OPEN-TEXT UNDERWATER DETECTION: GROUNDING RICH DESCRIPTIONS ACROSS OBJECTS AND CONTEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Underwater object detection becomes challenging when object appearance alone is insufficient to distinguish target instances, while relations and environmental context provide complementary evidence. We study open-text underwater detection, where rich descriptions of appearance, relations, and context are used to localize all matching instances. We introduce **AquaGround**, to our knowledge the largest benchmark for open-text underwater detection, containing 224K images, 893K object instances, 2,363 categories, and 7.52M query records. AquaGround provides four progressively enriched query levels and five structured environmental attributes, with each query explicitly mapped to all valid matching instances. We further propose **AquaDetect**, a cue-guided open-text underwater detector that combines target semantics, underwater context, and detail–redundancy cues to guide progressive visual aggregation while preserving spatial correspondence for localization. Experiments on AquaGround show that AquaDetect reaches 58.94 AP, outperforming the SOTA method by 5.22 points.

1 INTRODUCTION

Underwater object detection is a key perception component for marine science and industry, supporting applications such as robotic exploration and ecological monitoring. Yet challenging underwater imaging conditions often degrade object appearance and local visual evidence. Under such conditions, category labels and object appearance alone may provide insufficient evidence for distinguishing target instances, while relations to nearby organisms or structures, habitat, and spatial context can provide complementary cues. For example, "*an orange clownfish above another fish near an anemone*" narrows the matching set beyond the category "clownfish" by combining appearance, relation, and context. We therefore study **open-text underwater detection: localizing all instances that match rich descriptions of appearance, relations, and context.**

From the data perspective, investigating this question requires object-centered supervision that links rich semantic information to individual objects and each query to all of its valid matching instances. Recent underwater resources provide useful annotations, including fine-grained categories, instance descriptions, relations, environmental information, grounding, and question answering (Zheng et al., 2024; Li et al., 2026b;a; Xu et al., 2025; Zhang et al., 2025; Xue et al., 2026). However, each resource typically captures only part of the object-related information relevant to open-text detection. The same object is rarely annotated with complementary information spanning appearance, relations, environment, and surrounding context.

Therefore, we introduce **AquaGround**, to our knowledge the largest benchmark for open-text underwater detection, containing 224K images, 893K object instances, 2,363 categories, and 7.52M query records. AquaGround is constructed through four stages: enrich objects, enrich semantics, filter objects and refine expressions. The resulting annotations organize each object into four progressively enriched query levels: **Category**, **Object Attributes**, **Relations & Environment**, and **Background & Position**, together with five structured environmental attributes: Water Color, Water Clarity, Illumination, Habitat, and Occlusion. Each query is explicitly mapped to all valid matching instances, supporting both single- and multi-instance queries.

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.

The richer supervision in AquaGround also raises a modeling question. Different queries may rely on different combinations of target appearance, contextual evidence, and local visual detail, and their importance can vary with the query and scene. To address this, we introduce **AquaDetect**, a cue-guided open-text underwater detector. AquaDetect organizes the available guidance into three complementary cue families: **Target-Aligned Semantic Cues**, **Underwater Context Cues**, and **Detail–Redundancy Cues**, and adaptively combines them to score visual evidence. At three successive backbone stages, high-scoring tokens are retained as aggregation centers, while lower-priority redundant tokens are progressively merged. Spatial correspondences are recorded during aggregation and used to recover multi-scale features for localization. This design allows semantic and environmental information to guide visual processing while preserving the spatial information required for accurate detection.

Experiments show that AquaDetect reaches 58.94 AP on open-text detection task, outperforming the second-best method by 5.22 points. On visual grounding task, it achieves 37.47 Acc@0.9, with larger gains under stricter localization criteria. Ablations further show that both underwater-context and detail–redundancy guidance contribute to the final performance, and that using the same semantic and environmental information through direct embedding is less effective than the proposed cue-guided aggregation.

Our contributions are threefold:

- **AquaGround**, to our knowledge the largest benchmark for open-text underwater detection, with 224K images, 893K object instances, 2,363 categories, and 7.52M query records. It provides four progressively enriched query levels, structured underwater attributes, and explicit query–instance associations.
- **AquaDetect**, a cue-guided open-text underwater detector that adaptively combines target semantics, underwater context, and detail–redundancy cues to guide progressive visual aggregation while maintaining spatial correspondence for localization.
- Extensive experiments demonstrate gains on both open-text detection and visual grounding, with particularly clear improvements under strict localization criteria. Component ablations further verify the contributions of multi-cue guidance and spatial recovery.

2 Related Work

2.1 Underwater Datasets and Vision–Language Resources

Underwater benchmarks have been developed for a broad range of perception tasks, including object detection, instance segmentation, counting, and fine-grained marine recognition (Liu et al., 2024; Cheng et al., 2024; Wei et al., 2026; Huang et al., 2025b; Zheng et al., 2024). FathomNet provides broad marine concept coverage for ocean visual recognition (Pan et al., 2025), while more recent resources increasingly extend underwater perception beyond category labels. MarineInst introduces instance-level visual descriptions (Zheng et al., 2024); MARIS and related underwater segmentation benchmarks expand fine-grained and open-vocabulary marine perception (Li et al., 2026b;a); and NAUTILUS, UWBench, and UVLM introduce language-mediated scene understanding, grounding, environmental reasoning, and question answering (Xu et al., 2025; Zhang et al., 2025; Xue et al., 2026). AquaticCLIP further scales image–text supervision for underwater vision–language representation learning (Zong et al., 2025), while recent work also explores dense annotation generation for underwater scenes (Carion et al., 2025).

These resources provide increasingly rich semantic information, but their annotation scope differs substantially. Consequently, the same object is rarely associated with these complementary forms of supervision under a unified object-centered annotation structure. A detailed comparison of semantic supervision, annotation scale, and construction strategies is provided in Appendix A.

2.2 Open-Vocabulary and Open-Text Object Detection

Open-vocabulary object detection extends conventional detectors beyond fixed category sets by aligning visual representations with language. Recent approaches such as OWL-ViT, Grounding DINO, YOLO-World, and YOLOE enable detection from open category vocabularies or flexible

textual prompts (Minderer et al., 2023; Liu et al., 2024; Cheng et al., 2024; Wang et al., 2025). LLMDet further introduces large-language-model supervision through region-level and image-level descriptions to strengthen open-vocabulary detection (Yang et al., 2025). Beyond category-level prompts, recent grounding methods increasingly address richer linguistic structure, including complex instructions, attributes, relations, compositional expressions, and multiple valid referents(Zong et al., 2025; Hu et al., 2025; Garosi et al., 2025; Cai et al., 2025; Liu et al., 2023; Luo et al., 2025).

Related developments have also appeared in aerial and remote-sensing imagery. OpenRSD and Locate Anything on Earth study open-prompt or open-vocabulary detection (Bai et al., 2025; Reimers & Gurevych, 2019), while AerialVG emphasizes positional relations in visual grounding(Liu et al., 2025). Most closely related to our setting, OTA-Det (Wei et al., 2025) formulates open-text detection by combining rich textual descriptions with multi-instance object detection (Katija et al., 2022). These studies establish increasingly flexible language interfaces for detection and grounding; our work studies open-text detection in underwater scenes, where rich descriptions incorporate structured relational and environmental information and may correspond to multiple valid instances.

3 AQUAGROUND DATASET

To support object-centered supervision for open-text underwater detection, we construct AquaGround, containing 224,632 images, 893,799 object instances across 2,363 categories, and 7,522,059 query records. AquaGround provides multi-level textual descriptions, underwater attributes, and explicit query–instance associations for each annotated object.

3.1 TASK AND ANNOTATION STRUCTURE

AquaGround defines open-text underwater detection as localizing all annotated instances that satisfy a textual query. Each query is associated with its own set of valid bounding boxes, allowing both single- and multi-instance matches.

Each annotated object has a persistent identifier linking its geometric and semantic information. AquaGround organizes these annotations into four progressively enriched query levels: **Category**, **Object Attributes**, **Relations & Environment**, and **Background & Position**. The **Relations & Environment** level includes relations to nearby entities together with five structured environmental attributes: *Water Color*, *Water Clarity*, *Illumination*, *Habitat*, and *Occlusion*. The three descriptive levels are produced by the Self, Relative, and Absolute Interfaces, respectively. The query levels and matching rules are detailed in Appendix B.1, and the environmental fields in Appendix B.2.

3.2 CONSTRUCTION FRAMEWORK

AquaGround is constructed through four stages, as illustrated in Figure 1.

I. Enrich Objects. Underwater images and available annotations are collected from resources as documented in Appendix B.3. We supplement source boxes with screened, category–prompted SAM 3 proposals (Carion et al., 2025) for selected sources and normalize category labels to establish object anchors. Object scale and absolute position are computed from bboxes in the original image.

II. Enrich Semantics. Starting from each object's category label, Qwen3-VL (Bai et al., 2025) progressively enriches the object annotation through three semantic interfaces. The **Self Interface** combines the precomputed scale with visually annotated color and geometric appearance, forming the *Object Attributes* level. The **Relative Interface** then incorporates relations to nearby entities together with the structured underwater environmental attributes, forming the *Relations & Environment* level. Finally, the **Absolute Interface** adds local background information and the precomputed absolute position, forming the *Background & Position* level.

Each interface receives the annotations produced by the preceding interfaces as context, while introducing only the additional semantic information assigned to that interface. Object crops provide appearance evidence for the Self Interface, whereas local-context views support the Relative and Absolute Interfaces. The object category and geometry remain fixed throughout the semantic-enrichment process. The stage-specific annotation procedure and expert checks are detailed in Appendix B.4.

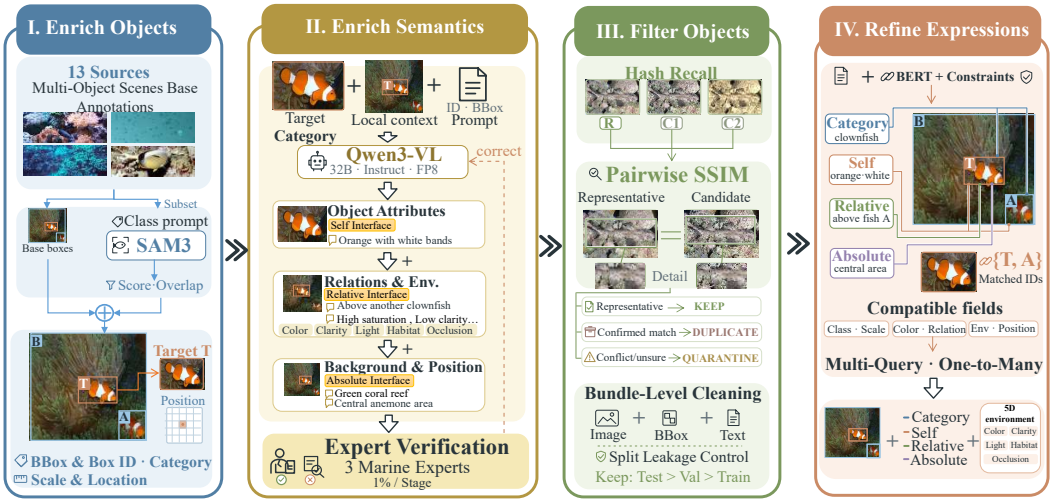

Figure 1: **AquaGround construction pipeline.** The four-stage framework establishes object anchors, progressively enriches semantic descriptions, verifies duplicate data, and constructs query–instance associations. Marine-expert spot checks are applied during semantic annotation.

III. Filter Objects. Verified duplicates are removed together with their associated geometric and semantic annotations, while conflicting or uncertain cases are quarantined.

IV. Refine Expressions Within each image, Sentence-BERT similarities (Reimers & Gurevych, 2019) and structured field constraints associate descriptions with compatible object records. Category and scale require consistency, while color, relations, environmental attributes, and position are checked under their respective compatibility rules. The resulting associations define the query-specific matching sets without modifying the descriptions or bounding boxes.

3.3 Quality Control and Dataset Splits

We check object records and query–instance associations for structural consistency, and all retained queries have nonempty GT mappings. SAM-assisted boxes are manually spot-checked for target localization. After each semantic annotation stage, three reviewers with marine-related academic training or professional experience randomly inspect approximately 1% of the images and their object annotations, checking identity, attributes, relations, environment, and position. Annotations identified as incorrect are removed, and a new random sample is then reviewed. This process is repeated until the observed error rate in the reviewed annotations falls below 1%; the first round yielded an error rate of approximately 1.67%.

Official source partitions are used when available. For the remaining sources, we deterministically partition samples within each dataset/subset group. Verified duplicate overlap across splits is removed before the final partitions are formed, reducing leakage between training and evaluation data. The resulting dataset contains 190,409 training, 22,690 validation, and 11,533 test images. The full duplicate audit and split protocol are reported in Appendix B.5.

3.4 Dataset Statistics and Coverage

Figure 2 summarizes the scale and diversity of AquaGround. The dataset contains 224,632 images, 893,799 object instances, and 7.52M query records from 13 data sources, covering at least 1,270 biological labels. Its annotations span diverse object scales and challenging underwater conditions, including color attenuation, low contrast, turbidity, backscatter, occlusion, low light, and complex backgrounds. The five structured environmental attributes form 662 observed combinations. Scale definitions and instance counts are provided in Appendix B.6.

AquaGround further provides four semantic query levels, with 95.67% of queries belonging to the three fine-grained descriptive levels. Query length increases with semantic richness, while 1.49M



Figure 2: **Statistics and semantic coverage of AquaGround.** The figure summarizes dataset scale, object-size and underwater-condition coverage, together with the four semantic query levels and their multi-instance matching characteristics. The case study illustrates how increasingly detailed queries describe and distinguish targets within the same underwater scene.

queries match multiple valid instances. All retained queries have valid GT mappings, supporting both fine-grained and multi-instance open-text underwater detection. The complete query inventory, matching cardinalities, and text-length statistics are reported in Appendix B.7.

4 AquaDetect Method

Underwater images often contain complex backgrounds and substantial visual interference, making it difficult to identify the image regions relevant to a given description. Target appearance alone may also be insufficient when similar objects occur in the same scene, while relations and environmental information provide additional cues. Motivated by these challenges, we propose AquaDetect, which uses target and environmental cues to guide progressive visual aggregation while maintaining the spatial correspondence required for detection, as illustrated in Figure 3.

4.1 Overall Framework

Given an underwater image I and a textual query q, AquaDetect further uses auxiliary textual and structured conditions provided by AquaGround. The image is encoded by a DINOv3 TinyPlus backbone (Siméoni et al., 2025; Huang et al., 2025a), producing the initial visual tokens X_0. The query and auxiliary texts are encoded separately by a frozen SigLIP 2 encoder (Tschannen et al., 2025), while the structured conditions are processed by a dedicated condition encoder:

$$T_q = \Phi_{\text{txt}}(q), \qquad T_a = \Phi_{\text{txt}}(\mathcal{A}_{\text{txt}}), \qquad E = \Phi_{\text{env}}(\mathcal{A}_{\text{env}}). \tag{1}$$

We denote $T = (T_q, T_a, E)$ as the encoded conditions used to construct the routing guidance. The supplied conditions and their encodings are detailed in Appendix C.1.

The encoded conditions guide three successive routers in the visual backbone, producing X_1, X_2, X_3 together with their patch-to-token correspondences M_1, M_2, M_3. The routed representations are subsequently recovered into multi-scale features for detection.

4.2 Multi-Cue Guidance

At each routing stage, AquaDetect evaluates the visual tokens using six guidance scores grouped into three cue families. From the encoded conditions, we use category, appearance, and relation repre-

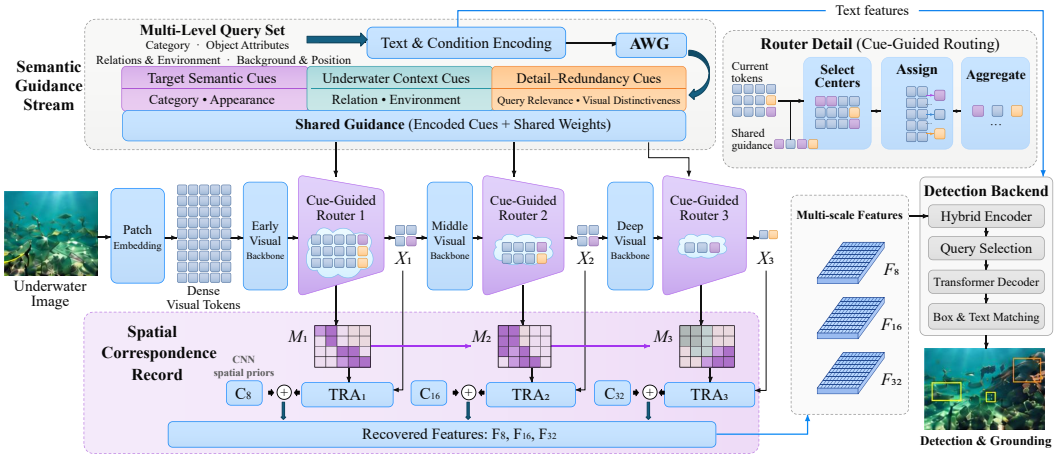

Figure 3: **Overview of AquaDetect.** Multi-cue guidance steers three stages of visual aggregation, followed by correspondence-based multi-scale recovery for detection and grounding.

sentations t_{cls}, t_{attr}, and t_{rel}, together with the environmental representation e_{env} and the complete-query reference q_{obj}. For clarity, we omit the stage superscript in the following definitions and write $v_i = v_i^{(s)}$. Let $c_i(z) = \kappa(v_i, z)$ denote the cosine similarity between visual token v_i and a guidance representation z. The six scores are

$$\begin{aligned} S_{\text{cls}}(i) &= c_i(t_{\text{cls}}), \quad & S_{\text{attr}}(i) &= \frac{1}{N_a}\sum_{\ell=1}^{N_a} c_i(t_{\text{attr}}^{(\ell)}), \\ S_{\text{rel}}(i) &= c_i(t_{\text{rel}}), \quad & S_{\text{env}}(i) &= c_i(e_{\text{env}}), \\ S_{\text{obj}}(i) &= c_i(q_{\text{obj}}), \quad & S_{\text{vis}}(i) &= 1 - \max_{j \neq i} \kappa(v_i, v_j). \end{aligned} \tag{2}$$

The first pair measures agreement with the target category and appearance, the second pair incorporates relational and underwater-context information, and the last pair characterizes query relevance and visual distinctiveness. A lower S_{vis} indicates that a visually similar alternative token is already present. Detailed score definitions and projection spaces are provided in Appendix C.2.

The relative contribution of the six cues is determined by adaptive weighting (AWG). Restoring the stage index, let $\mathbf{s}_i^{(s)} \in \mathbb{R}^6$ denote the six guidance scores defined above, and let $\mathbf{w}^{(s)} \in \mathbb{R}^6$ denote their corresponding AWG weights. The final guidance score is

$$G_i^{(s)} = \mathbf{w}^{(s)\top} \mathbf{s}_i^{(s)}. \tag{3}$$

The AWG architecture and stage-sharing configuration are detailed in Appendix C.3.

4.3 Progressive Visual Aggregation

AquaDetect inserts three cue-guided routers at successive stages of the visual backbone. Let $\mathcal{B}_s$ and R_s denote the backbone block group and router at stage s, respectively:

$$H_s = \mathcal{B}_s(X_{s-1}), \qquad (X_s, M_s) = R_s(H_s, T, M_{s-1}), \quad s = 1, 2, 3. \tag{4}$$

Here M_0 is the original one-to-one patch mapping.

At each stage, the guidance score $G_i^{(s)}$ ranks the visual tokens. The highest-scoring tokens are selected as aggregation centers, while each remaining token is assigned to one of its eight spatially nearest centers according to visual similarity and spatial distance. Assigned features are then merged into their centers by mass-weighted averaging. Information is progressively condensed in three stages: visual tokens are selected by guided cues, with the condensation intensity rising from shallow to deep layers. Assignment, aggregation, and correspondence tracking are detailed in Appendix C.4.

Table 1: Open-text detection Results(Protocol A)

Method	Source	Average precision						Average recall					
		AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AR@1	AR@10	AR@100	AR_s	AR_m	AR_l
BSR5-DETR	TPAMI'25	9.66	12.72	10.76	5.45	5.11	10.78	25.76	37.15	38.69	23.78	26.60	43.03
YOLO-World	CVPR'24	48.04	57.17	52.72	13.95	28.65	59.42	38.43	60.53	61.02	17.74	37.05	**72.19**
OTA-Det-M	ICML'26	53.72	65.09	59.09	27.34	39.67	60.71	48.13	58.11	58.18	30.17	43.28	64.73
AquaDetect	–	**58.94**	**69.76**	**63.77**	**32.76**	**44.50**	**65.89**	**52.44**	**62.73**	**62.89**	**35.23**	**47.82**	69.61

Table 2: Top-1 visual grounding Results (Protocol B)

Method	Source	Acc@0.5	Acc@0.6	Acc@0.7	Acc@0.8	Acc@0.9	mIoU	cIoU
YOLO-World	CVPR'24	49.45	48.72	47.14	43.17	31.22	45.20	64.10
Grounding DINO	ECCV'24	53.72	52.22	50.01	46.01	34.58	49.12	64.18
MARIS	CVPR'26	58.35	56.47	52.29	43.03	25.52	51.10	68.52
OTA-Det-M	ICML'26	64.39	62.03	58.39	51.68	31.57	57.09	72.45
AquaDetect	–	**64.98**	**62.99**	**59.71**	**53.92**	**37.47**	**58.37**	**73.90**

4.4 Spatial Recovery and Detection

After each routing stage, the token-recovery adapter (TRA) uses the recorded correspondence M_s to map the aggregated tokens X_s back to the original patch grid, producing a dense feature map U_s. A lightweight residual block further refines the recovered features, yielding Y_s, and Y_1, Y_2, Y_3 are resized to strides 8, 16, and 32. In parallel, a CNN spatial-prior branch extracts local features directly from the input image I. At each scale, the spatial-prior and recovered features are concatenated and fused to produce F_8, F_{16}, and F_{32}, which are passed to the detection backend (Wei et al., 2026) for box prediction and text matching. We retain its query-level and attribute-level alignment objectives together with the standard localization losses, and train the routing and recovery modules jointly with the detector. Spatial recovery and multi-scale fusion are detailed in Appendix C.4; optimization and gradient propagation are described in Appendix C.5.

5 Experiments

5.1 Experimental Setup

Comparison Methods. AquaDetect is compared with representative methods selected from complementary detection settings. YOLO-World V2.1-L (Cheng et al., 2024) and Grounding DINO Swin-L (Liu et al., 2024) are included as general-purpose detectors, BSR5-DETR-512 (Zhou et al., 2025) represents underwater object detection, MARIS (Li et al., 2026b) represents instance segmentation, and OTA-Det-M (Wei et al., 2026) represents aerial open-text detection. For a fair comparison, all methods are adapted to AquaGround and retrained on the same training split, while identical textual inputs are used during evaluation.

Training configuration. AquaDetect is initialized with pretrained DINOv3 TinyPlus visual weights (Siméoni et al., 2025; Huang et al., 2025a) and SigLIP 2 text-encoder weights (Tschannen et al., 2025), with the text encoder kept frozen. We use AdamW with a learning rate of 2.5×10^{-5} for the visual backbone and 5×10^{-4} for the remaining trainable modules. Weight decay is set to 10^{-4}. Training is performed for 64 epochs with four images per GPU and an effective batch size of 32. Training images are resized to 800×800 and further processed with multi-scale training, while evaluation uses a fixed resolution of 800×800.

For the remaining comparison methods, retraining and fine-tuning on AquaGround follow their official implementations and recommended configurations, using official pretrained checkpoints where applicable. All experiments are conducted on eight NVIDIA RTX 3090 GPUs

Table 3: **Component ablations for open-text detection (Protocol A)**

	Average precision						Average recall					
Variant	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AR@1	AR@10	AR@100	AR_s	AR_m	AR_l
w/o context	56.77	67.62	61.75	30.05	42.91	63.20	51.05	60.35	60.96	32.35	45.62	67.95
w/o detail–redun.	56.03	67.04	60.72	28.95	42.20	64.21	49.56	59.78	60.48	31.98	46.32	67.21
w/o TRA	56.78	68.04	60.90	30.12	42.77	63.71	50.96	60.97	61.46	32.80	46.24	67.83
direct embedding	55.06	67.31	60.34	28.91	41.98	63.14	49.67	59.86	60.23	31.52	45.25	67.04
AquaDetect	**58.94**	**69.76**	**63.77**	**32.76**	**44.50**	**65.89**	**52.44**	**62.73**	**62.89**	**35.23**	**47.82**	**69.61**

Table 4: **Component ablations for visual grounding (Protocol B)**

Variant	Acc@0.5	Acc@0.6	Acc@0.7	Acc@0.8	Acc@0.9	mIoU	cIoU
w/o context	64.43	62.45	58.87	52.38	35.30	57.57	73.15
w/o detail–redun.	64.23	61.93	58.22	51.96	33.64	57.10	72.43
w/o TRA	64.61	62.19	58.73	52.27	34.90	57.47	73.25
direct embedding	64.13	61.97	58.13	51.45	33.43	56.86	72.51
AquaDetect	**64.98**	**62.99**	**59.71**	**53.92**	**37.47**	**58.37**	**73.90**

5.2 Evaluation Protocols

Protocol A: Open-Text Detection. We evaluate detection using the standard COCO bounding-box AP and AR metrics (Lin et al., 2014). To adapt the evaluator to open-text detection, each image–query pair is assigned a unique evaluation category ID, such that AP and AR are aggregated over query-specific evaluation units rather than over a fixed semantic vocabulary. The top 100 scored predictions are retained globally per image before evaluation; all remaining settings follow the standard COCO protocol, with details provided in Appendix D.1.

Protocol B: Visual Grounding. We follow the standard Top-1 visual grounding protocol and report Acc@$\{0.5, 0.6, 0.7, 0.8, 0.9\}$. For each example, the highest-scoring prediction $\hat{b}_i$ is evaluated against the valid ground-truth referents. Unlike conventional single-referent visual grounding, an AquaGround description may correspond to multiple valid targets. We therefore extend the standard protocol as

$$u_i = \max_{g \in \mathcal{G}_i} \mathrm{IoU}(\hat{b}_i, g), \tag{5}$$

where $\mathcal{G}_i$ contains all valid referents for the description. Acc@ τ counts an example as correct when $u_i \geq \tau$; hence, only one valid referent needs to be localized. We additionally report mIoU and cIoU as complementary overlap measures. Additional metric definitions and interpretation are provided in Appendix D.2.

5.3 Comparison with Baselines

Open-text detection. As shown in Table 1, AquaDetect achieves 58.94 AP, the best overall performance among the compared methods. Its advantage is consistent across different IoU criteria and object scales, with leading results on AP_{50}, AP_{75}, AP_s, AP_m, and AP_l. Overall, AquaDetect achieves the best performance on 11 of the 12 reported metrics, demonstrating a broad and consistent advantage in open-text underwater detection.

Visual grounding. As shown in Table 2, AquaDetect achieves the best performance across all localization thresholds and overlap metrics. More importantly, its advantage becomes larger as the IoU threshold increases. While the methods are relatively close at lower IoU thresholds, the differences become more evident when more accurate box localization is required. AquaDetect reaches 37.47 Acc@0.9, suggesting that its main advantage lies in more precise grounding rather than simply identifying a valid referent.

5.4 Qualitative Results

Figure 4 presents representative predictions of AquaDetect under different open-text queries. In multi-object scenes, AquaDetect can localize multiple valid instances when they satisfy the same query, while more specific descriptions further distinguish targets through appearance, relations, and contextual information. The examples also show that the model remains effective in challenging underwater scenes containing small objects, visual clutter, occlusion, and degraded appearance.

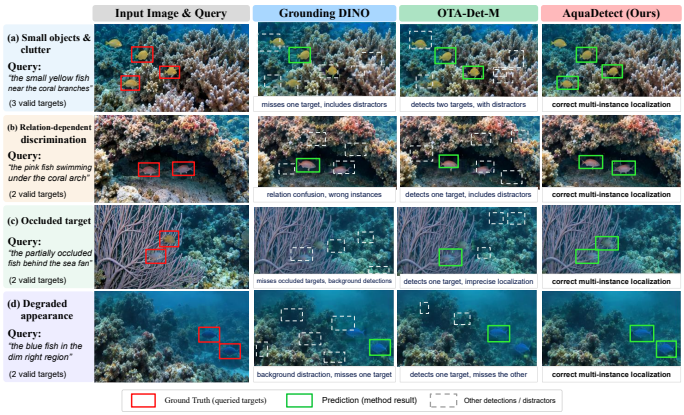

Figure 4: **Qualitative results on AquaGround.**

5.5 Ablation Study

We evaluate four controlled variants of AquaDetect under the same training and evaluation settings. *w/o context* and *w/o detail–redun.* remove the underwater-context and detail–redundancy cue families, respectively, while retaining target-aligned semantic guidance. *w/o TRA* replaces the learned spatial recovery with direct scatter/unmerge, and *direct embedding* replaces the structured cue-guidance formulation with direct visual–semantic matching while retaining the same semantic and environmental inputs. Table 3 evaluates their effects on open-text detection, while Table 4 examines visual grounding.

Open-text detection. Table 3 shows that both cue families contribute to the full model. Removing underwater-context guidance produces a relatively consistent degradation across object scales, whereas removing detail–redundancy guidance has a larger effect on small-object detection. This difference is consistent with the intended roles of the two cue families: contextual information provides complementary evidence in ambiguous underwater scenes, while detail–redundancy guidance helps preserve fine visual information during progressive aggregation.

Visual grounding. Table 4 shows a clearer difference under strict localization criteria. The ablated variants remain relatively close to the full model at Acc@0.5, but the gaps become larger as the IoU threshold increases. This pattern is particularly clear for direct embedding and for the variant without TRA. Since direct embedding receives the same semantic and environmental information as the full model, the result shows that the gain does not come from access to richer conditions alone, but from how this information is used to guide visual processing. The degradation without TRA further supports the importance of spatial recovery for accurate localization after progressive aggregation.

6 Conclusion

We introduced AquaGround, an open-text underwater detection benchmark with four levels of text queries, underwater environmental attributes, and explicit mappings from queries to matching instances. This structure supports evaluation of category, appearance, relation, and context-dependent detection in multi-object scenes. We further introduced AquaDetect, which uses target, context, and detail–redundancy cues to guide progressive visual aggregation and spatial recovery. On the full validation set, AquaDetect achieves 58.94 AP and 37.47% Acc@0.9, leading the general-purpose and underwater-specific methods under both open-text detection setting and visual grounding setting. Ablations show that both context and detail–redundancy guidance contribute, with the latter having a larger effect under strict localization. Together, AquaGround and AquaDetect provide a benchmark and a model for studying how rich language queries guide underwater detection.

AI USE STATEMENT

Generative AI tools were used in two aspects of this work. First, ChatGPT was used during manuscript preparation to assist with language polishing, wording refinement, table formatting, and consistency checking. All AI-assisted content was subsequently reviewed, discussed, and revised by the authors, who take full responsibility for the final manuscript. Second, Qwen3-VL-32B was used as part of the AI-assisted annotation pipeline for AquaGround to generate structured visual attributes and contextual descriptions from underwater imagery. These model-assisted annotations were manually reviewed and verified before being incorporated into the dataset. Generative AI was not used to determine experimental results or draw scientific conclusions.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or other sensitive personal data. AquaGround is constructed from existing underwater imagery collected from multiple data sources. We respect the licenses and usage conditions of the original data sources and provide appropriate attribution to them. For source images whose licenses do not permit redistribution, we release the corresponding annotations and processing information but require users to obtain the original images from their respective sources. We do not identify any additional ethical concerns specific to this work beyond responsible use of the released data and models in accordance with the applicable licenses.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the AquaGround construction and annotation pipeline, the AquaDetect architecture, training settings, evaluation protocols, and ablation studies in the main paper and appendix. The annotation procedure, including the AI-assisted attribute generation process, is documented to facilitate reproduction of the dataset construction pipeline. An anonymous repository containing the implementation and supporting materials is provided for review. AquaGround annotations and the associated processing code will be released publicly, while source images subject to redistribution restrictions must be obtained from their original providers. In accordance with the applicable release policy, the complete source code will be made publicly available upon acceptance.

REFERENCES

D. Akkaynak and T. Treibitz. A Revised Underwater Image Formation Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6723–6732, 2018. doi: 10.1109/CVPR.2018.00703.

S. Bai, Y. Cai, R. Chen, et al. Qwen3-VL Technical Report. arXiv preprint arXiv:2511.21631, 2025. URL https://arxiv.org/abs/2511.21631.

Zhixi Cai, Fucai Ke, Simindokht Jahangard, Maria Garcia de la Banda, Reza Haffari, Peter J. Stuckey, and Hamid Rezatofighi. NAVER: A Neuro-Symbolic Compositional Automaton for Visual Grounding with Explicit Logic Reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24078–24089, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Cai_NAVER_A_Neuro-Symbolic_Compositional_Automaton_for_Visual_Grounding_with_Explicit_ICCV_2025_paper.html.

N. Carion, L. Gustafson, Y. T. Hu, et al. SAM 3: Segment Anything with Concepts. arXiv preprint arXiv:2511.16719, 2025. URL https://arxiv.org/abs/2511.16719.

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. doi: 10.1109/CVPR52733.2024.01599. URL https://doi.org/10.1109/cvpr52733.2024.01599.

Marco Garosi, Alessandro Conti, Gaowen Liu, Elisa Ricci, and Massimiliano Mancini. Compositional Caching for Training-free Open-vocabulary Attribute Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15098–15107, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Garosi_Compositional_Caching_for_Training-free_Open-vocabulary_Attribute_Detection_CVPR_2025_paper.html.

Rui Hu, Lianghui Zhu, Yuxuan Zhang, Tianheng Cheng, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. GroundingSuite: Measuring Complex Multi-Granular Pixel Grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23105–23114, October 2025. doi: 10.1109/ICCV51701.2025.02145. URL https://openaccess.thecvf.com/content/ICCV2025/html/Hu_GroundingSuite_Measuring_Complex_Multi-Granular_Pixel_Grounding_ICCV_2025_paper.html.

S. Huang, Y. Hou, L. Liu, X. Yu, and X. Shen. Real-Time Object Detection Meets DINOv3. arXiv preprint arXiv:2509.20787, 2025a. URL https://arxiv.org/abs/2509.20787.

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, October 2025b. doi: 10.1109/ICCV51701.2025.00785. URL https://openaccess.thecvf.com/content/ICCV2025/html/Huang_OpenRSD_Towards_Open-prompts_for_Object_Detection_in_Remote_Sensing_Images_ICCV_2025_paper.html.

K. Katija, E. C. Orenstein, B. Schlining, et al. FathomNet: A Global Image Database for Enabling Artificial Intelligence in the Ocean. *Scientific Reports*, 12:15914, 2022. doi: 10.1038/s41598-022-19939-2.

Bingyu Li, Tao Huo, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Exploring the Underwater World Segmentation without Extra Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39879–39889, June 2026a. URL https://arxiv.org/abs/2511.07923.

Bingyu Li, Feiyu Wang, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. MARIS: Marine Open-Vocabulary Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24939–24949, June 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_MARIS_Marine_Open-Vocabulary_Instance_Segmentation_CVPR_2026_paper.html.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48.

C. Liu, H. Ding, and X. Jiang. GRES: Generalized Referring Expression Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23592–23601, 2023. doi: 10.1109/CVPR52729.2023.02259.

Junli Liu, Qizhi Chen, Zhigang Wang, Yiwen Tang, Yiting Zhang, Chi Yan, Dong Wang, Xuelong Li, and Bin Zhao. AerialVG: A Challenging Benchmark for Aerial Visual Grounding by Exploring Positional Relations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5177–5187, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liu_AerialVG_A_Challenging_Benchmark_for_Aerial_Visual_Grounding_by_Exploring_ICCV_2025_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 *European Conference on Computer Vision (ECCV)*, pp. 38–55, 2024. doi: 10.1007/978-3-031-72970-6_3. URL https://doi.org/10.1007/978-3-031-72970-6_3.

Zhuoyan Luo, Yinghao Wu, Tianheng Cheng, Yong Liu, Yicheng Xiao, Hongfa Wang, Xiao-Ping Zhang, and Yujiu Yang. CoHD: A Counting-Aware Hierarchical Decoding Framework for Generalized Referring Expression Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22685–22694, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Luo_CoHD_A_Counting-Aware_Hierarchical_Decoding_Framework_for_Generalized_Referring_Expression_ICCV_2025_paper.html.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling Open-Vocabulary Object Detection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 72983–73007, 2023. doi: 10.52202/075280-3191. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e6d58fc68c0f3c36ae6e0e64478a69c0-Abstract-Conference.html.

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://doi.org/10.1609/aaai.v39i6.32672.

N. Reimers and I. Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

O. Siméoni, H. V. Vo, M. Seitzer, et al. DINOv3. arXiv preprint arXiv:2508.10104, 2025. URL https://arxiv.org/abs/2508.10104.

M. Tschannen, A. Gritsenko, X. Wang, et al. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. arXiv preprint arXiv:2502.14786, 2025. URL https://arxiv.org/abs/2502.14786.

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, October 2025. doi: 10.1109/ICCV51701.2025.02280. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_YOLOE_Real-Time_Seeing_Anything_ICCV_2025_paper.html.

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Guoting Wei, Yu Liu, Xia Yuan, Xizhe Xue, Linlin Guo, Yifan Yang, Chunxia Zhao, Zongwen Bai, Haokui Zhang, and Rong Xiao. OS-W2S: An Automatic Labeling Engine for Language-Guided Open-Set Aerial Object Detection. arXiv preprint arXiv:2505.03334, 2025. URL https://arxiv.org/abs/2505.03334.

Guoting Wei, Xia Yuan, Yang Zhou, Haizhao Jing, Yu Liu, Xianbiao Qi, Chunxia Zhao, Haokui Zhang, and Rong Xiao. Open-Text Aerial Detection: A Unified Framework for Aerial Visual Grounding and Detection. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://openreview.net/forum?id=lgzkDjvVwP.

Wei Xu, Cheng Wang, Dingkang Liang, Zongchuang Zhao, Xingyu Jiang, Peng Zhang, and Xiang Bai. NAUTILUS: A Large Multimodal Model for Underwater Scene Understanding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 28367–28393, 2025. doi: 10.52202/085713-0843. URL https://papers.nips.cc/paper_files/paper/2025/hash/23a081d2dbdeb531080ced0c2942103a-Abstract-Conference.html.

Xizhe Xue, Yang Zhou, Dawei Yan, Lijie Tao, Junjie Li, Ying Li, Haokui Zhang, and Rong Xiao. UVLM: Benchmarking Video Language Model for Underwater World Understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(14):11532–11540, 2026. doi: 10.1609/aaai.v40i14.38136. URL https://ojs.aaai.org/index.php/AAAI/article/view/38136.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer Is Better but Not Necessary in Vision Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802, 2025. doi: 10.1109/CVPR52734.2025.01843.

Da Zhang, Chenggang Rong, Bingyu Li, Feiyu Wang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. UW-Bench: A Comprehensive Vision-Language Benchmark for Underwater Understanding. arXiv preprint arXiv:2510.18262, 2025. URL https://arxiv.org/abs/2510.18262.

Ziqiang Zheng, Yiwei Chen, Huimin Zeng, Tuan-Anh Vu, Binh-Son Hua, and Sai-Kit Yeung. MarineInst: A Foundation Model for Marine Image Analysis with Instance Visual Description. In *European Conference on Computer Vision (ECCV)*, pp. 239–257, 2024. doi: 10.1007/978-3-031-72627-9_14. URL https://doi.org/10.1007/978-3-031-72627-9_14.

J. Zhou, Z. He, D. Zhang, S. Liu, X. Fu, and X. Li. Spatial Residual for Underwater Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4996–5013, 2025. doi: 10.1109/TPAMI.2025.3548652.

Yongshuo Zong, Qin Zhang, Dongsheng An, Zhihua Li, Xiang Xu, Linghan Xu, Zhuowen Tu, Yifan Xing, and Onkar Dabeer. Ground-V: Teaching VLMs to Ground Complex Instructions in Pixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24635–24645, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zong_Ground-V_Teaching_VLMs_to_Ground_Complex_Instructions_in_Pixels_CVPR_2025_paper.html.

APPENDIX CONTENTS

The entries below summarize the supplementary material and indicate where to find each topic.

Section and contents		Page
A	**Comparison with Existing Resources** Semantic supervision, geometric annotation volume, and construction methodologies.	15
B	**Dataset Definitions and Construction Details** Definitions, construction procedures, source composition, and dataset statistics.	17
B.1	Query Levels and Matching Sets *Four query levels and rules for associating each query with its matching objects.*	17
B.2	Structured Environmental Attributes *Object-linked environmental fields and the counting of observed combinations.*	17
B.3	Source Composition and Annotation Completeness *Retained source collections, annotation completeness, and image totals.*	17
B.4	Progressive Semantics *Staged descriptions, conditional habitat refinement, and expert spot checks.*	18
B.5	Duplicate Verification, Cleaning, and Split Protocol *Split construction, pairwise duplicate verification, and the source-level audit.*	18
B.6	Scale Coverage *Absolute bounding-box area bins and object-scale counts.*	19
B.7	Query Statistics *Query counts by split and level, matching cardinalities, and word lengths.*	19
C	**Condition Encoding, Routing, and Optimization** Supplied inputs, routing and aggregation, spatial recovery, and joint optimization.	21
C.1	Supplied Conditions and Encodings *Text and structured-condition encodings, field masking, and the query reference.*	21
C.2	Routing Scores *Six guidance scores, their computation, and their roles in center selection.*	21
C.3	Adaptive Cue Weighting *AWG inputs, shared cue weights, stage-wise center selection, and token budgets.*	22
C.4	Spatial Correspondence and Feature Recovery *Token assignment, mass-weighted aggregation, TRA, and multi-scale feature fusion.*	23
C.5	Optimization and Output Interface *Detection outputs, joint losses, and gradient paths through discrete routing.*	23
D	**Evaluation Protocol Details** Detection and localization metrics, their settings, and interpretation.	24
D.1	Protocol A: Query-Conditioned COCO Evaluation *AP and AR definitions, detection limits, and absolute-area size ranges.*	24
D.2	Protocol B: Multi-Referent Top-1 Localization *Any-target success, mIoU and cIoU aggregation, and the distinction from AR@1.*	25

A Comparison with Existing Resources

Tables 5–7 compare representative underwater resources with AquaGround in terms of semantic supervision, geometric annotation scale, and dataset construction.

Table 5: **Semantic supervision and matching structure.** Descriptive 1:N denotes that a non-category-only description may supervise multiple object instances or masks. Object-level environmental fields are structured environmental attributes attached to individual instances. Language levels denote explicitly separated language-supervision granularities. A dash denotes absence in the specified annotation form, not a model capability limit.

Benchmark	Publication	Multiple objects	Object descriptions	Relation text	Environment text	Language levels	Object-level env. fields	Descriptive 1:N
FathomNet [22]	Scientific Reports 2022	✓	–	–	–	–	–	–
RUOD [13]	Neurocomputing 2023	✓	–	–	–	–	–	–
UIIS [69]	ICCV 2023	✓	–	–	–	–	–	–
IOCfish5K [68]	CVPR 2023	✓	–	–	–	–	–	–
MarineInst20M [2]	ECCV 2024	✓	✓	✓	–	✓	–	✓
USIS10K [23]	ICML 2024	✓	–	–	–	–	–	–
CoralMask [70]	CVPR 2024	✓	✓	–	–	–	–	–
SeaClear [24]	Scientific Data 2024	✓	–	–	–	–	–	–
NautData [25]	NeurIPS 2025	✓	✓	✓	✓	✓	–	–
SynTIDE [71]	CVPR 2025	✓	–	–	–	–	–	–
MARIS [3]	CVPR 2026	✓	–	–	–	–	–	–
AquaOV255 [4]	CVPR 2026	✓	–	–	–	–	–	–
UVLM [5]	AAAI 2026	✓	✓	✓	✓	–	–	–
AquaticCLIP [66]	IEEE TNNLS 2026	✓	✓	–	–	✓	–	–
AquaGround	**This work**	✓	✓	✓	✓	✓	✓	✓

Table 6: **Geometric annotation inventory.** Counts follow the units and snapshots reported by the respective sources; boxes, masks, center points, generated dense labels, and detector-derived regions are not directly interchangeable annotation units.

Dataset	Publication	Images	Spatial annotations	Classes / concepts
FathomNet [22]	Scientific Reports 2022	84,454	175,873 localizations (boxes)	2,244
RUOD [13]	Neurocomputing 2023	14,000	74,903 boxes	10
UIIS [69]	ICCV 2023	4,628	Pixel-level instance masks	7
IOCfish5K [68]	CVPR 2023	5,637	659,024 center-point annotations	–
MarineInst20M [2]	ECCV 2024	2,420,851	∼19.2M instance masks	–
USIS10K [23]	ICML 2024	10,632	18,873 pixel-level salient-instance masks	7
CoralMask [70]	CVPR 2024	41,297	330,144 coral masks	–
SeaClear [24]	Scientific Data 2024	8,610	Instance masks and boxes	40
NautData [25]	NeurIPS 2025	∼158K	Detection/grounding boxes + multitask spatial supervision	–
SynTIDE [71]	CVPR 2025	∼54.51K	Semantic masks + depth maps	–
MARIS [3]	CVPR 2026	16,151	Pixel-accurate instance masks	158
AquaOV255 [4]	CVPR 2026	20,723	Instance-level pixel masks	255
UVLM [5]	AAAI 2026	2,109 videos; ∼0.86M frames	Frame-wise axis-aligned boxes	419
AquaticCLIP [66]	IEEE TNNLS 2026	2M image–text pairs	No GT spatial labels; 20.3M detector-derived regions	–
AquaGround	**This work**	**224,632**	**893,799 boxes**	**2,363**

Table 7: **Underwater dataset construction methodologies.** Representative resources differ in how geometry, semantics, supervision associations, and quality control are constructed.

Dataset	Publication	Data / geometry	Semantic construction	Supervision association	Quality control
FathomNet [22]	Scientific Reports 2022	Expert/community images; boxes + provenance	WoRMS taxonomy	Concept–box + acquisition metadata	Expert/community curation
RUOD [13]	Neurocomputing 2023	Real images; manual boxes	Predefined categories	Category–box	Unified LabelMe protocol
UIIS [69]	ICCV 2023	Underwater images; pixel-level instance masks	Category labels	Category–mask	COCO-style organization
IOCfish5K [68]	CVPR 2023	Video sampling; center-point labels	Counting supervision	Point–instance / image count	Filtering + consensus + double review
MarineInst20M [2]	ECCV 2024	Points/boxes/masks; human seeds + model expansion	Instance/image captions; multi-object instructions	Mask–description + multi-object association	Mask/sample filtering
USIS10K [23]	ICML 2024	Multi-source images; salient-instance masks	Saliency + instance categories	Instance–category–mask	Multi-annotator consensus + duplicate filtering
CoralMask [70]	CVPR 2024	Survey/web/public data; SAM-assisted expert masks	Coral groups + positive/negative instructions	Mask–instruction	Biologist review + negative pairs
SeaClear [24]	Scientific Data 2024	Multi-site/camera robotic acquisition; polygons	Category + material labels	Instance–category/material	Unified format; cross-site/camera organization
NautData [25]	NeurIPS 2025	Reused boxes and task annotations	Rules / template+MLLM / free-form instructions	Image/region/object QA + coordinates	Generate–review–regenerate
SynTIDE [71]	CVPR 2025	Text-conditioned image/depth/mask synthesis	Scene-text-driven dense labels	Text–image–depth–mask	Prompt dedup + cross-modal consistency
MARIS [3]	CVPR 2026	Multi-source masks; systematic re-annotation	Coarse → fine-grained taxonomy	Category–instance mask	Seen/overlap/unseen split design
AquaOV255 [4]	CVPR 2026	Web + existing data; ISAT/SAM-assisted masks	VLM-assisted fine-grained categories	Category–instance mask	Image cleaning + five-expert verification
UVLM [5]	AAAI 2026	Video clips; frame-wise object boxes	Biological/environmental, static/dynamic QA	Video/object/event–QA	General review + expert verification
AquaticCLIP [66]	IEEE TNNLS 2026	2M image–text pairs; detector-derived regions	Image- and instance-level descriptions	Image–text + instance-description supervision	Manual + CLIP-similarity filtering
AquaGround	**This work**	**Multi-source anchors; source boxes + screened SAM3; persistent IDs**	**Category → Attributes → Relation & Env. → Background & Position**	**Explicit query–instance sets via Sentence-BERT + constraints**	**Marine-expert checks + pairwise dedup + bundle cleaning**

Overall, the comparison shows that AquaGround jointly provides substantial annotation scale and rich object-centered supervision, spanning object descriptions, relations, environmental context, multiple language levels, and explicit one-to-many query–instance associations.

Table 8: **Four query levels in AquaGround.** Semantic scope increases across levels, while the number of matching instances is determined by each query and image.

Level	Main semantic content	Construction basis
Category	Object category.	Base object label.
Object Attributes	Category, scale, color, and geometric shape.	Anchored geometry and the first-stage attribute description.
Relations & Environment	Object attributes, relations to nearby entities, and local environmental context.	Second-stage relation and environment annotations.
Background & Position	Object attributes, relations, background context, and absolute position.	Third-stage background and position enrichment.

B Dataset Definitions and Construction Details

B.1 Query Levels and Matching Sets

Table 8 summarizes the four query levels and their construction basis. Query–instance associations are constructed within each image using `stsb-roberta-large` together with structured compatibility constraints. Category and scale must be consistent, while color and relation fields use a similarity threshold of 0.8; environmental and position fields are matched by their corresponding compatibility rules. The resulting matching set is query-specific and may contain multiple valid instances even for Background & Position queries.

B.2 Structured Environmental Attributes

Attribute semantics. AquaGround records five structured environmental attributes: *Water Color*, *Water Clarity*, *Illumination*, *Habitat*, and *Occlusion*. Water Color and Water Clarity are represented separately because visible coloration and visibility degradation reflect different aspects of underwater appearance (Akkaynak & Treibitz, 2018). Water Clarity distinguishes clear, slightly hazy, turbid, and unknown conditions; Illumination distinguishes natural, artificial, mixed, and unknown light sources; and Occlusion is categorized as none, slight, moderate, heavy, or unknown. Habitat describes the visible local environment, including open water, reef structures, crevices, hard bottoms, soft sediments, vegetation, rubble or shells, artificial structures, mixed habitats, and unknown conditions. These attributes describe visible semantic conditions rather than quantitative measurements of optical parameters, water quality, or biological habitat.

Object-level attachment. The five attributes are associated with individual objects through their `box_id`, rather than being stored only as image-level metadata. Objects in the same image may therefore share Water Color, Water Clarity, or Illumination while differing in local Habitat or Occlusion. The attributes contribute to the *Relations & Environment* and *Background & Position* query levels and can also be used for environment-conditioned analysis.

Combination counting. Environmental coverage is measured using unique per-image (*Water Color*, *Water Clarity*, *Illumination*, *Habitat*, *Occlusion*) tuples. Repeated identical tuples within the same image are counted once. AquaGround contains 662 distinct observed combinations, of which 479 contain no unknown or missing fields. These counts describe observed combinations rather than the Cartesian product of the five attribute vocabularies.

B.3 Source Composition and Annotation Completeness

Table 9 reports the source-collection inventory used to reconcile the benchmark total. The benchmark retains source identifiers and persistent object identifiers with geometric and semantic annotations. Existing boxes and supplementary SAM 3 proposals have separate provenance tags. The split sizes sum to the complete inventory: $190{,}409 + 22{,}690 + 11{,}533 = 224{,}632$. The source inventory is not a statement that every possible cross-source scene overlap has been eliminated.

Table 9: **Source inventory.** Active denotes the recorded image collection after preceding processing; complete denotes images with the required annotations. Excluded counts are annotation-completeness exclusions at this snapshot, not all duplicates removed earlier.

Source collection	Active	Complete	Excluded
AquaOV255	20,721	20,720	1
Jedi organism collection	7,151	7,147	4
MARIS	16,151	16,128	23
MKUO	10,043	10,043	0
MarineOrganism	19,488	19,475	13
NautData: underwater garbage	5,127	5,127	0
NautData: DeepFish	4,505	4,505	0
URPC / RUOD / DUO	20,216	20,208	8
UVLM	20,314	20,314	0
coralscapes	1,893	1,892	1
hkreeffish	729	729	0
ozfish	64,385	64,385	0
wildfish	33,966	33,959	7
Total	224,689	224,632	57

B.4 Progressive Semantics

Module II uses Qwen3-VL-32B-Instruct-FP8 (Bai et al., 2025) to construct the three descriptive query levels from an anchored object. Category, scale, and absolute position are retained as fixed fields, and each stage receives the preceding annotations as context while requesting only the additional semantic content required at that stage.

Category and Object Attributes. The first stage uses the target crop together with the category and scale to generate observable color and shape attributes and the corresponding `Object Attributes`. The prompt limits the description to at most 20 words and requests only attributes supported by the visible object, excluding annotation artifacts such as the displayed bounding box.

Relations & Environment. The second stage uses the local-context view and the preceding Object Attributes description to describe relations to nearby organisms, rocks, coral, anemones, or artificial structures and to assign the five structured environmental attributes, producing the `relative_caption`. The description is limited to at most 40 words. Relations use nearby entities as relational anchors rather than substituting whole-image position for an inter-object relation. When Habitat is labeled as `mixed`, a conditional Step 2.5 further refines its component habitats and their dominant or secondary roles. This refinement supplies additional context without introducing a fifth query level.

Background & Position. The third stage reuses the local-context view and preceding annotations to add nearby background types, colors, textures, and coverage together with the precomputed absolute position, forming the `absolute_caption`. Category queries are obtained directly from the source labels and therefore require no additional visual-generation stage.

Expert-check scope. After each semantic stage, three reviewers with marine-related academic training or professional experience inspect a random sample of approximately 1% of the images. They check object identity, visible attributes, relational referents, environmental labels, background descriptions, and consistency with the anchored position. Conditional mixed-habitat refinements are included in these checks when applicable. The sampling procedure screens for semantic errors and hallucinations but does not constitute manual verification of every generated description.

B.5 Duplicate Verification, Cleaning, and Split Protocol

Split construction. Source datasets with official train/validation/test partitions retain those assignments. Records from the remaining sources are partitioned within each dataset/subset group using

Table 10: **Object scale distribution.** Size ranges, instance counts, and percentages follow the values displayed in Figure 2. Scale is measured in the original image, regardless of the crop used for semantic annotation.

Scale	Bounding-box area (pixels^2)	Instances	Share
Tiny	$\leq 32^2$	49,500	5.54%
Small	32^2–96^2	184,100	20.60%
Medium	96^2–224^2	264,192	29.56%
Big	224^2–512^2	271,385	39.36%
Large	$> 512^2$	124,622	13.94%
Total		893,799	100%

a fixed-seed SHA-256 stable bucketing procedure. Cross-split duplicate conflicts are resolved after duplicate verification, with test taking precedence over validation and validation over training.

Duplicate verification. The same verification procedure is applied across source collections, with URPC/RUOD/DUO reported below as a detailed numerical audit. Exact hashes together with perceptual hashes (pHash and dHash) are used to retrieve candidate duplicates, which are then verified using strict pairwise SSIM checks (Wang et al., 2004). Each member of a duplicate group must match its retained representative directly; a match between A and B followed by a match between B and C is insufficient to group A with C. The confirmation rule requires zero pHash and dHash distance, grayscale SSIM of at least 0.9999, and RGB SSIM of at least 0.999. This direct-pair criterion prevents samples from being merged solely through transitive visual similarity.

Resolution and source audit. A retained representative keeps its own annotation bundle; annotations from different sources are not averaged or merged. Images, bounding-box annotations, semantic descriptions, and query–instance association records are handled together as a bundle, and archived or quarantined samples are excluded from subsequent association and training lists.

For the URPC/RUOD/DUO audit, 2,923 duplicate images were archived and 702 images were quarantined. The retained active snapshot contains 20,208 images with complete required annotations; Table 9 additionally identifies eight incomplete active images in this source collection.

The verification procedure targets near-identical image content rather than all visually similar scenes or consecutive frames. Accordingly, identity-duplicate removal should not be interpreted as establishing independence between related scenes or adjacent video frames.

B.6 Scale Coverage

Let $a = w_b h_b$ denote the absolute bounding-box area in square pixels, where w_b and h_b are the box width and height in the original image. Table 10 reports the five object-size groups, instance counts, and percentages displayed in Figure 2. The Tiny and Small counts sum to 233,600 instances (26.14%); the five counts sum to 893,799. Scale labels and the 5×5 absolute-position grid are computed before crop enlargement, so annotation crops do not alter either field.

B.7 Query Statistics

Counting unit and coverage. The final inventory contains 7,522,059 image–query records: 6,446,403 for training, 544,561 for validation, and 531,095 for testing. Each record is associated with a set of matching GT objects. Counts refer to query records rather than query–box pairs or globally unique text strings. Single-match and multi-match denote GT-set cardinality equal to one and greater than one, respectively. Every retained query has a valid, nonempty GT mapping; zero-match records total zero. This checks mapping completeness without implying that every semantic association has been manually verified.

Table 11: **Full query inventory and matching structure.** Panel (a) gives the four-level inventory by split; panel (b) aggregates matching cardinalities across all splits. Multi-match percentages use the corresponding row's total query count.

(a) Query counts by split

Semantic level	Train	Validation	Test	Total
Category	276,818	29,283	19,570	325,671
Object Attributes	2,394,832	205,344	187,652	2,787,828
Relations & Environment	1,512,813	124,312	131,673	1,768,798
Background & Position	2,261,940	185,622	192,200	2,639,762
Total	**6,446,403**	**544,561**	**531,095**	**7,522,059**

(b) Matching cardinalities across all splits

Semantic level	Single-match	Multi-match	Multi-match (%)
Category	196,709	128,962	39.60
Object Attributes	1,971,463	816,365	29.28
Relations & Environment	1,489,026	279,772	15.82
Background & Position	2,374,854	264,908	10.04
Total	**6,032,052**	**1,490,007**	**19.81**

Population scope. The validation and test inventories together contain 1,075,656 queries across all four levels. Excluding Category gives 515,278 validation and 511,525 test records, totaling 1,026,803 descriptive queries. Dataset-wide counts in this subsection include Category; the sample count for the reported Protocol B results is stated separately in Section 5.2. Dataset inventory size and the reported evaluation population are kept distinct.

Text length and semantic specificity. Table 12 reports word-length summaries for final query records. The Background & Position level has a mean of 19.38 words and a median of 10, showing that longer descriptions raise its average. The declining multi-match proportions across levels are aggregate observations; they do not establish a nested GT-set reduction for every individual object or query.

Table 12: **Word-length distributions of final query records.** The overall query-weighted mean is approximately 13.93 words, estimated from the rounded means below.

Semantic level	Mean	Median	Std	Min	Max
Category	1.04	1	0.22	1	3
Object Attributes	7.89	5	5.39	3	35
Relations & Environment	17.68	20	8.29	6	47
Background & Position	19.38	10	14.78	5	87

For per-level query counts n_ℓ and reported mean lengths $\bar{L}_\ell$, the overall estimate is

$$\widehat{\bar{L}} = \frac{\sum_{\ell=1}^{4} n_\ell \bar{L}_\ell}{\sum_{\ell=1}^{4} n_\ell} = \frac{104{,}765{,}596.96}{7{,}522{,}059} \approx 13.93 \text{ words.} \tag{6}$$

The numerator is a weighted reconstruction from rounded level means, not an exact integer word count. The three descriptive levels account for $7{,}196{,}388/7{,}522{,}059 = 95.67\%$ of the inventory.

C CONDITION ENCODING, ROUTING, AND OPTIMIZATION

C.1 SUPPLIED CONDITIONS AND ENCODINGS

AquaDetect uses the query together with auxiliary textual and structured conditions provided by AquaGround. As described in the main text, the query and auxiliary texts are encoded by the frozen SigLIP 2 encoder, while the structured conditions are processed by a dedicated condition encoder. The auxiliary textual input is constructed from four dataset fields: `category`, `self_caption`, `relative_caption`, and `absolute_caption`. These fields provide progressively enriched object-centered descriptions and are supplied directly from the dataset annotations.

The structured record contains seven fields: `scale`, `water_color`, `water_clarity`, `lighting`, `habitat`, `occlusion`, and `background`. In the final configuration, `scale` and `occlusion` are masked, leaving at most five active structured inputs: `water_color`, `water_clarity`, `lighting`, `habitat`, and `background`. Field values are mapped to identifiers by hashing and encoded by an independent `UnderwaterConditionEncoder`:

$$E = \Phi_{\text{env}}\big(\text{HashFields}(\mathcal{A}_{\text{env}}); m_{\text{env}}\big), \tag{7}$$

where m_{env} masks inactive fields.

The active structured inputs used by AquaDetect should not be confused with AquaGround's five annotated environmental dimensions. The dataset annotations include `occlusion`, whereas the structured-condition encoder masks `occlusion` and instead includes `background`. Masking a field from the structured-condition branch does not remove related information that may still appear in the supplied textual descriptions. All auxiliary conditions are dataset annotations available during both training and inference; they are not inferred online from the image branch or parsed from the input query.

For the set of valid queries associated with an image, AquaDetect constructs the query reference used by routing as

$$q_{\text{obj}} = \frac{\sum_q m_q t_q}{\sum_q m_q}, \tag{8}$$

where t_q is the frozen text feature of query q and m_q indicates whether the query is valid. Thus, multiple valid queries are summarized by a single pooled query reference for each image. This reference is computed before the visual backbone and is distinct from the learned object queries used later by the detection decoder.

C.2 ROUTING SCORES

The input text contains three major categories of cues: target semantic cues, underwater context cues, and detail-redundancy cues, amounting to six cues in total. Table 13 defines the corresponding six scores for the three cue families of AquaDetect.

For nonzero vectors, let $\kappa(u, v) = u^\top v/(\|u\|_2\|v\|_2)$ denote the raw cosine similarity. The stage index is omitted. For each score, v_i denotes the visual representation in the corresponding matching space, with any required projection implicit in this notation. The features t_{cls} and t_{rel} encode category and relation text, and $\{t_{\text{attr}}^{(\ell)}\}_{\ell=1}^{N_a}$ are attribute features. The attribute average applies when $N_a > 0$. For the query-relevance branch, this notation includes the projection from the 256-dimensional visual token to the 768-dimensional text space. No affine rescaling or clipping is applied in these score definitions.

Table 13: **Six routing scores in three complementary cue families.** The first five scores lie in $[-1, 1]$, whereas S_{vis} lies in $[0, 2]$. AWG weights their contributions to aggregation-center selection; visual similarity and spatial distance determine subsequent token-to-center assignment.

Cue family	Score and computation	Guidance	Routing role
Target-aligned semantic cues	$S_{\text{cls}}(i) = \kappa(v_i, t_{\text{cls}})$	Category text	Favor category-relevant centers.
	$S_{\text{attr}}(i) = \frac{1}{N_a}\sum_{\ell=1}^{N_a} \kappa(v_i, t_{\text{attr}}^{(\ell)})$	Appearance descriptions	Favor discriminative attribute evidence.
Underwater context cues	$S_{\text{rel}}(i) = \kappa(v_i, t_{\text{rel}})$	Relations and contextual descriptions	Favor relevant relational evidence.
	$S_{\text{env}}(i) = \kappa(v_i, e_{\text{env}})$	Active structured conditions	Account for supplied underwater context.
Detail–redundancy cues	$S_{\text{obj}}(i) = \kappa(v_i, q_{\text{obj}})$	Mean valid-query feature	Favor query-relevant visual content.
	$A_{ij} = \kappa(v_i, v_j)$ $S_{\text{vis}}(i) = 1 - \max_{j \neq i} A_{ij}$	Visual distinctiveness	Favor evidence with fewer similar alternatives.

A_{ij} is a visual similarity, whereas M_s records spatial correspondence. The maximum in S_{vis} excludes self-similarity. Small objects and boundary details can receive high importance through query relevance and visual distinctiveness, but their selection is determined by the combined ranking; no separate size- or boundary-based protection rule is applied. Grouping the scores into three families introduces no additional group-level softmax or dedicated family encoder.

C.3 Adaptive Cue Weighting

AWG produces one six-dimensional weight vector per image. The four descriptive-text levels are fused into a 768-dimensional vector and concatenated with the 256-dimensional structured-condition feature. A two-layer MLP, $\text{Linear}(1024, 256) \to \text{GELU} \to \text{Linear}(256, 6)$, maps this input to logits $a \in \mathbb{R}^6$. The weights are

$$w_k = \epsilon + (1 - 6\epsilon)\frac{\exp(a_k/\tau)}{\sum_{j=1}^{6} \exp(a_j/\tau)}, \qquad k = 1, \ldots, 6, \tag{9}$$

with $\epsilon = 0.010$ and $\tau = 1.0$, so $\sum_k w_k = 1$ and $w_k \geq 0.010$. The batch output has shape $B \times 6$ and is shared unchanged across the three stages. Current visual tokens, complete-query features, and the six scores are not direct AWG inputs. Holding the descriptive and structured conditions fixed while changing a localization query therefore leaves these weights unchanged, although the query-relevance scores can change.

At stage s, token importance and the selected centers are

$$I_s(i) = \sum_{k=1}^{6} w_k S_k^{(s)}(i), \qquad P_s = \text{TopKIndices}_i\big(I_s(i), K_s\big). \tag{10}$$

The fixed stage ratios $(0.875, 0.75, 0.50)$ are defined relative to the initial patch count, rather than applied recursively to the preceding stage. Input resolution determines the absolute counts. The routers are separate instances with their own pairing parameters. Changing visual features yields stage-specific scores, centers, and assignments despite shared AWG weights. In cue-family ablations, the removed scores are masked and the remaining active weights are renormalized.

C.4 Spatial Correspondence and Feature Recovery

Assignment and aggregation. For a non-center token i, let $\mathcal{N}_8(i)$ be its eight spatially nearest selected centers. Assignment maximizes

$$\ell_{ij} = \exp(\alpha_s)\kappa(x_i, x_j) - \mathrm{softplus}(\beta_s)\|p_i - p_j\|_2, \qquad j \in \mathcal{N}_8(i), \tag{11}$$

where p_i is the spatial centroid and α_s, β_s are the router's learned pairing parameters. The six guidance scores and AWG weights act through center selection and do not enter this pairing expression. Candidate restriction uses nearest neighbors, without an additional fixed distance threshold or forbidden region.

Let $\mathcal{C}_j$ include center j and all tokens assigned to it. The aggregated mass, feature, and centroid are

$$m'_j = \sum_{i \in \mathcal{C}_j} m_i, \qquad x'_j = \frac{\sum_{i \in \mathcal{C}_j} m_i x_i}{m'_j}, \qquad p'_j = \frac{\sum_{i \in \mathcal{C}_j} m_i p_i}{m'_j}. \tag{12}$$

Mass counts represented original patches and starts at one. Every patch token contributes to a center; selected centers also receive updates. Aggregation is lossy even though no source patch is excluded from the averages. The CLS token is retained separately and bypasses routing.

`TokenGridMapping` stores the original-grid-to-current-token assignment, representative-patch metadata, masses, centroids, and original grid dimensions. The initial assignment is the identity. Let $M_{s-1}(r)$ denote the current token assigned to original patch position r, and let π_s denote the old-token-to-new-token mapping produced by router s. The correspondence is updated by composition:

$$M_s(r) = \pi_s(M_{s-1}(r)), \tag{13}$$

while Equation 12 updates the associated token masses and spatial centroids.

Recovery at three stages. At every router output, TRA gathers the current token feature indexed by each original patch's assignment and reshapes the gathered sequence into a regular grid. Let $U_s = \mathrm{GridGather}(X_s, M_s)$ denote this recovered dense feature map. Patches belonging to the same aggregated token initially receive the same feature. TRA then applies a lightweight residual local refinement:

$$Y_s = U_s + \mathrm{PWConv}_{1\times1}(\mathrm{GELU}(\mathrm{DWConv}_{3\times3}(U_s))), \tag{14}$$

where $\mathrm{DWConv}_{3\times3}$ and $\mathrm{PWConv}_{1\times1}$ denote depthwise and pointwise convolutions, respectively. This operation restores a spatially organized representation suitable for localization, rather than reconstructing the independent pre-aggregation feature of each original patch.

For an 800×800 image, each stage yields a $256 \times 50 \times 50$ grid. The adapter bilinearly resizes the three stage outputs to 100×100, 50×50, and 25×25. An image-based CNN branch, `SpatialPriorModulev2`, independently extracts corresponding local features C_8, C_{16}, C_{32}. With $d_1 = 8, d_2 = 16, d_3 = 32$, the final construction is

$$F_{d_s} = \Psi_{d_s}([\mathrm{Resize}_{d_s}(Y_s); C_{d_s}]), \quad s = 1, 2, 3, \tag{15}$$

where concatenation is along channels and Ψ_{d_s} uses convolution and synchronized batch normalization to produce 256 channels. No early ViT block features are reused as skip connections, and no additional backbone blocks follow the third router.

C.5 Optimization and Output Interface

The detection features V produce bounding boxes $B = f_{\mathrm{box}}(V)$, complete-query scores $\mathbf{S}_{\mathrm{query}} = \mathcal{S}_q(V, T_q)$, and attribute scores $\mathbf{S}_{\mathrm{attr}} = \mathcal{S}_a(V, T_a)$. The output matrix $\mathbf{S}_{\mathrm{attr}}$ is distinct from the scalar routing score $S_{\mathrm{attr}}(i)$. The detection backend and associated loss components use the implementations in Wei et al. (2026). AquaDetect's cue-guided routing and spatial-recovery modules are trained with the joint objective

$$\mathcal{L} = \sum_{k \in \mathcal{K}} \lambda_k \mathcal{L}_k, \qquad \mathcal{K} = \{\mathrm{query}, \mathrm{attr}, \mathrm{box}, \mathrm{giou}, \mathrm{fgl}, \mathrm{ddf}\}. \tag{16}$$

Query and attribute losses supervise semantic alignment; the remaining terms provide box-localization supervision. SigLIP 2 remains frozen, and TRA uses no additional reconstruction loss.

Top-k indices are discrete. AWG receives surrogate gradients through a straight-through gate on selected tokens: the detection loss propagates through recovered features, selected-token gates, token importance, and the weighted score combination. Router pairing parameters receive gradients through a soft-assignment straight-through estimator. Feature aggregation and TRA use ordinary differentiable operations. The optimization settings and common checkpoint-selection rule are given in Section 5.1.

D Evaluation Protocol Details

D.1 Protocol A: Query-Conditioned COCO Evaluation

Protocol A uses the official COCO bounding-box evaluator (Lin et al., 2014), with one task-specific adaptation for open-text detection. Each image–query record is assigned a unique evaluation category ID, so its associated ground-truth boxes and query-conditioned predictions form an independent evaluation unit. Consequently, AP and AR are aggregated over query-specific units rather than over a fixed semantic category vocabulary. Before evaluation, predictions from all queries associated with an image are ranked by confidence and the top 100 are retained globally for that image. COCO matching, ignore rules, and precision–recall interpolation are otherwise left unchanged.

The IoU threshold set is

$$\mathcal{T} = \{0.50, 0.55, \ldots, 0.95\}.$$

AP averages precision over $\mathcal{T}$, while AP_{50} and AP_{75} use fixed IoU thresholds of 0.50 and 0.75. AR is evaluated over the same threshold set under the corresponding `maxDets` constraint. All values are reported on a 0–100 scale. Table 14 summarizes the exact settings.

Table 14: **COCO bounding-box metric settings for Protocol A.** $\mathcal{T}$ contains the ten IoU thresholds from 0.50 to 0.95.

Metric	IoU thresholds	Area range	`maxDets`
AP	$\mathcal{T}$	All	100
AP_{50}	0.50	All	100
AP_{75}	0.75	All	100
AP_{s}	$\mathcal{T}$	Small	100
AP_{m}	$\mathcal{T}$	Medium	100
AP_{l}	$\mathcal{T}$	Large	100
AR@1	$\mathcal{T}$	All	1
AR@10	$\mathcal{T}$	All	10
AR@100	$\mathcal{T}$	All	100
AR_{s}	$\mathcal{T}$	Small	100
AR_{m}	$\mathcal{T}$	Medium	100
AR_{l}	$\mathcal{T}$	Large	100

Detection limits and object scales. AR@1, AR@10, and AR@100 use COCO's standard `maxDets` settings after the image-level top-100 filtering described above. The COCO scale groups are defined by absolute bounding-box area in pixels: Small uses the range $[0, 32^2]$, Medium $[32^2, 96^2]$, and Large $[96^2, 10^{10}]$, with boundary handling following the official evaluator. The three COCO evaluation groups use a different grouping and naming convention from the five dataset-level size groups in Figure 2 and Table 10. Both schemes use absolute bounding-box area, but their group labels are not interchangeable. In particular, the dataset-level Small group is not the COCO Small group used for AP_{s} or AR_{s}.

D.2 Protocol B: Multi-Referent Top-1 Localization

For query i, let $\hat{b}_i$ denote the highest-scoring predicted box and $\mathcal{G}_i$ the set of all valid ground-truth referents associated with that query. We define

$$u_i = \max_{g \in \mathcal{G}_i} \mathrm{IoU}(\hat{b}_i, g), \qquad \mathrm{Acc}\,@\tau = \frac{100}{M} \sum_{i=1}^{M} \mathbf{1}[u_i \geq \tau].$$

Thus, Protocol B measures whether the top prediction localizes *any* valid referent for the query. It does not require recovering every matching instance; multi-instance detection is instead evaluated under Protocol A.

For the valid referent associated with the maximum overlap above, let I_i and U_i denote the intersection and union areas with $\hat{b}_i$. The reported overlap metrics are

$$\mathrm{mIoU} = \frac{100}{M} \sum_{i=1}^{M} \frac{I_i}{U_i}, \qquad \mathrm{cIoU} = 100 \frac{\sum_i I_i}{\sum_i U_i}. \tag{17}$$

mIoU gives equal weight to each query, whereas cIoU aggregates intersection and union areas across the evaluation set and therefore gives greater weight to cases with larger union areas.

AR@1 and Acc@τ should not be interpreted as the same quantity. AR@1 follows COCO detection matching and averages recall across the IoU threshold set $\mathcal{T}$ under a one-detection limit, whereas Acc@τ evaluates the top prediction of each query against any of its valid referents at a single specified IoU threshold.