What's New in Computer Vision?
CVPR 2025 Edition

- Foundation Models
- Multimodal
- Agents

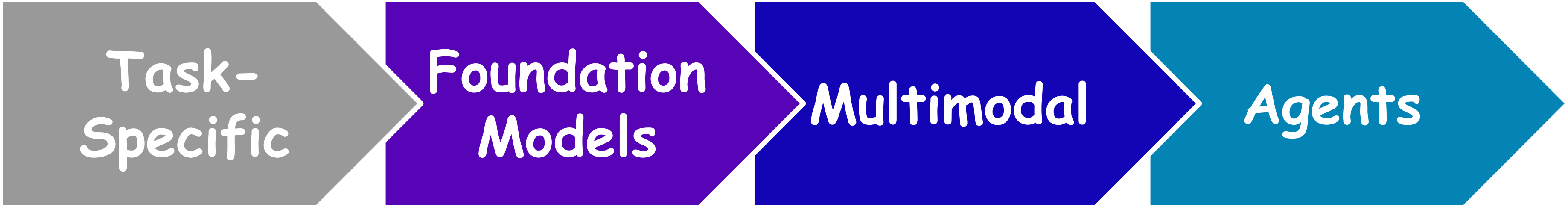
Heather Couture

June 26, 2025
11 am EDT

30 minutes + Q&A/Discussion

Pixel Scientia
LABS

Today's Focus


Task-Specific

Foundation Models

Multimodal

Agents

Who am I?

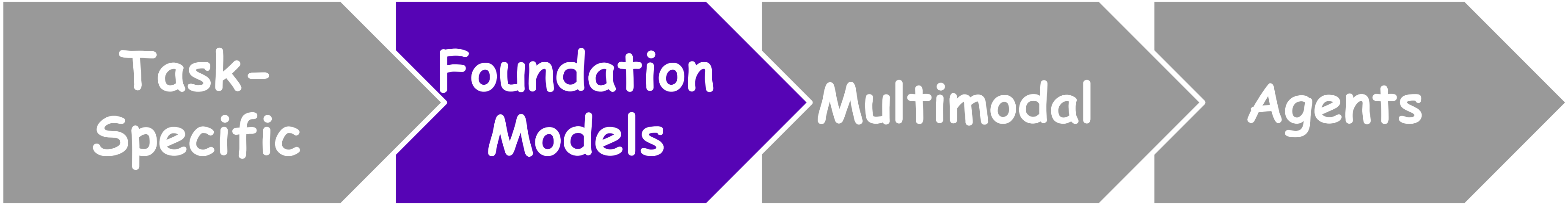
- Heather Couture
- MS from Carnegie Mellon University - autonomous science
- PhD from University of North Carolina - computational pathology
- Computer vision consultant

OWKIN DECIPHEX ULTIVUE WattTime Leica MICROSYSTEMS QRITIVE

- Keynote speaker at MICCAI workshop on computational pathology
- Contributor to Scientific American, The Pathologist, IEEE Spectrum
- Newsletter and podcast

Computer Vision Insights
by Pixel Scientia Labs

IMPACT AI
WITH HEATHER COUTURE



Task-Specific

Foundation Models

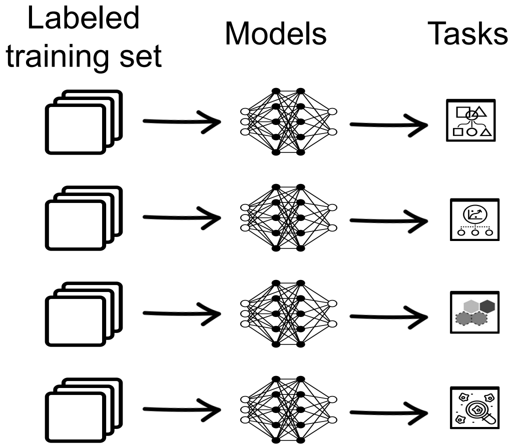
Multimodal

Agents

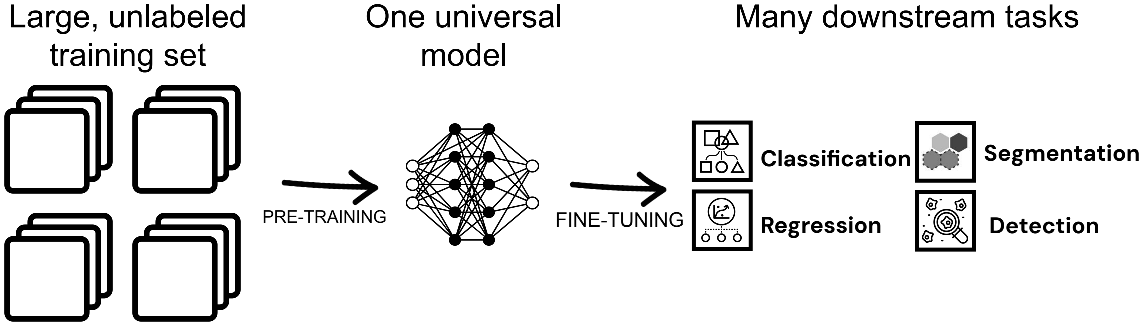
- New domains
- Domain-specific adaptations
- New datasets
- New benchmarks

The Paradigm Shift

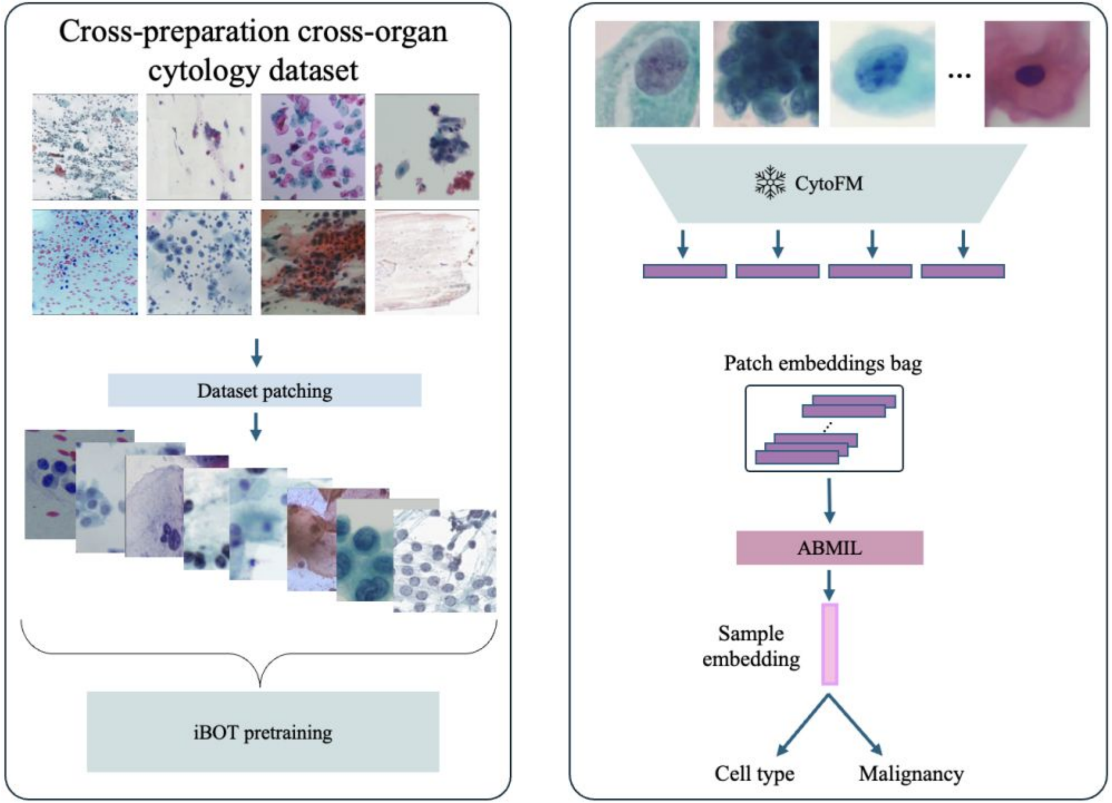
Traditional ML



Foundation Models



A Foundation Model for Cytology

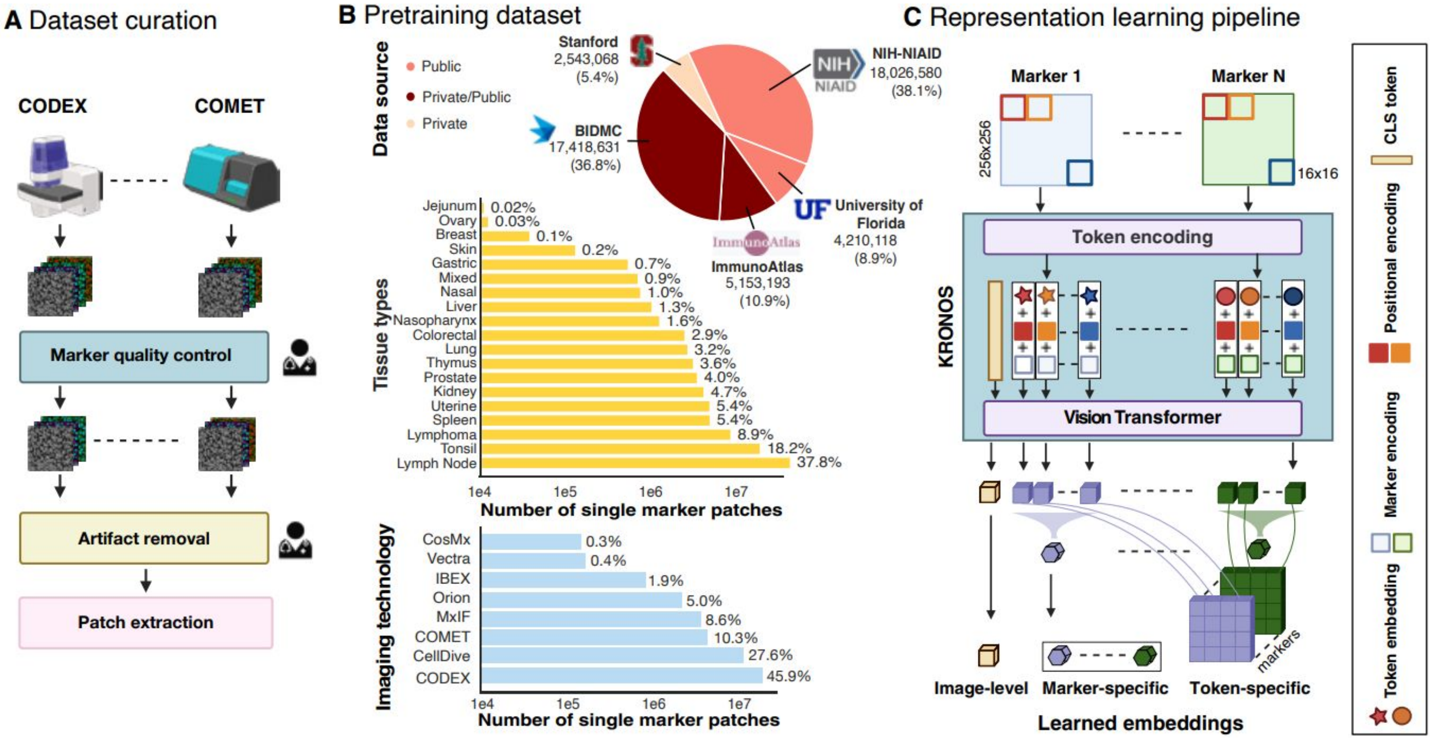


Ivezic, CytoFM: The first cytology foundation model, 2025

`CytoFM`

1.4 million image patches
8 datasets
7 institutions

A Foundation Model for Spatial Proteomics



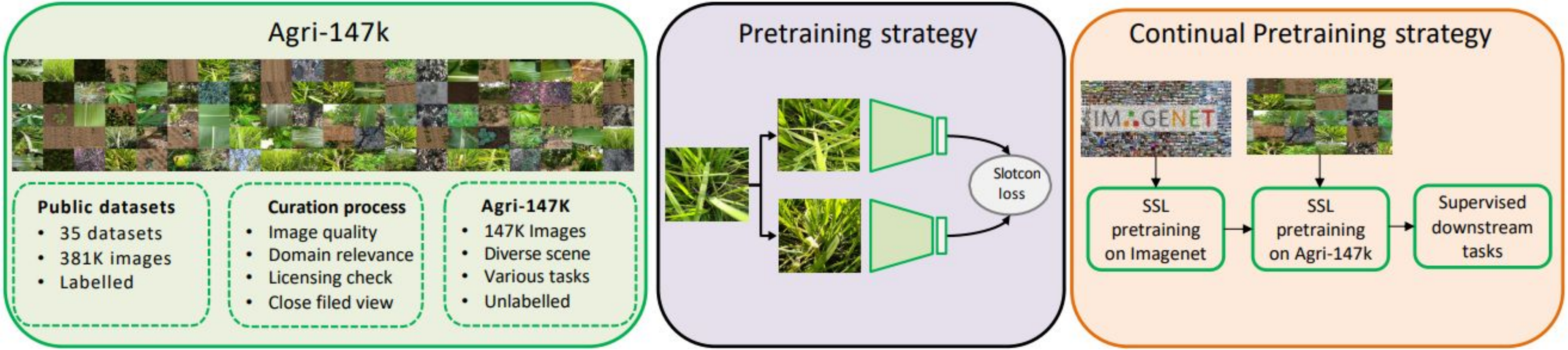
KRONOS

47 million image patches
175 protein markers
16 tissue types
8 imaging platforms

Shaban, A Foundation Model for Spatial Proteomics,2025

A Foundation Model for Agriculture

Agri-FM+



Agri-147k

Public datasets
- 35 datasets
- 381K images
- Labelled

Curation process
- Image quality
- Domain relevance
- Licensing check
- Close filed view

Agri-147K
- 147K Images
- Diverse scene
- Various tasks
- Unlabelled

Pretraining strategy

Slotcon loss

Continual Pretraining strategy

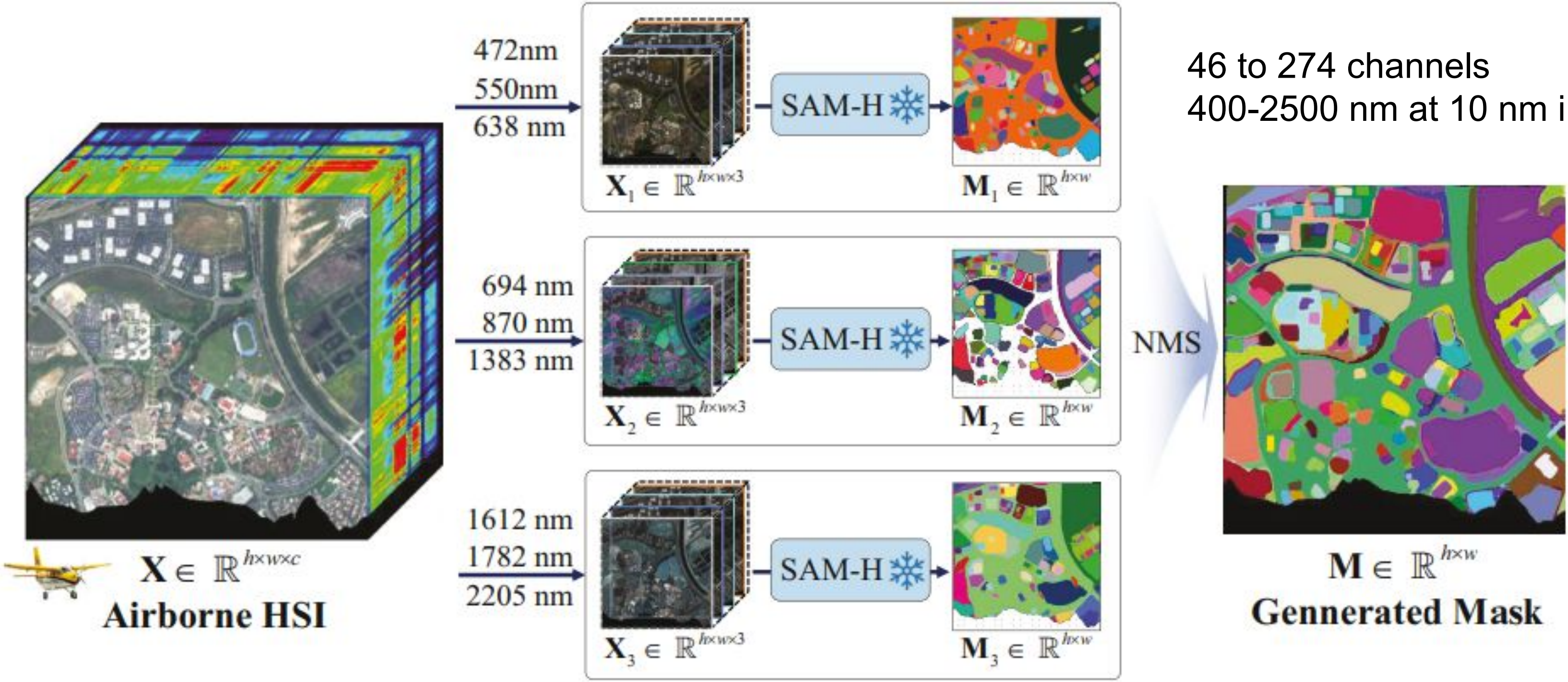
SSL pretraining on Imagenet → SSL pretraining on Agri-147k → Supervised downstream tasks

Nahian, Agri-FM+: A Self-Supervised Foundation Model for Agricultural Vision, 2025

A Foundation Model for Hyperspectral Imagery

`HyperFree`

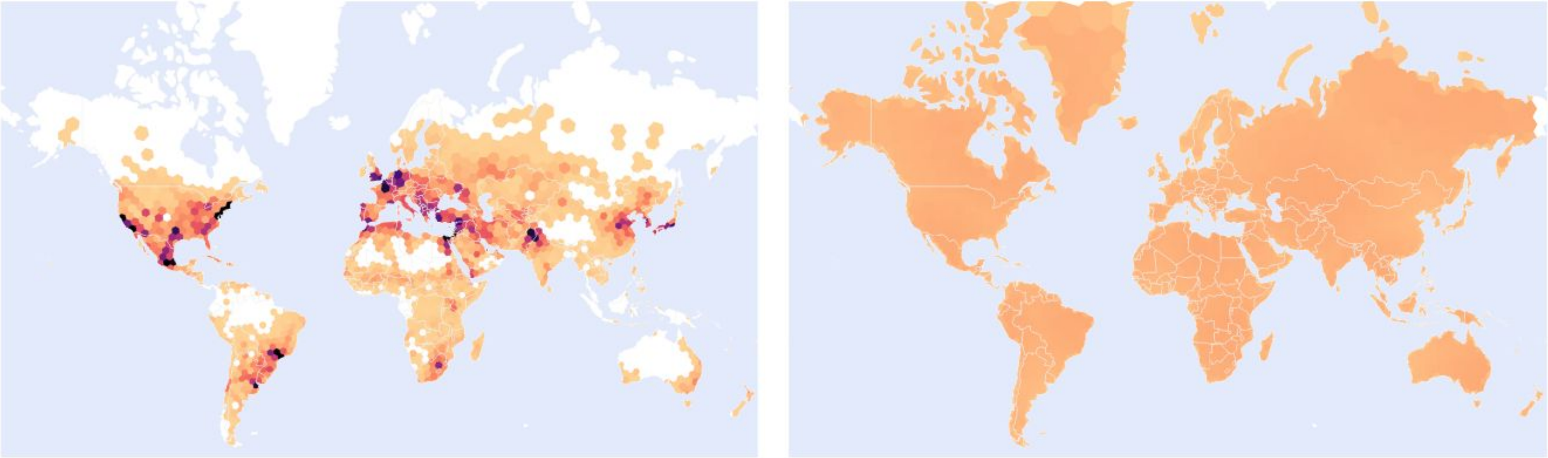
46 to 274 channels
400-2500 nm at 10 nm intervals



472nm
550nm
638 nm
$X_1 \in \mathbb{R}^{h\times w\times 3}$
SAM-H
$M_1 \in \mathbb{R}^{h\times w}$
694 nm
870 nm
1383 nm
$X_2 \in \mathbb{R}^{h\times w\times 3}$
SAM-H
$M_2 \in \mathbb{R}^{h\times w}$
1612 nm
1782 nm
2205 nm
$X_3 \in \mathbb{R}^{h\times w\times 3}$
SAM-H
$M_3 \in \mathbb{R}^{h\times w}$
NMS
$X \in \mathbb{R}^{h\times w\times c}$
Airborne HSI
$M \in \mathbb{R}^{h\times w}$
Gennerated Mask

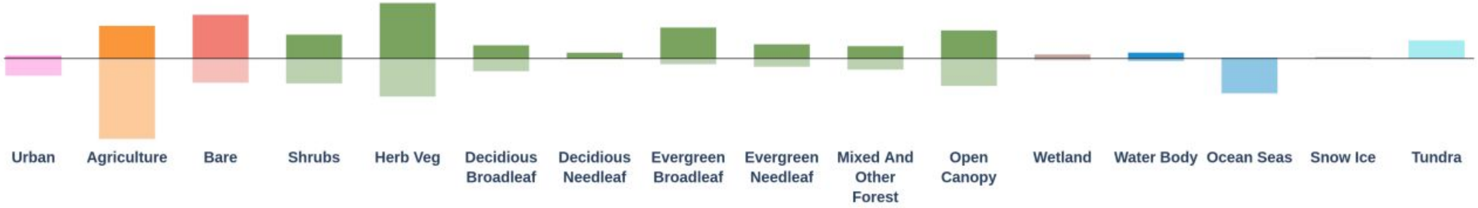
Li, HyperFree: A Channel-adaptive and Tuning-free Foundation
Model for Hyperspectral Remote Sensing Imagery, 2025

Remote Sensing Dataset: Global Distribution



(a) Spatial distribution of SSL4EO-S12 [89]

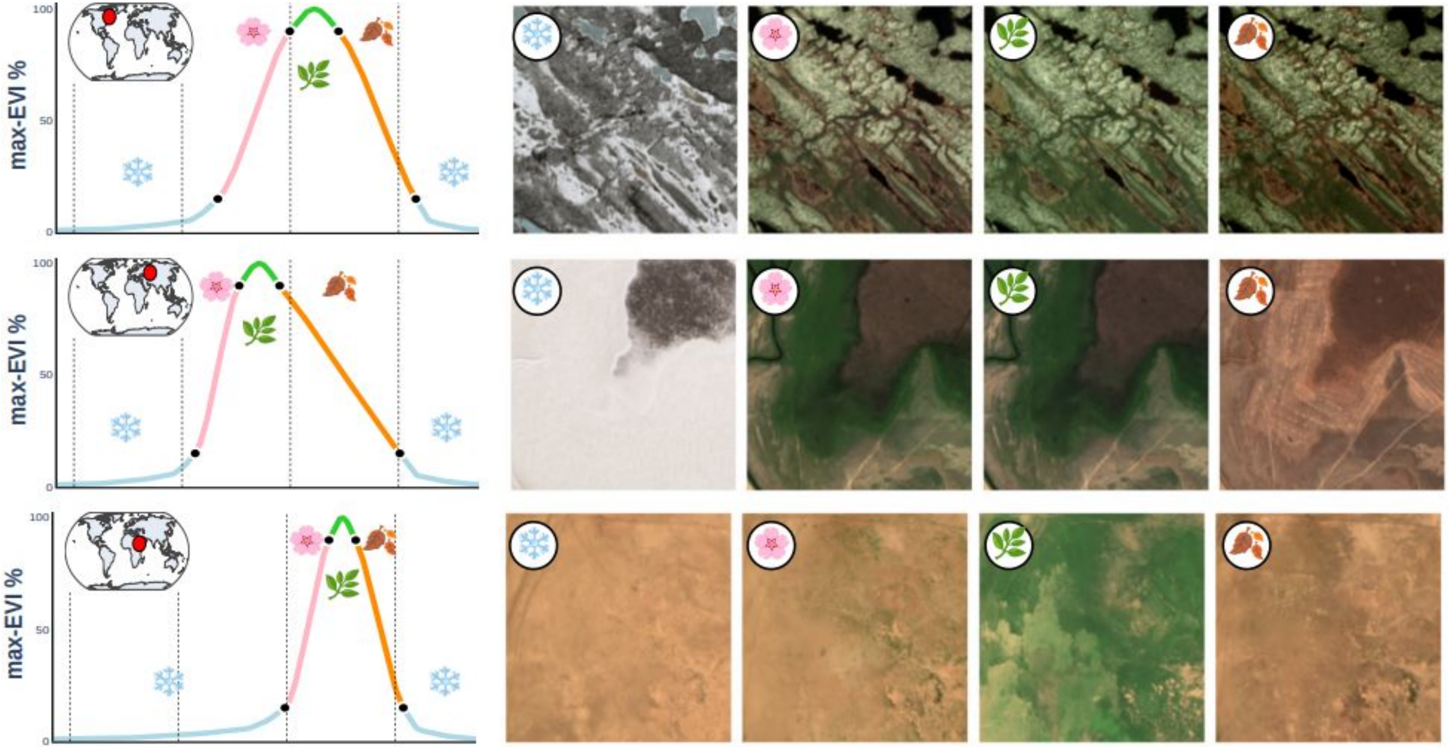
(b) Spatial distribution of SSL4Eco



(c) Copernicus land cover [55] distribution for SSL4Eco (upwards) and SSL4EO-S12 [89] (downwards)

Plekhanova, SSL4Eco: A Global Seasonal Dataset for Geospatial Foundation Models in Ecology, 2025

Remote Sensing Dataset: Seasonal Distribution



(a) EVI-based seasons

(b) Seasonal images

SSL4Eco

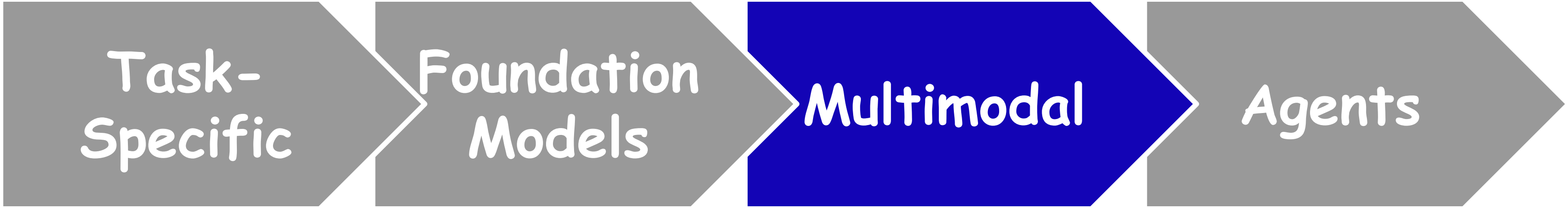
250k locations across entire landmass

Enhanced Vegetation Index-based seasonal sampling

SeCo-Eco

Seasonal contrastive learning

Plekhanova, SSL4Eco: A Global Seasonal Dataset for Geospatial Foundation Models in Ecology, 2025



Task-Specific

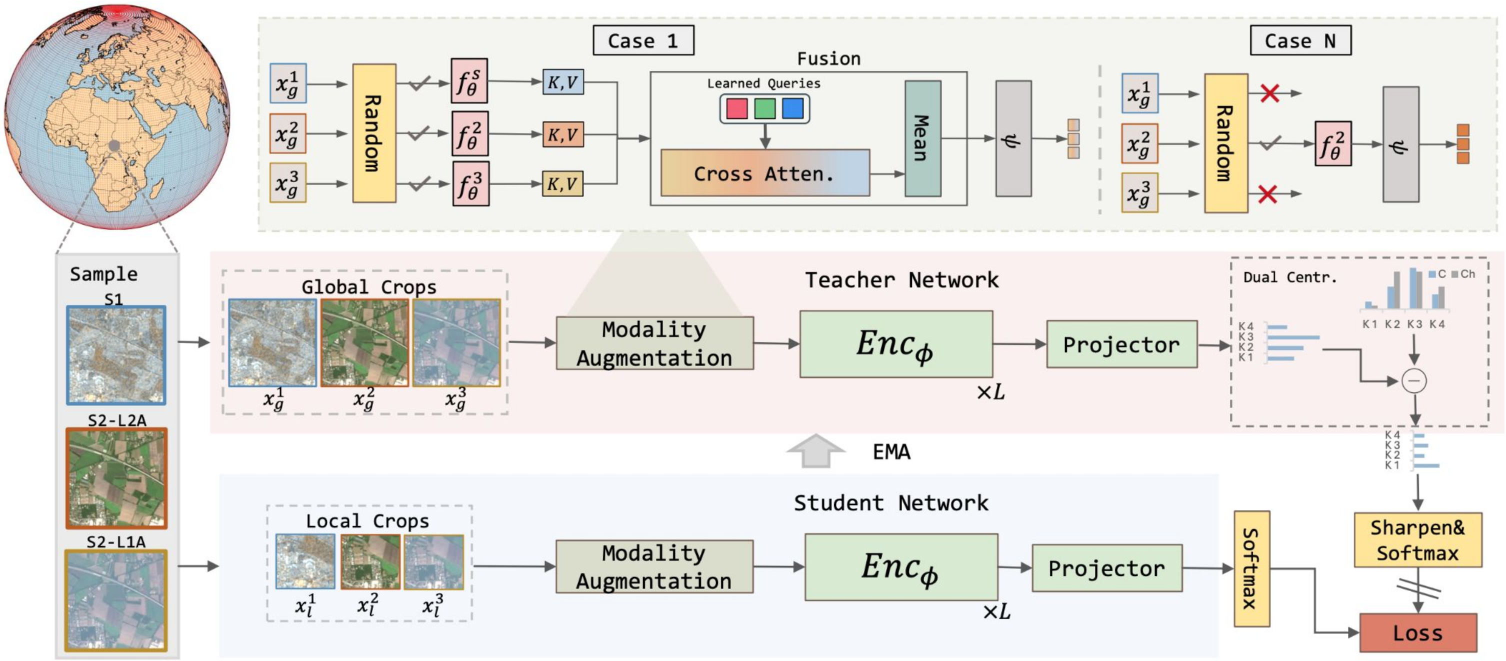
Foundation Models

Multimodal

Agents

- Open weights and open data
- Grounding
- VLMs - language is the “glue”
- Multi-stage pretraining
- Larger inputs
- Benchmarks

Remote Sensing: Optical + Radar

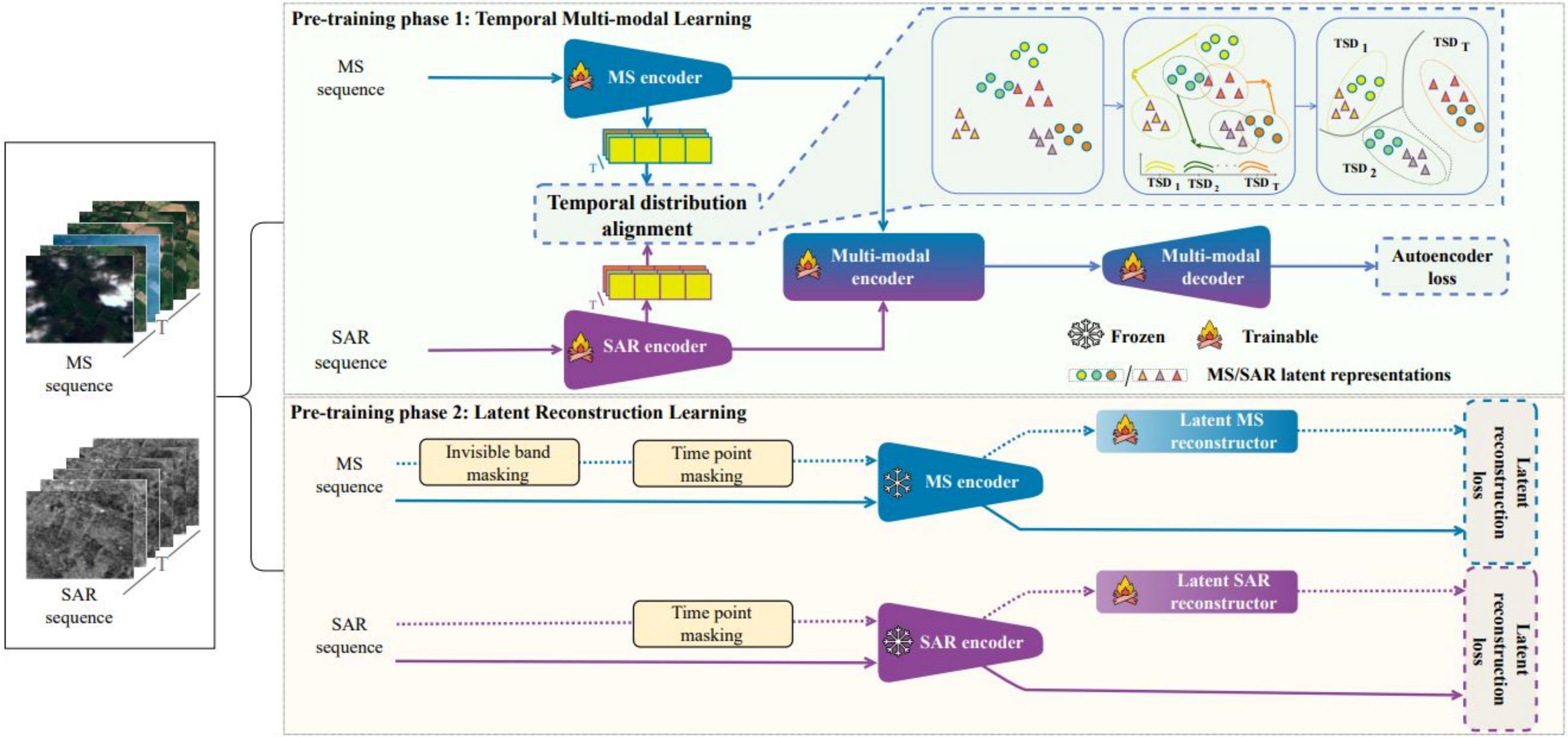


`TerraFM`

SAR and optical
18.7 million tiles
534x534 tiles

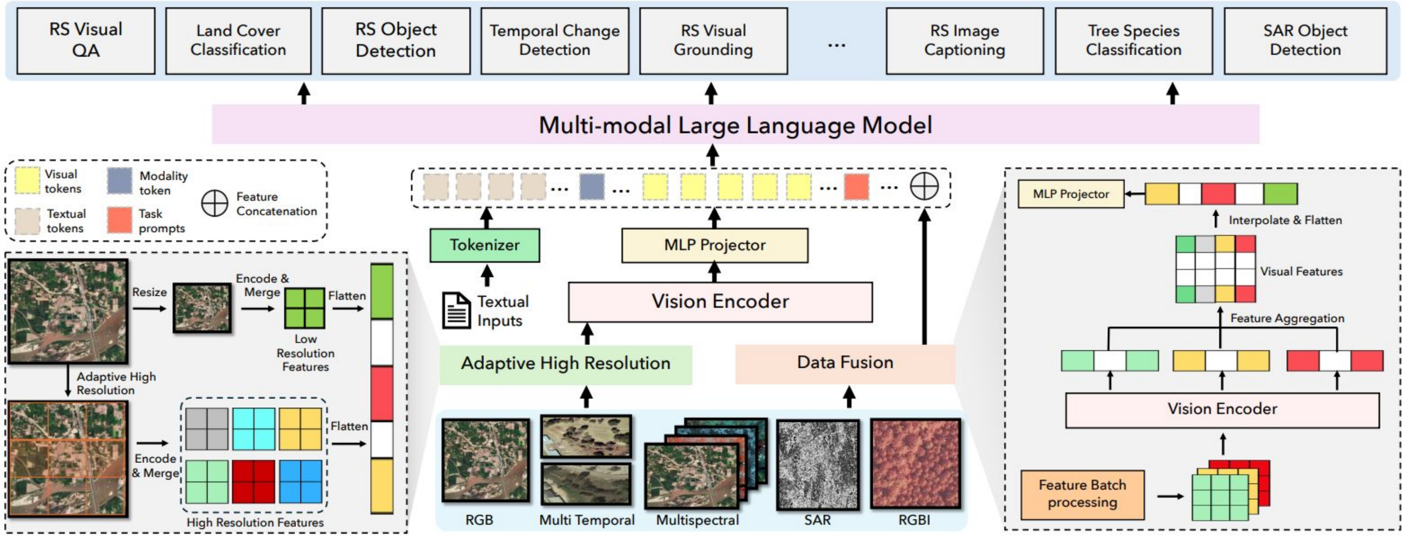
Danish, TerraFM: A Scalable Foundation Model for Unified Multisensor Earth Observation, 2025

Remote Sensing: Robust to Missing Data



Do, RobSense: A Robust Multi-modal Foundation Model for Remote Sensing with Static, Temporal, and Incomplete Data Adaptability, 2025

A VLM for Remote Sensing



`EarthDial`

RGB, multispectral, infrared, SAR
11 million instruction pairs

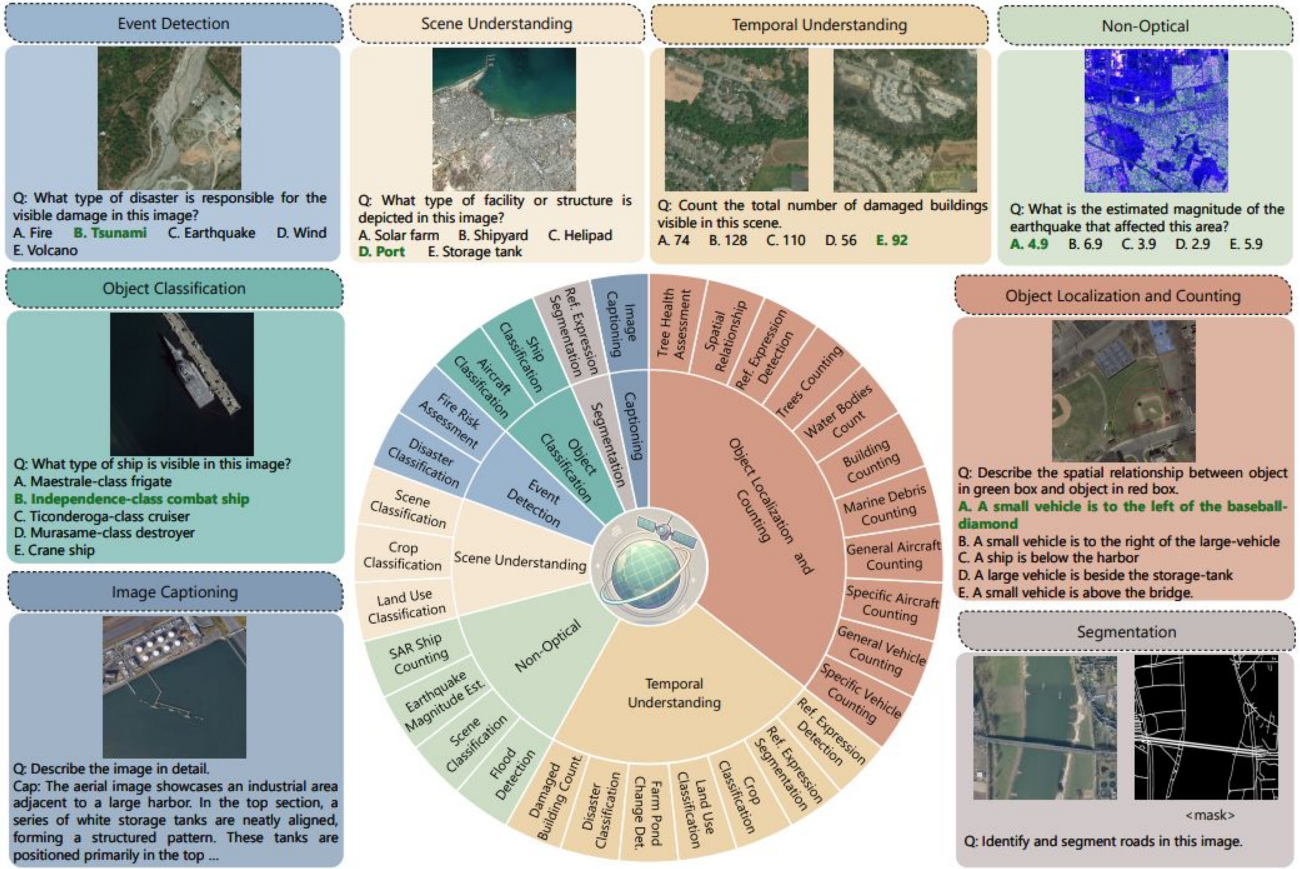
Soni, EarthDial: Turning Multi-sensory Earth Observations to Interactive Dialogues, 2025

Benchmarking VLMs for Remote Sensing

`GEOBench-VLM`

31 fine-grained tasks
8 categories
Optical, multispectral, SAR, temporal
10+ manually verified instructions
Multiple-choice format

Danish, GEOBench-VLM: Benchmarking Vision-Language Models for Geospatial Tasks, 2025



Event Detection

Q: What type of disaster is responsible for the visible damage in this image?
A. Fire **B. Tsunami** C. Earthquake D. Wind
E. Volcano

Scene Understanding

Q: What type of facility or structure is depicted in this image?
A. Solar farm B. Shipyard C. Helipad
D. Port E. Storage tank

Temporal Understanding

Q: Count the total number of damaged buildings visible in this scene.
A. 74 B. 128 C. 110 D. 56 **E. 92**

Non-Optical

Q: What is the estimated magnitude of the earthquake that affected this area?
A. 4.9 B. 6.9 C. 3.9 D. 2.9 E. 5.9

Object Classification

Q: What type of ship is visible in this image?
A. Maestrale-class frigate
B. Independence-class combat ship
C. Ticonderoga-class cruiser
D. Murasame-class destroyer
E. Crane ship

Object Localization and Counting

Q: Describe the spatial relationship between object in green box and object in red box.
A. A small vehicle is to the left of the baseball-diamond
B. A small vehicle is to the right of the large-vehicle
C. A ship is below the harbor
D. A large vehicle is beside the storage-tank
E. A small vehicle is above the bridge.

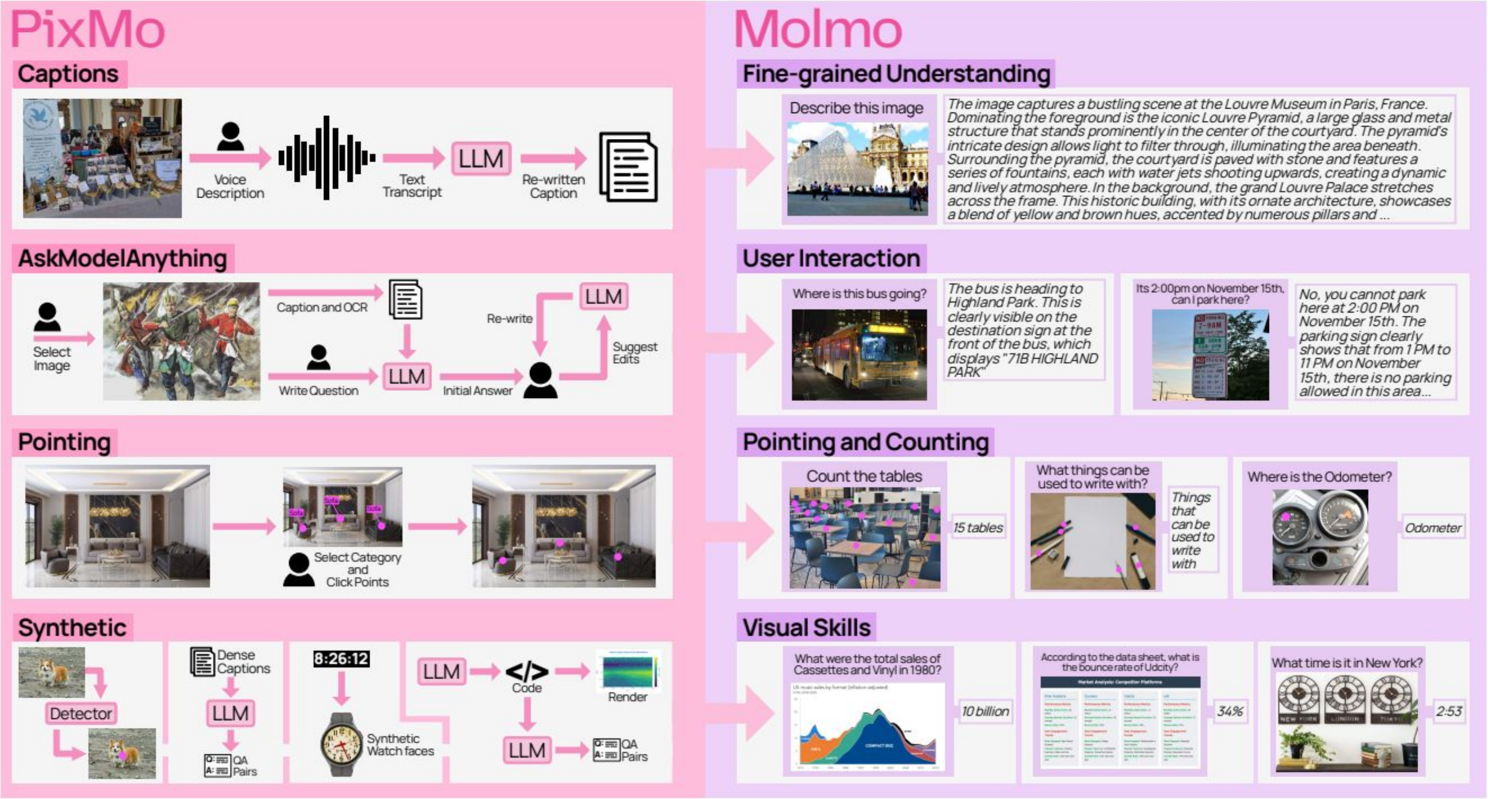
Image Captioning

Q: Describe the image in detail.
Cap: The aerial image showcases an industrial area adjacent to a large harbor. In the top section, a series of white storage tanks are neatly aligned, forming a structured pattern. These tanks are positioned primarily in the top ...

Segmentation

Q: Identify and segment roads in this image.

A VLM with Open Weights and Open Data



Describe this image

The image captures a bustling scene at the Louvre Museum in Paris, France. Dominating the foreground is the iconic Louvre Pyramid, a large glass and metal structure that stands prominently in the center of the courtyard. The pyramid's intricate design allows light to filter through, illuminating the area beneath. Surrounding the pyramid, the courtyard is paved with stone and features a series of fountains, each with water jets shooting upwards, creating a dynamic and lively atmosphere. In the background, the grand Louvre Palace stretches across the frame. This historic building, with its ornate architecture, showcases a blend of yellow and brown hues, accented by numerous pillars and ...

Where is this bus going?

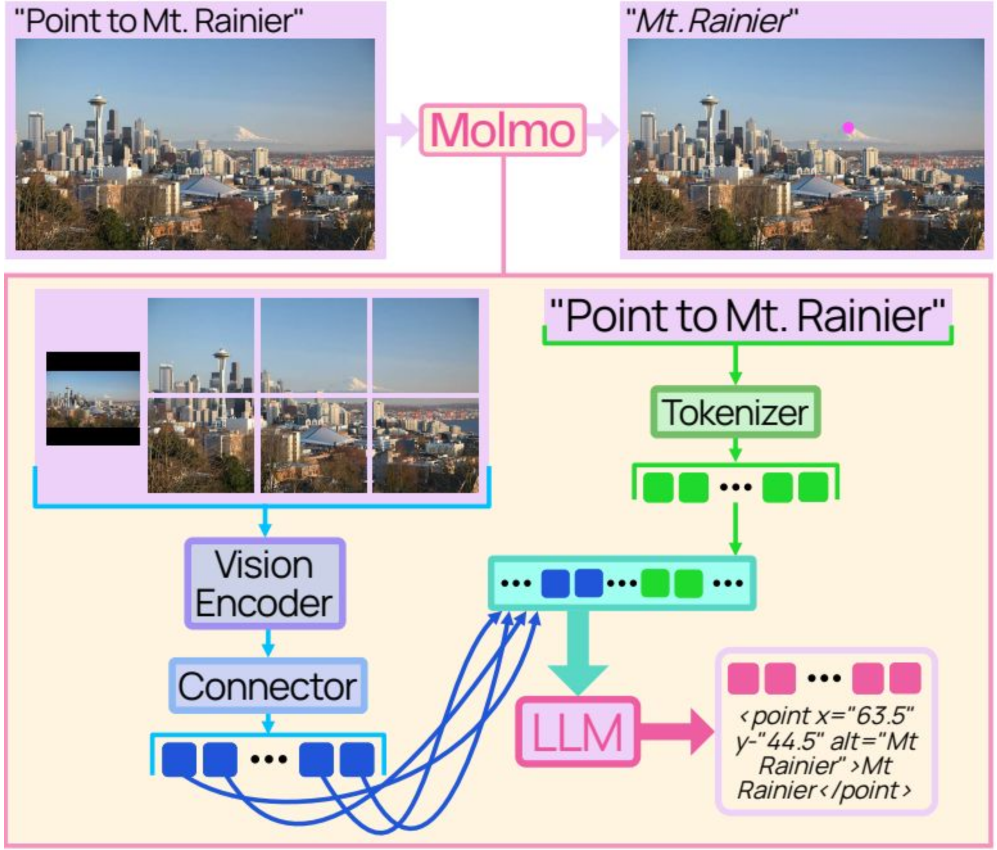
The bus is heading to Highland Park. This is clearly visible on the destination sign at the front of the bus, which displays "71B HIGHLAND PARK"

Its 2:00pm on November 15th, can I park here?

No, you cannot park here at 2:00 PM on November 15th. The parking sign clearly shows that from 1 PM to 11 PM on November 15th, there is no parking allowed in this area...

Deitke, Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models, 2025

A VLM with Open Weights and Open Data + Grounding

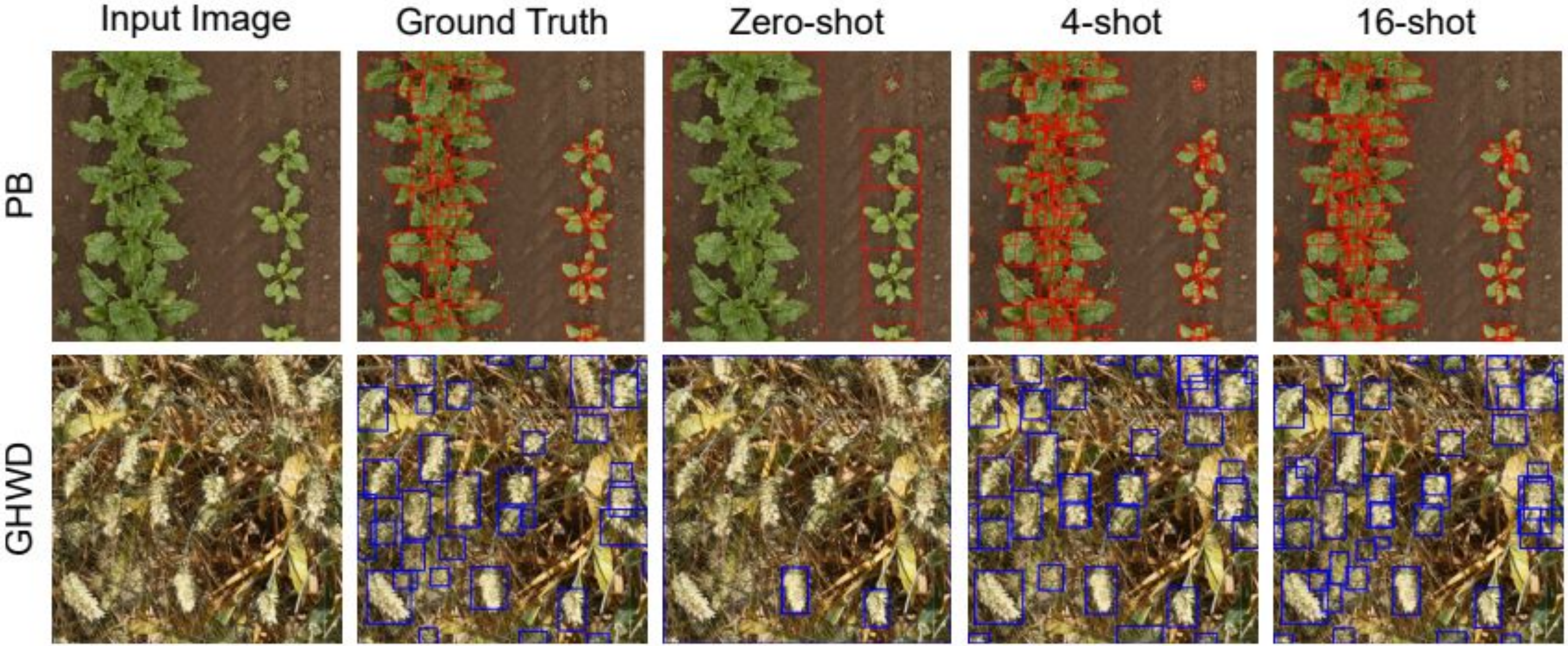


```
<point x="63.5"
y-"44.5" alt="Mt
Rainier" >Mt
Rainier</point>
```

Deitke, Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models, 2025

Grounding for Agriculture

Prompt: wheat spike head



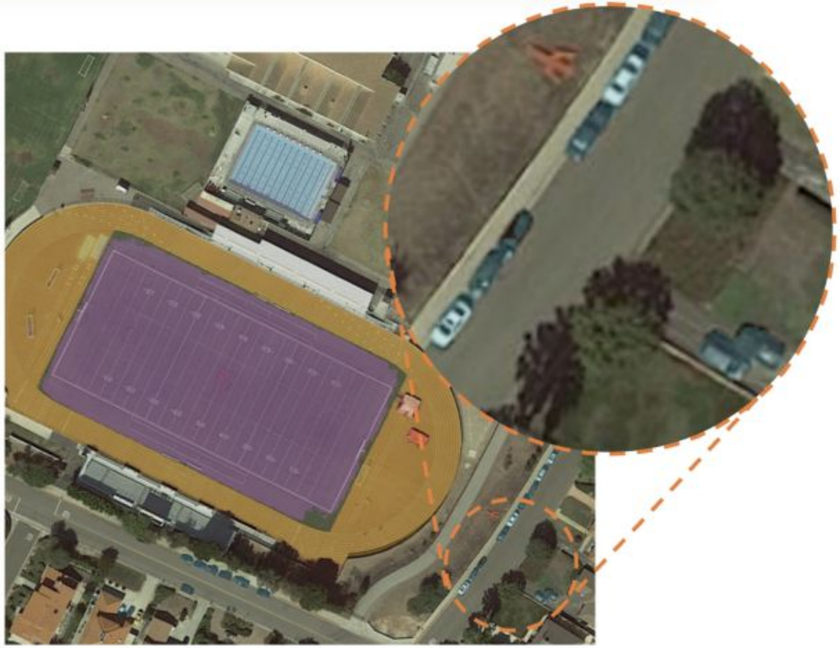
Singh, Few-Shot Adaptation of Grounding DINO for Agricultural Domain, 2025

Remote Sensing VLM with Grounding

GeoPixel

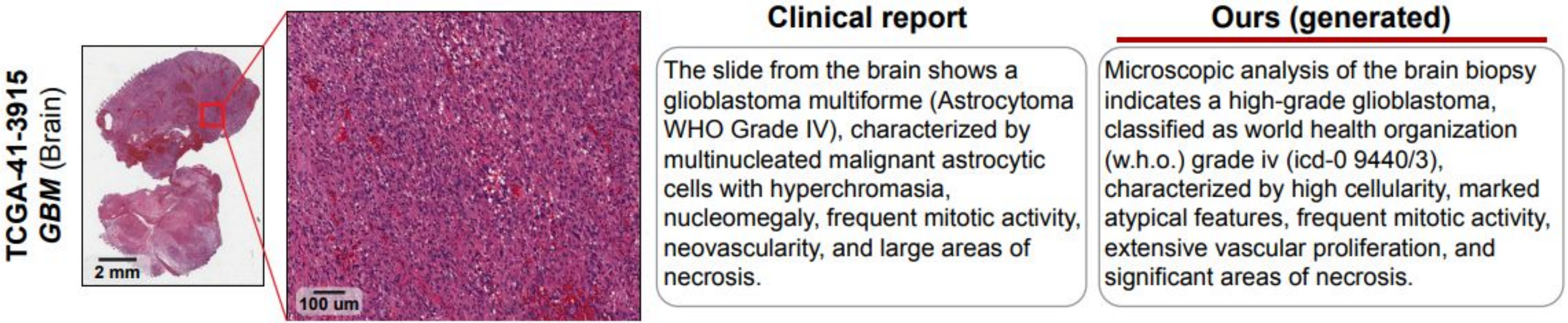
Can you give a thorough description of this image, including interleaved segmentation masks to highlight key objects?

The image is an aerial view of a sports complex featuring a running track and a swimming pool. The track has multiple lanes, and the field is marked for various sports. The surrounding area includes roads, buildings, and green spaces. The ground track field at the center is a large athletic venue with a multi lane running track encircling a grassy field marked for various sports. The swimming pool at the top is large, rectangular, and filled with a blue substance, likely water. Numerous small vehicles can be seen at different locations. A soccer field within the sports complex features a rectangular pitch with goals at each end, surrounded by a running track. The image gives a static and undisturbed appearance as no people or moving objects are visible.



Shabbir, GeoPixel: Pixel Grounding Large Multimodal Model in Remote Sensing, 2025

A VLM for Whole Slide Images - Pathology Reports



Clinical report

The slide from the brain shows a glioblastoma multiforme (Astrocytoma WHO Grade IV), characterized by multinucleated malignant astrocytic cells with hyperchromasia, nucleomegaly, frequent mitotic activity, neovascularity, and large areas of necrosis.

Ours (generated)

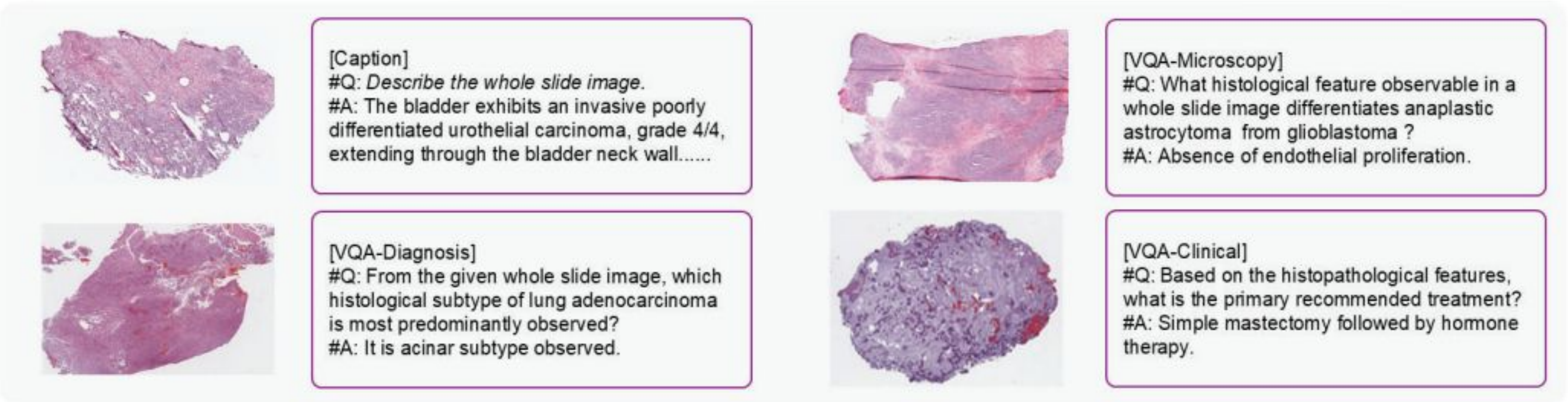
Microscopic analysis of the brain biopsy indicates a high-grade glioblastoma, classified as world health organization (w.h.o.) grade iv (icd-0 9440/3), characterized by high cellularity, marked atypical features, frequent mitotic activity, extensive vascular proliferation, and significant areas of necrosis.

Ding, Multimodal Whole Slide Foundation Model for Pathology, 2025

`TITAN`

335k WSIs with 423k synthetic captions
8192 x 8192 tiles
Three-stage training:
- SSL on patches
- Vision-language alignment
- Alignment with pathology reports

A VLM for Whole Slide Images - VQA



[Caption]
#Q: *Describe the whole slide image.*
#A: The bladder exhibits an invasive poorly differentiated urothelial carcinoma, grade 4/4, extending through the bladder neck wall......

[VQA-Microscopy]
#Q: What histological feature observable in a whole slide image differentiates anaplastic astrocytoma from glioblastoma ?
#A: Absence of endothelial proliferation.

[VQA-Diagnosis]
#Q: From the given whole slide image, which histological subtype of lung adenocarcinoma is most predominantly observed?
#A: It is acinar subtype observed.

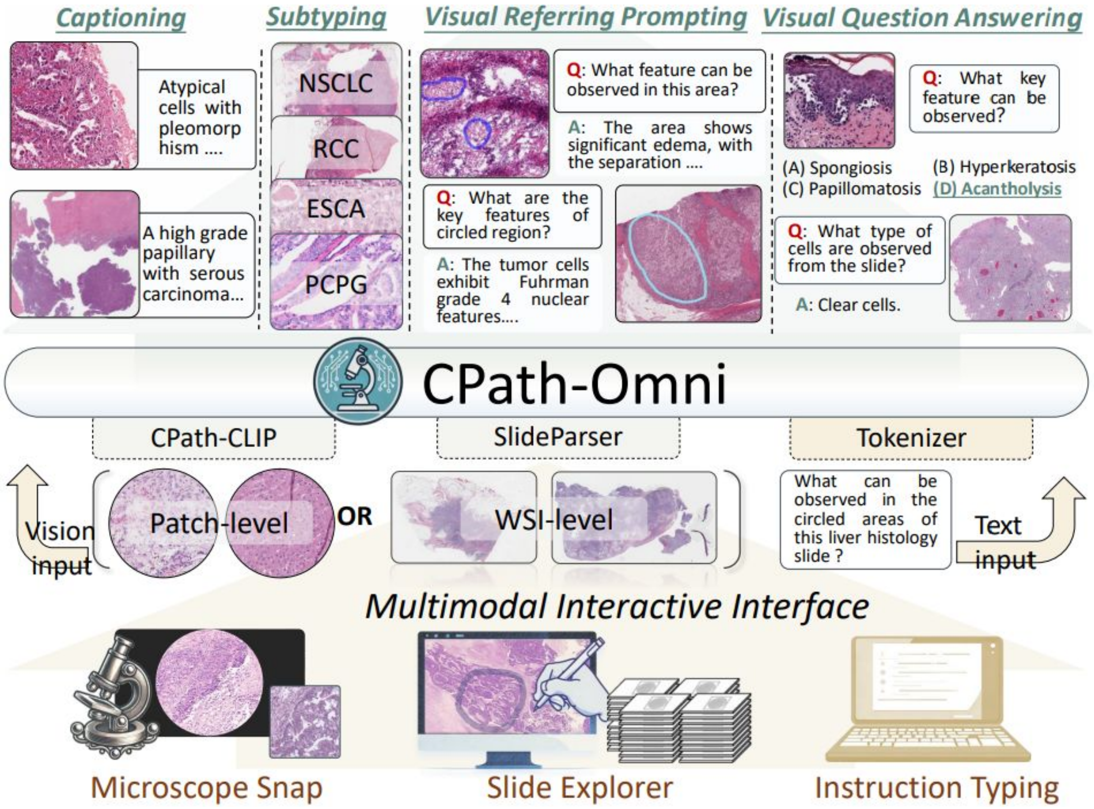
[VQA-Clinical]
#Q: Based on the histopathological features, what is the primary recommended treatment?
#A: Simple mastectomy followed by hormone therapy.

Chen, SlideChat: A Large Vision-Language Assistant for Whole-Slide Pathology Image Understanding, 2025

`SlideChat`

4k slide captions
176k question-answer pairs
Two-stage training:
  - Cross-domain alignment
  - Visual instruction learning

A VLM for Whole Slide Images - Multiple Use Cases



Atypical cells with pleomorphism ....

A high grade papillary with serous carcinoma...

Q: What feature can be observed in this area?

A: The area shows significant edema, with the separation ....

Q: What are the key features of circled region?

A: The tumor cells exhibit Fuhrman grade 4 nuclear features....

Q: What key feature can be observed?

(A) Spongiosis (B) Hyperkeratosis
(C) Papillomatosis (D) Acantholysis

Q: What type of cells are observed from the slide?

A: Clear cells.

What can be observed in the circled areas of this liver histology slide ?

Sun, CPath-Omni: A Unified Multimodal Foundation Model for Patch and Whole Slide Image Analysis in Computational Pathology, 2025

`CPath-Omni`

Virchow2 (DINOv2-based) + CLIP
Four-stage training:
- Vision-language alignment
- Patches: VQA, classification, captioning
- Whole slide pathology reports
- Slide and patch training

Task-Specific

Foundation Models

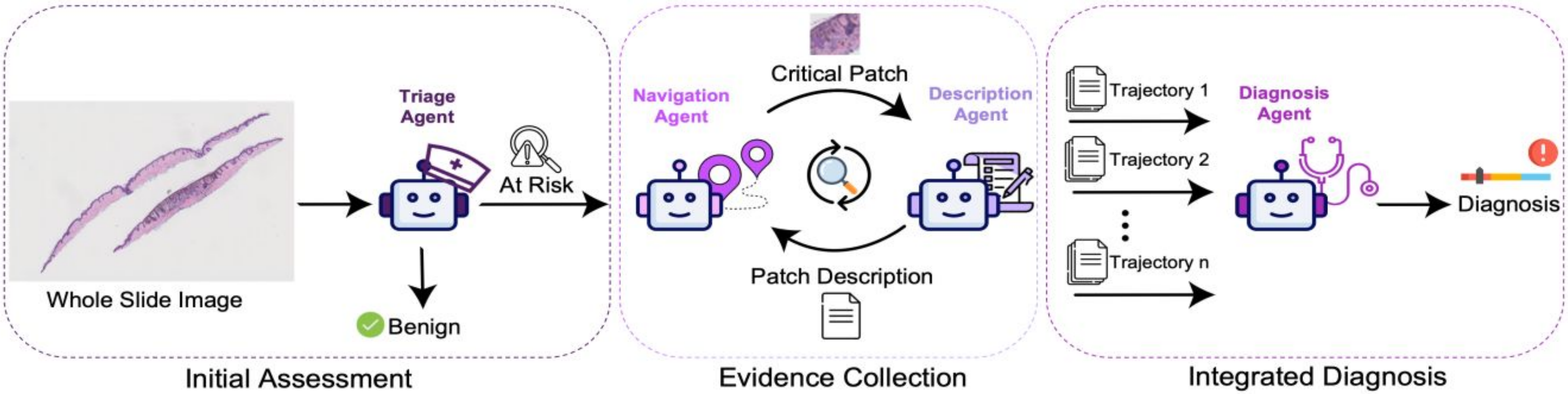
Multimodal

Agents

- Multi-agent systems
- Specialization
- Orchestration
- Mimic how human experts work

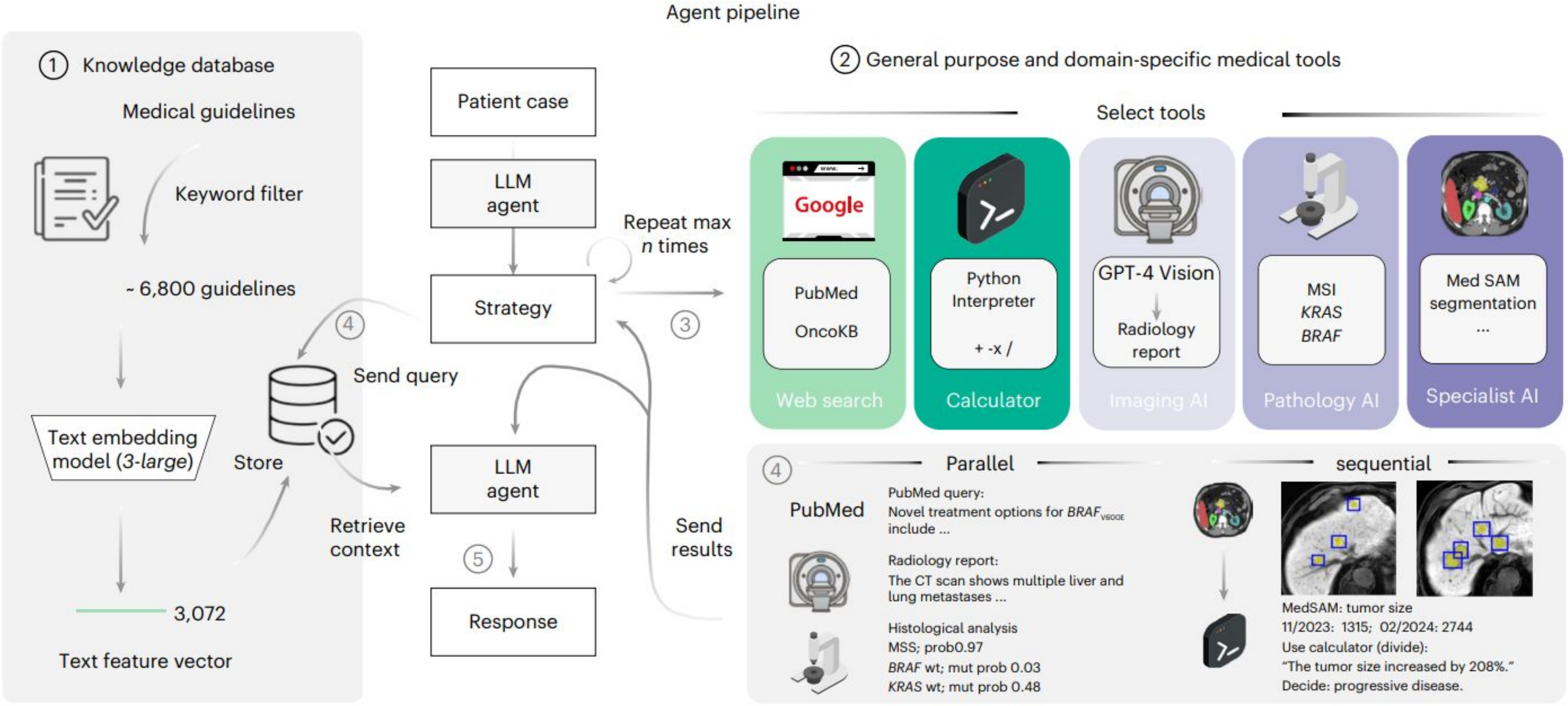
Agent for Histopathology Diagnosis

PathFinder



Ghezloo, PathFinder: A Multi-Modal Multi-Agent System for Medical Diagnostic Decision-Making Applied to Histopathology, 2025

Agent for Clinical Decision-Making in Oncology



Agent pipeline

① Knowledge database

Medical guidelines

Keyword filter

~ 6,800 guidelines

Text embedding model (*3-large*)

3,072

Text feature vector

Store

④ Send query

Retrieve context

Patient case

LLM agent

Strategy

LLM agent

⑤

Response

Repeat max *n* times

③

Send results

② General purpose and domain-specific medical tools

Select tools

PubMed
OncoKB
Web search

Python Interpreter
+ -x /
Calculator

GPT-4 Vision
Radiology report
Imaging AI

MSI
KRAS
BRAF
Pathology AI

Med SAM segmentation
...
Specialist AI

④ Parallel

PubMed

PubMed query:
Novel treatment options for $\mathit{BRAF}_{\mathrm{V600E}}$ include ...

Radiology report:
The CT scan shows multiple liver and lung metastases ...

Histological analysis
MSS; prob0.97
BRAF wt; mut prob 0.03
KRAS wt; mut prob 0.48

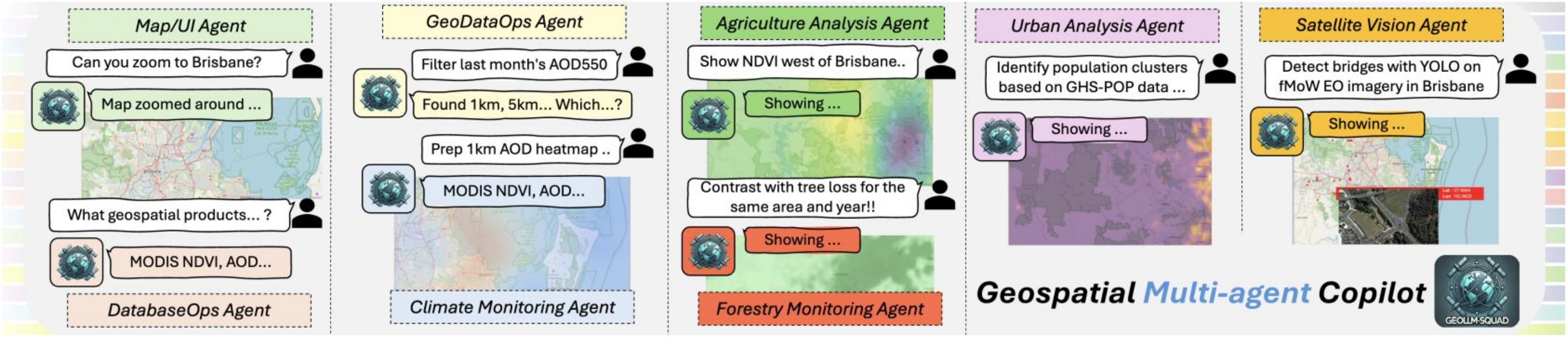
sequential

MedSAM: tumor size
11/2023: 1315; 02/2024: 2744
Use calculator (divide):
"The tumor size increased by 208%."
Decide: progressive disease.

Ferber, Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology, 2025

Geospatial Agent

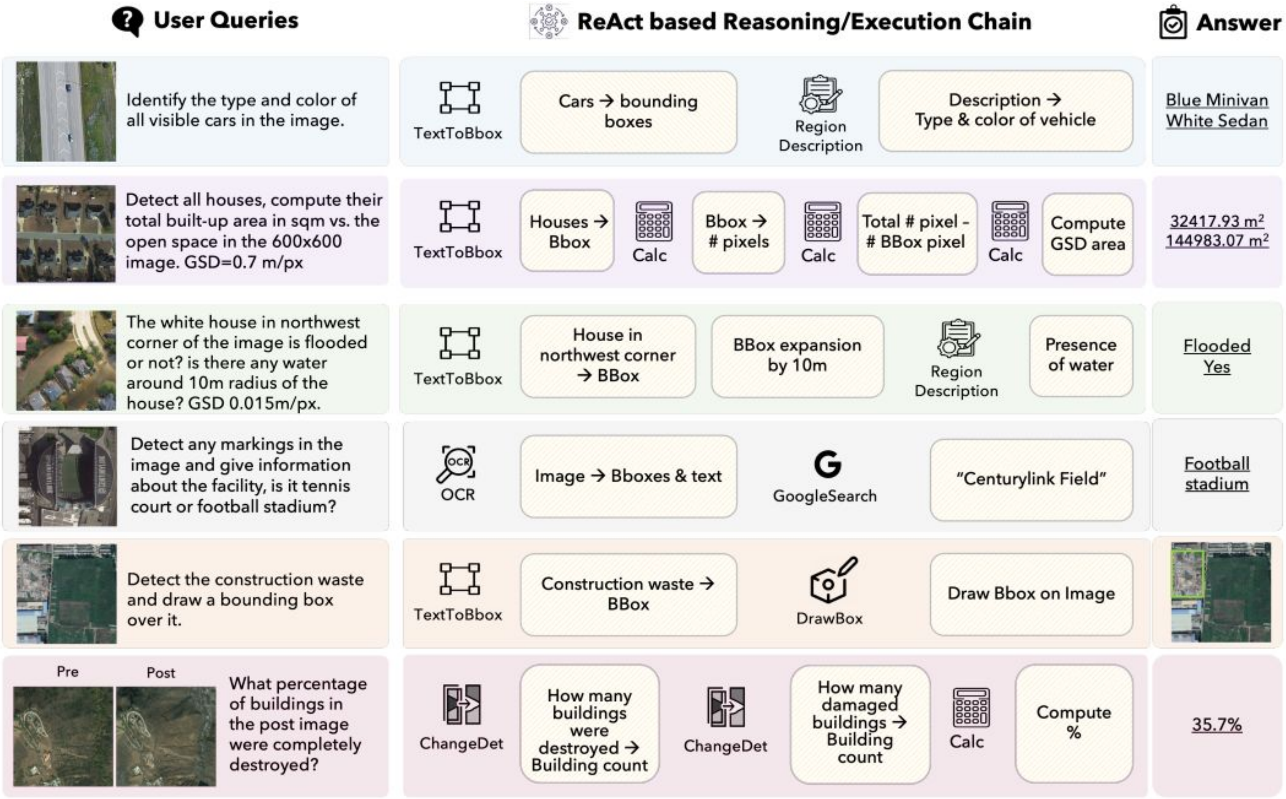
GeoLLM-Squad


Map/UI Agent
Can you zoom to Brisbane?
Map zoomed around ...
What geospatial products... ?
MODIS NDVI, AOD...
DatabaseOps Agent
GeoDataOps Agent
Filter last month's AOD550
Found 1km, 5km... Which...?
Prep 1km AOD heatmap ..
MODIS NDVI, AOD...
Climate Monitoring Agent
Agriculture Analysis Agent
Show NDVI west of Brisbane..
Showing ...
Contrast with tree loss for the same area and year!!
Showing ...
Forestry Monitoring Agent
Urban Analysis Agent
Identify population clusters based on GHS-POP data ...
Showing ...
Satellite Vision Agent
Detect bridges with YOLO on fMoW EO imagery in Brisbane
Showing ...
Geospatial Multi-agent Copilot
GEOLLM-SQUAD

Lee, Multi-Agent Geospatial Copilots for Remote Sensing Workflows, 2025

Benchmark for Geospatial Agents

ThinkGeo



User Queries	ReAct based Reasoning/Execution Chain	Answer
Identify the type and color of all visible cars in the image.	TextToBbox Cars → bounding boxes Region Description Description → Type & color of vehicle	Blue Minivan White Sedan
Detect all houses, compute their total built-up area in sqm vs. the open space in the 600x600 image. GSD=0.7 m/px	TextToBbox Houses → Bbox Calc Bbox → # pixels Calc Total # pixel - # BBox pixel Calc Compute GSD area	32417.93 m^2 144983.07 m^2
The white house in northwest corner of the image is flooded or not? is there any water around 10m radius of the house? GSD 0.015m/px.	TextToBbox House in northwest corner → BBox BBox expansion by 10m Region Description Presence of water	Flooded Yes
Detect any markings in the image and give information about the facility, is it tennis court or football stadium?	OCR Image → Bboxes & text GoogleSearch "Centurylink Field"	Football stadium
Detect the construction waste and draw a bounding box over it.	TextToBbox Construction waste → BBox DrawBox Draw Bbox on Image	
Pre Post What percentage of buildings in the post image were completely destroyed?	ChangeDet How many buildings were destroyed → Building count ChangeDet How many damaged buildings → Building count Calc Compute %	35.7%

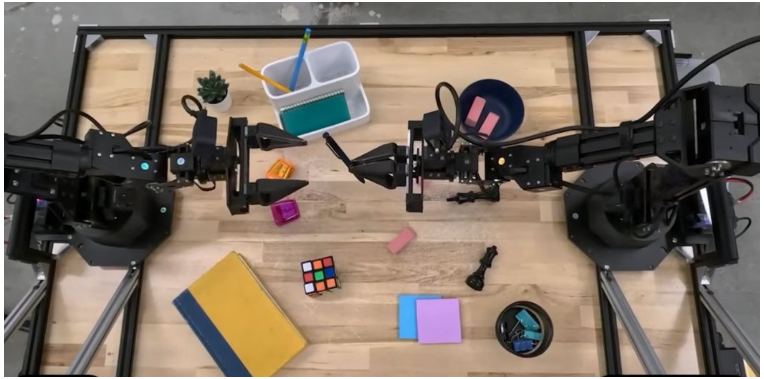
Shabbir, ThinkGeo: Evaluating Tool-Augmented Agents for Remote Sensing Tasks, 2025

Embodied Intelligence

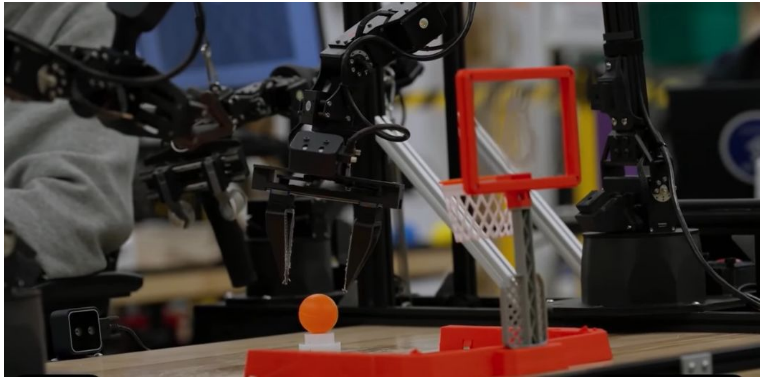
Generality, interactivity, and dexterity

“Can you pack me some trail mix?”

“Put the pen with the other pencils.”



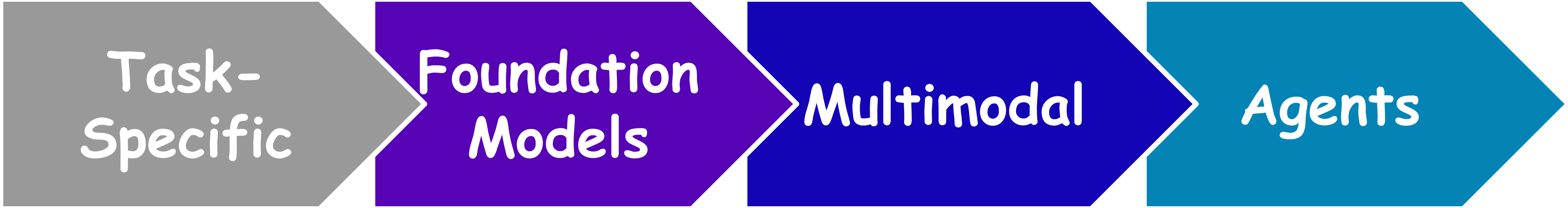
“Pick up the basketball and slam dunk it.”



“Put the objects in the right container.”

Keynote: Carolina Parada, Gemini Robotics, https://www.youtube.com/watch?v=o38k1k7f9Hk

Trends



Task-Specific

Foundation Models

Multimodal

Agents

- New domains
- Domain-specific adaptations
- New datasets
- New benchmarks

- Open weights and open data
- Grounding
- VLMs - language is the “glue”
- Multi-stage pretraining
- Larger inputs
- Benchmarks

- Multi-agent systems
- Specialization
- Orchestration
- Mimic how human experts work

Resources

Recordings of keynotes and workshops:

https://cvpr.thecvf.com/Conferences/2025/Videos

Voxel51’s “Best of CVPR” Series:

https://voxel51.com/events

Visual AI in Healthcare

June 27, 12-2 pm EDT

My talk: “Leveraging Foundation Models for Pathology: Progress and Pitfalls”

https://voxel51.com/events/visual-ai-in-healthcare-june-27-2025

Take the Next Step

Pixel Clarity Call

A free 30-minute call where we'll dive into your unique challenges and goals—whether you're seeking sharper models, deeper insights, or a new direction for your AI projects.

- Gain expert perspective on your current approach
- Discover high-leverage opportunities tailored to your mission

Book now: https://calendly.com/hdcouture/pixel-clarity-call

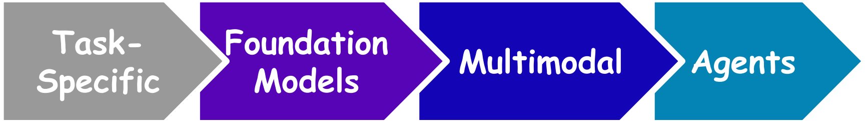
Q&A

Let’s flip the conversation:

1) If you attended CVPR or another recent conference, what is one thing you learned?

2) What imaging domain do you work with?

Which “phase” is your work in?



Is the current “phase” sufficient to solve a real world problem?

OR

What is blocking the path to the next “phase”?