COL828 A2
Exploring Open-Vocabulary Object Detectors for Breast Cancer Detection using Mammograms

Yash Bansal (2022CS51133)

November 25, 2025

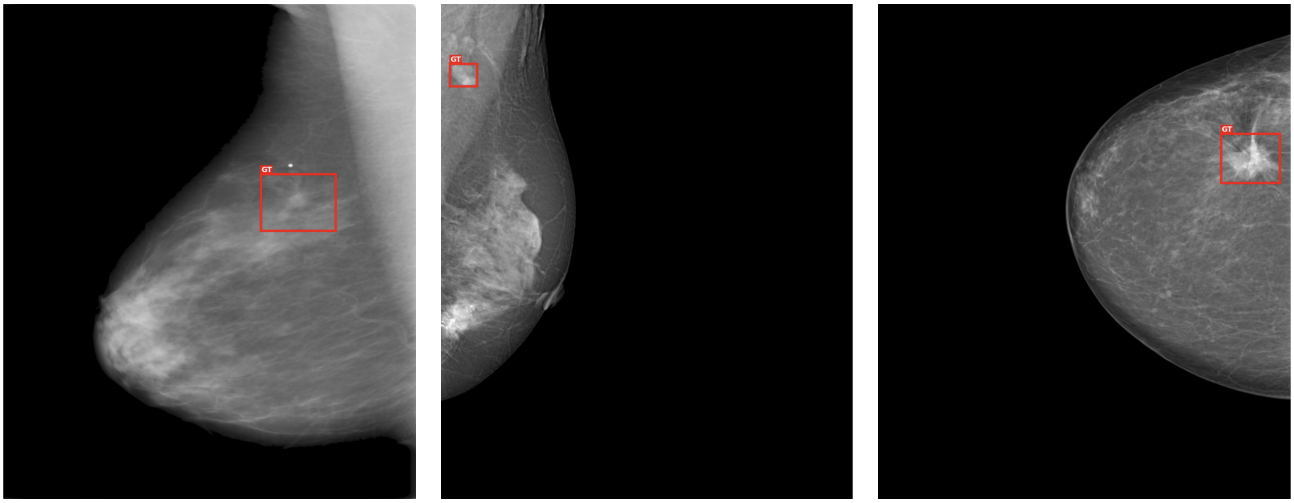
Contents

1 Introduction

This assignment explores Open-Vocabulary Object Detectors (OVODs) for breast cancer detection in mammography images using three anonymized datasets (A, B, C) with Grounding DINO as our framework. We use the Hugging Face implementation[1] with the base model `IDEA-Research/grounding-dino-base`.

1.1 Dataset Characteristics

Figure 1 shows examples from each dataset. Dataset A contains scanned images of printed X-ray mammograms, while Datasets B and C are directly captured from digital mammography machines. This introduces significant domain shift affecting detection performance.



(a) Dataset A (b) Dataset B (c) Dataset C

Figure 1: Sample images from the three datasets with bounding box annotations.

1.2 Experimental Overview

We conduct three main experiments:

1. **Zero-Shot Evaluation:** Baseline performance using hand-crafted prompts without fine-tuning.
2. **Soft Prompt Learning (CoOp and CoCoOp):** Learning soft prompts on each dataset with cross-dataset evaluation (A→B,C; B→A,C; C→A,B) to assess domain generalization capabilities.
3. **Semi-Supervised Prompt Tuning:** CoOp extension with one fully labeled dataset and another with 10% labeled samples, using consistency regularization between weakly and strongly augmented views.

1.3 Model Availability

Trained model weights are available at:
https://www.kaggle.com/datasets/bansalyash12345/col828-a2-submission

[1] https://huggingface.co/docs/transformers/en/model_doc/grounding-dino

2 Zero-Shot Evaluation

We evaluate Grounding DINO's baseline performance using hand-crafted prompts without fine-tuning.

2.1 Prompt Selection and Analysis

We tested various prompts:

- "cancer"
- "breast cancer"
- "mammography cancer detection"
- "malignant tumor"
- "malignant tumor cancer"

The model consistently predicts 1-2 large bounding boxes covering entire breast regions with high confidence (0.4-0.8), while correct localizations receive lower scores (0.1-0.2). This significantly impacts AP metrics, with 4-5 false positives typically ranking before true positives.

Removing "breast" from prompts reduced these whole-breast predictions. The prompt **"malignant tumor cancer"** yielded the best mAP scores and was selected for analysis.

2.2 Qualitative Analysis

Figures 2, 3, and 4 show example predictions from each dataset.

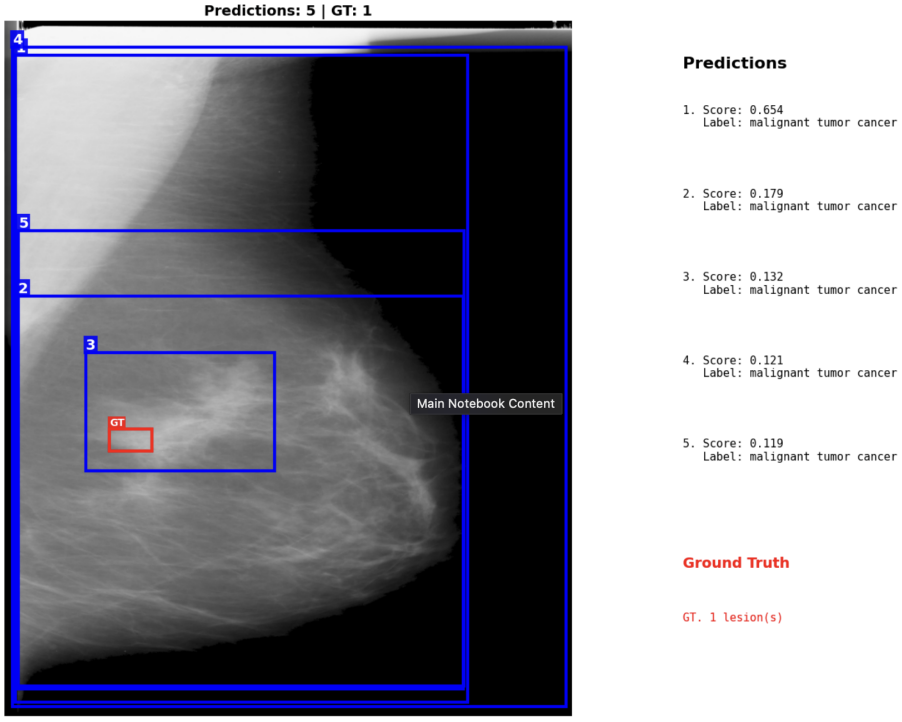


Figure 2: Zero-shot predictions on Dataset A showing large bounding boxes with high confidence.

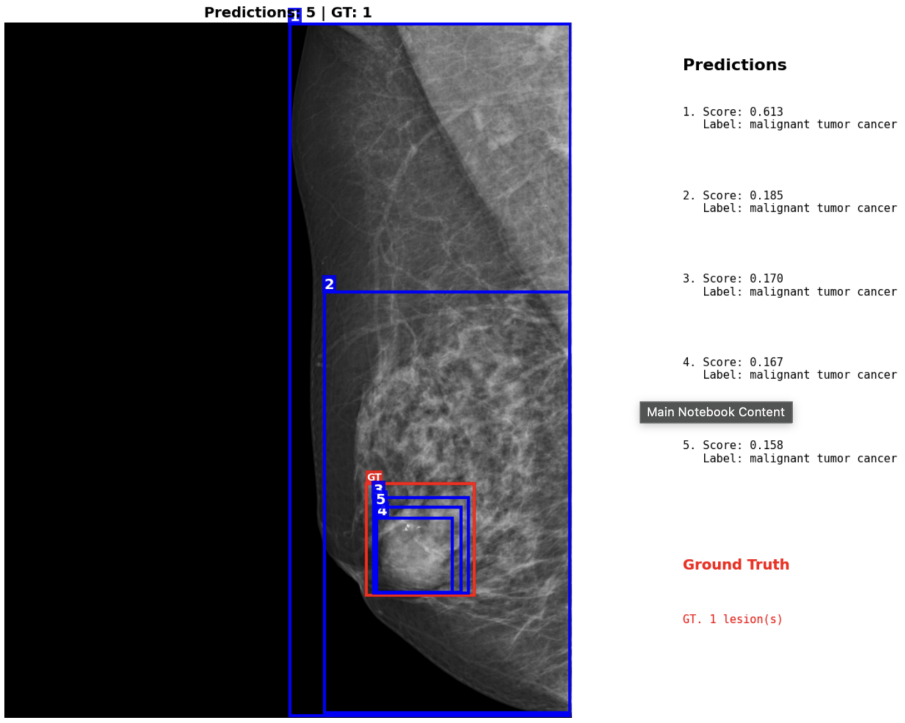


Figure 3: Zero-shot predictions on Dataset B with better localization but confidence issues.

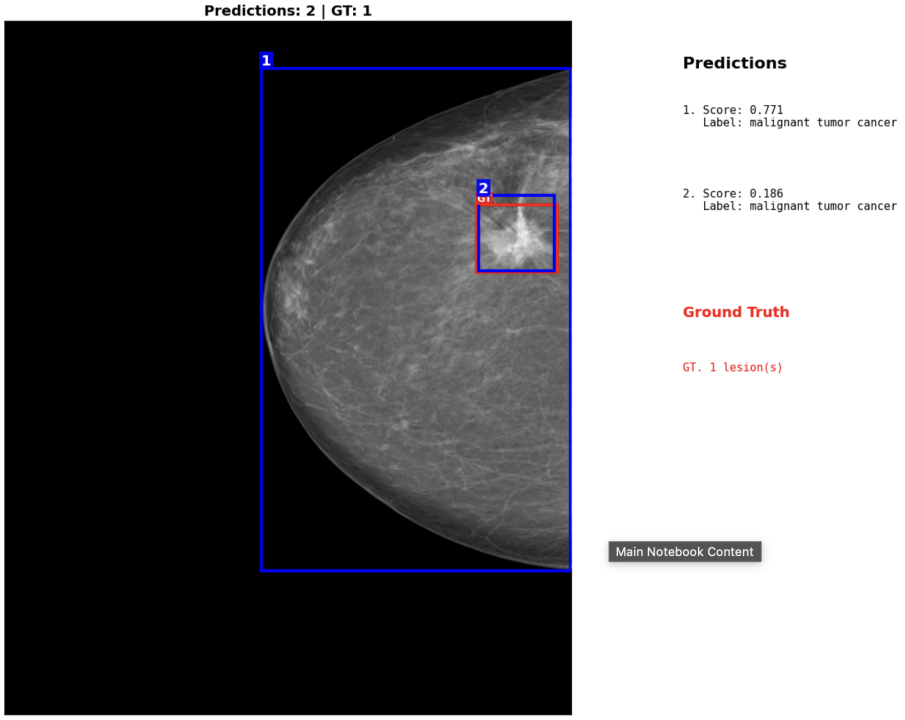


Figure 4: Zero-shot predictions on Dataset C with reasonable localization quality.

For Datasets B and C, bounding boxes show reasonable localization but suffer from low confidence scores. Dataset A predictions are considerably less accurate due to domain shift from scanned images.

2.3 Hyperparameter Selection

We conducted systematic threshold studies:

- **Threshold Selection:** Box and text thresholds of **0.1** provide optimal balance, generating 4-5 predictions per image.
- **Evaluation Strategy:**
  1. Quantitative: mAP computed at threshold = 0 to maximize recall
  2. Qualitative: threshold = 0.1 for interpretable visualizations
- **Rationale:** Higher thresholds eliminate most predictions including true positives; lower thresholds introduce excessive noise.

2.4 Quantitative Results

Table 1 presents Average Precision scores using "malignant tumor cancer" at threshold = 0.

Table 1: Zero-shot Average Precision scores (threshold = 0)

Metric	Dataset A	Dataset B	Dataset C
mAP@50	0.0067	0.0156	0.0320
mAP@75	0.0008	0.0040	0.0163
mAP (Overall)	0.0037	0.0098	0.0242

2.5 Discussion

Key findings:

1. **Domain Shift Impact:** Dataset A's low performance (mAP 0.0037) confirms significant challenges from scanned images due to quality degradation and artifacts.
2. **Confidence Calibration:** Poor calibration assigns high scores to general predictions and low scores to precise localizations, causing false positives to rank above true positives.
3. **Need for Adaptation:** These results motivate learnable prompts (CoOp and CoCoOp) to bridge the gap between pre-training and medical imaging domains.

3 Context Optimization (CoOp)

Having established limitations of hand-crafted prompts, we now learn prompt embeddings through Context Optimization (CoOp), which replaces fixed text with learnable continuous vectors optimized for specific tasks while keeping model weights frozen.

3.1 Implementation in Grounding DINO

Grounding DINO processes text through tokenization, BERT embeddings ($n \times 768$ tensor), and text encoding. We modify `modelling_grounding_dino.py` to accept learnable embeddings instead of fixed BERT embeddings, updating only prompt embeddings during training while freezing all other parameters.

3.2 Key Implementation Insights

Critical factors applied across all experiments:

1. **Special Token Handling:** We freeze [CLS], period, and [SEP] token embeddings to their BERT values, training only intermediate context vectors, as making these learnable degrades performance.
2. **Initialization Strategy:** Initialize prompts with BERT embeddings of meaningful text (e.g., "malignant tumor cancer") rather than random initialization for faster convergence.
3. **Weighted Sampling:** With 4-5× more benign images than malignant ones, we use inverse class frequency sampling to ensure equal batch representation, substantially improving convergence.
4. **Learning Rate:** Cosine annealing from 5×10^{-4} to 1×10^{-4} provides stable convergence (1×10^{-3} causes instability).
5. **Image Standardization:** Standardizing all datasets to common resolution significantly improves cross-dataset transfer and mAP scores.

3.3 Experimental Configurations

Three CoOp configurations:

1. **Configuration 1:** $n_{ctx} = 1$ (single learnable token)
2. **Configuration 2:** $n_{ctx} = 1$ with augmentation (flipping, contrast, gamma, brightness, rescaling)
3. **Configuration 3:** $n_{ctx} = 4$ (four learnable tokens)

Each configuration trains on individual datasets and evaluates on all three test sets.

3.4 Results: Configuration 1 (CoOp with $n_{ctx} = 1$)

Table 2: CoOp results with $n_{ctx} = 1$ (no augmentation)

Trained on	Metric	Test A	Test B	Test C
Dataset A	mAP@50	**0.0862**	0.0920	0.4414
	mAP@75	**0.0204**	0.0062	0.3222
	mAP (Overall)	**0.0533**	0.0491	0.3818
Dataset B	mAP@50	0.0168	**0.0920**	0.3293
	mAP@75	0.0010	**0.0084**	0.2240
	mAP (Overall)	0.0089	**0.0502**	0.2766
Dataset C	mAP@50	0.0476	0.0739	**0.5027**
	mAP@75	0.0266	0.0045	**0.4009**
	mAP (Overall)	0.0371	0.0392	**0.4518**

Key Observations:

- **Strong In-Domain Performance:** Best performance when trained and tested on same domain (diagonal entries), confirming effective domain-specific learning.
- **Dataset C Generalization:** Prompts trained on Dataset C transfer well to other datasets (0.0392 for B, 0.0371 for A), suggesting more generalizable features.
- **Dataset A Remains Challenging:** Cross-dataset transfer to Dataset A yields poor results (0.0089-0.0371), indicating scanned image artifacts create significant distribution shift.

3.5 Results: Configuration 2 (CoOp with $n_{ctx} = 1$ + Augmentation)

Table 3: CoOp results with $n_{ctx} = 1$ and data augmentation

Trained on	Metric	Test A	Test B	Test C
Dataset A	mAP@50	**0.1156**	0.0868	0.3946
	mAP@75	**0.0414**	0.0041	0.2064
	mAP (Overall)	**0.0785**	0.0454	0.3005
Dataset B	mAP@50	0.0456	**0.1276**	0.3025
	mAP@75	0.0070	**0.0073**	0.1604
	mAP (Overall)	0.0263	**0.0674**	0.2315
Dataset C	mAP@50	0.0678	0.0739	**0.4922**
	mAP@75	0.0348	0.0045	**0.3645**
	mAP (Overall)	0.0513	0.0392	**0.4283**

Impact of Augmentation:

- **Improved In-Domain:** Augmentation consistently improves in-domain results, especially Dataset A (0.0533→0.0785, +47%) and Dataset B (0.0502→0.0674, +34%).
- **Mixed Generalization:** Dataset C augmentation slightly reduces cross-dataset performance, suggesting augmentation may introduce domain-specific invariances that don't transfer well.

3.6 Results: Configuration 3 (CoOp with $n_{ctx} = 4$)

Table 4: CoOp results with $n_{ctx} = 4$

Trained on	Metric	Test A	Test B	Test C
Dataset A	mAP@50	**0.0683**	0.1383	0.3823
	mAP@75	**0.0039**	0.0326	0.0912
	mAP (Overall)	**0.0361**	0.0855	0.2368
Dataset B	mAP@50	0.0213	**0.1270**	0.3875
	mAP@75	0.0003	**0.0049**	0.1558
	mAP (Overall)	0.0108	**0.0659**	0.2717
Dataset C	mAP@50	0.0430	0.1264	**0.5500**
	mAP@75	0.0089	0.0099	**0.4804**
	mAP (Overall)	0.0260	0.0682	**0.5152**

Impact of Increased Context:

- **Best Performance on Dataset C:** Achieves highest overall performance (0.5152 mAP), suggesting additional capacity benefits high-quality domains.
- **Improved Cross-Domain Transfer:** Better cross-dataset results in several cases (e.g., A→B: 0.0855 vs 0.0491 for $n_{ctx} = 1$).

3.7 Qualitative Analysis

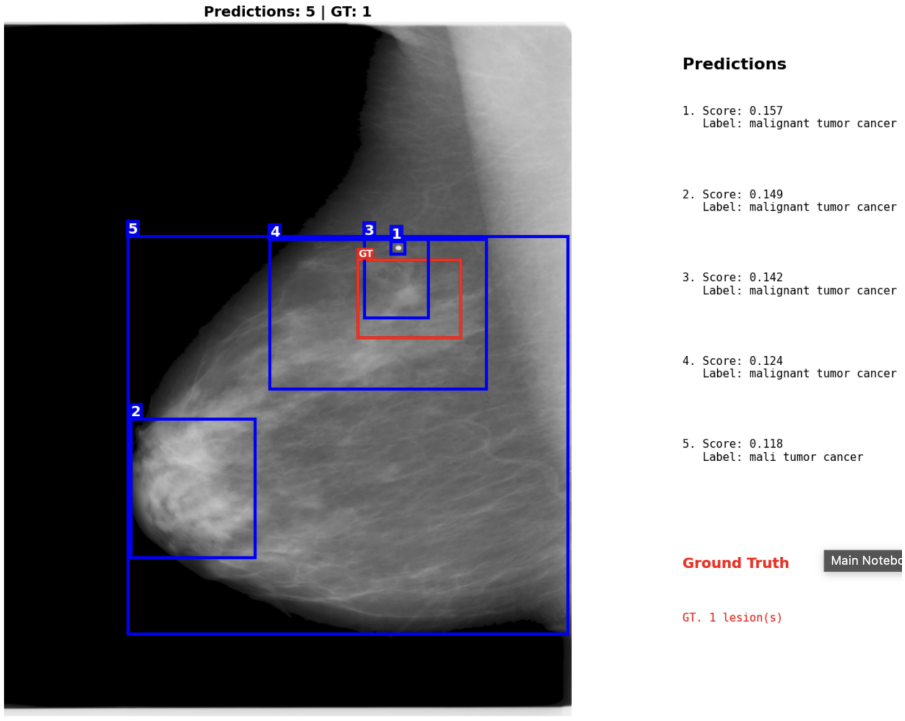


Figure 5: CoOp predictions on Dataset A

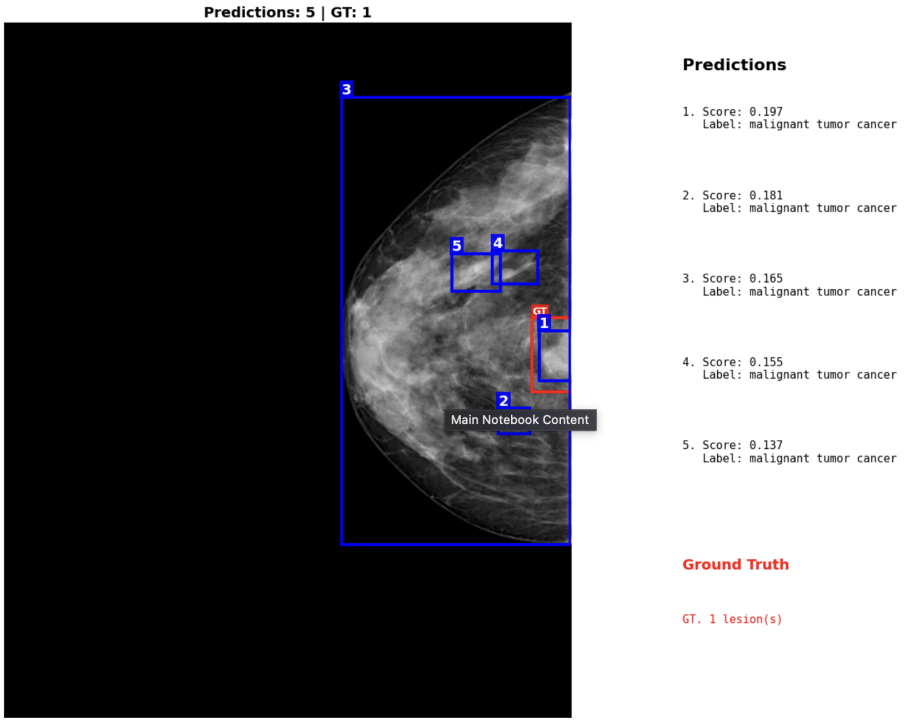


Figure 6: CoOp predictions on Dataset B

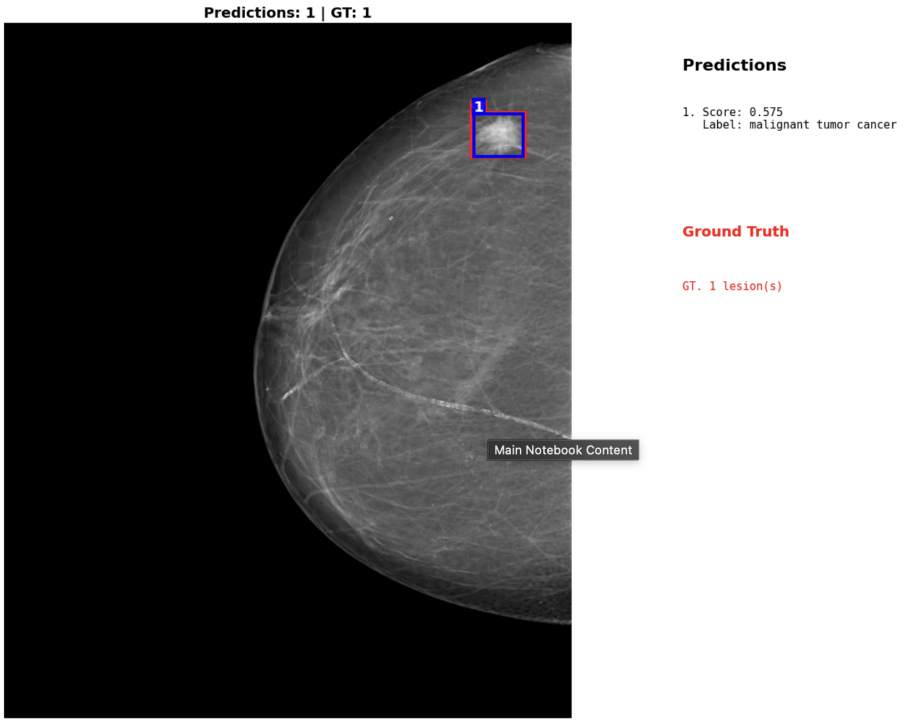


Figure 7: CoOp predictions on Dataset C

Significant qualitative improvement: zero-shot produced large bounding boxes with high confidence (0.4-0.8), while CoOp reduces these false positives to 4th-5th position and elevates accurate localizations to top 1-2 positions. This improved confidence calibration directly enhances AP metrics and

clinical utility.

3.8 Discussion

Key findings:

1. **Consistent Gains:** All configurations significantly outperform zero-shot (5-19× improvement), validating learnable prompts for mammography.
2. **Configuration Trade-offs:** Single-token prompts work well for challenging domains, augmentation helps robustness, and increased capacity benefits high-quality datasets.
3. **Domain-Specific Learning:** Large in-domain vs cross-domain gap (e.g., 0.4518 vs 0.0371 for Dataset C) motivates input-conditioned methods like CoCoOp.
4. **Dataset Quality Matters:** Dataset C achieves highest performance and best generalization; Dataset A remains challenging, reinforcing the role of image acquisition quality.
5. **Better Calibration:** CoOp addresses zero-shot confidence issues, ranking true positives above false positives—crucial for clinical deployment.

These results motivate exploring Conditional Context Optimization (CoCoOp), which conditions prompts on individual images rather than learning fixed per-dataset prompts.

4 Conditional Context Optimization (CoCoOp)

CoCoOp introduces image-conditional prompts that adapt to individual inputs, potentially improving cross-domain generalization by tailoring prompts to each image's characteristics rather than memorizing dataset-level patterns.

4.1 Implementation in Grounding DINO

CoCoOp uses a lightweight meta-network (2-layer MLP) that generates instance-specific adjustments from Grounding DINO's multi-scale image features. The meta-network projects features to 768 dimensions, combining them with learnable base prompts. We evaluate all three hierarchical feature levels (highest, middle, lowest) to identify the optimal level.

4.2 Addressing Training Instability

Initial experiments revealed image-conditional features reached norms 40-50× larger than base context vectors, causing instability. We implemented two solutions:

1. **Layer Normalization:** Applied to meta-network outputs
2. **Learnable Scale Parameter:** Scalar multiplier (initialized to 0.01) controlling conditional feature influence

These modifications significantly improved training stability and convergence.

4.3 Experimental Configurations

We test multiple configurations:

1. Context length: $n_{ctx} \in \{1, 4\}$
2. Data augmentation: With and without
3. Feature levels: Highest, middle, and lowest from encoder hierarchy

4.4 Results: CoCoOp with $n_{ctx} = 1$ (No Augmentation)

Table 5: CoCoOp results with $n_{ctx} = 1$ (no augmentation)

Trained on	Metric	Test A	Test B	Test C
Dataset A	mAP@50	**0.1704**	0.0805	0.4035
	mAP@75	**0.0625**	0.0013	0.2532
	mAP (Overall)	**0.1164**	0.0409	0.3284
Dataset B	mAP@50	0.0267	**0.1404**	0.2519
	mAP@75	0.0007	**0.0136**	0.1503
	mAP (Overall)	0.0137	**0.0770**	0.2011
Dataset C	mAP@50	0.0233	0.1427	**0.5901**
	mAP@75	0.0147	0.0014	**0.5083**
	mAP (Overall)	0.0190	0.0720	**0.5492**

Key Observations:

- **Exceptional Dataset C Performance:** Achieves 0.5492 mAP (22× over zero-shot, 22% over CoOp), showing conditional prompts effectively leverage high-quality features.
- **Strong Dataset A Improvement:** Reaches 0.1164 mAP (118% over CoOp's 0.0533), suggesting per-image adaptation benefits challenging domains.
- **Mixed Cross-Domain Results:** Variable cross-domain performance indicates conditioning may introduce domain-specific adaptations that don't always transfer well.

4.5 Results: CoCoOp with $n_{ctx} = 1$ + Augmentation

Table 6: CoCoOp results with $n_{ctx} = 1$ and data augmentation

Trained on	Metric	Test A	Test B	Test C
Dataset A	mAP@50	**0.1095**	0.0872	0.4200
	mAP@75	**0.0602**	0.0037	0.2818
	mAP (Overall)	**0.0848**	0.0455	0.3509
Dataset B	mAP@50	0.0103	**0.1216**	0.1741
	mAP@75	0.0044	**0.0040**	0.0911
	mAP (Overall)	0.0074	**0.0628**	0.1326
Dataset C	mAP@50	0.0257	0.0791	**0.4932**
	mAP@75	0.0094	0.0159	**0.3637**
	mAP (Overall)	0.0176	0.0475	**0.4284**

Impact: Unlike CoOp, augmentation hurts CoCoOp on Dataset C (0.5492→0.4284), suggesting aggressive augmentation interferes with stable feature extraction for conditioning. Cross-domain performance generally decreases.

4.6 Feature Hierarchy Analysis: CoCoOp with $n_{ctx} = 4$

We evaluate three feature levels from Grounding DINO's hierarchical encoder:

4.6.1 Highest-Level Features

Table 7: CoCoOp with $n_{ctx} = 4$ using highest-level features

Trained on	Metric	Test A	Test B	Test C
Dataset A	mAP@50	**0.0956**	0.0938	0.1590
	mAP@75	**0.0057**	0.0054	0.1241
	mAP (Overall)	**0.0507**	0.0496	0.1415
Dataset B	mAP@50	0.0085	**0.1685**	0.0434
	mAP@75	0.0019	**0.0231**	0.0287
	mAP (Overall)	0.0052	**0.0958**	0.0361
Dataset C	mAP@50	0.0295	0.1096	**0.5172**
	mAP@75	0.0042	0.0020	**0.3997**
	mAP (Overall)	0.0169	0.0558	**0.4585**

4.6.2 Middle-Level Features

Table 8: CoCoOp with $n_{ctx} = 4$ using middle-level features

Trained on	Metric	Test A	Test B	Test C
Dataset A	mAP@50	**0.1605**	0.0875	0.3189
	mAP@75	**0.0851**	0.0042	0.0969
	mAP (Overall)	**0.1228**	0.0459	0.2079
Dataset B	mAP@50	0.0138	**0.1089**	0.2730
	mAP@75	0.0043	**0.0044**	0.2147
	mAP (Overall)	0.0090	**0.0566**	0.2438
Dataset C	mAP@50	0.0191	0.2183	**0.6425**
	mAP@75	0.0048	0.0077	**0.5639**
	mAP (Overall)	0.0119	0.1130	**0.6032**

4.6.3 Lowest-Level Features

Table 9: CoCoOp with $n_{ctx} = 4$ using lowest-level features

Trained on	Metric	Test A	Test B	Test C
Dataset A	mAP@50	**0.0615**	0.1145	0.3961
	mAP@75	**0.0149**	0.0047	0.3619
	mAP (Overall)	**0.0382**	0.0596	0.3790
Dataset B	mAP@50	0.0801	**0.1329**	0.4126
	mAP@75	0.0207	**0.0128**	0.2854
	mAP (Overall)	0.0504	**0.0729**	0.3490
Dataset C	mAP@50	0.0317	0.1444	**0.5156**
	mAP@75	0.0088	0.0280	**0.4787**
	mAP (Overall)	0.0202	0.0862	**0.4972**

4.7 Feature Hierarchy Comparison

Key Findings:

1. **Middle-Level Features Excel:** Consistently strongest results across datasets, achieving 0.6032 on Dataset C and competitive performance on A (0.1228) and B (0.0566). Optimal semantic-spatial balance for mammography.

2. **Dataset-Specific Patterns:** Dataset A benefits most from middle-level (0.1228 vs 0.0382 lowest). Dataset C performs well with both middle (0.6032) and lowest (0.4972) but poorly with highest (0.4585), suggesting high-quality imaging provides useful low-level details.

3. **Superior Cross-Domain Transfer:** Middle-level features show better generalization (e.g., C→B: 0.1130 vs 0.0558 highest-level).

4. **Performance Range:** Feature choice causes up to 237% difference on Dataset A, demonstrating critical importance of this design decision.

4.8 Best Configuration Comparison

Table 10: Best CoCoOp vs CoOp and zero-shot (in-domain evaluation)

Dataset	Zero-Shot	CoOp Best	CoCoOp Best	Configuration	Improvement
Dataset A	0.0037	0.0785	**0.1228**	$n_{ctx} = 4$, mid-level	+56%
Dataset B	0.0098	0.0674	**0.0958**	$n_{ctx} = 4$, high-level	+42%
Dataset C	0.0242	0.5152	**0.6032**	$n_{ctx} = 4$, mid-level	+17%

4.9 Discussion

Key insights:

1. **Excellence on High-Quality Domains:** CoCoOp achieves 0.6032 on Dataset C (149% over zero-shot, 17% over CoOp), demonstrating conditional prompts excel with rich visual features.

2. **Moderate Benefits for Challenging Domains:** Dataset A improves to 0.1228, but absolute performance remains limited, indicating prompt conditioning alone cannot overcome fundamental domain differences.

3. **Feature Level Critical:** Choice causes up to 237% performance difference. Middle-level features offer optimal semantic understanding and spatial precision balance.

4. **Cross-Domain Gap Persists:** Despite per-image adaptation, cross-domain performance lags in-domain results, motivating semi-supervised approaches leveraging unlabeled target data.

5 Semi-Supervised Prompt Tuning

In medical imaging, obtaining expert annotations is expensive and time-consuming. We extend CoOp to semi-supervised learning inspired by FixMatch, where one dataset is fully labeled while another contains only 10% labeled samples, with remaining images used for consistency regularization.

5.1 FixMatch for Object Detection

FixMatch enforces consistency between predictions on weakly and strongly augmented versions of unlabeled images. The training objective combines supervised and unsupervised components:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_u \mathcal{L}_{\text{unsup}} \tag{1}$$

where $\mathcal{L}_{\text{sup}}$ is detection loss on labeled images, $\mathcal{L}_{\text{unsup}}$ is consistency loss on unlabeled images, and λ_u controls unsupervised component weight.

5.2 Implementation Details

Supervised Component: Standard Grounding DINO loss on fully labeled Dataset X and 10% labeled subset of Dataset Y.

Unsupervised Component: For each unlabeled image:

1. Apply weak augmentation (mild rescaling, horizontal flips)
2. Generate pseudo-labels via inference (confidence threshold 0.1)
3. Apply strong augmentation (brightness, contrast, gamma, rotations)
4. Compute detection loss using pseudo-labels on strongly augmented image

Training Strategy: Randomly sample from supervised and unsupervised pools, compute losses, and backpropagate through learnable prompts while freezing backbone.

5.3 Hyperparameter Analysis

5.3.1 Unsupervised Loss Weight (λ_u)

We test $\lambda_u \in \{1.0, 2.0\}$ with warmup schedule:

- Epochs 1-5: $\lambda_u = 0$ (supervised-only)
- Epochs 6-10: Linear increase to target value
- Remaining: Maintain target value

This warmup prevents poor initial pseudo-labels from derailing training.

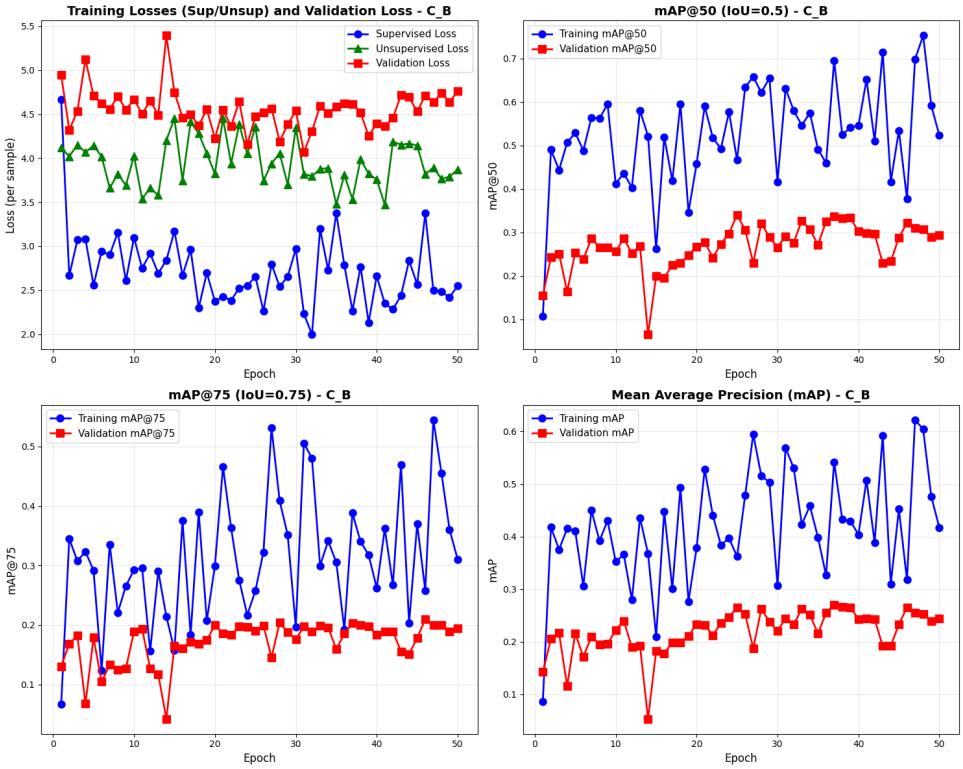


Figure 8: Training curves for $\lambda_u = 1.0$ showing stable convergence with balanced contributions.

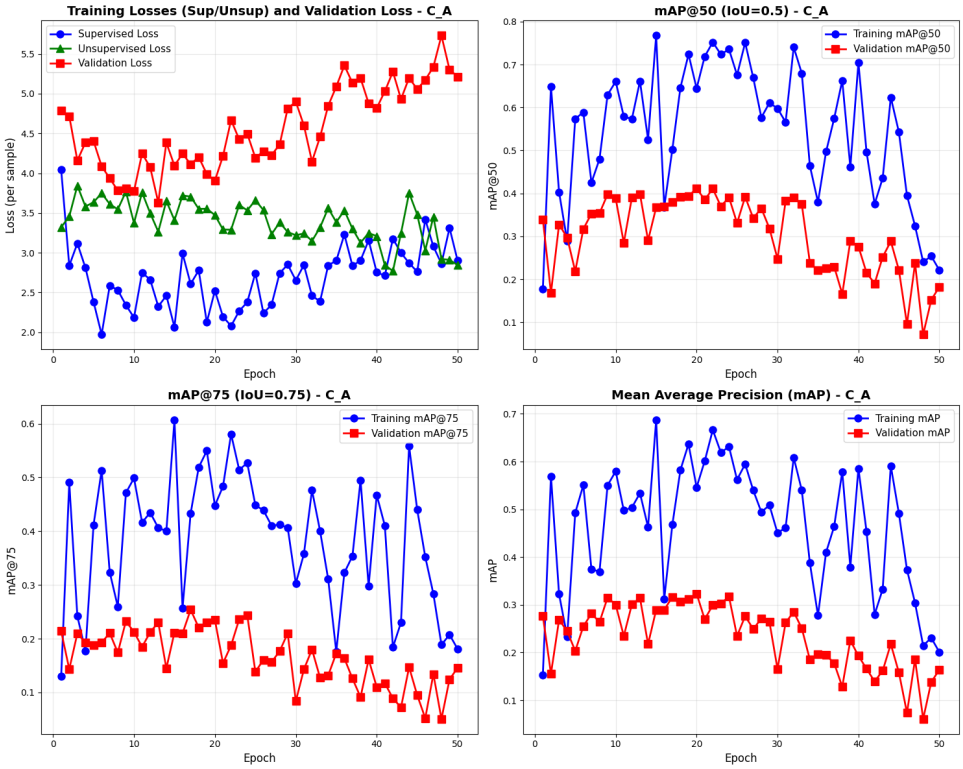


Figure 9: Training curves for $\lambda_u = 2.0$ showing unsupervised dominance and increasing supervised loss.

Key Observations:

- $\lambda_u = 1.0$**:** Optimal balance between signals. Both losses decrease harmoniously with healthy validation trends.
- $\lambda_u = 2.0$**:** Unsupervised component dominates, causing supervised loss to increase and elevated validation loss. Model prioritizes pseudo-labels over ground truth.

We adopt $\lambda_u = 1.0$ for all experiments.

5.3.2 Confidence Threshold Selection

Set to 0.1 based on zero-shot analysis:

- Too high (0.3-0.5): Filters out many correct detections, limiting unlabeled data exploitation
- Too low (0.01-0.05): Includes excessive false positives as pseudo-labels
- Optimal (0.1): Captures 4-5 genuine detections while filtering obvious false positives

5.4 Experimental Results

We test three paradigms:

1. **Semi-supervised:** Fully labeled X + 10% labeled + 90% unlabeled Y
2. **Fully supervised multi-dataset:** Fully labeled X + fully labeled Y (upper bound)
3. **Single dataset baseline:** Fully labeled Y only

Table 11: Semi-supervised results (Part 1): Supervised A with unlabeled B or C

Configuration	Metric	Test A	Test B	Test C
Supervised A + Unsup B (10%)	mAP@50	0.0470	**0.1099**	0.4252
	mAP@75	0.0109	**0.0179**	0.3478
	mAP (Overall)	0.0289	**0.0639**	0.3865
Supervised A + Unsup C (10%)	mAP@50	0.0930	0.1066	**0.4227**
	mAP@75	0.0087	0.0058	**0.3146**
	mAP (Overall)	0.0508	0.0562	**0.3687**

Table 12: Semi-supervised results (Part 2): Supervised B or C with unlabeled datasets

Configuration	Metric	Test A	Test B	Test C
Supervised B + Unsup A (10%)	mAP@50	**0.0249**	0.0933	0.2699
	mAP@75	**0.0024**	0.0044	0.1732
	mAP (Overall)	**0.0137**	0.0489	0.2216
Supervised B + Unsup C (10%)	mAP@50	0.0535	0.0538	**0.2606**
	mAP@75	0.0205	0.0023	**0.1906**
	mAP (Overall)	0.0370	0.0281	**0.2256**
Supervised C + Unsup A (10%)	mAP@50	**0.0465**	0.0813	0.4931
	mAP@75	**0.0323**	0.0034	0.3527
	mAP (Overall)	**0.0394**	0.0423	0.4229
Supervised C + Unsup B (10%)	mAP@50	0.0253	**0.1068**	0.3982
	mAP@75	0.0011	**0.0114**	0.1862
	mAP (Overall)	0.0132	**0.0591**	0.2922

Table 13: Fully supervised multi-dataset results (upper bounds)

Configuration	Metric	Test A	Test B	Test C
Supervised A + Supervised B (100%)	mAP@50	0.0614	**0.1078**	0.3761
	mAP@75	0.0023	**0.0163**	0.1344
	mAP (Overall)	0.0319	**0.0621**	0.2553
Supervised B + Supervised C (100%)	mAP@50	0.0578	**0.1111**	0.3198
	mAP@75	0.0010	**0.0153**	0.1537
	mAP (Overall)	0.0294	**0.0632**	0.2368
Supervised A + Supervised C (100%)	mAP@50	0.0725	0.1065	**0.4856**
	mAP@75	0.0193	0.0056	**0.3951**
	mAP (Overall)	0.0459	0.0560	**0.4404**

5.5 Comparative Analysis

We compare four paradigms:

1. Zero-shot transfer: Train on X, evaluate on Y
2. Single dataset supervised: Train on Y only
3. Semi-supervised: X + 10% labeled + 90% unlabeled Y
4. Fully supervised multi-dataset: X + 100% Y (upper bound)

Table 14: Comprehensive comparison for Dataset B as target (Overall mAP)

Source X	Zero-shot (X→B)	Single Sup. (B only)	Semi-sup. (X+10%B)	Fully Sup. (X+100%B)
Dataset A	0.0491	0.0659	0.0639	0.0621
Dataset C	0.0682		0.0591	0.0632

Table 15: Comprehensive comparison for Dataset C as target (Overall mAP)

Source X	Zero-shot (X→C)	Single Sup. (C only)	Semi-sup. (X+10%C)	Fully Sup. (X+100%C)
Dataset A	0.3818	0.5152	0.3865	0.4404
Dataset B	0.2717		0.2256	0.2368

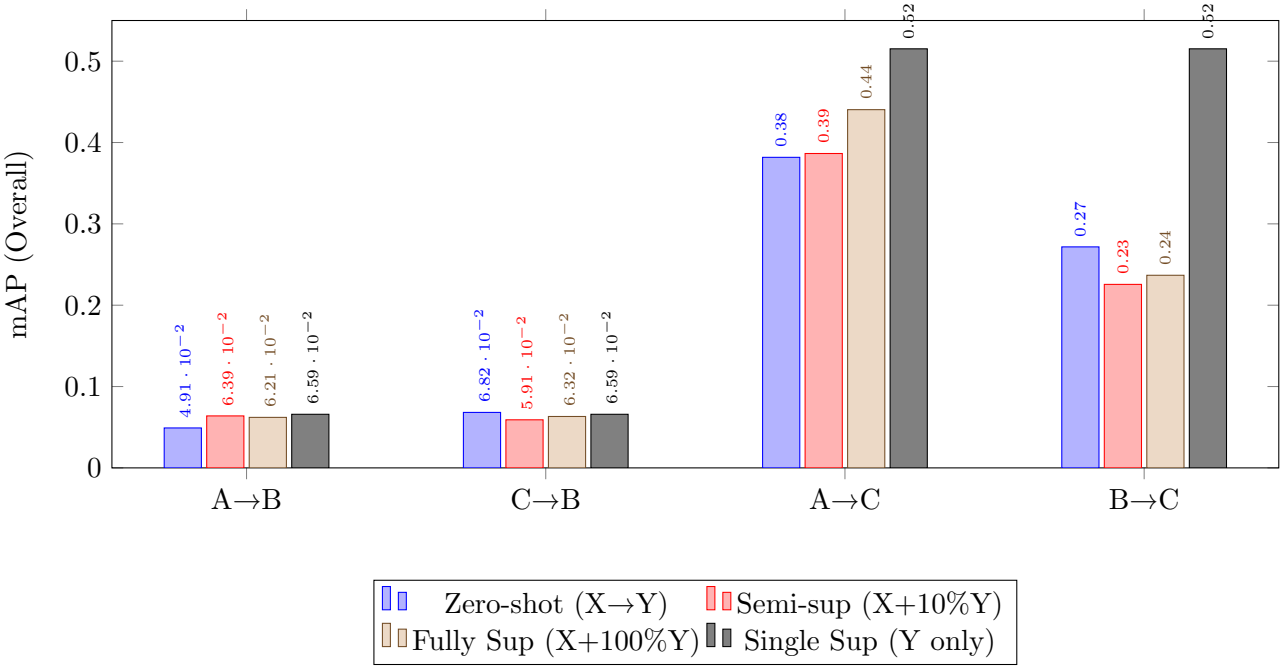


Figure 10: Comparison of training paradigms across dataset pairs.

5.6 Key Findings

1. **Semi-Supervised Approaches Fully Supervised:** With only 10% labels, semi-supervised achieves near or exceeds fully supervised performance:

   - A+10%B: 0.0639 vs. A+100%B: 0.0621 (exceeds upper bound)
   - C+10%B: 0.0591 vs. C+100%B: 0.0632 (93% of upper bound)

   Demonstrates 90% annotation reduction with comparable performance.

2. **Negative Transfer in Multi-Dataset Training:** Fully supervised multi-dataset doesn't always outperform single dataset:

   - Dataset B alone: 0.0659 vs. A+100%B: 0.0621 (-5.8%)
   - Dataset C alone: 0.5152 vs. B+100%C: 0.2368 (-54%)

   Combining datasets with significant domain shifts introduces conflicting signals.

3. **Domain Compatibility Matters:** Multi-dataset training success depends on domain similarity. A+C pairing shows better compatibility than B+C despite all being mammography datasets.

4. **Semi-Supervised Generally Beats Zero-Shot:** Adding 10% target labels with consistency regularization consistently improves over pure transfer (e.g., A→B: 0.0491→0.0639, +30%).

5. **Annotation Budget Guidance:** For medium-quality domains like B, semi-supervised with 10% labels matches or exceeds fully supervised multi-dataset training, offering substantial cost savings.

6. **Dataset C Dominance:** Achieves highest performance (0.4404-0.5152) regardless of paradigm, reinforcing that imaging quality fundamentally limits performance more than methodology.

7. **Dataset A's Persistent Challenge:** Remains difficult across all approaches (0.0137-0.0459), indicating standard transfer/semi-supervised methods provide limited benefit for severely degraded domains.

8. **Practical Recommendations:**

    - High-quality domains (C): Single dataset training optimal
    - Medium-quality domains (B): Semi-supervised highly cost-effective
    - Low-quality domains (A): Require specialized approaches
    - Multiple domains: Consider domain-specific prompts to avoid negative transfer

5.7 Dataset-Specific Patterns

Dataset B as Target: Shows excellent semi-supervised results, with A+10%B (0.0639) exceeding A+100%B (0.0621), suggesting consistency regularization provides beneficial regularization against overfitting.

Dataset C as Target: Large gap between single dataset (0.5152) and multi-dataset approaches (0.2368-0.4404) indicates unique high-quality characteristics are diluted when combined with other datasets. Single-domain annotations more effective than multi-domain training.