Filling Before Advancing: Capability-Gap-Driven Post-Training for Scenario-Specialized Remote Sensing MLLMs

Yuheng Zong,[1] Minghua Wang,[1,*] Xin Zhao,[1] Zhi-Hui Zhan,[1]
Antonio Plaza,[2] Jón Atli Benediktsson[3]

[1] The Institute of Robotics and Automatic Information System (IRAIS), the Tianjin Key Laboratory of Intelligent Robotics (tjKLIR), Nankai University, Tianjin 300071, China
[2] Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, Cáceres, Spain
[3] Faculty of Electrical and Computer Engineering, University of Iceland, Reykjavík, Iceland
[*]Corresponding author: wangminghua@nankai.edu.cn

Abstract

Remote sensing multimodal large language models (RS-MLLMs) have improved general aerial-image understanding. However, Earth observation applications require fine-grained scenario specialization, constrained by scarce high-quality scenario data and incomplete capability coverage. We formulate this adaptation as a capability-gap-driven post-training problem and propose *filling before advancing* (FBA). Rather than relying on single-stage supervised fine-tuning (SFT) over target-domain samples, FBA first fills prerequisite capability gaps before advancing toward scenario specialization. We instantiate FBA for coastal harbor understanding, a representative multi-source scenario, by constructing CPRS (Coastal-Port Remote Sensing), a three-layer supervision dataset coupled with three ordered stages: (1) RS semantic anchoring for overhead-view visual-language alignment; (2) domain-bridge convergence for shared RS priors across target and bridging scenarios under different modalities; and (3) evidence-grounded scenario tuning for downstream performance. We construct HarborEval, an eight-track diagnostic benchmark covering perception, spatial understanding, robustness, and generation. Under comparable training budgets, HarborEval increases from 57.95 with Direct-SFT to 70.29 with FBA on LLaVA-v1.5, and from 81.09 to 83.37 on Qwen3-VL. FBA also outperforms Collapsed-SFT and leads on harbor-related VRSBench/RSVQA subsets and OpenEval. Stage-wise and role-replacement analyses validate progressive gap filling and stage-specific roles. ***Public examples and release updates for CPRS, HarborEval, code, and trained weights are available at https://github.com/Z0ngL1ng/filling-before-advancing.***

Introduction

Multimodal large language models (MLLMs) have expanded visual understanding from closed-set recognition to open-world perception, logical reasoning, and instruction following (Radford et al. 2021; Alayrac et al. 2022; Li et al. 2023a; Dai et al. 2023; Liu et al. 2023; Bai et al. 2025b). This paradigm shift has also penetrated remote sensing (RS), spawning RS-MLLMs capable of aerial image captioning, visual question answering, region-level instruction following, and grounded geo-spatial interpretation (Hu et al. 2025; Kuckreja et al. 2024; Zhang et al. 2024a; Muhtar et al. 2024; Pang et al. 2025; Soni et al. 2025; Zhan, Xiong, and Yuan 2025). Instead of producing category labels or boxes, RS-MLLMs connect visual evidence with natural-language queries, steering RS toward general-purpose visual-language intelligence.

Real-world RS applications rely on the interpretation and understanding of targeted task scenarios, rather than broad scene categories (Li et al. 2024). Taking coastal harbor monitoring as a representative case, RS-MLLMs are expected to surpass simple and generic descriptions, such as water, roads, buildings, and ships. As illustrated in Fig. 1 (a), traditional RS-MLLMs just have the ability to show what exists in the scene. Nevertheless, scenario-specialized RS-MLLMs are required to generate a structured and evidence-grounded report of operationally meaningful harbor attributes, including functional zones, vessel scale, and spatial layout. This discrepancy discloses a gap between universal capability and scenario-specialized usability of RS-MLLMs.

A straightforward solution is to collect harbor instructions and directly fine-tune a general RS-MLLM, yet this is often insufficient due to two challenges, shown in Fig. 1 (b). # Challenge 1 (scarce high-quality scenario data): General MLLMs typically rely on massive natural image–text corpora for training (Alayrac et al. 2022). In contrast to readily accessible natural images, high-quality, expert-annotated supervision data from the target RS scenario remain scarce, because long-term scene observation, multi-sensor acquisition, and rigorous expert annotation are time-consuming and labor-intensive (Wang et al. 2024; Kuckreja et al. 2024). # Challenge 2 (difficult capability coverage): When migrating from general MLLMs to RS-MLLMs, gaps in alignment, modality, and task arise for specific RS scenarios, while single-stage training paradigms, typified by direct supervised fine-tuning (SFT), lack the capability to bridge them. These gaps result from the shift in viewing angles, the increment of different RS sensors, and the demand for task professionalization from natural to RS scenes (Liu et al. 2024a; Zhang et al. 2024a; Muhtar et al. 2024; Soni et al. 2025).

This motivates a novel *filling before advancing* (FBA) perspective of scenario-specialized RS-MLLM post-training, displayed in Fig. 1 (c). Here, *filling* refers to sequentially closing multi-level capability gaps, which encompasses overhead-view visual semantics, sensor-aware observability,

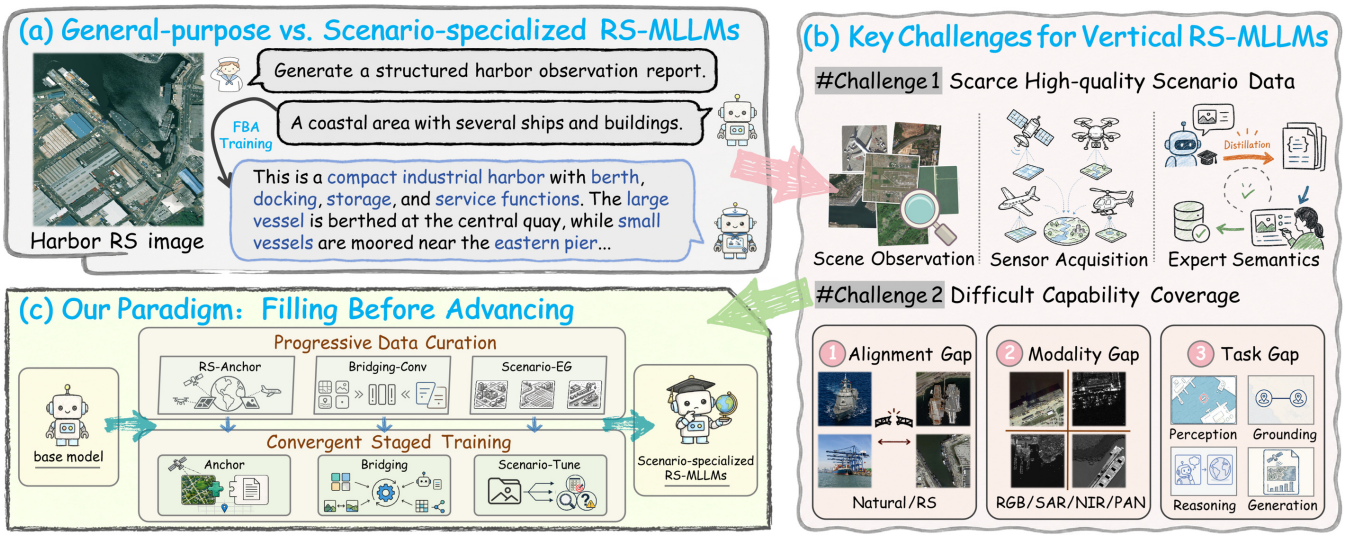

Figure 1: Motivation and paradigm of the proposed FBA.

target-bridging context, evidence grounding, and calibrated rejection. After filling these gaps, *advancing* entails the continuous evolution of model capabilities, shifting from generalized adaptation to the final specialized scenario expertise. The point is less to add data than to assign scarce and costly supervision to separable roles.

In this paper, we instantiate the proposed paradigm in the context of coastal harbor understanding. As illustrated in Fig. 2, we establish CPRS, a three-layer supervision dataset comprising RS-Anchor, Bridge-Conv, and Scenario-EG in alignment with the ordered adaptation route shown in Fig. 3. RS semantic anchoring establishes broad overhead-view visual-language alignment as the basis for subsequent specialization. Domain-bridge convergence learns RS priors shared across target and bridging scenarios under multiple modalities. Evidence-grounded scenario tuning focuses adaptation on harbor-specific downstream tasks and strengthens evidence-grounded responses under negative or ambiguous conditions.

An eight-track diagnostic benchmark is constructed and denoted as HarborEval, covering RGB, SAR, PAN, and NIR imagery from both harbor and non-harbor scenes. Under comparable training budgets, FBA consistently outperforms Direct-SFT and Collapsed-SFT with both LLaVA-v1.5 and Qwen3-VL backbones (Liu et al. 2024b; Bai et al. 2025a), as shown in Table 1. Further comparisons with existing RS-MLLMs on HarborEval, the harbor-related subsets of VRSBench and RSVQA (Li, Ding, and Elhoseiny 2024; Lobry et al. 2020), and OpenEval support the effectiveness of the proposed route (Table 2). Stage-wise analyses and role-replacement controls further verify the intended capability role of each stage and demonstrate progressive capability-gap filling along the ordered route (Tables 3 and 4).

The main contributions of this work are as follows:

- We formulate scenario-specialized RS-MLLM adaptation under scarce high-quality scenario supervision as a capability-gap-driven post-training problem.
- We propose FBA, a post-training route that progressively fills capability gaps through three ordered stages, including RS semantic anchoring, domain-bridge convergence, and evidence-grounded scenario tuning, respectively supported by the CPRS supervision layers RS-Anchor, Bridge-Conv, and Scenario-EG.
- We instantiate FBA for coastal harbor understanding and construct HarborEval, an eight-track diagnostic benchmark. Extensive experiments show that FBA consistently outperforms alternative post-training routes across multiple backbones, achieves competitive performance against existing RS-MLLMs, and progressively closes the targeted capability gaps across the ordered adaptation stages.

Related Work

Remote Sensing Vision-Language Foundations

RS vision-language research has narrowed the semantic gap between natural-image pretraining and RS imagery through captioning (Lu et al. 2018; Cheng et al. 2022), VQA (Lobry et al. 2020), cross-modal retrieval (Yuan et al. 2022), and large-scale geospatial image-text alignment (Wang et al. 2024; Zhang et al. 2024b; Liu et al. 2024a). These works provide broad RS semantic grounding for classification, localization, captioning, and open-ended understanding (Kuckreja et al. 2024; Li, Ding, and Elhoseiny 2024; Hu et al. 2025), and thus motivate the RS Semantic Anchoring stage to be considered in our route. However, they mainly address general overhead semantic alignment and rarely specify how to deal with scarce scenario-level supervision when target behaviors require modality-aware evidence, spatial grounding, uncertainty handling, and rejection.

RS-MLLMs and Scenario-Level Understanding

Recent RS-MLLMs adapt general multimodal foundations to the RS field through RS alignment, instruction tuning,

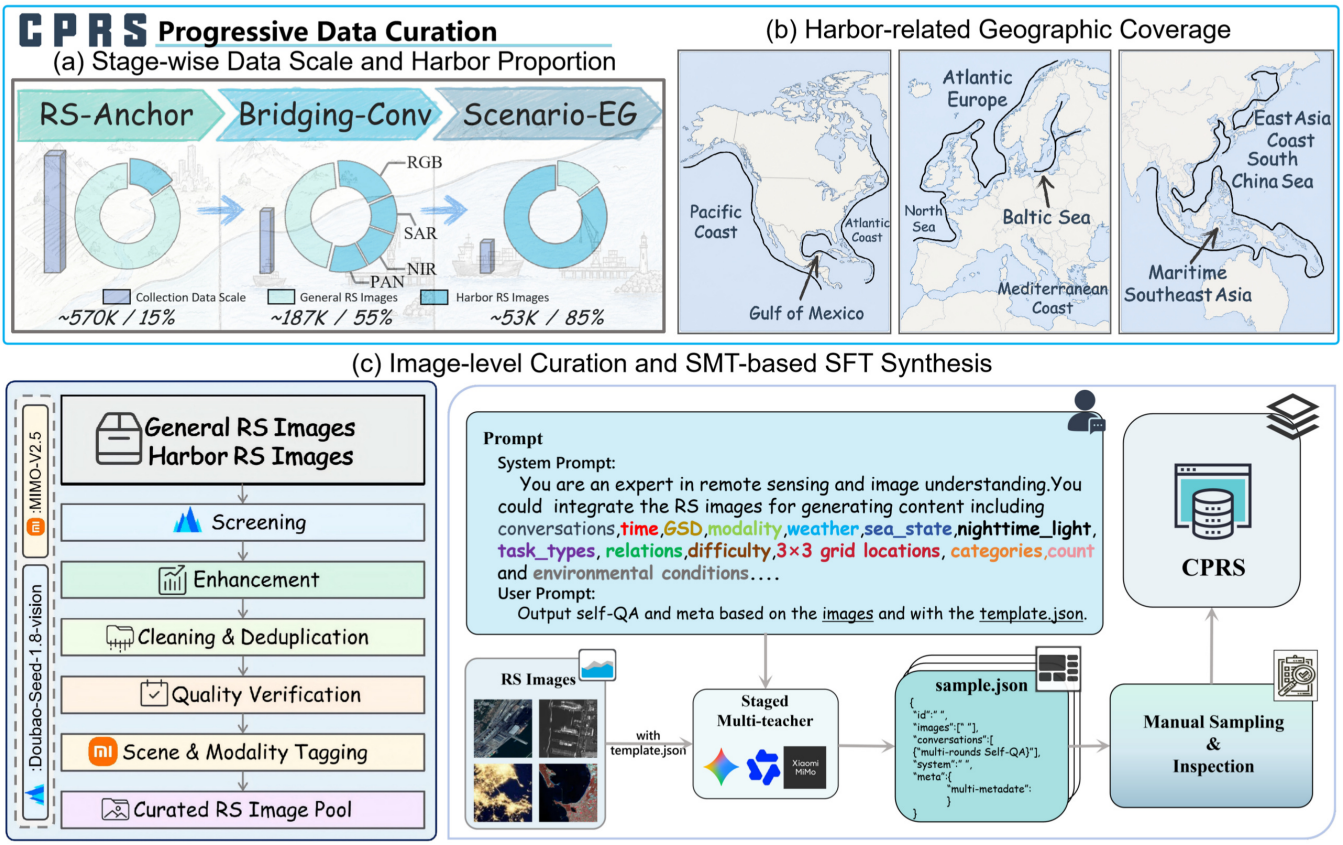

Figure 2: Progressive data curation of the CPRS dataset for the staged route.

grounded dialogue, and multi-source Earth-observation inputs (Kuckreja et al. 2024; Zhang et al. 2024a; Soni et al. 2025; Zhan, Xiong, and Yuan 2025). Such methodological advances boost general RS instruction following, captioning, VQA, region-level interpretation, and multi-source understanding (Hu et al. 2025; Luo et al. 2024; Li et al. 2025). Nevertheless, scenario-specific applications demand far more than broad RS competence. Taking coastal harbor analysis as a representative vertical task, reliable interpretation depends on functional zones, vessel layout, grid localization, non-harbor rejection, and evidence-grounded reporting across RGB, SAR, PAN, and NIR observations. Direct target-domain SFT can overburden limited high-quality scenario data, because the same samples must support both required capability adaptation and final scenario behavior.

Post-Training Routes for Scenario Specialization

Our work is also related to instruction tuning (Wei et al. 2022; Liu et al. 2023), curriculum learning (Bengio et al. 2009), and continual domain adaptation (Gururangan et al. 2020; Kirkpatrick et al. 2017; Rolnick et al. 2019). These paradigms show that training order and supervision design have a strong effect on adaptation, but their data partitioning strategies are primarily guided by example difficulty (Bengio et al. 2009) (Wang et al. 2023b), instruction diversity (Wang et al. 2023b), domain continuation (Gururangan et al. 2020), or task arrival (Kirkpatrick et al. 2017; Rolnick et al. 2019). In contrast, we organize post-training by the capabilities needed for scenario specialization. Each data layer is assigned a distinct role: broad RS semantic anchoring, multi-source domain-bridge convergence, and final evidence-grounded scenario tuning. This novel formulation recasts scenario specialization from target-only SFT as a capability-gap-filling process.

Methods

Progressive Data Curation

To demonstrate the practical realization of the proposed *FBA* paradigm through coastal harbor understanding, we first construct CPRS, a three-layer supervision dataset with broad geographic coverage across coastal and port regions, as illustrated in Fig. 2. The three supervision layers, denoted by D_1, D_2, and D_3, correspond to RS-Anchor, Bridge-Conv, and Scenario-EG, respectively. They contain 569,853 RGB RS image-caption pairs, 187,296 multi-source bridge-domain SFT samples, and 53,000 harbor-specialized instruction samples. The proportion of harbor-related RS images increases progressively from 15% in D_1 to 55% in D_2 and 85% in D_3. The construction of each supervision layer is detailed below.

To align overhead-view visual patterns with RS semantics, RS-Anchor is constructed from existing RGB RS captioning (Lu et al. 2018; Cheng et al. 2022; Ge et al. 2025),

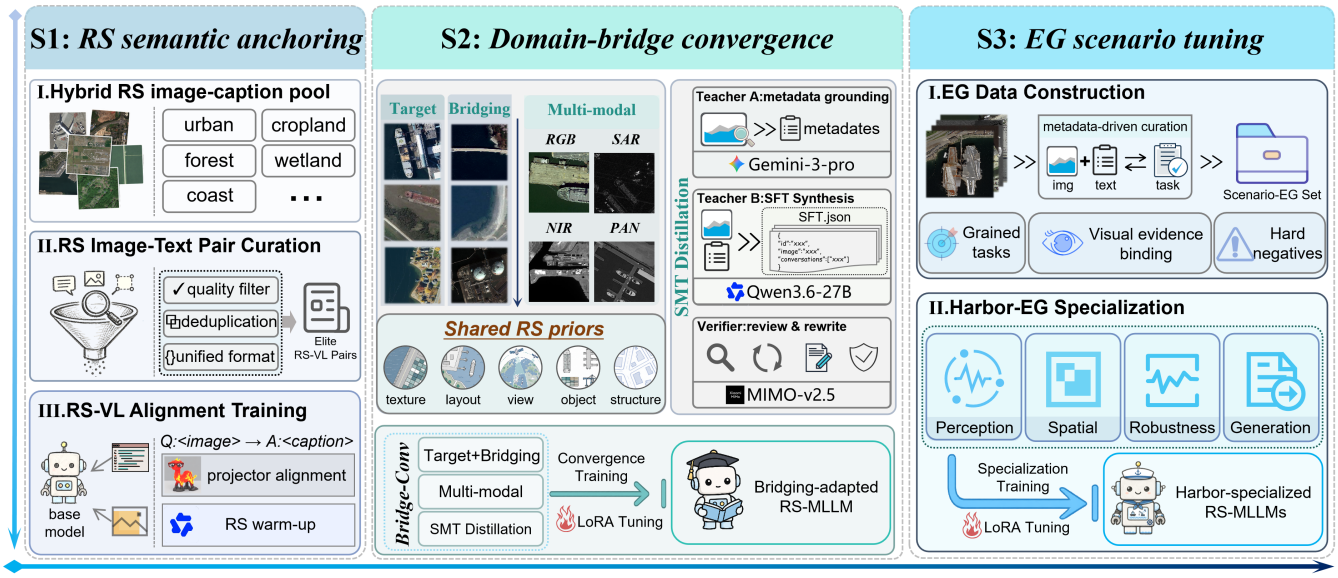

Figure 3: Convergent staged post-training route: S1 RS semantic anchoring, S2 domain-bridge convergence, and S3 evidence-grounded scenario tuning.

retrieval (Yuan et al. 2022), and geospatial image-text resources (Wang et al. 2024; Yuan et al. 2025; Soni et al. 2025). These data are screened, normalized, deduplicated, and quality-checked to form a large-scale, quality-controlled image-text supervision pool for initial RS visual-language alignment.

Following the initial RS visual-language alignment established by RS-Anchor, Bridge-Conv extends supervision beyond general RGB imagery to target harbor scenes together with coastal-port bridging scenes across RGB, SAR, PAN, and NIR. As illustrated in Fig. 2(c), we employ staged multi-teacher (SMT) distillation to construct D_2 as a high-quality instruction-tuning set from this multi-source image pool:

$$\begin{aligned} e_i &= T_{\text{meta}}(x_i, m_i, s_i), \\ \tilde{u}_i &= T_{\text{syn}}(x_i, e_i), \\ u_i &= T_{\text{ver}}(x_i, e_i, \tilde{u}_i), \\ D_2 &= \text{SMT}(\mathcal{X}_2) = \{(x_i, u_i) \mid x_i \in \mathcal{X}_2,\ u_i \neq \emptyset\}, \end{aligned} \tag{1}$$

where $\mathcal{X}_2$ represents the curated image pool that contains both target and bridging scenes. For each image x_i, m_i and s_i denote its modality and scene type, while e_i, $\tilde{u}_i$, and u_i are the normalized evidence, synthesized instruction response, and verified response, respectively. Here, SMT refers to the sequential application of three teachers: T_{meta} first normalizes the sensor, source, and scene evidence into e_i, T_{syn} then generates $\tilde{u}_i$ conditioned on the image and normalized evidence, and T_{ver} accepts, rewrites, or rejects the response according to its visual support and modality consistency.

Scenario-EG constitutes the final supervision layer D_3, comprising 53,000 harbor-specialized instruction samples that further concentrate supervision on evidence-grounded behavior in the target scenario. It covers diverse harbor-specific downstream behaviors and incorporates negative and uncertainty-aware supervision under stricter evidence-grounding criteria, thereby reducing overfitting to narrow scene patterns and improving response calibration when visual evidence is insufficient or ambiguous.

Overall, CPRS organizes supervision along a progressive trajectory from broad RGB RS semantics, through multi-source bridge-domain instruction tuning, to evidence-grounded harbor specialization, thereby supporting the subsequent convergent staged training. The specific SMT teacher configurations and prompt templates, together with the refinement and audit procedures for Scenario-EG, are detailed in Supplementary Sections S1–S2 and Tables S1–S2.

Convergent Staged Training

Given the three CPRS supervision layers D_1, D_2, and D_3, *FBA* organizes model adaptation into three ordered training stages, denoted by S_1, S_2, and S_3, as demonstrated in Fig. 3. These stages correspond to RS semantic anchoring, domain-bridge convergence, and evidence-grounded scenario tuning, respectively. Let $\mathcal{M}_0$ denote the initial general-purpose MLLM, with $\mathcal{M}_k$ representing the RS-MLLM model obtained after stage S_k. The staged adaptation is formulated as:

$$\mathcal{M}_k = \text{Adapt}\left(\mathcal{M}_{k-1}, D_k\right), \qquad k \in \{1, 2, 3\}, \tag{2}$$

where each stage initializes from the model produced by the preceding stage and leverages its corresponding supervision layer.

At stage S_1, RS semantic anchoring adapts the initial general-purpose MLLM $\mathcal{M}_0$ using RS-Anchor supervision D_1. This stage establishes the overhead-view visual-language correspondence required for RS interpretation, producing $\mathcal{M}_1$ with a broad RS semantic basis for the subsequent stages.

With the broad RS semantic basis established at S_1, stage S_2 performs domain-bridge convergence by further adapting

$\mathcal{M}_1$ using Bridge-Conv supervision D_2. As displayed in the S_2 panel of Fig. 3, this stage concurrently exposes the model to target harbor scenes and coastal and port-related bridging scenes across RGB, SAR, PAN, and NIR. Their joint supervision encourages the consolidation of RS priors shared across target and bridging scenes under multiple modalities, which we conceptually formulate as follows:

$$\begin{aligned} D_{2,\mathrm{t}}^{m} &= \{(x_i, u_i) \in D_2 \mid m_i = m,\ s_i = \text{target}\}, \\ D_{2,\mathrm{b}}^{m} &= \{(x_i, u_i) \in D_2 \mid m_i = m,\ s_i = \text{bridge}\}, \\ \mathcal{P}_{\text{shared}} &= \text{Shared}_{\text{RS}}\left(\left\{D_{2,\mathrm{t}}^{m}, D_{2,\mathrm{b}}^{m}\right\}_{m \in \mathcal{M}}\right), \end{aligned} \tag{3}$$

where $D_{2,\mathrm{t}}^{m}$ and $D_{2,\mathrm{b}}^{m}$ denote the target-scene and bridging-scene subsets of D_2 under modality m, respectively. $\mathcal{P}_{\text{shared}}$ represents the RS priors shared across the two scene groups and modalities, including texture, layout, viewpoint, object patterns, and scene structure. By leveraging these shared priors, S_2 strengthens the learning of harbor-relevant visual-language representations across multiple modalities and establishes the prerequisite capabilities for subsequent scenario-specialized tuning at S_3.

Finally, stage S_3 applies evidence-grounded scenario tuning to $\mathcal{M}_2$ using Scenario-EG supervision D_3. As shown in the S_3 panel of Fig. 3, this stage focuses training on harbor-specific downstream tasks across perception, spatial understanding, robustness, and generation, while strengthening the use of visual evidence and the handling of hard negative cases. Building on the RS semantic foundation established at S_1 and the multi-source shared priors learned at S_2, S_3 further aligns model responses with visual evidence from the target scenario, thereby improving reliability when such evidence is absent or ambiguous. The resulting $\mathcal{M}_3$ is the final harbor-specialized RS-MLLM.

Overall, the convergent staged training route progressively narrows the focus of supervision from broad RS semantic anchoring, through multi-source domain-bridge convergence, to evidence-grounded harbor specialization, with each subsequent stage building on the capabilities established in the previous stage. By assigning a distinct capability role to each stage, the ordered route yields the final harbor-specialized model $\mathcal{M}_3$.

Harbor-Scenario Evaluation Protocol

The evaluation of the proposed paradigm is conducted at two levels using various metrics. RS-VL Val. and MultiSource Val. assess intermediate capabilities acquired along the staged route, while HarborEval, together with the harbor-related subsets of VRSBench and RSVQA (Li, Ding, and Elhoseiny 2024; Lobry et al. 2020), and OpenEval evaluate final harbor-scenario performance. All scores are reported on a scale of 0–100 and are rounded to two decimal places. Further details on benchmark construction and split statistics, together with the per-track scoring protocols, including answer matching, caption description and expert scoring, and ambiguity resolution, are provided in the Supplementary Sections S3, S5, and S6.

Intermediate Capability Validation The intermediate capabilities associated with stages S_1 and S_2 are verified by RS-VL Val. and MultiSource Val., respectively. RS-VL Val. assesses overhead-view visual-language alignment and broad RS semantic grounding, whereas MultiSource Val. evaluates multi-source understanding across RGB, SAR, PAN, and NIR imagery. The above-mentioned capabilities are developed along the staged route prior to the final harbor-scenario evaluation.

HarborEval for Scenario Diagnosis HarborEval is an eight-track diagnostic benchmark spanning RGB, SAR, PAN, and NIR imagery from harbor and non-harbor scenes, with source records disjoint from CPRS training supervision. It assesses perception, spatial understanding, robustness, and generation through object recognition, functional-zone understanding, modality recognition, spatial relations, grid localization, negative-case handling, non-harbor rejection, and evidence-grounded reporting.

The overall HarborEval score H_{eval} is computed as the unweighted average of all eight diagnostic tracks after their normalization to a common 0–100 scale:

$$H_{\text{eval}} = \frac{1}{N_{\mathcal{T}}} \sum_{t \in \mathcal{T}} S_t, \quad \mathcal{T} = \{1, \ldots, N_{\mathcal{T}}\}. \tag{4}$$

where $\mathcal{T}$ denotes the set of diagnostic tracks, $N_{\mathcal{T}} = 8$, and S_t is the normalized score for track t. The seven structured tracks are evaluated using their corresponding task-specific metrics, whereas the open-ended reporting track is scored by a fixed image-grounded judge (Zheng et al. 2023) based on Doubao-Seed-1.8-Vision (ByteDance Seed Team 2025).

Public Benchmarks and Expert Evaluation To complement HarborEval with external validation, we derive harbor-related subsets from the public test splits of VRSBench and RSVQA. These subsets include both harbor and non-harbor samples relevant to the target scenario and evaluate harbor-related recognition and reasoning on public benchmark data. OpenEval further validates open-ended evidence-grounded responses and negative-case handling through manual scoring by domain experts.

Experiments and Analysis

Experimental Setup

We evaluate *FBA* on LLaVA-v1.5 and Qwen3-VL (Liu et al. 2023, 2024b; Bai et al. 2025a). Direct-SFT uses the harbor-specialization supervision in D_3, whereas Collapsed-SFT trains on $C_{2,3} = D_2 \cup D_3$ in a single stage, either directly or after S_1; *FBA* follows the ordered route $S_1 \rightarrow S_2 \rightarrow S_3$. For LLaVA-v1.5, the official model is reported as a backbone reference, while I_0 and G_0 denote the pre-alignment initialization and the natural-image–text aligned checkpoint, respectively. For Qwen3-VL, B_0 denotes the officially released checkpoint used for subsequent post-training.

Within each backbone family, all adapted routes share the same target-scenario supervision set D_3 but differ in the inclusion and ordering of prerequisite D_1 and D_2. LoRA configurations (Hu et al. 2022), optimization, and inference settings are otherwise aligned. Details appear in Supplementary Sections S4–S6.

Backbone	Training Route	Main	Perception			Spatial		Robustness		Generation
		Overall	Object	Zone	Modality	Relation	Grid	Negative	Reject	Report
LLaVA-v1.5 full-pipeline validation from multimodal initialization										
LLaVA-v1.5	Official Model (ref.)	46.22	71.42	51.67	48.54	51.12	32.13	47.15	40.24	27.47
	Direct-SFT	57.95	**75.07**	**66.67**	50.88	60.67	**48.37**	52.03	37.28	72.60
	$G_0 \rightarrow C_{2,3}$	55.34	42.41	52.22	61.40	68.09	45.86	52.85	58.69	61.20
	$I_0 \rightarrow S_1 \rightarrow C_{2,3}$	55.74	73.85	51.11	57.89	59.55	31.74	34.96	67.46	69.34
	FBA: $I_0 \rightarrow S_1 \rightarrow S_2 \rightarrow S_3$	**70.29**	73.47	67.22	**80.12**	**69.10**	43.82	**69.11**	**85.80**	**73.70**
Native MLLM adaptation from official visual-instruction-tuned base										
Qwen3-VL	B_0	70.37	80.06	74.44	76.02	40.91	65.81	65.85	94.67	65.24
	Direct-SFT	81.09	87.99	78.33	81.29	81.46	66.06	78.05	95.27	80.23
	$B_0 \rightarrow C_{2,3}$	72.84	88.06	39.44	58.48	82.58	65.70	74.80	**98.82**	69.25
	$B_0 \rightarrow S_1 \rightarrow C_{2,3}$	79.36	82.06	80.56	81.97	**84.27**	67.56	65.85	**98.82**	73.76
	FBA: $B_0 \rightarrow S_1 \rightarrow S_2 \rightarrow S_3$	**83.37**	**92.42**	**81.11**	**82.46**	83.15	**68.61**	**79.67**	98.22	**81.32**

Table 1: HarborEval diagnostic results for scenario-specialized RS-MLLMs under different training routes.

Model	Params	Data Scale	HarborEval	VRSBench	RSVQA	OpenEval
Existing RS-MLLMs						
GeoChat [CVPR'24]	7B	318K*	47.49	53.44	51.46	21.78
SkyEyeGPT [ISPRS'25]	7B	968K	28.28	36.72	36.72	12.33
LHRS-Bot-Nova [ISPRS'25]	7B	2.02M	39.73	28.40	31.15	22.96
SkySenseGPT [ISPRS'26]	7B	3.00M	47.24	42.99	42.98	35.78
Models trained with the proposed paradigm						
FBA (LLaVA-v1.5)	7B	**810K**	70.29	57.62	53.96	61.47
FBA (Qwen3-VL)	8B	**810K**	**83.37**	**67.77**	**63.00**	**76.67**

Table 2: Scenario-level comparison with existing RS-MLLMs on HarborEval, VRSBench, RSVQA, and OpenEval.
*GeoChat is initialized from LLaVA-1.5 and its reported data scale excludes the inherited ∼1.22M general-purpose samples for consistency.

Backbone	Route	RS-VL	MS	HE	VRS	RQA	OE
LLaVA-v1.5	I_0	N/A	N/A	N/A	N/A	N/A	N/A
	$+S_1$	69.71	52.97	29.44	49.54	37.00	42.23
	$+S_2$	87.53	**73.55**	35.61	54.88	47.55	58.01
	$+S_3$	**89.16**	68.04	**70.29**	**57.62**	**53.96**	**61.47**
Qwen3-VL	B_0	90.40	69.60	70.37	57.02	50.63	65.48
	$+S_1$	**95.29**	70.32	77.26	51.14	54.66	60.03
	$+S_2$	93.37	**79.77**	71.80	66.58	57.30	60.15
	$+S_3$	92.22	76.54	**83.37**	**67.77**	**63.00**	**76.67**

Table 3: Stage-wise capability trajectory. $+S_k$ denotes cumulative training. RS-VL and MS are intermediate diagnostics. HE, VRS, RQA, and OE denote HarborEval, VRSBench, RSVQA, and OpenEval and are reused in Table 4.

Main Results and Model Comparison

Training Route Comparison To investigate the performance gains arising from the ordered supervisions, we compare *FBA* with direct and collapsed training routes under matched settings across the two backbone families. Table 1 compares different training routes on HarborEval. For LLaVA-v1.5, *FBA* achieves an overall score of 70.29, substantially outperforming Direct-SFT (57.95), $G_0 \rightarrow C_{2,3}$ (55.34), and $I_0 \rightarrow S_1 \rightarrow C_{2,3}$ (55.74). For Qwen3-VL, *FBA* reaches 83.37, exceeding Direct-SFT (81.09), B_0 (70.37), $B_0 \rightarrow C_{2,3}$ (72.84), and $B_0 \rightarrow S_1 \rightarrow C_{2,3}$ (79.36). These results show that the ordered post-training route is more effective than either direct target-scenario tuning or collapsing bridge-domain and scenario supervision into a single stage. The gains of *FBA* are distributed across multiple capability dimensions, with particularly substantial improvements in zone understanding, modality recognition, negative-case handling, and report generation. Although several alternative routes remain competitive on individual tracks, *FBA* delivers the strongest overall, most balanced performance across both backbone families.

Comparison with Existing RS-MLLMs To assess whether the advantages of the proposed route extend beyond the controlled comparisons within each backbone family, we compare the resulting models with representative RS-MLLMs (Kuckreja et al. 2024; Zhan, Xiong, and Yuan 2025; Li et al. 2025; Luo et al. 2024) on HarborEval, the harbor-

Ctrl.	Variant	RS-VL	MS	HE	VRS	RQA	OE
D_1	Gen-IT D_1^{gen}	76.42	64.49	60.36	56.77	51.01	55.30
	RS-Anchor D_1^{anc}	**89.16**	**68.04**	**70.29**	**57.62**	**53.96**	**58.01**
D_2	Non-bridging D_2^{ori}	84.36	66.90	57.11	54.77	48.13	52.06
	Bridge-Conv D_2^{brg}	**89.16**	**68.04**	**70.29**	**57.62**	**53.96**	**58.01**
D_3	Non-EG D_3^{ori}	88.97	67.78	50.12	52.98	49.36	44.11
	Scenario-EG D_3^{eg}	**89.16**	**68.04**	**70.29**	**57.62**	**53.96**	**58.01**

Table 4: LLaVA-side role-replacement controls for testing capability-role assignment. Gen-IT denotes generic image-text data.

related subsets of VRSBench and RSVQA, and OpenEval. As shown in Table 2, both *FBA* variants outperform all compared RS-MLLMs across the four evaluation settings, with the Qwen3-VL variant achieving the strongest overall performance. Notably, these results are obtained using approximately 810K curated samples, fewer than those used by several compared models. The comparison further supports the effectiveness of the proposed paradigm for harbor-scenario specialization. Complementary analyses of hard-negative and visually ambiguous cases across RGB, SAR, PAN, and NIR imagery are provided in Supplementary Sections S7–S8 and Figures S2–S3.

Stage-wise and Role Analysis

To verify the intended capabilities developed by the ordered route across successive stages, we analyze the cumulative models obtained after S_1, S_2, and S_3, respectively. Table 3 reports their performance on the two intermediate diagnostics, RS-VL Val. and MultiSource Val., together with the four final harbor-scenario evaluations. The quantitative results show a clear stage-specific progression in capability. S_1 establishes a strong RS visual-language foundation, while S_2 yields the largest improvement in multi-source understanding and further strengthens performance on the public harbor benchmarks. With these prerequisite capabilities established, S_3 substantially enhances final harbor-scenario performance and achieves the best scores on HarborEval, VRSBench, RSVQA, and OpenEval for both backbone families. The intermediate diagnostics also remain robust as supervision becomes progressively focused on target-scenario behavior.

We further perform role-replacement controls on the LLaVA-v1.5 route to explore the contribution of each supervision layer. As shown in Table 4, RS-Anchor consistently outperforms generic image-text supervision across both intermediate and downstream evaluations, demonstrating the importance of RS semantic anchoring. Bridge-Conv improves all six metrics over non-bridging supervision, supporting the use of target-related bridging scenes for multi-source harbor adaptation. Scenario-EG delivers the largest gains on HarborEval and OpenEval while preserving the previously established intermediate capabilities, confirming its role in final evidence-grounded scenario specialization.

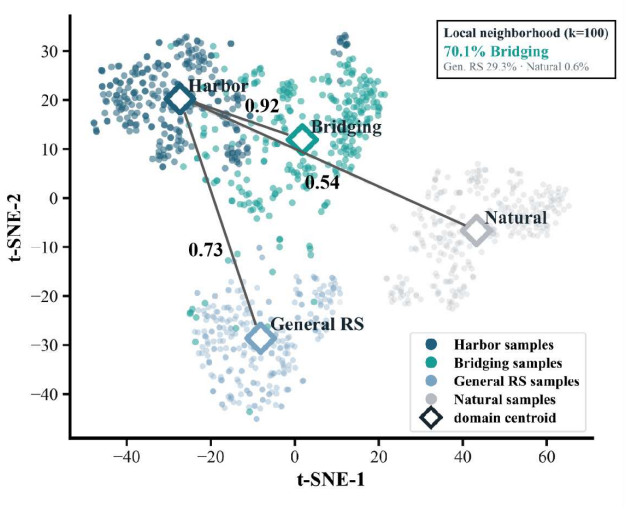

Figure 4: The representation proximity between the Harbor and the other domains. Cosine similarities and kNN statistics are computed in the original joint embedding space.

Together, these results validate the distinct and successive capability roles of the three supervision layers.

Bridging Transfer Analysis

We assess the suitability of Bridge-Conv scenes as an intermediate supervision domain by measuring their representational proximity to harbor imagery. Harbor, bridging, general RS, and natural-image samples are mapped to a shared visual-language embedding space. Figure 4 shows their two-dimensional t-SNE distributions (van der Maaten and Hinton 2008), while centroid cosine similarities and nearest-neighbor statistics are computed in the original embedding space.

In Fig. 4, the bridging centroid achieves a cosine similarity of 0.92 to the harbor centroid, exceeding that of general RS (0.73) and natural-image samples (0.54). A similar pattern appears in the local neighborhoods. Among the 100 nearest non-harbor neighbors retrieved for each harbor query, bridging samples account for 70.1%, compared with 29.3% general RS samples and 0.6% natural-image samples. These observations indicate that the design of Bridge-Conv and stage S_2 preserve stronger harbor-related visual-language priors than the broader comparison domains.

Conclusion

We present a capability-gap-driven staged route named FBA for scenario-specialized RS-MLLM adaptation. FBA first fills RS semantic anchoring and multi-source bridging before final evidence-grounded behaviors. The harbor instantiation serves as a practical example for scenario-specialized RS-MLLMs under limited high-quality supervision. Extensive experiments demonstrate that FBA achieves superior performance across comprehensive harbor diagnostics benchmarks over direct and collapsed alternatives on LLaVA-v1.5 and Qwen3-VL backbones.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grants 62571271 and 62201552; the Natural Science Foundation of Tianjin under Grant No.24JCQNJC01890, and the Fundamental Research Funds for the Central Universities.

Supplementary Material

Filling Before Advancing: Capability-Gap-Driven Post-Training for Scenario-Specialized Remote Sensing MLLMs

S1 Additional Data-Curation Details

S1.1 Unified Data Collection and Curation Pipeline

Although the stages differ in supervision format, they share a common curation pipeline. We normalize heterogeneous sources into image-text or ShareGPT-style instruction records, with explicit image references, modality labels, and task fields when available. For image-text data, we enforce image-level uniqueness and retain one caption per image. For instruction data, we center conversations on visible evidence and exclude audit-only metadata from the final training input.

The main filtering criterion is visual groundedness. We reduce weakly visual expressions, including place names, addresses, precise distances, geographic coordinates, unsupported attributes, and metadata-dependent answers. For non-RGB modalities, language is constrained by observability: synthetic aperture radar (SAR) samples allow conservative uncertainty, near-infrared (NIR) samples avoid RGB color assumptions, and panchromatic (PAN) samples emphasize grayscale contrast, geometry, texture, and spatial organization (Zhu et al. 2021; Vivone 2023). Model-assisted rewriting produces descriptions, visual question answering (VQA), localization, region understanding, relation reasoning, modality-aware questions, and concise reports. Fixed teacher roles are used for metadata grounding, instruction synthesis, and evidence verification. Their prompts, generation settings, and retry outcomes are retained only as audit information and are excluded from the exported student records. Subsequent checks remove malformed conversations, duplicated answers, missing image tokens, coordinate-style leakage, and modality-incompatible claims.

S1.2 Stage-Specific Progressive Design

Stage 1 constructs RS-Anchor, a broad RGB remote-sensing (RS) visual-language anchor. After caption cleaning, image deduplication, scene classification, and diversity-aware sampling, it retains 569,853 unique image-caption pairs from 3,135,250 original samples across eight public datasets (Lu et al. 2018; Yuan et al. 2022; Cheng et al. 2022; Ge et al. 2025; Yuan et al. 2025; Soni et al. 2025). The retained captions emphasize observable land cover, land use, spatial layout, object distribution, and coarse scene structure. Metadata-heavy descriptions are filtered to provide visually grounded semantics before instruction tuning and non-RGB exposure. The retained Stage 1 inventory comprises RSTeller (407,566 pairs), RSSRData (95,776), NWPU Caption (29,609), ChatEarthNet subsets (15,278), RSICD (10,219), EarthDial (8,806), UCM Captions (2,039), and Sydney Captions (560). Smaller caption datasets supply targeted coverage where available, while RSTeller preserves scale after strict visual-grounding filters.

Stage 2 constructs Bridging-Conv for Domain-Bridging Convergence. The final curated mixture contains 187,296 supervised fine-tuning (SFT) samples across RGB, SAR, NIR, and PAN. Its RGB subset contains 99,088 image-unique samples, including 59,453 water-, coast-, port-, dock-, or ship-related records. This bridging stage is scenario-aware without collapsing into the final harbor task. Its underlying sources include SAR text-anchored data, object and ship recognition datasets, multispectral land-cover corpora, public caption and VQA resources, near-domain maritime samples, and additional public harbor imagery. Representative documented sources in these families include OpenSARShip, SEN12MS, BigEarthNet, DOTA, DIOR, and EarthDial (Huang et al. 2018; Schmitt et al. 2019; Sumbul et al. 2019; Xia et al. 2018; Li et al. 2020; Soni et al. 2025). These sources are not concatenated as raw records; they are converted into image-centered SFT conversations, rewritten under modality constraints, and audited for visual groundedness. The non-RGB subsets follow modality-specific constraints: SAR emphasizes structural layout and uncertainty control, NIR avoids RGB color assumptions, and PAN stresses geometry, edges, grayscale contrast, and spatial organization.

Stage 3 constructs Scenario-EG for Evidence-Grounded Harbor Tuning. It contains 53,000 train-only ShareGPT samples associated with 8,703 RGB, SAR, PAN, and NIR images. Its supervision emphasizes relation reasoning, presence validation, multi-cell grid localization, and functional-zone descriptions that distinguish dominant, secondary, mixed, and uncertain interpretations. Controlled non-harbor negatives teach evidence-based rejection, while short-answer VQA, concise-response replay, and Stage 2 replay help preserve direct answering and broader RS behavior (Rolnick et al. 2019). The final pool is selected from a larger intermediate set and audited to remove malformed conversations, missing image references, metadata and task-field leakage, benchmark-specific traces, underscore labels, and unsupported identity, activity, cargo, location, or temporal claims. Table S1 summarizes the three training-supervision pools; the separately maintained evaluation packages are documented in Section S3.

Source-family inventory. The Stage 2 and Stage 3 pools draw on complementary sensor and task families rather than raw benchmark merges. SAR supervision combines text-anchored radar records with OpenSARShip and HRSID for ship, water, port, and structural-layout evidence under conservative uncertainty (Huang et al. 2018; Wei et al. 2020). PAN and multispectral supervision uses PANBench, SEN12MS, and BigEarthNet to emphasize grayscale structure, reflectance, texture, geometry, and land-cover context without importing RGB-only color assumptions (Wang et al. 2023a; Schmitt et al. 2019; Sumbul et al. 2019).

Object- and region-oriented records from DOTA, DIOR, and optical or SAR ship-recognition sources, including ShipRSImageNet, are converted into natural-language localization, relation, and region-understanding conversations (Xia et al. 2018; Li et al. 2020; Zhang et al. 2021). EarthDial, EarthGPT, and established RS caption corpora provide

Pool	Modality	Scale and audit focus
RS-Anchor (D_1)	RGB	569,853 image-caption pairs; image-level deduplication and visible-semantic filtering.
Bridging-Conv RGB (D_2)	RGB	99,088 SFT records, including 59,453 water-, coast-, port-, dock-, or ship-related records.
Bridging-Conv non-RGB (D_2)	SAR, NIR, PAN	SAR 29,984; NIR 28,475; PAN 29,749, with sensor-specific observability constraints.
Scenario-EG (D_3)	RGB, SAR, NIR, PAN	53,000 train-only SFT records over 8,703 images; metadata and benchmark traces removed.

Table S1: Training-supervision pools used by the progressive route. Evaluation records and answer-bearing fields are excluded from every pool.

description and question-answering material, while MME-RealWorld contributes only evidence-compatible rejection and contrastive repair cases (Soni et al. 2025; Zhang et al. 2024a; Lu et al. 2018; Zhang et al. 2025). Additional public harbor and maritime imagery expands variation in dock, quay, coastal industrial, and ship-dense layouts. All retained records undergo the same image deduplication, modality-aware rewriting, SMT verification, leakage removal, and train-only export described above.

Count traceability and data separation. The reported counts are derived from frozen construction inventories and independently checked against the final exported training pools. For each stage, the audit records preserve source-level counts, image-level deduplication outcomes, modality composition, missing-image checks, and the removal of metadata or benchmark-related fields. Superseded intermediate pools are excluded from the totals, preventing duplicate counting when a sample is rewritten or replaced. Before training, all retained records pass image deduplication, modality-aware rewriting, SMT verification, leakage removal, and train-only export. The resulting pools are therefore curated instruction data rather than raw benchmark merges, while source identifiers and collection notes remain available for internal audit.

Harbor imagery coverage. The harbor-oriented records used for Bridging-Conv, Scenario-EG, HarborEval, and OpenEval are designed to cover more than a single local port style. As summarized visually in main-paper Fig. 2(b), the collection spans Pacific and Atlantic coasts, the Gulf of Mexico, Atlantic Europe, the Baltic Sea, the Mediterranean Coast, the East Asia Coast, the South China Sea, and maritime Southeast Asia. Within East Asia, additional representative large-port imagery strengthens examples of coastal industrial harbor layouts, berthing zones, storage yards, inland-water connections, and ship-dense operational scenes. These geographic and functional additions are used as curation diversity rather than as location labels exposed to the student model.

Figure S1 gives a lexical sanity check of the harbor-oriented source vocabulary used during data construction and

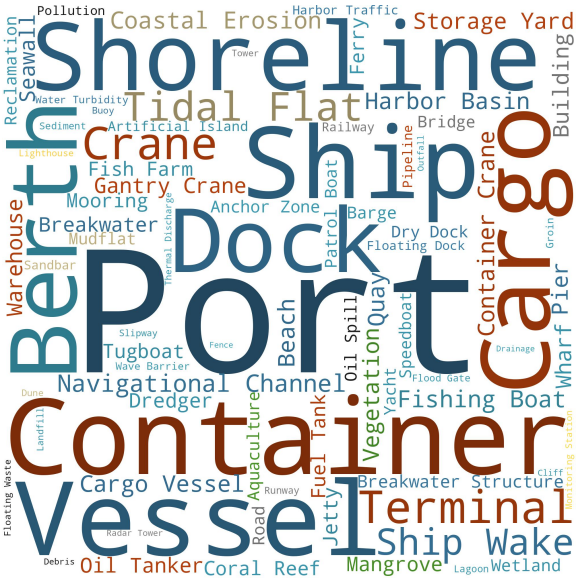

Figure S1: Word-cloud summary of harbor-oriented vocabulary used as a qualitative coverage check for scenario terms, functional-zone cues, maritime objects, and surrounding coastal evidence.

auditing; word size is used only for visualization, not as a reported quantitative statistic. The dominant terms cover core port infrastructure and operations, including ports, docks, berths, terminals, containers, cranes, cargo, vessels, and navigational channels. Medium- and small-frequency terms further cover coastal context and surrounding evidence, such as shoreline, tidal flats, breakwaters, seawalls, ferry/boat types, storage yards, wetlands, mangroves, aquaculture, oil tanks, and coastal erosion. This vocabulary spread supports the intended data role: the target scenario is represented not only by object names, but also by functional zones, maritime infrastructure, shoreline context, environmental surroundings, and evidence cues used for grounded reporting and rejection.

S2 Bridging-Conv Construction Workflow

The main paper provides the formal definition of Bridging-Conv; here we focus on its operational construction. For each source image, the pipeline first identifies the sensor modality and normalizes the available source and scene evidence under modality-specific observability constraints. RGB, SAR, PAN, and NIR inputs are then converted into their corresponding training representations and paired with target or neighboring bridging context, including harbor and port scenes, water, coasts, ships, docks, and related infrastructure. An instruction-synthesis teacher converts this grounded evidence into a candidate image–instruction–answer record, after which a stricter verifier either retains the record, rewrites unsupported wording, or discards the sample. The resulting audited records form the Stage 2 Bridging-Conv supervision pool. Source metadata, construction tags, and verifier rationales support curation and auditing only; they are never exposed to the student model as privileged training inputs.

Teacher and verifier identities, prompts, rubrics, and retry rules are retained as implementation records for reproducibility.

S2.1 Staged Multi-Teacher Distillation

We use staged multi-teacher (SMT) distillation as a construction-time mechanism, not as a separate model component. Its three ordered roles are metadata grounding, instruction synthesis, and evidence verification. Metadata grounding normalizes source and modality tags under sensor-specific observability constraints. Instruction synthesis then produces ShareGPT-style image–instruction–answer records, and the verifier retains, rewrites, or drops candidates under stricter visual-evidence rules. The student sees only the final audited records, while private metadata, verifier rationales, and teacher notes are removed.

For Bridging-Conv, SMT controls hallucination risks introduced by heterogeneous sources and sensor-dependent observability (Li et al. 2023b). For SAR, PAN, and NIR, verification suppresses unsupported color, material, fine-identity, and activity claims. Across modalities, it removes weakly visual answers, metadata-dependent statements, duplicate image references, benchmark traces, and unsupported operation descriptions. Retained samples emphasize observable scene structure, objects, spatial relations, modality-aware evidence, and calibrated uncertainty; unreliable samples are discarded rather than template-expanded.

S3 Evaluation Suite Construction and Audit

S3.1 Evaluation Package Roles and Separation

The paper uses five separately maintained evaluation packages, summarized in Table S3. RS-VL Val. and MultiSource Val. are frozen stage diagnostics that test whether the intended capabilities emerge after Stages 1 and 2. HarborEval is the principal scenario-diagnostic benchmark for final harbor understanding. Public Harbor tests transfer on public-source harbor subsets derived from VRSBench and RSVQA, while OpenEval examines open-ended grounded reporting under expert review. Keeping these packages separate prevents stage-level validation from being conflated with the final benchmark claim and clarifies the role of every reported score.

S3.2 HarborEval Construction and Audit

HarborEval is constructed as a scenario-diagnostic evaluation set rather than as another training corpus. Its purpose is to probe whether a model specialized for coastal harbor RS can combine object recognition, functional interpretation, spatial reasoning, sensor-aware observability, uncertainty handling, non-harbor rejection, and grounded reporting under the same input restrictions used for all compared models. The benchmark is maintained separately from the stage-wise training pools. In particular, the Stage 3 Scenario-EG pool is train-only, whereas HarborEval retains private answers, evidence, accepted labels, and scoring rubrics only for evaluation and auditing. Once a source record is assigned to HarborEval, its derived conversations are excluded from the training export; this source-record holdout is distinct from the field-level controls applied during inference.

S3.3 Evaluation Scope and Public/Private Separation

HarborEval contains 1,245 items over 471 unique images. Among them, 1,154 items are structured closed-form items and 91 items are open-ended description or rejection items. The average number of items per image is 2.64 and the maximum is 7. This item–image structure lets HarborEval probe several capability roles on the same visual source when appropriate, while keeping public inference records separate from answer-bearing audit records.

The benchmark is organized into public inference records and private answer records. During inference, a model receives only the image, question, and answer choices when the task is closed-form. Fields such as modality, scene, difficulty, metric type, track name, answer, reference answer, evidence objects, evidence relations, accepted grid cells, and scoring rubrics are not inserted into the prompt. The track field is retained only for evaluator-side grouping. The field-level audit found no disallowed private fields in the public inference records. This check complements, but does not replace, the source-record holdout described above.

S3.4 Track Construction

HarborEval contains eight diagnostic tracks. The first seven tracks focus on harbor-scenario understanding and grounded reporting, while the eighth track tests whether the model avoids forcing harbor interpretations onto non-harbor or near-domain scenes. The headline score in the main paper macro-averages all eight tracks after normalizing track-specific metrics to a common scale; for diagnostic interpretation, the T8 rejection track and the T7 open-ended description track are also inspected separately.

The tracks are tied to capability roles rather than to isolated task formats. T1 and T2 measure whether the model recognizes scenario-relevant objects and functional areas. T3 and T4 measure whether it can reason about relative layout and localize evidence without bounding-box supervision. T5 checks whether the model respects sensor-dependent observability constraints across RGB, SAR, PAN, and NIR. T6 tests whether the model can distinguish visible evidence from unsupported claims, including cases where identity, cargo type, operator, exact status, or location cannot be determined from imagery alone. T7 tests whether open-ended reports remain grounded and concise. T8 tests whether the model rejects false harbor interpretations in non-harbor or near-domain scenes.

S3.5 Construction and Cleaning Workflow

The construction process starts from harbor and non-harbor RS records and converts them into item-level diagnostic questions. For harbor records, we generate or audit object/scene questions, functional-zone labels, relation questions, grid-localization targets, modality-observability questions, evidence-calibrated yes/no/unknown judgments, and open-ended report prompts. For non-harbor records, we construct rejection-oriented questions that require the model to

SMT-P1 Evidence Grounding

```
Role: evidence-grounding teacher; produce audit tags, not final SFT answers.
Inputs: {image/rendering}, {modality: RGB|SAR|PAN|NIR}, {source fields}, {scenario context}.
Check: visible objects; functional zones; spatial layout; sensor-specific observability;
uncertainty risks.
Forbid: unsupported location, operator, cargo, identity, time, exact activity, and RGB-only
color claims for non-RGB images.
Output: {scene tags, visible evidence, spatial layout, modality limits, forbidden claims,
risk flags}.
```

SMT-P2 Instruction Synthesis

```
Role: synthesize one ShareGPT-style Bridging-Conv candidate.
Inputs: {image/rendering}, {evidence tags}, {task role}, {modality}, {target or neighboring
bridging context}.
Task menu: scene/object recognition; relation reasoning; spatial description; modality
judgment; calibrated uncertainty.
Constraint: ask and answer only from visible, modality-compatible evidence; never expose
metadata, labels, coordinates, rubrics, or source fields.
Output: {user turn, assistant turn, task type, modality, evidence summary, construction
tags}.
```

SMT-P3 Verify, Repair, or Drop

```
Role: strict verifier for evidence-grounded training records.
Inputs: {image/rendering}, {candidate SFT record}, {evidence tags}, {forbidden claims}.
Audit: visual support; modality compatibility; answer length; duplicate image references;
benchmark traces; metadata leakage; unsupported identity/cargo/operator/location/time claims.
Decision: KEEP if supported; REWRITE if the task is useful but wording is unsafe; DROP if
evidence is insufficient or the sample is unrecoverable.
Output: {decision, revised user, revised assistant, evidence status, flags, drop reason}.
```

Table S2: Semi-structured SMT prompt sketches for Bridging-Conv construction. The three roles decompose Stage 2 synthesis into evidence grounding, instruction generation, and verifier-based repair. Braced fields are populated during construction; private source fields, verifier rationales, and teacher notes are removed before ShareGPT export.

Package	Role	Modal.	Scale	Primary diagnostic or scoring focus
RS-VL Val.	Stage 1 semantic-alignment validation	RGB	356 images	Ten-way hard-negative retrieval and length-matched pairwise discrimination.
MultiSource Val.	Stage 2 cross-sensor validation	RGB / SAR / PAN / NIR	Frozen generation and choice tracks	Sensor-aware generation, modality recognition, retrieval, and pairwise discrimination.
HarborEval	Final scenario-diagnostic benchmark	RGB / SAR / PAN / NIR	1,245 items; 471 images	Eight tracks covering harbor understanding, grounding, calibration, reporting, and rejection.
Public Harbor	Public-source harbor transfer	RGB	580 samples; 200 images	430 VQA items and 150 captioning items derived from VRSBench and RSVQA.
OpenEval	Expert-scored open reporting	RGB / SAR / PAN / NIR	100 items	Grounded reporting, expected uncertainty, forbidden-claim control, and non-harbor rejection.

Table S3: Evaluation packages used throughout the paper. Stage diagnostics measure intermediate capability acquisition; HarborEval, Public Harbor, and OpenEval assess final scenario behavior and generalization.

Track	Role	Items	Question types
T1	Object/scene VQA	162	95 ML; 51 MC; 16 Y/N/U
T2	Functional zone	182	182 MC
T3	Relation reasoning	183	165 MC; 18 Y/N/U
T4	Grid grounding	164	164 grid-choice
T5	Observability	171	171 Y/N/U
T6	Evidence judgment	123	123 Y/N/U
T7	Report generation	79	79 open-ended
T8	Non-harbor rejection	181	113 MC; 56 Y/N/U; 12 open-ended

Table S4: HarborEval track composition. ML denotes multi-label, MC multiple-choice, and Y/N/U yes/no/unknown.

Audit item	Value	Interpretation
Closed and open split	1,154/91	Package scope.
Images and mean items	471/2.64	Item scores are correlated by image.
T5–T6 revised/replaced	294/64	Weakly visual items were revised.
T2 majority-class rate	32.97%	Context for functional-zone accuracy.
T4 multi-correct ratio	40.24%	Motivates accepted sets and soft grid F1.
T7 complete audit packets	79	All open reports retain private scoring support.
Public-field violations	0	Private answer and evidence fields are excluded.

Table S5: Selected HarborEval audit statistics used to interpret diagnostic scores.

avoid hallucinating docks, quays, cargo terminals, vessels, or port logistics when these elements are not supported by the image.

The cleaning workflow removes weakly visual questions, metadata-dependent answers, benchmark-specific traces, duplicated image references, malformed option sets, and modality-incompatible statements. For non-RGB items, we audit whether the wording relies on visible structure, grayscale contrast, backscatter/texture, reflectance, edges, or layout rather than RGB-only color assumptions. For evidence-calibrated questions, we remove or rewrite prompts that can be answered only from metadata, geographic knowledge, source filenames, hidden labels, or external facts. For open-ended items, forbidden-claim fields are kept in the private answer package to penalize unsupported port names, vessel identities, cargo types, coordinates, operational claims, and temporal claims during scoring.

This audit is intentionally conservative. It does not claim that every visually ambiguous item has a single uniquely correct answer; instead, the benchmark stores accepted alternatives only where the private evidence supports them. Functional-zone items can accept secondary or mixed labels when the image genuinely supports more than one coarse interpretation. Relation items remain stricter: accepted relation alternatives are added only when relation boundaries such as near versus adjacent or docked versus moored are visually defensible. Grid-grounding items use accepted grid-cell sets so that large or multi-cell objects are not penalized for crossing a single canonical cell boundary.

S3.6 Answer Distribution and Boundary Cases

The package is designed as a diagnostic benchmark, so answer balance is inspected at the track level rather than assumed globally. Several subtracks have asymmetric labels for substantive reasons: some auxiliary yes/no subsets are biased toward visible evidence, while the rejection track naturally contains more negative decisions because unsupported harbor-specific claims should be refused. We therefore report the headline track macro together with per-track scores and auxiliary strict or open-ended diagnostics instead of relying on a single item-level accuracy number.

T2 contains six coarse functional-zone classes. The largest class is water or navigation area with 60 items, followed by mixed or uncertain port area with 39 items, berthing area with 28 items, cargo storage area with 25 items, marina or small-boat harbor with 16 items, and industrial or logistics area with 14 items. This distribution reflects the visual structure of harbor scenes: water/navigation areas are frequent, but the task still requires distinguishing storage, berthing, industrial/logistics, marina, and mixed-use zones. T4 contains 164 grid-grounding items. Among them, 98 items have a single accepted grid cell and 66 have multiple accepted cells, with an average accepted-cell set size of 1.79 and a maximum of 9. The multi-correct ratio of 40.24% reflects the fact that vessels, docks, basins, and storage areas often span grid boundaries.

S3.7 Open-ended Audit

The T7 caption/report track contains 79 open-ended items. Each item includes private scoring support: a scoring rubric, forbidden claims, a reference answer, and key points. The rubric has priority over optional key points during frozen T7 judging. Environmental descriptors such as clear weather, calm sea state, daytime, nighttime, or modality names are treated as optional consistency cues unless explicitly required by the rubric. A model is therefore not penalized solely for omitting such cues when it correctly reports visible objects, relations, and layout. Conversely, hallucinated objects, unsupported port names, vessel identities, cargo types, coordinates, or operational details are penalized.

This design is important for RS reporting. Overhead imagery often supports functional and spatial interpretation, but it rarely supports exact vessel identity, cargo type, operator, port name, throughput, or timestamp-level claims. The open-ended audit therefore rewards grounded coverage and relation correctness while discouraging fluent but unsupported operational narratives. For T8 open-ended rejec-

tion items, the audit checks whether the response describes visible non-harbor scene elements and avoids unsupported harbor-specific claims such as quay, berth, terminal, docked vessel, cargo handling, or port logistics when those elements are absent.

S3.8 Image-level Correlation and Reporting

HarborEval is item-based because different diagnostic questions can be derived from the same image. This design improves coverage of capability dimensions without requiring a much larger image set, but it introduces image-level correlation. The package has 471 unique images, an average of 2.64 items per image, and a maximum of 7 items per image. Most tracks have approximately one item per image, while T8 has 181 items over 63 unique images, averaging 2.87 items per image. For this reason, the main paper reports item-level metrics for compactness and treats image-level macro scores as an auxiliary analysis.

S3.9 Remaining Benchmark Boundaries

The audit does not remove all sources of ambiguity. Some harbor functional zones are genuinely mixed, some spatial relations are boundary-dependent, and grid localization can remain ambiguous when objects span multiple cells. The accepted-answer mechanism reduces avoidable unfairness but does not replace visual inspection for borderline cases. Similarly, the T1 and T3 auxiliary yes/no subsets are retained for diagnostic completeness but are not used alone as evidence of scenario understanding. The T8 track is intentionally a rejection diagnostic and should be interpreted as a hallucination-control probe rather than as a general scene-understanding benchmark. These boundaries are consistent with the role of HarborEval in this paper: it is a compact diagnostic protocol for scenario-specific RS multimodal large language model (RS-MLLM) specialization, not an exhaustive benchmark for all harbor RS tasks.

S4 Implementation and Reproducibility Details

Backbones and training routes. The controlled experiments instantiate the same data-stage route on two representative multimodal large language model (MLLM) families. The LLaVA-style route uses Vicuna-7B, a CLIP ViT-L/14-336 vision tower, and a LLaVA-compatible multimodal projector (Radford et al. 2021; Liu et al. 2023, 2024b). The native route uses Qwen3-VL-8B and its built-in multimodal processor (Bai et al. 2025a). For each backbone, we compare the proposed staged route with direct Evidence-Grounded Harbor Tuning and Collapsed-SFT baselines. Collapsed-SFT denotes a single target-stage optimization run over the union of Bridging-Conv and Scenario-EG samples, rather than capability-ordered exposure. The S1+Collapsed variant first performs RS Semantic Anchoring and then collapses Bridging-Conv and Scenario-EG into one subsequent SFT stage. These routes keep the backbone, evaluation sets, semantic prompts, decoding policy, and scoring rules fixed within each family, so differences can be attributed to the ordering and composition of the post-training data rather than to unrelated model changes.

Frozen manifests and checkpoint policy. All controlled training uses frozen data manifests and a fixed random seed (default seed 42). Each stage follows a predeclared data exposure rather than selecting a checkpoint on a validation score. Except when resuming an interrupted run, the reported model is `checkpoint-final` after the planned epochs; intermediate `checkpoint-*` states are retained only for recovery and audit. Evaluation records and answer-bearing fields remain outside the training manifests. Public inference inputs never contain private answers, accepted alternatives, evidence annotations, scoring rubrics, verifier rationales, teacher metadata, or source-side construction notes.

Stage-wise exposure and learning rates. Stage 1 trains for one epoch on 569,853 RS-Anchor RGB image–caption pairs, corresponding to approximately 35.6K optimizer steps at effective batch size 16. The LLaVA route principally trains the multimodal projector with learning rate 1×10^{-3}, whereas Qwen3-VL uses LoRA with learning rate 1×10^{-4}. Stage 2 trains for one epoch on 187,296 Bridging-Conv SFT records (approximately 11.7K steps). Its modality composition is RGB 99,088 (52.9%), SAR 29,984 (16.0%), NIR 28,475 (15.2%), and PAN 29,749 (15.9%). LLaVA uses projector and LoRA learning rates of 2×10^{-4} and 1×10^{-4}, respectively; the final Qwen3-VL configuration uses the milder LoRA learning rate 3×10^{-5}. Stage 3 uses 53,000 train-only Scenario-EG ShareGPT records, or approximately 3.3K steps per epoch. Its modality proportions are RGB 70.1%, SAR 23.7%, NIR 2.2%, and PAN 4.0%; tasks cover presence validation, relation reasoning, grid localization, functional-zone understanding, and open rejection, with about 15% non-harbor or ambiguous hard negatives. Terminal routes use a frozen one- or two-epoch exposure according to the backbone-specific run manifest, with LLaVA and Qwen3-VL terminal LoRA learning rates of 1×10^{-4} and 3×10^{-5}, respectively.

Adapters and optimization. Unless otherwise specified by a frozen intermediate-stage manifest, LoRA adapters (Hu et al. 2022) target `q_proj`, `k_proj`, `v_proj`, and `o_proj` with rank 64, alpha 128, and dropout 0.05. The smaller rank-32, alpha-64 setting is restricted to selected Stage 2 intermediate adapters; terminal reported routes use rank 64 and alpha 128. All SFT runs use AdamW with $\beta = (0.9, 0.95)$, $\epsilon = 10^{-8}$, cosine scheduling, warmup ratio 0.03, weight decay 0.05, maximum gradient norm 1.0, and bf16 precision. LLaVA uses maximum sequence length 2048, per-device batch size 2, and gradient accumulation 8. Qwen3-VL uses maximum sequence length 2048, image-token budget 512, per-device batch size 1–2, and gradient accumulation 8–16. These settings keep the effective batch size at approximately 16.

Baseline and replay controls. Direct-SFT uses only D_3. Collapsed-SFT trains once on $D_2 \cup D_3$, and S1+Collapsed first loads the Stage 1 checkpoint before the same collapsed SFT exposure. Within each backbone family, optimizer, adapter, decoding, prompt, and normalization settings are

aligned with the full route. Stage 2/3 replay is limited to small retained components, including concise-response replay and Stage 2 bridge-observability replay during Stage 3, to reduce forgetting of short-answer behavior and multi-source observability without changing the primary stage objective.

Inference and scoring. All models use the same benchmark-specific semantic prompts within each evaluation set. HarborEval inputs include only the image, question, and answer choices when available; modality labels, task metadata, answers, evidence fields, accepted labels, and scoring rubrics are excluded from model prompts. Closed-form HarborEval tracks use deterministic decoding with sampling disabled, temperature 0, and beam size 1. Open-ended description and reporting tracks use the same deterministic policy with longer response budgets. Closed-form tracks are scored programmatically, while T7 uses a fixed image-grounded multimodal judge under a frozen rubric (Zheng et al. 2023); OpenEval uses expert scoring with anonymized responses. External RS-MLLMs are evaluated with their official checkpoints and recommended prompts when available, then normalized into the same prediction and scoring format.

Hardware and audit trail. Training is performed on NVIDIA RTX A6000 48GB GPUs. Each training run uses one GPU; two GPUs are used only to execute independent experiments in parallel. Wall-clock duration varies substantially by backbone and stage and is therefore not treated as a defining experimental condition. Instead, the retained run logs record start and end times, optimizer-step trajectories, model states, frozen manifests, preprocessing and decoding configurations, raw predictions, normalized outputs, dimension-level scores, and expert-scoring sheets for every reported run.

S5 Evaluation Protocol and Scoring Details

All model comparisons use the same public inference partition, images, semantic prompts, decoding settings, normalization rules, and scoring criteria within each evaluation set. HarborEval exposes only the item identifier, image reference, question, answer choices when applicable, question type, and track identifier used for evaluator-side grouping. Answer keys, accepted alternatives, reference answers, evidence annotations, modality labels, audit notes, and scoring rubrics remain private and are merged only after inference. For reproducibility, the evaluation record retains raw predictions, normalized predictions, closed-form and open-ended scores, hallucination flags, dimension-level assessments, and adjudication notes. This separation keeps the scoring process auditable while focusing the paper description on evaluator-visible logic.

S5.1 Prediction Normalization

The normalizer is used to remove superficial formatting loss rather than to correct semantic errors. It extracts option keys from variants such as "A", "A.", or "Option A", maps option text back to keys when the selected option is unambiguous, standardizes whitespace and Unicode variants, maps yes/no/unknown aliases to the canonical decision set, maps grid synonyms such as "upper-left" to *top-left*, and extracts explicit T5/T6 decisions from short explanatory answers. Each normalization action is logged so that raw and normalized scores can be compared. A large raw-normalized gap is treated as a format-sensitivity diagnostic, while the normalized score is used as the main reported score.

S5.2 Closed-form Track Scoring

HarborEval uses track-specific metrics normalized to $[0, 1]$. Multiple-choice and yes/no/unknown items are scored by accepted-answer accuracy:

$$s_i = \mathbf{1}[\hat{a}_i \in \mathcal{A}_i], \tag{S1}$$

where $\hat{a}_i$ is the normalized prediction and $\mathcal{A}_i$ is the accepted answer set. A strict score is also recorded by comparing $\hat{a}_i$ with the canonical answer only. This distinction matters for visually ambiguous functional-zone and calibrated-evidence items, where the accepted set may contain a small number of reviewer-approved alternatives.

Multi-label T1 object questions are scored by set F1. Given predicted set $\hat{Y}$ and gold set Y, precision, recall, and F1 are computed as

$$P = \frac{|\hat{Y} \cap Y|}{|\hat{Y}|}, \quad R = \frac{|\hat{Y} \cap Y|}{|Y|}, \quad F_1 = \frac{2PR}{P + R}, \tag{S2}$$

with the usual zero-handling when a predicted or gold set is empty. This choice discourages the degenerate strategy of selecting every visible category: recall may increase, but precision decreases.

T4 grid grounding uses a 3×3 grid with the canonical cells *top-left*, *top-center*, *top-right*, *middle-left*, *middle-center*, *middle-right*, *bottom-left*, *bottom-center*, and *bottom-right*. Because vessels, basins, piers, and storage regions can straddle cell boundaries, the main T4 metric is soft grid F1. Pairwise cell similarity is 1.0 for the same cell, 0.5 for edge-adjacent cells, 0.25 for diagonal-adjacent cells, and 0 otherwise. Each item stores one audited accepted-cell set. The scorer performs deterministic one-to-one matching from high to low similarity; ties follow the canonical row-major cell order. It then computes soft precision, soft recall, and soft F1:

$$P_g = \frac{m_g}{|\hat{G}|}, \quad R_g = \frac{m_g}{|G|}, \quad F_g = \frac{2P_gR_g}{P_g + R_g}, \tag{S3}$$

where m_g is the summed soft match score, $\hat{G}$ is the predicted grid-cell set, and G is the accepted grid-cell set. Exact grid F1, precision, recall, and exact match are retained as auxiliary diagnostics.

T5 and T6 use the same decision vocabulary, *Yes*, *No*, and *Cannot determine*, but probe different evidence roles. T5 asks whether a property is observable under the supplied modality and image quality; T6 asks whether a visually grounded claim is supported, contradicted, or not decidable from the image. The main T5/T6 score is semantic decision accuracy over accepted answers, while strict accuracy compares only with the canonical decision. During the final audit, weakly visual T5/T6 records are rewritten or replaced, including 64 visual replacements, so that these tracks emphasize visible evidence, modality observability, and calibrated rejection rather than text priors.

S5.3 Headline HarborEval Aggregation

For each track t, the track score S_t is the average of item-level scores in that track after applying the track metric above. For T7, S_7 is the normalized open-ended reporting score produced by the fixed image-grounded judging procedure. Let $\mathcal{T}_{\mathrm{HE}}$ denote the eight HarborEval diagnostic tracks. The headline score macro-averages the normalized track scores:

$$H_{\mathrm{eval}} = \frac{1}{|\mathcal{T}_{\mathrm{HE}}|} \sum_{t \in \mathcal{T}_{\mathrm{HE}}} S_t. \tag{S4}$$

We additionally report closed-form micro accuracy, strict variants, T5/T6 semantic and strict accuracies, T4 exact-grid diagnostics, and T7 rubric dimensions to make clear whether a gain comes from structured VQA, spatial grounding, evidence calibration, rejection, or open-ended reporting.

S5.4 Open-ended T7 and OpenEval Scoring

T7 evaluates whether a model can produce a grounded RS report rather than merely select an option. Each T7 scoring packet contains the image, question, reference answer, key visible objects, evidence or relation fields when available, positive criteria, forbidden claims, and an anonymized model response. A fixed image-grounded multimodal judge scores object coverage, object accuracy, spatial relation accuracy, functional scene understanding, modality or environment awareness, hallucination control, concision, and overall quality on a 0–5 scale, then normalizes the aggregate to the reporting scale. The scoring rubric has higher priority than optional key points: a response is rewarded for accurate visible evidence and penalized for unsupported port names, vessel identities, cargo types, coordinates, operating status, temporal claims, or objects absent from the image.

OpenEval is the broader expert-scored open-ended protocol used when a free-form answer must be assessed beyond closed-form matching. Its dimensions emphasize visual groundedness, relevant object and region coverage, functional-zone plausibility, spatial relation correctness, uncertainty or rejection behavior, hallucination control, and concise reporting. Model identities and training routes are hidden from the scoring sheet. The score record keeps the anonymized sample identifier, anonymous model identifier, dimension-level scores, hallucination or forbidden-claim flags, short reviewer notes, and adjudication status when a borderline or inconsistent case requires review by an additional expert.

S5.5 Public Harbor Scoring

Public Harbor is a VRSBench- and RSVQA-derived RGB RS evaluation subset designed to assess harbor-domain generalization (Lobry et al. 2020; Li, Ding, and Elhoseiny 2024). It contains 580 public test-style samples over 200 images, including 430 VQA items and 150 image-captioning items. The VQA portion covers object category, existence, quantity, color, shape, size, position, direction, scene type, and reasoning questions, while the captioning portion evaluates detailed image description.

Public Harbor is maintained as an evaluation-only partition separate from the stage-wise training pools and from HarborEval. Its reference answers and captions are hidden during inference and merged only by the scorer. This keeps the subset useful as an external public-source generalization check rather than as an additional source of training supervision.

For VQA, we follow the VRSBench semantic matching protocol: answers are first evaluated by relaxed substring matching, yes/no and numerical answers are evaluated by strict exact match, and remaining open-set answers are evaluated by an LLM semantic matcher (Li, Ding, and Elhoseiny 2024). For captioning, we use a CLAIR-style LLM score measuring semantic consistency between the generated and reference captions (Chan et al. 2023). Let n_{vqa} and n_{cap} be the numbers of VQA and captioning samples. The final Public Harbor score is computed as a sample-weighted aggregate after mapping both components to a 0–100 scale:

$$S_{\mathrm{PH}} = \frac{n_{\mathrm{vqa}} A_{\mathrm{VQA}} + n_{\mathrm{cap}} (100 C_{\mathrm{CLAIR}})}{n_{\mathrm{vqa}} + n_{\mathrm{cap}}}, \tag{S5}$$

where A_{VQA} is VQA accuracy on the 0–100 scale and C_{CLAIR} is the caption semantic-consistency score on the 0–1 scale. For Public Harbor, $n_{\mathrm{vqa}} = 430$ and $n_{\mathrm{cap}} = 150$.

S5.6 Stage-validation Scoring

RS-VL Val. and MultiSource Val. are retained only as stage-diagnostic metrics, not as separate public benchmark claims. RS-VL Val. measures Stage 1 RS semantic anchoring over RGB imagery, while MultiSource Val. measures Stage 2 adaptation to RGB/SAR/PAN/NIR observability. These sets are frozen before ablation inference, and all checkpoints receive identical images, prompts, and scoring rules. They support the intended interpretation of the staged route: RS-VL Val. diagnoses semantic anchoring, MultiSource Val. diagnoses cross-sensor bridging convergence, HarborEval diagnoses terminal scenario behavior, Public Harbor tests public-source generalization, and OpenEval tests open-ended grounded reporting.

We therefore interpret these metrics as checkpoint probes rather than leaderboards. Their role is to detect whether an intermediate stage has supplied the intended prerequisite before the final scenario objective is applied. A model can improve HarborEval after Stage 3 while still showing weak anchoring or weak sensor transfer; conversely, Stage 2 can improve cross-sensor observability without directly optimizing the final harbor-reporting score. Keeping the two validation sets separate makes these failure modes visible instead of hiding them inside a single terminal average.

Both stage-validation sets pair each image with one *retrieval* item and one *pairwise* item. Retrieval uses one positive caption and nine TF–IDF hard negatives; pairwise contrasts the positive with the length-matched hardest negative among the top-10 candidates, with A/B order randomized. For positive rank r, we report $\mathrm{R@}k = \mathbf{1}[r \le k]$, MRR $= 1/r$, and NDCG $= 1/\log_2(r+1)$, averaged over items. Pairwise accuracy is $\mathrm{Acc}_{\mathrm{pw}} = \frac{1}{N}\sum_i \mathbf{1}[\hat{y}_i = y_i]$, with A-rate bias checks retained.

The retrieval and pairwise items are paired by image so that one branch does not receive an easier image distribution than the other. Hard negatives are selected from the frozen

validation candidate pool rather than from the training pools, and the randomized A/B order is retained with the parsed prediction. We inspect A-rate bias and degenerate single-option behavior before aggregation; malformed or non-parseable answers are kept as errors rather than manually repaired.

RS-VL Val. RS-VL Val. scores RGB land-cover discrimination on 356 frozen images through retrieval and pairwise comparison. Retrieval reports R@1, R@5, MRR, and NDCG; pairwise reports $\mathrm{Acc}_{\mathrm{pw}}$. The headline score is $S_{\mathrm{RSVL}} = \frac{1}{2}(\mathrm{R@1} + \mathrm{Acc}_{\mathrm{pw}})$, equally weighting retrieval and forced-choice discrimination. Generative backbones emit a ranking directly, single-LLM backbones use per-candidate answer likelihood, and an auxiliary PMI ranking is reported only when image-conditioned likelihood behaves consistently. We use this diagnostic only for within-backbone stage trajectories; it is not intended to replace public RS retrieval or captioning benchmarks.

MultiSource Val. MultiSource Val. uses same-sensor hard negatives (RGB/SAR/PAN/NIR, ~ 80 images each) to probe sensor-specific evidence constraints. Its auxiliary *discrimination* track reuses retrieval and pairwise metrics per modality. Its primary *generation* track asks the model to describe each image without a modality label, then uses the same fixed image-grounded judging procedure to score scene understanding (d_1), object recognition (d_2), spatial–functional reasoning (d_3), modality-evidence use (d_4), evidence calibration (d_5), and clarity (d_6), each in $[0, 100]$. The private modality label is available only to the evaluator for enforcing sensor-specific evidence boundaries. The weighted composite is

$$S_{\mathrm{MS}}^{\mathrm{gen}} = \sum_{j=1}^{6} \omega_j d_j, \tag{S6}$$

with weights $\omega = (0.20, 0.25, 0.20, 0.20, 0.10, 0.05)$ for d_1–d_6. Clarity is down-weighted to avoid rewarding fluency over visual correctness. Severe scene errors, non-RGB color hallucinations, fabricated identities, and degenerate or refused answers are reflected in the corresponding dimension scores and retained as explicit failure flags. Macro scores average over present modalities, and modality-recognition accuracy is reported separately when the prompt requests a sensor guess. This modality-aware normalization prevents the larger RGB subset or fluent but sensor-incompatible reports from dominating the diagnostic.

S6 Prompt and Rubric Templates

All compared models receive the same semantic instruction within each task type. Model-specific chat wrappers may differ because each backbone has its own processor or conversation template, but the user-facing task, question text, choices, and decoding policy are fixed. Table S7 reports semi-structured inference templates that summarize the control logic used across HarborEval and Public Harbor. The placeholders are instantiated from public evaluation items; private answers, rubrics, accepted labels, evidence fields, and construction metadata are never inserted into the model input.

Dimension	**Reward**	**Penalty**
Groundedness	Visible or strongly supported evidence.	Unsupported names, coordinates, operators, cargo, activities, or timestamps.
Object coverage	Relevant visible vessels, docks, basins, storage, roads, buildings, vegetation.	Missing dominant evidence or adding absent objects.
Functional-zone understanding	Supported berthing, navigation, storage, logistics, mixed, uncertain, or non-harbor zones.	Confident zone labels for ambiguous or unsupported regions.
Spatial relation	Correct layout, adjacency, containment, grid location, and object-region relation.	Reversed locations or overstated geometry.
Uncertainty and rejection	Cannot determine or explicit rejection when evidence is insufficient.	Forced harbor claims under modality or resolution limits.
Concision	Useful, compact, non-redundant report.	Verbose generic text, speculation, or repeated boilerplate.

Table S6: Open-ended reporting rubric for expert OpenEval scoring and frozen T7 judging. Every dimension is scored on a 0–5 scale; forbidden claims are flagged separately.

Prompt normalization follows four practical rules. First, each prompt separates the task instruction from the answer channel, so that closed-form tracks can be parsed from an explicit final answer while open-ended tracks can still ask for concise supporting evidence. Second, the prompt never exposes modality labels, track names, accepted alternatives, evidence objects, or evaluator-side rubrics; these fields are merged only after inference. Third, prompts avoid dataset-specific wording and file-name cues, because the goal is to evaluate visual reasoning rather than benchmark memorization. Finally, all backbones use their own required image-token wrapper, but the semantic instruction inside the user turn remains unchanged.

For closed-form tracks, the scorer reads the first unambiguous final-answer token or option key. Multi-select questions accept a comma-separated option set only when the selected keys match the accepted set after normalization. Yes/no/unknown questions are intentionally phrased as evidence decisions: *Yes* requires clear positive evidence, *No* requires contradiction or modality-incompatible evidence, and *Cannot determine* is used when the target may be plausible but the image does not support the claim. This convention is important for SAR, PAN, and NIR inputs, where color, identity, cargo, ownership, operator, temporal status, and precise location claims are often not visually recoverable.

For open-ended reporting, the prompt asks for a compact visual report rather than a free-form story. Responses are rewarded for naming visible objects, functional zones, spatial

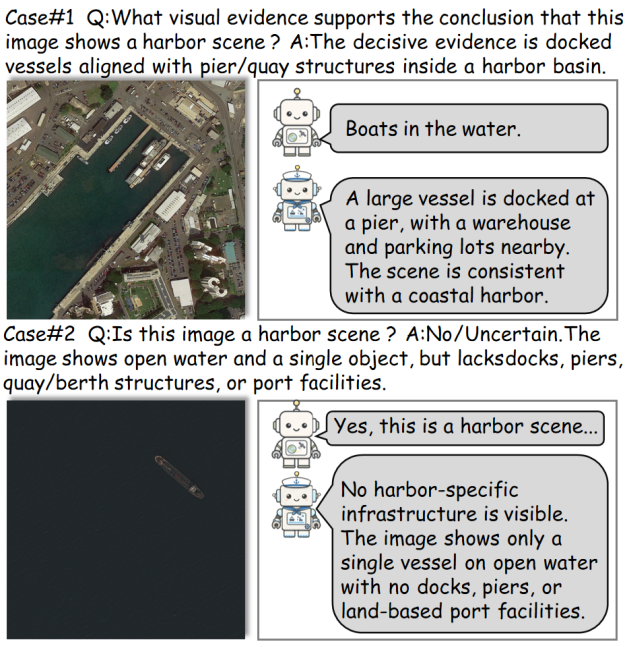

Figure S2: Positive harbor and water-adjacent hard-negative cases. The paired responses illustrate why water or a vessel alone is insufficient and how relational port infrastructure supports a grounded harbor decision.

relations, and uncertainty when appropriate, but penalized for unsupported port names, vessel identities, cargo types, operators, exact operations, or overconfident harbor interpretation under weak evidence. The shared evidence criteria provide a common auditing perspective for T7, T8 open reports, Public Harbor captions, and OpenEval despite their different output lengths.

Prompt instances are generated deterministically from the frozen public inference partition. The image, question, and options are inserted before applying the backbone-specific conversation wrapper. Decoding is fixed within each comparison and sampling is disabled for closed-form questions. The parser scores the explicit final token, retains any explanation for audit, and marks an unparseable response as ambiguous rather than repairing it manually. Thus, the semi-structured templates expose the invariant evidence, ambiguity, and output rules without implying that all backbones share an identical literal prompt. Evaluation inputs, model responses, parsed decisions, and score records are retained together so that abnormal results—especially on rejection and modality tracks—remain traceable to the exact response.

S7 Positive and Hard-Negative Evidence Cases

Figure S2 contrasts a positive harbor scene with a water-adjacent hard negative. Docked vessels aligned with pier or quay structures support the positive decision; the negative contains water and a vessel but lacks land-based port facilities. Water and vessel cues alone are therefore insufficient for a harbor decision.

The positive decision is supported by a conjunction rather than a single object: multiple vessels are berthed along linear pier or quay edges inside a bounded waterfront complex, with land-side buildings and service areas providing additional functional context. The hard negative deliberately preserves two tempting cues—open water and a vessel—while removing the land–water interface, docking geometry, basin organization, and port facilities needed for a defensible harbor interpretation.

This pair is an evidence audit of the Bridging-Conv curation rule, not an additional quantitative benchmark. During rewriting and verification, vessel or water keywords alone cannot license a harbor label. Positive records must retain visible relational support, whereas isolated ships, ambiguous coastlines, and low-information water scenes are assigned a negative or uncertainty-compatible response. This policy is designed to discourage shortcut reliance on frequent maritime nouns and to align the training language with the rejection behavior evaluated by HarborEval.

The comparison also makes the intended claim boundary explicit. The examples demonstrate how source records are screened for visually grounded harbor evidence; they do not establish that every port configuration must contain the same structures. Legitimate but atypical scenes may still require calibrated uncertainty when resolution, modality, occlusion, or crop boundaries hide decisive infrastructure.

S8 Qualitative Modality Examples

Figure S3 presents illustrative RGB, PAN, SAR, and NIR cases. These examples are not used as primary quantitative evidence; they show the modality-compatible observations rewarded by the diagnostic tracks, including functional-zone reporting, grid grounding, evidence-based VQA, and spatial-relation reasoning.

The RGB case illustrates the most semantically expressive setting. The response identifies a compact industrial harbor, a large vessel along the central quay, smaller vessels, dockside facilities, and storage buildings, and then organizes these visible elements into a combined berthing, service, and storage zone. Importantly, the answer remains at the level of visible structure and functional layout; it does not infer a named port, cargo category, operator, throughput, or current activity from appearance alone.

The PAN example instead emphasizes geometric localization. With spectral color unavailable, the answer relies on elongated bright structures, their alignment with the dockside, and proximity to a linear quay edge to select the supported grid cell. This is the intended role of the grid-grounding track: the prediction should reflect the spatial support of the target rather than generic confidence that vessels occur somewhere in the image. The example also shows why modality-blind prompts remain meaningful when the accepted evidence is defined geometrically.

The SAR and NIR cases expose two different evidence boundaries. In SAR, bright backscatter clusters distributed along the waterfront support the presence of dockside infrastructure and vessel-like targets beside a dark water surface, but they do not justify optical color or fine-grained appearance claims. In NIR, the response uses the coastal edge, adja-

P1 Closed-Form Recognition / Relation / Rejection

```
Goal: answer a structured harbor-diagnostic question from the image only.
Tracks: T1 object support; T2 dominant functional zone; T3 relation or geometry; T8
non-harbor rejection.
Inputs: {image}, {question}, {options when applicable}.
Rules: use visible evidence; ignore filenames, metadata, geography priors, and hidden labels;
choose all valid options only when the task is multi-select.
Output: FINAL = {option key(s) | Yes | No | Cannot determine}.
```

P2 Grid Grounding

```
Goal: localize the visually supported target without bounding boxes.
Grid: split the full image into a fixed 3x3 layout with standard cell names.
Inputs: {image}, {target or question}.
Rules: select every cell visibly occupied by the target; include multiple cells for spanning
objects; do not pad with neighboring cells for caution.
Output: FINAL = {comma-separated grid cells}.
```

P3 Evidence Decision and Modality Constraint

```
Goal: decide whether a claim is supported under the image modality.
Inputs: {image}, {modality-blind question}, {choices when applicable}.
Decision policy: Yes = clear positive evidence; No = clear contradiction or
image-incompatible claim; Cannot determine = plausible but insufficient evidence.
Forbid: cargo type, vessel identity, operator, ownership, exact location, time, throughput,
or operational status unless directly visible.
Output: FINAL = {Yes|No|Cannot determine}; EVIDENCE = {one concise visual reason}.
```

P4 Open Grounded Reporting

```
Goal: produce a concise report or short answer grounded in visible evidence.
Variants: T7 harbor description; T8 visible-scene report under rejection setting; Public
Harbor VQA; Public Harbor captioning.
Cover when defensible: major objects, functional zones, spatial relations,
modality-compatible cues, uncertainty, and missing evidence.
Avoid: named locations, identities, cargo, operators, precise operations, or overconfident
port interpretation when evidence is insufficient.
Output: DIRECT ANSWER = {grounded report or concise answer}.
```

Table S7: Semi-structured inference prompt templates used for HarborEval and Public Harbor. Backbone-specific chat wrappers add image tokens or conversation delimiters around the same semantic instructions.

cency between facilities and the waterfront, and orientation toward surrounding water to express a spatial relation. The relevant cue is therefore relational organization rather than natural-color appearance. Together, the two panels illustrate how the same harbor concept can be supported by different sensor-compatible observations.

Across all four panels, a successful answer combines a direct decision with a concise sensor-aware rationale.

The examples are included as an audit of response form and evidence use, not as cherry-picked substitutes for the quantitative results. They clarify the qualitative criterion applied throughout HarborEval and OpenEval: reward concise claims that can be connected to observable objects, regions, or relations, and penalize unsupported specificity even when the overall scene is plausibly maritime.

The panels deliberately use different task forms—captioning, grid grounding, evidence-based VQA, and relation reasoning—because cross-modal transfer should preserve a common grounding principle without forcing identical verbal detail from every sensor. No conclusion about relative modality difficulty or average performance is drawn from these four cases alone; those comparisons remain governed by the frozen benchmark scores and the modality-level diagnostics reported in the main paper.

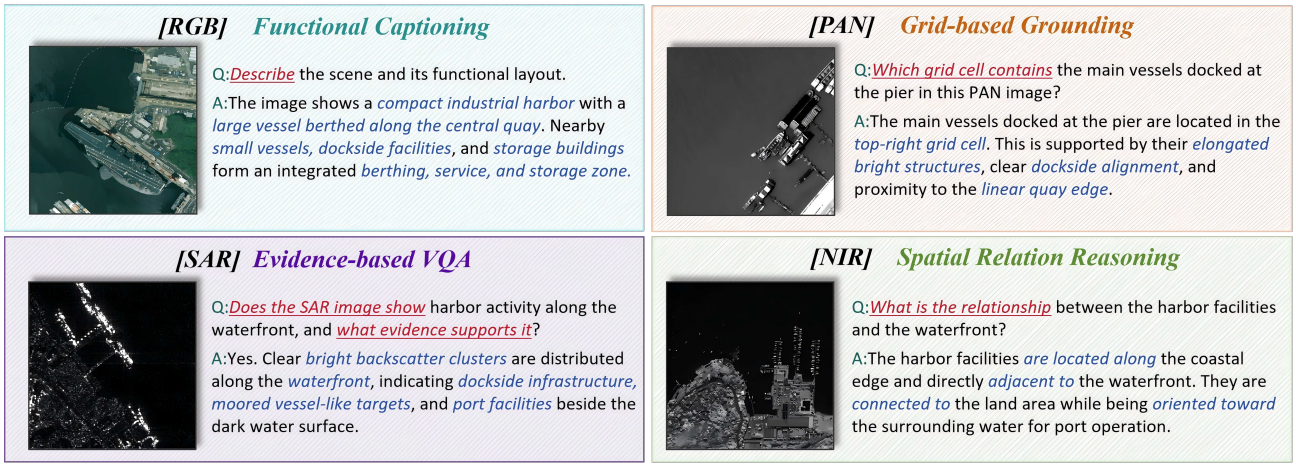

Figure S3: Qualitative examples of evidence-grounded harbor understanding across RGB, PAN, SAR, and NIR observations.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: A Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems 35*, 23716–23736.

Bai, S.; Cai, Y.; Chen, R.; Chen, K.; Chen, X.; Cheng, Z.; Deng, L.; Ding, W.; Gao, C.; Ge, C.; et al. 2025a. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025b. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum Learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 41–48.

ByteDance Seed Team. 2025. Seed1.8: A Generalized Agentic Model. Official model release. Accessed: 2026-07-14.

Chan, D.; Petryk, S.; Gonzalez, J.; Darrell, T.; and Canny, J. 2023. CLAIR: Evaluating Image Captions with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13638–13646.

Cheng, Q.; Huang, H.; Xu, Y.; Zhou, Y.; Li, H.; and Wang, Z. 2022. NWPU-Captions Dataset and MLCA-Net for Remote Sensing Image Captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–19.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-Purpose Vision-Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems 36*, 49250–49267.

Ge, J.; Zhang, X.; Zheng, Y.; Guo, K.; and Liang, J. 2025. RSTeller: Scaling up Visual Language Modeling in Remote Sensing with Rich Linguistic Semantics from Openly Available Data and Large Language Models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 226: 146–163.

Gururangan, S.; Marasovic, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8342–8360.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Hu, Y.; Yuan, J.; Wen, C.; Lu, X.; Liu, Y.; and Li, X. 2025. RSGPT: A Remote Sensing Vision Language Model and Benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 224: 272–286.

Huang, L.; Liu, B.; Li, B.; Guo, W.; Yu, W.; Zhang, Z.; and Yu, W. 2018. OpenSARShip: A Dataset Dedicated to Sentinel-1 Ship Interpretation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(1): 195–208.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming Catastrophic Forgetting in Neural Networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526.

Kuckreja, K.; Danish, M. S.; Naseer, M.; Das, A.; Khan, S.; and Khan, F. S. 2024. GeoChat: Grounded Large Vision-Language Model for Remote Sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27831–27840.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, 19730–19742.

Li, K.; Wan, G.; Cheng, G.; Meng, L.; and Han, J. 2020. Object Detection in Optical Remote Sensing Images: A Survey and A New Benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159: 296–307.

Li, X.; Ding, J.; and Elhoseiny, M. 2024. VRSBench: A Versatile Vision-Language Benchmark Dataset for Remote Sensing Image Understanding. In *Advances in Neural Information Processing Systems 37*, 3229–3242.

Li, X.; Wen, C.; Hu, Y.; Yuan, Z.; and Zhu, X. X. 2024. Vision-Language Models in Remote Sensing: Current Progress and Future Trends. *IEEE Geoscience and Remote Sensing Magazine*, 12(2): 32–66.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, X.; and Wen, J.-R. 2023b. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 292–305.

Li, Z.; Muhtar, D.; Gu, F.; He, Y.; Zhang, X.; Xiao, P.; He, G.; and Zhu, X. 2025. LHRS-Bot-Nova: Improved Multimodal Large

Language Model for Remote Sensing Vision-Language Interpretation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 227: 539–550.

Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; Ye, Q.; Fu, L.; and Zhou, J. 2024a. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–16.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024b. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26286–26296.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485*.

Lobry, S.; Marcos, D.; Murray, J.; and Tuia, D. 2020. RSVQA: Visual Question Answering for Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12): 8555–8566.

Lu, X.; Wang, B.; Zheng, X.; and Li, X. 2018. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4): 2183–2195.

Luo, J.; Pang, Z.; Zhang, Y.; Wang, T.; Wang, L.; Dang, B.; Lao, J.; Wang, J.; Chen, J.; Tan, Y.; and Li, Y. 2024. SkySenseGPT: A Fine-Grained Instruction Tuning Dataset and Model for Remote Sensing Vision-Language Understanding. *arXiv preprint arXiv:2406.10100*.

Muhtar, D.; Li, Z.; Gu, F.; Zhang, X.; and Xiao, P. 2024. LHRS-Bot: Empowering Remote Sensing with VGI-Enhanced Large Multimodal Language Model. In *Computer Vision – ECCV 2024*, 440–457. Springer Nature Switzerland.

Pang, C.; Weng, X.; Wu, J.; Li, J.; Liu, Y.; Sun, J.; Li, W.; Wang, S.; Feng, L.; Xia, G.-S.; and He, C. 2025. VHM: Versatile and Honest Vision Language Model for Remote Sensing Image Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6381–6388.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763.

Rolnick, D.; Ahuja, A.; Schwarz, J.; Lillicrap, T. P.; and Wayne, G. 2019. Experience Replay for Continual Learning. In *Advances in Neural Information Processing Systems 32*.

Schmitt, M.; Hughes, L. H.; Qiu, C.; and Zhu, X. X. 2019. SEN12MS: A Curated Dataset of Georeferenced Multi-Spectral Sentinel-1/2 Imagery for Deep Learning and Data Fusion. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume IV-2/W7, 153–160.

Soni, S.; Dudhane, A.; Debary, H.; Fiaz, M.; Munir, M. A.; Danish, M. S.; Fraccaro, P.; and Watson, C. D. 2025. EarthDial: Turning Multi-Sensory Earth Observations to Interactive Dialogues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14303–14313.

Sumbul, G.; Charfuelan, M.; Demir, B.; and Markl, V. 2019. BigEarthNet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding. In *2019 IEEE International Geoscience and Remote Sensing Symposium*, 5901–5904.

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9: 2579–2605.

Vivone, G. 2023. Multispectral and Hyperspectral Image Fusion in Remote Sensing: A Survey. *Information Fusion*, 89: 405–417.

Wang, S.; Zou, X.; Li, K.; Xing, J.; and Tao, P. 2023a. PanBench: Towards High-Resolution and High-Performance Pansharpening. *arXiv preprint arXiv:2311.12083*.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2023b. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 13484–13508.

Wang, Z.; Prabha, R.; Huang, T.; Wu, J.; and Rajagopal, R. 2024. SkyScript: A Large and Semantically Diverse Vision-Language Dataset for Remote Sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5805–5813.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2022. Finetuned Language Models Are Zero-Shot Learners. In *International Conference on Learning Representations*.

Wei, S.; Zeng, X.; Qu, Q.; Wang, M.; Su, H.; and Shi, J. 2020. HRSID: A High-Resolution SAR Images Dataset for Ship Detection and Instance Segmentation. *IEEE Access*, 8: 120234–120254.

Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2018. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3974–3983.

Yuan, Z.; Xiong, Z.; Mou, L.; and Zhu, X. X. 2025. ChatEarthNet: A Global-Scale Image-Text Dataset Empowering Vision-Language Geo-Foundation Models. *Earth System Science Data*, 17(3): 1245–1263.

Yuan, Z.; Zhang, W.; Fu, K.; Li, X.; Deng, C.; Wang, H.; and Sun, X. 2022. Exploring a Fine-Grained Multiscale Method for Cross-Modal Remote Sensing Image Retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–19.

Zhan, Y.; Xiong, Z.; and Yuan, Y. 2025. SkyEyeGPT: Unifying Remote Sensing Vision-Language Tasks via Instruction Tuning with Large Language Model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221: 64–77.

Zhang, W.; Cai, M.; Zhang, T.; Zhuang, Y.; and Mao, X. 2024a. EarthGPT: A Universal Multimodal Large Language Model for Multisensor Image Comprehension in Remote Sensing Domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–20.

Zhang, Y.-F.; Zhang, H.; Tian, H.; Fu, C.; Zhang, S.; Wu, J.; Li, F.; Wang, K.; Wen, Q.; Zhang, Z.; Wang, L.; Jin, R.; and Tan, T. 2025. MME-RealWorld: Could Your Multimodal LLM Challenge High-Resolution Real-World Scenarios that are Difficult for Humans? In *International Conference on Learning Representations*.

Zhang, Z.; Zhang, L.; Wang, Y.; Feng, P.; and He, R. 2021. ShipRSImageNet: A Large-Scale Fine-Grained Dataset for Ship Detection in High-Resolution Optical Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 8458–8472.

Zhang, Z.; Zhao, T.; Guo, Y.; and Yin, J. 2024b. RS5M and GeoRSCLIP: A Large-Scale Vision-Language Dataset and a Large Vision-Language Model for Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–23.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36*, 46595–46623.

Zhu, X. X.; Montazeri, S.; Ali, M.; Hua, Y.; Wang, Y.; Mou, L.; Shi, Y.; Xu, F.; and Bamler, R. 2021. Deep Learning Meets SAR: Concepts, Models, Pitfalls, and Perspectives. *IEEE Geoscience and Remote Sensing Magazine*, 9(4): 143–172.