



# Programmatic Querying of Large (Vision) Language Models for Research

Prateek Verma, Xintao Wu

Oct 02, 2024

Arkansas Integrative Metabolic Research Center

# Agenda and Objectives

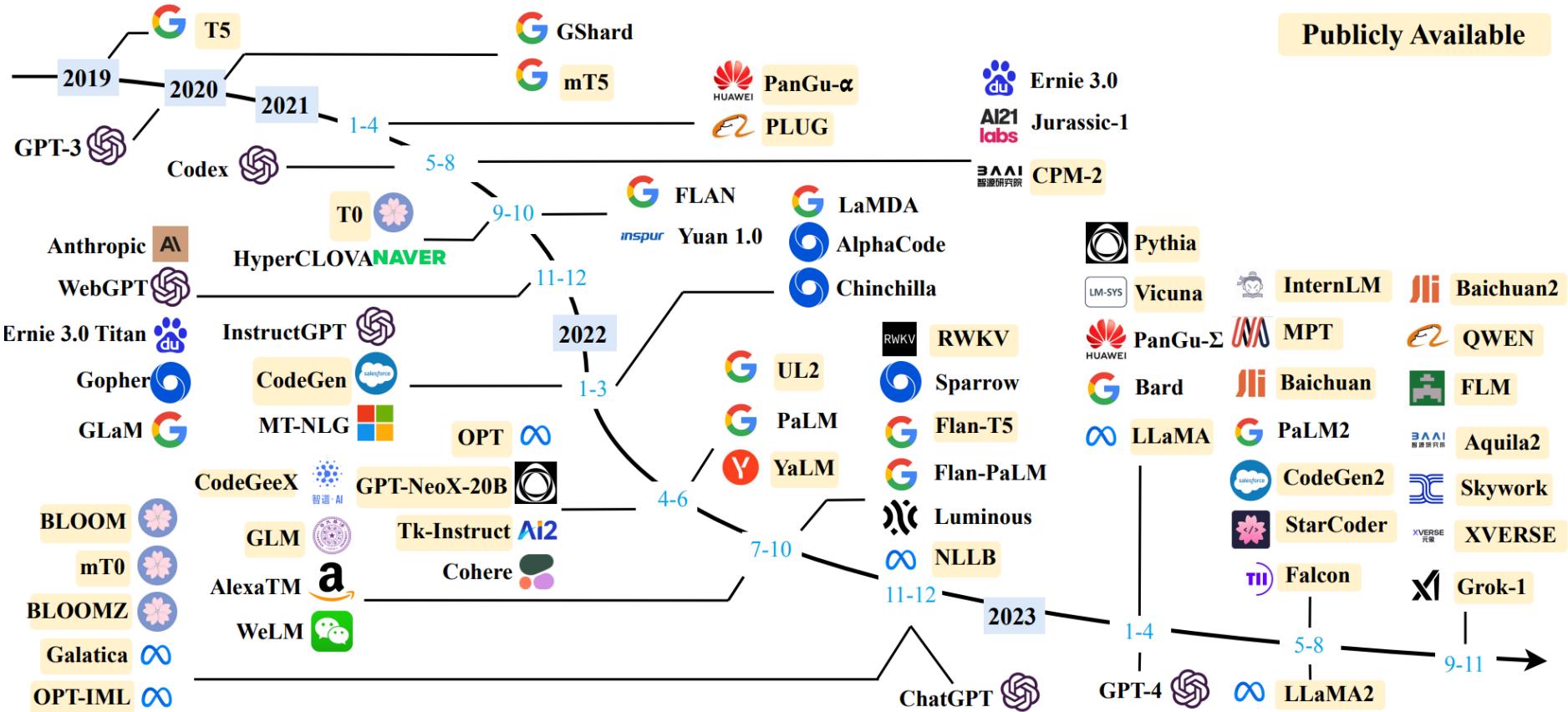
## □ Agenda

- Overview of LLMs and VLMs.
- Introduction to key VLM models:  
ChatGPT, BLIP, CLIP, LLaVA, SAM.
- Demonstrations and exercises.

# Understanding LLMs and VLMs

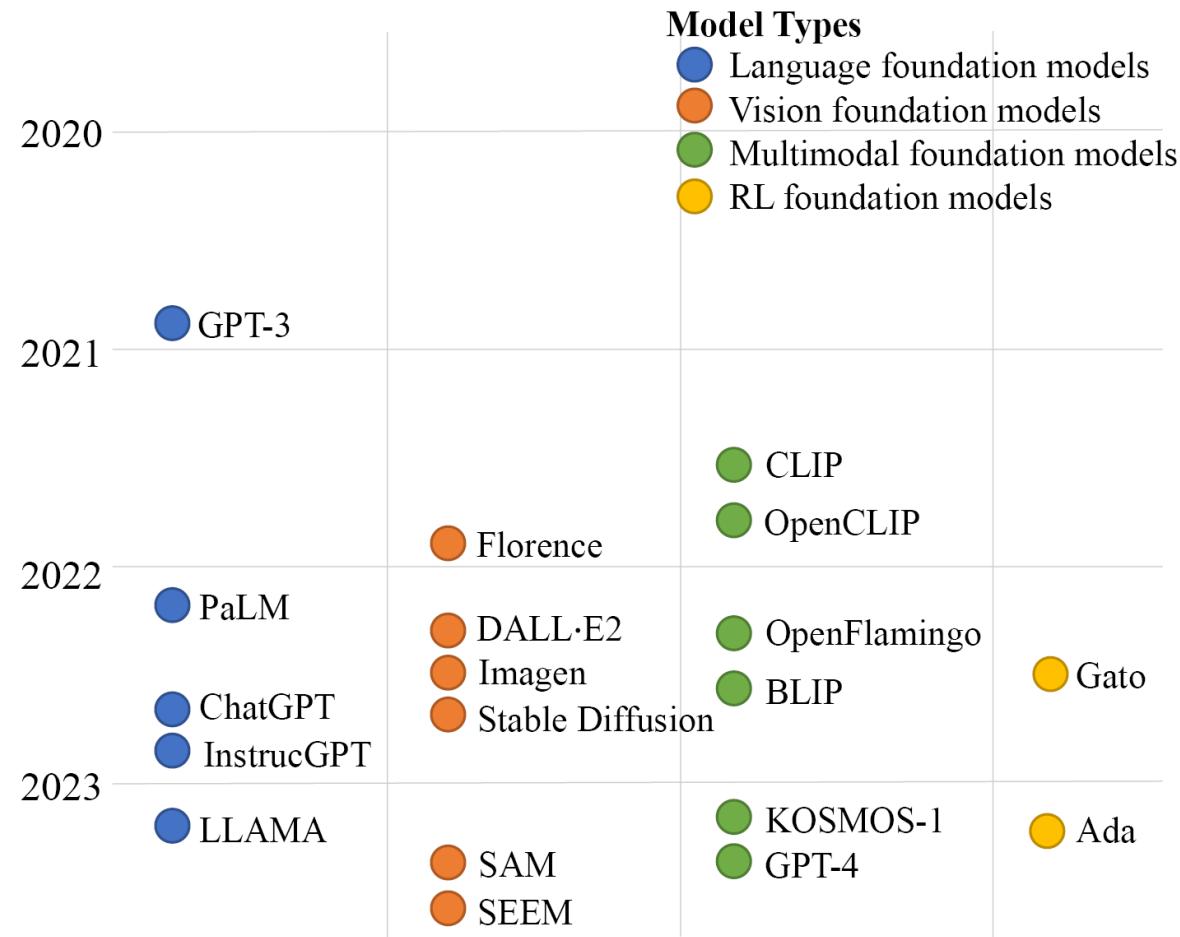
- **LLMs (Large Language Models)**
  - AI models trained on vast amounts of text data.
  - Examples: GPT, LLaMA, BERT.
  - Task examples: Text generation, translation, Q&A.
  
- **VLMs (Vision-Language Models)**
  - Combine visual and language understanding.
  - Can process both text and images for multimodal tasks.
  - Examples: BLIP, LLaVA, MiniGPT-4, SAM.

# Understanding LLMs and VLMs



# Understanding LLMs and VLMs

## Foundation Models Timeline



# Medical Foundation Models

## Language Models

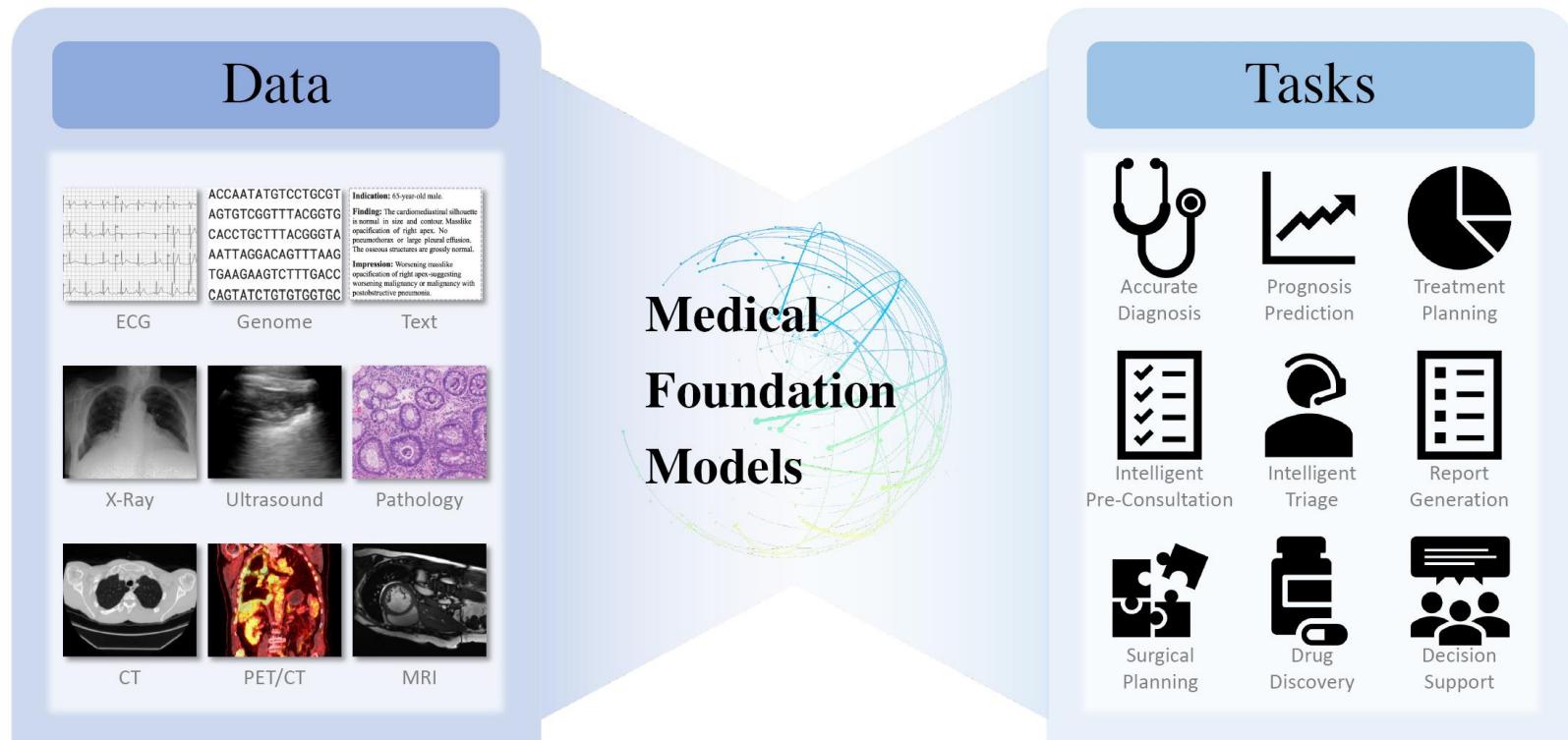
ChatGPT, GPT-4,  
Bard, PaLM

## Vision Language Models

GPT-4V, Flamingo,  
LLaVA

## Medical Vision Language Models

BiomedCLIP,  
Med-PaLM

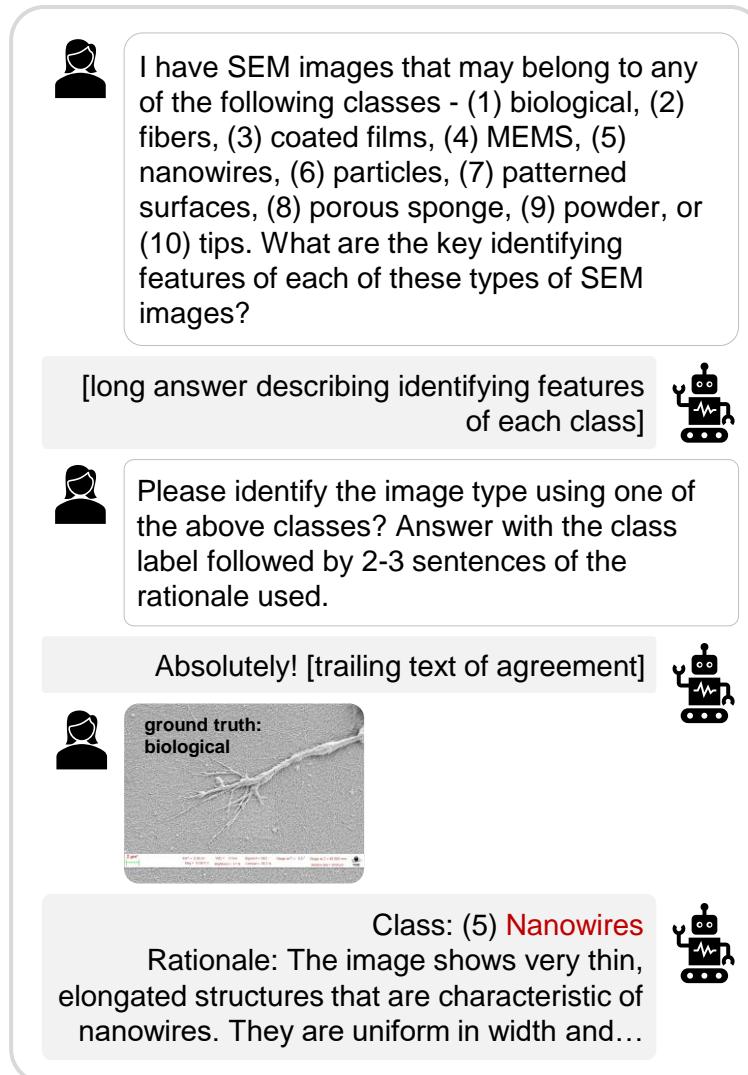


# Key Vision-Language Models

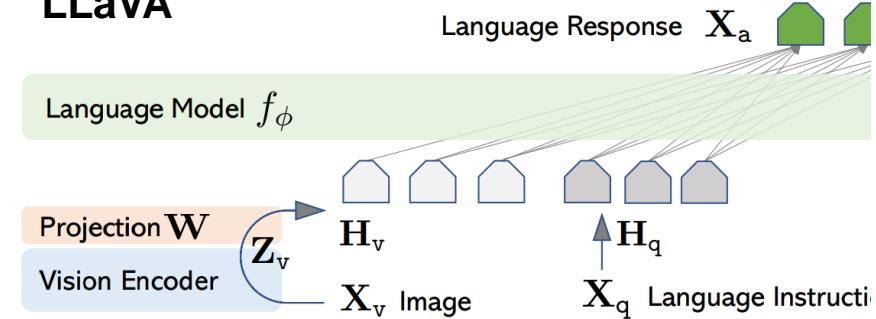
- ❑ **BLIP**: Bootstrapped Language-Image Pretraining model for image captioning.
- ❑ **LLaVA**: Large Language and Vision Assistant for visual question answering.
- ❑ **MiniGPT-4**: Lightweight version of GPT-4, combines vision and language for reasoning.
- ❑ **SAM**: Segment Anything Model for image segmentation tasks.
- ❑ **ChatGPT**: Large-scale LLM for natural language tasks.
- ❑ **OpenFlamingo**: Advanced vision-language model for few-shot learning.

# Key Vision-Language Models

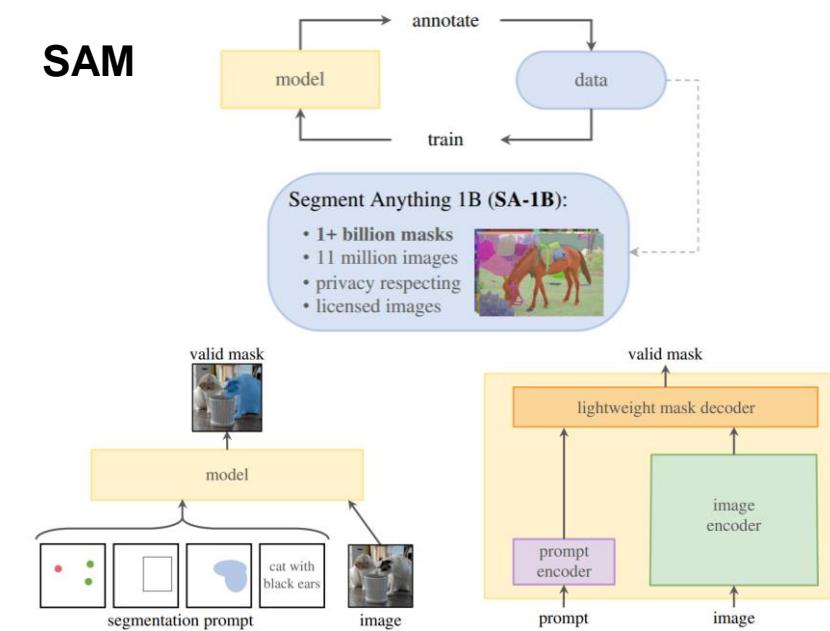
## ChatGPT



## LLaVA



## SAM



# Real-World Applications of LLMs and VLMs

## **LLM Applications:**

- Healthcare: Medical records summarization, patient Q&A.
- Robotics: Instruction following, environment understanding.
- Research: Natural language understanding and academic literature analysis.

## **VLM Applications:**

- VQA in autonomous vehicles and robotics.
- Image captioning in social media and accessibility tools.
- Image segmentation in medical imaging and satellite data analysis.

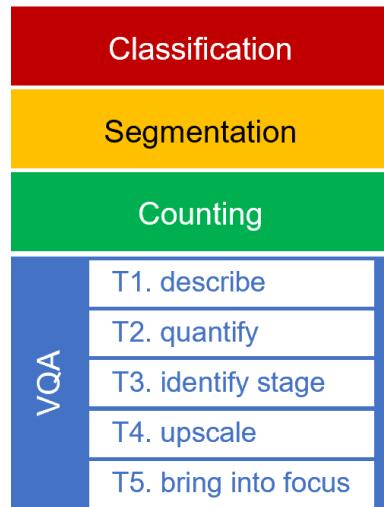
# Outline of Exercises and Demos

- Demo:** ChatGPT – Interactive VQA demo
- Exercise 1:** BLIP – Image captioning (Colab)
- Exercise 2:** CLIP – Image Classification (Colab)
- Demo:** LLaVA – Batch image classification
- Exercise 3:** SAM – Image segmentation (Colab)

# Research Overview



## Evaluation Tasks

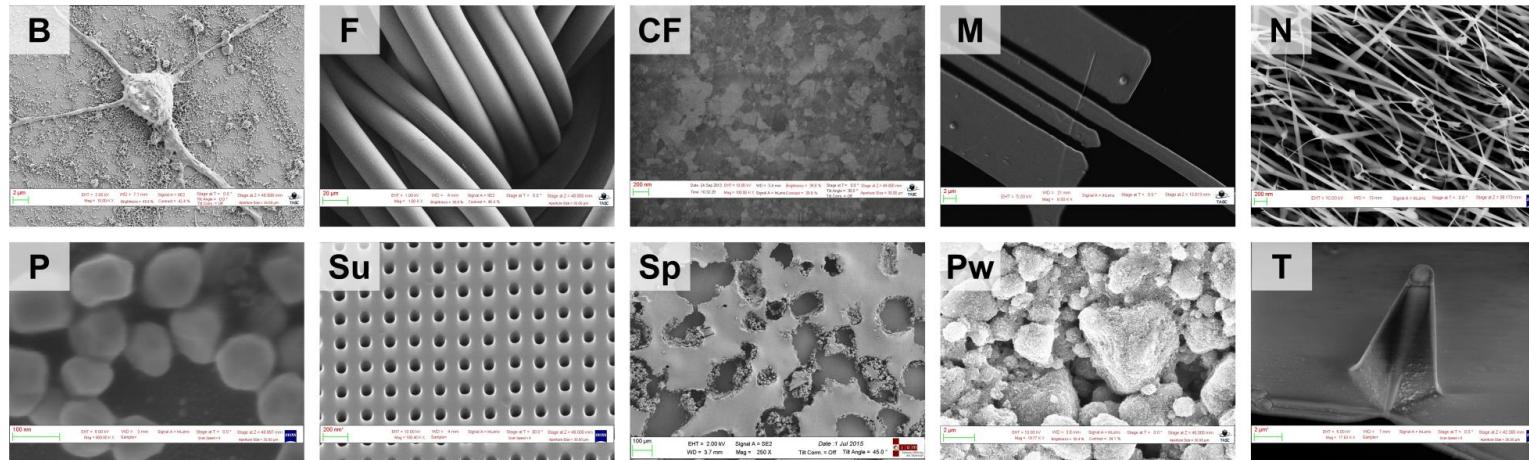




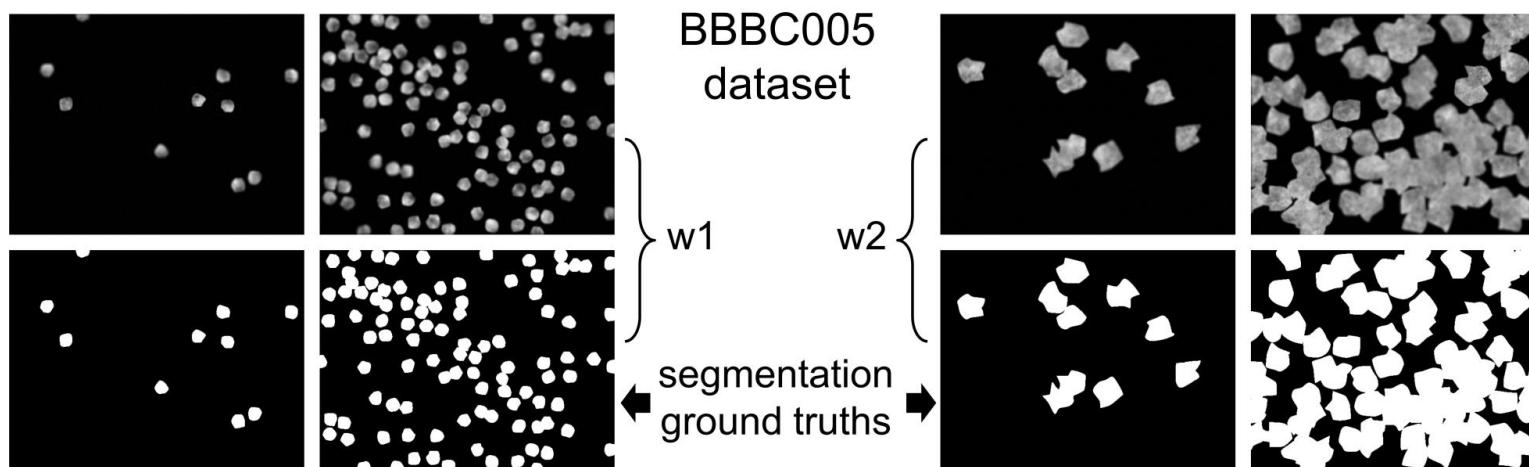
UNIVERSITY OF  
ARKANSAS

# Datasets

NFFA-Europe - 100% SEM dataset



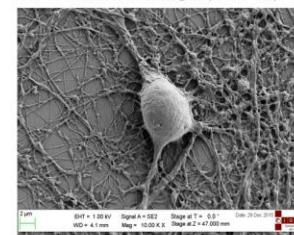
BBBC005  
dataset



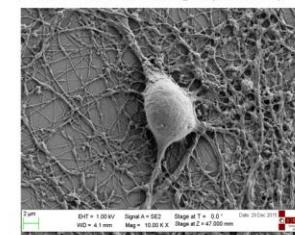


# Visual Question Answering

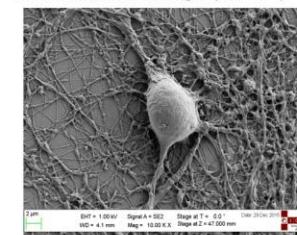
## T1 DESCRIBE



You Describe this image. (VQA-T1)



You Describe this image. (VQA-T1)



You Describe this image. (VQA-T1)

**ChatGPT** The image you've provided appears to be a scanning electron microscope (SEM) image... In this image, we see a large, rounded structure that could be a cell or a particle, surrounded by an intricate network of fibrous or filamentous structures. A mesh-like suggests a d material, possibly e rounded structure is than the surrounding meshwork, highlighting it as the central figure in this image.

IT =  
1.0  
acc  
(wo  
KX)  
vide  
technical information about how the SEM image was captured.  
The date stamped is "29 Dec 2015", indicating when the image was taken...

cell  
fibrous

**LLaVA** The image appears to be a microscopic view of a biological structure, possibly a cell or a tissue sample. It is a grayscale image with a high level of detail, showing a network of thin, thread-like structures that could be fibrous proteins or other components of the cell. At the center of the image, there is a spherical object that stands out from the surrounding network. This object has a smooth, rounded surface and is connected to the network by what looks like a thin stalk or membrane.

The image includes a label... The

biological

stage position at T = 0°, and the stage position at Z = 47,000 mm. The date of the image is also provided, which is December 29, 2015. The logo of the Istituto Officina dei Materiali...

## T2 QUANTIFY SIZE

You Can you visually estimate the size of this structure? (VQA-T2)

**ChatGPT** In the image, the scale bar is marked as 2  $\mu\text{m}$  (micrometers)... the rounded structure seems to be approximately the same length or slightly less than the scale bar in its longest dimension. Therefore, the diameter of the large, rounded structure can be estimated to be roughly around 1.5 to 2  $\mu\text{m}$ .

You Can you visually estimate the size of this structure? (VQA-T2)

**LLaVA** Based on the scale bar provided... I can estimate the size of the spherical object in the center... to be approximately half the length of the scale bar, so it is likely to be around 1  $\mu\text{m}$  in diameter. However, this is a rough estimate...

neuron

You Can you visually estimate the size of this structure? (VQA-T2)

**Gemini** ...Unfortunately, directly measuring the neuron's size from the image is difficult without a reference object for scale. However, we can make an estimation using the provided... Scale Bar... Soma Diameter: By visually comparing the soma to the scale bar, we can estimate its diameter. It appears to be roughly 1/4 to 1/3 the length of the scale bar, suggesting a soma diameter between 0.5  $\mu\text{m}$  and 0.75  $\mu\text{m}$ ...

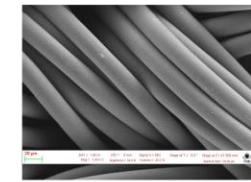


# Visual Question Answering

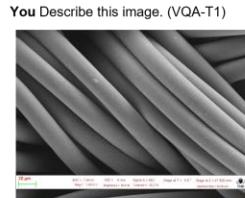
**T1**  
DESCRIBE



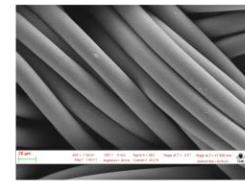
You Describe this image. (VQA-T1)



**ChatGPT** The structures could represent fibers, tubes, or filaments of some material, whether biological (like plant fibers or extracellular matrices) or synthetic (such as polymer fibers or nanotubes)... scale bar "20  $\mu\text{m}$ " (micrometers), indicating the scale for size comparison. Additionally, the image data...  
1.00  
accs  
(wor  
(mag  
  
**fibers**



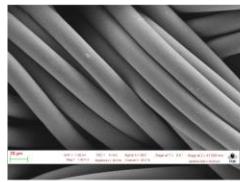
You Describe this image. (VQA-T1)



**LLaVA** The image... possibly a fiber or a bundle of fibers, captured using scanning electron microscopy (SEM)... are tightly packed together, creating a dense, layered pattern... label with various technical specifications... It includes the magnification (1.43kX), the signal (SE2), the stage position at T (0.00 mm), and the rotation (45.00°), and the voltage (4.00 kV), also me... an acronym for a specific model or feature of the SEM equipment used to capture the image....



You Describe this image. (VQA-T1)



**Gemini** The image displays a microscopic view of **interwoven fibers**, likely captured using a scanning electron microscope (SEM) due to the detailed surface texture and grayscale image. The fibers have a cylindrical structure and a relatively smooth

**interwoven  
fibers**

**T2**  
QUANTIFY  
DIAMETER

You Can you visually estimate the average diameter of the fibers? (VQA-T2)

**ChatGPT** ...Visually, if a fiber seems to cover about one-tenth to one-fifth of the scale bar's length, then the diameter of the fibers would be approximately 2  $\mu\text{m}$  to 4  $\mu\text{m}$ .

You Can you visually estimate the average diameter of the fibers? (VQA-T2)

**LLaVA** While I can provide a rough estimate, it's important to note that the actual diameter of the fiber would depend on the scale provided in the image, which is not visible in this case. However, based on the magnification given (1.43kX), I can make a rough estimate... the fiber appears to be about 1 pixel wide in the image, we can estimate the diameter of the fiber to be approximately 1.430  $\mu\text{m}$  (micrometers). This is a very rough estimate, as the actual scale of the image is not provided...

You Can you visually estimate the average diameter of the fibers? (VQA-T2)

**Gemini** Based on the provided scale bar and image, we can estimate the diameter of the fibers... scale bar... 20 micrometers ( $\mu\text{m}$ )... By visually comparing the fiber width to the scale bar, we can estimate that the diameter of a single fiber is... 4 to 5 micrometers ( $\mu\text{m}$ ).

**T2**  
QUANTIFY  
NUMBER

You Can you visually estimate the number of fibers? (VQA-T2)

**ChatGPT** ...I would estimate there could be around 10 to 20 fibers visible...

You Can you visually estimate the number of fibers? (VQA-T2)

**LLaVA** ...can give a rough estimate... it seems like there are dozens of them in the visible area of the image...

**10-20**

**dozens**

You Can you visually estimate the number of fibers? (VQA-T2)

**Gemini** Accurately counting the fibers in the image is challenging due to their interwoven nature and the limited field of view. However, we can make a rough estimate... page Fib... For inst... are app... fails... When extrapolate this to the entire image... Without specific measurements... we cannot provide a precise numerical estimate of the total fiber count.



# Visual Question Answering

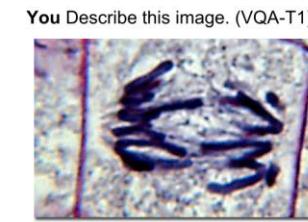
**T1**  
DESCRIBE  
(early anaphase)



**ChatGPT** The image appears to be a microscopic view of a cell, showing a structure during **metap** division. Chromosomes are aligned along the equator of the cell, preparing for segregation into daughter cells...



**LLaVA** The image appears to be a microscopic view of a biological structure or a tissue/cell. Mitochondria are referred to as the powerhouses of the cell due to their role in producing energy...



**Gemini** The image shows a cell undergoing **anaphase** of mitosis. The sister chromatids have opposite fibers that are pulling towards the poles.

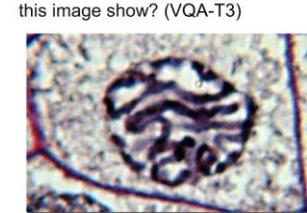
**mitosis**  
**anaphase**

**T3**  
IDENTIFY STAGE  
(late prophase)



You What stage is this? (VQA-T3)

**ChatGPT** The image displays a cell with condensed structures that are characteristic of chromosomes in a state of division. The chromosomes are more clustered and possibly starting to separate, suggesting that this is **anaphase**...



You What stage of cell division does this image show? (VQA-T3)



You What stage is this? (VQA-T3)

**Gemini** This is the **Prophase I** stage of Meiosis.

**prophase**

**T4**  
UPSCALE

You Can you upscale this image? (VQA-T4)  
**ChatGPT** <either responds saying that it cannot upscale or throws errors>

**errors**

You Can you upscale this image? (VQA-T4)

**LLaVA** <either responds saying that it cannot upscale or throws errors>

**can't**

You Can you upscale this image? (VQA-T4)

**Gemini** <either responds saying that it cannot upscale or throws errors>

**can't**

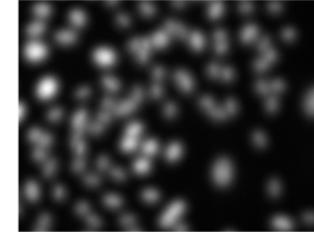
# Visual Question Answering

## T5 BRING IN FOCUS

**You** I have a microscope image of cells that are out of focus. Can you use some image processing techniques to bring the cells in focus or generate a corresponding image that would be in focus? (VQA-T5)

**ChatGPT** Unfortunately, I can't directly enhance or refocus an out-of-focus microscope image. The process of bringing an out-of-focus image into focus typically involves image processing techniques that might help, but I don't have access to current AI tools that have this capability. However, I can offer some general advice on techniques and software that might help you refocus or enhance the image manually...

**You** Can you perform deconvolution on this image?



**ChatGPT** Hmm...something seems to have gone wrong.

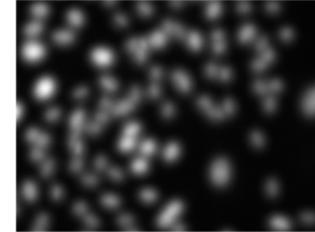
errors

**You** I have a microscope image of cells that are out of focus. Can you use some image processing techniques to bring the cells in focus or generate a corresponding image that would be in focus? (VQA-T5)

**LLaVA** Yes, it is possible to use image processing techniques to improve the focus of a microscope image. One

this purpose.  
<note that LLaVA currently lacks capability of images generation>

**You** Can you perform deconvolution on this image?



**LLaVA** The image you've provided appears to be a low-resolution or highly blurred image with numerous small white dots against a dark background. Deconv

used to restore a degraded image by removing noise. You could use image processing software or libraries that support deconvolution, such as ImageJ, MATLAB, or Python with libraries like scikit-image or OpenCV.

<note that LLaVA currently lacks capability of images generation>

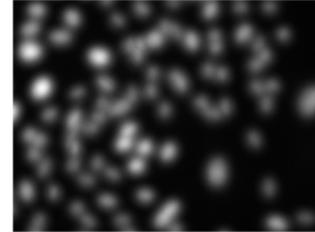
**You** I have a microscope image of cells that are out of focus. Can you use some image processing techniques to bring the cells in focus or generate a corresponding image that would be in focus? (VQA-T5)

**Gemini** Unfortunately, there is no perfect image processing technique to bring truly out-of-focus cells into sharp

has already been used to enhance the image and potentially improve the visibility of details... filtering... deconvolution... stacking

<note that Gemini currently lacks capability of images generation>

**You** Can you perform deconvolution on this image?



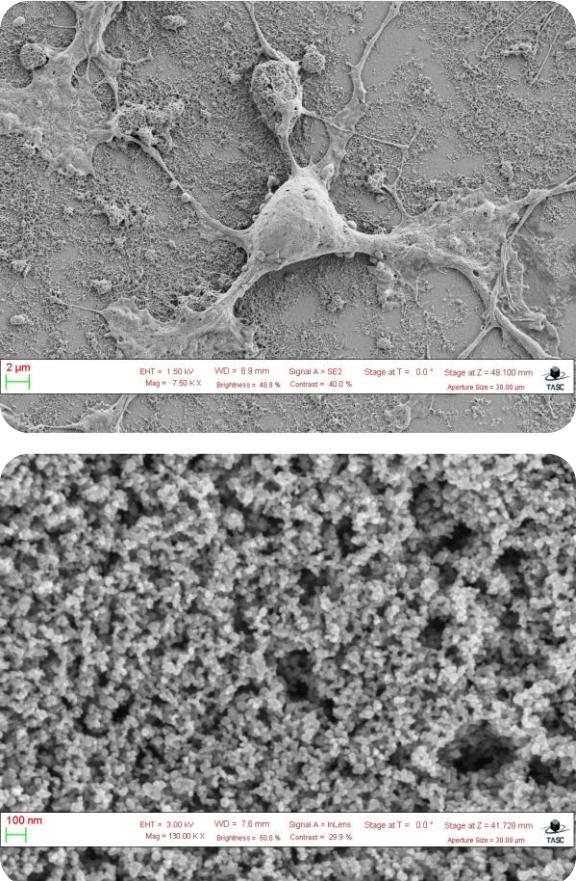
**Gemini** I can't directly perform deconvolution on the image you sent because I cannot access and process files.

Deconvolution is also a complex image processing task that requires more than what I can do with the image.

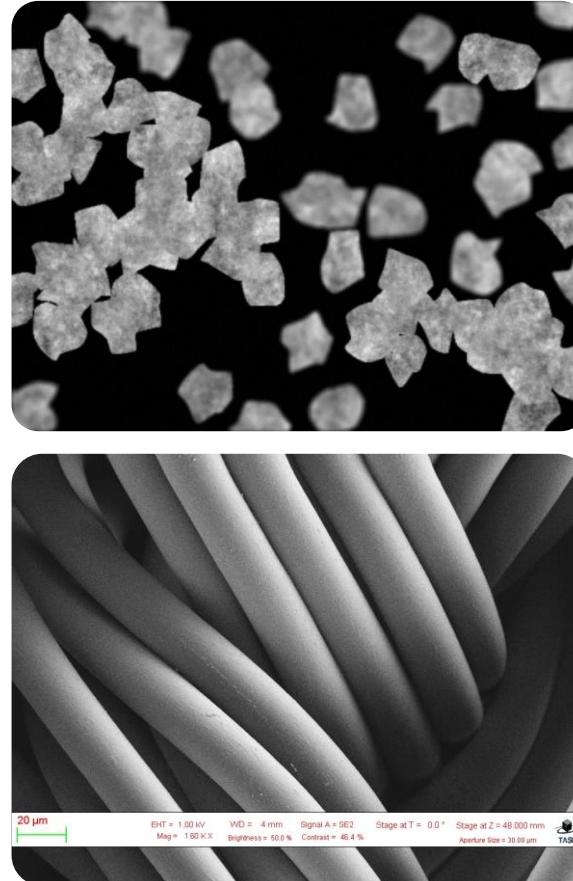
can't

# Interactive Demo: ChatGPT (10 mins)

describe



scale bar



count

segment

# The Core of LLMs: Attention and Transformers

## Attention Mechanism:

- Focuses on key parts of the input sequence.
- Helps models understand context in text.
- Example: "The cat sat on the mat" – Attention on "cat" and "mat."

## Transformers:

- Breakthrough architecture for LLMs.
- Processes input in parallel, unlike RNNs (Recurrent Neural Networks).
- Key components: Multi-Head Attention, Feedforward layers, Positional Encoding.

# Overview of LLaMA and GPT Models

## **LLaMA (Large Language Model Meta AI):**

- Developed by Meta for efficient LLM performance.
- Open-source and designed for smaller model sizes while maintaining strong performance.

## **GPT Models:**

- Developed by OpenAI, including GPT-3 and GPT-4.
- Trained on vast datasets for text generation, Q&A, and more.
- Known for scalability and few-shot learning capabilities.

# Bias, Hallucination, Scalability

## **Bias:**

- LLMs can inherit biases from training data.
- Example: Gender, racial, or cultural biases in generated text.

## **Hallucination:**

- LLMs sometimes generate false information confidently.
- Example: Inventing historical facts or references.

## **Scalability:**

- Larger models require immense computational resources.
- Balancing performance with resource constraints is critical.



# Zero-shot Evaluation

Performance of a model on  
**Tasks or data it has not been trained on**  
It demonstrates models' versatility and generalizability

This is different from the performance when testing is done on data that is similar to training data, such as in supervised deep learning.

# Few-Shot Prompting and Its Benefits

## **Few-shot Prompting:**

- Providing multiple examples in the prompt to improve model performance.
- Helps the model better understand task requirements.

## **Example:**

- Single-shot:** Provide one question and answer.
- few-shot:** Provide several Q&A pairs to show the model the pattern.

## **Benefits:**

- Increases accuracy by giving the model more context.
- Especially useful in tasks like image captioning and VQA.

# Exercise 1: Captioning with BLIP (10 mins)

Please go to our GitHub repo <https://github.com/pv-is-nrt/aimrc-data-science-core/> and navigate to `workshops > vlms > workshop-vlms.ipynb` and click the download button



Open the downloaded file in your Google Colab. Follow the instructions in the notebook file.

## Task:

- Load the BLIP model in Google Colab.
- Upload an image and generate a caption using the BLIP model.

## Steps:

- Open the provided Colab notebook.
- Run the cell to load BLIP from Hugging Face.
- Upload an image to the notebook.
- Run the model to generate an image caption.

# Vision-Language Models vs. Large Language Models

## LLMs:

- Only process text.
- Useful for natural language processing tasks (e.g., text generation, summarization)

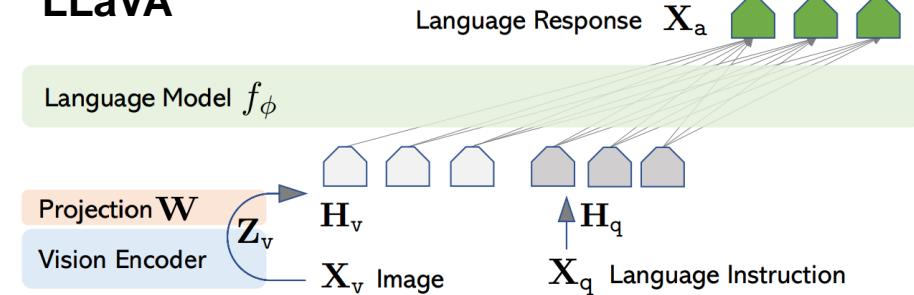
## VLMs:

- Process both text and images.
- Capable of multimodal tasks like image captioning, visual question answering (VQA), and image classification.

## Key Difference:

- VLMs combine visual features (from images) with language understanding, allowing for richer tasks.

## LLaVA



# Combining Vision and Language: Multimodal Learning

## How It Works:

- VLMs use **vision encoders** (e.g., CNNs, Vision Transformers) to extract features from images.
- A **language model** (e.g., GPT) processes text and integrates with visual features.
- The model generates a combined representation of both modalities for tasks like VQA or captioning.

## Example:

- Input: An image of a dog and the question "What color is the dog?"
- Output: The model generates "The dog is brown."

# Combining Vision and Language: Multimodal Learning

## Step-by-Step Process:

1. **Image Processing:** The image is passed through a visual encoder (e.g., CNN or Transformer).
2. **Question Processing:** The question is tokenized and passed to the language model.
3. **Linking:** The VLM links the visual features with the question's context.
4. **Answer Generation:** The language model generates a response based on the image and question.

## Example:

- Input: Image of a person holding a cup and the question, "What is the person holding?"
- Output: "A cup."

# Key Components of VLM Architectures

## Vision Encoders:

- ❑ Responsible for extracting features from images.
- ❑ Examples: **CNNs** (Convolutional Neural Networks) and **Vision Transformers**.

## Language Models:

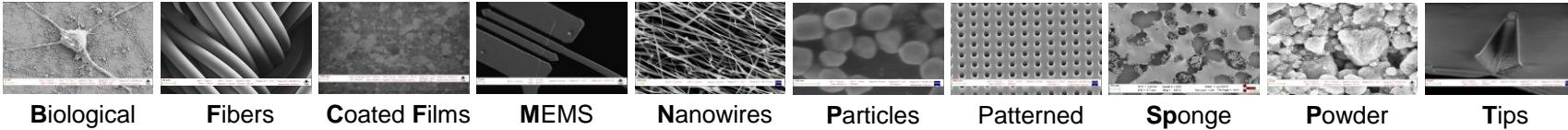
- ❑ Responsible for understanding and generating language based on text and visual features.
- ❑ Examples: GPT, BERT, LLaMA.

## How They Work Together:

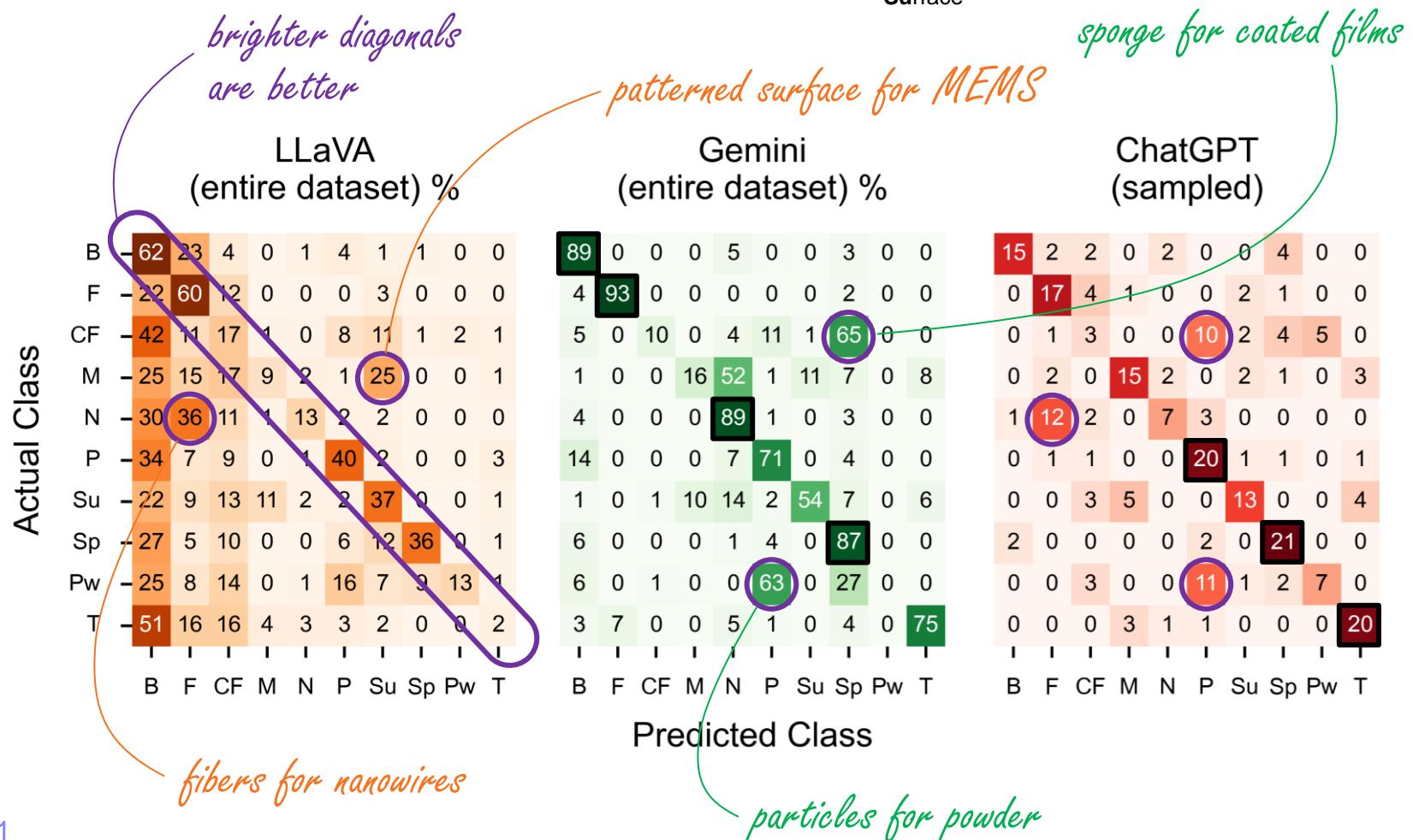
- ❑ The vision encoder extracts visual features, which are then passed to the language model.
- ❑ The language model integrates these visual features with text input to produce responses.



# Classification



Biological   Fibers   Coated Films   MEMS   Nanowires   Particles   Patterned Surface   Sponge   Powder   Tips





# Exercise 2: Classification with CLIP (5 mins)

Follow the instructions in the notebook file.

# Segment Anything Model (SAM): Image Segmentation

## Overview:

- ❑ SAM (Segment Anything Model) is designed for fast and accurate image segmentation.
- ❑ It can segment objects in any image without needing task-specific training.

## Key Features:

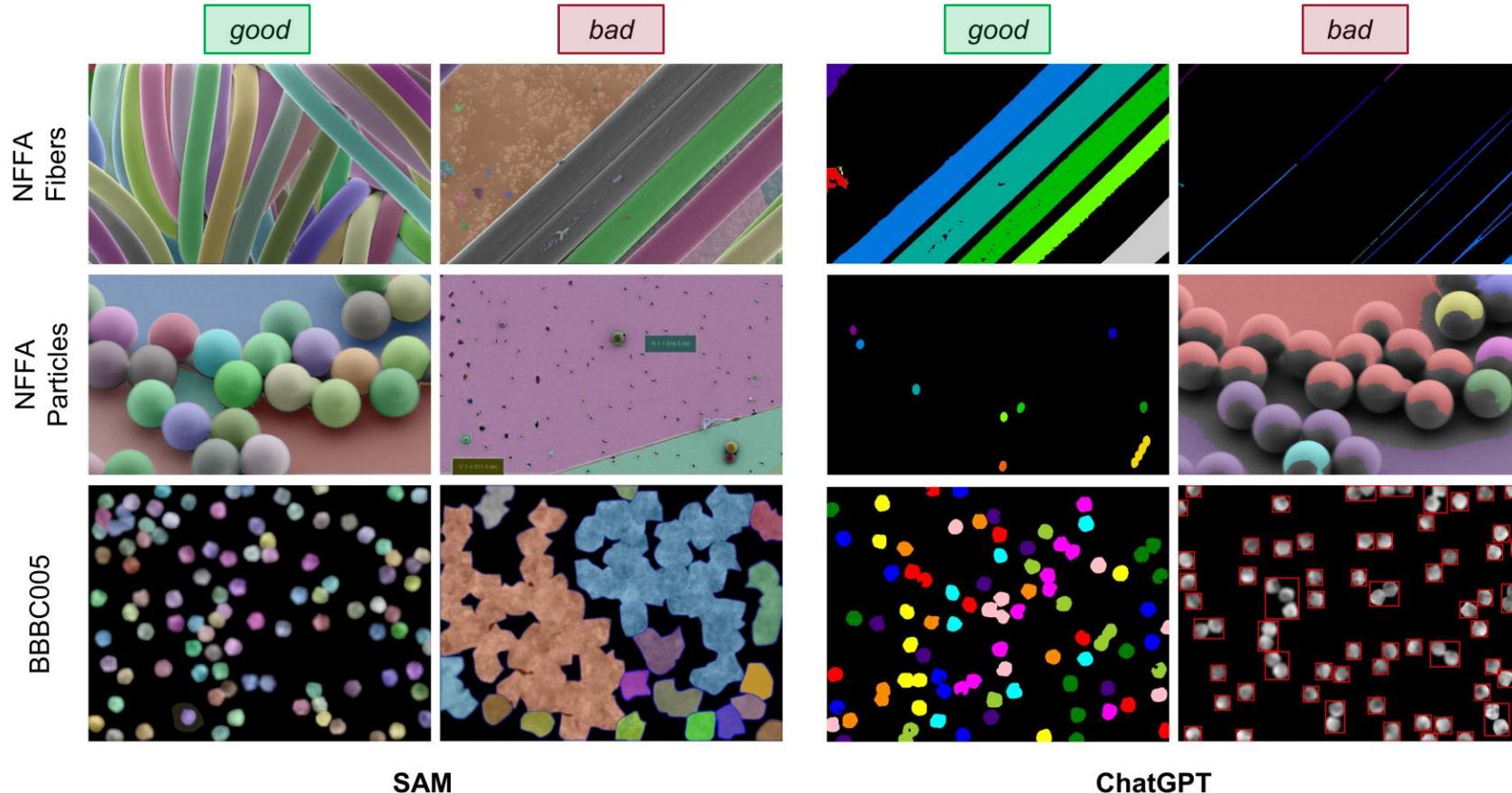
- ❑ Works on any image, even with complex backgrounds.
- ❑ Uses a combination of vision transformers to identify and segment objects.

## Use Cases:

- ❑ Medical imaging (e.g., tumor segmentation).
- ❑ Satellite image analysis.
- ❑ Object detection in robotics.



# Segmentation (manual)





UNIVERSITY OF  
ARKANSAS

# SAM Prompting

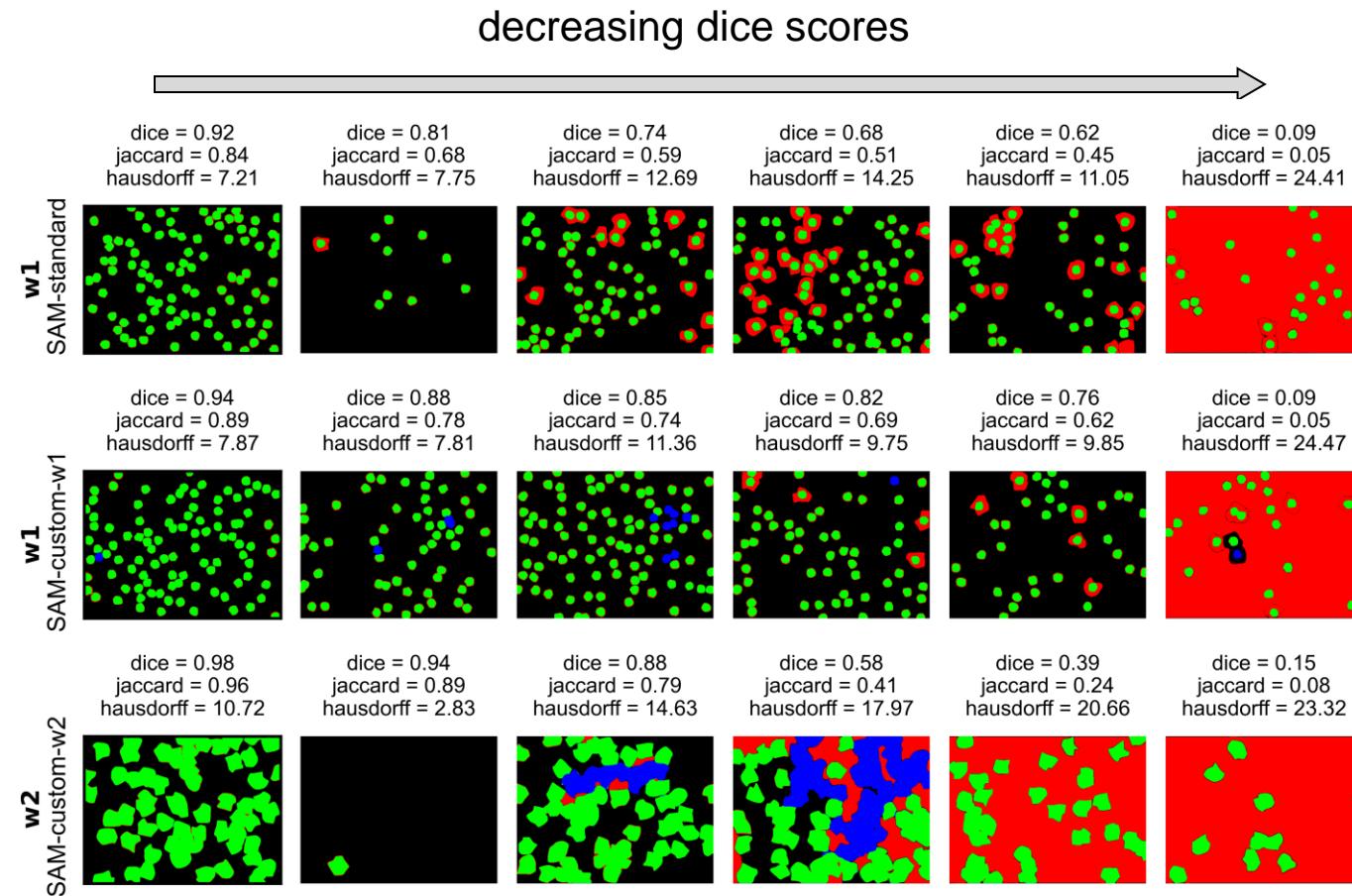


<https://segment-anything.com/>



# Segmentation

SAM segmentation quality  
measured using dice scores and other similar quantities

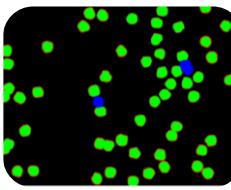
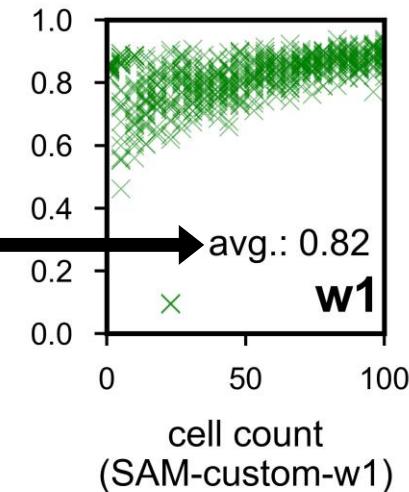
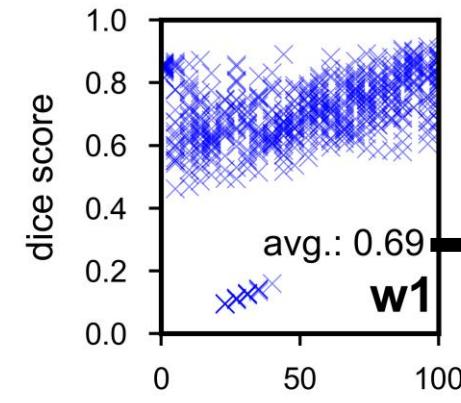
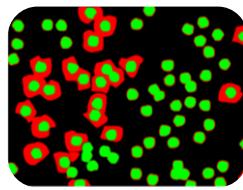


# SAM custom parameters

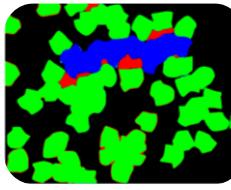
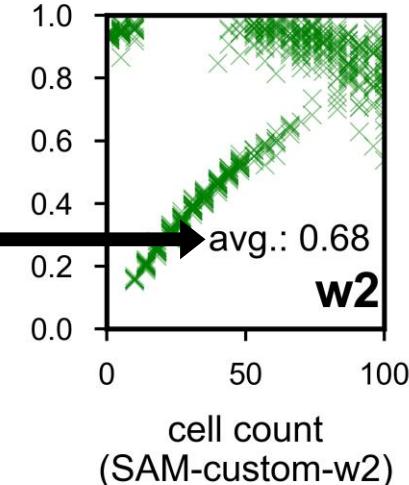
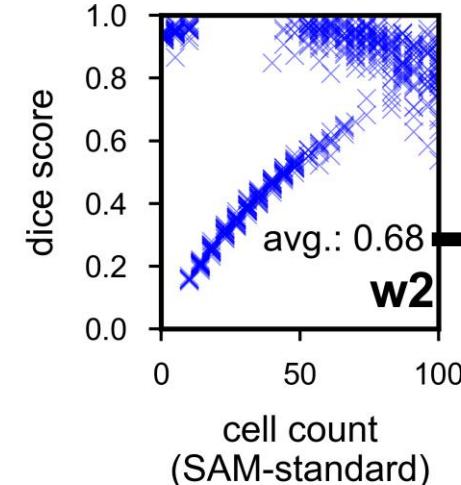
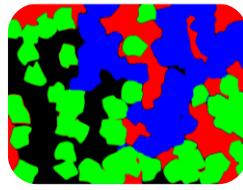
Parameter name	SAM-standard	SAM-custom -w1/w2
points_per_side	32	32
points_per_batch	64	64
<b>pred_iou_thresh</b>	0.88	0.98, 0.88
<b>stability_score_thresh</b>	0.95	0.98, 0.95
stability_score_offset	1.0	1.0
<b>box_nms_thresh</b>	0.7	0.9
crop_n_layers	0	0
<b>crop_nms_thresh</b>	0.7	0.95, 0.8
<b>crop_overlap_ratio</b>	512/1500	0.05, 0.3
crop_n_points_downscale_factor	1	1
point_grids	None	None
min_mask_region_area	0	0
output_mode	binary_mask	binary_mask



# Segmentation (automated)

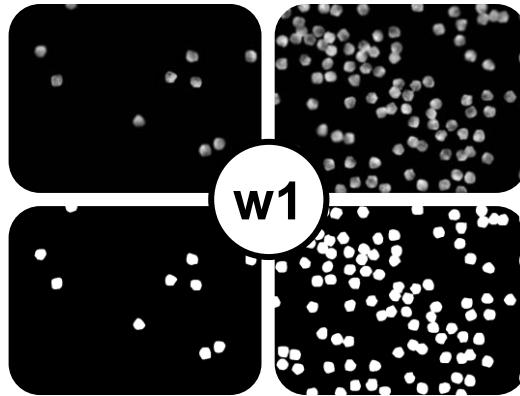


*through  
parameter  
tuning*

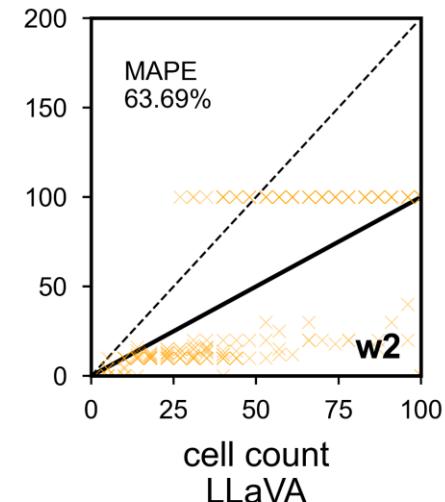
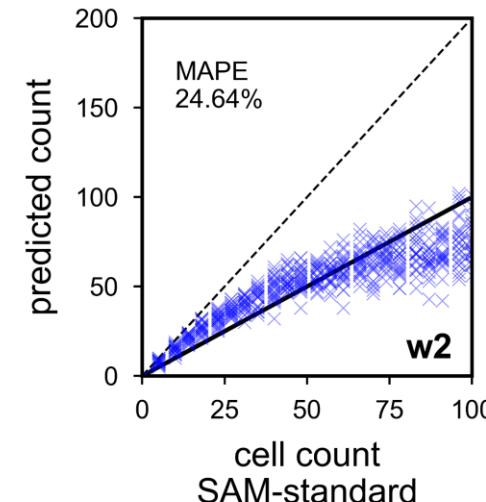
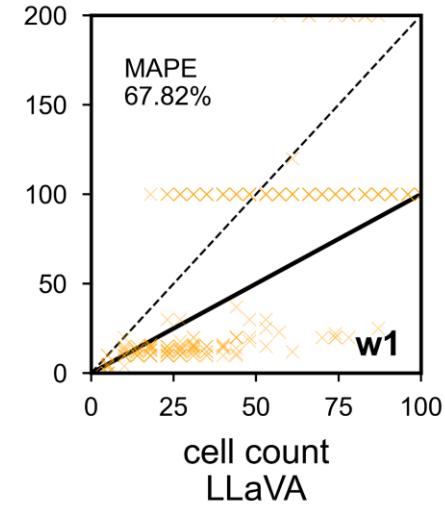
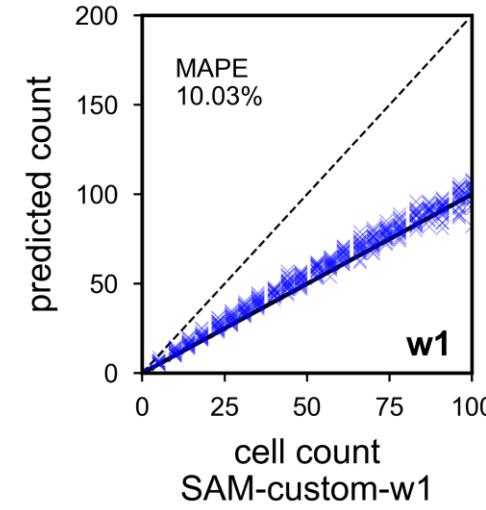
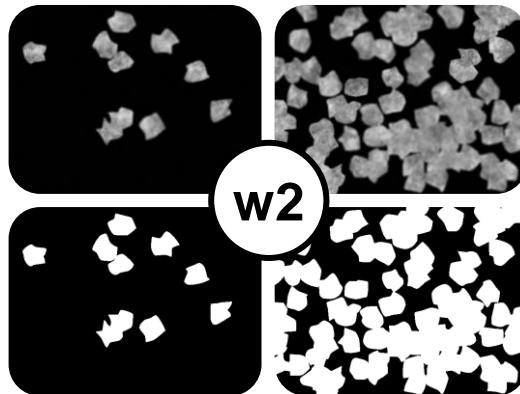




# Counting (automated)

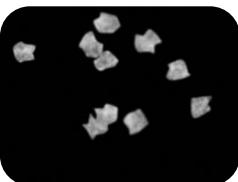
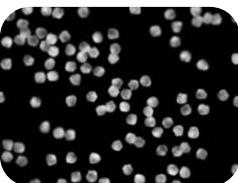
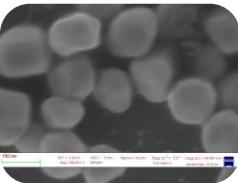
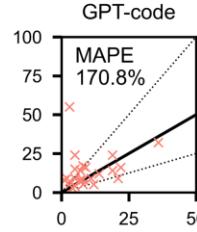
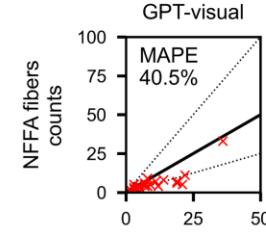
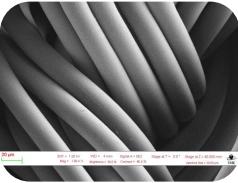


BBBC005 dataset





# Counting (assisted)





# Exercise 3: Segmentation (10 mins)

Follow the instructions in the notebook file.

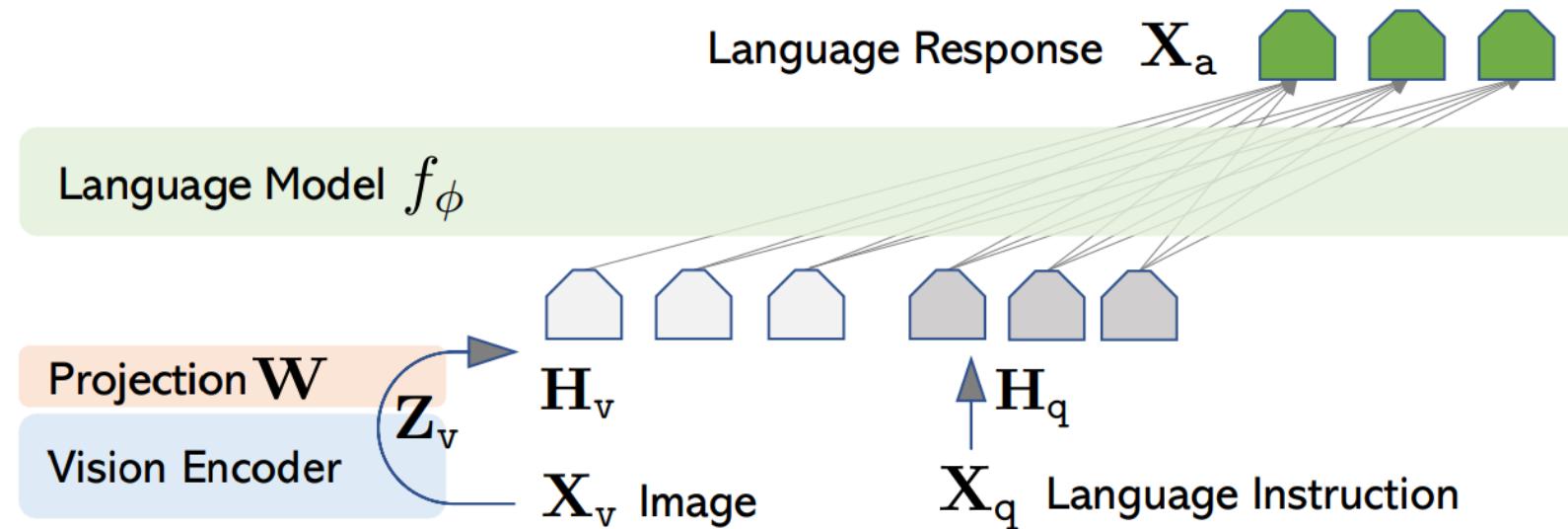
# LLaVA: Large Language and Vision Assistant

## Overview:

- LLaVA is designed specifically for answering questions about images.
- Combines a visual encoder with a language model to handle VQA tasks.

## Key Features:

- Capable of answering both simple and complex questions about images.





# Demo: Batch classification with LLaVA (5 mins)