



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

Prateek Verma, Minh-Hao Van, Xintao Wu

Oct 02, 2024

Arkansas Integrative Metabolic Research Center

# Agenda and Objectives

## □ Agenda

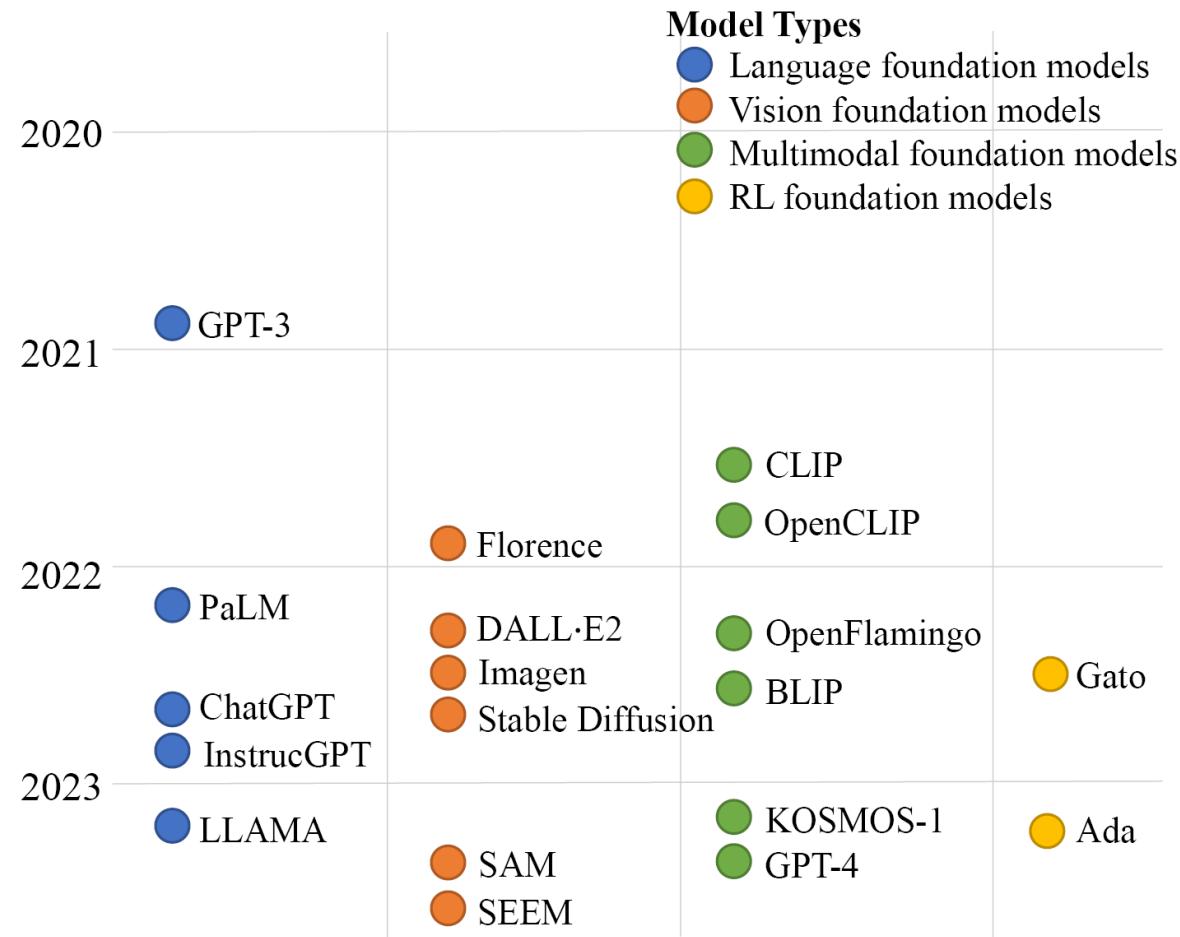
- Overview of LLMs and VLMs.
- Introduction to key VLM models:  
ChatGPT, BLIP, CLIP, LLaVA, SAM.
- Demonstrations and exercises.

# 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

## Foundation Models Timeline



# Medical Foundation Models

## Language Models

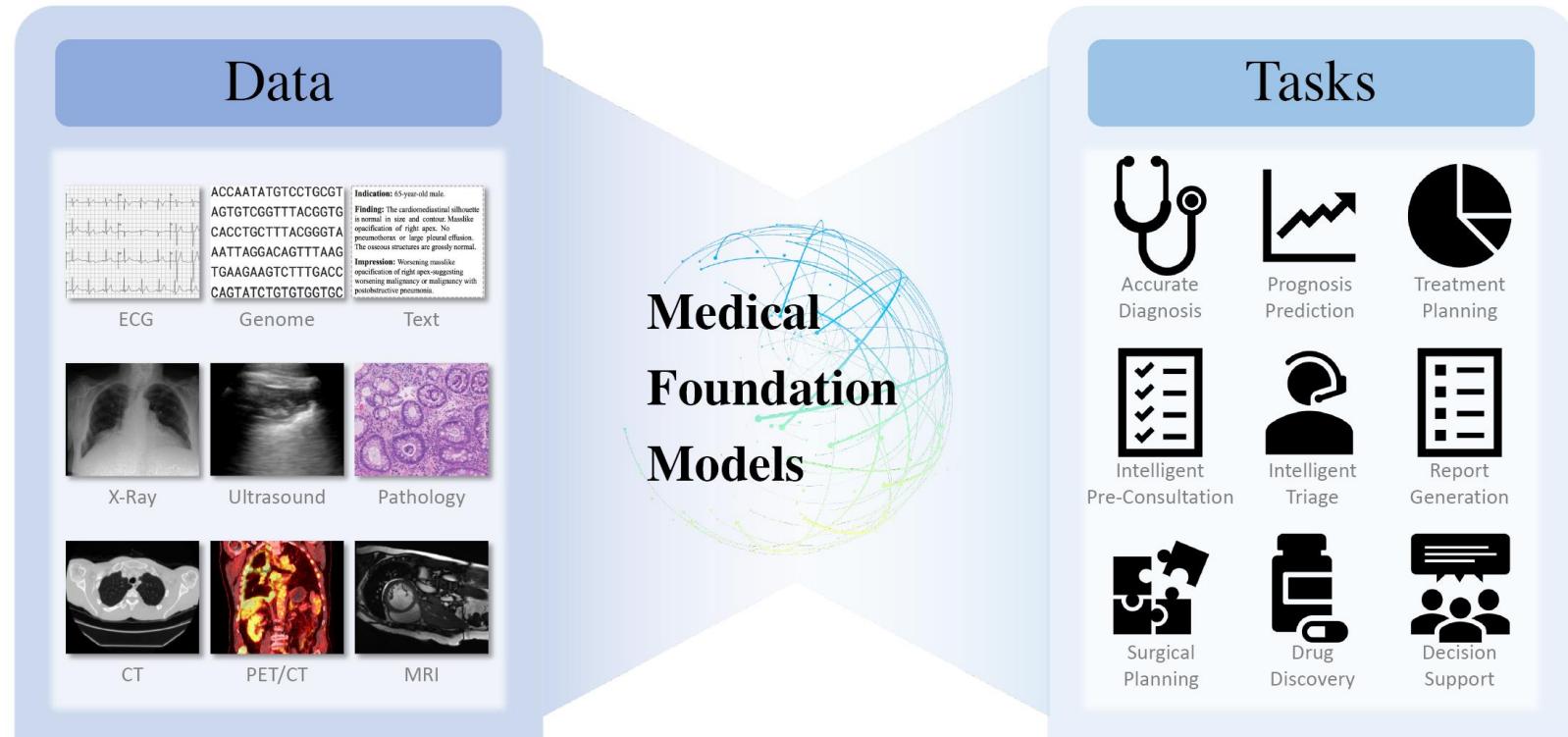
ChatGPT, GPT-4,  
Bard, PaLM

## 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.

# 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.

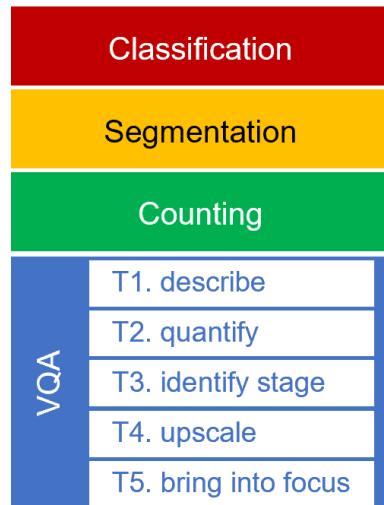
# Outline of Exercises and Demos

- 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 Example



## Evaluation Tasks

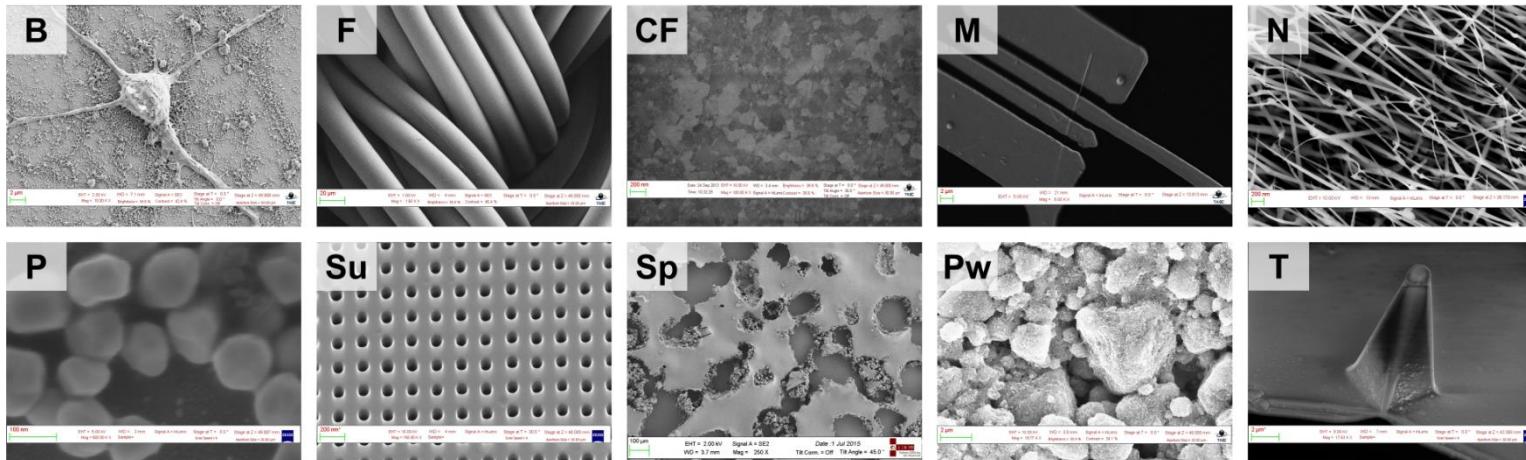




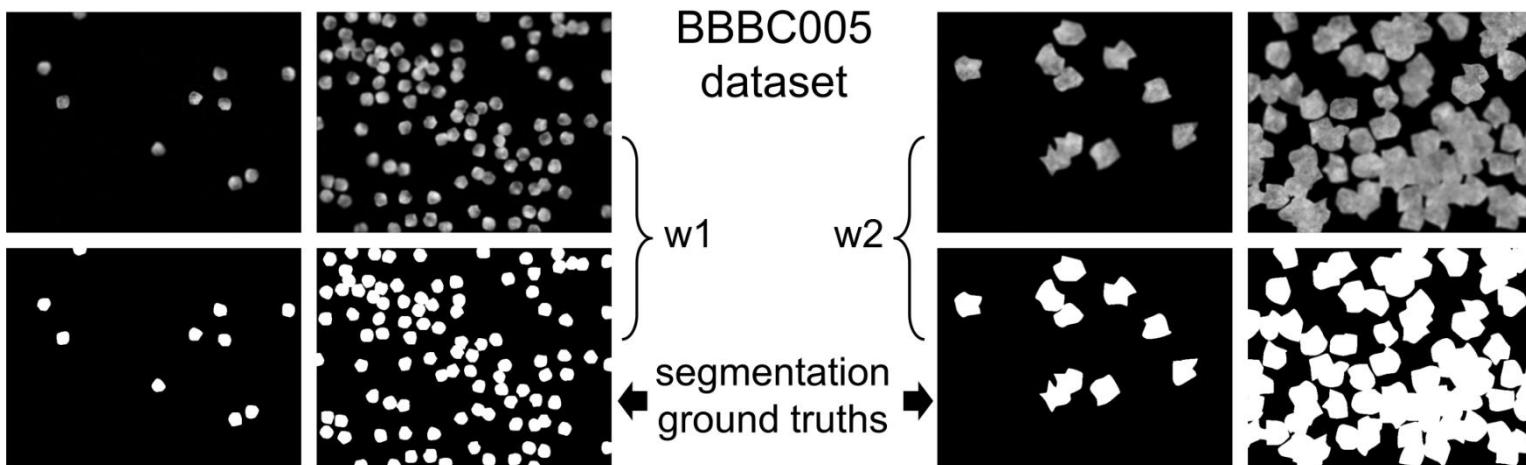
UNIVERSITY OF  
ARKANSAS

# Datasets

NFFA-Europe - 100% SEM dataset

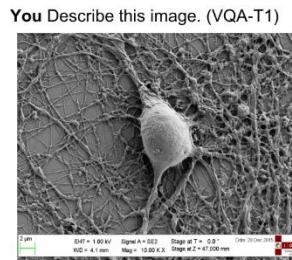


BBBC005  
dataset

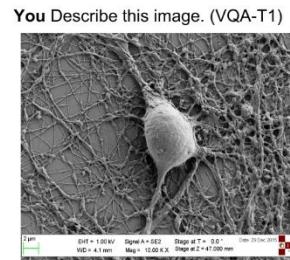


# Visual Question Answering

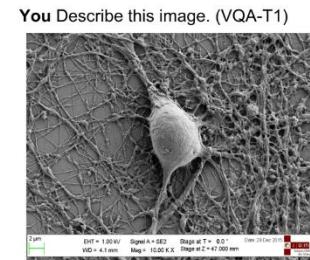
## T1 DESCRIBE



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 dashed rectangular box suggests a mesh-like material, possibly a kind of synthetic structure is larger than the surrounding meshwork, highlighting it as the central figure in this image.

IT = 1.00 acc (wo KX) provide 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...

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 μm (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 μm.

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 μm in diameter. However, this is a rough estimate...

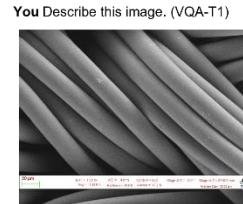
neuron

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

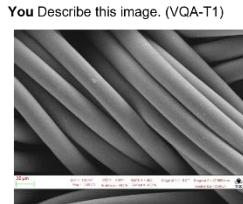
**Gemini** ...Unfortunately, directly measuring the neuron's size from the image is difficult without a reference object for scale. However, we can make an estimation using the provided... Scale Bar... Soma Diameter: By visually comparing the soma to the scale bar, we can estimate its diameter. It appears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

# Visual Question Answering

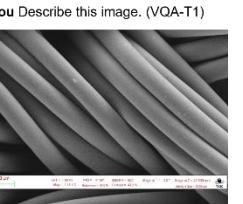
## T1 DESCRIBE



You Describe this image. (VQA-T1)



You Describe this image. (VQA-T1)



You Describe this image. (VQA-T1)

## T2 QUANTIFY DIAMETER

**fibers**

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  
accos  
(wor  
(mag  
he  
m"  
X"

**fibers**

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 stage position at T (0.00), and vol (4.00), and me (100%). So me is an acronym for a specific model or feature of the SEM equipment used to capture the image....

**interwoven  
fibers**

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

## T2 QUANTIFY NUMBER

**10-20**

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

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

**dozens**

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...

**fails**

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... Fibre... For inst... are app... then 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 and being prepared for segregation into daughter cells...

**T3**  
IDENTIFY STAGE  
(late prophase)



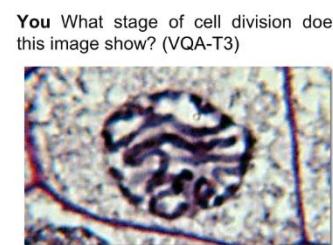
**ChatGPT** The image displays a cell with condensed structures that are characteristic of chromosomes in a state of **anaphase**. These are more clustered and possibly starting to separate, suggests that this is anaphase...

**T4**  
UPSCALE

**ChatGPT** <either responds saying that it cannot upscale or throws errors when trying to upscale>



**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...



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



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

**mitosis anaphase**

**You** What stage is this? (VQA-T3)



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

**prophase**

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

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

**can't**

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

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

**can't**

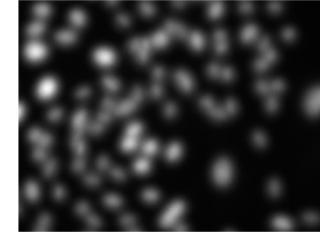
# Visual Question Answering

## 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 typically involves image processing techniques 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.

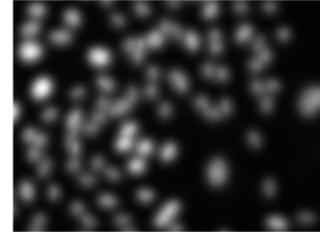
errors

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?



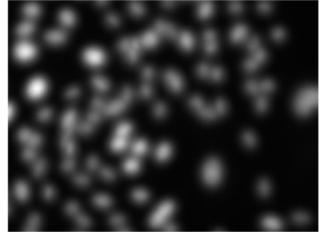
LLaVA The image you've provided appears to be a low-resolution or highly blurred image with numerous small white dots against a dark background. Deconvolution used to restore a degraded image from noise. You could use image processing software or libraries that support deconvolution, such as ImageJ, 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 here are some 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?



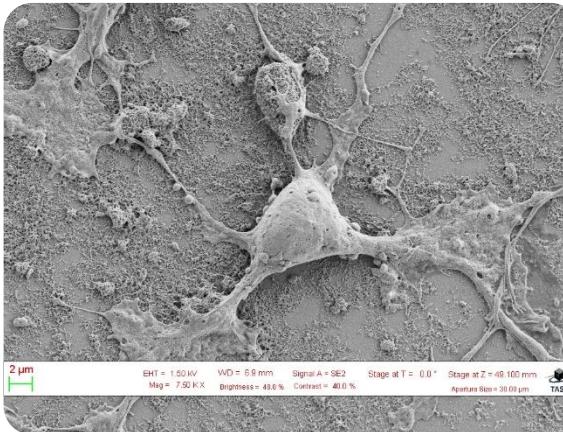
Gemini I can't directly perform deconvolution on the image you sent because I cannot access and process files. Deconvolution is also a complex image processing task that requires more than what I can do with the image I have.

can't

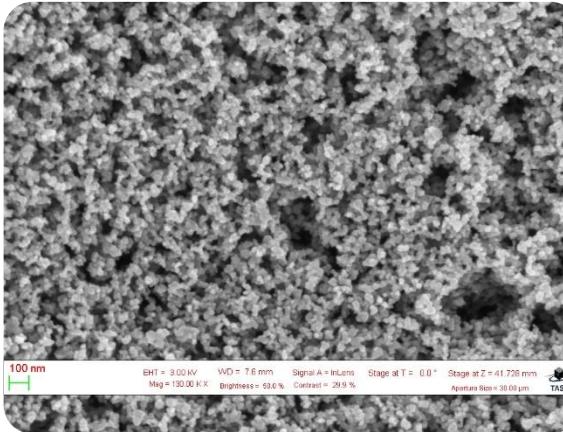
# Interactive Demo: ChatGPT (10 mins)

**NOTE:** A paid version of ChatGPT is required for the ability to upload images

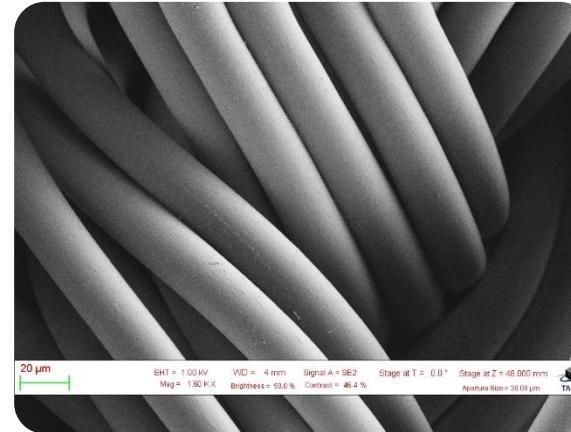
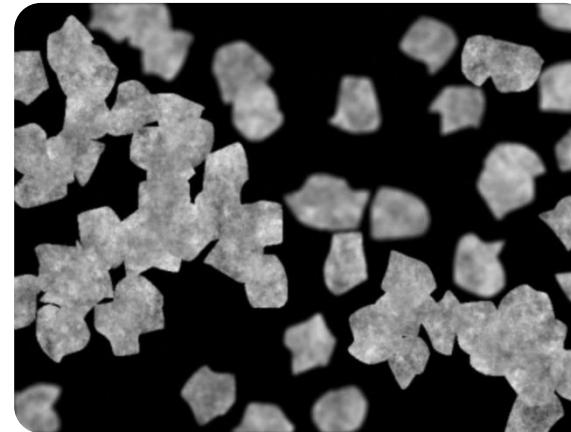
describe



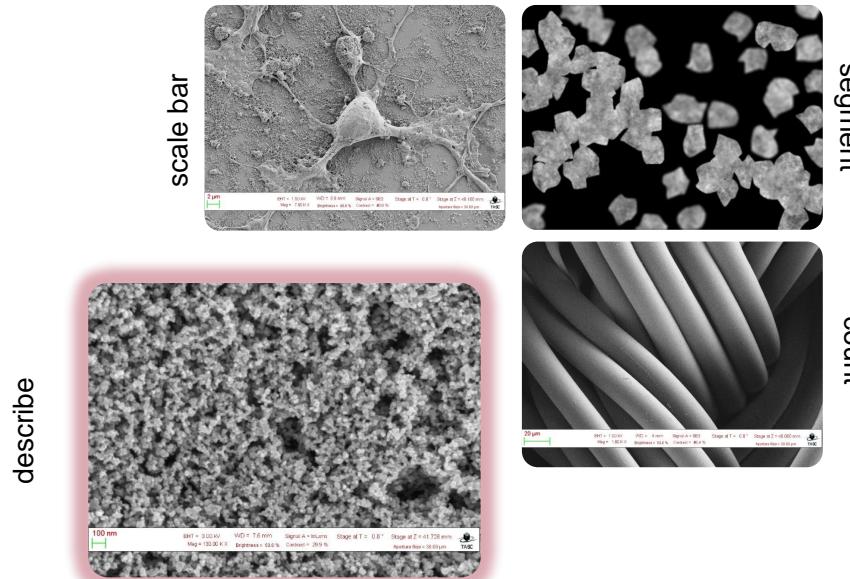
scale bar



segment count



# Interactive Demo: Gemini (2 mins)



# 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.

# 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.

## Compare with

- Training and test sets
- Fine-tuning of LLMs
- In-context and few-shot learning for LLMs

# 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.**

## BLIP-2

Gradio demo for BLIP-2, image-to-text generation from Salesforce Research. To use it, simply upload your image, or click one of the examples to load them.  
**Disclaimer:** This is a research prototype and is not intended for production use. No data including but not restricted to text and images is collected.

Paper: [BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models](#)

Code: BLIP2 is now integrated into GitHub repo: [LAVIS: a One-stop Library for Language and Vision](#)

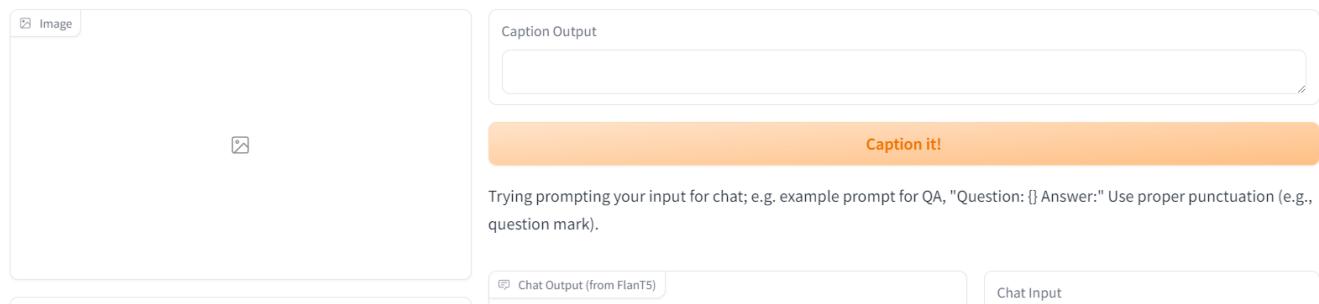
👉 `transformers` integration: You can now use `transformers` to use our BLIP-2 models! Check out the [official docs](#)

Project Page: [BLIP2 on LAVIS](#)

Description: Captioning results from BLIP2\_OPT\_6.7B. Chat results from BLIP2\_FlanT5xxl.

We have now suspended the official BLIP2 demo from March 23, 2023.

For example usage, see notebooks <https://github.com/salesforce/LAVIS/tree/main/examples>.



The screenshot shows the BLIP-2 Gradio demo interface. On the left, there's an 'Image' input field with a placeholder 'Upload image'. To its right is a 'Caption Output' text area. Below these is a large orange button labeled 'Caption it!'. At the bottom of the interface, there are two more text areas: 'Chat Output (from FlanT5)' on the left and 'Chat Input' on the right. A red arrow points to the 'Caption it!' button, indicating where to click.

**Why use Python instead of browser demo?**

# Vision-Language Models vs. Large Language Models

## LLMs:

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

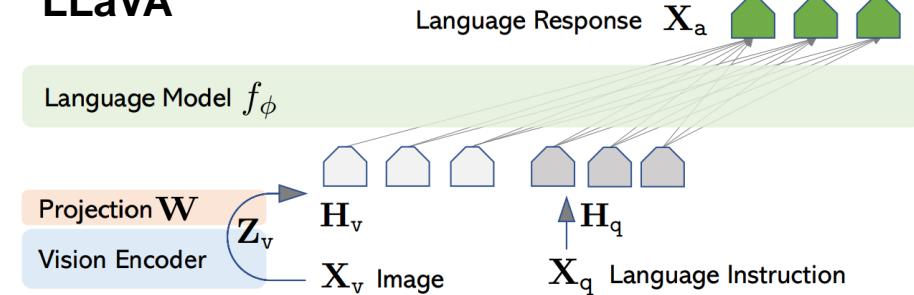
## VLMs:

- 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:

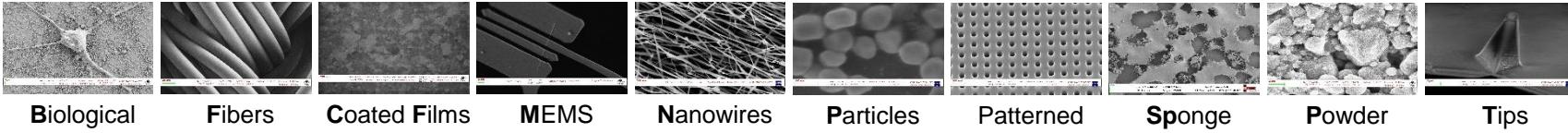
1. **Image Processing:** The image is passed through a visual encoder (e.g., CNN or Transformer).
2. **Question Processing:** The question is tokenized and passed to the language model.
3. **Linking:** The VLM links the visual features with the question's context.
4. **Answer Generation:** The language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

## Example:

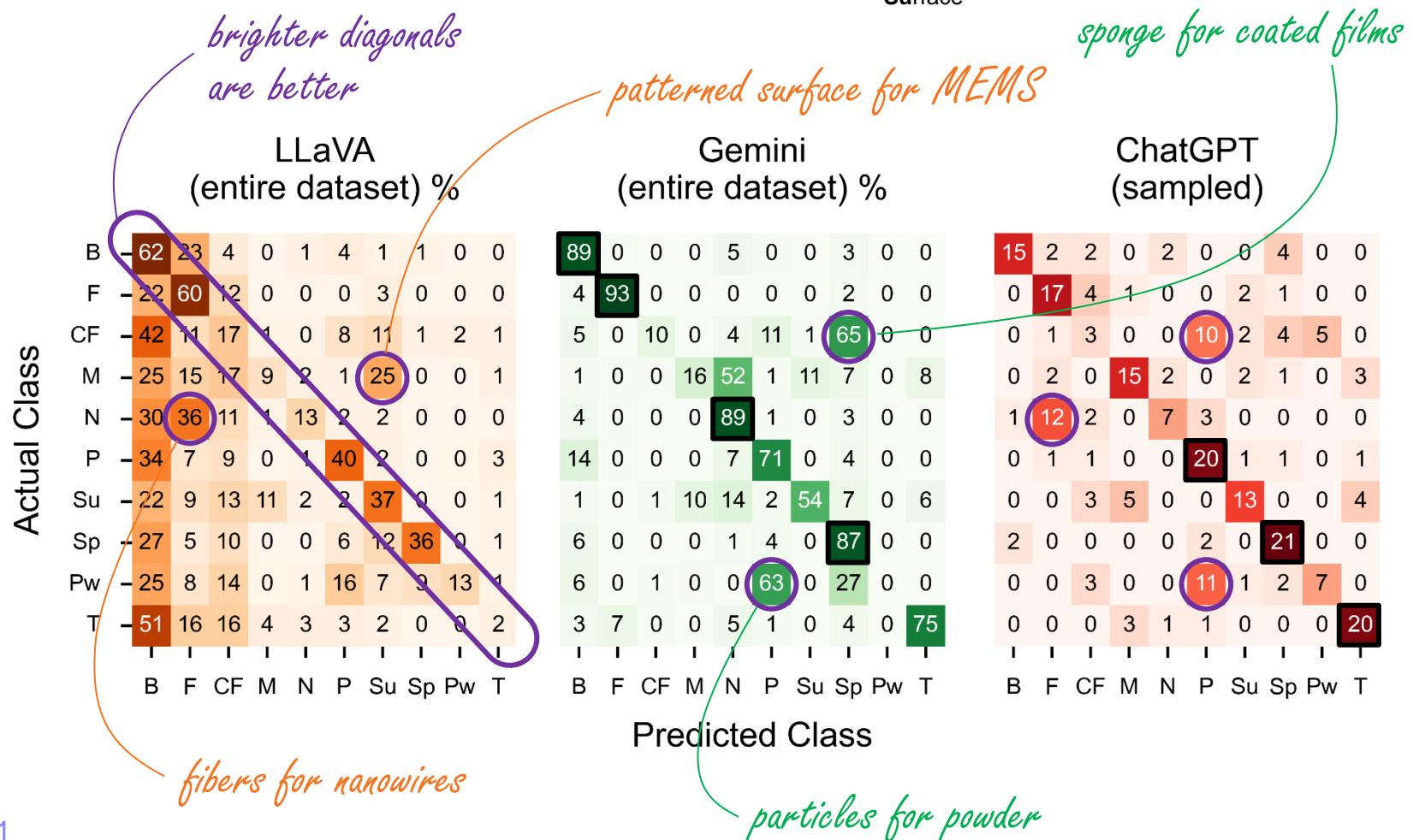
- Input: Image of a person holding a cup and the question, "What is the person holding?"
- Output: "A cup."



# 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.

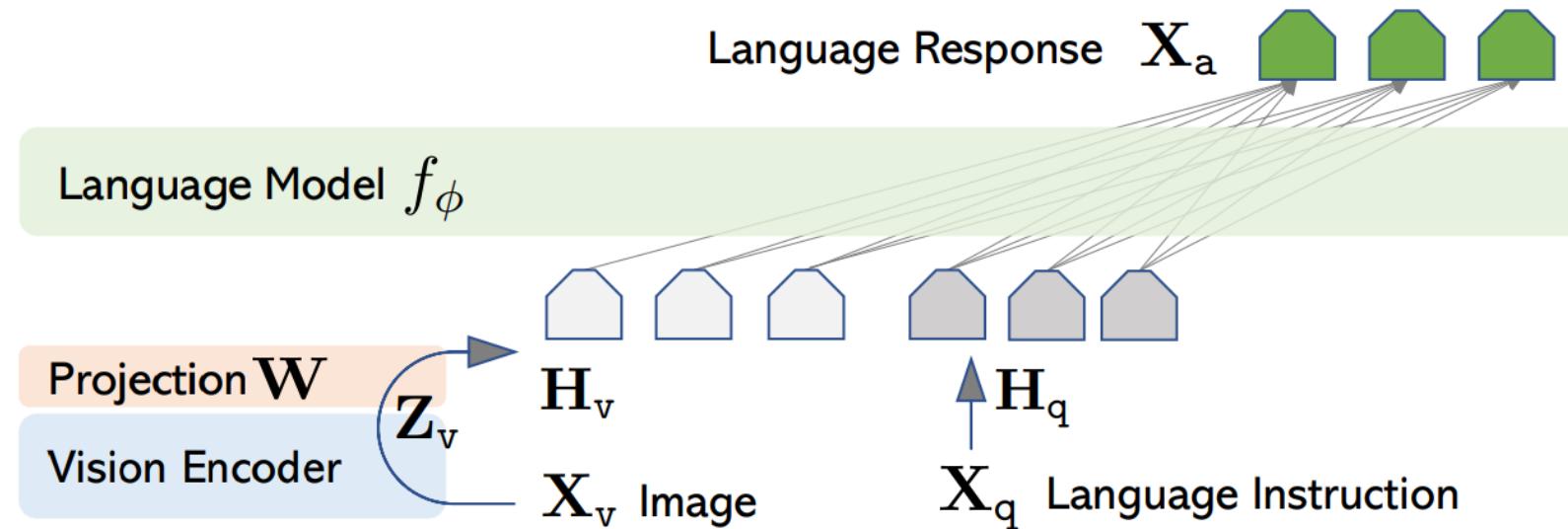
# LLaVA: Large Language and Vision Assistant

## 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)

 **LLaVA: Large Language and Vision Assistant**

[Project Page](#) | [Paper](#) | [Code](#) | [Model](#)

ONLY WORKS WITH GPU!

You can load the model with 4-bit or 8-bit quantization to make it fit in smaller hardwares. Setting the environment variable `bits` to control the quantization. Note: 8-bit seems to be slower than both 4-bit/16-bit. Although it has enough VRAM to support 8-bit, until we figure out the inference speed issue, we recommend 4-bit for A10G for the best efficiency.

Recommended configurations:

Hardware	T4-Small (16G)	A10G-Small (24G)	A100-Large (40G)
Bits	4 (default)	4	16

llava-v1.5-13b-4bit

Image

Drop Image Here  
- OR -  
Click to Upload

Describe this image

Examples

 What is unusual about this image?

The image features a close-up view of a human skull, focusing on intricate details of the bone structure. The skull appears to be a professional specimen used for educational purposes by professionals. The image provides a clear view of the skull's internal structures and the surrounding tissue.

## LLaVA-Interactive

### Image Chat, Segmentation and Generation/Editing -- All-in-one demo

Wei-Ge Chen, Irina Spiridonova, Jianwei Yang, Jianfeng Gao, Chunyuan Li

► Microsoft Research, Redmond

 arXiv  Code  Demo  Video

 As of Jun 10, 2024 the live demo or playground website is disabled. 

30

# 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.

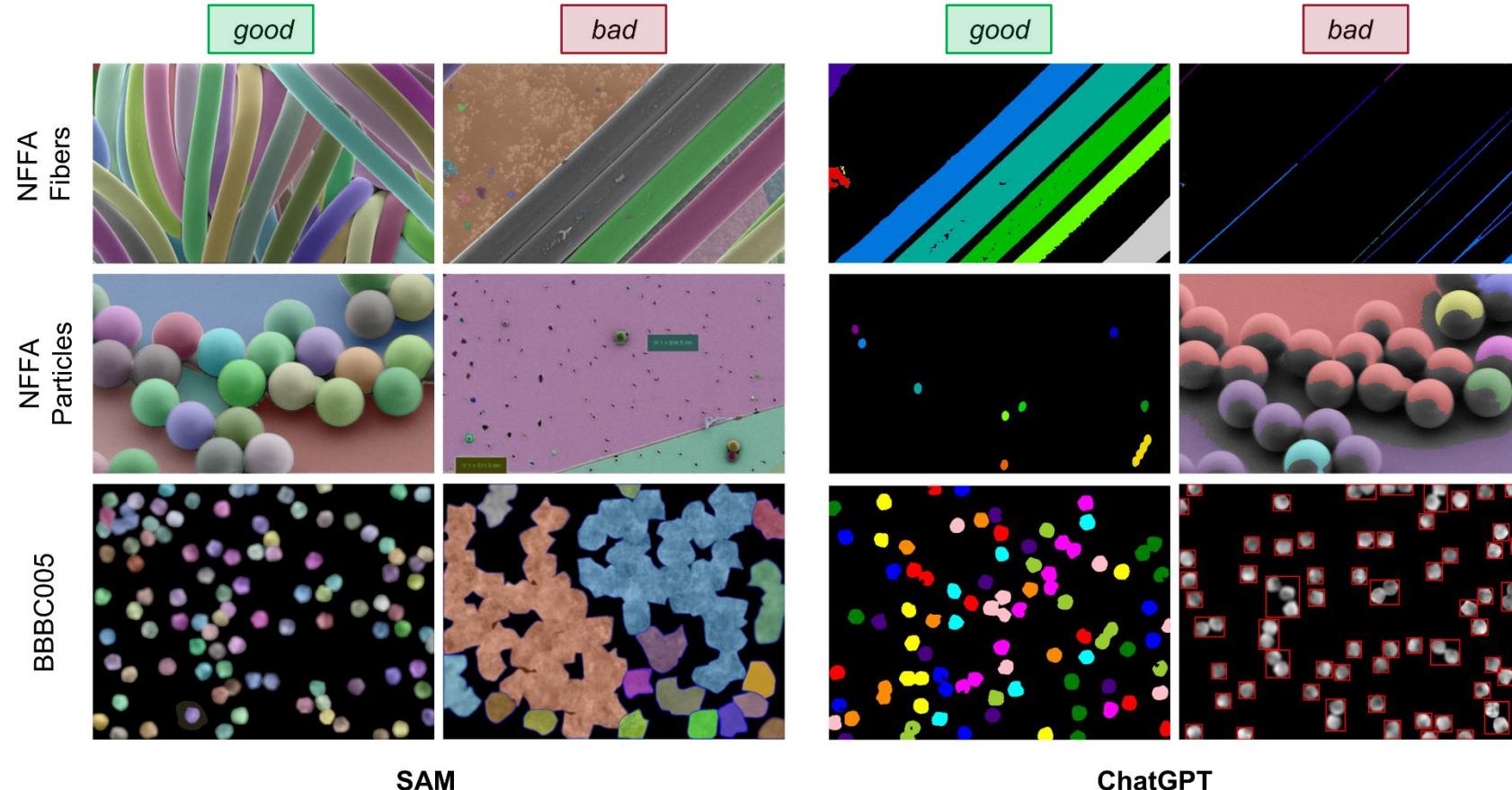
## Use Cases:

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



UNIVERSITY OF  
ARKANSAS

# Segmentation (manual)





UNIVERSITY OF  
ARKANSAS

# SAM Prompting

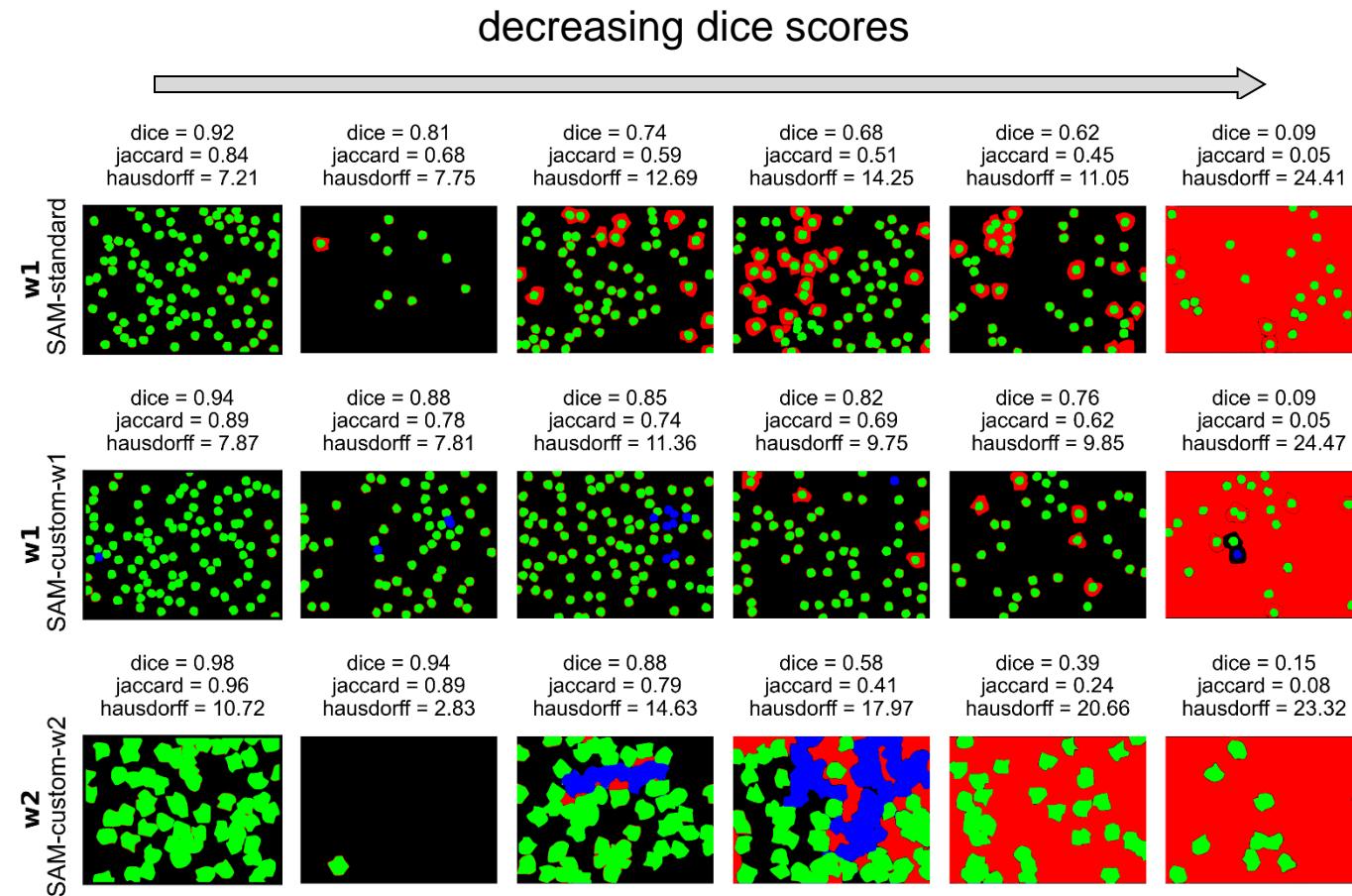


<https://segment-anything.com/>



# Segmentation

SAM segmentation quality  
measured using dice scores and other similar quantities

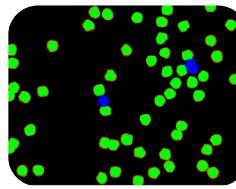
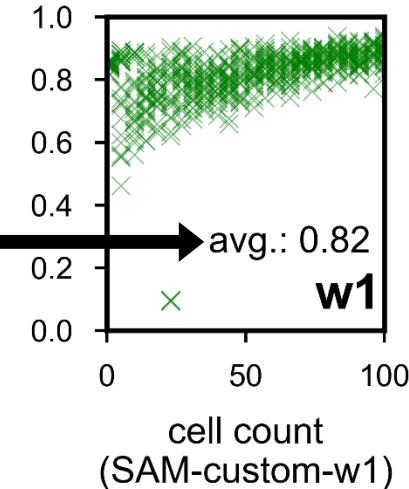
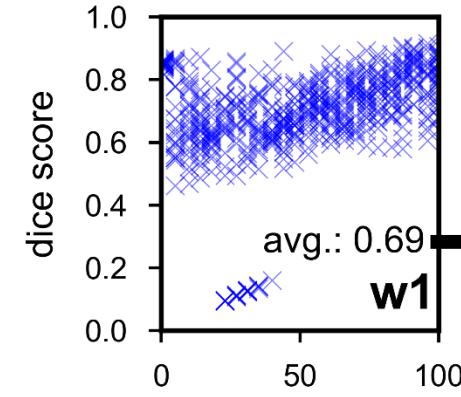
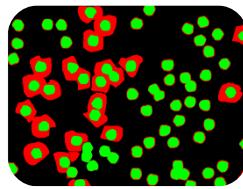


# SAM custom parameters

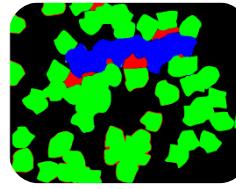
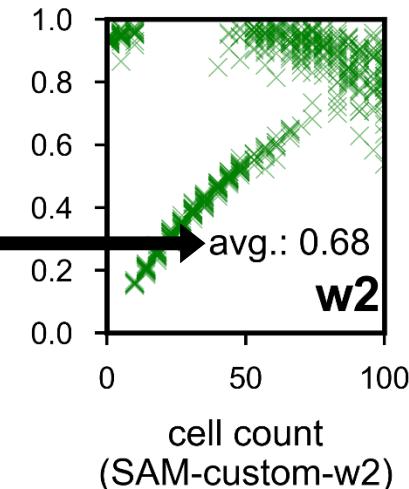
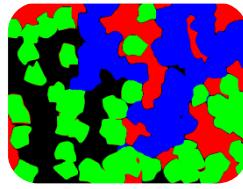
Parameter name	SAM-standard	SAM-custom -w1/w2
points_per_side	32	32
points_per_batch	64	64
<b>pred_iou_thresh</b>	0.88	0.98, 0.88
<b>stability_score_thresh</b>	0.95	0.98, 0.95
stability_score_offset	1.0	1.0
<b>box_nms_thresh</b>	0.7	0.9
crop_n_layers	0	0
<b>crop_nms_thresh</b>	0.7	0.95, 0.8
<b>crop_overlap_ratio</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)

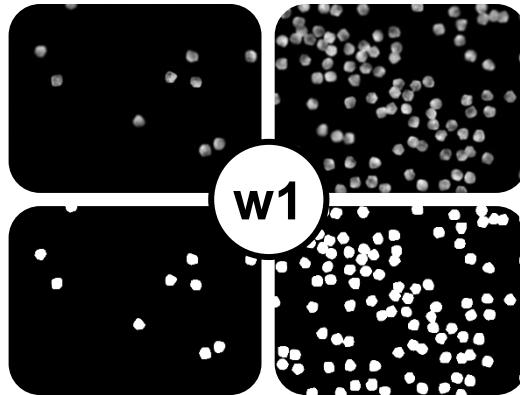


through  
parameter  
tuning

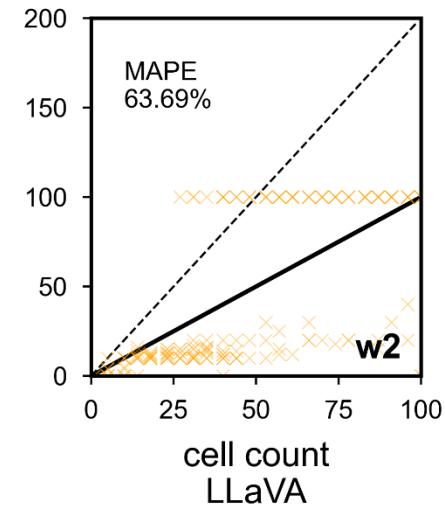
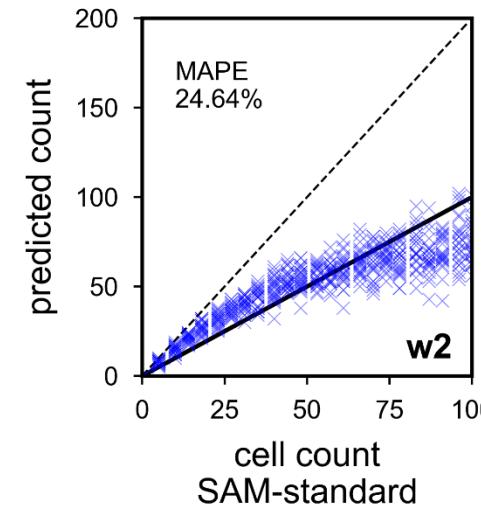
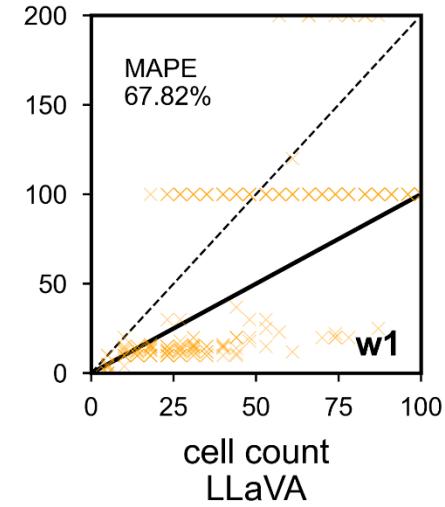
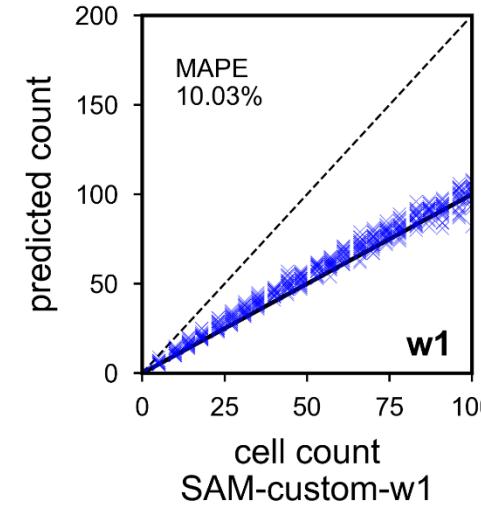
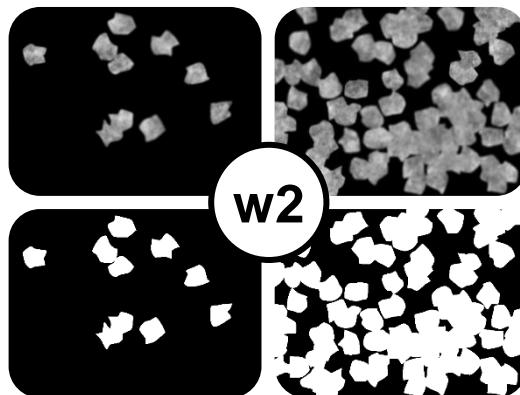




# Counting (automated)

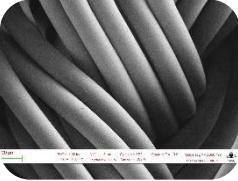


BBBC005 dataset

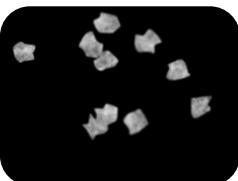
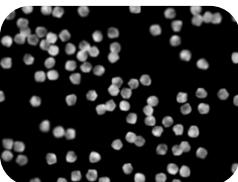
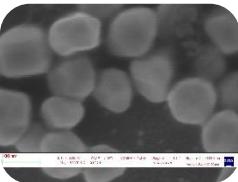
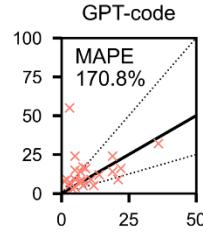
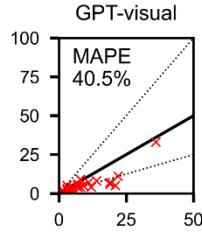




# Counting (assisted)

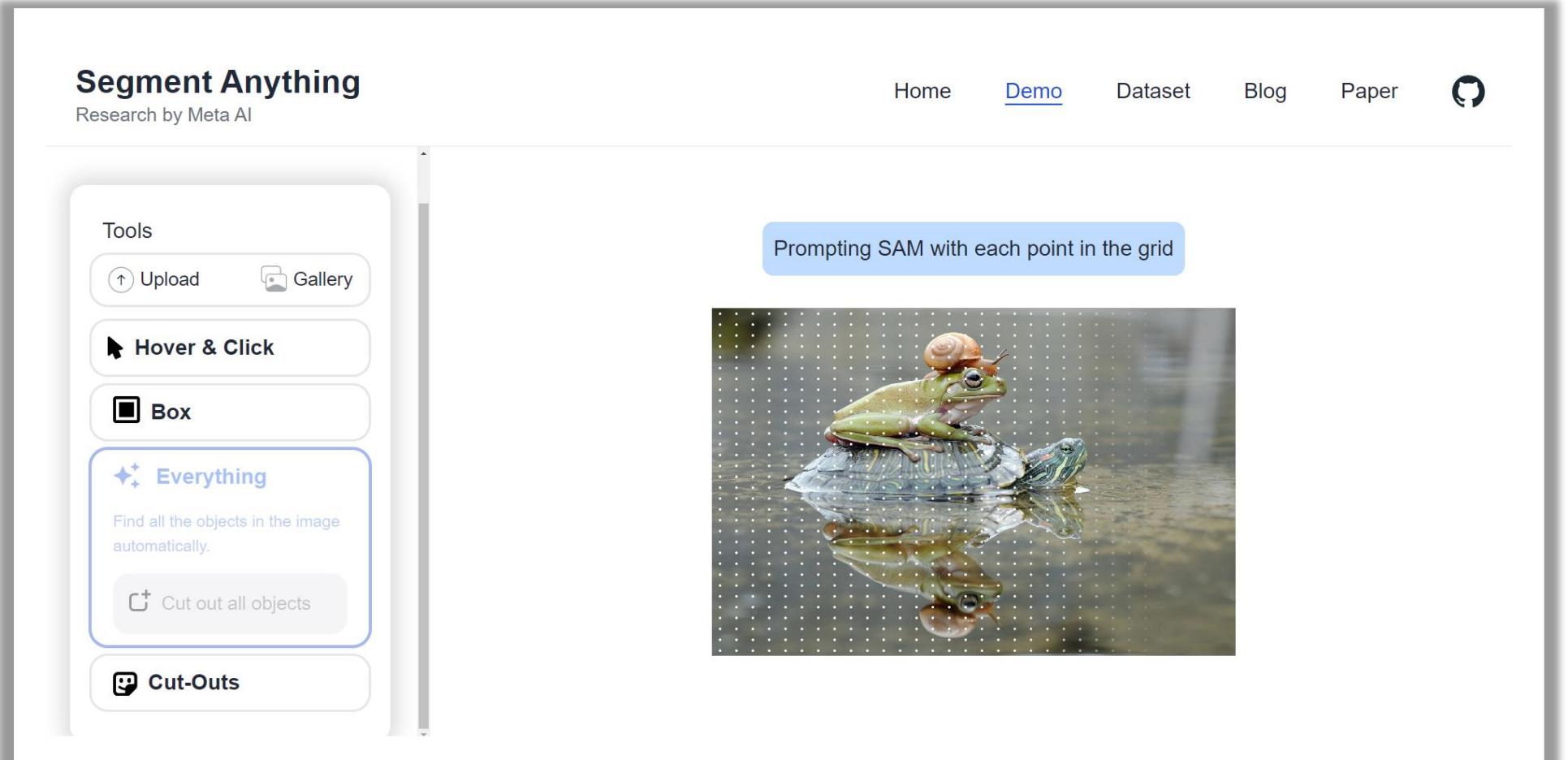


NFFA fibers counts



# Exercise 3: Segmentation (10 mins)

Follow the instructions in the notebook file.



The screenshot shows the Segment Anything demo interface. On the left, there's a sidebar titled "Tools" with options: "Upload" (with a file icon), "Gallery" (with a camera icon), "Hover & Click" (with a cursor icon), "Box" (with a square icon), "Everything" (selected, with a starburst icon), and "Cut-Outs" (with a person icon). A tooltip for "Everything" says: "Find all the objects in the image automatically." Below it is a button "Cut out all objects". The main area shows a photograph of a green frog sitting on a rock, with a small snail on its back. A white grid of points is overlaid on the image, indicating where the AI has detected objects. A blue callout bubble says "Prompting SAM with each point in the grid". At the top right, there are navigation links: Home, Demo (underlined), Dataset, Blog, Paper, and a user icon.