TIMESPOT: BENCHMARKING GEO-TEMPORAL UNDERSTANDING IN VISION–LANGUAGE MODELS IN REAL-WORLD SETTINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Geo-temporal understanding, the ability to identify the location, time, and contextual features of an image from visual cues alone, is a fundamental aspect of human intelligence with wide-ranging applications, from disaster response to autonomous navigation and geography education. While recent vision–language models (VLMs) have shown progress in image geo-localization using conspicuous cues like landmarks or road signs, their ability to understand temporal signals and related spatial reasoning cues remains underexplored. To address this gap, we introduce **TIMESPOT**, a comprehensive benchmark for evaluating real-world geo-temporal reasoning in VLMs. TIMESPOT comprises 1,455 images spanning 80 countries, where models must infer temporal attributes (season, month, time of day, daylight phase) and geolocation attributes (continent, country, climate zone, environment type, latitude–longitude coordinates) directly from the visual input. In addition, it includes spatial reasoning tasks that require integrating geographical, spatial, and temporal cues to solve complex understanding problems. Unlike prior benchmarks that emphasize obvious cues or iconic imagery, TIMESPOT prioritizes diverse and subtle settings, reflecting the difficulty of reasoning under real-world uncertainty. Our evaluation of state-of-the-art VLMs, including both open- and closed-source models, reveals consistently low performance across tasks, highlighting substantial challenges in achieving robust temporal and geographic reasoning. These findings underscore the pressing need for improved methods to enable reliable and trustworthy geo-temporal understanding in VLMs, paving the way for future research in this critical domain.

1 INTRODUCTION

Determining *where* and *when* a photograph was captured, using only its visual content, is a core aspect of human cognition, enabling situational awareness, memory coordination, and contextual reasoning. This capacity, which we term **geo-temporal understanding**, integrates a rich tapestry of visual cues: physical illumination and shadows; seasonal vegetation; architectural styles and materials; clothing and traffic patterns; and broader geographic regularities (Lin et al., 2013; Workman et al., 2015b; Arandjelovic et al., 2016; Hu et al., 2018; Hu & Lee, 2020). Beyond its cognitive significance, geo-temporal reasoning underpins critical applications in disaster response (Mirowski et al., 2018), environmental and climate monitoring (Zhai et al., 2017), autonomous navigation (Lynen et al., 2020; Sarlin et al., 2019), and media forensics (Tian et al., 2017).

In the geospatial domain, substantial progress has been made through cross-view and street-view benchmarks (Vo & Hays, 2016), evolving from early ground-to-aerial matching to large-scale, geographically diverse corpora and unified embeddings such as VIGOR (Zhu et al., 2021), GeoCLIP (Vivanco Cepeda et al., 2023), OpenStreetView 5M (Astruc et al., 2024), Global Streetscapes (Hou et al., 2024), CV, Cities (Huang et al., 2024), panoramic cross-view settings (Xia et al., 2025), and embedding advances (Cai et al., 2025). However, geo-temporal understanding—joint inference of geographic and temporal attributes from visual input—remains underexplored. Existing geolocation benchmarks typically assess where via retrieval accuracy or coordinate error (Hu & Lee, 2020), without requiring explicit when predictions (e.g., season, month, local time) or enforcing internal consistency (e.g., "July" paired with "winter" in the Northern Hemisphere). Temporal

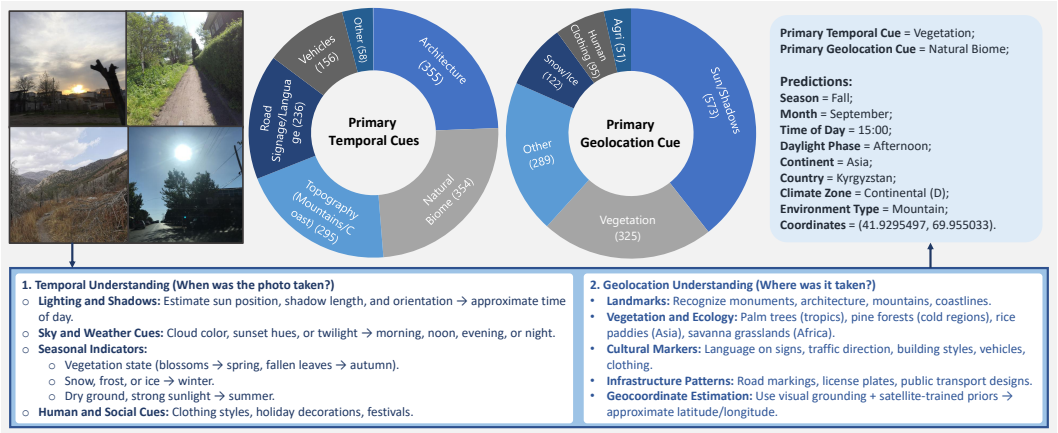


Figure 1: Illustration of the TIMESPOT benchmark for geo-temporal understanding. Models must infer temporal attributes (season, month, time of day, daylight phase) and geolocation attributes (continent, country, climate zone, environment type, coordinates) directly from visual input. Left: example images. Center: distributions of primary temporal cues (e.g., architecture, natural biome, topography) and geolocation cues (e.g., sun/shadows, vegetation, snow/ice). Right: an example prediction, showing how diverse and subtle cues must be integrated for reliable reasoning.

cues, when studied, are often isolated or reduced to proxy tasks such as navigation (Mirowski et al., 2018; Lynen et al., 2020), place recognition (Sarlin et al., 2019), or change detection (Sarlin et al., 2024), which do not yield structured temporal outputs. Recent geospatial VLM benchmarks in remote sensing (HRVQA (Li et al., 2024b), GEOBench-VLM (Danish et al., 2024)) focus on aerial imagery for classification, counting, and segmentation, but do not require fine-grained temporal fields or a comprehensive geographic schema (continent, country, climate, environment type, latitude, longitude) for ground-level photos. The field thus lacks a unified testbed evaluating time and place together, emphasizing subtle non-iconic cues over landmarks or text, and incorporating trustworthiness measures such as schema validity, calibration, and robustness to distribution shift (Astruc et al., 2024; Hou et al., 2024; Huang et al., 2024; Xia et al., 2025).

This gap is critical for three reasons. **First**, temporal cues are often decisive for disambiguation: solar geometry depends on hemisphere (Ye et al., 2024), vegetation phenology distinguishes climates at similar latitudes (Wang et al., 2024c), and daylight phases modulate urban lighting and activity patterns (Shi et al., 2020; Zhu et al., 2022). **Second**, reliable deployment in real-world settings demands *structured, verifiable outputs*—minute, level precision for local time, kilometer, range thresholds for coordinates, and cross, field consistency checks (e.g., rejecting "snow in July" for a Northern Hemisphere location)—which current retrieval, centric protocols lack, risking over-reliance on superficial cues. **Third**, robust global generalization requires *diversity and stress testing*: hemisphere flips, climate-region shifts, and out, of, distribution (OOD) splits expose biases that iconic or text-dependent benchmarks fail to reveal (Astruc et al., 2024).

To address these challenges, we introduce TIMESPOT, a benchmark for evaluating real-world geo-temporal understanding in VLMs. TIMESPOT comprises 1,455 natural, non-iconic images spanning 80 countries, curated to minimize landmark dependence and emphasize low-salience cues. For each image, models must predict a *structured schema* with four temporal attributes (season, month, local time, daylight phase) and five geographic attributes (continent, country, climate zone, environment type, latitude, longitude). This design encourages models to fuse illumination, material, and contextual evidence beyond textual shortcuts. Beyond per, field predictions, TIMESPOT includes *fusion questions* requiring integrated spatial and temporal reasoning, mirroring real-world scenarios where time and place are interdependent. Our key contributions are threefold:

1. **A challenging joint geo-temporal benchmark.** To our knowledge, TIMESPOT is the first to require simultaneous prediction of four temporal and five geographic attributes from natural images, producing *structured, interpretable outputs* that probe fine-grained spatiotemporal reasoning beyond retrieval or classification.

2. **Rigorous, verifiable evaluation of modern VLMs.** We evaluate diverse open, and closed, source models under uniform, metadata-free, and fully open, ended Q&A condition, en-

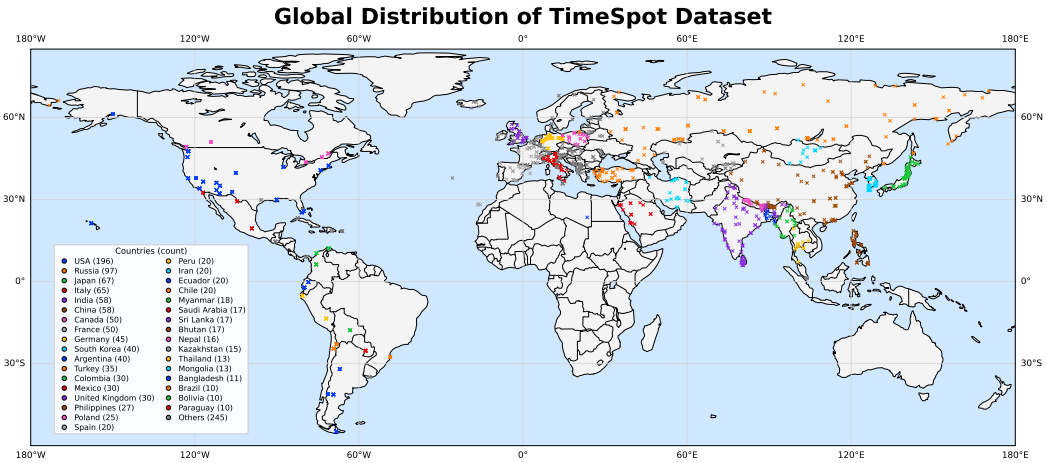


Figure 2: Global coverage of TIMESPOT. Locations of all 1,455 ground, level photos across 80 countries. Each marker denotes an image coordinate; colors indicate country (top contributors listed; "Others" aggregates the remainder). The dataset spans all inhabited continents and diverse climate zones and environments, providing non, iconic scenes for evaluating geo-temporal reasoning.

forcing precise metrics for time and coordinates, schema consistency (e.g., month/season vs. hemisphere; climate plausibility), and calibration analysis. Our evaluation reveals low temporal accuracy and frequent geo-temporal inconsistencies, highlighting that reliable joint understanding remains an unsolved scientific challenge.

3. **Comprehensive error analysis and root-cause investigation.** We systematically analyze model failures, identify patterns in temporal and spatial errors, and propose mitigation strategies to guide future research.

Our evaluation framework is verifiable, enforcing precise window accuracy for time and mean/thresholded distance for coordinates; constrained, ensuring schema consistency (e.g., month, season, hemisphere); robust, testing hard/OOD splits without auxiliary metadata; and calibrated, measuring expected calibration error and risk–coverage trade-offs. Applying it to state-of-the-art open- and closed-source VLMs reveals substantial gaps in joint geo-temporal reasoning: even the strongest models achieve 77.59% country accuracy (Gemini 2.5, Flash, Thinking) yet a median geodesic error of 892.54 km, indicating reliance on coarse heuristics. Time-of-day prediction reaches only 33.74% accuracy (GLM 4.1V, 9B, Thinking), highlighting fragile temporal grounding. TIMESPOT complements existing geolocation and remote-sensing benchmarks by offering joint, diversity-aware, trust-focused evaluation of time and place. Our results show that authentic geo-temporal understanding remains an open challenge and emphasize urgent directions for improving robustness, calibration, and real-world applicability.

2 PRELIMINARIES AND RELATED WORK

TIMESPOT targets the geo-temporal inference problem for ground, level scenes, where conspicuous textual cues are minimized and models must fuse weak, distributed signals. For each image, systems must output a structured schema with four temporal attributes (season, month, local time in HH:MM, daylight phase) and five geographic attributes (continent, country, climate zone, environment type, latitude and longitude), shifting evaluation from *retrieval* (cf. Vivanco Cepeda et al., 2023; Astruc et al., 2024; Hou et al., 2024; Huang et al., 2024) to *interpretable, verifiable fields*. Formally, we describe each image $x \in \mathcal{X}$ as containing observable cues (e.g., illumination, vegetation, materials, activity patterns) that map to a schema $y = (y^{\text{temp}}, y^{\text{geo}})$, with temporal fields $y^{\text{temp}} = (s, m, \tau, \phi)$ for season, month, local time, and daylight phase, and geographic fields $y^{\text{geo}} = (C, \kappa, z, e, (\lambda, \varphi))$ for continent, country, climate zone, environment type, and coordinates. The task is to evaluate a fixed vision, , language model via a mapping $f : \mathcal{X} \mapsto \widehat{\mathcal{Y}}$, where $\widehat{\mathcal{Y}}$ denotes predicted schemas in the same structured space. This design probes capabilities that *generic VLMs* (Radford et al., 2021; Li et al., 2022; Kim et al., 2021) and RS benchmarks do not explicitly test, reading solar/sky cues, phenological signatures, material styles, and context patterns to locate scenes in both space and time.

Benchmark / Dataset (year)	Temp	FineGeo	Subtle	HS/OOD	FusionQs	Schema	Calib	Verif	Globality
OpenStreetView–5M Astruc et al. (2024)	–	△	△	✓	–	–	–	✓	✓
Global Streetscapes Hou et al. (2024)	△	△	–	△	–	–	–	✓	✓
CV–Cities Huang et al. (2024)	–	△	–	✓	–	–	–	✓	✓
VIGOR Zhu et al. (2021)	–	△	–	✓	–	–	–	✓	△
CVACT Liu & Li (2019)	–	△	–	✓	–	–	–	✓	△
CVUSA Workman et al. (2015a)	–	△	–	✓	–	–	–	✓	–
University–1652 Zheng et al. (2020)	–	△	–	△	–	–	–	✓	–
GeoText–1652 Chu et al. (2024)	–	△	△	✓	–	△	–	✓	–
GeoCLIP Eval Vivanco Cepeda et al. (2023)	–	△	–	✓	–	–	–	✓	✓
Panoramic Cross–View Xia et al. (2025)	–	△	–	✓	–	–	–	✓	✓
HRVQA Li et al. (2024b)	–	–	–	△	△	–	–	–	✓
VRSBench Li et al. (2024c)	–	–	–	✓	△	–	–	–	✓
GEOBench–VLM Danish et al. (2024)	△	–	–	✓	△	△	–	–	✓
EarthVQA Wang et al. (2024b)	–	–	–	△	△	–	–	–	✓
FIT–RSFG / RSRC Luo et al. (2024)	–	–	–	△	△	△	–	–	✓
RemoteCount Liu et al. (2024)	–	–	–	△	△	–	–	–	✓
SkyEyeGPT / SkyBench Zhan et al. (2025)	–	–	–	△	△	–	–	–	✓
EO–VLM Benchmark Zhang & Wang (2024)	–	–	–	△	△	–	–	–	✓
RS5M Zhang et al. (2024)	–	–	–	–	–	–	–	–	✓
MapBench Hao et al. (2024)	–	–	–	✓	△	–	–	✓	–
MapQA Chang et al. (2022)	–	–	–	△	△	△	–	–	–
MapIQ Srivastava et al. (2025)	–	–	–	△	△	△	–	–	–
CulturalVQA Nayak et al. (2024)	△	△	△	△	△	–	–	–	✓
AMOS Time–lapse Jacobs et al. (2009)	△	–	–	–	–	–	–	–	✓
Transient Attributes Laffont et al. (2014)	△	–	–	–	–	–	–	–	✓
TimeSpot (Ours, 2025)	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison along TIMESPOT axes. *Temp*: Season/Month/Time/Daylight phase. *FineGeo*: Continent/Country/Climate/Environment/Lat–Lon. *Subtle*: non-iconic cue emphasis. *HS/OOD*: hemisphere sanity or hard/OOD splits. *FusionQs*: geo-temporal fusion tasks. *Schema*: structured field outputs. *Calib*: calibration/uncertainty metrics. *Verif*: GPS/OSM-verifiable scoring. *Globality*: multi-continental coverage. Symbols: ✓ = explicit support; △ = partial/limited; — = not present.

General multimodal benchmarks stress breadth of perception and reasoning across images, text, and diagrams, but offer limited leverage for *geo, , temporal* competence. Suites such as MMMU (Yue et al., 2024), M3Exam (Zhang et al., 2023), M4U (Wang et al., 2024a), MM, Vet (Yu et al., 2023), MME (Fu et al., 2023), MMBench (Liu et al., 2025), MMSTAR/Sphinx (Lin et al., 2023), SEED, Bench (Li et al., 2023), and SEED, Bench, 2 (Li et al., 2024a) probe diverse skills, from domain knowledge to chart/map understanding and free, form reasoning, yet they neither require predicting when and where a scene occurs nor enforce cross, field geographic constraints. Vision, language pretraining works (e.g., CLIP (Radford et al., 2021), ViT (Dosovitskiy et al., 2020), ViLT (Kim et al., 2021), BLIP (Li et al., 2022), MGP/SigLIP (Zhai et al., 2023), Long, CLIP (Zhang et al., 2025)) supply strong visual, textual priors but remain agnostic to solar geometry, phenology, or cartographic validity that determine time and place.

In TIMESPOT, we score categorical fields with top-1 accuracy; local time with minute-window accuracy and MAE; and coordinates with great circle distance (mean/median and thresholded ranges), following geolocation practice (Tian et al., 2017; Vo & Hays, 2016; Arandjelovic et al., 2016) for auditable spatial error while extending it to time. To promote *trustworthiness*, we add cross, field constraints and diagnostics, hemisphere, season agreement, daylight, phase compatibility with predicted time and coordinates, and climate plausibility at $(\texttt{lat}, \texttt{lon})$, as seen in large, scale cross, view/geolocation setups (Zhu et al., 2021; Astruc et al., 2024; Hou et al., 2024). We further report calibration (ECE, risk, coverage) to quantify confidence quality in multi, field outputs, addressing failure modes observed in cross, view and place, recognition systems (Lin et al., 2013; Workman et al., 2015b; Hu & Lee, 2020) where predictions can be confident yet inconsistent.

To assess generalization, TIMESPOT provides stratifications by continent, climate zone, and environment type, plus *hemisphere flips* and hard/OOD splits that reduce landmark shortcuts and amplify reliance on physical and ecological cues (Astruc et al., 2024; Hou et al., 2024; Huang et al., 2024; Xia et al., 2025). We also include fusion questions requiring models to integrate spatial and temporal evidence coherently, exposing independence assumptions that often surface in modular VLM decoders (Lin et al., 2013; Hu et al., 2018; Zhu et al., 2021). Compared to generic multimodal evaluations and RS, focused geospatial suites, TIMESPOT fills a specific and currently missing role: a *ground, level*, *joint* geo, , temporal benchmark with structured outputs, internal, consistency checks, calibration analysis, and robustness diagnostics. Empirically, we find that strong open/closed VLMs,

despite impressive progress on broad benchmarks (Vivanco Cepeda et al., 2023) and cross, view geolocation (Lin et al., 2013; Workman et al., 2015b; Hu & Lee, 2020), exhibit low accuracy on temporal fields and frequent geo-temporal inconsistencies, underscoring substantive headroom for physically grounded reasoning. In Appendix B, we provide an extended related work.

3 TIMESPOT BENCHMARK

TIMESPOT is designed to rigorously evaluate *geo-temporal reasoning* capabilities in vision, language models (VLMs) using natural, non, iconic, ground, level imagery. Unlike existing benchmarks that focus on either geolocation or isolated temporal cues, often relying on iconic landmarks or textual artifacts, TIMESPOT emphasizes subtle, low, salience cues such as illumination and shadow geometry, sky state, vegetation phenology, architectural style, and activity patterns. The benchmark comprises 1,455 photographs from 80 countries and requires models to output a structured schema spanning nine fields: four temporal attributes, *season, month, local time (HH:MM), daylight phase*, and five geographic attributes, *continent, country, climate zone* (Köppen, Geiger classes A, E (Peel et al., 2007)), *environment type*, and *latitude, longitude*. By coupling spatial with temporal inference, TIMESPOT stresses physically grounded reasoning under real, world uncertainty, extending geolocation evaluations beyond *retrieval accuracy* and *coordinate error* (Tian et al., 2017; Vo & Hays, 2016; Arandjelovic et al., 2016). Table 2 presents the dataset statistics of the TIMESPOT benchmark.

3.1 BENCHMARK CONSTRUCTION

To ensure scientific reproducibility and minimize annotation bias, we adopt a hybrid labeling methodology combining deterministic programmatic derivation with rigorous human verification. This dual approach ensures both scalability and semantic fidelity, particularly in ambiguous cases where automated rules must be reconciled with visual evidence.

Image collection and curation. We first collect non, iconic photographs from a mix of public internet sources and manually captured snapshots, ensuring broad global coverage. To avoid shortcut exploitation, we deliberately suppressed landmark, dominated and text, rich content during curation. Instead, we prioritized scenes where inference requires synthesizing fine, grained physical evidence, e.g., distinguishing summer vs. winter foliage, estimating solar elevation from shadow vectors, or inferring climate class from vegetation and built environment. Sampling followed principled coverage goals: balancing hemispheres, latitude bands, climate zones, and environment types. This design ensures diversity across geography and seasonality while avoiding overrepresentation of iconic settings that VLMs may have memorized during training.

Programmatic labels derivation. Ground, truth labels were derived deterministically from metadata and geographic priors. Months were obtained from capture dates; seasons were assigned via meteorological definitions with hemisphere correction; daylight phases were computed from solar elevation/azimuth using civil, nautical, and astronomical twilight thresholds; and local time was derived from time zone, adjusted ephemerides. Climate zones were mapped using a global Köppen, Geiger layer (Peel et al., 2007); continent and country boundaries were assigned from coordinates; and latitude, longitude were recorded as continuous targets. This deterministic procedure ensures reproducibility and provides physically consistent labels across the dataset.

Human verification. Labels undergo two, stage verification. Annotators first collect auxiliary geographic context (locality, land, water adjacency, road layout, Street View when available) and cross, check programmatic labels against visual cues (e.g., illumination, shadows, vegetation). A senior annotator resolves edge cases (e.g., sunrise/sunset ambiguity, artificial lighting) and validates labels against ephemeris data and scene consistency.

Schema and normalization. All ground truth labels are stored in a canonical JSON schema. To ensure consistency and enable exact per, field scoring, model outputs are normalized (e.g., "Autumn"→"Fall"; signed decimal degrees for coordinates). Checks include month, season, hemisphere consistency, daylight, phase compatibility, and climate plausibility at $(\texttt{lat}, \texttt{lon})$, providing auditable failure modes beyond retrieval accuracy.

3.2 ANNOTATION AND QUALITY CONTROL

Roles and pipeline. Although many targets are programmatically derived, human roles are explicit and sequential. (i) *Context collection:* trained annotators gather map, level context at the recorded

Axis	Field	Categories (top shown)
Temporal	Season	Summer (400), Fall (399), Spring (335), Winter (321)
	Daylight phase	Afternoon (584), Night (287), Sunset (210), Morning (203), Midday (124), Sunrise (47)
	Month	12 months represented; top: August (163), September (146), July (145), March (131)
	Hemispheric tag	Northern Hemisphere Summer (703), Northern Hemisphere Winter (615), Southern Hemisphere Winter (81), Southern Hemisphere Summer (56)
	Time coverage	Day (1182), Night (273)
	Hour range	Full 0–23; densest 08–18
Geography	Continents	Asia (529), Europe (430), North America (326), South America (170)
	Countries	82 unique; top: USA (196), Russia (97), Japan (67), Italy (65), China (58)
	Climate	Temperate (C) (582), Continental (D) (396), Tropical (A) (274), Arid (B) (180), Polar (E) (23)
	Environment type	Urban (648), Rural (202), Mountain (193), Coastal (181), Suburban (118), Desert (113)
	Lat/Lon span	lat -54.80 to 71.96, lon -173.24 to 170.31
Cues	Primary temporal cues	Sun/Shadows (573), Vegetation (325), Other (289), Snow/Ice (122), Human Clothing (95), Agricultural Activity (51)
	Primary geolocation cues	Architecture (355), Natural Biome (354), Topography (Mountains/Coast) (295), Road Signage/Language (236), Vehicles (156), Other (58)

Table 2: Dataset statistics. We report coverage across temporal, geographic, cue, and environment axes.

coordinates and note evidence relevant to temporal cues (e.g., horizon occlusion, canyoning). (ii) *Verification:* annotators validate programmatic labels against the image and context under written guidelines that formalize season, , hemisphere rules, solar, elevation thresholds with tolerated twilight windows, and the environment rubric; low, confidence items are flagged. (iii) *Expert timing review:* a senior annotator adjudicates flagged items, cross, checking ephemerides with observed shadow vectors and horizon occlusion; deterministic temporal labels remain tied to metadata/ephemerides unless a metadata error is documented, whereas non, deterministic fields (e.g., environment type) can be updated with rationale. (iv) *Constraint audit:* authors and annotators apply machine, enforced integrity constraints (month, , season, , hemisphere, phase, , time alignment within tolerance, climate plausibility at coordinates) and finalize a versioned JSON record containing ground truth, validation outcomes, human notes, adjudications, and the active constraint set.

Evaluation affordances. The schema supports *interpretable* outputs and *verifiable* metrics: top–1 accuracy for categorical fields (e.g., country, season); minute–window accuracy and mean absolute error (MAE) for local time; and great–circle distance (mean/median and thresholded ranges) for coordinates (Tian et al., 2017; Vo & Hays, 2016; Arandjelovic et al., 2016), aligning with spatial error practices in geolocation while extending them to time. We additionally report calibration (ECE, risk–coverage) and provide robustness stratifications—hemisphere flips; hard/OOD splits by continent, climate, and environment—to surface shortcutting behaviors that can arise in cross–view and global datasets (Zhu et al., 2021; Astruc et al., 2024; Hou et al., 2024; Huang et al., 2024).
By construction, TIMESPOT thus isolates capabilities that generic VLM benchmarks (e.g., CLIP (Radford et al., 2021), ViLT (Kim et al., 2021), BLIP (Li et al., 2022)) and RS–focused geospatial suites (e.g., GEOBench–VLM (Danish et al., 2024)) do not explicitly test—namely, coherent, calibrated prediction of *when* and *where* from subtle ground–level cues.

4 EXPERIMENTS AND EVALUATION

We benchmark a diverse set of vision–language models (VLMs) to provide a comprehensive assessment of geo-temporal reasoning capabilities. The models are grouped into four families: (i) *proprietary* VLMs (GPT-4o/mini, o3/o4-mini, Gemini-2/2.5-Flash, Claude 3.5 Haiku, Mistral Medium) (OpenAI, 2024; 2025; Team, 2025a; Anthropic, 2024); (ii) *open-source* VLMs spanning compact and large scales (InternVL3, Qwen2.5-VL, Llama-3.2-Vision, Gemma-3, GLM-4.5V) (Zhu et al., 2025; Bai et al., 2025; Grattafiori et al., 2024; Gemma Team, Google DeepMind, 2025; Team et al., 2025); (iii) *reasoning-augmented* variants exposing native "thinking" tokens but returning final structured answers (o3/o4-mini, Gemini-2/2.5-Flash-Thinking, GLM-4.1V-Thinking, Kimi-VL-A3B-Thinking, Step-3) (OpenAI, 2025; Team, 2025a; Team et al., 2025; Team, 2025b; StepFun, 2025). For open-source models we report two size buckets: *small* ($\leq$11B) and *large* ($>$11B). We evaluate with categorical accuracy (continent/country/climate/environment), local-time accuracy within $\leq$1 hour

Model	Geo-location Understanding							Temporal Understanding				
	Cnt.	**Cou.**	**Clim.**	**Env.**	**Lat.°**	**Long.°**	**Dist.(km)**	**Season**	**Month**	**Time**	**Time**	**DLP**
	Ac.(↑)	Ac.(↑)	Ac.(↑)	Ac.(↑)	MAE (↓)	MAE (↓)	MD (↓)	Ac.(↑)	Ac.(↑)	Ac.(↑)	MAE (↓)	Ac.(↑)
Proprietary Models												
GPT-4o-mini	82.68	49.14	50.93	57.87	12.40	24.70	2827.07	47.08	22.34	30.32	3:54	31.55
GPT-5-mini	83.62	68.27	**72.47**	60.01	4.72	15.64	1389.79	**58.43**	**34.27**	21.55	4:10	**44.60**
Gemini-2.0-Flash	89.07	76.91	68.52	60.96	3.32	11.23	994.30	49.76	22.89	27.35	4:22	30.24
Gemini-2.5-Flash	**90.51**	**77.25**	71.34	**64.32**	**3.05**	**10.38**	**917.61**	50.92	23.91	25.15	3:56	41.92
Claude 3.5 Haiku	77.25	55.53	61.86	55.74	6.85	27.51	2269.86	44.12	19.04	23.09	4:14	30.93
Mistral Medium 3.1	75.88	52.85	66.67	61.72	6.37	22.62	2045.61	36.84	15.26	**30.73**	**3:36**	36.01
Open-Source Models												
InternVL3.5-1B	43.02	14.15	32.50	53.54	44.68	4378.92	7700.42	30.65	3.78	7.77	11:45	35.80
InternVL3.5-2B	60.00	29.41	51.82	57.80	13.11	43.71	3959.29	36.29	5.70	27.80	4:30	24.05
Qwen-VL2.5-3B-Instruct	22.40	13.47	18.83	44.53	16.18	130.98	8231.18	27.49	9.96	22.06	4:34	8.52
InternVL3.5-4B	60.79	30.12	57.77	56.74	15.34	44.15	4236.77	37.55	12.03	29.33	4:10	41.61
Qwen-VL2.5-7B-Instruct	**85.70**	**73.96**	70.86	**75.21**	32.94	21.46	4719.95	**61.46**	**44.96**	25.68	3:47	**64.09**
Llama-3.2-11B-Vision-Instruct	74.22	55.73	57.12	57.61	5.85	26.57	2072.35	43.50	16.68	25.74	4:18	43.57
Gemma-3-27B-it	79.59	54.02	60.41	53.12	6.83	23.58	2063.93	44.81	17.11	26.34	4:28	30.86
Qwen-VL2.5-32B-Instruct	78.56	57.11	62.95	60.82	6.27	24.02	2010.12	44.81	17.86	**31.10**	**3:44**	44.54
Internvl3-78b	77.46	53.26	**71.61**	61.37	7.42	23.63	2180.29	45.91	16.43	29.64	4:07	34.91
Qwen-VL2.5-72B-Instruct	77.94	58.28	65.15	58.14	5.11	19.33	1711.42	44.47	18.28	28.71	4:00	36.84
Llama-3.2-90B-Vision-Instruct	78.08	53.54	63.85	59.04	7.05	26.79	2284.85	45.15	19.72	23.33	4:29	33.88
GLM-4.5V-106B-MoE	85.32	69.68	62.09	62.51	**4.23**	**14.09**	**1280.87**	57.55	36.04	30.51	4:09	42.45
Reasoning Models												
o4-mini	82.39	71.82	**73.06**	66.64	4.85	15.39	1359.96	**65.81**	**48.20**	23.91	4:04	**51.79**
Gemini-2-Flash-Thinking	88.66	76.22	66.73	59.93	3.44	11.70	1024.14	49.28	22.68	27.49	4:22	29.76
Gemini-2.5-Flash-Thinking	**90.31**	**77.59**	70.86	64.47	**3.04**	**9.85**	**892.54**	51.13	24.26	22.19	4:03	36.56
Kimi-VL-A3B-Thinking-2506	58.90	40.69	54.84	59.31	16.00	39.83	4034.15	39.72	12.65	32.23	4:18	25.70
GLM-4.1V-9B-Thinking	84.44	68.34	70.19	**68.54**	4.34	23.01	1788.77	58.02	38.88	**33.74**	**3:58**	47.76

Table 3: Performance of VLMs on **TIMESPOT** by questions. We bold and underline the best score within each model category. **Cnt.** → Continent, **Cou.** → Country, **Clim.** → Climate Zone; **Env.** → Environment Type, **Lat.°** → Latitude in degree, **Long.°** → Longitude in degree, **Dist.(km)** (MD) → mean distance from actual location in kilometers, **DLP** → Day-light phase. **Time** (Ac.) denotes accuracy, if the model predicted the time accurately within 1 hour window. **Time** (MAE) shows mean error in HH:MM format. Ac. denotes accuracy.

and MAE (minutes), coordinate MAE (lat/long, degrees) and mean geodesic distance MD (km), plus cross-field consistency diagnostics; full definitions appear in Appendix C.

5 RESULTS AND ANALYSIS

Table 3 reports per-field performance grouped by model family (best results within each family are highlighted in **bold** and underlined). Proprietary models lead on *country* and *MD*, but all families show large temporal errors and frequent geo–temporal inconsistencies. Large open-source models narrow spatial gaps yet remain weaker on *time* and *daylight phase*. Reasoning variants improve *month*/*season* modestly, indicating better use of low-salience cues, while *coordinate distance* remains challenging under open-ended generation. For ease of comparison in our results table, we use color coding: proprietary , open-source up to 11B , larger than 11B , proprietary reasoning , open-source reasoning .

Overall performance. Results presented in Table 3 suggests that proprietary models lead on spatial attributes and metric localization. Among non–*Thinking* models, **Gemini–2.5–Flash** performs best, achieving continent **90.51**%, country **77.25**%, climate **71.34**%, environment **64.32**%, and the lowest geodesic error (MD **917.61** km; latitude MAE **3.05**°, longitude MAE **10.38**°). Its *Flash–Thinking* variant further improves coordinate precision (latitude MAE **3.04**°, longitude MAE **9.85**°, MD **892.54** km) while maintaining high country accuracy (**77.59**%). Open–source models show more heterogeneous performance. **GLM–4.5V–106B–MoE** reaches competitive country accuracy (**69.68**%) with MD **1280.87** km, while **Qwen–VL2.5–7B–Instruct** shows strong categorical geography (continent **85.70**%, country **73.96**%, best environment **75.21**%) but struggles with precise coordinates (latitude MAE **32.94**°, MD **4719.95** km), highlighting a gap between place classification and metric localization. Temporal understanding remains difficult. Time-of-day accuracy remains low across all models (typically 22–34%), with MAE around four hours (3:36–4:30). Specialization varies: **o4–mini** leads calendar categoricals (season **65.81**%, month **48.20**%), **GLM–4.1V–9B–Thinking** achieves the top time-of-day accuracy (**33.74**%), and **Qwen–VL2.5–7B–Instruct** is strongest on daylight phase (**64.09**%). This suggests that while models can usually identify the correct daylight band, they struggle to infer precise local time from subtle cues like solar elevation, shadows, sky luminance, and artificial lighting.

Model	Phase & Time>1h (%)	Month & Season (%)	Season & Month (%)	Country & MD>200 km (%)	Country & MD<200 km (%)	Continent & Country (%)	MD>1000 km (%)
Gpt5-Mini	15.95	0.89	25.02	16.98	2.54	17.59	17.25
intern_vl3_78B	11.82	0.62	30.10	27.42	3.85	29.00	37.73
QwenVL-3B	0.21	0.82	18.35	12.78	0.00	8.93	95.19

Table 4: Consistency-violation and diagnostic rates across models on TIMESPOT. Lower is better.

Cross-Model Observation. Across models, high continent and country accuracy often coexists with large geodesic errors, indicating reliance on coarse cues (e.g., script, architecture, vegetation) rather than precise latitude and longitude. Models marketed as Thinking (e.g., Gemini–2.5–Flash–Thinking, o4–mini) show better consistency in coordinates and calendar attributes, suggesting improved use of low-salience cues such as vegetation phenology and illumination patterns. Among open-source models, larger systems improve classification and reduce distance errors but still lag behind proprietary models, likely due to pretraining diversity and geo-temporal supervision. Despite decent classification, precise localization remains challenging, reflected in kilometer-scale errors. Calendar fields (season, month) and daylight-phase predictions are moderately reliable, but fine-grained time-of-day reasoning is poor, indicating underuse of solar geometry, shadow, and lighting cues. These limitations affect downstream tasks like autonomous navigation, scene understanding, and context-aware reasoning, where fine-grained spatial and temporal perception is critical.

6 ERROR ANALYSIS

Qualitative Error Analysis We summarize the main empirical regularities on TIMESPOT (1,455 images); full diagnostics appear in Appendix A. **(i) Regional vs. national)** Continent accuracy is consistently high while country accuracy drops markedly, indicating under-use of micro-geo cues; see Table 3. **(ii) Temporal precision)** Minute-level time is weak despite moderate MAE, consistent with round-time anchoring; this holds across prompt variants (Table 3). **(iii) Spatial tail risk)** Models with similar mean MD can differ sharply in MD>1000 km mass (Table 3), which governs unusably large errors. **(iv) Cross-field incoherence)** Phase–time and time–longitude mismatches persist, revealing missing soft constraints across $\{$*phase, time, latitude*$\}$. **(v) Field-wise cues)** Phase $>$ time; climate/environment sit mid-band with Temperate/Urban defaults; month/season drift reflects hemispheric priors Constraint-aware joint decoding, micro-geo supervision with hard negatives, anti-anchor time regression, and hemisphere/biome-aware temporal targets are the most promising levers (details in Appendix A).

Quantitative Analysis Figure 3 presents representative successes and failures in geo–temporal reasoning on TIMESPOT. When salient physical cues are present—e.g., clean solar geometry in arid landscapes—the model aligns closely with ground truth (desert sunset: $|\Delta t| = 0{:}15$, MD $= 3.5$ km), indicating effective use of shadow direction and sky color for time/place inference. In contrast, scenes with low or ambiguous illumination degrade temporal accuracy: at *night*, predictions collapse to popular evening anchors (e.g., 20:30), yielding large time errors despite minor spatial drift; around *dawn/dusk*, symmetric chromatic cues trigger Sunrise↔Sunset flips (e.g., Uruguay→Argentina, $|\Delta t| \approx 11$ h), exposing weak phase disambiguation beyond hue.

Urban street views reveal a second pattern. Occlusions and canyon geometry distort shadow cues and compress apparent sun elevation, pushing hours later than reality (e.g., Turkey morning predicted as afternoon; $|\Delta t| \approx 4$ h). Spatially, the model often gets the *continent* right but swaps the *country* to a regional neighbor (e.g., Bangladesh→India), suggesting reliance on broad stylistic features over micro-geocues such as signage typography, license plates, and utility hardware. We also observe environment/climate drift under low light (Poland Night predicted as Morning; Continental $\rightarrow$ Temperate), and underuse of coastal topology, where visible sea–horizon boundaries are ignored, producing large MD when predictions move inland.

Overall, these examples highlight a divide between scenes with unambiguous physical constraints (clear shadows, distinctive biomes) and those requiring integration of subtle cues (phase at twilight, fine-grained regional markers, shoreline/elevation geometry). Improving robustness likely requires (i) explicit *solar-geometry* modules that couple latitude with day-of-year, (ii) *phase-aware* temporal heads to resist round-time collapse at night, and (iii) stronger *geo-linguistic/topographic* priors (script, plate formats, coastline/elevation fingerprints) to reduce neighbor substitutions and climate drift.

Figure 3: Qualitative results on TIMESPOT (Gemini-2.5-Flash). Each panel pairs the image with ground truth and model outputs—country, daylight phase, and local time—along with MD and $|\Delta t|$. Clear solar geometry yields accurate estimates (desert sunset). Night scenes and urban canyons cause round time anchoring and phase drift. Neighbor–country substitutions and limited use of coastline, topography, or micro-cues produce large MD.

7 CONCLUSION

Geo-temporal reasoning remains a major challenge for vision–language models in unconstrained, real-world images. TIMESPOT introduces 1,455 images across 80 countries with structured temporal (season, month, time, daylight) and geographic (continent, country, climate, environment, coordinates) fields. Evaluation reveals persistent weaknesses: even top models achieve only 77.6% country accuracy, 33.7% time-of-day accuracy, and median geodesic errors above 890 km, while weaker models fall below 50% country accuracy with extreme errors exceeding 4,700 km. Temporal predictions often conflict with solar and hemispheric constraints, and spatial predictions rely heavily on coarse priors, causing systematic neighboring-country swaps and climate misclassifications. Low-light, urban-canyon, and twilight scenes amplify failures, showing underuse of shadows, illumination gradients, and micro-geographic cues. TIMESPOT highlights the fragility of current VLMs and underscores that scaling or instruction-tuning alone is insufficient. Future work should target joint, constraint-aware reasoning, explicit solar modeling, and micro-geography supervision to improve real-world geo-temporal understanding.

8 ETHICS STATEMENT

TIMESPOT was created to study geo–temporal understanding in everyday, publicly observable scenes and to benchmark vision–language models on structured, verifiable outputs. Throughout curation we prioritized privacy and respectful representation. Images were sourced under licenses permitting research use or captured by the authors; items with unclear rights were excluded, and per-image license and attribution are recorded in the release. To mitigate privacy risks, we remove EXIF and other embedded metadata, blur or mask personally identifying details such as faces, license plates, and house numbers, and exclude scenes where sensitive content dominates. Geographic labels are reported at city or regional granularity, and exact dwelling locations are never included. Annotation guidelines prohibit demographic inference or stereotyping, and we audited geographic balance across continents, countries, climate zones, and environments to reduce bias. Annotators received training, quality checks, and fair compensation consistent with local norms, and could decline any image. We recognize dual-use risks in location inference; accordingly, the dataset license limits use to non-commercial research and prohibits attempts to identify individuals or private property, and we document salient failure modes to discourage deployment in safety-critical settings. To

our knowledge, the work does not constitute human-subjects research under institutional policy; if required by venue or institution, we will obtain approval prior to public release. A takedown procedure and contact channel are provided so that content owners or affected parties can request removal.

9 Reproducibility Statement

We aim for end-to-end reproducibility of all results reported in the paper. The release will include the full TimeSpot image set, annotations for spatial and temporal fields, and official train/validation/test splits with file hashes and a verification script. The code repository contains data loaders, unified evaluation utilities, prompt templates, and scripts to regenerate every table and figure directly from model outputs; a single configuration file specifies paths, random seeds, and metric options. Categorical fields are evaluated by accuracy; local time is assessed both by accuracy within a one-hour tolerance and by mean absolute error in minutes; geolocation uses mean geodesic distance in kilometers computed with the Haversine formula; and consistency diagnostics (e.g., daylight-phase vs. predicted time, hemisphere/month plausibility) are recomputable from raw predictions. For open-source models, we specify checkpoints, image preprocessing, resolution, batch size, and all inference hyperparameters; for API models, we fix temperature and decoding settings, cache responses, and, where nondeterminism may persist, report mean$\pm$standard deviation over repeated runs. The environment is documented via `requirements.txt` and a Dockerfile, with OS, Python, and library versions pinned; experiments reproduce on CPU with longer runtimes and on commodity GPUs with identical metrics. Upon acceptance, we will release dataset, code, prompts, configuration files, logs/CSVs used to produce the camera-ready results, and an executable reproduction script that regenerates the paper's tables and figures in a single command.

References

Anthropic. Model Card Addendum: Claude 3.5 Haiku and Upgraded Claude 3.5 Sonnet, October 2024. URL https://assets.anthropic.com/m/1cd9d098ac3e6467/original/Claude-3-Model-Card-October-Addendum.pdf. Accessed: 2025-09-11.

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5297–5307, 2016.

Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, et al. Openstreetview-5m: The many roads to global visual geolocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21967–21977, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Yakoub Bazi, Laila Bashmal, Mohamad Mahmoud Al Rahhal, Riccardo Ricci, and Farid Melgani. Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery. *Remote Sensing*, 16(9):1477, 2024.

Zhongang Cai, Yubo Wang, Qingping Sun, Ruisi Wang, Chenyang Gu, Wanqi Yin, Zhiqian Lin, Zhitao Yang, Chen Wei, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Jiaqi Li, Xiangyu Fan, Hanming Deng, Lewei Lu, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Has gpt-5 achieved spatial intelligence? an empirical study, 2025. URL https://arxiv.org/abs/2508.13142.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545*, 2022.

Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching. In *European Conference on Computer Vision*, pp. 213–231. Springer, 2024.

Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. *arXiv preprint arXiv:2411.19325*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394, 2023.

Gemma Team, Google DeepMind. Gemma 3 technical report, 2025. URL `https://storage.googleapis.com/deepmind-media/gemma/Gemma3Report.pdf`. Accessed: 2025-09-11.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song,

Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Xiaoshuai Hao, Mengchuan Wei, Yifan Yang, Haimei Zhao, Hui Zhang, Yi Zhou, Qiang Wang, Weiming Li, Lingdong Kong, and Jing Zhang. Is your hd map constructor reliable under sensor corruptions? *Advances in Neural Information Processing Systems*, 37:22441–22482, 2024.

Yujun Hou, Matias Quintana, Maxim Khomiakov, Winston Yap, Jiani Ouyang, Koichi Ito, Zeyu Wang, Tianhong Zhao, and Filip Biljecki. Global streetscapes—a comprehensive dataset of 10 million street-level images across 688 cities for urban science and analytics. *ISPRS Journal of Photogrammetry and Remote Sensing*, 215:216–238, 2024.

Sixing Hu and Gim Hee Lee. Image-based geo-localization using satellite imagery. *International Journal of Computer Vision*, 128(5):1205–1219, 2020.

Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7258–7267, 2018.

Gaoshuang Huang, Yang Zhou, Luying Zhao, and Wenjian Gan. Cv-cities: Advancing cross-view geo-localization in global cities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

Nathan Jacobs, Walker Burgin, Nick Fridrich, Austin Abrams, Kylia Miskell, Bobby H. Braswell, Andrew D. Richardson, and Robert Pless. The global network of outdoor webcams: Properties and applications. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL)*, pp. 111–120, November 2009. doi: 10.1145/1653771.1653789. Acceptance rate: 20.9%.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pp. 5583–5594. PMLR, 2021.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27831–27840, 2024.

Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on graphics (TOG)*, 33 (4):1–11, 2014.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Kun Li, George Vosselman, and Michael Ying Yang. Hrvqa: A visual question answering benchmark for high-resolution aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 214: 65–81, 2024b.

Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *arXiv preprint arXiv:2406.12384*, 2024c.

Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 891–898, 2013.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5624–5633, 2019.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pp. 216–233. Springer, 2025.

Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, et al. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint arXiv:2406.10100*, 2024.

Simon Lynen, Bernhard Zeisl, Dror Aiger, Michael Bosse, Joel Hesch, Marc Pollefeys, Roland Siegwart, and Torsten Sattler. Large-scale, real-time visual–inertial localization revisited. *The International Journal of Robotics Research*, 39(9):1061–1084, 2020.

Piotr Mirowski, Matt Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Andrew Zisserman, Raia Hadsell, et al. Learning to navigate in cities without a map. *Advances in neural information processing systems*, 31, 2018.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*, 2024.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

OpenAI. Openai o3 and o4-mini system card, April 2025. URL https://openai.com/index/o3-o4-mini-system-card/. System Cards. Accessed: 2025-09-11.

Murray C Peel, Brian L Finlayson, and Thomas A McMahon. Updated world map of the köppen-geiger climate classification. *Hydrology and earth system sciences*, 11(5):1633–1644, 2007.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12716–12725, 2019.

Paul-Edouard Sarlin, Eduard Trulls, Marc Pollefeys, Jan Hosang, and Simon Lynen. Snap: Self-supervised neural maps for visual positioning and semantic understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4064–4072, 2020.

Varun Srivastava, Fan Lei, Srija Mukhopadhyay, Vivek Gupta, and Ross Maciejewski. Mapiq: Benchmarking multimodal large language models for map question answering. *arXiv preprint arXiv:2507.11625*, 2025.

StepFun. Step-3 is large yet affordable: Model-system co-design for cost-effective decoding, 2025. URL https://arxiv.org/abs/2507.19427.

Gemini Team. Gemini: A family of highly capable multimodal models, 2025a. URL https://arxiv.org/abs/2312.11805.

Kimi Team. Kimi-vl technical report, 2025b. URL https://arxiv.org/abs/2504.07491.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL `https://arxiv.org/abs/2507.01006`.

Yicong Tian, Chen Chen, and Mubarak Shah. Cross-view image matching for geo-localization in urban environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3608–3616, 2017.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36:8690–8701, 2023.

Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 494–509. Springer, 2016.

Hongyu Wang, Jiayu Xu, Senwei Xie, Ruiping Wang, Jialin Li, Zhaojie Xie, Bin Zhang, Chuyan Xiong, and Xilin Chen. M4u: Evaluating multilingual understanding and reasoning for large multimodal models. *arXiv preprint arXiv:2405.15638*, 2024a.

Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5481–5489, 2024b.

Xiaolong Wang, Runsen Xu, Zhuofan Cui, Zeyu Wan, and Yu Zhang. Fine-grained cross-view geo-localization using a correlation-aware homography estimator. *Advances in Neural Information Processing Systems*, 36, 2024c.

Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 1–9, 2015a. doi: 10.1109/ICCV.2015.451. Acceptance rate: 30.3%.

Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3961–3969, 2015b.

Panwang Xia, Lei Yu, Yi Wan, Qiong Wu, Peiqi Chen, Liheng Zhong, Yongxiang Yao, Dong Wei, Xinyi Liu, Lixiang Ru, et al. Cross-view geo-localization with panoramic street-view and vhr satellite imagery in decentrality settings. *ISPRS Journal of Photogrammetry and Remote Sensing*, 227:1–11, 2025.

Junyan Ye, Qiyan Luo, Jinhua Yu, Huaping Zhong, Zhimeng Zheng, Conghui He, and Weijia Li. Sg-bev: Satellite-guided bev fusion for cross-view semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27748–27757, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 867–875, 2017.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Yang Zhan, Zhitong Xiong, and Yuan Yuan. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221:64–77, 2025.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pp. 310–325. Springer, 2025.

Chenhui Zhang and Sherrie Wang. Good at captioning, bad at counting: Benchmarking gpt-4v on earth observation data. *arXiv preprint arXiv:2401.17600*, 2024.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505, 2023.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM international conference on Multimedia*, pp. 1395–1403, 2020.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL `https://arxiv.org/abs/2504.10479`.

Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3640–3649, 2021.

Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1162–1171, 2022.

A More on Error Analysis

Metrics and diagnostics. Definitions for categorical accuracies, time $\leq$1 h vs. MAE, MD, and diagnostic rates (Continent✓ & Country$\times$, Country$\times$ & MD$<$200 km, Phase✓ & $|\Delta t|$ $>$120 min, MD$>$1000 km), plus time–MD correlations.

Failure modes. Neighboring-country substitutions and near-miss geography; heavy-tail spatial errors on non-iconic scenes; high-latitude phase/time anomalies; round-time anchoring.

Ablations and remedies. Joint, constraint-aware decoding; micro-geo cue heads (plates, lane/curb paint, shoreline typology, utility poles); hard negatives across borders and coast–lake lookalikes; anti-anchor minutes head with photometric perturbations; hemisphere- and biome-aware temporal targets.

B Extended Related Work

Recent geospatial VLM efforts primarily target aerial understanding (captioning, VQA, detection, change segmentation), providing valuable Earth observation coverage but leaving ground-level geo-temporal reasoning underconstrained. EarthVQA (Wang et al., 2024b), RS-LLaVA (Bazi et al., 2024), RSBench/VRSBench (Li et al., 2024c), GeoChat (Kuckreja et al., 2024), RemoteCLIP (Liu et al., 2024), RS5M/GeoRSCLIP (Zhang et al., 2024), and HRVQA (Li et al., 2024b) assess relational reasoning, captioning, or perception over remote sensing imagery but do not require joint prediction of season, month, time, daylight phase, continent, country, climate, environment, and (`lat`, `lon`). GEOBench-VLM (Danish et al., 2024) aggregates diverse geospatial tasks, including non-optical and segmentation, yet emphasizes remote sensing over ground-level photos where temporal cues are subtle. At ground level, cross-view and place-recognition research has advanced spatial localization via retrieval and matching—from early ground-to-aerial matching (Lin et al., 2013; Workman et al., 2015b; Tian et al., 2017) and NetVLAD (Arandjelovic et al., 2016) to large-scale, globally distributed datasets and methods (CVM/Net (Hu et al., 2018; Hu & Lee, 2020)). Benchmarks such as VIGOR (Zhu et al., 2021), OpenStreetView 5M (Astruc et al., 2024), Global Streetscapes (Hou et al., 2024), CV-Cities (Huang et al., 2024), and panoramic cross-view settings (Xia et al., 2025) expand geographic coverage and stress spatial reasoning under viewpoint and domain gaps, but typically score where (retrieval/coordinate error (Lin et al., 2013; Workman et al., 2015b)) without requiring calibrated predictions of when or validating month/season/hemisphere consistency and daylight plausibility (Hu & Lee, 2020).

C More on Evaluation Metrics

To ensure rigorous and reproducible evaluation, we adopt metrics tailored to both geographic and temporal prediction tasks. All metrics are applied uniformly across models, and malformed outputs (e.g., missing `HH:MM` fields or unsigned coordinates) are considered incorrect, thereby preventing models from gaining undue advantage through partial responses.

Geographic Metrics. For categorical geographic attributes—*continent*, *country*, *climate* (Köppen–Geiger A–E) Peel et al. (2007), and *environment*—we report top-1 accuracy. For continuous localization, we measure the mean absolute error (MAE) in degrees for latitude and longitude,

$$\mathrm{MAE}_{\text{lat}} = \tfrac{1}{N}\sum_{i=1}^{N}\big|\hat{\phi}_i - \phi_i\big|, \quad \mathrm{MAE}_{\text{lon}} = \tfrac{1}{N}\sum_{i=1}^{N}\big|\hat{\lambda}_i - \lambda_i\big|,$$

and the mean great-circle distance (MD, km) using the haversine formula,

$$\mathrm{MD} = \tfrac{1}{N}\sum_{i=1}^{N} R \cdot 2\arctan\Big(\sqrt{a_i}, \sqrt{1-a_i}\Big),\ a_i = \sin^2\tfrac{\Delta\phi_i}{2} + \cos\phi_i \cos\hat{\phi}_i \sin^2\tfrac{\Delta\lambda_i}{2},$$

with Earth radius R=6371 km and (ϕ, λ) denoting (lat, lon). These are standard geolocation metrics (Tian et al., 2017; Vo & Hays, 2016; Arandjelovic et al., 2016).

Temporal Metrics. For *season*, *month*, and *daylight phase* we report top-1 accuracy. For *local time* we report two complementary metrics: (i) window accuracy within ± 1 hour of ground-truth,

$$\mathrm{Acc}_{\pm 1\mathrm{h}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\big(|\hat{t}_i - t_i| \leq 60 \text{ min}\big) ,$$

and (ii) MAE in `HH:MM` after converting absolute minute errors to clock format.

D PROMPT TEMPLATE

TimeSpot : Direct : Prompt for Answering Model

```
You are a geo-spatio-temporal understanding assistant.
From the given image, answer every item below in an exact bullet list (one per line), with the format:
- field : value

Answer all fields regardless of certainty. Use one-word season and month where requested.
Fields and formats:
- season : name of the season in 1 word
- month : name of month in 1 word
- time_of_day : HH:MM (24-hour local time)
- daylight_phase : can choose one of: Sunrise, Morning, Midday, Afternoon, Sunset, Night
- continent : value
- country : value
- climate_zone : can choose one of: Tropical (A), Arid (B), Temperate (C), Continental (D), Polar (E)
- environment_type : can choose one of: Urban, Suburban, Rural, Coastal, Mountain, Desert
- coordinates_latitude : +/-DD.DDDDD (decimal degrees). DO NOT include N/S. Use +/-.
- coordinates_longitude : +/-DDD.DDDDD (decimal degrees). DO NOT include E/W. Use +/-.

Provide only the bullet list lines, nothing else.
```

TimeSpot : Prompt for Judging Model

```
You are a strict evaluator.
You will receive the ground truth and a model's answer (both in the same bullet-list format).
Compare each field. Treat abbreviations and long forms as equivalent (for example: USA == United States). Treat seasonal synonyms as
equivalent (for example: Fall == Autumn). DO NOT include E/W. Use +/- in coordinates.
Return a JSON object where each field maps to a nested object:
{
  "field_name": {
    "ground_truth": "<ground truth string>",
    "model_ans": "<model answer string>",
    "evaluation": 1 or 0
  },
  ...
}
Include all fields even if the model answer is malformed. Output only valid JSON.

Ground Truth:
{json_ground_truth}

Model Answer:
{model_response_text}
```

Figure 4: Prompts used for evaluation.

E DETAILED ANALYSIS

E.1 DAYLIGHT PHASE ANALYSIS (SEE TABLE 5)

P1: Best and worst per phase. From Table 5, Sunrise peaks with gemma 3 27B at 61.7%, while Intern_vl3_4B at 0.0% and qwen 2.5 32B instruct at 6.4% are lowest. Morning is led by gemini_flash_2.5_thinking at 65.0%, with kimi_vl_a3b_thinking at 7.4% the minimum. Midday strongly favors glm 4.5vs at 87.9% and kimi_vl_a3b_thinking at 74.2%, whereas gemini_flash_2.5_thinking drops to 3.2%. Afternoon tops with qwen 2.5 32B instruct at 70.6%, while kimi_vl_a3b_thinking at 14.4% and Llama 3.2 90B Vision Instruct at 18.0% trail. Sunset is highest for qwen 2.5 32B instruct at 55.7%. Night is uniformly difficult, with a 34.8% ceiling (glm 4.5vs and gpt5_mini) and a 20.9% floor (qwen 2.5 32B instruct).

P2: Phase difficulty profile. As summarized in Table 5, Night is the hardest regime for all models (all at or below mid 30s). Sunrise is volatile (0–61.7%), reflecting sensitivity to low angle illumination and color temperature. Midday splits systems into very strong versus very weak, consistent with reliance on explicit shadow geometry versus chroma heuristics.

P3: Model specialization. Table 5 indicates glm 4.5vs is solar geometry competent (dominant at Midday, competitive at Sunset and Night). qwen 2.5 32B instruct is afternoon centric (70.6%) and sunset strong (55.7%) but weak at Sunrise and Night, suggesting asymmetric priors. gemini_flash_2.5_thinking is morning specialized yet midday fragile.

P4: Balance versus peaks. Per Table 5, gpt5_mini shows the most balanced profile (roughly high 40s to low 50s across daylight phases, mid 30s at Night), avoiding catastrophic collapses. Several peers trade balance for sharp peaks (e.g., glm 4.5vs at Midday) with weaker neighboring phases.

P5: Morning versus afternoon asymmetry. Strong diurnal asymmetry in Table 5: gemini_flash_2.5_thinking at 65.0% Morning versus 34.8% Afternoon, and qwen 2.5 32B instruct at 20.2% Morning versus 70.6% Afternoon. This aligns with hemisphere or time of year priors and color temperature shortcuts rather than robust sun azimuth reasoning.

P6: Sunrise versus Sunset confusability. As seen in Table 5, qwen 2.5 32B instruct and glm 4.5vs are better at Sunset than Sunrise, consistent with warm toned spectra being easier than pre dawn blues. gemma 3 27B is a counter example with strong Sunrise performance, implying different colorimetric priors.

P7: Scale is not destiny. Table 5 shows Llama 3.2 90B Vision Instruct does not dominate; scores oscillate (e.g., Midday 54.8% vs Afternoon 18.0%). Instruction data and objectives appear more critical than parameter count.

P8: Two archetypes. We observe in Table 5 geometry driven models (sharp at Midday and often solid at Sunset) versus color cue models (better at Morning and Sunset but brittle at Midday and Night). glm 4.5vs and kimi_vl_a3b_thinking align with the geometry driven group; gemini_flash_2.5_thinking resembles the color cue group.

P9: Nighttime ceiling. The mid 30s ceiling at Night (Table 5) suggests under use of urban lighting, sky luminance gradients, and activity cues; absent shadows, models default to weak heuristics.

P10: Informative outliers. The 0.0% Sunrise of Intern_vl3_4B and 3.2% Midday of gemini_flash_2.5_thinking (Table 5) indicate brittle mode collapses likely due to data or finetune skew and prompt sensitivity, rather than uniformly low competence.

E.2 BROADER IMPLICATIONS AND ROOT CAUSES (ANCHORED TO TABLE 5)

Training corpora are likely day biased and lack explicit supervision for clock time and daylight phase, encouraging color temperature shortcuts that fail under distribution shift. Without objectives

Model	Sunrise	Morning	Midday	Afternoon	Sunset	Night
Intern_vl3_4B	0.00	12.81	27.42	65.69	37.62	28.92
Llama-3.2-90B-Vision-Instruct	23.40	42.36	54.84	17.98	46.67	28.57
gemini_flash_2.5_thinking	25.53	65.02	3.23	34.76	48.57	27.53
gemma_3_27B	61.70	56.16	33.06	22.26	31.90	23.69
glm_4.5vs	29.79	36.14	87.90	35.28	54.76	34.84
gpt5_mini	38.30	47.03	48.39	45.28	52.86	34.84
intern_vl3_78B	25.53	24.63	39.52	36.99	49.52	26.83
kimi_vl_a3b_thinking	21.28	7.39	74.19	14.38	45.24	27.18
o4_mini	14.89	51.72	31.45	26.71	49.52	28.57
qwen_2.5_32B_instruct	6.38	20.20	12.10	70.55	55.71	20.91

Table 5: Accuracy by daylight phase (DLP) for each model on TIMESPOT. Values are percentage accuracy; blank cells indicate insufficient samples.

tied to sun direction and elevation, models learn correlates (hue, saturation) instead of causes (solar geometry), explaining large morning versus afternoon asymmetries. Nighttime weakness points to limited nocturnal coverage and weak learning of artificial lighting semantics (streetlamp spacing, signage glow, skyglow). Instruction tuning may over emphasize textual plausibility and attenuate geometric priors, yielding inconsistent phase decisions in similar chromatic contexts. Parameter count alone is insufficient; domain targeted finetunes for shadow vectors, sun angle estimation, and radiometric invariance should help. Midday collapse in gemini_flash_2.5_thinking suggests color cast or white balance sensitivity; stronger photometric augmentation and cross camera normalization are warranted. High Midday scores for glm 4.5vs and kimi_vl_a3b_thinking indicate that short shadows and high solar zenith cues are learnable when emphasized and unconfounded. Sunrise remains challenging due to twilight spectra and long penumbras; curricula with pre dawn emphasis and temporal ordering can improve robustness. We recommend adding consistency constraints during training (e.g., jointly predicting phase, time within a one hour tolerance, and inferred sun geometry) to discourage plausible but contradictory outputs, as the discrepancies in Table 5 suggest.

E.3 SEASON ANALYSIS (SEE TABLE 6)

P1: Best and worst per season. From Table 6, Spring is led by gemma 3 27B at 44.48%, followed closely by gpt5 mini at 43.58% and glm 4.5vs at 42.99%. Summer is decisively won by glm 4.5vs at 84.92%, with intern vl3 78B at 79.75% and gemini_flash_2.5_thinking at 80.25% close behind. Autumn is uniformly 0.00% across models, indicating a pathological failure mode. Winter peaks at gpt5 mini 60.31% and llama 3.2 90B Vision Instruct 59.19%, while qwen 2.5 32B instruct trails at 36.45%.

P2: Seasonal difficulty profile. Summer is the easiest season for nearly all models (many above 70%), suggesting strong reliance on high-irradiance cues, saturated foliage, and short shadows. Spring is mid range (roughly 19–45%), indicating moderate confusion with summer and early autumn. Winter is moderately hard but tractable (36–60%) likely due to mixed signals across regions, variable snow presence, and low sun. Autumn is uniquely hard (0% for all), pointing to either systematic misclassification, label scarcity, or overlapping appearance with late summer and early winter.

P3: Model specialization. glm 4.5vs is a summer specialist with state of the art 84.92% and solid winter (55.76%), indicating competence with strong solar geometry and distinct phenology. gemma 3 27B is the spring frontrunner at 44.48% but comparatively weak in summer and winter, suggesting priors tuned to transitional vegetation but less to high-contrast lighting. gpt5 mini is consistently near the top for winter and competitive elsewhere, signaling a balanced seasonal prior.

P4: Balance versus peaks. We see a trade off between balanced models and seasonal specialists. gpt5 mini maintains stability across spring, summer, and winter without catastrophic drops. In contrast, glm 4.5vs exhibits a high summer peak with respectable winter, while several models (e.g., intern vl3 4B, kimi vl a3b thinking) oscillate more sharply between seasons.

P5: Spring versus summer asymmetry. Most models improve substantially from spring to summer (e.g., gemini_flash_2.5_thinking 33.43% to 80.25%, intern vl3 78B 23.58% to 79.75%). This suggests heavy reliance on saturated greens, clear skies, and short, well defined shadows, whereas spring's transitional foliage and variable cloud cover yield weaker, noisier cues.

P6: Winter distinctiveness. Winter results cluster in the 40–60% band, with gpt5 mini and llama 3.2 90B Vision Instruct leading. The spread implies partial use of low sun angles, bare trees, and snow cover when present. Failures likely arise in temperate and maritime climates where winter lacks snow and shares colorimetry with autumn.

P7: The autumn collapse. The 0% column for autumn across all models (Table 6) is a red flag. We suspect systematic confusion toward summer or winter, small or geographically skewed autumn sample sizes, or a mismatch between regional phenology and a four season taxonomy. This uniform failure differentiates seasonal recognition from daylight phase: here the bottleneck appears to be phenology and colorimetry rather than sun angle alone.

P8: Scale and instruction are not enough. Larger models do not guarantee seasonal robustness. llama 3.2 90B Vision Instruct excels in winter but is merely mid tier in spring and summer. Conversely, compact yet well tuned systems (e.g., gpt5 mini) remain competitive, emphasizing the role of targeted supervision and augmentation over sheer parameter count.

P9: Two archetypes. We observe phenology driven models that thrive in summer and hold in winter, and geometry driven models that leverage sun angle and shadow length across seasons. glm 4.5vs resembles the latter in summer, while gpt5 mini looks more balanced. Models with flatter spring performance likely underuse transitional vegetation cues.

P10: Informative outliers. intern vl3 4B's spring and winter gaps and kimi vl a3b thinking's modest spring suggest sensitivity to camera pipeline or color cast. qwen 2.5 32B instruct's weak winter (36.45%) despite decent summer points to priors tied to warm tone imagery and limited invariance to low irradiance scenes.

We hypothesize that pretraining corpora are season imbalanced with an overrepresentation of summer scenes and underrepresentation of transitional phenology, driving the summer peak and spring weakness. A uniform 0% in autumn suggests either severe class scarcity, regional misalignment of a four season taxonomy, or systematic bias in model priors that collapse autumn toward adjacent classes. Without explicit supervision on phenology cues (leaf senescence, canopy density, ground cover) and regional seasonality, models learn correlates like global hue and sky clarity instead of causal, region aware vegetation cycles. Photometric and phenology augmentations that simulate foliage transitions, overcast lighting, and precipitation could reduce reliance on saturated summer colorimetry. Curriculum learning that orders samples by phenological continuity (late summer $\rightarrow$ early autumn $\rightarrow$ mid autumn $\rightarrow$ early winter) may help models track gradual transitions rather than snap to summer or winter modes. Multi head objectives that jointly predict sun elevation, cloud regime, and vegetation state can disentangle lighting from plant phenology, improving spring and autumn reliability. Region conditioned season heads (e.g., temperate, mediterranean, monsoon) could mitigate taxonomy mismatch where autumn appears weak or absent. Finally, consistency losses that couple month, season, and daylight phase—while allowing regional exceptions—should reduce contradictory assignments and raise spring and autumn accuracy without harming summer or winter performance.

E.4 Climate Zone Analysis (see Table 7)

P1: Best and worst per zone. From Table 7, Tropical (A) is led by gemini_flash_2.5_thinking at 86.13%, while kimi vl a3b thinking is lowest at 59.85%. Arid (B) again favors gemini_flash_2.5_thinking at 83.89%, with llama 3.2 90B Vision Instruct the weakest at 58.33%. Temperate (C) is topped by o4 mini at 90.72% and has a broad high plateau above 80% for several models; Intern_vl3_4B trails at 62.89%. Continental (D) is difficult across the board; intern vl3 78B leads at 56.82% and gpt5 mini follows at 55.84%, while kimi vl a3b thinking collapses to 2.02%. Polar (E) peaks at 47.83% for gemini_flash_2.5_thinking; llama 3.2 90B Vision Instruct is the minimum at 4.35%.

Model	Spring	Summer	Autumn	Winter
Intern_vl3_4B	21.86	53.50	0.00	43.93
Llama-3.2-90B-Vision-Instruct	28.66	70.25	0.00	59.19
gemini_flash_2.5_thinking	33.43	80.25	0.00	52.02
gemma_3_27B	44.48	48.75	0.00	43.30
glm_4.5vs	42.99	84.92	0.00	55.76
gpt5_mini	43.58	78.95	0.00	60.31
intern_vl3_78B	23.58	79.75	0.00	40.81
kimi_vl_a3b_thinking	19.10	65.50	0.00	41.43
o4_mini	27.46	73.00	0.00	50.78
qwen_2.5_32B_instruct	36.42	62.50	0.00	36.45

Table 6: Accuracy by season category for each model on TIMESPOT.

P2: Difficulty profile by climate. Zones A, B, and C are comparatively easy, with many models above 70% and a standout at 90.72% in C. Zones D and E are consistently hard, rarely exceeding the mid 50s and often dipping below 25%. This suggests training and inductive biases tuned to lower latitude, higher population, and greener biomes.

P3: Temperate outlier. The 90.72% of o4 mini in Temperate (C) is a clear outlier. Several models cluster tightly between 82–86%, implying that C exhibits abundant, stable cues (deciduous phenology, moderate sky states, familiar built environments) that align with pretraining priors.

P4: Continental strain. Continental (D) shows the sharpest spread: two models near 56% and several below 25%, including a near floor at 2.02%. High seasonality, snow cover intermittency, and strong intra class variance likely erode the value of single cue heuristics such as vegetation greenness or shadow length.

P5: Polar brittleness. Polar (E) remains brittle. Even the best model is below 50%, and multiple systems fall into single digits or teens. Low sun elevations, snow albedo, and sparse vegetation reduce discriminative texture and color signals that these models usually exploit.

P6: A and B advantage. Strong results in Tropical (A) and Arid (B) for gemini_flash_2.5_thinking and intern vl3 78B suggest effective use of sky clarity, aridity signatures, and architectural style. Weak B performance for llama 3.2 90B Vision Instruct hints at limited invariance to high contrast radiometry and dust or haze.

P7: Balance versus specialization. gpt5 mini is broadly balanced, staying competitive in A through D and avoiding catastrophic failures. By contrast, kimi vl a3b thinking is excellent in C (85.40%) but collapses in D (2.02%), a hallmark of specialization without robust transfer.

P8: Scale is not sufficient. Model size does not ensure climate robustness. llama 3.2 90B Vision Instruct ranges from strong in C and Winter like contexts to very weak in E and B. Smaller but better tuned models keep pace or surpass it in difficult zones.

P9: Zone adjacency effects. High accuracy in C paired with low scores in D and E suggests over reliance on mid latitude priors. When cues deviate (long winter, low solar zenith, snow cover), models revert to incorrect temperate like assumptions.

P10: Informative extremes. The Temperate peak and the Continental Polar troughs together imply that current systems overfit to populous, well photographed regions and struggle where illumination physics and surface properties diverge from those priors.

Broader Implications and Root Causes (anchored to Table 7)

We hypothesize that pretraining corpora over represent temperate and tropical urban scenes, under represent continental winters and polar landscapes, and thus bias feature learning toward mid latitude

Model	A	B	C	D	E
Intern_vl3_4B	71.53	71.67	62.89	35.19	43.48
Llama-3.2-90B-Vision-Instruct	60.58	58.33	77.66	44.44	4.35
gemini_flash_2.5_thinking	86.13	83.89	80.24	41.92	47.83
gemma_3_27B	78.10	61.67	83.33	15.40	34.78
glm_4.5vs	76.84	62.57	82.27	23.23	43.48
gpt5_mini	75.91	75.00	82.47	55.84	43.48
intern_vl3_78B	76.28	81.67	78.69	56.82	13.04
kimi_vl_a3b_thinking	59.85	70.00	85.40	2.02	13.04
o4_mini	61.31	60.00	90.72	31.06	8.70
qwen_2.5_32B_instruct	70.44	74.44	85.91	21.46	17.39

Table 7: Accuracy by Köppen–Geiger climate zone (A–E) for each model on TIMESPOT.

colorimetry and textures. In zones D and E, low sun angles, snow and ice high albedo, and reduced vegetation suppress the chromatic and textural cues these models typically exploit, exposing gaps in shadow geometry reasoning and radiometric invariance. Domain shift is amplified by camera pipeline diversity, where auto white balance and HDR interact with snow and low angle illumination, degrading learned color priors. Instruction tuning likely emphasizes textual plausibility over physically grounded cues, weakening transfer to extreme climates. To mitigate, we recommend climate balanced curation, photometric augmentation targeted at low sun and high albedo conditions, and auxiliary heads for sun elevation and sky state to disentangle illumination from land cover. Region aware conditioning (e.g., Köppen class tokens during finetuning) can reduce miscalibration when moving from C to D or E. Finally, multi task consistency losses that link climate class with season, daylight phase, and time can regularize predictions in data poor regimes without sacrificing strengths in A through C.

E.5 ENVIRONMENT ANALYSIS (SEE TABLE 8)

P1: Best and worst per environment. From Table 8, Urban is led by gemini_flash_2.5_thinking at 75.93%, while gemma 3 27B is lowest at 54.17%. Suburban is broadly hard; glm 4.5vs tops at 46.61%, and gemini_flash_2.5_thinking bottoms out at 8.47%. Rural is dominated by gemma 3 27B at 82.67%, with Llama 3.2 90B Vision Instruct lowest at 46.53%. Coastal peaks with qwen 2.5 32B instruct at 61.33% and is weakest for gemma 3 27B and kimi vl a3b thinking at 45.86%. Mountain has a tie for best between Llama 3.2 90B Vision Instruct and gemini_flash_2.5_thinking at 72.54%, while gemma 3 27B is lowest at 38.34%. Desert is generally strong; intern vl3 78B leads at 78.76% and gemma 3 27B trails at 58.41%.

P2: Difficulty profile by environment. Suburban is the most challenging regime for nearly all models (mostly 20–40%), suggesting weak priors for transitional built areas. Rural, Mountain, and Desert are comparatively easier for several models, each with leaders above 70%. Urban is mid to high for many models (55–76%), indicating that dense man-made cues are exploitable. Coastal shows moderate spread (46–61%), pointing to varied coastline appearances and weather states.

P3: Urban versus Suburban asymmetry. Most models perform far better in Urban than Suburban (e.g., gemini_flash_2.5_thinking 75.93% vs 8.47%), implying that high-density architectural style and signage provide stronger anchors than mixed-density neighborhoods where cues dilute.

P4: Rural strengths and variance. Rural excels for gemma 3 27B (82.67%) and is solid for glm 4.5vs and intern vl3 78B (67–70%). We hypothesize these models leverage vegetation structure, field patterns, and road typology; Llama 3.2 90B Vision Instruct's weaker rural score suggests sensitivity to camera pipeline or vegetation colorimetry.

P5: Coastal specificity. qwen 2.5 32B instruct's lead at Coastal (61.33%) indicates effective use of shoreline geometry, horizon–waterline relations, and coastal infrastructure. Models at 46–50% likely underuse sky–sea radiometry and surf texture, or conflate coastal towns with generic urban–suburban scenes.

Model	Urban	Suburban	Rural	Coastal	Mountain	Desert
Intern_vl3_4B	60.65	27.35	54.46	46.96	60.62	77.88
Llama-3.2-90B-Vision-Instruct	54.94	27.97	46.53	55.80	72.54	76.99
gemini_flash_2.5_thinking	75.93	8.47	62.87	49.17	72.54	70.80
gemma_3_27B	54.17	27.12	82.67	45.86	38.34	58.41
glm_4.5vs	64.19	46.61	69.65	57.46	58.03	72.57
gpt5_mini	62.50	37.61	55.94	58.89	65.80	68.14
intern_vl3_78B	66.67	22.03	66.83	54.14	58.55	78.76
kimi_vl_a3b_thinking	68.83	20.34	57.43	45.86	59.07	70.80
o4_mini	63.27	32.20	66.83	50.28	54.40	70.80
qwen_2.5_32B_instruct	63.73	26.27	62.38	61.33	61.66	75.22

Table 8: Accuracy by environment type for each model on TIMESPOT.

P6: Mountain cues. The Mountain tie at 72.54% (Llama 3.2 90B Vision Instruct and gemini_flash_2.5_thinking) suggests successful exploitation of high-relief silhouettes, snow lines, and atmospheric perspective. Lower performers likely mis-handle low sun angles, haze, or mixed alpine–subalpine vegetation bands.

P7: Desert robustness. Desert is robust across models (68–79% leaders). Strong texture monotony, dune/ripples, arid sky states, and sparse vegetation act as distinctive cues; gemma 3 27B's relative drop (58.41%) hints at color-cast sensitivity or limited arid-scene exposure.

P8: Balanced versus specialist profiles. gpt5 mini is consistently mid-to-high across environments (59–66% except Suburban 37.61%), indicating balance. Specialists include gemini_flash_2.5_thinking (very high Urban and Mountain, very low Suburban) and gemma 3 27B (very high Rural, weak Mountain and Coastal).

P9: Intern family behavior. intern vl3 78B shows strong Desert (78.76%) and solid Rural/Urban, but low Suburban (22.03%). The 4B variant is surprisingly competitive in A-like arid settings (Desert 77.88%), yet Suburban remains a consistent weakness across the family.

P10: Informative extremes. The Suburban floor (8.47%) and Desert/Urban peaks (¿75%) reveal an over-reliance on either dense man-made cues or highly distinctive natural biomes, with failures where features blend (Suburban) or where coastal/mountain conditions perturb radiometry.

Broader Implications and Root Causes (anchored to Table 8)

We hypothesize that pretraining corpora over-represent iconic urban cores and visually distinctive biomes (deserts, alpine vistas), while under-representing transitional suburban morphologies; this drives sharp Urban and Desert peaks but Suburban collapses. Without objectives that disentangle structure (built-density gradients, street hierarchy) from appearance (color/contrast), models latch onto high-salience textures and signage and fail when cues mix. Rural success for some models suggests effective use of vegetation and road-layout priors, but sensitivity to color pipelines (white balance, HDR) still lowers robustness across cameras and seasons. Coastal variance reflects under-modeled sky–sea interactions, horizon geometry, and weather/season coupling; adding auxiliary heads for horizon detection, sky state, and water presence should help. Mountain performance benefits from silhouette and relief cues, yet drops when haze, low solar elevation, or snow cover shift radiometry—arguing for photometric augmentation tailored to altitude conditions. Suburban requires structure-aware supervision (e.g., density gradients, land-use mix) and region-aware conditioning to avoid misclassification into Urban or Rural. Finally, multi-task consistency that links environment type with daylight phase, climate zone, and season can regularize predictions in ambiguous scenes without eroding strengths in clearly distinctive environments.

E.6 ASIA COUNTRY ANALYSIS (SEE TABLE 9)

P1: Country leaders and laggards. From Table 9, the strongest single-country results include Singapore at 100.0% (gemini_flash 2.5 thinking), Bangladesh at 100.0% (glm 4.5vs), Japan at 92.54%

(gemini_flash 2.5 thinking), Turkey at 90.91% (gemini_flash 2.5 thinking), and India at 89.66% (gemini_flash 2.5 thinking). The weakest per-country minima include multiple near-zeros, e.g., Intern_vl3 4B in Saudi Arabia and Sri Lanka, kimi vl a3b thinking in Bangladesh, and gemma 3 27B in Kyrgyzstan.

P2: Model wins by country. gemini_flash 2.5 thinking leads in Japan, India, South Korea, Turkey, Philippines (tie with gpt5 mini), Sri Lanka, Mongolia, and Singapore. glm 4.5vs leads in Russia, China, Thailand, Bangladesh, Kyrgyzstan, and Myanmar. gpt5 mini leads in Kazakhstan and ties in Saudi Arabia and the Philippines. Intern_vl3 78B leads in Nepal, while Llama 3.2 90B Vision Instruct leads in Bhutan.

P3: Regional patterning. East and parts of Southeast Asia are broadly high performing across models (Japan, South Korea, China; Singapore, Philippines, Thailand). Central and high-altitude countries (Nepal, Bhutan, Kyrgyzstan, Mongolia) exhibit larger variance and more model-specific collapses. West and Central Asia show mid-band accuracy with occasional peaks (e.g., Saudi Arabia mid 60s for gemini_flash 2.5 thinking and gpt5 mini; Kazakhstan best with gpt5 mini).

P4: Balanced versus spiky profiles. gpt5 mini maintains stable mid-to-high performance across many countries (e.g., Russia 72.06, China 75.86, India 74.14, South Korea 85.00, Kazakhstan 66.67), with few catastrophic lows. In contrast, kimi vl a3b thinking and Intern_vl3 4B show spiky outcomes (good in select settings, near-zero in others), signaling sensitivity to scene distribution or prompt formatting.

P5: Large-scale is not decisive. Llama 3.2 90B Vision Instruct posts strong numbers in Japan and Bhutan but is mid-tier or weak elsewhere (e.g., Kazakhstan 20.00, Bangladesh 22.22), indicating that parameter count alone does not guarantee country-level robustness.

P6: Urbanized and coastal advantage. High-performing countries often coincide with distinctive urban or coastal cues (Japan, Singapore, Philippines). Models appear to leverage stable architectural style, traffic infrastructure, coastline geometry, and signage conventions, which are more uniform and thus easier to learn.

P7: Terrain and inland variance. Mountainous or landlocked countries (Nepal, Bhutan, Kyrgyzstan, Mongolia) provoke larger spreads across models, consistent with higher illumination variability, snowline effects, mixed vegetation bands, and fewer high-salience man-made anchors.

P8: Consistent leaders. gemini_flash 2.5 thinking is a frequent winner in populous and coastal contexts and remains competitive inland; glm 4.5vs excels in continental and mixed settings (Russia, China, Thailand, Bangladesh, Kyrgyzstan, Myanmar). Their complementary wins suggest different cue priors—color and structure in gemini_flash 2.5 thinking, geometry and terrain-texture in glm 4.5vs.

P9: Informative ties and splits. Ties (e.g., Saudi Arabia gemini_flash 2.5 thinking and gpt5 mini at 64.71, Philippines gemini_flash 2.5 thinking and gpt5 mini at 88.89) indicate shared strengths on specific country cues; large intra-country spreads (e.g., Bangladesh ranging from 0.00 to 100.0) expose brittleness to camera pipelines, lighting, or sub-region diversity.

P10: Error geography. Countries with mid-latitude or mixed biomes (e.g., Iran, Turkey, Mongolia) show moderate means but wide per-model dispersion, implying that models overfit to a subset of cues (seasonal colorimetry, skyline geometry) and falter when those cues shift within the country.

Broader Implications and Root Causes (anchored to Table 9)

We hypothesize that pretraining data density and visual homogeneity drive the country-level peaks: countries with abundant, consistent urban–coastal imagery yield higher scores, while terrain-diverse or sparsely documented countries induce variance. Absent explicit objectives for sun geometry, terrain context, and region-aware semantics, models lean on color and texture correlates that drift across sub-regions and seasons within the same country. Instruction-tuning likely amplifies textual

plausibility over physically grounded cues, causing collapses in countries where colorimetry or scene composition deviates from high-data priors. Camera pipeline differences (white balance, HDR) and weather regimes (haze, monsoon cloud fields, snow) further perturb radiometry, explaining the extreme lows. To mitigate, we recommend: (i) region-balanced curation within each country, (ii) photometric and phenology augmentations tied to local climate regimes, (iii) auxiliary heads for horizon detection, sun elevation, and terrain class to disentangle illumination from land cover, and (iv) consistency constraints that couple country with climate zone, season, and daylight phase. These steps should reduce intra-country variance while preserving strengths in highly photographed locales.

Table 9: Country-level accuracy (%) by model for all available countries in Asia.

Country	Intern_vl3_4B	Llama-3.2-90B-Vision-Instruct	gemini_flash 2.5_thinking	gemma_3_27B	glm_4.5vs	gpt5_mini	intern_vl3_78B	kimi_vl_a3b thinking	o4_mini	qwen_2.5_32B instruct
Russia	52.17	52.17	73.91	55.07	78.26	72.06	76.81	40.58	59.42	68.12
Japan	61.19	77.61	92.54	77.61	88.06	86.36	83.58	70.15	76.12	77.61
China	55.17	51.72	81.03	58.62	84.48	75.86	81.03	55.17	68.97	67.24
India	31.03	67.24	89.66	70.69	71.93	74.14	72.41	41.38	72.41	75.86
South Korea	25.00	75.00	87.50	67.50	67.50	85.00	62.50	60.00	65.00	62.50
Turkey	36.36	45.45	90.91	66.67	69.70	69.70	57.58	57.58	45.45	54.55
Philippines	14.81	48.15	88.89	74.07	59.26	88.89	33.33	44.44	74.07	51.85
Iran	5.26	45.00	70.00	35.00	50.00	65.00	50.00	25.00	50.00	35.00
Myanmar	5.56	5.56	22.22	33.33	55.56	27.78	16.67	11.11	16.67	11.11
Bhutan	5.88	64.71	58.82	47.06	52.94	47.06	35.29	29.41	35.29	52.94
Saudi Arabia	0.00	47.06	64.71	29.41	62.50	64.71	41.18	29.41	35.29	29.41
Sri Lanka	0.00	29.41	76.47	23.53	70.59	52.94	17.65	5.88	17.65	29.41
Nepal	18.75	31.25	68.75	56.25	62.50	50.00	75.00	56.25	56.25	56.25
Kazakhstan	0.00	20.00	60.00	40.00	60.00	66.67	6.67	13.33	40.00	13.33
Mongolia	23.08	61.54	76.92	46.15	46.15	69.23	53.85	15.38	53.85	61.54
Thailand	38.46	46.15	53.85	46.15	84.62	53.85	61.54	53.85	46.15	46.15
Singapore	20.00	80.00	100.00	90.00	90.00	90.00	70.00	20.00	90.00	80.00
Bangladesh	11.11	22.22	66.67	55.56	100.00	55.56	11.11	0.00	33.33	11.11
Kyrgyzstan	0.00	22.22	44.44	0.00	44.44	33.33	33.33	11.11	11.11	22.22

E.7 Europe Country Analysis (see Table 11)

E1: Standout highs and lows. Several microstates show ceiling effects with many models at or near 100% (Ireland, Switzerland, Netherlands; also Monaco and Malta for selected models). The sharpest dips include o4 mini in the United Kingdom (13.33%) and numerous zeros for Intern_vl3_4B across small countries (e.g., Denmark, Finland, Luxembourg).

E2: Western Europe profile. France, Italy, Spain, and the United Kingdom are consistently strong for gemini_flash 2.5 thinking (80–97% range) and often for gpt5 mini (60–100%). Llama 3.2 90B Vision Instruct is reliable in the UK (90%) and solid in Italy (60%) but more variable elsewhere (e.g., France 42%).

E3: Central Europe dispersion. Germany is a high-accuracy hub: gemini_flash 93.33%, intern_vl3 78B and qwen 86.67%, gemma 75.56%. Poland and Czechia show broader spread with mid-to-high wins for gemini_flash (84% in Poland) but mixed outcomes for others, indicating cue sensitivity to local appearance diversity.

E4: Nordics and Atlantic edge. Finland, Norway, Iceland, Ireland display very high plateaus for multiple models (often 100% for gemini_flash and gpt5 mini). That said, volatility remains for smaller models (e.g., Intern_vl3_4B at 0% in several Nordic cases) suggesting sample size and winter-lighting sensitivity.

E5: Balkans and microstates. Croatia has a triple 100% (glm 4.5vs, gpt5 mini, o4 mini), while North Macedonia remains uniformly low, and Slovenia centers around a single leader (gemini_flash 83.33%). Andorra, Vatican City, Luxembourg show sparse, highly discrete performance bands tied to tiny N.

E6: Coastal versus inland cues. Countries with strong coastal identity (Spain, Portugal, Norway) tend to deliver high marks for models that exploit horizon geometry, maritime infrastructure, and coastal skylines (gemini_flash, gpt5 mini). Inland nations with mixed terrain (Hungary, Slovakia, Lithuania) show more fragmentation and lower medians.

E7: Recurrent leaders. gemini_flash 2.5 thinking frequently tops major economies (UK 96.67%, Germany 93.33%, Spain 90%) and several small states (Estonia 100%, Monaco 100%). gpt5 mini repeatedly hits 100% in northern and microstate settings (Iceland, Ireland, Netherlands, Switzerland), pointing to robust priors for highly photographed European scenes.

E8: Model stability versus spikes. gpt5 mini offers steady high performance with few catastrophic failures. By contrast, o4 mini and Intern_vl3 variants show spikes and collapses depending on country (e.g., UK 13.33% for o4 mini versus 100% in Croatia), consistent with sensitivity to particular cue distributions.

E9: Alphabet and signage bias. High scores in countries with uniform Latin typography and standardized road signage (Netherlands, Ireland, UK) hint that text and iconography are strong anchors; failures in small-N countries likely arise when those cues are occluded or absent.

E10: Russia and mixed biomes. Russia exhibits mid-to-high accuracy for geometry-leaning systems (glm 75.00%) and balanced models (gpt5 mini 67.86%), but the spread across models suggests challenges with broad biome variety and seasonal changes even within a single country label.

Context and Likely Drivers (anchored to Table 11)

We read these Europe results as a blend of data-density effects and cue availability. Countries with abundant, iconic urban imagery and standardized visual systems (signage, architecture, transit) favor models that internalize stable mid-latitude priors. Small states and low-N splits produce quantization: a few successes yield 100%, a few misses yield 0%. Lighting regimes at high latitude (long twilight, low solar elevation) can either help (predictable skylight geometry) or hurt (photometric drift), depending on a model's radiometric invariance. Over-reliance on textual artifacts likely benefits the

UK, Ireland, Netherlands, but can falter where signage is atypical or occluded. Large parameter counts do not guarantee uniform wins; training mix and augmentation breadth appear more decisive. We recommend per-country balance checks (avoid microstate skew), photometric normalization for high-latitude light, and auxiliary heads that couple terrain, horizon, and sky state with country prediction. Consistency constraints linking country with climate, season, and daylight phase should curb implausible combinations and stabilize performance in heterogeneous interiors (e.g., Balkans, Central/Eastern Europe).

Table 11: Country-level accuracy (%) by model for all available countries in Europe (selected).

Country	Intern_vl3_4B	Llama-3.2-90B-Vision-Instruct	gemini_flash 2.5_thinking	gemma_3_27B	glm_4.5vs	gpt5_mini	intern_vl3_78B	kimi_vl_a3b thinking	o4_mini	qwen_2.5_32B instruct
Italy	24.62	60.00	81.54	58.46	69.23	72.31	46.15	50.77	64.62	56.92
France	40.00	42.00	80.00	42.00	60.00	60.00	46.00	62.00	42.00	50.00
Germany	33.33	57.78	93.33	75.56	75.56	55.56	86.67	53.33	66.67	86.67
United Kingdom	60.00	90.00	96.67	36.67	83.33	93.33	90.00	80.00	13.33	93.33
Russia	53.57	35.71	67.86	35.71	75.00	67.86	71.43	46.43	64.29	53.57
Poland	20.00	48.00	84.00	64.00	64.00	60.00	36.00	24.00	36.00	64.00
Spain	45.00	50.00	90.00	60.00	70.00	65.00	80.00	45.00	40.00	90.00
Vatican City	0.00	40.00	80.00	0.00	40.00	50.00	0.00	30.00	10.00	40.00
Andorra	0.00	50.00	83.33	16.67	83.33	66.67	0.00	0.00	0.00	16.67
Estonia	0.00	33.33	100.00	66.67	16.67	33.33	16.67	33.33	0.00	33.33
Iceland	0.00	66.67	83.33	83.33	83.33	100.00	83.33	66.67	66.67	83.33
Luxembourg	0.00	66.67	50.00	16.67	33.33	50.00	0.00	0.00	0.00	33.33
Malta	0.00	50.00	83.33	66.67	100.00	66.67	33.33	16.67	83.33	83.33
Monaco	0.00	66.67	100.00	100.00	83.33	66.67	16.67	0.00	66.67	66.67
North Macedonia	0.00	0.00	33.33	0.00	33.33	16.67	0.00	0.00	16.67	0.00
Slovenia	0.00	33.33	83.33	33.33	33.33	16.67	16.67	16.67	33.33	16.67
Croatia	0.00	50.00	75.00	50.00	100.00	100.00	50.00	50.00	100.00	50.00
Denmark	0.00	0.00	100.00	50.00	50.00	75.00	25.00	75.00	75.00	75.00
Ireland	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	75.00	100.00
Lithuania	0.00	25.00	50.00	0.00	25.00	25.00	25.00	0.00	0.00	25.00
Slovakia	0.00	25.00	25.00	0.00	50.00	0.00	0.00	0.00	25.00	25.00
Belgium	0.00	33.33	33.33	0.00	33.33	66.67	0.00	33.33	0.00	0.00
Czechia	0.00	33.33	33.33	66.67	66.67	33.33	33.33	33.33	0.00	33.33
Finland	0.00	100.00	100.00	66.67	66.67	100.00	0.00	0.00	33.33	66.67
Greece	33.33	0.00	66.67	66.67	33.33	33.33	0.00	0.00	33.33	0.00
Hungary	0.00	0.00	66.67	33.33	33.33	33.33	33.33	33.33	33.33	0.00
Norway	66.67	33.33	100.00	66.67	66.67	100.00	66.67	33.33	33.33	100.00
Portugal	33.33	33.33	66.67	66.67	33.33	66.67	33.33	66.67	33.33	33.33
Romania	0.00	33.33	100.00	0.00	100.00	66.67	0.00	0.00	33.33	0.00
Sweden	0.00	66.67	66.67	33.33	33.33	100.00	33.33	0.00	33.33	33.33
Netherlands	50.00	100.00	100.00	100.00	50.00	100.00	100.00	50.00	100.00	100.00
Serbia	0.00	50.00	50.00	0.00	50.00	100.00	50.00	0.00	50.00	50.00
Switzerland	100.00	100.00	100.00	100.00	100.00	100.00	100.00	50.00	100.00	100.00
Turkey	0.00	50.00	50.00	50.00	100.00	100.00	50.00	0.00	50.00	50.00

E.8 North America Country Analysis (see Table 13)

N1: Headline results. Table 13 shows the USA as a high-accuracy anchor across nearly all models (85–94%), with gemini_flash 2.5 thinking leading at 93.88%. Canada forms a second tier (54–82%), and Mexico is clearly tougher (best 56.67% with gemini_flash, 51.72% with glm 4.5vs).

N2: Island contrast. Cuba exhibits multiple 100% scores (llama 3.2 90B, gemini_flash, glm 4.5vs, gpt5 mini, o4 mini), suggesting easy-to-key cues or small-N quantization. The Dominican Republic is split: gemini_flash at 90% and gpt5 mini at 70% versus very low numbers for several others, indicating sharp model-specific sensitivity.

N3: Mesoamerican spread. Guatemala highlights divergent priors: gpt5 mini at 90% and glm 4.5vs at 80% outperform gemini_flash at 60%; other models lag at 20–30%. Panama is mid-range overall (most models 50–80%), with intern_vl3 78B dropping to 10%, a notable failure mode.

N4: Puerto Rico band. Puerto Rico sits in a tight mid band (40–60%) across models, implying useful but not decisive coastal or urban cues; no model dominates.

N5: Model leaders by country. gemini_flash 2.5 thinking leads the USA and ties or wins in several island settings; glm 4.5vs is consistently strong in Canada (76%) and also hits 100% in Cuba; gpt5 mini posts standout wins in Guatemala (90%) and remains competitive in the USA and Canada.

N6: Stability versus volatility. gpt5 mini is the steadiest performer (few collapses, many upper-mid scores). By contrast, intern_vl3 78B swings widely (e.g., 87.76% USA vs 0–10% in DR and Panama), and the 4B variant is frequently low (Mexico 3.33%, several zeros elsewhere).

N7: Coastal and signage cues. High marks in the USA and Cuba likely reflect rich coastal–urban signatures, standardized signage, and consistent road furniture. Mexico and parts of Central America may present greater intra-country heterogeneity (lighting regimes, street textures, informal signage), stressing models that lean on uniform text or skyline patterns.

N8: Large-scale isn't decisive. Parameter count does not settle the leaderboard: llama 3.2 90B is excellent on the USA and Cuba but trails or sits mid-pack in Canada and Mexico. Smaller, well-tuned systems (gpt5 mini, glm 4.5vs) keep pace or win outright in several countries.

N9: Informative outliers. Country–model extremes (e.g., o4 mini's strong USA 85.71% but only 20–50% in several smaller countries) suggest reliance on high-data priors; when those cues thin out, performance reverts to chance-like ranges.

N10: Practical takeaway. For deployment across North America, an ensemble that weights gemini_flash in the USA and islands, glm 4.5vs in Canada, and gpt5 mini in Central America would hedge against single-model brittleness.

Interpretation and Likely Drivers (anchored to Table 13)

We read these patterns as a mix of data-density effects and cue stability. The USA's consistently high numbers align with abundant, homogeneous training imagery and standardized visual systems (signage, lane markings, storefront styles). Canada's slightly lower scores likely reflect latitude-driven illumination changes and snow/overcast regimes that perturb color priors; models with better radiometric robustness (glm 4.5vs, gpt5 mini) benefit. Mexico and parts of Central America introduce more varied street morphology and photometric conditions (haze, intense sun, mixed materials), reducing the reliability of text and skyline heuristics. Island peaks (Cuba) may partly be small-N discretization, but also strong coastal silhouettes and maritime infrastructure that certain models exploit. The large swings of intern_vl3 variants point to prompt-format or finetune mismatch under domain shift. Mitigations include country-balanced sampling within each region, photometric augmentation for high-irradiance and overcast extremes, auxiliary heads for horizon/sky state to decouple lighting from semantics, and lightweight ensembling guided by per-country validation to avoid single-prior collapse.

Table 13: Country-level accuracy (%) by model for all available countries in North America.

Country	Intern_vl3_4B	Llama-3.2-90B-Vision-Instruct	gemini_flash 2.5_thinking	gemma_3_27B	glm_4.5vs	gpt5_mini	intern_vl3_78B	kimi_vl_a3b thinking	o4_mini	qwen_2.5_32B instruct
USA	68.37	92.35	93.88	88.27	88.78	89.80	87.76	69.39	85.71	87.24
Canada	16.00	54.00	82.00	40.00	76.00	64.00	54.00	22.00	72.00	64.00
Mexico	3.33	46.67	56.67	43.33	51.72	46.67	33.33	26.67	43.33	26.67
Cuba	30.00	100.00	100.00	90.00	100.00	100.00	80.00	40.00	100.00	70.00
Dominican Republic	20.00	60.00	90.00	20.00	20.00	70.00	0.00	10.00	50.00	10.00
Guatemala	0.00	30.00	60.00	30.00	80.00	90.00	20.00	20.00	20.00	30.00
Panama	0.00	40.00	80.00	70.00	70.00	70.00	10.00	30.00	50.00	60.00
Puerto Rico	0.00	40.00	60.00	50.00	60.00	50.00	30.00	40.00	30.00	50.00

E.9 South America Country Analysis (see Table 15)

Findings. *F1. Coverage snapshot.* Table 15 shows high scores in Chile and Brazil, moderate bands in Argentina, Colombia, Peru, and clear difficulty in Bolivia and Uruguay. *F2. Country leaders.* Chile peaks with gemini_flash 2.5 thinking at 90% (llama 3.2 90B at 80%, o4 mini and gpt5 mini at 75%). Brazil is co-led by gemini_flash 2.5 thinking and glm 4.5vs at 90%. Ecuador is strong for gemini_flash 2.5 thinking (85%) and glm 4.5vs (84.21%). *F3. Central Andean difficulty.* Bolivia is uniformly low with a best of 30% (gpt5 mini), many models at or near 0%. Uruguay is also challenging, topping at 40% (glm 4.5vs and gpt5 mini). *F4. Argentina and Peru mid tier.* Argentina is led by gemini_flash 2.5 thinking at 80%, with glm 4.5vs at 62.5%; Peru tops at 75% (gemini_flash 2.5 thinking), then 65% (glm 4.5vs). *F5. Colombia balanced high.* Colombia favors glm 4.5vs (76.67%), with gemini_flash 2.5 thinking (66.67%) and gpt5 mini (60%) close behind. *F6. Model-wise patterns.* gemini_flash 2.5 thinking is the most frequent winner across countries; glm 4.5vs often places first or second, particularly in Colombia, Brazil, and Ecuador; gpt5 mini rarely leads outright but posts stable upper–mid results and uniquely tops the low-signal Bolivia and ties for Uruguay. *F7. Volatility.* Intern_vl3_4B and kimi vl a3b thinking show repeated near-zeros (Argentina, Ecuador, Brazil, Uruguay), while intern_vl3_78B oscillates from 65% (Chile) to 0–2.5% (Argentina, Uruguay). *F8. Large scale is not decisive.* llama 3.2 90B is excellent in Chile (80%) and solid elsewhere (60% in Brazil, Peru), yet falls to 10% in Uruguay; parameter count alone does not explain the leaderboard.

Interpretation and Causes (anchored to Table 15)

C1. Cue availability. High results in Chile and Brazil align with strong coastal skylines, standardized road furniture, and dense urban signatures, which many models exploit reliably. *C2. Altitude and photometry.* Central Andean scenes (Bolivia, parts of Peru and Ecuador) feature high-elevation lighting, snowlines, and thin-atmosphere radiometry; models tuned to mid-latitude, sea-level appearances underperform. *C3. Heterogeneity within labels.* Mexico-like variance is echoed here: mixed street morphology, informal signage, and diverse building materials in inland South American settings reduce the utility of text and skyline heuristics. *C4. Robustness versus specialization.* gpt5 mini's steadiness suggests better invariance to camera pipelines and illumination shifts; by contrast, models with sharp peaks may rely on brittle chromatic or texture priors. *C5. Training mix effects.* The prevalence of web imagery from major coastal cities likely biases pretraining toward those cues, depressing performance in rural, highland, or low-contrast scenes. *C6. Small-*N *sensitivity.* Several near-zeros and sudden jumps (e.g., Uruguay) indicate discrete outcomes under limited samples; variance should be reported with per-country N and confidence intervals. *C7. What would help.* Climate- and altitude-aware augmentation (low sun, haze, high albedo, desaturated palettes), explicit objectives for horizon and sun geometry, and region-conditioned adapters are likely to reduce failures in Andean interiors. *C8. Deployment implication.* A simple ensemble—weighting gemini_flash 2.5 thinking for Brazil and Chile, glm 4.5vs for Colombia and Ecuador, and gpt5 mini as a stabilizer for Bolivia and Uruguay—would hedge single-model brittleness without heavy overhead.

Table 15: Country-level accuracy (%) by model for all available countries in South America.

Country	Intern_vl3_4B	Llama-3.2-90B-Vision-Instruct	gemini_flash 2.5_thinking	gemma_3_27B	glm_4.5vs	gpt5_mini	intern_vl3_78B	kimi_vl_a3b thinking	o4_mini	qwen_2.5_32B instruct
Argentina	0.0	40.00	80.00	35.0	62.50	55.0	2.50	12.5	22.50	37.50
Colombia	20.0	46.67	66.67	40.0	76.67	60.0	23.33	20.0	43.33	43.33
Chile	20.0	80.00	90.00	55.0	70.00	75.0	65.00	15.0	75.00	65.00
Ecuador	0.0	50.00	85.00	35.0	84.21	70.0	20.00	10.0	30.00	40.00
Peru	25.0	60.00	75.00	55.0	65.00	60.0	45.00	50.0	45.00	55.00
Bolivia	0.0	20.00	20.00	0.0	20.00	30.0	0.00	0.0	0.00	10.00
Brazil	0.0	60.00	90.00	20.0	90.00	60.0	50.00	20.0	50.00	50.00
Uruguay	0.0	10.00	30.00	10.0	40.00	40.0	0.00	0.0	0.00	0.00

F EXAMPLES OF TIMESPOT DATASET

Desert Sunset USA

Picture

Ground Truth

- Season: Summer,
- Month: June,
- Time_of_day: 19:00:00,
- Daylight_phase: Sunset,
- Continent: North America,
- Country: USA,
- Climate_zone: Arid (B),
- Environment_type: Desert,
- Coordinates_latitude: 32.2217,
- Coordinates_longitude: -110.9265

GPT-5-mini

- Season: Summer,
- Month: July,
- Time_of_day: 19:15,
- Daylight_phase: Sunset,
- Continent: North America,
- Country: United States,
- Climate_zone: Arid (B),
- Environment_type: Desert,
- Coordinates_latitude: +32.25000,
- Coordinates_longitude: -111.00000

Intern-VL3.5-2B

- Season: Winter
- Month: December
- Time_of_day: 17:30
- Daylight_phase: Sunset
- Continent: North America
- Country: United States (Arizona)
- Climate_zone: Arid (B)
- Environment_type: Desert
- Coordinates_latitude: +33.421
- Coordinates_longitude: -117.726

Intern-VL2.5-72B

- Season: Summer,
- Month: June,
- Time_of_day: 19:00,
- Daylight_phase: Sunset,
- Continent: North America,
- Country: United States,
- Climate_zone: Arid (B),
- Environment_type: Desert,
- Coordinates_latitude: +33.4484,
- Coordinates_longitude: -111.9385

Figure 5: Example of TIMESPOT dataset — Desert Sunset USA.

Mountain Town Morning In India

Picture

Ground Truth

- Season: Summer
- Month: June
- Time_of_day: 09:00:00
- Daylight_phase: Morning
- Continent: Asia
- Country: India
- Climate_zone: Arid (B)
- Environment_type: Rural
- Coordinates_latitude: 34.15258
- Coordinates_longitude: 77.57705

GPT-5-mini

- Season: Summer
- Month: August
- Time_of_day: 17:30
- Daylight_phase: Afternoon
- Continent: Asia
- Country: India
- Climate_zone: Arid (B)
- Environment_type: Mountain
- Coordinates_latitude: +34.15260
- Coordinates_longitude: +077.57700

Intern-VL3.5-2B

- Season: Spring
- Month: July
- Time_of_day: 12:00
- Daylight_phase: Midday
- Continent: Asia
- Country: Pakistan
- Climate_zone: Arid (B)
- Environment_type: Mountain
- Coordinates_latitude: +34.835
- Coordinates_longitude: +74.679

Intern-VL2.5-72B

- Season: Autumn
- Month: October
- Time_of_day: 16:00
- Daylight_phase: Afternoon
- Continent: Asia
- Country: India
- Climate_zone: Arid (B)
- Environment_type: Mountain
- Coordinates_latitude: +34.19999
- Coordinates_longitude: +77.58333

Figure 6: Example of TIMESPOT dataset — Mountain Town Morning In India.

Urban Morning Street Scene In Turkey

Picture

Ground Truth

- Season: Fall,
- Month: September,
- Time_of_day: 10:27,
- Daylight_phase: Morning,
- Continent: Asia,
- Country: Turkey,
- Climate_zone: Temperate (C),
- Environment_type: Urban,
- Coordinates_latitude: 39.896579640955,
- Coordinates_longitude: 32.860911162601

GPT-5-mini

- Season: Autumn,
- Month: October,
- Time_of_day: 11:30,
- Daylight_phase: Morning,
- Continent: Europe,
- Country: Turkey,
- Climate_zone: Temperate (C),
- Environment_type: Urban,
- Coordinates_latitude: +39.93340,
- Coordinates_longitude: +32.85970

Intern-VL3.5-2B

- Season: Winter
- Month: October
- Time_of_day: 12:00
- Daylight_phase: Midday
- Continent: Europe
- Country: Turkey
- Climate_zone: Temperate
- Environment_type: Urban
- Coordinates_latitude: 41.167
- Coordinates_longitude: 28.917

Intern-VL2.5-72B

- Season: Summer,
- Month: July,
- Time_of_day: 14:00,
- Daylight_phase: Afternoon,
- Continent: Europe,
- Country: Turkey,
- Climate_zone: Temperate (C),
- Environment_type: Urban,
- Coordinates_latitude: +41.02370,
- Coordinates_longitude: +28.98390

Figure 7: Example of TIMESPOT dataset — Urban Morning Street Scene In Turkey.

Urban Facade In Thailand Afternoon

Picture

Ground Truth

- Season: Summer,
- Month: August,
- Time_of_day: 20:00:00,
- Daylight_phase: Night,
- Continent: Asia,
- Country: Thailand,
- Climate_zone: Tropical (A),
- Environment_type: Urban,
- Coordinates_latitude: 12.242756,
- Coordinates_longitude: 102.5174734

GPT-5-mini

- Season: Autumn,
- Month: November,
- Time_of_day: 19:30,
- Daylight_phase: Night,
- Continent: Asia,
- Country: China,
- Climate_zone: Temperate (C),
- Environment_type: Urban,
- Coordinates_latitude: +31.23041,
- Coordinates_longitude: +121.47370

Intern-VL3.5-2B

- Season: Winter
- Month: December
- Time_of_day: 23:00
- Daylight_phase: Night
- Continent: Asia
- Country: China
- Climate_zone: Temperate (C)
- Environment_type: Urban
- Coordinates_latitude: 36.89000
- Coordinates_longitude: 117.33100

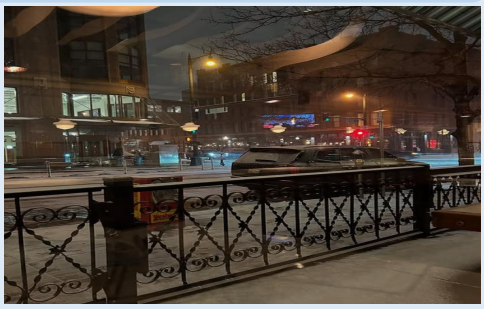
Intern-VL2.5-72B

- Season: Autumn,
- Month: October,
- Time_of_day: 20:00,
- Daylight_phase: Night,
- Continent: Asia,
- Country: China,
- Climate_zone: Temperate (C),
- Environment_type: Urban,
- Coordinates_latitude: +31.2304,
- Coordinates_longitude: +121.4737

Figure 8: Example of TIMESPOT dataset — Urban Facade In Thailand Afternoon.

Snowy Urban Night In USA

Picture



Ground Truth

- Season: Winter,
- Month: February,
- Time_of_day: 22:00:00,
- Daylight_phase: Night,
- Continent: North America,
- Country: USA,
- Climate_zone: Continental (D),
- Environment_type: Urban,
- Coordinates_latitude: 39.7392,
- Coordinates_longitude: -104.9903

GPT-5-mini

- Season: Winter,
- Month: January,
- Time_of_day: 20:30,
- Daylight_phase: Night,
- Continent: North America,
- Country: United States,
- Climate_zone: Continental (D),
- Environment_type: Urban,
- Coordinates_latitude: +41.87811,
- Coordinates_longitude: -087.62980

Intern-VL3.5-2B

- Season: Winter
- Month: January
- Time_of_day: 18:00
- Daylight_phase: Night
- Continent: North America
- Country: United States
- Climate_zone: Temperate (C)
- Environment_type: Urban
- Coordinates_latitude: -34.50167
- Coordinates_longitude: -73.85638

Intern-VL2.5-72B

- Season: Winter,
- Month: December,
- Time_of_day: 20:00,
- Daylight_phase: Night,
- Continent: North America,
- Country: United States,
- Climate_zone: Continental (D),
- Environment_type: Urban,
- Coordinates_latitude: +41.8781,
- Coordinates_longitude: -87.6298

Figure 9: Example of TIMESPOT dataset — Snowy Urban Night In USA.

Sunny Roadside Afternoon In Turkey

Picture

Ground Truth

- Season: Summer,
- Month: July,
- Time_of_day: 15:39,
- Daylight_phase: Afternoon,
- Continent: Asia,
- Country: Turkey,
- Climate_zone: Temperate (C),
- Environment_type: Urban,
- Coordinates_latitude: 40.189871,
- Coordinates_longitude: 25.895506

GPT-5-mini

- Season: Summer,
- Month: July,
- Time_of_day: 13:30,
- Daylight_phase: Midday,
- Continent: Asia,
- Country: Turkey,
- Climate_zone: Temperate (C),
- Environment_type: Rural,
- Coordinates_latitude: +38.35000,
- Coordinates_longitude: +27.20000

Intern-VL3.5-2B

- Season: Spring
- Month: April
- Time_of_day: 12:00
- Daylight_phase: Midday
- Continent: Asia
- Country: Israel
- Climate_zone: Temperate (C)
- Environment_type: Rural
- Coordinates_latitude: 31.854
- Coordinates_longitude: 34.915

Intern-VL2.5-72B

- Season: Summer,
- Month: July,
- Time_of_day: 12:00,
- Daylight_phase: Midday,
- Continent: Europe,
- Country: Greece,
- Climate_zone: Temperate (C),
- Environment_type: Rural,
- Coordinates_latitude: +37.95000,
- Coordinates_longitude: +23.71667

Figure 10: Example of TIMESPOT dataset — Sunny Roadside Afternoon In Turkey.

Rural River Afternoon In Russia

Picture

Ground Truth

- Season: Summer,
- Month: August,
- Time_of_day: 13:00:00,
- Daylight_phase: Afternoon,
- Continent: Asia,
- Country: Russia,
- Climate_zone: Continental (D),
- Environment_type: Rural,
- Coordinates_latitude: 56.793902,
- Coordinates_longitude: 118.273021

GPT-5-mini

- Season: Summer,
- Month: July,
- Time_of_day: 14:30,
- Daylight_phase: Afternoon,
- Continent: Asia,
- Country: Russia,
- Climate_zone: Continental (D),
- Environment_type: Rural,
- Coordinates_latitude: +60.50000,
- Coordinates_longitude: +101.50000

Intern-VL3.5-2B

- Season: Spring
- Month: May
- Time_of_day: 14:00
- Daylight_phase: Morning
- Continent: Europe
- Country: Ukraine
- Climate_zone: Continental (D)
- Environment_type: Mountain
- Coordinates_latitude: 38.4
- Coordinates_longitude: -116.6

Intern-VL2.5-72B

- Season: Summer,
- Month: June,
- Time_of_day: 12:00,
- Daylight_phase: Midday,
- Continent: Europe,
- Country: Russia,
- Climate_zone: Continental (D),
- Environment_type: Rural,
- Coordinates_latitude: +55.00000,
- Coordinates_longitude: +100.00000

Figure 11: Example of TIMESPOT dataset — Rural River Afternoon In Russia.

Urban Riverside Night In Ecuador

Picture

Ground Truth

- Season: Fall,
- Month: May,
- Time_of_day: 19:00:00,
- Daylight_phase: Night,
- Continent: South America,
- Country: Ecuador,
- Climate_zone: Tropical (A),
- Environment_type: Urban,
- Coordinates_latitude: -2.17,
- Coordinates_longitude: -79.9224

GPT-5-mini

- Season: Winter,
- Month: December,
- Time_of_day: 20:30,
- Daylight_phase: Night,
- Continent: Asia,
- Country: Thailand,
- Climate_zone: Tropical (A),
- Environment_type: Urban,
- Coordinates_latitude: +13.70850,
- Coordinates_longitude: +100.50420

Intern-VL3.5-2B

- Season: Winter
- Month: December
- Time_of_day: 22:00
- Daylight_phase: Night
- Continent: Asia
- Country: Japan
- Climate_zone: Temperate
- Environment_type: Urban
- Coordinates_latitude: 35.6995
- Coordinates_longitude: 139.7487

Intern-VL2.5-72B

- Season: Summer,
- Month: July,
- Time_of_day: 21:00,
- Daylight_phase: Night,
- Continent: North America,
- Country: United States,
- Climate_zone: Temperate (C),
- Environment_type: Urban,
- Coordinates_latitude: +25.7617,
- Coordinates_longitude: -80.1918

Figure 12: Example of TIMESPOT dataset — Urban Riverside Night In Ecuador.

Mountain Hillside Afternoon In Bulgaria

Picture

Ground Truth

- Season: Autumn,
- Month: November,
- Time_of_day: 20:08,
- Daylight_phase: Afternoon,
- Continent: Europe,
- Country: Bulgaria,
- Climate_zone: Continental (D),
- Environment_type: Mountain,
- Coordinates_latitude: 42.5492912,
- Coordinates_longitude: 23.0240361

GPT-5-mini

- Season: Autumn,
- Month: November,
- Time_of_day: 14:00,
- Daylight_phase: Afternoon,
- Continent: Europe,
- Country: Romania,
- Climate_zone: Continental (D),
- Environment_type: Mountain,
- Coordinates_latitude: 45.50000,
- Coordinates_longitude: 25.50000

Intern-VL3.5-2B

- Season: Autumn
- Month: October
- Time_of_day: 13:00
- Daylight_phase: Afternoon
- Continent: Europe
- Country: Romania
- Climate_zone: Temperate
- Environment_type: Mountain
- Coordinates_latitude: 45.56765
- Coordinates_longitude: 25.867

Intern-VL2.5-72B

- Season: Autumn,
- Month: October,
- Time_of_day: 16:00,
- Daylight_phase: Afternoon,
- Continent: North America,
- Country: United States,
- Climate_zone: Temperate (C),
- Environment_type: Rural,
- Coordinates_latitude: 35.00000,
- Coordinates_longitude: -105.00000

Figure 13: Example of TIMESPOT dataset — Mountain Hillside Afternoon In Bulgaria.

Cliffside Lake Afternoon In China

Picture

Ground Truth

- Season: Spring,
- Month: March,
- Time_of_day: 15:00:00,
- Daylight_phase: Afternoon,
- Continent: Asia,
- Country: China,
- Climate_zone: Continental (D),
- Environment_type: Suburban,
- Coordinates_latitude: 34.85714,
- Coordinates_longitude: 118.07065

GPT-5-mini

- Season: Autumn,
- Month: November,
- Time_of_day: 15:30,
- Daylight_phase: Afternoon,
- Continent: Asia,
- Country: China,
- Climate_zone: Continental (D),
- Environment_type: Mountain,
- Coordinates_latitude: +37.00000,
- Coordinates_longitude: +112.50000

Intern-VL3.5-2B

- Season: Autumn
- Month: September
- Time_of_day: 12:00
- Daylight_phase: Midday
- Continent: Asia
- Country: Iran
- Climate_zone: Arid (B)
- Environment_type: Mountain
- Coordinates_latitude: +31.0
- Coordinates_longitude: +48.9

Intern-VL2.5-72B

- Season: Autumn,
- Month: October,
- Time_of_day: 10:00,
- Daylight_phase: Morning,
- Continent: Asia,
- Country: China,
- Climate_zone: Continental (D),
- Environment_type: Mountain,
- Coordinates_latitude: +35.00000,
- Coordinates_longitude: +110.00000

Figure 14: Example of TIMESPOT dataset — Cliffside Lake Afternoon In China.

Forest Tram Midday In Russia

Picture

Ground Truth

- Season: Summer,
- Month: July,
- Time_of_day: 12:39,
- Daylight_phase: Midday,
- Continent: Asia,
- Country: Russia,
- Climate_zone: Continental (D),
- Environment_type: Rural
- Coordinates_latitude: 51.841950850008,
- Coordinates_longitude: 107.61454671992

GPT-5-mini

- Season: Summer,
- Month: June,
- Time_of_day: 17:30,
- Daylight_phase: Afternoon,
- Continent: Europe,
- Country: Russia,
- Climate_zone: Continental (D),
- Environment_type: Rural,
- Coordinates_latitude: +55.75000,
- Coordinates_longitude: +37.60000

Intern-VL3.5-2B

- Season: Summer
- Month: July
- Time_of_day: 10:00
- Daylight_phase: Morning
- Continent: Europe
- Country: Poland
- Climate_zone: Temperate (C)
- Environment_type: Rural
- Coordinates_latitude: +52.418
- Coordinates_longitude: +21.007

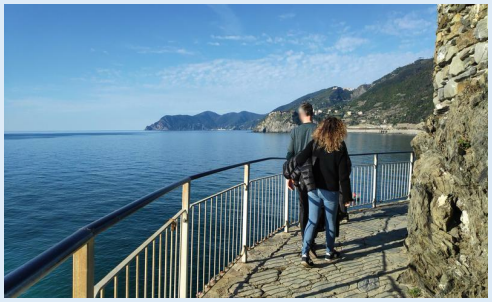
Intern-VL2.5-72B

- Season: Summer,
- Month: June,
- Time_of_day: 12:00,
- Daylight_phase: Midday,
- Continent: Europe,
- Country: Russia,
- Climate_zone: Continental (D),
- Environment_type: Rural,
- Coordinates_latitude: +55.75583,
- Coordinates_longitude: +37.61730

Figure 15: Example of TIMESPOT dataset — Forest Tram Midday In Russia.

Coastal Walkway Afternoon In Italy

Picture



Ground Truth

- Season: Winter,
- Month: December,
- Time_of_day: 18:11,
- Daylight_phase: Afternoon,
- Continent: Europe,
- Country: Italy,
- Climate_zone: Temperate (C),
- Environment_type: Coastal,
- Coordinates_latitude: 44.107313139433984,
- Coordinates_longitude: 9.725420480754678

GPT-5-mini

- Season: Autumn,
- Month: November,
- Time_of_day: 15:30,
- Daylight_phase: Afternoon,
- Continent: Europe,
- Country: Italy,
- Climate_zone: Temperate (C),
- Environment_type: Coastal,
- Coordinates_latitude: +44.10390,
- Coordinates_longitude: +9.70960

Intern-VL3.5-2B

- Season: Summer
- Month: July
- Time_of_day: 12:00
- Daylight_phase: Midday
- Continent: Europe
- Country: Italy
- Climate_zone: Temperate (C)
- Environment_type: Coastal
- Coordinates_latitude: +37.257
- Coordinates_longitude: +14.532

Intern-VL2.5-72B

- Season: Autumn,
- Month: October,
- Time_of_day: 16:00,
- Daylight_phase: Afternoon,
- Continent: Europe,
- Country: Italy,
- Climate_zone: Temperate (C),
- Environment_type: Coastal,
- Coordinates_latitude: +43.70000,
- Coordinates_longitude: +10.40000

Figure 16: Example of TIMESPOT dataset — Coastal Walkway Afternoon In Italy.

Rural Woodland Afternoon In China

Picture

Ground Truth

- Season: Winter,
- Month: February,
- Time_of_day: 17:00:00,
- Daylight_phase: Afternoon,
- Continent: Asia,
- Country: China,
- Climate_zone: Continental (D),
- Environment_type: Rural,
- Coordinates_latitude: 47.845639,
- Coordinates_longitude: 88.14023

GPT-5-mini

- Season: Autumn,
- Month: October,
- Time_of_day: 10:30,
- Daylight_phase: Morning,
- Continent: Asia,
- Country: China,
- Climate_zone: Arid (B),
- Environment_type: Desert,
- Coordinates_latitude: +41.85000,
- Coordinates_longitude: +100.23000

Intern-VL3.5-2B

- Season: Autumn
- Month: July
- Time_of_day: 13:00
- Daylight_phase: Morning
- Continent: Asia
- Country: South Korea
- Climate_zone: Temperate
- Environment_type: Rural
- Coordinates_latitude: 67.5676567,
- Coordinates_longitude: 90.67875

Intern-VL2.5-72B

- Season: Autumn,
- Month: October,
- Time_of_day: 16:00,
- Daylight_phase: Afternoon,
- Continent: Asia,
- Country: China,
- Climate_zone: Arid (B),
- Environment_type: Desert,
- Coordinates_latitude: +40.00000,
- Coordinates_longitude: +100.00000

Figure 17: Example of TIMESPOT dataset — Rural Woodland Afternoon In China.

Coastal Beachfront Afternoon In USA

Picture

Ground Truth

- Season: Summer,
- Month: August,
- Time_of_day: 13:00:00,
- Daylight_phase: Afternoon,
- Continent: North America,
- Country: USA,
- Climate_zone: Tropical (A),
- Environment_type: Coastal,
- Coordinates_latitude: 25.7617,
- Coordinates_longitude: -80.1918

GPT-5-mini

- Season: Summer,
- Month: July,
- Time_of_day: 13:00,
- Daylight_phase: Midday,
- Continent: North America,
- Country: United States,
- Climate_zone: Tropical (A),
- Environment_type: Coastal,
- Coordinates_latitude: +25.79065,
- Coordinates_longitude: -80.13005

Intern-VL3.5-2B

- Season: Summer
- Month: June
- Time_of_day: 12:00
- Daylight_phase: Midday
- Continent: North America
- Country: United States
- Climate_zone: Tropical
- Environment_type: Coastal
- Coordinates_latitude: +34.6071
- Coordinates_longitude: -82.2923

Intern-VL2.5-72B

- Season: Summer,
- Month: July,
- Time_of_day: 12:00,
- Daylight_phase: Midday,
- Continent: North America,
- Country: United States,
- Climate_zone: Tropical (A),
- Environment_type: Coastal,
- Coordinates_latitude: +25.7617,
- Coordinates_longitude: -80.1528

Figure 18: Example of TIMESPOT dataset — Coastal Beachfront Afternoon In USA.

Rural Forest Afternoon In Russia

Picture

Ground Truth

- Season: Summer,
- Month: June,
- Time_of_day: 17:00:00,
- Daylight_phase: Afternoon,
- Continent: Asia,
- Country: Russia,
- Climate_zone: Continental (D),
- Environment_type: Rural,
- Coordinates_latitude: 56.975553,
- Coordinates_longitude: 60.3645218

GPT-5-mini

- Season: Summer,
- Month: July,
- Time_of_day: 15:30,
- Daylight_phase: Afternoon,
- Continent: Europe,
- Country: Finland,
- Climate_zone: Continental (D),
- Environment_type: Rural,
- Coordinates_latitude: +63.50000,
- Coordinates_longitude: +25.00000

Intern-VL3.5-2B

- Season: Autumn
- Month: October
- Time_of_day: 12:00
- Daylight_phase: Morning
- Continent: North America
- Country: Canada
- Climate_zone: Temperate
- Environment_type: Mountain
- Coordinates_latitude: 49.1
- Coordinates_longitude: -99.4

Intern-VL2.5-72B

- Season: Summer,
- Month: July,
- Time_of_day: 15:00,
- Daylight_phase: Afternoon,
- Continent: Europe,
- Country: Russia,
- Climate_zone: Continental (D),
- Environment_type: Rural,
- Coordinates_latitude: +55.00000,
- Coordinates_longitude: +37.00000

Figure 19: Example of TIMESPOT dataset — Rural Forest Afternoon In Russia.

Rural Hillside Midday In Nepal

Picture

Ground Truth

- Season: Fall,
- Month: October,
- Time_of_day: 12:00:00,
- Daylight_phase: Midday,
- Continent: Asia,
- Country: Nepal,
- Climate_zone: Temperate (C),
- Environment_type: Rural,
- Coordinates_latitude: 28.5388196,
- Coordinates_longitude: 82.777235

GPT-5-mini

- Season: Autumn,
- Month: October,
- Time_of_day: 12:30,
- Daylight_phase: Midday,
- Continent: Asia,
- Country: Nepal,
- Climate_zone: Temperate (C),
- Environment_type: Mountain,
- Coordinates_latitude: +28.70000,
- Coordinates_longitude: +84.00000

Intern-VL3.5-2B

- Season: Spring
- Month: May
- Time_of_day: 11:00
- Daylight_phase: Morning
- Continent: Asia
- Country: Nepal
- Climate_zone: Arid (B)
- Environment_type: Rural
- Coordinates_latitude: +29.545
- Coordinates_longitude: +81.805

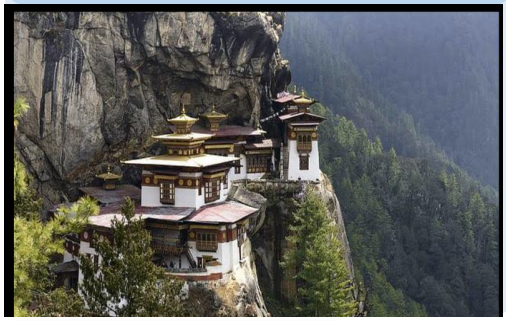
Intern-VL2.5-72B

- Season: Autumn,
- Month: October,
- Time_of_day: 14:00,
- Daylight_phase: Afternoon,
- Continent: Asia,
- Country: Nepal,
- Climate_zone: Temperate (C),
- Environment_type: Rural,
- Coordinates_latitude: +28.12500,
- Coordinates_longitude: +83.66667

Figure 20: Example of TIMESPOT dataset — Rural Hillside Midday In Nepal.

Mountain Monastery Morning In Bhutan

Picture



Ground Truth

- Season: Summer,
- Month: May,
- Time_of_day: 07:00:00,
- Daylight_phase: Morning,
- Continent: Asia,
- Country: Bhutan,
- Climate_zone: Temperate (C),
- Environment_type: Mountain,
- Coordinates_latitude: 27.4888,
- Coordinates_longitude: 89.3587

GPT-5-mini

- Season: Spring,
- Month: May,
- Time_of_day: 09:30,
- Daylight_phase: Morning,
- Continent: Asia,
- Country: Bhutan,
- Climate_zone: Temperate (C),
- Environment_type: Mountain,
- Coordinates_latitude: +27.49139,
- Coordinates_longitude: +89.36394

Intern-VL3.5-2B

- Season: Winter
- Month: January
- Time_of_day: 12:00
- Daylight_phase: Midday
- Continent: Asia
- Country: Nepal
- Climate_zone: Temperate (C)
- Environment_type: Mountain
- Coordinates_latitude: 27.7675
- Coordinates_longitude: 88.283

Intern-VL2.5-72B

- Season: Spring,
- Month: April,
- Time_of_day: 10:00,
- Daylight_phase: Morning,
- Continent: Asia,
- Country: Bhutan,
- Climate_zone: Temperate (C),
- Environment_type: Mountain,
- Coordinates_latitude: +27.10510,
- Coordinates_longitude: +90.48820

Figure 21: Example of TIMESPOT dataset — Mountain Monastery Morning In Bhutan.

Urban Plaza Sunset In Germany

Picture

Ground Truth

- Season: Spring,
- Month: May,
- Time_of_day: 17:48,
- Daylight_phase: Sunset,
- Continent: Europe,
- Country: Germany,
- Climate_zone: Temperate (C),
- Environment_type: Urban,
- Coordinates_latitude: 52.375299,
- Coordinates_longitude: 9.735393

GPT-5-mini

- Season: Spring,
- Month: May,
- Time_of_day: 18:45,
- Daylight_phase: Sunset,
- Continent: Europe,
- Country: Germany,
- Climate_zone: Temperate (C),
- Environment_type: Urban,
- Coordinates_latitude: +51.51363,
- Coordinates_longitude: +007.46530

Intern-VL3.5-2B

- Season: Summer
- Month: June
- Time_of_day: 13:00
- Daylight_phase: Afternoon
- Continent: Europe
- Country: Germany
- Climate_zone: Temperate
- Environment_type: Urban
- Coordinates_latitude: 52.518
- Coordinates_longitude: 13.404

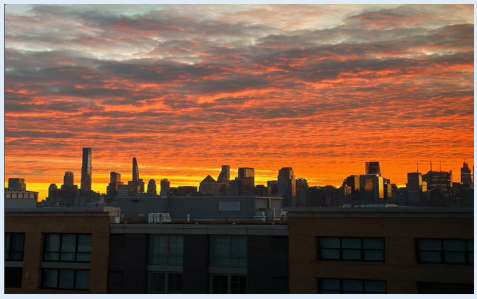
Intern-VL2.5-72B

- Season: Summer,
- Month: June,
- Time_of_day: 18:00,
- Daylight_phase: Sunset,
- Continent: Europe,
- Country: Germany,
- Climate_zone: Temperate (C),
- Environment_type: Urban,
- Coordinates_latitude: +52.5200,
- Coordinates_longitude: +13.4050

Figure 22: Example of TIMESPOT dataset — Urban Plaza Sunset In Germany.

Urban Skyline Sunset In USA

Picture



Ground Truth

- Season: Winter,
- Month: November,
- Time_of_day: 18:00:00,
- Daylight_phase: Sunset,
- Continent: North America,
- Country: USA,
- Climate_zone: Temperate (C),
- Environment_type: Urban,
- Coordinates_latitude: 40.7128,
- Coordinates_longitude: -74.006

GPT-5-mini

- Season: Autumn,
- Month: October,
- Time_of_day: 19:15,
- Daylight_phase: Sunset,
- Continent: North America,
- Country: United States,
- Climate_zone: Temperate (C),
- Environment_type: Urban,
- Coordinates_latitude: +40.75800,
- Coordinates_longitude: -073.98550

Intern-VL3.5-2B

- Season: Autumn
- Month: October
- Time_of_day: 18:00
- Daylight_phase: Sunset
- Continent: North America
- Country: United States
- Climate_zone: Temperate (C)
- Environment_type: Urban
- Coordinates_latitude: +38.961
- Coordinates_longitude: -77.628

Intern-VL2.5-72B

- Season: Fall,
- Month: October,
- Time_of_day: 18:30,
- Daylight_phase: Sunset,
- Continent: North America,
- Country: United States,
- Climate_zone: Temperate (C),
- Environment_type: Urban,
- Coordinates_latitude: +40.71278,
- Coordinates_longitude: -74.00597

Figure 23: Example of TIMESPOT dataset — Urban Skyline Sunset In USA.

Forest Road Afternoon In France

Picture

Ground Truth

- Season: Summer,
- Month: July,
- Time_of_day: 14:57,
- Daylight_phase: Afternoon,
- Continent: Europe,
- Country: France,
- Climate_zone: Temperate (C),
- Environment_type: Rural,
- Coordinates_latitude: 45.299533810116,
- Coordinates_longitude: 5.7525046159031

GPT-5-mini

- Season: Summer,
- Month: July,
- Time_of_day: 10:30,
- Daylight_phase: Morning,
- Continent: Europe,
- Country: Switzerland,
- Climate_zone: Temperate (C),
- Environment_type: Mountain,
- Coordinates_latitude: +46.80000,
- Coordinates_longitude: +008.20000

Intern-VL3.5-2B

- Season: Spring
- Month: May
- Time_of_day: 10:00
- Daylight_phase: Morning
- Continent: North America
- Country: Canada
- Climate_zone: Temperate
- Environment_type: Mountain
- Coordinates_latitude: +46.7379
- Coordinates_longitude: -111.6622

Intern-VL2.5-72B

- Season: Summer
- Month: July
- Time_of_day: 12:00
- Daylight_phase: Midday
- Continent: Europe
- Country: Germany
- Climate_zone: Temperate (C)
- Environment_type: Rural
- Coordinates_latitude: +47.50000
- Coordinates_longitude: +11.00000

Figure 24: Example of TIMESPOT dataset — Forest Road Afternoon In France.

Snowy Mountain Afternoon In Bhutan

Picture

Ground Truth

- Season: Autumn
- Month: October
- Time_of_day: 13:00:00
- Daylight_phase: Afternoon
- Continent: Asia
- Country: Bhutan
- Climate_zone: Continental (D)
- Environment_type: Mountain
- Coordinates_latitude: 27.8242
- Coordinates_longitude: 89.2692

GPT-5-mini

- Season: Autumn
- Month: October
- Time_of_day: 08:30
- Daylight_phase: Morning
- Continent: Asia
- Country: Nepal
- Climate_zone: Continental (D)
- Environment_type: Mountain
- Coordinates_latitude: 28.69667
- Coordinates_longitude: 83.48778

Intern-VL3.5-2B

- Season: Winter
- Month: January
- Time_of_day: 12:00
- Daylight_phase: Midday
- Continent: Asia
- Country: Nepal
- Climate_zone: Polar (E)
- Environment_type: Mountain
- Coordinates_latitude: +28.9
- Coordinates_longitude: +83.9

Intern-VL2.5-72B

- Season: Winter
- Month: December
- Time_of_day: 10:00
- Daylight_phase: Morning
- Continent: Asia
- Country: Nepal
- Climate_zone: Continental (D)
- Environment_type: Mountain
- Coordinates_latitude: 28.12500
- Coordinates_longitude: 85.00000

Figure 25: Example of TIMESPOT dataset — Snowy Mountain Afternoon In Bhutan.

Coastal Beach Afternoon In Philippines

Picture

Ground Truth

- Season: Winter
- Month: February
- Time_of_day: 17:00:00
- Daylight_phase: Afternoon
- Continent: Asia
- Country: Philippines
- Climate_zone: Tropical (A)
- Environment_type: Coastal
- Coordinates_latitude: 14.7191773
- Coordinates_longitude: 121.0583076

GPT-5-mini

- Season: Summer
- Month: May
- Time_of_day: 17:30
- Daylight_phase: Afternoon
- Continent: Asia
- Country: Philippines
- Climate_zone: Tropical (A)
- Environment_type: Coastal
- Coordinates_latitude: 11.00000
- Coordinates_longitude: 123.00000

Intern-VL3.5-2B

- Season: Summer
- Month: June
- Time_of_day: 11:00
- Daylight_phase: Morning
- Continent: Asia
- Country: Thailand
- Climate_zone: Tropical
- Environment_type: Coastal
- Coordinates_latitude: +1.862
- Coordinates_longitude: 99.593

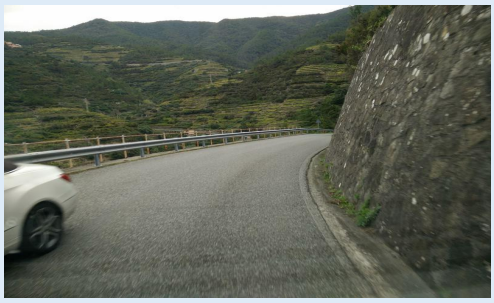
Intern-VL2.5-72B

- Season: Summer
- Month: June
- Time_of_day: 16:00
- Daylight_phase: Afternoon
- Continent: Asia
- Country: Indonesia
- Climate_zone: Tropical (A)
- Environment_type: Coastal
- Coordinates_latitude: -6.17500
- Coordinates_longitude: 106.82830

Figure 26: Example of TIMESPOT dataset — Coastal Beach Afternoon In Philippines.

Coastal Road Sunset In Italy

Picture



Ground Truth

- Season: Fall
- Month: October
- Time_of_day: 18:55
- Daylight_phase: Sunset
- Continent: Europe
- Country: Italy
- Climate_zone: Temperate (C)
- Environment_type: Coastal
- Coordinates_latitude: 44.10809959
- Coordinates_longitude: 9.73384485

GPT-5-mini

- Season: Spring
- Month: April
- Time_of_day: 16:30
- Daylight_phase: Afternoon
- Continent: Europe
- Country: Italy
- Climate_zone: Temperate (C)
- Environment_type: Mountain
- Coordinates_latitude: 44.12700
- Coordinates_longitude: 9.70800

Intern-VL3.5-2B

- Season: Autumn
- Month: October
- Time_of_day: 12:00
- Daylight_phase: Afternoon
- Continent: Europe
- Country: Spain
- Climate_zone: Temperate (C)
- Environment_type: Mountain
- Coordinates_latitude: 40.358
- Coordinates_longitude: 0.538

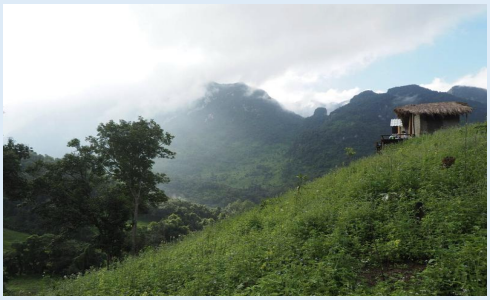
Intern-VL2.5-72B

- Season: Spring
- Month: April
- Time_of_day: 14:00
- Daylight_phase: Afternoon
- Continent: Europe
- Country: Spain
- Climate_zone: Temperate (C)
- Environment_type: Mountain
- Coordinates_latitude: 37.77490
- Coordinates_longitude: -0.45500

Figure 27: Example of TIMESPOT dataset — Coastal Road Sunset In Italy.

Mountain Hillside Afternoon In Thailand

Picture



Ground Truth

- Season: Fall
- Month: October
- Time_of_day: 15:00:00
- Daylight_phase: Afternoon
- Continent: Asia
- Country: Thailand
- Climate_zone: Tropical (A)
- Environment_type: Mountain
- Coordinates_latitude: 19.367814
- Coordinates_longitude: 98.9649024

GPT-5-mini

- Season: Summer
- Month: August
- Time_of_day: 08:30
- Daylight_phase: Morning
- Continent: Asia
- Country: Thailand
- Climate_zone: Tropical (A)
- Environment_type: Mountain
- Coordinates_latitude: 18.69700
- Coordinates_longitude: 98.48700

Intern-VL3.5-2B

- Season: Spring
- Month: April
- Time_of_day: 12:00
- Daylight_phase: Midday
- Continent: Asia
- Country: Myanmar
- Climate_zone: Tropical
- Environment_type: Mountain
- Coordinates_latitude: 22.8
- Coordinates_longitude: 81.6

Intern-VL2.5-72B

- Season: Summer
- Month: June
- Time_of_day: 08:00
- Daylight_phase: Morning
- Continent: Asia
- Country: Nepal
- Climate_zone: Tropical (A)
- Environment_type: Mountain
- Coordinates_latitude: 28.12500
- Coordinates_longitude: 83.75000

Figure 28: Example of TIMESPOT dataset — Mountain Hillside Afternoon In Thailand.

Mountain Road Sunset In Italy

Picture

Ground Truth

- Season: Winter
- Month: December
- Time_of_day: 17:40
- Daylight_phase: Sunset
- Continent: Europe
- Country: Italy
- Climate_zone: Continental (D)
- Environment_type: Mountain
- Coordinates_latitude: 46.641320366734
- Coordinates_longitude: 11.691309903245

GPT-5-mini

- Season: Winter
- Month: January
- Time_of_day: 15:30
- Daylight_phase: Afternoon
- Continent: Europe
- Country: Austria
- Climate_zone: Continental (D)
- Environment_type: Mountain
- Coordinates_latitude: +47.30000
- Coordinates_longitude: +012.00000

Intern-VL3.5-2B

- Season: Winter
- Month: January
- Time_of_day: 10:00
- Daylight_phase: Morning
- Continent: Europe
- Country: Austria
- Climate_zone: Temperate (C)
- Environment_type: Mountain
- Coordinates_latitude: 46.2
- Coordinates_longitude: 14.5

Intern-VL2.5-72B

- Season: Winter
- Month: December
- Time_of_day: 16:00
- Daylight_phase: Afternoon
- Continent: Europe
- Country: Switzerland
- Climate_zone: Continental (D)
- Environment_type: Mountain
- Coordinates_latitude: +46.50000
- Coordinates_longitude: +8.50000

Figure 29: Example of TIMESPOT dataset — Mountain Road Sunset In Italy.