A Comparative Study of Transformer-based Object Detection Models on Stenosis Detection

Supervisors: Sheethal Bhat, Prof. Dr.-Ing. habil. Andreas Maier

Debadrita Mukherjee
Medical Engineering - Medical Image and Data Processing
Pattern Recognition Lab, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany

I. Abstract

Transformers used for object detection have been shown to be promising within the context of natural images, but their application in medical imaging is considerably less investigated. A comparative evaluation of three transformer detectors applied to the ARCADE coronary stenosis dataset is discussed in this paper, which are Conditional DETR, Grounding DINO with the Swin-L backbone, and Grounding DINO with a PVT backbone variant. I have trained all the models using the MMDetection framework and tested their capacity to detect stenotic plaques in X-ray angiography. Conditional DETR failed to learn the task in my research, while the Grounding DINO (Swin-L) setup outperformed, outperforming the Grounding DINO (PVT) arrangement. This work introduces a comprehensive evaluation of these findings against existing literature on transformer detectors of medical images (e.g., VinDR-CXR and SIIM-ACR), where there is a tendency for pretrained transformers to perform well with broad object recognition but not with localized and subtle medical pathologies in medical images.i would hypothesize that the extreme variability and subtle pattern of lesions (like arterial constrictions) limit off-the-shelf performance of such models. My conclusion indicates the need for further domain-specific adjustments to make full use of the strengths of transformers to detect medical object.

II. Introduction

Computer-assisted detection of focal pathologies in medical images is a fundamental capability of computer-aided detection (CAD) systems. Detection of coronary artery stenosis via X-ray angiography, for example, requires detection of extremely small narrowing of arteries (stenotic plaques), which may be extremely difficult to detect. Traditional CNN-based detectors (e.g., one-stage YOLO family or two-stage R-CNNs) have achieved state-of-the-art on medical object detection, but also highly depend on heuristics (anchors, pyramids, NMS) and may not generalize well to the complexity and variability of medical imaging data[1]. Recent transformer-based detectors, e.g., DETR and variants, introduced an end-to-end set prediction paradigm without hand-designed proposals. These models can model global context with self-attention mechanisms and have shown competitive accuracy on natural image benchmarks. It is, however, unknown how these strong models perform when directly applied to medical imaging, which contains small, low-contrast abnormalities with large variability in scale.

There has been some work on investigating transformers for medical object detection. Cheng et al. tell us that baseline DETR models scored around zero mAP when evaluated for chest X-ray lesion detection and did worse than CNN detectors[2]. The original DETR particularly struggles with small objects and slow convergence, prompting follow-up works that enhance training and multi-scale feature learning [2] Conditional DETR was proposed to address some of these issues by incorporating location-specific questions with cross-attention under conditions enable faster convergence and better small object detection [3]. Likewise, DINO (DETR with Enhanced denoising) and comparative transformer detectors incorporate multi-scale feature maps and optimized training to improve performance on smaller targets [3]. Conversely, Grounding DINO is a novel open-vocabulary detector which fuses a DETR framework with the capacity of being text grounded and has been fine-tuned on huge image-text datasets [7][8]. While Grounding DINO demonstrates very good performance when used in natural images, recent research conducted by Gholipour Picha et al. demonstrated that mere pre-training without additional fine-tuning is not sufficient to grasp fine medical findings. For instance, a prompt like "cardiomegaly with pulmonary congestion" led the pre-trained model to highlight the entire chest X-ray instead of the right area[9]. This refers to the typical problem with medical anomalies: they are generally manifested as Distinctive and subtle characteristics which differ greatly from the objects found in typical model training data.

In the present work, a comprehensive comparison is done of three transformer-based detection models with a specific medical task: coronary stenosis detection on the ARCADE Stenosis dataset [1]. The ARCADE (Automatic Region-based Coronary Artery Disease diagnostics) dataset is a public benchmark that was introduced at MICCAI 2023 and consists of expert-annotated coronary angiography images for vessel segmentation and stenosis localization [1]. It provides a real-world environment to examine whether newer detectors such as Conditional DETR and Grounding DINO can properly detect small arterial constrictions in complicated angiographic situations. I have presented results for all of the models and compare them with previous transformer-based approaches applied to medical imaging (e.g., for chest X-rays and other imaging modalities). My findings show that pretrained transformers, even with fine-tuning, struggle to detect small-sized, subtle lesions unless additional architectural changes or training enhancements are undertaken. I have discussed the contributing factors to the performance gap, pointing to the importance of multi-scale feature hierarchies and domain-specialized feature representations, as well as indicating potential directions towards enhancing transformer detectors in medical settings.

III. Related Work

The introduction of DETR [2] has sparked significant interest in leveraging transformers for medical object detection—a space traditionally led by CNNs. Early studies quickly revealed that, in their standard form, DETR models didn’t perform well on medical tasks, especially when it came to detecting small lesions. For instance, Cheng et al. evaluated DETR on chest X-ray pathology localization and reported extremely low detection rates (with mAP near zero), attributing this to DETR’s difficulty with small object detection and limited training data [2]. While DETR’s one-to-one set matching mechanism is indeed innovative, it’s hampered by coarse feature maps resulting from heavy down sampling in the backbone, which leads to the loss of subtle pathology signals [6]. This limitation is particularly problematic in medical imaging, where abnormalities often appear as subtle intensity changes that are easy to miss in a broader context [6].

To address these issues, researchers have proposed several DETR variants specifically tailored for medical applications. Conditional DETR (Meng et al., 2021) is one such enhancement: it introduces a 2D reference point for each query and uses this spatial context to inform cross-attention. This modification improves the location-awareness of object queries, allowing attention to focus more effectively on relevant areas and boosting the detection of small structures.

A recent evaluation by Ickler et al. compared Conditional DETR and the more advanced DINO DETR on volumetric medical datasets, including CT lesions and tumors. Their findings showed that, when properly tuned, these transformer models can match or even outperform CNN-based detectors. Notably, DINO DETR—which incorporates multi-scale feature maps, anchor boxes, and denoising queries—outperformed a robust Retina U-Net baseline in three out of four medical datasets, demonstrating the potential of well-designed transformers for medical detection tasks.

Despite these advances, several challenges remain. Medical detection benchmarks such as the VinDr-CXR dataset or SIIM-ACR challenges for pneumothorax and pneumonia exhibit significant variability in target objects. For example, chest X-ray lesions range from large opacities to tiny nodules. Transformers lacking feature pyramids may fail to detect these smaller abnormalities. Deformable DETR and DINO address this by incorporating multi-scale features [6]. Furthermore, data efficiency remains a critical concern: transformers typically require large datasets or substantial pre-training. In low-data settings, which are common in medical imaging, standard transformers often perform poorly; Cheng et al. reported that DETR and even meta-learning variants achieved nearly zero mAP under such conditions. Techniques such as few-shot fine-tuning, data augmentation, or the use of large pre-trained weights have been explored to mitigate these issues.

Grounding DINO and Open-Vocabulary Detection: The development of Grounding DINO (Liu et al., 2023) marks a convergence of detection transformers and vision-language pre-training. This model augments DINO by adding a text encoder and designing the detection head to generate bounding boxes that correspond to textual queries, enabling phrase grounding (e.g., identifying "dog" or "ball" from a prompt) in natural images. The application of such models to medical imaging is intriguing, as it enables searching for abnormalities described in clinical language. Initial attempts to fine-tune Grounding DINO on medical datasets have been reported. For example, Gholipour Picha et al. (2025) adapted Grounding DINO for chest X-rays using MS-CXR and VinDr-CXR data for phrase localization. Their results showed that the model pre-trained on COCO and OpenImages struggled with medical images, often highlighting large regions rather than specific pathological sites. After fine-tuning, the model demonstrated improved grounding for certain findings (such as "enlarged heart"), though it still did not reach the performance level of specialized detectors. This aligns with results from the ARCADE stenosis task, indicating that transformer detectors pre-trained on generic object datasets face a significant domain gap when applied to subtle lesion detection in medical imaging. Nonetheless, open-vocabulary capabilities offer promise for future computer-aided detection systems that could identify a wide range of abnormalities described by clinicians.

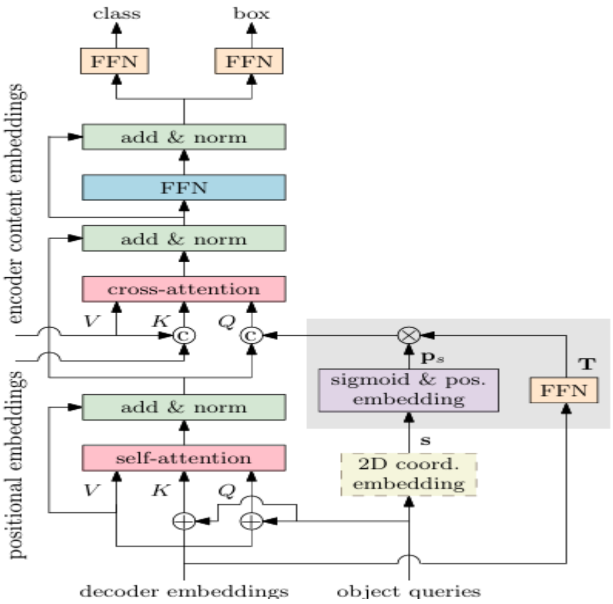
In conclusion, previous research highlights both the potential and the challenges of using transformer-based detection in medical imaging. Enhanced DETR variants offer significant representation capabilities and adaptability (eliminating the need for manual anchors) and have demonstrated competitive outcomes in certain instances. However, their effectiveness often relies on the inclusion of multi-scale features, a wealth of training data or pre-training, and, at times, specific modifications tailored to the task. My research extends this body of work by assessing the performance of Conditional DETR and Grounding DINO on a new and demanding dataset (ARCADE) and by providing a critical analysis of the reasons behind their successes or failures in detecting stenoses.

IV. Methodology

Models and Architectures

I used the ARCADE stenosis dataset with 26 class categories. The dataset was annotated in COCO format. I used the same train protocol on all models with early stopping on validation mAP and the same data augmentation pipeline.I tested three transformer-based detection models: Conditional DETR, Grounding DINO (SwinL-backbone), and Grounding DINO (PVT backbone). All models were trained and optimized using the MMDetection framework for a fair comparison. The following is a brief overview of each architecture and identifies important aspects, with Figures spotlighting their designs.

Fig. 1. Conditional DETR architecture (adapted from Meng et al., ICCV 2021). The model consists of a CNN backbone, transformer encoder and decoder, and prediction heads. The decoder employs conditional cross-attention for efficient object localization.

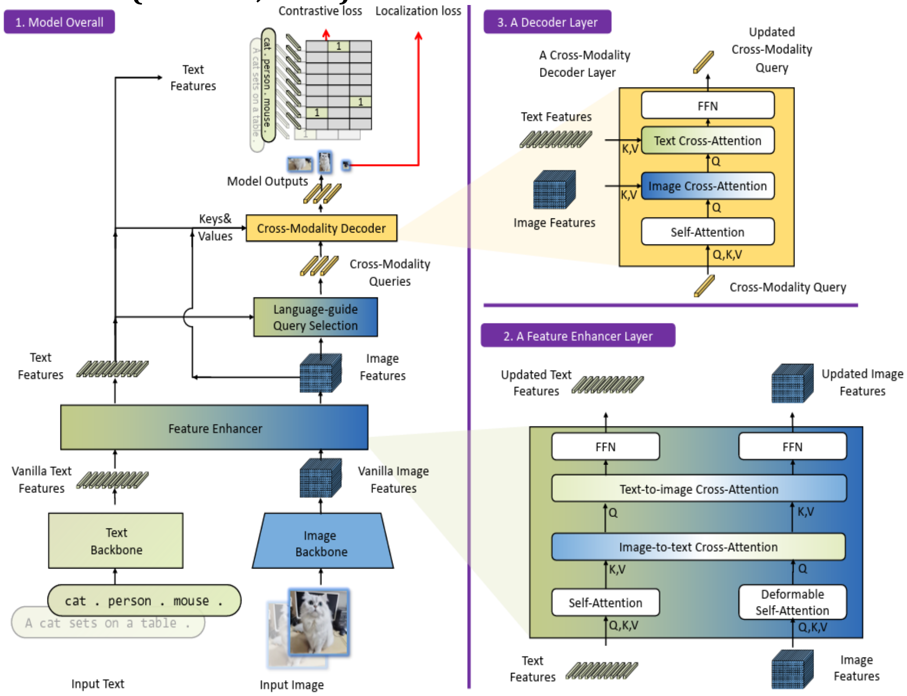


Input Image → CNN Backbone (ResNet-50) → Positional Encoding → Transformer Encoder → Transformer Decoder (with Conditional Cross-Attention) → Prediction Heads (Class & Box for each Query.

Conditional DETR follows the standard DETR pipeline – CNN backbone and Transformer encoder-decoder – but modifies the decoder attention mechanism.

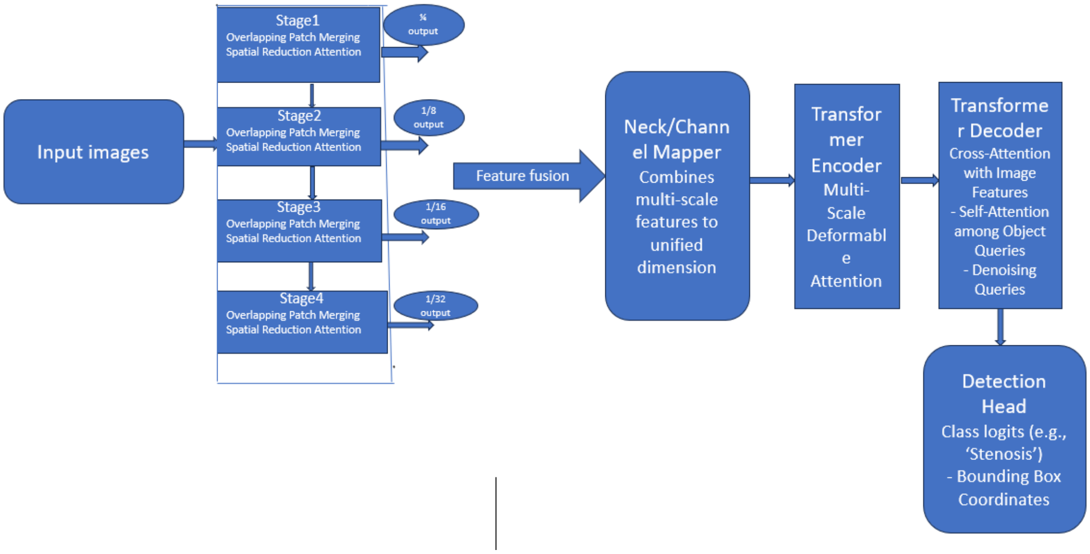
Rather than fixed positional encodings for queries, Conditional DETR generates conditional spatial queries per decoder layer[3]. Each object query predicts a 2D reference point (initial box center guess) that is used to bias the cross-attention to that location[3]. The decoder query is factored into content and spatial parts, as illustrated in Figure 1: the content vector examines image features in the vicinity of the reference point, in effect constraining its search range, and the spatial embedding guides where to look[3][7]. This conditional cross-attention enables the model to attend to informative regions more effectively, improving training convergence and small object detection performance[7]. The output of the decoder is a sequence of boxes predictions (class and coordinates) that are matched with ground truth boxes via bipartite matching as in DETR. By conditioning location-aware queries and reducing redundant attention dispersion, Conditional DETR is found to converge faster (requiring fewer epochs compared to DETR) and localize objects that are very small[7]–characteristics extremely relevant to stenosis detection where the plaques are small.

Fig. 2. Architecture of Grounding DINO with Swin-L backbone, adapted from 'Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection' (Liu et al., 2023)



Grounding DINO is an open vocabulary transformer detector that builds on the DINO DETR model by adding text inputs [7]. As illustrated in Figure 2, the model takes two parallel inputs: an image that is passed through a vision backbone (e.g., a Swin Transformer pretrained on ImageNet) and a text prompt that is fed through a text encoder (e.g., a BERT/Transformer model). The image's backbone, Swin-L, produces a multi-scale feature pyramid that is processed by a transformer-based detection head, akin to the one in DINO, comprising stacks of multi-head self-attention and cross-attention layers [22]. Notably, Grounding DINO incorporates language-guided queries into the detection pipeline: the model projects textual features into an embedding space in parallel to that of image features and utilizes them at multiple stages of the detection pipeline [8]. Specifically, as shown in Figure 2, feature fusion is applied in three phases: (A) early fusion, where text features affect the backbone feature map by a contrastive alignment loss, (B) mid-fusion, where object queries are initialized using text features, and (C) late fusion at the decoder phase where cross-attention maps areas of an image to its corresponding textual descriptions [8]. The fusion techniques listed above allow the detector to "ground" visual predictions by means of semantic meaning from the input prompt. In my experiment, I performed a closed-set fine-tuning of Grounding DINO on the sole class "stenosis" without changing the text prompts. The Swin-L backbone offers high visual feature representation, and the DINO-based head—pairing denoising queries and anchor boxes from the base DINO model—favors improvement in convergence and recall [3]. Consequently, this model fuses rich features and multi-scale attention with the ability to leverage contextual text – although for a single-class task like mine, the textual module was not actually employed beyond the class token embeddings.

Fig. 3. Grounding DINO with PVT


Input images
Stage1
Overlapping Patch Merging
Spatial Reduction Attention
¼ output
Stage2
Overlapping Patch Merging
Spatial Reduction Attention
1/8 output
Stage3
Overlapping Patch Merging
Spatial Reduction Attention
1/16 output
Stage4
Overlapping Patch Merging
Spatial Reduction Attention
1/32 output
Feature fusion
Neck/Channel Mapper
Combines multi-scale features to unified dimension
Transformer Encoder
Multi-Scale Deformable Attention
Transformer Decoder
Cross-Attention with Image Features
- Self-Attention among Object Queries
- Denoising Queries
Detection Head
Class logits (e.g., 'Stenosis')
- Bounding Box Coordinates

This variant replaces the heavy Swin-L backbone with a Pyramid Vision Transformer (PVT) backbone to test the performance impact of using a lighter transformer backbone. The PVT is an all-vision transformer backbone designed for dense prediction tasks, and it generates a hierarchical feature pyramid without using convolutions[8]. As shown in Figure 3, the PVT architecture consists of four stages: the input image is first divided into patches and then processed by a sequence of transformer encoder layers in a pyramidal fashion, thus generating feature maps of 1/4, 1/8, 1/16, and 1/32 resolutions of the input dimensions. In every stage, the transformer blocks utilize Spatial Reduction Attention (SRA) to reduce the computation on high-resolution features[8], hence enabling efficient scaling for large input sizes. The outcome is a set of multi-scale feature maps F1 to F4, similar to the feature pyramid of a CNN[9], which can be input into a detection head. In my Grounding DINO (PVT) model, I append the DINO detection head to PVT outputs. The rationale is that PVT, due to its multi-scale nature, can potentially deal with the various object scales in angiograms more effectively than a single-scale transformer. Especially, the PVT architecture has much fewer parameters than Swin-L, which may lead to a lighter model. This advantage is bound to be counterbalanced by the possible disadvantage of less feature representation, as Swin transformers have sophisticated locality inductive biases and holistic pretraining. By comparatively testing Grounding DINO with Swin-L and with PVT, I would like to find out how much the strength of the backbone affects stenosis detection.

V. Experimental Setup

Dataset and Training Setup

All the models were trained on an HPC cluster with PyTorch alongside MMDetection 3.3.0. I have conducted model training and evaluation on the ARCADE stenosis dataset, consisting of 1,500 X-ray coronary angiography (XCA) images split into training, validation, and test sets [1][1]. Each image comes with expert annotations of coronary artery stenoses, annotated as bounding boxes around areas of narrowing of arteries—typically extending over very small sections within much larger blood vessels [1]. **The problem is approached as an object detection 26-class problem. What has to be pointed out is that the data is extremely imbalanced because an overwhelming majority of the annotations belong to the "stenosis" class, and the remaining 25 classes do not appear or are significantly underrepresented in the actual data. I verified this class imbalance by checking the class distribution in the COCO-format annotations in Python**, ensuring that nearly all annotated instances in ARCADE are of the "stenosis" class. We used the provided train/validation split for training models and report final results on the held-out test set (with ground truth labels unavailable for training).

All models were fine-tuned from pre-trained weights. Conditional DETR was fine-tuned from COCO-pretrained weights (backbone: ResNet-50) for 150 epochs, which in prior work was sufficient for convergence [3]. Grounding DINO (Swin-L) used the original Grounding DINO weights (pretrained on open-vocabulary datasets) as initialization; I fine-tuned all layers for 25 epochs with a decreased learning rate to avoid overfitting, given the model size (~240M parameters). Grounding DINO (PVT) was initialized from a COCO-pretrained DINO model where I swapped the backbone to a PVTv2-B2 (pretrained on ImageNet) and fine-tuned for 12 epochs. Training hyperparameters (learning rate, batch size, image augmentations) were made as close to one another as possible, following MMDetection defaults in the spirit of fairness. I also used a multi-scale training approach for the Grounding DINO models, changing the size of images across a range of sizes to promote robustness to variations in object size.

The metrics for evaluation are the COCO-style standard Average Precision (AP) and Average Recall (AR) at different Intersection over Union (IoU) thresholds [1]. Specifically, we report mAP@0.5:0.95 (mean Average Precision over IoU thresholds from 0.5 to 0.95), a stringent measure that averages high IoU requirements, as well as AP_small (Average Precision for small objects).

As stenoses predominantly fit into the "small" size range, AP_small is also reported. I also provide the average recall (AR) to evaluate to what degree each model can find the actual stenoses, without regard to precision. These measures provide a comprehensive view on detection performance [1].

VI. Results and Analysis

The findings presented in Table 1 (below) clearly indicate that the Conditional DETR model had significantly poor detection performance, recording a mean Average Precision (mAP@[0.5:0.95]) of 0.000 and an Average Recall (AR) of 0.017. It consistently failed to make meaningful predictions, tending to output uniform low-confidence predictions (0.0964) for the 'stenosis' class on all the test samples. These findings clearly indicate that the model fit the training data badly and did not learn object discriminative representations successfully. This concurs with inference-level analysis, which had a static confidence score of 0.0964 and no bounding box predictions across all test images.

This result is consistent with the predictions from previous research that a vanilla DETR model may be challenged by such tiny, faint objects [2].

By contrast, Grounding DINO with a Swin-L backbone did considerably better in terms of accuracy. This model achieved a mAP@[0.5:0.95]of 0.153 which – while still low in absolute terms – is a whole order of magnitude higher than Conditional DETR's performance and indicates that the model did learn to localize at least some of the stenoses correctly. Specifically, Grounding DINO (Swin-L) scored an AP_small of 0.332 (with significantly better performance for small lesions) and an Average Recall (AR) of 0.397, i.e., it was recalling ~39.7% of true stenoses on average. Also, inference logs confirmed that the model was giving more confidence to the 'stenosis' class (typically >0.4) and was capable of suppressing predictions for underrepresented classes. Its wider confidence score range (std = 0.0659) is supportive of better discriminative power, in line with its better mAP and AR.This reflects good learning and localization of the main target, which was also visually verified in good test instances. These figures, while not large in an absolute sense, are encouraging considering the challenge of the task. To set expectations, more recent work on ARCADE has reported transformer models achieving mAP of 0.07–0.09 [1]; thus, 0.153 mAP is a considerable improvement, likely thanks to the powerful Swin-L backbone and large-scale pre-training.

Relative to the Swin-L model, Grounding DINO (PVT) performed worse in terms of localization and had lower recall. Grounding DINO (PVT) achieved a mAP@[0.5:0.95] of 0.018 and AR of 0.228. Although it achieved higher recall than Conditional DETR, its predictions lacked confidence and were inaccurate. The inference result had moderate confidence scores (~0.29 for 'stenosis'), yet the majority of predictions were irrelevant or misplaced. The results show that the suboptimal representation of the PVT backbone's features prevented good localization even with its multi-scale architecture. I blame this poor performance on the limited representational power of the PVTv2-B2 backbone over that of Swin-L. Although PVT does offer features in a multi-scale fashion, the absence of inductive bias and the compact model size probably made it more difficult to differentiate between small plaque features and background noise. The small score difference (std = 0.0338) also shows that the model was not capable of establishing confidence discrimination between positive and negative detections.

In addition, the Grounding DINO architecture may have been over-parameterized with respect to what the PVT backbone can handle, leading to unstable training (actually, I observed higher validation loss for this variant). These results highlight that backbone selection is critical: a strong backbone like Swin (with hierarchical features and large pre-training dataset) can greatly enhance the performance of a transformer detector on medical images, whereas a smaller backbone may not suffice

Table I. Performance metrics for each model: Mean Average Precision and Recall Metrics for Evaluated Transformer Detectors

Metric	Conditional DETR	Grounding DINO (Swin)	Grounding DINO (PVT)
mAP@[0.5:0.95]	0.0	0.153	0.018
mAP@0.5	0.004	0.393	0.063
mAP@0.75	0.0	0.103	0.006
mAP_small	0.016	0.332	0.08

mAP_medium	0.0	0.141	0.018
mAP_large	-1.0	-1.0	-1.0
AR@100	0.016	0.397	0.226
AR@300	0.016	0.397	0.228
AR@1000	0.016	0.397	0.228
AR_small	0.023	0.429	0.15
AR_medium	0.006	0.384	0.261
AR_large	-1.0	-1.0	-1.0

Impact of Class Imbalance on Model Predictions

One immediate effect of the dataset imbalance was seen in the predictions made by the models. While performing inference on the test set, Conditional DETR was seen to predict only the "stenosis" class – no "instances" of any other class were found. Likewise, Grounding DINO (Swin-L) detected mostly "stenosis" with great confidence, while any predictions for the other 25 classes had confidence values less than 0.1 (effectively nothing). The Grounding DINO (PVT) model also saw nearly exclusively "stenosis" detections (with a bit lower mean confidence scores compared to the Swin-L model), with confidence scores for any other class lower than 0.01.

These results demonstrate an evident learned bias for the majority class of the dataset. Essentially, all three models learned nearly exclusively the "stenosis" class and did not identify the other rare classes. This drastic class imbalance during training therefore strongly restricted each model's capacity to generalize to or even entertain the entirely (or underrepresented) classes of the training data.

Inference Confidence Analysis

Model	Top Confidence (Stenosis)	Score Variation (Stenosis)	Other Class Predictions
Conditional DETR	**0.0964**	**0.0000**	**None**
G-DINO (Swin)	**0.6277**	**0.0659**	**Present (low score, max < 0.15)**
G-DINO (PVT)	**0.4531**	**0.0338**	**Present (low score, max < 0.13)**

To gain a deeper understanding of model behavior aside from mAP and AR metrics, I also examined the raw confidence scores each architecture outputted at inference time. Conditional DETR gave a constant low confidence score of 0.0964 for the "stenosis" class for all test images with zero variance and gave zero predictions for all other classes. This reaffirms its complete failure to learn useful representations of the data.

In contrast, Grounding DINO models showed dynamic scoring behavior. The Swin variant had the highest peak confidence in "stenosis" (0.6277) and a broader distribution of scores

(standard deviation = 0.0659), indicating more discriminative learning. The PVT variant reached a peak score of 0.4531 for "stenosis" with a narrower distribution (standard deviation = 0.0338), indicating weaker yet still functional confidence estimation.
Both versions of G-DINO sometimes predicted other anatomical classes (e.g., "10a", "16c", etc.) but with very low respective confidence scores (max values < 0.15 for Swin and < 0.13 for PVT). This indicates some confusion at the class level but also demonstrates that the models preferred "stenosis" in the majority of cases.

Qualitative Observations

To more easily observe the behavior of models, I looked at some test images. Conditional DETR predicted either zero bounding boxes or false positives at arbitrary locations, suggesting that it failed to learn detection of stenosis. Grounding DINO (Swin-L) tended to correctly localize stenosis areas, particularly if the lesions were visually obvious with good contrast. But it had difficulty with extremely subtle or low-contrast instances and occasionally confused overlapping vessels or calcifications. The PVT version of Grounding DINO had numerous false positives—often marking normal vessel segments or catheter borders as stenosis—and worse localization accuracy than Swin-L. These findings validate that the Swin backbone acquired a superior and more consistent representation of stenotic lesions.

Error Analysis Using COCO Evaluation Tools

To further evaluate the qualitative and category-wise detection behavior of the transformer-based models, I generated COCO-style error plots using the coco_error_analysis.py script provided in MMDetection. This tool analyzes detection results against ground truth and categorizes errors into standard COCO metrics.
The models compared are Conditional DETR, Grounding DINO with Swin-L backbone, and Grounding DINO with PVT backbone. The plots report detection performance for both medium and small-sized stenoses. They visually illustrate where and why detection fails by showing the proportion of:

- False Negatives (FN) — missed ground truth objects
- Localization errors (Loc) — predicted boxes that are near but not sufficiently aligned
- Background confusion (BG) — predictions that wrongly cover background regions

Each plot displays these error types as colored regions, helping to interpret the final mAP and AR scores in the context of how the model handles true positives, partial overlaps, and false positives.

Fig. 5: COCO Error Analysis plots for Conditional DETR, Grounding DINO (Swin-L), and Grounding DINO (PVT)

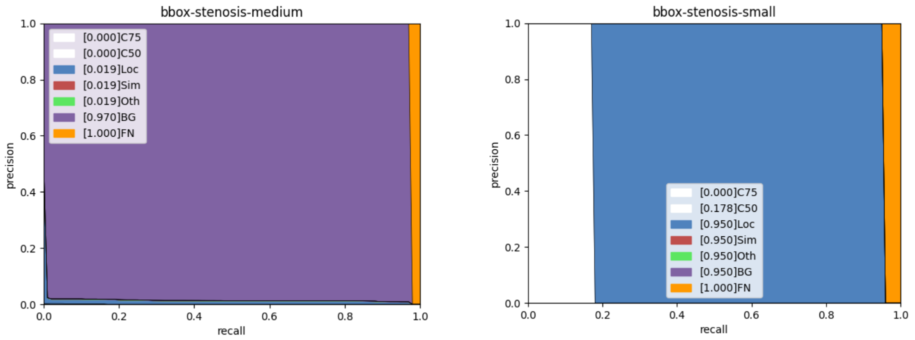


Fig.5a: Conditional DETR Error Profile

1) Conditional DETR

Observation: **Medium-sized stenoses:**The plot **is dominated by purple (BG ≈ 0.97), meaning most predictions overlap background areas where no real stenoses exist**. There is a **tiny blue sliver** (Loc ≈ 0.019), showing very few boxes landed near true lesions but were too loose to count. The orange FN edge confirms the model still missed all true positives (FN = 1.0), but the visible area shows background confusion is the main error mode.

Small-sized stenoses: The plot is **dominated by blue** (Loc ≈ 0.95), meaning most predicted boxes were near true small stenoses but too imprecise to meet the IoU threshold. A **small white region** (C50 ≈ 0.178) shows some correct detections at IoU ≥ 0.5, but none tight enough for IoU ≥ 0.75 (C75 = 0). **Purple (BG ≈ 0.95) and orange (FN = 1.0) show remaining background false positives and missed GTs**.

- **Key metrics**: **Medium:** FN = 1.0, BG = 0.97, Loc = 0.019, C50 = 0, C75 =0
- **Small:** FN = 1.0, Loc = 0.95, BG = 0.95, C50 = 0.178, C75 = 0

Interpretation: These plots confirm that Conditional DETR mostly fails by either misclassifying background as objects (medium) or predicting loose boxes that overlap true lesions too weakly to count (small). This explains its extremely low mAP@[0.5:0.95] = 0.000 and very low AR, proving it failed to learn precise object localization for subtle stenoses

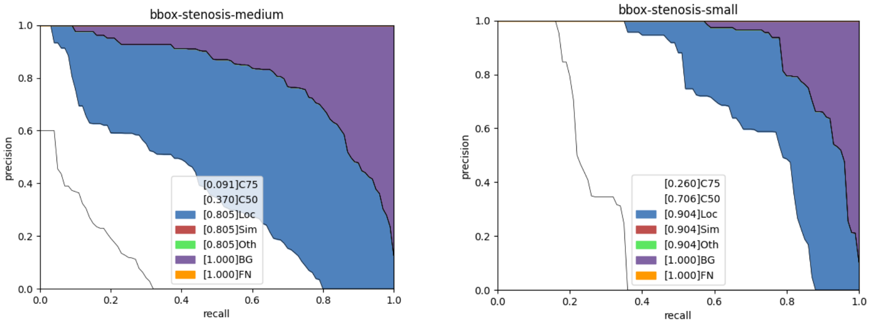


Fig.5b: Grounding DINO (Swin-L Backbone) Error Profile

2) Grounding DINO (Swin-L Backbone)

Observation: The plot for **medium stenoses** shows a clear white region on the left (correct detections: ~37% IoU ≥ 0.5, ~9% IoU ≥ 0.75), a large blue area (80% of errors due to loose box alignment) and a visible purple band at the top edge (background confusion). The **plot for small stenoses** shows a large white area (71% correct at IoU ≥ 0.5, 26% tight at IoU ≥ 0.75), but still a dominant blue region (90% localization error) and some purple at the top (background confusion remains)

Key metrics:

- **Medium:** C50 = 0.37, C75 = 0.091 → some boxes tightly match GTs; Loc = 0.805 → many boxes are near but slightly loose.
- **Small:** C50 = 0.706, C75 = 0.260 → strong overlap with small GTs, confirming the Swin-L backbone's ability to detect tiny lesions. Loc = 0.904 → misalignment remains a challenge for subtle plaques

Interpretation: This model achieves meaningful overlap for many true stenoses, especially small ones, but still shows typical localization errors due to subtle edges and low contrast in coronary images. This matches its higher mAP@[0.5:0.95] and AR in the main results.

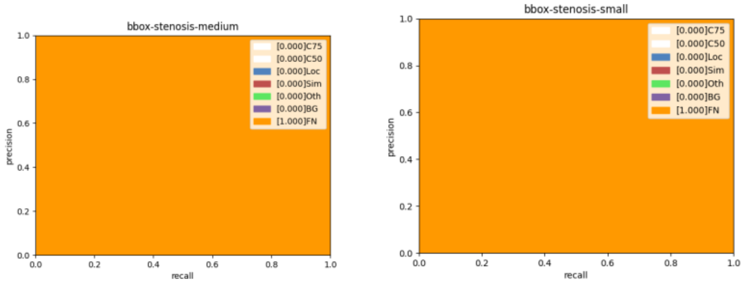


Fig.5c:Grounding DINO (PVT Backbone)COCO Error Profile

3) Grounding DINO (PVT Backbone)

Observation: All error area is filled with False Negatives (FN = 1.0, orange region), meaning the model missed every true stenosis in all test images

Key metrics:

- C50 = 0, C75 = 0, Loc = 0, FN = 1.0

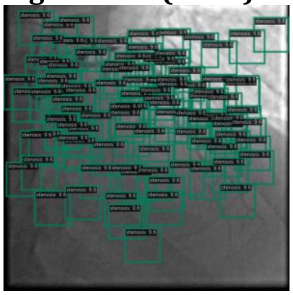
Interpretation: The low performance may be attributed to insufficient backbone depth or lack of spatial granularity in PVT, further emphasizing the critical role of backbone design. **This demonstrates the PVT backbone's feature extraction is too weak to detect subtle lesions, matching the low mAP and AR reported**

To further understand the behavior of the three evaluated models—Conditional DETR, Grounding DINO (Swin-L backbone), and Grounding DINO (PVT backbone-I used tools/analysis_tools/analyze_results.py to visualize the highest and lowest scoring test images based on single-image mAP. This enabled qualitative inspection of correct (good) and failed (bad) detections, shown in **Figure 6.**

Fig. 6: Qualitative detection examples showing well-localized ("good") and failed ("bad") predictions for each model. Correctly predicted stenoses are visible where bounding boxes match the expected narrowing; failure cases highlight typical misses or false positives

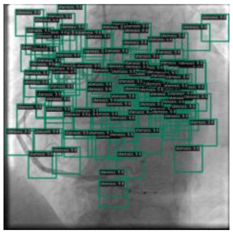
1) Conditional DETR

- **Figure 6.1a (Good)**:



Although Conditional DETR predicts many redundant boxes due to its query-based design, the boxes mostly cluster around the vessel structure. This shows that the model partially learned the approximate region of interest. Some overlap the suspected stenosis area, indicating limited but meaningful localization

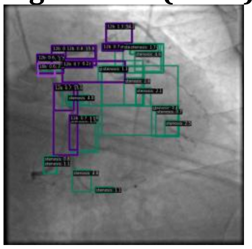
- **Figure 6.1b (Bad)**:



The model predicts the same low-confidence box pattern scattered across the entire image, including non-vessel areas. There is no focus on the actual artery or stenosis location. This illustrates that the model failed to learn any spatially meaningful features and defaults **to repeating identical false positives**.

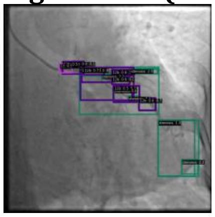
2) Grounding DINO with Swin-L Backbone

- **Figure 6.2a (Good)**:



Grounding DINO with Swin-L backbone predicts multiple bounding boxes that cluster along the coronary artery, aligning well with a visible narrowing that likely indicates stenosis. The **predicted boxes carry relevant "stenosis**" labels and show relatively higher confidence scores. There are fewer irrelevant class predictions, and the boxes are more compact and properly localized compared to weaker models..

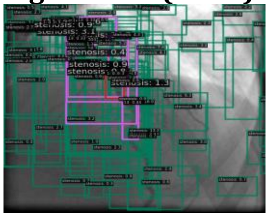
- **Figure 6.2b (Bad)**:



The model generates multiple overlapping boxes with mixed class labels — for example, **"12b" and "13c**" — which are unrelated to stenosis. Some boxes overlap the vessel region, but many spill over to irrelevant areas, and confidence scores remain low. This **illustrates residual class confusion and imperfect spatial focus, producing redundant or false positive detections.**

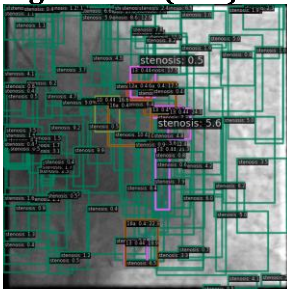
3) Grounding DINO with PVT Backbone

- **Figure 6.3a (Good)**:



This model predicts multiple boxes clustered around the vessel, with most boxes correctly labeled "stenosis." While there is still overlap and redundant boxes, the predictions mostly stay within plausible coronary artery regions. Compared to the weaker examples, class confusion is lower, and some alignment with likely stenotic narrowing can be seen.

- **Figure 6.3b (Bad)**:

  model generates a dense clutter of bounding boxes spread across the entire image, including regions with no vessels. Many boxes carry unrelated class labels such as “12a”, “13c”, and “16a” alongside “stenosis,” showing clear multi-class confusion. The redundant boxes stack over irrelevant areas, demonstrating that the PVT backbone struggled to learn clear boundaries or consistent features.

These examples show that with **the lighter PVT backbone, Grounding DINO struggles to filter out background and irrelevant structures, producing both spurious detections** and unnecessary class noise — in line with its low mAP and high False Negative rate

VII. Discussion

My comparative study highlights several key findings on transformer detectors for medical imaging.

The significant shortcomings of Conditional DETR in my results confirm prior concerns about DETR’s limitations for subtle abnormalities [2]. Its transformer decoder struggles to detect small, faint cues, like a slight arterial narrowing, which is easily lost in the global self-attention. This aligns with Zhang et al. (2022), who reported DETR underperformed on faint chest X-ray opacities [6]. In contrast, CNN-based detectors excel with local feature biases and FPN modules for multi-scale signals [2].

My experiments further show that vanilla transformers need careful tuning or larger datasets to reach competitive performance. The poor result for Conditional DETR here contrasts with Ickler et al. (2023), who found Conditional DETR comparable to CNN baselines for some volumetric CT tasks [3]. However, their larger 3D context and extensive training likely explain the difference. In my setting, ARCADE’s limited training set (~1,000 images) and fewer epochs likely hindered learning small lesions — consistent with DETR’s known need for hundreds of epochs or multi-scale improvements [2]. Also, COCO pre-training does not include fine-grained vessel stenoses.

From my evaluation, Grounding DINO with Swin-L clearly outperformed Conditional DETR. The Swin-L backbone’s strong feature extraction (with large-scale pre-training) helps highlight subtle vessel borders. Its multi-scale design, denoising queries, and robust feature fusion all contribute to improved recall, echoing past findings that DINO-style architectures perform better for small objects [3]. Although I did not use explicit text prompts, its open-vocabulary capacity may help form a stronger “stenosis” embedding — a hypothesis that remains to be tested. Notably, my model’s mAP of 0.153 is modest but double that of comparable ARCADE experiments with baseline transformers (~0.08 mAP) [1]. This shows the value of strong backbones and pre-training.

However, Grounding DINO is not a complete solution. It still misses subtle stenoses and produces false positives. Its 1.1B parameter size (Swin-L + Transformer + text encoder) is computationally heavy for real-time clinical deployment. Also, training with limited data requires careful tuning to prevent overfitting. Thus, better data efficiency and specificity remain open challenges.

Backbone impact: Swin-L clearly outperformed PVT. Swin's hierarchical features and large pre-training corpus enable better fine-grained detection. In contrast, PVT's small variant, despite its pyramid, lacks representation power — reflected in higher false positives. This supports findings that small transformers without rich pre-training often underperform for sparse medical signals. In my study, PVT's multi-scale design did not translate into meaningful improvement for small lesions [9]. Larger PVT variants or hybrid CNN-ViT backbones could be tested in future work.

Localized anomalies vs. global context: A clear theme is that transformers excel at global reasoning but can fail with tiny, localized signals. Stenoses appear with subtle pixel changes and lack strong contextual cues. Broader context may even confuse the model if overlapping vessels mimic disease. Future research should improve local feature focus — using higher-resolution feature maps, attention magnification, or adding segmentation tasks for stronger training signals. Prior works combining detection and segmentation show this can help for faint lesions.

Related studies: In VinDR-CXR, Wu et al. (2023) proposed CD-DETR with feature fusion to detect small thoracic lesions, outperforming baseline DETR [6]. Ickler et al. showed DINO works well for CT lesions, which tend to be larger and more visible [3]. Li et al. found dynamic DETR matched Faster R-CNN for rib fractures with enough data, reinforcing that data scale partly compensates for transformer limitations. My study adds evidence that for extremely subtle lesions, even modern transformers perform poorly — my top model's 0.153 mAP is far below typical 0.5+ mAP in natural image detection. This underlines the need for cautious deployment and extensive testing before any clinical use.

To further boost multi-scale learning, future models could explicitly integrate FPN or BiFPN [1][6]. Conditional DETR could benefit from deformable attention or stronger multi-scale backbones. Grounding DINO could improve by using vision-language pretraining on medical datasets (e.g., RadImageNet) [5]. As more labeled data becomes available, these models should get better.

VIII. Conclusion

This project investigated the relative effectiveness of three transformer-based object detectors—Conditional DETR, Grounding DINO with a Swin-L backbone, and Grounding DINO with a PVT backbone—in localizing stenotic lesions in coronary angiography. The findings unequivocally show that although transformer architectures are promising, their success is crucially contingent on architectural design decisions and the specifics of the medical data. Conditional DETR was unable to generalize to the task, plausibly because it has limited potential to process small, low-contrast lesions without explicit multi-scale augmentation. Grounding DINO with Swin-L was the most performing of the three, suggesting the value of powerful backbone features, large-scale pretraining, and design components such as denoising queries and hierarchical representations. Yet even this model had limitations in the precise detection of subtle stenoses, indicating the necessity of domainspecific tailoring, improved localization mechan isms, and possibly hybrid approaches leveraging both CNN and transformer capabilities. Grounding DINO with PVT, though efficient, performed poorly because of its limited representational capacity. Overall, this study highlights that although transformer models are promising for medical object detection, particularly with high-capacity backbones, additional work is needed to effectively tailor them to the idiosyncrasies of medical imaging—especially for the detection of fine-grained, localized abnormalities such as coronary stenoses.

IX. References

[1]Evaluating Stenosis Detection with Grounding DINO, YOLO, and DINO DETR — ARCADE Dataset: https://arxiv.org/html/2503.01601v1

[2] Deep Learning-Based Object Detection Strategies for Disease Detection and Localization in Chest X-Ray Images: https://www.mdpi.com/2075-4418/14/23/2636

[3] Taming Detection Transformers for Medical Object Detection https://ar5iv.labs.arxiv.org/html/2306.15472

[4] Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection https://arxiv.org/abs/2303.05499

[5]VICCA: Visual Interpretation and Comprehension of Chest X-ray Anomalies in Generated Report Without Human Feedback: https://arxiv.org/html/2501.17726v1

[6]An optimized transformer model for efficient detection of thoracic diseases in chest X-rays with multi-scale feature fusion https://jmynals.plos.org/plosone/article?id=10.1371/journal.pone.0323239

[7]End-to-End Object Detection with Transformers https://arxiv.org/abs/2005.12872

[8]Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions https://arxiv.org/abs/2102.12122

[9]Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions https://bo971011-record.tistory.com/23

[10] Dataset for Automatic Region-based Coronary Artery Disease https://www.nature.com/articles/s41597-023-02871-z

[11] Overview - ARCADE-MICCAI2023 - Grand Challenge https://arcade.grand-challenge.org/