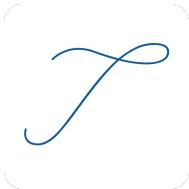
Fine-tune multimodal Vision models

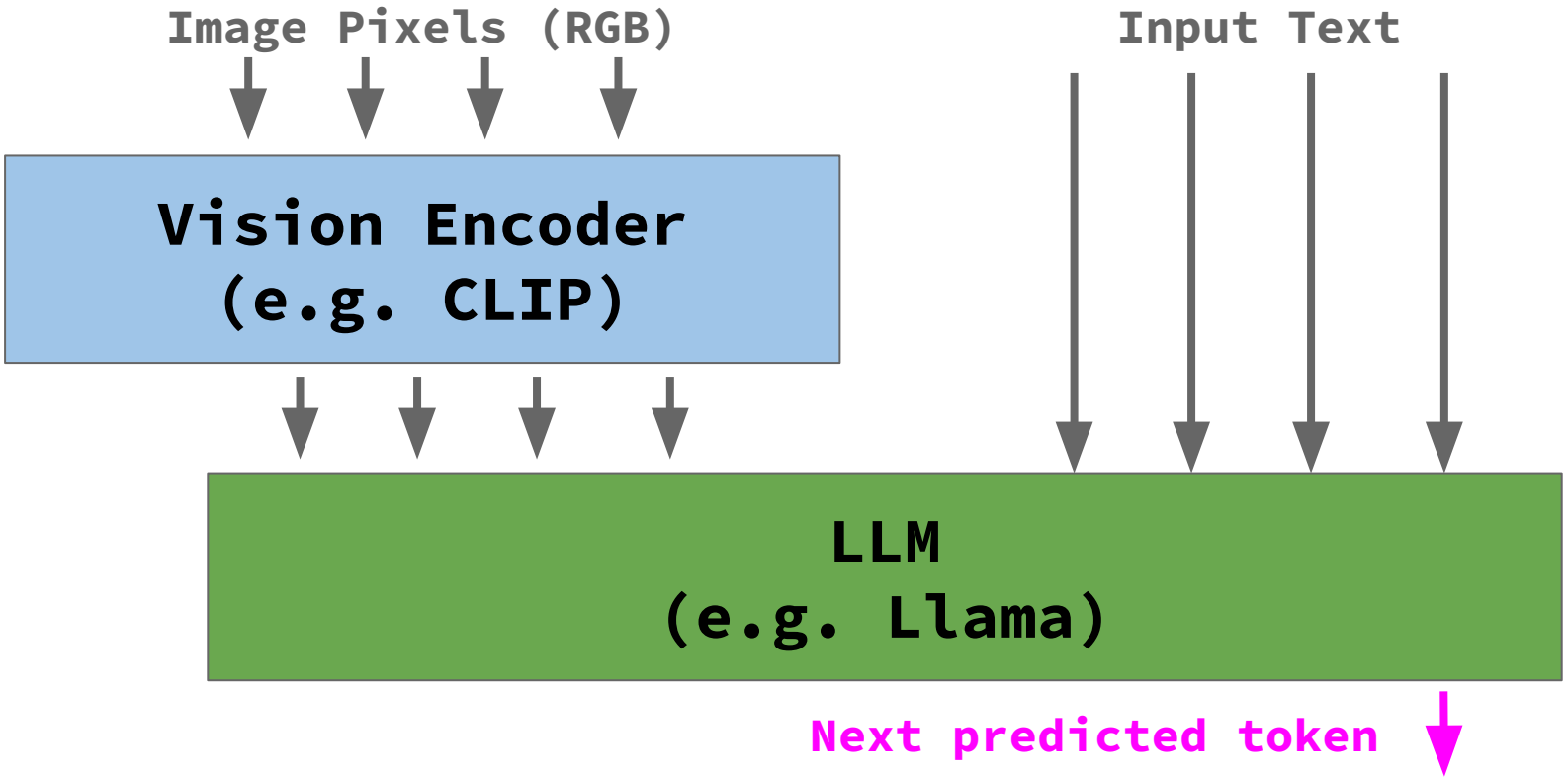
Trelis Research

Overview

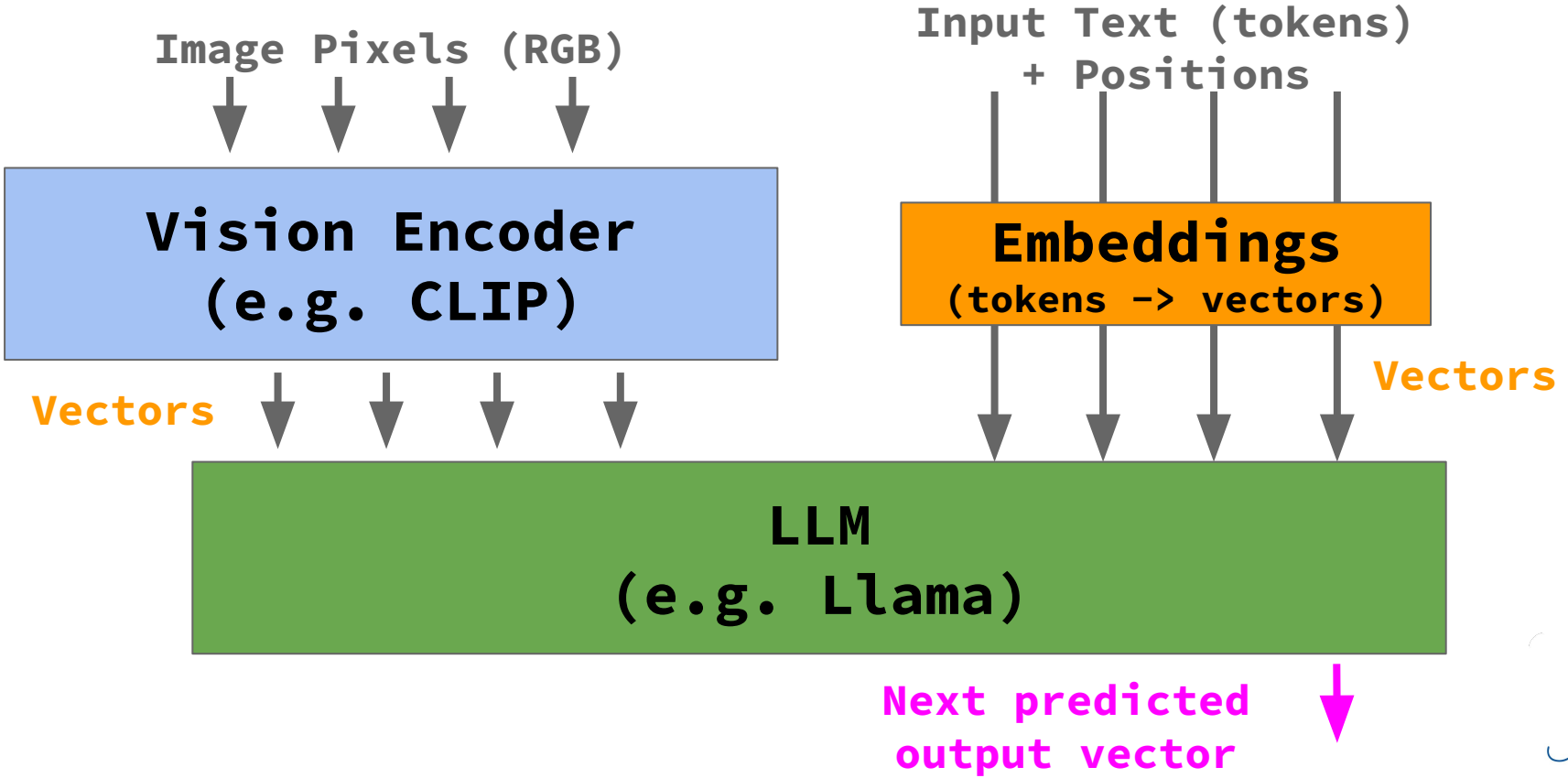
1. Demo
2. Applications.
3. Building a vision + text model.
4. Three specific architectures - LLaVA 1.5, LLaVA 1.6. IDEFICS.
5. Preparing a fine-tuning dataset.
6. Fine-tuning LLaVA 1.6
    a. Mistral 7B version
    b. Yi 34B
7. Final thoughts



Vision + Text models



Vision + Text models



Vision encoder

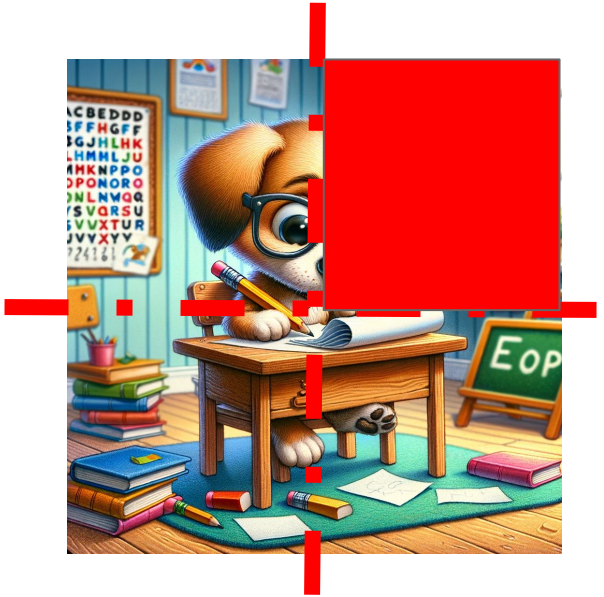
Image Pixels (RGB) + Positions

Embeddings
(pixels -> vectors)

Input Vectors

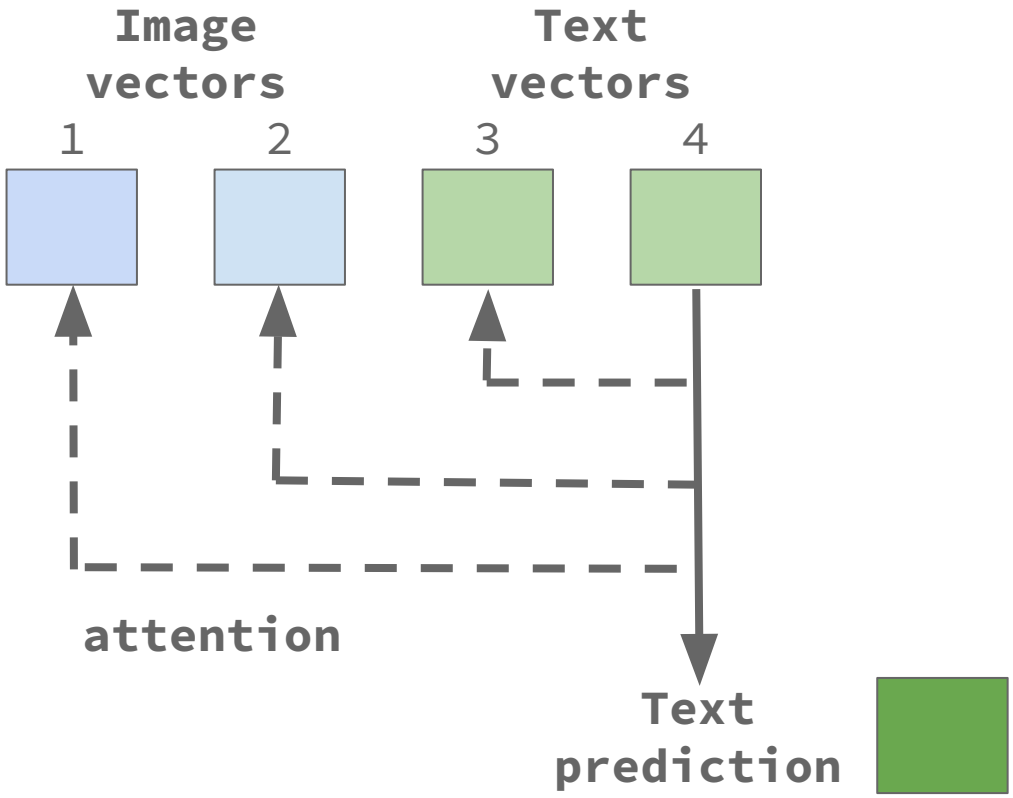
Vision Encoder
(Multi-layer neural net with “attention”)

Output Vectors

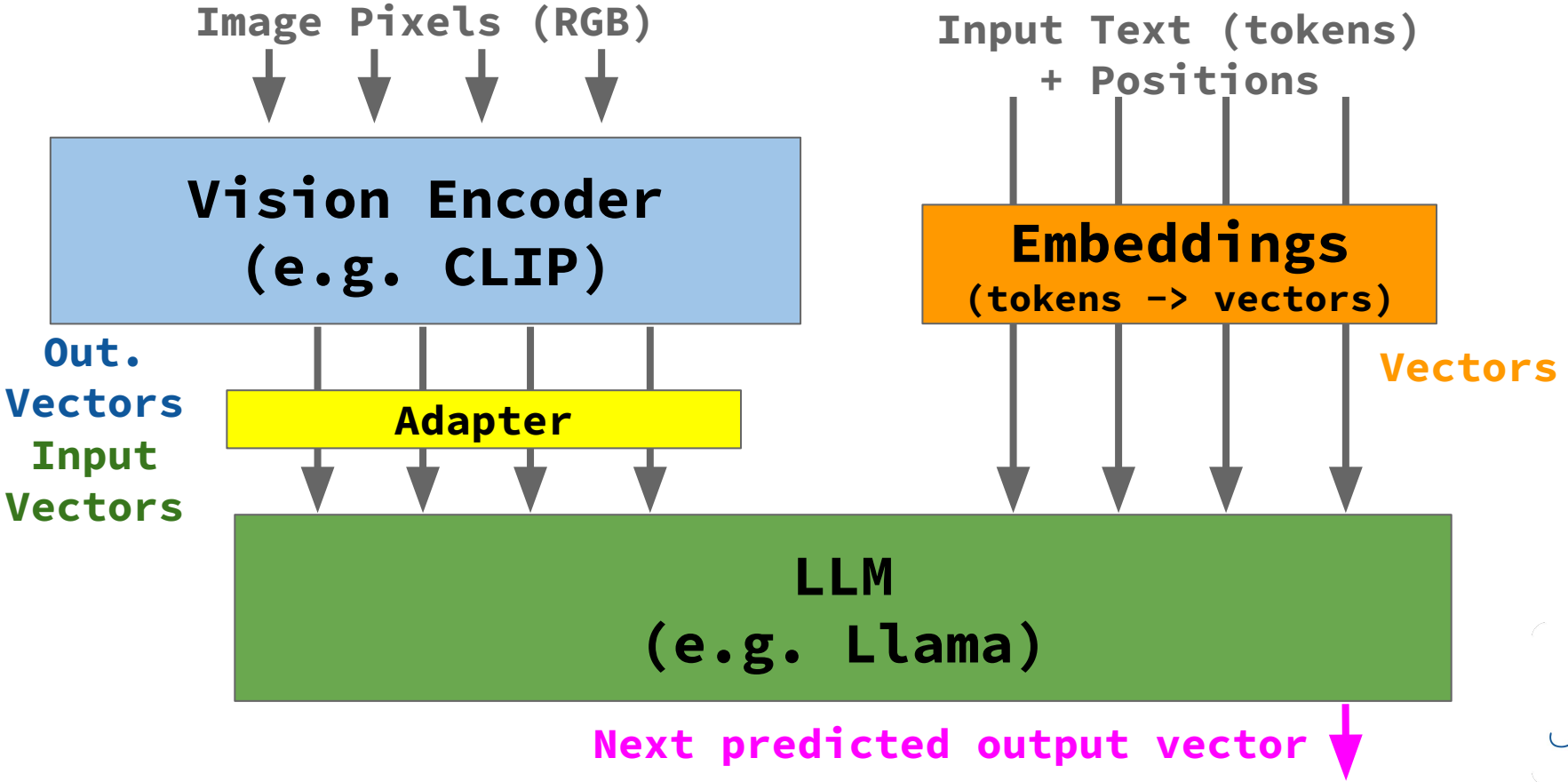


Each tile (16x16px)
= 1 input vector

Language Model decoder



Adding an adapter!



LLAVA 1.5 FORMULA

1. CLIP ENCODER + LLAMA 2 LLM
2. FREEZE THE VISION ENCODER + LLM => TRAIN THE ADAPTER
    A. USE IMAGE + TEXT DATASETS
3. UNFREEZE EVERYTHING => TRAIN EVERYTHING
    A. USE SYNTHETIC* IMAGE + TEXT DATA

*CHATGPT (NON VISION) IS USED TO WRITE DETAILED DESCRIPTIONS OF IMAGES GIVEN DETAILS OF THEIR CONTENTS.

LLAVA 1.6 FORMULA

1. CLIP ENCODER + MISTRAL 7B OR YI 34B LLM
2. USE AN MLP INSTEAD OF A SIMPLE LINEAR LAYER (E.G. LINEAR LAYER + ACTIVATION LAYER).
3. USE A LARGER VISION MODEL 336X336 NOT 224X224.

*CHATGPT (NON VISION) IS USED TO WRITE DETAILED DESCRIPTIONS OF IMAGES GIVEN DETAILS OF THEIR CONTENTS.

IDEFICS FORMULA

1. CLIP ENCODER + LLAMA -1 LLM.
2. FLAMINGO ARCHITECTURE: INJECT VISION VECTORS THROUGHOUT THE LLM.
3. TRAINING USING LONG MULTI-IMAGE DOCUMENTS.

FLAMINGO PAPER: HTTPS://ARXIV.ORG/PDF/2204.14198.PDF

IDEFICS MODEL (80B):
HTTPS://HUGGINGFACE.CO/HUGGINGFACEM4/IDEFICS-80B-INSTRUCT