

# GeoRC: A Benchmark for Geolocation Reasoning Chains

Mohit Talreja, Joshua Diao, Jim Thannikary James, Radu Casapu,  
Tejas Santanam, Ethan Mendes, Alan Ritter, Wei Xu, James Hays

{mtalreja6, jdiao6, jimjames, rcasapu3, tsantanam, emendes3, hays}@gatech.edu,  
{alan.ritter, wei.xu}@cc.gatech.edu

Georgia Institute of Technology, Atlanta, GA, U.S.A

[Project Home](#), [GitHub](#), [HuggingFace](#)

## Abstract

Vision Language Models (VLMs) are good at recognizing the global location of a photograph – their geolocation prediction accuracy rivals the best human experts. But many VLMs are startlingly bad at *explaining* which image evidence led to their prediction, even when their location prediction is correct. In this paper, we introduce GeoRC, the first benchmark for geolocation reasoning chains sourced directly from Champion-tier GeoGuessr experts, including the reigning world champion. This benchmark consists of 800 “ground truth” reasoning chains across 500 query scenes from GeoGuessr maps, with expert chains addressing hundreds of different discriminative attributes, such as soil properties, architecture, and license plate shapes. We evaluate LLM-as-a-judge and VLM-as-a-judge strategies for scoring VLM-generated reasoning chains against our expert reasoning chains and find that Qwen 3 LLM-as-a-judge correlates best with human-expert scoring. Our benchmark reveals that while large, closed-source VLMs such as Gemini and GPT 5 rival human experts at predicting locations, they still lag behind human experts when it comes to producing auditable reasoning chains. Small open-weight VLMs such as Llama and Qwen catastrophically fail on our benchmark – they perform only slightly better than a baseline in which an LLM hallucinates a reasoning chain with oracle knowledge of the photo location but *no visual information at all*. We believe the gap between human experts and VLMs on this task points to VLM limitations at extracting fine-grained visual attributes from high resolution images. We open source our benchmark for the community to use.

## 1 Introduction

The task of determining the location of a photo has been of interest for more than a century. For example, Frederick Cook claimed to be the first person to climb Denali in 1906 and offered a photograph of

the purported summit to support his claim. Cook’s claim was discredited by the geolocation of that photo to a different mountain (Washburn, 1956) among other evidence. Beyond photo forensics, global photo geolocation has long been seen as a fun brain teaser, e.g. Condé Nast Traveler’s “Where are you?” competition dating to 1993 (Condé Nast Traveler, 2011).

Today, photo geolocation is still widely relevant as a forensics task, e.g. work by investigative journalists such as Bellingcat and others in the OSINT community, and as a game, e.g. GeoGuessr (GeoGuessr AB, 2013). In both cases, human experts demonstrate extraordinary skill at using subtle image evidence to determine the location of photographs.

In the last two decades, machine learning approaches have made enormous progress on the global image geolocation task. Starting from im2gps (Hays and Efros, 2008, 2015), methods improved with the introduction of deep learning (Weyand et al., 2016; Vo et al., 2017; Seo et al., 2018; Müller-Budack et al., 2018), then with foundation models such as CLIP (Radford et al., 2021; Clark et al., 2023; Vivanco Cepeda et al., 2023; Haas et al., 2023a,b; Luo et al., 2022), and recently with large scale Vision Language Models (VLMs) (Zhang et al., 2024; Mendes et al., 2024; Yerramilli et al., 2025; Zhang et al., 2025). VLMs, with no fine-tuning, appear to roughly match the performance of bespoke geolocation methods.

After two decades of machine learning investigation into geolocation, it is still not clear whether machines or human experts are better at this task. Pigeon (Haas et al., 2023b) claims that their geolocation method outperforms the best GeoGuessr players. Our experiments suggest that humans still reign supreme. The difference is small and probably depends on the particular experimental setup. It may be the case that the strongest possible geolocation system is a hybrid combination of humans and

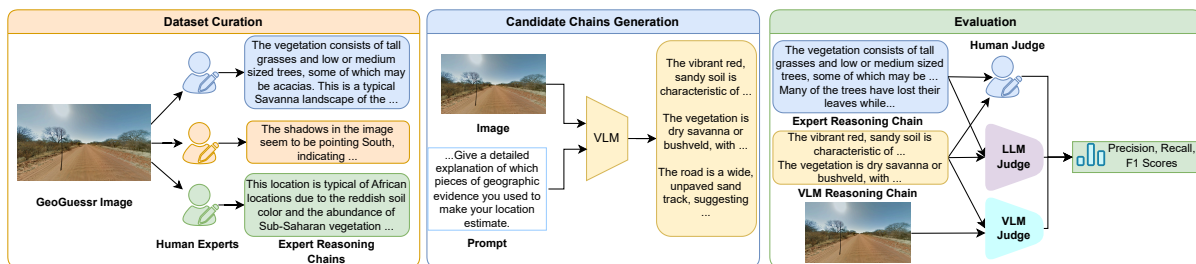


Figure 1: GeoRC Benchmark. We curate a dataset of GeoGuessr challenges and associated reasoning chains from three human experts. We then generate reasoning chains from open-weight and proprietary VLMs, and evaluate them through several proposed judging methods.

machines, analogous to the face recognition task where human “superrecognizers” have been used as verifiers for machine learning methods (Phillips et al., 2018).

While the gap in geolocation *accuracy* between human experts and machines may be small, we claim that the gap in *explainability and auditability* is large. When asked to geolocate a photo, human experts can support their decision with specific image evidence related to infrastructure, vegetation, architecture, and a litany of additional attributes that led them to their conclusion. These reasoning chains help establish *trust* – even a non-expert can verify the presence of the lane markings, writing, or terrain attributes mentioned in the reasoning chain. These explanations also *teach* non-experts to perform the task themselves.

Large generative Vision Language Models can also be asked to *explain their reasoning* at the geolocation task. Generally, they can produce a list of evidence that is qualitatively similar to human expert reasoning chains. However, these VLM reasoning chains often contain *hallucinations*, miss *fine-scale image details*, and exhibit tunnel vision in *rationalizing* the decision made by the VLM.

In this work, we introduce the first purpose-built benchmark of *geolocation reasoning chains*, elicited directly from credentialed *Champion-tier GeoGuessr experts*, including the reigning GeoGuessr world champion, Radu Casapu (RC). We propose a grading scheme to assess how well a candidate reasoning chain agrees with a “ground truth” human expert reasoning chain. We deploy an “LLM-as-a-Judge” approach to apply our grading scheme to numerous VLMs – smaller, open weight models such as Gemma (Team et al., 2025), Llama (Dubey et al., 2024), and Qwen (Yang et al., 2025) as well as large, closed weight models such as Gemini (DeepMind, 2025) and ChatGPT (Ope-

nAI, 2025b). We find a large spread in capability among these models, and we find that the best model (GPT-4.1) still lags behind human experts.

The contributions of our work include:

- The first purpose-built dataset of human expert geolocation reasoning chains (Section 2)
- A grading protocol for humans and machines to evaluate reasoning chain agreement in terms of precision and recall (and summarized with F1 score) (Section 3.1)
- An investigation into various LLM-as-a-judge and VLM-as-a-judge methods for assessing reasoning chain quality (Section 3.2)
- The first quantification of VLM reasoning chain quality and a characterization of the dominant errors observed – misattribution, hallucinations, false tool use, axiomatic irrelevance, and missed details (Section 4)
- We open source our expert reasoning chains and our best LLM-as-a-judge benchmark for the community to use.

## 2 Geolocation Reasoning Chains (GeoRC)

The aim of a geolocation reasoning chain is to detail the thought process of an expert guessing the location from an image. It describes the supporting evidence in an arbitrary order in the form of scene attributes consisting of, but not limited to, infrastructure, architecture, vegetation, climate, geology, terrain, culture, vehicles, and language. In this section, we first characterize a reasoning chain for this task, state its properties, and then explain the details of our dataset.

### 2.1 Characterization of GeoRC

Generally, we expect that “good” geolocation reasoning chains progressively refine an estimate of the location from available evidence, starting from a coarse level (e.g., hemisphere) to a fine level (e.g.,

city). Reasoning chains typically only contain discriminative scene attributes, rather than being exhaustive. Each attribute should also be associated with a statement of geographic support. For example, “Short bollards with a vertical reflector and a black ‘cap’ on top are found in Austria and former Yugoslav countries” cites bollards as a scene attribute alongside relevant geographic regions. Each step in the chain may also represent the confidence level or the degree with which the attribute aids in progressing to the next step in the chain. The “conclusion” statement of a reasoning chain consists of the final guess consisting of the country and city or region along with an informal degree of confidence.

## 2.2 The GeoRC Dataset

We curate a dataset, titled GeoRC, consisting of 800 reasoning chains generated by three expert GeoGuessr players. Our first two experts, Joshua Diao (JD) and Tejas Santanam (TS), are ranked in the Champion Division (top 0.01% of players), while our final expert is the 2025 GeoGuessr World Cup champion and a professional GeoGuessr player, Radu Casapu (RC). Our 800 reasoning chains are generated from 100 GeoGuessr challenges, each comprising 5 unique locations. We task our experts with generating reasoning chains that fit the properties laid out in Section 2.1, with each expert writing reasoning chains for a shared set of 150 locations, while the remaining 350 locations are divided among all three experts. See Appendix A.1 for detailed instructions provided to our GeoGuessr players.

## 2.3 Geographic Scene Attribute Categories

The non-exhaustive reasoning chains in the GeoRC dataset cite a wide variety of geographic scene attributes, which can be broadly categorized. The most cited categories are the infrastructure visible (*e.g.* poles, bollards), followed by vegetation and architecture. One unique category is “meta information,” which includes GeoGuessr and Street View-specific cues such as the Street View car, camera quality, and map coverage. The language category typically refers to signs with visible language. The least cited category happens to be the culture that is unique to a specific region or country. Additional categories include terrain, climate, vehicles, and geology. After having our experts write reasoning chains, we ask an LLM to categorize and label each point in a reasoning chain into a maximum of three categories. Figure 3 shows the count of these

categories across our 800 expert chains.

## 2.4 Geographical Distribution

Distribution of chosen locations in our dataset are derived from popular maps on GeoGuessr, such as GeoGuessr Saturday, An Arbitrary World, and An Arbitrary Urban World, all of which are inherently conditioned upon the distribution of official Google Street View coverage. These maps are curated through a random sampling of latitude and longitude, followed by selecting the nearest street view image to the randomly sampled location coordinates (Vercel). Sampling of the maps is also parameterized, such as “urban” for the Arbitrary Urban World map to indicate urban city landscapes (slashP, 2023). The geographical distribution of our dataset is represented in Figure 4.

## 3 Evaluating Reasoning Chains

We aim to develop an automated method for assessing candidate reasoning chains against human expert chains. In this section, we describe the process of human grading, which we use for calibration, and propose three methods for grading candidate reasoning chains. The first two approaches utilize an LLM to measure the relationship between each point in the candidate chain and points in the ground truth chain. The third approach utilizes an open source VLM in addition to an LLM to score the candidates. In-depth algorithm pseudocode can be found in Appendix A.3.3.

### 3.1 Human Grading

We asked our expert GeoGuessr players to evaluate 150 candidate reasoning chains by adopting the one-to-all bipartite strategy. In this strategy, each candidate step is compared to all steps in the reference chain. The overall score is computed by averaging across all points. Each comparison direction gives two measures, precision and recall, which are then used to calculate the F1 score. An illustration of the approach on a candidate reasoning chain is shown in Figure 12. To avoid bias, graders were assigned reasoning chains they did not write. Exact grading guidelines can be found in Appendix A.3.2.

### 3.2 Approach 1: One-to-all LLM-as-a-judge Evaluation

In this approach, we utilize only an LLM judge to determine the degree of similarity of each point in the candidate reasoning chain with the points in

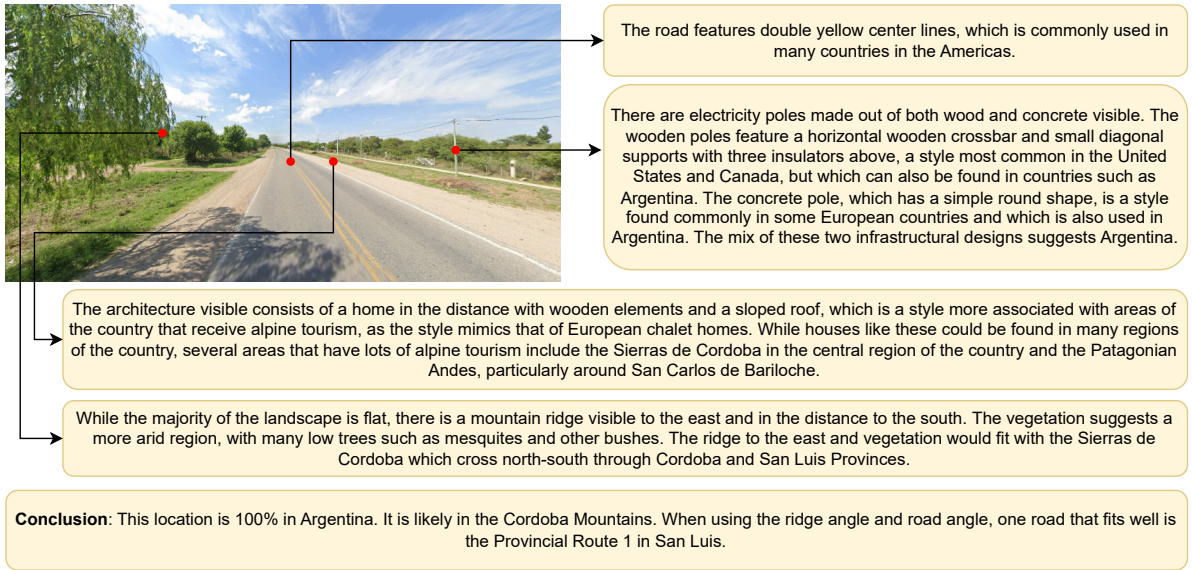


Figure 2: **Example of Expert Reasoning Chain.** Experts typically note discriminative visual features in a coarse-to-fine manner before concluding with a country guess. Expert chains are usually non-exhaustive, only requiring a small number of keypoints to localize the image to a country or region.

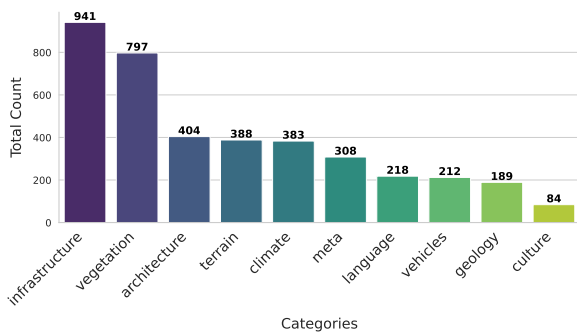


Figure 3: Categories of Geographic Scene Attributes cited by our GeoGuessr experts. Infrastructure and vegetation are the top cited scene attributes.

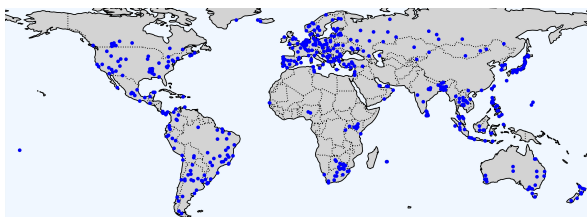


Figure 4: Geographical Distribution of GeoRC Dataset that is drawn from popular GeoGuessr world maps inherently conditioned upon Google Street View coverage.

the ground truth reasoning chain. In our prompt to score a given candidate point against the complete ground truth chain, the LLM judge is provided the context, a set of rules to adhere to, the complete ground truth chain, and a single candidate point and is asked to respond with a similarity score out of 100. We request the LLM judge over each can-

didate point and then compute an average score for the complete chain. This results in the precision score for our grading. Similarly, when iterating through each ground truth point and the complete candidate chain, we obtain the recall score for our grading. Using the precision and recall together, we compute the F1 score. The algorithm for this approach is shown in Algorithm 1 with chain 1 being the candidate chain and chain 2 the ground truth chain for precision and vice versa for recall. We had also experimented with simpler strategies, such as both the chains scored in their entirety, but these deviated from the human grading expectations.

### 3.3 Approach 2: Key Points Guided LLM-as-a-judge Evaluation

To further enhance the reliability, consistency, and to mitigate biases with LLM-as-a-judge ((Gu et al., 2026), (Fujinuma, 2025)), we propose this key points guided approach. The complex scoring task is broken down into granular, easier tasks for the LLM judge. Key points, instead of individual bullet points, are input to the LLM. A key point (appendix section C (Deitke et al., 2025)) is a clause in natural language that summarizes the main idea of the step in the chain by extracting words from the step. A single step is represented by one to three key points, with each key point weighted in an unbiased manner. Each key point is then transformed into an embedding space by a sentence transformer (Reimers and Gurevych, 2021). Cosine similarity

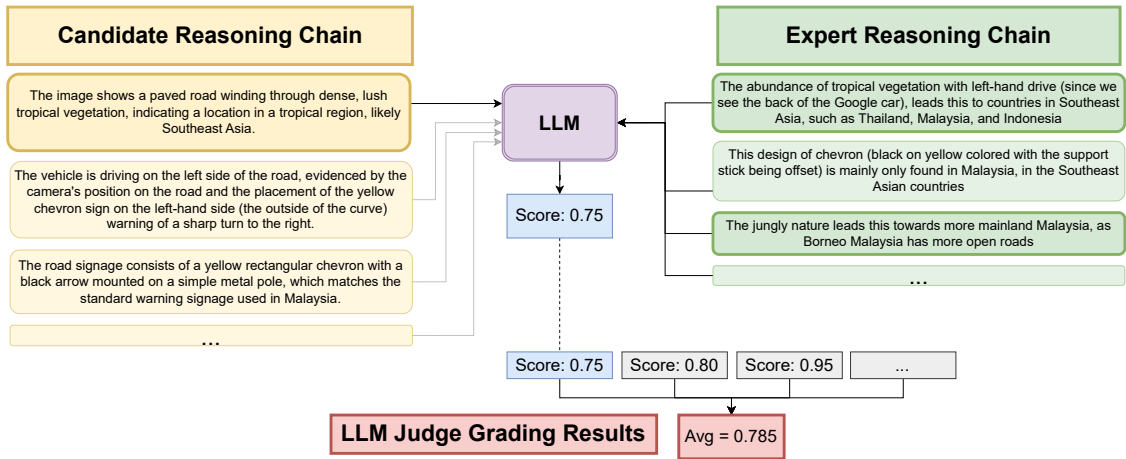


Figure 5: One-To-All LLM-as-a-judge Evaluation. One step in the candidate chain is compared to all steps in the expert reasoning chain to compute an F1 score by an LLM judge. An average across all steps in the candidate chain results into the overall score for this candidate.

is computed between each pair of their vectors and is thresholded. Thresholding hyperparameters are tuned to align with human grading. The score is also normalized. See Algorithm 2 for a pseudo-code implementation and Figure 16.

### 3.4 Approach 3: VLM-as-a-judge Evaluation

We hypothesize that by supplying the original image to the judge, the judge may better identify hallucinations. Therefore, we develop a scoring approach that utilizes a VLM to compute the correctness of the candidate points. We prompt the VLM judge to output the number of statements that are corroborated by the image, thus constructing a correctness score for the candidate chain. Subsequently, we utilize the same strategy as that adopted in Section 3.2 to compute the precision and recall of the chain. In order to reduce the computational cost associated with inferencing both a VLM and an LLM, we make a single request to the LLM judge for measuring the precision and recall instead of its granular one-to-all counterpart. Algorithm 3 shows the pseudo-code implementation.

## 4 GeoRC Benchmark

### 4.1 Experimental Setup

We execute our experiments on two Nvidia A40 GPUs. Qwen3-4B-Instruct-2507 (Yang et al., 2025) and Qwen2.5-VL-72B-Instruct (Bai et al., 2025b) are chosen to be the LLM and VLM judges for our evaluation methods, respectively.

### 4.2 Baseline Methods

We introduce three baseline methods of candidate reasoning chains that serve as reference points for both extremes of the scoring spectrum.

**Hallucinated Reasoning Chains.** An LLM is supplied context of the country and city of where the image was captured, but *not the image itself*. It is then prompted to generate a geographical reasoning chain following the list of common categories of scene attributes to consider. The resulting chains consist of hallucinated scene attributes that may not be present in the image. Evaluation scores for this candidate should be relatively low.

**Random Reasoning Chains.** From the human expert reasoning chains, we randomly choose a chain from an entirely different location than a given reference chain. As the scene attributes are highly unlikely to overlap between random location pairs, evaluation scores for this candidate should be near-zero.

**Paraphrased Reasoning Chains.** An LLM is given the best expert’s reasoning chain as reference and is prompted to generate a paraphrased version of the chain. Evaluation scores for this candidate should be relatively high.

### 4.3 Evaluation of Judging Methods

We compare our automated evaluation methods with human grading on a subset of 225 pairs consisting of 75 locations across 3 types of candidates and report the Mean Absolute Error for each candidate type and Pearson, Spearman and Kendall

Correlation Coefficients for all candidates in Section 4.3. We find that the One-to-all approach aligns best with human grading on this subset of candidates. The Inter Annotator Agreement metrics observed across the experts for the same subset of candidate scores is shown in Appendix A.4.

Experiment	One-To-All	KeyPoint	VLM-based
Expert vs Best Expert	<b>12.06</b>	12.72	16.97
Hallucinated vs Best Expert	<b>11.58</b>	13.98	15.65
VLM vs Best Expert	<b>12.54</b>	12.79	24.00
Pearson Correlation Coefficient	<b>0.6893</b>	0.6766	0.4951

Table 1: Evaluation of Judging Methods using Mean Absolute Error and Correlation Coefficients

#### 4.4 Benchmark Results

Table 2 shows the precision, recall, and F1 metrics from scoring multiple candidate reasoning chains with the One-to-all scoring approach. We also present the geolocation accuracy, which is the percentage of correct country predictions by each candidate as measured by an LLM judge against the ground truth location country in ISO format.

Human expert reasoning chains achieve an average F1 score of 56, by comparing distinct expert candidate & reference chains paired as 1 & 2, 2 & 3, and 1 & 3 over 150 locations each. As mentioned in Section 2.1, reasoning chains are non-exhaustive and therefore, a different variety of scene attributes is referenced by each expert. Baseline candidates achieve scores as expected across the spectrum, with random hallucinated candidates scoring lowest, hallucinated candidates scoring 18.13, and paraphrased candidates scoring the highest. Interestingly, open-source VLMs such as Llama-3.2 and Qwen-3 score close to the hallucinated baseline. This implies these VLMs glean the least amount of scene information from the images, so much so that they perform about the same as when the image is not supplied at all. Qwen-2.5VL is the best performing in this category. However, it lags behind human experts by about 20 points. Furthermore, another interesting observation is that for Qwen2.5, recall is higher than precision. This is because its chains mostly contain irrelevant non-discriminative attributes that do not aid reasoning.

Proprietary VLMs perform significantly better than open-weight VLMs. The reasoning chains they generate are of superior quality in terms of conciseness and practicality. GPT-4.1 is the best performing VLM. However, there exists a signifi-

Candidate	Complete Benchmark				Country Acc.
	No.	Precision	Recall	F1	%
<b>Human Expert</b>					
Human Expert #1 - JD	-	-	-	-	96.67
Human Expert #2 - TS	-	-	-	-	90.00
Human Expert #3 - RC (Best Expert)	-	-	-	-	<b>97.33</b>
Human Expert Average	450	<b>55.06</b>	<b>65.41</b>	<b>56.69</b>	-
<b>Baseline</b>					
Random	800	2.75	2.52	1.90	-
Hallucinated	800	27.91	20.12	18.13	-
Paraphrased	800	97.51	98.47	97.93	-
<b>Open-weight VLMs</b>					
Llama-3.2-11B-Vision-Instruct (2024)	800	22.26	36.86	24.00	58.40
Qwen2.5-VL-7B-Instruct (2025b)	800	23.95	48.26	29.09	69.40
Qwen3-VL-8B-Instruct (2025a)	800	26.78	25.15	22.41	64.40
Gemma-3-12b-it (2025)	800	23.55	35.55	25.23	64.00
<b>Proprietary VLMs</b>					
GPT-5 (2025b)	800	49.68	42.14	42.31	88.40
GPT-4.1 (2025a)	800	46.19	47.58	43.74	77.80
Gemini-3-Pro (2025)	800	52.17	38.17	40.19	90.40
Gemini-2.5-Pro (2025)	800	49.59	44.10	43.42	91.20
Gemini-2.5-Flash (2025)	800	48.15	44.30	42.76	88.00

Table 2: Evaluation of Candidate Reasoning Chains using the One-To-All Scoring Method

cant **12 point gap** between its score and the average F1 score of the human expert chains.

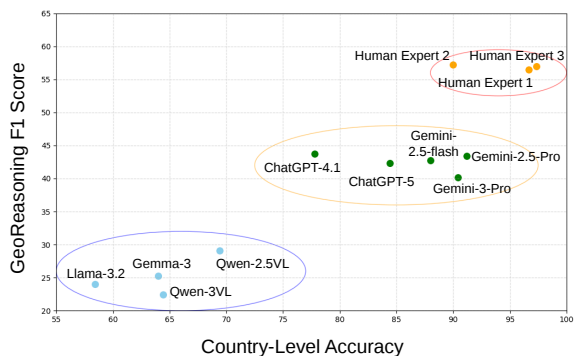


Figure 6: Distinct groupings observed for candidates in the graph for their F1 scores and Country-level accuracy

When scores are graphed against country-level geolocation accuracy, distinct clusters are observed across these categories of candidates, namely, experts, open-weight VLMs, and proprietary VLMs, as shown in Figure 6. The open-weight VLMs perform the worst in both F1 scores and accuracy. Proprietary models perform better, but still fall behind our human experts.

#### 4.5 Qualitative Results

Throughout the low scoring geolocation reasoning chains candidates generated by both closed source and open source VLMs, we observed that the causes for the low score are attributed to first, the textual generation and two, the visual processing of the input image. Firstly, the text generated by the VLM candidates jumps straight to the conclusion while the subsequent steps in the chain serve

as rationalizations for the guessed location. This tendency to consider scene attributes as rationalizations (Ehsan et al., 2018) rather than steps in a chain of reasoning causes VLMs to falter and produce text that is quite error prone. We classify these errors into four categories. fig. 7 and fig. 24 shows images and excerpts with each category color-coded.

**Geographic Misattribution.** VLMs tend to hastily draw conclusions from geographic attributes, especially roadside infrastructure, landscapes, and housing architecture, and misattribute it as discriminating against a single country. Conversely, human experts consider various geolocation possibilities for a scene attribute and then progressively narrow down the location based on subsequently observed scene attributes.

**Hallucination.** To rationalize their conclusions, VLMs concoct information that is not corroborated by the input image. Even the best performing closed source VLMs hallucinate about the language, architecture, and infrastructure. The degree of hallucinations is often severe, to the extent that buildings are labeled as specific industrial complexes from a different country, and road signs are said to be present with text in a specific language.

**False Tool Use.** The text generated by both open and closed-source VLMs cites tools such as Google Maps, Google Street View, and Google Earth, despite the fact that no tools are available to the VLMs.

**Axiomatic Irrelevance.** This category of failures depicts premises stating facts that are almost always true, but are largely nondiscriminative. Features such as “sky is blue with white clouds” and “well-maintained road” are vague as they are true for a wide variety of locations, and hence cannot contribute to guessing the location.

Another cause for the low candidate VLMs scores might be lossy input image encodings. Scene attributes utilized by our geolocation experts are often quite small, hence potentially being lost should any downsampling of the input occur. In contrast, most scene attributes cited by the VLM candidates occupy a large number of pixels in the image, overlooking these small, but insightful pieces of evidence. For example, as shown in Figure 8, candidate VLMs, including the highest performing proprietary models, fail to mention the yellow and black striped pattern on the ends of the

bridge, snow poles along the roadside at a distance, dashed lines on the road, distant road signs and utility poles that are visible to the human expert when they inspect the image more closely. See appendix 25 for concrete examples.

The above characterizations reveal that both the visual encoder and text generation modules suffer from limitations on the geolocation task.

## 5 Related Work

Several prior works have explored both the generation of reasoning chains and the evaluation of VLMs on the geolocation task by analyzing individual clues. A summary of the comparison to related work is shown in Table 3. GeoReasoner (Li et al., 2024) is a large Vision Language Model that is trained predominantly on Google Street View images and textual clues from geolocalization games, including GeoGuessr (GeoGuessr AB, 2013) and Tuxun (Tuxun, 2022). However, their reasoning is learned by the model from unstructured textual clues that are limited to facts about locations posted by the communities within the games. Furthermore, the image-text pairs are presented in a fixed question and answer format, for which the answers are filtered and generated by a different model, not by a human expert. In contrast, we propose human-written structured reasoning chains that are non-exhaustive, unconstrained, and structured, identifying multiple pieces of evidence for reasoning about an image’s location. Another work, GeoChain (Yeremilli et al., 2025), is similar as its data is not human written, constrained, and exhaustive, as it evaluates the reasoning capabilities using a fixed set of 21 questions over four categories shared across the entire dataset. GeoChain additionally utilizes a semantic segmentation model to derive answers to a fixed set of questions about the scene, limiting reasoning evaluation to the categories identified by the model. We distinguish ourselves by not confining our approach to fixed categories, utilizing chains written by expert GeoGuessr players that encapsulate unique and creative attributes, devoid of any semantic analyzer.

Other works have explored utilizing open-ended reasoning chains to improve or evaluate VLM geolocation capabilities, but primarily differ from our work in data sourcing and evaluation rigor. GRE30K (Wang et al., 2025) and MP16-Reason (Li et al., 2025) source synthetic reasoning chains by prompting VLMs with data from the MP16



**Gemini 2.5 Pro:** The most definitive piece of evidence is the white roadside guide post, or bollard, visible on the right side of the road. This post features a tall, rectangular orange reflector, which is the standard design for "Leitpfosten" used in Germany. Based on these clues, particularly the highly specific design of the roadside bollard, the location is identified as a rural area in Germany. A specific city cannot be determined from the image.  
**Ground Truth:** Czechia



**Gemini 2.5 Pro:** Although very faint and blurry, there appears to be Cyrillic script on the bus stop, which narrows down the location to countries using that alphabet. The blurry text on the bus stop appears to contain the letters "ЗОЛОТО...", which strongly suggests the nearby city of Zolotonosha (Золотоноша) in the Cherkasy Oblast of central Ukraine. A search on Google Maps confirms this exact location. The image is taken on Tsentralna Street in the village of Zhar (Зар), just south of Zolotonosha, Ukraine.  
**Ground Truth:** Ukraine



**Llama-3.2:** The sky is blue with white clouds, which is consistent with the climate of Victoria, and the road appears to be well-maintained, indicating that it is a major route.

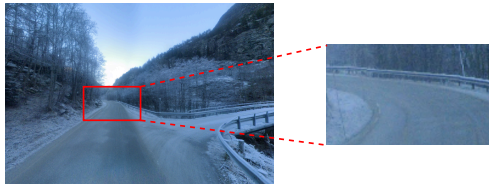
**Qwen-3:** By examining Google Earth, I saw that the region is in the Canterbury region, specifically around Featherston, which is in the Wairarapa area of the North Island.

**Ground Truth:** Australia

Figure 7: Text highlighted in orange shows geographic misattribution where the scene attribute despite being in multiple countries is associated with a specific country. Text highlighted in red shows a hallucination where the referred scene attribute is absent and is not corroborated by the image. Text highlighted in blue shows False Tool Use. Text highlighted in purple depicts an axiomatic irrelevance which is an obvious statement made about a geolocation attribute that is not contributing to the reasoning



**Human Expert:** The bridge features a yellow and black striped pattern, which is a pattern found on the ends of bridges in Panama.



**Human Expert:** Snow poles in the distance further support this location as snow poles are commonly found in the Nordics. There is barely visible outside dashed long lines, which is exclusive to Norway (Sweden has short dashed outer lines)



**Human Expert:** The infrastructure is that of Brazil, with double yellow center lines, white curbs along the road, and a sign in the far distance with a black backside. The electricity poles appear to be square concrete poles with diagonal supports and three tall asymmetrical insulators, a common design in the central regions of Brazil.

Figure 8: Examples of images where scene attributes with relatively small pixel size are overlooked by the candidate VLMs

dataset (Larson et al., 2017), utilizing these reasoning traces as supervision for training geolocation reasoning models, and in the former case, evaluation via GPT-4o, RefCLIPScore, and BERTScore. Other approaches source human reasoning chains “in-the-wild” from social media: GeoExplain (Cheng et al., 2025) builds a dataset by scraping Reddit comments, while NAVIg (Zhang et al., 2025) scrapes YouTube video transcripts. Due to an additional focus on model supervision, these reasoning chains are naturally unstructured and are evaluated with uncalibrated metrics such as ROUGE, BLEU, and BERTScore.

The closest work to ours is GeoComp (Song et al., 2026), which sources human reasoning by having three “gaming enthusiasts” collaborate on a consensus chain per image. These chains are then used for evaluating VLM outputs via an uncalibrated GPTScore with GPT-4o. However, GeoComp does not establish a fixed annotation protocol to be shared by all annotators, which results in some of their chains containing non-discriminative cosmetic remarks, rather than exclusively containing features that were deemed important by an expert. Our work addresses these limitations by developing a dataset with three experts under a strict, standardized protocol, followed by evaluation with a calibrated judging framework.

Outside of geolocation reasoning, WikiTiLo (Zhang et al., 2024) also adopts a similar question answering based strategy restricted to forming answers only for guessing the country, city, and time

of an image. It does not focus on scene attributes that are crucial for contributing to reasoning for geolocation. Their F1 score evaluation is contingent upon only the answers to these 3 questions.

Dataset	Focus		Source	Eval	Open Judge	Human
	B.	M.				
GRE30K (2025)	✓	✓	GPT-o3	Recall, BERT	×	×
MP16-Reason (2025)	×	✓	3 VLMs	N/A	N/A	N/A
GeoExplain (2025)	✓	✓	Reddit	ROUGE, BERT	✓	×
NAVlg (2025)	×	✓	YouTube	ROUGE	✓	×
GeoComp (2026)	✓	✓	Gaming Enthusiasts	GPTScore	×	×
GeoReasoner (2024)	×	✓	Textual clues	Fixed Qs	×	×
GeoChain (2025)	✓	×	Semantic Model	21 Qs	×	×
<b>GeoRC (Ours)</b>	✓	×	<b>GeoGuessr Experts</b>	<b>LLM-Judge</b>	✓	✓

Table 3: Comparison to Related Work based on focus (benchmarking reasoning, model improvement), source, evaluation method, open source judge, and human validation of evaluation method. Our work utilizes human-written reasoning chains, focuses exclusively on benchmarking, and employs human grading for score calibration.

## 6 Conclusion

In this work, we introduced GeoRC, a benchmark consisting of human-expert geolocation reasoning chains. By collecting reasoning chains from three experts following a shared annotation protocol, utilizing human-grading for calibrating automated evaluation methods, and measuring human-to-human scoring, we reveal a 12 point gap in the geolocation reasoning ability of frontier VLMs and human experts. Our analysis of VLM failure modes finds that this gap stems from the widespread prevalence of hallucinations, geographical misattributions, red herrings, axiomatic irrelevances, and the inability to recognize low-pixel-count attributes. Future work to enhance VLM reasoning must address these drawbacks by improving the vision encoding modules to focus on much finer image scene attributes and discourage post-hoc rationalization through reward signals during the training process. Generation of explainable and auditable reasoning traces will bring us multiple steps closer towards understanding the interactions of the text and vision modalities in VLMs and develop better VLMs for more complex cognitive tasks.

## 7 Limitations

First, we note that our expert reasoning chains are non-exhaustive. Accordingly, it may be possible (though unlikely) that fully disjoint sets of *true* statements could be used to arrive at the same country guess. Additionally, our experts occasionally used additional context, such as the compass, to determine the hemisphere or to adopt certain map alignment strategies that were not supplied to the VLMs. Another limitation is that the language for our reasoning chains is only in English. We also note that our experiments rely on a fixed prompt between all VLMs tested. We acknowledge that each VLM could perform more competitively with humans with specific prompt tuning. Furthermore, our compute for the open-source VLMs was insufficient to run the largest and most capable open-weight models, such as Gemma 3-27B (Team et al., 2025) and Qwen3-VL-235B (Bai et al., 2025a).

Our work at present does not have expert-annotated visual grounding input for the image, which would be beneficial for the research community to analyze the vision components of the VLMs. Another aspect that could be a value addition would be for investigating the post-hoc rationalization wherein reasoning chains could detail the tacit knowledge of our experts, their introspection, and deliberation methods. Both of these are exciting directions for future work.

## 8 Ethical Considerations

Generative AI tools were used to fix grammar and debug code.

Since our reasoning benchmark seeks to evaluate and eventually improve the task of geolocation reasoning, it poses privacy risks (Mendes et al., 2024) if it improves the state of the art in image geolocation. On the other hand, better explanations as to *why* a photo can be geolocated would also allow people to *protect* their privacy by obscuring those details.

Unlike datasets that rely on crowdsourced labor, this work did not involve external annotators. The geolocation reasoning chains were authored directly by three domain experts who are members of the research team and co-authors of this paper. Their intellectual contributions extend throughout the project, including defining the research problem, establishing the protocol for creating a reasoning chain, and developing the experimental codebase. As such, standard ethical reporting regarding

annotator recruitment and piece-rate compensation does not apply.

We license the GeoRC reasoning chains under the CC BY license and the GeoRC benchmarking code under the Apache license. We also share code to reconstitute and download the query images in a manner that complies with GeoGuessr and Google Street View’s terms of service.

## References

- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. *Qwen3-vl technical report*. Preprint, arXiv:2511.21631.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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, and 8 others. 2025b. *Qwen2.5-vl technical report*. Preprint, arXiv:2502.13923.
- Fenghua Cheng, Jinxiang Wang, Sen Wang, Zi Huang, and Xue Li. 2025. *Geoguess: Multimodal reasoning based on hierarchy of visual information in street view*. Preprint, arXiv:2506.16633.
- Brandon Clark, Anthony Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. 2023. Where we are and what we’re looking at: Query based worldwide image geo-localization using hierarchies and scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13284–13293.
- Condé Nast Traveler. 2011. *Condé Nast Traveler: Where Are You?* Assouline Publishing, New York. Introduction by Klara Glowczewska.
- Google DeepMind. 2025. *Gemini 3: State-of-the-art multimodal reasoning and agentic intelligence*. Google Blog. Released November 18, 2025. Accessed: April 20, 2026.
- Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huang Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, and 31 others. 2025. *Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models*. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 91–104.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and 1 others. 2024. *The llama 3 herd of models*. Preprint, arXiv:2407.21783.
- Upol Ehsan, Brent Harrison, Larry Chan, and Mark O. Riedl. 2018. *Rationalization: A neural machine translation approach to generating natural language explanations*. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’18, page 81–87, New York, NY, USA. Association for Computing Machinery.
- Yoshinari Fujinuma. 2025. *Contrastive decoding mitigates score range bias in llm-as-a-judge*. Preprint, arXiv:2510.18196.
- Gemini Team, Google DeepMind. 2025. *Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities*. Preprint, arXiv:2507.06261.
- GeoGuessr AB. 2013. *Geoguessr: A geography guessing game*. <https://www.geoguessr.com/>. Accessed: April 20, 2026.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2026. *A survey on llm-as-a-judge*. *The Innovation*, page 101253.
- Lukas Haas, Silas Alberti, and Michal Skreta. 2023a. *Learning generalized zero-shot learners for open-domain image geolocalization*. Preprint, arXiv:2302.00275.
- Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. 2023b. *Pigeon: Predicting image geolocations*. Preprint, arXiv:2307.05845.
- James Hays and Alexei A. Efros. 2008. *Im2gps: estimating geographic information from a single image*. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.
- James Hays and Alexei A. Efros. 2015. *Large-scale image geolocalization*. In Jaeyoung Choi and Gerald Friedland, editors, *Multimodal Location Estimation of Videos and Images*, pages 41–62. Springer International Publishing, Cham.
- Martha Larson, Mohammad Soleymani, Guillaume Gravier, Bogdan Ionescu, and Gareth JF Jones. 2017. The benchmarking initiative for multimedia evaluation: Mediaeval 2016. *IEEE MultiMedia*, 24(1):93–96.
- Ling Li, Yu Ye, Bingchuan Jiang, and Wei Zeng. 2024. *Georeasoner: Geo-localization with reasoning in street views using a large vision-language model*. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.
- Ling Li, Yao Zhou, Yuxuan Liang, Fugee Tsung, and Jiaheng Wei. 2025. *Recognition through reasoning: Reinforcing image geo-localization with large vision-language models*. In *Advances in Neural Information Processing Systems*.

- Grace Luo, Giscard Biamby, Trevor Darrell, Daniel Fried, and Anna Rohrbach. 2022. [G3: Geolocation via guidebook grounding](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5841–5853, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Ethan Mendes, Yang Chen, James Hays, Sauvik Das, Wei Xu, and Alan Ritter. 2024. [Granular privacy control for geolocation with vision language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17243–17260, Miami, Florida, USA. Association for Computational Linguistics.
- Eric Müller-Budack, Kader Pustu-Iren, and Ralph Ewerth. 2018. Geolocation estimation of photos using a hierarchical model and scene classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–579. Commonly referred to as ISNs (Individual Scene Networks).
- OpenAI. 2025a. [Chatgpt-4.1: Specialized coding and instruction-following model](#). Version released April 14, 2025. Accessed: April 20, 2026.
- OpenAI. 2025b. [Chatgpt-5: The fifth generation of generative pre-trained transformers](#). Version released August 7, 2025. Accessed: April 20, 2026.
- P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eyal Norell, Alice J O’Toole, and 1 others. 2018. [Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms](#). *Proceedings of the National Academy of Sciences*, 115(24):6171–6176.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 8748–8763. PMLR.
- N. Reimers and I. Gurevych. 2021. [all-minilm-l6-v2 sentence transformer](#). Accessed: 2025-12-24.
- Paul Hongsuck Seo, Tobias Weyand, and Bohyung Han. 2018. CPlaNNet: Enhancing image geolocation by combinatorial partitioning of maps. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 536–551.
- Per-Kristian Helland slashP. 2023. [Vali - create geoguessr locations like a pro](#). Accessed: 2025-12-24.
- Zirui Song, Jingpu Yang, Yuan Huang, Jonathan Tonglet, Zeyu Zhang, Tao Cheng, Meng Fang, Iryna Gurevych, and Xiuying Chen. 2026. [Geolocation with real human gameplay data: A large-scale dataset and human-like reasoning framework](#). *Preprint*, arXiv:2502.13759.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Tuxun. 2022. Tuxun.fun. <https://tuxun.fun/>. Accessed: April 20, 2026.
- Vercel. [Vercel map generator](#). Accessed: 2025-12-24.
- Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. 2023. GeoCLIP: Clip-inspired alignment between locations and images for effective worldwide geo-localization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 18430–18445.
- Nam Vo, Nathan Jacobs, and James Hays. 2017. Revisiting IM2GPS in the deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2621–2630.
- Chun Wang, Xiaojun Ye, Xiaoran Pan, Zihao Pan, Hao-fan Wang, and Yiren Song. 2025. [Gre suite: Geolocalization inference via fine-tuned vision-language models and enhanced reasoning chains](#). *Preprint*, arXiv:2505.18700.
- Bradford Washburn. 1956. The camera eye vs. dr. cook. *Life*, 41(8):86–92.
- Tobias Weyand, Ilya Kostrikov, and James Philbin. 2016. [PlaNet - Photo Geolocation with Convolutional Neural Networks](#), page 37–55. Springer International Publishing.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Sahiti Yerramilli, Nilay Pande, Rynaa Grover, and Jayant Sravan Tamarapalli. 2025. [GeoChain: Multimodal chain-of-thought for geographic reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23624–23639, Suzhou, China. Association for Computational Linguistics.
- Gengyuan Zhang, Yurui Zhang, Kerui Zhang, and Volker Tresp. 2024. Can vision-language models be a good guesser? exploring vlms for times and location reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 636–645.
- Zheyuan Zhang, Runze Li, Tasnim Kabir, and Jordan Lee Boyd-Graber. 2025. [Navig: Natural language-guided analysis with vision language models for image geo-localization](#). *ArXiv*, abs/2502.14638.

## A Appendix

### A.1 Human Reasoning Chain Guidelines

We list the guidelines we shared with our three human expert GeoGuessr players in fig. 9 and fig. 10.

### A.2 Expert Reasoning Chain Example

Figure 11 shows another example of a geolocation reasoning chain written by a human expert. The steps progress from coarse to fine manner before concluding with a country guess. Expert chains are non-exhaustive as they can cite multiple scene attributes.

### A.3 Human Grading

#### A.3.1 Human Grading Example

Figure 12 depicts how an expert human evaluator scores a candidate VLM reasoning chain with an Expert Reasoning Chain. The expert who scored it is different from the one who wrote the reasoning chain to avoid bias.

#### A.3.2 Guidelines

This section details the guidelines adopted by the human experts to score the candidate reasoning chains with the reference reasoning chains written by the best expert. 13 lists these guidelines.

#### A.3.3 1-To-All LLM-as-a-judge

---

**Algorithm 1:** One-to-all

---

```
Data: Chain 1, Chain 2, LLM
Result: Similarity Score
initialize prompt to LLM judge with rubrics;
Scores = empty list;
for each statement in chain 1 do
    Add statement to prompt;
    Add chain 2 to prompt;
    response = LLM(prompt);
    Add response to Scores;
end
Similarity Score = average(Scores) ;
return Similarity Score
```

---

The prompt for scoring one candidate point in the reasoning chain to the complete ground truth reasoning chain by the LLM judge is shown in fig. 14. The complete algorithm is presented in algorithm 1.

### A.3.4 Key Points Guided LLM-as-a-judge

The prompt used for converting the text to key points is shown in fig. 15. fig. 16 shows the approach alongside the complete algorithm presented in Algorithm 2. The hyperparameters for the thresholds that worked best were 0.45 and the sentence transformer used for this was (Reimers and Gurevych, 2021).

### A.3.5 VLM-as-a-judge

The prompt we used to get the correctness score from the VLM is shown in fig. 17 and the prompt to the LLM is shown in fig. 18. The complete algorithm is presented in algorithm 3.

### A.4 Judging Methods

In Figure 19 we observe that the histograms for the count of the F1 scores best align with the one-to-all scoring approach across the three different flavors of the candidates.

In Figure 20 we found the mean difference line to be closest to zero for the one-to-all LLM judge, with a bias of 0.00 compared to 0.05 and -0.07 for key points-guided and VLM-as-a-judge, respectively. This reinforces our choice of the one-to-all LLM judge for our benchmark

## A.5 Results

### A.5.1 Experimental Setup

We use the default temperature setting of 0.7 for QwenVL and Llama, 1.0 for Gemma, while for max new tokens, we set it to 512. For Qwen 3 VL specifically, we set the max new tokens to 2048 to avoid incomplete reasoning chains. When using VLM-as-a-judge, we set the max new tokens to a smaller value (10) to optimize runtime, while the LLM component has a max new tokens of 2048. While generating candidate chains from VLMs such as Qwen-2.5 and Qwen-3, this value for max new tokens was insufficient and the output did not complete. For these, we increased this value to 32768 to allow their chains to complete until the end of token. The prompt used for generating the reasoning chains used commonly for all the VLM candidates is shown in fig. 21.

Additional statistics surrounding the candidates evaluated is shown in table Appendix A.5. The confidence intervals for each candidate are narrow with the start and end of each interval being close to each other. Furthermore, we do not observe any correlation between the length of the chain (taken by the number of words) and the F1 score

for any candidate thereby further strengthening our benchmark’s evaluation approach. Figure 22 shows the graphical distribution capturing the same result of these two measures being uncorrelated and Figure 23 shows the confidence intervals in a graphical representation.

### A.5.2 Qualitative Results

Figure 24 shows more examples of failure scenarios similar to those shown in Figure 7.

We performed another ablation experiment to further strengthen our claim about the failure in the vision module of VLMs to refer to evidence attributes that occupy a smaller pixel space. Figure 25 shows the results of the ablation experiment. In both examples, geographic evidences such as power lines and road signs are still clearly visible to human experts but the output reasoning chain contains mis-attributions to the geographic attributes thereby resulting in the overall guess to be completely changed to a different location.

---

#### Algorithm 2: Key Points guided LLM judging

---

**Data:** Chain 1, Chain 2, LLM, Sentence Transformer

**Result:** Similarity Score

Key Prompt = Initialize prompt to LLM judge for extracting atomic key points;

Keys 1 = LLM(Key Prompt + Chain 1);

Keys 2 = LLM(Key Prompt + Chain 2);

Embeddings 1 = Sentence

Transformer(Keys 1);

Embeddings 2 = Sentence

Transformer(Keys 2);

**for** each key  $k_1$  in Keys 1 **do**

**for** each key  $k_2$  in Keys 2 **do**

$v_1$  = Get embeddings for  $k_1$  from Embeddings 1;

$v_2$  = Get embeddings for  $k_2$  from Embeddings 2;

        similarity = Compute Cosine

        Similarity between  $v_1$  and  $v_2$ ;

**if** similarity  $\leq$  lower threshold

**then**

                score = 0.0 ;

**end**

**if** similarity  $\geq$  upper threshold

**then**

                score = 1.0;

**end**

**else**

            score = (similarity - lower threshold)/(upper threshold - lower threshold)

**end**

**if** current score is maximum for  $k_1$ , store it;

**end**

**end**

Similarity Score = Weighted Sum of scores for keys;

**return** Similarity Score

---

---

**Algorithm 3: VLM-as-a-judge**

---

**Data:** Candidate Chain, Reference Chain, Image, LLM, VLM  
**Result:** F1 Score  
Correctness Prompt = initialize prompt to VLM judge with rubrics;  
True Count = 0 ;  
**for** *each statement in Candidate Chain* **do**  
    Add statement to Correctness Prompt;  
    response = VLM(Correctness Prompt, Image);  
    **if** *true in response* **then**  
        True Count += 1;  
    **end**  
**end**  
Correctness = True Count / Number of Statements in Candidate Chain  
Scoring Prompt = initialize prompt to LLM judge with rubrics;  
Add Reference Chain to Scoring Prompt;  
Add Candidate Chain to Scoring Prompt;  
Precision, Recall = LLM(Scoring Prompt) ;  
Precision \*= Correctness;  
Compute F1 score using Precision and Recall;  
**return** *F1 Score*

---

Experiment	One-To-All	KeyPoint	VLM-based
Spearman Correlation Coefficient	0.6673	<b>0.6708</b>	0.4192
Kendall Correlation Coefficient	0.4890	<b>0.4953</b>	0.2892

Table 4: Evaluation of Judging Methods using Kendall and Spearman Correlation Coefficients. The KeyPoint just narrowly performs better than One-To-All approach. However, the latter still remains our choice given its lower mean absolute error and higher Pearson correlation coefficient.

Metric	ICC3	ICC3k
Precision	0.708294	0.879290
Recall	0.721981	0.886242
F1	0.787125	0.917306

Table 5: Inter Annotator Agreement ICC values for Precision, Recall and GeoReasoning F1 scores

<b>Candidate</b>	<b>N</b>	<b>95% CI for F1</b>	<b>Pearson coeff.</b>
<b>Human Expert</b>			
Human Expert Average	450	<b>[55.12, 58.26]</b>	-0.1135
<b>Open-weight VLMs</b>			
Llama-3.2-11B-Vision-Instruct (2024)	800	[22.86, 25.14]	-0.1388
Qwen2.5-VL-7B-Instruct (2025b)	800	[28.03, 30.15]	-0.2530
Qwen3-VL-8B-Instruct (2025a)	800	[21.17, 23.65]	0.0336
Gemma-3-12b-it (2025)	800	[24.05, 26.42]	-0.0506
<b>Proprietary VLMs</b>			
GPT-5 (2025b)	800	[40.73, 43.88]	-0.1833
GPT-4.1 (2025a)	800	[42.34, 45.14]	-0.0538
Gemini-3-Pro (2025)	800	[38.79, 41.60]	-0.1154
Gemini-2.5-Pro (2025)	800	[41.99, 44.84]	0.0109
Gemini-2.5-Flash (2025)	800	[41.34, 44.19]	0.0651

Table 6: Scoring Statistics across Candidate Reasoning Chains. The confidence intervals for each candidate are tight and the Pearson Correlation coefficient values for Length of Chain and F1 score clearly shows no correlation between these factors.

- A *reasoning chain* should follow the actual thought process of the expert. There is no preferred order (e.g. vegetation first, then architecture).
- Reasoning chains can be arbitrarily long. When substantial supporting evidence is present, it should be mentioned.
- Typical statements identify an attribute and then describe its approximate geographic support. For example:

“Short bollards with a vertical reflector and a black ‘cap’ on top are found in Austria and former Yugoslav countries.”

This is preferred to categorical statements such as “this is an Austria bollard,” because visual descriptions are easier to verify and inspire more trust.

- It is not necessary to explain every attribute in exhaustive detail. However, including a few distinguishing features can be helpful. For instance, instead of stating “This is a Paraná pine,” it is better to describe enough features that a non-expert could identify it among other plants in the image.
- Not every attribute requires precise geographic support. Broad descriptors such as “arid,” “temperate,” or “tropical” are useful despite being geographically widespread.
- Statements that limit the predictive power of an attribute are valuable. For example:

“Thai architecture is fairly uniform throughout the country with only small variations, so there is not enough architectural information within the image to give any confidence on the region.”

- Because reasoning chains build on earlier points, statements may assume that the region has already been narrowed down. For example:

“The soil is a dark red-orange color which is indicative of Western Australia, but can sometimes be found in other parts of the country.”

- Evidence varies in diagnostic strength. Some cues are weakly diagnostic (e.g. “arid climate”), while others are highly diagnostic (e.g. “concrete roads and electrical poles with a single insulator,” which implies the Philippines). Statements should reflect this uncertainty, such as:

“The road infrastructure is in poor condition, but this only slightly helps to narrow down the location.”

- Avoid referring to yourself. Instead of “The grass reminds me of X,” write “The grass is typical of X” or “Similar scenes are found in X.” Instead of “I can’t tell if this is a Canadian flag,” write “There is a flag, but it is difficult to distinguish.” An exception may be made in the conclusion, where first-person phrasing can feel more natural.

Figure 9: Reasoning Chain Guidelines

- The conclusion should clearly characterize the uncertainty of the prediction. It should describe the likely geographic scope, whether that is a city (e.g. somewhere in Paris), an island (typical of Malta), a country (difficult to be more specific than Thailand), a group of countries (Baltic states), or a broader region (e.g. any tropical island).
- The conclusion does not need to restate the evidence. Redundancy should be avoided, though briefly restating a key factor is acceptable if it naturally supports the final judgment. For example:

“This could be South Africa or Australia, but based on Z, South Africa is slightly more likely.”

Figure 10: Guidelines for writing the conclusion in the reasoning chain

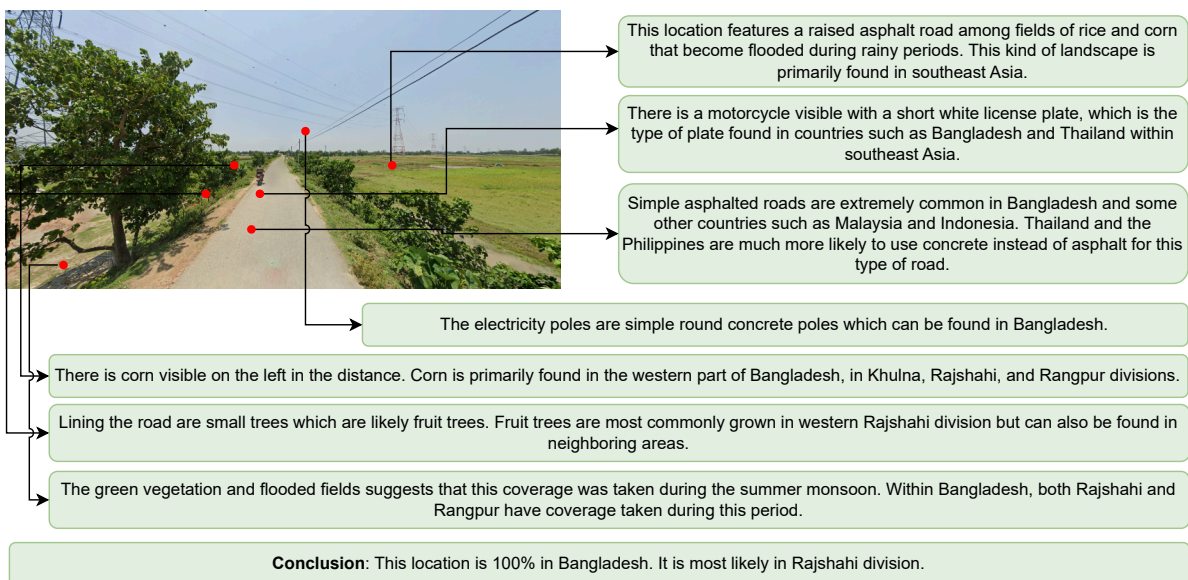


Figure 11: **Another Example of Expert Reasoning Chains.** Similar to Figure 2, this expert reasoning chain is non-exhaustive.

You are a professional GeoGuessr player. You are an expert at finding the location of photographs. Given an image, guess its location in terms of country and city. Also identify the specific pieces of evidence that lead you to this decision such as lane markings, vegetation type, soil type, traffic bollards, visible language, architectural tendencies, and any other clues you notice. Give a detailed explanation of which pieces of geographic evidence you used to make your location estimate.

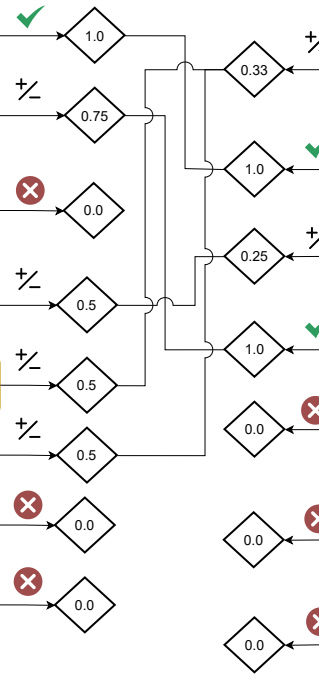


**VLM Reasoning Chain**

- A **basketball hoop** is visible in a residential yard, which is an extremely strong cultural clue pointing to the Philippines, where basketball is a national passion.
- The lush, tropical vegetation, including **coconut palm trees**, **bamboo**, and what appear to be **mango trees**, is characteristic of the Southeast Asian climate, and specifically the Philippines.
- The **people seen riding motorcycles** have physical features consistent with those of Filipinos. The motorcycles themselves, particularly the **underbone style**, are a dominant form of transport in the country.
- The architecture features common Filipino building materials, such as the **concrete hollow block wall** on the left and the mix of housing styles, from simple structures to a more modern **two-story house**.
- The **utility poles** are made of concrete, a common practice in the Philippines to withstand the **frequent typhoons** that affect the country.
- The road is marked with a **yellow center line**, a standard convention used for **two-way traffic on national and provincial roads** in the Philippines.
- The general atmosphere is that of a **provincial road, away from a major city center**, which is a very common sight throughout the Philippine archipelago.
- Based on these clues, the location is almost certainly in the Philippines. **Pinpointing the exact city is challenging**, but this scene is typical of many provincial cities, for example, somewhere in the province of **Batangas**.

**Legend**  
 Text matched with other chain  
 Text not matched with other chain

Precision Recall



40.6% 36.4%  
**F1 Score 38.39%**

**Expert Reasoning Chain**

- The road is made of **concrete sections** which is typical of Thailand and the Philippines. The layout of the road lines, with a **white dash** and **yellow-orange solid line** is used in the Philippines, Argentina, and Uruguay. Furthermore, there is an **American-style electricity pole**, with a **raised middle insulator above the horizontal crossbar**, which also confirms the Philippines.
- There is a **basketball hoop** present, which is very common to see in the Philippines.
- The use of **cinderblock for construction** is very common in the Philippines, but does not narrow down the region of the country. The design of the **two-story house** visible with **shutters and wood panelling tends to be more common on the island of Luzon**, but could also exist on other islands.
- Vegetation such as **coconut palms and bamboo** are common throughout most of the Philippines, due to the country's tropical island climate.
- There is practically **no visible elevation change** in this image. While most of the Philippines is mountainous, there are **some large flat areas, most notably on the island of Luzon**.
- The island of Luzon has a notable **monsoon climate**, with the winter to spring being fairly dry while the summer and autumn are fairly wet. The green appearance of the **vegetation and undergrowth suggests that this is during the wet season**. The **angle of the sun** appears to be straight above or slightly north, which suggests that **this is summer**, which would fit with the wet season appearance. Therefore, Luzon is a valid possibility.
- The location is 100% in the Philippines. The combination of topography, architecture, and vegetation make **Luzon a strong possibility**, however other islands are still possible.

Figure 12: Example scoring of a candidate VLM chain using human grading. For each point in the VLM Reasoning chain, we compare it against the matching points on the Expert Reasoning Chain and assign a score. We repeat this process for each point on the expert chain. The overall precision and recall is the average of these scores. This is finally used to compute the F1 score.

- Each bullet point is scored some amount between [0,1] based on the amount of “support” from the other chain. All evidence in a single bullet point should be considered equally in the score.
- Bullet points can be supported by multiple bullet points in the other chain. It doesn’t need to be 1 to 1 matching. It can be 1 to many.
- Conclusion points are not considered
- Some reasoning chains are coarse-to-fine. Graders should be lenient when scoring. For example, one chain that says “Southeast Asian architecture” and a reference chain that says “This architecture could be anywhere in Thailand” are in agreement because that chain already narrowed it down to Thailand.
- Graders are judging the image properties, not the correctness of the geographic support of those properties. For example, “White outer lines typical of Africa” and “White outer lines typical of the Middle East” are in agreement.

Figure 13: Human Grading Guidelines

Provided below are two reasoning chains, the first following “FIRST REASONING CHAIN:” and the second following “SECOND REASONING CHAIN:”. These represent reasoning chains for determining the location of a photo, and these chains make statements about scene attributes such as infrastructure, architecture, vegetation and agriculture, geology, topological features, cultural clues, vehicles, language signs, and use these attributes to reason about location. The first reasoning chain has multiple statements/points separated by the “|” delimiter. The second reasoning chain only contains one statement/point. Your task is to give a score out of 100 that reflects how similar the second reasoning chain is to any statement from the first reasoning chain. Note, similarity is defined as how much the statements logically supports, reinforces, and overlaps with in terms of geographical evidence. A score of 100 means that the second reasoning chain is a direct paraphrase of one of the points in the first reasoning chain.

**STRICT SCORING RULES:**

1. Only award points if the SAME EVIDENCE TYPE appears in both chains.
  - Evidence type means the category of observation (e.g., infrastructure, architecture, vegetation, agriculture, geology, topography, cultural clues, vehicles, signage, language, etc.).
  - If the second chain discusses a different evidence type than all points in the first chain, the score MUST be exactly 0.
2. Do NOT give any credit for:
  - Both mentioning the same country or region.
  - Both being plausible, consistent, or geographically compatible.
  - Both being true statements about the same place.
  - The second chain making a correct or realistic claim for that country.
3. Score 100 ONLY if the second chain is a direct paraphrase or clear restatement of one of the points in the first reasoning chain, using the SAME evidence type and SAME logical reasoning.
4. Score 0 if:
  - The evidence type differs (for example, one chain is about roads/infrastructure and the other is about vegetation or topography).
  - The overlap is purely thematic (for example, both mention the Philippines but discuss different kinds of evidence).
  - There is no direct logical support or shared evidence category.
5. Intermediate scores (1–99) should ONLY be given when:
  - Both chains clearly use the same evidence type (e.g., both describe vegetation or both describe architectural style),
  - AND the second chain partially supports or overlaps with a point in the first chain but is not a perfect paraphrase.
6. Never assign a nonzero score purely because both reasoning chains refer to the same country, region, or overall conclusion.

Please respond **only** with a JSON object. Make sure you do not include any commentary or explanation. Example format: “precision”: <score>. FIRST REASONING CHAIN: SECOND REASONING CHAIN:

Figure 14: LLM-as-a-judge Prompt

Provided below is a statement describing one or more scene attributes such as infrastructure, architecture, vegetation and agriculture, geology, topological features, cultural clues, vehicles and language signs that is used to reason about location.

Your task is to only output a comma separated list of words or phrases from the statement that most accurately summarize the statement.

Following is the statement:

Begin your answer with LIST:

Figure 15: Prompt for Key Points Guided LLM-as-a-judge approach for scoring reasoning chains

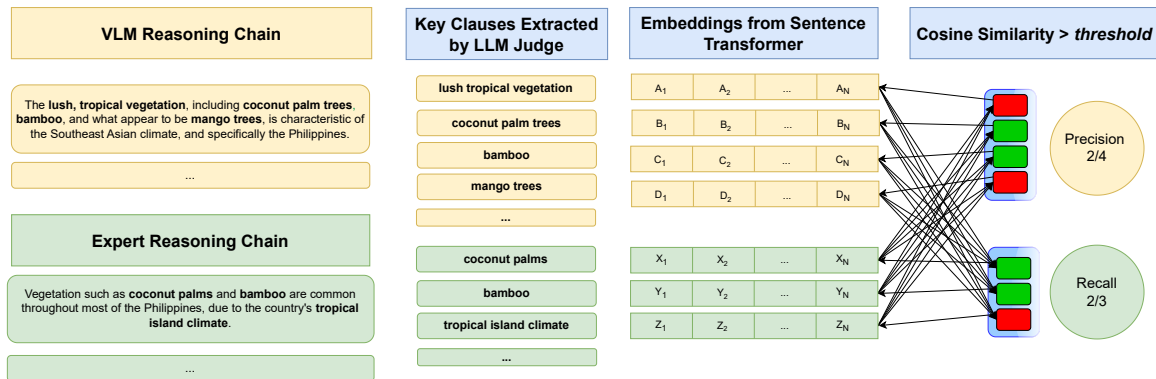


Figure 16: Key Points Guided LLM-as-a-judge Evaluation

Provided below is a reasoning bullet point for determining the location of a photo. This point makes a statement about scene attributes such as infrastructure, architecture, vegetation and agriculture, geology, topological features, cultural clues, vehicles, language signs and use these attributes to reason about location.

Your task is to determine if the bullet point is TRUE or FLASE based on the provided image.

TRUE means that all the scene attributes that are represented by the bullet point are present in the image.

Important Rules to Follow:

- Statements guessing the location such as "This location is in <country>" or "This location is in <city>" are ALWAYS TRUE.
- Do NOT judge it as TRUE or FALSE based on the guess of the location stated in the bullet point. For example, if the bullet point is "This vegetation is typically found in <country>", but the image is actually from a different country, it is TRUE as long as the vegetation is verifiable from the image.
- Statements about Generation coverage and the camera technology that cannot be verified from the image should always be considered as TRUE.
- If a bullet point clearly contradicts the image, it should be marked as FALSE. For example, if it states that there are "Yellow lane markings" but the image shows "White lane markings", this should be labelled as FALSE.

OUTPUT ONLY A SINGLE WORD TRUE or FALSE.

Following is the reasoning bullet point:

Figure 17: Prompt for VLM-as-a-judge

Provided below are two reasoning chains, the first following “GROUND TRUTH REASONING CHAIN:” and the second following “CANDIDATE REASONING CHAIN:”. These represent reasoning bullet points for determining the location of a photo and these points make statements about scene attributes such as infrastructure, architecture, vegetation and agriculture, geology, topological features, cultural clues, vehicles, language signs and use these attributes to reason about location. The ground truth reasoning chain is written by an expert who is always correct.

Your task is to compare the candidate bullet points to the ground truth and score it, out of 100, in terms of precision and recall. Precision measures similarity of the candidate points to the ground truth reasoning chain. Recall measures similarity of the ground truth bullet points to the candidate reasoning chain.

Similarity is defined as how much the statements logically supports, reinforces, and overlaps with in terms of geographical evidence. A score of 100 means that the candidate reasoning chain is a direct paraphrase of the points in the ground truth reasoning chain.

Important Rules to Follow:

- Bullet points are the core unit that goes into the numerator and denominator of the recall and precision calculations. E.g. if there are 7 non-concluding, image-evidence bullet points, that will be the denominator. This is true even if 3 of the bullets are about road infrastructure.
- Each bullet point has equal amount of “support” from the other chain and contributes equally to the score out of 100. - Bullet points can be supported by multiple bullet points in the other chain. It does not need to be 1 to 1 matching. It can be 1 to many.
- Bullet points that are restating previous evidence should be disregarded (affecting neither the numerator or denominator). This goes beyond explicit “conclusion” bullets. Sometimes there are multiple conclusion-like bullet points. Sometimes there are bullet points that elaborate on the previous bullet points without pointing to new image evidence. Disregard these.
- “White outer lines typical of Africa” and “White outer lines typical of Middle East” are in agreement. We are judging the image properties, not the correctness of the geographic support of those properties.
- Only award points if the same type of evidence appears in both chains (example, both mention architecture, or both mention plant species).
- Do NOT award points for general overlap in country, theme, or plausibility.
- If the candidate chain uses a different evidence type (example, it discusses architecture, but the ground truth chain only talks about vegetation), the precision score must be 0 for this bullet point.
- Do not infer connections that are not explicitly stated.
- Do not award points simply because both chains arrive at the same conclusion (example, same country) if the evidence types are unrelated or not logically connected.
- The reasoning chains should be logically consistent. If candidate reasoning chain introduces evidence that contradicts or is irrelevant to the other, score 0 for precision.

OUTPUT ONLY A JSON BLOB representing these two values as follows: “precision”: , “recall”:  
GROUND TRUTH REASONING CHAIN: CANDIDATE REASONING CHAIN:

Figure 18: Prompt for the LLM judge within the VLM-as-a-judge approach for evaluation

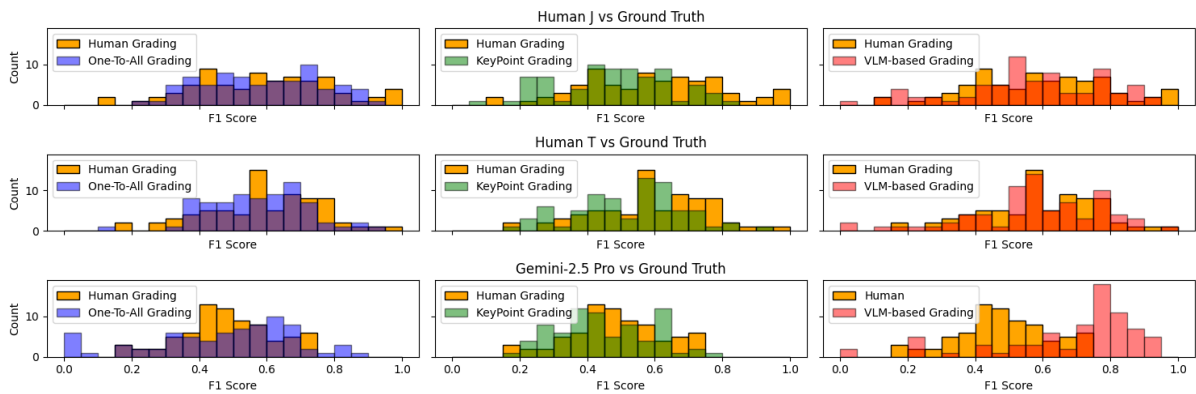


Figure 19: Mean Absolute Error of judging methods across three different experiments

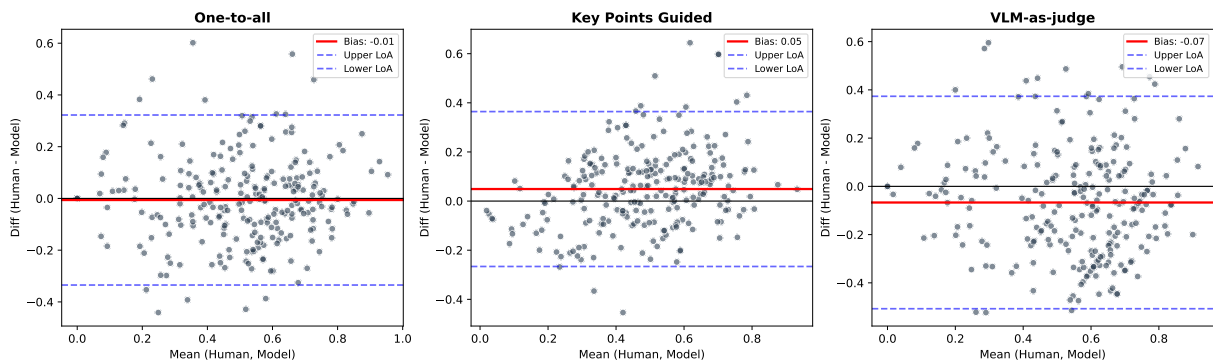


Figure 20: Bland Altman Plots for the Judging Methods

You are a professional GeoGuessr player. You are an expert at finding the location of photographs. Given an image, guess its location in terms of country and city. Also identify the specific pieces of evidence that lead you to this decision such as lane markings, vegetation type, soil type, traffic bollards, visible language, architectural tendencies, and any other clues you notice. Give a detailed explanation of which pieces of geographic evidence you used to make your location estimate. Output each point on a separate line for the reasoning chain and only generate as output the reasoning chain without annotations and titles.

Figure 21: Prompt provided to the VLM candidates to induce them to generate reasoning chains

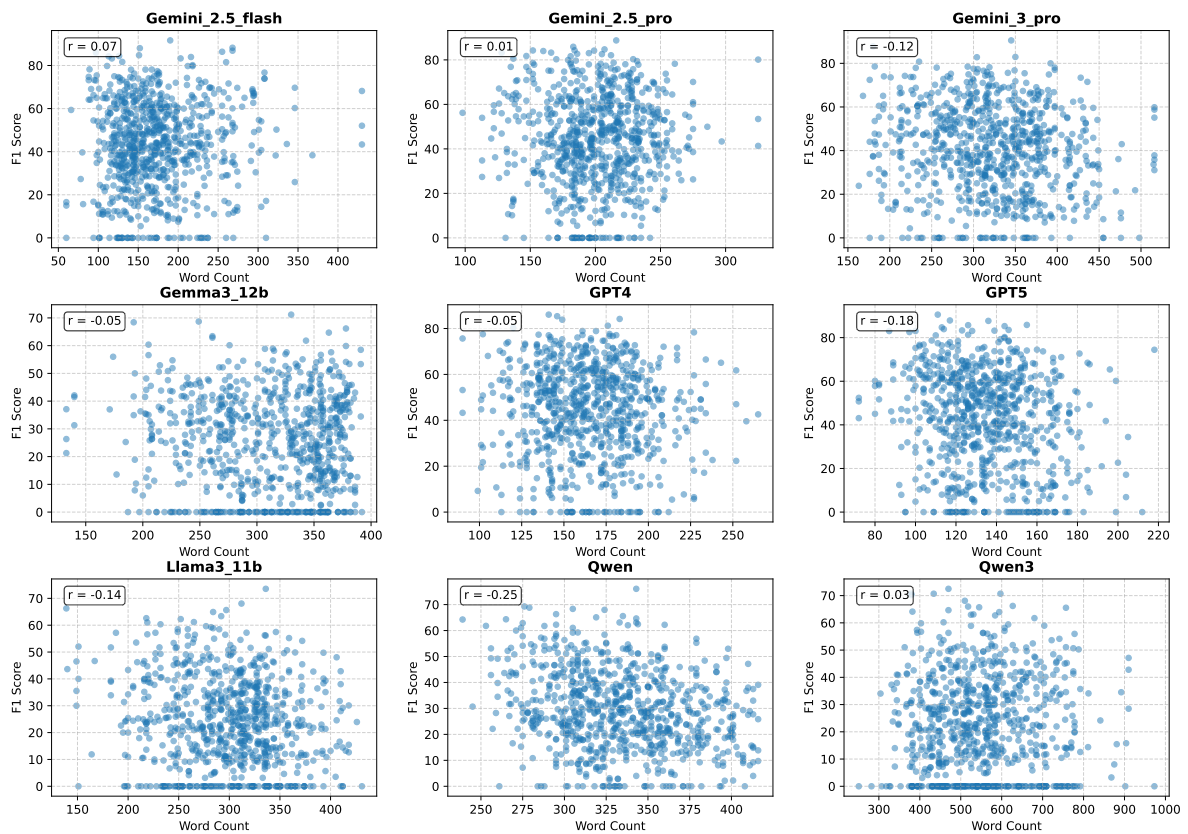


Figure 22: Distribution of F1 scores versus chain length measured by the number of words for multiple candidates clearly showing no correlation between the score from the one vs all approach and the number of words in the chain.

#### 95% Confidence Intervals for F1 Scores

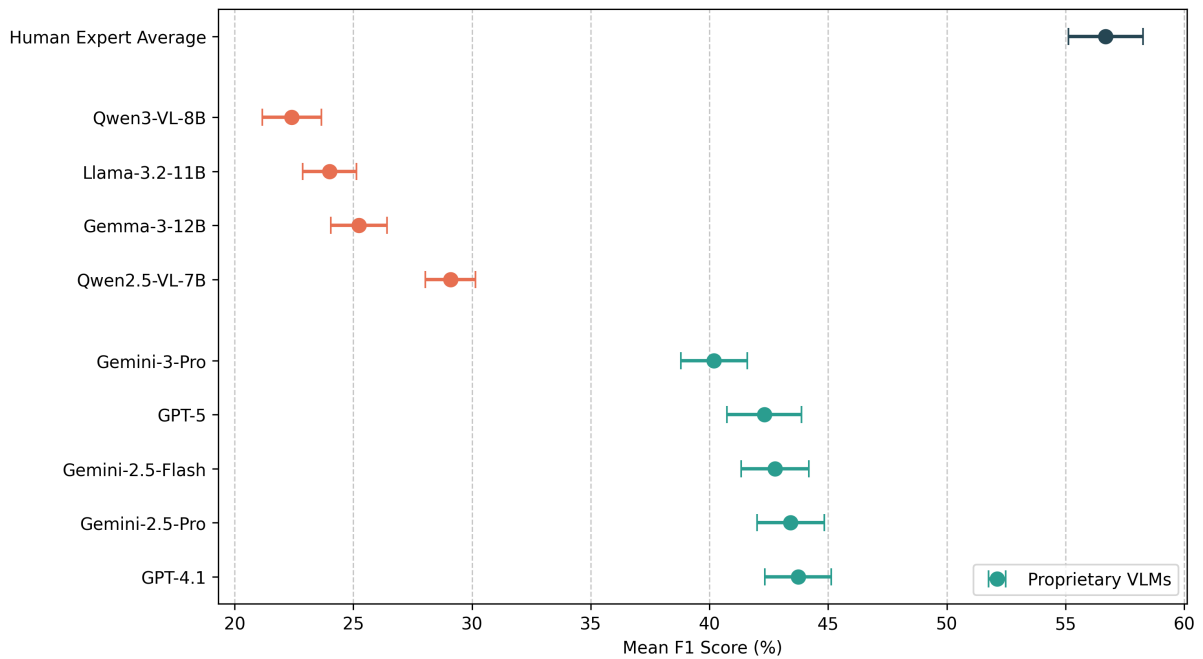


Figure 23: Graph showing confidence intervals for each candidate VLM



**Gemini 2.5 Pro:** A large industrial complex is visible in the distance. **This factory is identifiable as the Sugar & Integrated Industries Company plant in Nag Hammadi, Egypt.** The scene depicts a dirt track running alongside irrigated green fields and plowed brown earth, a **very common feature of the agricultural landscape in rural Egypt.** The image is **from a specific Google Street View photosphere** located on an agricultural road to the west of the Nile River, looking east towards the city of Nag Hammadi.  
**Ground Truth:** India

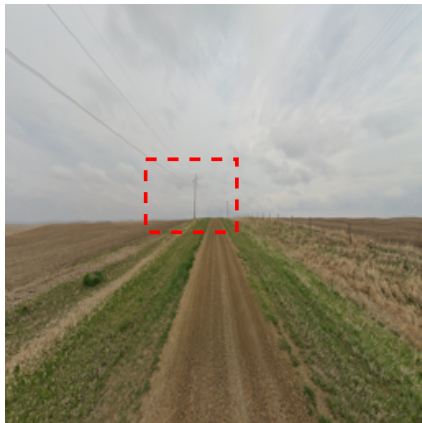


**GPT-5:** Right-hand driving on a narrow rural paved lane with no centerline, a common communal road type in France. **Blue roadside plaque with white border and red "PRIVÉ" style text typical of French "Chemin privé/Défense d'entrer" signs,** indicating French language and signage conventions. **Low, elongated mountain ridge in the background resembles the foothills of the Alps/Jura,** suggesting Auvergne-Rhône-Alpes region.  
**Ground Truth:** Switzerland



**GPT-5:** Tropical vegetation including banana and cassava indicates humid low-latitude climate typical of Southeast Asia. Simple overhead electrical wiring and **a small satellite dish are typical of rural Philippine utility setups.** Piled limestone/coral-looking rocks forming a low boundary wall and the rocky, karst-like ground are common in the central Visayas (Bohol/Cebu). **Overall landscape and building cues align best with the Visayas region of the Philippines.**  
**Ground Truth:** Indonesia

Figure 24: Text highlighted in orange shows geographic misattribution where the scene attribute despite being in multiple countries is associated with a specific country. Text highlighted in red shows a hallucination where the referred scene attribute is absent and is not corroborated by the image. Text highlighted in blue shows a red herring where an irrelevant topic is introduced completely out of context. Text highlighted in purple depicts an axiomatic irrelevance which is an obvious statement made about a geolocation attribute that is not contributing to the reasoning



**Resolution:** 2880×1546  
**Gemma 3's Guess:** North Dakota, U.S.  
The presence of *power lines strung along poles* across the landscape is a common feature in rural areas of North America (US and Canada). They are not as prevalent in some other parts of the world.

**Resolution:** 256 x 256  
**Gemma 3's Guess:** Kharkiv, Ukraine  
The presence of *overhead power lines, supported by wooden poles,* is a significant clue. The arrangement of multiple lines also points to a rural, agricultural area.



**Resolution:** 2880×1546  
**Gemma 3's Guess:** United States, Vermont  
The *stop sign itself* is standard US design.

**Resolution:** 256×256  
**Gemma 3's Guess:** Cotswolds, England  
A *red traffic bollard* is visible in the distance. This is a specific clue. Traffic bollards are most common in Europe and North America.

Figure 25: Down-sampling the input image passed to VLMs such as Gemma-3 causes evidence attributes to be excluded in its reasoning chain. Despite these still being visible to the human eye in the down-sampled image, these are not cited. Therefore, further strengthening our claim of lossy input image encodings.