# Benchmarking Multimodal LLMs on Closed-World Image Geolocation

Maria Di Gisi[1,2], Giuseppe Fenza[2], Mariacristina Gallo[2,*] and Maria Palomba[2]

[1]IMT School for Advanced Studies Lucca, 55100 Lucca, Italy
[2]Department of Management and Innovation Systems, University of Salerno, 84084 Fisciano (SA), Italy

**Abstract**

Geolocating an image solely from its visual content is a critical capability for multiple cybersecurity and OSINT applications, yet current multimodal Large Language Models (LLMs) are rarely evaluated on closed-world data (i.e., images that do not appear in any public dataset). This work presents a systematic benchmark of nine state-of-the-art multimodal LLMs (including five versions of Gemini, two of LLaMA, Qwen 3, and Gemma 3) on a closed-world geolocation task, using four key prompting strategies: Zero-Shot, Chain-of-Thought (CoT), Zero-Shot CoT, and ReAct. Results reveal a complex relationship between prompting and model architecture. While standard CoT often degrades performance due to noisy output, the Zero-Shot CoT strategy achieves peak accuracy by enforcing an internalized reasoning process while suppressing textual output. This demonstrates that performance degradation in standard CoT stems from output generation rather than the reasoning mechanism itself. Top-performing models, Gemini 3 Pro and Gemini 2.5 Pro, achieved a peak of 66% exact-location accuracy. These findings underscore both the potential and current limitations of multimodal LLMs for real-world geolocation, highlighting their applicability in content verification tasks (e.g.,mis- and dis- information detection) as well as in Open Source Intelligence (OSINT) contexts.

**Keywords**

Large Language Models (LLMs), Geolocation, OSINT

## 1. Introduction

The ability to determine the geographic location of an image solely from visual content has become increasingly relevant in multiple domains, including cybersecurity, open-source intelligence (OSINT), digital forensics, and threat verification. Automated geolocation can support situation awareness in crisis events, assist analysts in validating user-generated content, and help detect mis- or disinformation campaigns involving repurposed or manipulated imagery.

The rapid emergence of multimodal Large Language Models (LLMs) has renewed interest in image geolocation. These models combine high-capacity visual encoders with advanced language-based reasoning, raising the question of whether they can infer location cues from previously unseen images. This work performs a systematic evaluation of state-of-the-art multimodal LLMs on a closed-world geolocation task (i.e., using images that do not appear in any public dataset and were never published online). This setup eliminates the risk of unintentional memorization, resulting in a more accurate assessment of the models' ability to perform visual geospatial inference.

The proposed benchmark includes five versions of Gemini (3.0 Pro Preview, 2.5 Flash, 2.5 Pro, 2.0 Flash Lite, and 2.0 Flash), two versions of LLaMA (4 Scout and 4 Maverick), Qwen 3, and Gemma 3. It also compares four prompting strategies (Zero-Shot, Chain-of-Thought (CoT), Zero-Shot CoT, and ReAct) to understand how structured reasoning impacts performance. The results reveal substantial variability between the models and the prompting methods.

Joint National Conference on Cybersecurity (ITASEC & SERICS 2026), February 09-13, 2026, Cagliari, IT
*Corresponding author.
maria.digisi@imtlucca.it (M. Di Gisi); gfenza@unisa.it (G. Fenza); mgallo@unisa.it (M. Gallo); m.palomba21@studenti.unisa.it (M. Palomba)
0009-0003-5434-5426 (M. Di Gisi); 0000-0002-4736-0113 (G. Fenza); 0000-0002-5474-2697 (M. Gallo)

## 2. Related Works

Recent work has explored the geolocation capabilities of Large Language Models (LLMs) and Vision-Language Models (VLMs), as well as frameworks designed to systematically evaluate them across diverse input modalities. A growing body of research focuses on establishing standardized benchmarks. Chen et al. [1] introduce IMAGEO-Bench, a comprehensive framework that assesses models along four axes (accuracy, distance error, geospatial bias, and reasoning quality), offering a structured protocol for consistent comparison. Liu et al. [2] conduct a systematic evaluation of LLaVA and GPT-4o in zero-shot and few-shot settings and propose ETHAN, a Chain-of-Thought–based framework that significantly improves geolocation accuracy over direct inference. Zhang et al. [3] examine multiple multimodal foundation models on a global city localization dataset, finding that Gemini-2.5-Pro achieves the strongest performance, with nearly 49% of predictions falling within 1 km. Complementing these static datasets, GeoArena [4] offers a dynamic, user-driven platform that ranks model predictions through human voting, emphasizing privacy, transparency, and real-world evaluation.

Several studies specifically investigate geolocation from images. Wang et al. [5] construct a large-scale Street View corpus spanning multiple countries and perform both training-free and fine-tuned evaluations on a wide range of open- and closed-source multimodal models. Jay et al. [6] build a geographically balanced benchmark from Google Street View to evaluate single-image geolocation, demonstrating that modern foundation VLMs can achieve median errors below 300 km, with agentic variants that leverage external tools achieving an additional 30% reduction. Img2Loc [7] proposes a novel formulation of image geolocation as a text-generation task, combining CLIP-based retrieval with large multimodal models such as GPT-4V and LLaVA to directly generate coordinates without training.

Beyond general-purpose image geolocation, Yin et al. [8] introduce an LLM-enhanced disaster geolocation approach that fuses explicit and implicit spatial cues from multimodal data. Their system integrates LLM reasoning with external map services to produce geographically grounded predictions in crisis scenarios, highlighting the utility of LLM-based geolocation in time-sensitive real-world applications.

Building on these efforts, this work benchmarks a diverse set of models across closed-image datasets and varying prompt types. Its goal is to systematically analyze the reasoning patterns underlying LLM-based geolocation and to understand how different prompting strategies influence model predictions, revealing the cognitive and visual cues exploited by these systems.

## 3. Methodology

This section describes the methodological framework adopted to evaluate the geolocation capabilities of several state-of-the-art Large Language Models (LLMs) when provided with a single input photograph. The experimental setup is outlined, including the dataset used, the large language models, and the prompting strategies employed in the study.

### 3.1. Dataset

The dataset used in this study consists of 100 privately collected photographs. All images were taken directly by the authors or by individuals within their personal network. The dataset comprises images taken from multiple European countries (Austria, Italy, Sweden, Denmark, Switzerland, Germany, Greece, the United Kingdom, Spain, Luxembourg, France, Bulgaria, and Romania) and a smaller subset from North America, specifically Mexico. The photographs encompass a diverse range of visual environments, including both rural and urban areas, with and without identifiable landmarks. Natural landscapes and rural areas encompass coastal regions, rural environments, mountains, and vegetation-rich areas, whereas urban scenes are characterized by streets, buildings, and architectural structures.

The dataset is intentionally composed of non-iconic and personally captured scenes, which reduces the likelihood that models rely on memorized internet imagery. This makes the task particularly challenging and suitable for assessing the true visual reasoning capabilities of modern multimodal

LLMs. Each image was annotated with its ground-truth location at the precise place, city, and country level. No metadata (such as EXIF geotags) was used or provided to the models.

### 3.2. Models

Multiple LLM families were evaluated, accessed through different inference providers:

- **Google Gemini** (accessed via Vertex AI API): *Gemini 3 Pro Preview*, *Gemini 2.5 Pro*, *Gemini 2.5 Flash*, *Gemini 2.0 Flash*, *Gemini 2.0 Flash Lite*.
- **LLaMA** (accessed via Groq API[1]): *LLaMA-4 Scout*, *LLaMA-4 Maverick*.
- **Qwen** (served locally through Ollama[2]): *Qwen 3 8B*.
- **Gemma** (served locally through Ollama): *Gemma 3 4B*.

To guarantee consistent comparison across all models, a temperature of 0.1 was set.

### 3.3. Prompting strategies

In this subsection, the considered prompting paradigms are detailed.

**Zero-shot prompting** refers to a setting where the model is asked to perform the geolocation task without receiving examples, reasoning instructions, or multi-step guidance. This configuration evaluates the model's native ability to recognize visual cues such as landmarks, architecture, vegetation, or climate features, without the influence of externally imposed reasoning structures.

Zero-Shot Prompt

You are a simple, precise visual geolocation algorithm. Your only task is to identify the location where the photo was taken. Crucially, your output MUST contain ONLY the location name (Place, City, Region, Country).
What is the location of this image?

**Chain-of-Thought (CoT) prompting** requires the model to articulate a sequence of intermediate reasoning steps before producing the final geolocation prediction. The prompt explicitly instructs the model to describe the visual cues observed in the image, infer climate and cultural indicators, narrow down plausible regions, and only then provide a final location hypothesis. This setup evaluates whether structured, human-like reasoning improves the model's ability to deduce geographical information from visual input.

CoT Prompt

You are a professional visual geolocation analyst. Your goal is to determine where this photo was taken using visual reasoning and geographical deduction.
Think step-by-step:
1. Describe what you see (environment, buildings, vegetation, people, signs, climate).
2. Infer the likely climate zone and region type.
3. Identify cultural or linguistic hints (signs, vehicles, architecture).
4. Analyze any unique geographic or topographic markers.
5. Narrow down possible countries, regions, cities, place.
6. Give your final location hypothesis.

[1] https://groq.com/
[2] https://ollama.com/

**Zero-shot CoT prompting** is introduced by Kojima et al. [9] as the most effective prompt strategy in tasks that involve arithmetic, common sense reasoning, and symbolic reasoning. It refers to a setting that preserves the rigid output-formatting instruction of the Zero-Shot prompt and combines it with the implicit reasoning trigger of Chain-of-Thought. The goal is to force the model to perform a multi-step internal deductive analysis while producing only the final result as output, thereby removing inaccurate intermediate reasoning that could degrade overall accuracy.

Zero-Shot CoT Prompt

You are a simple, precise visual geolocation algorithm. Your only task is to identify the location where the photo was taken. Crucially, your output MUST contain ONLY the location name (Place, City, Region, Country).
What is the location of this image?
Let's think step by step.

**ReAct prompting (Reason + Act)** alternates structured reasoning steps with explicit hypothesis updates. The prompt instructs the model to follow a Thought/Action loop: first, describe the evidence extracted from the image (Thought), then derive its geographical implication (Action). This cycle is repeated multiple times until the model formulates a final location estimate. The ReAct formulation enables the assessment of whether iterative reasoning dynamics improve the model's geolocation performance compared to linear reasoning or unguided inference.

ReAct Prompt

You are a professional visual geolocation analyst. Your task is to infer where this image was taken by alternating between reasoning (Thought) and deduction (Action).
Follow this reasoning loop:
Thought: Describe what you notice in the image (environment, architecture, vegetation, signs, climate, people, etc.).
Action: State the inference you derive from that observation.
Repeat the Thought/Action cycle several times, extracting new clues each time. When you have enough evidence, conclude with your final reasoning and hypothesis.
Use this output format:
Thought 1: ...
Action 1: ...
Thought 2: ...
Action 2: ...
...
Final Reasoning: ...
Final Answer: [Most likely place, city, region, country]

## 4. Experimental Results

This section presents the results of the geolocation experiments using multimodal LLMs. It reports both quantitative metrics and qualitative observations, highlighting differences across models and prompting effective strategies.

### 4.1. Quantitative Results

For each model and prompting strategy, the predicted locations were evaluated against ground-truth labels at the exact position, area, and country level. The main metrics for evaluation were:

- **Exact Position Accuracy**: Correct prediction of the precise location (city or precise place).
- **Area Accuracy**: A prediction falling within 50 km of the ground-truth location. It is computed automatically for each image. It is computed using the geographical (or geodetic) distance, which is the shortest arc length along the Earth's surface. This distance is calculated from geographical coordinates (latitude and longitude) using the Haversine formula [10], which estimates the great-circle distance between two points. The distance was calculated between the ground-truth coordinates and the model's predicted location (converted to latitude/longitude using geocoding when necessary). The geopy[3] Python library was used to calculate both geographical coordinates and distance.
- **Country Accuracy**: Correct identification of the country.

**Table 1**
Geolocation accuracy for all models under the Zero-Shot prompting strategy.

| Model | Version | Exact Position | Area ($\leq$ 50 km) | Country |
|---|---|---|---|---|
| Gemini | 3 Pro | 66% | 78% | 93% |
| Gemini | 2.5 Pro | 63% | 83% | 85% |
| Gemini | 2.5 Flash | 51% | 73% | 83% |
| Gemini | 2.0 Flash Lite | 41% | 66% | 83% |
| Gemini | 2.0 Flash | 39% | 71% | 85% |
| LLaMA | 4 Scout | 27% | 41% | 66% |
| LLaMA | 4 Maverick | 27% | 39% | 71% |
| Qwen | 3 8B | 22% | 39% | 51% |
| Gemma | 3 4B | 22% | 34% | 76% |

**Table 2**
Geolocation accuracy for all models under the Chain-of-Thought (CoT) prompting strategy.

| Model | Version | Exact Position | Area ($\leq$ 50 km) | Country |
|---|---|---|---|---|
| Gemini | 3 Pro | 63% | 76% | 90% |
| Gemini | 2.5 Pro | 66% | 76% | 85% |
| Gemini | 2.5 Flash | 49% | 66% | 80% |
| Gemini | 2.0 Flash Lite | 37% | 56% | 73% |
| Gemini | 2.0 Flash | 39% | 61% | 78% |
| LLaMA | 4 Scout | 22% | 37% | 63% |
| LLaMA | 4 Maverick | 24% | 41% | 66% |
| Qwen | 3 8B | 27% | 37% | 51% |
| Gemma | 3 4B | 15% | 29% | 49% |

As in Tables 1, 2, 3, 4 Gemini 3 Pro and Gemini 2.5 Pro consistently emerge as the top-performing models. Gemini 3 Pro achieved the highest country Accuracy (93% with Zero-Shot and ReAct) and 95% with Zero-Shot CoT, demonstrating superior global inference. Furthermore, while Gemini 2.5 Pro peaked at 66% exact-position accuracy with CoT, Gemini 3 Pro matched this performance (66%) using both the Zero-Shot and the Zero-Shot CoT strategies. This strongly suggests that for the newest Gemini architecture, the need for structured prompting to unlock maximum performance has been mitigated, allowing both Zero-Shot and the constrained hybrid strategy to deliver top results through efficient native processing. Performance decreases gradually across smaller Gemini versions, with Flash and Flash Lite showing expected degradation due to reduced model size and capability.

[3] https://geopy.io

**Table 3**
Geolocation accuracy for all models under the Zero-Shot CoT prompting strategy.

| Model | Version | Exact Position | Area ($\leq$ 50 km) | Country |
|---|---|---|---|---|
| Gemini | 3 Pro | 66% | 76% | 95% |
| Gemini | 2.5 Pro | 63% | 80% | 85% |
| Gemini | 2.5 Flash | 63% | 76% | 85% |
| Gemini | 2.0 Flash Lite | 51% | 73% | 83% |
| Gemini | 2.0 Flash | 46% | 73% | 83% |
| LLaMA | 4 Scout | 24% | 41% | 68% |
| LLaMA | 4 Maverick | 32% | 49% | 76% |
| Qwen | 3 8B | 27% | 41% | 58% |
| Gemma | 3 4B | 17% | 34% | 58% |

**Table 4**
Geolocation accuracy for all models under the ReAct prompting strategy.

| Model | Version | Exact Position | Area ($\leq$ 50 km) | Country |
|---|---|---|---|---|
| Gemini | 3 Pro | 61% | 76% | 93% |
| Gemini | 2.5 Pro | 63% | 76% | 83% |
| Gemini | 2.5 Flash | 46% | 63% | 78% |
| Gemini | 2.0 Flash Lite | 34% | 58% | 76% |
| Gemini | 2.0 Flash | 39% | 61% | 78% |
| LLaMA | 4 Scout | 27% | 39% | 68% |
| LLaMA | 4 Maverick | 24% | 44% | 71% |
| Qwen | 3 8B | 17% | 37% | 51% |
| Gemma | 3 4B | 12% | 27% | 46% |

Crucially, the introduction of the Zero-Shot CoT strategy identified a significant opportunity to enhance performance in lower-capacity models. Among open-source models, LLaMA-4 Maverick shows the most competitive performance, reaching a new peak of 32% exact-position accuracy with Zero-Shot CoT (compared to 27% with Zero-Shot), along with 49% area accuracy and 76% country accuracy. Qwen 3 8B, similarly, achieved its peak area accuracy of 41% with the Zero-Shot CoT strategy. This result suggest that the internal reasoning mechanism of CoT, when decoupled from the output generation noise, significantly benefits models with more limited multimodal training.

To facilitate a quick assimilation of the quantitative results and the relative impact of the prompting strategies, Figure 1 provides a visual comparison of the Exact Position Accuracy across all evaluated models.

## 4.2. Qualitative Results

To complement the quantitative evaluation, a qualitative analysis was conducted to identify the visual cues used by the models when inferring geographical locations. For a subset of representative images, the visual elements referenced in the models' reasoning traces (e.g., landmarks, architectural patterns, vegetation characteristics, signage, or terrain morphology) were manually annotated. The annotated images in Figures 2,3,4,5 highlight the specific cues that the models explicitly mentioned during inference, illustrating how different prompting strategies influence the attention and interpretative focus of each model.

Since Gemini 2.5 Pro was the top-performing model in exact position identification with reasoning prompts according to the quantitative results, its qualitative reasoning traces are reported to provide a detailed illustration of how the best model utilizes visual information. Zero-shot and Zero-Shot

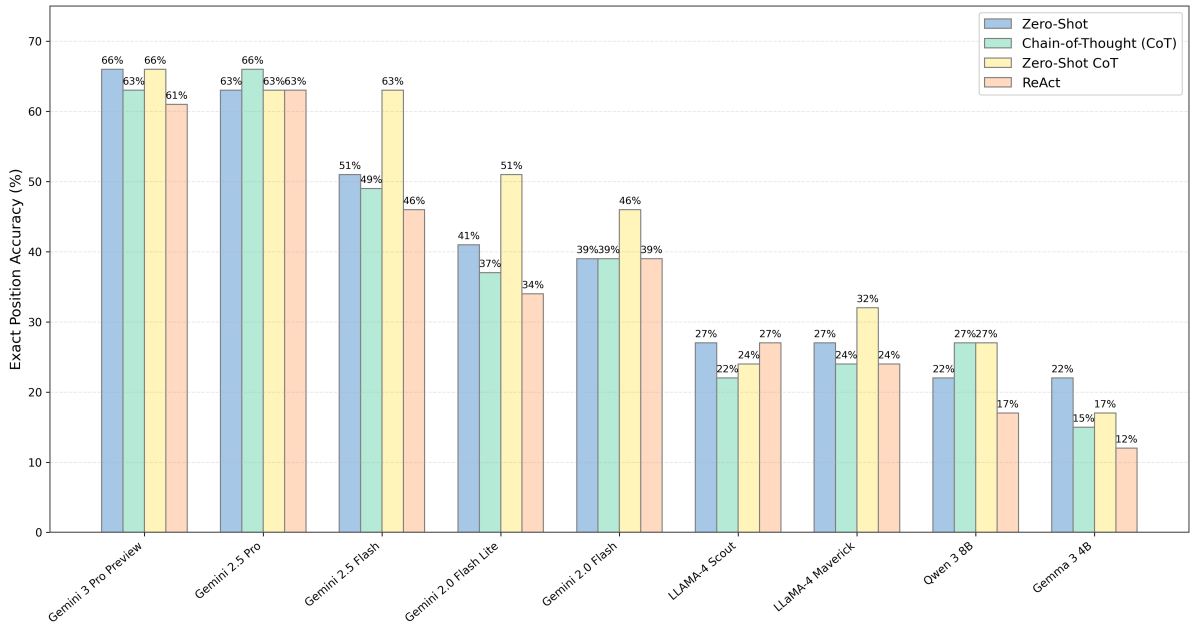


**Figure 1:** Exact Position Accuracy (%) across Models and Prompting Strategies

CoT promptings were excluded from the qualitative analysis because they produces no interpretable reasoning steps. Qualitative comparisons, therefore, focus exclusively on CoT and ReAct, which provide explicit visual cues in their reasoning traces.

(a) Clues with Chain-of-Thought Prompt

(b) Clues with ReAct Prompt

**Figure 2:** Annotated visual cues extracted by Gemini 2.5 pro during the geolocation process for an image captured in Isla Holbox, Quintana Roo, Mexico.

For Figure 2, both prompting strategies identify the scene as a tropical coastal environment, noting the shallow turquoise lagoon, white sandbar, and the seabird on weathered wooden posts. Their reasoning, however, diverges in how they use visual evidence. The CoT prompt focuses on broad environmental cues (water coloration, the Laughing Gull, and the flat coastline), quickly generalizing

to a wide geographic region (Caribbean/Gulf of Mexico) and ultimately selecting Holbox through global associations with sandbars and calm lagoons. By contrast, the ReAct trace integrates the same environmental and faunal details with a more fine-grained inspection of distinctive elements: the milky jade-green water linked to limestone sediment, the specific curvature of the sandbar, and the characteristic wooden posts often photographed in this exact spot.

(a) Clues with Chain-of-Thought Prompt

(b) Clues with ReAct Prompt

**Figure 3:** Annotated visual cues extracted by Gemini 2.5 pro during the geolocation process for an image captured in the Nørrebro district, Copenhagen, Denmark.

As shown in Figure 3, the CoT prompt enables the model to identify key visual and textual clues (the Danish mural "DINE PENGE MIG I RØVEN," the "Nordic Abaya" shop, and the characteristic architecture and streetscape of Copenhagen) and combine them directly to infer the correct location. In contrast, the ReAct trace alternates observations and verification steps: it extracts visual cues (buildings, bus stop, bicycles, trees) and textual elements (the mural, "Nordic Abaya," and the bus stop display "NØRREBROGADE"), then systematically cross-checks each clue with external resources such as maps and Street View. Through this iterative process, ReAct confirms both the street and the precise viewpoint.

Figure 4 shows that with the CoT prompt, the model focused on the Hooded Crow and distinctive lampposts, leading to an incorrect match with Giardini Indro Montanelli due to insufficient attention to pond layout and fences. In contrast, ReAct integrates natural and man-made cues (i.e., vegetation, seasonal indicators, bird species, pond structure, and fencing) and iteratively verifies them against external resources, correctly identifying the location as Parco Sempione, Milan. This example highlights how the two prompting strategies differ in their attention to and integration of visual evidence.

Figure 5 shows that both prompting strategies correctly identify the double-cove formation and the steep Mediterranean terrain leading to the twin beaches. The CoT prompt relies on high-level landscape features to infer Porto Timoni after a broad Mediterranean regional guess. ReAct, instead, combines these global cues with a detailed, evidence-driven analysis of vegetation, shoreline geometry, and viewpoint characteristics.

### 4.3. Error Analysis

To provide an example of geolocation failure, in Figure 6 the CoT prompt leads the model to incorrectly localize the image in Tyre, Lebanon. Non-discriminative visual cues (the flat-roofed, arched building,

(a) Clues with Chain-of-Thought Prompt

(b) Clues with ReAct Prompt

**Figure 4:** Annotated visual cues extracted by Gemini 2.5 pro during the geolocation process for an image captured in Parco Sempione, Milan, Italy.

(a) Clues with Chain-of-Thought Prompt

(b) Clues with ReAct Prompt

**Figure 5:** Annotated visual cues extracted by Gemini 2.5 pro during the geolocation process for an image captured in Porto Timoni, Corfù, Greece.

palm trees, warm sunset lighting, and a dry Mediterranean atmosphere) were over-interpreted as characteristic of the Eastern Mediterranean. Additional elements, like the small harbor with recreational boats and the jet ski marked '1370 YK', and common Mediterranean vehicles (e.g., Peugeot Partner vans), further reinforced the false association with the Al Fanar promontory in Tyre.

Using the ReAct strategy, the model focused on a limited set of cues: the red building on the offshore islet, perceived as a Greek chapel (Chapel of Afentis Christos, Crete), the palm trees indicating a warm

(a) Clues with Chain-of-Thought Prompt

(b) Clues with ReAct Prompt

**Figure 6:** Annotated visual cues extracted by Gemini 2.5 pro during the geolocation process for an image captured in Marzamemi, Sicily, Italy.

Mediterranean climate, and the arrangement of boats and floating docks, linked to Greek tourist marinas. A misinterpreted boat registration number also contributed to the error. Collectively, these observations led ReAct to misidentify the scene as Sarantari Beach near the Creta Maris Beach Resort in Hersonissos, Crete, Greece.

## 5. Discussion and Challenges

The quantitative results highlight a fundamental divergence in how different LLM architectures benefit from explicit reasoning. For the majority of models tested, including smaller Gemini versions and open-source models (LLaMA, Qwen, Gemma), the standard Zero-Shot prompting proved to be the superior or equivalent strategy. This trend supports the hypothesis that the articulation of reasoning in standard Chain-of-Thought (CoT) and ReAct introduces 'Output Noise'. For models with limited parameter counts or training data, forcing multi-step textual output can lead to inaccurate intermediate deductions that ultimately override the model's native, high-capacity visual encoding. The use of the Zero-Shot CoT strategy validated this hypothesis by demonstrating that the benefit of CoT's internal reasoning can be isolated from the output noise. This hybrid prompt, which triggers the internal reasoning process but strictly suppresses the textual output, unlocked significant performance gains for intermediate models, such as LLaMA-4 Maverick reaching a new peak of 32% Exact Position Accuracy.

**Limitations.** A key limitation of this study lies in the size of the evaluation dataset. Because all images originate from private user collections, strict privacy safeguards were necessary during data curation. This resulted in the removal of a substantial number of photographs that contained sensitive personal details. Another limitation concerns in dataset's restricted geographic diversity. Although the collection includes images from multiple regions, it ultimately covers only a limited number of countries. As a result, the benchmark provides only partial geographic coverage and does not support conclusions about model performance at a global scale.

## 6. Conclusion

This work provides a systematic benchmark for image geolocation on an innovative closed-world dataset, a critical requirement for evaluating the true geospatial inference capabilities of Multimodal Large Language Models. This work demonstrates that geolocation accuracy varies significantly across model architectures and prompting strategies. The success of the Zero-Shot CoT strategy supports the hypothesis that performance degradation in standard CoT and ReAct approaches is primarily due to noise introduced during the explicit textual generation of reasoning steps, rather than limitations in the internal reasoning mechanisms themselves. Moreover, the demonstrated ability of top LLMs to rapidly geolocate non-iconic, user-generated content with minimal prompting has direct implications for Open-Source Intelligence (OSINT) and disinformation containment, as it can be leveraged for content verification to counter the spread of manipulated or misleading visual information. Notably, this benchmark establishes a foundation for future work aiming to detect mismatches between images and associated news content, providing a systematic framework to assess and improve the reliability of multimodal verification tools.

## Acknowledgements

This work was partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

## Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly in order to: Grammar and spelling check, Paraphrase, and reword. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

## References

[1] L. Li, R. Yu, Q. Hu, B. Li, M. Deng, Y. Zhou, X. Jia, From pixels to places: A systematic benchmark for evaluating image geolocalization ability in large language models, arXiv preprint arXiv:2508.01608 (2025).

[2] Y. Liu, G. Deng, J. Ding, Y. Li, T. Zhang, W. Sun, Y. Zheng, J. Ge, Mission: Impossible–image-based geolocation with large vision language models, Proceedings on Privacy Enhancing Technologies (2025).

[3] X. Zhang, X. Cheng, Evaluation of geolocation capabilities of multimodal large language models and analysis of associated privacy risks, arXiv preprint arXiv:2506.23481 (2025).

[4] P. Jia, Y. Zhang, X. Zhao, S. Li, Geoarena: An open platform for benchmarking large vision-language models on worldwide image geolocalization, 2025. URL: https://arxiv.org/abs/2509.04334. arXiv:2509.04334.

[5] Z. Wang, D. Xu, R. M. S. Khan, Y. Lin, Z. Fan, X. Zhu, Llmgeo: Benchmarking large language models on image geolocation in-the-wild, 2024. arXiv:2405.20363.

[6] N. Jay, H. M. Nguyen, T. D. Hoang, J. Haimes, Evaluating precise geolocation inference capabilities of vision language models, 2025. URL: https://arxiv.org/abs/2502.14412. arXiv:2502.14412.

[7] Z. Zhou, J. Zhang, Z. Guan, M. Hu, N. Lao, L. Mu, S. Li, G. Mai, Img2loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 2749–2754. URL: https://doi.org/10.1145/3626772.3657673. doi:10.1145/3626772.3657673.

[8] W. Yin, Y. Xue, Z. Liu, H. Li, M. Werner, Llm-enhanced disaster geolocalization using implicit geoinformation from multimodal data: A case study of hurricane harvey, International Journal of Applied Earth Observation and Geoinformation 137 (2025) 104423.

[9] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022.

[10] C. De Maio, G. Fenza, M. Gallo, V. Loia, A. Volpe, A perceived risk index leveraging social media data: assessing severity of fire on microblogging, Cognitive Computation 16 (2024) 2724–2734.