ONE OBSERVATION, ONE IDENTITY: JOINT ASSOCIATION AND TRUTH DISCOVERY FOR REMOTE-SENSING AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Remote-sensing agents must determine whether observations across modalities and time describe the same object, which reported attributes are correct, and how much to trust each tool output. However, approaches that fix correspondence before attribute inference or associate tool reports independently do not explicitly preserve the shared identity of an observation while resolving conflicting evidence. Such separation can propagate early matching errors or assign reports from a single detection to different objects. We introduce OneID-Agent, a framework for joint identity association and truth discovery built on the principle of *one observation, one identity*. A variational formulation ties all claims from an observation to a shared latent identity and alternates identity and attribute-truth updates using context-conditioned reliability. Attribute evidence can thus revise an observation's identity and reallocate all its claims together. Physical candidate constraints restrict admissible associations, while provenance budgets limit the influence of related reports. A separate evidence check guides answering, further acquisition, or abstention. Further, we construct RSTIBench to evaluate tool execution, complemented by independently annotated identities and attribute truths on optical–SAR and temporal imagery. On the 94 held-out image-pair samples, the shared identity method achieves a pair F1 of 0.7320 and a pair recall of 0.5820.

1 INTRODUCTION

Remote sensing target interpretation (RSTI) is evolving from scene-level recognition and single-image perception toward object-level reasoning over multiple observations. Modern Earth observation systems must not only locate objects, compare their attributes, and identify temporal changes across optical, synthetic aperture radar (SAR), and multi-temporal imagery, but also determine whether different observations refer to the same object and whether their conclusions are supported by consistent evidence. This shift makes cross-modal and cross-temporal interpretation more than a multi-source input problem.

Different observation sources provide complementary information but also introduce distinct uncertainties. Optical imagery offers rich spectral, textural, and appearance cues, yet is sensitive to clouds, atmospheric conditions, and illumination. SAR imagery supports all-weather and day-night observation but is affected by speckle noise and geometric distortions. Multi-temporal imagery reveals object-state changes, while seasonal effects, acquisition variations, and background changes can be mistaken for actual target changes. Consequently, relying on a single modality may lead to conclusions supported by insufficient evidence, whereas naively combining multiple modalities may obscure cross-source inconsistencies instead of resolving them.

Existing work has advanced along two directions. First, cross-modal identity correspondence methods address object matching across sensors and acquisition times, including optical–SAR re-identification with HOSS ReID and MOS (Wang et al., 2025a; Zhao et al., 2026b). Second, RS observation models, agents, and benchmarks advance multimodal understanding and executable reasoning, including Copernicus-FM and EarthDial (Wang et al., 2025c; Soni et al., 2025), tool-augmented Earth-Agent and OpenEarthAgent (Feng et al., 2026a; Shabbir et al., 2026), and the GeoMMBench and EO-Gym evaluation frameworks (Xiao et al., 2026; Ma et al., 2026).

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.

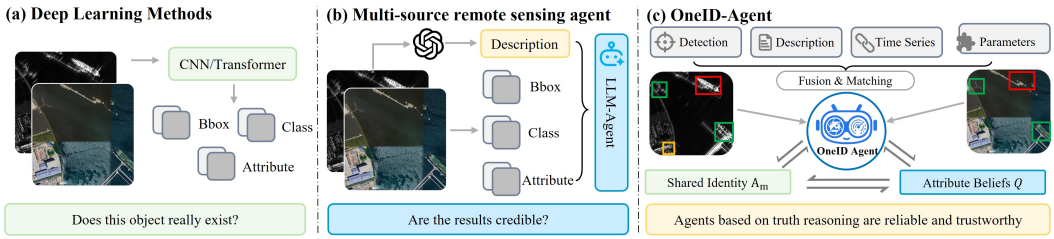

Figure 1: Three correspondence strategies. Target re-recognition method for directly outputting results; text-guided correction revises claims independently; OneID-Agent jointly updates a shared observation-level identity A_m and attribute beliefs Q across optical, SAR, and temporal evidence.

Despite this progress, most approaches still treat cross-modal and cross-temporal interpretation primarily as a multi-source availability problem: the system receives several observations and predicts an answer. Reliable object-level interpretation instead requires two coupled capabilities. The first is joint identity association: determining whether claims from different observations belong to the same latent object. The second is truth discovery: estimating that object's true attributes or states when claims are noisy, redundant, or contradictory. Identity association determines which evidence can be aggregated, while truth discovery determines how it should be weighted; errors in either direction can propagate to the other. So the key question in existing research is: how can a remote sensing agent jointly infer object identities and attribute truths across modalities and time while deciding whether the evidence is sufficient for a reliable conclusion?

We propose OneID-Agent under the principle of "one observation, one identity." Multiple claims produced by the same local detection or bounded temporal event share a common latent identity instead of being independently assigned to different targets. OneID-Agent alternates between identity association and truth discovery: the identity distribution determines how claims contribute to object-specific truths, while inferred attribute truths provide semantic feedback that refines identity assignments. Orientation, fine-grained attributes, and temporal states can therefore disambiguate identities, resolve conflicts, and reallocate other claims from the same observation. An independent evidence-state branch models support, conflict, and uncertainty across modality and time requirements, enabling the agent to answer, acquire additional evidence, or abstain when evidence is insufficient. Meanwhile, we build RSTIBench to evaluate tool execution and evidence acquisition in remote sensing agents, and annotate real identities and attributes, so that it can be used to evaluate identity reasoning. Our contributions are:

- **OneID Object-Level Evidence Modeling.** Binding multiple attribute declarations generated from a single observation to a unified latent object identity addresses the challenges of traditional hard matching, which is difficult to correct.
- **Joint identity association and truth discovery.** We develop a physically constrained inference mechanism that alternates identity and truth updates using semantic feedback, reliability, capacity, null identities, and provenance budgets.
- **RSTIBench and systematic evaluation.** We construct RSTIBench, a benchmark covering object grounding, spatial reasoning, fine-grained attributes, optical–SAR interpretation, and multi-temporal change analysis, and the paired images were also marked with ground truth attributes and identities.

2 Related Work

Deep learning for remote sensing. Deep-learning methods have substantially improved remote-sensing (RS) interpretation across sensors, resolutions, and acquisition times. Cross-modal correspondence methods such as HOSS ReID and MOS learn transformer-based optical–SAR ship re-identification and narrow the modality gap (Wang et al., 2025a; Zhao et al., 2026b). Foundation and observation models extend this paradigm to heterogeneous Earth observation: Copernicus-FM, Prithvi-EO-2.0, EarthDial, SkyMoE, SkySense, TESSERA, and billion-scale models learn multi-sensor, multi-temporal, and vision-language representations (Wang et al., 2025c; Szwarcman et al., 2026; Soni et al., 2025; Liu et al., 2026a; Guo et al., 2024; Feng et al., 2026b; Cha et al., 2026), and

multimodal pretext objectives align optical, SAR, and other modalities (Nedungadi et al., 2024). These methods provide strong perception, but they do not maintain a shared observation-level identity while allocating multiple conflicting attribute claims (Figure 1).

Remote-sensing agents. Remote-sensing agents extend multimodal perception with tool orchestration, grounded interaction, and executable reasoning. Earth-Agent and OpenEarthAgent build reasoning trajectories over RS tools (Feng et al., 2026a; Shabbir et al., 2026); OpenEarth-Agent and RemoteAgent add tool creation and reinforcement-learned planning (Zhao et al., 2026a; Yao et al., 2026a); and ZoomEarth performs active perception (Liu et al., 2026b). Benchmarks evaluate expert understanding, interactive evidence gathering, and operational reasoning, including GeoMMBench, EO-Gym, GeoDisaster, and TerraBench (Xiao et al., 2026; Ma et al., 2026; Hasan et al., 2026; Nguyen et al., 2026). OneID-Agent instead updates identity and attribute truth jointly, then checks evidence sufficiency, while RSTIBench isolates this shared-identity inference from tool-use accuracy.

Truth discovery. Truth discovery estimates facts from conflicting reports by modeling source reliability, dependence, and uncertainty (Wang et al., 2025b). Classical estimator aggregation and crowdsourcing models learn annotator error rates, expertise, and correlation (Dawid & Skene, 1979; Raykar et al., 2010; Whitehill et al., 2009; Li et al., 2019); social-sensing methods extract and verify claims from text, images, and videos (Shao et al., 2021), and two-stage assessment extends this to disaster false claims (Yao et al., 2026b). These studies reason over noisy claims but not the observation-level coupling needed to allocate conflicting tool reports while RS identity remains uncertain. OneID-Agent instead couples association and attribute-truth discovery through a shared latent identity per observation, so attribute evidence can revise the identity and reallocate its other claims.

3 METHOD

Taking optical–SAR cross-interpretation as an example, we must address three questions: (1) Do the detections refer to the same object? (2) Are the reported object attributes consistent? (3) How reliable are the tool outputs? These questions are coupled: attribute agreement can support identity association only when the reports are reliable, while apparent disagreement may arise from different objects, erroneous reports, or genuine state changes across acquisition times. Resolving any one question in isolation can therefore produce a conclusion undermined by uncertainty in the other two. This motivates the variational inference framework underlying OneID-Agent (Figure 2): it maintains approximate beliefs over shared observation identities and object–attribute–time truths, and alternately updates them using context-conditioned tool reliability.

3.1 OBSERVATIONS, CLAIMS, AND SHARED IDENTITY

A local optical detection can yield several reports, such as an object category and a heading. These reports describe the same detected instance, but the persistent object to which that instance corresponds remains to be inferred. We therefore distinguish the observation that produced the reports, the claims that record their content, and the object-specific facts whose truths are to be estimated.

We represent this local detection as a primitive observation o_m; the same representation also covers SAR detections and scoped temporal events. The index set $\mathcal{I}_m$ identifies the claims attached to o_m. For each $i \in \mathcal{I}_m$, the evidence record is $\gamma_i = (m(i), a_i, \tau_i, s_i, v_i, \mathbf{x}_i, \mathcal{G}_i)$. Using an optical image as an example, $m(i)$ denotes the local detection associated with claim i; a_i is the reported attribute, such as object category or heading; τ_i is the acquisition time; s_i identifies the reporting tool; v_i is the output of the tool; $\mathbf{x}_i$ contains visibility-related covariates, such as visible fraction and pixel size; and $\mathcal{G}_i$ records provenance parents, such as the source image or preceding claims.

A claim records what a tool reported about the detection; a fact specifies which object's attribute is being estimated. If the optical detection corresponds to candidate object k, its heading report concerns that object's heading at acquisition time τ. We index this fact by $j = (k, a, \tau)$, where $k \in \mathcal{C}_m$ denotes a candidate persistent object, a specifies the reported attribute, and τ specifies the acquisition time. The same object observed at a different time therefore corresponds to a different fact. Under candidate identity k, claim i is associated with $j(i, k) = (k, a_i, \tau_i)$. All claims in $\mathcal{I}_m$

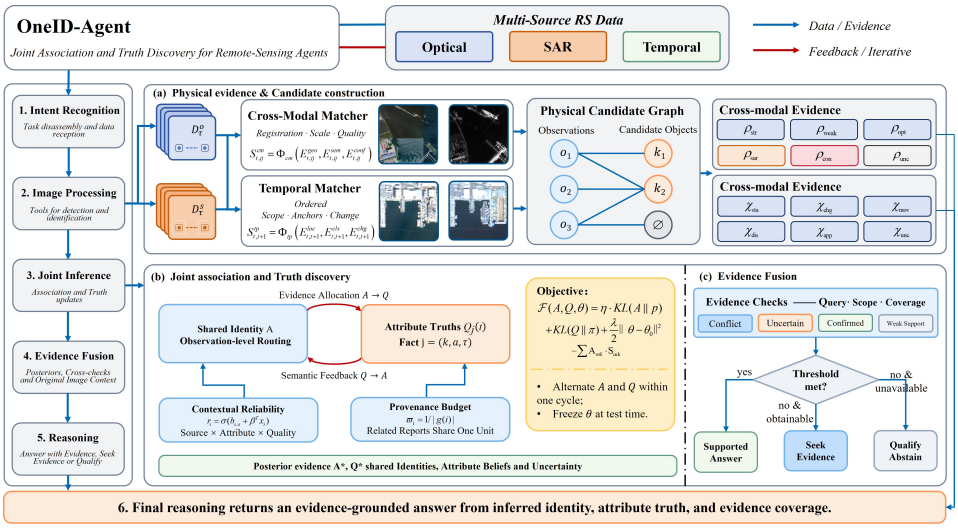

Figure 2: Overall framework of OneID-Agent. The agent routes a query to tools and skills, organizes cross-modal and temporal observations in persistent object-level records, and uses intermediate evidence to guide further execution and the final response.

share $Z_m \in \mathcal{C}_m \cup \{\varnothing\}$, while each fact j has a latent truth T_j. The inference model maintains an association belief A, truth beliefs Q, and context-conditioned claim correctness values r:

$$\begin{aligned} A_{mk} &= q(Z_m = k), \qquad Q_j(t) = q(T_j = t), \\ r_i &= P\big(v_i = T_{j(i,k)} \mid Z_{m(i)} = k, s_i, \mathbf{x}_i\big), \qquad k \neq \varnothing. \end{aligned} \tag{1}$$

Here, A_{mk} is the belief that observation o_m corresponds to candidate object k, and $Q_j(t)$ is the belief that the truth of fact j is t. The quantity r_i is the probability that the reported value agrees with the associated fact, conditioned on its reporting tool and observation context. Thus, association beliefs determine how each report contributes to object-specific truths, while truth beliefs provide semantic support for candidate identities. The distribution q denotes the approximate posterior beliefs described in Appendix C.7. Because r_i is shared across admitted candidate identities, candidate uncertainty is represented through A_{mk} rather than candidate-specific reliability scores.

One observation, one identity. The preceding representation separates what a detection reports from which persistent object those reports describe. We therefore use the primitive observation, rather than an individual claim, as the allocation unit: all claims indexed by $\mathcal{I}_m$ are routed through the same association row A_m, including its null component. For an optical detection that reports both a category and a heading, the two claims consequently receive the same candidate-identity weights, while retaining separate object–attribute–time facts. If the heading provides evidence that distinguishes candidates, its update to A_m changes how the category claim is allocated to those facts as well. This shared routing, rather than a direct assumption that attributes predict one another, is the source of the cross-attribute reallocation characterized below.

3.2 Physical Candidate Graph

The shared row does not make every object a possible match. We first construct a sparse candidate graph from physical evidence: overlap, displacement, scale, and registration for spatial observations, plus ordered scope and compatible anchors for temporal events. Candidate membership is fixed within an inference cycle and is not altered by the inferred attribute values. This separation gives the semantic updates a physically grounded support set rather than allowing a report to create a new identity.

The physical prior uses pair features $\mathbf{f}_{mk}$ and weights $\mathbf{d}, d_\varnothing$ calibrated on independent matched/unmatched observations and then frozen:

$$p_{mk} = \frac{\exp(\mathbf{d}^\top \mathbf{f}_{mk})}{\exp(d_\varnothing) + \sum_{h \in \mathcal{C}_m} \exp(\mathbf{d}^\top \mathbf{f}_{mh})}, \qquad p_{m\varnothing} = 1 - \sum_{k \in \mathcal{C}_m} p_{mk} > 0. \tag{2}$$

The feasible set $\mathcal{A}$ contains nonnegative, row-normalized matrices with zero off-graph mass. For each acquisition/primitive-channel block B, distinct detections obey $\sum_{m\in B} A_{mk} \le 1$ for real objects k. The constraint is imposed on detections, not attributes: a category and a heading from one detection consume one identity slot together. Different acquisitions have separate blocks, while the null component has unlimited capacity (Appendix C.5). Deterministic restatements of the physical features remain measurements of the same evidence and are not added as semantic likelihood terms.

3.3 Joint Association and Truth Discovery

Contextual reliability and evidence budgets. Given this candidate graph, the remaining question is how much each report should influence identity and truth. We model that influence through context-conditioned correctness rather than through candidate-specific reliability. Source–attribute intercepts capture channel-specific accuracy, while observation covariates adjust it for visibility and acquisition conditions. With sigmoid σ and $\theta = \{b_{s,a}, \boldsymbol{\beta}_a\}$,

$$r_i = \sigma(b_{s_i,a_i} + \boldsymbol{\beta}_{a_i}^{\mathsf{T}} \mathbf{x}_i), \qquad L_i(t;\theta) = \begin{cases} r_i, & v_i = t, \\ (1-r_i)/(K_{a_i}-1), & v_i \neq t. \end{cases} \tag{3}$$

The same correctness model is evaluated under every admitted candidate identity. Thus, r_i describes the reporting channel and local observation context; it cannot directly encode the candidate identity that the report is meant to help infer. We include missingness indicators in $\mathbf{x}_i$ so that the reliability model can distinguish an observed context from an unavailable one; the resulting reliability channel is bounded as specified in Appendix C.7.

Reports from the same observation or processing chain should not count as independent votes. After removing exact replays and deterministic restatements, local provenance groups connect claims that share a primitive, input roots, or a derivation path. We therefore share one unit of weight among the claims in a group, assigning $\varpi_i = 1/|g(i)|$ so that each group contributes one evidence unit and $\sum_{i\in\mathcal{I}_m} \varpi_i \le 1$; repeated or co-derived reports cannot dominate the update merely by increasing their number. This is a conservative control on repeated evidence, not an assumption that all reports are independent (Appendix C.6).

Objective. The model now combines the four ingredients in one objective. The association KL term keeps each shared row near the physical prior p_m, with η controlling the strength of that anchor. The truth KL term prevents unsupported object–attribute–time facts from drifting away from their declared priors π_j. The reliability regularizer keeps any adapted parameters near the independently calibrated reference θ_0. Finally, the evidence term rewards an assignment when the reports attached to that observation are compatible with the candidate-specific truths. The resulting optimization allows semantic evidence to revise identity, but only within the physical graph and under the declared evidence controls.

Let $\ell_{ik} = \mathbb{E}_{Q_{j(i,k)}} \log L_i(T_{j(i,k)};\theta)$ and $S_{mk} = \sum_{i\in\mathcal{I}_m} \varpi_i \ell_{ik}$. Null uses a budgeted uniform background, $S_{m\varnothing} = \sum_{i\in\mathcal{I}_m} \varpi_i \log(1/K_{a_i})$. For truth priors π_j uniform over each attribute's declared value domain, physical-prior strength $\eta \ge 1$, and regularization $\lambda > 0$ toward calibrated θ_0, minimize

$$\begin{aligned} \mathcal{F}(A,Q,\theta) = \eta \sum_m \mathrm{KL}(A_m \| p_m) + \sum_j \mathrm{KL}(Q_j \| \pi_j) + \frac{\lambda}{2}\|\theta - \theta_0\|_2^2 \\ - \sum_m \sum_{k\in\mathcal{C}_m \cup \{\varnothing\}} A_{mk} S_{mk}, \qquad A \in \mathcal{A},\ \theta \in \Omega. \end{aligned} \tag{4}$$

Within one inference cycle, observations, claim attachments, provenance groups, priors, and contexts are fixed; only the belief blocks are updated. Appendix B.1 reports the shared hyperparameters.

Association update. Equation (4) couples the association and truth blocks through the evidence term, and the updates inherit this coupling. The association block uses semantic scores computed from Q, while the truth block uses responsibilities from A. We therefore minimize the same objective by alternating block updates, rather than assigning identities once and then aggregating attributes. Reliability parameters are updated only in the declared fitting or adaptation phase. For fixed Q, θ, semantic compatibility reweights the physical prior and the result is projected onto the

feasible set:

$$\widetilde{A}_{mk} \propto p_{mk} \exp(S_{mk}/\eta), \qquad A \leftarrow \operatorname*{arg\,min}_{A^{\mathrm{trial}} \in \mathcal{A}} \sum_{m} \mathrm{KL}(A_m^{\mathrm{trial}} \| \widetilde{A}_m). \tag{5}$$

The KL/Bregman projection enforces row normalization, the candidate graph, and detection capacities, while retaining unlimited null capacity. Consequently, semantic support can redistribute a row only among admissible candidates and cannot bypass the one-detection-one-slot constraint.

Truth update. Holding A, θ fixed, the truth update performs the complementary operation. For $j = (k, a, \tau)$, let $\mathcal{E}_j$ contain the claims with attribute a and scope τ that could describe candidate k. Update

$$Q_j(t) \propto \pi_j(t) \exp\left(\sum_{i \in \mathcal{E}_j} \varpi_i A_{m(i)k} \log L_i(t; \theta) \right). \tag{6}$$

Each report is weighted by the responsibility of its entire observation for candidate k. Thus, when an attribute changes the shared identity row, every claim attached to that observation is reallocated through the same row, including claims that did not drive the change. With no incoming responsibility, $Q_j = \pi_j$ remains prior-only belief.

Interpreting feedback. For fixed Q, θ and real candidates k, h with inactive capacities, the shared row satisfies

$$\log \frac{A_{mk}}{A_{mh}} = \log \frac{p_{mk}}{p_{mh}} + \frac{1}{\eta} \sum_{i \in \mathcal{I}_m} \varpi_i w_i \big[Q_{j(i,k)}(v_i) - Q_{j(i,h)}(v_i) \big], \tag{7}$$

where $w_i = \log((K_{a_i} - 1) r_i / (1 - r_i))$. The first term is the physical log-odds; the second is the aggregate semantic correction supplied by all claims from the observation. A report at chance level, $r_i = 1/K_{a_i}$, has $w_i = 0$ and cannot distinguish candidates. Above chance, agreement with a candidate increases its odds; below chance, the channel is anti-correlated and reverses that direction. The semantic correction in one update is bounded by $\max_{i \in \mathcal{I}_m} |w_i| / \eta$; active capacities add the corresponding dual terms. This bound protects a physical margin, not the correctness of the resulting identity. The candidate graph still preserves off-graph zeros, while provenance weights bound a group's direct contribution to truth. Formal statements are given in Appendix C.7.

Calibration and test protocol. Independent calibration estimates θ_0 and the physical-prior parameters. The default system evaluates r_i from the frozen calibrated model θ_0; at test time no reliability parameter is updated and inference alternates only A and Q. An auxiliary-fitting variant estimates supported reliability parameters before test inference and then freezes them. Calibration details and identifiability conditions are given in Appendices C.8 and B.1.

3.4 Evidence States and Answering

Inference yields beliefs over identity and truth, while answerability is assessed separately. For each query-candidate pair, $z_{q,k} \in \{\textit{Conflict}, \textit{Uncertain}, \textit{Confirmed}, \textit{Weak Support}\}$ summarizes whether the identity, truth, modality, and temporal requirements are satisfied (Appendix C.4). The agent answers only when these conditions pass development-set thresholds; a prior-only belief is never sufficient. Otherwise, it acquires additional admissible evidence, updates the candidate graph and provenance groups, and reruns inference. If no informative action remains, it reports the unresolved scope or abstains (Appendix C.9).

4 RSTIBench: Tasks and Inference Reference

RSTargetInterpretation-Bench (RSTIBench) adds tasks, object-level annotations, and execution trajectories to public or licensed imagery. Its 13,908 training and 1,414 test trajectories cover grounding, spatial measurement, fine-grained attributes, optical–SAR interpretation, and temporal change (Figure 3).

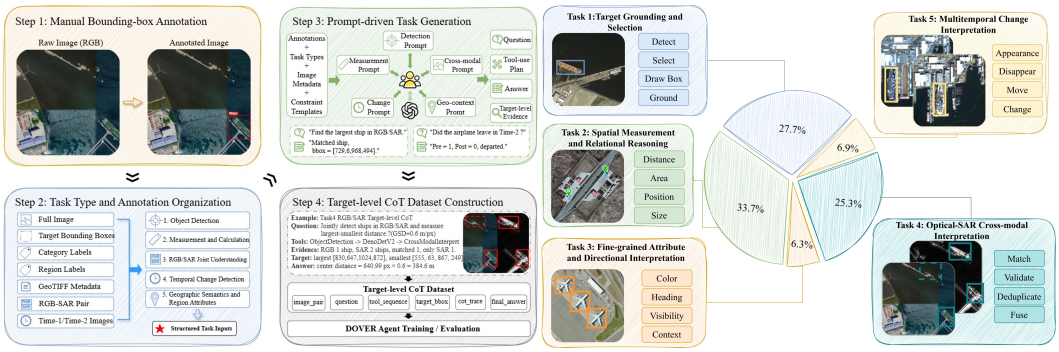

Figure 3: RSTIBench construction (left) and task coverage (right), with proportions computed over the combined training and test trajectories.

Sources and construction. Sources include GF-2, GF-3, DIOR (Li et al., 2020), DOTA (Xia et al., 2018), SSDD (Zhang et al., 2021), and SIVED (Lin et al., 2023). Annotators author the tasks and execution trajectories from annotated inputs, retaining actions, intermediate evidence, provenance, and answers; GPT-5 mini is used only for structured consistency checks and issue flagging, not for authoring tasks, trajectories, or final annotations. Appendix A.1 details construction.

Splits and annotation. Splits precede construction and keep source images, scene/object groups, and paired derivatives together. Two annotators author and review each sample, with a third adjudicating disagreements; GPT-5 mini performs structured consistency checks on trajectories but does not establish final annotations. Appendix A.2 reports distributions and examples.

Inference reference. Identity and truth are separately adjudicated on the same imagery, independently of evaluated tool outputs, with scene/acquisition isolation across splits. The reference includes unmatched observations, quality metadata, and provenance. Identity is scored per primitive, truth per object–attribute–time fact, and decisions per query, not by trajectory counts or *Confirmed* labels (Appendix B.10).

5 Experiments

5.1 Experimental Setup

Our experiments address two questions: (1) Can the agent correctly invoke tools, obtain usable results, and integrate module outputs into an answer? (2) Does shared-identity inference improve object association and attribute-truth estimation?

For agent execution, we use the step-level and end-to-end evaluation protocols(Shabbir et al., 2026) on RSTIBench. Step-level metrics assess action validity, tool selection, argument agreement, and answer correctness; end-to-end metrics assess tool-call Macro-F1, sequence and argument agreement, and the overall E2E score. All baselines share the atomic-action schema and are evaluated against reference trajectories; agreement measures reference fidelity rather than the uniqueness of a valid execution.

For shared-identity inference, Table 2 evaluates the 94 held-out image-pair samples and their detection boxes, with independently annotated identities and attribute truths. Ablations fix observations, candidate graphs, reports, and calibration access. Precision, Recall, F1, and Pair Brier evaluate pairs; ID Acc. and ID ECE evaluate observation identity; Truth Acc. evaluates object–attribute–time facts. Lower Brier and ECE are better; higher values are better for the other metrics. Appendix B.11 defines scoring and the 374/94 holdout-sites split; Tables 12–13 evaluate query-level answering separately.

Qwen3-4B-Instruct-2507 (Yang et al., 2025) is fine-tuned with assistant-only LoRA supervision using MS-Swift (Zhao et al., 2025). Frozen LAE-DINO (Pan et al., 2025) and DenoDetV2 (Ni et al., 2026) supply optical/SAR detections. Hyperparameters, calibration, and training details are provided in Appendices B.1–B.2.

Table 1: Execution on RSTIBench under step-level and end-to-end protocols. Bold/underline: best/second-best. Attempt/Blocked/Steps are the mean successful-call rate, rejected-call rate, and executed steps per trajectory; Steps is reported for reference and not ranked.

Method	Step						End-to-end							
	Inst.	Tool	Arg_1	$\text{Arg}_{\text{strict}}$	Arg_2	Answer	Macro	Seq.	Tool Rec.	Norm. Arg.	E2E	Attempt	Blocked	Steps
Qwen3-4B-Instruct	97.63	71.71	59.63	27.20	27.20	<u>59.25</u>	39.46	7.21	8.56	1.62	28.88	**84.83**	**3.28**	1.74
Qwen3-8B (No Thinking)	97.32	<u>75.19</u>	<u>61.17</u>	<u>27.45</u>	<u>27.45</u>	53.47	<u>51.99</u>	<u>14.78</u>	21.43	4.80	<u>31.96</u>	83.33	6.33	3.00
Qwen2.5-7B-Instruct	93.82	68.00	56.09	26.39	26.41	55.10	41.33	8.63	18.18	3.55	25.01	65.61	21.24	3.73
Qwen2.5-3B-Instruct	95.61	62.41	49.29	18.64	18.75	36.25	36.96	12.52	17.11	1.33	24.89	57.21	34.31	4.19
Llama-3.1-8B-Instruct	<u>99.50</u>	71.16	57.56	23.27	23.41	47.80	47.80	8.42	<u>27.65</u>	<u>5.19</u>	27.51	70.24	18.83	4.64
InternLM3-8B-Instruct	93.27	65.42	54.50	25.25	25.25	50.10	35.52	5.10	6.25	1.20	24.95	83.50	4.70	1.61
GLM-4-9B-Chat	97.88	68.15	54.93	23.27	23.33	52.57	37.00	5.80	7.00	1.10	26.20	<u>83.90</u>	<u>4.20</u>	1.68
OneID-Agent	**99.67**	**79.78**	**64.46**	**46.69**	**46.71**	**78.90**	**78.18**	**42.72**	**58.06**	**31.94**	**44.90**	83.74	15.44	4.28

Table 2: Shared-identity ablations with 374 samples used for fitting and 94 held-out samples used for evaluation. Columns: identity association A, semantic feedback S_{mk}, capacities/null $\mathcal{A}/\varnothing$, evidence weights ϖ_i, reliability r_i (θ_0: frozen calibrated context model; cal.: recalibrated table; θ: additionally fitted context model). Bold/underline: best/second-best; ↓: lower is better. n/a: pair-level metrics are inapplicable to U's claim-level identity.

	Modules					Performance						
Method	A	S_{mk}	$\mathcal{A}/\varnothing$	ϖ_i	r_i	Precision	F1	Recall	Pair Brier ↓	Truth Acc.	ID Acc.	ID ECE ↓
H	hard	—	—	1	θ_0	0.3401	0.4553	**0.6885**	0.2618	0.6885	0.8852	0.1482
S	prior	—	—	1	θ_0	0.5232	0.5788	<u>0.6475</u>	0.2440	0.7064	0.9021	0.1324
U	claim	✓	—	1	θ_0	*n/a*	*n/a*	*n/a*	*n/a*	0.7187	0.9165	0.1206
J	shared	✓	—	1	θ_0	0.7273	0.6516	0.5902	0.2225	0.7356	0.9430	0.1015
J+	shared	✓	—	1	cal.	0.8182	0.6857	0.5902	0.2020	0.7402	0.9504	0.0908
J+N	shared	✓	✓	1	cal.	0.8987	0.7065	0.5820	0.1872	0.7430	0.9550	0.0837
J+NP	shared	✓	✓	$1/\lvert g(i)\rvert$	cal.	0.9722	0.7216	0.5738	0.1748	0.7448	0.9592	<u>0.0769</u>
J_θ	shared	✓	—	1	θ	0.8824	0.7246	0.6148	0.1910	0.7424	0.9563	0.0805
J_θNP	shared	✓	✓	$1/\lvert g(i)\rvert$	θ	<u>0.9733</u>	**0.7411**	0.5984	0.1776	<u>0.7455</u>	**0.9670**	**0.0644**
F−P	shared	✓	✓	1	θ_0	0.9452	0.7077	0.5656	0.1958	0.7419	0.9522	0.1216
F	shared	✓	✓	$1/\lvert g(i)\rvert$	θ_0	**0.9861**	<u>0.7320</u>	0.5820	**0.1689**	**0.7471**	<u>0.9621</u>	0.1042

5.2 Can the Agent Correctly Invoke Tools and Integrate Module Outputs?

Table 1 shows improvements in both tool invocation and answer production. At the step level, OneID-Agent achieves 79.78 tool agreement and 46.69 strict argument agreement, exceeding the strongest baselines by 4.59 and 19.24 points, respectively. Answer accuracy reaches 78.90, a gain of 19.65 points over the best baseline. Under end-to-end execution, tool-call Macro-F1 rises from 51.99 to 78.18 and E2E from 31.96 to 44.90, gains of 26.19 and 12.94 points. These results show that the fine-tuned agent invokes tools and produces the final answer more reliably, providing a foundation for shared-identity verification.

5.3 How Do Shared Identity and Its Components Affect Performance?

Table 2 ablates eleven configurations: H and S use hard and soft physical association, U adds per-claim semantic feedback, J shares one identity row per observation, J+, J+N, and J+NP add calibrated reliability, capacities/null, and budgeting, the J_θ variants refit the context model, and F is the full fixed-reliability stack with F−P its unbudgeted control.

The paired tests quantify the gains: the soft prior adds 1.79 points of Truth Acc. (95% CI [0.67, 2.91]) and shared identity 1.69 points ([0.75, 2.63]), while calibrated reliability, capacities/null, contextual reliability, and full-stack budgeting lower Pair Brier by 0.0205, 0.0148, 0.0110, and 0.0269. Identity calibration follows the same ordering, with ID ECE falling from 0.1482 (H) to 0.0644 (J_θNP) and dropping under budgeting from 0.0837 to 0.0769 and from 0.1216 to 0.1042. Shared identity converts recall into precision: H trades precision 0.3401 for the highest recall 0.6885, whereas F attains the highest precision 0.9861 and lowest Pair Brier 0.1689 at recall 0.5820 and J_θNP maximizes F1 (0.7411) at recall 0.5984. Finally, we adopt the precision-oriented F because a false merge propagates to every claim of an observation while a missed pair only withholds evidence.

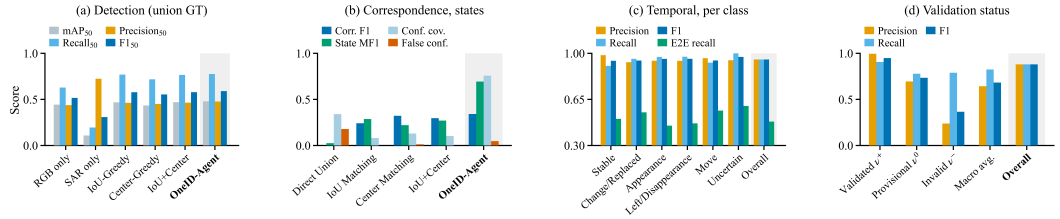

Figure 4: Cross-modal association summary. Panels report detection recall by modality, correspondence and confirmed coverage with false confirmations, and temporal performance separating conditional scores on recovered candidates from end-to-end recall.

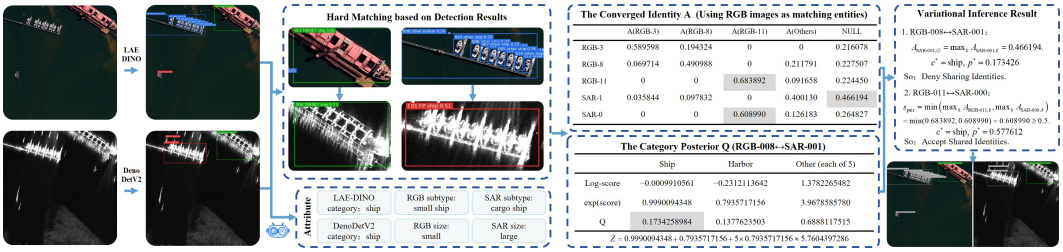

Figure 5: A calculation example of cross-interpretation based on variational inference.

Of nine pre-specified paired tests (Table 10; Appendix B.10), seven remain significant after Holm correction; semantic feedback and calibrated-stack budgeting are positive but not significant, and Appendix B.11 reports calibration and the independent diagnostics.

5.4 Remaining Problems and Limitations

Detection and candidate coverage limit grounded inference. Figure 4 shows that grounded inference is constrained by the upstream detection and candidate-generation pipeline. The SAR-only condition has markedly lower recall than RGB, and the temporal panel shows a clear gap between conditional performance on recovered candidates and lower end-to-end recall, indicating that accurate inference over available candidates cannot remove the effect of missed evidence. The SAR detector is a principal contributor to this ceiling: our SAR-only branch reaches mAP_{50} 0.1085 and recall 0.1955, far below the RGB-only branch (0.4436 and 0.6290), so many true targets never enter the candidate set. SAR-specific speckle, low target-to-clutter contrast, small object scale, and view-dependent scattering make these detections intrinsically harder than their optical counterparts. Because the shared-identity stage can only reweight candidates admitted by the physical graph, improving upstream SAR detection and cross-modal candidate generation would raise the achievable ceiling more than further inference-side refinement.

5.5 Qualitative Case Study

Figure 5 traces an optical–SAR case from detector outputs to an association decision. Hard matching based on detection geometry can pair an RGB harbor region with a false-positive SAR ship detection. Incorporating category, subtype, and size reports into the joint updates of identity beliefs A and attribute-truth beliefs Q allows the agent to reject this unsupported pairing while retaining a better-supported correspondence. The rejected pairing has shared-identity support of only 0.0978, below the displayed 0.5 decision threshold; the retained correspondence reaches 0.6090.

6 Conclusion

OneID-Agent jointly infers a shared identity for each observation and object–attribute–time truths. The full system achieves the highest pair precision and lowest Pair Brier among the evaluated configurations, together with strong identity accuracy. Its performance depends on candidate availability and evidence quality: missing candidates cannot be recovered through feedback, and misleading reports can redirect assignments. Future work will improve candidate acquisition, robustness to misleading reports, and calibration.

AI USE STATEMENT

Generative AI assisted literature analysis and manuscript writing and revision. GPT-5 mini was used only as a validation aid for RSTIBench: it performed structured consistency checks on human-authored tasks and trajectories and flagged potential issues during quality review, but it did not author tasks or trajectories. At evaluation time, answer correctness was graded semantically by `deepseek-v4-flash`. Two annotators authored and reviewed each sample, with a third adjudicating disagreements; final annotations were established through this human review process (Section 4). The authors designed the methods, ran the experiments, and take responsibility for the manuscript, claims, and results, including AI-assisted content.

ETHICS STATEMENT

The study uses attributed public or licensed imagery and no human-subject data. Object association across sensors and times can support remote-sensing analysis but also enable surveillance or other harmful uses. Incorrect association may attribute an observation or event to the wrong object; outputs require human verification before consequential use. Third-party imagery remains subject to its source-specific authorization and attribution requirements.

REPRODUCIBILITY STATEMENT

Section 3 specifies the inference objective and updates. Assumption 1, Algorithm 1, and Appendix C.7 give the calibration conditions, execution procedure, and derivations. Appendices B.1 and B.2 report the fixed hyperparameters of the shared-identity objective, the association thresholds, the calibration-reference construction, and the training settings; Appendices A.1 and B.10 describe data construction, annotation units, and evaluation protocols.

REFERENCES

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. arXiv:1412.5154, 2014. URL https://arxiv.org/abs/1412.5154.

Keumgang Cha, Junghoon Seo, and Taekyung Lee. A billion-scale foundation model for remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 19:16363–16378, 2026.

Alexander P. Dawid and Allan M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

Peilin Feng, Zhutao Lv, Junyan Ye, Xiaolei Wang, Xinjie Huo, Jinhua Yu, Wanghan Xu, Wenlong Zhang, Lei Bai, Conghui He, et al. Earth-agent: Unlocking the full landscape of earth observation with agents. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=dkIXAbWuxO.

Zhengpeng Feng, Sadiq Jaffer, Ira Shokar, Jovana Knezevic, Mark Elvers, Clement Atzberger, Robin Young, Aneesh Naik, Niall Robinson, Andrew Blake, et al. Tessera v2: Scaling pixel-wise earth foundation models. *arXiv preprint arXiv:2607.03949*, 2026b.

Songtao Guo, Xin Luna Dong, Divesh Srivastava, and Remi Zajac. Record linkage with uniqueness constraints and erroneous values. *Proceedings of the VLDB Endowment*, 3(1–2):417–428, 2010. doi: 10.14778/1920841.1920897.

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27672–27683, 2024.

Maram Hasan, Aman Verma, Savitra Roy, Hariseetharam Gunduboina, Daksh Jain, Muhammad Haris Khan, Subhasis Chaudhuri, and Biplab Banerjee. Geodisaster: Benchmarking orchestrated agents for operational disaster geo-intelligence. *arXiv preprint arXiv:2606.17246*, 2026.

Jiacheng Huang, Yao Zhao, Wei Hu, Zhen Ning, Qijin Chen, Xiaoxia Qiu, Chengfu Huo, and Weijun Ren. Trustworthy knowledge graph completion based on multi-sourced noisy data. In *Proceedings of the ACM Web Conference 2022*, pp. 956–965, 2022. doi: 10.1145/3485447.3511938.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Generalized variational inference: Three arguments for deriving new posteriors. arXiv:1904.02063, 2019. URL https://arxiv.org/abs/1904.02063.

Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020.

Yuan Li, Benjamin Rubinstein, and Trevor Cohn. Exploiting worker correlation for label aggregation in crowdsourcing. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3886–3895, 2019. URL https://proceedings.mlr.press/v97/li19i.html.

Xin Lin, Bo Zhang, Fan Wu, Chao Wang, Yali Yang, and Huiqin Chen. Sived: A sar image dataset for vehicle detection based on rotatable bounding box. *Remote Sensing*, 15(11):2825, 2023.

Jiaqi Liu, Ronghao Fu, Lang Sun, Haoran Liu, Xiao Yang, Weipeng Zhang, Xu Na, Zhuoran Duan, and Bo Yang. Skymoe: A vision-language foundation model for enhancing geospatial interpretation with mixture of experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7168–7178, 2026a.

Ruixun Liu, Bowen Fu, Jiayi Song, Kaiyu Li, Wanchen Li, Lanxuan Xue, Hui Qiao, Weizhan Zhang, Deyu Meng, and Xiangyong Cao. Zoomearth: Active perception for ultra-high-resolution geospatial vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34877–34888, 2026b.

Sai Ma, Zhuang Li, Sichao Li, Xinyue Xu, Ruibiao Zhu, Tony Boston, and John A Taylor. Eo-gym: A multimodal, interactive environment for earth observation agents. *arXiv preprint arXiv:2605.01250*, 2026.

Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning. In *European Conference on Computer Vision*, pp. 164–182. Springer, 2024.

Dat Tien Nguyen, Thao Nguyen, Fadillah Adamsyah Maani, Huy M Le, Muhammad Umer Sheikh, Numan Saeed, Muhammad Haris Khan, and Salman Khan. Terrabench: Can agents reason over heterogeneous earth-system data? *arXiv preprint arXiv:2606.13148*, 2026.

Kang Ni, Minrui Zou, Yuxuan Li, Xiang Li, Kehua Guo, Ming-Ming Cheng, and Yimian Dai. Denodet v2: Phase-amplitude cross denoising for sar object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8142–8150, 2026.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6281–6289, 2025.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(43): 1297–1322, 2010. URL https://jmlr.csail.mit.edu/papers/v11/raykar10a.html.

Akashah Shabbir, Muhammad Umer Sheikh, Muhammad Akhtar Munir, Hiyam Debary, Mustansar Fiaz, Muhammad Zaigham Zaheer, Paolo Fraccaro, Fahad Shahbaz Khan, Muhammad Haris Khan, Xiao Xiang Zhu, et al. Openearthagent: A unified framework for tool-augmented geospatial agents. In *Proceedings of the European Conference on Computer Vision*, 2026.

Huajie Shao, Dachun Sun, Shuochao Yao, Lu Su, Zhibo Wang, Dongxin Liu, Shengzhong Liu, Lance Kaplan, and Tarek Abdelzaher. Truth discovery with multi-modal data in social sensing. *IEEE Transactions on Computers*, 70(9):1325–1337, 2021. doi: 10.1109/TC.2020.3008561.

Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, et al. Earthdial: Turning multi-sensory earth observations to interactive dialogues. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14303–14313, 2025.

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Orsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique De Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, et al. Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–20, 2026.

Han Wang, Shengyang Li, Jian Yang, Yuxuan Liu, Yixuan Lv, and Zhuang Zhou. Cross-modal ship re-identification via optical and sar imagery: A novel dataset and method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7873–7883, 2025a.

Shuang Wang, He Zhang, Quan Z. Sheng, Xiaoping Li, Zhu Sun, Taotao Cai, Wei Emma Zhang, Jian Yang, and Qing Gao. A survey on truth discovery: Concepts, methods, applications, and opportunities. *IEEE Transactions on Big Data*, 11(2):314–332, 2025b. doi: 10.1109/TBDATA.2024.3423677.

Yi Wang, Zhitong Xiong, Chenying Liu, Adam J. Stewart, Thomas Dujardin, Nikolaos Ioannis Bountos, Angelos Zavras, Franziska Gerken, Ioannis Papoutsis, Laura Leal-Taixé, and Xiao Xiang Zhu. Towards a unified copernicus foundation model for earth vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9888–9899, October 2025c.

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, volume 22,

2009. URL https://papers.nips.cc/paper_files/paper/2009/file/f899139df5e1059396431415e770c6dd-Paper.pdf.

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3974–3983, 2018.

Aoran Xiao, Shihao Cheng, Yonghao Xu, Yexian Ren, Hongruixuan Chen, and Naoto Yokoya. Geommbench and geommagent: Toward expert-level multimodal intelligence in geoscience and remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34843–34853, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Wei Yang, Yaowen Fu, and Xiang Li. Joint target tracking and classification via RFS-based multiple model filtering. *Information Fusion*, 18:101–106, 2014. doi: 10.1016/j.inffus.2013.05.010.

Liang Yao, Shengxiang Xu, Fan Liu, Chuanyi Zhang, Bishun Yao, Rui Min, Yongjun Li, Chaoqian Ouyang, Shimin Di, and Min-Ling Zhang. Remoteagent: Bridging vague human intents and earth observation with rl-based agentic mllms. *arXiv preprint arXiv:2604.07765*, 2026a.

Ruichen Yao, Tejna Dasari, Gulshat Baispay, Aizhan Zaurbek, Yifan Liu, Yaokun Liu, Zelin Li, and Dong Wang. Beyond truth discovery: A two-stage framework to assess the severity of false claim during disasters. *arXiv preprint arXiv:2608.20983*, 2026b. URL https://arxiv.org/abs/2608.20983.

Tianwen Zhang, Xiaoling Zhang, Jianwei Li, Xiaowo Xu, Baoyou Wang, Xu Zhan, Yanqin Xu, Xiao Ke, Tianjiao Zeng, Hao Su, et al. Sar ship detection dataset (ssdd): Official release and comprehensive data analysis. *Remote Sensing*, 13(18):3690, 2021.

Sijie Zhao, Feng Liu, Xueliang Zhang, Hao Chen, Xinyu Gu, Zhe Jiang, Fenghua Ling, Ben Fei, Wenlong Zhang, Junjue Wang, et al. Openearth-agent: From tool calling to tool creation for open-environment earth observation. *arXiv preprint arXiv:2603.22148*, 2026a.

Yujian Zhao, Hankun Liu, and Guanglin Niu. Mos: Mitigating optical-sar modality gap for cross-modal ship re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30335–30345, 2026b.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, et al. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 29733–29735, 2025.