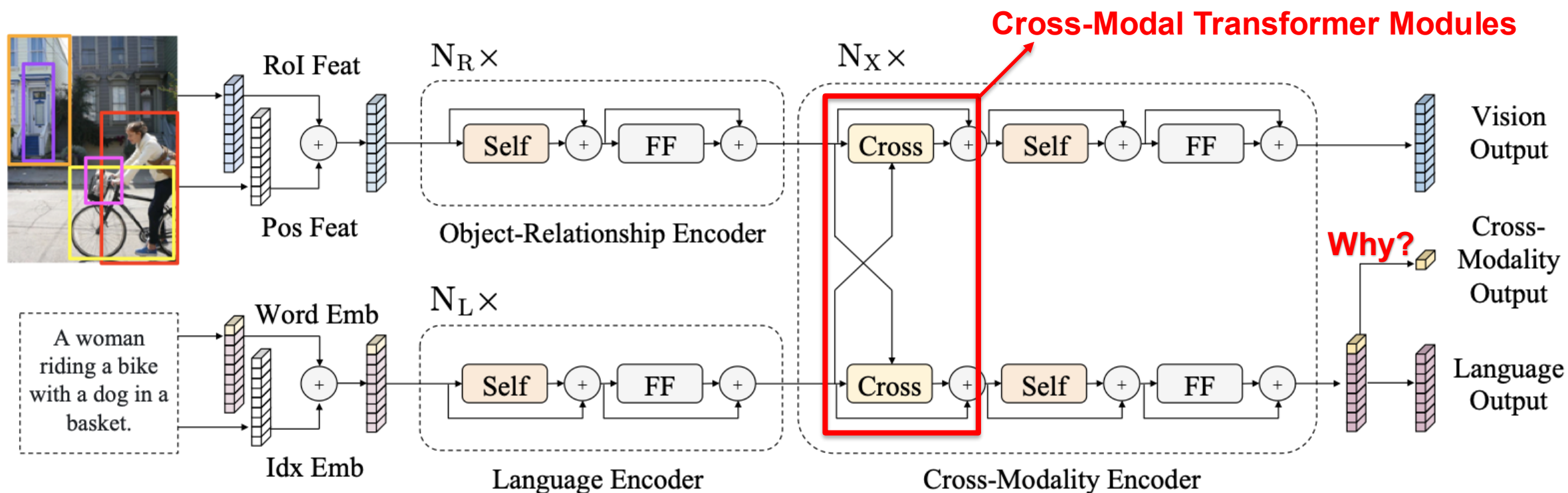
Tsai et al., Multimodal Transformer for Unaligned Multimodal Language Sequences, ACL 2019

Example: ViLBERT is using Cross-Modal Transformer Module



Lu, Jiasen, et al. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." *arXiv* (August 6, 2019).

Another Example: LXMERT



Tan, Hao, and Mohit Bansal. "Lxmert: Learning cross-modality encoder representations from transformers." *arXiv* (August 20, 2019).

2nd VLM Generation: Patch-based Visual Encoders

Visual-Language Models: From Objects to Patches...

of Visual Tokens

~ 30 object / image

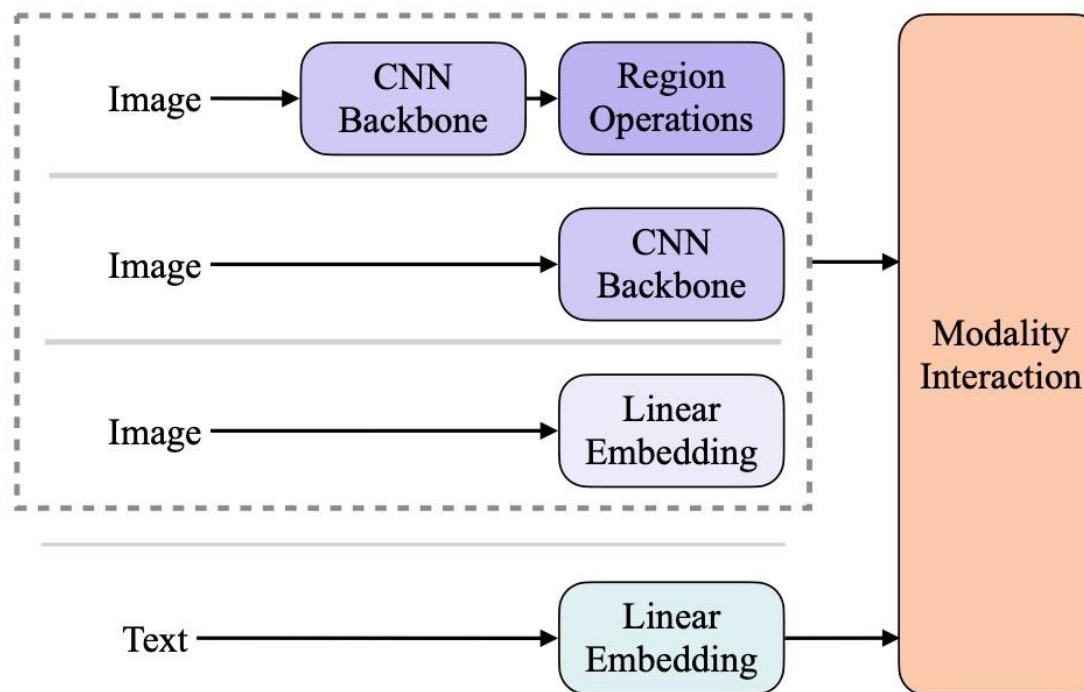
Region Feature
(ViLBERT, UNITER, ...)

Grid Feature
(Pixel-BERT)

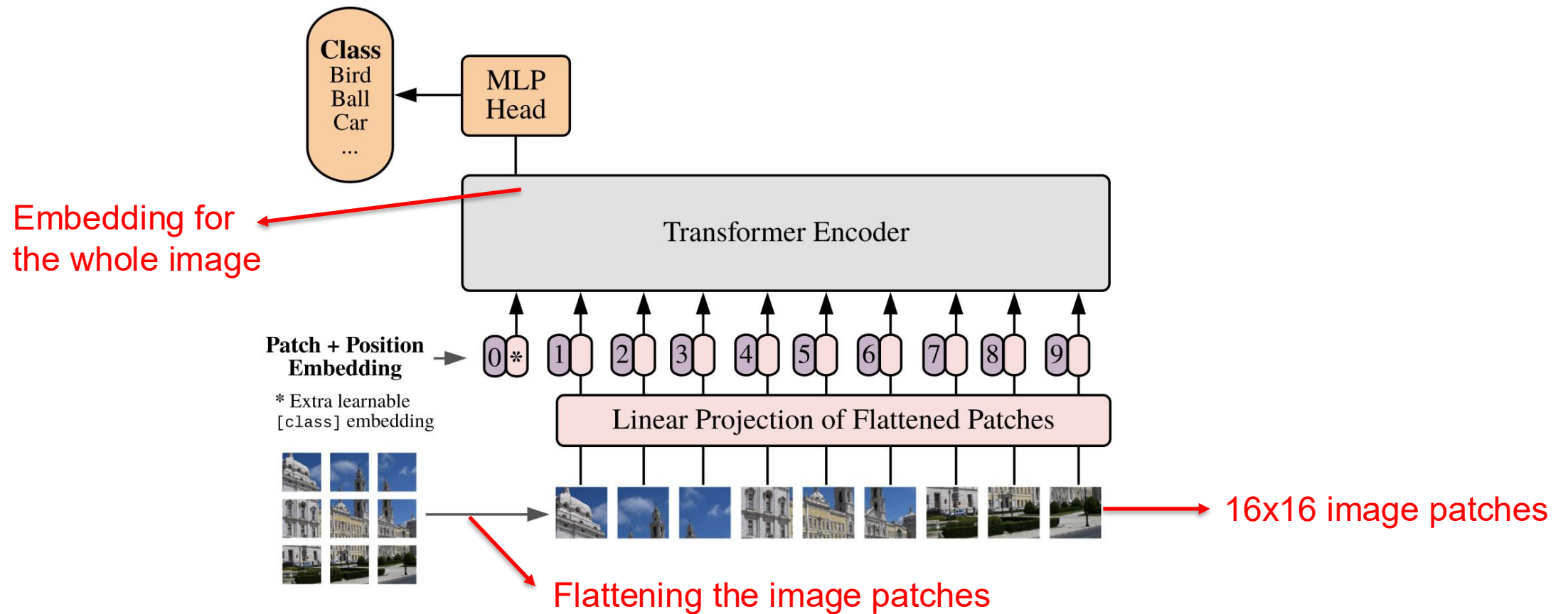
~ 500 patches / image

Patch Projection

Visual Embedding Schema

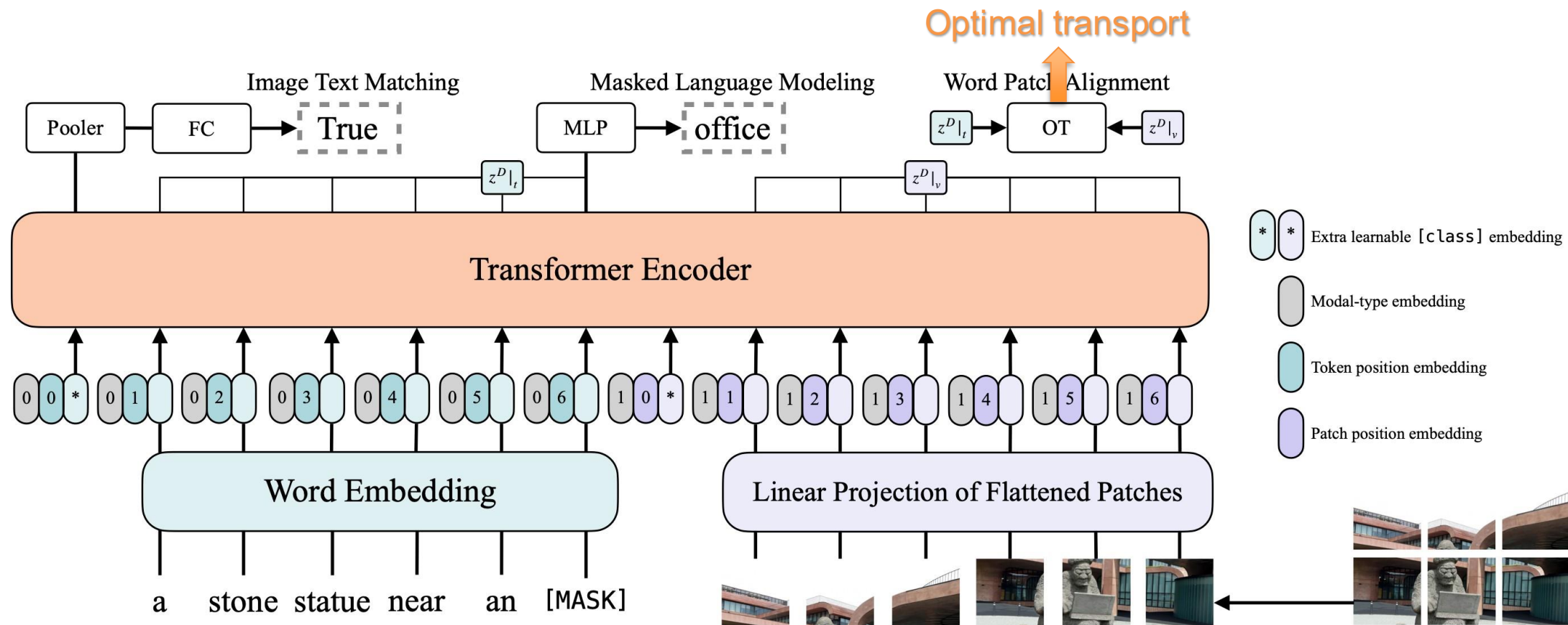


Reminder: Vision Transformer (ViT)



Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv* (2020).

Visual-and-Language Transformer (ViLT) ($\approx$ BERT + ViT)



Visual-and-Language Transformer (ViLT)

Example of alignment between modalities:



a display of **flowers** growing out and over the retaining **wall** in front of **cottages** on a **cloudy** day.



flowers



wall



cottages



cloudy



a room with a **rug**, a **chair**, a **painting**, and a **plant**.



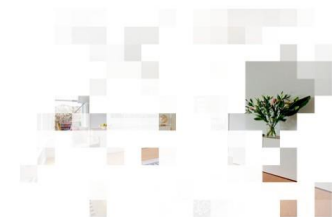
rug



chair

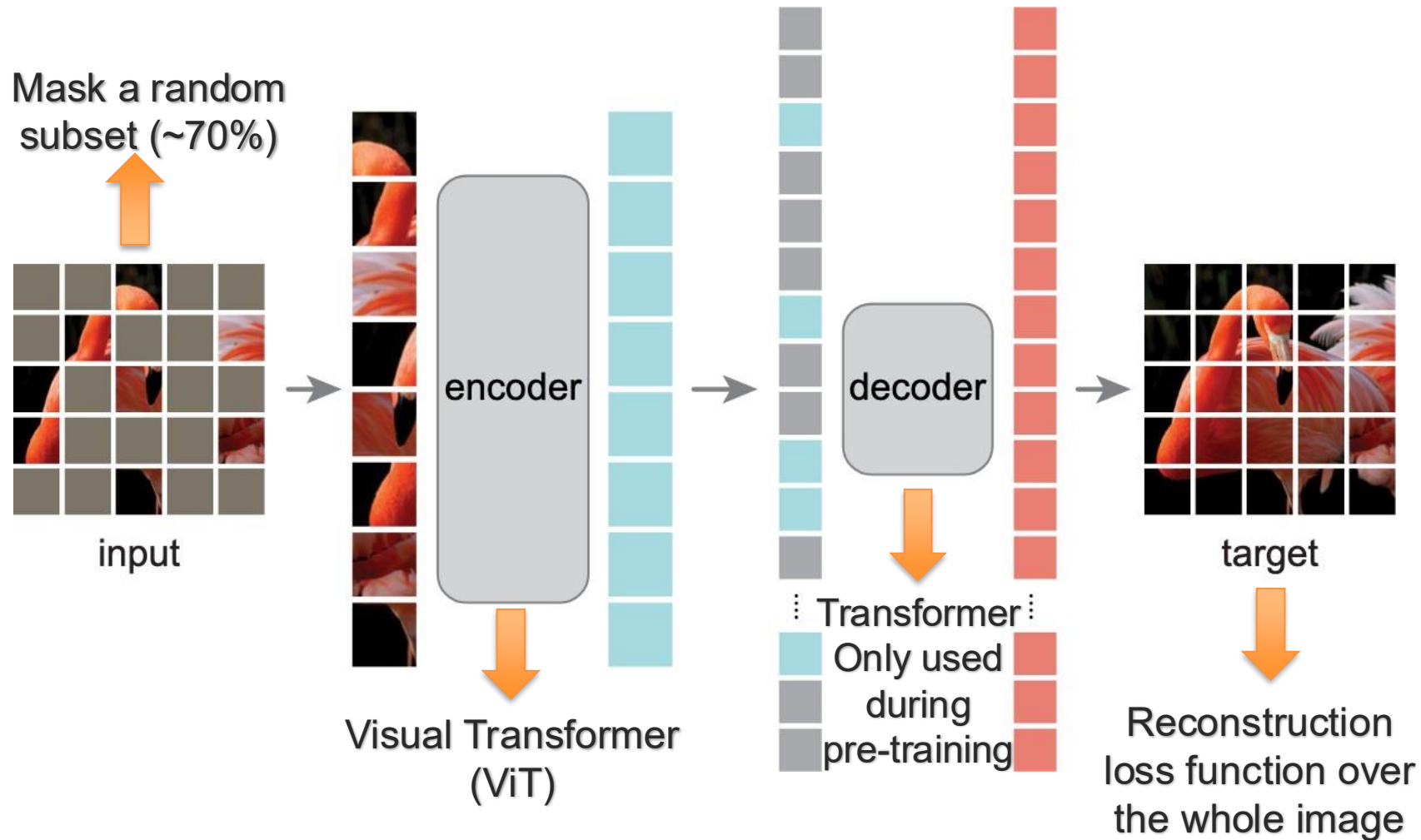


painting



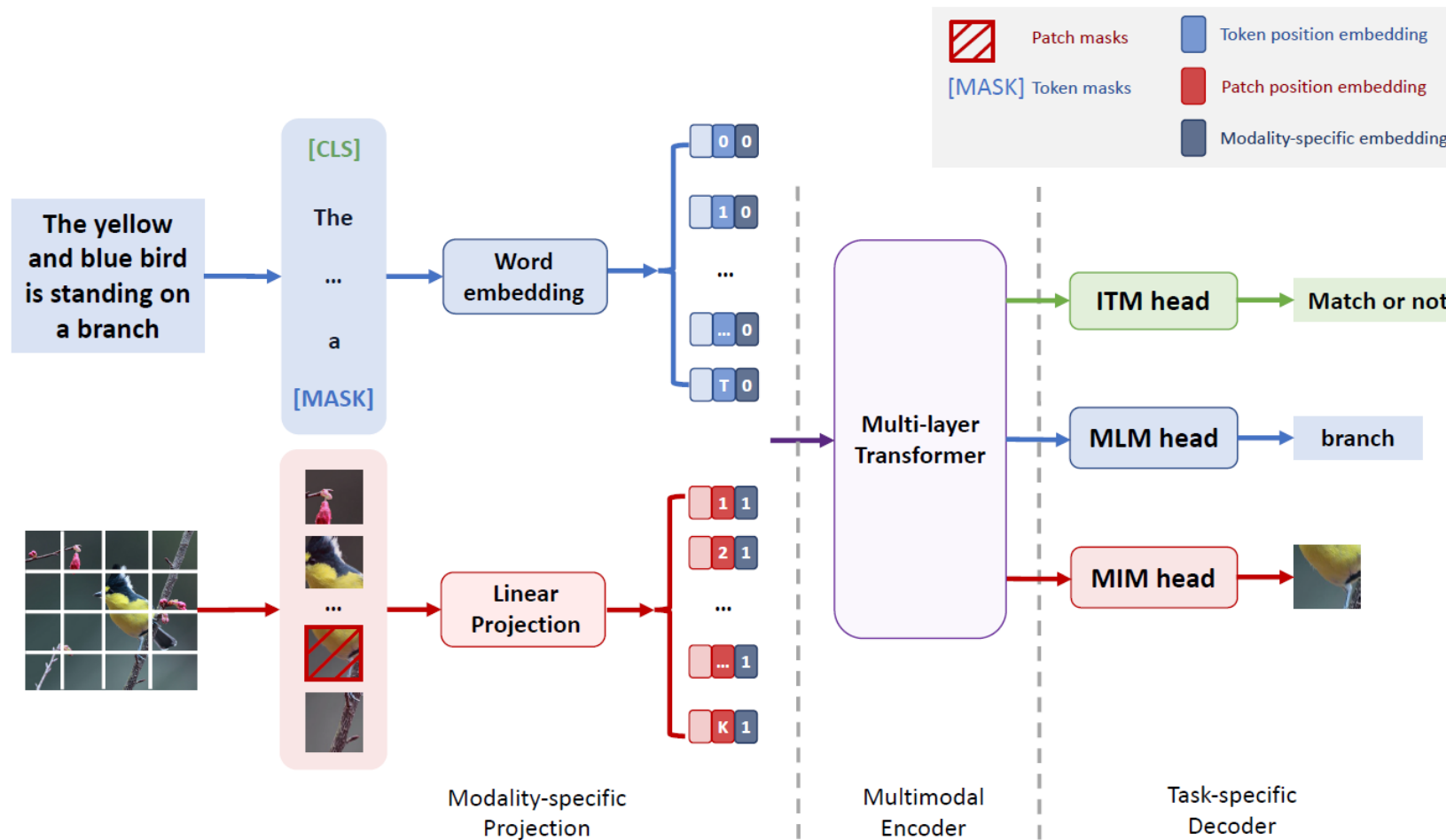
plant

Reminder: Masked Auto-Encoder (MAE)



He et al., Masked Autoencoders Are Scalable Vision Learners, CVPR 2022

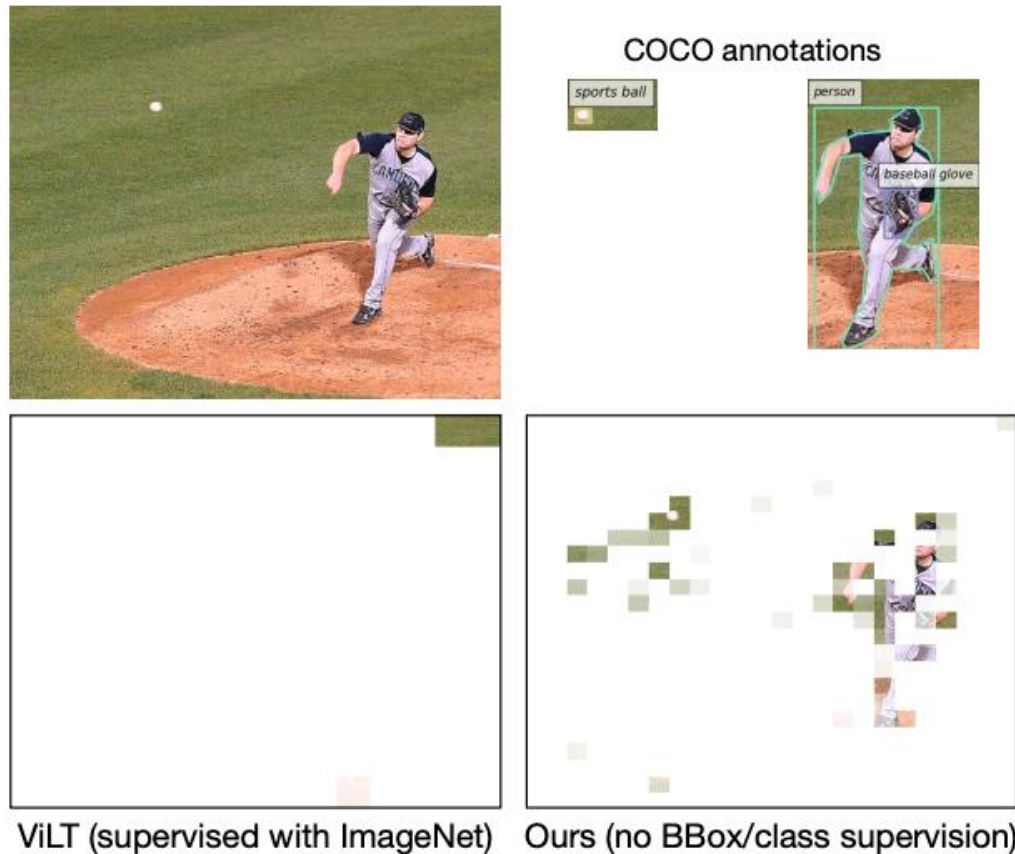
Vision-Language from Captions (VLC)



Vision-Language from Captions (VLC)

What are we learning?

A pitcher at a baseball game who has just **thrown** the ball.



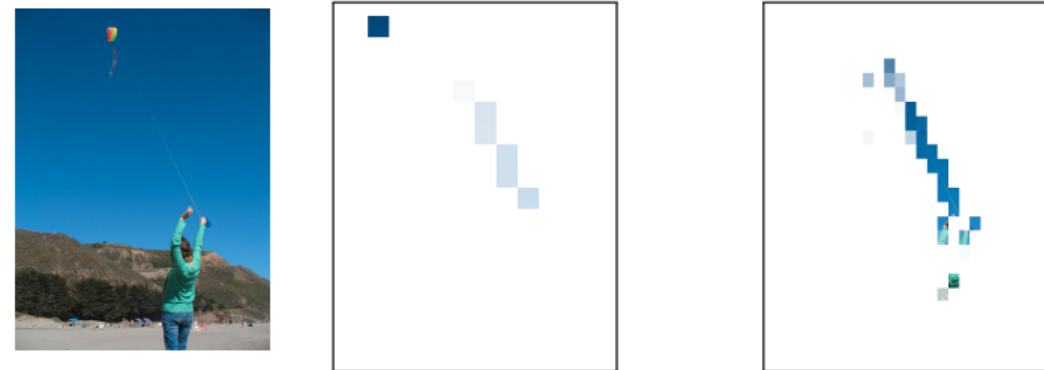
Caption with **focus**

Original Image

ViLT

VLC

A person on a beach holding a kite **string** and a kite is in the air

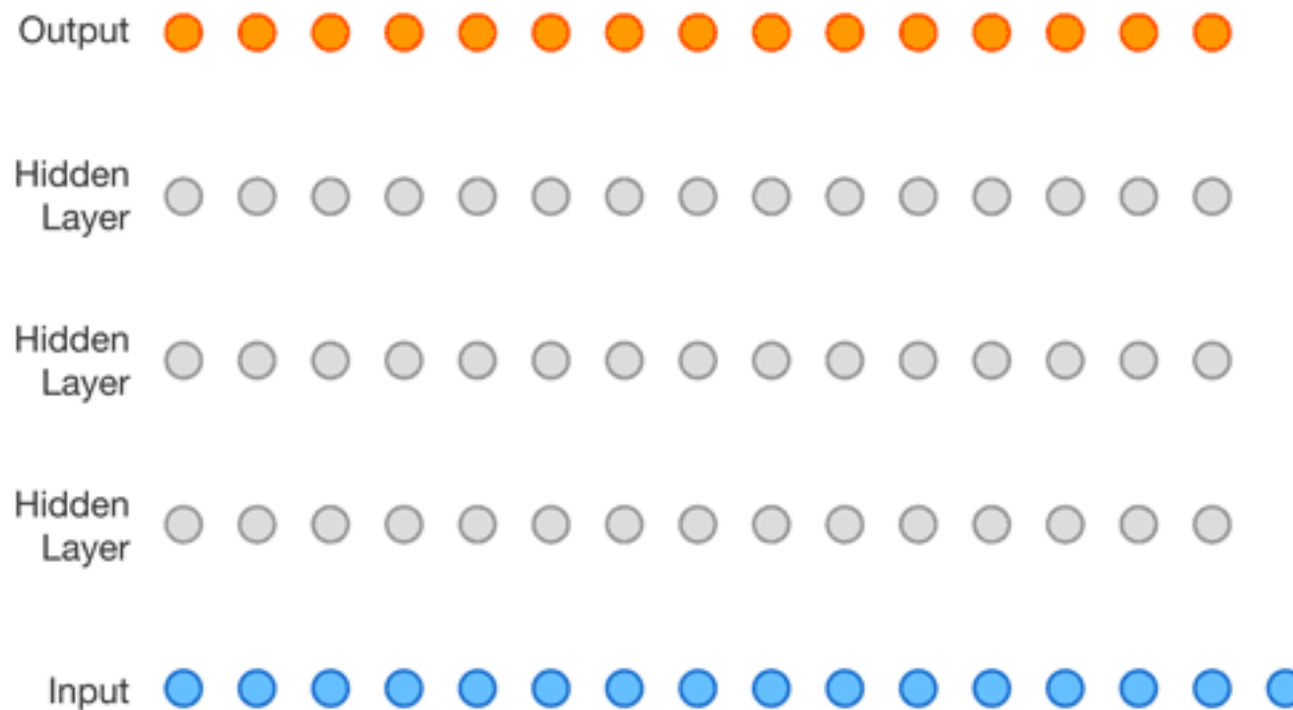


3rd VLM Generation : Auto-regressive Multimodal LLMs

Autoregressive Models

Autoregressive audio generation models

$$p(\mathbf{x}) = \prod_{t=1}^T p(x_t | x_1, \dots, x_{t-1})$$

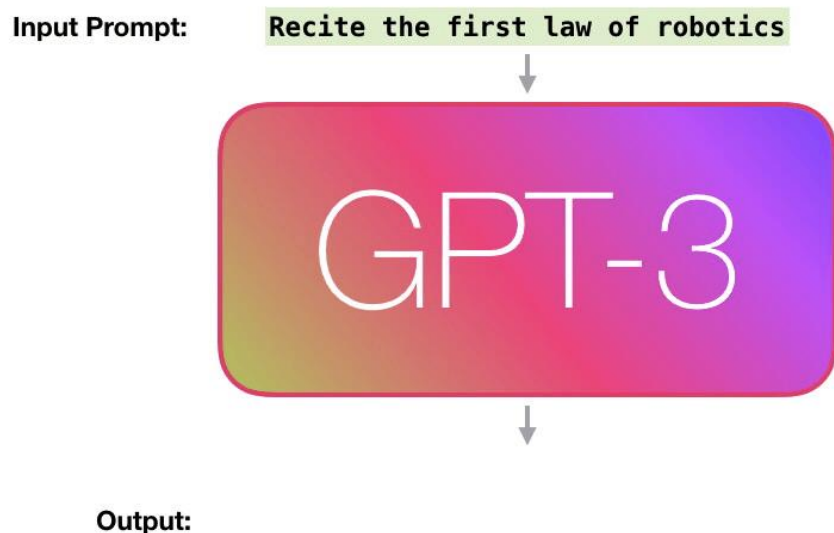


[van den Oord et al., WaveNet: A Generative Model for Raw Audio. ICML 2016]

Autoregressive Large Language Models

Autoregressive language models

$$p(\mathbf{x}) = \prod_{t=1}^T p(x_t \mid x_1, \dots, x_{t-1})$$



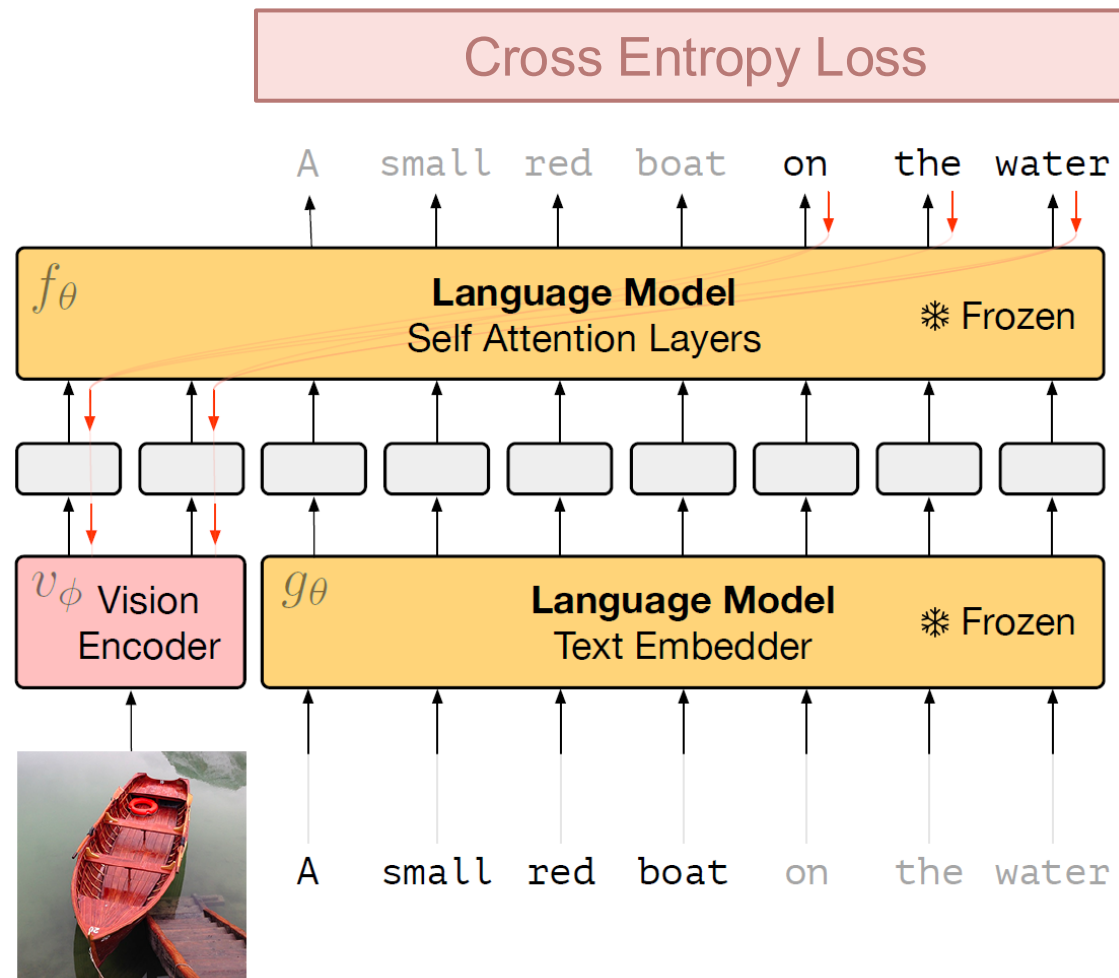
Multimodal LLM: Replaces the text encoder with a generative LLM

- Main advantages: open-ended output, in-context learning, world knowledge.

[Brown et al., Language Models are Few-shot Learners. NeurIPS 2020]

A Simple Approach for Multimodal LLM: Prefix Tuning

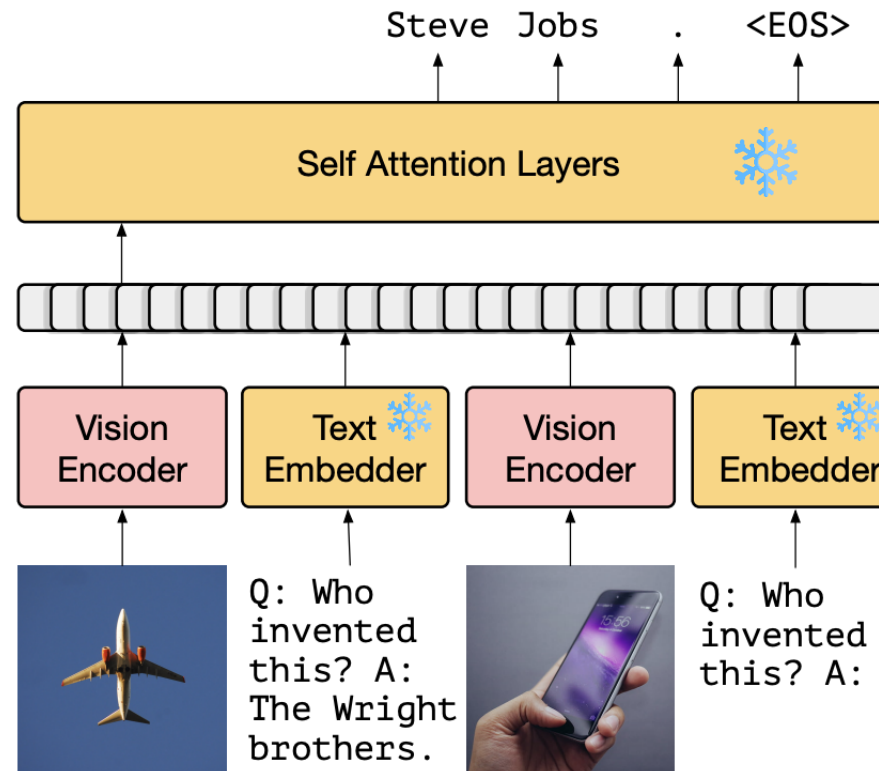
Train a visual encoder to map images into a pre-trained LM's hidden space, using captioning data



[Frozen, Tsimpoukelli et al., 2021]

A Simple Approach for Multimodal LLM: Prefix Tuning

LLM retains its knowledge and in-context learning abilities

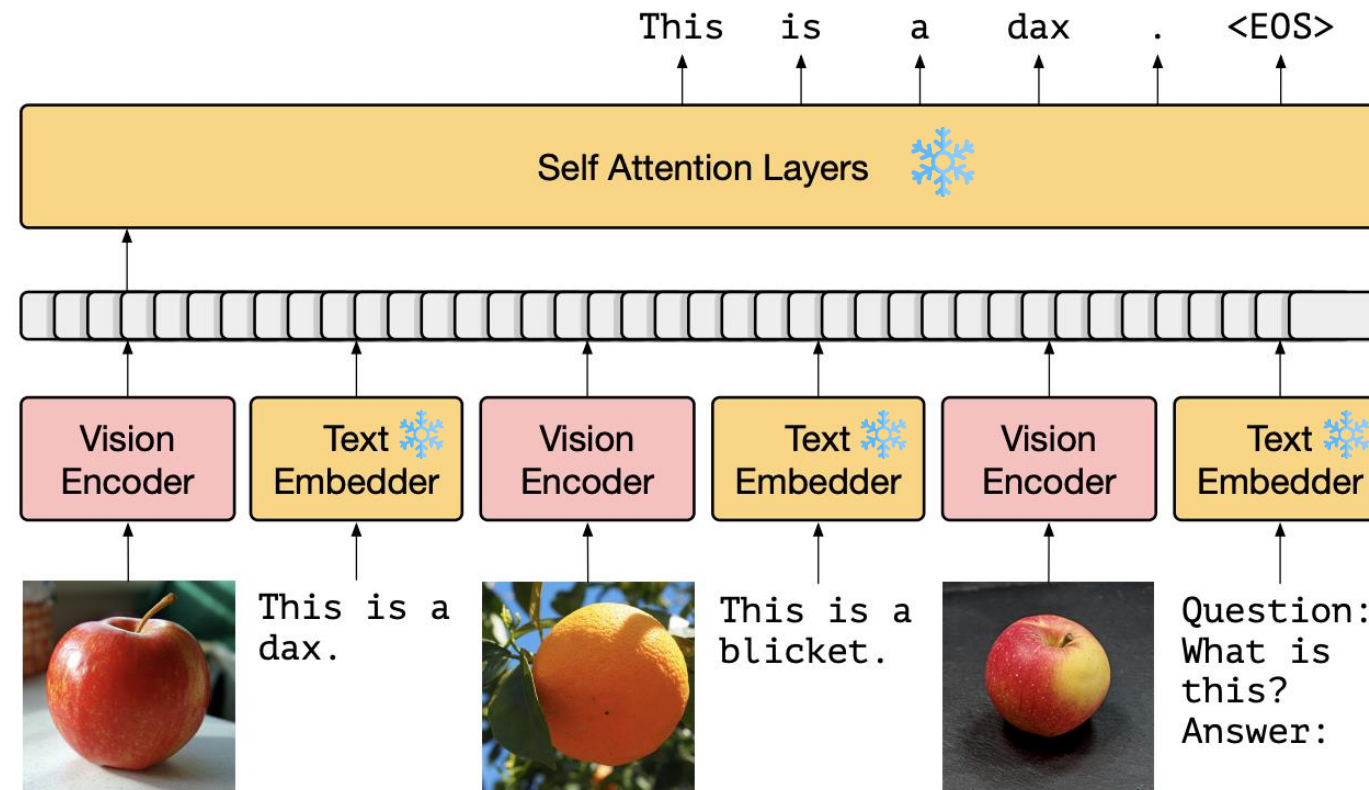


[Frozen, Tsimpoukelli et al., 2021]

<https://arxiv.org/abs/2106.13884>

A Simple Approach for Multimodal LLM: Prefix Tuning

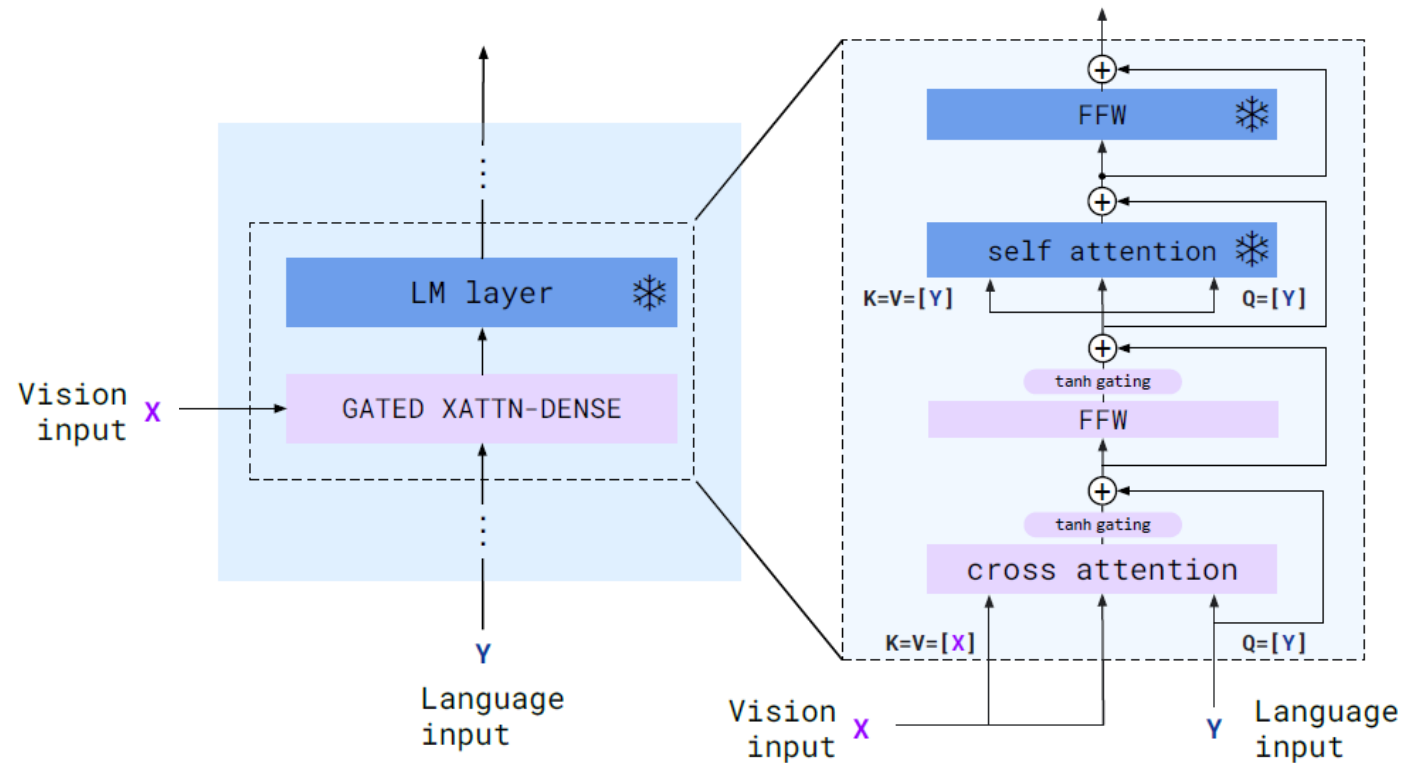
LLM retains its knowledge and in-context learning abilities



[Frozen, Tsimpoukelli et al., 2021] <https://arxiv.org/abs/2106.13884>

A Second Approach: Cross-Attention Layers (e.g., Flamingo)

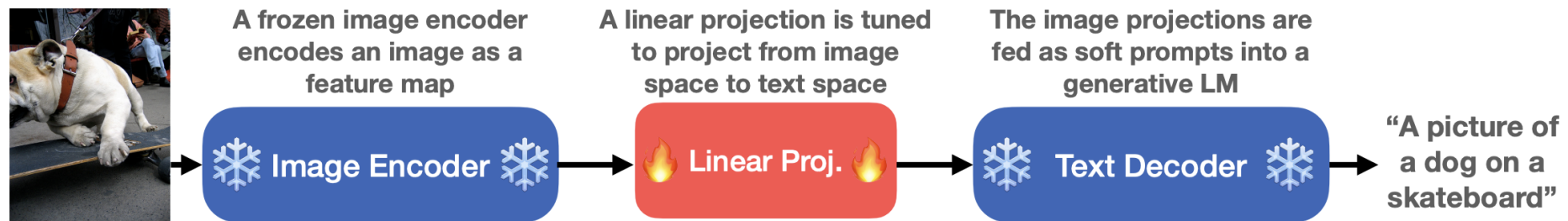
Interlaced cross-attention transformer layers:



[Alayrac et al. Flamingo: a Visual Language Model for Few-Shot Learning. NeurIPS 2022]

A Third Approach: Linear Projection (e.g., Limber)

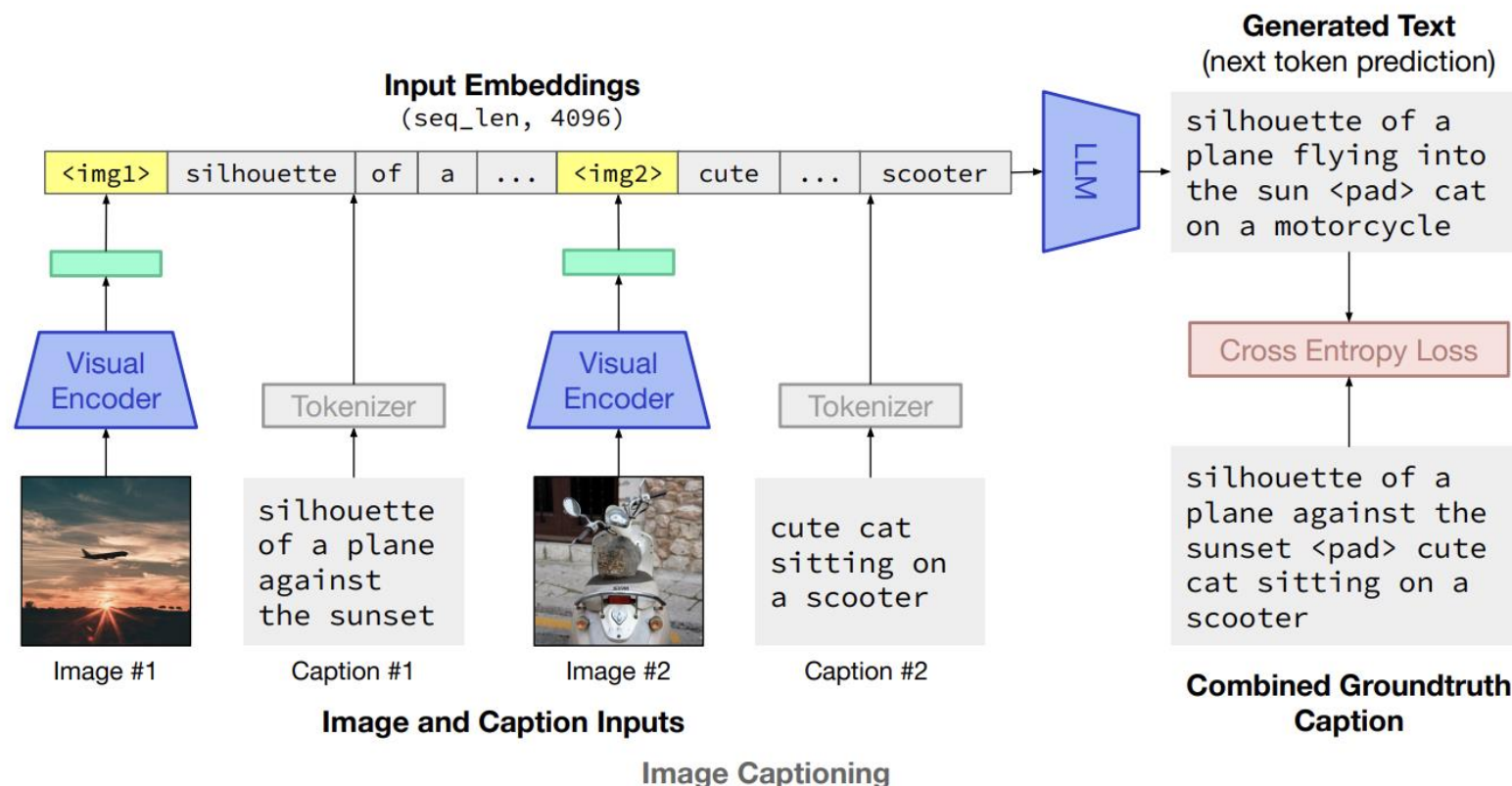
Maybe we can also use a frozen image encoder?



[Merullo et al., Linearly Mapping from Image to Text Space (LiMBer). ICLR 2023]

A Third Approach: Linear Projection (e.g., Limber)

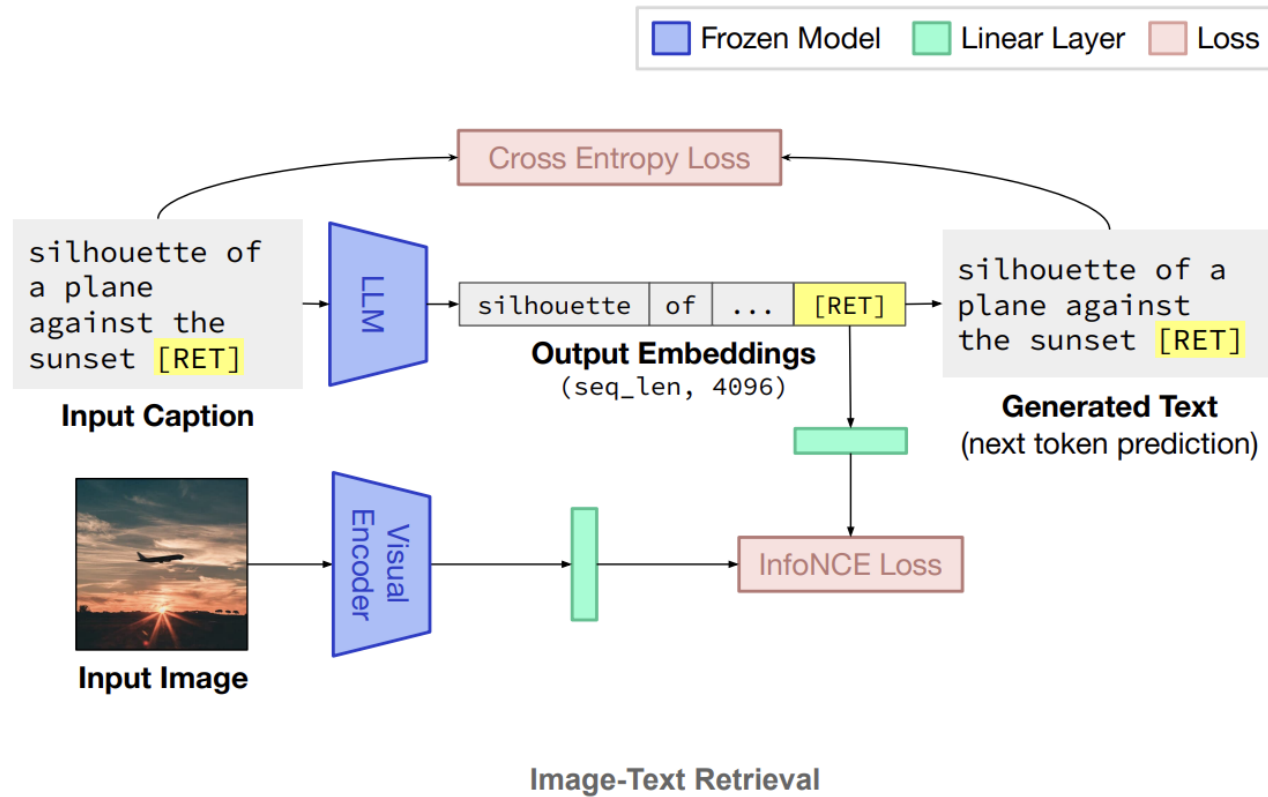
... can be trained with cross-entropy and image caption data



[Merullo et al., Linearly Mapping from Image to Text Space (LiMBer). ICLR 2023, image from Koh et al., 2023]

A Third Approach: Linear Projection + CLIP Loss (e.g., Fromage)

- CLIP, with a frozen LLM. Trainable in 1 day on 1 GPU



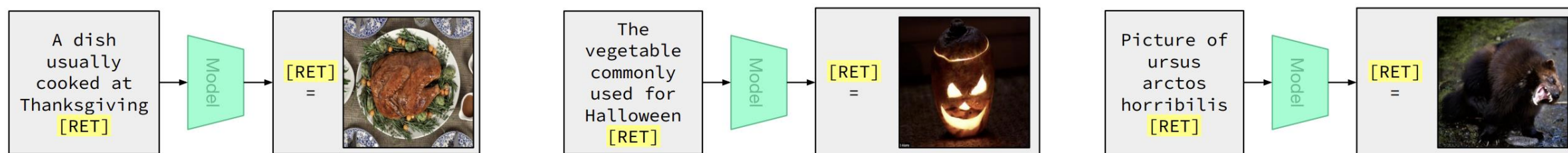
[Koh et al., Grounding Language Models to Images for Multimodal Inputs and Outputs. ICML 2023]

A Third Approach: Linear Projection + CLIP Loss (e.g., Fromage)

- Freezing the LM allows generalizing to novel tasks and domains



Concept composition

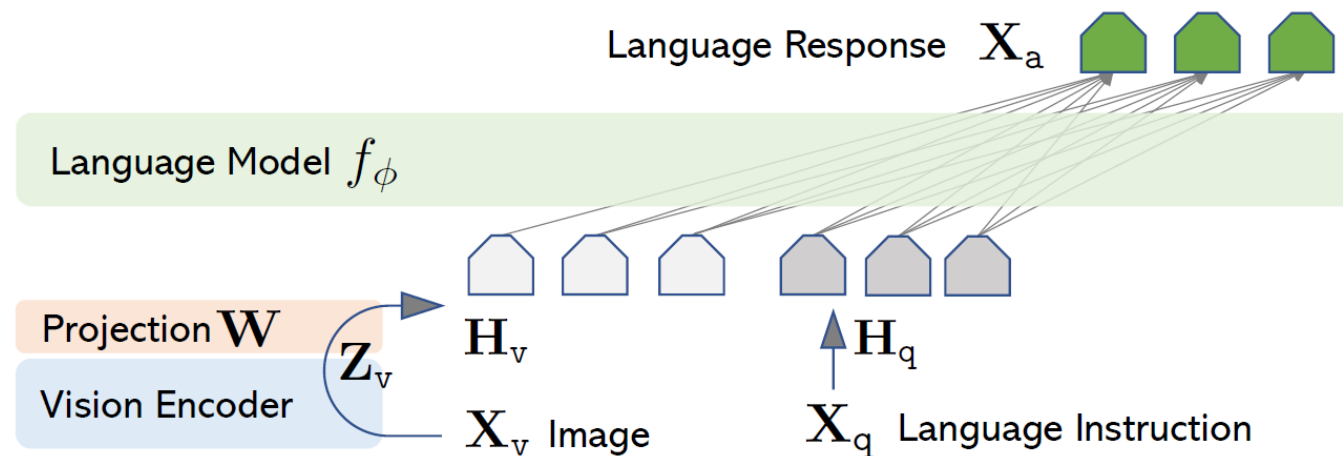


Knowledge-aware retrieval

A Fourth Approach: Visual Instruction Tuning (e.g., LLaVA)

Three main parts of the model architecture (*similar to Limber*):

- Vision encoder: CLIP ViT-L/14 (*frozen*)
- Visual connector: a single linear projection (v1.5: 2-layer MLP)
- Language model: pre-trained LLM (Vicuna)



[Liu et al., LLaVA: Large Language and Vision Assistant. Neurips 2023]

A Fourth Approach: Linear Projection + Instruction Tuning (LLaVA)

Training Stage 1: **Feature alignment** (*~595K image–caption pairs*)

- Goal: land visual features inside the LLM's existing embedding space
 - Trainable parameters: the linear projection only.
- Captions reformatted as instructions: *"Describe this image briefly"*

Training Stage 2: **Visual instruction tuning** (*158K generated examples*)

- Data written by a text-only GPT-4 that never sees the image:
 - Given only the image's captions and bounding boxes as input
 - Multi-turn conversation, detailed description, complex reasoning
- Trainable parameters: projection + LLM

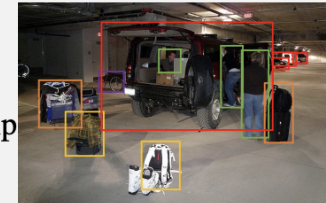
Only one training objective: next-token prediction!

A Fourth Approach: Linear Projection + Instruction Tuning (LLaVA)

Example of data created by GPT-4 with only the captions and boxes:

Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.



Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

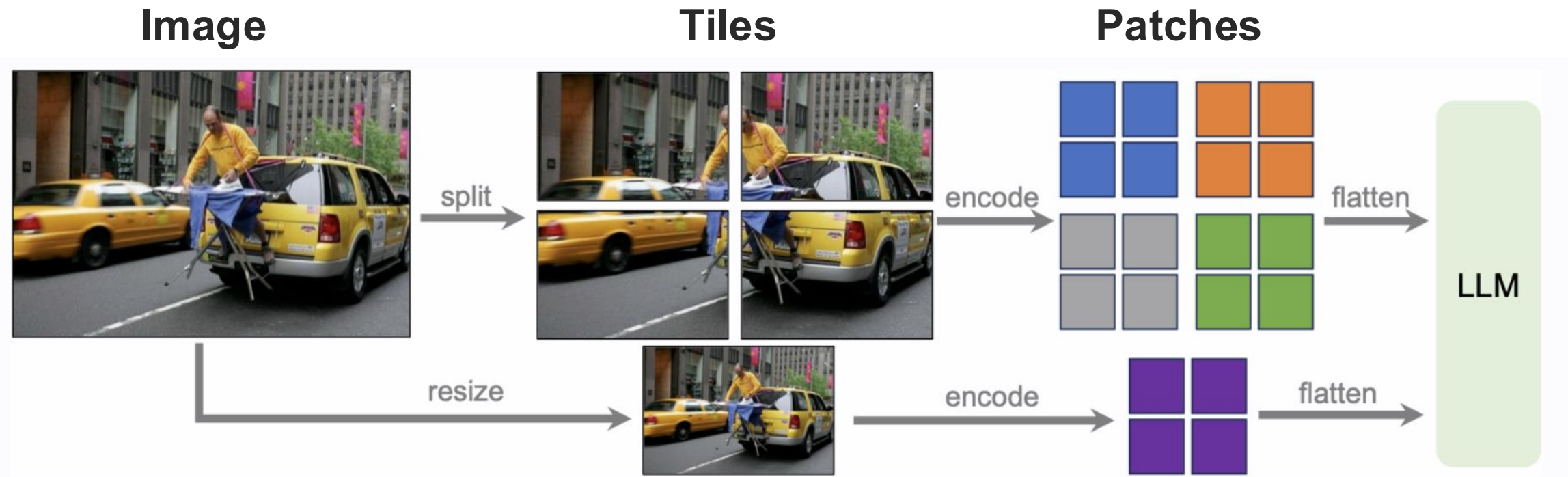
Response type 3: complex reasoning

Question: What challenges do these people face?

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

Data wins!!

LLaVA-NeXT: Scaling Data, Patches and LLM Parameters



Number of patches: ~3000 patches (from 4 tiles + 1 global view)

<https://lava-vl.github.io/blog/2024-01-30-llava-next/>