NeSy-Route: A Neuro-Symbolic Benchmark for Constrained Route Planning in Remote Sensing

Ming Yang[1,2], Zhi Zhou[1*], Shi-Yu Tian[1,2], Kun-Yang Yu[1,2], Lan-Zhe Guo[1,3], and Yu-Feng Li[1,2⋆]

[1] National Key Laboratory for Novel Software Technology, Nanjing University, China
[2] School of Artificial Intelligence, Nanjing University, China
[3] School of Intelligence Science and Technology, Nanjing University, China
{yangm,zhouz,tiansy,yuky}@lamda.nju.edu.cn,
{guolz,liyf}@nju.edu.cn

Abstract. Remote sensing underpins crucial applications such as disaster relief and ecological field surveys, where systems must understand complex scenes and constraints and make reliable decisions. Current remote-sensing benchmarks mainly focus on evaluating perception and reasoning capabilities of multimodal large language models (MLLMs). They fail to assess planning capability, stemming either from the difficulty of curating and validating planning tasks at scale or from evaluation protocols that are inaccurate and inadequate. To address these limitations, we introduce NeSy-Route, a large-scale neuro-symbolic benchmark for constrained route planning in remote sensing. Within this benchmark, we introduce an automated data-generation framework that integrates high-fidelity semantic masks with heuristic search to produce diverse route-planning tasks with provably optimal solutions. This allows NeSy-Route to comprehensively evaluate planning across 10,821 route-planning samples, nearly 10 times larger than the largest prior benchmark. Furthermore, a three-level hierarchical neuro-symbolic evaluation protocol is developed to enable accurate assessment and support fine-grained analysis on perception, reasoning, and planning simultaneously. Our comprehensive evaluation of various state-of-the-art MLLMs demonstrates that existing MLLMs show significant deficiencies in perception and planning capabilities. We hope NeSy-Route can support further research and development of more powerful MLLMs for remote sensing.The dataset and code are available at https://mingyang1010.github.io/NeSy-Route/.

Keywords: Remote Sensing · Neuro-Symbolic AI · Route Planning

1 Introduction

In recent years, multimodal large language models (MLLMs) [1, 2, 7, 31, 34] have achieved remarkable progress in visual perception [27, 50, 51] and complex reasoning [23, 32, 35, 44, 45, 52]. Remote sensing imagery, a crucial source for ob-

⋆ Corresponding authors.

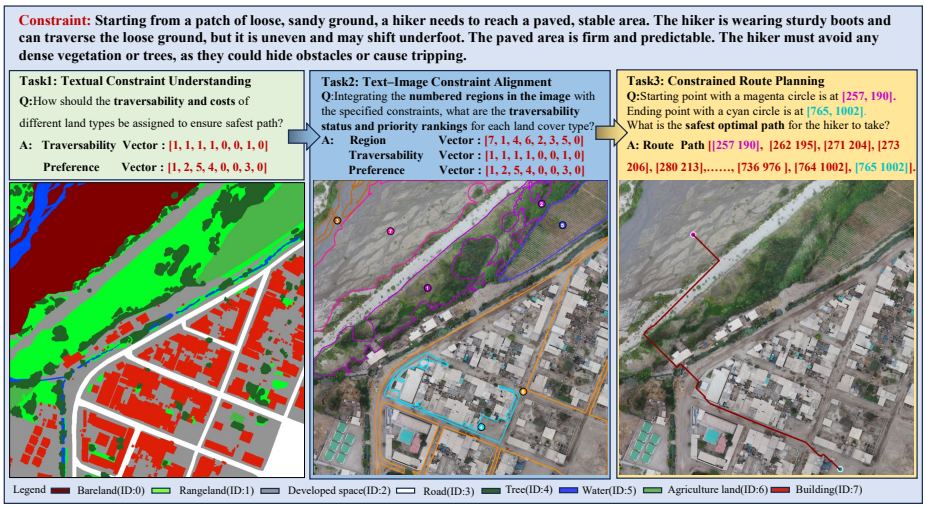

Fig. 1: A typical example from NeSy-Route Benchmark. NeSy-Route evaluates MLLMs through a hierarchical reasoning pipeline consisting of three integrated tasks. Task 1 involves extracting symbolic traversability and cost vectors from the provided scenario. Task 2 focuses on anchoring these symbolic constraints to specific identified regions within the remote sensing image. Task 3 assesses the capability to generate a sparse waypoint trajectory that avoids obstacles and minimizes the cumulative cost.

serving the Earth's surface, supports applications such as environmental monitoring [19, 28], disaster assessment [6], agricultural management [48], and urban planning [53]. Among these applications, route planning is important for tasks such as emergency response and resource allocation. However, conventional route-planning systems rely on costly terrestrial surveying [12] to construct road maps, making them vulnerable in disaster-stricken or poorly mapped regions.

MLLMs offer a potential alternative by processing remote sensing imagery for route planning, yet existing remote-sensing benchmarks [20, 37] mainly evaluate perception and reasoning, providing little insight into constrained planning abilities, which stem either from the difficulty of curating and validating planning tasks at scale or from evaluation protocols that are inaccurate and inadequate.

To address these challenges, we introduce NeSy-Route, a neuro-symbolic evaluation benchmark for constrained route planning built on OpenEarthMap [43], designed to hierarchically assess route planning capability of MLLMs in remote sensing. The benchmark organizes its task across three levels:

1. **Textual Constraint Understanding Task.**This task evaluates the capacity of MLLMs to decode natural language instructions into formal symbolic logic through 3,607 problems, establishing a critical prerequisite for performing constrained route planning.
2. **Text–Image Constraint Alignment Task.** This task challenges models to anchor textual constraints onto 12,975 samples with identified visual re-

gions, requiring joint reasoning to determine terrain-specific traversability and priority rankings.
3. **Constrained Route Planning Task.** This task assesses end-to-end planning solver performance by requiring the generation of executable waypoint sequences between designated start and end points that strictly honor both topological barriers and land type constraints across 10,821 samples.

In summary, our main contributions are as follows:

1. We propose NeSy-Route, the first neuro-symbolic evaluation benchmark for route planning in remote sensing. It hierarchically assesses the constrained planning capabilities of MLLMs with a symbolic evaluator across three dimensions: textual constraint understanding, text–image constraint alignment, and constrained route planning.
2. We design an automated, symbolized data generation framework and develop a symbolized route planning evaluator, establishing a closed-loop framework that integrates neuro generation with symbolic verification.
3. We conduct a comprehensive evaluation of state-of-the-art MLLMs and reveal significant limitations in their perception, reasoning, planning capabilities simultaneously in real remote sensing scenarios. We further discuss potential directions for MLLMs in advancing planning capabilities.

2 Related Work

2.1 Remote Sensing Benchmark

In recent years, the rapid advancement of MLLMs has accelerated the transition of remote sensing research from fundamental perception tasks to complex reasoning scenarios featuring smaller targets or higher spatial resolutions. While early benchmarks primarily addressed semantic segmentation [4,40,43], object detection [16,18] and classification [42], more recent efforts have expanded into VQA task such as spatial relationship reasoning [20, 25, 29, 36], temporal change detection [17], and climate-related risk prediction [38]. However, these datasets are predominantly designed for descriptive reasoning regarding specific objects and their mutual relationships, thereby neglecting the challenges of route planning within complex remote sensing contexts. XLRS-Bench [37]manually annotates 16 sub-tasks and is the first to introduce the Route Planning task in the remote sensing domain, which requires the model to select the correct route description from several candidate options. However, such an evaluation paradigm cannot quantitatively assess planning capability in real remote sensing environments, where routes must be generated through spatial reasoning rather than selected from predefined candidate options.

2.2 Multimodal Large Language Model

MLLMs have advanced rapidly in recent years, and most of them now support inputs at native image resolution. Representative closed-source models in-

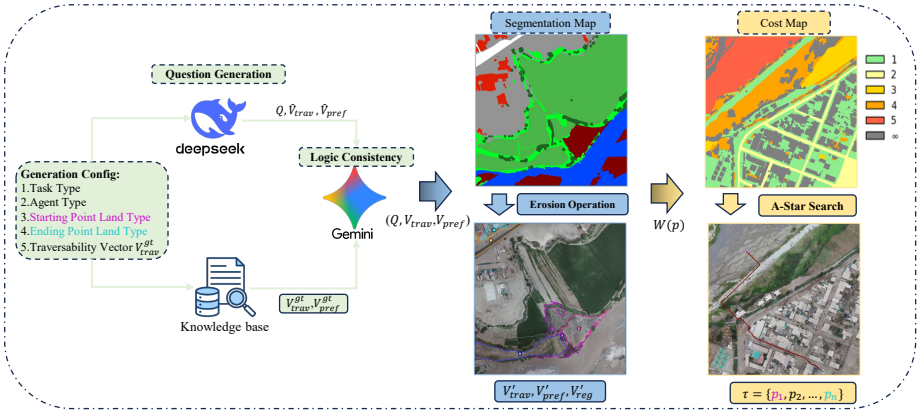

Fig. 2: An overview of automated data generation framework. The pipeline integrates dual-LLM logical verification for query synthesis, morphological erosion for semantic visual grounding, and constrained A-Star search algorithm for deriving mathematically optimal trajectories under symbolic rules.

clude Gemini-3-Pro [7], GPT-5.1 [34], and Qwen3-VL-Plus [2], which demonstrate strong capabilities in perception and reasoning. Among open-source models, dense architectures include LLaVA-OneVision [1], the InternVL-3.5-8B [41], Qwen3-VL-8B [2], and Qwen3.5-27B [31], while mixture-of-experts (MoE) architectures include Qwen3-VL-30B-A3B [2] and Qwen3-VL-235B. These open-source models also achieve competitive performance.MLLMs have also shown promising progress in the remote sensing domain. Geochat [15], built on the LLaVA-1.5 architecture [22], is trained to enable multi-task conversational capabilities. SkySenseGPT [8], also based on LLaVA-1.5, achieves improved performance across a variety of remote sensing tasks through large-scale training. EarthGPT [47] focuses on understanding of multi-sensor remote sensing data.

3 Automated Data Generation Framework

As illustrated in Fig. 2, the automated framework generates image-text pairs, GT vectors, and optimal trajectories for sub-tasks through a three-stage process.

3.1 Knowledge Base Construction

Following OpenEarthMap [43] standards, we define our knowledge base(KB) constraints across eight land-cover types: Bareland (ID:0), Rangeland (ID:1), Developed space (ID:2), Road (ID:3), Tree (ID:4), Water (ID:5), Agriculture land (ID:6), and Building (ID:7). Let $\mathcal{C} = \{c_0, c_1, \ldots, c_{L-1}\}$ be the set of $L = 8$ predefined land-cover classes. Specifically, we define three levels of traversability: always traversable, conditionally traversable, and non-traversable:

Table 1: Agent-Specific Traversability Definitions

Agent	$\mathcal{T}_A$	$\mathcal{T}_C$	$\mathcal{T}_N$
Pedestrian	$\{c_2, c_3\}$	$\{c_0, c_1, c_4, c_6\}$	$\{c_5, c_7\}$
Car	$\{c_2, c_3\}$	$\{c_0, c_1, c_6\}$	$\{c_4, c_5, c_7\}$
Drone	$\{c_0, c_1, c_2, c_3, c_5, c_6\}$	$\{c_4, c_7\}$	$\emptyset$
Boat	$\{c_5\}$	$\emptyset$	$\{c_0, c_1, c_2, c_3, c_4, c_6, c_7\}$

- **Always Traversable ($\mathcal{T}_A$)**: Terrains that are always passable by default. These areas are considered accessible unless explicitly restricted by specific problem constraints.
- **Conditionally Traversable ($\mathcal{T}_C$)**: Terrains that are initially impassable. These areas are considered blocked unless specific conditions for passage are met, as defined by the problem constraints.
- **Non-traversable ($\mathcal{T}_N$)**: Terrains that are permanently impassable. These areas are permanent obstacles that cannot be crossed under any circumstances, regardless of task-specific requirements.

Based on the land-cover physical properties, we define four agent types (pedestrian, vehicle, drone, and boat) with traversability constraints in Tab. 1.

Table 2: Agent-Specific Task Priority Rankings

Agent	**Fastest**	**Comfort**	**Safest**	**Shortest**
Pedestrian	$c_0 = c_1 = c_2 = c_3 = c_4 = c_6$	$c_2 > c_3 > c_6 > c_1 > c_0 > c_4$	$c_2 > c_3 > c_6 > c_1 > c_0 > c_4$	$c_0 = c_1 = c_2 = c_3 = c_4 = c_6$
Car	$c_3 > c_2 > c_6 > c_1 > c_0$	$c_3 > c_2 > c_6 > c_1 > c_0$	$c_3 > c_2 > c_6 > c_1 > c_0$	$c_0 = c_1 = c_2 = c_3 = c_6$
Drone	$c_0 = c_1 = c_2 = c_3 = c_5 = c_6 = c_7 > c_4$	$c_0 = c_1 = c_2 = c_3 = c_6 = c_7 > c_5 > c_4$	$c_0 = c_1 = c_2 = c_3 = c_6 > c_4 = c_5 = c_7$	$c_0 = c_1 = c_2 = c_3 = c_5 = c_6 = c_7 > c_4$
Boat	c_5	c_5	c_5	c_5

Additionally, we define four routing objectives: shortest, fastest, safest, and most comfortable in Tab. 2. For each routing objective, a priority ranking of land cover types is established, which determines the preferred terrains for each agent depending on the specific goal.

3.2 Problem Formulation

Given a remote sensing visible light image $I \in \mathbb{R}^{H \times W \times 3}$ and a natural language query Q describing a route planning scenario, we formulate the constrained route planning task as a hierarchical evaluation process. The objective is to derive

symbolic constraints from Q, align them with the visual contexts in I, and generate an optimal trajectory τ^* that minimizes cost function derived from these constraints.

Task 1: Textual Constraint Understanding Given Q, the model extracts symbolic traversability and preference vectors via $f_{\text{text}} : Q \to (\mathbf{V}_{\text{trav}}, \mathbf{V}_{\text{pref}})$.

- $\mathbf{V}_{\text{trav}} \in \{0,1\}^L$: Binary vector where $V_{\text{trav}}[i] = 1$ if class c_i is traversable for the specific agent, and 0 otherwise.
- $\mathbf{V}_{\text{pref}} \in \{0,\ldots,K\}^L$: Ordinal vector representing the priority of class c_i under the specified task objective, where K is the maximum priority tier. The higher the value, the higher the priority.

Task 2: Text–Image Constraint Alignment Given Q and the image I annotated with regions and their corresponding IDs, the model outputs symbolic region, traversability, and preference vectors: $f_{\text{align}} : (Q, I) \to (\mathbf{V}'_{\text{reg}}, \mathbf{V}'_{\text{trav}}, \mathbf{V}'_{\text{pref}})$.

- $\mathbf{V}'_{\text{reg}} \in \mathbb{N}^L$: A vector where $V_{\text{reg}}[i] = \text{Region\_ID}$ if land cover c_i is present and identified in the image, and 0 otherwise.
- $\mathbf{V}'_{\text{trav}}$ and $\mathbf{V}'_{\text{pref}}$: Represent the traversability and preference status specifically for the detected regions to evaluate whether the model can correctly apply symbolic rules to the finite set of terrains visible in I.

Based on $\mathbf{V}'_{\text{pref}}$, we define cost map $W(p)$ for every pixel $p \in I$ corresponding to its land-cover class $c(p)$:

$$W(p) = \begin{cases} \infty, & \text{if } V_{\text{trav}}[c(p)] = 0 \\ \max(\mathbf{V}'_{\text{pref}}) - V_{\text{pref}}[c(p)] + 1, & \text{if } V_{\text{trav}}[c(p)] = 1 \end{cases} \tag{1}$$

Task 3: Constrained Route Planning Given Q, I, start point $S \in \mathbb{N}^2$ and end point $E \in \mathbb{N}^2$, the objective is to generate a trajectory $\tau = \{p_1, p_2, \ldots, p_n\}$ that minimizes the cumulative spatial cost, defined as $f_{\text{plan}} : (I, Q, S, E) \to \tau$:

$$\min_{\tau} \sum_{i=1}^{n} W(p_i) \tag{2}$$

subject to $p_1 = S$ and $p_n = E$.

3.3 Symbolized Data Generation

Controllable Symbolic Query Synthesis We initiate the pipeline by sampling a configuration tuple $\sigma = (\text{agent}, \text{task}, \mathcal{C}_{\text{start/end}}, \mathcal{C}_{\text{allowed}})$. The ground-truth symbolic vectors are deterministically derived via KB, which ensures that

the underlying constraint is axiomatically governed by the rules defined in Sec. 3.1.

$$f_{\mathrm{KB}}(\sigma) \rightarrow (\mathbf{V}^{gt}_{\mathrm{trav}}, \mathbf{V}^{gt}_{\mathrm{pref}}) \tag{3}$$

We utilize DeepSeek-V3.2 [21] to synthesize the natural language query Q based on σ. To enforce semantic alignment, the model is required to perform self-inference, re-deriving the logic vectors from its own generated text:

$$Q = \mathcal{M}(\sigma), \quad (\hat{\mathbf{V}}_{\mathrm{trav}}, \hat{\mathbf{V}}_{\mathrm{pref}}) = \mathcal{P}_{\mathrm{self}}(Q) \tag{4}$$

where $\mathcal{M}$ denotes the generative mapping and $\mathcal{P}_{\mathrm{self}}$ represents the self-inference process, which ensures self-inferred vectors match the KB's GT. In addition, we utilize Gemini-3-Pro [7] to verify that the textual descriptions in Q for all $L = 8$ land-cover classes are strictly compliant with the formal KB definitions, minimizing hallucinations and ensuring high-fidelity alignment between textual reasoning and symbolic constraints.

Semantic Visual Grounding and Region Identification We ground the symbolic constraints from task 1 into segmentation masks provided by OpenEarthMap [43].To maintain the structural integrity of critical terrains, we perform selective morphological filtering. Small isolated regions are filled with the majority class of their neighborhood, with the exception of functionally narrow classes such as Water (ID: 5), Road (ID: 3), and Building (ID: 7). This preserves the topological connectivity required for path planning while reducing visual clutter. To ensure that each identified region provides a pure visual signal for Task 2, we apply iterative morphological erosion to large-scale land cover patches. Formally, for a region R_i of class c_k:

$$\mathcal{E}(R_i) = \{p \in R_i \mid \forall p' \in \mathcal{N}(p), c(p') = c_k\} \tag{5}$$

where $\mathcal{N}(p)$ denotes the local neighborhood, ensuring the probe area is a maximum inscribed pure region. Instead of random pairing, we ensure that the visual scene I can fully support the logic in Q. A query Q is matched with an image I only if the set of land covers present in the image $\mathcal{C}_I$ is a superset of the allowed and start/end classes specified in the query:$\mathcal{C}_I \supseteq \mathcal{C}_{\mathrm{allowed}}$. Specifically, we select the maximum area of each class present in I as the primary region of interest, which forces the model to reason across diverse ground conditions and complex topological structures.

Constrained Trajectory Generation and Optimization To ensure the feasibility of the planning mission under specific agent constraints, we first construct a region-level connectivity graph $\mathcal{G} = (V, E)$. Each node $v_i \in V$ represents a unique land-cover region R_i identified in Stage 2. We construct an initial adjacency matrix $\mathbf{A} \in \{0, 1\}^{N \times N}$, where $A_{ij} = 1$ if regions R_i and R_j are connected, and 0 otherwise. To incorporate the symbolic constraints, we derive a constrained adjacency matrix $\tilde{\mathbf{A}}$ by masking $\mathbf{A}$ with $\mathbf{V}^{gt}_{\mathrm{trav}}$:

$$\tilde{A}_{ij} = A_{ij} \cdot V^{gt}_{\mathrm{trav}}[c(R_i)] \cdot V^{gt}_{\mathrm{trav}}[c(R_j)] \tag{6}$$

where $c(R_i)$ denotes the class ID of region R_i. A candidate pair of (R_s, R_e) is validated if a path exists between them. To ensure a non-trivial navigation challenge and mitigate spatial biases, we specifically prioritize pairs (R_s, R_e) that are spatially distant yet topologically connected.

Upon confirming reachability, we project the symbolic vectors onto the pixel grid to construct the cost map $W(p)$ following Eq. (1). We then implement the A-Star search algorithm [10] to find the optimal path. For any pixel p, the total estimated cost $f(p)$ is defined as:

$$f(p) = g(p) + h(p) \tag{7}$$

where $g(p)$ is the actual cumulative cost from the start S to p, and $h(p)$ is the heuristic function estimating the cost from p to the destination E. We employ the Euclidean distance as the heuristic $h(p) = \|p - E\|_2$. The optimality of the generated path is guaranteed by the following properties:

- Admissibility: Since the minimum cost for any traversable pixel is $W(p) \geq 1$ (following our cost definition), the Euclidean distance always satisfies $h(p) \leq d^*(p, E)$, where d^* is the actual minimum cost.
- Consistency: Given that $W(p) \geq 1$ and the Euclidean distance satisfies the triangle inequality, the heuristic is consistent, meaning $h(p) \leq W(p, q) + h(q)$ for any adjacent pixels p and q.

Because $h(p)$ is both admissible and consistent, the A-Star algorithm is guaranteed to find the optimal trajectory τ^*.

4 NeSy-Route Benchmark

4.1 Dataset Statistics and Difficulty Stratification

The NeSy-Route benchmark provides a large-scale collection of hierarchical Q-A pairs. Task 1 comprises 3,607 symbolic samples derived from the knowledge base, specifically designed to evaluate the model's comprehension of textual constraints. For Task 2, the reasoning complexity escalates with the semantic density of the visual scene; thus, we stratify the dataset into three difficulty tiers based on the number of annotated land-cover classes M present in the image. This subset is partitioned into Easy ($1 \leq M \leq 4$) with 7,659 samples, Medium ($5 \leq M \leq 6$) with 3,712 samples, and Hard ($7 \leq M \leq L$) with 1,604 samples. We quantify the route planning challenge in Task 3 using Complexity Score ($\mathcal{D}$):

$$\mathcal{D} = \lambda_1 H_{\text{inter}} + \lambda_2 H_{\text{intra}} + \lambda_3 H_{\text{count}} + \lambda_4 \mathcal{C}_{\text{topo}} \tag{8}$$

where $H_{\text{inter}} = -\sum u_k \log_2 u_k$ and $H_{\text{count}} = -\sum v_k \log_2 v_k$ measure diversity via area proportions u_k and region counts v_k; $H_{\text{intra}} = \frac{1}{L} \sum_k \frac{-\sum_j p_{kj} \log_2 p_{kj}}{\log_2 N_k}$ quantifies intra-class fragmentation with $p_{kj} = a_{kj}/A_k$ and region count N_k; and $\mathcal{C}_{\text{topo}} = \min(1.0, \frac{2|E|}{|V|(|V|-1)})$ evaluates adjacency graph density. Based on $\mathcal{D}$, Task 3 is stratified into Easy (6,492 samples), Medium (2,705 samples), and Hard (1,624 samples), ensuring a granular assessment across complex landscapes.

Table 3: Comparison of NeSy-Route with existing remote sensing benchmarks.

Dataset	Key Statistic	Evaluation Paradigm			Construction Pipeline	
	Volume	Planning	Hierarchical	Symbolic	GT Optimality	Automated
RSVQA [24]	-	✗	✓	✗	✓	✓
RSIVQA [49]	-	✗	✗	✗	✓	✓
EarthVQA [39]	-	✗	✗	✗	✓	✗
LRSVQA [26]	-	✗	✓	✗	✗	✗
VRS-Bench [20]	-	✗	✗	✗	✗	✗
FIT-RSRC [25]	-	✗	✓	✗	✗	✗
XLRS-Bench [37]	1,130	✓	✓	✗	✓	✗
NeSy-Route	**10,821**	✓	✓	✓	✓	✓

4.2 Symbolized Evaluator

A typical example in NeSy-Route is shown in Fig. 1. To quantitatively evaluate the performance across three tasks, we employ a comprehensive set of metrics to assess the model's textual constraint understanding, text-image constraint alignment, and the feasibility and optimality of the generated trajectories.

Task 1: Textual Constraint Understanding For Task 1, we employ three primary metrics to assess the precision of symbolic constraint extraction from the query Q. First, the Traversability Matching (TM) measures the exact alignment of the predicted binary traversability vector with the ground truth:

$$TM = (1/N) \sum_{k=1}^{N} \mathbb{I}(\mathbf{V}_{\text{trav},k} = \mathbf{V}^{gt}_{\text{trav},k}) \tag{9}$$

where $\mathbb{I}$ is the indicator function and N is the total number of samples. Second, the Preference Ranking Correlation (PR) assesses the ordinal alignment of the preference vector. Let $M = \|\mathbf{V}^{gt}_{\text{trav}}\|_1$ denote the number of land cover classes permissible for navigation. We adopt the Kendall Tau coefficient [14] as the primary indicator, whcih has been used in many tasks [9, 11, 13]:

$$PR = (N_c - N_d)/\binom{M}{2} \tag{10}$$

where N_c and N_d are scalar counts representing the concordant and discordant pairs among the $\binom{M}{2}$ possible combinations. Note that PR is strictly evaluated only for terrains where $V^{gt}_{\text{trav}}[i] = 1$ to ensure the ranking reflects the optimization logic among traversable surfaces. Finally, the Fully Matching Accuracy (FM) requires simultaneous mastery of both traversability rules and preference rankings.:

$$FM = (1/N) \sum_{k=1}^{N} \mathbb{I}((\mathbf{V}_{\text{trav},k} = \mathbf{V}^{gt}_{\text{trav},k}) \wedge (\mathbf{V}_{\text{pref},k} = \mathbf{V}^{gt}_{\text{pref},k})) \tag{11}$$

This allows for a granular analysis of how models internalize the symbolic logic embedded in textual constraints.

Task 2: Text–Image Constraint Alignment For Task 2, we evaluate the text-image constraint alignment using the Region Matching Rate (RM), TM, and PR. The RM measures the global accuracy of region identification:

$$RM = (1/(N \cdot L)) \sum_{k=1}^{N} \sum_{i=0}^{L-1} \mathbb{I}(\mathbf{V}'_{\text{reg},k}[i] = \mathbf{V}^{gt}_{\text{reg},k}[i]) \tag{12}$$

The TM and PR metrics are applied to the localized vectors $\mathbf{V}'_{\text{trav}}$ and $\mathbf{V}'_{\text{pref}}$ following Eq. (9) and Eq. (10). This allows for the precise identification of failure bottlenecks within the cross-modal reasoning process.

Task 3: Constrained Route Planning For Task 3, we evaluate the global trajectory planning by reconstructing a continuous dense path $\hat{\tau}$ from the predicted sparse waypoints τ. Let $S_A = \{k \in \{1, \ldots, N\} \mid \forall p \in \tau_k, W(p) < \infty\}$ denote the subset of samples where all waypoints reside in traversable regions, and $S_B = \{1, \ldots, N\} \setminus S_A$ be the complementary subset of non-compliant samples. The Adherence Rate (AR) is thus defined as:

$$AR = |S_A|/N \tag{13}$$

For samples in S_A, the dense trajectory $\hat{\tau}$ is reconstructed by connecting adjacent waypoints using the A-Star search algorithm based on the cost map $W(p)$. For samples in S_B, we utilize the Bresenham line algorithm [5] for interpolation. The Cost Ratio (CR) is calculated for S_A to assess the optimality relative to τ^*:

$$CR = (1/|S_A|) \sum_{k \in S_A} (\sum_{p \in \hat{\tau}_k} W(p) / \sum_{q \in \tau^*_k} W(q)) \tag{14}$$

Conversely, for the non-compliant samples in S_B, the Violation Ratio (VR) quantifies the proportion of pixels in the reconstructed path that infringe upon non-traversable regions:

$$VR = (1/|S_B|) \sum_{k \in S_B} (|\{p \in \hat{\tau}_k \mid W(p) = \infty\}|/|\hat{\tau}_k|) \tag{15}$$

Finally, to evaluate the geometric proximity between the reconstructed dense trajectory $\hat{\tau}$ and the optimal trajectory τ^*, we utilize the Chamfer Distance [3] (CD), a standard metric widely adopted in autonomous driving [30, 33, 46] for trajectory comparison, across all N samples:

$$CD(\tau, \tau^*) = (1/N) \sum_{k=1}^{N} [(1/|\tau_k|) \sum_{p \in \hat{\tau}_k} \min_{q \in \tau^*_k} \|p - q\|_2 + (1/|\tau^*_k|) \sum_{q \in \tau^*_k} \min_{p \in \tau_k} \|q - p\|_2] \tag{16}$$

This multi-dimensional metrics system enables a rigorous quantification of both logical adherence and spatial optimality in trajectory evaluation.

Table 4: Performance comparison on Task 1: Textual Constraint Understanding. Metrics include Traversability Matching (TM), Preference Ranking Correlation (PR), and Fully Matching Accuracy (FM). The best and second-best results are highlighted in bold and underlined, respectively.

Method	**TM↑**	**PR↑**	**FM↑**
Close-source Models			
GPT-5.1	77.02	<u>0.959</u>	<u>69.20</u>
Gemini-3-Pro	**98.34**	**0.962**	**92.24**
Qwen3-VL-Plus	72.97	0.923	66.23
Open-source Models			
LLaVA-OneVision	26.68	0.187	9.32
InternVL-3.5-8B	52.48	0.141	21.18
Qwen3-VL-8B	53.01	0.068	27.67
Qwen3-VL-30B-A3B	66.20	0.752	44.44
Qwen3-VL-32B	69.75	-0.616	31.69
Qwen3-VL-235B-A22B	72.91	0.872	52.98
Qwen3.5-27B	<u>80.32</u>	0.906	63.52

4.3 Comparison with Existing Benchmarks

Leveraging the comparative analysis in Tab. 3, NeSy-Route distinguishes itself through three key innovations. First, it significantly expands the scale of remote sensing planning tasks, achieving an order-of-magnitude increase over XLRS-Bench [37] with 10,821 rigorously constrained samples. Second, its unique hierarchical and symbolic evaluation paradigm decouples perception, reasoning, and planning, enabling researchers to precisely trace failure bottlenecks to specific cognitive stages. Finally, by integrating high-fidelity semantic segmentation with heuristic search, our automated data generation framework ensures that every ground-truth trajectory represents a mathematically proven global optimum, establishing an objective and highly extensible gold standard for route planning.

5 Experiment

5.1 Experimental Setup

To evaluate the reasoning and planning capabilities of MLLMs on the NeSy-Route benchmark, we evaluate a diverse suite of state-of-the-art models grouped into two primary categories: (a) proprietary frontier models representing the current performance ceiling, including GPT-5.1 [34], Gemini-3-Pro [7], and Qwen3-VL-Plus [2]; (b) open-source MLLMs spanning various parameter scales and architectures to investigate performance in route planning tasks, including the Qwen3-VL series (8B-Instruct, 30B-A3B-Instruct, 32B-Instruct, and 235B-A22B-Instruct) [2], Qwen-3.5-27B [31], InternVL-3.5-8B [41], and LLaVA-OneVision

Table 5: Evaluation results for Task 2: Text–Image Constraint Alignment. The model performance is assessed via Region Matching Rate (RM), Traversability Matching (TM), and Preference Ranking Correlation (PR). The best and second-best results are indicated in bold and underlined, respectively.

Method	Easy			Medium			Hard			Avg		
	RM↑	TM↑	PR↑	RM↑	TM↑	PR↑	RM↑	TM↑	PR↑	RM↑	TM↑	PR↑
Close-source Models												
GPT-5.1	69.79	66.37	0.737	48.90	22.84	<u>0.567</u>	50.09	16.72	0.510	56.26	35.31	0.605
Gemini-3-Pro	**78.16**	58.14	0.582	**63.17**	10.78	0.402	**71.70**	13.59	0.404	**71.01**	27.50	0.463
Qwen3-VL-Plus	<u>72.44</u>	<u>69.50</u>	**0.778**	46.58	25.85	0.582	53.56	16.67	0.541	57.53	37.34	<u>0.634</u>
Open-source Models												
LLaVA-OneVision	24.75	24.27	-0.001	16.41	9.93	0.064	12.84	9.36	0.157	18.00	14.52	0.073
InternVL-3.5-8B	57.99	56.53	0.011	35.48	16.35	0.052	32.69	7.56	0.117	42.05	26.81	0.060
Qwen3-VL-8B	33.21	41.89	-0.107	16.88	11.97	-0.024	15.95	13.07	0.086	22.01	22.31	-0.015
Qwen3-VL-30B-A3B	72.27	65.16	0.280	51.00	<u>40.55</u>	0.365	48.66	<u>22.65</u>	<u>0.617</u>	57.31	<u>42.79</u>	0.421
Qwen3-VL-32B	67.31	62.62	-0.190	37.48	22.15	-0.176	39.26	21.04	-0.154	48.02	35.27	-0.173
Qwen3-VL-235B-A22B	64.25	**71.38**	<u>0.749</u>	33.81	**43.80**	**0.667**	32.92	**43.20**	**0.690**	43.66	**52.79**	**0.702**
Qwen3.5-27B	71.80	63.95	0.576	<u>52.13</u>	16.04	0.483	<u>57.74</u>	19.48	0.505	<u>60.56</u>	33.16	0.521

[1]. For a rigorous and fair comparison, Task 1 is evaluated in a few-shot setting with fixed format demonstrations, whereas Tasks 2 and 3 are evaluated in a zero-shot setting with only output-format instructions, ensuring that the results reflect models' inherent perception, reasoning, and planning capabilities. During the experiments we observe that both Geochat [15] and SkySenseGPT [8] are built upon the LLaVA-1.5 [22] architecture. Neither model was trained with data requiring individual coordinate points as outputs, and both were designed for specific remote sensing tasks. As a result, their instruction-following capability is extremely limited, making them unable to complete the evaluation process of NeSy-Route. Therefore, the performance of these domain-specific models is not reported in our experimental results.

5.2 Main Results

Task 1 reveals a clear performance gradient across models in textual constraint understanding. The results summarized in Tab. 4 further illustrate this trend across the evaluated models. Close-source Models, exemplified by Gemini-3-Pro, demonstrate constraint understanding capability. This indicates that frontier models are highly capable of precisely parsing nested hard constraints for heterogeneous agents and constructing rigorous logical chains of priority. Concurrently, Qwen-3.5-27B stands out as a representative open-source model, attaining a TM of 80.32%, even surpassing certain closed-source counterparts which validates the constraint understanding capabilities.

Task 2 indicates that effective text–image constraint alignment remains challenging for current MLLMs and critically relies on strong

Table 6: Experimental results on Task 3: Constrained Route Planning. The trajectory quality is measured by Adherence Rate (AR), Violation Ratio (VR), Cost Ratio (CR), and Chamfer Distance(CD). The top two methods are denoted by bold and underlined formats, respectively.

Method	Easy				Medium				Hard				Avg			
	AR↑	VR↓	CR↓	CD↓	AR↑	VR↓	CR↓	CD↓	AR↑	VR↓	CR↓	CD↓	AR↑	VR↓	CR↓	CD↓
Close-source Models																
GPT-5.1	25.90	31.70	1.47	66.71	14.90	35.8	2.71	74.79	12.10	38.90	1.32	82.12	17.63	35.47	1.83	74.54
Gemini-3-Pro	27.40	**28.30**	**1.20**	58.09	15.40	**32.70**	**1.25**	70.17	13.30	**35.90**	1.28	77.96	18.70	**32.30**	**1.24**	68.74
Qwen3-VL-Plus	42.80	35.30	7.16	**53.78**	25.80	37.80	6.61	**63.36**	20.40	39.60	2.49	**68.60**	29.67	37.57	5.42	**61.91**
Open-source Models																
LLaVA-OneVision	37.64	39.47	12.42	103.76	21.11	38.24	9.24	122.91	16.26	37.54	13.63	160.41	25.00	38.42	35.29	129.03
InternVL-3.5-8B	34.80	36.50	14.82	109.88	19.40	38.70	15.06	127.07	15.10	41.30	21.38	146.55	23.10	38.83	17.09	127.83
Qwen3-VL-8B	40.30	37.20	13.58	170.86	24.20	39.20	17.68	181.78	18.40	41.60	10.64	232.09	27.63	39.33	13.97	194.91
Qwen3-VL-30B-A3B	36.70	35.70	9.36	95.35	21.30	38.50	8.36	107.58	16.90	40.70	6.12	130.98	24.97	38.30	7.95	111.30
Qwen3-VL-32B	**46.00**	36.50	7.39	62.60	**28.80**	38.40	14.86	75.36	**23.30**	40.40	16.35	75.60	**32.70**	38.43	12.87	71.19
Qwen3-VL-235B-A22B	40.90	35.50	10.78	58.44	24.00	37.30	10.14	64.91	18.50	39.10	5.19	69.50	27.80	37.30	8.70	64.28
Qwen3.5-27B	30.10	31.90	2.29	67.81	16.50	35.70	6.96	86.89	13.70	38.30	**1.26**	184.52	20.10	35.30	10.51	113.07

textual reasoning capabilities. Compared with performance in Task 1, all models experience a precipitous decline in performance upon the introduction of visual features, as shown in Tab. 5, which underscores that while MLLMs may internalize textual constraint, their perception ability remains remarkably deficient. Although Gemini-3-Pro demonstrated the highest level of visual identification with RM of 71.01%, it lagged significantly behind Qwen3-VL-Plus. This misalignment between visual perception and constraint validates the necessity of decoupled evaluation in Task 2. Notably, the Mixture-of-Experts (MoE) based Qwen3-VL-235B-A22B exhibited exceptional performance in logical alignment, with an average PR of 0.702 and TM of 52.79%, both substantially surpassing GPT-5.1. Furthermore, empirical data show that model with poor logic parsing in Task 1 like LLaVA-OneVision maintains extremely low performance in Task 2, confirming that textual constraint understanding is the cornerstone of text-image constraint alignment.

Task 3 demonstrates that constrained route planning remains a major challenge for current MLLMs, revealing a clear gap between perceptual understanding and effective planning solutions. As shown in Tab. 6, closed-source models such as Gemini-3-Pro exhibit a distinct pattern of high-precision planning. Although the overall adherence rate of 18.70% for Gemini-3-Pro is not the highest due to rigorous logical alignment requirements, its performance on compliant paths demonstrates strong optimality. Specifically, the average CR of 1.24 and CD of 68.74 for this model are the best compared with the optimal trajectory. In contrast, open-source models represented by Qwen3-VL-32B achieve a leading adherence rate of 32.70%, demonstrating robust awareness of local obstacle avoidance. However, the average CR for these models remains at an elevated level of 12.87. This disparity suggests that while open-source models are often successful at remaining within traversable regions, they lack coherent global strategies, resulting in highly redundant and ineffi-

cient trajectories. Furthermore, performance metrics deteriorate significantly as environmental complexity increases from easy to hard levels. For example, the adherence rate of GPT-5.1 decreases from 25.90% to 12.10% as the violation ratio increases. These results confirm that path planning is a high-level cognitive capability that extends beyond simple semantic perception and reasoning. Even with a foundation in terrain perception and reasoning, models still struggle to produce complex and efficient solutions for route planning.

Hierarchical Analysis Correlation analysis across the three integrated tasks confirms that strong perception and reasoning abilities do not necessarily translate into planning ability. The experimental results demonstrate that comprehensive perception and reasoning abilities constitute a necessary but insufficient condition for the successful execution of complex planning tasks. A deficiency in textual constraint understanding during the first task leads to certain failure in subsequent stages, whereas strong performance in the first and second tasks does not ensure success in the third task. This disparity indicates that models must bridge a fundamental cognitive gap between perceptual attribute recognition and constrained planning.

Discussions Our experiments reveal two limitations of current MLLMs: 1)They struggle to incorporate land-type constraints, likely because existing training data emphasizes object recognition while neglecting land-type textures and geological characteristics. 2)Current remote sensing MLLMs rely on outdated architectures, limiting their ability to handle complex reasoning and planning tasks, which necessitates the development of powerful remote sensing models based on more advanced architectures.

6 Conclusion

In this paper, we introduce NeSy-Route, a large-scale neuro-symbolic benchmark designed to evaluate constrained route planning in remote sensing. The benchmark assesses the perception, reasoning and planning capabilities of MLLMs simultaneously through three hierarchical stages comprising textual constraint understanding, the text to image constraint alignment Task, and the constrained route planning Task. Leveraging high-fidelity semantic segmentation masks and heuristic search algorithms, we construct an algorithmically derived optimal route for each sample, which serves as an objective reference for evaluating model performance. Experimental results indicate that while current MLLMs demonstrate competence in pure constraint understanding, they face significant bottlenecks in text-image preception and planning capabilities. NeSy-Route focuses on 2D semantic route planning over remote-sensing imagery, leaving agentic planning workflows and real-world physical interaction to future work. We hope NeSy-Route can support further research on MLLMs that better integrate perception, reasoning, and planning in complex geographic environments.

Acknowledgements

This research was supported by the Jiangsu Science Foundation (BG2024036 and BK20243012) and the Natural Science Foundation of China (624B2068, 62576162, and 62576174). It was also supported by the Fundamental Research Funds for the Central Universities (022114380023).

References

1. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Wu, C., Tan, H., Li, C., Yang, J., Yu, J., Wang, X., Qin, B., Wang, Y., Yan, Z., Feng, Z., Liu, Z., Li, B., Deng, J.: Llava-onevision-1.5: Fully open framework for democratized multimodal training. arXiv preprint arXiv:2509.23661 (2025)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Barrow, H.G., Tenenbaum, J.M., Bolles, R.C., Wolf, H.C.: Parametric correspondence and chamfer matching: Two new techniques for image matching. In: IJCAI. pp. 659–663 (1977)
4. Boguszewski, A., Batorski, D., Ziemba-Jankowska, N., Dziedzic, T., Zambrzycka, A.: Landcover. ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1102–1110 (2021)
5. Bresenham, J.E.: Algorithm for computer control of a digital plotter. IBM Syst. J. **4**(1), 25–30 (1965). https://doi.org/10.1147/sj.41.0025
6. Dell'Acqua, F., Gamba, P.: Remote sensing and earthquake damage assessment: Experiences, limits, and perspectives. Proc. IEEE **100**(10), 2876–2890 (2012)
7. Google DeepMind: Gemini 3 pro: Model card. Model Card (2025), https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf, accessed: 2026-01-28
8. Guo, X., Lao, J., Dang, B., Zhang, Y., Yu, L., Ru, L., Zhong, L., Huang, Z., Wu, K., Hu, D., He, H., Wang, J., Chen, J., Yang, M., Zhang, Y., Li, Y.: Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 27672–27683 (June 2024)
9. Hamed, K.: Effect of persistence on the significance of kendall's tau as a measure of correlation between natural time series. The European Physical Journal Special Topics **174**(1), 65–79 (2009)
10. Hart, P.E., Nilsson, N.J., Raphael, B.: A formal basis for the heuristic determination of minimum cost paths. IEEE Trans. Syst. Sci. Cybern. **4**(2), 100–107 (1968)
11. He, X., Wang, J., Su, Y., Liu, D., Zhao, J., Lu, G.: Monotonic rank knowledge distillation via kendall correlation. IEEE Trans. Circuit Syst. Video Technol. (2026)
12. Hu, R., Bai, S., Wen, W., Xia, X., Hsu, L.T.: Towards high-definition vector map construction based on multi-sensor integration for intelligent vehicles: Systems and error quantification. IET Intell. Transp. Syst. **18**(8), 1477–1493 (2024)
13. Huang, T., Niu, S., Zhang, F., Wang, B., Wang, J., Liu, G., Yao, M.: Correlating gene expression levels with transcription factor binding sites facilitates identification of key transcription factors from transcriptome data. Frontiers in Genetics **15**, 1511456 (2024)

14. Kendall, M.G.: A new measure of rank correlation. Biometrika **30**(1-2), 81–93 (1938)
15. Kuckreja, K., Danish, M.S., Naseer, M., Das, A., Khan, S., Khan, F.S.: Geochat: Grounded large vision-language model for remote sensing. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 27831–27840 (2024)
16. Lee, D.H., Hong, J.H., Seo, H.W., Oh, H.: Kfgod: A fine-grained object detection dataset in kompsat satellite imagery. Remote Sens. **17**(22), 3774 (2025)
17. Li, K., Dong, F., Wang, D., Li, S., Wang, Q., Gao, X., Chua, T.S.: Show me what and where has changed? question answering and grounding for remote sensing change detection. arXiv preprint arXiv:2410.23828 (2024)
18. Li, K., Wan, G., Cheng, G., Meng, L., Han, J.: Object detection in optical remote sensing images: A survey and a new benchmark. ISPRS J. Photogramm. Remote Sens. **159**, 296–307 (2020)
19. Li, T., Wang, C., de Silva, C.W., et al.: Coverage sampling planner for uav-enabled environmental exploration and field mapping. In: IEEE/RSJ Int. Conf. Intell. Robots Syst. pp. 2509–2516. IEEE (2019)
20. Li, X., Ding, J., Elhoseiny, M.: Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. Adv. Neural Inform. Process. Syst. **37**, 3229–3242 (2024)
21. Liu, A., Mei, A., Lin, B., Xue, B., Wang, B., Xu, B., Wu, B., Zhang, B., Lin, C., Dong, C., et al.: Deepseek-v3. 2: Pushing the frontier of open large language models. arXiv preprint arXiv:2512.02556 (2025)
22. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 26296–26306 (2024)
23. Liu, R., Fu, B., Song, J., Li, K., Li, W., Xue, L., Qiao, H., Zhang, W., Meng, D., Cao, X.: Zoomearth: Active perception for ultra-high-resolution geospatial vision-language tasks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 34877–34888 (2026)
24. Lobry, S., Marcos, D., Murray, J., Tuia, D.: Rsvqa: Visual question answering for remote sensing data. IEEE Trans. Geosci. Remote Sens. **58**(12), 8555–8566 (2020)
25. Luo, J., Pang, Z., Zhang, Y., Wang, T., Wang, L., Dang, B., Lao, J., Wang, J., Chen, J., Tan, Y., et al.: Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. arXiv preprint arXiv:2406.10100 (2024)
26. Luo, J., Zhang, Y., Yang, X., Wu, K., Zhu, Q., Liang, L., Chen, J., Li, Y.: When large vision-language model meets large remote sensing imagery: Coarse-to-fine text-guided token pruning. In: Int. Conf. Comput. Vis. pp. 9206–9217 (2025)
27. Lv, S.L., Chen, Y.Y., Zhou, Z., Yang, M., Guo, L.Z.: Bmip: Bi-directional modality interaction prompt learning for vlm. arXiv preprint arXiv:2501.07769 (2025)
28. Marsocci, V., Jia, Y., Bellier, G.L., Kerekes, D., Zeng, L., Hafner, S., Gerard, S., Brune, E., Yadav, R., Shibli, A., et al.: Pangaea: A global and inclusive benchmark for geospatial foundation models. arXiv preprint arXiv:2412.04204 (2024)
29. Muhtar, D., Li, Z., Gu, F., Zhang, X., Xiao, P.: Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In: Eur. Conf. Comput. Vis. pp. 440–457. Springer (2024)
30. Palladin, E., Brucker, S., Ghilotti, F., Narayanan, P., Bijelic, M., Heide, F.: Self-supervised sparse sensor fusion for long range perception. In: Int. Conf. Comput. Vis. pp. 27498–27509 (2025)
31. Qwen Team: Qwen3.5: Towards native multimodal agents (February 2026), `https://qwen.ai/blog?id=qwen3.5`, accessed: 2026-06-26

32. Shao, J.J., Zhang, B.W., Yang, X.W., Chen, B., Han, S., Wei, W.D., Cai, G., Dong, Z., Guo, L.Z., Li, Y.F.: Chinatravel: An open-ended benchmark for language agents in chinese travel planning. In: Workshop on Scaling Environments for Agents (2025)
33. Shi, G., Wu, C.C., Hsu, C.H.: Error concealment of dynamic lidar point clouds for connected and autonomous vehicles. In: IEEE Glob. Commun. Conf. pp. 5409–5415. IEEE (2023)
34. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
35. Tian, S.Y., Zhou, Z., Dong, W., Yu, K.Y., Yang, M., Cheng, Z.J., Guo, L.Z., Li, Y.F.: Tabularmath: Understanding math reasoning over tables with large language models. arXiv preprint arXiv:2505.19563 (2025)
36. Tian, S.Y., Zhou, Z., Yu, K.Y., Yang, M., Chen, Y., Shang, Z., Guo, L.Z., Li, Y.F.: Last: Leveraging tools as hints to enhance spatial reasoning for multimodal large language models. arXiv preprint arXiv:2604.09712 (2026)
37. Wang, F., Wang, H., Guo, Z., Wang, D., Wang, Y., Chen, M., Ma, Q., Lan, L., Yang, W., Zhang, J., et al.: Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14325–14336 (2025)
38. Wang, J., Xuan, W., Qi, H., Liu, Z., Liu, K., Wu, Y., Chen, H., Song, J., Xia, J., Zheng, Z., et al.: Disasterm3: A remote sensing vision-language dataset for disaster damage assessment and response. arXiv preprint arXiv:2505.21089 (2025)
39. Wang, J., Zheng, Z., Chen, Z., Ma, A., Zhong, Y.: Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In: AAAI. vol. 38, pp. 5481–5489 (2024)
40. Wang, J., Zheng, Z., Ma, A., Lu, X., Zhong, Y.: Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. arXiv preprint arXiv:2110.08733 (2021)
41. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
42. Xia, G.S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X.: Aid: A benchmark data set for performance evaluation of aerial scene classification. IEEE Trans. Geosci. Remote Sens. **55**(7), 3965–3981 (2017)
43. Xia, J., Yokoya, N., Adriano, B., Broni-Bediako, C.: Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In: IEEE/CVF Winter Conf. Appl. Comput. Vis. pp. 6254–6264 (2023)
44. Yang, X.W., Shao, J.J., Guo, L.Z., Zhang, B.W., Zhou, Z., Jia, L.H., Dai, W.Z., Li, Y.F.: Neuro-symbolic artificial intelligence: Towards improving the reasoning abilities of large language models. arXiv preprint arXiv:2508.13678 (2025)
45. Yu, K.Y., Zhou, Z., Tian, S.Y., Yang, X.W., Jia, Z.Y., Yang, M., Cheng, Z.J., Guo, L.Z., Li, Y.F.: Thinking with tables: Enhancing multi-modal tabular understanding via neuro-symbolic reasoning. arXiv preprint arXiv:2603.24004 (2026)
46. Zhang, L., Xiong, Y., Yang, Z., Casas, S., Hu, R., Urtasun, R.: Copilot4d: Learning unsupervised world models for autonomous driving via discrete diffusion. arXiv preprint arXiv:2311.01017 (2023)
47. Zhang, W., Cai, M., Zhang, T., Zhuang, Y., Mao, X.: Earthgpt: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. IEEE Trans. Geosci. Remote Sens. **62**, 1–20 (2024)

48. Zhang, X., Sun, Y., Shang, K., Zhang, L., Wang, S.: Crop classification based on feature band set construction and object-oriented approach using hyperspectral images. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. **9**(9), 4117–4128 (2016)
49. Zheng, X., Wang, B., Du, X., Lu, X.: Mutual attention inception network for remote sensing visual question answering. IEEE Trans. Geosci. Remote Sens. **60**, 1–14 (2021)
50. Zhou, Z., Guo, L.Z., Jia, L.H., Zhang, D., Li, Y.F.: Ods: Test-time adaptation in the presence of open-world data shift. In: Int. Conf. Mach. Learn. pp. 42574–42588. PMLR (2023)
51. Zhou, Z., Yang, M., Shi, J.X., Guo, L.Z., Li, Y.F.: Decoop: Robust prompt tuning with out-of-distribution detection. arXiv preprint arXiv:2406.00345 (2024)
52. Zhou, Z., Yuhao, T., Li, Z., Yao, Y., Guo, L.Z., Li, Y.F., Ma, X.: A theoretical study on bridging internal probability and self-consistency for llm reasoning. Adv. Neural Inform. Process. Syst. **38**, 87380–87413 (2026)
53. Zhu, Z., Zhou, Y., Seto, K.C., Stokes, E.C., Deng, C., Pickett, S.T., Taubenböck, H.: Understanding an urbanizing planet: Strategic directions for remote sensing. Remote Sens. Environ. **228**, 164–182 (2019)