SETOPD: FROM FEW VISUAL EXEMPLARS TO MULTI-MODAL CANDIDATE SETS FOR REMOTE-SENSING OPEN-PROMPT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-prompt detection (OPD) allows users to specify targets with text, visual exemplars, or both. Existing multimodal OPD typically compresses multiple exemplars into a class-level prototype and combines text and vision in prompt or representation space, leaving modality-specific candidate states unavailable for explicit cross-modal comparison. We formulate these two limitations from a set perspective. SetOPD first introduces Base–Residual prompting, which reads boxed exemplars in scene context and maps pooled support evidence into a fixed-capacity anchor–correction visual memory that drives a dedicated visual pathway. We further introduce Paired-Query Arbitration (PQA), in which text and visual readers decode from a shared prompt-conditioned initialization, producing naturally paired candidate states that are arbitrated within each pair and subsequently refined as a set. Across 11 heterogeneous remote-sensing sources, including four fully held-out datasets, PQA outperforms representation-level joint prompting on every source and raises Macro11 AP from 54.00 to 55.57. It also improves over the text and Base–Residual visual readers by 1.66 and 2.02 AP, respectively, while adding only 0.63M trainable parameters. Detection-level analysis further shows that the two readers recover complementary objects and that PQA preserves a substantial fraction of this modality-exclusive evidence.

1 INTRODUCTION

Open-prompt detection (OPD) allows a user to tell a detector what to find through category names, visual exemplars, or both (Huang et al., 2025). It complements open-vocabulary detection (OVD) (Zareian et al., 2021): OVD asks what supervision a category received during training, whereas OPD asks through which interface a user specifies the target at inference time. This interface matters in remote sensing, where detectors are increasingly trained on unions of heterogeneous datasets with inconsistent taxonomies (Pan et al., 2025; Huang et al., 2025), so a base/novel split inside one dataset no longer indicates whether a concept was truly unseen and direct target specification becomes a central requirement (Huang et al., 2025; Yang et al., 2026).

Visual exemplars are the most direct way to specify a target: a few real instances show the detector what to look for, including appearance that a category name cannot express—which matters in remote sensing, where the same category varies substantially with scale, orientation, imaging conditions, and background. We argue that current open-prompt detectors underuse this modality in two ways, and that both are most naturally understood from a set perspective (Figure 1). The input to visual prompting is a support set, an unordered set of K boxed exemplars; when text is also provided, each modality can form its own set of candidate detections; and the output is itself a set, predicted by DETR-style detectors with a permutation-invariant loss (Carion et al., 2020; Zhang et al., 2023).

The first issue is how a support set becomes a prompt. Exemplar-based category interfaces are well established (Zang et al., 2022; Li et al., 2024; Jiang et al., 2024; Wang et al., 2025; Fu et al., 2026b), but given multiple exemplars, existing approaches typically compress them into a single class-level embedding through averaging, selection, or prototype aggregation (Yang et al., 2026; Jiang et al., 2024; Qian et al., 2026). The visual prompt is thereby textualized: one vector plays the role of a category name, often aligned with or injected into the text pathway (Jiang et al., 2024; Xu et al., 2023;

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.

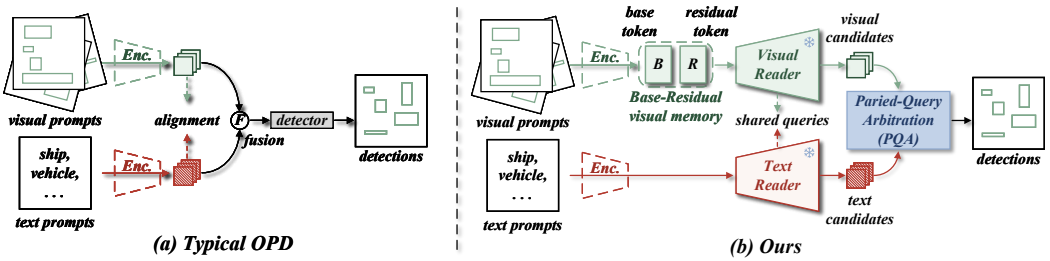

Figure 1: **Typical open-prompt detectors versus SetOPD.** (a) Existing detectors compress the support set into a single class-level embedding—the visual prompt is textualized—and interact in prompt or representation space. (b) SetOPD reads a few exemplars in scene context into a fixed-capacity visual memory, lets text and visual readers decode modality-specific candidate states from a shared prompt-conditioned initialization, and combines the two modalities through candidate-state collaboration.

Wu et al., 2025). From a set perspective, the issue is not that pooling produces a single vector—any fixed-size prompt must pool—but that the pooled vector is consumed as a category-level prototype, whose effectiveness can be sensitive to support-set size and composition in the few-example regime.

We instead keep evidence gathering visual and refuse the prototype use of its result. Base–Residual prompting reads every boxed support instance in its full scene context, so the evidence retains target position, scale, and surrounding context, and encodes any support set into a permutation-invariant prompt of $R + 1$ tokens: the pooled evidence is decomposed into a Base anchor plus a learnable Residual correction rather than used directly as a category vector, and it drives its own detection pathway rather than acting as an accessory to the text pathway. With no more than ten exemplars, this decomposition suffices: one correction raises Visual-only Macro11 AP by 6.4 points over the anchor alone, whereas routing examples into multiple competing slots brings no further gain.

The second issue is how to combine text and visual modalities. Text provides stable category semantics, while exemplars capture appearance under the current imaging conditions, but existing methods combine them early through visually enhanced language queries, aligned prompt spaces, joint multimodal prompts, or encoder/decoder-level interaction (Yang et al., 2026; Jiang et al., 2024; Xu et al., 2023; Wu et al., 2025; Guan et al., 2025; Qin et al., 2026), leaving modality-specific decoded candidate states unavailable for explicit comparison and arbitration. Recent remote-sensing results also show that representation-level multimodal fusion can degrade visual prompting when textual cues are mismatched (Yang et al., 2026). Complementarity between modalities therefore does not imply that their representations should be merged directly.

Starting from a shared prompt-conditioned proposal initialization, we preserve separate text- and visual-conditioned decoding streams and perform explicit cross-modal state interaction only after modality-specific candidate states have been formed. This turns fusion into a set problem: deciding which text candidate should be compared with which visual candidate is itself an assignment problem. Paired-Query Arbitration (PQA) resolves it structurally—text and visual readers decode from the same initial queries and reference boxes, so the i-th candidates form a pair without any matching step, preserved under arbitrary query permutations (Proposition 2, Appendix B). A learned gate reads both hidden states, category confidence, and geometric disagreement to arbitrate within each pair, followed by a permutation-equivariant module that reasons over the fused candidate set. Our detection-level analysis further shows that the two readers recover non-overlapping ground-truth objects and that PQA preserves a substantial fraction of these modality-exclusive detections across diverse sources.

In short, SetOPD treats multimodal open-prompt detection as candidate-state collaboration: from a shared prompt-conditioned proposal initialization, text and visual memories produce modality-specific detection states, which PQA arbitrates pairwise and refines jointly as a candidate set. The two components follow the same principle—retain modality-specific pathways until candidate detections are formed: exemplars are treated as visual evidence rather than category names, and text and vision are combined as two candidate sets rather than collapsed into one fused representation.

Our contributions are threefold:

- **Set-based formulation.** We identify two structural bottlenecks in multimodal OPD—prototype-style visual prompting and representation-level cross-modal fusion—and formulate support encoding and multimodal collaboration as two set problems.
- **Base–Residual prompting.** We map few boxed exemplars into a fixed-capacity anchor–correction visual memory that drives a dedicated visual pathway; the default two-token representation raises Visual-only Macro11 AP from 47.11 to 53.55 over the Base-only control.
- **Paired-Query Arbitration.** We exploit shared initialization to pair modality-specific candidate states, arbitrate each pair, and refine the fused set; PQA outperforms representation-level joint prompting on all 11 sources ($+1.57$ Macro11 AP) with only 0.63M trainable parameters.

2 Related Work

2.1 Open-Vocabulary and Open-Prompt Detection

Open-vocabulary detection uses vision–language pretraining to recognize categories that do not receive bounding-box supervision (Zareian et al., 2021). Language-conditioned detectors such as GLIP and Grounding DINO build transferable category interfaces through region–text alignment (Li et al., 2022; Liu et al., 2024), and remote-sensing detectors increasingly train on unions of heterogeneous datasets (Pan et al., 2025; Huang et al., 2025). Open-prompt detection instead focuses on the interface through which a target is specified. OV-DETR and OWL-ViT condition detection on image exemplars (Zang et al., 2022; Minderer et al., 2022); DINOv, T-Rex2, and YOLOE build visual or unified prompt interfaces (Li et al., 2024; Jiang et al., 2024; Wang et al., 2025); OpenRSD defines remote-sensing open-prompt detection with text and image prompts (Huang et al., 2025); and RS-MPOD combines text, visual instances, and their joint representation to specify remote-sensing targets (Yang et al., 2026).

2.2 Visual Prompting and Set Encoding

Visual prompting has been extended to remote sensing, real-time unified detection, Grounding-DINO-style frameworks, and few-image open-set perception (Huang et al., 2025; Wang et al., 2025; Fu et al., 2026b; Zhang et al., 2025). Given multiple exemplars, existing approaches commonly compress them by aggregation, prototype construction, or selection (Yang et al., 2026; Jiang et al., 2024; Qian et al., 2026). DETR-ViP further emphasizes prompt consistency, discriminability, and selective fusion (Qian et al., 2026). These methods treat exemplars as a group of objects to summarize; we instead treat them as a set whose encoding should satisfy explicit structural properties (permutation invariance and fixed capacity).

Set learning. Permutation-invariant set functions admit sum decompositions (Zaheer et al., 2017), whose simplest instance underlies prototype-based classification (Snell et al., 2017); learned variants include seeded attention (Lee et al., 2019), slot competition (Locatello et al., 2020), and residual aggregation around learnable centers (Arandjelovic et al., 2016). Base–Residual prompting reuses these building blocks but differs in three ways: it pools deviations from a support-dependent anchor rather than raw inputs or residuals from fixed centers; it targets few-shot support sets with at most ten exemplars and one to three deterministic slot queries, so the default two-token prompt is an anchor and one correction rather than competing object slots; and it is trained only through detection losses and consumed directly as a detection prompt.

2.3 Fusing Text and Visual Evidence

Existing multimodal detectors typically combine text and visual evidence through prompt-space alignment, visually enhanced language queries, or feature-level interaction (Xu et al., 2023; Jiang et al., 2024; Wu et al., 2025; Yang et al., 2026; Guan et al., 2025; Fu et al., 2026b; Qin et al., 2026). Unlike these representation-level approaches, PQA delays explicit cross-modal state interaction until modality-specific candidate states have been formed.

Related ideas also appear in later-stage evidence fusion outside this exact setting. F-VLM fuses region scores from different models (Kuo et al., 2023), DeepInteraction preserves modality-specific repre-

sentations while allowing interaction (Yang et al., 2022), and MS-DETR models sensor contributions through loosely coupled fusion (Xing et al., 2024).

Set prediction. DETR formulates detection as set prediction with permutation-invariant matching (Carion et al., 2020), and DINO further develops this formulation through denoising and improved anchor boxes (Zhang et al., 2023). Relation Networks and Learning NMS show that relations among candidate hypotheses can affect final decisions (Hu et al., 2018; Hosang et al., 2017). PQA combines these ideas at the candidate-set level: it operates on two candidate sets, establishes pairing through shared initialization rather than an explicit matching step, and applies a permutation-equivariant module to reason over the fused set.

3 METHOD

3.1 OPEN-PROMPT DETECTION AS A SET-TO-SET MAPPING

For a query image I and task category set $\mathcal{C}$, each category c is specified by a text description t_c, a support set of K boxed exemplars

$$\mathcal{S}_c = \{(I_c^k, b_c^k)\}_{k=1}^K,$$

or both. We treat $\mathcal{S}_c$ as a set: its order carries no information, while K is chosen by the user. The detector maps these prompts to a detection set through three set-valued stages:

$$\begin{aligned} V_c &= \Phi_{\mathrm{BR}}(\mathcal{S}_c), \qquad |V_c| = R+1 \quad \forall K, \\ \{s_i^T\}_{i=1}^N &= \Psi_T\big(I, \{t_c\}_{c\in\mathcal{C}}, Q^0\big), \qquad \{s_i^V\}_{i=1}^N = \Psi_V\big(I, \{V_c\}_{c\in\mathcal{C}}, Q^0\big), \\ Q^F &= \Psi_{\mathrm{PQA}}\big(\{(s_i^T, s_i^V)\}_{i=1}^N\big), \qquad \hat{\mathcal{Y}} = \mathrm{Head}(Q^F). \end{aligned} \tag{1}$$

Here Φ_{BR} is the Base–Residual support-set encoder (Sec. 3.2); the text reader Ψ_T and the visual reader Ψ_V decode the same image from a shared initialization Q^0 of $N = 900$ queries, conditioned respectively on the text descriptions $\{t_c\}_{c\in\mathcal{C}}$ and the Base–Residual memories $\{V_c\}_{c\in\mathcal{C}}$; s_i^T and s_i^V are the detection states produced from the i-th initial query; and Ψ_{PQA} fuses paired states into the set Q^F, from which the prediction head outputs at most 300 detections (Sec. 3.3).

Figure 2 shows the overall architecture of SetOPD. This set view yields three requirements, verified in Appendix B: (a) Φ_{BR} should be invariant to exemplar order and its output size should not depend on K; (b) correspondence between the two candidate sets should not depend on query order; and (c) Ψ_{PQA} should be permutation equivariant, so the model output is a set and the Hungarian set loss is independent of query order.

3.2 BASE–RESIDUAL SUPPORT-SET ENCODING

Box-conditioned instance encoding. For each boxed support instance, a frozen prompt-independent image encoder extracts full-image multi-scale features $F_{ck}^{P3:P5}$. Beyond appearance, the box conveys target position, scale, and shape through the six-dimensional geometry descriptor

$$m_{ck} = [c_x, c_y, w, h, \log(w/h), \log(wh/(WH))],$$

where W and H are the image width and height. The first four normalized coordinates serve as reference points for multi-scale deformable attention (Zhu et al., 2021), which reads instance evidence from the full-image features; because each instance is read in its complete scene context rather than from an independently encoded crop, the readouts are context aware:

$$\begin{aligned} q_{ck}^{\mathrm{box}} &= \mathrm{LN}(\mathrm{MLP}_{\mathrm{box}}(m_{ck})), \\ e_{ck}^{(\ell,g)} &= \mathrm{LN}\Big(W_e\ \mathrm{MSDeformAttn}^{\ell,g}(q_{\mathrm{content}}, q_{ck}^{\mathrm{box}}, F_{ck}^{P3:P5}, m_{ck,1:4})\Big), \end{aligned} \tag{2}$$

with $\ell = 1, \ldots, L$ feature levels ($L = 3$ for P3–P5) and $g = 1, \ldots, G$ sampling points per level. Instead of collapsing the $L \cdot G$ sampled responses of an exemplar into a single token, the encoder retains them, after a shared projection W_e and normalization, as evidence vectors; the support set thus produces $E_c = \{e_{cn}\}_{n=1}^{N_e}$ with $N_e = KLG$, padded positions masked out.

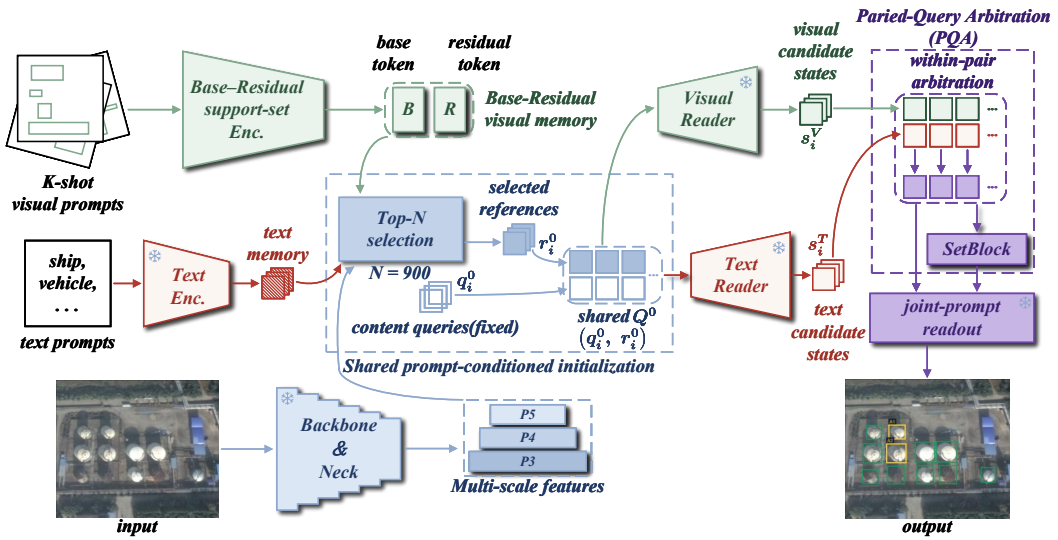

Figure 2: **SetOPD overview.** A support set of K boxed exemplars is encoded by the box-conditioned Base–Residual encoder into a fixed-capacity visual memory. Text and visual readers decode from a shared prompt-conditioned initialization, so their candidates are paired by index; PQA arbitrates within each pair, reasons over the fused candidate set, and the frozen prompt-conditioned head produces the final detections. The visual memory also supports the native Visual-only pathway.

Base–Residual support-set representation. The Base token summarizes the evidence set, and each Residual slot pools deviations from the Base:

$$\mu_c = \frac{1}{N_e}\sum_{n=1}^{N_e} e_{cn}, \qquad B_c = \mathcal{B}(\mu_c), \qquad R_{cr} = \mathcal{R}_r\Big(\{e_{cn} - \mathrm{sg}(B_c)\}_{n=1}^{N_e}\Big), \tag{3}$$

where $\mathcal{B}$ is a single-seed mean encoder (Lee et al., 2019; Zaheer et al., 2017) and $\mathcal{R}_r$ is a slot update that follows the normalization pattern of Slot Attention (Locatello et al., 2020) but pools deviations from a support-dependent anchor with deterministic slot queries and residual feed-forward steps rather than a GRU; $\mathrm{sg}(\cdot)$ denotes stop-gradient, and full parameterizations are given in Appendix A.7. The visual memory is

$$V_c = [B_c, R_{c1}, \ldots, R_{cR}], \qquad |V_c| = R + 1, \tag{4}$$

so the prompt size is always $R+1$, regardless of K. Our default is $R = 1$, giving two tokens (T2). T1, T3, and T4 use $R = 0, 2, 3$, respectively, and serve as structural controls in Table 1B.

Readout and the Visual-only pathway. For candidate query z_i, the visual memory yields a category score through gated Residual responses,

$$s^V_{ic} = z_i^\top B_c + \lambda_R \frac{\sum_r g_{icr}\, z_i^\top R_{cr}}{\sum_r g_{icr} + \epsilon}, \qquad g_{icr} = \sigma\left(\frac{z_i^\top R_{cr}}{\sqrt{d}\,\tau}\right), \tag{5}$$

with $\lambda_R = 0.5$ and $\tau = 1.0$; for the default $R = 1$, this reduces to the fixed combination $z_i^\top(B_c + \lambda_R R_{c1})$. The structured visual memory plugs directly into the detector's native prompt-conditioned pathway and drives category-dependent candidate scoring, selection of the top 900 queries, decoder–prompt interaction, and final classification, forming a complete Visual-only detector.

3.3 Paired-Query Arbitration over Candidate Sets

Shared prompt-conditioned initialization. Text and visual prompts are jointly used only to select the top-N proposal references, while the learnable content queries remain fixed. Both readers then decode from the same initialization $Q^0 = \{(q_i^0, r_i^0)\}_{i=1}^N$ against their own prompt memories, producing modality-specific states $s_i^m = (h_i^m, r_i^m, p_i^m, b_i^m)$ (hidden state, reference, class probabilities, and box), $m \in \{T, V\}$. Hence, same-index text and visual candidates form natural initialization pairs

without explicit cross-modal matching. This pairing is a structural prior rather than a hard object-level correspondence; an audit in Appendix A.3 shows that it is nevertheless highly consistent for valid detections.

Within-pair arbitration. For pair i, the arbitrator receives both hidden states, their absolute difference, the maximum class confidence from each branch, and the difference between the two references in inverse-sigmoid space—a 774-dimensional descriptor ($3 \times 256 + 2 + 4$, with hidden dimension $d = 256$)—and produces a scalar text weight g_i that modulates both the semantic state and the geometric reference:

$$\begin{aligned} x_i &= [h_i^T, h_i^V, |h_i^T - h_i^V|, \max_c p_i^T(c), \max_c p_i^V(c), \operatorname{logit}(r_i^T) - \operatorname{logit}(r_i^V)], \\ g_i &= \sigma(\mathrm{MLP}(x_i)), \qquad \mathrm{MLP} : 774 \to 128 \to 1, \\ h_i^P &= g_i h_i^T + (1 - g_i) h_i^V, \qquad r_i^P = \sigma\big(g_i \operatorname{logit}(r_i^T) + (1 - g_i) \operatorname{logit}(r_i^V)\big). \end{aligned} \tag{6}$$

A constant gate recovers fixed rules as special cases: text only ($g_i = 1$), visual only ($g_i = 0$), and mean fusion ($g_i = 1/2$); a saturated gate selects one candidate from each pair. The learned gate chooses among these behaviors according to the evidence in x_i.

From pairs to set-level reasoning. Pairwise arbitration processes each pair independently and therefore cannot account for other candidates, for example when two pairs describe the same object. We therefore treat the N fused states as a set (Carion et al., 2020; Zhang et al., 2023) and apply one set-reasoning block,

$$H^F = \mathrm{SetBlock}(H^P), \qquad H^P = [h_1^P, \ldots, h_N^P], \tag{7}$$

where $\mathrm{SetBlock}$ is one eight-head self-attention layer without index-dependent positional encoding followed by a $256 \to 512 \to 256$ FFN, both with residual connections and LayerNorm (full form in Appendix A.7); the references remain r_i^P.

Frozen joint-prompt readout. The final classifier does not combine the original unimodal reader scores. Instead, both modality-conditioned probabilities are recomputed from the same fused state h_i^F; box offsets are decoded from (h_i^F, r_i^P) by the frozen regression branch, where $R^P = [r_1^P, \ldots, r_N^P]$:

$$\begin{aligned} p_{ic}^{T,F} &= \frac{1}{|\mathcal{T}_c|} \sum_{j \in \mathcal{T}_c} \sigma\big(\ell^T(h_i^F, t_j)\big), \qquad p_{ic}^{V,F} = \sigma\big(\ell^V(h_i^F, V_c)\big), \\ s_{ic}^F &= \operatorname{logit}\Big[\operatorname{clip}\Big(\tfrac{1}{2} p_{ic}^{T,F} + \tfrac{1}{2} p_{ic}^{V,F},\ 10^{-7},\ 1 - 10^{-7}\Big)\Big], \end{aligned} \tag{8}$$

where ℓ^T is the frozen text-conditioned classifier averaged over the text tokens $\mathcal{T}_c$ of category c, and ℓ^V is the frozen Base–Residual readout of Eq. (5). The final 1:1 fusion is performed in probability space on this single fused state, and category semantics remain carried entirely by the prompt tokens.

Structural properties. The Base–Residual encoder is permutation invariant to exemplar order and has a fixed output size of $R + 1$ tokens for any K. With the default $R = 1$, it implements a mean-dependent anchor–correction factorization rather than instance-level routing. PQA is permutation equivariant with respect to the shared query ordering, and pairing by initialization costs $O(N)$ instead of the $O(N^3)$ assignment that aligning two independently decoded candidate sets would require. Formal statements and proofs are given in Appendix B.

3.4 Learning Objective

Training objective. We use the standard DETR Hungarian detection loss. In addition, for each matched fused query, we penalize its detection error when it exceeds that of the better paired unimodal reader:

$$\mathcal{L} = \mathcal{L}_{\mathrm{det}} + 0.1 \ \mathrm{mean}_i \max\big(0,\ e_i^F - \min(e_i^T, e_i^V)\big).$$

The unimodal errors are stop-gradient targets, so this term constrains only the fused state; full loss definitions and the staged training protocol, which concludes with a quality-focal-loss continuation producing the final model, are given in Appendix A.6.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Data. We construct an 11-source remote-sensing OPD evaluation benchmark under a unified 153-category semantic space. The training corpus and category harmonization follow the multi-source construction introduced in prior work (Hu et al., 2026) (175,644 training images over 153 categories, integrating sources with different taxonomies, annotation granularities, and scene distributions), while the evaluation protocol in this paper uses seven in-domain test sources—DIOR, DOTA-v2.0, FAIR1M, SIMD, SODA-A, ShipRSImageNet, and WHU-Buildings—and four fully held-out sources that are never used during training: HRSC2016, HRRSD, RSOD, and VEDAI. We therefore report Macro11 over seven in-domain test sources and four held-out sources. Main experiments report COCO-style AP, and comparisons with external methods also report AP50 when available. Support exemplars and query images always come from non-overlapping scenes. The experiments address two questions: how a few visual exemplars can form an effective visual prompt for a frozen detector (Sec. 4.3), and whether, beyond joint prompting, modality-specific candidate states expose complementary detection evidence that can be explicitly arbitrated (Sec. 4.4).

Implementation details. Training has three stages: Stage 0 adapts Grounding DINO (Liu et al., 2024) with a Swin-T backbone and BERT text encoder on the RS153 corpus into the prompt-neutral base endpoint, which is frozen thereafter; Stage I trains only the Base–Residual visual-prompt encoder; and Stage II trains only the PQA pair gate and one-layer set adapter. We use $K = 5$ boxed exemplars per category. Full optimization details—iterations, batch sizes, optimizer, seeds, and the final quality-focal-loss continuation that produces the reported model—are given in Appendix A.6.

4.2 COMPARISON WITH OPEN-PROMPT DETECTORS

We compare SetOPD with existing open-prompt detectors on DIOR and DOTA-v2.0. Because published methods use different training data and evaluation protocols, the complete reported-results comparison, organized by text, visual, and text+visual prompting modes, is provided in Appendix A.4 (Table 6). Within SetOPD, text+visual prompting raises Macro11 AP over both unimodal readers (55.57 vs. 53.92 and 53.55): on DIOR it improves over both (57.7/58.2 $\rightarrow$ 60.0), and on DOTA-v2.0 it improves over the visual reader (46.0 $\rightarrow$ 49.0) while remaining on par with the strong text reader (49.4). In the reported values, SetOPD also obtains the highest DOTA-v2.0 AP among the methods for which AP is available.

4.3 ABLATION AND ANALYSIS OF VISUAL PROMPT ENCODING

Table 1 contains two complementary studies: Part A analyzes sensitivity to the number of support exemplars, while Part B performs structural ablations of the Base–Residual representation; the equal-token mean-prototype control matches the prompt length of Base+R1 without the anchor–correction decomposition.

Three conclusions suffice. First, a single exemplar already reaches 94% of the $K = 5$ Macro11 AP (50.45 vs. 53.55), and performance saturates around $K = 5$ while the prompt length stays fixed at two tokens. Second, the anchor–correction factorization is substantially more effective than directly using or duplicating the pooled prototype in our implementation: Base alone reaches 47.11 AP, adding one Residual correction raises it to 53.55 ($+6.4$ AP), and the equal-token mean-prototype control reaches only 47.68 AP, so prompt length alone does not explain the gap. Third, multi-Residual routing does not help: Base+2R and Base+3R score no higher than Base+R1. The mean-prototype and Base-only variants suffer severe performance degradation on the held-out RSOD source, where Base+R1 restores detection; even excluding RSOD, Base+R1 still improves over Base-only by 2.83 AP on average across the remaining ten sources. This suggests that the anchor–correction parameterization improves both average accuracy and robustness across source domains.

4.4 CANDIDATE-STATE COLLABORATION OUTPERFORMS REPRESENTATION-LEVEL FUSION

To compare candidate-set-level late fusion with common representation-level fusion and candidate-relation strategies, we report two unimodal reference endpoints followed by collaboration baselines

Table 1: Ablation and analysis of Base–Residual visual prompting. Part A studies support-set size; Part B isolates the contributions of Base, Residual, prompt length, and multi-Residual routing. All structural variants are independently trained from the same frozen Stage-0 endpoint with the same optimization budget and support-sampling protocol. All values are Visual-only AP.

Setting	DIOR	DOTA	FAIR1M	HRSC	SIMD	SODA-A	Ship	WHU	HRRSD	RSOD	VEDAI	Macro11
A. Support-example efficiency												
Mean prototype, $K = 1$	55.4	42.6	23.5	63.3	59.5	37.9	47.3	51.3	26.7	0.4*	55.3	42.11
Base–Residual, $K = 1$	56.0	41.3	25.9	70.5	68.9	39.0	48.0	78.9	30.1	42.7	53.7	50.45
Base–Residual, $K = 3$	57.3	45.4	27.5	74.7	70.2	39.5	57.0	78.7	31.3	40.2	51.1	52.08
Base–Residual, $K = 5$	58.2	46.0	28.8	78.4	71.6	39.5	58.2	78.9	32.4	42.7	54.4	53.55
Base–Residual, $K = 10$	58.4	48.3	28.5	73.5	71.6	39.4	58.9	79.2	32.1	43.0	56.4	53.57
B. Structural attribution ($K = 5$)												
Base+R1	58.2	46.0	28.8	78.4	71.6	39.5	58.2	78.9	32.4	42.7	54.4	53.55
Residual only	36.0	25.9	13.3	41.4	43.1	33.2	17.8	27.2	21.7	36.8	31.0	29.76
Mean prototype $\times 2$	58.1	47.6	27.7	74.2	68.6	43.7	57.3	66.8	28.3	0.44*	51.73	47.68
Base only	57.8	47.1	27.5	77.8	70.6	39.0	56.7	63.5	28.8	0.1*	49.3	47.11
Base+2R (T3)	58.1	46.3	28.4	79.8	71.1	39.3	59.7	78.8	31.9	42.2	53.4	53.55
Base+3R (T4)	58.2	46.4	28.8	79.1	71.5	39.6	58.1	79.0	32.1	42.4	52.2	53.40

*Entries with severe performance degradation.

Table 2: Comparison of unimodal readers and text–visual collaboration strategies on all 11 sources (AP).

Method	DIOR	DOTA	FAIR	HRSC	SIMD	SODA	ShipRS	WHU	HRRSD	RSOD	VEDAI	M11
Text Reader (Stage-0, frozen)	57.75	**49.45**	29.78	78.10	72.08	39.36	60.38	79.41	33.05	40.49	53.22	53.92
Base–Residual Visual Reader (Ours)	58.11	47.30	28.99	78.84	71.82	39.61	58.89	79.02	31.35	41.67	53.46	53.55
Joint Prompt Representation	58.90	48.93	29.25	79.24	71.98	39.66	60.68	79.27	32.16	39.96	54.00	54.00
Relation-Aware Selection	58.35	46.71	29.90	79.18	71.18	39.80	61.90	78.51	**33.44**	41.17	54.91	54.10
SetOPD (Ours)	**59.97**	49.01	**30.54**	**79.89**	**72.54**	**44.59**	**62.66**	**83.06**	32.19	**41.91**	**54.93**	**55.57**

within the same base detection framework, with every collaboration baseline built on the same Base–Residual visual reader. The text reader is the frozen Stage-0 text pathway of the adapted Grounding DINO (Liu et al., 2024) base detector, and the Base–Residual visual reader is our own visual branch (Sec. 3.2), decoded under the shared query initialization of the joint pathway rather than run as the native Visual-only detector of Table 1—its per-source APs therefore differ slightly from the Visual-only rows (e.g., 47.30 vs. 46.0 on DOTA-v2.0). Joint Prompt Representation (Yang et al., 2026; Xu et al., 2023) fuses text and vision at the prompt/representation level, and Relation-Aware Selection (Hu et al., 2018; Hosang et al., 2017) models object relations at the candidate level. Three further adaptations—Sparse Query Fusion (Carion et al., 2020), MS-DETR (Xing et al., 2024), and DeepInteraction (Yang et al., 2022)—were adapted on only four shared sources and are reported in Appendix A.5.

Table 2 first reports the two unimodal readers as reference endpoints, followed by three text–visual collaboration strategies. SetOPD reaches 55.57 Macro11 AP across all 11 sources, improving over the text reader (53.92), Base–Residual visual reader (53.55), Joint Prompt Representation (54.00), and Relation-Aware Selection (54.10) by 1.66, 2.02, 1.57, and 1.48 AP, respectively. At the source level, SetOPD obtains the highest AP in the table on nine of 11 sources, including the held-out RSOD and VEDAI, indicating that the overall improvement is not driven by a single source. Beyond its average gain, SetOPD also exceeds the *stronger* of the two unimodal readers on 9 of 11 sources and remains within 1 AP of it on the remaining two (DOTA-v2.0 and HRRSD), indicating that candidate-state collaboration often produces gains beyond simply selecting the better modality.

More importantly, relative to representation-level Joint Prompt Representation, PQA achieves higher AP on all 11 sources and raises Macro11 from 54.00 to 55.57. The gain is not confined to in-domain sources: PQA improves over Joint Prompt Representation by 1.96 AP on the seven in-domain sources and by 0.89 AP on the four fully held-out sources. RS-MPOD's cross-dataset and fine-grained experiments likewise show that representation-level multimodal prompting can underperform visual prompting on some external datasets, which its authors attribute to semantic noise from mismatched text cues during fusion (Yang et al., 2026). Because the training data and evaluation protocols differ, we do not make direct numerical comparisons to those results. Our controlled comparison within a

shared base framework instead shows that delaying explicit cross-modal state fusion until text and vision have independently formed candidate detections more consistently preserves useful evidence from both modalities.

Diagnosing the DOTA-v2.0 sensitivity. DOTA-v2.0 exposes a residual sensitivity in the final scoring stage. Simply continuing PQA training brings only a marginal gain, whereas using the frozen Text-reader boxes recovers noticeably more AP. Removing the best-reader preservation term has almost no additional effect, while quality-aware classification provides a further improvement. More importantly, a zero-training counterfactual that retains the PQA candidate list but replaces its scores with Text-reader scores recovers substantially more AP. This indicates that the residual DOTA-v2.0 degradation is dominated by score/query-class ranking and its coupling with localization, rather than by the absence of useful multimodal candidate states. A diagnostic ablation further localizes the sensitivity to the final score/readout stage; zero-training readout probes provide additional evidence that useful multimodal candidates are already present (Appendix A.1).

Detection-level analysis further confirms this behavior: both readers contribute modality-exclusive correct detections, while PQA retains 53.8–100% of their union across four audited sources (Appendix A.2).

5 Conclusion

SetOPD shows that multimodal open-prompt detection need not collapse text and visual evidence in representation space. Across 11 heterogeneous remote-sensing sources, preserving modality-specific candidate states and reasoning over them as a set consistently improves over representation-level joint prompting. Together with the observed detection-level complementarity, these results support candidate-set reasoning as an alternative design principle for multimodal prompt integration.

AI use statement

Generative AI tools were used to assist with language editing, including improving the clarity, concision, and organization of the manuscript; literature search and organization; and code implementation and debugging. All AI-assisted content and code were reviewed and verified by the authors, who take full responsibility for the final manuscript and implementation.

Reproducibility statement

Section 3 and Appendices A.6–A.7 specify the model architecture, layer dimensions, training objectives, hyperparameters, and the staged optimization protocol, including $N = 900$, $R = 1$, $\lambda_R = 0.5$, $\tau = 1.0$, detection-loss weights $1/5/2$, the $\mathcal{L}_{\text{keep}}$ weight 0.1, iteration counts, batch sizes, optimizer, and seeds. Appendix B proves Propositions 1 and 2. The supplementary material additionally provides anonymized implementation notes and module-level pseudocode documenting the interfaces and tensor flows of the proposed components; the complete source code and configuration files will be released upon acceptance.

References

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5297–5307, 2016. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Arandjelovic_NetVLAD_CNN_Architecture_CVPR_2016_paper.html.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, volume 12346 of *Lecture Notes in Computer Science*, pp. 213–229. Springer, 2020. doi: 10.1007/978-3-030-58452-8_13. URL https://arxiv.org/abs/2005.12872.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16901–16911, 2024.

Shenghao Fu, Yukun Su, Fengyun Rao, Jing Lyu, Xiaohua Xie, and Wei-Shi Zheng. Wedetect: Fast open-vocabulary object detection as retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20377–20387, 2026a.

Weifu Fu, Jinyang Li, Bin-Bin Gao, Jialin Li, Yuhuan Lin, Hanqiu Deng, Wenbing Tao, Yong Liu, and Chengjie Wang. PET-DINO: Unifying visual cues into Grounding DINO with prompt-enriched training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13039–13048, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Fu_PET-DINO_Unifying_Visual_Cues_into_Grounding_DINO_with_Prompt-Enriched_Training_CVPR_2026_paper.html.

Yuchen Guan, Chong Sun, Canmiao Fu, Zhipeng Huang, Chun Yuan, and Chen Li. Text-guided visual prompt DINO for generic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21288–21298, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Guan_Text-guided_Visual_Prompt_DINO_for_Generic_Segmentation_ICCV_2025_paper.html.

Jan Hosang, Rodrigo Benenson, and Bernt Schiele. Learning non-maximum suppression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4507–4515, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Hosang_Learning_Non-Maximum_Suppression_CVPR_2017_paper.html.

Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3588–3597, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Relation_Networks_for_CVPR_2018_paper.html.

Jinlong Hu, Yi Zhang, Zhiqi Xia, Yikang Zhou, and Shunping Ji. Hi-OPD: Hierarchy-aware open-prompt detection for remote sensing images. *arXiv preprint arXiv:2609.25584*, 2026. URL https://arxiv.org/abs/2609.25584.

Ziyue Huang, Yongchao Feng, Ziqi Liu, Shuai Yang, Qingjie Liu, and Yunhong Wang. OpenRSD: Towards open-prompts for object detection in remote sensing images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8384–8394, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Huang_OpenRSD_Towards_Open-prompts_for_Object_Detection_in_Remote_Sensing_Images_ICCV_2025_paper.html.

Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-Rex2: Towards generic object detection via text-visual prompt synergy. In *Computer Vision – ECCV 2024*, volume 15091 of *Lecture Notes in Computer Science*, pp. 38–57. Springer, 2024. doi: 10.1007/978-3-031-73414-4_3. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/4804_ECCV_2024_paper.php.

Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-VLM: Open-vocabulary object detection upon frozen vision and language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://research.google/pubs/f-vlm-open-vocabulary-object-detection-upon-frozen-vision-and-language-models/.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Feng Li, Qing Jiang, Hao Zhang, Tianhe Ren, Shilong Liu, Xueyan Zou, Huaizhe Xu, Hongyang Li, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Visual in-context prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12861–12871, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_Visual_In-Context_Prompting_CVPR_2024_paper.html.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10965–10975, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Li_Grounded_Language-Image_Pre-Training_CVPR_2022_paper.html.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, 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 *Computer Vision – ECCV 2024*, volume 15105 of *Lecture Notes in Computer Science*, pp. 38–55. Springer, 2024. doi: 10.1007/978-3-031-72970-6_3. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/6319_ECCV_2024_paper.php.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with Slot Attention. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11525–11538, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/8511df98c02ab60aea1b2356c013bc0f-Abstract.html.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection. In *Computer Vision – ECCV 2022*, volume 13670 of *Lecture Notes in Computer Science*, pp. 728–755. Springer, 2022. doi: 10.1007/978-3-031-20080-9_42. URL https://link.springer.com/chapter/10.1007/978-3-031-20080-9_42.

Jiancheng Pan, Yanxing Liu, Yuqian Fu, Muyuan Ma, Jiahao Li, Danda Pani Paudel, Luc Van Gool, and Xiaomeng Huang. Locate anything on earth: Advancing open-vocabulary object detection for remote sensing community. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6): 6281–6289, 2025. doi: 10.1609/aaai.v39i6.32672. URL https://ojs.aaai.org/index.php/AAAI/article/view/32672.

Bo Qian, Dahu Shi, and Xing Wei. DETR-ViP: Detection transformer with robust discriminative visual prompts. In *International Conference on Learning Representations (ICLR)*, pp. 93513–93535, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/97011c648eda678424f9292dadeae72e-Abstract-Conference.html.

Yitong Qin, Lihua Zhou, Jiwei Wei, Ran Ran, Shiyuan He, Zeyu Ma, Shuaifeng Li, Nianxin Li, and Heng Tao Shen. ViTPrompt: Training-free prompt refinement with visual tokens for open-vocabulary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3111–3121, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Qin_ViTPrompt_Training-Free_Prompt_Refinement_with_Visual_Tokens_for_Open-Vocabulary_Detection_CVPR_2026_paper.html.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html.

Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOE: Real-time seeing anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24591–24602, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_YOLOE_Real-Time_Seeing_Anything_ICCV_2025_paper.html.

Yongjian Wu, Yang Zhou, Jiya Saiyin, Bingzheng Wei, and Yan Xu. Visual textualization for image prompted object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20900–20910, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wu_Visual_Textualization_for_Image_Prompted_Object_Detection_ICCV_2025_paper.html.

Yinghui Xing, Shuo Yang, Song Wang, Shizhou Zhang, Guoqiang Liang, Xiuwei Zhang, and Yanning Zhang. MS-DETR: Multispectral pedestrian detection transformer with loosely coupled fusion and modality-balanced optimization. *IEEE Transactions on Intelligent Transportation Systems*, 25(12):20628–20642, 2024. doi: 10.1109/TITS.2024.3450584. URL https://arxiv.org/abs/2302.00290.

Yifan Xu, Mengdan Zhang, Chaoyou Fu, Peixian Chen, Xiaoshan Yang, Ke Li, and Changsheng Xu. Multi-modal queried object detection in the wild. In *Advances in Neural Information Processing Systems*, volume 36, pp. 4452–4469, 2023. doi: 10.52202/075280-0198. URL https://papers.nips.cc/paper_files/paper/2023/hash/0e3af444e7d82d29871804de476d1fbe-Abstract-Conference.html.

Shuai Yang, Ziyue Huang, Jiaxin Chen, Qingjie Liu, and Yunhong Wang. Beyond open vocabulary: Multimodal prompting for object detection in remote sensing images. *arXiv preprint arXiv:2602.01954*, 2026. URL https://arxiv.org/abs/2602.01954.

Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. DeepInteraction: 3D object detection via modality interaction. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1992–2005, 2022. doi: 10.52202/068431-0145. URL https://arxiv.org/abs/2208.11112.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep Sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary DETR with conditional matching. In *Computer Vision – ECCV 2022*, volume 13669 of *Lecture Notes in Computer Science*, pp. 106–122. Springer, 2022. doi: 10.1007/978-3-031-20077-9_7. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/1248_ECCV_2022_paper.php.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14393–14402, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zareian_Open-Vocabulary_Object_Detection_Using_Captions_CVPR_2021_paper.html.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2203.03605.

Jinrong Zhang, Penghui Wang, Chunxiao Liu, Wei Liu, Dian Jin, Qiong Zhang, Erli Meng, and Zhengnan Hu. Just a few glances: Open-set visual perception with image prompt paradigm. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9):9969–9976, 2025. doi: 10.1609/aaai.v39i9.33081. URL https://ojs.aaai.org/index.php/AAAI/article/view/33081.

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 (ICLR)*, 2021. URL https://arxiv.org/abs/2010.04159.

A Supplementary Experiments and Details

This appendix reports diagnostic analyses of PQA, detection-evidence complementarity, the initialization-pairing audit, comparisons with external open-prompt detectors, adapted collaboration baselines, and full training-objective and implementation details.

A.1 Diagnostic Ablation and Readout Analysis on DOTA-v2.0

We further diagnose why candidate-state collaboration is more sensitive on DOTA-v2.0. The analysis separates four possible factors: insufficient optimization, the localization path used by the fused candidates, the best-reader preservation loss, and classification-score calibration. All trained controls start from the same PQA checkpoint and use the same additional training budget and optimization protocol. We additionally include two zero-training readout counterfactuals to test whether useful candidate states already exist before score repair.

Table 3 shows that additional optimization alone changes AP only marginally. Replacing the fused localization path with the frozen Text-reader boxes gives a larger recovery, whereas removing the best-reader preservation loss provides almost no further gain. With the box policy and preservation setting fixed, QFL classification adds another 0.30 AP. The strongest evidence comes from the zero-training controls: retaining the PQA candidate list while replacing its scores with Text-reader scores reaches 50.14 AP, and logit averaging with Text-reader boxes reaches 50.17 AP. Thus, the DOTA-v2.0 gap is primarily associated with score/query-class ranking and score–localization coupling rather than failure to form useful multimodal candidates.

Table 3: Diagnostic ablation and zero-training readout analysis on DOTA-v2.0. The first group isolates optimization, localization policy, the preservation loss, and QFL classification; the second group probes the final scoring/readout without additional training.

Setting	DOTA AP	Δ AP
Diagnostic ablations		
PQA checkpoint	48.25	–
+ continued optimization	48.33	+0.08
+ Text-reader localization	48.67	+0.34
− preservation loss	48.71	+0.04
+ QFL classification (Final SetOPD)	49.01	+0.30
Zero-training readout probes		
PQA candidate list + Text-reader scores	50.14	$+1.89^{\dagger}$
Text/PQA logit averaging + Text-reader boxes	50.17	$+1.92^{\dagger}$

For the first five rows, Δ AP is relative to the immediately preceding matched control. †Zero-training counterfactual; Δ is relative to the PQA checkpoint.

A.2 Detection-Evidence Complementarity

At IoU threshold 0.5, we partition ground-truth objects into those detected only by the text reader (Text-only GT), only by the visual reader (Visual-only GT), and by both readers (Common GT). The modality-exclusive set is the union of the two exclusive subsets, denoted Exclusive GTs, and PQA retention is the fraction of these exclusive ground-truth objects that PQA still detects. Counting unique ground-truth objects rather than raw detections ensures that duplicate detections of the same object are not counted multiple times. This analysis directly quantifies complementarity at the detection level.

Table 4: Detection-level complementarity between the text and visual readers at IoU 0.5, counted over unique ground-truth objects. Exclusive GTs is the union of Text-only and Visual-only ground-truth objects; PQA retained counts how many of them PQA still detects.

Dataset	Text-only GT	Visual-only GT	Common GT	Exclusive GTs	PQA retained
DIOR	698	1,716	116,758	2,414	1,763 / 2,414 (73.03%)
DOTA-v2.0	1,876	2,196	134,463	4,072	2,191 / 4,072 (53.81%)
HRSC2016	1	18	1,209	19	19 / 19 (100.00%)
RSOD	12	71	7,103	83	70 / 83 (84.34%)

Both readers contribute modality-exclusive ground-truth objects on all four sources, and PQA retains them unevenly: from 100.00% on HRSC2016 down to 73.03% on DIOR and 53.81% on DOTA-v2.0. Retention below 100% quantifies how much single-reader evidence is lost when disagreeing pairs are resolved into a single decision. Notably, the visual reader contributes more modality-exclusive ground-truth objects than the text reader on all four audited sources, further indicating that visual prompting supplies non-redundant instance-level evidence. Figure 3 shows qualitative examples: the visual reader recovers storage tanks missed by the text reader, and in another image the two readers detect complementary airplane instances; PQA retains both.

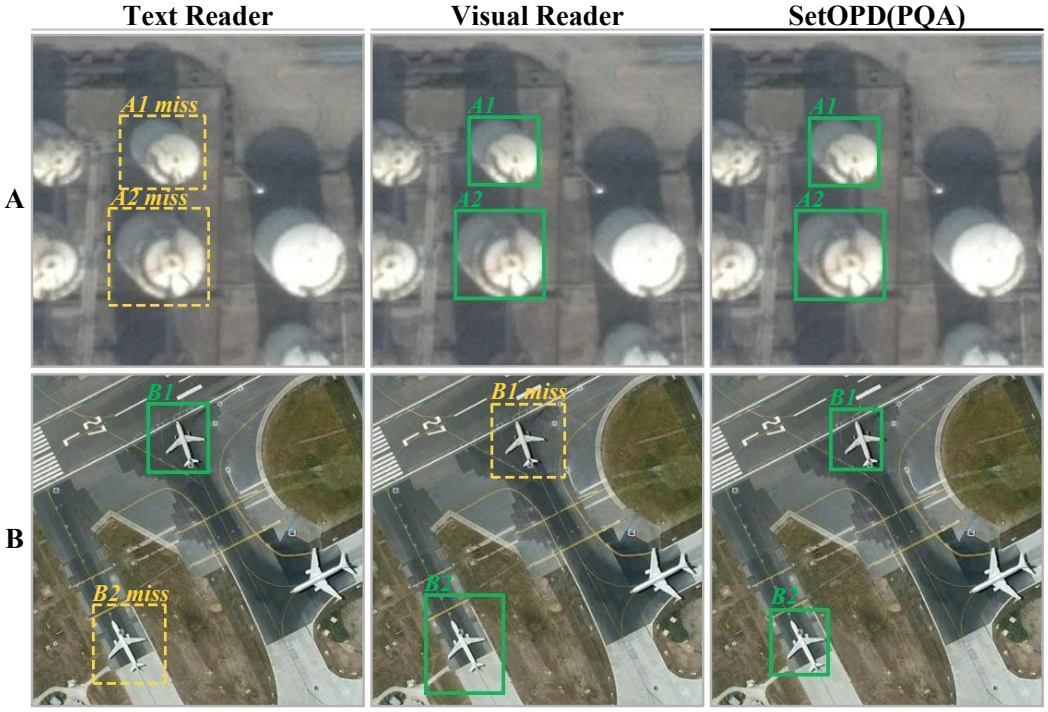

Figure 3: **Qualitative detection-level complementarity and PQA arbitration.** Columns show ground truth and the predictions of each reader. Green boxes denote matched predictions, while yellow dashed boxes indicate selected ground-truth objects missed by the corresponding reader. Top: the visual reader recovers storage tanks missed by the text reader, and PQA preserves them. Bottom: the two readers recover complementary airplane instances, while PQA retains both. All columns within each row use the same image crop.

A.3 Initialization-Pairing Audit

To quantify how reliably shared initialization induces object-level correspondence, we audit the text reader and the Base–Residual visual reader on three sources (DIOR, DOTA-v2.0, and HRSC2016; 128 images each, fixed sampling). Table 5 reports three quantities. First, conditioned on both branches producing class-correct true positives (IoU ≥ 0.5), same-index pairs refer to the same ground-truth object in at least 99.86% of the qualifying pairs on each audited source. Second, same-index reference

boxes are spatially well aligned (mean IoU 0.767–0.857). Third, compared with a box-IoU-optimal one-to-one Hungarian reassignment within the Top-50 selected queries, the original index pairing is retained for 83.94–94.03% of indices.

Same-GT is a conditional measure: it should not be interpreted as correspondence accuracy over all selected queries, since most queries do not satisfy the strict verification condition that both branches be class-correct true positives. Likewise, index pairing agrees with box-IoU-optimal assignment for most Top-50 candidates, while the remaining 5.97–16.06% shows that shared initialization is not a universal correspondence guarantee. Shared initialization therefore provides a strong index-wise correspondence prior for valid detections, and the residual disagreement is one reason PQA performs within-pair arbitration and set-level reasoning instead of averaging same-index states directly.

Table 5: Initialization-pairing audit over three sources (128 images each).

Dataset	Same-GT given both TP	Same-index box IoU	Index retained after Hungarian
DIOR	99.90%	0.854	94.03%
DOTA-v2.0	99.86%	0.857	93.97%
HRSC2016	100.00%	0.767	83.94%

Same-GT is conditioned on both branches being class-correct TPs at IoU ≥ 0.5. It should not be interpreted as correspondence accuracy over all selected queries.

A.4 Reported Results of External Open-Prompt Detectors

Table 6 compares SetOPD with existing open-prompt detectors on DIOR and DOTA-v2.0 using results reported in their source papers.

Table 6: Reported-results comparison with existing open-prompt detectors on DIOR and DOTA-v2.0. Dashes indicate metrics not reported in the source papers.

Method	DIOR AP	DIOR AP50	DOTA-v2.0 AP	DOTA-v2.0 AP50
Grounding DINO (Liu et al., 2024) (Text)	–	78.7	–	71.8
LAE-DINO (Pan et al., 2025) (Text)	–	85.5	46.8	–
OpenRSD (Huang et al., 2025) (Text)	–	76.7	–	71.8
OpenRSD (Huang et al., 2025) (Visual)	–	76.7	–	69.8
Hi-OPD (Hu et al., 2026) (Text)	–	79.7	48.2	72.3
Hi-OPD (Hu et al., 2026) (Visual)	–	75.2	46.1	69.5
WeDetect-tiny (Fu et al., 2026a)	–	76.5	42.6	65.6
YOLO-World-L (Cheng et al., 2024)	–	73.2	–	58.0
RS-MPOD (Yang et al., 2026) (Text)	–	81.3	46.1	73.8
RS-MPOD (Yang et al., 2026) (Text+Visual)	–	83.6	48.0	74.6
RS-MPOD (Yang et al., 2026) (Visual, 32 inst.)	–	76.2	44.1	70.0
SetOPD (Text)	57.7	80.3	49.4	73.6
SetOPD (Visual, 5 inst.)	58.2	80.6	46.0	68.8
SetOPD (Text+Visual)	**60.0**	81.7	**49.0**	71.6

A.5 Adapted Baselines on Four Shared Sources

Sparse Query Fusion is an adaptation built from the sparse-query set formulation of DETR (Carion et al., 2020), rather than a method name from that work; MS-DETR Adaptation (Xing et al., 2024) and DeepInteraction Adaptation (Yang et al., 2022) transfer, respectively, loosely coupled multispectral fusion and Camera–LiDAR two-stream interaction to text/visual prompting. They were adapted on only four shared sources (DIOR, DOTA-v2.0, HRSC2016, and RSOD), so Macro11 is not reported; SetOPD outperforms all three on these sources (Table 7).

Table 7: Adapted collaboration baselines on the four shared sources (AP).

Method	DIOR	DOTA	HRSC	RSOD
Sparse Query Fusion (Carion et al., 2020)	18.85	11.59	42.79	13.14
MS-DETR Adaptation (Xing et al., 2024)	27.94	25.37	30.92	37.62
DeepInteraction Adaptation (Yang et al., 2022)	53.04	43.66	70.45	38.72
SetOPD (Ours)	**59.97**	**49.01**	**79.89**	**41.91**

A.6 Training-Objective Details

Training has three stages. Stage 0 adapts Grounding DINO (Liu et al., 2024) with a Swin-T visual backbone and BERT text encoder on the RS153 training corpus, producing the prompt-neutral base endpoint, which is frozen in all subsequent stages. We use $K = 5$ boxed exemplars per category to build the default Base+R1 two-token visual memory. Stage I trains only the Base–Residual visual-prompt encoder for 10,000 iterations with effective batch size 16. Stage II trains only the PQA pair gate and one-layer set adapter for 1,200 iterations with effective batch size 4 (about 0.63M parameters). Both stages use AdamW with learning rate 10^{-4} and weight decay 10^{-4}; random seeds are 0 and 71, respectively.

PQA is trained with DETR-style Hungarian matching and the standard classification, ℓ_1, and GIoU losses, weighted by 1, 5, and 2:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{cls}} + 5\mathcal{L}_{\text{L1}} + 2\mathcal{L}_{\text{GIoU}}. \tag{9}$$

For a fused query i matched to ground-truth object j, we compare only the text, visual, and fused states that share the same index i. The matched detection error of each path is

$$e_i^m = -\log p_i^m(c_j) + 5\|b_i^m - b_j\|_1 + 2\big(1 - \text{GIoU}(b_i^m, b_j)\big), \qquad m \in \{T, V, F\}. \tag{10}$$

We then use

$$\mathcal{L}_{\text{keep}} = \text{mean}_i \,\text{ReLU}\big(e_i^F - \min(e_i^T, e_i^V)\big)\,, \tag{11}$$

$$\mathcal{L} = \mathcal{L}_{\text{det}} + 0.1\mathcal{L}_{\text{keep}}. \tag{12}$$

The unimodal errors e_i^T and e_i^V are treated as constants through stop-gradient, so $\mathcal{L}_{\text{keep}}$ constrains only the fused state. A fused query is penalized when it fits its matched target worse than the better of its two paired readers. The term encourages PQA to preserve stronger unimodal evidence, but does not guarantee a higher final AP. After this objective converges, training concludes with the fused candidates localized by frozen text-reader boxes and a quality-focal-loss (QFL) classification continuation, with all other components frozen.

A.7 Full Parameterizations

Base transformation $\mathcal{B}$. The Base token is a learnable transformation of the mean μ_c computed with a learnable query q_B and feed-forward refinement; structurally, it is a single-seed PMA with uniform attention weights (Lee et al., 2019), i.e., a Deep-Sets-style encoding of the mean (Zaheer et al., 2017):

$$\tilde{b}_c = \text{LN}(\mu_c + q_B), \qquad B_c = \text{LN}\Big(\tilde{b}_c + \text{FFN}_B(\tilde{b}_c)\Big)\,. \tag{13}$$

Residual slot update $\mathcal{R}_r$. Each Residual slot r starts from a learnable query a_r and pools evidence deviations $\delta_{cn} = e_{cn} - \text{sg}(B_c)$. Attention logits are first normalized across slots so the slots compete for each evidence vector, and are then renormalized over valid evidence vectors, following the normalization pattern of Slot Attention (Locatello et al., 2020). Unlike Slot Attention, the slots pool deviations from a support-dependent anchor, use deterministic slot-specific queries rather than sampled initial states, and are updated by residual feed-forward steps rather than a GRU. After two iterations,

$$\alpha_{rn} = \text{softmax}_r\left(\frac{\text{LN}(a_r)^\top \text{LN}(\delta_{cn})}{\sqrt{d}}\right), \qquad w_{rn} = \frac{\alpha_{rn}}{\sum_{n'} \alpha_{rn'}}, \tag{14}$$

$$a_r \leftarrow \text{LN}\left(a_r + W_u \sum_n w_{rn}\delta_{cn}\right), \qquad a_r \leftarrow \text{LN}(a_r + \text{FFN}_s(a_r))\,, \qquad R_{cr} = \text{LN}(W_o a_r). \tag{15}$$

With a single Residual slot, the slot-wise softmax equals one and the weights are uniform, so R_{c1} is a learnable function of the offset $\mu_c - B_c$ (Appendix B).

Set-reasoning block. SetBlock applies one eight-head self-attention layer without index-dependent positional encoding, followed by a per-element $256 \to 512 \to 256$ FFN, both with residual connections and LayerNorm:

$$\widetilde{H} = \text{LN}\big(H^P + \text{MHA}(H^P, H^P, H^P)\big)\,, \qquad H^F = \text{LN}\Big(\widetilde{H} + \text{FFN}(\widetilde{H})\Big)\,. \tag{16}$$

B Structural Properties and Proofs

This section collects the formal statements of the structural properties discussed in Secs. 3.2 and 3.3. The properties themselves are elementary; their role is to make the design constraints explicit.

Proposition 1 (support-set encoding). For any $K \geq 1$, $\Phi_{\mathrm{BR}}(\mathcal{S}_c)$ contains $R+1$ tokens, and for any permutation π of the K exemplars,

$$\Phi_{\mathrm{BR}}(\pi \cdot \mathcal{S}_c) = \Phi_{\mathrm{BR}}(\mathcal{S}_c).$$

Analysis: the default prompt is an anchor–correction factorization. Because attention logits are normalized across slots, the softmax over the single Residual slot of the default prompt ($R = 1$, T2) is identically one, so the Residual weights are uniform over valid evidence vectors and $V_c = [B_c, R_{c1}]$ is a deterministic function of the mean μ_c, with R_{c1} a learnable correction computed from the offset $\mu_c - B_c$. Although the default Base+R1 prompt therefore contains no information beyond the first-order pooled statistic μ_c, it is not equivalent to a raw mean prototype: it learns a nonlinear two-token re-parameterization $F(\mu_c) = [B_c, R_{c1}]$ that maps the pooled visual evidence into a detector-compatible prompt space. What it cannot represent is higher-order structure of the support set—two support sets with identical means necessarily produce identical Base+R1 prompts; such distribution-sensitive routing becomes possible only when multiple Residual slots compete for individual evidence vectors ($R \geq 2$). Section 4.3 shows that this extra capacity is unnecessary in the few-shot regime: with at most ten exemplars, Base+2R and Base+3R score no higher than Base+R1. The tested improvement over the mean-prototype controls therefore does not require instance-level multi-slot routing; in this few-shot setting, the simpler mean-dependent anchor–correction factorization is sufficient.

Proposition 2 (paired candidate sets). Assume both readers are permutation equivariant with respect to their queries, as in DETR-style decoders. If the initial queries are reordered by a permutation π, then (i) the multiset of initialization pairs $\{(s_i^T, s_i^V)\}$ is unchanged, and (ii) Ψ_{PQA} is permutation equivariant:

$$\Psi_{\mathrm{PQA}}(\pi \cdot P) = \pi \cdot \Psi_{\mathrm{PQA}}(P),$$

where $P = [(s_i^T, s_i^V)]_{i=1}^N$. Hence the final detection set is independent of query order except for score ties, and the Hungarian loss is also query-order invariant.

Pairing by initialization costs $O(N)$. Aligning two independently decoded candidate sets would instead require choosing a cross-modal matching cost and solving an assignment problem, with $O(N^3)$ complexity for the Hungarian algorithm. Proposition 2 also clarifies the role of the set module: pairwise arbitration is an elementwise equivariant map but cannot make one candidate depend on another, whereas the attention layer introduces exactly this interaction while preserving equivariance.

We use the notation from Section 3. A permutation π acts on a support set by reordering exemplars and on a list with N elements by reordering entries; Π denotes the corresponding permutation matrix.

Proof of Proposition 1. The instance encoder and the projection that produces evidence vectors use shared parameters and act independently on each exemplar. Reordering the exemplars therefore reorders the evidence vectors in groups of $L \cdot G$. The mean μ_c is invariant to any permutation of the evidence vectors, so B_c is also invariant. The Residual logits

$$\frac{\mathrm{LN}(a_r)^\top \mathrm{LN}(\delta_{cn})}{\sqrt{d}}$$

depend on n only through δ_{cn}. Reordering evidence vectors therefore permutes the columns of the slot-by-evidence logit matrix and permutes the padding mask in the same way. Normalization across slots is performed independently within each column, while the subsequent normalization over evidence vectors depends on the columns only through a sum over n. Hence the pooled vector $\sum_n w_{rn}\delta_{cn}$ is invariant. Every slot update depends on the evidence only through these pooled vectors, so every iteration and every final R_{cr} is unchanged. By construction,

$$V_c = [B_c, R_{c1}, \ldots, R_{cR}]$$

contains exactly $R+1$ entries for any K. □

Proof of Proposition 2. (i) Reordering the initial queries by π and applying permutation-equivariant readers places $s^T_{\pi(i)}$ and $s^V_{\pi(i)}$ at position i. The pair at position i is therefore $(s^T_{\pi(i)}, s^V_{\pi(i)})$, so the multiset of pairs is unchanged.

(ii) The arbitrator uses shared parameters and computes x_i, g_i, h^P_i, and r^P_i only from pair i, so it commutes with π. Self-attention without index-dependent positional encoding satisfies

$$\mathrm{MHA}(\Pi H, \Pi H, \Pi H) = \Pi\ \mathrm{MHA}(H, H, H),$$

because the attention-logit matrix transforms as $\Pi M \Pi^\top$, row-wise softmax commutes with this simultaneous row/column permutation, and $\Pi^\top \Pi = I$. LayerNorm, FFN, and the prediction head act independently on each element. The composition of equivariant maps remains equivariant, so the prediction head outputs the same detection multiset regardless of query order. The Hungarian loss, which minimizes over all assignments, also retains the same value. □