Lecture 5.3 - Perceiver IO an more generalized Multimodal Models

Generative AI Teaching Kit

This lecture

- Multimodal Models Recap
- GPT-4V and LLaVa Multimodal Assistants
- Perceiver Architectures

Multimodal Models

Recap on what modalities we’ve seen

Multimodal Model Types

So far we have looked at the following modalities:

- Text
- Audio
- Images

These three (along with video) make up the bulk of problems that are associated with multimodalities.

We have seen these as inputs to many models with transformers as the backbone.

GPT/BERT

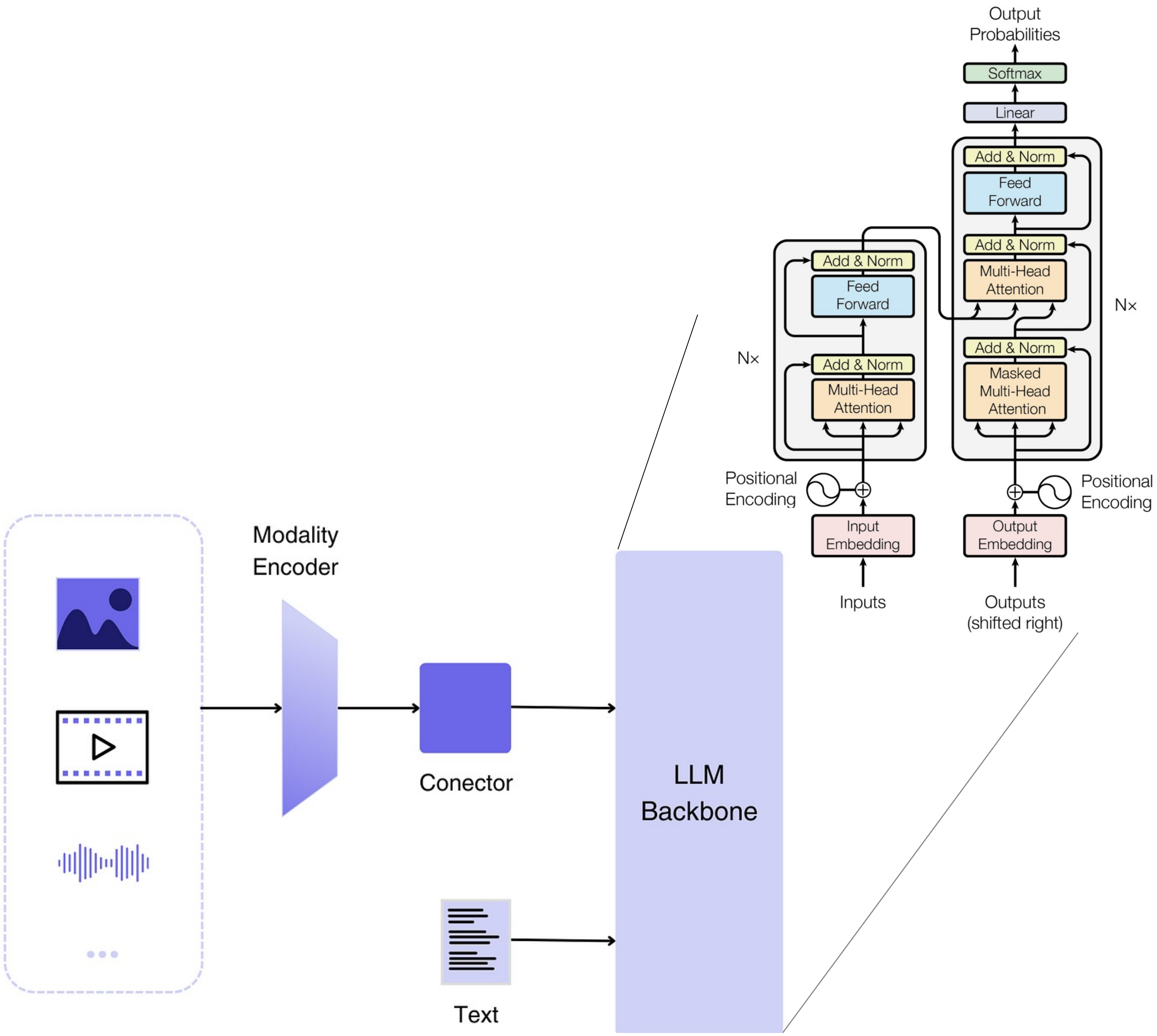
Input: text | Output: **text**

Whisper

Input: Audio | Output: **text**

ViT/CLIP

Input: Images | Output: **text**

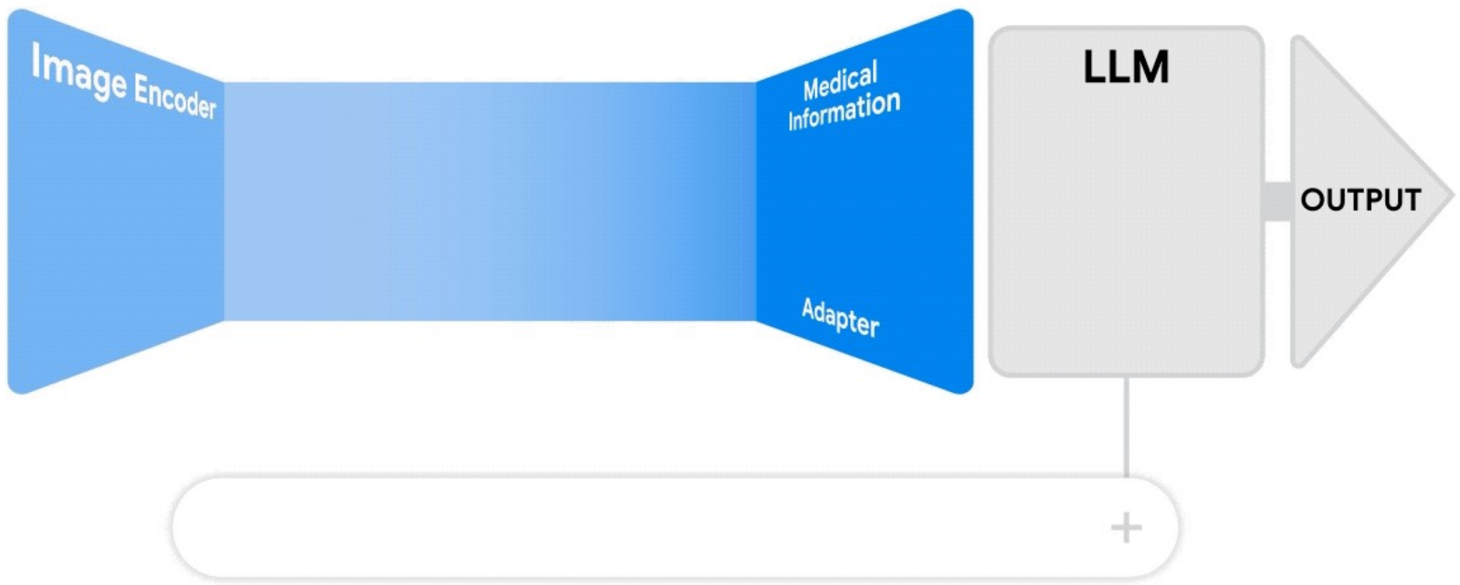


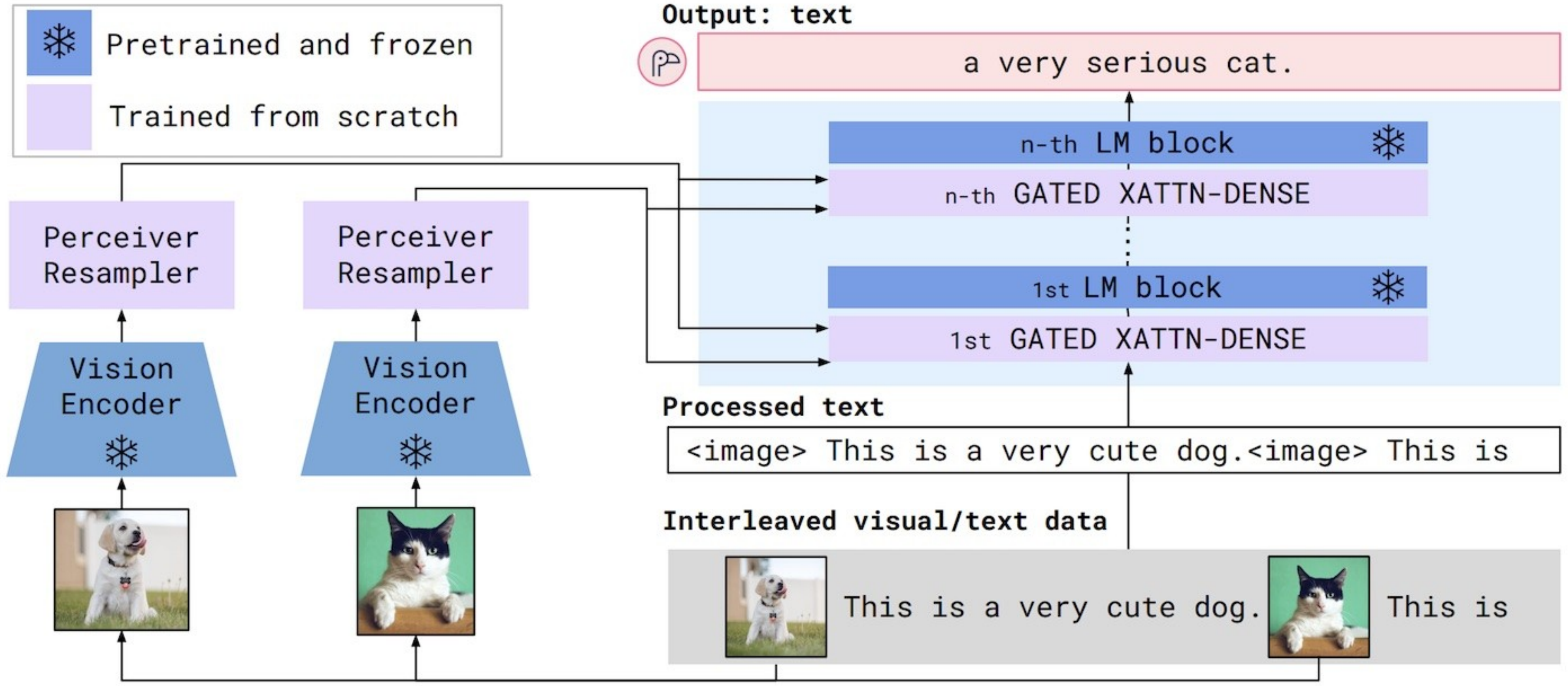
Multimodal Model Types – Image to Text

We have seen these models that can produce text that classifies, captions, and answers questions to what is in an image.

Some examples of these include:

- Vision Transformer
- Contrastive Language-Image Pretraining
- Flamingo
  - This model extends on some of the previous work in CLIP by allowing the model to generate text rather than align text to an image. This allows Flamingo to be used like BLIP as a VQA model answering questions of input images and text.



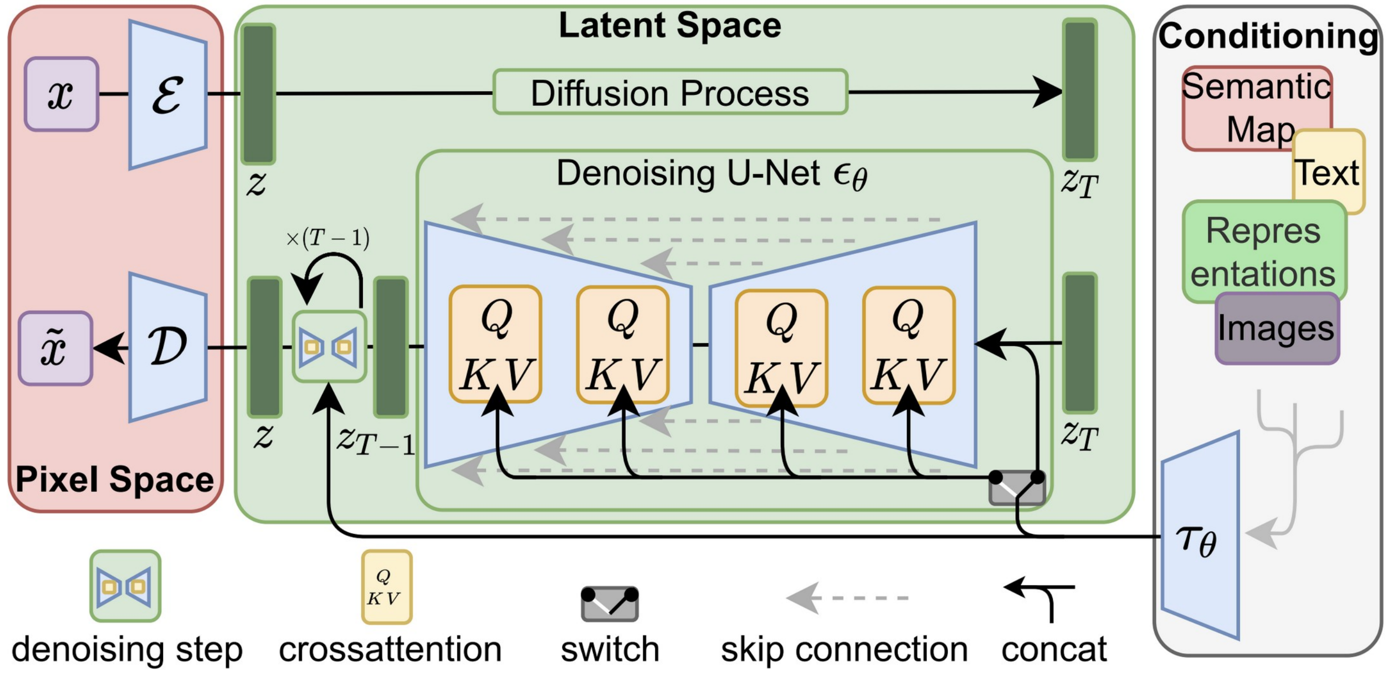


Flamingo Model

Multimodal Model Types – Text/Image to Image

To generate images, typically diffusion models are used in modern application. These models, which take text as input and output images, older models like cycle/styleGAN can also generate images based on image, and potentially text, as input.

In Module 6 we will cover image generation and diffusion models in much more detail, but they are worth mentioning here as a form of multimodal models.



Multimodal Model Types - Audio

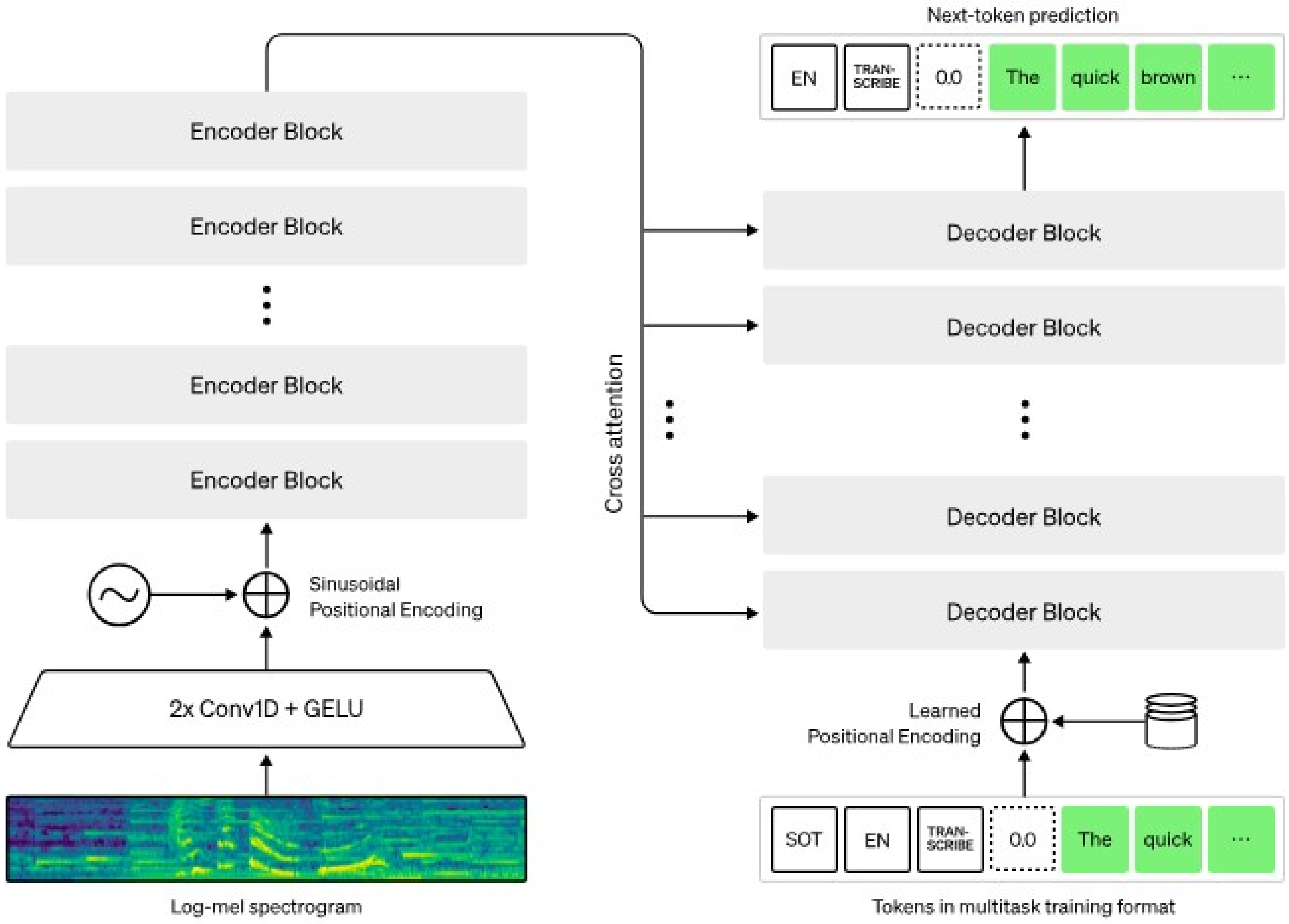
The Whisper model uses a transformer with some additional encoders to make use of audio signals and text generation for audio transcription.

Encoder:

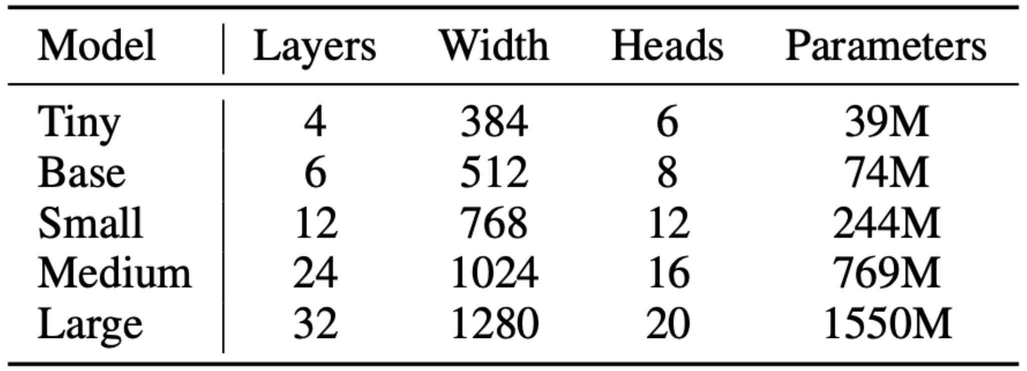
- 12-24 transformer blocks (size dependent)
- Multi-head cross-attention
- Layer normalization strategy

Decoder:

- Causal masking for autoregressive generation
- Cross-attention to audio features
- Vocabulary of 50k tokens (similar to GPT-3)
- Special tokens for tasks/languages



As with all approaches to transformer-based multimodal modelling, the key challenge is encoding the inputs into a latent space before being processed.



Model	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

Table 1. Architecture details of the Whisper model family.

GPT-4V and LLaVa

Chatting with modalities

Adding Vision to ChatGPT

With the release of GPT-4o, ChatGPT was able to interpret images and other modalities in a single model.

GPT-4V – OpenAI’s Vision Transformer

Prior to this, GPT-4V was an image description model that would take as input images and generate text output based on conditioned inputs.

While the exact technical details of GPT-4V were not released, speculation as to the architecture included a variation of the CLIP-based models (CLIP also being an OpenAI model) with a vision encoder, likely a Vision Transformer, conditioned on image-text pairs.

ChatGPT with GPT-4o

GPT-4o, however, is further trained with Reinforcement learning with Human Feedback to allow the model to chat along side processing image information.

September 25, 2023

GPT-4V(ision) system card

May 13, 2024

Hello GPT-4o

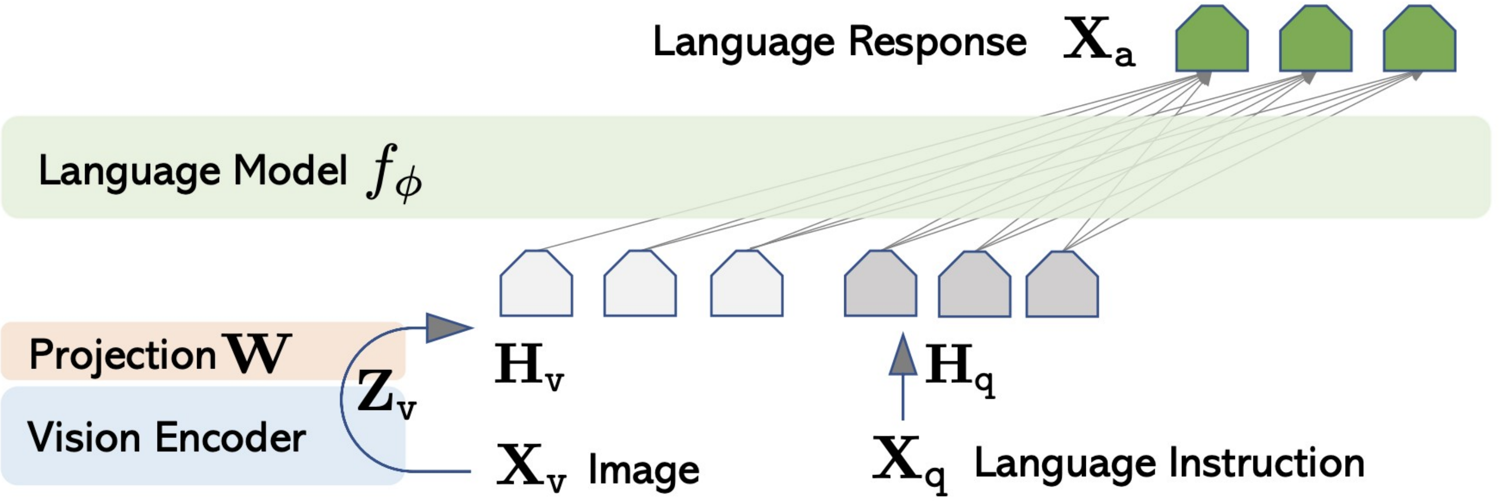
We’re announcing GPT-4o, our new flagship model that can reason across audio, vision, and text in real time.

Note: while ChatGPT technically can produce images, this is done by parsing the inputs from the user and sending relevant prompts to a Dall-E model which uses diffusion modeling to produce the image.

Community response to GPT-4V - LLaVA

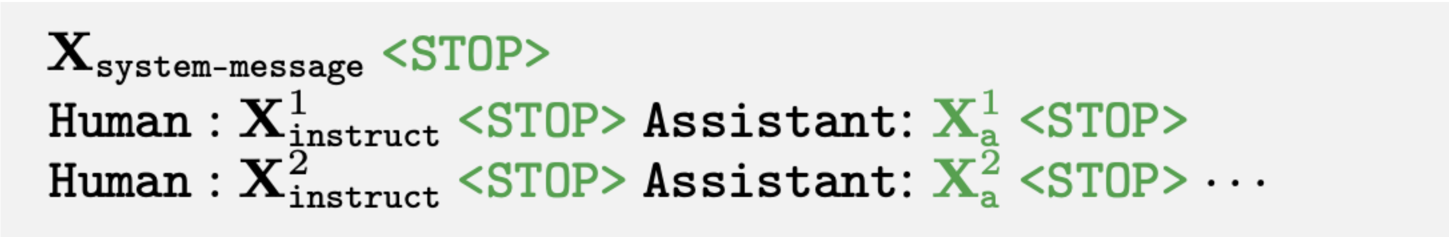
LLaVA: Large Language and Vision Assistant

LLaVA is an open-source alternative to GPT4-V and -4o. This model utilizes both images and text as input and outputs text.



Language Response $\mathbf{X}_{\mathrm{a}}$
Language Model f_{ϕ}
Projection $\mathbf{W}$
$\mathbf{Z}_{\mathrm{v}}$
$\mathbf{H}_{\mathrm{v}}$
$\mathbf{H}_{\mathrm{q}}$
Vision Encoder
$\mathbf{X}_{\mathrm{v}}$ Image
$\mathbf{X}_{\mathrm{q}}$ Language Instruction

Figure 1: LLaVA network architecture.



$\mathbf{X}_{\texttt{system-message}}$ <STOP>
Human : $\mathbf{X}^{1}_{\texttt{instruct}}$ <STOP> Assistant: $\mathbf{X}^{1}_{\mathrm{a}}$ <STOP>
Human : $\mathbf{X}^{2}_{\texttt{instruct}}$ <STOP> Assistant: $\mathbf{X}^{2}_{\mathrm{a}}$ <STOP> $\cdots$

Input: Vision and Language Encoding

Images are processed using a vision encoder, typically a ViT, and are projected into the same space as the text encoder, thereby enabling a combined multimodal encoding.

Output: Text-Language

Text outputs are trained in an autoregressive manner, in conjunction with Instruction/Chat modeling to enable an assistant-type interaction.

Further Improvements to LLaVA

LLaVA 1.5 – Simple Changes and Large Improvements

6 Months after the release of the LLaVA 1.0 model, the authors improved upon the design by releasing LLaVA 1.5 (May 2024). This update included:

- Adding a MLP layer instead of a linear projection layer after the vision encoder to form richer interactions with the language encoder.
- Using a more advanced CLIP-ViT image encoder
- Training with VQA academic tasks

With these changes, LLaVA 1.5 achieved SOTA in various benchmarks and strongly surpassed the performance of LLaVA 1.0.

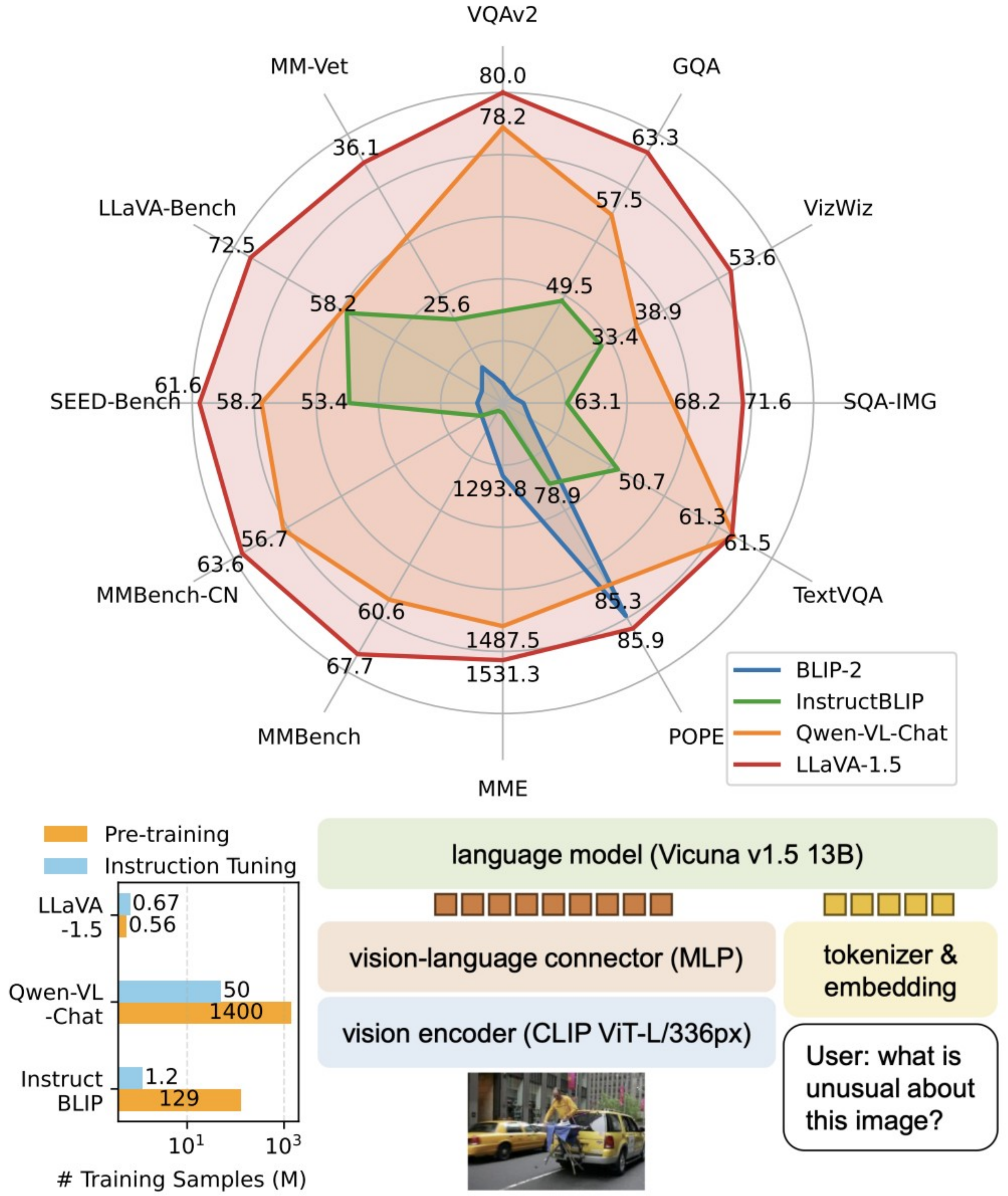
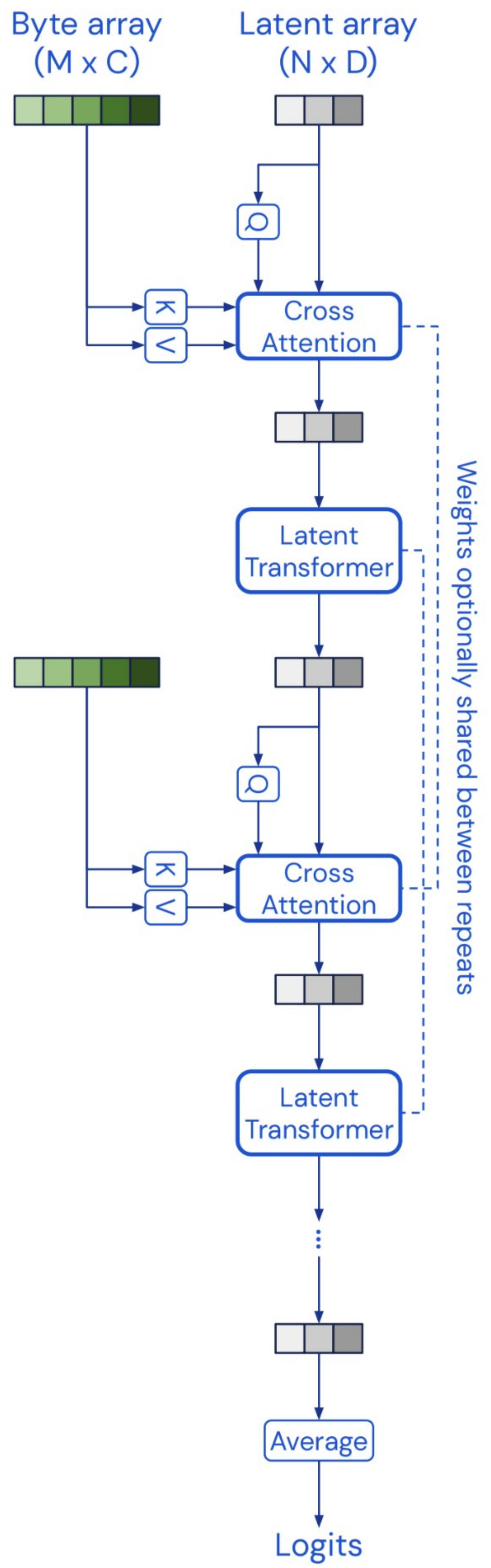


Figure 1. **LLaVA-1.5** achieves SoTA on a broad range of 11 tasks (Top), with high training sample efficiency (Left) and simple modifications to LLaVA (Right): an MLP connector and including academic-task-oriented data with response formatting prompts.

Perceiver Models

Enabling more than just text outputs with generalized attention

Perceiver – Utilizing agnostic attention architecture



Flexible Multimodal Architectures

The Perceiver (Jun 2021) is designed to very flexibly handle a wide range of inputs out of the box even if they come from very different modalities, including high-bandwidth ones such as images and audio

Concept of Latent Bottleneck

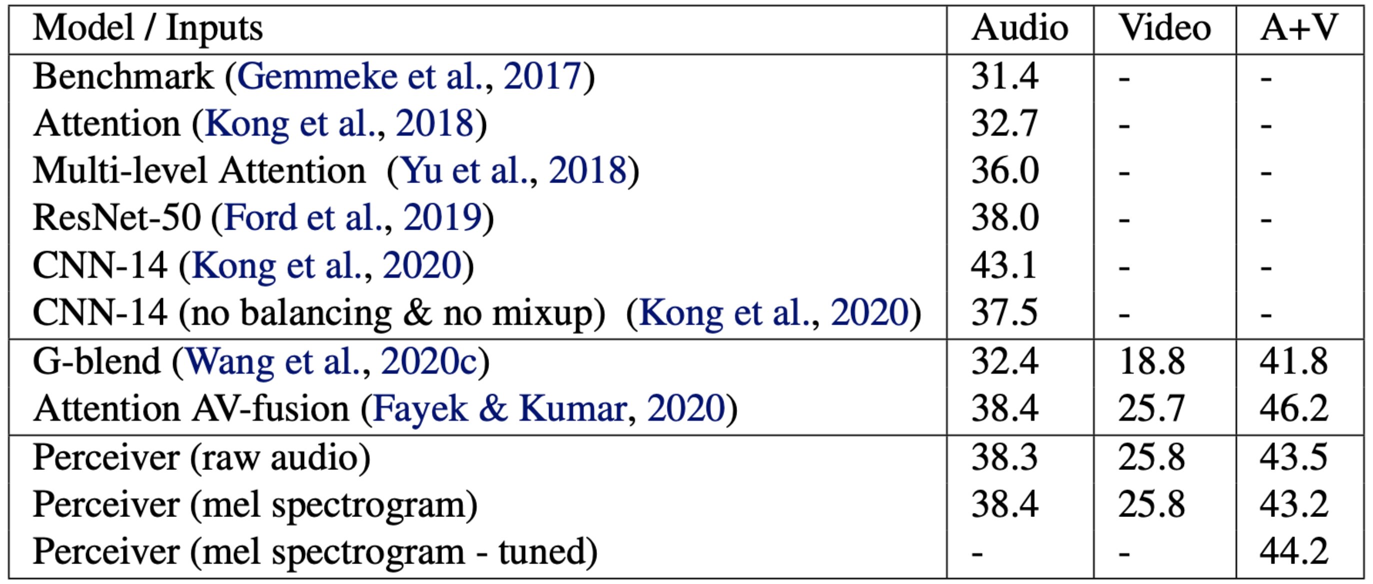
- The core idea is to introduce a small set of latent units that forms an attention bottleneck through which the inputs must pass.
- This eliminates the quadratic scaling problem of all-to-all attention of a classical Transformer and decouples the network depth from the input's size, allowing the construction of very deep models.
- By attending to the inputs iteratively, the Perceiver can channel its limited capacity to the most relevant inputs, informed by previous steps.

Perceiver – Utilizing agnostic attention architecture

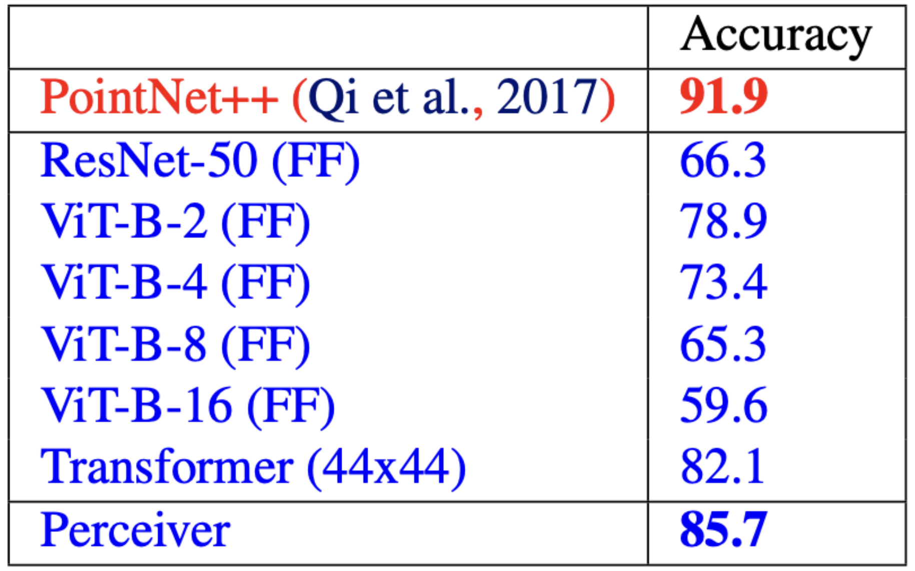
The Perceiver architecture can be trained on several different modalities.

In their work, the authors show that across images, audio-video, and point-cloud data, the Perceiver performs at or above state-of-the-art purpose built models.

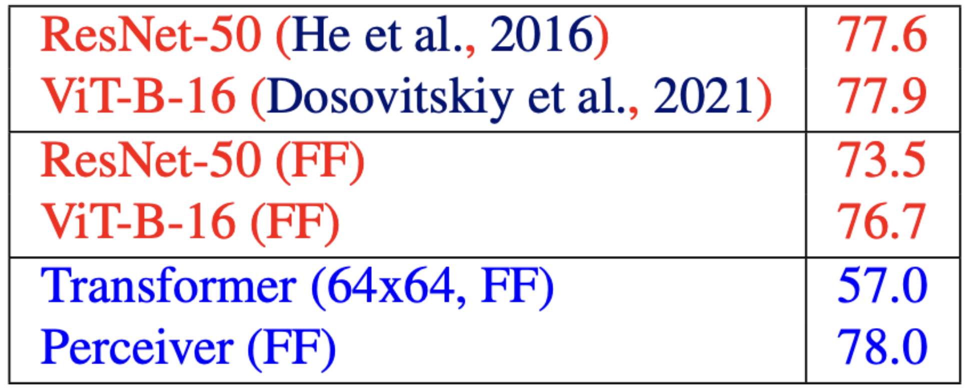
However, while this model can take arbitrary inputs, the outputs are still designed for text-only output.



Model / Inputs	Audio	Video	A+V
Benchmark (Gemmeke et al., 2017)	31.4	-	-
Attention (Kong et al., 2018)	32.7	-	-
Multi-level Attention (Yu et al., 2018)	36.0	-	-
ResNet-50 (Ford et al., 2019)	38.0	-	-
CNN-14 (Kong et al., 2020)	43.1	-	-
CNN-14 (no balancing & no mixup) (Kong et al., 2020)	37.5	-	-
G-blend (Wang et al., 2020c)	32.4	18.8	41.8
Attention AV-fusion (Fayek & Kumar, 2020)	38.4	25.7	46.2
Perceiver (raw audio)	38.3	25.8	43.5
Perceiver (mel spectrogram)	38.4	25.8	43.2
Perceiver (mel spectrogram - tuned)	-	-	44.2



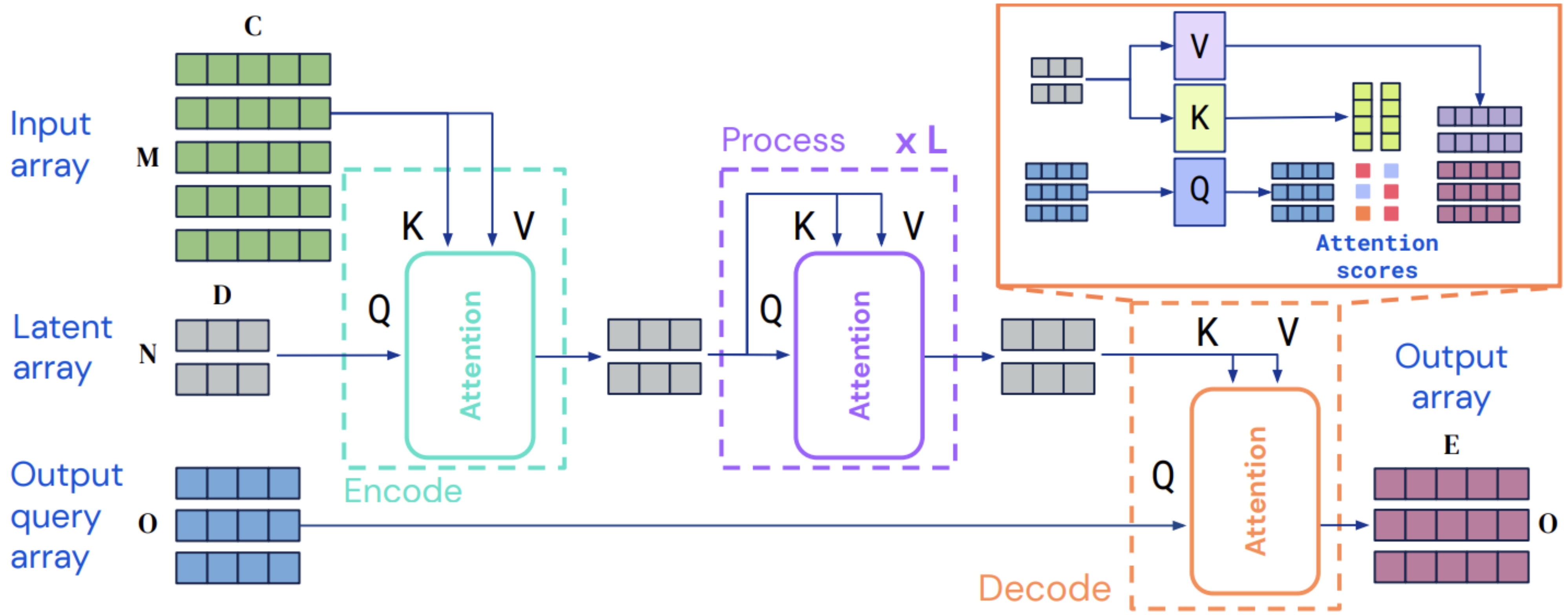
	Accuracy
PointNet++ (Qi et al., 2017)	**91.9**
ResNet-50 (FF)	66.3
ViT-B-2 (FF)	78.9
ViT-B-4 (FF)	73.4
ViT-B-8 (FF)	65.3
ViT-B-16 (FF)	59.6
Transformer (44x44)	82.1
Perceiver	**85.7**



ResNet-50 (He et al., 2016)	77.6
ViT-B-16 (Dosovitskiy et al., 2021)	77.9
ResNet-50 (FF)	73.5
ViT-B-16 (FF)	76.7
Transformer (64x64, FF)	57.0
Perceiver (FF)	78.0

Perceiver IO – Enabling multimodal output

To achieve multimodal output as well as input, the authors released an update to Perceiver in Mar 2022, Perceiver IO.



Output Arbitrary Query Array

Perceiver IO maps an arbitrary input array to an arbitrary output array in a domain agnostic process.

The bulk of the computation takes place in a reduced latent space where data is compressed, which enables more efficient processing of even high resolution input data.

Perceiver IO – Enabling multimodal output

The Perceiver IO architecture can be split into three main components. These are all roughly similar in design, but make use of different vectors to attend to.

Encode

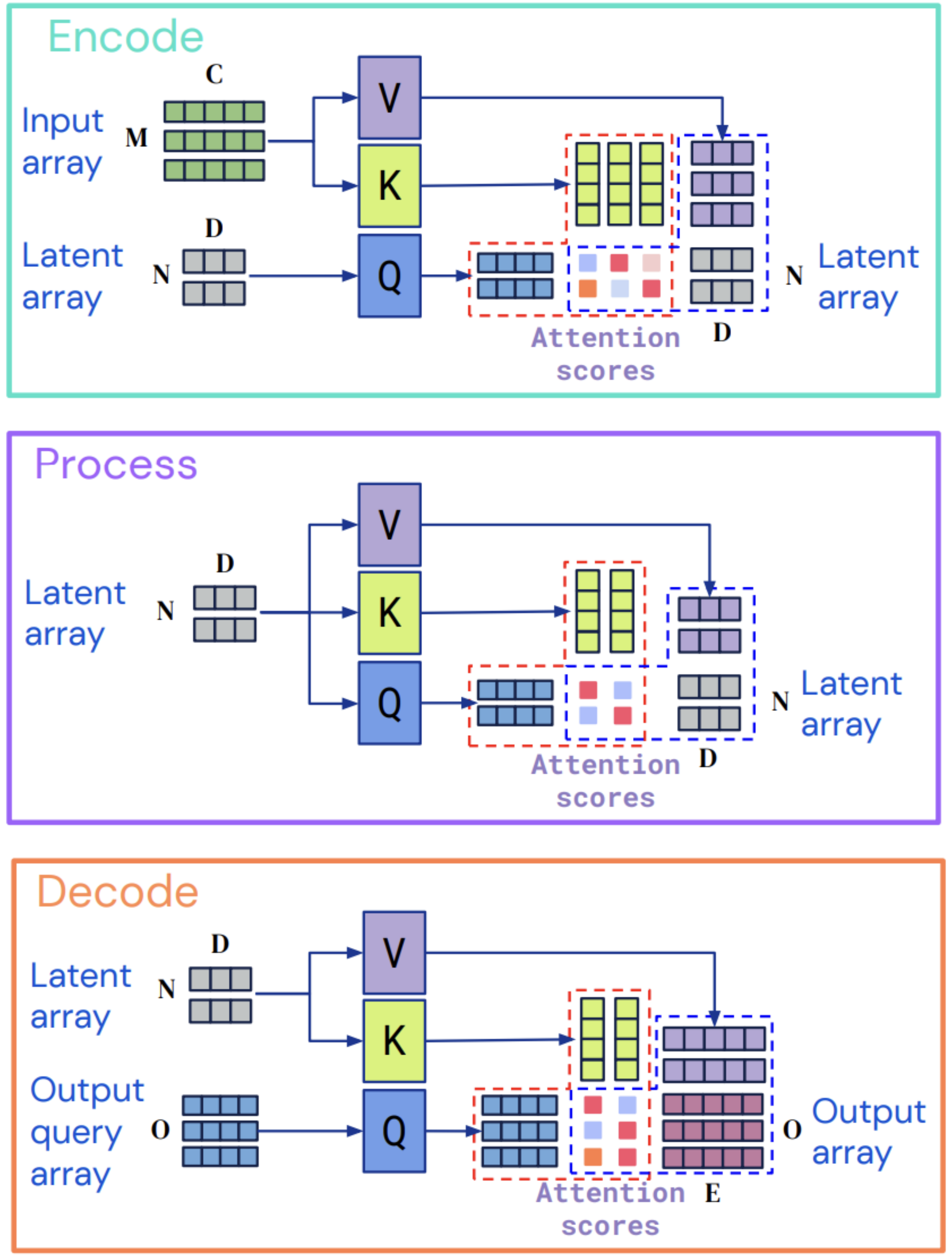
- Here inputs are mapped to a latent space, typically with a smaller dimension relative to the input.

Process

- Self-attention blocks in the style of regular transformers are used to produce highly enriched vectors.

Decode

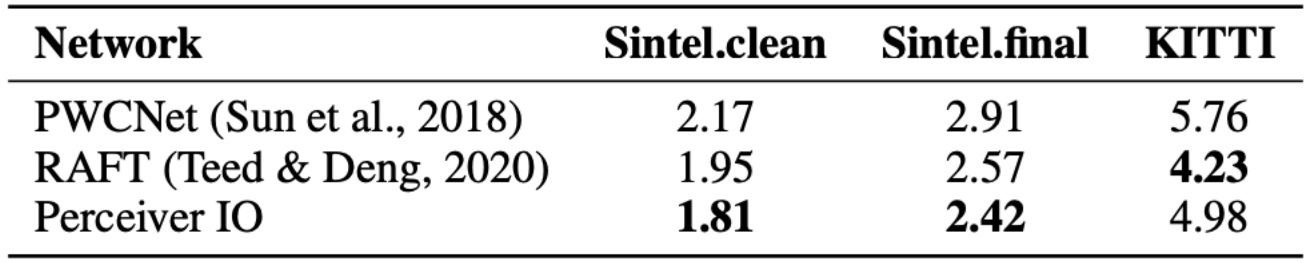
- The latent space is mapped to the output space, which also has a larger dimensionality than the latent space



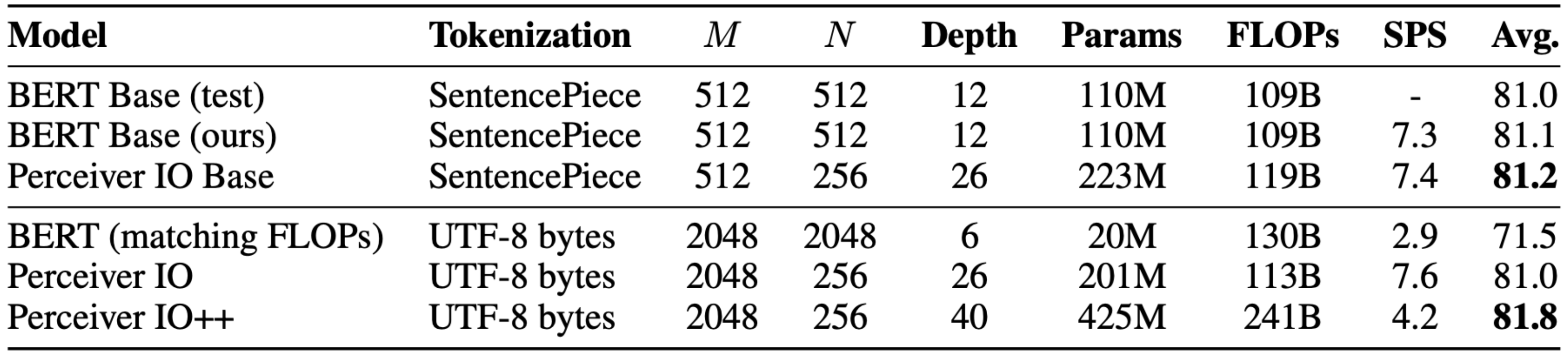
Perceiver IO – Enabling multimodal output

With the development of the output query vector, Perceiver IO can then be used to process and produce multimodal output. In the paper, the authors show SOTA or near SOTA in tasks like:

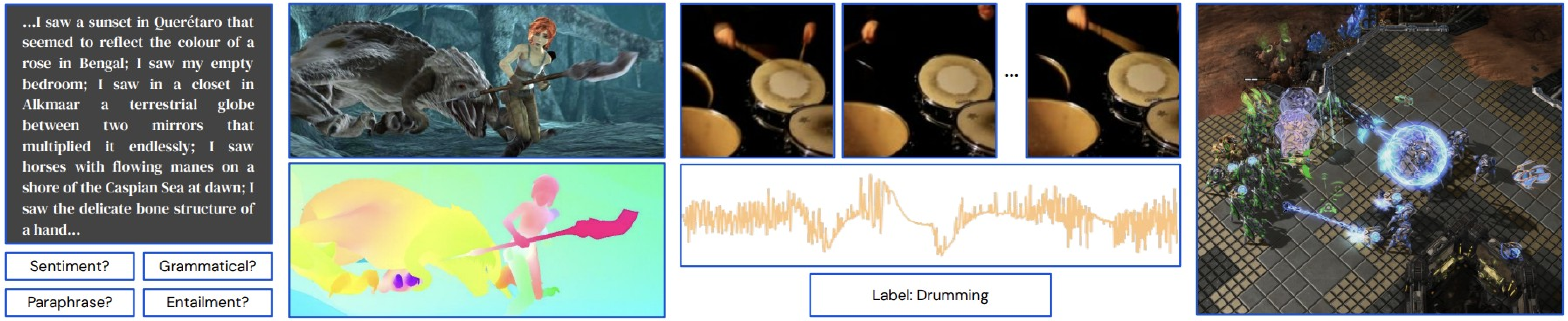
- Optical Flow
- Playing Starcraft
- Language Tasks
- Audio transcription
- Image captioning/labeling



Network	Sintel.clean	Sintel.final	KITTI
PWCNet (Sun et al., 2018)	2.17	2.91	5.76
RAFT (Teed & Deng, 2020)	1.95	2.57	**4.23**
Perceiver IO	**1.81**	**2.42**	4.98



Model	Tokenization	M	N	Depth	Params	FLOPs	SPS	Avg.
BERT Base (test)	SentencePiece	512	512	12	110M	109B	-	81.0
BERT Base (ours)	SentencePiece	512	512	12	110M	109B	7.3	81.1
Perceiver IO Base	SentencePiece	512	256	26	223M	119B	7.4	**81.2**
BERT (matching FLOPs)	UTF-8 bytes	2048	2048	6	20M	130B	2.9	71.5
Perceiver IO	UTF-8 bytes	2048	256	26	201M	113B	7.6	81.0
Perceiver IO++	UTF-8 bytes	2048	256	40	425M	241B	4.2	**81.8**

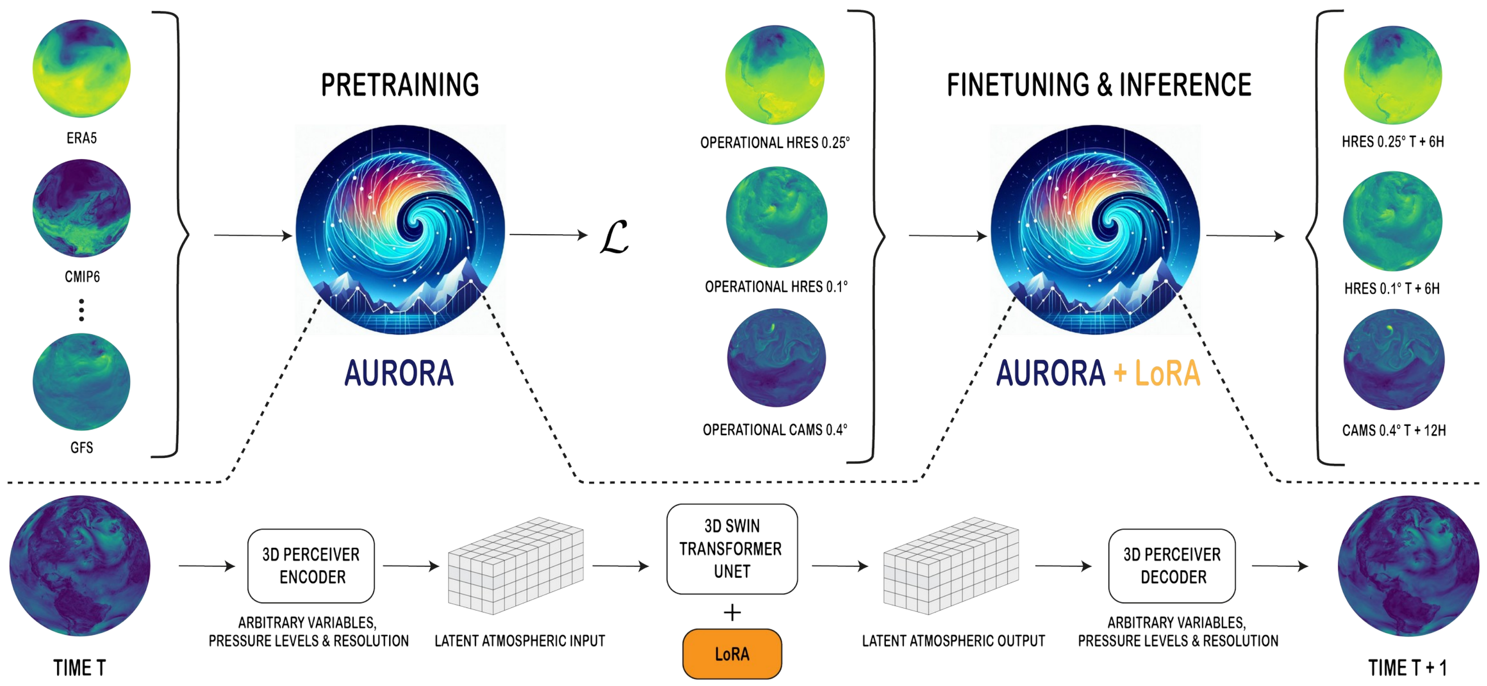


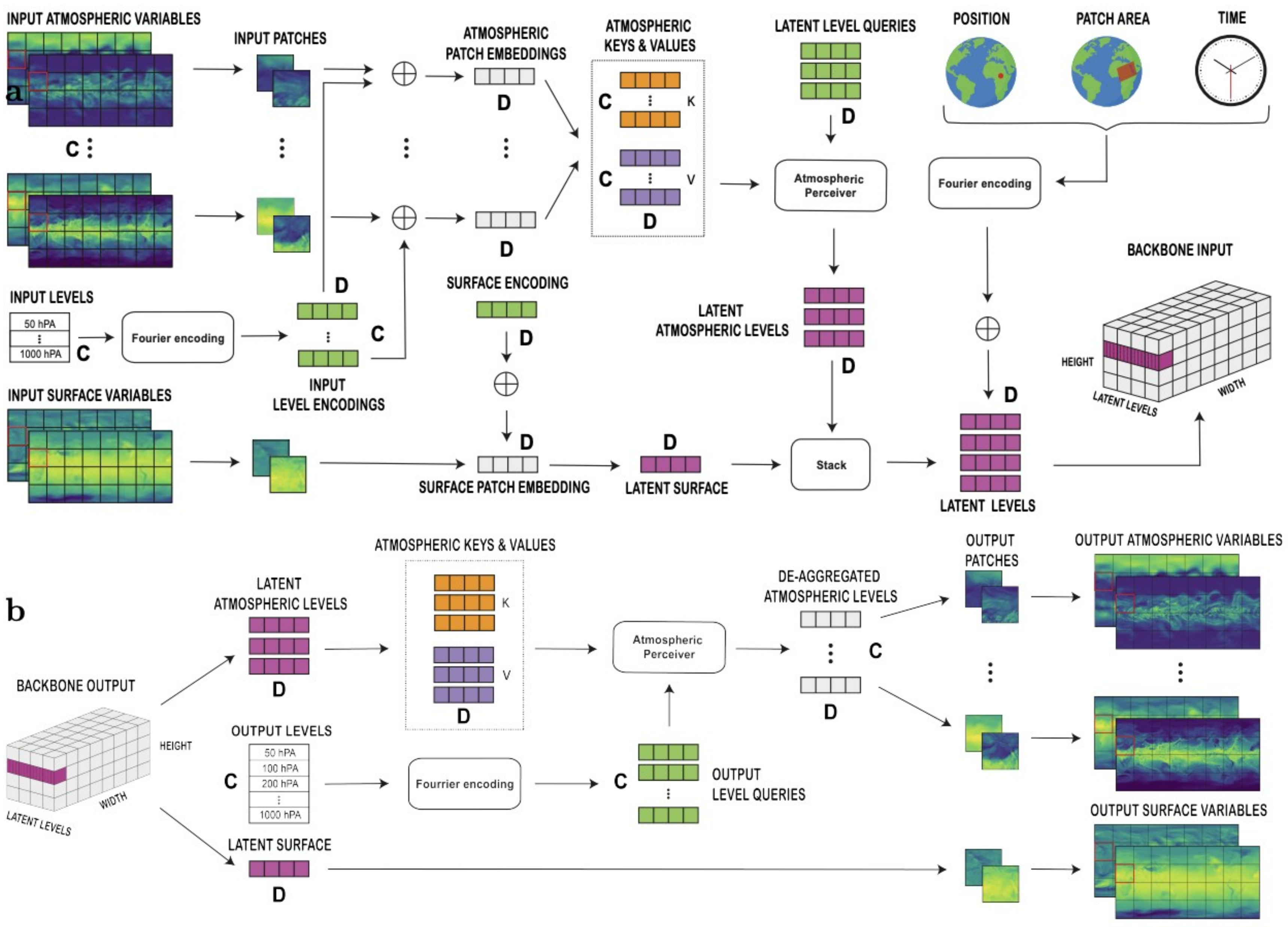
...I saw a sunset in Querétaro that seemed to reflect the colour of a rose in Bengal; I saw my empty bedroom; I saw in a closet in Alkmaar a terrestrial globe between two mirrors that multiplied it endlessly; I saw horses with flowing manes on a shore of the Caspian Sea at dawn; I saw the delicate bone structure of a hand...

Application of Perceiver IO - Aurora

Aurora is a climate forecasting model that utilizes Perceiver IO as a backbone, in conjunction with a UNet for time stepping, and enables production of multiple geospatial fields over time. This model is a massive foundation model that can be fine tuned to predict different climate variables across the globe.

Code here: https://github.com/microsoft/aurora

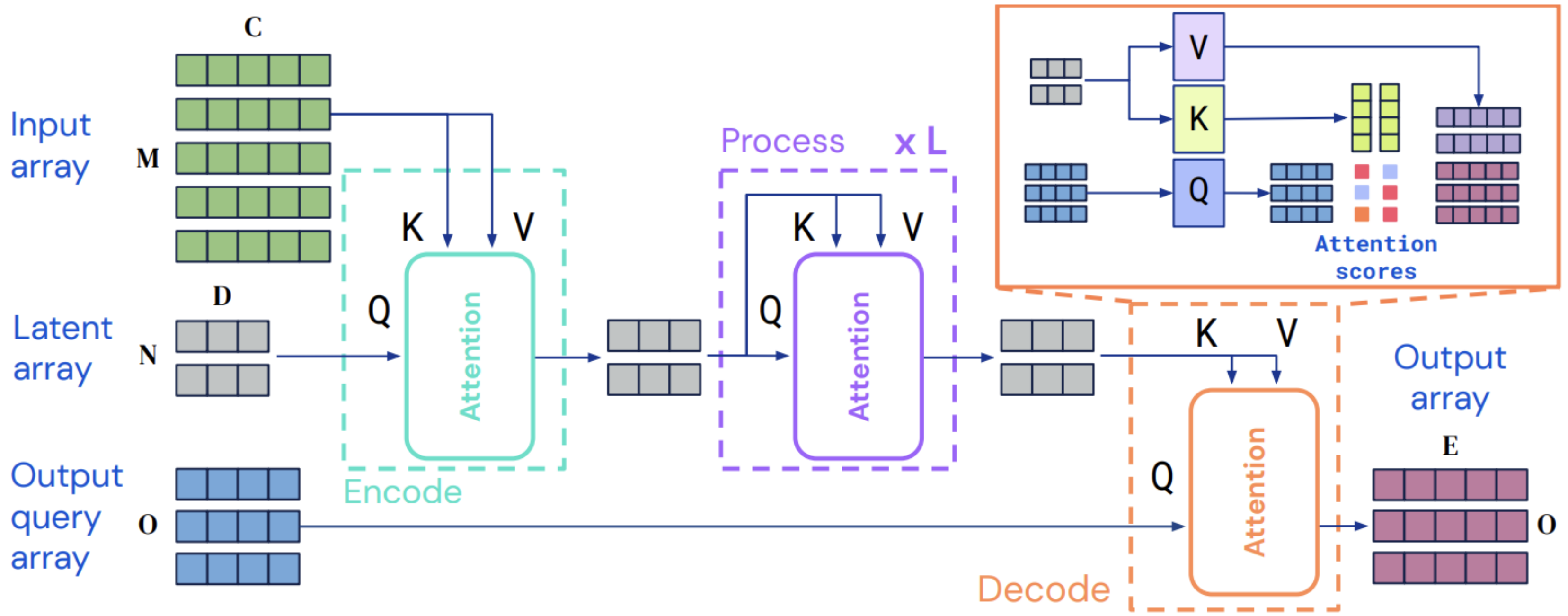




Wrap Up

Perceiver IO and more generalized Multimodal Models

- Today we wrapped up or multimodal module
- We saw the development of vison and language assistants with LLaVA 1.0/1.5 which makes use of vision and language encoders
- The Perceiver architecture which abstracts out the input modality to generalized attention processes, was introduced
- Perceiver IO, a further development of the Perceiver model to allow for multimodal outputs with the introduction of an output query vector.
- We explored an application of the Perceiver IO model to model climate simulations with the Aurora model.



Thank you!