VLMs and Multi-Modal RAG: What Does the Future Hold?

Natdhanai Praneenatthavee
AI engineer at AXONS, MSc. Aritificial Intelligence

Me?

Natdhanai Praneenatthavee (JOB)

ประวัติ
Msc. Artificial Intelligence (KMITL)
ตีพิมพ์ 2 papers ระดับ นานาชาติ ด้าน VLMs
AI Software engineer (Research + Production > 5 ปี)
Speaker Tensorflow, Thaipy
สอนนักเรียนปริญญาโทด้าน DataSci ที่อังกฤษ
สอนกลุ่ม เกี่ยวกับ Computer Vision
ผู้ช่วยสอน LLMs ที่ KMITL, SEAGATE
เจ้าของเพจ ตื่นมาโค้ด python
Part-time Medium Writer

สามารถพบเจอได้ในงาน AI, Tech เป็นครั้งคราว

Agenda

9:00 – 9:45	Introduction LLMs, VLMs
9:45 – 10:00	Q&A Break
10:00 – 10:45	Multi Modal RAG
10:45 – 11:15	Example Research(Code) , Q&A
11:15 – 12:00	Discussing, AI VLMs Career

Participate

Just chill and relax

ทำตัวให้สบายๆ ไม่ต้องกดดันแล้วมาฟังสิ่งที่ยิ่งใหญ่

AI TOOLS

ChatGPT

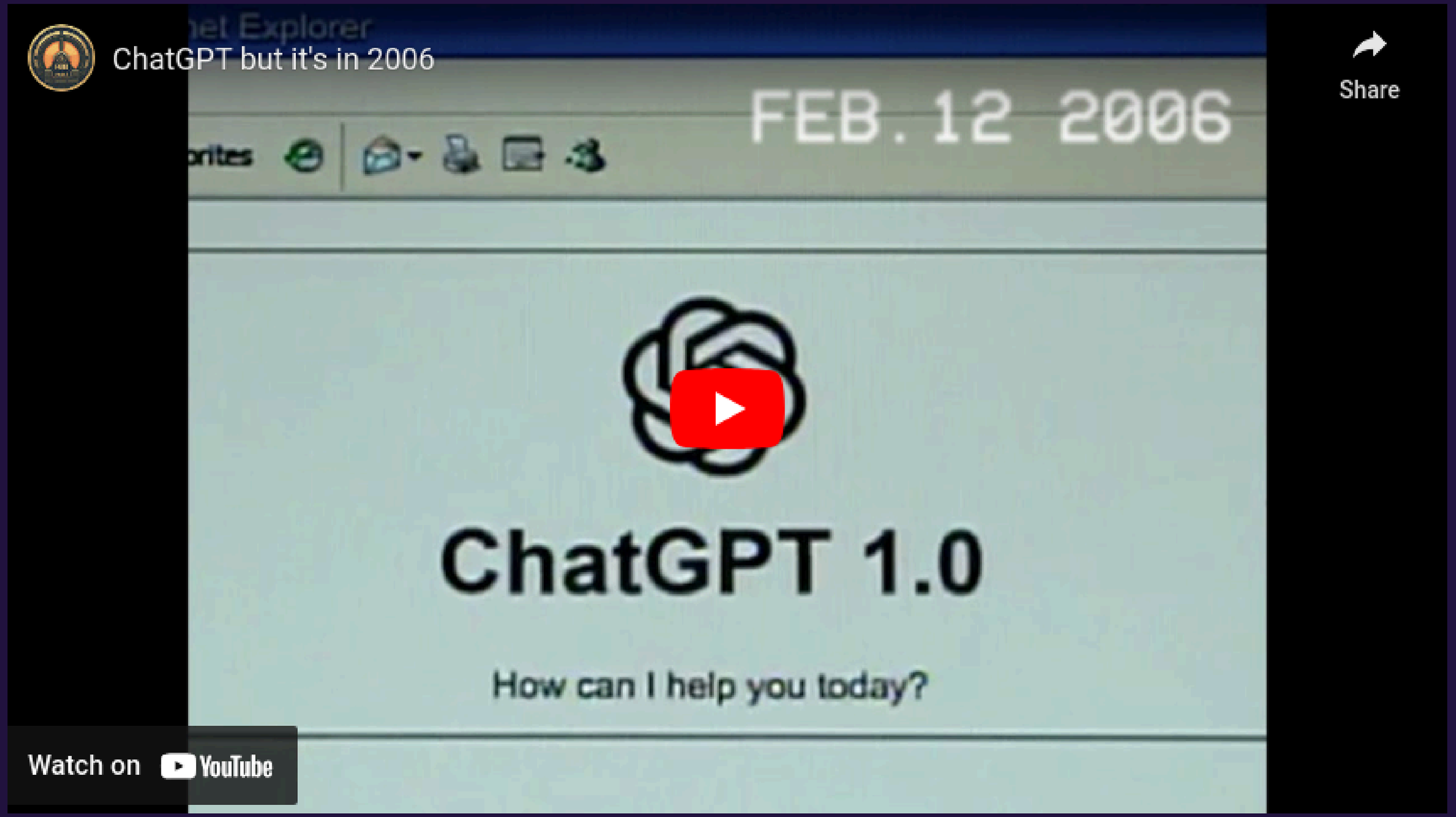
Claude

BY ANTHROP\C

perplexity

Gemini

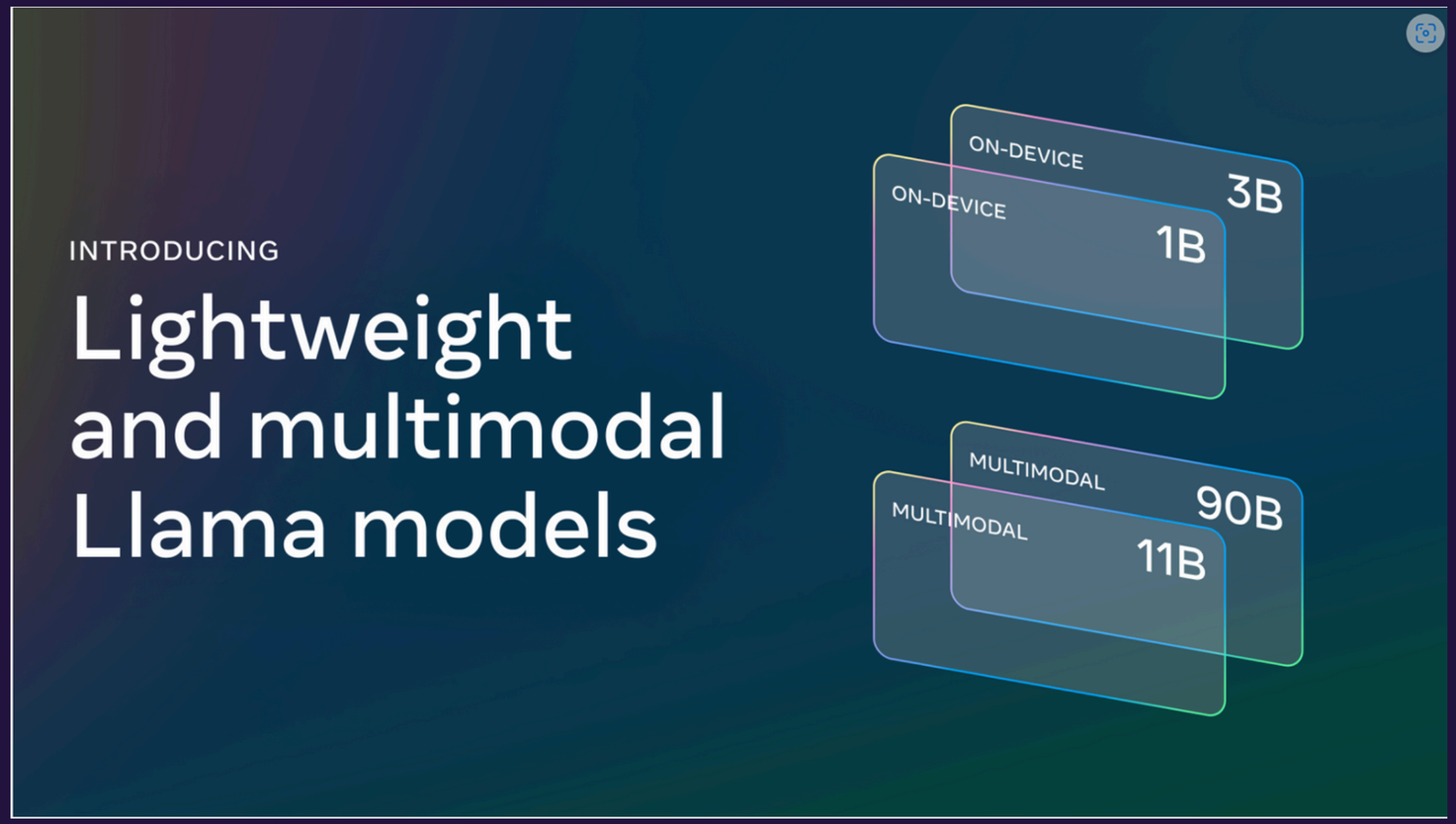
???


ChatGPT but it's in 2006
Share
FEB. 12 2006
ChatGPT 1.0
How can I help you today?
Watch on YouTube

Deep Drive into AI

Model	Architecture	Parameter count	Training data	Release date	Training cost
GPT-1	12-level, 12-headed Transformer decoder (no encoder), followed by linear-softmax.	117 million	BookCorpus:[35] 4.5 GB of text, from 7000 unpublished books of various genres.	June 11, 2018[9]	30 days on 8 P600 GPUs, or 1 petaFLOP/s-day.[9]
GPT-2	GPT-1, but with modified normalization	1.5 billion	WebText: 40 GB of text, 8 million documents, from 45 million webpages upvoted on Reddit.	February 14, 2019 (initial/limited version) and November 5, 2019 (full version)[36]	"tens of petaflop/s-day",[37] or 1.5e21 FLOP.[38]
GPT-3	GPT-2, but with modification to allow larger scaling	175 billion[39]	499 billion tokens consisting of CommonCrawl (570 GB), WebText, English Wikipedia, and two books corpora (Books1 and Books2).	May 28, 2020[37]	3640 petaflop/s-day (Table D.1[37]), or 3.1e23 FLOP.[38]
GPT-3.5	Undisclosed	175 billion[39]	Undisclosed	March 15, 2022	Undisclosed
GPT-4	Also trained with both text prediction and RLHF; accepts both text and images as input. Further details are not public.[34]	Undisclosed. Estimated 1.7 trillion.[40]	Undisclosed	March 14, 2023	Undisclosed. Estimated 2.1×10^{25} FLOP.[38]

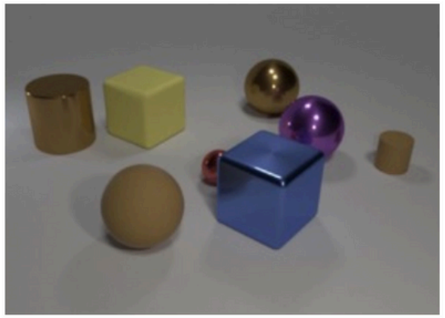
LLMs VS VLMs


INTRODUCING
Lightweight and multimodal Llama models
ON-DEVICE
3B
ON-DEVICE
1B
MULTIMODAL
90B
MULTIMODAL
11B


Qwen2-VL


OpenAI
GPT-4o
Audio
vision
Text

Applications?



Prompt: How many small spheres are the same color as the big rubber cube?

Ovis: There are no small spheres the same color as the big rubber cube.

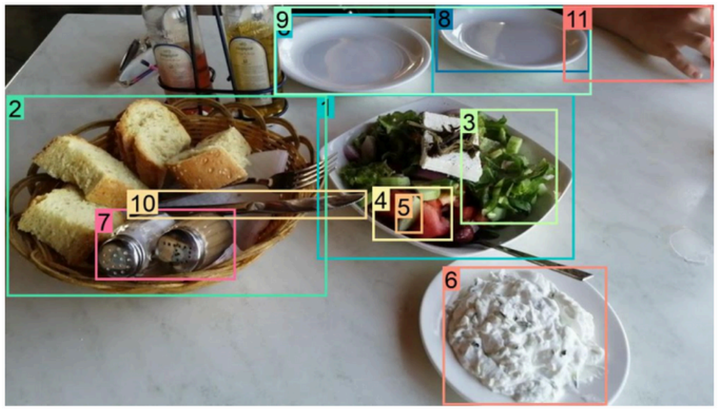
Prompt: Who is the person wearing jersey number 23 in the picture?

Ovis: The person wearing jersey number 23 is LeBron James.

Prompt: Count the number of books in the image.

Ovis: There are five books in the image.

OVIS



The image shows a table with a basket of bread and a plate of salad[1]. The basket[2] is made of woven straw and has several slices of bread in it. Next to the basket, there is a small bowl[1] of salad with a variety of vegetables, including lettuce[3], tomatoes[4], cucumbers[5], and feta cheese[6]. There are also two salt[7] and pepper shakers[7] on the table. On the right side of the table, there are two white plates[8,9] with a dollop of white sauce on them. The table is covered with a white tablecloth and there are a few other dishes[9] and utensils[10] scattered around. A person"s hand[11] can be seen in the top right corner of the image.

Florence2

Let’s Try (Code After)

Concept first coding after

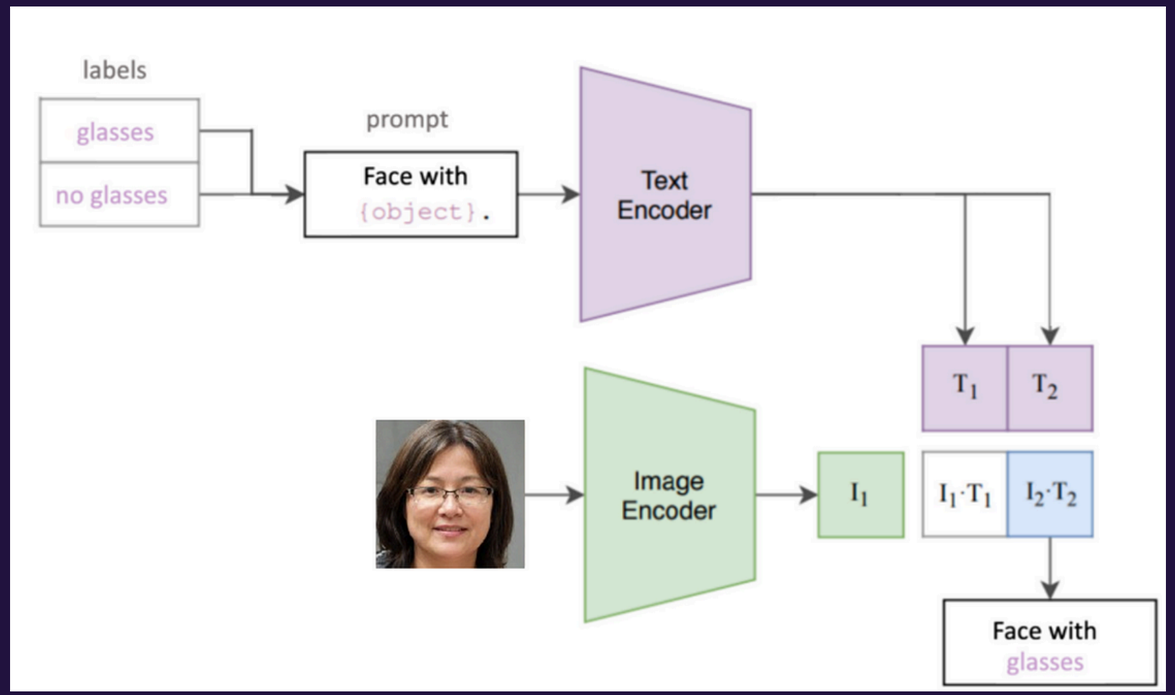
LLama3.2

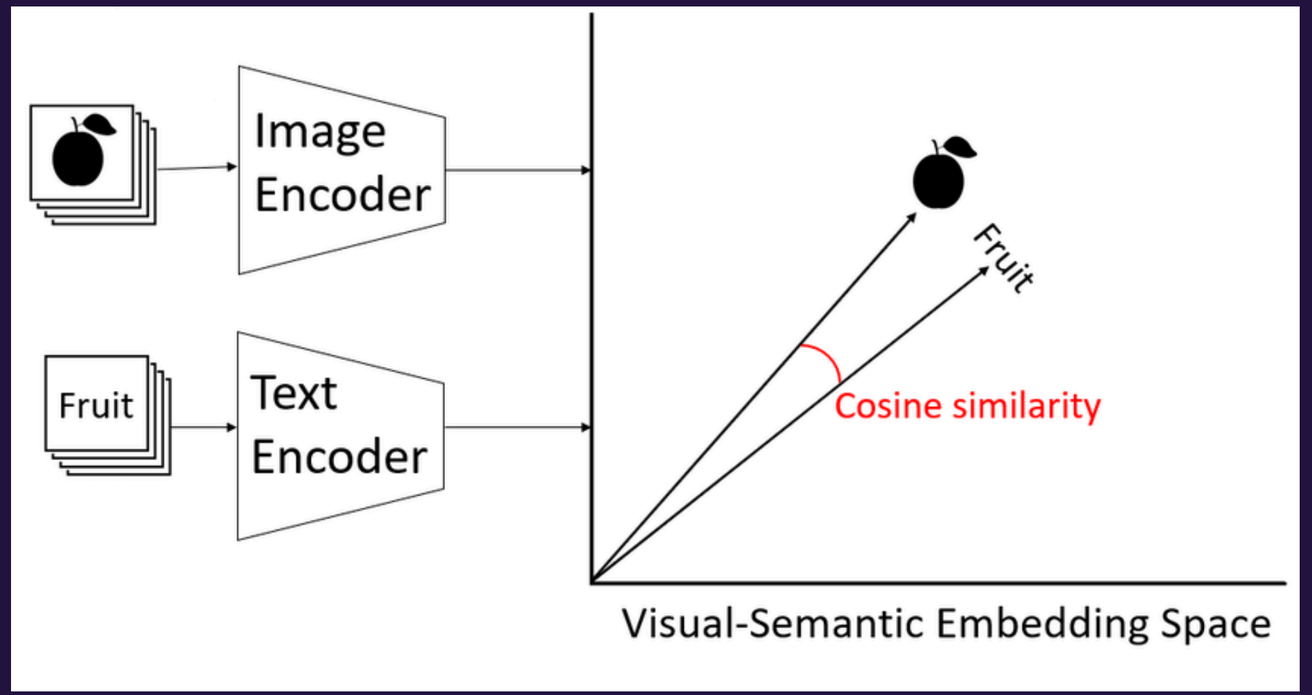
Ovis

Qwen2VL

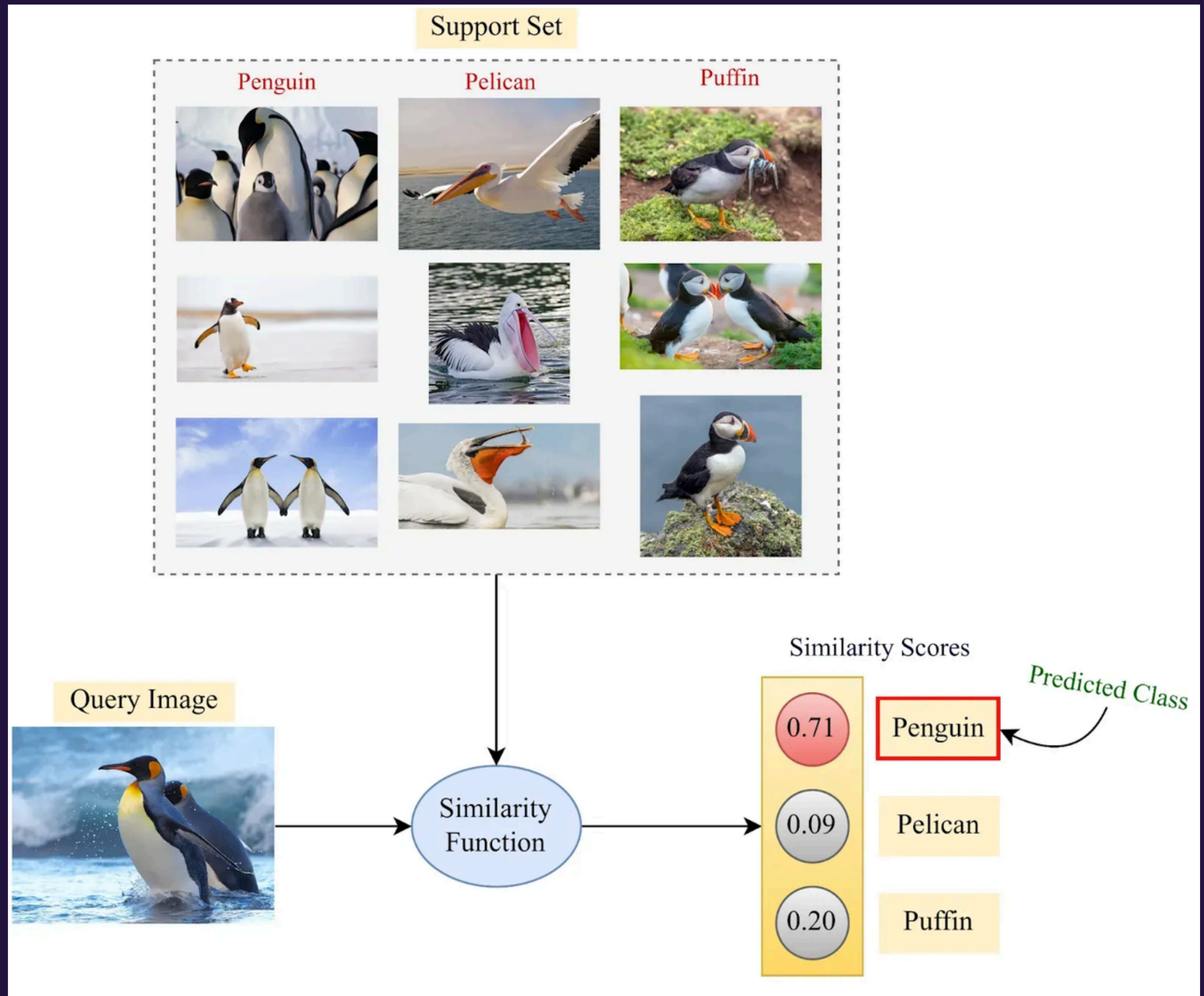
Florence2

Zero-Shot Learning

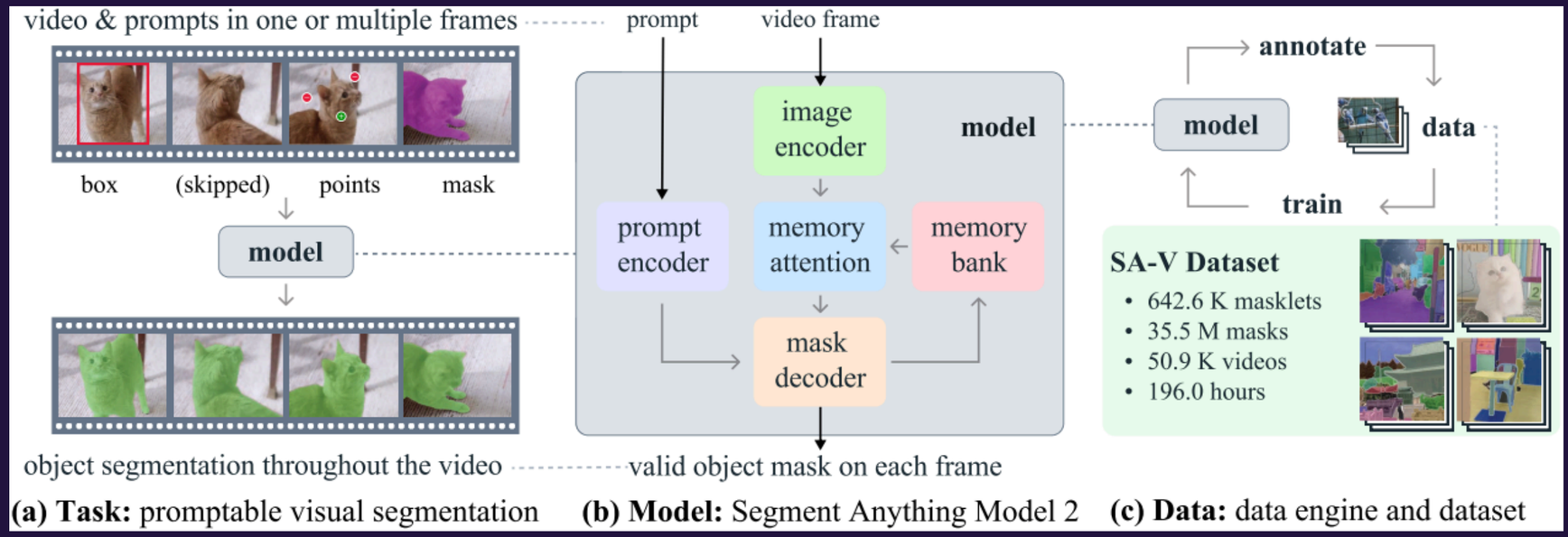


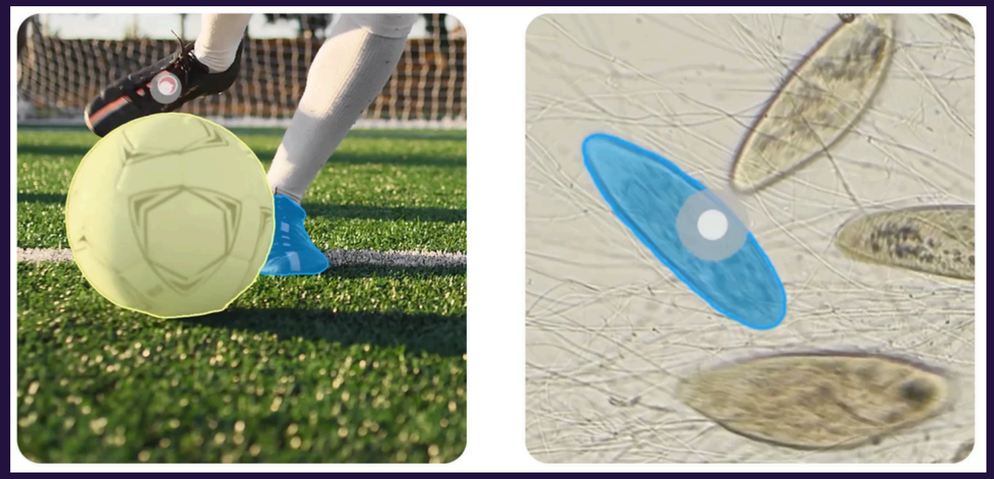


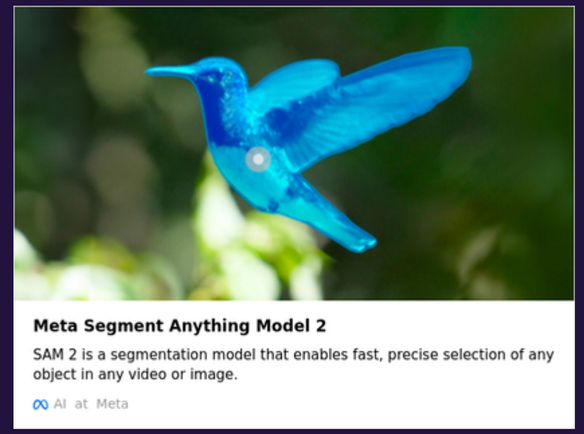
Few-Shot Learning



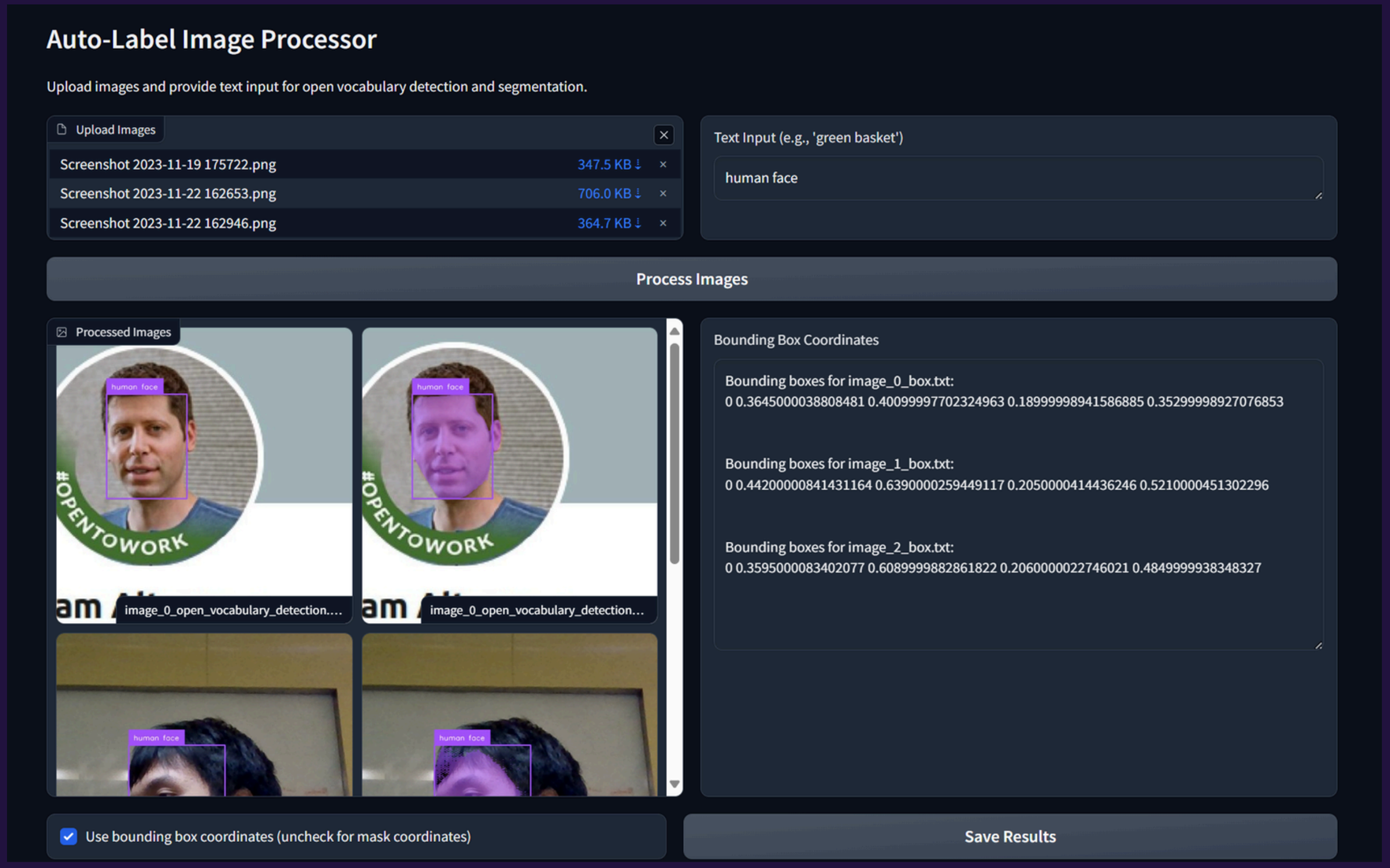
Segment Anything Model2!!



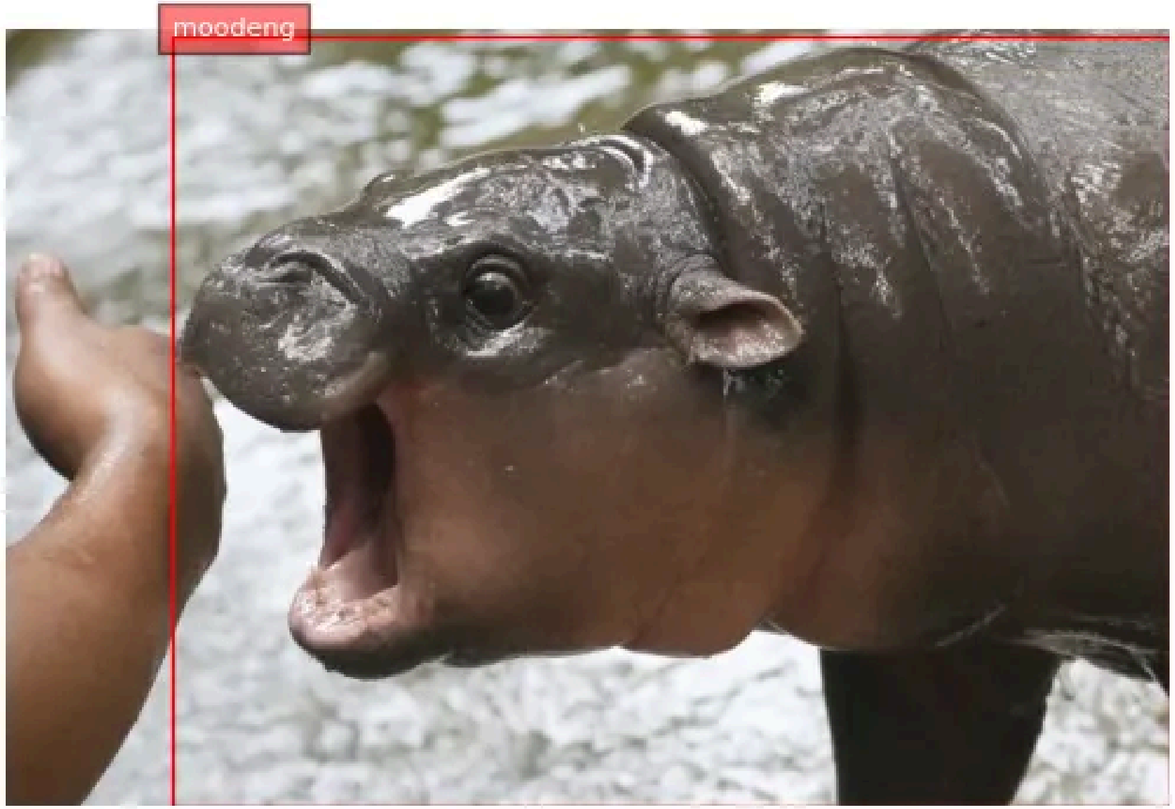




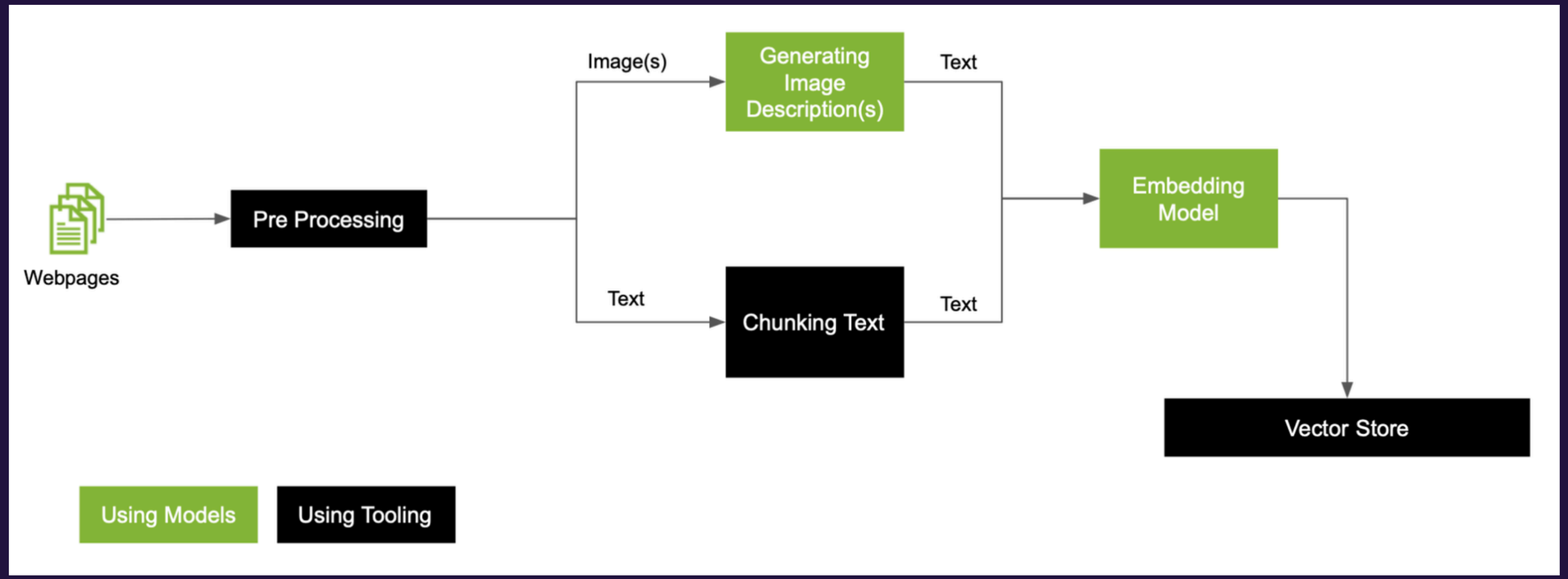
Auto-Labeling?



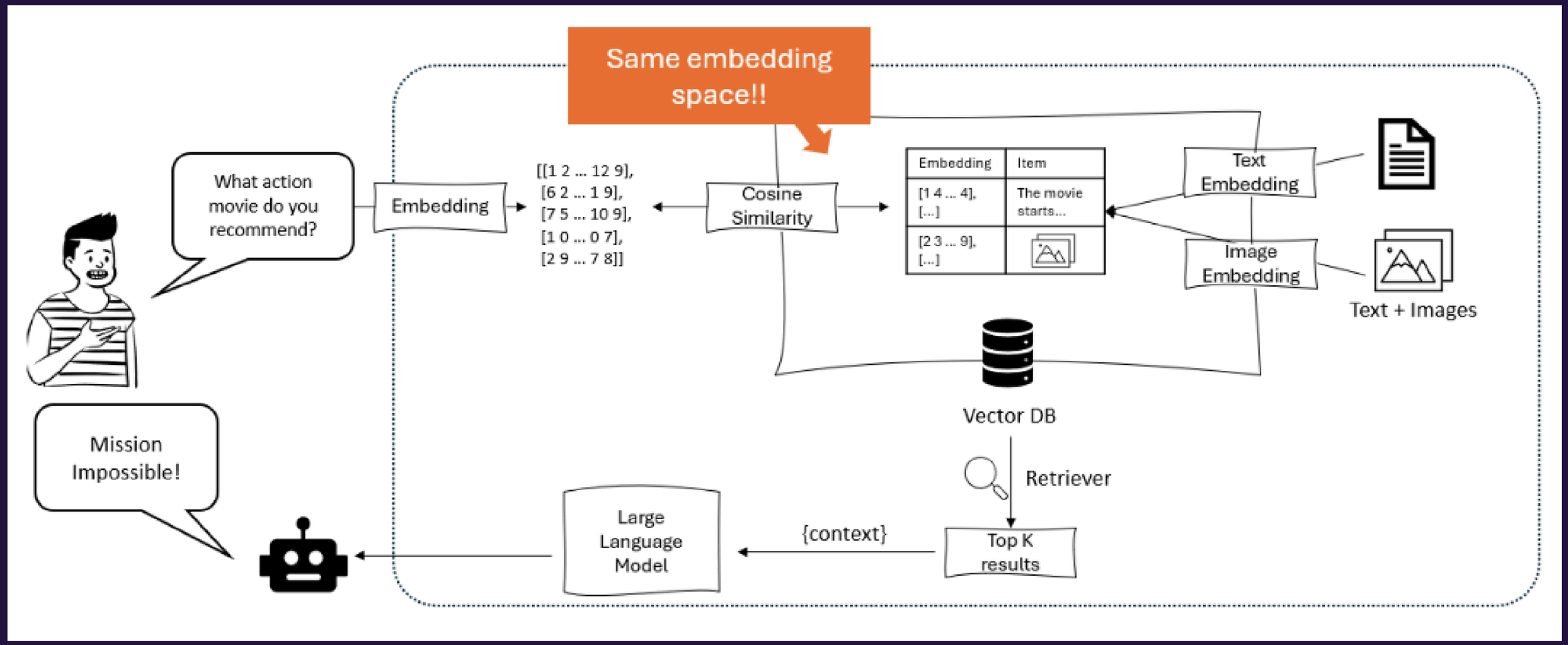
Q&A Break



Multi Modal RAG



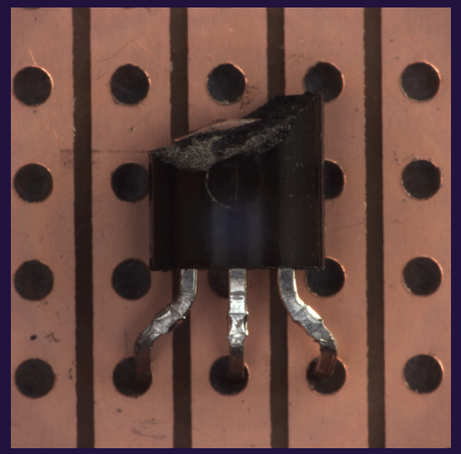
Multi Modal RAG

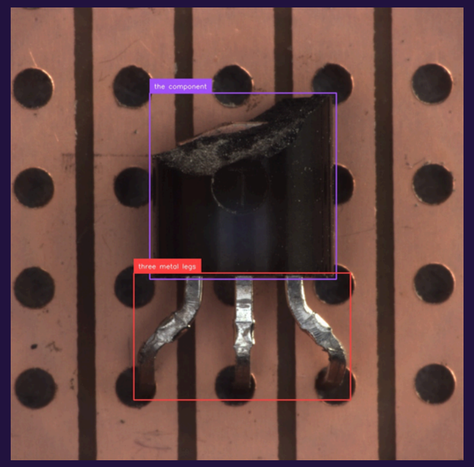


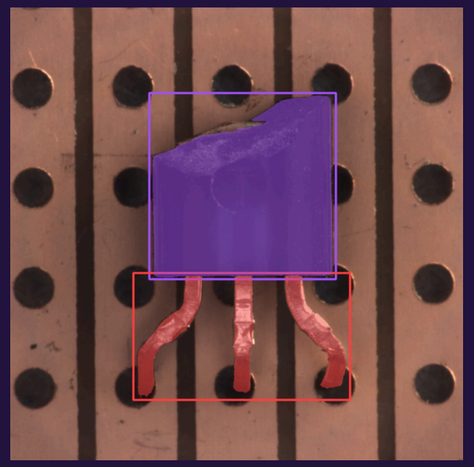
Let’s Cook

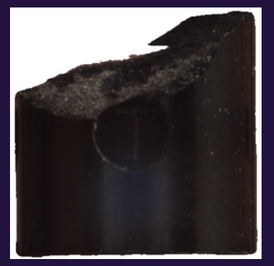
Code Section

Zero-/Few-Shot Anomaly Classification for Transistor Using Multimodal CLIP Retrieval-Augmented

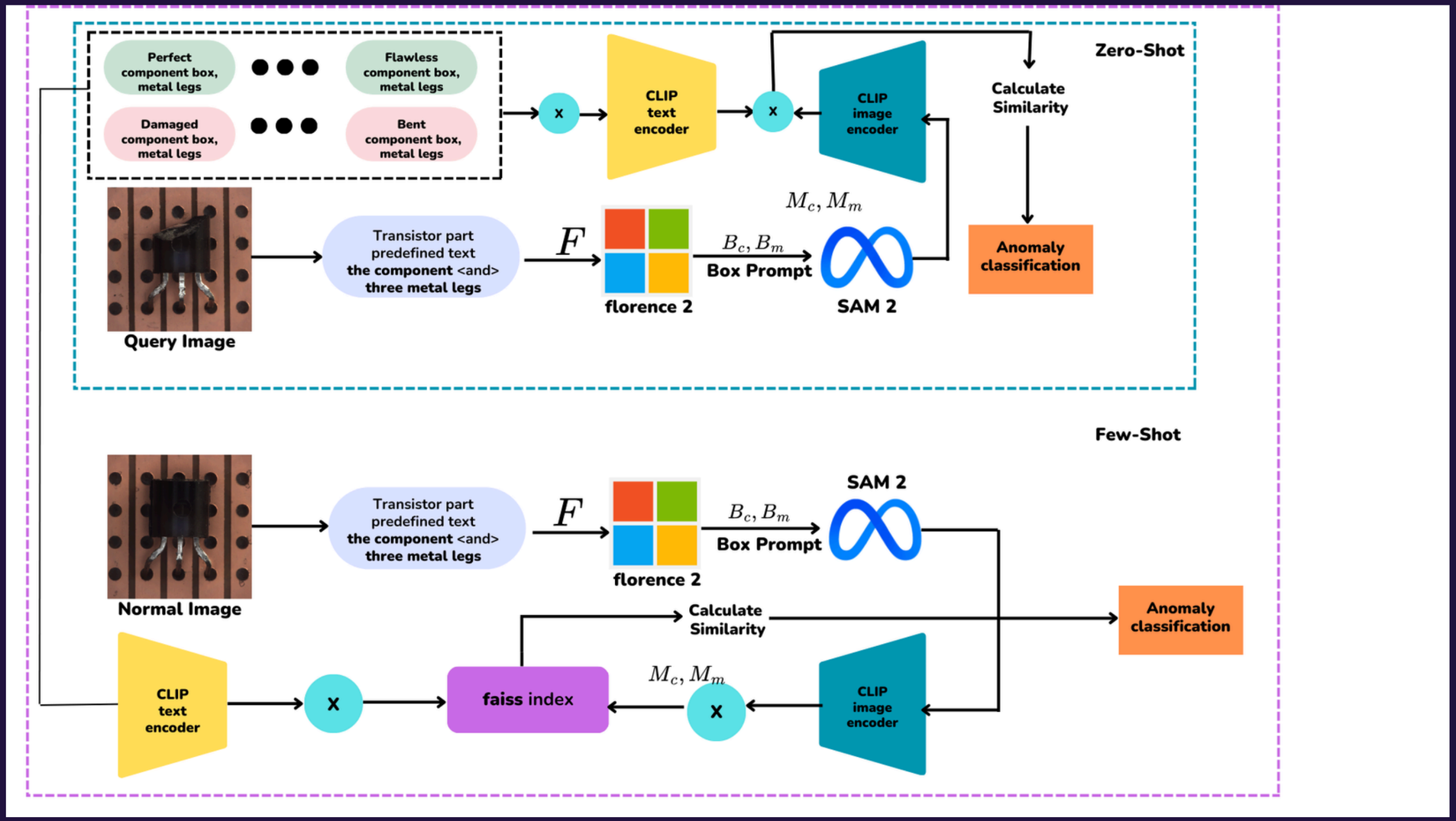








Let’s Cooking



AI VLMs Carrer?

Meta

AI Research Engineer, VLLM (vision large language models) - Generative AI

Menlo Park, CA · 1 week ago · 58 applicants

$85.10/yr - $251K/yr · Full-time

AMD

ML Frameworks Software Development Engineer - vLLM

Austin, TX · 1 day ago · 54 applicants

$160.8K/yr - $241.2K/yr · On-site · Full-time · Mid-Senior level

Refrences

What is Few-Shot Learning?

In this blog post, we discuss what few-shot learning is, architectural approaches for implementing few-shot learning, and specific implementations of few-shot learning techniques.

Roboflow Blog / Apr 15

CLIP: Creating Image Classifiers Without Data

A hands-on tutorial explaining how to generate a custom Zero-Shot image classifier without training...

Towards Data Science / Mar 1, 2023

FIGURE 1: Schematic view of the zero-shot classification methodology of...

Download scientific diagram | Schematic view of the zero-shot classification methodology of CLIP from publication: Implicit Stereotypes in Pre-Trained Classifiers | Pre-trained deep learning models underpin...

ResearchGate