This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Geo3DVQA: Evaluating Vision-Language Models for 3D Geospatial Reasoning from Aerial Imagery

Mai Tsujimoto[1] Junjue Wang[1] Weihao Xuan[1,2] Naoto Yokoya[1,2,†]
[1]The University of Tokyo [2]RIKEN AIP †Corresponding author

Abstract

Three-dimensional geospatial analysis is critical to applications in urban planning, climate adaptation, and environmental assessment. Current methodologies depend on costly, specialized sensors (e.g., LiDAR and multispectral), which restrict global accessibility. Existing sensor-based and rule-driven methods further struggle with tasks requiring the integration of multiple 3D cues, handling diverse queries, and providing interpretable reasoning. We hereby present Geo3DVQA, a comprehensive benchmark for evaluating vision-language models (VLMs) in height-aware, 3D geospatial reasoning using RGB-only remote sensing imagery. Unlike conventional sensor-based frameworks, Geo3DVQA emphasizes realistic scenarios that integrate elevation, sky view factors, and land cover patterns. The benchmark encompasses 110k curated question–answer pairs spanning 16 task categories across three complexity levels: single-feature inference, multi-feature reasoning, and application-level spatial analysis. The evaluation of ten state-of-the-art VLMs highlights the difficulty of RGB-to-3D reasoning. GPT-4o and Gemini-2.5-Flash achieved only 28.6% and 33.0% accuracy respectively, while domain-specific fine-tuning of Qwen2.5-VL-7B achieved 49.6% (+24.8 points). These results reveal both the limitations of current VLMs and the effectiveness of domain adaptation. Geo3DVQA introduces a new frontier of challenges for scalable, accessible, and holistic 3D geospatial analysis. The dataset and code will be released upon publication at https://github.com/mm1129/Geo3DVQA.

1. Introduction

Recent advancements in vision-language models (VLMs) [1–3] have exhibited noteworthy capabilities in two-dimensional (2D) visual understanding. However, their potential for three-dimensional (3D) spatial reasoning from widely available RGB imagery remains largely unexplored. This gap is particularly important given the growing importance of 3D geospatial analysis in addressing contemporary global challenges.

The Intergovernmental Panel on Climate Change (IPCC) projects an increasing frequency and intensity of climate extremes under global warming scenarios [4]. These risks are particularly amplified in urban areas, where three-dimensional morphology drives local microclimates and intensifies Urban Heat Island effects [5, 6]. Accurate assessment of spatial properties, such as sky view factors, building heights, and urban morphology, is essential for applications ranging from heat mitigation to disaster risk assessment, energy planning, and quality-of-life improvements.

Nevertheless, current approaches to three-dimensional geospatial analyses face significant accessibility barriers. Traditional methods depend on costly specialized sensors such as LiDAR (often costing $100,000 or more) and multispectral sensors [7–9]. Beyond acquisition costs, these approaches require considerable technical expertise for sensor operation, data processing workflows, and cross-feature integration. These limitations restrict advanced geospatial reasoning to well-funded institutions, creating a substantial accessibility gap in areas where climate adaptation analyses are most urgently needed.

Conversely, aerial RGB imagery is widely accessible through platforms such as Google Earth and national mapping agencies at minimal or no cost. This contrast naturally leads to the question: *Can vision-language models perform sophisticated 3D spatial reasoning directly from ubiquitous RGB imagery alone?* We studied their ability to learn and infer imperceptible 3D properties directly from the observable RGB patterns. Among these properties, the Sky View Factor (SVF) serves as a fundamental metric: it represents the ratio (0–1) of the visible sky from a point and is essential for urban microclimate and solar radiation analysis [10].

To advance the development of VLMs with robust 3D spatial reasoning, we introduce **Geo3DVQA**, a dedicated VQA (Visual Question Answering) benchmark specifically designed to evaluate height-aware spatial reasoning from aerial RGB imagery alone. As shown in Figure 1, our framework systematically evaluates the models' ability to extract 3D spatial properties from 2D visual cues. Here we employed a concise three-tier taxonomy (single-feature, multi-feature, and application-level).

Geo3DVQA focuses on coarse, decision-oriented, height-

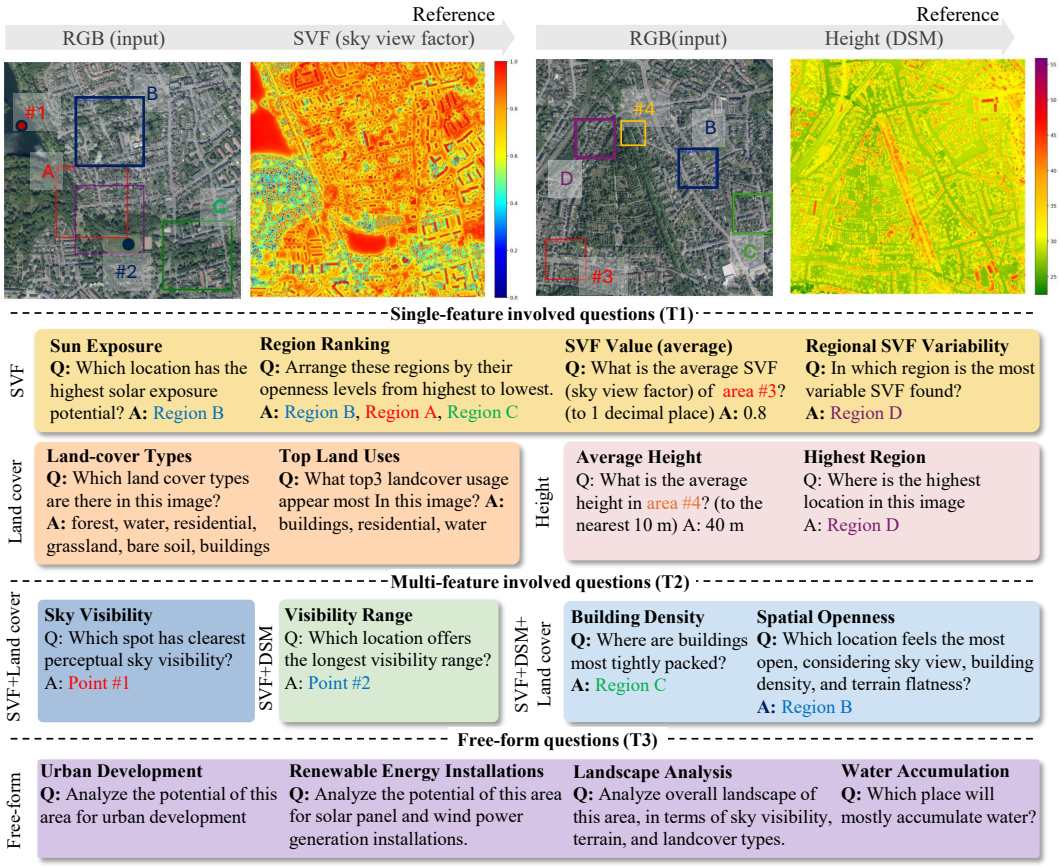


Figure 1. **Geo3DVQA Task Taxonomy and Evaluation Framework.** Our benchmark comprises 16 task categories organized into three complexity tiers to evaluate RGB-to-3D spatial reasoning capabilities. **Tier 1 (Single-feature)**: Direct inference of individual properties across SVF, land cover, and height from RGB patterns. **Tier 2 (Multi-feature)**: Composite reasoning tasks requiring the integration of multiple spatial attributes (e.g., sky visibility combines SVF with building penalties). **Tier 3 (Application-level)**: Free-form reasoning for real-world applications in urban planning, renewable energy, landscape analysis, and water accumulation. Ground truth generation uses multimodal reference data (SVF, DSM, land cover), whereas the models are evaluated using only RGB imagery. The colored regions indicate the locations of multiple-choice options, with answers highlighted for visualization.

aware reasoning (e.g., relative comparisons and bin-level distinctions). It is designed to be realistic for RGB-only inputs where precise 3D recovery from monocular RGB is inherently ill-posed, while coarse height distinctions remain feasible through learned 2D cues, such as shadows, occlusions, and perspective [8, 11–14].

Our main contributions are as follows:

1. We introduce Geo3DVQA, a dedicated benchmark for RGB-to-3D geospatial reasoning with a clear three-tier taxonomy, containing 110k carefully constructed question–answer pairs across **16** task categories.
2. We systematically evaluated ten leading VLMs and revealed fundamental limitations in height-aware spatial reasoning under RGB-only inputs.
3. We established baselines via domain-specific instruction tuning, achieving +24.8% accuracy with improvements in land cover and SVF tasks.

2. Related Work

2.1. Vision-Language Models for Remote Sensing and the Height-Aware Gap

Recent advances in vision-language models (VLMs) have enabled strong visual understanding and natural language reasoning across general-purpose tasks [1–3]. In remote sensing, large-scale resources have catalyzed progress in captioning, grounding, and VQA research. RSVQA [15] pioneered VQA (visual question answering) from RS data using OSM-derived (OpenStreetMap) supervision, HRVQA [16] scaled VQA to high-resolution imagery, and VRSBench [17] broadened evaluations beyond VQA to detailed captioning and visual grounding. RSIEval (RSGPT) [18] further supports evaluation of rich remote-sensing captions. DisasterM3 [19] targets bi-temporal disaster assessment with optical/SAR pairs, DynamicVL [20] benchmarks MLLMs for long-term

Dataset	Primary Focus	Scale
RSVQA [15]	VQA with OSM supervision (2D)	–
HRVQA [16]	High-resolution VQA (2D)	1.07M
VRSBench [17]	Versatile VL tasks (2D)	123K
RSIEval (RSGPT)	RS-VLM caption evaluation (2D)	–
GeoChat [22]	Unified instruction following	318K
TEOChat [21]	Temporal EO understanding	554K
DisasterM3 [19]	Disaster assessment (bi-temporal)	123K
Geo3DVQA (Ours)	**Height-aware 3D inference (RGB-only)**	110K QA

Table 1. Representative remote sensing benchmarks and our focus.

multi-temporal city understanding, and TEOChat [21] extends to temporal instruction following.

Systems such as GeoChat [22] and GEOBench-VLM [23] standardize RS instruction following and evaluation, and EarthVQA [24] addresses 2D geospatial relational reasoning. Despite these advances, existing datasets and models primarily focus on 2D recognition and relational reasoning rather than 3D remote sensing inference from 2D RGB inputs. This creates a significant gap in geospatial understanding, a challenge our study seeks to address.

2.2. Inferring 3D Properties from RGB

For monocular height estimation, a rich line of work estimates the height or DSM from single-view imagery. IM2HEIGHT [25] pioneered learning-based height regression from monocular images, while subsequent approaches leveraged shadow geometry and sun elevation for building height estimation [14] and extended to IM2ELEVATION-style pipelines for DSM/height prediction [13]. Most of these methods target continuous value map generation with explicit height supervision.

The SVF can be delivered either from DSM/LiDAR or street-view fisheye imagery with semantic segmentation [10, 26]. Estimating continuous SVF from RGB data alone is a challenge. Some attempts have been made to use satellite shadow cues [27], but robust inference generally relies on 3D data. Furthermore, practical applications require an integrated understanding of multiple spatial properties beyond isolated SVF estimation. This need motivated our investigation into VLMs' ability to infer height-aware properties from RGB patterns. Our evaluation bridges 2D semantic understanding and 3D geometric reasoning through a language-based assessment without specialized sensors.

3. Geo3DVQA Dataset

3.1. Dataset Construction and Composition

Geo3DVQA is built on GeoNRW [28], which provides orthorectified aerial imagery, LiDAR elevation models, and semantic segmentation for North Rhine-Westphalia, Germany. The GeoNRW dataset contains 7,783 image triplets that were curated at 1 m spatial resolution, covering 10 land

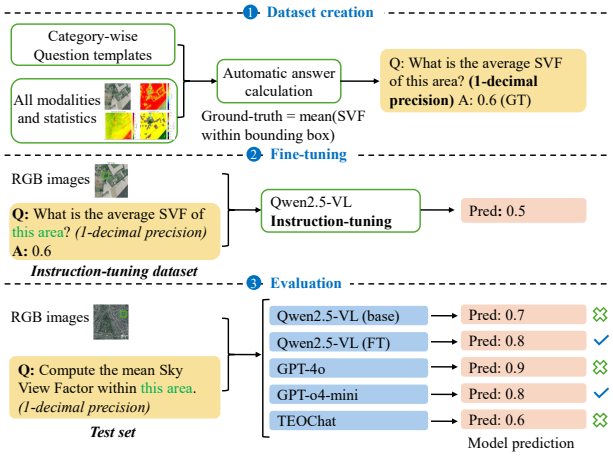


Figure 2. **Geo3DVQA Dataset Construction and Evaluation Pipeline.** (1) Dataset Creation: Question templates were generated for each category, with ground-truth answers automatically derived from multimodal data (SVF, DSM, land cover) and their statistics. (2) Fine-tuning: Qwen2.5-VL learns height-aware spatial reasoning from RGB inputs using the constructed dataset. (3) Evaluation: Models are tested using only RGB inputs, with predictions compared against the ground truth. The fine-tuned model leverages domain knowledge from step (2).

cover classes.

Spatial data processing. Digital Surface Models (DSM) provide elevation information vital for spatial analysis [29, 30]. From the DSM data, we computed the SVF values using the Urban Multi-scale Environmental Predictor (UMEP) toolkit with ray casting [10]. This approach is widely adopted in urban climate analysis and demonstrates strong agreement with fisheye imagery and high-resolution 3D models (R^2>0.91, RMSE $\approx$ 0.07) [26, 31, 32]. Our processing pipeline (Fig. 2) aligned these multimodal reference data (i.e., SVF, elevation, and segmentation data), extracted spatial indicators, and generated question–answer pairs.

Although the SVF mainly quantifies vertical openness, it can overestimate perceived openness because horizontal obstructions are not fully captured [6, 32, 33]. To address this limitation, we employed a composite *sky visibility* metric that integrates SVF with land cover information, providing a more realistic representation of perceived openness [34–36]. The reliability of these metrics and the derived results was checked via statistical validation and multimodal visual inspection.

Geo3DVQA comprises 110k QA pairs across 16 task categories. The question formats included multiple-choice, ranking, numerical estimation, and open-ended questions. Quality control ensures balanced answer distributions and spatial diversity (see Fig. 3).

Task Design and Generation. We organize tasks into three tiers. *T1* focuses on single-feature inference directly com-

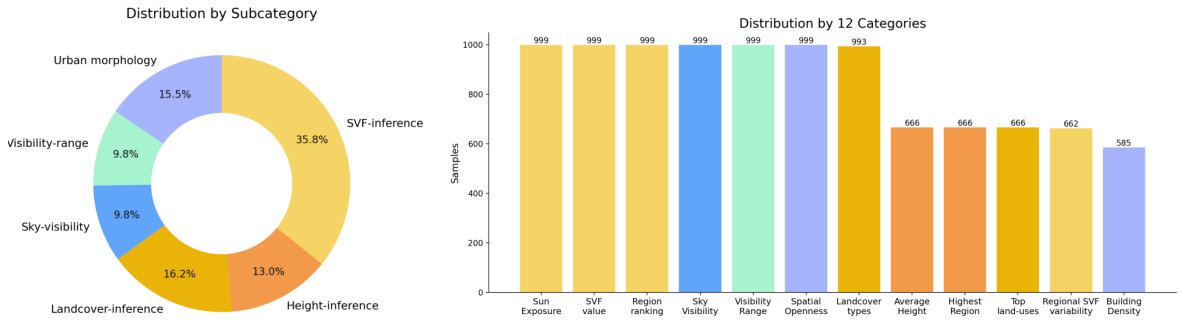


Figure 3. Task composition of Geo3DVQA evaluation set (left). Subcategory-level distribution (right panel). Distribution across 12 task categories. Tier 3 was excluded (116 questions in total).

putable from a single modality: SVF inference (sun exposure, region ranking, regional SVF variability, mean SVF), segmentation inference (land cover type, land use), and DSM inference (height average, highest region). *T2* requires multi-feature reasoning that integrates SVF, DSM, and land cover to produce a unified score for sky visibility, spatial openness, building density, and visibility range. Scoring adheres to established geospatial analysis methodologies from prior studies (e.g., [33, 34, 36–39]). For instance, the visibility range is computed from the viewshed, SVF, and terrain to balance sightline reach, sky access, and terrain blocking. Spatial openness combines SVF statistics, terrain flatness, and the complement of building density, which together reflect sky access, gentle relief, and low obstructions. The full details and scientific rationale for each metric, as well as the sensitivity study for weights and prompts, are provided in Appendix D.

T3 covers application-oriented free-form analyses (urban development, renewable energy, landscape, water), where answers are generated from statistics distilled from the reference modalities for QA construction. During the evaluation, the models received only RGB, and spatial queries used normalized coordinates (x%, y%) following TEOChat [21].

3.2. Dataset Creation Pipeline

Short-answer (T1/T2). After preparing the required inputs (SVF, DSM, and segmentation) and passing suitability checks, we generated questions from category-specific templates. Correct answers and distractors were sampled from a shared candidate pool to ensure balanced scores and spatial diversity. Option labels were randomized to avoid bias, and sampled Q&As were manually verified for correctness and scientific validity.

Free-form (T3). We compiled scene-level statistics from the reference modalities and used them to generate answers to rubric-constrained prompts with GPT-4.1-mini. The drafts were automatically checked against the statistics and then reviewed by human annotators, and only validated pairs were retained in the evaluation set. The templates use a structured format for observations that describe SVF, elevation, and land cover, as well as conclusions that provide actionable recommendations. Training and test separation are enforced by strict geographic splits.

Fig. 4 shows open-ended tasks where renewable energy assessments combine sky view factors, elevation, and land cover, whereas water accumulation focuses on low-elevation, low-SVF regions. Additional visualizations of the Tier-3 answer space, including a word cloud, are provided in Appendix C.

Supporting details. Appendix D provides the SVF computation settings and validation evidence, along with the scientific rationale for each metric. Appendix E details the QA construction pipeline and near-equal frequencies for each option label.

4. Benchmark Experiments

Implementation Setting. To comprehensively assess 3D spatial reasoning capabilities, we benchmarked ten VLMs. These cover open-source models (LLaVA-OneVision [2], InternVL3-8B [40], Qwen2.5-VL variants [3]), commercial systems (GPT-4o [41], o4-mini, GPT-4.1-mini, Gemini-2.5-Flash), and domain-specific remote sensing VLMs (TEOChat [21], GeoChat [22]). For fine-tuning, we investigated two configurations: a *10K* and a *100K* set of short-answer Q&As, both mixed with 1 K free-form Q&As.
Evaluation Metrics. For short-answer QA, we report per-category accuracy across the 12 tasks defined in T1+T2 [2]. Multi-label categories rely on Jaccard similarity with a 0.7 threshold, whereas height estimations require magnitude-aware tolerances to capture relative and absolute errors (see Appendix B for the threshold study). Free-form responses were scored on a 1–5 rubric in four application domains, guided by structured "observation" and "conclusion" outputs. The observation prompt explicitly asked for SVF, elevation,

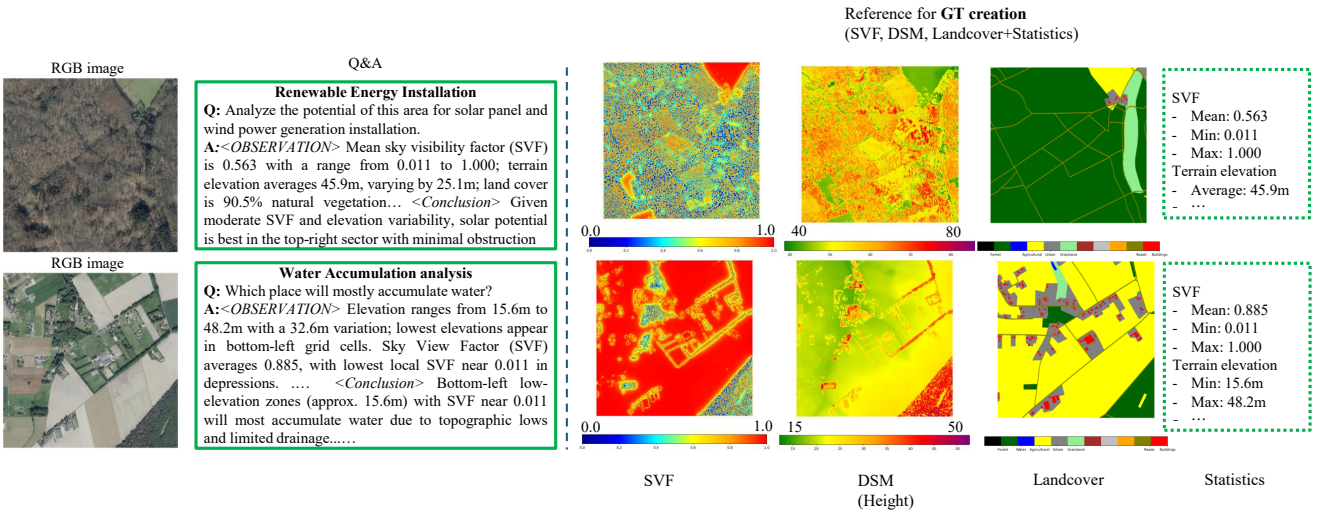


Figure 4. The shortened examples of Tier-3 free-form Q&A with **RGB input** and reference for ground truth. Statistics were first calculated from the reference modalities (SVF, DSM, segmentation) and input into GPT-4.1-mini to generate answers corresponding to pre-defined categories. These answers were validated by cross-checking them against the reference statistics using GPT-5. Q&As that passed the validation are then verified by humans, and the final Q&A pairs are used for evaluation.

Model	Short-answer (%)					Free-form (1–5)	
	SVF	Height	Land Use/LC	Multi	Overall	Total	Conclusion
• *Commercial models*							
*o4-mini	20.2	11.7	15.3	23.7	19.2	2.05	2.41
GPT-4o	30.7	27.8	17.5	32.4	28.6	2.27	2.57
GPT-4.1-mini	28.5	23.7	25.1	31.5	28.1	2.53	3.02
Gemini-2.5-Flash	32.9	19.4	40.0	35.6	33.0	1.96	2.34
• *Remote sensing VLMs*							
TEOChat	17.5	21.4	4.0	31.5	19.3	1.45	1.66
GeoChat	20.2	21.0	22.1	25.4	21.9	1.15	1.67
• *Open-source models*							
LLaVA-OneVision	22.4	24.5	21.6	19.3	21.8	1.48	2.01
InternVL3-8B	21.8	24.9	20.7	22.0	22.1	2.22	2.62
Qwen2.5-VL-3B	21.3	22.5	33.0	24.4	24.2	1.38	2.11
Qwen2.5-VL-7B Base	23.5	21.3	34.5	22.6	24.8	2.04	2.23
Fine-tuned models							
Qwen2.5-VL-7B FT (10K)	34.24	33.56	71.25	39.30	41.43	2.69	3.01
Qwen2.5-VL-7B FT (100K)	**44.61**	**41.07**	**77.09**	**45.45**	**49.63**	**2.89**	**3.11**
(FT − Base)	↑21.1	↑19.8	↑42.6	↑22.9	↑24.8	↑0.85	↑0.88

Table 2. Aggregated Evaluation Across Major Categories. Accuracy (%) and free-form scores (1–5 scale) in the last two columns. SVF, Height, and Land Use/LC are single-feature inference (T1). Multi represents multi-feature inference (T2). Overall covers all short-answer categories (T1+T2). Free-form evaluation (T3) uses rubric-based scoring for structured-reasoning tasks. FT(10 K) and FT(100 K) in the result tables denote fine-tuning with 10 K and 100 K instruction-tuning samples, respectively. 'FT–Base' indicates the performance improvements achieved by fine-tuning Qwen-2.5-VL on 100K Q&As ('Qwen2.5-VL-7B FT (100K)') compared to the baseline results ('Qwen2.5-VL-7B Base').
*Format errors (repeated questions without answers): o4-mini 47.2%, Gemini-2.5-Flash 16.3%, others <2%.

and land cover descriptions, as shown in Table 5. To test generalization, we evaluated free-form QA using prompts that differed from those used during fine-tuning.

All models were decoded with nearly deterministic settings (temperature of 0.0 or the lowest commercial API setting). Appendix A details the implementation setup, Appendix B explains the metric rationale and sensitivity checks, and Appendix C describes the full evaluation protocol.

Model	SVF-aware				Height-aware	
	SVF val.	Reg. rank	SVF var.	Sun exp.	Hgt. avg.	Highest
• *Commercial models*						
GPT-4o	11.7	30.6	36.7	35.0	26.4	29.1
GPT-4.1-mini	1.9	33.8	37.3	31.3	18.9	28.4
Gemini-2.5-Flash	7.2	36.2	37.3	39.9	9.0	29.9
• *Remote sensing VLMs*						
TEOChat	0.7	15.1	34.4	22.4	18.9	23.9
GeoChat	6.5	16.7	35.7	25.4	17.3	24.6
• *Open-source models*						
LLaVA-OneVision	3.9	18.4	33.4	27.2	16.8	32.1
InternVL3-8B	2.9	19.0	33.5	26.3	18.5	31.2
Qwen2.5-VL-3B	5.5	19.3	32.2	23.2	18.0	27.0
Qwen2.5-VL-7B Base	6.8	19.5	34.0	28.6	9.9	32.7
Fine-tuned models						
Qwen2.5-VL-7B FT (10K)	27.9	29.6	39.4	35.3	34.8	32.3
Qwen2.5-VL-7B FT (100K)	**42.7**	**39.7**	**51.4**	**43.9**	**42.2**	**39.9**
(FT − Base)	↑35.9	↑20.2	↑17.4	↑15.3	↑32.3	↑7.2

Table 3. Category-wise Accuracy (Part I): SVF-aware and Height-aware categories. Abbreviations: SVF val. = SVF value, Reg. rank = region ranking, SVF var. = regional SVF variability, Sun exp. = sun exposure, Hgt. avg. = height average, Highest = highest region
Note. Part II (Table 4) continues with land cover and multi-factor categories.

Model	Land cover		Multi-factor			
	Top uses	LC type	Spatial open.	Sky vis.	Bldg. dens.	Vis. range
• *Commercial models*						
GPT-4o	16.1	18.4	41.4	36.3	39.7	24.2
GPT-4.1-mini	14.6	32.1	41.2	32.8	43.8	22.9
Gemini-2.5-Flash	15.6	56.4	45.1	38.1	46.3	26.8
• *Remote sensing VLMs*						
TEOChat	9.6	0.2	20.3	38.8	28.2	26.0
GeoChat	7.1	32.1	22.0	24.0	26.3	26.1
• *Open-source models*						
LLaVA-OneVision	19.5	23.1	32.7	11.6	31.5	19.9
InternVL3-8B	14.6	24.8	31.2	15.8	31.8	22.4
Qwen2.5-VL-3B	17.0	43.7	30.1	19.7	32.5	24.4
Qwen2.5-VL-7B Base	11.6	49.9	32.2	13.1	39.8	22.0
Fine-tuned models						
Qwen2.5-VL-7B FT (10K)	47.2	87.4	40.6	48.5	45.0	26.8
Qwen2.5-VL-7B FT (100K)	**56.5**	**90.9**	**47.6**	**55.3**	**50.4**	**32.7**
(FT − Base)	↑44.9	↑41.0	↑15.3	↑42.2	↑10.6	↑10.7

Table 4. Category-wise Accuracy (Part II): Land cover and multi-factor categories. Abbreviations: Top uses = top land uses, LC type = land cover type, Spatial open. = spatial openness, Sky vis. = sky visibility, Bldg. dens. = building density, Vis. range = visibility range.
Most multiple-choice questions have four options (25% baseline), but 'regional SVF variability' has three options and 'region ranking' varies by question nature.

Detailed Evaluation Protocol The pipeline enforces structured output formats while maintaining parsing tolerance to whitespace and reasoning chains. We shuffled multiple-choice options to reduce positional bias and only reported format errors when no valid answer could be recovered from the response.

4.1. Comparative Results

Domain gap and scaling effects. The aggregated results in Table 2 reveal clear domain gaps in 3D geospatial reasoning for all models. Commercial models generally outperform their open-source counterparts because of their larger pre-training scale and stronger alignment. However, scaling the parameters within the family (Qwen 3B to 7B Base) yielded limited gains without 3D geospatial data (24.2% to 24.8%). Notably, o4-mini exhibited 47.2% format errors (repeated questions without answers), while Gemini-2.5-Flash showed 16.3% errors, compared to <2% for other models. These format failures are likely due to thinking conventions and conversational defaults, and detailed analysis is provided in Appendix C.

Fine-tuned models achieve substantial improvements. The fine-tuned Qwen2.5-VL-7B secured the best overall short-answer accuracy (49.63%) and the highest free-form scores (Total: 2.89, Conclusion: 3.11 in Table 2). Increasing the size of the instruction-tuning corpus from 10 K to 100 K improved short-answer accuracy by 8.2 pp and slightly increased free-form scores (+0.20 in total and +0.10 in conclusion). This result supports the hypothesis that scaling short-answer supervision improves both the short-answer

Model	SVF	Land Cover	Elevation
• *Commercial models*			
o4-mini	2.03	2.12	1.73
Gemini-2.5-Flash	1.50	2.05	1.22
GPT-4o	2.12	2.27	1.62
GPT-4.1-mini	2.29	2.57	2.39
• *Remote sensing VLMs*			
TEOChat	1.25	1.50	1.38
GeoChat	1.20	1.23	1.19
• *Open-source models*			
LLaVA-OneVision	1.10	1.78	1.39
InternVL3-8B	1.81	2.22	1.92
Qwen2.5-VL-3B	1.29	1.36	1.27
Qwen2.5-VL-7B Base	1.57	2.18	1.73
Fine-tuned models			
Qwen2.5-VL-7B FT (10K)	3.22	2.80	2.69
Qwen2.5-VL-7B FT (100K)	**3.40**	**2.91**	**2.85**
(FT − Base)	↑1.83	↑0.73	↑1.12

Table 5. Free-form Analysis (Part I): Specialized feature agreement (1–5 scale).

Model	Urban	Energy	Landscape	Water
• *Commercial models*				
o4-mini	1.97	2.21	2.07	1.97
Gemini-2.5-Flash	2.10	2.07	2.28	1.38
GPT-4o	2.14	2.52	2.38	2.03
GPT-4.1-mini	2.38	2.86	2.28	2.59
• *Remote sensing VLMs*				
TEOChat	1.45	1.34	1.83	1.17
GeoChat	1.14	1.00	1.45	1.00
• *Open-source models*				
LLaVA-OneVision	1.62	1.24	1.66	1.41
InternVL3-8B	2.17	2.52	2.28	1.90
Qwen2.5-VL-3B	1.31	1.21	1.86	1.14
Qwen2.5-VL-7B Base	2.03	2.14	2.24	1.76
Fine-tuned models				
Qwen2.5-VL-7B FT (10K)	2.62	3.00	2.38	**2.76**
Qwen2.5-VL-7B FT (100K)	**2.93**	**3.17**	**2.76**	2.69
(FT − Base)	↑0.90	↑1.03	↑0.52	↑0.93

Table 6. Free-form Analysis (Part II): Category-specific performance (1–5 scale).
Urban = Urban Development, Energy = Renewable Energy, Water = Water Accumulation.

and downstream reasoning tasks.Table 7 details the effects of combining 100 K short-answer training with free-form QA, which constitutes our main 'FT' configuration. This combination lifted the overall short-answer accuracy from 49.0% to 49.6%, showing the advantages of supervision learned through open-ended questions.

Category-wise performance variations. Land-use questions saw the greatest improvement with instruction tuning (top land uses: +44.9 pp), confirming that 2D visual learning is significantly improved through domain adaptation. The SVF-related inference also showed sizable gains (SVF value: +35.9 pp), which indicates that sky view factors are learnable from RGB patterns (Tables 4 and 3). On the other hand, height-related tasks remained challenging despite our coarse decision-oriented design (relative comparisons and bin-level distinctions). This may be due to the ambiguity inherent in inferring 3D spatial properties from monocular 2D RGB inputs, where scale and depth cues are fundamentally underdetermined [8, 11, 12]. The multi-factor categories improved to 45.5% (+22.9 pp) after instruction tuning, which reflects better integration of multiple spatial cues (Table 2).

Upper-bound and Two-stage Baseline Analysis. To rigorously assess the value of the end-to-end VLM approach, we conducted additional experiments using inputs other than RGB. First, we evaluated a configuration with access to various ground truth modalities (DSM/SVF/segmentation) at inference. Using only the necessary modalities for each question tended to improve category-specific accuracy. The best overall accuracy reached **57.35%** (Height Inference: 69.25%) with agent-style routing, which selected the necessary modalities for each question. Second, we evaluated a two-stage pipeline that used three pretrained UNet-based monocular predictors on GeoNRW to estimate the DSM/SVF/segmentation from RGB, followed by agent-style inference. This achieved only **33.91%** overall accuracy (Height Inference: **14.97%**), which reflects the error propagation in the two-stage pipeline. Future work could combine improved RGB-to-DSM estimation with VLMs to bridge the gap toward upper-bound performance. Please refer to Appendix B for detailed modality ablation tables.

Category	Base	10K+free	100K	100K+free
Major categories				
SVF inference	23.5	34.2	43.9	44.6
Height inference	21.3	33.6	38.0	41.1
Land cover	34.5	71.3	77.5	77.1
Multi-factor	22.6	39.3	45.5	45.5
Overall	**24.8**	**41.4**	**49.0**	**49.6**

Table 7. Comprehensive comparison of Qwen2.5-VL-7B variants showing synergistic effects of combining short-answer and free-form training. All values indicate accuracy (%). '100 K' here means purely 100 K short-answer Q&As, and 10 K+free and 100 K+free are '10 K' and '100 K' recipes in other tables respectively, mixed with 1 K free-form Q&As.

4.2. Detailed Analysis

VLM spatial reasoning challenges and architectural constraints. The results point to the fundamental limitations of current vision-language model architectures in spatial tasks requiring coordinate-level understanding. For instance, VLMs achieve only 32.7% accuracy in the visibility range, which highlights the constraints of patch-based vision encoders. Even with advanced high-resolution techniques such as AnyRes (LLaVA-OneVision) and dynamic resolution processing (Qwen2.5-VL), images are still partitioned into discrete patches rather than interpreted in continuous coordinates [40, 42, 43]. This leads to systematic errors, such as

misinterpreting patch boundaries as actual coordinate positions [44], implying that the challenge is rooted in architectural design rather than insufficient training. In contrast, sky visibility achieved 55.3%, likely because SVF + land cover reasoning aligns more naturally with patch-level recognition than SVF + DSM inference, given that land cover varies less frequently within a patch than elevation.

Evidence for height estimation learning from RGB patterns. Domain-specific training still improved the performance on height-aware tasks by +19.8 pp (Table 2). This confirms that despite the challenges [45], VLMs can learn some aspects of 3D reasoning from RGB patterns through supervised learning. Prior work on monocular height estimation [13, 14] has demonstrated that deep learning models can learn height estimation from 2D visual cues. Our substantial performance gains demonstrate the effective conversion of established mappings from visual cues to geometric properties within the VLM architecture when sufficient training data and task-specific supervision are available.

Error patterns and robustness analysis. Models without fine-tuning typically exhibit weak spatial grounding, generate vague answers in free-form tasks, and often fail to integrate multiple spatial cues (e.g., SVF and elevation) effectively. In the multi-label land cover category, domain-specific models such as TEOChat and GeoChat often collapse into a single label, indicating limitations in following instructions that may be rooted in VQA training with single answers.

Performance comparison with specialized methods. VLMs achieved 41.1% and 44.6% accuracy on height- and SVF-related tasks, respectively (Table 2), which reveals substantial gaps compared with specialized sensor-based approaches. However, VLMs offer unique advantages including natural language interfaces and integrated reasoning across multiple spatial attributes. Our free-form evaluation achieved a 3.11/5 conclusion score and demonstrates compositional reasoning potential that specialized estimators cannot achieve. Although a direct comparison with traditional monocular depth estimation would be valuable, our work focused on evaluating VLMs within a unified framework to provide a consistent assessment of diverse spatial reasoning tasks. We will leave the explicit baseline comparisons with specialized estimators plus rule-based QAs for future work.

Prompt sensitivity analysis. To test the model's robustness to prompt phrasing, we reevaluated our best fine-tuned model using a version of the benchmark in which all question texts were paraphrased by GPT-4, and semantic equivalence was checked on 1% of the samples. The accuracy dropped from 49.6% to 40.8%, whereas the format error rate remained very low (0.09%). This performance still largely exceeds both the Qwen-base model (24.8%) and the best inference-only model (33.0%), indicating that the fine-tuned model retains advantages even under prompt variations. Full results are in Appendix D.

4.3. Limitations and Future Directions

Current limitations. Despite domain-specific improvements, our best models achieved only $<$50% accuracy on complex spatial reasoning tasks and remain insufficient for reliable deployment. Additionally, the evaluation was confined to North Rhine-Westphalia, which limits generalization claims. A systematic cross-regional evaluation is left for future work. Furthermore, performance varies significantly across task types, and precision for coordinate-level tasks remains problematic, with an accuracy rate of 32.7%. Although option labels were balanced and choices were randomized during dataset creation to reduce positional bias, a dedicated sensitivity study of multiple-choice option ordering is left for future work. Finally, our geospatial specialization may impact general vision-language performance and requires future assessment of this trade-off.

Future research directions. In summary, our error analysis highlights the main challenges of Geo3DVQA and suggests several future research directions. 1) *Architectural innovations*: Development of spatial encoders with 3D structure understanding and coordinate-aware attention mechanisms to better integrate 2D patches with continuous spatial positions. 2) *Generalization*: Broader geographic and seasonal training data for real-world robustness. 3) *Explainability*: Attention visualization and counterfactuals to reveal the actual model reasoning processes for height-aware inference and failure modes. Pursuing these directions could lead to robust 3D reasoning from RGB without relying on costly specialized sensor data and, in turn, make real-world deployment more scalable.

5. Conclusion

We introduced Geo3DVQA, the first VQA benchmark that evaluates height-aware 3D spatial reasoning directly from RGB imagery. Our evaluation across 16 categories and three complexity tiers demonstrates that domain-specific instruction tuning substantially improves reasoning from RGB images alone. Specifically, fine-tuned Qwen2.5-VL-7B attained 49.6% overall accuracy compared to 24.8% for the base model, and surpassed the 28.6% achieved by GPT-4o. Improvements were particularly pronounced in SVF-aware and multi-factor reasoning tasks, as well as in land cover tasks.

Despite this achievement, precise coordinate-level estimation remains a challenge. This necessitates targeted architectural innovations to achieve reliable spatial precision in these models. As a unified 3D inference benchmark based solely on RGB imagery, Geo3DVQA highlights both the current strengths and remaining limitations while identifying concrete opportunities for advancing accessible 3D geospatial understanding without specialized sensors.

Acknowledgements

This work was supported by JST NEXUS, Japan Grant Number JPMJNX25CA. Weihao Xuan is supported by the RIKEN Junior Research Associate (JRA) program

References

[1] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34892–34916, 2023. 1, 2

[2] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, *et al.*, "LLaVA-OneVision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024. 4

[3] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, *et al.*, "Qwen2.5-VL technical report," *arXiv preprint arXiv:2502.13923*, 2025. 1, 2, 4

[4] P. Arias, N. Bellouin, E. Coppola, R. Jones, G. Krinner, J. Marotzke, V. Naik, M. Palmer, G.-K. Plattner, J. Rogelj, *et al.*, "Climate Change 2021: the physical science basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change; technical summary," 2021. 1

[5] J. Unger, "Intra-urban relationship between surface geometry and urban heat island: review and new approach," *Climate research*, vol. 27, no. 3, pp. 253–264, 2004. 1

[6] A. Middel, J. Lukasczyk, R. Maciejewski, M. Demuzere, and M. Roth, "Sky View Factor footprints for urban climate modeling," *Urban Climate*, vol. 25, pp. 120–134, 2018. 1, 3

[7] Y. Li, W. Li, S. Tang, W. Darwish, Y. Hu, and W. Chen, "Automatic Indoor as-Built Building Information Models Generation by Using Low-Cost RGB-D Sensors," *Sensors*, vol. 20, no. 1, p. 293, 2020. 1

[8] H. Lu, S. Xu, and S. Cao, "SGTBN: Generating dense depth maps from single-line lidar," *IEEE Sensors Journal*, vol. 21, pp. 19091–19100, 2021. 2, 7

[9] N. Takhtkeshha, G. Mandlburger, F. Remondino, and J. Hyyppä, "Multispectral light detection and ranging technology and applications: a review," *Sensors*, vol. 24, no. 5, p. 1669, 2024. 1

[10] F. Lindberg, C. Grimmond, A. Gabey, B. Huang, C. W. Kent, T. Sun, N. E. Theeuwes, L. Järvi, H. C. Ward, I. Capel-Timms, Y. Chang, P. Jonsson, N. Krave, D. Liu, D. Meyer, K. F. G. Olofson, J. Tan, D. Wästberg, L. Xue, and Z. Zhang, "Urban Multi-scale Environmental Predictor (UMEP): An integrated tool for city-based climate services," *Environmental Modelling & Software*, vol. 99, pp. 70–87, 2018. 1, 3

[11] Y. Cai, B. Li, Z. Jiao, H. Li, X. Zeng, and X. Wang, "Monocular 3D Object Detection with Decoupled Structured Polygon Estimation and Height-Guided Depth Estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 10478–10485, 2020. 2, 7

[12] F. Lourenço and H. Araujo, "Intel RealSense SR305, D415 and L515: Experimental Evaluation and Comparison of Depth Estimation," in *VISIGRAPP (4: VISAPP)*, pp. 362–369, 2021. 7

[13] C.-J. Liu, V. A. Krylov, P. Kane, G. Kavanagh, and R. Dahyot, "IM2ELEVATION: Building height estimation from single-view aerial imagery," *Remote Sensing*, vol. 12, no. 17, p. 2719, 2020. 3, 8

[14] G. Liasis and S. Stavrou, "Satellite images analysis for shadow detection and building height estimation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 119, pp. 437–450, 2016. 2, 3, 8

[15] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "RSVQA: Visual question answering for remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8555–8566, 2020. 2, 3

[16] K. Li, G. Vosselman, and M. Y. Yang, "HRVQA: A Visual Question Answering benchmark for high-resolution aerial images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 214, pp. 65–81, 2024. 2, 3

[17] X. Li, J. Ding, and M. Elhoseiny, "VRSBench: A versatile vision-language benchmark dataset for remote sensing image understanding," *Advances in Neural Information Processing Systems*, vol. 37, pp. 3229–3242, 2024. 2, 3

[18] Y. Hu, J. Yuan, C. Wen, X. Lu, Y. Liu, and X. Li, "RSGPT: A Remote Sensing Vision Language Model and Benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 224, pp. 272–286, 2025. 2

[19] J. Wang, W. Xuan, H. Qi, Z. Liu, K. Liu, Y. Wu, H. Chen, J. Song, J. Xia, Z. Zheng, *et al.*, "DisasterM3: A Remote Sensing Vision-Language Dataset for Disaster Damage Assessment and Response," *arXiv preprint arXiv:2505.21089*, 2025. 2, 3

[20] W. Xuan, J. Wang, H. Qi, Z. Chen, Z. Zheng, Y. Zhong, J. Xia, and N. Yokoya, "DynamicVL: Benchmarking Multimodal Large Language Models for Dynamic City Understanding," *arXiv preprint arXiv:2505.21076*, 2025. 2

[21] J. A. Irvin, E. R. Liu, J. C. Chen, I. Dormoy, J. Kim, S. Khanna, Z. Zheng, and S. Ermon, "TEOChat: A large Vision-Language assistant for temporal earth observation data," in *International Conference on Learning Representations*, 2025. 3, 4

[22] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "GeoChat: Grounded Large Vision-Language Model for Remote Sensing," in *2024*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 27831–27840, IEEE, jun 16 2024. 3, 4

[23] M. Danish, M. A. Munir, S. R. A. Shah, K. Kuckreja, F. S. Khan, P. Fraccaro, A. Lacoste, and S. Khan, "GEOBench-VLM: Benchmarking Vision-Language models for geospatial tasks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7132–7142, October 2025. 3

[24] J. Wang, Z. Zheng, Z. Chen, A. Ma, and Y. Zhong, "EarthVQA: Towards queryable earth via relational reasoning-based remote sensing visual question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 5481–5489, 2024. 3

[25] L. Mou and X. X. Zhu, "IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network," *arXiv preprint arXiv:1802.10249*, 2018. 3

[26] J. Liang, J. Gong, J. Sun, J. Zhou, W. Li, Y. Li, J. Liu, and S. Shen, "Automatic sky view factor estimation from street view photographs—A big data approach," *Remote Sensing*, vol. 9, no. 5, p. 411, 2017. 3

[27] M. Hodul, A. Knudby, and H. C. Ho, "Estimation of continuous urban sky view factor from landsat data using shadow detection," *Remote Sensing*, vol. 8, no. 7, p. 568, 2016. 3

[28] G. Baier, A. Deschemps, M. Schmitt, and N. Yokoya, "GeoNRW," 2020. 3

[29] P. L. Guth, A. Van Niekerk, C. H. Grohmann, J.-P. Muller, L. Hawker, I. V. Florinsky, D. Gesch, H. I. Reuter, V. Herrera-Cruz, S. Riazanoff, *et al.*, "Digital elevation models: Terminology and definitions," *Remote Sensing*, vol. 13, no. 18, p. 3581, 2021. 3

[30] U. Alganci, B. Besol, and E. Sertel, "Accuracy assessment of different digital surface models," *ISPRS International Journal of Geo-Information*, vol. 7, no. 3, p. 114, 2018. 3

[31] Z. Jiao, H. Ren, X. Mu, J. Zhao, T. Wang, and J. Dong, "Evaluation of four sky view factor algorithms using digital surface and elevation model data," *Earth and Space Science*, vol. 6, pp. 222 – 237, 2019. 3

[32] M. Dirksen, R. Ronda, N. Theeuwes, and G. Pagani, "Sky view factor calculations and its application in urban heat island studies," *Urban Climate*, vol. 30, p. 100498, 2019. 3

[33] C. Miao, S. Yu, Y. Hu, H. Zhang, X. He, and W. Chen, "Review of methods used to estimate the sky view factor in urban street canyons," *Building and Environment*, vol. 168, p. 106497, 2020. 3, 4

[34] M. Daramola and I. Balogun, "Analysis of the urban surface thermal condition based on sky-view factor and vegetation cover," *Remote Sensing Applications: Society and Environment*, 2019. 3, 4

[35] Y. Xia, N. Yabuki, and T. Fukuda, "Sky view factor estimation from street view images based on semantic segmentation," *Urban Climate*, 2021.

[36] A. Bradley, J. Thornes, and L. Chapman, "A method to assess the variation of urban canyon geometry from sky view factor transects," *Atmospheric Science Letters*, vol. 2, no. 1-4, pp. 155–165, 2001. 3, 4

[37] J. Ha, S. Lee, and C. Park, "Temporal Effects of Environmental Characteristics on Urban Air Temperature: The Influence of the Sky View Factor," *Sustainability*, vol. 8, no. 9, 2016.

[38] J. Li, H. Zhang, and E. Xu, "A two-level nested model for extracting positive and negative terrains combining morphology and visualization indicators," *Ecological Indicators*, vol. 109, p. 105842, 2020.

[39] S. Hoechstetter, U. Walz, N. X. Thinh, *et al.*, "Effects of topography and surface roughness in analyses of landscape structure-A proposal to modify the existing set of landscape metrics," *Landscape Online*, pp. 3–3, 2008. 4

[40] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao, *et al.*, "InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025. 4, 7

[41] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, *et al.*, "GPT-4o System Card," *arXiv preprint arXiv:2410.21276*, 2024. 4

[42] J. Jiao, Y. Tang, K.-L. C. Lin, Y. Gao, J. Ma, Y. Wang, and W.-S. Zheng, "DilateFormer: Multi-scale dilated transformer for visual recognition," *IEEE Transactions on Multimedia*, vol. 25, pp. 8906–8919, 2023. 7

[43] H. Tang, D. Liu, and C. Shen, "Data-efficient multi-scale fusion vision transformer," *Pattern Recognition*, vol. 161, p. 111305, 2025. 7

[44] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-Language Models for Vision Tasks: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, pp. 5625–5644, 2023. 8

[45] X. Hou, W. Gan, and N. Yokoya, "Enhancing monocular height estimation from aerial images with street-view images," *arXiv preprint arXiv:2311.02121*, 2023. 8