



ARTICLE

Adapting GroundingDINO for Remote Sensing Object Detection via Lightweight Feature Refinement and Concise Category-Aware Prompts

Angran Li and Chao Tang*

Northwest Institute of Nuclear Technology, Xi'an, China

*Corresponding Author: Chao Tang. Email: tangchao10@163.com

Received: 17 June 2026; Accepted: 27 August 2026; Published: 22 October 2026

ABSTRACT: Text-guided vision-language detectors provide a natural-language interface for remote sensing (RS) object detection, but their adaptation to RS imagery remains limited by degraded small-object performance and inefficient text-vision alignment. This paper presents a systematic study on adapting GroundingDINO to RS object detection through end-to-end training on DOTA v1.0 (Dataset for Object Detection in Aerial Images). We introduce Lightweight Feature Refinement GroundingDINO (LFR-GDINO), which integrates a lightweight feature-refinement module (v3)—combining Efficient Channel Attention (ECA) with InnerMPDIoU loss—and Remote Sensing Category-Aware Prompt Engineering (RS-CAPE), a concise prompt strategy appending minimal RS-relevant keywords. On DOTA v1.0, LFR-GDINO achieves 0.420 mAP (mean average precision) and 0.695 mAP₅₀, attaining the highest mAP₅₀ among compared methods under the text-guided horizontal bounding box (HBB) protocol. Prompt experiments point to a detection–precision trade-off: verbose prompts improve mAP₅₀ but reduce mAP₇₅, whereas concise 2–3-token prompts achieve the best balance, supported by a training-free Bidirectional Encoder Representations from Transformers (BERT) attention analysis. Transfer experiments on DIOR (Detection in Optical Remote-sensing images), DIOR-RSVG (DIOR remote sensing visual grounding), VisDrone-DET (VisDrone object detection), and a Swin-T backbone comparison further characterize the adaptation behavior. A Seen/Unseen split shows that short fine-tuning improves Seen-class adaptation but does not preserve Unseen-category transfer. These findings offer practical guidelines for adapting text-guided vision-language detectors to RS imagery.

KEYWORDS: Text-guided object detection; vision-language detector adaptation; vision-language model; remote sensing; prompt engineering; GroundingDINO; DOTA; small object detection

1 Introduction

Closed-set detectors anchored in predefined categories—e.g., oriented bounding box (OBB) networks with selective kernel attention on DOTA (Dataset for Object Detection in Aerial Images)—must be retrained to recognize novel classes. This reliance on retraining constrains the efficiency of remote sensing (RS) interpretation and creates a demand for scalable solutions that circumvent massive annotation efforts.

Vision-Language Models (VLMs) have pioneered a text-guided detection paradigm to address this limitation. GroundingDINO, leveraging a DEtection TRansformer (DETR) architecture with Bidirectional Encoder Representations from Transformers (BERT)-based text encoding, achieves remarkable zero-shot performance on Common Objects in Context (COCO), offering a viable pathway to mitigate the scarcity of annotated RS data. However, systematic adaptation of GroundingDINO to RS remains underexplored. Direct application faces two challenges: (1) **small object bottleneck**—RS imagery contains abundant small objects

where GroundingDINO shows degraded performance; and (2) **text-vision alignment inefficiency**—bare category names underutilize the text-guided capability, yet verbose descriptions can degrade performance.

In this paper, we present an empirical engineering study that addresses both challenges under practical resource constraints (dual graphics processing units (GPUs), 15-epoch schedule). Rather than proposing fundamentally new modules, we focus on *diagnosis-driven adaptation*: identifying where and why a pre-trained VLM detector underperforms on RS imagery, and applying targeted, lightweight corrections examined through controlled ablation. Our contributions are:

- We establish an end-to-end training baseline, diagnose scale-dependent performance bottlenecks through per-scale AP decomposition, and derive a lightweight adaptation pipeline (v3) whose design choices—Efficient Channel Attention (ECA) over the Convolutional Block Attention Module (CBAM), InnerMPDIoU with $r = 0.7$ —are motivated by ablation evidence under the adopted budget rather than architectural novelty alone.
- We introduce Remote Sensing Category-Aware Prompt Engineering (RS-CAPE), a prompt strategy that appends concise domain keywords to each category name, and empirically characterize a detection–precision trade-off associated with prompt design. A training-free token-level diagnostic on frozen BERT provides supporting evidence for the recommended 2–3-token operating point.
- We evaluate the adaptation across four RS benchmarks (DOTA, Detection in Optical Remote-sensing images (DIOR), DIOR remote sensing visual grounding (DIOR-RSVG), VisDrone object detection (VisDrone-DET)) under a unified protocol, reporting both positive gains and negative findings (e.g., large-object regression, loss of unseen-category transfer) to provide practical guidelines for future text-guided RS detection research.

Terminology: The architecture enhancement (ECA + InnerMPDIoU) is referred to as v3; the prompt strategy as RS-CAPE; the full pipeline (v3 + RS-CAPE) as Lightweight Feature Refinement GroundingDINO (LFR-GDINO).

2 Related Work

2.1 Object Detection in Remote Sensing

Remote sensing object detection has evolved from hand-crafted features to sophisticated deep learning paradigms. Architectures such as Faster Region-based Convolutional Neural Network (Faster R-CNN) [1] and its derivatives (e.g., Region of Interest (RoI) Transformer [2], Oriented R-CNN [3]) utilize region proposals to set high accuracy standards, whereas single-stage detectors like You Only Look Once (YOLO) [4] and RetinaNet [5] trade off some precision for superior inference rates. Facing the complexities of scale and density, Large Selective Kernel Network (LSKNet) [6] employs selective kernel attention to capture multi-scale contexts, while Convex-hull Feature Adaptation (CFA) [7] utilizes convex-hull feature adaptation for dense, rotated objects. Despite these advancements, such models remain constrained by a closed-world paradigm, recognizing only the categories present in the training set.

2.2 Vision-Language Models for Object Detection

Vision-language pre-training has enabled open-vocabulary detection. Grounded Language-Image Pre-training (GLIP) [8] unifies phrase grounding and detection through grounded language-image pre-training. Open-World Vision Transformer (OWL-ViT) [9] extends Contrastive Language-Image Pre-training (CLIP) [10] for open-vocabulary detection. GroundingDINO [11] tightly couples the detection prowess of DETR with Improved deNoising anchor boxes (DINO) [12] with BERT’s linguistic capacity, establishing a new benchmark for text-prompted detection. Efficient Grounding DINO [13] further introduces efficient

cross-modality fusion and label assignment strategies tailored for RS visual grounding, reporting competitive results with reduced computational cost. A dedicated investigation into how training schedule, architectural bottlenecks, and prompt design jointly affect RS detection performance remains absent; the present work complements these efforts with a diagnosis-driven adaptation study under constrained resources.

2.3 Prompt Engineering for Vision-Language Models

The performance of VLMs hinges critically on prompt engineering. Although methods like Context Optimization (CoOp) [14] and Conditional Context Optimization (CoCoOp) [15] have explored continuous prompt learning for CLIP, and DetPro [16] has adapted these techniques for detection tasks, GroundingDINO demonstrates the most direct interaction through its positive_map mechanism. Nevertheless, the application of these techniques to remote sensing remains superficial. Given that RS categories suffer from semantic ambiguity and varying levels of specificity, the impact of prompt design on detection stability has not been rigorously examined.

3 Methodology

3.1 Preliminary: GroundingDINO Architecture

GroundingDINO performs text-guided object detection through tight vision-language fusion. Given image I and text prompt T (category names separated by periods): (1) ResNet-50 extracts multi-scale features $\{F_i\}$ while BERT encodes T into embeddings $E \in \mathbb{R}^{N_t \times d}$; (2) a ChannelMapper neck projects the visual features to a unified dimension, and a Feature Enhancer then performs cross-modality fusion through stacked layers of image-to-text and text-to-image cross-attention, mutually enhancing the visual and textual representations [11]; (3) Language-Guided Query Selection uses the text embeddings to select the top- k enhanced encoded features as the initial object queries, which the deformable cross-attention decoder layers subsequently refine rather than generate [11]; (4) a *positive_map* $P \in \{0, 1\}^{C \times N_t}$ maps categories to token positions for text-guided classification.

Fig. 1 illustrates the overall framework of LFR-GDINO, and Fig. 2 details the internal structure of each enhanced module, highlighting the v3 architecture enhancements and the RS-CAPE text strategy.

3.2 Architecture Enhancement: v3

The baseline diagnosis (Section 4.1) reveals that AP_s lags behind AP_m and AP_l by 12.8 and 19.5 percentage points (pp), respectively, pointing to two remediable bottlenecks: (i) the ChannelMapper neck lacks inter-channel interaction, and (ii) Generalized Intersection over Union (GIoU) loss provides weak gradients for small boxes with low overlap. We address each with a targeted, low-overhead modification selected through ablation (Section 4.2) rather than exhaustive architectural search.

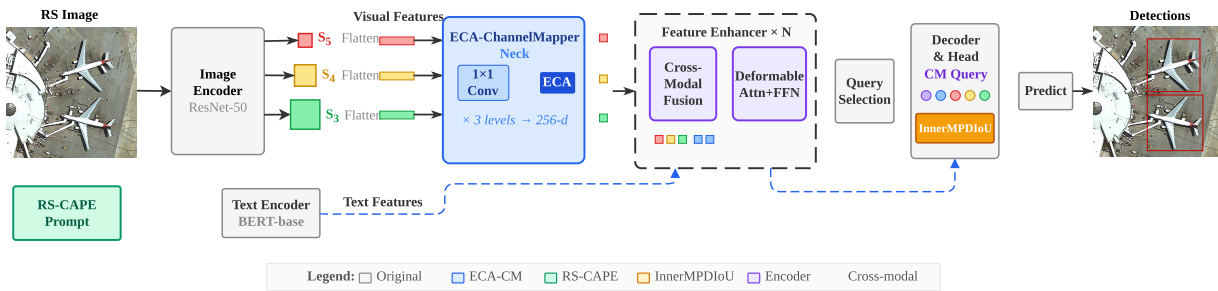


Figure 1: Overall framework of LFR-GDINO (module details in Fig. 2).

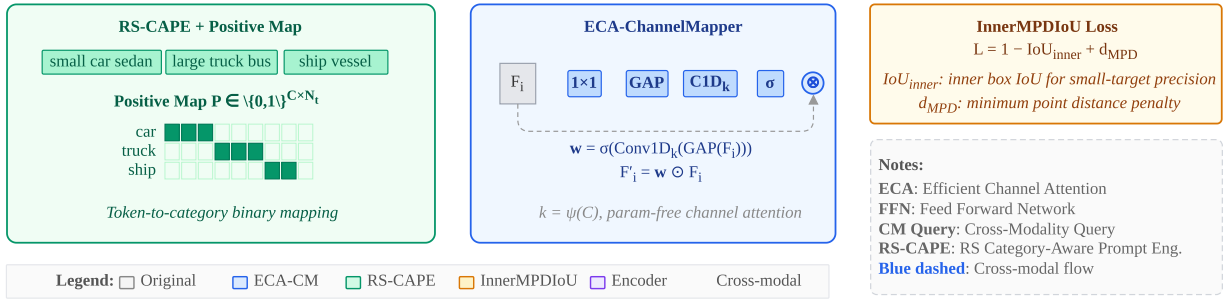


Figure 2: Module details of Lightweight Feature Refinement GroundingDINO (LFR-GDINO): Remote Sensing Category-Aware Prompt Engineering (RS-CAPE) with positive_map, Efficient Channel Attention (ECA)-ChannelMapper, and Inner Minimum Point Distance Intersection over Union (InnerMPDIoU) loss.

3.2.1 ECA Channel Attention in Neck

The standard ChannelMapper neck performs dimension projection without inter-channel interaction. We augment it with ECA [17], which captures cross-channel dependencies through a 1D convolution on channel-wise statistics with <0.3 KB additional parameters per level:

$$w = \sigma(\text{Conv1D}_k(\text{GAP}(F_i))), \quad \hat{F}_i = w \odot F_i \quad (1)$$

where GAP denotes global average pooling, k is the adaptive kernel size, and σ is the sigmoid function. Unlike CBAM [18], ECA avoids spatial attention, which is redundant with Deformable Self-Attention in the DETR encoder.

3.2.2 InnerMPDIoU Loss

We replace GIoU loss [19] with InnerMPDIoU, combining the Inner-IoU auxiliary box framework [20] with the minimum point distance penalty from MPDIoU [21]. An inner box $\hat{b}_{\text{inner}}$ is constructed by shrinking $\hat{b}$ toward its center by factor $r = 0.7$:

$$\mathcal{L}_{\text{InnerMPDIoU}} = 1 - \text{IoU}(\hat{b}_{\text{inner}}, b) + \frac{d^2(p_{\hat{b}}, p_b)}{d_{\text{union}}^2} \quad (2)$$

where $d(p_{\hat{b}}, p_b)$ is the closest-corner distance and d_{union}^2 is the squared diagonal of the enclosing box. This formulation provides stronger gradients for small objects where overlap signals are weak.

3.3 RS-CAPE: Category-Aware Prompt Engineering

3.3.1 Motivation

In conventional training pipelines for Grounding DINO, object categories are flattened into a simple token sequence (e.g., “plane. ship. ...”), where each label is mapped to spatial predictions through the positive_map. Yet, this approach reduces rich remote sensing concepts to mere lexical tokens. We argue that augmenting these labels with contextually relevant supplementary keywords will bridge the semantic gap, thereby optimizing the cross-modal alignment for complex RS scenes.

3.3.2 Design Principle: Conciseness over Verbosity

Counterintuitively, our early trials indicated that appending lengthy keyword chains to each category name actually degraded localization precision. This prompted us to explore a more restrained

approach, ultimately yielding two prompt configurations with markedly different token budgets: **RS-CAPE v1 (Verbose)** lets each category entry run to 5–8 tokens, padding the bare name with synonyms and descriptors (e.g., “small vehicle sedan automobile car compact”), whereas **RS-CAPE v2 (Concise)** caps each entry at 2–3 tokens, appending at most two domain-relevant keywords (e.g., “small car”). Representative entries of both variants and the full v2 template are given in [Section 3.4](#), and the complete per-category mappings for every dataset are tabulated in [Appendix A](#).

3.3.3 Token Capacity Effect

The divergence between v1 and v2 may be associated with how cross-modal attention is distributed across the tokens that the `positive_map` assigns to each category: as a category claims more tokens, the average attention weight per token tends to decrease, which may sharpen detection sensitivity (higher mAP_{50}) while diluting the focus needed for precise localization (lower mAP_{75}). RS-CAPE v2 therefore stays within the ≤ 3 -token regime to preserve localization fidelity; [Section 5.3](#) provides a quantitative token-level diagnostic.

3.4 Experimental Setup

3.4.1 Datasets

Our evaluation spans four datasets, each chosen to probe a different aspect of adaptation behavior. DOTA v1.0 [22] serves as the primary benchmark: 2806 images across 15 horizontal bounding box (HBB) categories, cropped into 1024×1024 patches with 200-pixel overlap (15,749 train/2649 val patches). DIOR [23] (23,463 images, 20 categories, 0.5–30 m resolution) functions as an *in-domain* transfer target, sharing 9 semantically overlapping categories with DOTA (plane $\leftrightarrow$ airplane, baseball-diamond $\leftrightarrow$ baseball-field, bridge, ground-track-field, ship, tennis-court, basketball-court, storage-tank, harbor). For cross-task evaluation, the remote sensing visual grounding (RSVG) benchmark DIOR-RSVG [24] (17,402 images, 38,320 language–target pairs) provides free-form text descriptions rather than bare category names, testing whether architectural gains carry over to open-ended text input. VisDrone-DET [25] (drone-captured, 10 categories) rounds out the evaluation with a cross-domain perspective ([Section 4.12](#)).

3.4.2 Implementation Details

All experiments use MMDetection 3.2.0 [26] with ImageNet-pretrained ResNet-50 [27] backbone and BERT-base-uncased [28] text encoder. We train for 15 epochs with AdamW ($\text{lr} = 1 \times 10^{-4}$, weight decay = 1×10^{-4} , MultiStepLR decay at epoch 13, gradient clipping norm 0.1). Data augmentation includes six-scale random resize ($\{800, 900, 1000, 1100, 1200, 1344\}$) and horizontal flip ($p = 0.5$). Training uses dual NVIDIA RTX 4090 GPUs with batch size 4 per GPU (effective 8). All reported results correspond to a single training run with a fixed random seed (seed = 42); repeated-run statistics are not available due to computational budget constraints, and we therefore do not make statistical-significance claims for differences between single runs—all observed gaps should be read as indicative trends rather than confirmed effects. Regarding DOTA preprocessing, the original oriented-box annotations are converted to horizontal boxes by taking the axis-aligned enclosing box of each oriented box; images are then split into 1024×1024 patches with 200-pixel overlap, and an object is kept in a patch when its box center falls inside that patch, with boxes clipped at patch borders. At evaluation, predictions from overlapping patches are merged back onto the original large images following the standard DOTA protocol, and overlapping detections are suppressed by non-maximum suppression. The complete RS-CAPE v2 prompt template for DOTA is: “airplane aircraft. diamond baseball field. long bridge. oval track field. small car. large truck bus. ship boat. tennis court. basketball court. circular storage tank. soccer football field. circular roundabout. harbor port. swimming

pool. helicopter rotorcraft.” The RS-CAPE v1 (verbose) variant enumerates 5–8 tokens per category, e.g., “small vehicle sedan automobile car compact”; the DIOR and VisDrone-DET experiments follow the same concise principle (1–2 supplementary keywords per category), e.g., DIOR: “airplane aircraft”; VisDrone-DET: “walking pedestrian”. The complete per-category mappings and the concatenated prompt strings of all datasets are given in Appendix A, so that every prompt used in this study can be reproduced verbatim. Because each enriched category entry replaces the bare class name and is tokenized as a single phrase by BERT, the supplementary keywords are included in the corresponding positive_map token span; the positive_map itself is constructed automatically by the MMDetection GroundingDINO implementation (BERT tokenization → token-index mapping to classification head slots). Metrics: COCO-style mAP, mAP₅₀, AP_s/AP_m/AP_l for detection; Pr@*k* and mean/cumulative IoU for RSVG.

4 Results

4.1 Baseline Establishment and Diagnosis

Table 1 reports the impact of training strategies on baseline performance.

Table 1: Impact of training strategies on GroundingDINO baseline (DOTA v1.0). Best results in **bold**.

Configuration	Epochs	Multi-Scale	mAP	mAP ₅₀	mAP ₇₅	AP _s	AP _m	AP _l
Original	12	Single 1024	0.398	0.651	0.457	0.270	0.402	0.444
+Multi-scale	12	6-scale	0.409	0.662	0.478	0.286	0.406	0.462
+Extended training	15	6-scale	0.418	0.670	0.483	0.290	0.418	0.485

Multi-scale training contributes +1.1 pp mAP, primarily from AP_s (+1.6 pp). The final baseline (mAP = 0.418) shows a pronounced small-object gap: AP_s (0.290) is 12.8 pp below AP_m and 19.5 pp below AP_l, motivating the targeted optimizations below.

4.2 Ablation Study: Neck Structure and Loss Function

Table 2 reports the ablation results for neck structure and loss function.

Table 2: Ablation study on neck structure and loss function (DOTA v1.0, 15 epochs). CM denotes the ChannelMapper neck. Best results in **bold**.

Method	Neck/Loss	mAP	mAP ₅₀	mAP ₇₅	AP _s	AP _m	AP _l
Baseline	CM/GIoU	0.418	0.670	0.483	0.290	0.418	0.485
+CBAM(P3)	CBAM + CM/GIoU	0.417	0.667	0.483	0.288	0.416	0.476
+CBAM + InnerMPDIoU	CBAM + CM/InnerMPDIoU	0.416	0.667	0.482	0.283	0.422	0.461
+ECA + InnerMPDIoU (v3)	ECA + CM/InnerMPDIoU	0.419	0.670	0.483	0.294	0.423	0.469

Two patterns stand out. First, the CBAM variants show no improvement over the baseline under the single-run protocol, which is consistent with the view that CBAM’s spatial attention is redundant with the Deformable Self-Attention in the DETR encoder; the channel-only ECA + InnerMPDIoU configuration (v3) yields small AP_s (+0.4 pp) and AP_m (+0.5 pp) gains. We note, however, that the adopted ablation does not include the full 2 × 2 factorial design (the ECA + Giou and ChannelMapper + InnerMPDIoU configurations were not run under the current budget), so the individual contributions of ECA and InnerMPDIoU cannot be fully separated; v3 is therefore reported as a combined configuration, and we do not attribute the observed gains to either component in isolation. Second, all modifications exhibit a scale trade-off: AP_s/AP_m

improvements co-occur with reduced AP_l (v3: -1.6 pp), which may reflect feature competition within DETR’s shared multi-scale representation, although this explanation remains qualitative (Section 5.2).

4.3 Ablation Study: RS-CAPE Text Strategy

Table 3 compares the two RS-CAPE variants against the baseline.

Table 3: Ablation study on text prompt strategy (DOTA v1.0, 15 epochs). Best results in **bold**.

Text Strategy	mAP	mAP ₅₀	mAP ₇₅	AP _s	AP _m	AP _l
Bare names (Baseline)	0.418	0.670	0.483	0.290	0.418	0.485
RS-CAPE v1 (verbose, 5–8 tok.)	0.414	0.693	0.467	0.292	0.418	0.467
RS-CAPE v2 (concise, 2–3 tok.)	0.418	0.682	0.476	0.281	0.425	0.487

Both RS-CAPE variants improve mAP₅₀ over bare names (v1: $+2.3$ pp, v2: $+1.2$ pp), consistent with a *detection-precision trade-off* associated with prompt design (both token length and keyword content vary across configurations, so no causal attribution is implied). Verbose prompts (v1) attain the highest mAP₅₀ (0.693) but show a 1.6 pp mAP₇₅ deficit, consistent with reduced localization precision. Concise prompts (v2) preserve baseline mAP while gaining on mAP₅₀, AP_m, and AP_l, achieving a better overall balance. Fig. 3 illustrates the training dynamics.

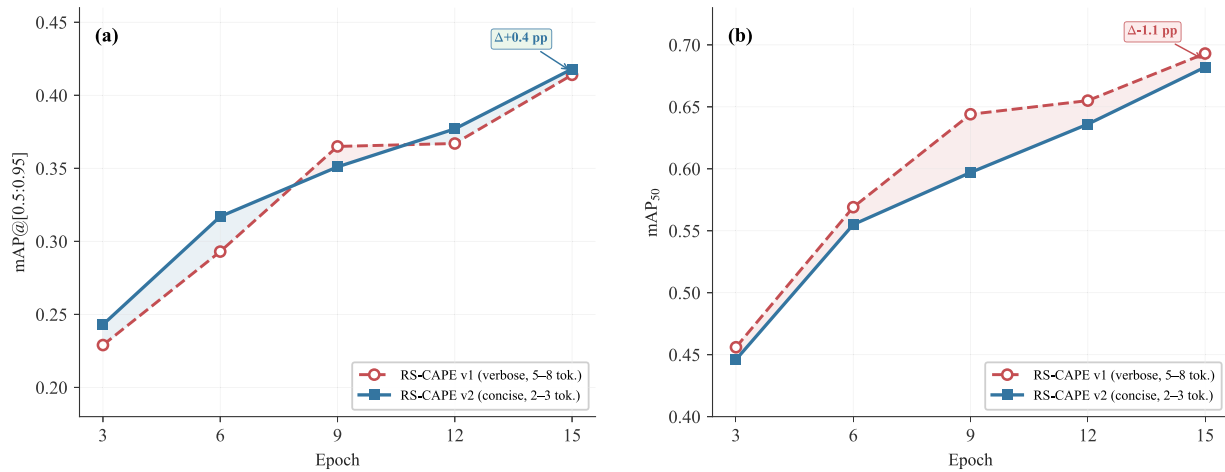


Figure 3: Training dynamics of RS-CAPE v1 (verbose) and v2 (concise) on DOTA v1.0. (a) mAP; (b) mAP₅₀. Shaded region: Performance gap; Δ: Per-epoch difference.

4.4 Progressive Integration

Table 4 presents the combined effect of v3 architecture and RS-CAPE.

The final LFR-GDINO configuration achieves 0.420 mAP and 0.695 mAP₅₀, representing $+0.2$ pp mAP and $+2.5$ pp mAP₅₀ over the baseline. Given the single-run protocol, we refrain from interpreting either difference as statistically confirmed: the larger mAP₅₀ gap may indicate a meaningful trend, whereas the mAP gap is plausibly within run-to-run variation. Under this caveat, the v3 configuration coincides with improved small/medium-object localization, while RS-CAPE coincides with higher detection confidence. EMA-based strategies (ExpMomentumEMA, ShortEpochEMA) were also tested but provided no measurable benefit under the short 15-epoch schedule. The scale trade-off persists in the combined model ($AP_l = 0.467$).

vs. baseline 0.485), which may be associated with feature competition within DETR’s shared multi-scale representation (Section 5.2).

Table 4: Progressive integration of optimization modules (DOTA v1.0, 15 epochs). Best results in **bold**.

Configuration	ECA Neck	InnerMPDIoU	RS-CAPE	mAP	mAP ₅₀	AP _s	AP _m	AP _l
Baseline				0.418	0.670	0.290	0.418	0.485
+v3	✓	✓		0.419	0.670	0.294	0.423	0.469
+RS-CAPE			✓	0.418	0.682	0.281	0.425	0.487
LFR-GDINO (v3 + RS-CAPE)	✓	✓	✓	0.420	0.695	0.286	0.423	0.467

4.5 Multi-Scale Performance Analysis

Fig. 4 compares the four key configurations across five metrics. v3 expands AP_s/AP_m (localization enhancement), while RS-CAPE v2 extends mAP₅₀ (confidence boost). LFR-GDINO covers the largest area overall, and the mAP₅₀ margin (+2.5 pp) is the widest of the five metrics; as all values originate from single runs, this pattern is read as an indication that v3 and RS-CAPE may address complementary aspects of detection performance rather than as a confirmed effect.

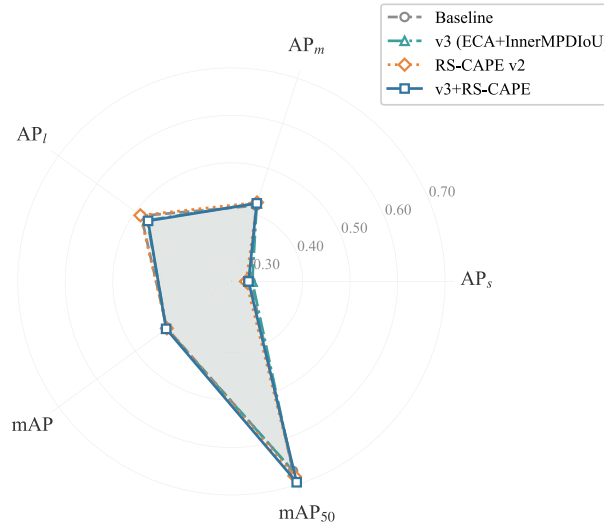


Figure 4: Radar chart of five metrics across key configurations on DOTA v1.0. LFR-GDINO achieves the largest coverage area.

4.6 Reference Comparison under HBB AP₅₀ Protocol

Table 5 compares LFR-GDINO with representative detectors on DOTA v1.0.

LFR-GDINO achieves 69.5% mAP₅₀, the highest among compared methods, with the best per-class AP₅₀ on 8/15 categories (SV, LV, SH, TC, ST, HA, SP, HC). Compared with our own baseline in Table 1 (67.0%), LFR-GDINO improves by +2.5 pp under the same evaluation protocol. We note that the “Grounding DINO” row (62.17%) follows the upstream implementation’s original training recipe, which differs from our extended 15-epoch multi-scale recipe in Table 1; the two numbers therefore correspond to different training protocols and are not interchangeable. Closed-set CNN detectors retain advantages on well-represented static categories (PL, BC, SBF), yet our method ranks second on all such categories.

Table 5: Reference comparison on DOTA v1.0 (AP₅₀, %). All methods use ResNet-50. **Red/blue:** best/second-best per column. Category abbreviations: PL, plane; BD, baseball-diamond; BR, bridge; GTF, ground-track-field; SV, small-vehicle; LV, large-vehicle; SH, ship; TC, tennis-court; BC, basketball-court; ST, storage-tank; SBF, soccer-ball-field; RA, roundabout; HA, harbor; SP, swimming-pool; HC, helicopter.

Method	Params	GFLOPs	FPS	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP ₅₀
<i>CNN-based</i>																			
RetinaNet [5]	36.62	215.15	29.9	87.00	58.10	36.90	25.40	43.90	68.80	58.60	88.90	52.50	55.30	38.70	56.50	68.80	50.60	9.70	53.32
Faster R-CNN [1]	41.42	211.09	30.9	89.50	72.30	47.80	67.70	47.30	77.60	61.10	90.70	62.00	62.00	60.90	61.40	75.20	53.30	55.20	65.60
<i>Transformer-based</i>																			
Deformable DETR [29]	40.36	205.36	20.6	87.50	62.30	44.20	55.10	55.20	74.60	77.70	89.00	54.00	59.50	50.60	52.40	70.80	49.20	58.00	62.67
DINO [12]	47.57	284.01	15.4	85.00	61.00	41.20	44.40	64.80	77.70	83.40	86.70	45.70	71.60	47.10	55.00	70.40	54.90	52.70	62.77
<i>Vision-language-based</i>																			
Efficient G-DINO [13]	169.3	309.12	8.4	84.40	62.70	36.20	48.40	63.80	78.30	84.30	87.20	49.80	71.60	51.80	49.20	68.90	56.50	46.30	62.63
GroundingDINO [11]	172.4	414.79	7.7	84.30	63.00	40.40	50.40	67.20	78.00	83.90	87.80	40.80	72.30	44.60	52.00	69.50	55.10	43.30	62.17
Enhanced G-DINO [30]	160.0	208.76	12.6	84.60	61.70	37.60	42.20	66.90	79.80	84.30	85.40	42.30	72.40	45.60	53.90	65.70	56.40	55.00	62.28
<i>Ours</i>																			
LFR-GDINO	169.3	309.0	13.1	89.10	69.60	47.40	59.50	69.70	85.10	86.60	91.00	55.30	73.70	53.50	59.00	75.90	60.10	66.50	69.50

Note: Our mAP₅₀ uses COCO 101-point interpolation; reference methods may use VOC 11-point (<1 pp difference). All “Vision-language-based” rows are reproduced by the authors using the MMDetection framework with a ResNet-50 backbone; the “Grounding DINO” row follows the upstream implementation’s original recipe, whereas Efficient G-DINO and Enhanced G-DINO are reproduced under the schedules reported in their respective papers; CNN-based and Transformer-based results are taken from the cited literature under their originally reported protocols.

4.7 Per-Category Analysis

Table 6 reports per-class AP for the four configurations. v3 yields the largest gains on helicopter (+6.5 pp) and swimming-pool (+2.0 pp), where small-to-medium objects may benefit from InnerMPDIoU. RS-CAPE excels on tennis-court (+1.4 pp) and harbor (+1.4 pp), where supplementary keywords align well. LFR-GDINO achieves the best AP on basketball-court (0.414, +3.1 pp) and ship (0.610). The baseline retains advantages on ground-track-field and small-vehicle, suggesting these categories benefit from bare text prompts.

Table 6: Per-category AP@ [0.5:0.95] comparison on DOTA v1.0 (15 epochs). Best results per category in **bold**. Per-category AP values are obtained from a dedicated class-wise evaluation pass (each category queried individually), whereas the overall mAP row reports the standard COCO-style all-class metric; because the two aggregation protocols differ, the overall row cannot be recovered by averaging the per-category columns.

Category	Baseline	v3	RS-CAPE	LFR-GDINO
Plane	0.629	0.630	0.639	0.630
Baseball-diamond	0.392	0.405	0.368	0.393
Bridge	0.221	0.222	0.221	0.227
Ground-track-field	0.420	0.390	0.370	0.380
Small-vehicle	0.425	0.410	0.409	0.409
Large-vehicle	0.617	0.621	0.624	0.623
Ship	0.603	0.608	0.606	0.610
Tennis-court	0.804	0.796	0.818	0.803
Basketball-court	0.383	0.380	0.391	0.414

(Continued)

Table 6 (continued)

Category	Baseline	v3	RS-CAPE	LFR-GDINO
Storage-tank	0.448	0.451	0.447	0.449
Soccer-ball-field	0.412	0.401	0.397	0.389
Roundabout	0.296	0.294	0.296	0.296
Harbor	0.430	0.400	0.444	0.413
Swimming-pool	0.228	0.248	0.236	0.239
Helicopter	0.268	0.333	0.305	0.329
Overall mAP	0.418	0.419	0.418	0.420

4.8 Cross-Task Transfer: DIOR-RSVG

We evaluate v3 on DIOR-RSVG visual grounding (RS-CAPE is *not applicable* here, as the text input is free-form descriptions). [Table 7](#) reports the results.

Table 7: Cross-task transfer on DIOR-RSVG visual grounding. Best results in **bold**.

Configuration	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	meanIoU	cumIoU
Baseline	82.41	79.76	74.40	63.45	42.68	72.96	82.07
v3 (ECA + InnerMPDIoU)	82.64	79.83	74.53	63.87	42.17	73.10	82.17
Δ	+0.23	+0.07	+0.13	+0.42	−0.51	+0.14	+0.10

v3 attains numerically greater values for most metrics reported (Pr@0.5–0.8, meanIoU, and cumIoU), notably at Pr@0.8 (+0.42 pp), suggesting a possible gain in localization precision. The −0.51 pp difference at Pr@0.9 cannot be distinguished from run-to-run variation without repeated experiments. Under the single-run protocol, these differences should be read as indicative trends rather than confirmed gains.

4.9 In-Domain Transfer: DIOR Object Detection

To examine whether gains are DOTA-specific, we train and evaluate on DIOR (satellite imagery, 20-category prompt). Module-level isolation is included. [Table 8](#) reports the results.

v3-only retains the small-/medium-object gains from DOTA (AP_s +1.8 pp, AP_m +1.5 pp), suggesting that ECA + InnerMPDIoU may transfer across RS datasets. *RS-CAPE-only* produces minor shifts, consistent with DOTA's observation that prompt enrichment primarily boosts mAP_{50} . The *combined LFR-GDINO* improves mAP_{50} by +0.5 pp yet regresses on mAP (−1.2 pp) due to AP_l degradation (−1.5 pp) and an unexpected AP_s drop, suggesting a mild negative interaction between v3 and RS-CAPE on DIOR's broader category taxonomy. The RS-CAPE prompt enumerates all 20 category phrases ([Appendix A](#)), and the four configurations share identical annotation files, so the comparison among them is unaffected.

Table 8: In-domain transfer on DIOR object detection (satellite imagery, 15 epochs). Module-level isolation is included alongside the combined pipeline. Metrics are averaged over the 19 DIOR categories that carry ground truth in the converted annotations used here. Best results in **bold**.

Configuration	mAP	mAP ₅₀	AP _s	AP _m	AP _l
Baseline	0.578	0.775	0.175	0.392	0.704
v3-only (ECA + InnerMPDIoU)	0.576	0.772	0.193	0.407	0.699
RS-CAPE-only (v2)	0.571	0.775	0.183	0.399	0.694
LFR-GDINO	0.566	0.780	0.163	0.402	0.689
Δ (LFR-GDINO vs. Baseline)	-1.2 pp	+0.5 pp	-1.2 pp	+1.0 pp	-1.5 pp

4.10 Feature Activation Analysis

Fig. 5 visualizes Neck output activation maps for Baseline and LFR-GDINO. v3 + RS-CAPE exhibits two distinct patterns: (1) suppressed background activation in non-target regions (water, vegetation, rooftops), and (2) spatially concentrated high-intensity responses at object locations (ships, vehicles, harbor). These qualitative observations are consistent with the observed mAP₅₀ and AP_m changes, suggesting that ECA channel re-weighting sharpens feature contrast between targets and background.

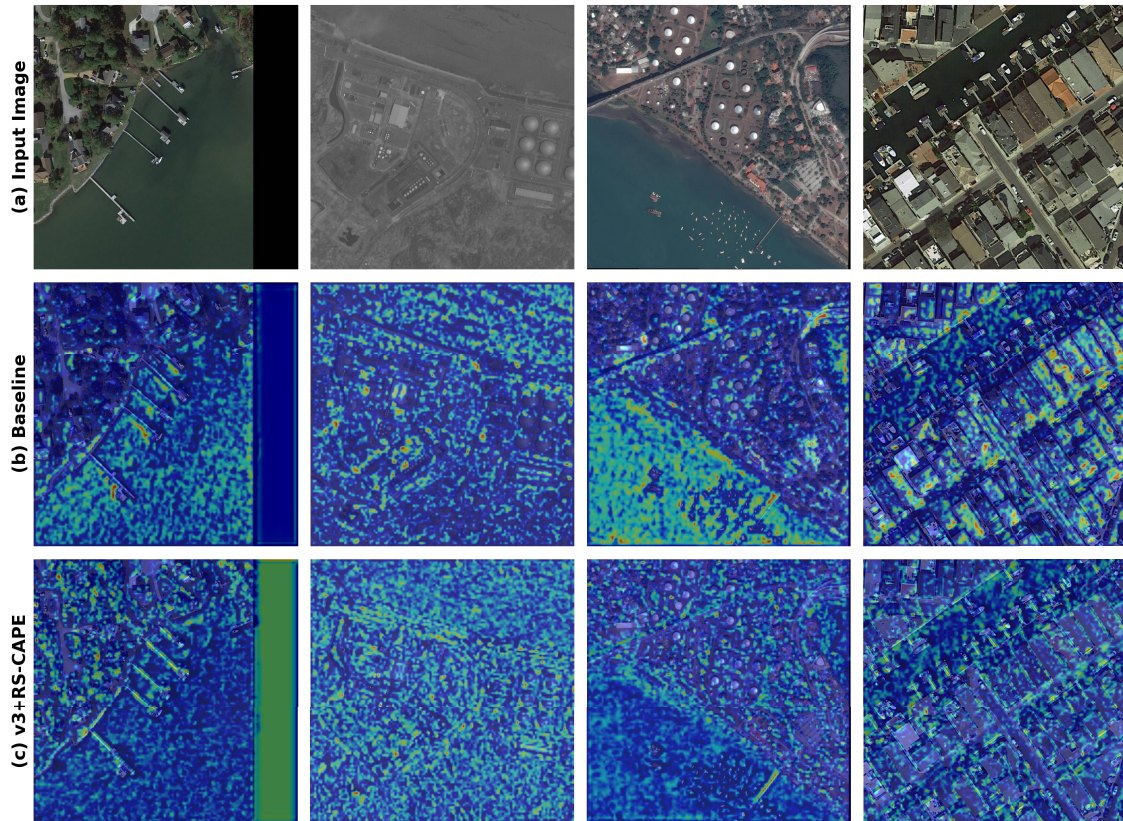


Figure 5: Neck feature activation heatmaps on four DOTA v1.0 patches. (a) Input; (b) Baseline; (c) v3 + RS-CAPE. Warm colors indicate high activation.

4.11 Seen/Unseen Split Analysis

To diagnose whether short-schedule RS fine-tuning preserves text-guided transfer, we partition DOTA’s 15 categories into **10 Seen** (all except the five below) and **5 Unseen** (ground-track-field, basketball-court, soccer-ball-field, roundabout, swimming-pool). Training filters annotations to Seen classes while updating *all* model components with no frozen parameters; evaluation supplies the full 15-class text prompt (including the five Unseen names, exactly as in the standard protocol) and collects predictions for all classes. [Table 9](#) reports the results.

Table 9: Seen/Unseen split evaluation on DOTA v1.0 (10 Seen/5 Unseen split, 15 epochs). Seen mAP is the mean over the 10 training categories; Unseen mAP is the mean over the 5 held-out categories, evaluated using their category names as text prompts. Overall mAP/mAP₅₀ are the standard COCO-style all-class metrics computed in a single 15-class evaluation, whereas the Seen/Unseen means are aggregated from per-class APs obtained in a separate class-wise evaluation; because the two aggregation protocols differ, the overall values are not the weighted averages of the Seen and Unseen columns. Best results in **bold**.

Configuration	Overall mAP	Overall mAP ₅₀	AP _s	AP _m	Seen mAP	Unseen mAP
Baseline (10 Seen)	0.297	0.474	0.238	0.306	0.476	≈0.000
LFR-GDINO (10 Seen)	0.311	0.502	0.244	0.315	0.498	≈0.000
Δ	+1.4 pp	+2.8 pp	+0.6 pp	+0.9 pp	+2.2 pp	—

Two findings emerge. **(1)** Seen-class gains persist: LFR-GDINO improves Seen mAP by +2.2 pp and mAP₅₀ by +2.8 pp, in line with the full 15-class trend. **(2)** Unseen-category transfer is not preserved: both configurations achieve ≈0 AP on all 5 Unseen categories. A plausible explanation is that updating all parameters on Seen-only supervision drives representation drift in the image encoder, text encoder, and fusion layers: Seen category tokens are repeatedly reinforced by gradient updates, whereas Unseen names receive no training signal and lose discriminative alignment with visual features. The positive_map itself is a fixed token-index routing structure and is not learned; the specialization occurs in the encoder and fusion parameters that it indexes into. This interpretation remains a hypothesis derived from the observed behavior; potential mitigations such as freezing the text encoder, open-vocabulary regularization, or RS-specific vision-language pre-training are left for future work.

4.12 Cross-Domain Evaluation: VisDrone-DET

We evaluate LFR-GDINO on VisDrone-DET [25] (drone-captured urban imagery, 10 categories), which differs from DOTA in viewing angle, resolution, and object density. [Table 10](#) reports the results.

Table 10: Cross-domain results on VisDrone-DET (15 epochs). Best results in **bold**.

Configuration	mAP	mAP ₅₀	AP _s	AP _m	AP _l
Baseline	0.237	0.413	0.159	0.334	0.513
LFR-GDINO	0.246	0.441	0.160	0.349	0.480
Δ	+0.9 pp	+2.8 pp	+0.1 pp	+1.5 pp	−3.3 pp

The mAP_{50} gain again suggests the improvement is not DOTA-specific, though the AP_l regression suggests that the benefits may be dataset-dependent. The recurring gap between mAP_{50} gains and stagnant COCO mAP reflects that mAP_{50} measures detection sensitivity, whereas COCO mAP penalizes imprecise localization; although InnerMPDIoU directly targets box regression during training, the single-run results show no consistent improvement in high-IoU localization, so the observed gains may be dominated by detection confidence rather than by box regression.

4.13 Backbone Study: Swin Transformer vs. ResNet-50

To study backbone effects, we replace ResNet-50 with Swin-T [31] (ImageNet-22k pre-trained) under the same 15-epoch schedule. Table 11 reports the results.

Table 11: Backbone study on DOTA v1.0 (15 epochs, LFR-GDINO). The params column reports backbone-only parameter counts; the full LFR-GDINO model comprises approximately 169.3 M parameters in total (Table 5). Best results in **bold**.

Backbone	Backbone Params (M)	mAP	mAP_{50}	mAP_{75}	AP_s	AP_m	AP_l
ResNet-50 (main)	25.6	0.420	0.695	0.479	0.286	0.423	0.467
Swin-T	28.3	0.314	0.538	0.355	0.231	0.323	0.353
Δ	+2.7	-10.6 pp	-15.7 pp	-12.4 pp	-5.5 pp	-10.0 pp	-11.4 pp

Swin-T underperforms by -10.6 pp mAP, with training curves still improving at epoch 15 (0.039 $\rightarrow$ 0.314), indicating under-fitting rather than a capacity ceiling. The domain gap between ImageNet-22k and RS imagery, combined with Swin-T’s window attention adding complexity without benefit over the existing deformable attention, may explain this result. Realizing stronger backbones’ potential would require RS-domain pre-training or longer schedules (36–50 epochs). Note that the parameter counts in Table 11 refer only to the backbone (25.6 M for ResNet-50 and 28.3 M for Swin-T), whereas the approximately 169.3 M parameters reported in Table 5 count the entire LFR-GDINO model, including the BERT text encoder, fusion modules, and detection head.

5 Discussion

5.1 Positioning within the RS Detection Landscape

LFR-GDINO operates in a different paradigm from closed-set OBB detectors (Oriented R-CNN, LSKNet): it offers a text-query interface at inference time without retraining. However, after the adopted fine-tuning the model does not retain reliable detection capability for arbitrary categories outside the trained set (Section 4.11); the text-query interface is therefore meaningful mainly for categories covered during adaptation. Because the two paradigms employ incompatible evaluation protocols (OBB vs. HBB) and distinct detection heads, numerical comparisons are cross-paradigm reference points rather than direct rankings; Table 12 provides such a reference.

5.2 Scale Trade-off and Computational Overhead

A recurring observation across all experiments is that AP_s/AP_m improvements co-occur with AP_l degradation. We hypothesize that this stems from DETR’s shared multi-scale representation: ECA’s channel re-weighting amplifies features discriminative for small objects, which occupy the high-frequency channels, at the partial expense of the low-frequency channels that encode large-object context. This explanation

remains qualitative; a definitive account would require per-level gradient analysis or attention disentanglement, which we leave for future work. A practical mitigation—decoupled per-level attention with independent parameters—was not explored under the current budget.

Table 12: Cross-paradigm reference on DOTA v1.0. Closed-set OBB methods use the official OBB evaluation; text-guided HBB methods use COCO HBB protocol. Results are **not directly comparable**. Best results in **bold**.

Method	Paradigm	Backbone	Box Type	mAP ₅₀	Text-Query
Oriented R-CNN [3]	Closed OBB	R-50	Oriented	75.87	×
LSKNet-S [6]	Closed OBB	LSKNet	Oriented	81.64	×
Baseline (Ours)	Text-guided HBB	R-50	Horizontal	67.0	✓
LFR-GDINO (Ours)	Text-guided HBB	R-50	Horizontal	69.5	✓

Regarding computational overhead, LFR-GDINO uses 169.3 M parameters and 309.0 GFLOPs vs. 172.4 M/414.8 GFLOPs for the original GroundingDINO [11]; the ECA module adds <0.3 KB per feature level (negligible), and InnerMPDIoU modifies only the loss computation. The FPS increase (7.7 → 13.1) is attributable to the Efficient G-DINO backbone implementation [13] rather than to our modifications. Under the same base configuration, Baseline, v3, RS-CAPE, and LFR-GDINO differ only by the ECA kernels and the text input, so their parameter counts, FLOPs, and inference cost are essentially identical; the row differences in Table 5 stem from different implementation codebases. Peak GPU memory and training time were not logged per variant under the current budget, which we acknowledge as a limitation.

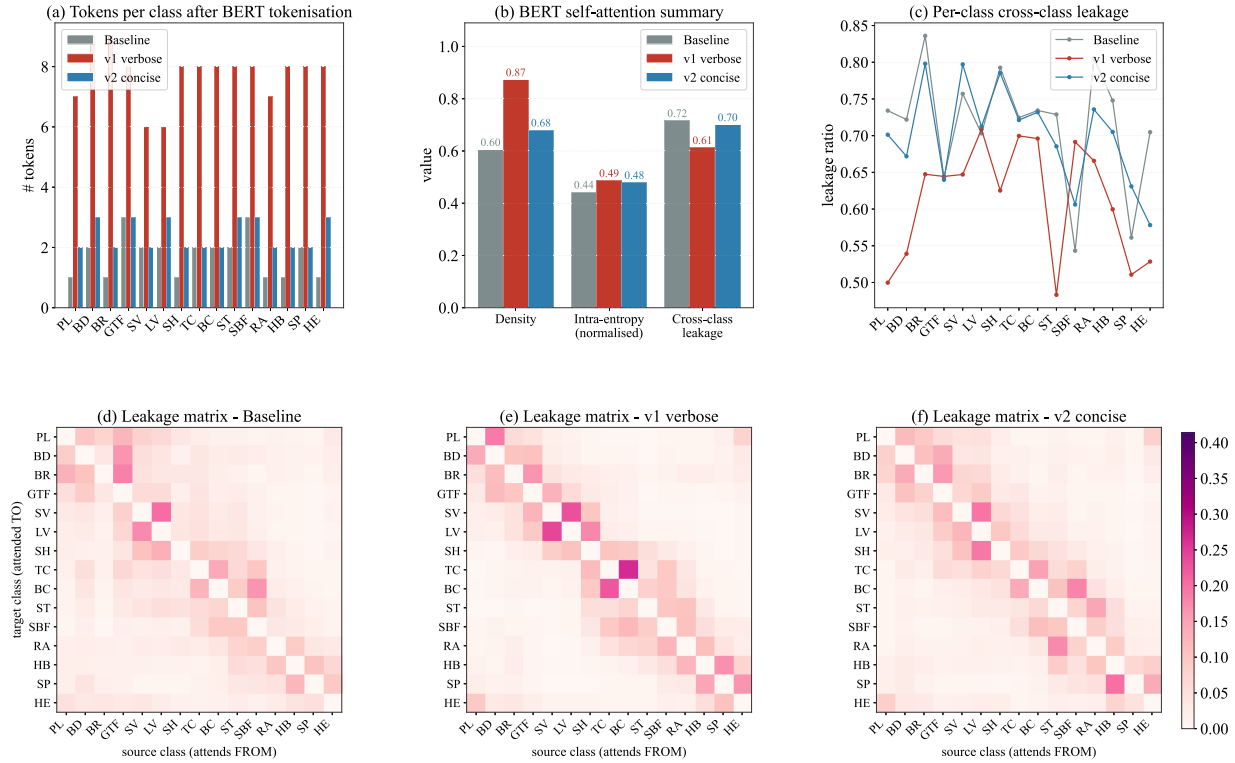
5.3 Token Capacity Diagnostic on BERT Self-Attention

A plausible mechanism behind the trade-off is attention dilution: GroundingDINO concatenates all category names into a single text sequence, so as a category occupies more tokens, the per-token cross-modal attention weight tends to decrease. To examine this independently of full detector training, we encode the three prompt regimes with the same frozen BERT-base-uncased [28] and inspect the last-layer self-attention. We define three metrics: **Density** $\rho = n_{\text{cls}}/L$ (fraction of input length occupied by class tokens); **Intra-class entropy** H (length-normalised Shannon entropy of each class-token’s outgoing attention, averaged); and **Cross-class leakage** ℓ (fraction of incoming attention from other categories’ tokens).

Three observations are consistent with the detection–precision trade-off (Table 13, Figs. 6 and 7); we emphasize that these are correlational findings from attention visualization, not causal evidence; controlled intervention experiments (e.g., attention masking) would be needed to establish causality. First, v1 inflates the sequence length by 3.1× ($\rho = 0.872$) with higher intra-class entropy ($H = 0.486$ vs. 0.442), consistent with diffuse per-token attention coinciding with reduced mAP₇₅. Second, v1 achieves the lowest leakage (0.612) due to a “crowding-out” effect, yet this alone does not improve overall detection. Third, v2 occupies an intermediate position on every axis, mirroring its balanced detection trade-off and suggesting 2–3 tokens per class as a favourable operating point under the current architecture. We note that this diagnostic covers only three manually designed prompt regimes; whether the trends generalize to finer-grained token budgets should be verified with additional controlled configurations in future work.

Table 13: Token capacity diagnostics on frozen BERT-base-uncased over the 15-category DOTA prompt. Detection metrics (mAP_{75}) are from Table 3. Best results in **bold**.

Strategy	Seq. Len. L	Class Tok.	Density ρ	Entropy H	Leakage ℓ	mAP_{75}
Baseline	43	26	0.605	0.442	0.716	0.483
v1 verbose (5–8)	133	116	0.872	0.486	0.612	0.467
v2 concise (2–3)	53	36	0.679	0.479	0.700	0.476

**Figure 6:** Token capacity diagnostics on frozen BERT-base-uncased. (a) Per-class token counts. (b) Aggregated diagnostics. (c) Per-class leakage curves. (d–f) Cross-class leakage matrices. Axis labels use the category abbreviations of Table 5, except that harbor and helicopter appear as HB and HE in the panel axes.

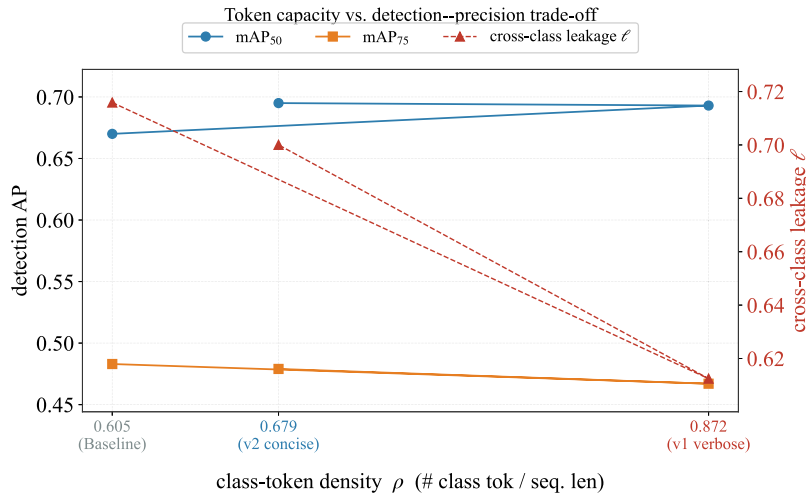


Figure 7: Token capacity vs. detection-precision trade-off on DOTA v1.0. Class-token density ρ is plotted against mAP₅₀, mAP₇₅, and the cross-class leakage ℓ . v1 maximises mAP₅₀ at the cost of mAP₇₅; v2 recovers much of the localisation precision while retaining a positive mAP₅₀ gain over the baseline.

6 Conclusions

This paper presents an empirical engineering study on adapting GroundingDINO to remote sensing imagery under practical resource constraints (dual-GPU, 15-epoch schedule). Through end-to-end training on DOTA v1.0, we observe that the targeted architectural modifications of v3 (ECA neck attention with InnerMPDIoU loss) combined with the concise RS-CAPE prompts improve the adapted detector in the single runs reported here, attaining the highest mAP₅₀ among the compared methods under the text-guided HBB protocol and the best per-class AP₅₀ on 8/15 DOTA categories; because these values come from single runs, they are reported as indicative outcomes rather than statistically confirmed gains.

A detection-precision trade-off associated with prompt design is empirically observed: concise 2–3-token supplements per class yield the most favourable balance, supported by a training-free token-capacity diagnostic (interpreted as correlational evidence rather than causal proof). Transfer experiments on DIOR, DIOR-RSVG, and VisDrone-DET suggest that the mAP₅₀ gain is not DOTA-specific, although overall mAP exhibits dataset-dependent variation. A Seen/Unseen split shows that short-schedule fine-tuning improves Seen-class adaptation but does not preserve Unseen-category transfer, and the backbone study indicates that Swin-T requires longer schedules to outperform ResNet-50 under the adopted budget.

Future directions include extended schedules or RS-domain pre-training for larger backbones, automated prompt search via learnable soft prompts, oriented bounding box detection, and preserving unseen-category transfer through open-vocabulary objectives. Beyond numerical gains, this work contributes a practical prompt guideline (2–3 tokens per class), a training-free token-level diagnostic framework, and an empirical characterization of short-schedule text-guided adaptation. We acknowledge that all reported results are from single training runs; repeated experiments with confidence intervals would strengthen the statistical basis of these findings.

Acknowledgement: Not applicable.

Funding Statement: This research received no external funding.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Angran Li; methodology, Angran Li; software, Angran Li; validation, Angran Li; formal analysis, Angran Li; investigation, Angran Li; resources, Angran Li; data curation, Angran Li; writing—original draft preparation, Angran Li; writing—review and editing, Chao Tang; supervision, Chao Tang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data presented in this study are derived from the following resources available in the public domain:

- DOTA v1.0 [22]: <https://captain-whu.github.io/DOTA/>
- DIOR [23]: <https://gcheng-nwpu.github.io/#Datasets>
- DIOR-RSVG [24]: <https://github.com/ZhanYang-nwpu/RSVG-pytorch>
- VisDrone-DET [25]: <https://github.com/VisDrone/VisDrone-Dataset>

No new data were created in this study.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A Complete RS-CAPE Prompt Templates

RS-CAPE substitutes each bare category name in the text prompt with a short domain-specific phrase and leaves the remainder of the pipeline unchanged. For a dataset with C categories, the phrases are concatenated in the class order of the corresponding annotation file using the period separator adopted by GroundingDINO, that is, “phrase₁. phrase₂. . . . phrase_C.”. The substitution is realised as a data-pipeline transform placed in both the training and the test pipeline, so that training and evaluation always operate on identical prompts; a category that is absent from a mapping table keeps its original name. Because the positive_map is rebuilt from the BERT tokenization of the concatenated prompt at run time, no token index has to be maintained manually. Tables A1–A3 list every mapping used in this study. Entries whose bare class name is already an unambiguous RS term (e.g., tennis-court, swimming-pool) are retained unchanged apart from the removal of the hyphen, which is why a few phrases coincide with the original class name.

Table A1: Complete RS-CAPE mappings for the 15 DOTA v1.0 categories. Column 2 corresponds to the verbose variant (v1) and column 3 to the concise variant (v2) adopted by LFR-GDINO.

Category	RS-CAPE v1 (Verbose)	RS-CAPE v2 (Concise)
Plane	Plane airplane aircraft jet fighter aerial aviation	Airplane aircraft
Baseball-diamond	Baseball diamond field sports batting infield ground area	Diamond baseball field
Bridge	Bridge overpass span crossing viaduct structure river road	Long bridge
Ground-track-field	Ground track field athletics oval running stadium course	Oval track field
Small-vehicle	Small vehicle sedan automobile car compact	Small car
Large-vehicle	Large vehicle truck bus transport freight	Large truck bus
Ship	Ship vessel boat cargo marine watercraft sailing	Ship boat
Tennis-court	Tennis court rectangular sports ground playing surface net	Tennis court
Basketball-court	Basketball court rectangular sports ground playing area hoop	Basketball court
Storage-tank	Storage tank circular container petroleum fuel oil depot	Circular storage tank
Soccer-ball-field	Soccer football field sports grass pitch ground rectangular	Soccer football field
Roundabout	Roundabout circular traffic intersection road junction ring	Circular roundabout
Harbor	Harbor port dock marina wharf pier berth waterfront	Harbor port

(Continued)

Table A1 (continued)

Category	RS-CAPE v1 (Verbose)	RS-CAPE v2 (Concise)
Swimming-pool	Swimming pool rectangular water facility blue aquatic leisure	Swimming pool
Helicopter	Helicopter rotorcraft chopper rotor aircraft aerial landing	Helicopter rotorcraft

Table A2: Complete RS-CAPE mappings for the 20 DIOR categories, listed as two category–phrase pairs per row. Following the original DIOR class list, ESA and ETS abbreviate the expressway service area and the expressway toll station, respectively.

Category	RS-CAPE Phrase	Category	RS-CAPE Phrase
Airplane	Airplane aircraft	Groundtrackfield	Oval track field
Airport	Airport runway	Harbor	Harbor port
Baseballfield	Baseball diamond field	Overpass	Overpass flyover
Basketballcourt	Basketball court	Ship	Ship vessel
Bridge	Long bridge	Stadium	Stadium arena
Chimney	Tall chimney smokestack	Storagetank	Circular storage tank
Dam	Dam reservoir	Tenniscourt	Tennis court
ESA	Expressway service area	Trainstation	Train railway station
ETS	Expressway toll station	Vehicle	Small vehicle car
Golffield	Golf course field	Windmill	Wind turbine windmill

Table A3: Complete RS-CAPE mappings for the 10 VisDrone-DET categories, listed as two category–phrase pairs per row.

Category	RS-CAPE Phrase	Category	RS-CAPE Phrase
Pedestrian	Walking pedestrian	Truck	Large truck
People	Standing person	Tricycle	Tricycle three-wheel
Bicycle	Bicycle bike	Awning-tricycle	Covered awning tricycle
Car	Small car sedan	Bus	Long bus
Van	Van minibus	Motor	Motorcycle motorbike

For direct reuse, the concatenated prompt strings are reproduced below; the DOTA v1.0 RS-CAPE v2 string is already given in [Section 3.4](#).

- **DOTA v1.0, RS-CAPE v1 (verbose):** “plane airplane aircraft jet fighter aerial aviation. baseball diamond field sports batting infield ground area. bridge overpass span crossing viaduct structure river road. ground track field athletics oval running stadium course. small vehicle sedan automobile car compact. large vehicle truck bus transport freight. ship vessel boat cargo marine watercraft sailing. tennis court rectangular sports ground playing surface net. basketball court rectangular sports ground playing area hoop. storage tank circular container petroleum fuel oil depot. soccer football field sports grass pitch ground rectangular. roundabout circular traffic intersection road junction ring. harbor port dock marina wharf pier berth waterfront. swimming pool rectangular water facility blue aquatic leisure. helicopter rotorcraft chopper rotor aircraft aerial landing.”

- **DIOR, RS-CAPE (concise):** “airplane aircraft. airport runway. baseball diamond field. basketball court. long bridge. tall chimney smokestack. dam reservoir. expressway service area. expressway toll station. golf course field. oval track field. harbor port. overpass flyover. ship vessel. stadium arena. circular storage tank. tennis court. train railway station. small vehicle car. wind turbine windmill.”
- **VisDrone-DET, RS-CAPE (concise):** “walking pedestrian. standing person. bicycle bike. small car sedan. van minibus. large truck. tricycle three-wheel. covered awning tricycle. long bus. motorcycle motorbike.”

For DIOR-RSVG no category mapping applies, because the text input consists of free-form referring expressions supplied by the dataset rather than of category names; the prompts are therefore used exactly as released with the dataset.

References

1. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *Adv Neural Inf Process Syst.* 2015;28(6):1–9. doi:10.1109/tpami.2016.2577031.
2. Ding J, Xue N, Long Y, Xia GS, Lu Q. Learning RoI transformer for oriented object detection in aerial images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2019 Jun 3–7; Long Beach, CA, USA. p. 2849–58.
3. Xie X, Cheng G, Wang J, Yao X, Han J. Oriented R-CNN for object detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; 2021 Oct 19–23; Montreal, QC, Canada. p. 3520–9.
4. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2016 Jun 27–30; Las Vegas, NV, USA. p. 779–88.
5. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; 2017 Oct 22–29; Venice, Italy. p. 2980–8.
6. Li Y, Hou Q, Zheng Z, Cheng MM, Yang J, Li X. Large selective kernel network for remote sensing object detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; 2023 Oct 1–6; Paris, France. p. 16794–805.
7. Guo Z, Liu C, Zhang X, Jiao J, Ji X, Ye Q. Beyond bounding-box: convex-hull feature adaptation for oriented and densely packed object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2021 Jun 20–25; Nashville, TN, USA. p. 8792–801.
8. Li LH, Zhang P, Zhang H, Yang J, Li C, Zhong Y, et al. Grounded language-image pre-training. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2022; New Orleans, LA, USA. p. 10965–75.
9. Minderer M, Gritsenko A, Stone A, Neumann M, Weissenborn D, Dosovitskiy A, et al. Simple open-vocabulary object detection. In: *Proceedings of the European Conference on Computer Vision (ECCV)*; 2022 Oct 23–27; Tel Aviv, Israel. p. 728–55.
10. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: *Proceedings of the International Conference on Machine Learning (ICML)*; 2021 Jul 18–24; Virtual, pp. 8748–63.
11. Liu S, Zeng Z, Ren T, Li F, Zhang H, Yang J, et al. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In: *Proceedings of the European Conference on Computer Vision (ECCV)*; 2024 Sep 29–Oct 4; Milan, Italy.
12. Zhang H, Li F, Liu S, Zhang L, Su H, Zhu J, et al. DINO: DETR with improved deNoising anchor boxes for end-to-end object detection. In: *Proceedings of the International Conference on Learning Representations (ICLR)*; 2023 May 1–5; Kigali, Rwanda.
13. Hu Z, Gao K, Zhang Z, Yang Z, Wang J, Zhu Z, et al. Efficient Grounding DINO: efficient cross-modality fusion and efficient label assignment for visual grounding in remote sensing. *IEEE Trans Geosci Remote Sens.* 2025;63:1–14.

14. Zhou K, Yang J, Loy CC, Liu Z. Learning to prompt for vision-language models. *Int J Comput Vis.* 2022;130(9):2337–48. doi:10.1007/s11263-022-01653-1.
15. Zhou K, Yang J, Loy CC, Liu Z. Conditional prompt learning for vision-language models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2022 Jun 19–24; New Orleans, LA, USA. p. 16816–25.
16. Du Y, Wei F, Zhang Z, Shi M, Gao Y, Li G. Learning to prompt for open-vocabulary object detection with vision-language model. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2022 Jun 19–24; New Orleans, LA, USA. p. 14084–93.
17. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-Net: efficient channel attention for deep convolutional neural networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2020 Jun 13–19; Seattle, WA, USA. p. 11534–42.
18. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: *Proceedings of the European Conference on Computer Vision (ECCV)*; 2018 Sep 8–14; Munich, Germany. p. 3–19.
19. Rezaatofghi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized intersection over union: a metric and a loss for bounding box regression. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2019 Jun 15–20; Long Beach, CA, USA. p. 658–66.
20. Zhang H, Xu C, Zhang S. Inner-IoU: more effective intersection over union loss with auxiliary bounding box. *arXiv:2311.02877.* 2023.
21. Ma S, Xu Y. MPDIoU: a loss for efficient and accurate bounding box regression. *arXiv:2307.07662.* 2023.
22. Xia GS, Bai X, Ding J, Zhu Z, Belongie S, Luo J, et al. DOTA: a large-scale dataset for object detection in aerial images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 3974–83.
23. Li K, Wan G, Cheng G, Meng L, Han J. Object detection in optical remote sensing images: a survey and a new benchmark. *ISPRS J Photogramm Remote Sens.* 2020;159:296–307.
24. Zhan Y, Xiong Z, Yuan Y. RSVG: exploring data and models for visual grounding on remote sensing data. *IEEE Trans Geosci Remote Sens.* 2023;61:1–13.
25. Zhu P, Wen L, Du D, Bian X, Fan H, Hu Q, et al. Detection and tracking meet drones challenge. *IEEE Trans Pattern Anal Mach Intell.* 2021;44(11):7380–99. doi:10.1109/tpami.2021.3119563.
26. Chen K, Wang J, Pang J, Cao Y, Xiong Y, Li X, et al. MMDetection: open MMLab detection toolbox and benchmark. *arXiv:1906.07155.* 2019.
27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2016 Jun 26–Jul 1; Las Vegas, NV, USA. p. 770–8.
28. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT*; 2019 Jun 3–5; Minneapolis, MN, USA. p. 4171–86.
29. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: Deformable transformers for end-to-end object detection. In: *Proceedings of the International Conference on Learning Representations (ICLR)*; 2021 May 3–7; Virtual.
30. Hu Z, Gao K, Wang J, Yang Z, Zhang Z, Cheng H, et al. Enhanced grounding DINO: efficient cross-modality block for open-set object detection in remote sensing. *IEEE J Sel Top Appl Earth Obs Remote Sens.* 2025;18:15291–303.
31. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; 2021 Oct 11–17; Montreal, QC, Canada. p. 10012–22.