Benchmarking Vision-Language Models for Object Detection in Satellite Imagery

Devansh Lodha
23110091@iitgn.ac.in

Nupoor Assudani
23110224@iitgn.ac.in

Nipun Batra
nipun.batra@iitgn.ac.in

Sustainability Lab   IIT Gandhinagar

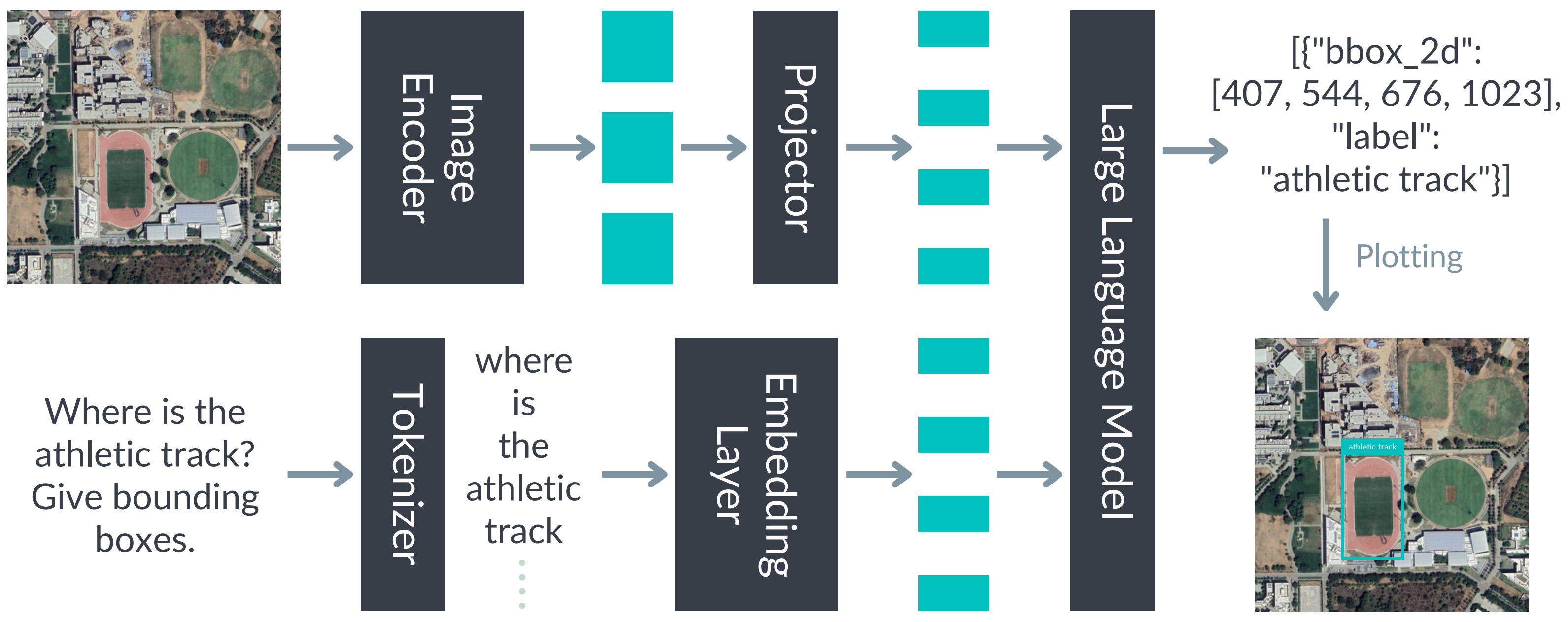
Motivation

Problem: Air pollution is a major issue in India (1.7M deaths/year)[1], with brick kilns contributing significantly (14%).[2]

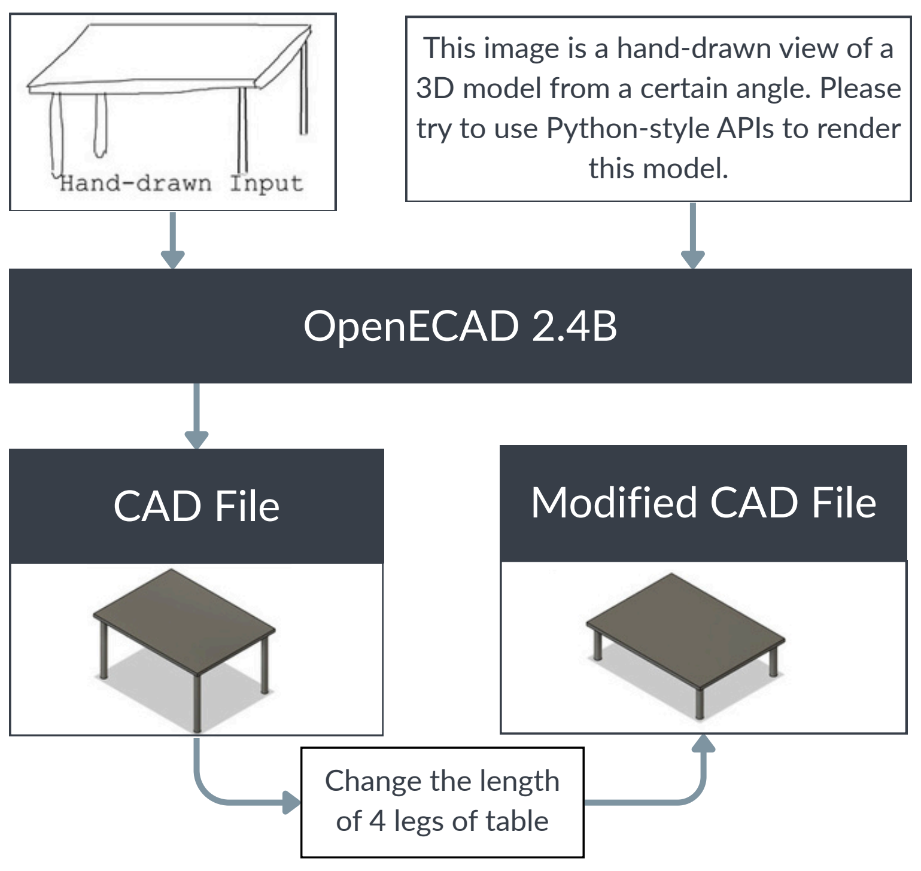
Challenge: Detecting these kilns is difficult as it is an unorganized sector. Manual monitoring is unscalable.

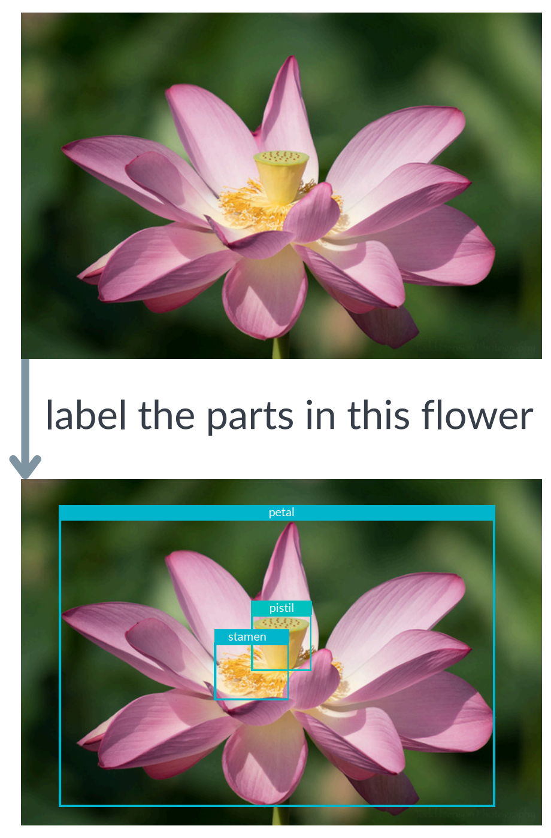
Technical Gap: Traditional object detection models (like YOLO) require large amounts of labelled domain-specific data, which is often scarce for satellite imagery tasks. They also struggle with domain adaptation (perform poorly when tested on regions different from their training data).

Vision-Language Models



VLMs combine visual and text understanding


Hand-drawn Input
This image is a hand-drawn view of a 3D model from a certain angle. Please try to use Python-style APIs to render this model.
OpenECAD 2.4B
CAD File
Modified CAD File
Change the length of 4 legs of table



Models Tested

PaliGemma 3B

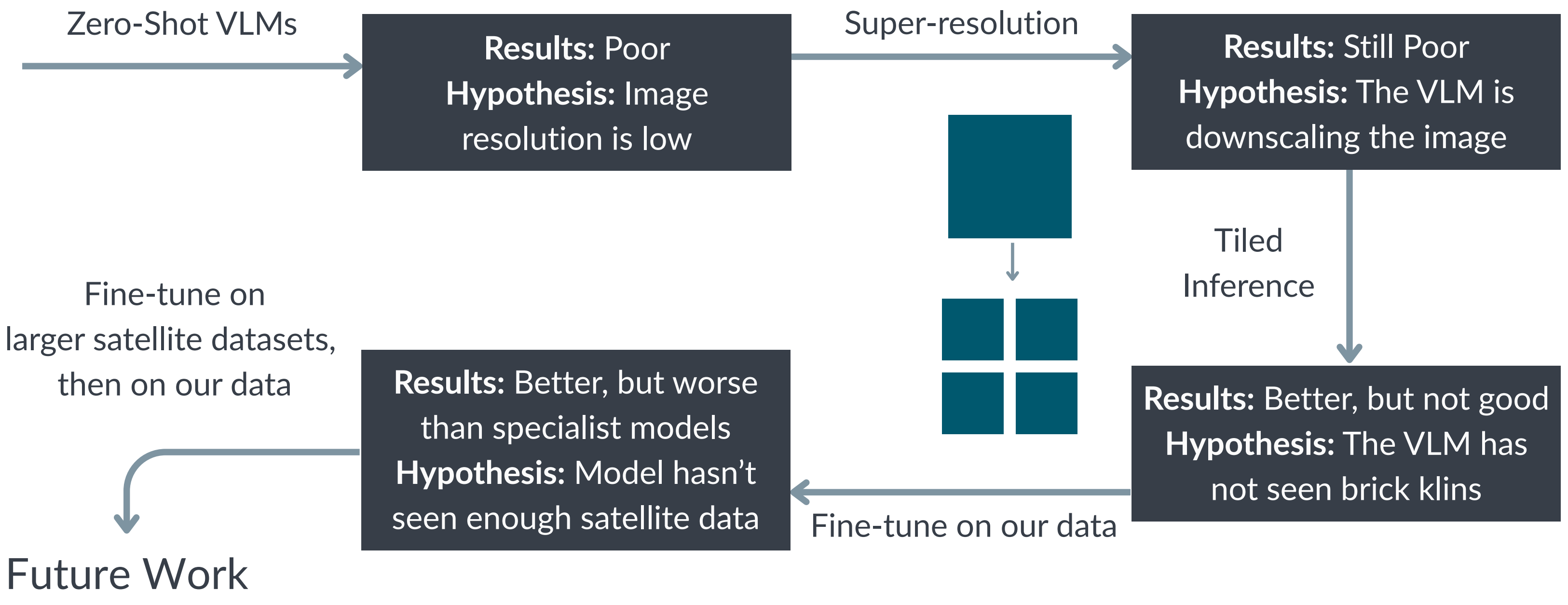
Florence2 large: 770M

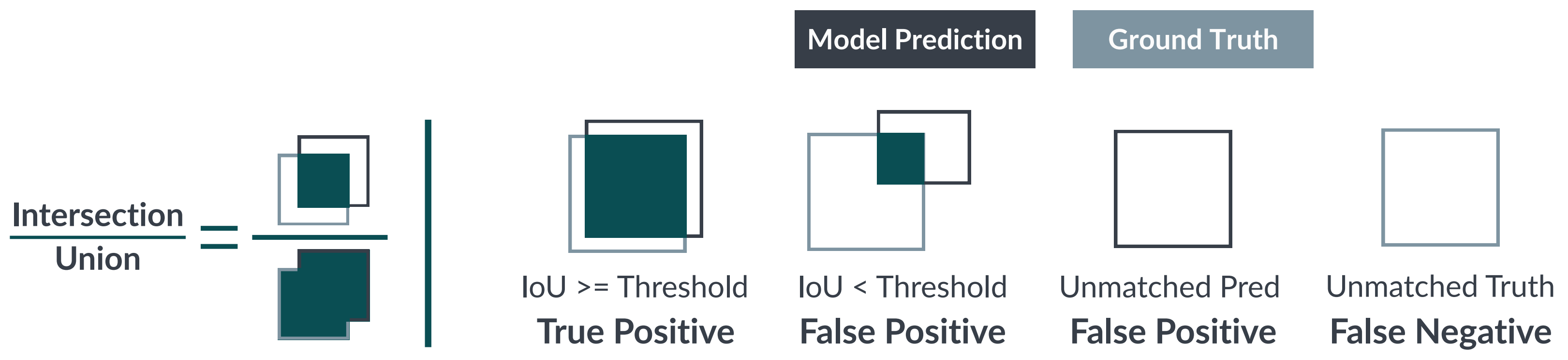
Deepseek-VL2 27B

Qwen2.5-VL 72B

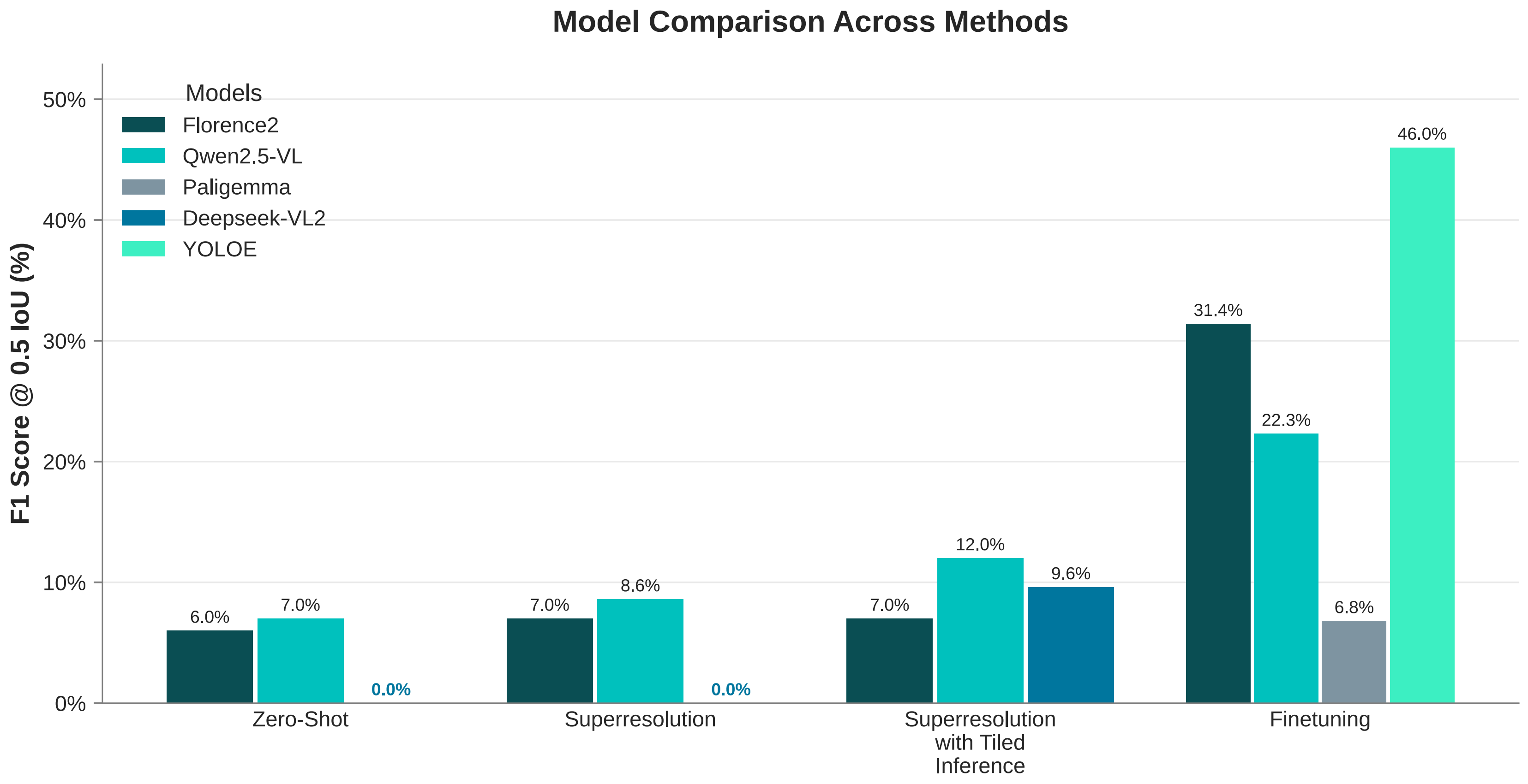
YOLOE 43.6M

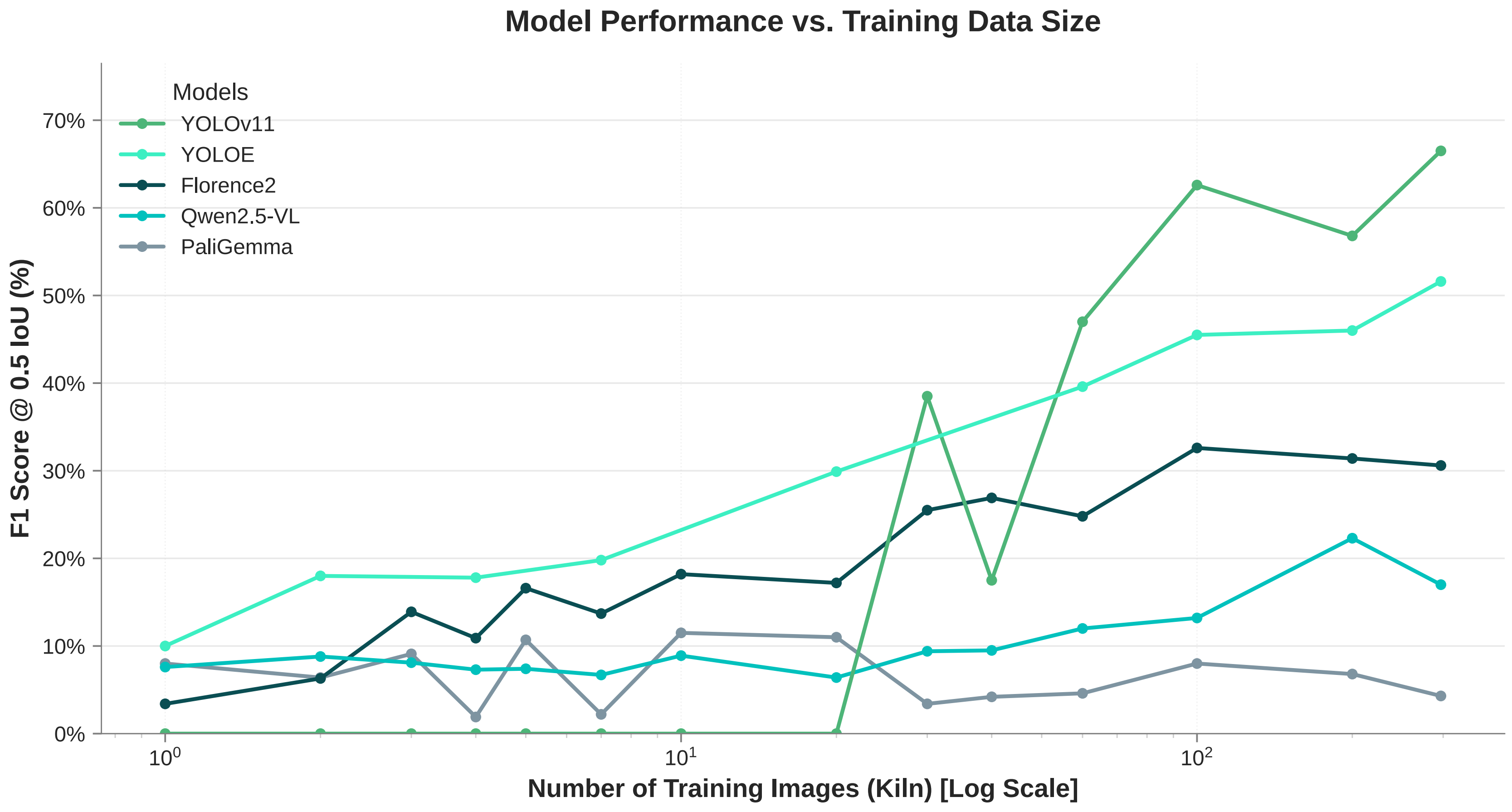
Hypothesis Driven Approach


Zero-Shot VLMs
Results: Poor
Hypothesis: Image resolution is low
Super-resolution
Results: Still Poor
Hypothesis: The VLM is downscaling the image
Tiled Inference
Results: Better, but not good
Hypothesis: The VLM has not seen brick klins
Fine-tune on our data
Results: Better, but worse than specialist models
Hypothesis: Model hasn't seen enough satellite data
Fine-tune on larger satellite datasets, then on our data
Future Work

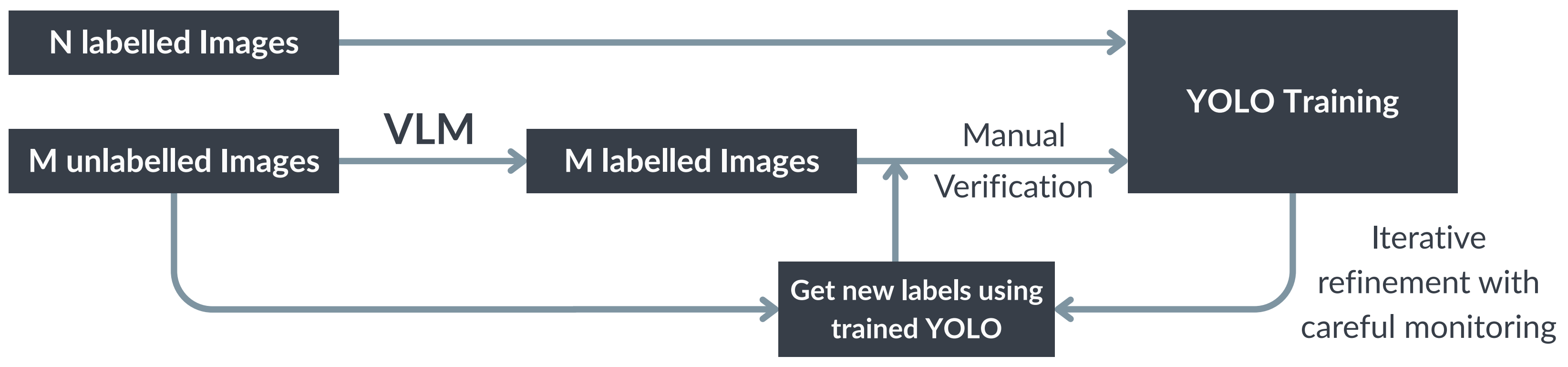


Key Results





Conclusions and Future Directions



- VLMs show potential but require adaptation for satellite object detection.
- Image resolution, model architecture, and model vocabulary affect VLM performance
- Fine-tuning VLMs improves results significantly

Acknowledgments

We gratefully acknowledge the guidance and support of Prof. Nipun Batra, Rishabh Mondal, and Zeel Patel. Special thanks to Shataxi Dubey for her invaluable contributions and collaboration on this project.

References

[1] A. Pandey et al., "Health and economic impact of air pollution in the states of India: the Global Burden of Disease Study 2019," The Lancet Planetary Health, vol. 5, no. 1, Dec. 2020, doi: https://doi.org/10.1016/S2542-5196(20)30298-9.
[2] "Brick kilns embrace zigzag design to cut pollution and boost efficiency," Mongabay India, Aug. 2024. [Online]. Available: https://india.mongabay.com/2024/08/brick-kilns-embrace-zigzag-design-to-cut-pollution-and-boost-efficiency/. Accessed: Apr. 17, 2025.
[3] Planet Labs PBC, "Planet Application Program Interface: In Space for Life on Earth." [Online]. Available: https://api.planet.com. Accessed: 2024.
[4] A. Steiner et al., "PaliGemma 2: A Family of Versatile VLMs for Transfer," arXiv preprint arXiv:2412.03555, Dec. 2024. [Online]. Available: https://arxiv.org/abs/2412.03555
[5] B. Xiao et al., "Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2024, pp. 4818-4829.
[6] X. Chen et al., "DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding," arXiv preprint arXiv:2412.10302, Dec. 2024. [Online]. Available: https://arxiv.org/abs/2412.10302
[7] S. Bai et al., "Qwen2.5-VL Technical Report," arXiv preprint arXiv:2502.13923, Feb. 2025. [Online]. Available: https://arxiv.org/abs/2502.13923
[8] K. Kuckreja et al., "GeoChat: Grounded Large Vision-Language Model for Remote Sensing," arXiv preprint arXiv:2311.15826, Nov. 2023. [Online]. Available: https://arxiv.org/abs/2311.15826
[9] G. Jocher and J. Qiu, "Ultralytics YOLO11," ver. 11.0.0, 2024. [Online]. Available: https://github.com/ultralytics/ultralytics. Accessed: Apr. 17, 2025.
[10] A. Wang et al., "YOLOE: Real-Time Seeing Anything," arXiv preprint arXiv:2503.07465, Mar. 2025. [Online]. Available: https://arxiv.org/abs/2503.07465
[11] Z. Yuan, J. Shi, and Y. Huang, "OpenECAD: An efficient visual language model for editable 3D-CAD design," Comput. Graph., vol. 124, Art. no. 104048, 2024, doi: 10.1016/j.cag.2024.104048.