

ENC-Bench: A Benchmark for Evaluating Multimodal Large Language Models in Electronic Navigational Chart Understanding

Ao Cheng¹ Xingming Li¹ Xuanyu Ji¹ Xixiang He¹ Qiyao Sun¹

Chunping Qiu² Runke Huang³ Qingyong Hu^{2*}

¹National University of Defense Technology ²Intelligent Game and Decision Lab

³The Chinese University of Hong Kong

Abstract

*Electronic Navigational Charts (ENCs) are the safety-critical backbone of modern maritime navigation, yet it remains unclear whether multimodal large language models (MLLMs) can reliably interpret them. Unlike natural images or conventional charts, ENCs encode regulations, bathymetry, and route constraints via standardized vector symbols, scale-dependent rendering, and precise geometric structure—requiring specialized maritime expertise for interpretation. We introduce **ENC-Bench**, the first benchmark dedicated to professional ENC understanding. ENC-Bench contains 20,490 expert-validated samples from 840 authentic National Oceanic and Atmospheric Administration (NOAA) ENCs, organized into a three-level hierarchy: **Perception** (symbol and feature recognition), **Spatial Reasoning** (coordinate localization, bearing, distance), and **Maritime Decision-Making** (route legality, safety assessment, emergency planning under multiple constraints). All samples are generated from raw S-57 data through a calibrated vector-to-image pipeline with automated consistency checks and expert review. We evaluate **10** state-of-the-art MLLMs such as GPT-4o, Gemini 2.5, Qwen3-VL, InternVL-3, and GLM-4.5V, under a unified zero-shot protocol. The best model achieves only 47.88% accuracy, with systematic challenges in symbolic grounding, spatial computation, multi-constraint reasoning, and robustness to lighting and scale variations. By establishing the first rigorous ENC benchmark, we open a new research frontier at the intersection of specialized symbolic reasoning and safety-critical AI, providing essential infrastructure for advancing MLLMs toward professional maritime applications.*

1. Introduction

Maritime transportation underpins over 90% of global trade, yet over 26,000 marine casualties occurred in EU waters

between 2014 and 2023 [1, 15], with catastrophic incidents exceeding USD 400 million in losses [8]. **Electronic Navigational Charts** (ENCs) are now mandatory for commercial vessels under International Maritime Organization regulations, with paper charts being phased out by 2030 [20, 43]. As maritime AI reaches 4.13 billion in market value with demonstrated potential to reduce collision risks by 75% [34, 46], reliable AI interpretation of ENCs becomes critical for next-generation safety systems.

However, ENCs represent a **highly specialized vertical domain** that fundamentally differs from general visual content [4, 21, 47]. Unlike natural images or statistical charts, ENCs encode safety-critical information through three distinctive characteristics that demand professional maritime expertise: (1) **standardized symbolic systems** (International Hydrographic Organisation, IHO S-57) [25] where regulatory information, bathymetric features, and navigational aids are represented through formal vector symbols with precise semantic meanings; (2) **scale-dependent multi-layer rendering** where cartographic features dynamically appear, disappear, or transform across zoom levels following cartographic generalization principles; and (3) **multi-constraint spatial geometry** requiring simultaneous evaluation of water depth, vessel draft, restricted zones, traffic separation schemes, and environmental conditions. This domain specificity necessitates systematic evaluation of whether AI systems possess the requisite symbolic grounding and structured reasoning capabilities.

Multimodal Large Language Models (MLLMs) [10, 24, 41, 54, 56, 60] have demonstrated remarkable capabilities in general visual understanding tasks, excelling on benchmarks spanning everyday photographs, common objects, and natural scenes [17, 18, 23, 30, 37, 51, 66]. Yet research consistently reveals a critical limitation: models trained predominantly on general-domain visual data struggle when transferred to specialized domains requiring domain-specific symbolic systems and structured spatial reasoning [5, 11, 49, 58, 63, 65]. This raises a fundamental question: *Can current MLLMs bridge the gap between gen-*

*Corresponding author.

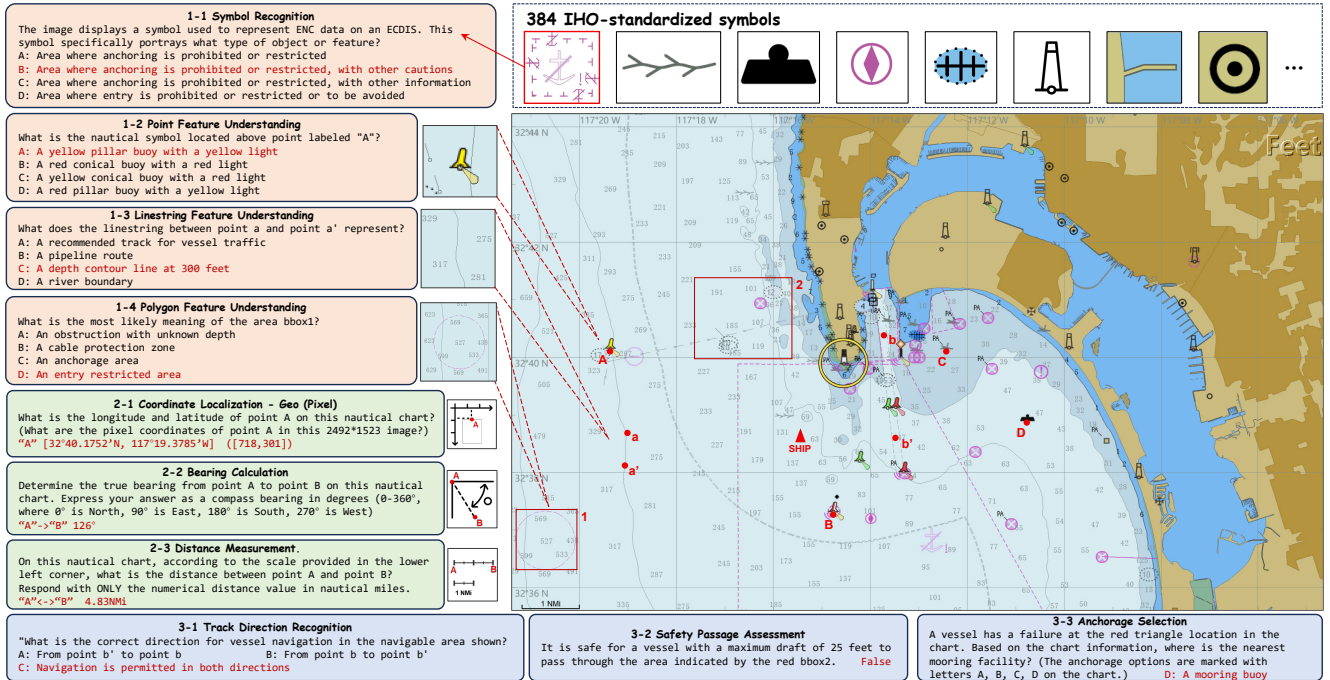


Figure 1. **Overview of ENC-Bench.** Our benchmark evaluates MLLMs across three hierarchical tiers: Perception (L-1), Spatial Reasoning (L-2), and Maritime Decision-Making (L-3). Example tasks are shown on the left with corresponding visual elements on an authentic NOAA chart (right), demonstrating the progression from basic symbol interpretation to complex multi-constraint decision-making required in professional maritime navigation.

eral visual understanding and the symbolic, structured, and safety-critical domain of ENC’s

Unfortunately, existing benchmarks fail to tackle this issue. Chart understanding benchmarks [22, 27, 38, 40, 59] focus on statistical plots with unstructured visual elements rather than formal geospatial symbology. Document benchmarks [12, 33, 39] emphasize text extraction and layout understanding over geometric spatial reasoning. Geographic reasoning benchmarks [50, 52, 62] explore location identification but lack the intersection of **standardized symbolic interpretation**, **multi-scale cartographic reasoning**, and **multi-constraint safety decision-making** that defines professional maritime navigation. No benchmark systematically evaluates whether MLLMs possess the specialized cognitive faculties required for ENC comprehension—a critical capability gap as AI systems increasingly integrate into maritime safety infrastructure.

To bridge this gap, we introduce **ENC-Bench**, the first comprehensive benchmark for evaluating MLLM capabilities in ENC’s understanding. Designed to mirror the cognitive pipeline of certified maritime navigators—from symbol recognition to safety-critical decision-making under multiple constraints—ENC-Bench provides a rigorous testbed for assessing model readiness in specialized, high-stakes visual domains. Our contributions are:

- **Professional-Grade Dataset.** We construct 20,490 sam-

ples from 840 authentic NOAA charts conforming to IHO S-57 standards. Each sample undergoes rigorous quality control through automated consistency validation and systematic expert review, ensuring alignment with professional maritime practice and safety protocols.

- **Hierarchical Evaluation Framework.** We design a three-tier assessment pipeline (Perception → Spatial Reasoning → Maritime Decision-Making) that decomposes ENC understanding from basic symbol recognition through coordinate-based geometric reasoning to complex, multi-constraint safety judgments—mirroring the cognitive hierarchy required in maritime navigation.
- **Comprehensive Evaluation and Analysis.** We conduct extensive zero-shot evaluation of **10** state-of-the-art MLLMs including GPT-4o, Gemini 2.5, Qwen3-VL, GLM-4.5V, InternVL-3, and Llama-4, revealing severe capability gaps: the best model achieves only 47.88% accuracy. Through in-depth error analysis, we identify three fundamental limitations—*symbolic grounding bottleneck* (failure to interpret formal notation like coordinate grids and scale bars), *multi-constraint reasoning deficiency* (greedy local optimization rather than explicit global constraint satisfaction), and *lack of robustness* across lighting modes and scale variations—providing concrete directions for advancing vision-language models toward deployment readiness in professional domains.

Table 1. Comparison of **ENC-Bench** with existing benchmarks. ✓, ▲, and ✗ separately represent **full support** (supports the core capability), **partial support** (e.g., informal/math symbols, Euclidean/layout reasoning, or multi-resolution), and **no support**.

Benchmark	Domain	Standardized Symbol Recognition	Precise Geospatial Reasoning (Numerical)	Multi-Scale (Cartographic)	Multi-Light (Rendered Modes)
MME [16]	General	✗	✗	✗	✗
MMBench [32]	General	✗	✗	✗	✗
MMMU [64]	General	▲	▲	✗	✗
DocVQA [39]	Chart/Doc	✗	▲	✗	✗
ChartQA [38]	Chart/Doc	✗	✓	✗	✗
MathVista [36]	Geometry/Math	▲	✓	✗	✗
GeoQA [6]	Geometry/Math	▲	✓	✗	✗
MapQA [13]	Map	✗	▲	✗	✗
MapEval [14]	Map	✗	✓	✗	✗
RSVQA [35]	Remote Sensing	✗	▲	▲	✗
SkyScript [57]	Remote Sensing	✗	✗	▲	✗
ENC-Bench (Ours)	<i>Maritime</i>	✓	✓	✓	✓

2. Related Work

Multimodal LLMs and General Visual Understanding.

Recent MLLMs [2, 9, 31, 44, 53, 60] have achieved remarkable success on general visual understanding benchmarks such as MME [16], MMBench [32], and MMMU [64], which evaluate perception and reasoning over natural images and everyday scenes [17, 23, 30]. However, research consistently reveals their limitations when transferred to specialized professional domains requiring formal symbolic systems and domain-specific expertise [49, 58, 65].

Chart and Document Understanding. Benchmarks including DocVQA [39], ChartQA [38], and InfoVQA [40] assess information extraction from business documents and statistical visualizations. Specialized models such as Pix2Struct [29] and ChartLlama [19] advance these capabilities through targeted training. However, these works focus on *pixel-rendered data visualizations* where values are encoded through informal visual geometry (bar heights, pie angles)—not *vector-based standardized symbology* governed by international regulatory frameworks encoding legal constraints and safety-critical information.

Geospatial Visual Understanding. Remote sensing benchmarks (RSVQA [35], SkyScript [57], LHRBench [42]) evaluate satellite imagery interpretation, while map understanding benchmarks (MapQA [13], MapEval [14]) assess reasoning on consumer web maps. Domain-adapted models like GeoChat [28] and RS-LLaVA [3] demonstrate improved performance through specialized pretraining. Yet these works evaluate either *observational photographic data* or *informal consumer cartography* with approximate spatial representations—neither addresses *safety-certified professional navigation charts* em-

ploying standardized symbology (IHO S-57), legally mandated positional accuracy, and regulatory compliance for international maritime operations.

Structured Visual Reasoning. Geometry-focused benchmarks including GeoQA [6], UniGeo [7], and MathVista [36] evaluate mathematical reasoning over educational diagrams and abstract geometric problems. While these assess structured reasoning capabilities, they operate within *idealized Euclidean coordinate systems* suitable for textbook scenarios—distinct from *spherical geodetic coordinate systems* required for real-world nautical computations involving Haversine distance, meridian convergence, and DMS notation critical for maritime route planning.

Our Positioning: Safety-Critical Symbolic Reasoning in Maritime Navigation. Existing benchmarks (Table 1) evaluate general vision, informal visualizations, geospatial maps, or abstract geometry—none address professional maritime navigation charts requiring regulatory compliance. ENC-Bench uniquely combines four capabilities essential for safety-critical ENC interpretation: (1) *Standardized Symbolic Recognition*—IHO S-57 regulated symbology encoding legal constraints; (2) *Precise Geospatial Reasoning*—exact coordinate computation with nautical accuracy requirements; (3) *Multi-Scale Cartographic Rendering*—scale-dependent feature aggregation following professional cartographic principles; and (4) *Multi-Lighting Operational Modes*—robustness across day/dusk/night rendering modes. By simultaneously evaluating these dimensions in authentic navigational contexts where interpretation errors lead to maritime casualties, ENC-Bench establishes essential infrastructure for advancing MLLMs from general-purpose understanding toward deployment in high-stakes professional domains.

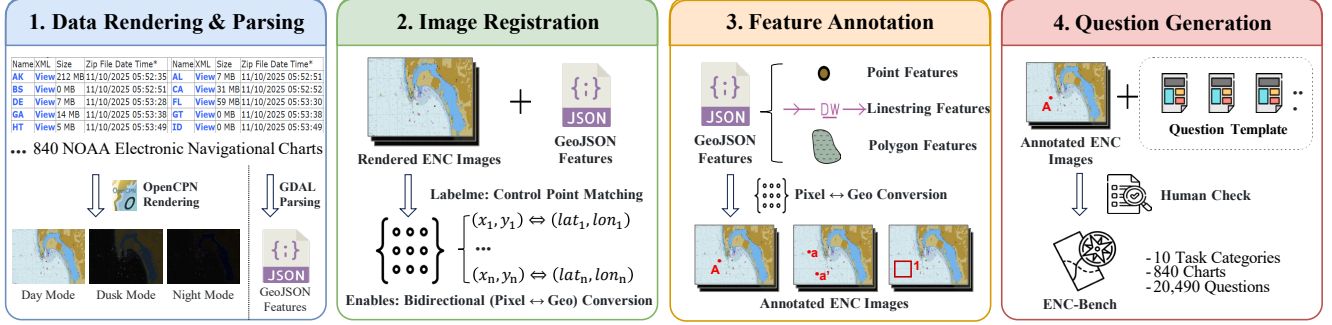


Figure 2. **ENC-Bench Data Generation Pipeline.** Four-stage process transforms 840 NOAA S-57 charts into 20,490 validated samples: (1) **Rendering & Parsing** produces multi-condition images and extracts GeoJSON features; (2) **Image Registration** establishes pixel-to-geo coordinate conversion; (3) **Feature Annotation** marks point/line/polygon features with expert verification; (4) **Question Generation** applies templates to create structured QA pairs with validated ground truth.

3. ENC-Bench

We introduce ENC-Bench, a comprehensive benchmark with 20,490 expert-validated samples from 840 authentic NOAA S-57 charts, organized into three tiers (Perception $\rightarrow$ Spatial Reasoning $\rightarrow$ Maritime Decision-Making) with systematic variations across lighting modes and scale levels.

3.1. Dataset Construction

3.1.1. Data Source and Coverage

We collected 840 Electronic Navigational Charts from the National Oceanic and Atmospheric Administration (NOAA) [43], which provides publicly accessible, quality-assured official ENCs actively used in operational navigation. All charts conform to the IHO S-57 standard [25], the globally recognized format for official ENCs. The dataset encompasses operational scenarios: shallow harbors, deep-water shipping lanes, traffic separation schemes, restricted zones, and ecologically sensitive areas, ensuring comprehensive representation of real-world navigation contexts.

3.1.2. Data Generation Pipeline

We developed a semi-automated four-stage pipeline to transform raw S-57 vector data into structured question-answer pairs (Figure 2):

Stage 1: Data Rendering & Parsing. S-57 binary charts are rendered via OpenCPN [45] across three lighting modes (day/dusk/night) following Electronic Chart Display and Information System (ECDIS) operational standards [26] and six scale levels (1:50k, 1:70k, 1:100k, 1:130k, 1:200k, 1:300k). Charts are parsed into GeoJSON features using GDAL [48], with abbreviated attribute codes mapped to human-readable descriptions via official IHO lookup tables [25]. Cross-referenced features are merged based on LNAM identifiers to maintain semantic integrity.

Stage 2: Image Registration. We establish bidirectional pixel-to-geographic coordinate conversion through control

point matching. Two geographic control points are manually labeled per chart in pixel space using Labelme [55], enabling affine transformation matrices for precise coordinate conversion required by spatial reasoning tasks.

Stage 3: Feature Annotation. Point, linestring, and polygon features are systematically annotated. Point features are grouped using graph-coloring algorithms to prevent visual overlap, with manual refinement for ambiguous cases. Linear features are marked at endpoints. Polygons use axis-aligned bounding boxes. Density control ensures visual clarity with a median of 8 features per image. All annotations are verified through expert review to ensure correctness and nautical plausibility.

Stage 4: Question Generation. We apply task-specific templates to generate 20,490 questions across 10 categories. Ground truth for spatial reasoning tasks is computed using validated nautical formulas: Haversine distance for nautical mile calculations, bearing computation via arctangent of coordinate differences adjusted for true north, and affine transformation for coordinate localization. Multiple-choice distractors are systematically generated based on common navigation errors (e.g., reversed bearings, incorrect unit conversions, overlooked constraints).

3.1.3. Quality Control

We implement a rigorous two-stage validation protocol. **Stage 1 (Automated Verification):** Each question’s answer is cross-checked against original chart file attributes, validating coordinates through transformation matrices, depth values via S-57 SOUNDG objects, feature classifications through object class codes, and computed spatial metrics using reference implementations. **Stage 2 (Expert Review):** All generated questions undergo review by maritime navigation professionals for formatting consistency, linguistic clarity, and plausibility of multiple-choice options, ensuring alignment with professional practice.

3.2. Benchmark Design

We structure ENC-Bench into three tiers with 10 task dimensions reflecting the cognitive hierarchy of maritime navigation: perception, spatial reasoning, and decision-making, as illustrated in Figure 1.

3.2.1. L-1 Perception (4 Tasks)

Perception tasks assess the ability to interpret standardized maritime symbology in authentic chart contexts, testing both isolated symbol recognition and contextual feature understanding across three geometry types.

Symbol Recognition. Models identify IHO-standardized symbols from isolated visual appearance, testing direct visual-to-semantic mapping of maritime notation.

Point Feature Understanding. Models extract structured attributes from point features (buoys, lighthouses) embedded in complete chart scenes.

Linestring Feature Understanding. Models identify types and directional semantics of linear features (tracks, contours) marked with endpoints in full chart contexts.

Polygon Feature Understanding. Models interpret regional semantics and safety attributes of area features (zones, restricted areas) marked with bounding boxes.

3.2.2. L-2 Spatial Reasoning (3 Tasks)

Spatial reasoning demands quantitative geometric computation grounded in cartographic principles, testing numerical reasoning with coordinate systems and scale conversion.

Coordinate Localization. Given a marked point, models predict geographic location through two approaches: outputting decimal-degree coordinates converted to pixel space, or directly predicting pixel coordinates. Both are evaluated using pixel-space error, enabling comparison of symbolic notation interpretation versus visual localization.

Bearing Calculation. Given two points, models compute compass bearing (0–360°, true north = 0°) between them, requiring point identification, vector angle calculation, and alignment with chart orientation.

Distance Measurement. Given two points, models calculate distance in nautical miles, extracting scale information, measuring pixel distance, and applying unit conversion.

3.2.3. L-3 Maritime Decision-Making (3 Tasks)

The highest tier evaluates synthesis of perceptual and spatial information into actionable navigation decisions under real-world multi-constraint scenarios.

Track Direction Recognition. Given a marked route with endpoints, models determine legal navigation direction: “from a to a' ”, “from a' to a ”, or “non-navigable”, interpreting traffic separation schemes and routing regulations.

Safety Passage Assessment. Given a vessel’s draft and passage, models judge if safe transit is possible (True/False), requiring extraction of depth soundings, identification of minimum depth, and application of safety margins.

Table 2. Key statistics of the ENC-Bench benchmark. *Excluding 384 Symbol Recognition samples (isolated symbols without contextual rendering). Contextual tasks total: 20,106 samples.

Statistics	Number
Total Questions	20,490
Total Charts	840
Task Categories	10
Task Hierarchy	
Perception	14,886 (72.6%)
- Symbol Recognition	384
- Point Features	7,830
- Linestring Features	2,562
- Polygon Features	4,110
Spatial Reasoning	4,338 (21.2%)
- Coordinate Localization	2,016
- Bearing Calculation	1,161
- Distance Measurement	1,161
Maritime Decision-Making	1,266 (6.2%)
- Track Direction Recognition	141
- Safety Passage Assessment	702
- Anchorage Selection	423
Lighting Modes (Contextual Tasks)*	
Day Mode	6,702 (33.3%)
Dusk Mode	6,702 (33.3%)
Night Mode	6,702 (33.3%)
Scale Levels (Contextual Tasks)*	
Large Scale (1:50k, 1:70k)	6,507 (32.4%)
Intermediate Scale (1:100k, 1:130k)	7,557 (37.6%)
Small Scale (1:200k, 1:300k)	6,042 (30.0%)

Anchorage Selection. When a vessel encounters an emergency, models select the nearest safe anchorage from four candidates: minimizing distance while ensuring adequate depth and verifying non-restricted status.

3.3. Dataset Statistics

Scale and Distribution. Table 2 presents key statistics. ENC-Bench comprises 20,490 samples with hierarchical distribution: Perception (72.6%), Spatial Reasoning (21.2%), and Maritime Decision-Making (6.2%). This pyramid structure reflects cognitive complexity progression—models may excel at perception yet fail at spatial computation or multi-constraint judgment.

Multi-Condition Rendering. Contextual samples are evenly distributed across three lighting modes and six scale levels. *Lighting modes* include day (bright backgrounds), dusk (reduced contrast), and night (dark backgrounds, high-contrast symbols). *Scale levels* span large scales (1:50k, 1:70k) with maximum detail, intermediate scales (1:100k, 1:130k) with moderate density, and small scales (1:200k, 1:300k) with cartographic generalization. This systematic variation tests model robustness across conditions mariners encounter when adjusting display settings.

Table 3. Performance on Perception and Decision-Making tasks. Results in accuracy (%). **Bold** indicates best performance.

Model	Symbol	Perception			Decision-Making			Average
		Point	Line	Polygon	Track	Safety	Anchor	
Gemini-2.5-Pro [10]	69.53	45.38	30.05	39.95	63.12	57.55	29.55	47.88
Gemini-2.5-Flash [10]	63.80	41.89	29.08	39.44	59.57	55.84	24.35	44.85
GPT-4o [24]	50.78	36.39	21.58	23.97	45.39	45.44	20.57	34.87
Qwen3-VL-235B-Instruct [61]	57.03	51.79	29.70	29.93	74.47	58.97	26.48	46.91
Qwen3-VL-235B-Thinking [61]	54.17	49.34	28.18	27.83	75.18	61.54	30.50	46.68
Qwen3-VL-32B-Instruct [61]	50.26	48.77	29.39	26.86	68.79	58.40	17.26	42.82
Qwen3-VL-32B-Thinking [61]	46.88	35.93	28.45	24.67	73.76	55.98	18.44	40.59
GLM-4.5V [54]	38.80	43.44	20.92	21.61	53.19	65.67	26.24	38.55
InternVL-3-38B [56]	55.99	27.36	19.87	30.29	51.06	54.13	20.57	37.04
Llama-4-Maverick-17B [41]	47.66	42.75	19.83	18.54	53.90	57.55	14.89	36.44
Random Chance	25.00	25.00	25.00	25.00	33.33	50.00	25.00	29.76

Table 4. Performance on Spatial Reasoning tasks. Results in Acc@T (%) and Mean Error. **Bold** indicates best performance.

Model	Coordinate (Geo)		Coordinate (Pixel)		Direction		Distance	
	Acc@200px (%) ↑	Err (px) ↓	Acc@200px (%) ↑	Err (px) ↓	Acc@20° (%) ↑	Err (°) ↓	Acc@0.2 (%) ↑	Err (%) ↓
Gemini-2.5-Pro [10]	17.36	495.2	21.43	480.5	46.86	29.95	25.67	42.31
Gemini-2.5-Flash [10]	16.77	502.1	21.03	466.0	28.77	66.89	25.93	42.72
GPT-4o [24]	12.20	540.5	6.94	552.1	25.15	73.88	19.29	48.78
Qwen3-VL-235B-Instruct [61]	11.31	570.4	12.20	510.9	36.95	33.53	16.71	56.36
Qwen3-VL-235B-Thinking [61]	11.01	555.2	12.90	501.7	55.64	34.15	20.33	51.02
Qwen3-VL-32B-Instruct [61]	5.26	630.3	5.16	635.3	26.44	52.11	12.75	48.70
Qwen3-VL-32B-Thinking [61]	6.94	615.1	7.34	610.2	44.10	52.08	4.91	58.20
GLM-4.5V [54]	9.52	580.6	14.38	497.9	28.60	32.22	9.04	63.03
InternVL-3-38B [56]	13.19	512.7	10.52	565.1	41.09	54.80	13.09	47.95
Llama-4-Maverick-17B [41]	9.13	590.3	12.90	574.4	14.90	74.22	7.41	68.99

4. Experiments

4.1. Experimental Setup

Evaluated Models. We evaluate ten state-of-the-art MLLMs spanning diverse architectures and parameter scales, categorized into two types: 1) *Closed-source models* include GPT-4o [24], Gemini 2.5 Pro, and Gemini 2.5 Flash [10], representing commercial frontier systems; 2) *Open-source models* comprise Qwen3-VL-235B and Qwen3-VL-32B [61] (each available in both Instruct and Thinking variants), InternVL-3-38B [56], GLM-4.5V [54], and Llama-4-Maverick-17B-128E [41]. This selection spans 17B to 235B parameters, enabling analysis across model sizes, architectural paradigms (dense vs. MoE), and reasoning strategies (instruct vs. thinking).

Evaluation Metrics. Model performance is evaluated under zero-shot setting with uniform prompts across all models. Perception and Decision-Making tasks use multiple-choice format with chart images and natural language questions, while Spatial Reasoning tasks require numeric outputs with explicit unit specifications. Closed-source models are accessed via official APIs; open-source models via HuggingFace Transformers. Perception and Decision-Making tasks use **Accuracy (%)**. Spatial Reasoning tasks employ **Accuracy at Tolerance (Acc@T)**—predictions within error thresholds—and **Mean Error** in task-specific units. Co-

ordinate localization evaluates geographic coordinates (latitude/longitude converted to pixels) and direct pixel prediction, both with 200px tolerance. Bearing and distance use 20° and 20% relative error thresholds. Full prompts and additional results are in the Appendix.

4.2. Main Results

We present evaluation results across the three tiers. Table 3 reports performance on Perception and Decision-Making tasks, while Table 4 presents Spatial Reasoning results.

Results on Perception Tasks. Perception tasks reveal a substantial performance gap between isolated and contextual understanding. Symbol Recognition on isolated symbols proves most accessible, with Gemini-2.5-Pro achieving approximately 70% accuracy. However, when these same symbols appear embedded in complete chart scenes, performance degrades dramatically—contextual feature recognition drops to a range of 30%-52% depending on feature type and visual complexity. This degradation demonstrates that overlapping layers, dense bathymetry, and competing visual elements fundamentally challenge current vision architectures trained predominantly on natural images. All models substantially exceed random baseline performance, yet the best average across perception tasks reaches only 48%, falling far short of professional maritime requirements where near-perfect accuracy is expected.

Table 5. Average model performance across lighting conditions. Delta shows change relative to day mode baseline. **green** indicates improvement, **Red** indicates degradation.

Task Category	Day	Dusk	Night
<i>Perception & Decision-Making (Accuracy % ↑)</i>			
Point Features	42.88	41.76 (-1.12)	42.27 (-0.61)
Linestring Features	27.01	24.86 (-2.15)	25.26 (-1.75)
Polygon Features	27.32	29.92 (+2.60)	27.70 (+0.38)
Track Direction	65.35	61.27 (-4.08)	58.92 (-6.43)
Safety Assessment	57.80	55.99 (-1.81)	57.52 (-0.28)
Anchorage Selection	24.63	22.18 (-2.45)	21.85 (-2.78)
<i>Spatial Reasoning (Mean Error ↓)</i>			
Coordinate (Geo) Error (px)	592.4	587.6 (-4.8)	490.5 (-101.9)
Coordinate (Pixel) Error (px)	558.5	539.1 (-19.4)	520.6 (-37.9)
Direction Error (°)	55.11	51.26 (-3.85)	44.78 (-10.33)
Distance Error (%)	53.01	52.46 (-0.55)	52.94 (-0.07)

Results on Spatial Reasoning Tasks. Spatial Reasoning exposes fundamental limitations in visually-grounded numerical computation (Table 4). Coordinate localization reveals a counterintuitive pattern: models achieve better accuracy through direct pixel prediction than through geographic coordinate conversion, despite the latter being theoretically more precise for cartographic localization. This pattern reveals a critical bottleneck—models struggle more with interpreting formal coordinate notation systems than with visual feature localization itself. The geographic approach requires reading grid annotations, interpolation, and unit conversion, with errors cascading at each step. Direction calculation shows that extended reasoning variants substantially outperform standard models, while distance measurement proves universally challenging with best performance around 26% accuracy and mean relative errors exceeding 40%. The consistently poor results suggest fundamental architectural limitations in maintaining precision through multi-step numerical reasoning chains.

Results on Decision-Making Tasks. Decision-Making tasks reveal divergent difficulty levels that directly correlate with constraint complexity. Simpler tasks with constrained decision spaces achieve moderate success, reaching approximately 65%-75% accuracy. In stark contrast, Anchorage Selection—requiring simultaneous optimization of distance, depth, and regulatory constraints—proves catastrophic, with best performance barely exceeding random baseline. Extended reasoning mechanisms consistently improve performance, with gains ranging from minimal on simple tasks to substantial on complex ones. However, these improvements prove insufficient for operational deployment. The results expose a fundamental architectural limitation: current models lack explicit mechanisms for constraint enumeration, independent verification, and multi-objective synthesis required in safety-critical professional navigation scenarios.

Table 6. Average model performance across scale levels. Delta shows change relative to large scale baseline. **green** indicates improvement, **Red** indicates degradation.

Task Category	Large Scale (1:50k, 1:70k)	Intermediate (1:100k, 1:130k)	Small Scale (1:200k, 1:300k)
<i>Perception Tasks (Accuracy % ↑)</i>			
Point Features	42.24	47.72 (+5.48)	38.19 (-4.05)
Linestring Features	28.83	20.23 (-8.60)	20.26 (-8.57)
Polygon Features	28.93	26.73 (-2.20)	25.37 (-3.56)
<i>Decision-Making Tasks (Accuracy % ↑)</i>			
Track Direction	68.16	58.33 (-9.83)	52.30 (-15.86)
Safety Assessment	51.19	59.72 (+8.53)	48.96 (-2.23)
Anchorage Selection	16.47	20.34 (+3.87)	11.22 (-5.25)

4.3. Ablation Studies

We analyze model performance robustness across two critical rendering conditions: lighting modes and scale levels. These dimensions directly impact operational ECDIS usage as mariners regularly adjust display settings based on ambient conditions and navigation phases.

Performance Across Lighting Modes. Table 5 presents average performance across three lighting modes, revealing divergent task-specific dependencies on color information. Perception and Decision-Making tasks favor day mode, degrading up to 6% under night mode as color-dependent features lose discriminability. In stark contrast, Spatial Reasoning tasks paradoxically improve under night mode rendering. Geographic coordinate localization shows dramatic improvement, substantially exceeding direct pixel localization gains. This counterintuitive pattern reveals an underlying trade-off: high-contrast monochromatic rendering enhances symbolic notation reading by reducing visual ambiguity, while simultaneously impairing semantic feature classification that relies on color discrimination. Direction calculation shows substantial improvements, while distance measurement demonstrates minimal sensitivity.

Performance Across Scale Levels. Table 6 analyzes performance across large, intermediate, and small scales, exposing adaptation failures to cartographic generalization. Perception tasks exhibit complex non-monotonic patterns as scale decreases: linear and polygon features degrade consistently with each scale reduction, while point features paradoxically improve at intermediate scales before declining. This demonstrates models cannot adjust interpretation strategies as rendering simplifies at smaller scales—a capability humans employ intuitively. Decision-Making shows greater instability, with Track Direction suffering severe degradation (approaching 16%), while other tasks fluctuate unpredictably. Overall, Perception degrades progressively while Decision-Making maintains stability at intermediate scales before collapsing. These variations fundamentally undermine reliability for operational use, where mariners routinely adjust zoom levels.

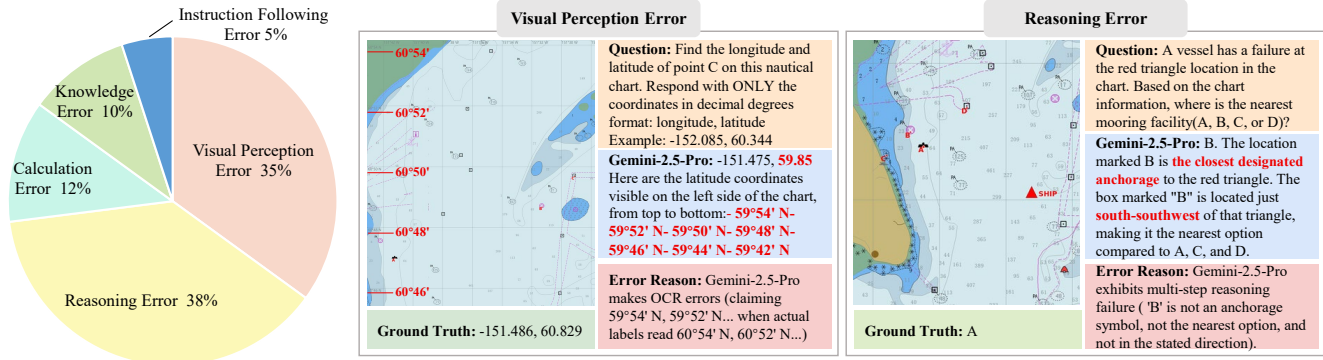


Figure 3. Error distribution of Gemini-2.5-Pro's incorrect results and example errors in its responses.

4.4. Error Analysis and Discussions

Error Analysis. We conduct error analysis on 200 randomly sampled incorrect predictions from Gemini-2.5-Pro, as visualized in Figure 3. Reasoning errors and visual perception errors dominate failures, accounting for 38% and 35% of total errors respectively. Reasoning errors primarily manifest as multi-step failures in spatial judgment and constraint integration, where models correctly perceive individual elements but fail to reason about their relationships. For instance, in anchorage selection tasks, models misidentify feature types, misjudge spatial proximity, and provide incorrect directional descriptions—failing across multiple reasoning dimensions simultaneously. Visual perception errors commonly involve OCR failures on coordinate grid annotations: models consistently misread latitude values by approximately one degree, directly explaining why geographic coordinate prediction underperforms direct pixel localization. Calculation errors occur when numerical computations fail despite correct visual perception, while knowledge errors and instruction-following issues represent secondary failure modes likely stemming from insufficient domain-specific pretraining.

Findings and Discussions. We discuss key findings from our experiments to inspire future work: **1)** Current MLLMs fall significantly short of professional maritime requirements, with best performance reaching only 47.88% compared to near-perfect accuracy expected in operational settings. The gap between perception and decision-making capabilities indicates that symbolic recognition alone is insufficient—structured reasoning remains the critical bottleneck. **2)** Geographic coordinate prediction consistently underperforms direct pixel localization despite being theoretically more precise, revealing that models struggle more with formal notation systems than with visual feature detection. This exposes a fundamental limitation in symbolic-to-spatial translation. **3)** Extended reasoning mechanisms substantially improve performance on decomposable geometric tasks but fail at multi-constraint optimization. Think-

ing modes excel at procedural decomposition yet lack explicit constraint verification mechanisms for complex decisions. **4)** Models show limited robustness across rendering variations. Non-monotonic patterns across scale levels and paradoxical improvements under night mode demonstrate inability to adjust interpretation strategies—a capability humans employ intuitively. **5)** Multi-constraint decision-making reveals architectural limitations of current models. Tasks requiring simultaneous optimization of distance, depth, and regulatory compliance demand structured reasoning capabilities absent in existing models. These findings demonstrate that professional maritime applications require innovations in symbolic grounding, verifiable reasoning, and rendering-aware processing beyond merely scaling existing architectures.

5. Conclusion

We introduce ENC-Bench, the first comprehensive benchmark for evaluating multimodal large language models on Electronic Navigational Chart understanding. ENC-Bench addresses the gap between general visual understanding and professional maritime navigation through 20,490 expert-validated samples from 840 NOAA charts, organized into a three-tier framework spanning perception, spatial reasoning, and maritime decision-making. Evaluation of 10 state-of-the-art MLLMs reveals severe capability gaps, with the best model achieving only 47.88% accuracy and catastrophic failures in multi-constraint reasoning (30.50% on anchorage selection). Error analysis identifies fundamental limitations in symbolic notation interpretation, constraint satisfaction, and adaptive visual processing across rendering variations. Looking ahead, our benchmark methodology extends to other professional domains requiring standardized symbolic systems—aviation charts, infrastructure blueprints, and scientific visualization. We anticipate that ENC-Bench will inspire research in domain-specific visual understanding and benefit the development of deployable AI systems for highly specialized vertical domains.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62306331 and CAAI Youth Talent Lifting Project under Grant CAAI2023-2025QNRC001.

References

- [1] Allianz Commercial. Safety and shipping review 2025. Accessed: 2025-05-01. [1](#)
- [2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. [3](#)
- [3] Yakoub Bazi, Laila Bashmal, Mohamad Mahmoud Al Rahhal, Riccardo Ricci, and Farid Melgani. RS-LLaVA: A large vision-language model for joint captioning and question answering in remote sensing imagery. *Remote Sensing*, 16(9): 1477, 2024. [3](#)
- [4] Simon Blindheim and Tor Arne Johansen. Electronic navigational charts for visualization, simulation, and autonomous ship control. *IEEE Access*, 10:3716–3737, 2021. [1](#)
- [5] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: a survey. *arXiv preprint arXiv:2402.12451*, 2024. [1](#)
- [6] Jiaqi Chen, Jianheng Li, Jinghui Lv, Wenjie Ding, Yuanping Zhang, Yuqin Xiao, Songtao Lin, Wenhao Wang, Jiajun Wu, and Lixin Wang. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523, 2021. [3](#)
- [7] Jiaqi Chen, Tong Lv, Lei Lin, Jinghui Lai, Jianheng Pan, Jiajun Wu, Yufeng Xiao, Yuqi Wu, and Lixin Wang. Uni-Geo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*, 2022. [3](#)
- [8] Jian Chen, Hao Wang, et al. Multi-source data-driven bayesian network for risk analysis of maritime accidents in the high sea. *Frontiers in Marine Science*, 12, 2025. [1](#)
- [9] Zhe Chen, Jiannan Wu, Wenhao Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. [3](#)
- [10] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blisstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. [1](#), [6](#)
- [11] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 958–979, 2024. [1](#)
- [12] Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, et al. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1135–1159, 2025. [2](#)
- [13] Shuaichen Deng, Zhiqiu Hu, Zheng Yuan, Yiqun Qin, Ali Mansourian, and Qunying Zhou. MapQA: A benchmark for question answering on choropleth maps. *arXiv preprint arXiv:2501.01626*, 2025. [3](#)
- [14] Mahir Labib Dihan, MD Tanvir Hassan, MD TANVIR PARVEZ, Md Hasebul Hasan, Md Almash Alam, Muhammad Aamir Cheema, Mohammed Eunus Ali, and Md Rizwan Parvez. Mapeval: A map-based evaluation of geo-spatial reasoning in foundation models. In *Forty-second International Conference on Machine Learning*, 2025. [3](#)
- [15] European Maritime Safety Agency. Annual overview of marine casualties and incidents 2024. Technical report, European Maritime Safety Agency, 2024. Accessed: 2024-12-17. [1](#)
- [16] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yongdong Qin, Mengdan Zhang, Xing Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2024. [3](#)
- [17] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. [1](#), [3](#)
- [18] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. [1](#)
- [19] Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhou. ChartLlama: A multimodal LLM for chart understanding and generation. *arXiv preprint arXiv:2311.16483*, 2024. [3](#)
- [20] Hartis Organization. Electronic navigational charts (enc), 2024. Accessed: 2024. [1](#)
- [21] Horst Hecht. *The electronic chart: functions, potential and limitations of a new marine navigation system*. Geomares publishing, 2006. [1](#)
- [22] Muye Huang, Han Lai, Xinyu Zhang, Wenjun Wu, Jie Ma, Lingling Zhang, and Jun Liu. Evchart: A benchmark and a self-training approach towards real-world chart understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3680–3688, 2025. [2](#)
- [23] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional

- question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1, 3
- [24] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1, 6
- [25] S IHO. 57. iho transfer standard for digital hydrographic data. *International Hydrographic Organisation*, 2000. 1, 4
- [26] International Maritime Organization. Electronic nautical charts (enc) and electronic chart display and information systems (ecdis), 2024. Accessed: 2024. 4
- [27] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016. 2
- [28] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. GeoChat: Grounded large vision-language model for remote sensing. *arXiv preprint arXiv:2311.15826*, 2024. 3
- [29] Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. *International Conference on Machine Learning*, pages 18893–18912, 2023. 3
- [30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 3
- [31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2024. 3
- [32] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer, 2024. 3
- [33] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024. 2
- [34] Lloyd’s Register. The rapid rise of ai in maritime, 2024. Accessed: 2024. 1
- [35] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. RSVQA: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020. 3
- [36] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. *International Conference on Learning Representations*, 2024. 3
- [37] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 1
- [38] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022. 2, 3
- [39] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 2, 3
- [40] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 2, 3
- [41] AI Meta. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>, checked on, 4 (7):2025, 2025. 1, 6
- [42] Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. LHRs-Bot: Empowering remote sensing with VGI-enhanced large multimodal language model. In *European Conference on Computer Vision*, pages 437–455. Springer, 2024. 3
- [43] NOAA Office of Coast Survey. NOAA ENC – electronic navigational charts, 2024. Accessed: 2024. 1, 4
- [44] OpenAI. GPT-4V(ision) system card. *arXiv preprint arXiv:2303.08774*, 2023. 3
- [45] OpenCPN Project. Opencpn: Open source chart plotter and navigation software, 2025. Version 5.12.4, accessed November 12, 2025. 4
- [46] Orca AI. Can ai chart a new course for the maritime industry?, 2025. Accessed: 2025-06-10. 1
- [47] Athanasios Palikaris and Athanasios K Mavraeidopoulos. Electronic navigational charts: International standards and map projections. *Journal of Marine Science and Engineering*, 8(4):248, 2020. 1
- [48] Even Rouault, Frank Warmerdam, Kurt Schwehr, Andrey Kiselev, Howard Butler, Mateusz Łoskot, Tamas Szekeres, Etienne Tourigny, Martin Landa, Idan Miara, Ben Elliston, Kumar Chaitanya, Lucian Plesea, Daniel Morissette, Ari Jolma, Nyall Dawson, Daniel Baston, Craig de Stigter, and Hiroshi Miura. GDAL, 2025. 4
- [49] Luca M Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, 7(1):96–106, 2025. 1, 3
- [50] Fatemeh Shiri, Xiao-Yu Guo, Mona Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21440–21455, 2024. 2
- [51] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, pages 8317–8326, 2019. 1

- [52] Varun Srivastava, Fan Lei, Srija Mukhopadhyay, Vivek Gupta, and Ross Maciejewski. Mapiq: Evaluating multimodal large language models for map question answering. *arXiv preprint arXiv:2507.11625*, 2025. 2
- [53] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2024. 3
- [54] V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihang Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazhen Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. 1, 6
- [55] Kentaro Wada. Labelme: Image Polygonal Annotation with Python. 4
- [56] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 1, 6
- [57] Zhecheng Wang, Rajanie Prabhu, Tianyuan Huang, Jiajun Wu, and Karttikeya Mangalam. SkyScript: A large and semantically diverse vision-language dataset for remote sensing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5508–5516, 2024. 3
- [58] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE, 2023. 1, 3
- [59] Renqiu Xia, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Botian Shi, Junchi Yan, and Bo Zhang. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *IEEE Transactions on Image Processing*, 2025. 2
- [60] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 1, 3
- [61] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 6
- [62] Sahiti Yerramilli, Nilay Pande, Rynaa Grover, and Jayant Sravan Tamarapalli. Geochain: Multimodal chain-of-thought for geographic reasoning. *arXiv preprint arXiv:2506.00785*, 2025. 2
- [63] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12): nwae403, 2024. 1
- [64] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. *arXiv preprint arXiv:2311.16502*, 2024. 3
- [65] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024. 1, 3
- [66] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 1