GAEA: A Geolocation Aware Conversational Assistant

Ron Campos*, Ashmal Vayani*, Parth Parag Kulkarni*, Rohit Gupta, Aizan Zafar,
Aritra Dutta, Mubarak Shah

University of Central Florida

Abstract

Image geolocalization, in which an AI model traditionally predicts the precise GPS coordinates of an image, is a challenging task with many downstream applications. However, the user cannot utilize the model to further their knowledge beyond the GPS coordinates; the model lacks an understanding of the location and the conversational ability to communicate with the user. In recent days, with the tremendous progress of large multimodal models (LMMs), both proprietary and open-source researchers have attempted to geolocalize images using LMMs. However, the issues remain unaddressed; beyond general tasks, for more specialized downstream tasks, such as geolocalization, LMMs struggle. In this work, we propose solving this problem by introducing a conversational model, GAEA*, that provides information regarding the location of an image as the user requires. No large-scale dataset exists that enables the training of such a model. Thus, we propose* GAEA-1.4M*, a comprehensive dataset comprising over 800k images and approximately 1.4M question-answer pairs, constructed by leveraging OpenStreetMap (OSM) attributes and geographical context clues. For quantitative evaluation, we propose a diverse benchmark,* GAEA-Bench*, comprising 3.5k image-text pairs to evaluate conversational capabilities equipped with diverse question types. We consider 11 state-of-the-art open-source and proprietary LMMs and demonstrate that* GAEA *significantly outperforms the best open-source model, LLaVA-OneVision, by 18.2% and the best proprietary model, GPT-4o, by 7.2%. Our dataset, model and codes are available* ***https://ucf-crcv.github.io/GAEA***.

1. Introduction

Image geolocalization [25, 37, 58, 64, 69] is a notoriously challenging task, in which conventional AI models predict the precise GPS coordinate of an image taken anywhere on Earth. Moreover, seasonal changes, geographical and climatic diversity, changes in solar zenith angle, and lack of diverse image distributions make the geolocalization task more challenging. Although challenging, geolocalization has direct applications in multiple domains, including tourism [16], navigation, urban planning [51], and security [22, 58], among others.

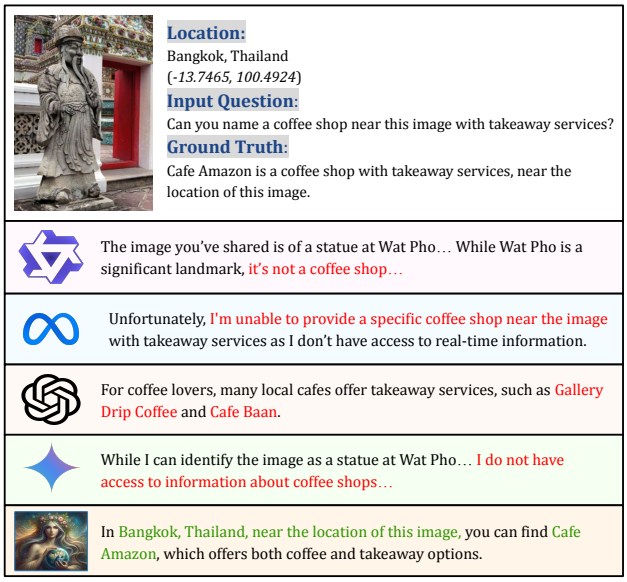

Figure 1. **Performance comparison** of various LMMs on the geographically-grounded visual-question-answering task (included in our new GAEA-Bench benchmark). Most LMMs can describe the Wat Pho statue, but only GAEA, our **G**eolocation **A**war**E** **A**ssistant, retrieves the correct nearby cafe, *Cafe Amazon*.

Recently, the CLIP-inspired image-to-GPS retrieval approach, GeoCLIP [58], has shown significant performance in global-scale image geolocalization. To further mitigate the performance gap, and to increase the generalization capacity of the models, interestingly, a new wave of works infuse *human-level cognition and inference capacity* in their model training [25, 37, 64]. E.g., PIGEON is trained on data from the popular geolocalization game GeoGuesser [2]; a recent vision-language model, GeoReasoner [37] uses user- and administrator-maintained approximately 3K textual clues from GeoGuessr and Tuxun gaming platforms.

These focused geolocalization models lack a geographical understanding of the predicted locations beyond their

*equally contributing first authors

GPS coordinates. They cannot provide additional information that might be invaluable for applications such as tourism, navigation, urban planning, etc. Even if the models possess that understanding, they do not possess the conversational ability to convey that information and fail to meet the user's needs. In contrast, despite having the conversational capability, visually and textually prompted large language models (LLMs) [21, 56, 65] and their multimodal variants, popularly referred to as large multimodal models (LMMs) [10, 11, 14, 40, 55], fail to capture fine-grained nuances from an image in specialized downstream tasks such as geolocalization, making their predictions vastly imprecise and worse than random guesses in many cases(Fig. 1).

Motivated by these aspects, in this paper, we propose GAEA, an open-source conversational model with global-scale geolocalization capability. To the best of our knowledge, this is the first work in the ground-view geolocalization domain that introduces an open-source conversational chatbot, where the user can obtain image geolocalization, a relevant description of the image, and engage in a meaningful conversation about the surrounding landmarks, natural attractions, restaurants or coffee shops, medical or emergency facilities, and recreational areas.

However, training an open-source LMM with conversational capacity is not straightforward. These models are data-hungry, and their training is computationally intensive. Unfortunately, no dataset can facilitate the training of such a model. To this end, we meticulously curate a GAEA-1.4M–a high-quality conversational VQA pair equipped with diversity in scene understanding and image captions for training and instruction tuning the LMMs on the street-level geolocalization task. GAEA-1.4M is a comprehensive dataset consisting over 800k images from *MP-16* [34], *GLD-v2* [63], and CityGuessr68k [33] covering locations around the Earth. We augment these images with rich metadata from OpenStreetMap (OSM) [43] at a 1 km radius, a first effort of its kind. OSM attributes contain details about the surrounding area, nearby landmarks, accessible services, and the historical buildup of the region. Additionally, the geolocalizable explanatory captions set contains 385K image-QA pairs and is equipped with their corresponding knowledge and reasoning captions. These knowledge and reasoning captions are constructed using a set of geographical context clues from GeoGuessr [2] that enable the model to gain a holistic understanding of the location. Taken together, GAEA-1.4M is the largest and most comprehensive collection of geolocalizable and conversational QA pairs. Using this, we train our conversational chatbot, GAEA.

To quantitatively evaluate the conversational capability of LMMs and address the scarcity of benchmark datasets in a geolocalization setting, we propose GAEA-Bench, a diverse set of 3.5K conversational question samples.

GAEA-bench comprises multiple-choice (MCQs) and true/false (T/Fs) for checking a model's understanding and choosing capability, short questions (SVQAs) for testing a model's knowledge, and long questions (LVQAs) for evaluating a model's descriptive and in-depth explanation ability about the location in question.

We summarize the main contributions as follows:

- We propose GAEA-1.4M (Section 3), a new dataset for training conversational image geolocalization models.
- For evaluating conversational capabilities in a geolocalization setting (Section 5), we propose GAEA-Bench, a novel benchmark of 3.5K samples with various question types, including short, long, MCQs, and T/F.
- We propose GAEA, a conversational chatbot (Section 4) that extends beyond global-scale geolocalization, providing information about the location described by an image.
- We quantitatively compare the performance of our model to 8 state-of-the-art open-source and 3 proprietary LMMs, including GPT-4o [9] and Gemini-2.0-Flash [55].

2. Related Work

Large Multimodal Models (LMMs) have been at the forefront of computer vision research; geo-localizable LMMs are in their nascent stages. Multimodal learning unifies different modalities by a common representation. By *contrastively* fitting text and images into the same feature space, CLIP [44] has revolutionized multimodal learning. LLMs like GPT2 [45] made strides in representing text and next token prediction. Visual question answering (VQA) was of interest before, but after LLaVA [39] and BLIP2 [36] combined the conversational aspects of LLMs and the representational capabilities of CLIP-like models, many problems of VQA are addressed. After that, numerous modern works emerged, such as GeoChat [32], Qwen2.5-VL [12], LLaMA-3.2 Vision [10], and LLaVA-OneVision [35], X-Former[52] as well as proprietary models like GPT4 [9] and Gemini [55]. Although most of these models are excellent for general VQA, they perform poorly on specialized tasks in fields like medicine, statistics, and geolocalization. This inspires the need for specialized LMMs for specific tasks.

Geolocalization is a crucial field in vision research with essential applications in forensics, social media, and exploration; see [16, 51, 58]. On a global scale, Weyand et al. [62] first introduced a classification-based approach on the Im2GPS [26] dataset. Vo et al. [59] introduced classification in multiple hierarchies, while CPlaNet [50] introduced a combinatorial partitioning technique for combining coarse hierarchies to predict finer ones. Over the years, many other works like ISNs [28], TransGeo [69], TransLocator [60], and GeoDecoder [15] have made significant advancements in this classification-based worldwide geolocalization by introducing scene-based specialized encoders and hierarchical evaluation, auxiliary scene recognition, and twin encoder approach, and a query-based encoder-decoder archi-

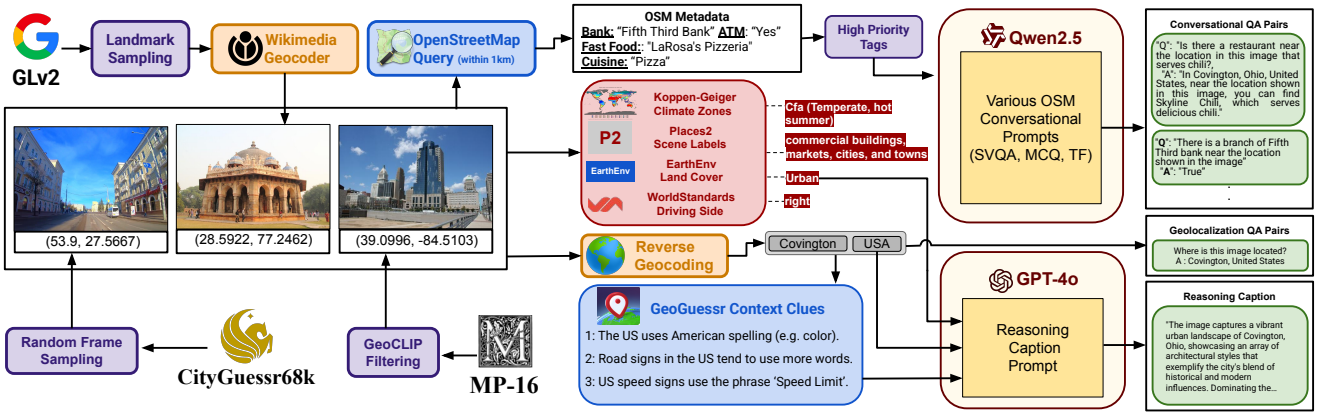

Figure 2. **Data Collection and Annotation Pipeline.** GAEA-1.4M includes geographically diverse visual samples from various data sources, such as MP-16 [34], GLD-v2 [63], and CityGuesser68k [33] *(left)*. We also incorporate additional metadata and auxiliary context for each image from OpenStreetMap (OSM), ranging from climate zones to geographical clues about the country *(middle)*. Using open-source LLMs and GPT-4o, we generate four diverse question-answer pairs across geolocation, reasoning, and conversational subsets *(right)*.

tecture, respectively. PIGEON [25], the most recent work, leverages the image representation capabilities of the CLIP vision encoder and a unique clustering approach to improve its geolocalization performance. Image-to-image retrieval models tend to be more accurate than their classification-based counterparts, but are infeasible on a global scale due to their requirement for large reference image galleries. GeoCLIP [58] was the first work to incorporate the contrastive multimodal learning between images and raw GPS information, which revolutionized this domain by introducing a more accurate retrieval-based model for a global scale.

These specialized models work well for worldwide image geolocalization but lack the conversational aspect that can aid an individual in gaining a holistic understanding of a location portrayed in an image. GeoReasoner [37] attempts to incorporate an inherent geospatial understanding into an LMM by looking at specific information displayed in the image. It also introduces the idea of *locatability*, which can determine the extent of that information present in the image, which may improve the reasoning capability of the model. The model, however, lacks the *conversational aspect*, and the locatability-based filtering of data might hurt its generalization capability. We address these issues in GAEA by primarily focusing on its conversational ability.

The generalizability of GAEA comes from its training data. All specialized geolocalization global-scale models are trained on MP-16 [34], which is a large-scale worldwide dataset. However, it lacks the verbal context required in LMM training. Hence, we introduce a new conversational dataset GAEA-1.4M; see details in Section 3. Additionally, inspired by different task-specific LMM evaluation benchmarks[27, 30, 31, 53, 54], we introduce the first conversational benchmark in Section 4 to evaluate Geolocalization LMMs and an evaluation pipeline to judge the efficacy of such models.

3. Dataset Curation and Annotation

3.1. GAEA-1.4M

The GAEA-1.4M dataset provides comprehensive global coverage, featuring both rich conversational and diverse geolocalization sets. It includes various QA formats, such as MCQs, True/False, and open-ended VQA (long and short), from more than 234 countries/territories, grouped under conversational and geolocalization groups. Spanning 40k cities across 7 continents, GAEA-1.4M is structured into two key groups: conversational and geolocalization. With over 1.4 million QA pairs, it captures the geographical diversity of both underrepresented and widely recognized regions worldwide. Fig. 2 shows our complete curation pipeline, which we will discuss piecewise.

Acquiring Diverse Geo-localizable Images. We sample geographically diverse visual data from MediaEval 2016 (MP-16) [34], Google Landmarks v2 (GLDv2) [63], and CityGuessr [33] to curate GAEA-1.4M.

MP-16 contains over 4.6 million geotagged Flickr images, including indoor and outdoor scenes. For our street-view geolocalization subset, we filter out indoor images, retaining 3 million outdoor images. However, some of these images are non-geolocalizable, such as close-up shots of doors, grass, or wires, which are excluded from the final dataset. To filter out non-geolocalizable images, we process all 3 million samples using GeoCLIP [58], which is trained on the full MP-16 dataset and effectively identifies non-geolocalizable outlier images. GeoCLIP assigns a confidence score based on its ability to predict GPS coordinates, with higher scores indicating geo-localizability. We set a confidence threshold of 0.75 and computed the distance between the ground truth MP-16 GPS coordinates and the GeoCLIP's predicted location. We retain the images if this distance is less than 500 km; see additional ablations on

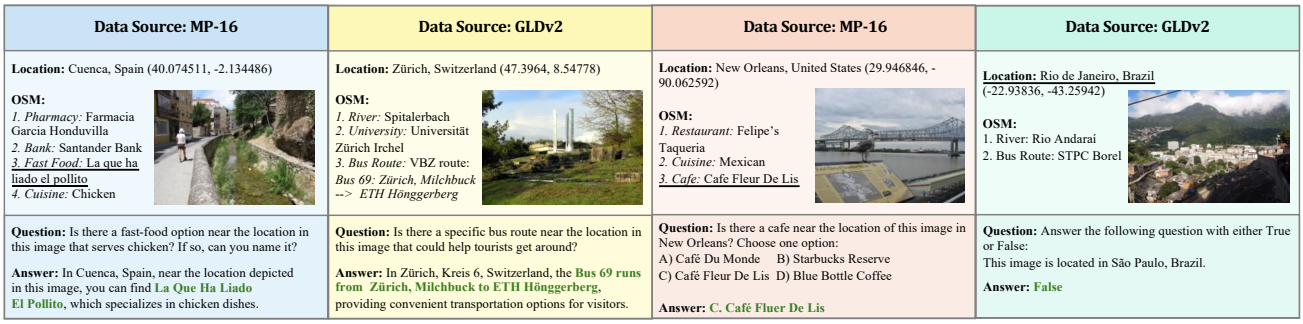

Figure 3. **Qualitative samples** from our GAEA-1.4M training set showcasing various question-types, including multiple-choice, true/false, short and long VQAs generated using a proprietary model, GPT-4o [42]. We carefully select geographical tags from OSM metadata to generate conversational question-answer pairs.

different thresholds and distance metrics in the Appendix.

To achieve a balanced geographical distribution in GAEA-1.4M, we use the 10th hierarchy of S2-Cells [5] to partition our filtered MP-16 dataset into 16,753 spatial grid cells. S2-Cells enable hierarchical spatial indexing, ensuring diverse global coverage while preventing overrepresenting densely imaged regions. We randomly sample up to 200 images from each cell, resulting in a final set of over 750k distinct samples.

GLDv2 [63] is a fine-grained landmark recognition dataset featuring natural and human-made landmarks across diverse time zones, climates, and lighting conditions. Given the significance of landmark geolocation for real-world applications, we randomly sample 50K distinct landmarks from GLDv2. These highly recognizable landmarks offer rich geographic and cultural context. Each image is linked to Wikipedia metadata, from which we extract GPS coordinates using the `Wikimedia` [6] API. We then apply the `reverse_geocoder` Python library to determine each landmark's corresponding city and country.

CityGuessr68k [33] focuses on global video-based geolocalization emphasizing urban regions and hierarchical prediction across 166 major cities. To incorporate this diversity, we randomly sample one frame from each of the 54k training videos and include them in our dataset. These three sources provide over 852k geographically diverse geolocalizable images, forming GAEA-1.4M dataset.

3.1.1. Meta-data curation for dataset annotation

After acquiring all visual samples for our GAEA-Conversational Assistant, we churn the metadata for each image for a comprehensive QA-pair generation.

Churning OSM metadata. OpenStreetMap (OSM) [43] is a collaborative open-source mapping platform that provides extensive geographical data. In our work, OSM plays a central role by enriching geolocalization and conversational capabilities. We retrieve metadata from a 1 km radius around the GPS coordinates of 850k images, leveraging OSM's detailed, publicly annotated tags. These tags cover many real-world elements, including amenities, transportation, hotels, and restaurants, making them invaluable for our ground-view geolocalization and QA generation.

OSM data is multilingual, which is a key challenge. To ensure accessibility, we use GPT-4o [9] to translate these annotations into English. Additionally, many retrieved tags consisted of plain numbers or non-meaningful entries, which we systematically filtered out to retain only informative and contextually relevant metadata. To our knowledge, this is the first work to utilize OSM's rich metadata to develop a conversational chatbot for ground-view geolocalization. Fig. 3 shows high-priority meta-tags derived from OSM churned data.

Curating Country-Specific Geographical Clues. We web-crawled diverse clues from `Plonkit`[4], an open-source community resource for the GeoGuessr [2] game, which has over 65 million players. Similar datasets have been used in recent works [25, 37]. We obtained 129 country clues but found gaps for some countries, such as New Zealand and France. To address this, we curated clues for 58 additional countries using GPT-4o, aligning them with `Plonkit`'s style, resulting in altogether 187 countries. These clues are incorporated into our dataset for generating reasoning-based Q&As. For examples of the type of clues utilized, see Fig. 9 in the Supplementary material.

Additional Metadata. For auxiliary context, we group our country-specific, geographically diverse dataset in 31 Köppen-Geiger climate zones [13]. We obtain the traffic direction data through WorldStandards [7] and *Land Cover Use* statistics from EarthEnv [1]. Additionally, we compute scene labels for each image using the Places2 [68] database.

3.1.2. Question-Answer (QA) Pairs Generation

GAEA-1.4M is carefully curated to enhance ground-view geolocalization through diverse, context-rich QA pairs; see Fig. 2. Comprising over 800k distinct images and around 1.4 million QA pairs, it stands as the largest and most com-

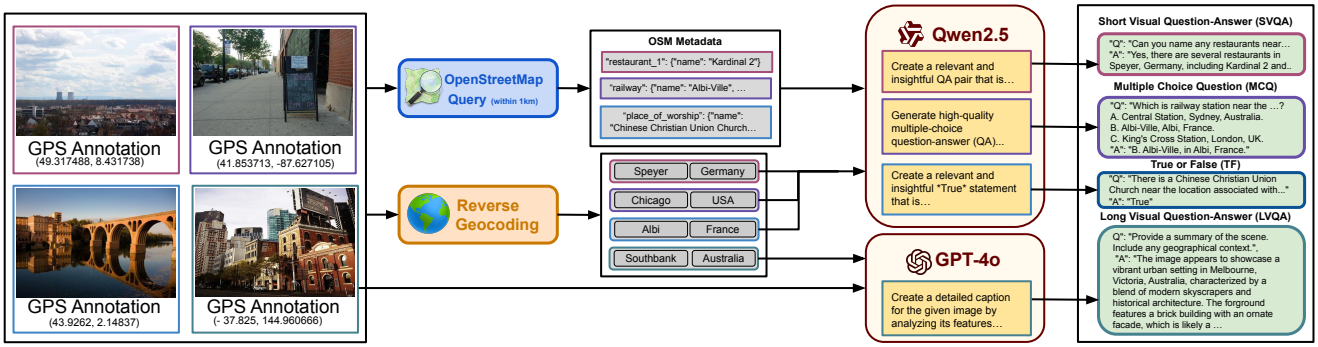

Figure 4. **Overview of GAEA-Bench.** GAEA-Bench is designed to evaluate the conversational abilities of various LMMs across different question types, including MCQs, T/F, and both short and long VQAs. We have carefully selected a subset of 3.5k samples from MP-16 [34] and generated corresponding OSM metadata to generate QA pairs using GPT-4o [42]. GAEA-Bench aims to fill the gap in conversational benchmarks by incorporating geolocalization capabilities.

prehensive dataset for this task; see Fig. 3. Unlike existing works, such as [20, 37], which are limited to JSON structures and fewer question types, our work emphasizes the conversational capabilities of the model, providing a broader range of QA formats. The dataset is divided into three subsets—Conversational, Reasoning, and Geolocalization, each designed to capture different aspects of geographic understanding. These subsets feature various question formats, including multiple-choice [41], true/false, and open-ended questions (SVQA and LVQA) [57]. Below, we detail the curation process for each subset.

Conversational QA Generation. We generate conversational QA pairs using OSM metadata from the sampled MP-16 and GLDv2 subsets. We prompt Qwen-2.5-14B [65] with enriched OSM attributes to create diverse question formats, including short-form, multiple-choice, and true/false questions. These OSM tags cover various categories such as amenities, food places, financial institutions, government offices, accommodation, transportation, healthcare, religious sites, education, and waterways. This subset comprises over 380k questions. Fig. 3 showcases qualitative samples of QA pairs curated using OSM churned data.

Geolocalization Questions. To enhance the geolocalization capabilities of our GAEA model, we introduce large-scale meta-geographic information through geolocation-specific QA pairs. This subset consists of 820k image-question pairs designed to help the model predict the correct location of an image. We curate 50k geolocation questions from GLDv2, each corresponding to a distinct landmark, leveraging their global recognition to improve location-based reasoning. Additionally, we incorporate 54K geolocation QA pairs from CityGuessr, which focuses on urban environments, and 720K from MP-16, ensuring broad geographic coverage. This results in a diverse and well-distributed geolocation QA dataset spanning 234 countries and territories, 40k cities, and 7 continents.

Reasoning Questions. We generate detailed image-caption QA pairs (Long-VQA) to enhance fine-grained reasoning in our GAEA model. We prompt GPT-4o [9] with each image, its scene labels, and country-specific geographical attributes, including GeoGuessr clues, traffic-side driving information, Köppen-Geiger climate zone, and land cover data. While scene labels are unique to each image, the other attributes provide country-level context. GPT-4o integrates this information to generate contextually rich and highly correlated captions with the provided geographic labels. These reasoning-based captions strengthen the model's geolocalization and conversational capabilities and induce a rich semantic understanding in our model by infusing *human-level cognition and inference capability*, enabling the model to emphasize why particular image features might be associated with specific geographical contexts, reducing disinformation [47, 48]. In total, we curate 237k knowledge-driven LVQA pairs.

3.2. GAEA-Bench

Existing benchmarks for evaluating geolocalization tasks mainly focus on retrieval and classification-based methods, such as IM2GPS [26], IM2GPS3k [59], and GSW15k [18], which assess the distance between ground-truth and predicted GPS coordinates. However, there is a lack of conversational benchmarking datasets to evaluate the geolocalization and conversational capabilities of LMMs. We introduce GAEA-Bench, a geographically diverse and conversationally rich multimodal benchmark to address these shortcomings. GAEA-Bench is designed to assess LMMs across various question types, including MCQs, true/false, and long and short VQAs while integrating geolocalization tasks. It includes 3.5k image-text QA pairs that provide a rich geographical context for each image.

GAEA-Bench Curation. We curate a non-overlapping subset of highly geolocalizable MP-16 images, manually filtering out the non-geolocalizable ones. Using OpenStreetMaps (OSM), we generate metadata within a 1km

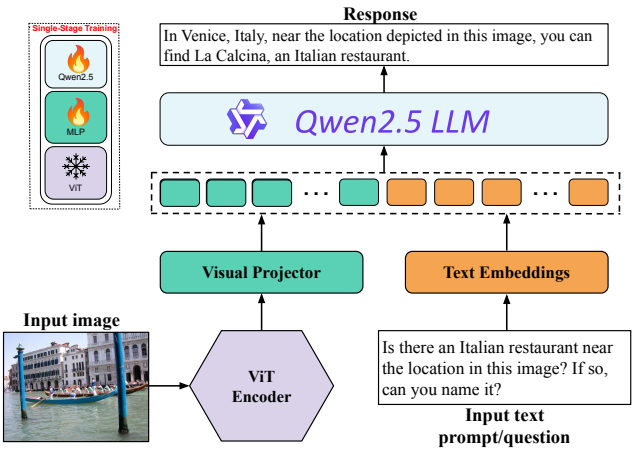

Figure 5. **Overview of the GAEA model architecture and workflow.** An input image is first processed by a Vision Transformer (ViT) encoder, whose output is projected through a visual projector to obtain visual embeddings. Simultaneously, the input text prompt is converted into text embeddings. The combined visual and textual embeddings are then fed into the Qwen2.5 LLM space, which generates a response based on the multimodal input. We follow the single-stage training approach, unfreezing MLP, and performing LoRA fine-tuning in the same stage.

radius and curate 975 short-form (SVQA), 978 multiple-choice (MCQ), and 978 true/false (T/F) questions. For long-form questions (LVQAs), we follow a similar process for generating reasoning questions in GAEA-Bench, resulting in an additional 383 questions. In total, we curate 3,314 diverse image-text QA pairs. To ensure that the GAEA-Bench remains independent of the training set, we select geographically distinct locations for its 3.5k samples. We show our GAEA-Bench annotation and curation process in Fig. 4. The OSM metadata are fetched for each image and are passed to Qwen2.5-14B for generating several QA pairs, including SVQA, MCQ, and T/F.

4. Model Architecture

GAEA follows the architecture of the open-source model, Qwen2.5-VL [12], which seamlessly integrates (1) a vision encoder, (2) a vision-to-language projector, and (3) a language model. The re-engineered vision-transformer (ViT) architecture incorporates 2D-RoPE and window attention. The projector is a two-layer multi-layer perception (MLP) to align raw patch features from the ViT and provides the final representation $\mathbf{E}^{\text{Joint}}$ by concatenating the image embeddings, $\mathbf{E}^{\text{Img}}$ with the text embeddings, $\mathbf{E}^{\text{Text}}$ such that $\mathbf{E}^{\text{Joint}} = [\mathbf{E}^{\text{Img}}, \mathbf{E}^{\text{Text}}]$; see Fig. 5.

Training Details. We perform single-stage fine-tuning of Qwen2.5VL on our GAEA Conversational Assistant dataset. The model is trained across all three subsets, *geolocalization, reasoning, and conversational*, covering both open-ended QA formats (short and long answers) and decision-based questions (multiple-choice and true/false). This fine-tuning process enables the model to integrate rich geographical cues, contextual metadata, and image-specific attributes, enhancing its spatial reasoning, location inference, and multimodal conversational capabilities. We employ LoRA fine-tuning [29] with a rank of $r = 16$ and $\alpha = 32$ along with the unfrozen vision-to-language MLP projector. To handle varying image resolutions, we apply dynamic resolution processing: Images below 448×448 are upsampled, while those exceeding 1000×1000 are downsampled, similar to [12]. The model is trained for one epoch over 12,600 steps.

5. Benchmarking and Evaluations

GAEA-1.4M training set comprises four distinct question types: Multiple Choice Questions (MCQs), True/False (T/F), and Short and Long Visual Question Answering (VQA). GAEA is meticulously trained to ensure conversational fluency while possessing the capability to geolocalize visual samples. Current evaluation frameworks primarily focus on standard geolocalization datasets, measuring accuracy using distance-based metrics at various scales, including Street (1 km), City (25 km), Region (200 km), Country (750 km), and Continent (2,500 km). However, these methods fail to assess the conversational capabilities of LMMs. To address this gap, we define our evaluation process in three key dimensions: (a) Conversational accuracy, (b) Quantitative geolocalization accuracy, and (c) Classification accuracy.

5.1. Evaluation and Metrics

Conversational Evaluation. Most geolocation-specific models operate as "black box" systems, providing GPS coordinates without offering any reasoning or justification behind their outputs. In contrast, GAEA is the first model of its kind, explicitly trained on 1.4 million instructions, which include a significant number of knowledge-reasoning question-answer pairs. This enables GAEA to integrate world knowledge, such as geographical clues, conversational meta-tags, and advanced reasoning capabilities, making its geolocation predictions more transparent and insightful. To address the challenges of complex conversational evaluation, we benchmark 12 state-of-the-art open-source and closed-source LMMs on GAEA-Bench, meticulously curated to evaluate LMMs on diverse question types, including multiple-choice, true/false, and open-ended questions (short and long VQAs). See Section 8.1 for the baselines used in this work.

We employ different prompts for each type of question. We use GPT-4o as a judge and prompt it to score responses to various types of questions with different criteria. We use *accuracy* for MCQs and T/F, *correctness* for SVQA, and *consistency, relevance, and geographical correctness*

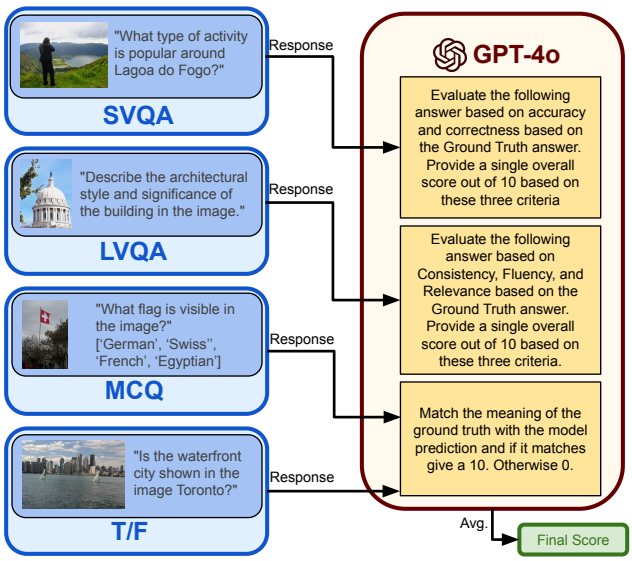

Figure 6. **Evaluation pipeline** for conversational benchmarking on GAEA-Bench, highlighting various question types we introduce in our GAEA-Bench. Each question type is evaluated with various defined criteria using GPT-4o as a judge. For instance, SVQA is evaluated against *accuracy* and *correctness*, and LVQA is evaluated on *Consistency*, *Fluency*, and *Relevancy* criteria.

for long VQAs (LVQAs); see the evaluation pipeline in Fig. 6. Here, *correctness* refers to how closely the model's output matches the location and the correct answer in the ground-truth response [57]. For LVQA, the *consistency* metric evaluates the fluency and readability of the model's prediction [49, 56, 57], while *geographical correctness* assesses whether the model's prediction accurately identifies the correct city and country, directly matching the ground-truth answer (Further discussed in Section 8.2, Fig. 12).

Quantitative Geolocalization Evaluation. We compared the performance of GAEA against six state-of-the-art (SoTA) geolocalization models, namely PlaNet [62], CPlaNet [50], ISNs [28], TransLocator [60], GeoDecoder [18], and PIGEON [25] on three standard geolocalization benchmarks including IM2GPS [26], IM2GPS3k [59], GWS15k [18]. We prompt various LMMs to output the corresponding city and country to which the image belongs. We retrieve GPS coordinates using GeoPy [3] and compute distance with ground truth. We compare the output with distance thresholds of 1 km, 25 km, 200 km, 750 km, and 2,500 km; see Table 2.

Classification Accuracy. Fig. 7 illustrates the *classification* accuracy at the city and country levels and *distance* threshold accuracy computation pipeline. For this evaluation, we benchmark on three datasets: GeoDE [46], DollarStreet [23], and CityGuessr68k [33]. From GeoDE, we sampled 22k images based on 16 meta-tags having geolocalizable features. From DollarStreet, we manually sampled 1.3k images, removing indoor and non-geolocalizable

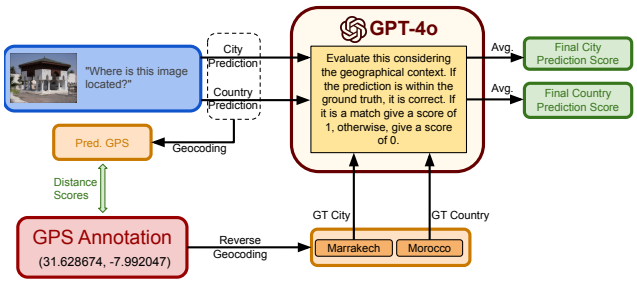

Figure 7. **Classification and distance threshold accuracy computation pipeline** simultaneously evaluates geolocalization performance at *city* and *country* level by comparing model predictions with ground truth annotations derived from reverse-geocoding GPS coordinates and accuracy at different distance thresholds by geocoding predictions of the model.

samples. Since its metadata contains only country-level information, we evaluate this dataset solely for country classification. Additionally, we use the validation set of 14k images from CityGuessr and all 22k GeoDE samples for city and country classification tasks.

5.2. Results and Discussion

GAEA-Bench Evaluation. Table 1 presents the performance of 12 recent LMMs on GAEA-Bench. The results offer several insights: (*i*) Our proposed model, GAEA, achieves the highest average performance across decision-making questions (T/F and MCQs) and Short VQAs. Among proprietary models, GPT-4o [9] overall performs the best, with an accuracy of 60.3%, excelling particularly in Long VQAs, outperforming GAEA by 1.9% in this category. However, both open-source and proprietary models struggle with short-form questions. E.g., GPT-4o's accuracy drops from 66.3% on long questions to 49.5% on short questions. (*ii*) GAEA outperforms all LMMs with an average accuracy of 67.5%, surpassing GPT-4o by 7.2% and outperforming the second-best open-source model, LLaVA-OneVision [35], by 18.2%. (*iii*) Several open-source models, including LLaMA-3.2-11B [21], GLM-4V-9B [24], and Phi-3.5-Vision [8], achieve comparable overall performance. (*iv*) LMMs perform better on decision-making questions (MCQs and T/F) than open-ended questions; see Fig. 17. E.g., LLaVA-OneVision experiences a 26.5% drop in accuracy on SVQA compared to MCQ questions. The low performance on free-form questions underscores the challenge of using short questions in effectively assessing conversational capabilities in the GAEA-Bench. Figs. 14,15,16 shows comparisons with several LMMs.

Standard Geolocalization Evaluation. Table 2 compares GAEA's performance with various specialized encoder-only methods across three standard geolocalization benchmarks. While GAEA is trained on a large-scale conversational dataset with geolocalization capabilities, it achieves competitive results against specialized models. We evaluate

Model Name	Performance on Different QA Formats				
	LVQA	SVQA	MCQ	TF	Average
GeoChat-7B [32]	23.9 −40.5	16.4 −36.2	54.5 −18.2	32.1 −46.4	33.2 −34.3
LLaVA-Next-Mistral-7B [38]	48.0 −16.4	23.2 −29.4	29.2 −43.5	56.7 −21.8	37.7 −29.8
Phi-3.5-Vision-Instruct [8]	48.7 −15.7	14.3 −38.3	54.7 −18.0	57.4 −21.1	42.9 −24.6
LLaMA-3.2-Vision-11B [10]	48.6 −15.8	29.9 −22.7	53.2 −19.5	47.1 −31.4	44.0 −23.5
GLM-4V-9B [24]	41.6 −22.8	29.7 −22.9	56.7 −16.0	50.6 −27.9	45.2 −22.3
Qwen2.5-VL [12]	57.0 −7.4	29.9 −22.7	48.1 −24.6	59.4 −19.1	47.1 −20.4
InternVL2-8B [66]	54.5 −9.9	31.0 −21.6	55.7 −17.0	56.8 −21.7	48.6 −18.9
LLaVA-OV-7B [35]	54.1 −10.3	31.5 −21.1	58.0 −14.7	56.4 −22.1	49.3 −18.2
Gemini-2.0-Flash [55]	58.3 −6.1	34.7 −17.9	57.2 −15.5	56.7 −21.8	50.5 −17.0
GPT-4o-mini [9] *	61.8 −2.6	34.1 −18.5	54.9 −17.8	33.9 −44.6	43.4 −24.1
GPT-4o [9] *	**66.3** +1.9	49.5 −3.1	59.4 −13.3	69.6 −8.9	60.3 −7.2
GAEA (Ours)	64.4	**52.6**	**72.7**	**78.5**	**67.5**

Table 1. **Benchmarking of 11 open-source and proprietary LMMs on GAEA-Bench.** Notably, GAEA outperforms all open-source models and fares higher than the proprietary models on decision-making questions *(MCQs and TFs)*. We provide the relative performance change for each model compared to GAEA. *- We use GPT-4o as a judge for evaluation, and it has been documented that LLMs as judges prefer their long-form output [61, 67], hence the scores for these models are likely overestimated.

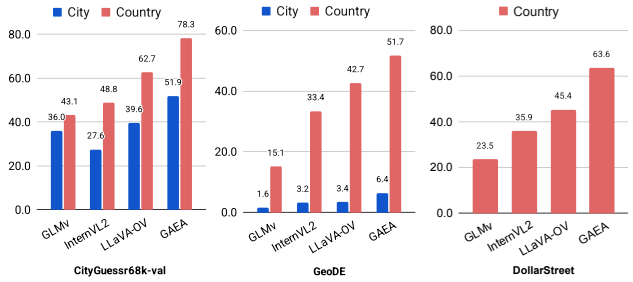

Figure 8. **Classification accuracy** for both city and country labels, where GAEA surpasses several recent LMMs in performance.

against GaGA [20], which is trained on a dataset five times larger than ours, on IM2GPS and IM2GPS3k. However, we exclude comparisons on GWS15k due to differences in dataset curation. We contacted the authors of [18] for the original GWS15k benchmark for fair evaluation.

On the IM2GPS3k benchmark, GAEA consistently outperforms all existing LMMs, including the domain-specialized PIGEON model [25], across all four distance thresholds. Notably, GAEA surpasses PIGEON by 2.2% at the 200 km threshold and by approximately 1.4% at the continent level. Compared to GeoCLIP [58], GAEA achieves gains of 5.3% at the region level and 2.5% at the city level. Additionally, GAEA significantly outperforms the open-source models, GaGA [20], with improvements of 6.1% at the country level and 8.0% at the regional level, and GeoReasoner [37] with 10.5% at city, 15.6% at region, and 13.9% at a 2500 km threshold. On IM2GPS, GAEA outperforms both the specialized LMMs, GeoReasoner, and GaGA on all four distance metrics. It also surpasses PIGEON and GeoCLIP models at the city level by 2.1% and 1.2%, respectively, while maintaining competitive performance across other thresholds. We also evaluate GAEA on GWS15k, one of the most challenging datasets, which includes non-geolocalizable landmarks. GAEA outperforms GeoCLIP [58] and GeoDecoder [18] on city-level distance and achieves comparable performance at the region and country levels. Fig. 8 presents GAEA's **Classification Accuracy** on three new datasets: CityGuessr68k-val [33], GeoDE [46], and DollarStreet [23]. GAEA outperforms recent LMMs, including LLaVA-OneVision [35], InternVL [17], and GLM-4V-9B [24], on both city- and country-level classification. These results highlight GAEA's extensive geographical coverage and strong geolocation capabilities.

Benchmark	Model	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
IM2GPS [26]	PlaNet [62]	24.5	37.6	53.6	71.3
	CPlaNet [50]	37.1	46.4	62.0	78.5
	ISNs [28]	43.0	51.9	66.7	80.2
	TransLocator [60]	48.1	64.6	75.6	86.7
	GeoCLIP [58]	41.8	60.8	77.2	89.9
	GeoDecoder [18]	**50.2**	**69.0**	80.0	89.1
	PIGEON [25]	40.9	63.3	**82.3**	**91.1**
	GeoReasoner [37]	24.9	48.1	65.8	82.3
	GaGA [20]	38.8	54.8	75.1	87.7
	GAEA (Ours)	43.0	57.4	77.2	89.5
IM2GPS3k [59]	PlaNet [62]	24.8	34.3	48.4	64.6
	CPlaNet [50]	26.5	34.6	48.6	64.6
	ISNs [28]	28.0	36.6	49.7	66.0
	TransLocator [60]	31.1	46.7	58.9	80.1
	GeoDecoder [18]	33.5	45.9	61.0	76.1
	GeoCLIP [58]	34.5	50.7	69.7	83.8
	PIGEON [25]	36.7	53.8	72.4	85.3
	GeoReasoner [37]	26.5	40.4	57.7	72.8
	GaGA [20]	33.0	48.0	67.1	82.1
	GAEA (Ours)	**36.9**	**56.0**	**73.2**	**86.7**
GWS15k [18]	ISNs [28]	0.6	4.2	15.5	38.5
	TransLocator [60]	1.1	8.0	25.5	48.3
	GeoDecoder [18]	1.5	8.7	26.9	50.5
	GeoCLIP [58]	3.1	**16.9**	**45.7**	**74.1**
	GAEA (Ours)	**3.7**	16.7	43.3	73.5

Table 2. **Benchmarking the performance of various specialized models on standard geolocation datasets.** GAEA demonstrates competitive results, outperforming GaGA on multiple distance thresholds in both IM2GPS and IM2GPS3k.

6. Conclusion

We introduced GAEA, the first interactive conversational model with specialized geolocation capabilities, explicitly trained on a large-scale conversational dataset, GAEA-1.4M. We meticulously designed the dataset to enhance GAEA's reasoning, conversational abilities, and geolocation accuracy. We curated geolocalizable images from MP-16, GLDv2, and CityGuessr68k, enriching them with auxiliary context and metadata, such as geographic clues and climate zones. In addition to a high-quality instruction set, we present GAEA-Bench, a comprehensive benchmark that evaluates LMMs across multiple question types, including MCQs, True/False, short- and long-VQAs. Our results show that GAEA outperforms recent LMMs on GAEA-Bench, demonstrating strong geolocation and conversational capabilities by leveraging OpenStreetMap (OSM) data. These findings establish GAEA as a strong baseline for future research in geolocalization.

Acknowledgements

This work was supported by "MFC Lockheed Martin, Orlando". We would also like to thank David Shatwell, Manu S Pillai, Praveen Tirupattur, Brian Dina, Gaurav Kumar Nayak, and Suranadi Dodampagan-mage for their insightful discussions and contributions.

References

[1] . EarthEnv. https://www.worldstandards.eu/cars/list-of-left-driving-countries/, . 4, 12

[2] . GeoGuessr. https://www.geoguessr.com/, . 1, 2, 4, 12

[3] . GeoPy. https://geopy.readthedocs.io/en/stable/, . 7

[4] . Plonkit. https://www.plonkit.net/, . 4, 12

[5] . S2-Cells. https://code.google.com/archive/p/s2-geometry-library/, . 4

[6] . WikiMedia. https://commons.wikimedia.org/w/api.php, . 4

[7] . WorldStandards. https://www.worldstandards.eu/cars/list-of-left-driving-countries/, . 4, 12

[8] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 7, 8, 15

[9] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 4, 5, 7, 8, 15

[10] Meta AI. Llama 3.2: Vision and edge models. Meta AI Blog, 2024. 2, 8, 15

[11] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in neural information processing systems*, pages 23716–23736, 2022. 2

[12] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 6, 8, 15

[13] Hylke E Beck, Niklaus E Zimmermann, Tim R McVicar, Noemi Vergopolan, Alexis Berg, and Eric F Wood. Present and future köppen-geiger climate classification maps at 1-km resolution. *Scientific data*, 5(1):1–12, 2018. 4, 12

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in neural information processing systems*, pages 1877–1901, 2020. 2

[15] Denis Carriere. Geocoder: Simple, consistent. https://geocoder.readthedocs.io/. Accessed: [Insert Date]. 2

[16] Athanasios Chalvatzaras, Ioannis Pratikakis, and Angelos A Amanatiadis. A survey on map-based localization techniques for autonomous vehicles. *IEEE Transactions on intelligent vehicles*, 8(2):1574–1596, 2022. 1, 2

[17] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 8, 15

[18] Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. 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*, pages 23182–23190, 2023. 5, 7, 8, 15

[19] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022. 15

[20] Zhiyang Dou, Zipeng Wang, Xumeng Han, Chenhui Qiang, Kuiran Wang, Guorong Li, Zhibei Huang, and Zhenjun Han. Gaga: Towards interactive global geolocation assistant. *arXiv preprint arXiv:2412.08907*, 2024. 5, 8, 15

[21] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2, 7

[22] Aritra Dutta, Srijan Das, Jacob Nielsen, Rajatsubhra Chakraborty, and Mubarak Shah. Multiview aerial visual recognition (mavrec): Can multi-view improve aerial visual perception? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22678–22690, 2024. 1

[23] William Gaviria Rojas, Sudnya Diamos, Keertan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. *Advances in Neural Information Processing Systems*, 35:12979–12990, 2022. 7, 8, 15

[24] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024. 7, 8, 15

[25] Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. Pigeon: Predicting image geolocations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12893–12902, 2024. 1, 3, 4, 7, 8, 15

[26] James Hays and Alexei A Efros. Im2gps: estimating geographic information from a single image. In *2008 ieee conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 2, 5, 7, 8, 15

[27] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 3

[28] Zheng Hong, Yiwei Yin, Zhe Luo, and Jiebo Luo. Isns: Image-specific neural style transfer for image geolocation. *arXiv preprint arXiv:2106.11593*, 2021. 2, 7, 8, 15

[29] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. 6, 15

[30] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. 3

[31] Younggun Kim, Sirnam Swetha, Fazil Kagdi, and Mubarak Shah. Safe-llava: A privacy-preserving vision-language dataset and benchmark for biometric safety. *arXiv preprint arXiv:2509.00192*, 2025. 3

[32] Kartik Kuckreja, Muhammad S. Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad S. Khan. Geochat: Grounded large vision-language model for remote sensing. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 8, 15

[33] Parth Parag Kulkarni, Gaurav Kumar Nayak, and Mubarak Shah. Cityguessr: City-level video geo-localization on a global scale. In *European Conference on Computer Vision*, pages 293–311. Springer, 2024. 2, 3, 4, 7, 8, 15

[34] Martha Larson, Mohammad Soleymani, Guillaume Gravier, Bogdan Ionescu, and Gareth JF Jones. The benchmarking initiative for multimedia evaluation: Mediaeval 2016. *IEEE MultiMedia*, 24(1):93–96, 2017. 2, 3, 5

[35] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. 2, 7, 8, 15

[36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2

[37] Ling Li, Yu Ye, Bingchuan Jiang, and Wei Zeng. GeoReasoner: Geo-localization with reasoning in street views using a large vision-language model. In *Proceedings of the 41st International Conference on Machine Learning*, pages 29222–29233, 2024. 1, 3, 4, 5, 8, 15

[38] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024. 8, 15

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

[40] Muhammad Maaz, Hanoona Rasheed, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M Anwer, Tim Baldwin, Michael Felsberg, and Fahad S Khan. Palo: A polyglot large multimodal model for 5b people. *arXiv preprint arXiv:2402.14818*, 2024. 2

[41] Vishal Narnaware, Ashmal Vayani, Rohit Gupta, Swetha Sirnam, and Mubarak Shah. Sb-bench: Stereotype bias benchmark for large multimodal models. *arXiv preprint arXiv:2502.08779*, 2025. 5

[42] OpenAI. Gpt-4o mini: Our affordable and intelligent small model for fast, lightweight tasks, 2024. Available at: https://platform.openai.com/docs/models/o1. 4, 5

[43] OpenStreetMap contributors. Openstreetmap, 2024. [Data set]. OpenStreetMap Foundation. Available as open data under the Open Data Commons Open Database License (ODbL). 2, 4, 12

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. 2

[45] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019. 2

[46] Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems*, 36:66127–66137, 2023. 7, 8, 15

[47] Shaina Raza, Rizwan Qureshi, Anam Zahid, Joseph Fioresi, Ferhat Sadak, Muhammad Saeed, Ranjan Sapkota, Aditya Jain, Anas Zafar, Muneeb Ul Hassan, et al. Who is responsible? the data, models, users or regulations? responsible generative ai for a sustainable future. *arXiv preprint arXiv:2502.08650*, 2025. 5

[48] Shaina Raza, Ashmal Vayani, Aditya Jain, Aravind Narayanan, Vahid Reza Khazaie, Syed Raza Bashir, Elham Dolatabadi, Gias Uddin, Christos Emmanouilidis, Rizwan Qureshi, et al. Vldbench: Vision language models disinformation detection benchmark. *arXiv preprint arXiv:2502.11361*, 2025. 5

[49] Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39, 2022. 7

[50] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 536–551, 2018. 2, 7, 8, 15

[51] Qiaomu Shen, Wei Zeng, Yu Ye, Stefan Müller Arisona, Simon Schubiger, Remo Burkhard, and Huamin Qu. Streetvizor: Visual exploration of human-scale urban forms based on street views. *IEEE transactions on visualization and computer graphics*, 24(1):1004–1013, 2017. 1, 2

[52] Swetha Sirnam, Jinyu Yang, Tal Neiman, Mamshad Nayeem Rizve, Son Tran, Benjamin Yao, Trishul Chilimbi, and Mubarak Shah. X-former: Unifying contrastive and reconstruction learning for mllms. In *Computer Vision – ECCV 2024*, pages 146–162. Springer Nature Switzerland, 2025. 2

[53] Sirnam Swetha, Rohit Gupta, Parth Parag Kulkarni, David G Shatwell, Jeffrey A Chan Santiago, Nyle Siddiqui, Joseph Fioresi, and Mubarak Shah. Implicitqa: Going beyond frames towards implicit video reasoning. *arXiv preprint arXiv:2506.21742*, 2025. 3

[54] Sirnam Swetha, Hilde Kuehne, and Mubarak Shah. Timelogic: A temporal logic benchmark for video qa. *arXiv preprint arXiv:2501.07214*, 2025. 3

[55] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2, 8, 15

[56] Omkar Thawakar, Ashmal Vayani, Salman Khan, Hisham Cholakal, Rao M Anwer, Michael Felsberg, Tim Baldwin, Eric P Xing, and Fahad Shahbaz Khan. Mobillama: Towards accurate and lightweight fully transparent GPT. *arXiv preprint arXiv:2402.16840*, 2024. 2, 7

[57] Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, et al. All languages matter: Evaluating lmms on culturally diverse 100 languages. *arXiv preprint arXiv:2411.16508*, 2024. 5, 7

[58] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geolocalization. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 8

[59] Nam Vo, Nathan Jacobs, and James Hays. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2621–2630, 2017. 2, 5, 7, 8, 15

[60] Bingxian Wang, Ying Chen, Xin Zhang, Haojie Wang, Ziqiang Wang, and Wen Xu. Translocator: A transformer-based large-scale image geolocalization approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2558–2566, 2022. 2, 7, 8, 15

[61] Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in LLM-as-a-judge, 2025. 8

[62] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 37–55. Springer, 2016. 2, 7, 8, 15

[63] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020. 2, 3, 4

[64] Yibo Yan and Joey Lee. Georeasoner: Reasoning on geospatially grounded context for natural language understanding. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4163–4167, 2024. 1

[65] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 2, 5

[66] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models, 2024. 8

[67] Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*, 2025. 8

[68] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 4, 12

[69] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1162–1171, 2022. 1, 2