acm India Council

IIT Gandhinagar
ACM Summer School 2025
AI for Social Good

Vision-Language Models

From Foundations to Fine-Tuning

Shataxi Dubey, Devansh Lodha
IIT Gandhinagar

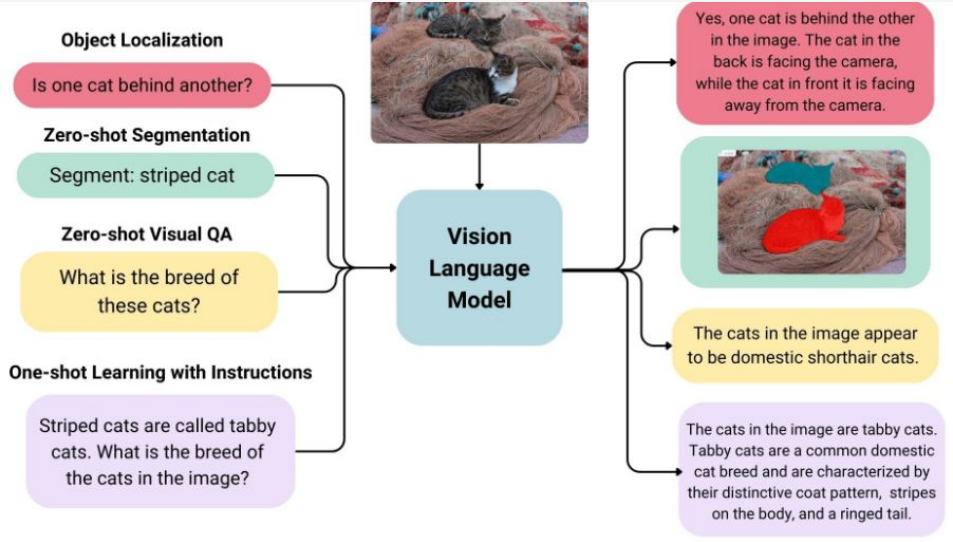
SUSTAINABILITY LAB

Seeing the World, Understanding with Words

VLMs: AI that *sees, understands,* and *localizes* through language.

Goal: Joint understanding, grounding language in visual specifics.

Impact: Precise interaction, detailed scene analysis, robotics.



Learning Objectives for This Tutorial

Today, you will:

- Understand key VLM concepts & their evolution.
- Perform inference with pre-trained VLMs (Zero-shot Classification, Captioning, VQA).
- Grasp the basics of fine-tuning VLMs.
- Gain hands-on experience fine-tuning a VLM.

Key Tools & Concepts

Core Idea: Joint Image-Text Embeddings

Key Architectures:

- Transformers (Self-Attention)
- Vision Transformers (ViT)
- Dual Encoders (e.g., CLIP)
- Encoder-Decoders / LLM-based (e.g. Qwen2.5-VL)

Tools: Python, PyTorch, Hugging Face transformers, Supervision, Unsloth

Image credits: https://blog.ngxson.com/my-journey-in-vision-models