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

Final Project Progress Update

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 Summary of Current Efforts

In this stage of the project, we set up an initial evaluation to check how well Grounding DINO can generalize to new domains in a Zero-shot setting. We ran baseline tests on three different types of data: Industrial Defect Detection (MVTec AD), Medical Imaging (Chest X-ray), and Digital Pathology (Camelyon16). These early results helped us understand where the model works well and where it struggles, especially when the task requires recognizing very detailed texture patterns.

1.1 Project Status Recap

Our main goal is to check whether the open-set detector Grounding DINO can handle specialized domains without any extra training. In addition, for tasks where the zero-shot results are weak, we also plan to improve the performance through few-shot tuning as a follow-up step.

- **Industrial Domain:** We tested all 15 categories in the MVTec AD dataset. The results show a clear gap between object-focused categories (such as *Bottle*) and texture-focused categories (such as *Carpet*).
- **Medical & Pathological Domains:** We further evaluated the model on Chest X-ray images (NIH) and small pathology patches (PCam). These tests showed that the model has difficulty telling normal structures apart from real abnormalities, and often reacts too strongly to harmless patterns in a Zero-shot scenario.

1.2 Technical Implementation

We built a modular inference pipeline based on Grounding DINO with a Swin-T backbone. The system can process different datasets in a consistent way, handling differences in image size and format, such as high-resolution industrial images versus small pathology patches, so the comparison remains fair.

Prompt Design Strategy

Because Grounding DINO relies heavily on text prompts, we created simple prompts for each category to guide the detector. For this first round of experiments, we used short descriptive phrases that combine the category name with common defect terms.

Table 1 shows the prompts we used. For industrial data, we used direct descriptions of the object. For medical data, especially X-rays, we used a "negative listing" style (listing disease terms like "pneumonia" instead of organs) to reduce false alarms on normal regions.

Table 1: Baseline Text Prompts for Zero-shot Evaluation

Domain	Category Example	Text Prompt Used (Baseline)
Industrial	Bottle	`bottle . broken bottle . glass defect .`
(MVTec AD)	Carpet	`carpet . hole . cut . color stain .`
Medical	Chest X-ray	`pneumonia . lung opacity . effusion .`
Pathology	Cell Patch	`cancer . tumor . metastasis .`

These prompts are only our starting point. In the next stage, we plan to do more focused **prompt engineering** to see how different levels of detail (from broad terms to very specific descriptions) change the model's performance.

Evaluation Protocols

Because Zero-shot models do not produce pixel-level segmentation masks for these datasets, we treated the task as an **image-level anomaly detection problem**. We used the highest-confidence bounding box in each image as the anomaly score. We then evaluated the model using:

- **Image-level AP & AUC:** These measure how well the model ranks defective images above normal ones.
- **Maximum Confidence Score:** We compare the highest confidence on normal vs. abnormal images to check how often the model “hallucinates“ defects on healthy images.

2 Cross-Domain Analysis

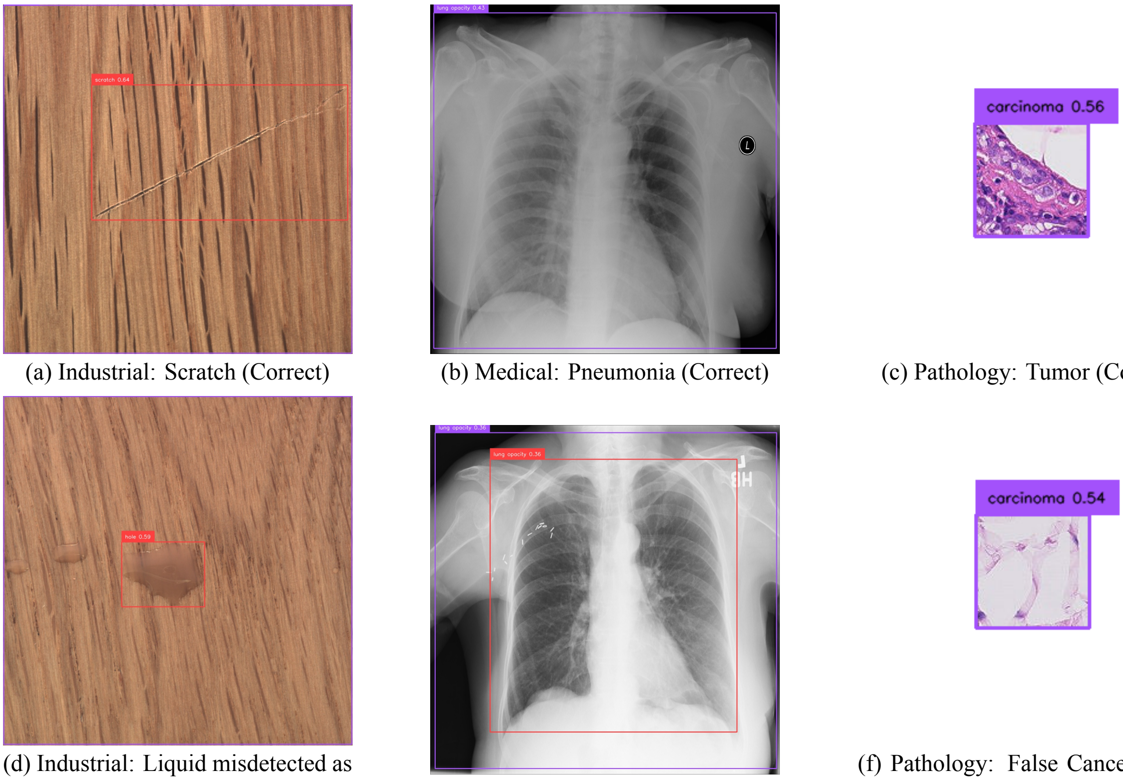
Figure 1 shows examples from three domains. The goal is to highlight where Grounding DINO works well and where it fails. Successful cases show clear localization of defects or medical findings. Failure cases come from situations where the model confuses harmless patterns as anomalies, or reacts to visual noise instead of real defects.

Table 2 further confirms several consistent patterns across domains. Object-centric categories such as *leather*, *tile*, and *hazelnut* show high AP scores and a clear confidence gap between defective and normal samples, indicating that Grounding DINO can correctly react to true anomalies while remaining stable on normal images. In contrast, texture-heavy categories such as *carpet*, *grid*, and *wood* exhibit small or even reversed confidence gaps, revealing frequent hallucinations where normal patterns trigger stronger responses than actual defects. A similar trend appears in medical and pathology data, where the maximum confidence on normal images is comparable to or higher than that on abnormal samples, suggesting that the model lacks domain-specific cues and cannot reliably distinguish meaningful abnormalities. Overall, the table highlights that Grounding DINO performs well when the defect has a strong object-level structure, but becomes unstable in fine-texture or high-ambiguity domains.

2.1 Future Work: Few-Shot Fine-tuning

The cross-domain results reveal clear strengths and weaknesses in Grounding DINO’s Zero-shot behavior. These observations naturally guide the next steps of our study. Rather than relying solely on prompt design, we will explore **Few-Shot Fine-tuning** to inject minimal domain knowledge and reduce the model’ s tendency to fire high-confidence detections on normal samples. This lightweight adaptation is expected to improve stability on texture-heavy industrial categories and medical/pathology images, where Zero-shot performance remains unreliable.

In addition, the current evaluation is based on sampled subsets from each dataset. Our next stage will perform a more complete, full-dataset analysis to obtain a more reliable and statistically meaningful view of model behavior. We also plan to compare Grounding DINO with additional baseline methods, enabling a clearer understanding of how its Zero-shot and Few-shot performance stands relative to other cross-domain anomaly detection approaches.



(a) Industrial: Scratch (Correct)

(b) Medical: Pneumonia (Correct)

(c) Pathology: Tumor (Correct)

(d) Industrial: Liquid misdetected as Hole

(e) Medical: Over-sensitivity

(f) Pathology: False Cancer Detection

Figure 1: Qualitative comparison across Industrial, Medical, and Pathological datasets. The top row shows correct predictions, while the bottom row shows failure cases.

Table 2: Results across Industrial and Medical datasets.

Dataset / Category	**Image AP**	**Image AUC**	**Max Conf (Anom.)**	**Max Conf (Norm.)**
Industrial Domain				
bottle	0.8981	0.6937	0.6886	0.6457
cable	0.5864	0.3902	0.4933	0.4798
capsule	0.8427	0.5104	0.5650	0.5590
carpet	0.6020	0.1220	0.5084	0.5104
grid	0.7322	0.4678	0.5537	0.5201
hazelnut	0.8535	0.7021	0.8877	0.8577
leather	0.9748	0.9171	0.5770	0.4336
metal_nut	0.6846	0.1935	0.7817	0.7883
pill	0.8510	0.5596	0.6126	0.5216
screw	0.7618	0.5488	0.5972	0.5877
tile	0.9169	0.7615	0.6376	0.5561
toothbrush	0.8565	0.6278	0.7350	0.5978
transistor	0.4780	0.4788	0.6566	0.6303
wood	0.6648	0.2649	0.6985	0.7111
zipper	0.7464	0.3884	0.6165	0.5879
Medical Domain				
ChestXray				
lung	0.6343	0.5950	0.4347	0.4218
Pathology				
cell	0.4305	0.3400	0.5603	0.5773