CANDIDATE-BOX SELECTION MATTERS FOR CROSS-DOMAIN FEW-SHOT OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-domain few-shot object detection (CD-FSOD) aims to transfer knowledge from abundant source-domain pretraining to an unseen target domain with only a few labeled samples, while overcoming domain shift and limited target supervision. Grounding DINO has become a mainstream baseline, which selects a fixed number of high-confidence candidate boxes by semantic matching before decoder refinement. In this paper, unlike current works, we identify the selected candidate boxes as an underexplored bottleneck: under domain shift, high-quality boxes may be ranked below the selection budget and discarded, while top-ranked boxes may fail to enclose objects. By comparing in-domain and cross-domain cases, we find that this issue is more severe in cross-domain settings and trace it to vision-language misalignment in candidate selection rather than candidate generation. Motivated by this finding, we propose Support-Conditioned Local Evidence (SCLE), a lightweight and plug-and-play refinement of candidate selection. SCLE constructs foreground and background prototypes from few support samples, builds a local evidence map, and evaluates candidates by in-box purity and inside-outside contrast. A fusion-quota strategy then combines semantically ranked and local-evidence-ranked candidates to better match the target-domain distribution without modifying the detector architecture, adding learnable parameters, or using external data. SCLE can be plugged into different Grounding-DINO-based methods. Extensive experiments validate its effectiveness and demonstrate consistent improvements across state-of-the-art CD-FSOD approaches.

1 INTRODUCTION

Cross-domain few-shot object detection (CD-FSOD) seeks to overcome the constraints of domain shift and scarce target annotation. It transfers the knowledge acquired from abundant source-domain pretraining to an unfamiliar target domain that provides only a few labeled samples, and reliably detect objects in these domains. Its central difficulty therefore comes from both the feature discrepancy between domains and the limited supervision in the target domain.

Currently, Grounding DINO (Liu et al., 2023) has become the mainstream baseline for its strong performance, where the model follows a fixed pipeline (Fig. 1a): given an image and a text prompt, it produces dense visual features, and the detection head outputs a large number of candidate boxes together with their semantic matching scores; before the candidates reach the decoder for iterative refinement, the model selects a fixed number of high-confidence candidates by semantic score and uses them as the initial boxes. Recent works optimize this framework from many aspects, including refined fine-tuning and decoding strategies, multi-modal prototypes, and target-domain data augmentation with retrieval or diffusion models (Yu et al., 2026; Li et al., 2025; Zhuang et al., 2026; Wang et al., 2026). However, in this paper, we find another aspect that largely degrades performance: the selected candidate boxes. As shown in Fig. 1b, some of the candidates that rank first fail to enclose the object, while a box that better covers it is ranked below the selection budget and never reaches the decoder. Since the selected candidate boxes are the basis of the subsequent object detection, this problem would largely influence the performance, but is seldom explored so far.

To handle this problem, we first compare the selected candidate boxes between the in-domain data and the cross-domain data, and find that this problem is more severe in the cross-domain cases than in-domain ones, indicating the domain gap may be the cause. Then, we look into both the candidate-box generation and the candidate-box selection phase to see which process is harmed by the domain

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

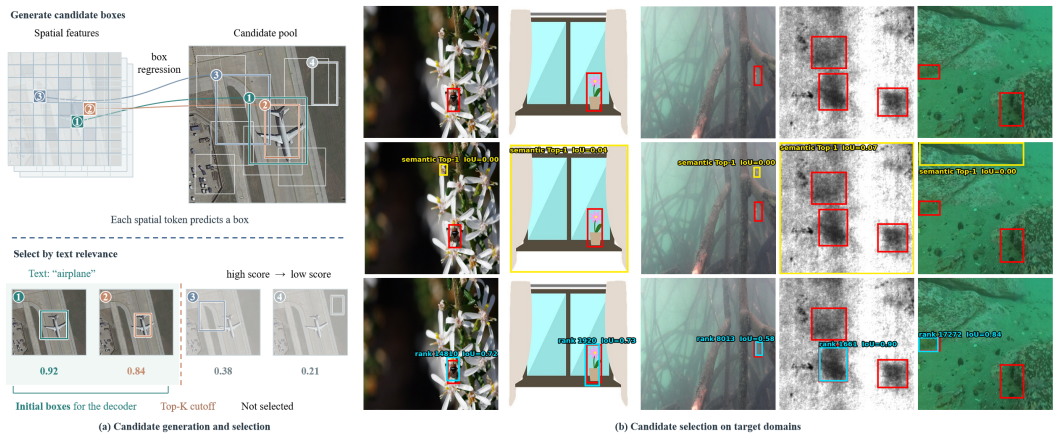


Figure 1: (a) A brief illustration of the current Grounding-DINO-based works. (b) We find that compared with the ground-truth bounding box (red), for target domains, the top-ranked candidate boxes (yellow) are less likely to tightly bound the objects, where the better candidate boxes (cyan) may be ranked outside the Top-900 quota. Unlike current work, we identify the selected candidate boxes as another bottleneck for the CD-FSOD task, which we aim to address in this paper.

gap, and find that the model can itself generate candidate boxes no worse than the in-domain cases, but the selection phase is more influenced. As the selection is performed by comparing the semantic descriptions and each local visual token, since current work (Zhao et al., 2026) has pointed out that the alignment between text and local visual tokens is more influenced by domain gaps, we conclude that it is the vision-language misalignment that leads to the unsatisfactory selected candidate boxes.

Guided by this, as the vision-language alignment may not be reliable for the CD-FSOD task, to address this issue, we look back at the vision features from the support and query samples themselves. As the support and query samples are from the same domain, their visual features are inherently better aligned than the vision-language features, which can be used for better candidate-box selection. Based on this intuition, we propose SCLE (Support-Conditioned Local Evidence), which introduces target-domain local visual evidence to correct biases in semantic ranking without modifying the detector architecture or adding learnable parameters. It constructs foreground and background prototypes from few support samples to generate a local evidence map. Candidate boxes are evaluated using in-box purity and inside-outside contrast. A fusion-quota strategy combines semantically ranked candidates and local-evidence candidates to form a candidate set adapted to the target-domain distribution, mitigating the loss of high-quality candidates under cross-domain settings.

In summary, our contributions can be listed as follows.

- We find that the selected candidate boxes inherited by Grounding DINO from its source-domain pretraining constitute another underexplored bottleneck for the CD-FSOD task. By experiments, we find that the problem lies in the vision-language misalignment caused by the domain gap.
- To handle this, we propose SCLE, a lightweight and plug-and-play refinement of candidate selection. It brings local visual evidence into the candidate ranking and repairs the selection bias caused by the vision-language misalignment. It needs no additional parameters, no structural modification, and no external data, and can be plugged into different Grounding-DINO-based methods.
- Comprehensive experiments validate the effectiveness of SCLE, demonstrating consistent improvements across different state-of-the-art works.

2 Related Work

Cross-domain few-shot object detection. CD-FSOD transfers a source-pretrained detector to a target domain with distinct appearance, background and imaging conditions, using only a small number of annotated samples per class (Xiong, 2023). Existing work falls into three branches. The first optimizes feature representations to mitigate foreground and background feature shift under target-domain astigmatism (Jiang et al., 2026). The second refines training and decoding strategies by investigating fine-tuning schemes and adopting parallel decoders instead of serial ones (Yu et al.,

2026). The third augments training data via retrieval-guided compositional synthesis or diffusion-based generation (Li et al., 2025; Zhuang et al., 2026). Multimodal prototypes are also adopted for category transfer (Wang et al., 2026). For the related cross-domain few-shot learning task, transfer failure has been analyzed via attention collapse, CLIP tail alignment and CLS-token properties (Yi et al., 2026a;b; Zou et al., 2024). All these methods improve detector computation over fixed candidate sets while keeping the original candidate-selection rule for decoder input.

Candidate selection in language-guided detectors. GLIP and Grounding DINO fuse visual and language representations, enabling open-set recognition and localization with text prompts instead of fixed categories. Their transformer pipeline adopts multi-scale encoder proposals from Deformable DETR and DINO (Li et al., 2022; Liu et al., 2023; Zhu et al., 2021; Zhang et al., 2022). Within these detectors, the text score jointly predicts categories and decides which candidates enter the decoder; the former is the target design, while the latter is inherited from prior interfaces. The candidate-selection rule has stayed consistent across this line of work. DETR uses content-agnostic learned queries without encoder-side candidate selection (Carion et al., 2020). Deformable DETR selects top-scoring encoder features as initial boxes for faster convergence (Zhu et al., 2021). DINO introduces mixed query selection, where query positions are derived from score-ranked encoder features (Zhang et al., 2022), and Grounding DINO maintains this pipeline by keeping the top-K candidates ranked by text score (Liu et al., 2023). A shared trait is that selection relies on classification confidence rather than localization quality. Such confidence correlates well with localization under the pretraining distribution, yet this correlation originates from data instead of model architecture. We investigate whether this holds for CD-FSOD and redesign the selection stage accordingly.

3 PRELIMINARIES

The cross-domain few-shot object detection task. Cross-domain few-shot object detection (CD-FSOD) (Xiong, 2023) transfers a source-domain pretrained detector to a target domain whose appearance, background, and imaging conditions differ clearly from the source, using a few labeled samples per category. In the N-shot setting, each target category provides N labeled samples for fine-tuning, and evaluation is performed on the full target test set; we adopt the 1/5/10-shot configuration in this paper.

Baseline: Grounding DINO. Our baseline is Grounding DINO, which takes an image and a textual category description as input and outputs the corresponding boxes. The detection process is a three-stage mapping,

$$\{(B_i, C_i)\}_{i=1}^{N} = H(E(X)), \qquad \mathcal{I}_K = \mathcal{S}_K(C_1, \ldots, C_N), \qquad Y = D(\{B_i\}_{i\in\mathcal{I}_K}), \tag{1}$$

where the encoder E outputs multi-scale dense features, the detection head H regresses an initial box for each token and computes its text score, $\mathcal{S}_K$ is the pre-decoder selection rule that picks K candidates, and the decoder D iteratively refines the initial boxes into the final predictions.

We improve only the selection module $\mathcal{S}_K$. The baseline ranks candidates by text score alone:

$$s_i^{\text{sem}} = \max_t C_{i,t}, \qquad \mathcal{I}_K = \text{TopK}_i(s_i^{\text{sem}}). \tag{2}$$

Here $C_{i,t}$ is the score that token i gives at the t-th text position. The box coordinates B_i do not appear in Eq. 2: a candidate is discarded once its text score falls outside the top K, even if it covers the target well. Since the decoder refines only the selected initial boxes, a discarded candidate cannot take part in the subsequent localization; the decoder then refines them over several layers into the final predictions.

4 ANALYSIS

4.1 DIAGNOSTIC PROTOCOL

Existing methods on this framework improve what the detector computes from the candidates they are given, but they do not examine how a domain shift affects the candidates themselves. We measure that effect here. Rather than examining the shape of the top-ranked proposals, we check one by one whether each ground-truth box is retained. At an IoU threshold θ, a ground-truth box counts as covered if and only if the candidate set contains a candidate whose IoU with it is at least θ. For a given candidate-selection rule S, we define

$$\text{cov}(S) = \frac{1}{|X|}\sum_{x\in X}\frac{1}{|g_x|}\sum_{g\in g_x} \mathbb{1}\left[\max_{p\in S(x)} \text{IoU}(p, g) \geq \theta\right], \qquad \theta = 0.75, \tag{3}$$

where X is the set of test images and g_x the ground-truth set of image x. The quantity first computes the fraction of covered ground-truth boxes within each image and then averages over images. Two settings of the rule S are of interest. Setting $S = \mathcal{I}_K$, the semantic top-K selection of Eq. 2, measures how many targets the candidates actually sent to the decoder can cover; we refer to this quantity as (a). Setting $S = \mathcal{P}$, all encoder candidates of the image, measures how many targets the detector has localized somewhere in its candidate pool; we refer to it as (b). The first is what the decoder receives, and the second is what it could have received. The threshold 0.75 requires a tight overlap between candidate and target, keeping the metric from being dominated by loose boxes.

4.2 THE COVERAGE GAP

Tab. 1 reports (a), (b), and their difference on the pretrained model, for the source domain and for typical target domains. From this table, we can have the following findings.

Table 1: Per-object coverage at IoU ≥ 0.75 on the pretrained model. (a) and (b) are the two quantities defined in Sec. 4; Abs. is their difference (b)−(a), and Rel. expresses that difference as a share of the attainable upper bound. Arrows mark the direction of interest.

	Size		Coverage ↑		Gap ↓	
Domain	Images	GT	(a) Sem.	(b) Pool	Abs.	Rel. (%)
Source domain						
COCO	5952	36334	0.874	0.890	0.016	1.8
Target domains						
NEU-DET	360	834	0.741	0.900	0.159	17.7
UODD	506	3218	0.525	0.615	0.090	14.7
DIOR	5000	28810	0.773	0.832	0.059	7.1
DeepFish	909	3029	0.803	0.833	0.030	3.6

(1) The problem in selected candidate boxes is more severe in target domains. We first compare the selected candidate boxes between in-domain and cross-domain cases. In Tab. 1, we can see that for the *(a) Sem* column, the source domain (COCO) shows, on average, higher coverage than the target domains, indicating better selected candidate boxes. Since the domain discrepancy leads to differences in the quality of selected boxes, we hold that the domain gap leads to the problem. Therefore, we then study in what aspect the domain gap influences the selected boxes.

(2) Compared with candidate-box generation, selection is more influenced by domain gaps. In Tab. 1, as the *(b) Pool* column means the coverage of all possible candidate boxes among the ground-truth boxes, it measures the upper limit of the model's ability to generate candidate boxes. Therefore, the difference between the *(a) Sem.* and the *(b) Pool* column measures the selection of candidate boxes from all possible candidate boxes. Although the upper limit of the model's ability is restricted by the complexity in each domain, we can see that target domains show much higher decreases in the selection of boxes. Therefore, we argue that the selection of candidate boxes is more influenced by domain gaps.

(3) The selection mechanism is misled by the vision-language misalignment. Currently, the selection of candidate boxes is mainly based on the similarity of the semantic descriptions and visual tokens, which is largely relevant to the vision-language alignment. However, some work (Zhao et al., 2026) has pointed out that this alignment, especially for the local alignment between each visual token and the language, is severely influenced by the domain gap. Therefore, we conclude that it is the misalignment between vision and language caused by the domain gap that leads to the unsatisfactory selection of candidate boxes.

In summary, to handle the bottleneck in candidate-box selection, we need to alleviate the vision-language misalignment caused by domain gaps.

5 METHOD

5.1 OVERVIEW

Section 4 identifies candidate selection, rather than candidate pool generation, as the cross-domain detection bottleneck stemming from domain-induced vision–language misalignment. We propose SCLE, which introduces a support-set local-evidence signal prior to decoding to reorder initial candidates. Instead of purely using text similarity, it selects candidates based on intra-box target evidence and background separability. SCLE only revises the construction rule of $\mathcal{S}_K$ in Eq. 1, with no extra parameters or architectural modifications. Prototypes are learned from fine-tuning support samples via gradient-free exponential moving average, frozen at inference to generate a dense evidence map. Candidates are scored by interior purity and inside-versus-ring contrast, and a quota mechanism aggregates three candidate groups for the decoder (Figure 2).

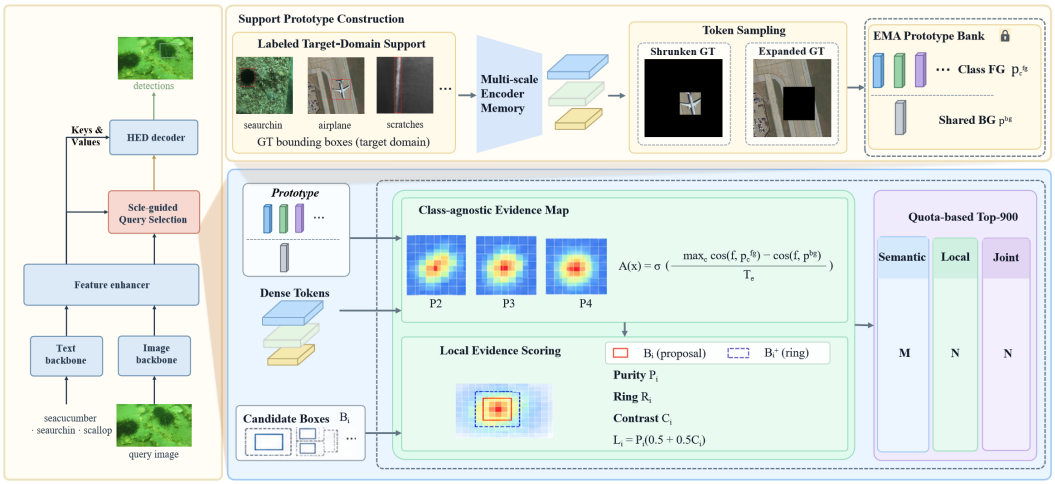


Figure 2: SCLE only modifies the pre-decoder selection stage of the language-guided detector. *Left*: detection pipeline, where SCLE-driven query selection replaces text-score ranking before the HED decoder. *Upper right*: class-wise foreground prototypes and one shared background prototype, updated via exponential moving average from labeled support samples. *Lower right*: the prototype bank produces a class-agnostic evidence map. Candidates are scored using interior purity P_i, ring evidence R_i and contrast C_i. A quota divides the Top-900 budget into semantic, local-evidence and joint channels. Only the indices fed into the decoder are changed.

5.2 Support-conditioned prototype bank

For a target-domain training image, let $f(x) \in \mathbb{R}^d$ denote the post-encoder feature of a valid multi-scale token at normalized coordinate x. We average features within foreground and background regions of support samples to obtain class-wise foreground prototypes and one shared background prototype:

$$\bar{p}_c^{\mathrm{fg}} = \frac{1}{|\Omega_c^{\mathrm{fg}}|} \sum_{x \in \Omega_c^{\mathrm{fg}}} f(x), \qquad \bar{p}^{\mathrm{bg}} = \frac{1}{|\Omega^{\mathrm{bg}}|} \sum_{x \in \Omega^{\mathrm{bg}}} f(x), \tag{4}$$

where Ω_c^{fg} contains positions inside shrunk ground-truth boxes of class c, and Ω^{bg} collects positions outside all expanded ground-truth boxes. Foreground boxes are scaled down by ρ, while background boxes are expanded by β and clipped to the image boundary. This design isolates object regions for foreground prototypes and avoids object contamination in background prototypes.

All prototypes are updated via normalized exponential moving average:

$$p \leftarrow \mathrm{norm}\big(\mu p + (1-\mu)\,\mathrm{norm}(\bar{p})\big). \tag{5}$$

where p is the accumulated prototype, $\mu \in [0, 1)$ the momentum, and $\mathrm{norm}(\cdot)$ L_2 normalization. Foreground prototypes are stored per class, and one background prototype is shared across all classes. The EMA stabilizes background representations, which would otherwise be biased by limited context in individual support batches. Prototypes are updated solely from labeled target-domain support samples, saved as buffers in the checkpoint, and fixed during inference. SCLE requires no extra annotations and has no learnable variables at test time.

5.3 Dense foreground evidence map

For the feature $f(x)$ of each token, we define the foreground evidence

$$A(x) = \sigma\left(\frac{\max_{c \in \mathcal{C}} \cos(f(x), p_c^{\mathrm{fg}}) - \cos(f(x), p^{\mathrm{bg}})}{T_e}\right), \tag{6}$$

where $\mathcal{C}$ denotes the set of target classes available in fine-tuning, $\cos(\cdot,\cdot)$ the cosine similarity, and T_e the evidence temperature. Three design points are noted. First, the maximum over classes yields a class-agnostic localization signal; category semantics come purely from the original language logits, and SCLE is not involved in classification. Second, subtracting the background similarity avoids rewarding object regions and domain context indiscriminately, suppressing tokens resembling the

common target-domain background. Third, the sigmoid σ restricts $A(x)$ to $[0, 1]$, matching the value range for subsequent pooling and contrast calculations. The two terms handle distinct failure cases: the foreground term supplements local target evidence missed by semantic ranking, while the background term suppresses distracting domain context. Independent of category semantics, both signals work alongside the original language classification branch without interference.

5.4 Proposal purity and inside-versus-ring contrast

The dense evidence map must be converted into scores for each candidate box. For candidate B_i, we pool evidence at its corresponding feature level:

$$P_i = \mathrm{RoIAlign}(A, B_i), \qquad O_i = \mathrm{RoIAlign}(A, B_i^+), \tag{7}$$

where B_i^+ denotes B_i expanded by factor r around its center and clipped to the image boundary. P_i is interior purity and O_i the average evidence inside the expanded box. Since B_i^+ consists of B_i and its surrounding ring, the ring evidence R_i is calculated by area subtraction:

$$R_i = \frac{|B_i^+|O_i - |B_i|P_i}{|B_i^+| - |B_i|}. \tag{8}$$

We define the contrast and local-evidence score as

$$C_i = \sigma\left(\frac{P_i - R_i}{T_c}\right), \qquad L_i = P_i(0.5 + 0.5C_i). \tag{9}$$

where σ is the sigmoid and T_c the contrast temperature. C_i distinguishes boxes lying inside an object from boxes enclosing the whole object. P_i alone cannot tell them apart: two boxes may share the same P_i with equal interior evidence density, even if one covers only part of the object and the other extends beyond it. C_i checks whether evidence decays at box boundaries. For an inner box, the ring still belongs to the object, so $R_i \approx P_i$ and the contrast is neutral; for a box wrapping the object, the ring falls on background with low R_i and high contrast; contrast approaches zero when the box interior is background while the ring contains the object.

A multiplicative formulation is adopted instead of addition. An additive $L_i = P_i + C_i$ may incorrectly rank background candidates with clean ring evidence above valid targets with slightly lower P_i. The product lets P_i dominate ranking while C_i only differentiates candidates of similar purity. The term $0.5 + 0.5C_i$ linearly maps $C_i \in [0, 1]$ to $[0.5, 1.0]$, limiting the maximum penalty to halving P_i. This soft penalty accounts for unreliable ring evidence under low resolution, occlusion and crowded objects; high P_i does not guarantee the box lies inside the object and should not be discarded outright.

5.5 Quota-based Top-K selection

Semantic and local-evidence scores differ in scale and error patterns. Instead of fusing raw values, we rank candidates within each image and normalize ranks to $[0, 1]$, with the best candidate as 1 and the worst as 0. Using $\hat{s}_i = \mathrm{rank}(s_i^{\mathrm{sem}})$ and $\hat{l}_i = \mathrm{rank}(L_i)$, the joint score reads

$$J_i = w\hat{s}_i + (1 - w)\hat{l}_i. \tag{10}$$

where $w \in [0, 1]$ is the fusion weight. Let K denote the total candidate budget for the decoder. The final candidate set comes from three disjoint channels with $\alpha_s K$, $\alpha_l K$ and $\alpha_j K$ slots, where $\alpha_s + \alpha_l + \alpha_j = 1$. For simplicity, we set $\alpha_l = \alpha_j$. The semantic channel sorted by s_i^{sem} preserves semantic recall; the local-evidence channel sorted by L_i recovers candidates discarded by semantic ranking; the joint channel sorted by J_i selects candidates agreed by both signals. We pick top candidates from each channel sequentially. Selected candidates are not re-counted, and the joint score fills remaining slots if the total number is less than K.

We adopt a quota mechanism rather than a single fused score because the vision-language misalignment varies drastically across domains. A unified fused score may suppress useful semantic candidates when misalignment is slight, yet lack enough capacity for local evidence on hard domains. The quota strategy allocates independent capacity to each signal instead of relying on one global weight for all domains.

Table 2: Main results (mAP) on the six CD-FSOD target domains under the 1-, 5-, and 10-shot protocols. † marks the baseline we re-trained under our own protocol. In each column the best entry of the shot block is bold, with ties resolved in favour of our row.

	Method	Backbone	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Avg.
1-SHOT	DE-ViT-FT (Zhang et al., 2024)	DINOv2-L	10.5	13.0	14.7	19.3	0.6	2.4	10.1
	CD-ViTO (Fu et al., 2024)	DINOv2-L	21.0	17.7	17.8	20.3	3.6	3.1	13.9
	VFM-DETR (Liu et al., 2025)	DINOv2-L	26.1	20.1	20.6	24.2	9.1	9.0	18.2
	CDFormer (Meng et al., 2025)	DINOv2-L	36.0	54.0	16.3	34.5	7.4	12.7	26.8
	ETS (Pan et al., 2025)	Swin-B	28.1	55.8	12.7	39.3	11.7	18.9	27.8
	Domain-RAG (Li et al., 2025)	Swin-B	**57.2**	56.1	18.0	38.0	12.1	20.2	33.6
	Grounding DINO (Liu et al., 2023)	Swin-B	26.3	55.3	14.8	36.4	9.3	15.9	26.3
	Grounding DINO + SCLE	Swin-B	47.3	56.5	22.0	39.0	13.7	23.5	33.7
	HED† (Yu et al., 2026)	Swin-B	42.9	56.9	23.5	42.9	15.2	23.2	34.1
	HED + SCLE	Swin-B	48.3	**58.0**	**23.5**	**43.3**	**17.0**	**23.6**	**35.6**
5-SHOT	DE-ViT-FT (Zhang et al., 2024)	DINOv2-L	49.0	38.1	23.4	21.2	7.8	5.0	24.1
	CD-ViTO (Fu et al., 2024)	DINOv2-L	47.9	41.1	26.9	22.3	11.4	6.8	26.1
	VFM-DETR (Liu et al., 2025)	DINOv2-L	63.3	45.1	32.1	29.5	19.0	19.6	34.7
	CDFormer (Meng et al., 2025)	DINOv2-L	65.0	58.9	28.1	31.7	23.8	23.8	38.6
	ETS (Pan et al., 2025)	Swin-B	64.5	59.7	29.3	42.1	23.5	27.7	41.1
	Domain-RAG (Li et al., 2025)	Swin-B	70.0	59.8	31.5	43.8	24.2	26.8	42.7
	Grounding DINO (Liu et al., 2023)	Swin-B	68.4	57.6	29.6	41.6	19.7	25.6	40.4
	Grounding DINO + SCLE	Swin-B	74.2	58.9	33.9	42.8	23.8	28.1	43.6
	HED† (Yu et al., 2026)	Swin-B	74.9	59.6	34.9	45.5	26.4	26.6	44.7
	HED + SCLE	Swin-B	**76.5**	**60.5**	**35.6**	**45.8**	**27.0**	**28.7**	**45.7**
10-SHOT	DE-ViT-FT (Zhang et al., 2024)	DINOv2-L	49.2	40.8	25.6	21.3	8.8	5.4	25.2
	CD-ViTO (Fu et al., 2024)	DINOv2-L	60.5	44.3	30.8	22.3	12.8	7.0	29.6
	VFM-DETR (Liu et al., 2025)	DINOv2-L	71.3	49.9	37.8	34.1	23.7	22.1	39.8
	CDFormer (Meng et al., 2025)	DINOv2-L	68.7	59.0	32.5	35.5	18.1	26.4	40.0
	ETS (Pan et al., 2025)	Swin-B	70.6	60.8	37.5	42.8	26.1	28.3	44.4
	Domain-RAG (Li et al., 2025)	Swin-B	73.4	61.1	39.0	41.3	26.3	31.2	45.4
	Grounding DINO (Liu et al., 2023)	Swin-B	73.0	58.6	37.2	38.5	25.5	30.3	43.9
	Grounding DINO + SCLE	Swin-B	78.2	61.2	**41.9**	44.1	26.1	30.7	47.0
	HED† (Yu et al., 2026)	Swin-B	79.0	61.1	41.4	46.0	28.1	31.9	47.9
	HED + SCLE	Swin-B	**79.9**	**61.7**	41.7	**46.7**	**28.2**	**33.2**	**48.6**

6 EXPERIMENTS

6.1 SETUP

Datasets and metrics. We follow the CD-FSOD benchmark protocol (Xiong, 2023) and evaluate on six target domains with substantial appearance shift from COCO: ArTaxOr (Drange, 2020), clipart1k (Inoue et al., 2018), DIOR (Li et al., 2020), DeepFish (Saleh et al., 2020), NEU-DET (Song & Yan, 2013), and UODD (Jiang et al., 2021). Each domain is evaluated under 1-, 5-, and 10-shot adaptation, and we report COCO-style box mAP.

Implementation details. The main experiments use plain Grounding DINO with a Swin-B backbone, and we additionally validate the method on HED with a parallel decoder. SCLE changes neither the detector architecture nor the number of learnable parameters, so it applies to both model types directly. All experiments fix 900 encoder candidates. The prototype bank is updated from labeled target-domain samples and frozen at test time. Because the selection rule changes the initial boxes of the decoder and takes effect during training, all models are retrained end to end. Default hyperparameters: foreground shrink $\rho = 0.60$, background expansion $\beta = 1.35$, prototype momentum $\mu = 0.90$, evidence temperature $T_e = 0.20$, contrast temperature $T_c = 0.15$, ring expansion ratio $r = 1.45$, and fusion weight $w = 0.50$. The candidate quota is set to 450 for the semantic channel and 225 each for the local-evidence and joint channels.

6.2 MAIN RESULTS

Table 2 reports 1-, 5-, and 10-shot detection mAP on the six CD-FSOD targets. Within each shot block, the upper rows are reference results reported in the literature and the lower rows are the two baselines together with our method; the two configurations of our method are shaded.

Because SCLE changes only the selection rule, the informative comparison is against the two baselines retrained under the same protocol. SCLE raises the six-domain average at every budget: on Grounding DINO from 26.3 to 33.7, from 40.4 to 43.6, and from 43.9 to 47.0 under 1, 5 and 10 shots, and on HED from 34.1 to 35.6, from 44.7 to 45.7, and from 47.9 to 48.6. The gain falls as supervision grows (+7.4, +3.2 and +3.1 points on Grounding DINO, and +1.5, +1.0 and +0.7 on

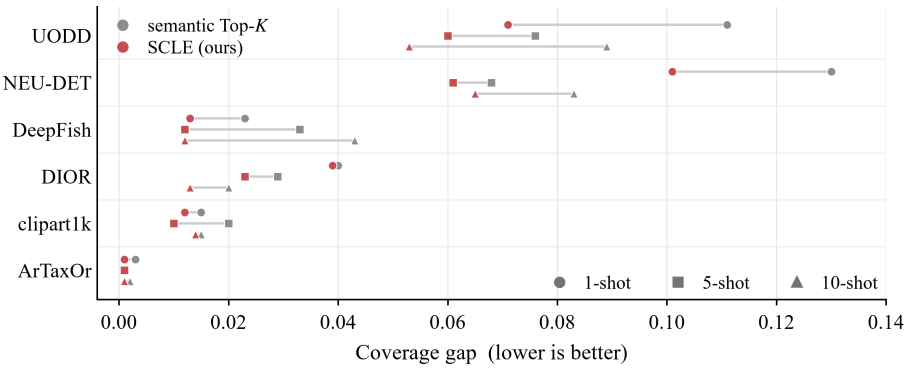


Figure 3: Coverage gap of the two selection rules for the same detector across all six domains and three shot budgets under identical training. SCLE narrows the gap in the vast majority of the configurations, and gains most where semantic selection performs worst.

HED), which is the trend Section 4 predicts: more labeled support samples let fine-tuning correct more of the mis-ranking, leaving less for the selection rule to repair.

6.3 Proposal-level verification

Because SCLE acts at the pre-decoder selection stage, end-to-end mAP can only verify that the strategy takes effect during training; it does not directly reflect a genuine improvement in selection quality. We therefore use the per-object coverage metric of Section 4 to compare the candidate coverage of pure semantic Top-K selection against SCLE directly, under all 1/5/10-shot settings of the six target domains. The two strategies are trained independently on the same data and network, so their candidate pools differ only marginally, which rules out any interference from the pool. Figure 3 reports the coverage gap for all eighteen settings, and SCLE narrows it in the vast majority of them without ever increasing it. The effect is concentrated: 1-shot UODD falls from 0.111 to 0.071, a reduction of 0.040 or 36% of the initial gap, 1-shot NEU-DET from 0.130 to 0.101, and UODD, NEU-DET and DeepFish together account for 0.208 of the 0.239 reduction summed over all settings. The remaining three domains contribute 0.031 and their gaps are at most 0.020 to begin with, so little headroom is left—the 1-shot DIOR gap of 0.039, which the method barely moves, is the exception. This completes the verification that SCLE raises candidate coverage quality rather than merely replacing the candidate set.

Table 3: Ablation of the two evidence-map terms, reporting six-domain mAP (%) for each shot budget. *Semantic Top-900* disables the local-evidence channel and corresponds to the selection rule in Eq. 2. The other three variants retain different subsets of evidence-map terms, and the shaded row denotes the full model. The best entry within each shot group per column is bolded; ties are broken in favour of the full method. All variants use identical code and training settings.

	Evidence map	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Avg.
1-SHOT	Semantic Top-900	42.9	56.9	23.5	42.9	15.2	23.2	34.1
	Foreground only	44.3	58.0	23.5	42.9	16.1	23.0	34.6
	Background only	45.7	57.2	23.2	43.1	17.0	23.1	34.9
	Foreground − background	**48.3**	**58.0**	**23.5**	**43.3**	**17.0**	**23.6**	**35.6**
5-SHOT	Semantic Top-900	74.9	59.6	34.9	45.5	26.4	26.6	44.7
	Foreground only	76.5	60.5	34.6	45.1	25.7	26.6	44.8
	Background only	75.5	60.0	33.5	45.8	27.0	28.7	45.1
	Foreground − background	**76.5**	**60.5**	**35.6**	**45.8**	**27.0**	**28.7**	**45.7**
10-SHOT	Semantic Top-900	79.0	61.1	41.4	46.0	28.1	31.9	47.9
	Foreground only	78.7	61.0	40.4	46.1	27.5	33.2	47.8
	Background only	79.5	61.3	41.7	46.3	26.2	32.9	48.0
	Foreground − background	**79.9**	**61.7**	**41.7**	**46.7**	**28.2**	**33.2**	**48.6**

6.4 Ablations

We separate the two sources of the evidence map. Table 3 compares the two terms used alone against their combination and against Semantic Top-900, all under the same code and training protocol.

Used on its own, each term is close to Semantic Top-900 and better than it on the domains where the gap is large, but neither is uniformly ahead of it; combining the two gives the best average at all three

Table 4: Hyperparameter sensitivity on UODD under the 1-shot setting. Each parameter is swept independently while others remain at default values. All values report mAP (%) on the UODD 1-shot test set. The baseline configuration $\beta = 1.35$, $r = 1.45$, $\rho = 0.60$ yields 23.6. Performance is stable across the tested range: no parameter yields a monotonic trend, and the performance variation is smaller than the cross-domain gap.

Hyperparameter	Sweep range	Best	Worst	Range
Background expansion β	1.1–1.8	23.6	22.1	1.5
Ring expansion ratio r	1.1–1.8	23.6	22.6	1.0
Foreground shrink ρ	0.4–1.0	23.6	22.9	0.7

Table 5: Effect of semantic-channel quota on 1-shot mAP (%). The three channels share a fixed total budget $K = 900$. A semantic quota of 900 disables the local-evidence and joint channels, reverting to the vanilla selection rule in Eq. 2. The best value in each column is bolded.

Semantic quota $\alpha_s K$	UODD	NEU-DET	clipart1k
900 (no quota, = semantic Top-900)	23.2	15.2	56.9
600	22.7	16.7	56.9
450 (default)	**23.6**	**17.0**	**58.0**
300	22.9	15.7	57.7

budgets. This agrees with the mechanism of Section 5: the foreground term supplies a complementary signal where semantic ranking misses local target evidence, and the background term supplies a suppression signal where semantic ranking is attracted by domain-level context. Removing either term loses one of the two signals, so the value of using both lies in their complementarity rather than in either being stronger on its own.

6.5 Sensitivity analysis

We ablate the background expansion β, the ring expansion ratio r, and the foreground shrink ρ one at a time on the 1-shot UODD setting, changing only one parameter per run. Tab. 4 gives the best and worst value attained by each parameter over its sweep. None of the three produces a monotone trend, no sweep moves mAP by more than 1.5 points, the best value shows no discernible ordering, and this variation is smaller than the spread between target domains. This matches what the method predicts: the three parameters control only the smoothness of the evidence map, so as long as the evidence map separates the object from its background, changing these scales has limited effect on the selection outcome. The main experiments therefore use the default values and are not tuned per domain.

The quota allocation is analysed separately, since it directly controls how many candidate slots the local-evidence channel can occupy, with the total over the three channels fixed at K. We set the semantic quota $\alpha_s K$ to 300, 450, and 600, together with 900, which uses no quota and degenerates to pure semantic Top-900, as a control. Table 5 shows the resulting 1-shot mAP on three datasets. All datasets attain their best performance at the default quota, and both increasing and decreasing it cause a decline. Disabling the quota altogether costs 1.8, 1.1, and 0.4 points on NEU-DET, clipart1k, and UODD respectively, which confirms that giving every slot to semantic ranking discards attainable candidates.

7 Conclusion

We studied candidate selection before decoding in language-guided CD-FSOD. Semantic ranking retains candidates by text similarity, which says nothing about coverage; under a fixed budget it discards candidates that could cover the target, and this gap varies with the domain. SCLE redesigns this selection with support-conditioned local evidence: prototypes from the few labeled support samples induce a per-token evidence map, boxes are scored by interior purity and inside-versus-ring contrast, and a quota merges semantic and local-evidence candidates. It adds no learnable parameters and changes no detector architecture. On six target domains it never increases the coverage gap and narrows it in the vast majority of the settings we test.

AI USE STATEMENT

This work used generative AI tools to assist with manuscript preparation, including improving the readability of the text, organizing tables and references, and drafting and revising portions of the text. We did not use generative AI tools for method design, experimental design, code implementation, or the analysis and interpretation of results, nor did we use AI to generate any reported numerical values. All AI-assisted text was reviewed and verified by the authors against the underlying code and experimental logs. We take responsibility for the final content of this work, including the text, the conclusions, and all associated artifacts.

REPRODUCIBILITY STATEMENT

The pre-decoder selection rule of SCLE is implemented as a detector subclass that overrides only the Top-900 query-selection indices; the detector architecture, detection losses, denoising training, bipartite matching, and decoder layers are inherited unchanged from the baseline, so the reported differences are attributable to the selection rule alone. Prototype buffers form part of the checkpoint state and are updated only from labeled target-domain training samples and frozen at inference. The full configuration is stated in Section 6: the six target domains, the 1/5/10-shot budgets, the 900-candidate budget, the number of decoder layers, and the values of every hyperparameter of the selection rule. All six target datasets and both baselines are publicly available through the references cited in Section 6. The proposal-level coverage metric used in Section 4 is defined explicitly in Eq. 3 and can be recomputed from a trained checkpoint without any additional annotation.

REFERENCES

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 2020.

Geir Drange. Arthropod taxonomy orders object detection dataset. Dataset, 2020.

Yuqian Fu, Yu Wang, Yixuan Pan, Lian Huai, Xingyu Qiu, Zeyu Shangguan, Tong Liu, Yanwei Fu, Luc Van Gool, and Xingqun Jiang. Cross-domain few-shot object detection via enhanced open-set object detector. In *European Conference on Computer Vision*, pp. 247–264. Springer, 2024.

Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5001–5009, 2018.

Lihao Jiang, Yi Wang, Qi Jia, Shengwei Xu, Yu Liu, Xin Fan, Haojie Li, Risheng Liu, Xinwei Xue, and Ruili Wang. Underwater species detection using channel sharpening attention. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4259–4267, 2021. doi: 10.1145/3474085.3475563.

Yongwei Jiang, Yixiong Zou, Yuhua Li, and Ruixuan Li. Remedying target-domain astigmatism for cross-domain few-shot object detection. *arXiv preprint arXiv:2603.18541*, 2026.

Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020. doi: 10.1016/j.isprsjprs.2019.11.023.

Liunian Harold Li, Peggy Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yi Zhong, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Yu Li, Xingyu Qiu, Yuqian Fu, Jie Chen, Tianwen Qian, Xu Zheng, Danda Pani Paudel, Yanwei Fu, Xuanjing Huang, Luc V Gool, and Yu-Gang Jiang. Domain-RAG: Retrieval-guided compositional image generation for cross-domain few-shot object detection. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 147594–147614. Curran Associates, Inc., 2025. doi: 10.52202/085713-4935. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d8af9a23bf25dedfb32433597c462045-Paper-Conference.pdf.

Chang-Han Liu, Xunzhi Xiang, Zixuan Duan, Wenbin Li, Qi Fan, and Yang Gao. Don't need retraining: A mixture of DETR and vision foundation models for cross-domain few-shot object detection. In *Advances in Neural Information Processing Systems*, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2023.

Boyuan Meng, Xiaohan Zhang, Peilin Li, Zhe Wu, Yiming Li, Wenkai Zhao, Beinan Yu, and Hui-Liang Shen. CDFormer: Cross-domain few-shot object detection transformer against feature confusion. In *IEEE International Conference on Multimedia and Expo*, 2025.

Jiancheng Pan, Yanxing Liu, Xiao He, Long Peng, Jiahao Li, Yuze Sun, and Xiaomeng Huang. Enhance then search: An augmentation-search strategy with foundation models for cross-domain few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1548–1556, 2025.

Alzayat Saleh, Issam H. Laradji, Dmitry A. Konovalov, Michael Bradley, David Vazquez, and Marcus Sheaves. A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific Reports*, 10(1):14671, 2020. doi: 10.1038/s41598-020-71639-x.

Kechen Song and Yunhui Yan. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Applied Surface Science*, 285:858–864, 2013. doi: 10.1016/j.apsusc.2013.09.002.

Wanqi Wang, Jingcai Guo, Yuxiang Cai, and Zhi Chen. Learning multi-modal prototypes for cross-domain few-shot object detection. *arXiv preprint arXiv:2602.18811*, 2026. doi: 10.48550/arXiv.2602.18811.

Wuti Xiong. CD-FSOD: A benchmark for cross-domain few-shot object detection. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096216.

Shuai Yi, Yixiong Zou, Yuhua Li, and Ruixuan Li. Addressing exacerbated attention sink for source-free cross-domain few-shot learning. *arXiv preprint arXiv:2605.25799*, 2026a.

Shuai Yi, Yixiong Zou, Yuhua Li, and Ruixuan Li. Improving CLIP adaptation by breaking tail alignment for source-free cross-domain few-shot learning. *arXiv preprint arXiv:2605.29776*, 2026b.

Xuanlong Yu, Youyang Sha, Longfei Liu, Xi Shen, and Di Yang. A closer look at cross-domain few-shot object detection: Fine-tuning matters and parallel decoder helps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Hao Zhang, Feng Li, Shilong Liu, Lu Zhang, Hang Su, Jun Zhu, Dapeng Ni, and Heung-Yeung Shum. Dino: DETR with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations*, 2022.

Xinyu Zhang, Yuhan Liu, Yuting Wang, and Abdeslam Boularias. Detect everything with few examples. In *Conference on Robot Learning*, 2024.

Yaze Zhao, Yixiong Zou, Yuhua Li, and Ruixuan Li. Interpretable cross-domain few-shot learning with rectified target-domain local alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

Zijian Zhuang, Yixiong Zou, Yuhua Li, and Ruixuan Li. Free-lunch augmentation by revisiting diffusion-based data generation for cross-domain few-shot object detection. *arXiv preprint arXiv:2608.04394*, 2026.

Yixiong Zou, Shuai Yi, Yuhua Hu, and Ruixuan Li. A closer look at the CLS token for cross-domain few-shot learning. In *Advances in Neural Information Processing Systems*, 2024.