SONARLLM: A NATIVE SONAR–OPTICAL MULTIMODAL LARGE LANGUAGE MODEL FOR UNDERWATER PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable underwater perception requires complementary sensing under variable visibility. Optical cameras capture appearance and semantics but degrade rapidly with turbidity, whereas imaging sonar preserves geometry while exhibiting distinct range–azimuth structure and acoustic artifacts. Existing MLLMs, built primarily on optical encoders, are therefore ill-suited to model sonar or adaptively exploit sonar–optical complementarity. We propose SonarLLM, a sonar–optical MLLM that treats sonar as a native perceptual modality. It combines a sonar-specific encoder, modality-specific physics-aware feature enhancement, and reliability-aware hierarchical fusion to align acoustic structure with optical semantics and dynamically adjust their contributions as sensing quality changes. We also introduce SonarBench, a paired benchmark that spans four tasks—recognition, counting, visual question answering, and captioning—and, across the benchmark, three input settings: sonar-only, optical-only, and fusion. By fixing the scene and sonar observation while varying optical degradation, SonarBench enables controlled measurement of cross-modal complementarity. SonarLLM achieves 72.0% macro accuracy across sonar-only recognition, counting, and VQA, outperforming the strongest baseline by 34.4 percentage points, and 68.7% under fusion, exceeding the best baseline by 25.1 points. For recognition and counting, the fusion-over-optical gain grows from 6.0 to 36.0 points as turbidity increases, indicating the increasing complementary value of sonar under controlled optical degradation. Together, these results show that robust heterogeneous perception depends not only on adding sonar, but on representing and weighting it according to its sensing characteristics.

1 INTRODUCTION

Light does not travel far underwater. Absorption and scattering progressively erase color, texture, contrast, and object boundaries as turbidity increases (Li et al., 2020)—precisely the cues that modern multimodal large language models (MLLMs) depend on. Underwater models such as MarineGPT (Zheng et al., 2023) and NAUTILUS (Xu et al., 2025) extend open-ended visual understanding to marine captioning, question answering, and recognition, yet they remain overwhelmingly optical: when the camera loses the scene, the model loses it too. Imaging sonar faces the opposite trade-off. Largely independent of illumination, it preserves object contours, spatial structure, and range information in the same poor visibility (Neupane & Seok, 2020), but its range–azimuth images, speckle, and acoustic shadows resemble nothing in any optical training set (Li et al., 2025). The two sensors are thus complementary—appearance and semantics on one side, stable geometry on the other—as Fig. 1 shows for a single scene: as the water turns turbid, the diver fades from every optical frame while the sonar return barely changes.

Exploiting this complementarity, however, requires more than adding sonar as another visual input; it poses three distinct obstacles. **Representation:** sonar–optical systems target task-specific detection or tracking (Li et al., 2025; Wu et al., 2026) rather than language reasoning, and natural-image encoders are poorly matched to sonar's range–azimuth geometry, speckle, reverberation, and range-dependent propagation loss (Neupane & Seok, 2020; Steiniger et al., 2022). **Degradation:** the two modalities fail through different physics—scattering for optics, reverberation and attenuation

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

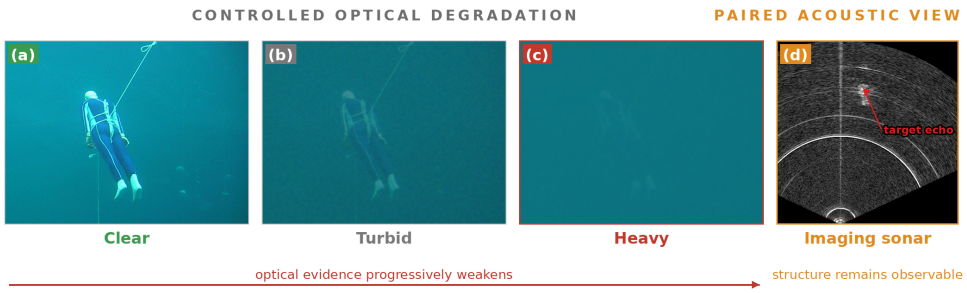


Figure 1: One scene under increasing turbidity. From left to right, the diver disappears from the optical record, while the sonar observation preserves the target echo and its range arcs.

for acoustics—so correction must be modality-specific. **Reliability:** neither sensor is trustworthy everywhere; optical evidence may collapse under turbidity while sonar is corrupted by acoustic noise, making fixed fusion vulnerable (Park et al., 2025). A sonar–optical MLLM must therefore solve representation, degradation, and reliability jointly.

We propose **SonarLLM**, a sonar–optical MLLM that treats imaging sonar as a native perceptual modality rather than an auxiliary image, with one mechanism answering each obstacle. For representation, SonarLLM retains the pretrained Qwen3-VL-8B optical encoder (Bai et al., 2025) and introduces an independent sonar encoder with a multi-scale Sonar Stem and range–azimuth positional encoding. For degradation, dedicated Visual Feature Enhancement (VFE) modules operate per modality: Optical-VFE targets scattering-related corruption, while Acoustic-VFE targets reverberation and range attenuation. For reliability, AGFM predicts quality-aware modality weights and dual-stream hierarchical DeepStack delivers the reweighted optical and sonar features to multiple language-model layers. A four-stage curriculum then establishes sonar-domain representations, cross-modal correspondence, reliability learning, and instruction following in turn.

We further introduce **SonarBench**, a paired benchmark spanning recognition, counting, VQA, and captioning under sonar-only, optical-only, and fusion settings, with optical inputs from clear to heavily turbid and captioning serving as a separate open-ended diagnostic. Its defining protocol is an intervention: the scene, the question, and the sonar observation stay fixed while only the optical image degrades—dimming the lights while holding the world constant. Any accuracy change is then attributable to sensing quality rather than sample composition, separating stronger unimodal modeling from genuinely complementary sonar evidence.

Three findings emerge: **72.0%** sonar-only macro accuracy across recognition, counting, and VQA—**34.4 points** above the strongest baseline, with substantially larger 27B–35B MLLMs not closing the gap, so scale is not the missing ingredient; **68.7%** under fusion, exceeding the best baseline by **25.1 points**; and a fusion-over-optical gain that grows from **6.0** to **36.0 points** as turbidity increases while fusion performance stays comparatively stable—the murkier the water, the more the second sensor is worth.

Contributions. Our contributions follow this arc: **SonarLLM**, which casts sonar–optical language understanding as heterogeneous sensing and pairs sonar-specific representation with per-modality enhancement, reliability-aware gating, and hierarchical interaction in a shared language model; **SonarBench**, whose controlled intervention—fixed scene, fixed sonar observation, degrading optics—isolates unimodal sonar capability from cross-modal complementarity; and evidence from controlled comparisons, representation analysis, and structural ablations that native sonar modeling provides capabilities not recovered by model scale or instruction tuning alone, while reliability-aware fusion grows more valuable as optical evidence deteriorates.

2 Related Work

On the language side, MarineGPT (Zheng et al., 2023) applies domain-specific instruction tuning to captioning, question answering, and recognition; AquaticCLIP (Alawode et al., 2025) learns underwater vision–language representations through contrastive pretraining; NAUTILUS (Xu et al., 2025)

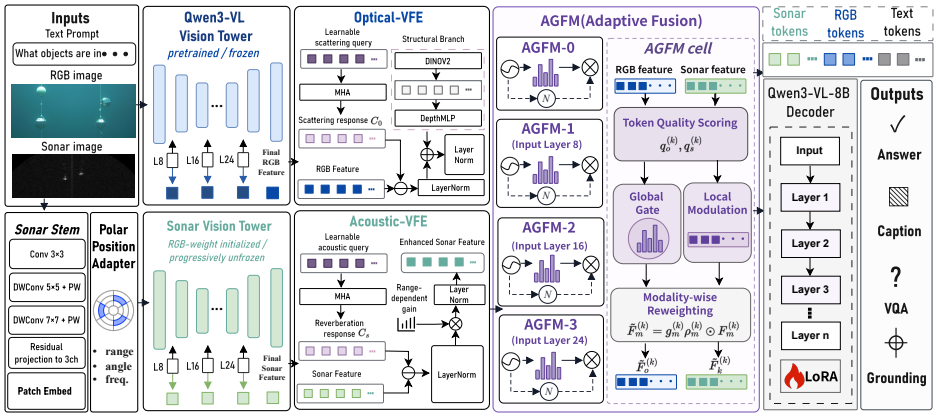


Figure 2: Overall architecture of SonarLLM, comprising sonar-native representation, modality-specific feature enhancement, reliability-aware hierarchical fusion, and progressive training.

incorporates physics-aware enhancement; OceanGPT (Bi et al., 2024) targets broader ocean-domain knowledge; and OceanPile (Xue et al., 2026) and OceanGym (Xue et al., 2025) provide ocean multimodal data and embodied evaluation. On the sonar side, research addresses detection, classification, segmentation, and tracking (Neupane & Seok, 2020; Steiniger et al., 2022), supported by UATD (Xie et al., 2022) and SCTD (Zhang et al., 2022), while paired datasets such as RGBS50 (Li et al., 2025), UMOD (Wu et al., 2026), and SOVIS (Chen et al., 2026) demonstrate sonar–optical fusion for specific discriminative tasks. Both lines stop short of this paper's question: connecting acoustic representations to open-ended language reasoning, and establishing when sonar contributes complementary evidence under controlled optical degradation.

General-purpose MLLMs connect visual and linguistic representations through diverse mechanisms: LLaVA (Liu et al., 2023b;a) uses learned projection, BLIP-2 (Li et al., 2023) a Q-Former, Flamingo (Alayrac et al., 2022) cross-layer attention, and Qwen3-VL (Bai et al., 2025) hierarchical visual features through DeepStack. Gated and reliability-aware fusion methods (Arevalo et al., 2017; Park et al., 2025) additionally adjust modality contributions according to sensor quality. Nevertheless, these approaches generally assume visually homogeneous inputs or ignore sensors with different imaging geometry, degradation processes, and environment-dependent reliability; SonarLLM instead combines modality-specific representation and enhancement with reliability-aware hierarchical interaction.

Existing underwater evaluation resources are similarly divided: UIEB (Li et al., 2020) and EUVP (Islam et al., 2020) focus on optical enhancement; UATD and SCTD support sonar perception; RGBS50, UMOD, and SOVIS target specific paired-sensor tasks; and NautData (Xu et al., 2025) and UWBench (Zhang et al., 2025) remain centered on optical vision-language evaluation, so none can isolate whether multimodal gains arise from stronger unimodal modeling or from complementary sonar evidence.

3 Method

3.1 Overall Architecture

SonarLLM instantiates these three mechanisms within a single language model. As shown in Fig. 2, given an optical image I_o, an imaging-sonar observation I_s, and a textual instruction x, it preserves the pretrained optical encoder $\mathcal{E}_o$ of Qwen3-VL-8B (Bai et al., 2025) and introduces an independent sonar encoder $\mathcal{E}_s$, treating sonar as a native perceptual modality rather than an auxiliary image.

Each visual encoder produces a final feature F_m and three intermediate features $\{F_m^{(k)}\}_{k=1}^{3}$ (extracted after blocks 8, 16, and 24; $m \in \{o, s\}$). Final features are enhanced by modality-specific Visual Feature Enhancement (VFE) modules, projected into the language space, and reweighted by AGFM_0; intermediate features bypass the VFEs, are reweighted by AGFM_k, and are injected into language-model layers 8, 16, and 24 through dual-stream DeepStack.

3.2 SONAR-NATIVE VISUAL REPRESENTATION

We initialize a Polar-aware Sonar Vision Transformer (PSVT) from the Qwen3-VL visual tower—retaining transferable visual priors despite the domain gap—and adapt it through a multi-scale Sonar Stem and explicit geometric positional encoding.

Multi-scale Sonar Stem. Before patch embedding, the Sonar Stem captures echoes, boundaries, and acoustic shadows at different receptive fields:

$$\Delta I_s = \mathrm{Conv}_{1\times1}\left[f_7\left(f_5\left(f_3(I_s)\right)\right)\right], \qquad \widetilde{I}_s = I_s + \gamma \Delta I_s, \tag{1}$$

where f_3, f_5, and f_7 denote convolutional transformations with different receptive fields. The zero-initialized coefficient γ gradually introduces sonar-specific structure without disrupting the transferred representation at initialization.

Range–Azimuth Positional Adaptation. For a sonar patch centered at range r_i and azimuth θ_j, we augment the transferred positional representation as:

$$z_{ij}^{s} = z_{ij}^{\mathrm{base}} + \lambda_p\left[\phi_r(r_i) + \phi_\theta(\theta_j)\right], \tag{2}$$

where ϕ_r and ϕ_θ encode physical range and azimuth, respectively, letting PSVT retain pretrained visual priors while adapting to sonar imaging geometry.

3.3 MODALITY-SPECIFIC FEATURE ENHANCEMENT

Optical and sonar observations degrade through different physics—absorption and backscattering for optics, spherical spreading and acoustic attenuation for sonar (image-formation abstraction in Appendix B). Motivated by these distinct attenuation and interference terms, separate Optical-VFE and Acoustic-VFE modules correct the final high-level features in feature space—rather than inverting the raw image-formation processes—while intermediate DeepStack features bypass the VFEs.

Optical-VFE. A learnable query e_b estimates feature corruption related to scattering in F_o, while a frozen DINOv2-L encoder (Oquab et al., 2023) provides a generic structural reference D_o:

$$\begin{aligned} &C_o = \mathrm{MHA}(e_b, F_o, F_o), \qquad F_o^c = \mathrm{LN}\left(F_o - W_b C_o\right), \\ &\overline{F}_o = \mathrm{LN}\left[F_o^c + \Psi_o(D_o)\right], \end{aligned} \tag{3}$$

where W_b is a learned channel projection and Ψ_o maps D_o to the token shape of F_o^c.

Acoustic-VFE. A learnable acoustic query e_r estimates a feature component associated with reverberation, while token range parameterizes learned compensation for range-dependent attenuation:

$$\begin{aligned} &C_s = \mathrm{MHA}(e_r, F_s, F_s), \qquad F_s^c[i] = F_s[i] - \sigma(g_r) W_s C_s, \\ &G(r_i) = \exp\left[2\log\left(\frac{r_i}{r_{\min}}\right) + 2\alpha_c(r_i - r_{\min})\right], \quad \overline{F}_s[i] = \mathrm{LN}\left(G(r_i) F_s^c[i]\right), \end{aligned} \tag{4}$$

where W_s is a learned channel projection, σ the sigmoid, and g_r a scalar gate; the learned coefficient $\alpha_c = \mathrm{softplus}(\widehat{\alpha}_c)$ is deliberately distinct from the physical α_{phy}, and using each token's physical range r_i retains the native range–azimuth geometry.

3.4 RELIABILITY-AWARE HIERARCHICAL FUSION

Observation reliability varies across environments and spatial regions, so SonarLLM combines the Adaptive Gated Fusion Module (AGFM) with dual-stream DeepStack to model global sensor reliability and local token quality at multiple levels. At level k (k=0 for the enhanced final representation, $k \in \{1, 2, 3\}$ for the features after blocks 8, 16, 24), let $F_m^{(k)} = \{f_{m,i}^{(k)}\}_{i=1}^{N_m}$; AGFM estimates token quality and aggregates it into modality-level reliability:

$$\begin{aligned} &q_{m,i}^{(k)} = h_m^{(k)}\left(\mathrm{LN}\left(f_{m,i}^{(k)}\right)\right), \qquad \bar{q}_m^{(k)} = N_m^{-1}\sum_{j=1}^{N_m} q_{m,j}^{(k)}, \\ &\overline{\mathbf{q}}^{(k)} = \left[\bar{q}_o^{(k)}, \bar{q}_s^{(k)}\right], \qquad \mathbf{g}^{(k)} = \mathrm{softmax}\left(\overline{\mathbf{q}}^{(k)}/\tau_k\right), \end{aligned} \tag{5}$$

Table 1: Progressive training schedule. CE denotes category cross-entropy.

Stage	Data/objective	Trainable modules
I	Unlabeled sonar / MAE	PSVT, decoder
II	Labeled sonar / CE	PSVT/Stem, classifier
III	Paired / $\mathcal{L}_{\text{align}}$	Sonar path, VFEs, AGFM
IV	Instructions / $\mathcal{L}_{\text{inst}}$	LoRA, interfaces

where $h_m^{(k)} : \mathbb{R}^{d_k} \to \mathbb{R}$ is a token-shared scalar scorer and $\mathbf{g}^{(k)} = [g_o^{(k)}, g_s^{(k)}]$ holds relative modality weights; each τ_k is learnable, lower-bounded at 0.05, initialized to 2.0.

AGFM further captures spatially non-uniform quality through a mean-preserving per-token modulation (Appendix B): the global factor $g_m^{(k)}$ allocates reliability across sensors, while the token-level factor $\rho_{m,i}^{(k)}$ redistributes importance without changing mean scale, and AGFM reduces to an identity mapping for single-modality input.

AGFM reweights rather than merges the two modality sequences: the reweighted optical and sonar tokens form the initial multimodal context at the final level, while at the three intermediate levels they are projected by modality-specific DeepStack mergers and injected into language layers 8, 16, and 24, respectively:

$$H^{(\ell_k)} \leftarrow H^{(\ell_k)} + \mathcal{J}_o^{(k)}\left[\mathcal{P}_o^{(k)}\left(\widetilde{F}_o^{(k)}\right)\right] + \mathcal{J}_s^{(k)}\left[\mathcal{P}_s^{(k)}\left(\widetilde{F}_s^{(k)}\right)\right], \tag{6}$$

where $\mathcal{P}_m^{(k)}$ and $\mathcal{J}_m^{(k)}$ are the modality-specific merger and injection operators; keeping the two streams separate before injection avoids premature compression of their heterogeneous representations.

3.5 Progressive Training Strategy

We train SonarLLM in four stages that progressively establish sonar-domain representations, acoustic semantics, cross-modal reliability, and language-level reasoning (Table 1). Direct joint optimization would require limited paired data to simultaneously resolve domain shift, semantic organization, sensor correspondence, gate calibration, and instruction following; the staged curriculum instead stabilizes acoustic representations first and then introduces cross-modal and language supervision, reducing interference among these heterogeneous objectives.

Stage I: Sonar-Domain Adaptation. PSVT is adapted by masked reconstruction on unlabeled sonar images, weighting masked patches by local variation so that low-response background does not dominate ($\mathcal{L}_{\text{adapt}}$; Appendix A); only the sonar pathway is optimized, and the decoder is discarded afterward.

Stage II: Acoustic Semantic Learning. Category supervision organizes the adapted representations by acoustic semantics ($\mathcal{L}_{\text{sem}}$); all other modules stay frozen.

Stage III: Cross-Modal Alignment and Reliability Learning. Synchronized sonar–optical pairs with stochastic optical degradation (clear pairs retained as anchors) combine global contrastive alignment (InfoNCE), hierarchical cosine alignment across levels, BCE gate supervision toward $\pi_s(\eta) = \text{clamp}(0.5 + 0.4\eta, 0, 1)$—shifting supervision toward sonar as degradation increases and making g_s a supervised degradation proxy rather than an emergent reliability estimate—and a negative-entropy term against early gate collapse ($\mathcal{L}_{\text{align}}$; Appendix A). Modality-missing examples preserve compatibility with unimodal inputs.

Stage IV: Multimodal Instruction Tuning. Finally, sonar-only, optical-only, and paired instructions are trained with an answer-only autoregressive loss over the LoRA parameters (Hu et al., 2022) and multimodal interfaces ($\mathcal{L}_{\text{inst}}$), transferring the learned representations to open-ended recognition, counting, question answering, and captioning.

4 EXPERIMENTS

Our evaluation asks three questions: *can scale or instruction tuning substitute for sonar-native representation*, tested against substantially larger MLLMs and a LoRA control trained on identical instruction data; *when does the second sensor pay off*, isolated by the paired degradation protocol and probed through AGFM's response; and *which components carry the gains*, traced by representation analysis and ablations (cost in Appendix F)? Unless otherwise specified, all models share task prompts, input protocols, and deterministic decoding.

4.1 EXPERIMENTAL SETUP

Training Data. Stage I pools approximately 98K candidate sonar images from RGBS50, UATD, SCTD, DeeperSense, FLC+FLS, and OceanGym (Xue et al., 2025), plus sonar-filtered frames from OceanInstruct and OceanInstruction (Xue et al., 2026); after filtering, byte-level deduplication, and brightness-balanced sampling, 40K unlabeled images remain. Stage II uses 23-class sonar annotations aggregated from these sources, Stage III uses synchronized RGBS50 pairs with online optical degradation, and Stage IV uses a 635K-sample instruction mixture (533K sonar-related, 102K optical; unimodal and paired inputs).

SonarBench. SonarBench evaluates recognition, counting, VQA, and captioning: the first three use sonar-only, optical-only, and fusion inputs, and captioning uses sonar-only and optical-only inputs. Optical and fusion settings span clear, turbid, and heavily turbid conditions, while the scene, question, and sonar observation stay fixed and only the paired RGB image changes—a controlled stress test of optical reliability rather than a simulation of natural turbidity. Table 5 (Appendix D) summarizes the structure.

Evaluation samples come from held-out RGBS50 and UMOD sequences (25 subsets, 150 QA instances each; zero training overlap confirmed by a sequence–frame audit, Appendix C). The three optical conditions are rendered by the same generator used for training degradation, isolating response to a known reliability shift rather than out-of-family generalization.

Metrics and Judging. Qwen3.8-Max (Qwen Team, 2026b) judges the semantic correctness of recognition and VQA outputs, counting uses integer parsing with exact match, and captioning reports the semantic GOOD rate with NLTK METEOR (Banerjee & Lavie, 2005) ($\times 100$) as a lexical diagnostic; macro averages cover the three accuracy tasks, captioning separately. Re-evaluating 300 outputs with Kimi-K3 (Kimi Team, 2026) yields 94.3% agreement (Cohen's $\kappa = 0.87$) with unchanged rankings; Appendix C gives the full protocol.

Baselines. We evaluate Qwen3-VL-8B, InternVL3.5-8B (Wang et al., 2025), and MiniCPM-V 4.5 (Yu et al., 2025), the larger Qwen3.6-35B-A3B and Qwen3.8-27B models (Qwen Team, 2026a;c), and the underwater-domain OceanGPT-o-7B and NAUTILUS-7B models. Qwen3-VL-8B+LoRA uses the same Stage-IV data and LoRA configuration as SonarLLM, testing whether instruction tuning alone explains the gains. For fusion, all baselines use their native multi-image interface with fixed modality order and labels.

Implementation Details. SonarLLM uses Qwen3-VL-8B with 448×448 inputs and freezes the language model in Stages I–III; full hyperparameters and the parameter accounting appear in Appendix C.

4.2 MAIN RESULTS

Native Sonar Understanding. The answer to the first question is unambiguous. SonarLLM obtains 72.0% sonar-only macro accuracy, exceeding the strongest baseline by 34.4 points. The gain is consistent across recognition (64.7%), counting (84.7%), and VQA (66.5%), with a caption GOOD rate of 33.3% versus 12.0% for the best baseline. Larger capacity does not close the gap—Qwen3.6-35B-A3B and Qwen3.8-27B reach only 28.9% and 29.1%, and underwater-domain models are substantially weaker—showing the need for sonar-native representation rather than scale or domain knowledge alone.

Table 2: SonarBench results (%; 25 subsets, n=150 each). Recognition/VQA use semantic judging; counting uses exact match; ‡ denotes caption GOOD rate; averages exclude captioning; best per row in bold.

Modality	Task (Condition)	Qwen3-VL -8B	Qwen3-VL -8B+LoRA	InternVL 3.5-8B	MiniCPM-V 4.5	NAUTILUS -7B	OceanGPT -o-7B	Qwen3.6 -35B-A3B	Qwen3.8 -27B	**Sonar LLM**
Sonar	Recognition	25.3	28.7	29.3	23.3	14.0	2.0	27.3	32.0	**64.7**
	Counting	40.7	51.3	33.3	47.3	32.7	14.0	35.3	32.7	**84.7**
	VQA	30.2	32.6	27.5	21.7	25.8	20.8	24.2	22.5	**66.5**
	Caption‡	8.0	4.0	7.3	4.7	12.0	5.3	10.0	4.7	**33.3**
Optical	Recognition (Clear)	41.3	38.7	38.7	36.0	34.7	10.7	44.0	40.0	**77.3**
	Recognition (Turbid)	37.3	36.0	36.0	35.3	36.0	12.7	40.7	38.7	**63.3**
	Recognition (Heavy)	38.0	35.3	33.3	29.3	30.7	9.3	34.7	39.3	**44.7**
	Counting (Clear)	71.3	69.3	60.7	66.0	59.3	40.7	**73.3**	72.7	66.0
	Counting (Turbid)	37.3	**39.3**	26.0	32.0	24.0	22.7	29.3	22.7	37.3
	Counting (Heavy)	8.0	16.0	6.7	10.0	6.0	10.0	10.0	6.0	**36.0**
	VQA (Clear)	40.9	37.0	37.0	41.1	32.6	22.9	38.7	36.8	**47.5**
	VQA (Turbid)	34.7	32.5	30.8	32.0	28.2	21.5	32.8	28.5	**44.7**
	VQA (Heavy)	28.1	28.2	26.3	22.9	20.7	20.3	24.1	24.7	**41.2**
	Caption (Clear)‡	23.3	26.7	25.3	22.0	18.7	16.0	26.7	22.0	**29.3**
	Caption (Turbid)‡	24.7	26.0	24.7	25.3	22.7	12.0	24.7	21.3	**28.7**
	Caption (Heavy)‡	**15.3**	9.3	10.0	10.7	14.0	6.0	13.3	12.0	14.7
Fusion	Recognition (Clear)	37.3	30.7	34.7	34.7	34.0	15.3	40.7	40.7	**83.3**
	Recognition (Turbid)	36.0	30.0	36.0	39.3	33.3	15.3	38.7	40.0	**76.7**
	Recognition (Heavy)	39.3	30.0	33.3	32.0	30.7	10.7	37.3	38.7	**80.7**
	Counting (Clear)	66.7	36.0	65.3	62.0	62.7	52.7	23.3	63.3	**72.0**
	Counting (Turbid)	44.7	36.0	59.3	51.3	63.3	52.0	25.3	61.3	**70.7**
	Counting (Heavy)	47.3	36.0	48.7	36.0	62.0	55.3	24.0	40.0	**72.0**
	VQA (Clear)	39.1	30.3	33.9	38.0	38.0	11.6	25.6	38.5	**54.6**
	VQA (Turbid)	33.7	31.5	34.4	26.9	28.9	15.5	24.7	36.1	**53.5**
	VQA (Heavy)	26.4	30.1	28.6	20.4	22.6	13.5	20.3	34.5	**55.2**
Sonar Average		32.1	37.5	30.0	30.8	24.2	12.3	28.9	29.1	**72.0**
Optical Average		37.4	36.9	32.8	33.8	30.2	19.0	36.4	34.4	**50.9**
Fusion Average		41.2	32.3	41.6	37.8	41.7	26.9	28.9	43.7	**68.7**

Optical and Fusion Performance. The second sensor pays off—but only for models that can represent it. SonarLLM achieves 50.9% on optical input and ranks first on every recognition and VQA subset; its main optical weakness is clear/turbid counting. With both sensors, SonarLLM reaches 68.7%, 25.1 points above the strongest baseline, with fusion recognition between 76.7% and 83.3% at every degradation level. Naive multi-image input is instead unreliable—Qwen3-VL-8B+LoRA drops from 36.9% optical accuracy to 32.3% under fusion, and Qwen3.6-35B-A3B from 36.4% to 28.9%—so dedicated heterogeneous representation and interaction are what unlock the second sensor.

Controlling Instruction Tuning and Sampling Uncertainty. Relative to the Qwen3-VL-8B+LoRA control, SonarLLM improves sonar, optical, and fusion accuracy by 34.4, 14.0, and 36.5 points, so instruction tuning alone does not reproduce the gains; 10,000-replication scene-clustered bootstrap intervals are $[28.1, 40.5]$, $[11.3, 16.7]$, and $[34.0, 39.0]$, all with positive lower bounds, quantifying test-set sampling uncertainty but not variation across independent training runs.

4.3 Controlled Degradation and Reliability-Aware Fusion

Complementarity, however, is not an assumption to be asserted but a property to be measured. Because optical and fusion evaluations share scenes, questions, and sonar observations, fusion minus optical accuracy provides a paired estimate of the benefit of adding sonar. Table 4 in Appendix D summarizes this controlled comparison: the fusion-over-optical gain expands from 6.0 to 36.0 points for recognition and counting and from 7.1 to 14.0 for VQA, while clear-to-heavy fusion changes stay within ± 2.6 points despite much larger optical declines. For recognition, fusion also exceeds the stronger individual sensor by 6.0, 12.0, and 16.0 points across clear, turbid, and heavy conditions, showing that sonar increasingly complements rather than merely replaces optical sensing.

We next probe AGFM's internal response on 150 independent RGBS50 scenes at five degradation levels (750 observations). Mean sonar weight rises monotonically with severity (Fig. 3(a): 0.491 to 0.795; Spearman $\rho = 0.749$; 95.3% of scenes non-decreasing). Because Stage III explicitly supervises the gate with degradation strength, we read this as verification of the intended internal response rather than emergent reliability estimation.

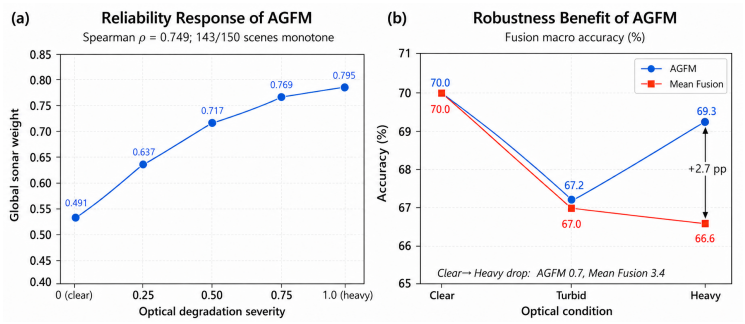


Figure 3: Reliability-aware fusion analysis on 150 held-out scenes. **(a)** Sonar weight increases with optical degradation. **(b)** Under the same checkpoint, AGFM's advantage over equal weighting grows from 0.0 to 2.7 points.

Replacing AGFM by equal weighting at inference, the accuracy advantage is 0.0/0.2/2.7 points under clear/turbid/heavy conditions, and clear-to-heavy degradation is 0.7 points with AGFM versus 3.4 with equal weighting. Complementarity is thus not an unconditional advantage of using more sensors: recognition benefits consistently because appearance and acoustic geometry provide distinct evidence, whereas fusion counting and VQA do not always exceed sonar-only performance when answers depend primarily on geometry or degraded optical cues distract. AGFM is therefore not an oracle that selects the best modality for every question but acts mainly as a degradation buffer: the sonar pathway determines which acoustic evidence reaches language reasoning, while AGFM controls its influence as reliability shifts. The controlled protocol does not cover non-uniform scattering, dynamic suspended particles, sonar-specific failures, or temporal misalignment; naturally degraded paired data remains necessary.

4.4 Ablation and Representation Analysis

The final question concerns provenance: which components carry the gains? Ablations use a diagnostic split disjoint from training and the main test set (25 modality–task–condition combinations, 150 samples each, N=3,750), scored by rules rather than the semantic judge of Table 2, so absolute values are not directly comparable.

Domain Adaptation and Representation Formation. Stage-I MAE with the same $r = 128$ adaptation improves the sonar R/C average from 72.0% to 74.7% and METEOR by 4.1 (Table 6, Appendix E); MAE-$r64$ reaches only 69.3%, so adapted features also need language-side capacity. Masking and data-scale sensitivity (Table 7) favor an intermediate ratio and more data; we use 0.60 masking with 40K images.

Figure 4 shows that the silhouette score of sonar categories rises from -0.18 under the pretrained optical tower to 0.16 after Stages I–II, yet matched and mismatched sonar–optical pairs remain nearly indistinguishable ($\Delta_{\text{align}} = 0.001$). Stage III raises the alignment margin to 0.660, supporting that cross-modal correspondence is learned beyond unimodal sonar structure; after Stage IV the margin remains substantial at 0.423, indicating that instruction tuning relaxes strict feature similarity while preserving task-relevant correspondence.

Structural Components. Table 3 identifies sonar geometry as the dominant factor: removing Polar PE costs 13.5/10.1 fusion/sonar-macro points and removing the Sonar Stem 8.2/7.5. The VFEs show the intended modality specificity—Optical-VFE mainly reduces optical accuracy and Acoustic-VFE mainly sonar-macro—and their joint removal causes consistent losses on all inputs, confirming complementary corrections; removing DeepStack costs 6.2 sonar-macro points of intermediate acoustic structure.

Quality-aware weighting matters most for robustness: Mean Fusion and w/o gate loss reduce static fusion accuracy by only 1.3/0.9 points but increase the clear-to-heavy decline from 1.3 to 3.8/3.0 points, supporting AGFM—with the same-checkpoint comparison in Fig. 3—as a quality-aware adaptation mechanism. The Sonar Stem and Polar PE variants require front-end retraining and are

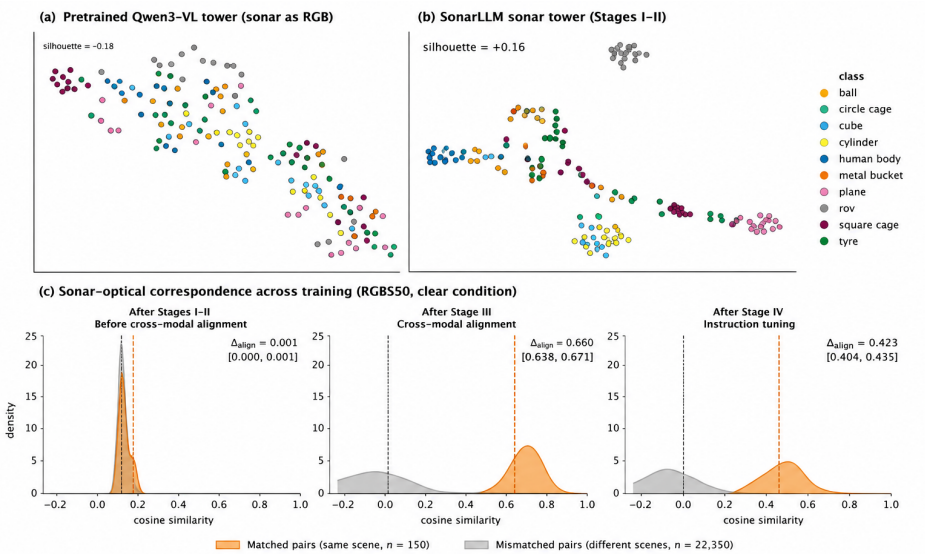


Figure 4: Progressive representation formation. **(a)** Sonar frames processed by the pretrained optical tower. **(b)** Class structure after Stages I–II. **(c)** Similarity of matched and mismatched sonar–optical pairs before and after alignment and instruction tuning. Panel (c) uses 150 matched same-scene pairs and 22,350 mismatched cross-scene pairs under the clear condition; intervals denote 95% confidence intervals for the mean-similarity margin.

Table 3: Structural ablation on the rule-scored validation split (%). Fusion, Optical, and Sonar Macro average recognition, counting, and VQA. *Drop*: clear-to-heavy fusion decline. †: sonar front-end retraining required. Mean Fusion is an independently retrained equal-weight variant.

Variant	Fusion	Drop	Optical	Sonar Macro
SonarLLM	**82.1**	**1.3**	**72.5**	**65.8**
Mean Fusion	80.8	3.8	72.3	65.6
w/o gate loss	81.2	3.0	72.7	64.9
w/o both VFE	79.8	3.1	67.2	61.4
w/o Optical-VFE	80.7	2.6	68.8	65.4
w/o Acoustic-VFE	80.4	2.4	72.1	62.0
w/o DeepStack	78.9	2.8	71.8	59.6
w/o Sonar Stem†	73.9	5.3	71.6	58.3
w/o Polar PE†	68.6	7.1	71.4	55.7

not strict inference-time ablations, but their consistent losses across sonar, fusion, and degradation sensitivity support their necessity.

5 Conclusion

SonarLLM treats sonar–optical language understanding as heterogeneous sensing, combining sonar-native representation with reliability-aware hierarchical fusion in a shared language model. On SonarBench it achieves 72.0% sonar-only and 68.7% fusion accuracy, with the fusion-over-optical gain rising from +6.0 to +36.0 points as visibility declines—gains attributable to domain-adapted sonar features and adaptive interaction rather than scale. The broader lesson mirrors the architecture: seeing through the murk is not always possible, but knowing which sensor to trust, at every level of representation, is. Future work will evaluate natural turbidity, temporal observations, broader sensors, and deployment robustness.

AI USE STATEMENT

In this work, we used generative AI tools to assist in implementing parts of the code. We have not used generative AI tools for generating synthetic datasets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on the research methodology or experiments, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results, and the remaining required-disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work: the AI-assisted code was checked by two authors to ensure its correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

SonarLLM is intended to support underwater perception research and applications such as environmental monitoring, infrastructure inspection, and robotic exploration. The same capability could, however, be repurposed for intrusive surveillance, military targeting, or autonomous decisions in safety-critical operations. Perception errors may also lead to incorrect identification or localization, creating risks for people, wildlife, equipment, and the environment. Deployment should therefore include task- and environment-specific validation, human oversight for consequential decisions, access controls, and operational fail-safes rather than treating model outputs as authoritative.

SonarBench draws on the research datasets listed in Section 4.1. Any release or redistribution should preserve the original licenses and provenance and should remove identifying or sensitive metadata where applicable. Because part of the evaluation relies on large-language-model judges, their decisions may inherit model biases or inconsistencies; the reported agreement checks and manual inspection reduce but do not eliminate this limitation.

REPRODUCIBILITY STATEMENT

The construction of SonarBench, including data sources, sequence-level partitioning, degradation rendering, metrics, and judge validation, is described in Section 4.1 and Appendix C. The SonarLLM architecture, its four progressive training stages with all loss functions, and implementation details such as input resolution, learning rates, and LoRA configuration are specified in Sections 3 and 4.1 and Appendices A and C; the ablation protocol is described in Section 4.4, and additional ablation details and efficiency measurements are provided in Appendices E and F.

REFERENCES

Basit Olakunle Alawode, Iyyakutti Iyappan Ganapathi, Sajid Javed, Naoufel Werghi, Mohammed Bennamoun, and Arif Mahmood. AquaticCLIP: A vision-language foundation model for underwater scene analysis, 2025.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 23716–23736, 2022.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. Gated multimodal units for information fusion. In *ICLR 2017 Workshop Track*, 2017.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report, 2025.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. OceanGPT: A large language model for ocean science tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3357–3372. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.184. URL https://aclanthology.org/2024.acl-long.184/.

Weitung Chen, Phil Tinn, Per Gunnar Auran, Martin Ludvigsen, and Peter Halland Haro. A sonar-visual dataset for cross-modal underwater robot perception, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Muhammad Jamil Islam, Youya Xia, and Junaed Sattar. Fast underwater image enhancement for improved visual perception. *IEEE Robotics and Automation Letters*, 5(2):3227–3234, 2020.

Kimi Team. Kimi K3: Open frontier intelligence, 2026.

Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, 29:4376–4389, 2020. doi: 10.1109/TIP.2019.2955241.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 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 (ICML)*, pp. 19730–19742. PMLR, 2023.

Yunfeng Li, Bo Wang, Jiuran Sun, Xueyi Wu, and Ye Li. Rgb-sonar tracking benchmark and spatial cross-attention transformer tracker. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(3):2260–2275, 2025. doi: 10.1109/TCSVT.2024.3497214.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023b.

Dhiraj Neupane and Jongwon Seok. A review on deep learning-based approaches for automatic sonar target recognition. *Electronics*, 9(11):1972, 2020. doi: 10.3390/electronics9111972. URL https://www.mdpi.com/2079-9292/9/11/1972.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2023.

Konyul Park, Yecheol Kim, Daehun Kim, and Jun Won Choi. Resilient sensor fusion under adverse sensor failures via multi-modal expert fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6720–6729, 2025.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all. Official model release, April 2026a. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b. Accessed: 2026-08-24.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork. Official model release, August 2026b. URL https://qwen.ai/blog?id=qwen3.8. Accessed: 2026-08-24.

Qwen Team. Qwen3.8-27B. Official model card, August 2026c. URL https://huggingface.co/Qwen/Qwen3.8-27B. Accessed: 2026-08-24.

Yannik Steiniger, Dieter Kraus, and Tobias Meisen. Survey on deep learning based computer vision for sonar imagery. *Engineering Applications of Artificial Intelligence*, 114:105157, 2022.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025.

Yujie Wu, Wenling Wang, Cong Lin, Mingxin Hou, and Mingxin Liu. Towards multimodal underwater object detection: A bidirectional feature recomposition network and visual-sonar dataset. *Expert Systems with Applications*, 316:131710, 2026. doi: 10.1016/j.eswa.2026.131710.

Kaibing Xie, Jian Yang, and Kang Qiu. A dataset with multibeam forward-looking sonar for underwater object detection. *Scientific Data*, 9(1):739, 2022. doi: 10.1038/s41597-022-01854-w. URL https://www.nature.com/articles/s41597-022-01854-w.

Wei Xu, Cheng Wang, Dingkang Liang, Zongchuang Zhao, Xingyu Jiang, Peng Zhang, and Xiang Bai. NAUTILUS: A large multimodal model for underwater scene understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Yida Xue, Mingjun Mao, Xiangyuan Ru, Yuqi Zhu, Baochang Ren, Shuofei Qiao, Mengru Wang, Shumin Deng, Xinyu An, Ningyu Zhang, Ying Chen, and Huajun Chen. OceanGym: A benchmark environment for underwater embodied agents, 2025.

Yida Xue, Ningyu Zhang, Tingwei Wu, Zhe Ma, Daxiong Ji, Zhao Wang, Guozhou Zheng, and Huajun Chen. OceanPile: A large-scale multimodal ocean corpus for foundation models, 2026.

Tianyu Yu, Zefan Wang, Chongyi Wang, et al. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe, 2025.

Da Zhang, Chenggang Rong, Bingyu Li, Feiyu Wang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. UWBench: A comprehensive vision-language benchmark for underwater understanding, 2025.

Peng Zhang, Jinsong Tang, Heping Zhong, Mingqiang Ning, Dandan Liu, and Ke Wu. Self-trained target detection of radar and sonar images using automatic deep learning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:4701914, 2022. doi: 10.1109/TGRS.2021.3096011.

Ziqiang Zheng, Jipeng Zhang, Tuan-Anh Vu, Shizhe Diao, Yue Him Wong Tim, and Sai-Kit Yeung. MarineGPT: Unlocking secrets of ocean to the public, 2023.

A Training Objectives

This section gives the full training objectives of the four progressive stages described in Section 3.5.

Stage I: Sonar-Domain Adaptation. Because low-response background occupies a large portion of each sonar frame, normalized masked patches $\bar{x}_i$ are weighted by their local variation w_i:

$$\mathcal{L}_{\text{adapt}} = \frac{\sum_{i\in\mathcal{M}} w_i \left\|\mathcal{D}(z_i) - \bar{x}_i\right\|_2^2}{\sum_{i\in\mathcal{M}} w_i}, \tag{7}$$

where $\bar{x}_i = (x_i - \mu(x_i))/(\sigma(x_i) + \varepsilon)$ and $w_i = \text{clip}(\sigma(x_i), \sigma_{\min}, \sigma_{\max})$.

Stage II: Acoustic Semantic Learning. Category supervision uses

$$\mathcal{L}_{\text{sem}} = -\frac{1}{B}\sum_{n=1}^{B}\sum_{c=1}^{C} y_{n,c} \log p_{s,c}^{(n)}, \qquad p_s^{(n)} = \text{softmax}\left(W_c h_s^{(n)} + b_c\right), \tag{8}$$

where $h_s^{(n)} = \text{Pool}[\mathcal{E}_s(\widetilde{I}_s^{(n)})]$.

Stage III: Cross-Modal Alignment and Reliability Learning. Training combines global contrastive alignment, hierarchical feature alignment, BCE-based gate supervision, and gate regularization:

$$\begin{aligned}
&\mathcal{L}_{\mathrm{NCE}} = \mathrm{InfoNCE}(z_o, z_s), \qquad \mathcal{L}_{\mathrm{hier}} = \sum_{k=1}^{3} \left[1 - \cos\left(z_o^{(k)}, z_s^{(k)}\right)\right], \\
&\mathcal{L}_{\mathrm{gate}} = \frac{1}{4}\sum_{k=0}^{3} \mathrm{BCE}\left(g_s^{(k)}, \pi_s(\eta)\right), \qquad \mathcal{L}_{\mathrm{ent}} = -\frac{1}{4}\sum_{k=0}^{3} \mathcal{H}_b\left(g_s^{(k)}\right), \\
&\mathcal{L}_{\mathrm{align}} = \mathcal{L}_{\mathrm{NCE}} + 0.5\mathcal{L}_{\mathrm{hier}} + 0.2\mathcal{L}_{\mathrm{gate}} + 0.1\mathcal{L}_{\mathrm{ent}}.
\end{aligned} \tag{9}$$

Here, z_m and $z_m^{(k)}$ are pooled final and intermediate representations; InfoNCE uses temperature 0.07, and $\mathcal{H}_b$ denotes binary entropy.

Stage IV: Multimodal Instruction Tuning. For visual input $\mathcal{I} \in \{I_s, I_o, (I_s, I_o)\}$, prompt x, and target answer y, we minimize the answer-only autoregressive objective:

$$\mathcal{L}_{\mathrm{inst}} = -\frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}} \log p_\Theta\left(y_t \mid x, \mathcal{I}, y_{<t}\right), \tag{10}$$

where $\mathcal{A}$ contains answer-token positions and Θ includes the trainable LoRA parameters and multimodal interfaces.

B Additional Architectural Details

Image-Formation Abstraction. Optical and sonar observations undergo different physical degradations. A compact abstraction of their image-formation processes is:

$$\begin{aligned}
I_o(u) &= J(u)e^{-\beta d(u)} + B_\infty\left[1 - e^{-\beta d(u)}\right], \\
S(r,\theta) &= \frac{S_0 TS(\theta)}{r^2} e^{-2\alpha_{\mathrm{phy}} r} + R(r,\theta),
\end{aligned} \tag{11}$$

where $d(u)$ is scene range, β optical attenuation, $J(u)$ undegraded scene radiance, B_∞ backscattered light, S_0 a source-level constant, α_{phy} acoustic attenuation, $TS(\theta)$ target strength, and $R(r,\theta)$ reverberation.

Local Token Modulation. Beyond the global gate, AGFM captures spatially non-uniform quality through mean-preserving token modulation:

$$\begin{aligned}
u_{m,i}^{(k)} &= \frac{1 + \sigma\left(q_{m,i}^{(k)}\right)}{2}, \qquad \bar{u}_m^{(k)} = N_m^{-1}\sum_{j=1}^{N_m} u_{m,j}^{(k)}, \\
\rho_{m,i}^{(k)} &= \frac{u_{m,i}^{(k)}}{\bar{u}_m^{(k)}}, \qquad \widetilde{f}_{m,i}^{(k)} = g_m^{(k)}\rho_{m,i}^{(k)} f_{m,i}^{(k)}.
\end{aligned} \tag{12}$$

The token-level factor $\rho_{m,i}^{(k)}$ redistributes importance among tokens without changing their mean scale.

C Additional Evaluation and Implementation Details

Metrics and Judging. Qwen3.8-Max judges the semantic correctness of recognition and VQA outputs against reference answers. Counting uses integer parsing and exact match; unparseable outputs (0.9%) are incorrect. VQA equally macro-averages attribute, existence, and counting scores, so its values need not be integer multiples of $1/150$; reported improvements are computed from unrounded aggregate scores. To assess possible same-family judge bias, we independently re-evaluate a stratified set of 300 recognition/VQA outputs with Kimi-K3; the judges achieve 94.3% agreement and Cohen's $\kappa = 0.87$, and all model rankings remain unchanged after manual inspection of disagreements.

Implementation Details. Stage I uses a 0.60 masking ratio. Stage II optimizes PSVT and the Sonar Stem at a learning rate of 5×10^{-5}. Stage III degrades the optical input with probability 0.7, samples $\eta \sim \mathcal{U}(0.1, 1.0)$ for degraded examples, and jointly trains the sonar pathway, both VFEs, and AGFM at a learning rate of 10^{-4}. Stage IV uses LoRA with $r = 128$, $\alpha = 256$, and learning rate 10^{-4}. The architecture and Stage-IV LoRA adapters add 882.4M and 349.2M parameters (10.3B total vs. 8.77B for the backbone, including the frozen DINOv2-L structural-prior encoder).

Degradation Rendering. The three optical conditions provide a controlled contrast in observation quality. Clear images remain unchanged; turbid images combine reduced brightness, additive white noise, colored veiling, and Gaussian blur; and heavy-turbid images retain the same photometric attenuation while replacing white noise with multi-scale correlated disturbances.

Evaluation Composition. The 25 evaluation subsets comprise 4 sonar-only, 12 optical-only, and 9 fusion, each with 150 QA instances. An underlying sequence–frame may recur across tasks, modalities, and degradation conditions to support paired comparisons. Source sequences are partitioned before all training stages, and the sequence–frame audit confirms zero overlap between training and evaluation.

D Additional Analysis Results

Table 4: Controlled degradation summary derived from Table 2 (percentage points). Δ(F–O) is fusion minus optical accuracy; C/T/H denote clear/turbid/heavy conditions.

Task	Δ(F–O) C/T/H	Optical C→H	Fusion C→H
Recognition	$+6.0/+13.4/+36.0$	-32.6	-2.6
Counting	$+6.0/+33.4/+36.0$	-30.0	0.0
VQA	$+7.1/+8.8/+14.0$	-6.3	$+0.6$

Table 4 supports the controlled degradation analysis discussed in Section 4.3.

Table 5: SonarBench evaluation structure. S/O/F denote sonar, optical, and fusion inputs; C/T/H denote clear, turbid, and heavy optical conditions.

Task	Input conditions	Metric	Subsets
Recognition	S; O/F×C/T/H	Sem. accuracy	7
Counting	S; O/F×C/T/H	Exact accuracy	7
VQA	S; O/F×C/T/H	Sem. accuracy	7
Captioning	S; O×C/T/H	GOOD / METEOR	4
Total			25

Table 5 summarizes the SonarBench evaluation structure described in Section 4.1.

E Additional Ablation Details

Table 6: Sonar front-end adaptation on the rule-scored validation split. R/C avg. is the mean of recognition and counting accuracy; METEOR is reported in %.

Variant	R/C avg.	Rec.	Count	METEOR
Transplant (no MAE)	72.0	64.0	**80.0**	45.1
MAE-$r64$	69.3	61.3	77.3	46.8
MAE-$r128$	**74.7**	**71.3**	78.0	**49.2**

Table 6 details the sonar front-end adaptation. In addition to the summary in Section 4.4, the transplant variant without Stage-I MAE retains the strongest counting accuracy (80.0%) but lags in recog-

nition and METEOR, indicating that reconstruction-based adaptation mainly strengthens semantic organization of sonar features rather than low-level appearance fidelity.

Table 7: Sensitivity of the sonar R/C average (%) to Stage-I masking and data scale.

Factor	Tested values	R/C avg.
Mask ratio	0.50 / 0.60 / 0.75	73.3 / **74.7** / 69.3
Unlabeled images	10K / 20K / 40K	70.0 / 72.7 / **74.7**

Table 7 reports the full sensitivity grid for Stage-I masking ratio and unlabeled data scale summarized in Section 4.4.

F Inference Efficiency

Table 8: Inference efficiency on one A100-80GB GPU. Results use BF16, FlashAttention-2, batch size 1, two 448×448 images, 512 visual tokens, and 128 generated tokens; SonarLLM latency and throughput are reported as mean$\pm$std over 10 runs.

Model	Params	Peak mem.	Prefill (ms)	Decode (tok/s)
Qwen3-VL-8B	8.77B	17.7GB	77.2 ± 0.2	38.4 ± 0.1
SonarLLM	10.3B	21.6GB	137.5 ± 1.1	21.1 ± 1.2

Table 8 shows the cost of native sonar processing: parameters and peak memory rise by 17.4% and 22.0%, prefill latency by 78.1%, and decoding throughput falls by 45.1%. Most added capacity is inherited: the 736.9M sonar tower is pretrained and frozen DINOv2-L contributes 0.30B parameters; only 25.1M non-LoRA adaptation and fusion parameters are randomly initialized.