ICML
International Conference
On Machine Learning

TimeSpot: Benchmarking Geo-Temporal Understanding in Vision–Language Models in Real-World Settings

Azmine Toushik Wasi, Shahriyar Zaman Ridoy, Koushik Ahamed Tonmoy, Kinga Tshering, S. M. Muhtasimul Hasan, Wahid Faisal, Tasnim Mohiuddin, Md Rizwan Parvez

Computational Intelligence and Operations Lab

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

TimeSpot-Gt.github.io

Inferring where and when, from pixels alone

Humans read **subtle, distributed physical cues** — shadow geometry, seasonal vegetation, materials, clothing, sky — to reconstruct a scene's place and moment.

WHEN

- Illumination & shadows
- Vegetation phenology
- Sky luminance gradient

WHERE

- Architecture & materials
- Natural biome & terrain
- Signage, vehicles, roads

Underpins disaster response · navigation · environmental monitoring · world modeling

Benchmarks measure where — and ignore when

WHAT WE HAVE

Mature spatial localization

- Cross-view & street-view retrieval, scored by rank or coordinate error
- Large, geographically diverse datasets and embedding frameworks
- Driven by landmarks, road signs, and other iconic, text-like cues

WHAT IS MISSING

Temporal & joint reasoning

- Explicit time — season, month, daylight phase, local clock time
- Cross-field consistency — no **"snow in July"** in the North
- Non-iconic scenes that demand subtle physical inference

TimeSpot: a diagnostic geo-temporal benchmark

1,455
ground-level images

80
countries · 6 continents

4+5
temporal + geographic fields

Structured, verifiable, physically-grounded prediction — auditable for consistency, not just ranked by retrieval.

One image → nine structured fields

TEMPORAL · 4

Season

Month

Local time

HH:MM

Daylight phase

Scored by windowed accuracy (≤1h) and minute-level MAE

GEOGRAPHIC · 5

Continent

Country

Climate zone

Köppen–Geiger

Environment

Latitude–longitude

Scored by categorical accuracy and geodesic distance (km)

Automated audits enforce month–season–hemisphere alignment · phase–time compatibility · climate plausibility at (lat, lon)

Objective physical labels, not subjective tags

01

Collect & curate

Public & captured imagery; landmark- and text-rich scenes suppressed.

02

Derive labels

Seasons, daylight phase & local time from timestamps and **solar ephemerides**; climate & place from coordinates.

03

Human verify

Two-stage check by 5 trained annotators; senior adjudication of twilight & lighting edge cases.

04

Audit consistency

Constraint checks across all nine fields produce auditable, reproducible records.

~576 h primary annotation, over 6 weeks

no crowdsourcing engineering graduates with geospatial expertise

Strong on place, weak on time

WHERE · coarse spatial

77.6%

country accuracy for the best model — yet a **median geodesic error of 892 km**. High recognition, weak metric grounding.

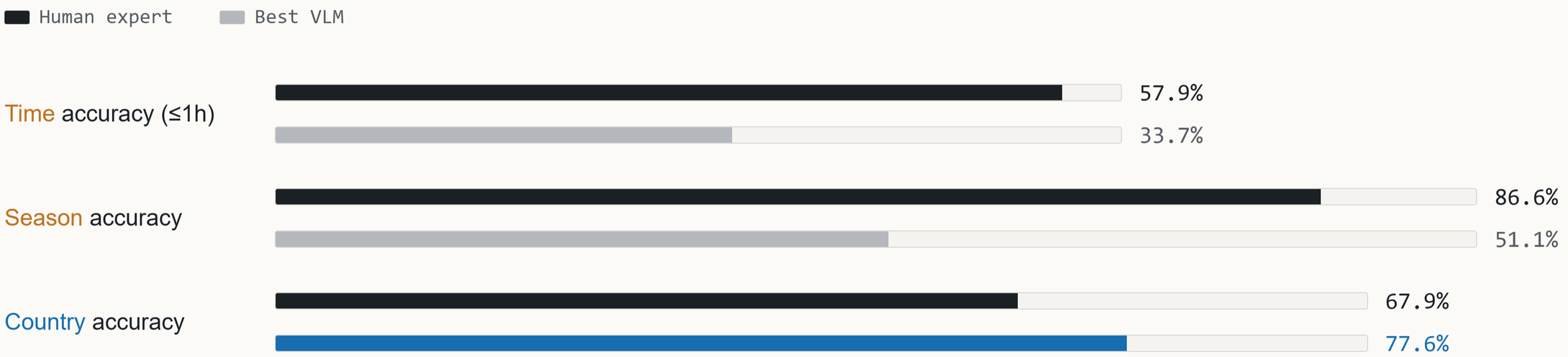
WHEN · local time

~34%

peak time-of-day accuracy — most models cluster at **22–34%**, with mean errors near **four hours**.

Best overall: Gemini-2.5-Flash-Thinking · 25 open- & closed-source VLMs evaluated under uniform, metadata-free prompting

Models match humans on place — not on time



VLMs lag experts on every temporal field — best Time MAE 3:36 vs. expert 1:36 — while matching them on coarse spatial memorization.

Failures are systematic, not random

WHEN

Round-time anchoring

Minute-level predictions collapse onto familiar clock anchors — 15-minute marks, "20:30" at night.

WHEN

Sunrise–sunset flips

Symmetric chromatic sky cues confuse dawn with dusk — a ~12-hour error despite little spatial drift.

WHERE

Neighbor-country confusion

Right continent, wrong country — shared regional style defaults to larger neighbors (Bangladesh→India).

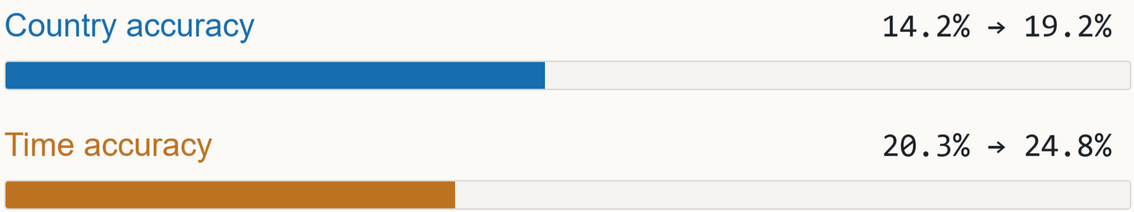
JOINT

Cross-field inconsistency

Phase–time–longitude conflicts; autumn collapses entirely; climate & environment default to "Temperate, Urban".

Fine-tuning helps — but space and time compete

Fine-tuning Qwen2.5-VL-3B on a 40% split lifts both tasks — but single-task tuning **degrades the other**, and joint tuning stays below either peak.



GRADIENT CONFLICT

Country prediction leans on illumination-invariant structure; time leans on illumination-sensitive shadows and sky.

Under shared LoRA parameters the two objectives interfere — motivating constraint-aware and RL-based training.

What TimeSpot reveals

01 High spatial scores can **mask weak physical grounding** — coarse place recognition coexists with large metric and temporal errors.

02 Temporal reasoning is **not ancillary** — it is essential for coherent, physically-plausible world modeling.

03 The path forward: explicit solar-geometry biases, topographic priors, and constraint-aware supervision.

TimeSpot — benchmarking when as rigorously as where.

ICML
International Conference
On Machine Learning

Thank You!

If you have any questions, please ask!



Computational
Intelligence and
Operations Lab

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

TimeSpot-Gt.github.io