CoMODAL: Cooperative Foundation Models For Object Detection Active Learning

Group 11 Zizheng Zhou

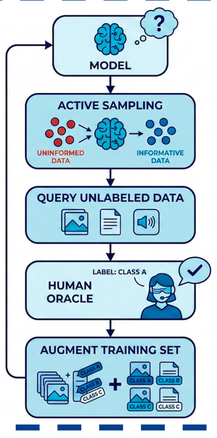
Introduction

What is Active Learning?

Traditional ML: Randomly labels massive amounts of data (High cost, lots of redundancy).

Active Learning (AL): Instead of random selection, the system automatically identifies the unlabeled data it is most "confused" about. It then effectively picks the most valuable samples for a human to label.

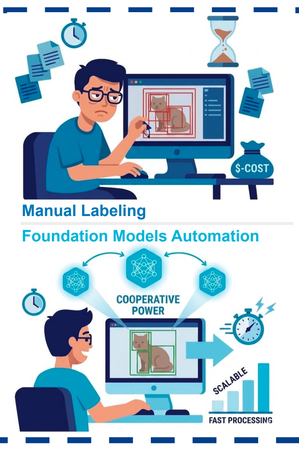
The Goal: Achieve maximum model performance with minimum human annotation effort.



Motivation

While Active Learning reduces the number of images to process, the manual effort of drawing precise bounding boxes remains a **slow**, **expensive,** and **non-scalable** bottleneck.

Motivated by the strong zero-shot capabilities of **Foundation Models**, we seek to leverage their cooperative power to automate this heavy lifting.



Key Contributions

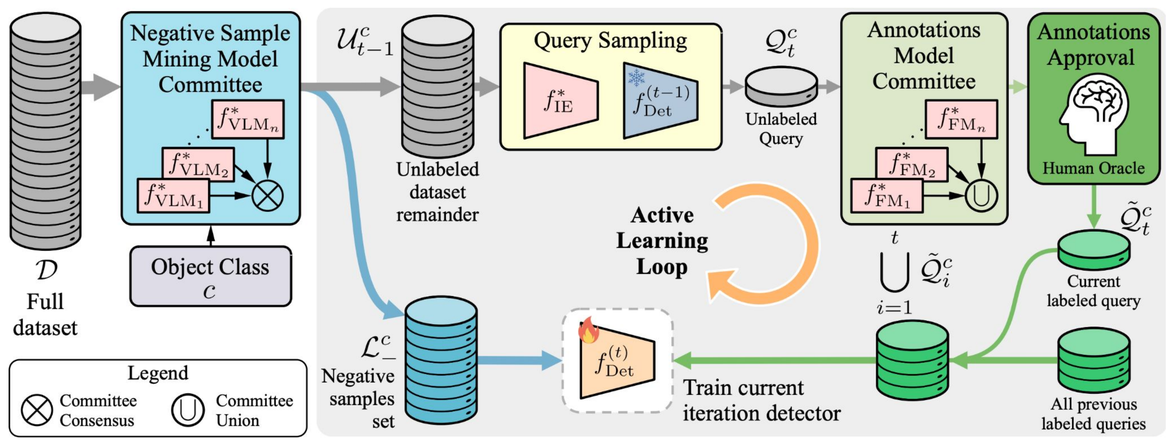
Negative Sample Mining: Utilizing an ensemble of VLMs to reach a consensus on negative samples, removing them from the human labeling queue entirely.

Dual-Objective Query Sampling: By leveraging DINOv3 patch embeddings, we sample diverse high-uncertainty predictions (via k-Means) and simultaneously identify high-confidence false positives (via cosine similarity).

Accelerated Annotation: Instead of drawing boxes manually, a Human Oracle simply provides binary (approve/reject) feedback on bounding box candidates proposed by foundational models.

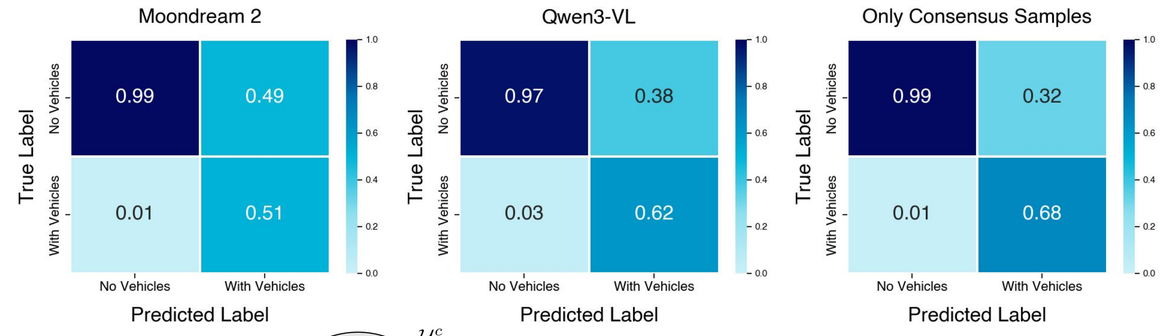
Method

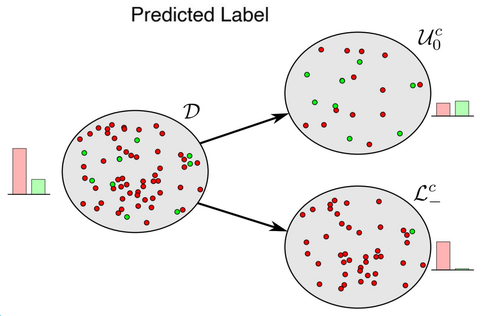
Overview



Filtering: A VLM committee automatically extracts negative samples, bypassing human labeling.
Sampling: The system leverages Foundation Model features to strategically sample the most diverse and informative queries.
Annotation: Models propose bounding boxes; humans simply approve or reject them instead of drawing.
Model Update: The detector is fine-tuned on both mined negatives and newly approved queries.

Negative Sample Mining

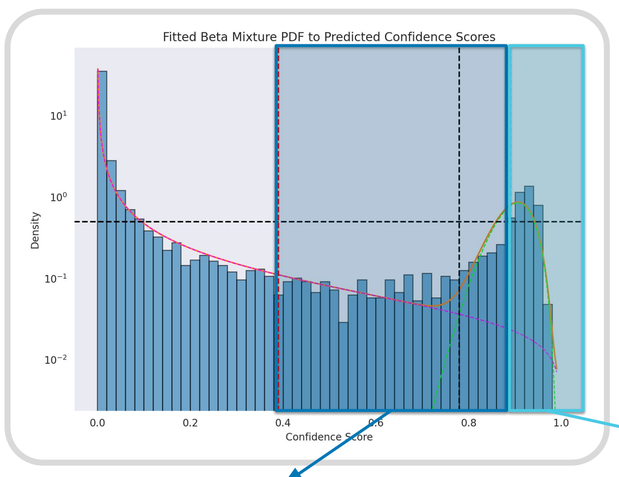




Zero-Shot Filtering: We utilize an ensemble of diverse Vision-Language Models to independently classify whether an image contains the target object.

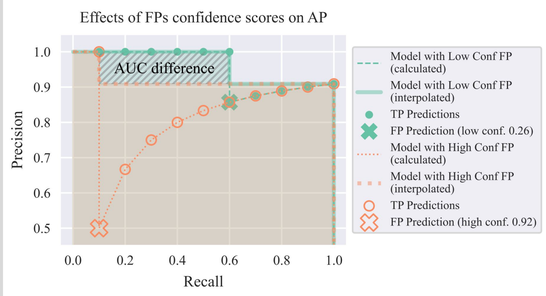
The Consensus Strategy: An image is only added to the negative sample set if all models in the committee agree it lacks the target object. As shown in the confusion matrices, while individual models may struggle, the "Consensus" approach achieves near-perfect precision (e.g., 0.99 for True Negatives).

Query Sampling



Why high-confidence FPs?

They are bad for AP!



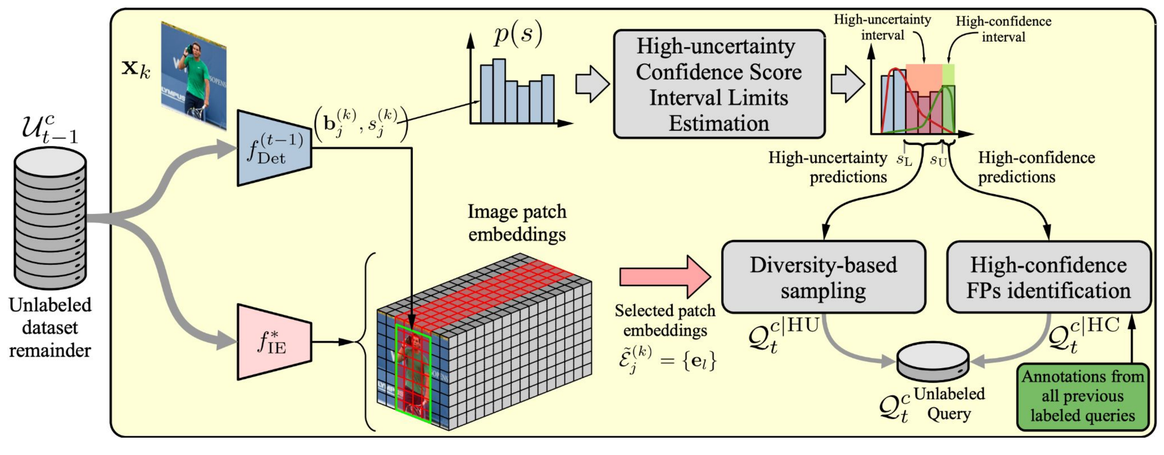
High Uncertainty Predictions
Goal: sample a diverse and representative subset

- Detector predictions on unlabeled data
- DINOv3 patch embeddings of the predictions
- Apply k-Means to the selected embeddings
- Get the representative embeddings closest to each cluster centroid

High Confidence Predictions
Goal: identify potential FPs

- Detector predictions on unlabeled data → DINOv3 patch embedding of preds
- Previous query annotations → DINOv3 patch embeddings of anns

Find the prediction patches as far as possible from the annotation patches in cosine distance



Results