Analyzing and Improving the Generalization Capabilities of Grounding DINO on Unseen Domains

Final Project Proposal

Group Members:

黃皓群 (B123245009)
周霖 (B123245028)
陳圖億 (B123040002)

Group Number: 8

Department of Computer Science and Engineering
National Sun Yat-sen University

November 2025

1. Problem Definition and Motivation

The field of object detection is experiencing a major shift from closed-set recognition to open-set scenarios, where models aim to detect arbitrary objects described through human language rather than a fixed set of predefined categories. A notable milestone in this shift is Grounding DINO, an architecture that combines the Transformer-based detector DINO with grounded pre-training to achieve state-of-the-art results. Although the model has shown impressive generalization on standard benchmarks such as COCO and ODinW, its real-world applicability and adaptability to specialized domains have not yet been thoroughly tested. This project aims to conduct a evaluation of Grounding DINO's adaptation capabilities, focusing on its Zero-Shot performance boundaries and the effectiveness of Few-Shot fine-tuning in a novel task setting.

Current Landscape & Limitations:

Traditional object detection models, such as **DETR** or the **YOLO** family, follow a *closed-set* framework. They can only recognize object categories that were defined during training, making them unable to detect new or unseen objects without significant retraining. Although recent progress in *open-set object detection* has introduced models that can identify arbitrary objects using natural language descriptions, their reliability and adaptability in specialized domains remain major challenges for current research.

The SOTA solution:

Grounding DINO, introduced by Liu et al. (2024), represents the current state-of-the-art in open-set object detection. By integrating the Transformer-based detector DINO with grounded pre-training, it achieves remarkable performance —including 52.5 AP on the COCO zero-shot benchmark and a new record on ODinW. Its architecture tightly combines vision and language through a feature enhancer and language-guided query selection, enabling the model to detect virtually any object described by text.

Research gap (Motivation):

Despite its strong results, the original paper also acknowledges several limitations. For example, in fine-grained tasks such as Referring Expression Comprehension (REC), the model performs suboptimally without targeted fine-tuning. Moreover, while it achieves strong results on common benchmarks like **COCO**, **LVIS**, and **ODinW**, its **zero-shot** transferability to specialized domains with large domain gaps —such as industrial defect detection, medical imaging, or remote sensing —remains largely unexplored. We hypothesize that even with extensive pre-training on datasets like **O365** and **GoldG**, the model may struggle to generalize effectively to these distinct visual modalities without additional adaptation.

Project goal:

This project aims to conduct a critical evaluation of Grounding DINO' s domain adaptation boundaries. Specifically, we seek to examine how well its zero-shot capabilities generalize to a novel and unseen task, and to quantify the effectiveness of few-shot fine-tuning as a potential bridge for improving domain transfer.

2. Proposed Project and Approach

We propose a systematic study to evaluate the transferability of the Grounding DINO model on a novel dataset. Rather than modifying the model architecture, our approach focuses on experimental evaluation and transfer learning strategies. The project is structured into two main phases:

Phase 1: Zero-Shot Evaluation Boundary Analysis (Baseline)

We use the official pre-trained checkpoints to test Grounding DINO on a new domain with domain-specific text prompts. This phase examines whether open-set detectors fail to handle fine-grained data without fine-tuning, as suggested in the paper. We expect a noticeable performance drop caused by semantic and visual domain gaps and will qualitatively analyze failure cases—such as false positives or hallucinations—to determine whether errors arise from ambiguous prompts or visual misalignment.

Few-Shot Fine-Tuning Adaptation (Solution)

To overcome the limits of zero-shot performance, we apply a few-shot fine-tuning strategy. Using the model's language-aware architecture, we fine-tune it on small subsets of the target data (e.g., 1-shot, 10-shot, 50-shot). During this phase, we will also explore prompt engineering to refine text inputs and improve alignment between visual and linguistic features. We evaluate performance growth to verify that even limited fine-tuning helps the model adapt to the new domain. We expect to show that a small amount of domain-specific data, combined with optimized prompts and Grounding DINO's pre-trained knowledge, can achieve strong detection performance—demonstrating a practical recipe for adapting SOTA models to specialized domains.

3. Key References

[1] **Main Paper:** Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., & Zhang, L. (2024). Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *Proceedings of the European Conference on Computer Vision (ECCV)*.

[2] **Official Codebase:** IDEA-Research. (2023). Grounding DINO GitHub Repository. Available at: `https://github.com/IDEA-Research/GroundingDINO`

[3] **Target Dataset (Industrial):** Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2019). MVTec AD —A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Available at: `https://www.mvtec.com/company/research/datasets/mvtec-ad`

[4] **Target Dataset (Medical):** Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. *CVPR*. Available at: `https://www.kaggle.com/datasets/nih-chest-xrays/data`

[5] **Target Dataset (Pathology):** Bejnordi, B. E., et al. (2017). Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*. (Camelyon16 Challenge). Available at: `https://camelyon16.grand-challenge.org/Data/`