Supplementary Material for: 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,†]

Overview of Supplementary Materials

Appendix A details implementation (models, decoding, infrastructure), Appendix B describes evaluation protocols and scoring rules (with sensitivity analyses), Appendix C presents extended results and analyses, Appendix D formalizes the spatial metrics and methodology, Appendix E provides extended qualitative discussions and worked examples, and Appendix F documents data/code availability and release scope.

A. Technical Implementation Details

We summarize technical implementation details for reproducibility, including the experimental setup, dataset construction methodology, evaluation protocols, and supplementary reference materials.

A.1. Experimental Implementation

A.1.1. Model Selection and Training Configurations

Our model selection represents the current state-of-the-art capabilities across different paradigms: commercial models provide performance benchmarks, open-source models enable systematic analysis of architectural constraints (high-resolution processing via AnyRes, dynamic resolution handling), and domain-specific models offer comparisons with specialized approaches for temporal and spatial reasoning in remote sensing contexts.

Detailed Instruction-tuning Setups. Learning rate: 3e-5, batch size: 1 (effective 8 with gradient accumulation), optimizer: AdamW, epochs: 3, schedule: cosine, warmup: 10% of total steps, weight decay: 0.01, gradient clipping: 1.0, adaptation: LoRA (rank 8, target modules: all transformer attention and feedforward layers).

A.1.2. Infrastructure and Reproducibility

Software Environment.

- Python: 3.9.7
- PyTorch: 2.2.2+cu121
- Transformers: 4.21.1
- CUDA: 12.1
- CUDNN: 8
- Additional libraries: NumPy 1.23.5, Pandas 1.3.3, Scikit-learn 1.0.2, aenum, nptyping, rasterio, nvidia-pyindex, openmim

A.1.3. Inference and Decoding Configuration

Open-source (HuggingFace) models. Unless otherwise noted, inference uses greedy decoding:

- `do_sample=False`, `num_beams=1`
- `temperature=None`, `top_p=None`, `top_k=None` (ignored when `do_sample=False`)
- `max_new_tokens=256`
- Evaluation loop wrapped with `with torch.no_grad():` and `model.generate(**inputs, ...)`

Commercial APIs. To ensure deterministic conditions across all models, we configured endpoints with temperature set to 0.0 (or the minimum allowable value) and disabled sampling, while maintaining token limits comparable to open-source experiments. However, achieving complete uniformity across API providers presents inherent challenges due to provider-specific constraints. Context window limits, internal token budget allocations, and distributed inference mechanisms vary across different models. Furthermore, commercial models (e.g., GPT-4o, Gemini) may exhibit subtle non-deterministic behavior even at temperature 0.0, attributed to internal optimizations and distributed execution architectures [1] [2]. While we minimized these variations through consistent use of minimum temperature settings and comparable token budgets, readers should note that minor reproducibility differences may emerge from these API-level factors, representing inherent limitations in commercial-to-open-source model comparisons.

A.1.4. Model Endpoints and Checkpoints

We list all models, endpoints, and checkpoints used for evaluation and fine-tuning.

---

[1] University of Tokyo, Graduate School of Information Science and Technology [2] RIKEN AIP † Corresponding author: yokoya@k.u-tokyo.ac.jp Emails: maitsuji@g.ecc.u-tokyo.ac.jp

- GPT-4o: OpenAI API (`gpt-4o-2024-08-06`)
- GPT-4.1-mini: OpenAI API (`gpt-4.1-mini-2025-04-14`)
- GPT-5: OpenAI API (`gpt-5-2025-08-07`)
- GPT-4: OpenAI API (`gpt-4-0613`)
- Qwen2.5-VL-7B (Base): `Qwen/Qwen2.5-VL-7B-Instruct` (HuggingFace)
- Qwen2.5-VL-7B (FT-100K): LoRA-adapted checkpoint from 100K QA fine-tuning (ours)
- GeoChat: `MBZUAI/GeoChat-7B` (HuggingFace)
- TeoChat: `TeoChat/TeoChat-7B` (HuggingFace)

A.1.5. Question Generation Framework

Our question-generation framework incorporates several key design principles to ensure systematic coverage of reasoning capabilities while maintaining diversity and practical relevance.

Difficulty Stratification Methodology. Questions were systematically categorized into difficulty levels through a multi-stage classification process. The stratification framework ensures a broad evaluation of model capabilities across different complexity levels, from basic spatial queries to complex multi-feature reasoning tasks and complete deep reasoning analyses.

Modality-Specific Design Principles. Each task category is designed to leverage specific data modalities effectively (SVF, DSM, RGB, and semantic segmentation), ensuring that multi-feature integration provides meaningful advantages over single-feature approaches. The design process followed established geospatial analysis methodologies while adapting them for vision-language model evaluation.

Template-Based Generation Strategy. Following established methodologies from prior VLM benchmarking research, we employed a systematic template-based approach with manually curated question templates for each category. During dataset construction, a GPT-4 family model (default: `gpt-4o-mini`) optionally paraphrased these base templates to increase linguistic diversity while preserving task semantics and answer formats. Our template system includes: (1) Spatial Comparison Templates for location (region or coordinate-based) comparative analysis; (2) Quantitative Analysis Templates for precise value extraction and calculation; (3) Grid-Based Templates for structured spatial analysis tasks; and (4) Free-form Description Templates for comprehensive landscape analysis.

A.1.6. Free-form Generation Implementation

Statistical Feature Extraction. For Tier 3 free-form questions, we computed comprehensive scene-level statistics from all four modalities (SVF, DSM, RGB, and semantic segmentation). The statistical extraction process includes:

Summary Statistics: Mean, standard deviation, minimum, maximum, and quartiles for continuous data modalities (SVF, DSM).

Land Cover Analysis: Ratio calculations for different semantic classes, spatial distribution patterns, and transition zone analysis.

Color Space Analysis: RGB histogram distributions, dominant color identification, and color diversity metrics.

Grid Analysis: 3x3 grid-based analysis for each metric related to the questions, including statistical computation and reference generation for optimal locations (e.g., SVF, DSM, land cover, suitable places for the question).

Open-ended Q&As Creation Workflow. The structured statistics served as input to GPT-4.1-mini to generate contextually relevant free-form questions and answers. The processing workflow follows these steps:

1. Statistical feature compilation into a structured JSON-style format from SVF, DSM, segmentation, and RGB inputs
2. Template-based question generation using pre-defined categories and base prompts
3. Answer generation based on statistical evidence and domain knowledge using GPT-4.1-mini and human expert prompt refinement
4. Iterative refinement with an LLM-based verifier that checks answers against the SVF/DSM/segmentation/RGB statistics and a 2×2 multimodal panel, combined with rule-based consistency checks
5. Human expert validation on a sampled subset for accuracy and consistency with colorized SVF, DSM, RGB, and segmentation data

A.1.7. Quality Assurance and Validation

Bias Mitigation Implementation. Our generation framework includes systematic bias detection and mitigation strategies to ensure a fair evaluation across different spatial patterns and urban environments. The bias mitigation process includes:

- Systematic sampling across different urban morphologies
- Geographic diversity within the GeoNRW dataset coverage area
- Statistical validation of answer distribution patterns

Automated Validation Procedures. The automated validation system implements multiple checking mechanisms.

Answer Consistency Verification. We use automated rule based and LLM based verifiers to check answer consistency. These verifiers compare `<OBSERVATION>` and `<CONCLUSION>` against precomputed SVF, DSM, and land cover statistics. They also check against a 2×2 multimodal panel that includes RGB, SVF, DSM, and segmentation data. We supplement these automated checks with human inspection of a sampled subset.

B. Evaluation Methodology and Protocols

Deterministic decoding and repeats. Unless otherwise noted, we used deterministic decoding (temperature$= 0$; no sampling).

Multiple-choice option order. Choices are shuffled at the generation time by a bias-free shuffler that prevents positional skew.

B.0.1. Short-answer Scoring Methodology

Normalization and category-specific criteria. We lowercase the predictions, trim whitespace, and canonicalize comma-separated lists. The category-specific rules are as follows:

- **landcover_type**: Jaccard similarity ≥ 0.7 between predicted and ground-truth label sets
- **land_use**: order-independent exact set match
- **height_average**: mixed tolerance with 10 m quantization; exact for 0 m; ± 10 m if ≤ 30 m; $\pm 30\%$ if > 30 m
- **hard_pixel**: absolute error ≤ 0.05

For multi-label categories (e.g., *landcover_type*), we used Jaccard similarity (intersection-over-union) between the predicted label set and the ground-truth set, and counted an answer as correct when the similarity was at least 0.7. Jaccard is a standard set-similarity metric in classification and segmentation evaluation and is supported in common ML libraries for multilabel settings (e.g., scikit-learn's `jaccard_score`) [3]. This threshold strikes a balance between strictness and leniency, being stricter than the IoU@0.5 threshold used in the VOC style but not as unforgiving as an exact set match [4, 5].

For height estimation (*height_average*), numeric answers were compared to the ground truth with 10 m quantization and magnitude aware tolerances. We require exact match for 0 m. For ground truth values up to 30 m, we allow ± 10 m deviation. For values above 30 m, we use $\pm 30\%$ relative tolerance.

This mixed absolute and relative rule reflects measurement error characteristics in remote sensing elevation and canopy height products. Satellite DEMs such as SRTM frequently report absolute vertical errors on the order of 5 to 10 m [6]. Canopy height estimation methods such as GEDI exhibit substantial relative errors (25 to 45%) [7]. Moreover, relative percentage errors are known to be ill behaved near zero [8]. Hence, absolute tolerances are preferred for small magnitudes, and relative tolerances are appropriate for larger values.

For SVF value estimation (*hard_pixel*), we applied an absolute error threshold of ≤ 0.05. This threshold was calibrated to the SVF value range [0.0, 1.0] with typical quantization at 0.1 increments. In relative terms, this represents approximately 5–10% tolerance (0.05/0.5–1.0), which is comparable to the height estimation tolerances when accounting for the different value ranges and measurement characteristics. The different tolerance criteria across tasks (absolute for SVF, mixed absolute/relative for height) reflect the distinct error characteristics of the underlying data sources: SVF values are normalized ratios with bounded ranges, whereas height measurements exhibit scale-dependent uncertainties that require magnitude-aware processing.

B.0.2. Threshold Sensitivity Analysis

In addition to the default Jaccard threshold of 0.7 for multi-label land cover categories, we also evaluated a stricter setting of 0.8. Table 1 summarizes accuracies (%) for a subset of models, comparing three summary metrics covering *landcover_type*, *land_use_landcover_type*, and *overall* accuracy across all short-answer categories. Most of the models showed similar performance between the two thresholds (within a 2% drop in overall accuracy), but Gemini-2.5-Flash showed a significant drop in accuracy (approximately 11 percentage points, from 33.0% to 22.0% in the overall metric). This may reflect its stronger performance on intermediate-complexity tasks, and the task complexity at the 0.7 threshold was best suited to this model.

Model	**landcover_type**		**land_use_landcover_type**		**Overall**	
	0.7	**0.8**	**0.7**	**0.8**	**0.7**	**0.8**
GPT-4o	18.43	4.83	17.50	9.34	28.60	27.29
GPT-4.1-mini	32.12	11.28	25.10	12.60	28.10	26.06
o4-mini	—	7.25	15.30	8.32	19.20	18.06
Gemini-2.5-Flash	56.39	21.85	40.00	19.89	33.00	22.04
LLaVA-one-vision	23.06	10.27	21.60	13.98	21.80	20.52
InternVL3-8B	24.77	5.54	20.70	9.16	22.10	20.20
Qwen2.5-VL-7B Base	49.85	21.45	34.50	17.48	24.80	22.03
Qwen2.5-VL-7B FT (10K)	87.41	68.18	71.25	59.73	41.43	39.56
Qwen2.5-VL-7B FT (100K)	90.94	71.10	77.09	65.22	49.63	47.70
(FT − Base)	↑41.09	↑49.65	↑42.59	↑47.74	↑24.83	↑25.67

Table 1. Sensitivity to Jaccard threshold for multi-label land cover categories. Accuracies (%) under the default 0.7 criterion and the stricter 0.8 criterion.

B.0.3. Free-form Evaluation Protocol

Open-ended responses were scored by a well-prompted GPT evaluator with temperature$= 0.0$ using a rubric, which was repeatedly adjusted by human validation. Domain scores (*SVF*, *Land cover*, *Elevation*) are assigned *NaN* only when the ground truth lacks that domain entirely. We saved raw evaluator outputs and per-question summaries for auditability.

B.0.4. Modality Ablations and Two-Stage Pipeline Analysis

We conduct modality ablation experiments to quantify how each input type and training recipe contributes to Geo3DVQA performance. These analyses complement the main benchmark results by contrasting our end-to-end VLM with (i) an upper-bound oracle that has access to ground-truth DSM/SVF at inference and (ii) a conventional two-stage pipeline that first predicts DSM/SVF and segmentation from RGB and then answers questions from these predictions.

Terminology (definitions) for Modality Ablations

- **Modality**: The type of input image provided to the model. Options include:

  - **RGB**: Optical aerial imagery capturing surface appearance
  - **DSM**: Digital Surface Model representing elevation data
  - **SVF**: Sky View Factor measuring visible sky proportion (0–1)
  - **Seg**: Semantic segmentation map of land cover classes
- **Legend**: A color bar overlay on DSM colormaps showing minimum and maximum elevation values. Because colormaps alone do not convey an absolute scale, adding legends substantially improves the height estimation accuracy.
- **Arrays (numeric image)**: Raw numeric representation of DSM/SVF data (DSM in meters, SVF as 0–1 values) without colormap conversion. This contrasts with the colormap-based visualization.
- **Colormap**: Visualization that converts numeric DSM/SVF values into color gradients for visual interpretation by VLMs.
- **Necessary/needed modalities**: A routing strategy where each question is analyzed (via rules + AI) to determine which modalities are required, providing only the relevant inputs rather than all available data.

Summary (inference only). Table 2 reports overall and domain-wise accuracies for non-fine-tuned inference under different input settings, using 3000 QA samples from the test set. The metrics cover overall accuracy, height inference, land use/land cover (LULC), Sky View Factor (SVF) inference, and multi-factor reasoning, and provide a high-level view of how each modality combination affects performance.

Detailed categories (inference only). Table 3 expands this analysis to all task categories, grouping SVF-aware, height-aware, land-cover, and multi-factor metrics to show where each modality is most effective.

Summary (fine-tuned) and Two-Stage Comparison. Table 4 provides a comprehensive breakdown of accuracy across fine-tuned configurations. To rigorously evaluate the benefits of an end-to-end approach, we benchmarked a **Two-Stage Pipeline** in which a U-Net (trained on GeoNRW) first predicts DSM, SVF, and segmentation maps from RGB, followed by a rule-based agent fine-tuned on ground truth data. This two-stage baseline achieved only 33.91% overall accuracy, lagging significantly behind our end-to-end VLM (50.10% with RGB FT) and oracle upper bound (57.35%). This substantial gap is primarily driven by error propagation in height inference (14.97% vs. 41.07% for the VLM), underscoring the resilience of the unified VLM architecture.

Detailed categories (fine-tuned). Table 5 reports the complete category-level results for the fine-tuned settings with different modality inputs during fine-tuning and inference.

Key observations on modality alignment. Across both non-fine-tuned and fine-tuned settings, the highest accuracy is achieved when the input modality matches the queried attribute. For example, calibrated DSMs improved accuracy for height-aware questions, SVF images improved accuracy for SVF tasks, and segmentation, RGB, or all modality combinations improved accuracy for land-cover and appearance-based reasoning. In non-fine-tuned inference, adding a legend to DSM inputs roughly doubles the accuracy of height inference for the Qwen2.5-VL-7B base model (from about 20% to 50.0% in Table 2), and it would be because the legend resolves absolute scale ambiguities that cannot be recovered from colormaps alone. Fine-tuning further amplifies this benefit. The agent-style recipe ('FT (necessary)'), which routes each question to its required modalities, achieves strong overall accuracy (53.40%) and dominates the performance for the height-aware categories (Table 5).

Oracle performance with necessary-only routing. The upper-bound capability of the system appears when RGB is excluded and only the ground-truth modalities required for each question are provided ('necessary-only (no rgb)' in Table 5). This configuration achieves the best overall accuracy of **57.35%**, with particularly strong height-aware performance (height inference 69.25%, height average 93.05%) and near-perfect land-cover accuracy (99.64%). These results confirm the importance of strict alignment between modality and question type. DSMs should be used for height-aware reasoning, segmentation maps for semantic land cover-related tasks, and SVFs for sky view factor questions.

Visualizations versus raw arrays We observed that using DSM or SVF as direct numeric arrays (without colormap rendering) generally degraded performance compared with calibrated colormap visualizations, and overall accuracy dropped from roughly 22% to 19–20% in Table 2. This pattern supports the hypothesis that current multimodal transformers are optimized for extracting features from visual patterns rather than raw numerical grids, validating our choice of calibrated colormaps. A small exception occurs for height inference with 'dsm' in the legend-free setting where arrays slightly outperform colormaps, but DSM with legends still yields much higher height-aware accuracy overall.

Role of VLM-based reasoning Although raw height values can be obtained from sensors, many benchmark questions require interpreting spatial relationships and composing multiple cues. Our fine-tuned VLM attains a high free-form conclusion score (3.11/5 in Table 6), which demonstrates that the VLM can synthesize spatial relationships and explain height-dependent phenomena in context. Such compositional reasoning is difficult to replicate with purely rule-based or retrieval-oriented systems, highlighting the value of VLMs for 3D geospatial question answering.

Model	Modality at inference	Overall	Height inf.	LULC	SVF inf.	Multi inf.
Qwen2.5-VL-7B Base (no FT)						
Qwen2.5-VL-7B Base	all	25.49	19.79	**44.66**	22.75	21.04
Qwen2.5-VL-7B Base	dsm (no legend)	22.54	20.86	28.21	22.37	20.08
Qwen2.5-VL-7B Base	svf	22.23	20.32	26.07	**23.74**	18.03
Qwen2.5-VL-7B Base	rgb	24.79	18.98	38.03	23.05	**22.40**
Qwen2.5-VL-7B Base	rgb+dsm (no legend)	23.79	18.18	34.83	22.37	22.13
Qwen2.5-VL-7B Base	rgb+dsm (legend)	**27.84**	**50.00**	35.68	21.83	22.27
Qwen2.5-VL-7B Base	dsm (legend)	25.59	**50.00**	21.15	23.36	19.95
Qwen2.5-VL-7B Base (arrays, no FT)						
Qwen2.5-VL-7B Base	all (arrays)	**25.14**	20.59	**44.44**	21.76	**21.17**
Qwen2.5-VL-7B Base	dsm (arrays)	20.46	**21.93**	15.81	22.67	18.72
Qwen2.5-VL-7B Base	svf (arrays)	19.31	19.25	11.32	**23.21**	17.49
GPT-4o (no FT)						
GPT-4o	all	**28.07**	25.40	**23.50**	32.06	29.64
GPT-4o	dsm (legend)	23.67	**29.95**	2.78	29.39	27.32
GPT-4o	rgb	26.77	24.06	18.59	30.00	**31.83**
GPT-4o	svf	24.43	25.67	2.99	**33.82**	24.59

Table 2. Modality ablation (inference only) — Summary metrics. Accuracy (Overall), height inference, LULC (land cover & land use), SVF inference, and multi-factor inference (%).

Sensitivity to RGB-to-DSM estimation errors A critical insight emerges when comparing the oracle setup with the two-stage pipeline: using explicit RGB-to-DSM predictors results in significantly lower accuracy (33.91%) than using ground-truth modalities (57.35%). This suggests that current VLMs are sensitive to estimation artifacts in upstream predictions, which further supports the end-to-end learning approach. In this approach, the model can potentially learn robust features directly from RGB or fuse noisy inputs more effectively.

Generalization risks of colormap dependence The strong gains from legends and specific colormap choices also expose potential generalization risks, as models become accustomed to particular visualization schemes. Our plasma (SVF) and terrain (DSM) colormaps follow domain conventions, but substantial changes in color mapping or legend design could degrade performance. Future work should investigate colormap-invariant encoders and hybrid representations that jointly embed raw numeric arrays and their visualizations, thereby reducing reliance on any single color palette while preserving the benefits of calibrated legends.

C. Comprehensive Results Analysis

C.0.1. Free-form Evaluation Results

We evaluated *free-form* categories under a controlled protocol: training used 1000 free-form QA pairs, and evaluation used a disjoint set of 100 free-form QA pairs covering four categories (Urban Development, Renewable Energy, Landscape Analysis, Water Accumulation). An automated rubric-based assessor (GPT-4) scored the responses on a 1–5 scale along the general and domain axes. Subsequently, a sampled subset was evaluated by human annotators to validate the automated scoring.

Note on evaluator/template affinity. Because our prompt templates were developed using GPT-4.1-mini and the rubric judge is a GPT-4–family model, the evaluation may exhibit shared-family bias: the judge could assign slightly higher scores to outputs that match stylistic patterns common in the same model family. Accordingly, we treat rubric-based metrics as relative rather than absolute, and iteratively refined the rubric until its judgments aligned with human annotators on a validation subset. The full rubric and evaluation prompts will be released in our project repository to ensure transparency. Nevertheless, the Geo3DVQA fine-tuned Qwen2.5-VL-7B surpasses GPT-4.1-mini and other baselines across most categories, indicating that our main conclusions are robust to this potential bias.

C.0.2. Synergistic Training Effects Analysis

Table 7 demonstrates the synergistic effects of combining short-answer and free-form QA training, showing systematic improvements with notable gains in the overall accuracy.

C.0.3. Format Error Analysis

Output formatting and decoding behavior. The short-answer evaluation pipeline first analyzes raw model outputs to design task-specific parsing rules that follow each model's

(a) SVF-aware and Height-aware categories

Model	Modality	SVF val.	Reg. rank	R.SVF.v	Sun.e	Hgt.avg	Highest
Qwen2.5-VL-7B Base (no FT)							
Qwen2.5-VL-7B Base	all	**10.60**	18.86	36.02	24.91	5.88	33.69
Qwen2.5-VL-7B Base	dsm w/o leg.	8.48	18.51	32.26	27.76	8.02	33.69
Qwen2.5-VL-7B Base	svf	9.89	20.28	33.87	27.40	7.49	33.16
Qwen2.5-VL-7B Base	rgb	5.65	17.79	35.48	**30.25**	3.74	34.22
Qwen2.5-VL-7B Base	rgb+dsm (no leg.)	9.89	18.86	32.80	24.20	1.60	34.76
Qwen2.5-VL-7B Base	rgb+dsm (leg.)	9.54	17.79	32.26	23.84	64.17	**35.83**
Qwen2.5-VL-7B Base	dsm (legend)	**10.60**	17.44	33.87	28.11	**64.71**	35.29
Qwen2.5-VL-7B Base	all (arrays)	4.59	19.93	32.80	25.98	7.49	33.69
Qwen2.5-VL-7B Base	dsm (arrays)	6.36	20.28	**37.63**	27.40	9.63	34.22
Qwen2.5-VL-7B Base	svf (arrays)	7.77	**22.42**	36.56	25.62	4.81	33.69
GPT-4o (no FT)							
GPT-4o	all	15.19	28.47	45.16	**35.23**	16.58	34.22
GPT-4o	dsm (legend)	15.55	**30.25**	43.55	26.69	**20.86**	**39.04**
GPT-4o	rgb	11.66	28.11	41.94	**35.23**	20.32	27.81
GPT-4o	svf	**27.56**	29.54	**45.70**	32.38	17.11	34.22

(b) Land cover and Multi-factor categories

Model	Modality	Land use	LC type	Open.	Sky vis.	Bldg.d.	Vis.rng
Qwen2.5-VL-7B Base (no FT)							
Qwen2.5-VL-7B Base	all	26.60	**56.79**	27.96	14.23	35.29	19.22
Qwen2.5-VL-7B Base	dsm w/o leg.	8.51	41.43	28.32	12.81	31.18	20.64
Qwen2.5-VL-7B Base	svf	3.72	41.07	**30.82**	12.10	26.47	18.86
Qwen2.5-VL-7B Base	rgb	12.23	55.36	30.47	12.81	**42.94**	19.57
Qwen2.5-VL-7B Base	rgb+dsm (no leg.)	14.89	48.21	29.75	14.23	36.47	**21.35**
Qwen2.5-VL-7B Base	rgb+dsm (leg.)	15.43	49.29	29.39	13.52	40.00	20.28
Qwen2.5-VL-7B Base	dsm (legend)	6.38	31.07	30.47	**14.95**	26.47	21.00
Qwen2.5-VL-7B Base	all (arrays)	**30.32**	53.93	29.39	12.46	38.24	19.57
Qwen2.5-VL-7B Base	dsm (arrays)	1.06	25.71	26.88	14.59	27.06	17.79
Qwen2.5-VL-7B Base	svf (arrays)	4.79	15.71	28.32	10.32	28.82	17.79
GPT-4o (no FT)							
GPT-4o	all	**23.94**	**23.21**	**40.86**	29.89	**42.35**	21.71
GPT-4o	dsm (legend)	5.85	0.71	35.84	24.20	35.29	**25.62**
GPT-4o	rgb	21.81	16.43	37.28	**32.74**	41.76	24.91
GPT-4o	svf	6.91	0.36	37.99	23.13	37.65	18.15

Table 3. Modality ablation (inference only) — Detailed categories grouped by SVF-aware, height-aware, land cover, and multi-factor tasks. Abbreviations: SVF val. = SVF value, Reg. rank = region ranking, R.SVF.v = regional SVF variability, Sun.e = sun exposure, Hgt.avg = height average, Highest = highest region, Land use = top land uses, LC type = land cover type, Open. = spatial openness, Sky vis. = sky visibility, Bldg.d. = building density, Vis.rng = visibility range. Qwen2.5-VL-7B Base (no FT) and GPT-4o are evaluated under different input modalities. "LULC (land cover & land use)" denotes combined land use/land cover accuracy.

typical output tendencies (e.g., stripping chain-of-thought prefixes, extracting the last line as the final answer, or canonicalizing comma-separated lists). We counted an output as a formatting error only when no valid answer could be recovered under these tolerant rules. Quantitatively, o4-mini exhibited a high overall rate of invalid outputs, whereas Gemini-2.5-Flash maintained a lower overall error rate but showed pronounced spikes in specific categories (e.g., *SVF value*). The o4-mini often repeats questions without answers for difficult questions, whereas Gemini-2.5-Flash sometimes exceeds the total token limits (including thinking tokens), and increasing the limits leads to redundant answers without following answer instructions. For such "thinking-only" responses that never commit to a final answer string, we classify the outputs as formatting errors in our statistics. Both models had large error rates in the SVF value, which estimates the absolute average SVF value in the region, likely because absolute inference is difficult. The combined summary is presented in Table 8. The per-category error counts in this table sum to the overall error counts (e.g., 4833/10232

Recipe	Modality at inference	Overall	Height inf.	LULC	SVF inf.	Multi inf.
FT with arrays for DSM/SVF (used colormap at inference)						
FT (arrays)	all	44.83	18.72	75.00	40.23	**47.13**
FT (arrays)	dsm w/o legend	39.18	20.32	50.00	37.86	44.26
FT (arrays)	rgb	45.01	27.54	**77.56**	38.17	45.36
FT (arrays)	svf	39.08	15.24	42.74	**41.60**	44.40
FT (arrays)	dsm with legend	**46.32**	**32.35**	76.92	39.01	46.99
FT with RGB						
FT (rgb)	rgb	50.10	44.12	**83.76**	41.37	47.27
FT (rgb)	dsm+rgb with legend	**50.52**	**50.53**	79.70	**41.83**	**47.40**
FT with RGB+DSM (visualized)						
FT (rgb+dsm)	all	**48.40**	48.20	76.23	39.49	**46.60**
FT (rgb+dsm)	dsm	42.73	48.71	41.19	**40.88**	43.98
FT (rgb+dsm)	dsm+rgb	47.27	**50.00**	75.82	37.87	44.37
FT (rgb+dsm)	rgb	46.77	44.85	**76.64**	37.87	44.50
FT (rgb+dsm)	svf	39.03	19.33	43.65	40.15	44.11
FT with necessary modalities (agent-selected)						
FT (necessary)	all	53.40	58.82	88.03	41.91	49.04
FT (necessary)	dsm	45.11	64.17	42.74	39.16	47.54
FT (necessary)	rgb	41.57	27.54	55.13	37.48	47.40
FT (necessary)	seg	45.98	26.74	**90.60**	34.96	46.99
FT (necessary)	svf	43.59	41.18	41.24	43.13	47.13
FT (necessary)	dsm+rgb	48.37	63.64	63.03	39.01	47.95
FT (necessary-only)	necessary-only (no rgb)	**57.35**	**69.25**	88.46	**46.03**	**51.64**
Two-Stage Baseline (Prediction → Agent)						
Two-Stage Pipeline	RGB → Pred. DSM/SVF	33.91	14.97	40.81	33.59	39.75

Table 4. Modality ablation (fine-tuned) — Summary metrics. Accuracy (overall), height inference, LULC, SVF inference, and multi-factor inference (%). The Two-Stage Pipeline serves as a baseline for explicit depth estimation approaches.

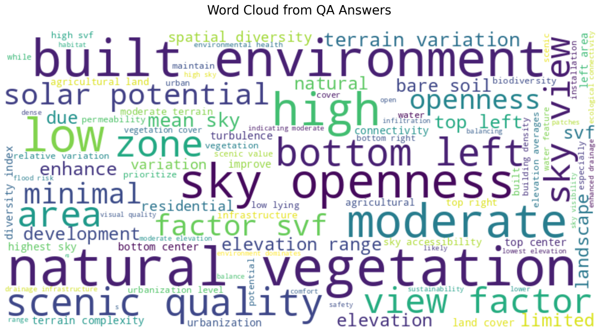


Figure 1. Word cloud of ground-truth answers for Tier-3 free-form Q&As.

for o4-mini).

Performance analysis insights. Among the non-fine-tuned models, Gemini-2.5-Flash attained the highest overall accuracy (33.0%), likely because of its hybrid reasoning architecture with thinking capabilities. Without fine-tuning, Gemini-2.5-Flash performs strongly across short-answer categories, except for height inference. We assume that pre-existing 2D geospatial datasets and deliberate option-by-option comparisons achieved by reasoning conventions improved selection accuracy among candidates.

D. Spatial Metrics and Scientific Methodology

D.0.1. Scientific Rationale for Spatial Metrics

Rationale and weight selection. We derived the weighting schemes for our composite indices from established geospatial and environmental design literature. Primary physical determinants dominate each score, whereas proxy and auxiliary indicators contribute only marginally.

Urban density assessment relies on coverage and volume based metrics. Building coverage ratio (BCR) and floor area ratio (FAR) serve as fundamental descriptors of built form [9]. We assign BCR the highest weight (0.5) because surface occupancy directly shapes street level compactness. FAR receives a weight of 0.25 because it exhibits less per-

(a) SVF-aware and Height-aware categories

Recipe	Modality	SVF val.	Reg. rank	R.SVF.v	Sun.e	Hgt.avg	Highest
FT (arrays for DSM/SVF at inference)							
FT (arrays)	all	**39.22**	33.10	44.09	43.06	0.00	37.43
FT (arrays)	dsm	32.86	33.81	42.47	40.57	1.07	**39.57**
FT (arrays)	rgb	35.34	31.32	45.16	39.86	20.32	34.76
FT (arrays)	svf	31.80	**37.72**	**48.39**	**45.20**	0.00	30.48
FT (arrays)	dsm (legend)	34.98	33.81	45.70	41.99	**25.67**	39.04
FT with RGB							
FT (rgb)	rgb	**40.28**	35.59	45.70	43.06	49.73	38.50
FT (rgb)	dsm+rgb	37.10	**37.01**	**47.31**	**43.77**	**62.03**	**39.04**
FT with RGB+DSM (visualized)							
FT (rgb+dsm)	all	39.73	31.85	45.08	**41.44**	56.19	**40.21**
FT (rgb+dsm)	dsm	38.70	**36.64**	48.19	**41.44**	58.76	38.66
FT (rgb+dsm)	dsm+rgb	39.73	29.79	41.97	39.73	**60.31**	39.69
FT (rgb+dsm)	rgb	**40.75**	29.45	44.04	38.70	54.64	35.05
FT (rgb+dsm)	svf	35.27	34.59	**50.78**	**41.44**	10.31	28.35
FT with necessary modalities (agent-selected)							
FT (necessary)	all	**42.76**	38.43	42.47	39.86	74.33	43.32
FT (necessary)	dsm	35.69	33.81	45.16	36.30	83.42	44.92
FT (necessary)	rgb	33.22	32.03	41.40	41.64	24.06	31.02
FT (necessary)	seg	30.04	23.84	41.40	38.79	22.99	30.48
FT (necessary)	svf	38.87	39.86	48.92	44.13	49.20	33.16
FT (necessary)	dsm+rgb	35.34	33.81	44.09	36.65	85.03	42.25
FT (necessary)	necessary	37.81	**44.13**	**50.54**	**45.91**	**93.05**	**45.45**

(b) Land cover and Multi-factor categories

Recipe	Modality	Land use	LC type	Open.	Sky vis.	Bldg.d.	Vis.rng
FT (arrays for DSM/SVF at inference)							
FT (arrays)	all	54.26	88.93	43.01	**52.67**	**51.76**	**38.79**
FT (arrays)	dsm	22.87	68.21	41.22	51.25	45.88	36.30
FT (arrays)	rgb	**55.32**	**92.50**	41.58	51.60	49.41	36.65
FT (arrays)	svf	22.34	56.43	**47.31**	51.96	43.53	37.37
FT (arrays)	dsm (legend)	53.72	**92.50**	40.86	**52.67**	51.76	38.43
FT with RGB							
FT (rgb)	rgb	**65.43**	**96.07**	43.73	**55.16**	**54.12**	35.23
FT (rgb)	dsm+rgb	60.11	92.86	**45.88**	**55.16**	**54.12**	**35.59**
FT with RGB+DSM (visualized)							
FT (rgb+dsm)	all	57.14	89.04	41.24	**54.27**	**53.07**	34.93
FT (rgb+dsm)	dsm	19.39	55.82	41.92	50.17	38.55	**41.10**
FT (rgb+dsm)	dsm+rgb	59.69	86.64	39.52	53.58	39.11	38.36
FT (rgb+dsm)	rgb	**56.63**	**90.07**	38.49	51.88	50.28	33.56
FT (rgb+dsm)	svf	19.39	59.93	**42.27**	51.88	42.46	37.33
FT with necessary modalities (agent-selected)							
FT (necessary)	all	71.81	98.93	46.24	54.09	**55.88**	39.86
FT (necessary)	dsm	18.09	59.29	46.95	55.87	44.12	41.28
FT (necessary)	rgb	20.21	78.57	40.50	53.38	54.12	37.37
FT (necessary)	seg	**76.60**	**100.00**	43.01	51.25	54.12	38.43
FT (necessary)	svf	19.68	55.71	45.88	52.67	52.94	38.08
FT (necessary)	dsm+rgb	29.79	85.36	46.95	53.38	51.76	40.21
FT (necessary)	necessary	71.81	99.64	**53.41**	**56.58**	**55.88**	**44.13**

Table 5. Modality ablation (fine-tuned Qwen2.5-VL-7B) — Detailed categories grouped by SVF-aware, height-aware, land cover, and multi-factor tasks.

Model	Total	Observation	Logic	Conclusion
Gemini-2.5-Flash	1.96	1.92	2.66	2.34
GPT-4o	2.27	2.27	3.23	2.57
GPT-4.1-mini	2.53	2.52	3.37	3.02
o4-mini	2.05	1.98	2.69	2.41
LLaVA-one-vision	1.48	1.00	1.89	2.01
InternVL3-8B	2.22	2.08	2.94	2.62
Qwen2.5-VL-3B	1.38	1.00	1.90	2.11
Qwen2.5-VL-7B Base	2.04	2.06	2.72	2.23
TeoChat	1.45	1.39	1.91	1.66
GeoChat	1.15	1.00	1.68	1.67
Qwen2.5-VL-7B FT	**2.89**	**2.91**	**3.41**	**3.11**
(FT − Base)	↑0.85	↑0.85	↑0.69	↑0.88

Table 6. General criteria (free-form questions). Higher is better.

Category	Base	10K+free	100K	100K+Free
SVF-aware inference				
SVF value	6.8	27.9	42.5	42.7
region ranking	19.5	29.6	37.8	39.7
regional SVF variability	34.0	39.4	53.3	51.4
sun exposure	28.6	35.3	42.3	43.9
Height-aware inference				
height average	9.9	34.8	35.7	42.2
highest region	32.7	32.3	40.2	39.9
Land cover inference				
top land uses	11.6	47.2	58.6	56.5
land cover type	49.9	87.4	90.2	90.9
Multi-factor inference				
spatial openness	32.2	40.6	46.5	47.6
sky visibility	13.1	48.5	56.0	55.3
building density	39.8	45.0	49.7	50.4
visibility range	22.0	26.8	32.5	32.7
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**

This table demonstrates the synergistic effects of combining short-answer and free-form QA training. Key observations: (1) Height-aware tasks benefit significantly from free-form integration (+3.1 pp from 100 K to 100 K+Free); (2) Most categories show consistent improvements with free-form augmentation; (3) The 100 K+Free configuration achieves optimal balance between discrete supervision and structured reasoning.

Table 7. Comprehensive comparison of Qwen2.5-VL-7B variants showing synergistic effects of combining short-answer and free-form training. All values show 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 respectively, mixed with 1 K free-form Q&As.

ceptual salience despite its importance in planning contexts.

Secondary indicators complement these primary metrics. The SVF complement captures geometric enclosure [10, 11] with a moderate weight (0.15). Edge density reflects morphological fragmentation, while local daytime RGB brightness and shadowing provide only a weak proxy for built intensity. Dense built-up areas tend to appear darker during daytime imagery due to cast shadows, higher impervious surface ratios, and reduced vegetation cover [12–14]. Both indicators therefore receive minimal weights (0.05), given their strong dependence on illumination conditions, sensor viewing geometry, and surface properties rather than urban form alone.

Category	o4-mini		Gemini-2.5-Flash	
	Errors	Rate (%)	Errors	Rate (%)
Overall	4833/10232	47.23	1666/10232	16.28
SVF_value	692	69.27	811	81.18
region_ranking	640	64.06	36	3.60
height_average	299	44.89	305	45.80
highest_region	369	55.41	128	19.22
top_land_uses	151	22.67	4	0.60
landcover_type	575	57.91	85	8.56
spatial_openness	443	44.34	50	5.01
regional_svf_variability	110	16.62	14	2.11
sky_visibility	436	43.64	100	10.01
sun_exposure	356	35.64	66	6.61
building_density	225	38.46	20	3.42
visibility_range	537	53.75	47	4.70

Table 8. Format error counts and rates across categories for o4-mini and Gemini-2.5-Flash.

Spatial openness assessment emphasizes geometric enclosure over surface coverage. The topographic Openness Index [15] dominates the score (0.5) because it directly captures elevation based enclosure. SVF complements this geometric descriptor (0.25) by quantifying sky access from surrounding structures and vegetation [16].

Because building density correlates with SVF, we assign it a lower weight (0.15) to avoid double counting occlusion effects. Terrain flatness (0.05) accounts for slope's influence on viewsheds [17], while visual simplicity (0.05) reflects evidence that spatial complexity reduces perceived openness [18].

Sky visibility assessment relies on SVF to determine radiative exchange and vertical openness [16]. However, it can overestimate perceived openness when horizontal obstructions and near field barriers are significant [19–21]. We therefore assign SVF the highest coefficient (0.7) while explicitly combining it with building penalties (0.3). These penalties encode canyon geometry effects and local building occupancy that constrain sky access and urban climate processes [10, 22, 23].

Edge penalties remain small (0.05 to 0.025) because skyline roughness affects the subjective continuity of sky views but contributes less to the physical sky fraction [24].

These hierarchical weighting reflects both geometric theory and empirical evidence. Primary indices receive weights of at least 0.5, and secondary complements range from approximately 0.15 to 0.3, whereas proxies receive weights of 0.05 or less. This structure emphasizes the theoretical dominance of geometric descriptors while incorporating human perception and urban climate evidence.

Supplementary Explanation for the Scientific Rationale and the Weighting Schemes.

Sky View Factor and Sky Visibility Methodology. The Sky View Factor (SVF) quantifies the hemispherical sky fraction visible from a given point. This geometric metric has become foundational in urban climatology [25, 26] with applications covering energy efficiency [27, 28] and solar potential analysis [29]. Our computational methodology follows established geometric algorithms for hemispherical sky visibility calculations, incorporating digital surface model data for accurate obstruction modeling.

However, SVF captures primarily vertical openness. Recent reviews note that it may overestimate perceived openness when lateral obstructions are significant [19–21]. Street canyon studies link lower SVF values to stronger horizon obstruction and heat island intensity [16, 30]. Yet lateral barriers such as building walls, tree belts, and near field structures can block sky access even where SVF remains high.

We therefore model sky visibility as a composite measure governed by SVF but modulated by local land cover and building configuration. This approach better approximates perceptual sky access by combining SVF with land cover derived building occupancy and edge based penalties.

To consider perceptual sky visibility, we introduce two stabilizing terms. First, a *building penalty* based on normalized building occupancy in a local window approximates sub hemispherical occlusions from nearby massing, consistent with obstruction aware SVF estimators [31, 32]. Second, an *edge penalty* discourages boundary adjacent points where horizon estimation and classification transitions are unstable due to façade edges and material transitions.

The resulting score $V = 0.7\,\mathrm{SVF} - 0.3\,\mathrm{BuildingPenalty} - w_e\,\mathrm{EdgePenalty}$ favors locations that jointly maximize sky access and minimize the immediate occlusion risk. We use a small edge penalty coefficient ($w_e \in \{0.05,\ 0.025\}$) to keep skyline roughness as a secondary modifier consistent with the weighting scheme above.

Spatial openness (Openness Assessment). Perceived spatial openness reflects the balance of sky access, low obstruction, gentle terrain, and low visual clutter. We first derive an *OpennessIndex* from the mean SVF and its variability. This index serves as the dominant geometric descriptor in the openness score.

We then combine this index with the mean SVF, the complement of building density, terrain flatness, and a visual simplicity term derived from edge variance. These components use the weights described above. Prior work connects SVF and obstruction to perceived spaciousness and comfort in streetscapes [27, 33] and relates terrain smoothness to functional and visual quality [34]. Our exponential penalties for height variability and edge variance reflect the diminishing openness under rugged relief and textural clutter.

Visibility range. The visibility range extends beyond local openness to incorporate the geometric sightline length across the terrain. Standard viewshed analysis on DSMs models horizon blocking and line of sight constraints. This approach is widely used for environmental monitoring and path planning [35].

Landscape visibility surveys emphasize that horizon structure and local relief shape visible extent and perceived view quality [36]. These studies recommend multi criteria formulations that combine viewshed extent with openness proxies. We therefore mix the viewshed (60%), local SVF context (25%), and terrain roughness (15%) to balance the raw sightline reach, occlusion likelihood, and terrain comfort.

Urban density. Urban density is characterized by coverage and volume based metrics. The building coverage ratio (BCR) and floor area ratio (FAR) are canonical in urban planning and daylight regulation. We complement these with an SVF component ($1 - \overline{\mathrm{SVF}}$) that captures enclosure from local massing. This term is clamped to zero in nature dominant regions to avoid conflating canopy cover with urban density.

Prior studies have connected building density, enclosure, and visual or environmental quality [37–39]. Edge density and RGB brightness modestly increase the density score as proxies for fragmentation and surface shadowing. Both receive small weights, consistent with their role as weak modifiers in the index.

D.0.2. Metric Computation Framework

Each category employs explicit scoring functions based on standard geospatial metrics:

- **Urban Density**: $D = 0.5\,\mathrm{BCR} + 0.25\,\mathrm{FAR} + 0.15\,\mathrm{SVF\_comp} + 0.05\,\mathrm{EdgeDensity} + 0.05\,(1 - \mathrm{Brightness})$, where $\mathrm{SVF\_comp} = 1 - \overline{\mathrm{SVF}}$ with a correction to 0 under nature-dominant regions
- **Openness Assessment**: $O = 0.5\,\mathrm{OpennessIndex} + 0.25\,\overline{\mathrm{SVF}} + 0.15\,(1 - \mathrm{BuildingDensity}) + 0.05\,\mathrm{TerrainFlatness} + 0.05\,\mathrm{VisualSimplicity}$
- **Sky Visibility**: $V = 0.7\,\mathrm{SVF} - 0.3\,\mathrm{BuildingPenalty} - w_e\,\mathrm{EdgePenalty}$, with $w_e \in \{0.05,\ 0.025\}$ for standard/hard settings

To evaluate the reliability of these composite metrics and the derived results, we combined statistical validation with multimodal visual inspection. We performed correlation and sensitivity analyses between the indices and their constituent SVF, DSM, and land cover components. We also manually inspected overlaid RGB, DSM, SVF, and segmentation maps for representative scenes to verify that high or low scores corresponded to intuitively open or enclosed locations.

Qualitative weight sensitivity. To assess the robustness of these handcrafted weights, we ran a small sensitivity study on the *sky_visibility* metric by perturbing individual coefficients by ± 0.1 (e.g., the SVF weight, building-penalty weight, window size normalization, and edge-penalty weights) over 30 randomly sampled questions. Across all perturbations, the top-1 region changed in at most 6.7% of the sampled cases (with many settings exhibiting 0% change), indicating that the discrete argmax decisions were qualitatively stable under modest weight variations.

Parameterization and derived terms

- $\mathrm{OpennessIndex} = (\overline{\mathrm{SVF}} + 0.5\,\mathrm{std(SVF)})/1.25$
- $\mathrm{TerrainFlatness} \approx \exp(-\mathrm{std(height)}/5.0)$, $\mathrm{VisualSimplicity} \approx \exp(-\mathrm{edge\_var}/500)$
- $\mathrm{BuildingPenalty} = 0.3\,\mathrm{building\_ratio\_in\_window} \times \mathrm{norm}$ (window normalization)
- Viewshed is the mean of max line-of-sight distances over 8 directions normalized by an upper bound; $\mathrm{TerrainRoughness} \approx \exp(-(\mathrm{std(height)} - 10)^2/(2 \cdot 10^2))$
- $\mathrm{FAR} \approx \mathrm{BCR} \times \mathrm{avg\_floors}$, $\mathrm{avg\_floors} \approx \min(20, \mathrm{height}/3.5\,\mathrm{m})$, normalized by 5

The mapping from height (in meters) to the approximate number of stories adopts a representative 3.5 m per floor, which is commonly used in urban analytics for large-scale building-height approximation [40].

D.0.3. Task Category Definitions

We summarize the benchmark categories according to the three-tier taxonomy.

Tier 1 (T1): Single-Feature Analysis

Only SVF Categories:

- **Sun Exposure**: Identification of locations with highest solar exposure potential based on SVF values
- **Region Ranking**: Arrangement of regions by their openness levels from highest to lowest
- **Regional SVF variability**: Identification of regions with highest SVF standard deviation
- **Average SVF value**: Calculation of precise average SVF values for specific areas (1-decimal precision)

Only landcover map Categories:

- **Landcover type**: Identification of land-use types present in the image using landcover maps
- **Land Use**: Analysis of predominant landcover usage in specific areas

Only DSM Categories:

- **Height inference**: Calculation of average height in images using DSM data (10 m precision)
- **Highest Region**: Identification of highest elevation locations using DSM

Tier 2 (T2): Multi-Feature Analysis

SVF + landcover map Categories:

- **Sky Visibility**: Assessment of 'perceived' unobstructed sky view using SVF and landcover data

SVF + DSM Categories:

- **Visibility Range**: Determination of locations with longest visibility range using multiple modalities

SVF + DSM + landcover map Categories:

- **Spatial Openness**: Assessment of most expansive open locations considering SVF, terrain, and landcover
- **Building Density**: Evaluation of building concentration using SVF penalties and building height data

Tier 3 (T3): Free-form Caption

- **Urban Development/land use Application**: Analysis of urban development potential and recommendations for improving scenic quality, safety, and human–natural coexistence
- **Renewable Energy Installation**: Assessment of potential for solar panel and wind power generation installation
- **Landscape Analysis**: Comprehensive analysis of landscape characteristics including sky visibility, terrain, and landcover types
- **Water accumulation**: Analysis of water accumulation risk based on terrain and ground characteristics

D.1. Evaluation Templates and Coordinate System

We standardized the evaluation prompts across all models. All coordinates were normalized to a percentage in the range [0,100].

Coordinate Guide:

- Point: (x%, y%) where (0,0) is top-left. x is horizontal and y is vertical.
- Region: [xmin%, ymin%, xmax%, ymax%] defining a rectangle; require xmin < xmax, ymin < ymax.

Prompt-embedded scoring definitions. For all Tier 2 categories (*visibility_range*, *sky_visibility*, *spatial_openness*, and *building_density*), we explicitly embedded the scoring method and coordinate system description in the user prompt to remove ambiguity. The scoring definitions are exactly aligned with those in Table 9.

Format-Constrained Answers. We specified strict target formats to minimize invalid outputs, but the evaluation used a tolerant parser rather than treating each deviation as an error. The parser first normalizes whitespace and case, canonicalizes comma-separated lists, and strips common reasoning prefixes or suffixes to recover the final answer string. Outputs were only counted as formatting errors when no valid answer could be reliably extracted (e.g., missing numeric value, no region/point selected, or responses that contained only meta-reasoning without a final answer).

- Numeric (SVF mean): X.X in [0.0,1.0], 1 decimal.
- Height: "X m" with 10-meter increments.
- Category: exact label strings from the provided vocabulary (case-sensitive).
- Ranking: "Region X, Region Y, Region Z" (comma-separated, exact case).

Category	**Scoring Function**	**Question Type**	**Modalities**	**ID**
		Tier 1		
Sun Exposure	100% SVF value	Multiple choice	SVF only	(sun_exposure)
Average SVF value	Mean SVF of a region	Numeric (1-dec.)	SVF only	(SVF_value)
Region Ranking	Multi-region SVF ranking	near multiple choice (format specified)	SVF only	(region_ranking)
Regional SVF variability	Regional SVF std/coverage	Multiple choice	SVF only	(regional_svf_variability)
Land cover type	N/A (categorical)	Multi-label	Segmentation only	(landcover_type)
Land use	N/A (categorical)	Multi-label	Segmentation only	(top_land_uses)
Height inference	Mean height (10 m bins)	Numeric (”X m”)	DSM only	(height_average)
Highest Region	Max elevation	Multiple choice	DSM only	(highest_region)
		Tier 2		
Sky Visibility	$0.7\,\mathrm{SVF} - 0.3\,\mathrm{BuildingPenalty} - w_e\,\mathrm{EdgePenalty}$	Multiple choice	SVF + Segmentation	(sky_visibility)
Spatial Openness	$0.5\,\mathrm{OpennessIndex} + 0.25\,\overline{\mathrm{SVF}} + 0.15\,(1 - \mathrm{BuildingDensity}) + 0.05\,\mathrm{TerrainFlatness} + 0.05\,\mathrm{VisualSimplicity}$	Multiple choice	SVF + DSM + Segmentation	(spatial_openness)
Visibility Range	Viewshed (60%) + SVF (25%) + terrain (15%)	Multiple choice	SVF + DSM	(visibility_range)
Building Density	$0.5\,\mathrm{BCR} + 0.25\,\mathrm{FAR} + 0.15\,\mathrm{SVF\_comp} + 0.05\,\mathrm{EdgeDensity} + 0.05\,(1 - \mathrm{Brightness})$	Multiple choice	SVF + DSM + Segmentation	(building_density)

Table 9. Category taxonomy with scoring functions, output formats, modality requirements, and identifiers.

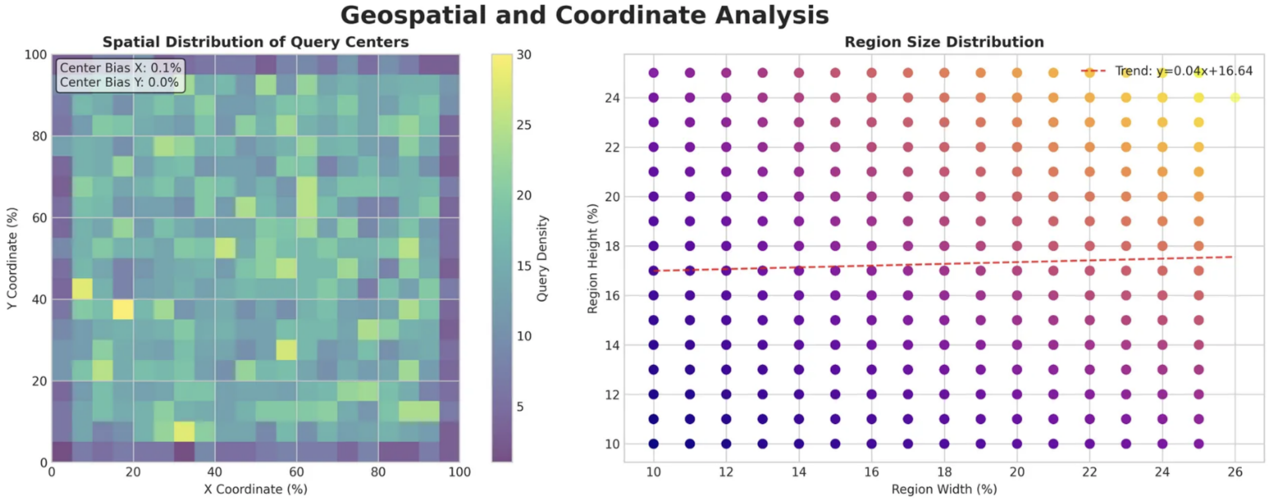


Figure 2. Coordinates of the options were uniformly sampled in the two-dimensional space, and the bounding boxes were selected to cover a wide range of scales. Points are denoted as $(x\%, y\%)$ with $(0, 0)$ at the top-left corner; regions are represented as $[x_{\min}\%, y_{\min}\%, x_{\max}\%, y_{\max}\%]$ with $x_{\min} < x_{\max}$ and $y_{\min} < y_{\max}$.

- Region choice: ”Region A”/”Region B”/”Region C”/”Region D” only.
- Point choice: ”Point (x.x%, y.y%)” exactly as listed.
- Balanced choices: correct-option indices are balanced across A/B/C/D and choices are shuffled without positional bias.
- Diversity: candidate selection enforces minimal similarity to avoid ambiguous ties.

D.1.1. Prompt paraphrasing robustness

To evaluate robustness to natural language variations, we created a paraphrased version of the benchmark in which GPT-4 rewrote each question based on the same task templates. We manually verified a random subset of ∼100 items to ensure that the paraphrases preserved the original semantics and the answer labels. Using this paraphrased set, we re-ran the evaluation for our best RGB-only model, Qwen2.5-VL-7B fine-tuned with the 100K+free recipe.

Setup	SVF	Height	LULC	Multi	Overall
Qwen Base (std)	23.5	21.3	34.5	22.6	24.8
FT (100K+free, std)	44.6	41.1	77.1	45.5	49.6
FT (100K+free, para)	32.7	33.6	71.0	39.9	40.8
FT(para) − FT(std)	-11.9	-7.5	-6.1	-5.6	-8.8

Table 10. Prompt paraphrasing robustness for Qwen2.5-VL-7B FT (100K+free). Accuracy (%) for major-category summaries and overall accuracy under the original evaluation templates vs GPT-4-paraphrased prompts. ”std” stands for standard templates, ”para” stands for paraphrased prompts. LULC denotes the combined land use/land cover major category and Multi denotes the multi-factor major category (cf. Table 7).

Table 10 summarizes the major-category accuracies for Qwen2.5-VL-7B Base and the fine-tuned 100K+free model under the original vs paraphrased prompts. Paraphrasing moderately degraded the fine-tuned model (8.8 pp drop in overall accuracy), but instruction tuning still yielded large gains over the baseline model under the original templates (approximately +20–40 pp across SVF, height, LULC, multi-factor, and overall). Even under paraphrased prompts, the fine-tuned model remained clearly above the baseline model evaluated using the standard templates. The format-error rate remains negligible (0.09% overall; 0.91% only in *land-cover_type*), indicating that our tolerant parser and format constraints generalize well to linguistically diverse prompts.

E. Extended Discussion and Worked Examples

We provide extended discussions and worked examples of the benchmark tasks.

E.0.1. Feasibility and Scope of RGB-to-3D Reasoning

Feasibility. While precise metric 3D estimation from monocular RGB is ill-posed owing to scale ambiguity, prior work has shown that *coarse* height and relative elevation cues can be learned from 2D indicators such as shadow geometry, occlusion boundaries, perspective distortions, and structural regularities [41–45]. Geo3DVQA targets this feasible regime.

Scope of the benchmark. We evaluate height-aware reasoning at decision-oriented granularity rather than full 3D reconstruction.

- **Categorical height distinctions:** accuracy is calculated on 10 m bins.
- **Relative comparisons:** which region is higher; which point has more sky access.
- **SVF pattern recognition:** open vs. enclosed spaces from RGB-visible urban density cues.
- **Multi-feature integration:** combining height, openness, and land cover for composite spatial metrics.

Practical applications. This coarse level supports (i) urban heat risk screening (tall + low SVF zones), (ii) solar siting pre-screening via sky openness, (iii) evacuation planning through open corridors and cluster identification, and (iv) accessibility planning via coarse terrain opennesswithout requiring specialized sensors. These applications prioritize actionable stratification over centimeter-level accuracies.

Geographic generalization and scalable deployment. As acknowledged in the main text limitation, our evaluation is currently limited to North Rhine-Westphalia (NRW) in Germany, which constrains generalization claims. However, the framework's design principles support scalable deployment and appropriate geographic expansion. The current GeoNRW limitation reflects data availability constraints rather than fundamental methodological limitations. For scalable real-world deployment, the following requirements should be addressed: (1) *Geographic diversity*: training data should encompass multiple regions with varying urban morphologies, terrain types, and climate zones to reduce geographic bias; (2) *Seasonal and temporal variation*: incorporating multi-temporal imagery across different seasons would improve robustness to phenological changes and weather conditions; and (3) *Adaptive fine-tuning*: region-specific fine-tuning or few-shot adaptation could bridge domain gaps when deploying to new geographic areas. Within the current NRW constraint, the framework has already demonstrated practical utility for prescreening, coarse stratification, and initial risk assessment tasks that do not require centimeter-level precision. The scalable deployment pathway involves the systematic expansion of training data across diverse geographic contexts, which is a natural next step for production systems rather than a fundamental limitation of the approach.

E.0.2. Dataset Statistics and Visualization

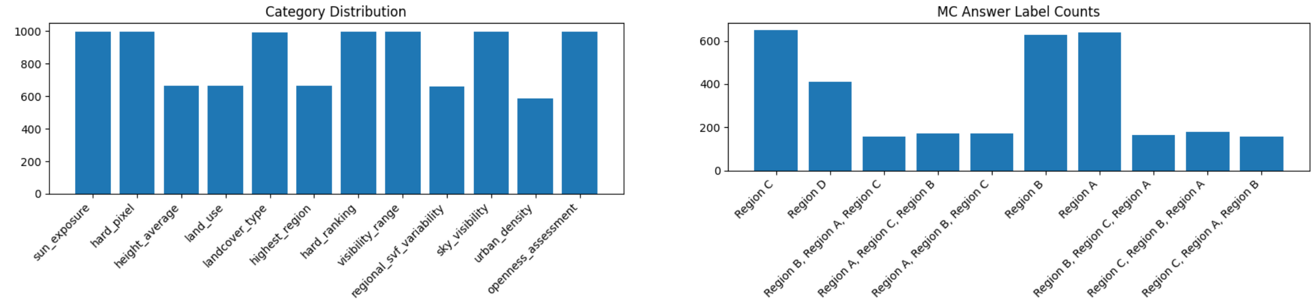
E.0.3. Model Size Constraints and Future Scaling

All local experiments were bounded to models with at most 10 B parameters owing to server memory limits for both training and inference. This constraint influences our architectural choices and may understate the attainable ceiling for elevation- and SVF-aware reasoning. Future work will investigate the scaling behavior with respect to free-form coherence, format compliance under constrained decoding, and multi-image conditioning for improved coordinate fidelity.

E.0.4. Input/Output Normalization and Examples

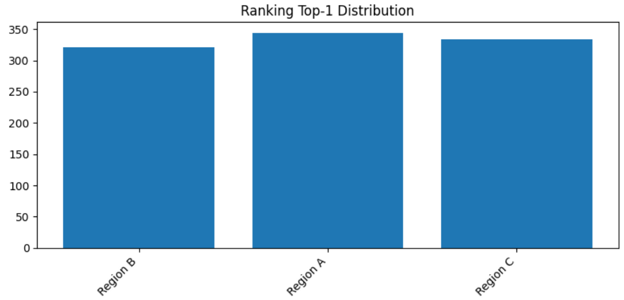
Input/Output Normalization Rules.

- Coordinates and Regions: See the Coordinate Guide for normalized coordinate definitions and region format.
- Output labels: fixed vocabulary for LU/LC; case-sensitive matching.
- Numerical outputs: clipped to valid ranges and rounded to specified precision.



(a) Per-category question counts (train/test).

(b) Correct label frequencies in multiple-choice.



(c) Top-1 region distribution in ranking tasks.

Figure 3. Geo3DVQA dataset statistics showing task-level dataset properties: (a) per-category question counts, (b) label frequencies in multiple-choice, (c) Top-1 region in ranking tasks.

F. Feasibility and Learnability Analysis

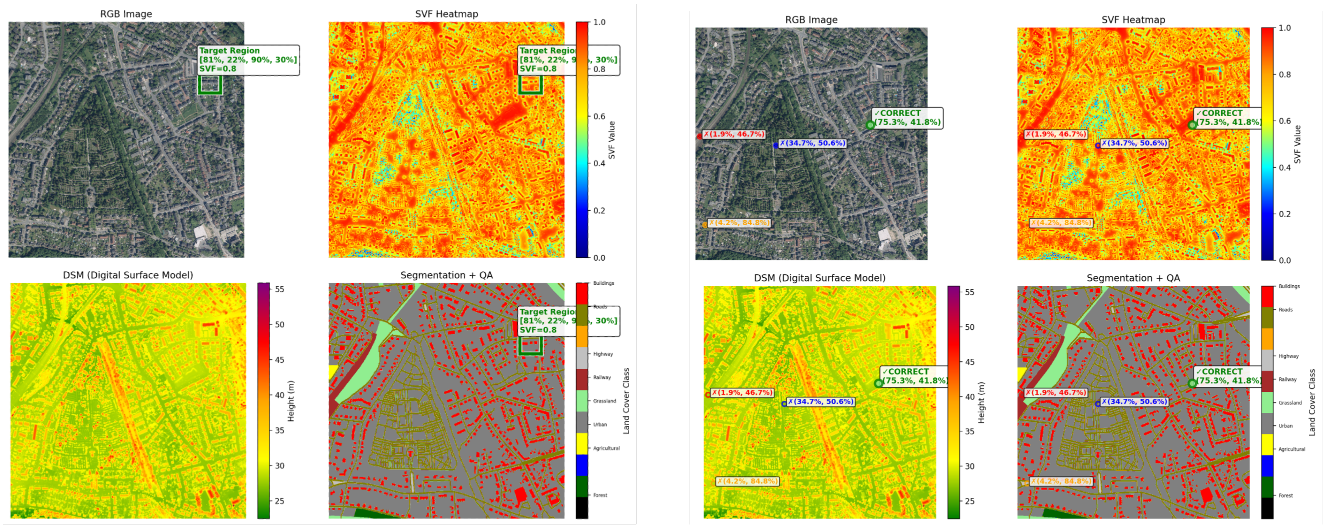
F.1. Feasibility and Learnability

Chance Level vs. Performance. While an overall accuracy of ∼50% might seem moderate, it must be contextualized against the random-chance level. For 4-choice questions, chance is 25%, and for multi-label or ranking tasks, it is significantly lower. The performance of the fine-tuned model (approx. double the chance level) confirms that it has learned meaningful geospatial patterns rather than relying on hallucinations.

Relative vs. Absolute Reasoning. Our analysis shows that the models perform better on relative tasks (e.g., Ranking, Highest Region) than on absolute metric regression. This aligns with the inherent ambiguities of monocular vision. For urban planning applications, relative accuracy (e.g., “Area A is denser than Area B”) is often sufficient for initial screening, supporting the practical feasibility of the approach despite the limitations of absolute metric precision.

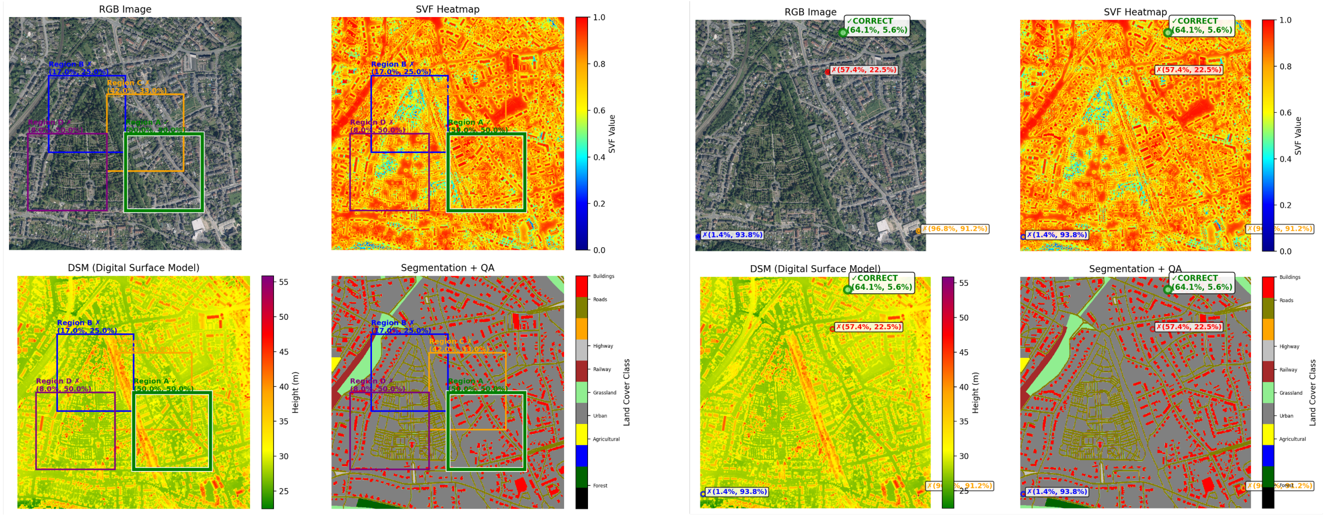
Semantic-Geometric Integration. A key advantage of VLMs is their ability to leverage semantic context. For instance, a pure depth model could have difficulty distinguishing a flat gray roof from gray pavement if height cues are ambiguous. However, a VLM can recognize an object as a ”school building” based on contextual cues, such as shape and the presence of a playground, and infer that it must be elevated. This semantic-geometric integration allows the model to resolve ambiguities that purely geometric methods cannot, which further supports the end-to-end paradigm.

Worked Examples.



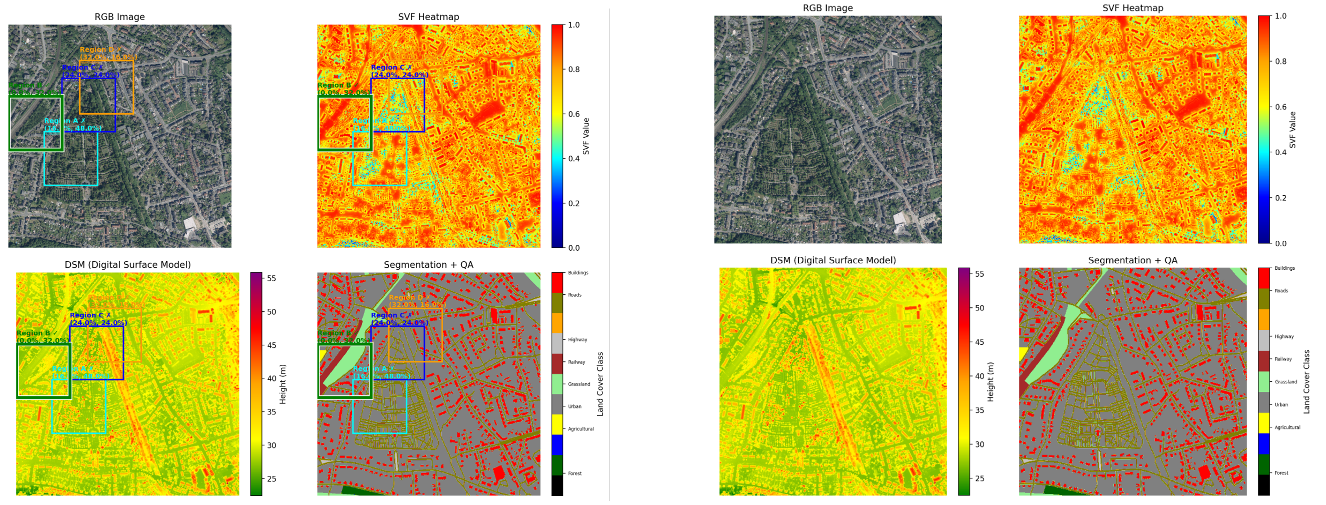
(a) **Q30 Mean SVF (Region)**. Q: Calculate the mean SVF within [81%, 22%, 90%, 30%]. A: 0.8

(b) **Q8 Visibility Range**. Q: Which location has the most comprehensive sightlines? A: Point (75.3%, 41.8%)



(c) **Q18 Urban Density**. Q: Which area has the most crowded urban layouts? A: Region A

(d) **Q22 Sky Visibility**. Q: Where is sky access the most unrestricted? A: Point (64.1%, 5.6%)



(e) **Q23 Spatial Openness**. Q: Which area demonstrates maximum openness with minimal obstruction? A: Region B

(f) **Q3 Height Average**. Q: Calculate the mean elevation within [19%, 47%, 25%, 53%]. A: 30 m

Figure 4. Visualized QA pairs for a sample scene. Each subfigure shows RGB, DSM, SVF, and segmentation overlays. Subcaptions include the question summary and the ground-truth answer.

Prompt and Output Listings

Listing 1. Example statistical input JSON

```
{
  "scene_id": "geonrw_sample_001",
  "svf_statistics": {
    "mean": 0.4237,
    "std": 0.2841,
    "min": 0.0023,
    "max": 0.9876,
    "quartiles": [0.1892, 0.3945, 0.6234]
  },
  "height_statistics": {
    "mean": 12.34,
    "std": 8.92,
    "building_coverage_ratio": 0.31,
    "max_height": 45.2
  },
  "land_cover_statistics": {
    "building_ratio": 0.28,
    "vegetation_ratio": 0.35,
    "road_ratio": 0.23,
    "water_ratio": 0.02,
    "other_ratio": 0.12
  },
  "spatial_statistics": {
    "edge_density": 0.156,
    "patch_count": 23,
    "largest_patch_ratio": 0.187
  },
  "rgb_statistics": {
    "brightness_mean": 127.3,
    "contrast_std": 31.2,
    "dominant_colors": ["#4A5D23", "#8B7355", "#A3A3A3"]
  },
  "derived_metrics": {
    "urban_density_score": 0.67,
    "openness_index": 0.43,
    "solar_potential": 0.58
  }
}
```

Listing 2. FT example: Urban Development (QID 1: koeln_364_5633_dem)

```
Question ID: 1 (koeln_364_5633_dem)
Question (excerpt): Analyze the potential of this area for urban development.

Ground truth sections (excerpt):
<OBSERVATION>Built environment coverage was 20.8%, with vegetation dominating at 66.5% (agricultural
    ↪ 57.4%, natural vegetation 9.2%). Mean SVF is 0.915. Elevation ranges 43.4-76.8 m (delta 33.4 m),
    ↪  relative variation 0.60. Urbanization level was 2/5 and the spatial diversity index was 1.320.
    ↪ The highest sky accessibility was in the middle center, and the best scenic quality was in the
    ↪ middle right.</OBSERVATION>
<CONCLUSION>Low building density and high SVF support comfortable expansion; prioritize top-center
    ↪ infrastructure; preserve middle-right scenic zones; leverage elevation variation for diversified
    ↪  fabric; maintain vegetation >50% with green corridors.</CONCLUSION>

Model sections (FT prediction):
<OBSERVATION> The built environment coverage was 21.5%, with residential areas dominating at 18.7%.
    ↪ Agricultural land covers 69.1% of the area, whereas natural vegetation is minimal at 0.1%. Mean
    ↪ SVF is 0.851. Elevation ranges 48.1-60.1 m (delta 12.0 m), terrain complexity 0.24. The highest
    ↪ sky accessibility top-left; bottom-right offers the best development suitability.</OBSERVATION>
<CONCLUSION>Maintain low building density to preserve scenic quality and sky openness. Prioritize
    ↪ bottom-right development, enhance green corridors by increasing natural vegetation, and improve
    ↪ infrastructure connectivity while maintaining a safe elevation.</CONCLUSION>
```

```
Validation summary:
Word count: 131; Scores: Total 3.0, Obs. 3, Conc. 3, SVF 4, LC 3, Elev. 3, Logic 4
```

Listing 3. FT example: Landscape Analysis (QID 7: guetersloh_454_5748_dem)

```
Question ID: 7 (guetersloh_454_5748_dem)
Question (excerpt): Analyze the overall landscape in terms of sky visibility, terrain, and land cover.

Ground truth sections (excerpt):
<OBSERVATION>Natural vegetation 32.5%, built 7.8%, forest 28.8%, agriculture 56.7%. Elevation
    ↪ 67.9-100.7 m (mean 76.5 m, delta 32.8 m, var. 0.43). The mean sky openness was 0.783
    ↪ (0.011-1.000); highest top-right. Spatial diversity index: 1.174; urbanization: 1/5; scenic
    ↪ quality: 0.727.</OBSERVATION>
<CONCLUSION>Natural-agricultural balance with low built density, moderate relief supports connectivity,
    ↪  top-right openness enhances solar and visual quality, conserve forest patches, and focus
    ↪ development top-left to minimize fragmentation.</CONCLUSION>

Model sections (FT prediction):
<OBSERVATION>Natural vegetation 12.5% (forest 11.6%); agriculture 79.8%; built 0.1%. Elevation
    ↪ 100.1-141.1 m (mean 116.1 m, delta 41.0 m, var. 0.35). The mean sky openness was 0.811
    ↪ (0.011-1.000); highest top-left. The Shannon diversity was 0.641. Scenic quality 0.711.</
    ↪ OBSERVATION>
<CONCLUSION>Agriculture-dominant landscape with limited urbanization; terrain variation supports
    ↪ connectivity; high sky openness enhances visual quality; conserve forest patches to maintain
    ↪ biodiversity.</CONCLUSION>

Validation summary:
Word count: 129; Scores: Total 3.0, Obs. 2, Conc. 3, SVF 4, LC 2, Elev. 3, Logic 3
```

Listing 4. Worked example (Qwen Base)

```
Question ID: 1 (kleve_306_5742_dem)
Ground truth observation (excerpt): mean SVF 0.942; elevation 12.8-43.3 m; built 7.0%; bottom-right
    ↪ suitable; middle-left scenic.
Model sections:
<OBSERVATION></OBSERVATION>
<ANALYSIS> ... general discussion without coordinates ... </ANALYSIS>
<CONCLUSION> ... vague recommendations ... </CONCLUSION>
Scores: Total 2.0, Obs. 1, Anal. 3, Conc. 2, SVF 2, LC 2, Elev. 1, Logic 3
```

Listing 5. Numeric (Region mean SVF)

```
System: You are a precise geospatial assistant. Answer concisely.
User: Consider an RGB image. Answer using the specified format only.
Question: Calculate the mean Sky View Factor within the area
         [65%, 81%, 79%, 95%].
Note: Coordinates are the percentages of the image size [xmin%, ymin%, xmax%, and ymax%].
Answer format: X.X (a single number in [0.0, 1.0], rounded to 1 decimal)
```

Listing 6. Point Selection (Sky visibility / Visibility range)

```
System: You are a precise geospatial assistant. Answer concisely.
User: Which location has the highest sky visibility?
Hint: Areas with fewer obstacles have a higher sky view factor.
Please choose from:
Point (89.7%, 20.6%)
Point (81.7%, 37.0%)
Point (11.3%, 3.9%)
Point (57.8%, 9.7%)
Answer format: Point (x.x%, y.y%)
```

Listing 7. Point Selection (Visibility range with scoring and coordinates)

```
System: You are a precise geospatial assistant. Answer concisely.
User: Which location has the most comprehensive sight lines?
Hint: Areas with good visibility typically have a high sky view factor and fewer obstacles in the line
```

```
    ↪ of sight.
Scoring method: Locations were scored solely based on viewshed distance analysis (60%), Sky View Factor
    ↪  (25%), and terrain roughness variation (15%). Higher scores indicate a better visibility range
    ↪ with longer line-of-sight distances.
Coordinate system: Each point is specified by (x, y) coordinates as percentages of the image dimensions,
    ↪  where (0, 0) represents the top-left corner. 'x' represents the horizontal position (from left
    ↪ to right), and 'y' represents the vertical position (from top to bottom).
Please choose from:
Point (75.3%, 41.8%)
Point (1.9%, 46.7%)
Point (34.7%, 50.6%)
Point (4.2%, 84.8%)
Answer format: Point (x.x%, y.y%)
```

Listing 8. Region Ranking

```
System: You are a precise geospatial assistant. Answer concisely.
User: Compare the SVF values of the three regions. Reply with
      The order from highest to lowest was labeled A/B/C.
Regions (x1%, y1%, x2%, y2%):
  A: (46%, 60%, 65%, 75%)
  B: (44%, 48%, 63%, 63%)
  C: (28%, 32%, 46%, 48%)
Answer format:
"Region A, Region B, Region C"
```

Listing 9. Multi-Choice Region Selection

```
System: You are a precise geospatial assistant. Answer concisely.
User: Which region best matches this criterion? Choose one.

A: [22%, 67%, 41%, 87%]
B: [77%, 6%, 98%, 27%]
C: [70%, 46%, 91%, 67%]
D: [75%, 12%, 95%, 32%]

Please choose from:
Region A
Region B
Region C
Region D
Answer format:
Region X
```

Listing 10. Land Use / Landcover (Multi-Label)

```
System: You are a precise geospatial assistant. Answer concisely.
User: Which land use types are most frequent in the region
      (16%, 48%, 40%, 72%)? Choose from the list.
Choices: residential, agricultural, forest, grassland,
         railways, roads, bare_soil, buildings, water, other
Answer format:
Comma-separated labels in lowercase.
```

G. Data and Code Availability

Release scope and licensing constraints. We are committed to maximizing reproducibility within the GeoNRW data licensing restrictions. Upon publication, we plan to release the following:

Publicly released components (unrestricted access):

- Question–answer pairs with image identifiers (110k items) and splits
- Evaluation scripts and rubric-based assessors for free-form scoring
- Fine-tuned model checkpoints (Qwen2.5-VL-7B) and training code
- Documentation of task definitions, metrics, and validation procedures

Access-restricted components (require GeoNRW license):

- Raw aerial RGB imagery (1 m) and raw DSM/segmentation rasters

A project page with links to the repository and model weights will be announced on our GitHub repository.

References

[1] S. Ouyang, J. M. Zhang, M. Harman, and M. Wang, "An empirical study of the non-determinism of chatgpt in code generation," *ACM Transactions on Software Engineering and Methodology*, vol. 34, no. 2, pp. 1–28, 2025. 1

[2] M. Renze, "The effect of sampling temperature on problem solving in large language models," in *Findings of the association for computational linguistics: EMNLP 2024*, pp. 7346–7356, 2024. 1

[3] "jaccard_score — scikit-learn.org." `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.jaccard_score.html`. [Accessed 16-09-2025]. 3

[4] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010. 3

[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014. 3

[6] E. Rodriguez, C. S. Morris, and J. E. Belz, "A global assessment of the SRTM performance," *Photogrammetric Engineering & Remote Sensing*, vol. 72, no. 3, pp. 249–260, 2006. 3

[7] N. Lang, N. Kalischek, J. Armston, K. Schindler, R. Dubayah, and J. D. Wegner, "Global canopy height regression and uncertainty estimation from GEDI LIDAR waveforms with deep ensembles," *Remote sensing of environment*, vol. 268, p. 112760, 2022. 3

[8] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International journal of forecasting*, vol. 22, no. 4, pp. 679–688, 2006. 3

[9] M. B. Pont and P. Haupt, *Spacematrix: Space, Density and Urban Form-revised edition*. TU Delft OPEN Publishing, 2023. 7

[10] C. Grimmond, S. Potter, H. Zutter, and C. Souch, "Rapid methods to estimate sky-view factors applied to urban areas," *International Journal of Climatology: A Journal of the Royal Meteorological Society*, vol. 21, no. 7, pp. 903–913, 2001. 9

[11] L. Chen, E. Ng, X. An, C. Ren, M. Lee, U. Wang, and Z. He, "Sky view factor analysis of street canyons and its implications for daytime intra-urban air temperature differentials in high-rise, high-density urban areas of Hong Kong: a GIS-based simulation approach," *International Journal of Climatology*, vol. 32, no. 1, pp. 121–136, 2012. 9

[12] M. Morabito, A. Crisci, T. Georgiadis, S. Orlandini, M. Munafò, L. Congedo, P. Rota, and M. Zazzi, "Urban imperviousness effects on summer surface temperatures nearby residential buildings in different urban zones of parma," *Remote Sensing*, vol. 10, no. 1, p. 26, 2017. 9

[13] S. Bonafoni and C. Keeratikasikorn, "Land surface temperature and urban density: Multiyear modeling and relationship analysis using modis and landsat data," *Remote Sensing*, vol. 10, no. 9, p. 1471, 2018.

[14] E. S. Krayenhoff and J. A. Voogt, "Daytime thermal anisotropy of urban neighbourhoods: Morphological causation," *Remote Sensing*, vol. 8, no. 2, p. 108, 2016. 9

[15] R. Yokoyama, M. Shirasawa, and R. J. Pike, "Visualizing topography by openness: A new application of image processing to digital elevation models," *Photogrammetric engineering and remote sensing*, vol. 68, no. 3, pp. 257–266, 2002. 9

[16] T. R. Oke, "Canyon geometry and the nocturnal urban heat island: comparison of scale model and field observations," *Journal of climatology*, vol. 1, no. 3, pp. 237–254, 1981. 9, 10

[17] P. F. Fisher, "Algorithm and implementation uncertainty in viewshed analysis," *International Journal of Geographical Information Science*, vol. 7, no. 4, pp. 331–347, 1993. 9

[18] R. Ewing and S. Handy, "Measuring the unmeasurable: Urban design qualities related to walkability," *Journal of Urban design*, vol. 14, no. 1, pp. 65–84, 2009. 9

[19] 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. 9, 10

[20] 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.

[21] 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. 9, 10

[22] 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. 9

[23] Y. Xia, N. Yabuki, and T. Fukuda, "Sky view factor estimation from street view images based on semantic segmentation," *Urban Climate*, 2021. 9

[24] A. E. Stamps III, "Enclosure and safety in urbanscapes," *Environment and behavior*, vol. 37, no. 1, pp. 102–133, 2005. 9

[25] 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. 10

[26] B. Zheng and J. Li, "Evaluating the Annual Effect of the Sky View Factor on the Indoor Thermal Environment of Residential Buildings by ENVI-met," *Buildings*, vol. 12, no. 6, pp. 787–787, 2022. 10

[27] 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. 10

[28] J. Appelbaum and A. Aronescu, "View Factors of Flat Collectors, Including Photovoltaics, Visible to Partial Sky," *Energies*, vol. 15, no. 22, pp. 8742–8742, 2022. 10

[29] 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. 10

[30] 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. 10

[31] L. Zeng, J. Lu, W. Li, and Y. Li, "A fast approach for large-scale Sky View Factor estimation using street view images," *Building and Environment*, vol. 135, pp. 74–84, 2018. 10

[32] F.-Y. Gong, Z.-C. Zeng, F. Zhang, X. Li, E. Ng, and L. K. Norford, "Mapping sky, tree, and building view factors of street canyons in a high-density urban environment," *Building and Environment*, vol. 134, pp. 155–167, 2018. 10

[33] 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. 10

[34] 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. 10

[35] Z. Pan, J. Tang, T. Tjahjadi, Z. Wu, and X. Xiao, "A novel rapid method for viewshed computation on dem through max-pooling and min-expected height," *ISPRS Int. J. Geo Inf.*, vol. 9, p. 633, 2020. 10

[36] Z. Wu, Y. Wang, W. Gan, Y. Zou, W. Dong, S. Zhou, and M. Wang, "A survey of the landscape visibility analysis tools and technical improvements," *International Journal of Environmental Research and Public Health*, vol. 20, no. 3, p. 1788, 2023. 10

[37] Y. Liu, M. Chen, M. Wang, J. Huang, F. Thomas, K. Rahimi, and M. Mamouei, "An interpretable machine learning framework for measuring urban perceptions from panoramic street view images," *iScience*, vol. 26, no. 3, 2023. 10

[38] F. J. Abarca-Alvarez, F. S. Campos-Sánchez, and F. Osuna-Pérez, "Urban shape and built density metrics through the analysis of European urban fabrics using artificial intelligence," *Sustainability*, vol. 11, no. 23, p. 6622, 2019.

[39] S. Frank, C. F"urst, L. Koschke, A. Witt, and F. Makeschin, "Assessment of landscape aesthetics—Validation of a landscape metrics-based assessment by visual estimation of the scenic beauty," *Ecological Indicators*, vol. 32, pp. 222–231, 2013. 10

[40] H. Usui, "Comparison of precise and approximated building height: Estimation from number of building storeys and spatial variations in the Tokyo metropolitan region," *Environment and Planning B: Urban Analytics and City Science*, vol. 50, pp. 487–499, 2022. 11

[41] 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," pp. 10478–10485, 2020. 13

[42] 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.

[43] F. Lourenço and H. Araújo, "Intel RealSense SR305, D415 and L515: Experimental evaluation and comparison of depth estimation," pp. 362–369, 2021.

[44] 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.

[45] 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. 13