Skyfall: A Hierarchical Methodology for Probing the Limits of VLMs in Earth Observation

Devansh Lodha[1], Rishabh Mondal[1], Ujjwal Gupta[1,2], Nipun Batra[1]

[1]Indian Institute of Technology Gandhinagar, [2]Space Applications Centre, ISRO

SUSTAINABILITY LAB

Scan for the open-source code.

Problem Statement: Opaque Models

The deployment of Vision-Language Models (VLMs) in scientific domains such as Earth Observation is critically impeded by their black-box nature. Standard evaluation offers insights into what a model predicts, but fails to provide a formal, reproducible methodology to diagnose how and why it fails in its geospatial reasoning.

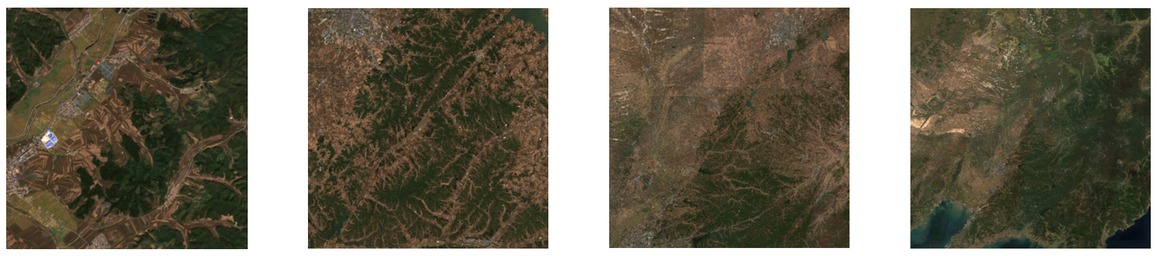
Our Contribution

We introduce **Skyfall**, a comprehensive, open-source framework designed to address this gap.

Skyfall Global Scene Packets

View Name	Assigned Tasks	Data Source	GSD (m)	Area (km²)
L1: Detail	Land Cover Classification	ESA WorldCover	10	5.12 × 5.12
L2: Local	State/Province Classification	FAO GAUL	100	51.2 × 51.2
L3: Regional	Biome, Climate Zone Classification	RESOLVE / Köppen-Geiger	500	256 × 256
L4: Synoptic	Country, Continent Classification	USDOS LSIB	1500	768 × 768

Views obtained from Sentinel - 2 (COPERNICUS/S2_SR_HARMONIZED)



L1: Detail | L2: Local | L3: Regional | L4: Synoptic

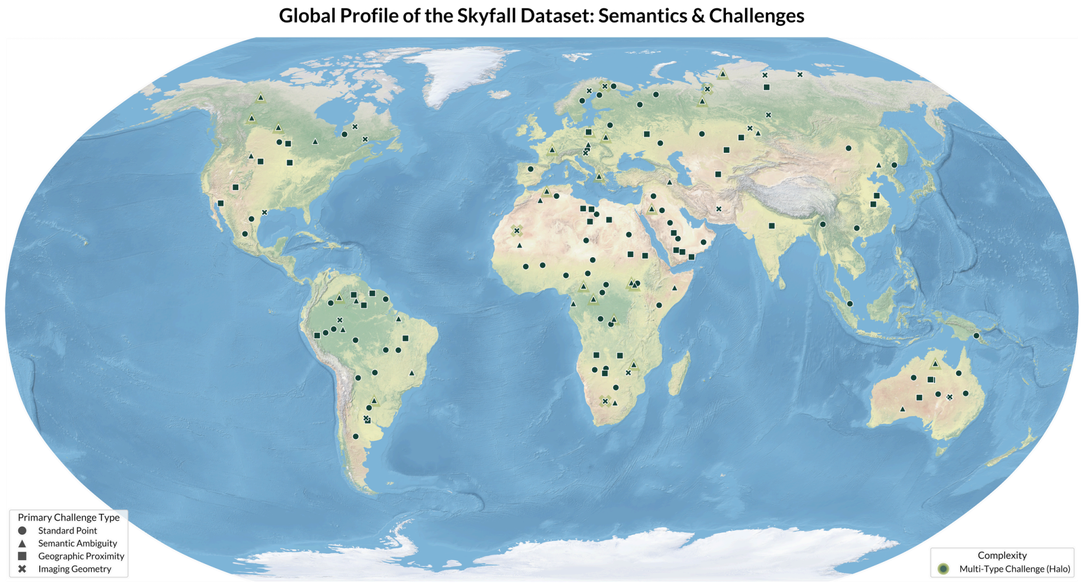
Skyfall Atlas

A knowledge graph containing the geographic, semantic, and probabilistic data including entity-level environmental distributions, a climate distance matrix, and real-world conditional probabilities.

Skyfall - X

A framework to quantify performance degradation against explicit, real-world challenges (e.g., long shadows, semantic ambiguity, mosaic artifacts).

Global Profile of the Skyfall Dataset: Semantics & Challenges



The Skyfall Evaluation Methodology

A model's final prediction, $\hat{y}_{i,t}$, is generated by composing the VLM ($\mathcal{V}$) with a Judge LLM ($\mathcal{J}$) that parses the raw output. This prediction is then compared to the ground truth y_i.

$$\hat{y}_{i,t} = (\mathcal{J} \circ \mathcal{V})(I_{i,t}, q_t)$$

Foundational Metrics

Strict Top-1 Accuracy measures the fraction of answers correct at the high-level `evaluation_group`. **Abstention Rate** measures the fraction of `Uncertain` responses.

$$\text{Accuracy}(T_t) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\text{group}(\hat{y}_{i,t}) = \text{group}(y_{i,t})) \qquad \text{AbstentionRate}(T_t) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\hat{y}_{i,t} = \texttt{Uncertain})$$

The Primary Metric: CASCADE Score

The **CASCADE** score, $S_i \in [0,1]$, provides nuanced partial credit for plausible errors by measuring them along geographic and semantic axes.

$$S_i(y_i, \hat{y}_i) = \begin{cases} 1.0 & \text{if } \text{group}(\hat{y}_i) = \text{group}(y_i) \\ 0.0 & \text{if } \hat{y}_i \in \{\texttt{Uncertain}, \texttt{Unparseable}\} \\ \text{score}_{\text{dist}}(y_i, \hat{y}_i) & \text{if } \text{group}(\hat{y}_i) \neq \text{group}(y_i) \end{cases}$$

A. Geographic Tasks (Country, State)

For incorrect answers, the score is a geometric mean of Proximity (S_{prox}) and Semantic Similarity (S_{sem}), ensuring dual competence.

$$\text{score}_{\text{dist}} = \sqrt{S_{\text{prox}}(y_{\text{entity}}, \hat{y}_{\text{entity}}) \cdot S_{\text{sem}}(y_{\text{point}}, \hat{y}_{\text{entity}})}$$

S_{prox} penalizes for physical distance (p_{phys}) and geopolitical hierarchy errors (p_{struct}). S_{sem} measures if the predicted entity's environment matches the image.

$$S_{\text{prox}} = 1 - (p_{\text{struct}} + (1 - p_{\text{struct}})p_{\text{phys}}) \qquad S_{\text{sem}} = \text{avg}(\text{Sim}_{\text{climate}}, \text{Sim}_{\text{biome}})$$

B. Semantic Tasks (Climate, Biome, Land Cover)

For incorrect answers, the **Semantic Plausibility Score (SPS)** is derived from conditional probabilities in the Atlas ($\mathcal{A}$). E.g., for a Biome error:

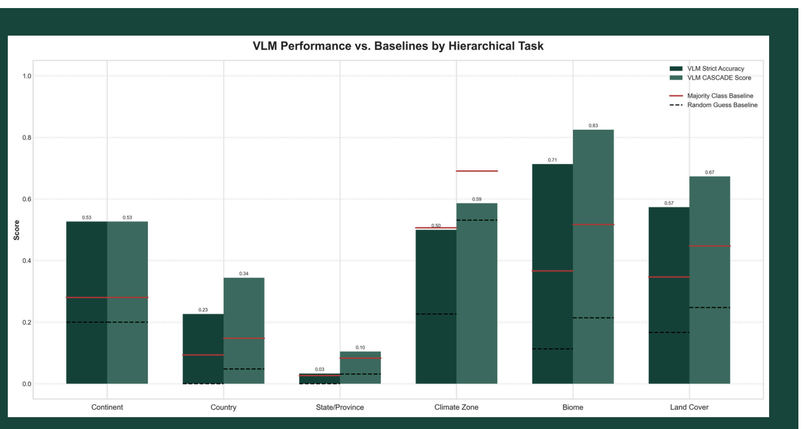
$$\text{SPS} = \text{P}(\text{Biome} = \hat{y}_{\text{biome}} | \text{Climate} = y_{\text{climate}})$$

Skyfall-X Robustness Analysis

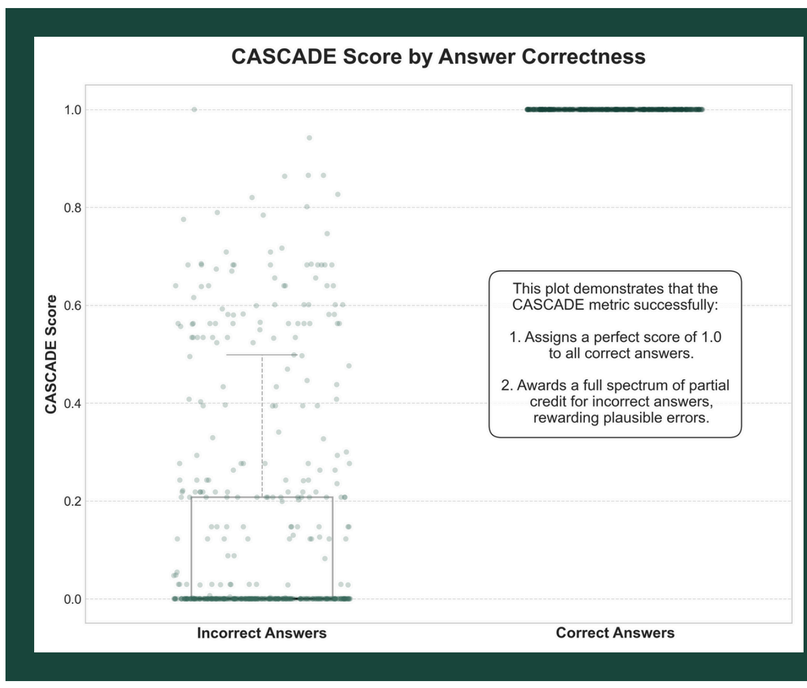
This framework quantifies model brittleness by measuring performance degradation Δ_k on challenging data subsets (e.g., scenes with `Extreme_Sun_Angle`) and by analyzing the correlation between performance and mosaic temporal artifacts.

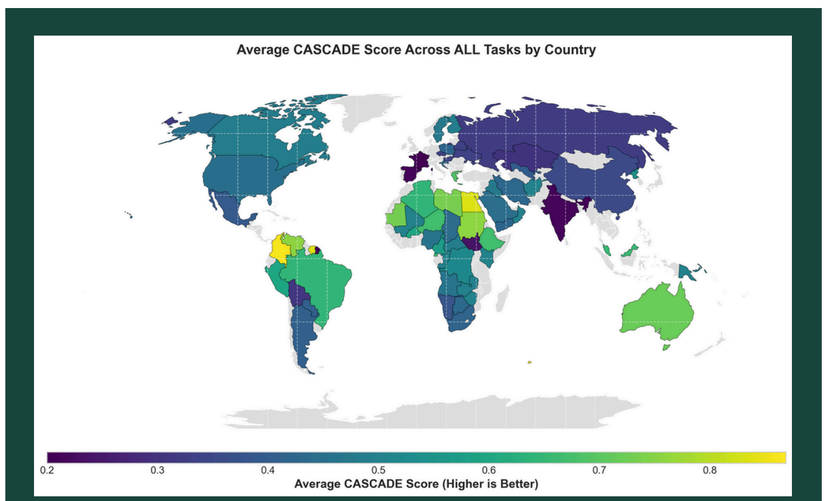
$$\Delta_k(M) = M(\mathcal{D}_{\neg k}) - M(\mathcal{D}_k)$$

Results and Discoveries

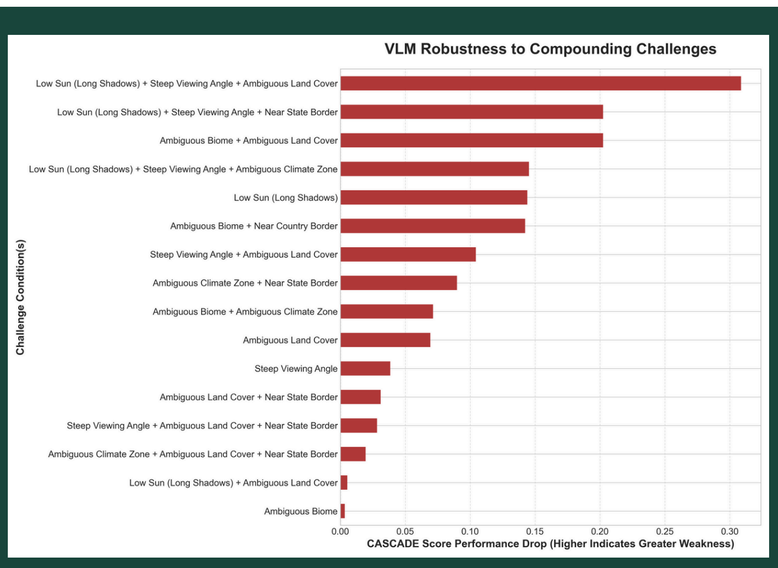


The VLM significantly outperforms baselines across all tasks, but performance degrades as geographic specificity increases from Continent to Country and State/Province.

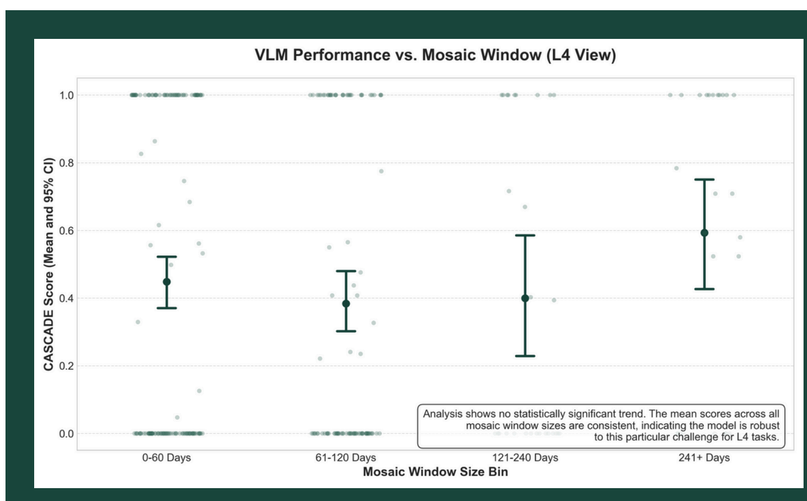




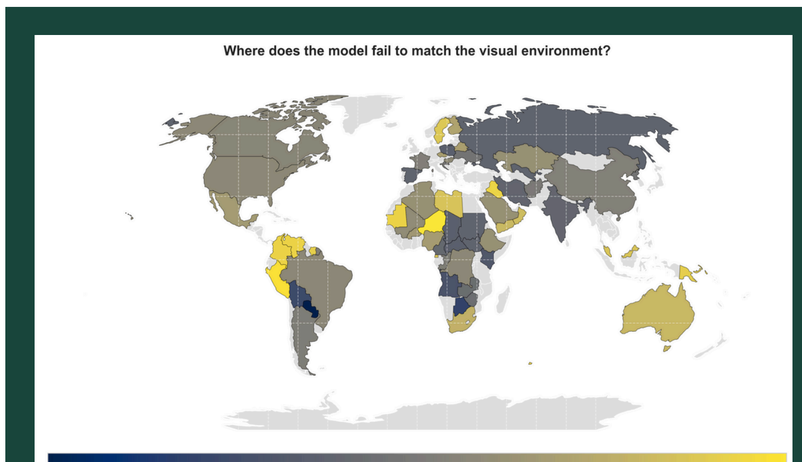
The model's performance reveals a strong training data bias, favoring Western nations while underperforming in diverse, non-Latin script regions like India, China, and Russia.



Model robustness degrades catastrophically under compounding challenges.



The model is robust to temporal incoherence in mosaics.



The model generally makes semantically plausible errors indicating strong visual but poor geographic reasoning.

Future Work: Probe object to scene grounding and temporal reasoning.

Reference: Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment, 202, 18-27.