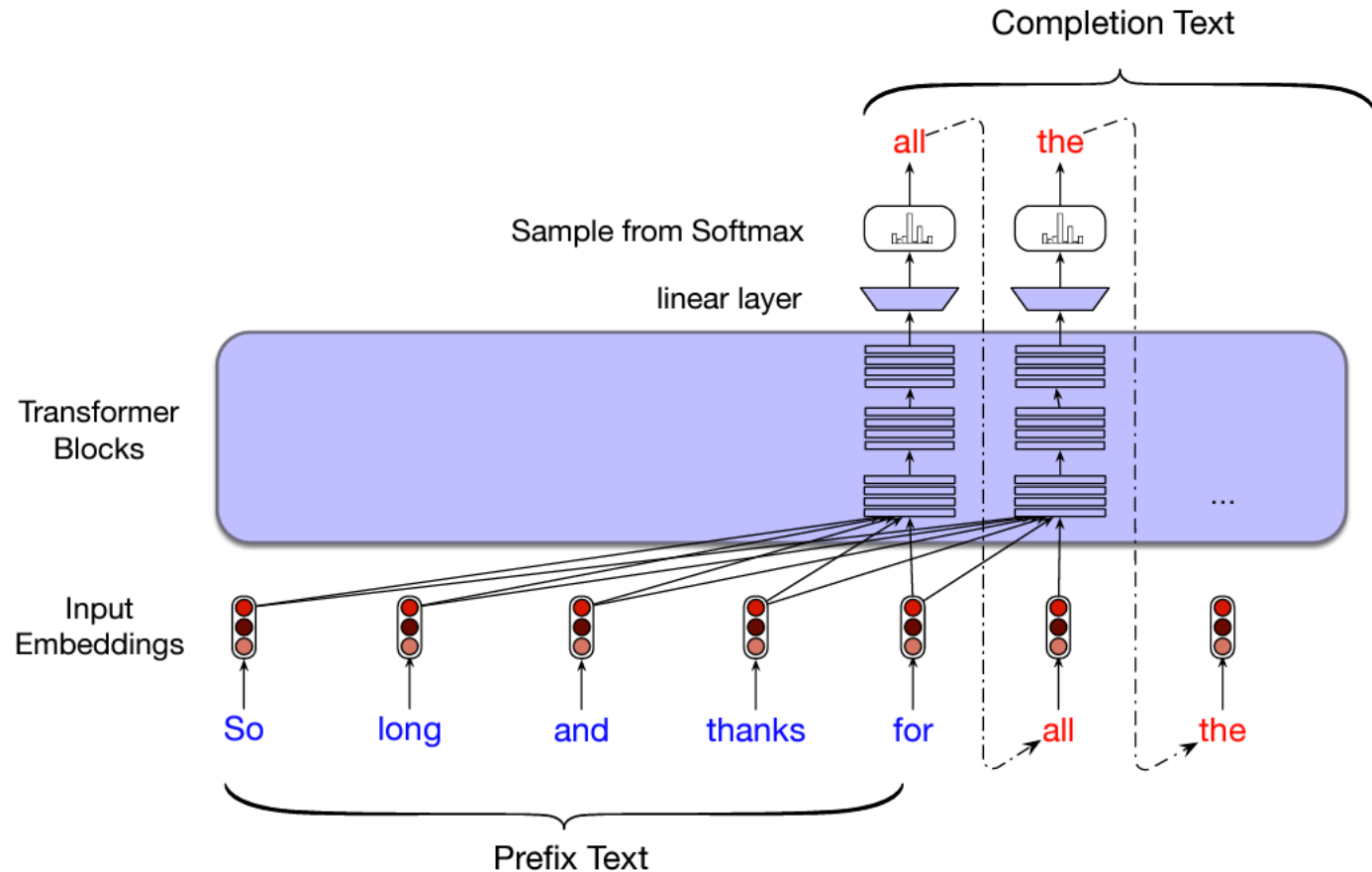




# EECS 230 Deep Learning

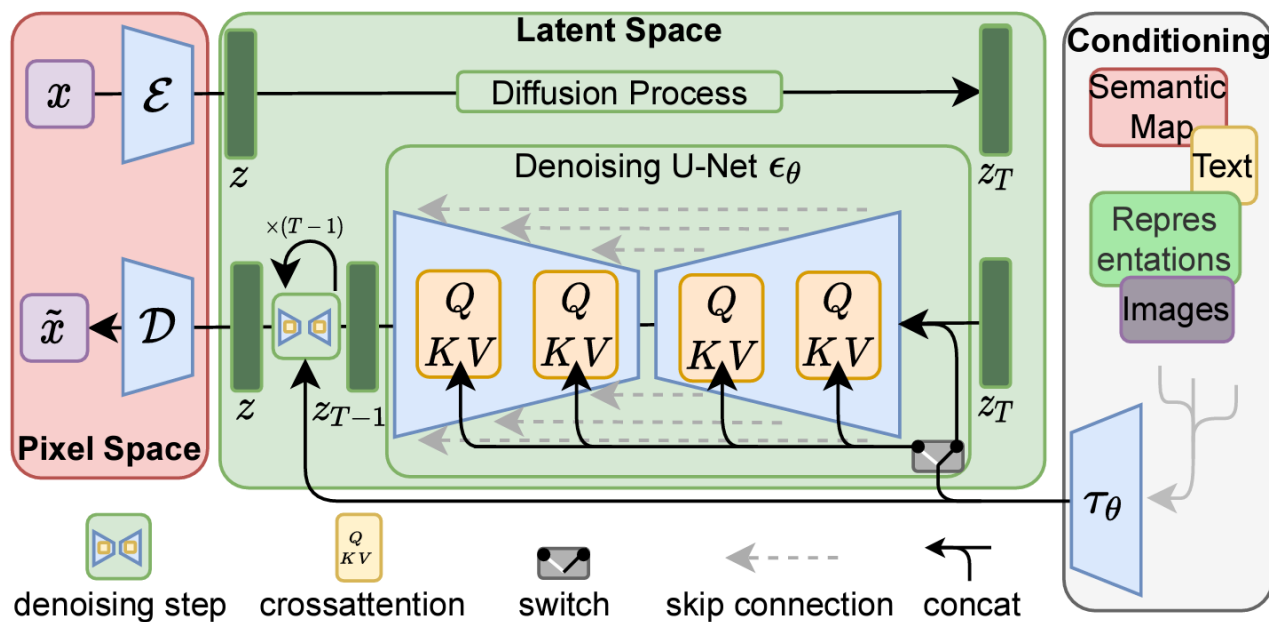
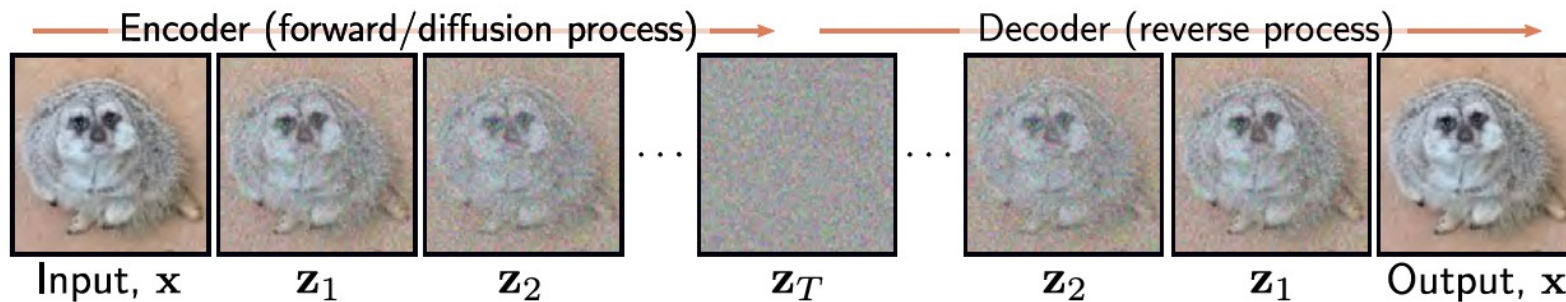
## Lecture 17: Multimodal Generative Model

# Recap: Large Language Model



# Recap: Image Generation Model

□ VAE, GAN, Diffusion model



# Multimodal Generative Model

- ❑ Generative model that takes multimodal data input or produces multimodal output
- ❑ Modalities: Video, image, text, audio, etc.
- ❑ Applications:
  - ❑ Visual Question Answering
  - ❑ Image Captioning
  - ❑ Text-to-image
  - ❑ Text-to-video

# Multimodal AI Glass



# Visual Question Answering

---

**Example of GPT-4 visual input:**

---

User      What is funny about this image? Describe it panel by panel.



Source: <https://www.reddit.com/r/hmmm/comments/ubab5v/hmmm/>

GPT-4      The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

---

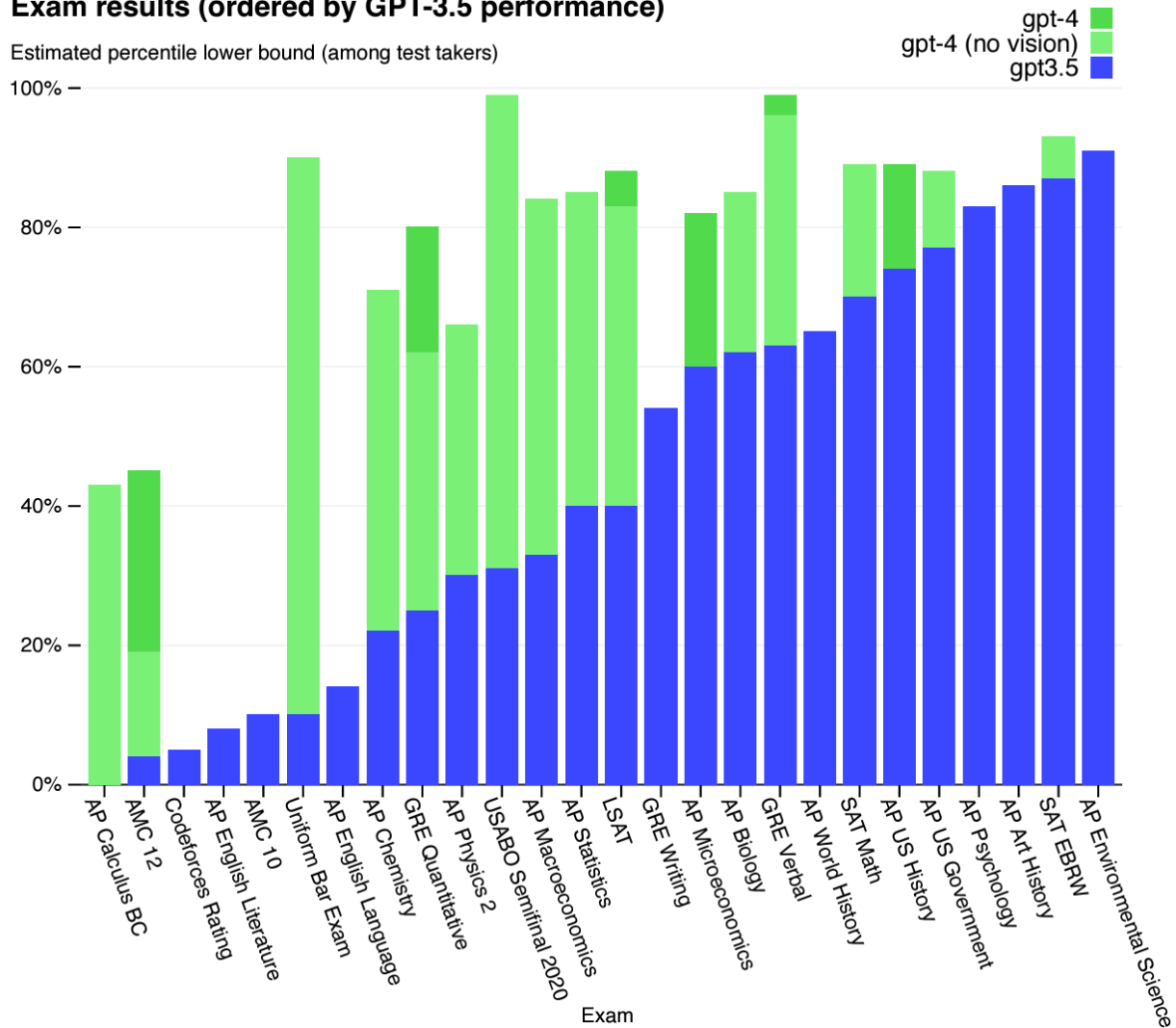
**Table 3.** Example prompt demonstrating GPT-4's visual input capability. The prompt consists of a question about an image with multiple panels which GPT-4 is able to answer.

---

# Multimodal model outperforms unimodal model

**Exam results (ordered by GPT-3.5 performance)**

Estimated percentile lower bound (among test takers)





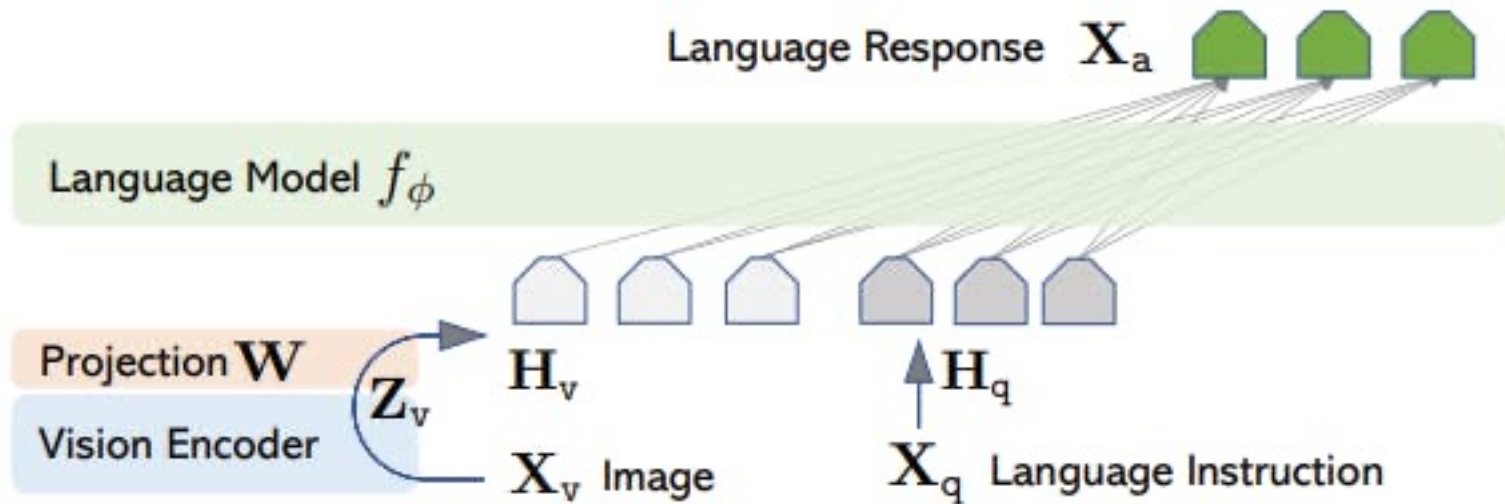
# List of Multimodal Models

Model	Year	Developer	Modality	Architecture	Key Features
<a href="#"><u>SORA</u></a>	2024	OpenAI	Video,Text	Image Encoder: Diffusion DiT	Generative Modeling,Text-to- Video
<a href="#"><u>Gemini V1.5</u></a>	2024	Google	Video,Text,Audio	Image Encoder: ViT,Text Encoder:Transformer	Generative Modeling,Long Context Window
<a href="#"><u>BLIP2</u></a>	2023	Salesforce Research	Image,Text	Q-Former: Bridging Modality Gap,Image Encoder: ViT-L/ViT- G,Text LLM Encoder: OPT/FlanT5	Generative Modeling,Image-to- Text,Visual Question Answering,Image-to- Text Retrieval
<a href="#"><u>GPT-4V</u></a>	2023	OpenAI	Image,Text	Text Encoder: GPT	Generative Modeling,Multimodal LLM,Visual Question Answering
<a href="#"><u>LLaVA</u></a>	2023	Microsoft	Image,Text	Text LLM Encoder: Vicuna,Image Encoder:CLIP visual ViT-L	Generative Modeling,Visual Instruction Generation
<a href="#"><u>FLAMINGO</u></a>	2022	DeepMind	Image,Text	Gated Cross Attention,Multitway Transformer,ViT-giant	VQA,Interleaved Visual and Textual Data
<a href="#"><u>BEiT-3</u></a>	2022	Microsoft	Image,Text	Text Encoder: OPT/FlanT5,Image Encoder:ViT-L/ViT-g	Object Detection,Visual Question Answering,Image Captioning
<a href="#"><u>CLIP</u></a>	2021	OpenAI	Image,Text	Text Encoder: Transformer; Image Encoder: ResNet/ViT	Multimodal Alignment,Zero-Shot Learning



# How to add image as input to LLM?

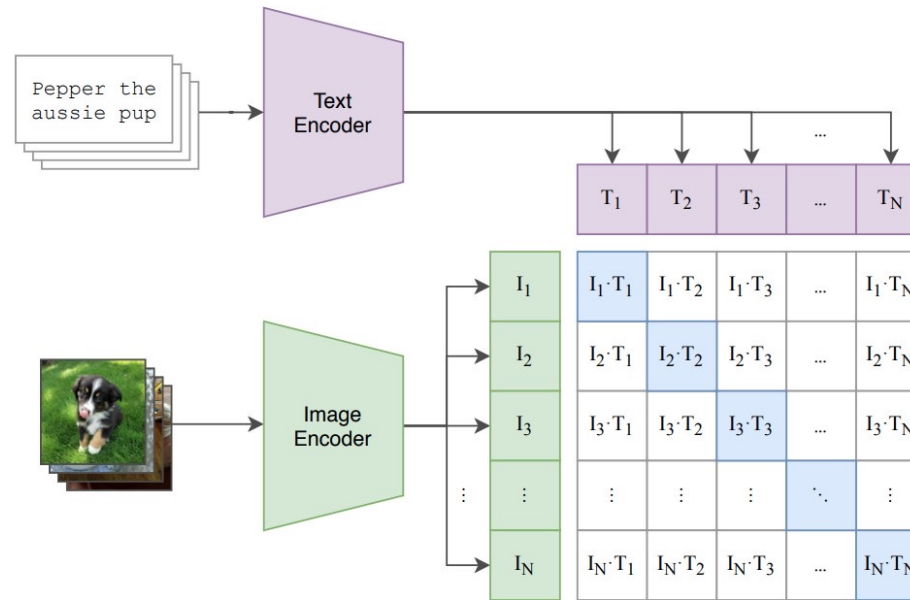
□ LLaVA (Large Language and Vision Assistant)



LLaVA architecture

# Recap: CLIP (Contrastive Language–Image Pre-training)

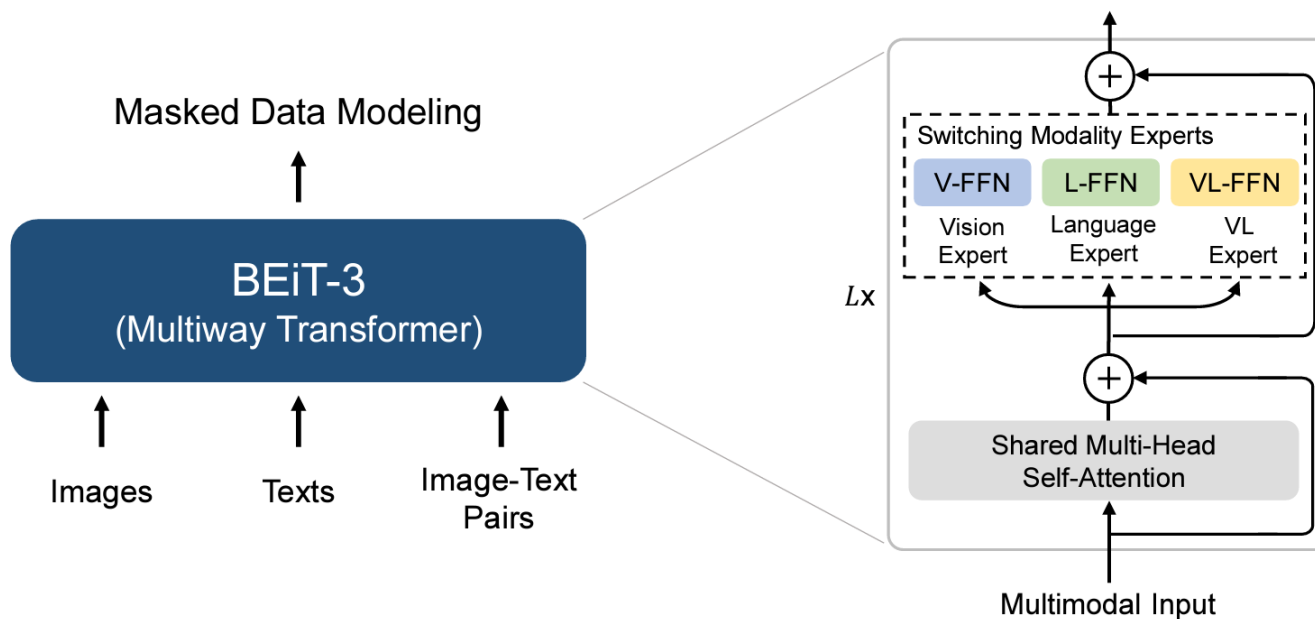
(1) Contrastive pre-training



**Figure:** Contrastive Pre-training of language-image pairs. The text encoder is a standard transformer encoder. The extracted feature is the embedding of the CLS token. The image encoder is either a ResNet-50 or a Vision Transformer (ViT).

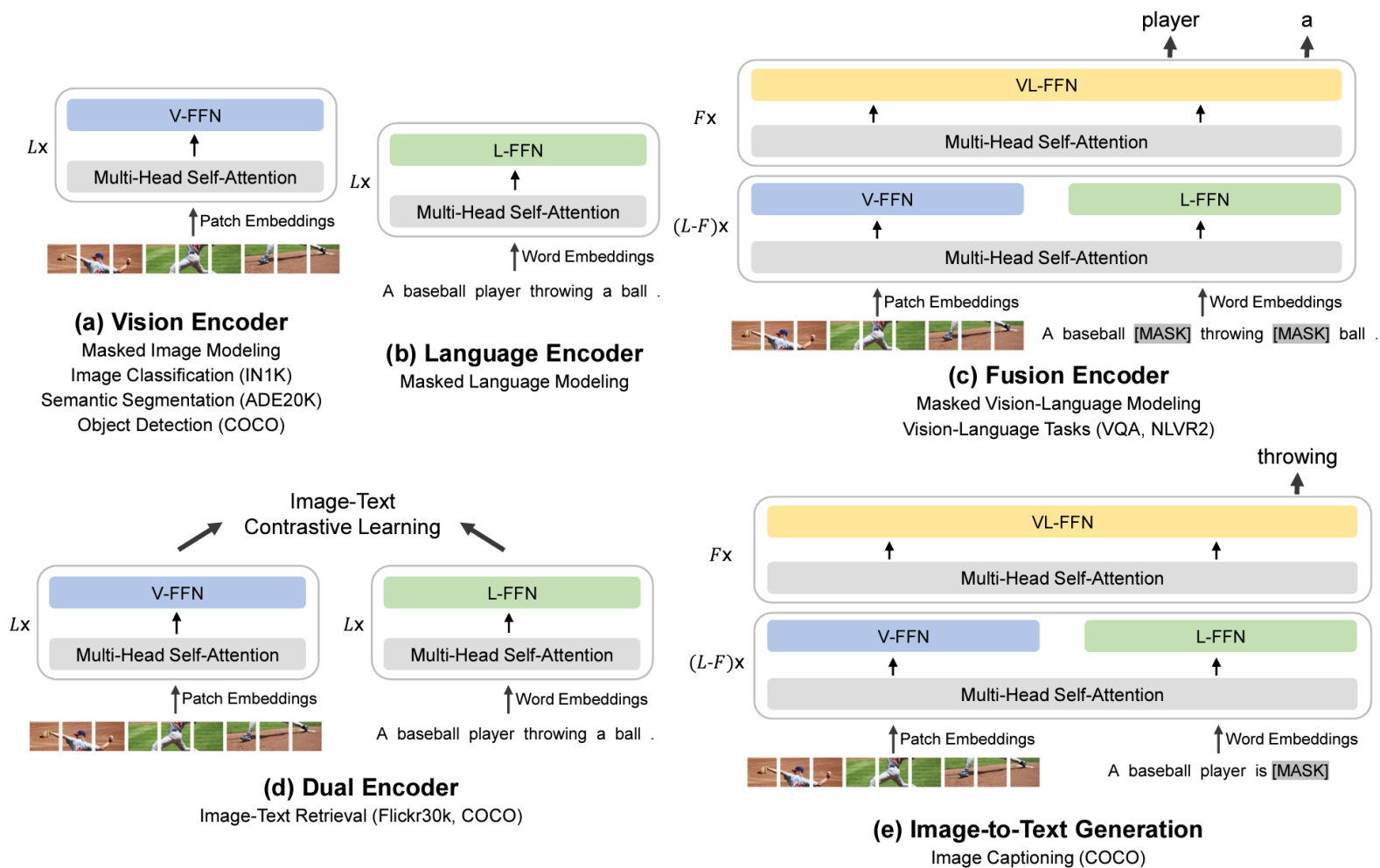
# Multiway-transformer

□ A shared Multiway Transformer as the backbone network



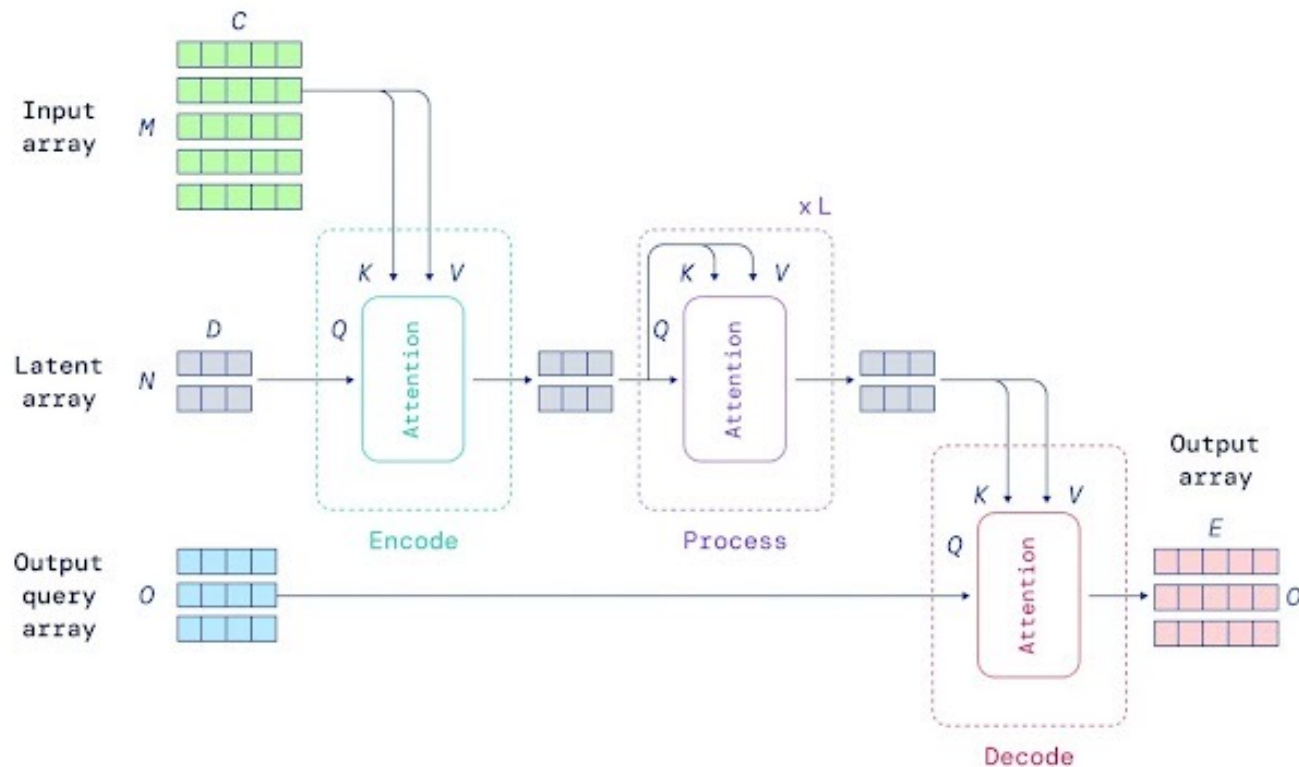
# Multiway-transformer

## □ A shared Multiway Transformer as the backbone network



# Multimodal architecture

□ Perceiver and Perceiver IO as general purpose architecture



# Flamminggo

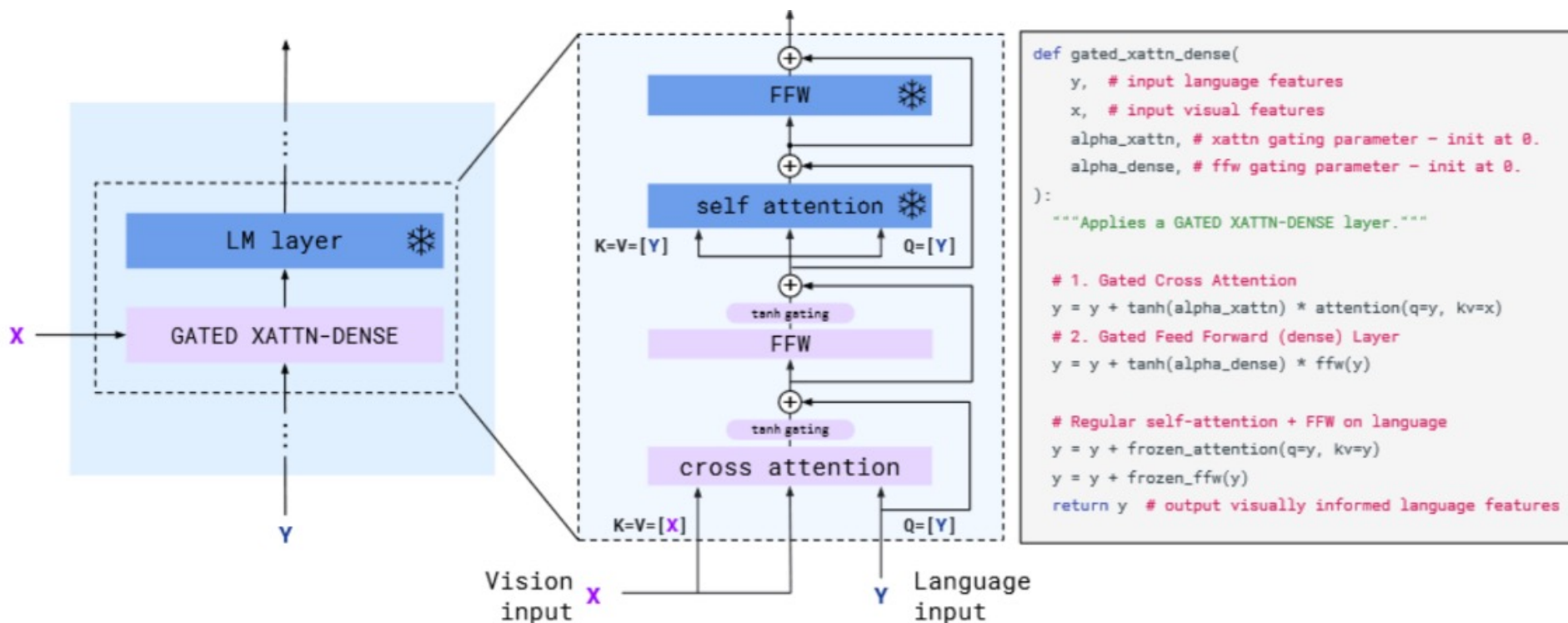


Figure 5 | **GATED XATTN-DENSE layers.** We insert new cross-attention layers, whose keys and values are obtained from the vision features while using language queries, followed by dense feed forward layers in between existing pretrained and frozen LM layers in order to condition the LM on visual inputs. These layers are *gated* so that the LM is kept intact at initialization for improved stability and performance.