




Street-Level Geolocalization Using Multimodal Large Language Models and Retrieval-Augmented Generation

Yunus Serhat Bıçakçı , Joseph Shingleton , Anahid Basiri 

Abstract—Street-level geolocalization from images is crucial for a wide range of essential applications and services, such as navigation, location-based recommendations, and urban planning. With the growing popularity of social media data and cameras embedded in smartphones, applying traditional computer vision techniques to localize images has become increasingly challenging, yet highly valuable. This paper introduces a novel approach that integrates open-weight and publicly accessible multimodal large language models with retrieval-augmented generation. The method constructs a vector database using the SigLIP encoder on two large-scale datasets (EMP-16 and OSV-5M). Query images are augmented with prompts containing both similar and dissimilar geolocation information retrieved from this database before being processed by the multimodal large language models. Our approach has demonstrated state-of-the-art performance, achieving higher accuracy compared against three widely used benchmark datasets (IM2GPS, IM2GPS3k, and YFCC4k). Importantly, our solution eliminates the need for expensive fine-tuning or retraining and scales seamlessly to incorporate new data sources. The effectiveness of retrieval-augmented generation-based multimodal large language models in geolocation estimation demonstrated by this paper suggests an alternative path to the traditional methods which rely on the training models from scratch, opening new possibilities for more accessible and scalable solutions in GeoAI.

Index Terms—Multimodal Large Language Models (MLLMs), Geospatial Artificial Intelligence (GeoAI), Retrieval-Augmented Generation (RAG), Image Localization, Street View Imagery (SVI).

I. INTRODUCTION

STREET-LEVEL geolocalization refers to the task of determining the precise geographic location of a scene from one single image [1]. This capability underpins numerous applications and services such as humanitarian aid and disaster relief [2, 3], mis/dis-information detection [4, 5], automated mapping [6, 7]. The growing popularity of social media platforms and the widespread availability of cameras in smartphones have generated vast amounts of user-created images, making the need for accurate and scalable geolocation methods more urgent than ever. [8, 9, 10]. However, the street

view imagery (SVI) can pose challenges to currently existing methods due to their viewpoint variance, cluttered urban environments, and sometimes limited metadata [11, 12, 13]. Despite recent progress, current methods still grapple with issues such as sensitivity to environmental factors [14, 15, 16], sparsely labeled data [11] that generalize effectively across diverse regions. Additionally, high computational costs and domain-specific challenges often prevent conventional image-based geolocation methods from achieving robust street-level accuracy in real-world scenarios.

Recent developments in large language models (LLMs) [17], multimodal large language models (MLLMs) [18] and Retrieval-Augmented Generation (RAG) [19, 20] offer promising solutions to these challenges [21, 22] in geospatial artificial intelligence (GeoAI). The process—initially driven by computer vision models [23, 24], advanced with the advent of language–image pre-training [25, 26, 27, 28], and most recently, predictions have been continually enhanced through the use of MLLMs. These models, already proficient across various tasks, are now capable of achieving high-accuracy geolocalization. In particular, RAG methods [20] can offer a powerful means of task-specific improvement by iteratively providing the model with existing data, thereby reducing both costs and time expenditure associated with large-scale re-training.

In this study, we explore how the integration of MLLMs and RAG techniques can further enhance geolocation estimation in SVI, and we present the state-of-the-art results and methodologies we have achieved.

Following the advent of the transformer architecture [29], we have witnessed rapid advancements in language modeling [30, 31, 17, 32, 33]. Subsequently, with the integration of image and text pairings into a unified representation space [25], MLLMs have emerged as a powerful tool for vision-language tasks [34, 35, 36, 37, 38, 39]. LLMs and MLLMs have become central to advancements in artificial intelligence, and their capabilities can be harnessed in GeoAI. The success of these models lies in their ability to understand and generalize both text and images. Equipped with the capacity to parse and generate both textual and visual content, MLLMs have achieved significant success in general domains such as image captioning [40, 41], visual question answering [42], and cross-modal information retrieval [43, 44]. However, their success in relatively specific or complex domains like geolocation

(Corresponding author: Yunus Serhat Bıçakçı.)

Yunus Serhat Bıçakçı is with the Vocational School of Social Sciences, Marmara University, İstanbul 34865, Türkiye and also with the Geospatial Data Science Group, School of Geographical & Earth Sciences, University of Glasgow, Glasgow G12 8QQ, Scotland, UK, (e-mail: yunus.serhat@marmara.edu.tr).

Joseph Shingleton and Anahid Basiri are with the Geospatial Data Science Group, School of Geographical & Earth Sciences, University of Glasgow, Glasgow G12 8QQ, Scotland, UK.

estimation remains under question [45]. This could be due to the inherent nature of MLLMs, which are trained to prioritize generalization over handling highly specific tasks such as geolocation estimation.

To accurately achieve geolocation estimation, models require a wide range of diverse information, such as landmarks [46] and architectural structures [47]. Nevertheless, purely classification-based and retrieval-based methods have only made incremental improvements in street-level geolocation accuracy [24, 48, 49, 50, 51, 52, 53, 54]. This may suggest that the successful performance bottlenecks remain in real-world scenarios.

While similar methods [45] have demonstrated the feasibility of retrieval-enhanced geolocation, this paper introduces a robust hybrid gallery—combining the Extended MediaEval Placing Tasks 2016 Dataset (EMP-16) [55, 56] and the OSV-5M dataset [13]—along with an open-weights MLLMs. This can address the challenges related to cost constraints, and scales to diverse data sources to provide accurate street-level geolocation. To do so, we transform these datasets into an embedding space using the SigLIP image encoder [26], forming a vector database that stores each image’s embedding alongside its geolocation information. Consequently, any new query image can be converted into an embedding via the same encoder, and the geolocation information of similar and dissimilar images is efficiently retrieved using the Faiss library [57, 58]. Then we prompt the MLLM with (i) the query image itself, and (ii) the retrieved geolocations from the most similar and most dissimilar images. Thus, the model should be able to execute the location estimation capability with an acceptable precision at the street-level, leveraging both general information about the image and location information from our database through the RAG approach. This approach enables us to compare the street-level location estimation results with benchmark datasets including IM2GPS [23], IM2GPS3k [1], and YFCC4k [59].

This paper makes contributions to GeoAI field, in particular to the task of estimating geolocation for SVI. Below, we summarize our key contributions:

- This study demonstrates that high geolocalization estimation accuracy can be achieved using MLLMs and RAG databases.
- The study provides cost and time savings by not performing pre-training and fine-tuning processes often required for other models.
- Our approach exhibits state-of-the-art performance at street-level on three benchmark datasets (IM2GPS, IM2GPS3k, and YFCC4k), and achieves very good results at all other levels.
- We offer performance comparisons with different datasets and models and present comparative tables illustrating which open-weights models excel in geolocation estimation.

II. RELATED WORK

While there are many geolocalization technologies and techniques such as GPS [60], Wi-Fi signals [61], or multispectral satellite imagery [62], with the availability and

popularity of social media platforms as well as miniaturization of smartphone cameras there is a growing need for accurate and scalable street-level image-based geolocalization. Street-level geolocalization refers to the challenge of inferring the geographic location where a single image was captured [1, 63].

The use of street-level images can be inherently complex due to factors like temporal variation [14], weather conditions [15], seasonal changes [16], and urban infrastructures [64]. Recent advancements in computer vision techniques, coupled with the widespread availability of SVI, have made it possible to map street-level features at high resolutions, facilitating geolocalization for a wide range of applications [11]. However, their scalability and accuracy can be still improved.

These have opened researchers with a wide range of opportunities and challenges to work on. For example, “IM2GPS” by Hays and Efros [23] is one of the foundational modern efforts in geolocation estimation that tackled one of the key challenges—limited labeled data—by leveraging a dataset of 100,000 images. Their classification-based approach (using support vector machines) illustrated how large-scale image repositories could be utilized for geolocation. While IM2GPS demonstrated the feasibility of learning geographic cues (e.g., terrain, vegetation, landmarks), the method struggled with scenes lacking salient features, highlighting the importance of handling diverse environmental conditions.

Building on this foundation, Tobias Weyand et al. [24] developed a model namely “PlaNet” using deep learning-based convolutional neural networks (CNNs) for geolocation estimation. They approached the geolocation estimation task as a classification problem by dividing the Earth’s surface into a series of geographic cells, which served as target classes. Subsequently, they employed CNNs to predict potential locations on the Earth’s surface using a continuous probability distribution. This method achieved significantly more accurate results compared to earlier models. Their advancements were further refined in the “CPlaNet” model [49], which predicted finer-grained geographic classes through the combinatorial partitioning of multiple geo-class sets. They developed a method that generated a distinct geo-class set for each classifier, incorporating features such as seasonal and weather conditions. This approach resulted in significantly improved performance compared to the previous model.

In 2018, Eric Müller-Budack and collaborators introduced the “ISNs” framework [48], which divided the Earth’s surface into a hierarchical structure of nested geographic cells. This approach enabled their model to leverage both global and local geographic cues for more precise predictions.

Similarly, the “TransLocator” model [50] approached the problem by segmenting the Earth’s surface into numerous geographic cells and assigning each image to one of these cells, treating geolocation estimation as a classification problem. However, TransLocator distinguished itself by employing a Transformer-based architecture, where two parallel Transformer branches interacted through a fusion strategy. This enabled the model to capture fine-grained details within images, further advancing geolocation estimation capabilities.

Clark et al. [51] introduced GeoDecoder, a model that learns specialized representations of different geographic levels and

scene types using a Swin Transformer-based architecture [65]. Integrating visual and scene information through a hierarchical cross-attention mechanism, the model captures features specific to each geographic category. However, training on the MediaEval Placing Task 2016 (MP-16) dataset led to lower performance compared to later models, especially in SVI tasks. Additionally, it has been shown that incorporating textual information and language localization can further enhance geolocalization.

PIGEOTTO by Haas et al. [66], is built upon the Contrastive Language-Image Pretraining (CLIP) image encoder. It is trained on over 4 million images sourced from the MP-16 dataset and Wikipedia. PIGEOTTO improves geolocalization estimation by leveraging semantic geocells, claiming superior results across a broader geographic scope compared to earlier models. The model demonstrates accuracy improvements at city and country levels, showcasing strong generalization capabilities. However, its performance in SVI tasks remains less competitive compared to our model and other MLLM-based approaches on benchmark datasets. Additionally, the diversity and scale of datasets used for PIGEOTTO’s training, while advantageous for accuracy, can pose computational cost challenges. Despite its architectural similarity to the PIGEON model developed by the same authors, PIGEOTTO achieves better results, supported by the use of distinct datasets in its training process.

GeoCLIP, proposed by Cepeda et al. [67], employs a CLIP-inspired architecture to align image features with global positioning system (GPS) coordinates. Unlike methods that divide the world into discrete classes, GeoCLIP the Earth as a continuous function through spatial encoding using random Fourier features, effectively creating a hierarchical representation to capture geographic information. The model utilizes a pre-trained CLIP model as its image encoder and introduces a location encoder to map GPS coordinates into the same embedding space as the image features. Leveraging the pre-trained CLIP backbone enables the model to handle text queries, though not as advanced as those processed by MLLMs. While the use of precomputed image features on large datasets reduces computational burden during inference, it imposes significant computational overhead during preprocessing. Additionally, the model’s flexibility for post-development modifications is limited, and relying on a single random Fourier feature value may not be sufficient to ensure optimal performance across both small and large geographic areas.

Img2Loc, proposed by Zhou et al. [45], adopts an approach to image geolocation similar to ours by leveraging MLLMs and image-based RAG. Instead of relying solely on retrieval or classification methods based on image content, the model uses a pre-trained CLIP model to create embeddings from images and their associated geolocations. These features are then utilized to retrieve relevant geographic information from a database of geotagged images. However, the model’s results were achieved using an application programming interface (API) that charges per query, raising concerns about the feasibility and scalability of this approach. Additionally, the dataset employed lacked sufficient SVI, leading to suboptimal performance on benchmark datasets for some geographic levels. The

reliance on an external, paid API for consistent performance and the dependency on the quality and scope of the RAG database—particularly its insufficient SVI content—limit the robustness and applicability of the method.

In summary, while many models including PlaNet, CPlaNet, ISNs, TransLocator, GeoDecoder, PIGEOTTO, GeoCLIP, and Img2Loc, represent significant progress in image-based geolocation, from coarse-grid classification to transformer architectures and continuous location encodings, there are still several important challenges to address. Specifically:

- **Model Development:** Methods like PIGEOTTO [66] lead to high computational overhead during training or pre-processing with image datasets. Additionally, making improvements or fine-tuning the models post-development with different images proves to be quite challenging.
- **SVI Performance:** Previous models exhibit subpar results in heavily urbanized or dense city environments, revealing a need for enhanced fine-grained feature extraction.
- **Reliance on External APIs:** Approaches like Img2Loc [45] depend on fee-based APIs, which can limit broad applicability and raise feasibility concerns.

These limitations highlight the need for more robust, scalable solutions that can effectively handle the challenges of SVI while minimizing reliance on expensive external resources. In light of these, this paper proposes a solution that combines an open-weight MLLM framework with a diverse, locally hosted RAG database to eliminate API dependencies, reduce training overhead, and enhance street-level results, as detailed in the next section.

III. METHODOLOGY

In light of these gaps—particularly the challenges of leveraging street-level data, balancing computational costs, and reducing reliance on external APIs—we propose a RAG approach that integrates open-weights MLLMs and includes both similar and dissimilar images geolocation information. This design is flexible allowing expansion to encompass further data modalities. Providing not only positive location cues but also contrasting negative examples, this approach can enhance the accuracy and robustness of geolocation estimations, especially for more complex street-level scenes.

For a given query image, our method for geolocation estimation begins by producing a high-dimensional numerical representation (or embedding) of the image using our chosen encoder. Motivated by the limitations highlighted in the literature review—where purely classification-based or retrieval-only approaches may overlook crucial contextual cues—we adopt a RAG strategy. Specifically, we leverage both similar and dissimilar images from our RAG database (constructed as shown in Figure 1) to form an augmented prompt.

The rationale for including both similar and dissimilar geolocations is twofold. First, similar geolocations provide positive location cues, such as landmarks, architectural styles, or environmental features, which can guide MLLMs toward a probable geolocation. Second, dissimilar geolocations serve as negative contexts, helping to clarify which visual or geographic features are absent in the query scene. This approach

RAG Database Construction

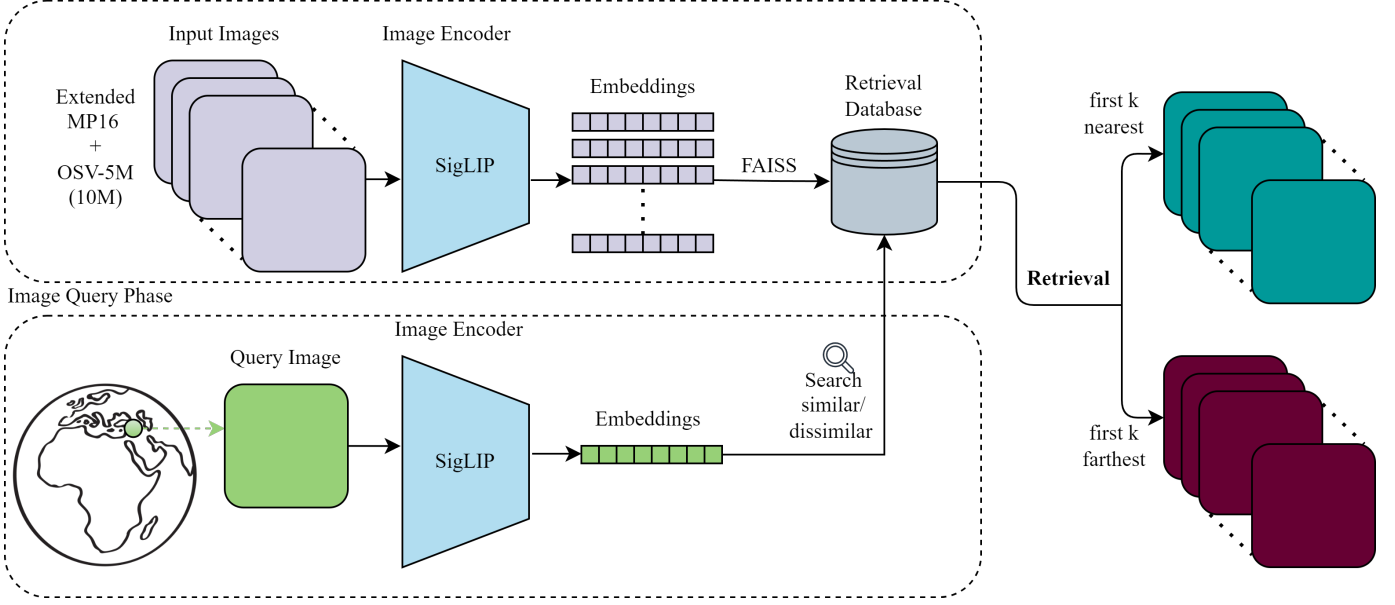


Fig. 1. The RAG database construction and image query pipeline (adapted from [45]).

is closely aligned with the contrastive training methods used by many image encoders within MLLMs.

By offering complementary geolocations from both the most similar and most dissimilar matches, this approach enriches the augmented prompt with relevant contextual information, enabling the MLLM to make more accurate geolocation estimations. A detailed step-by-step explanation of how embeddings and retrieved images are utilized is provided in the subsequent subsections.

A. RAG Database Construction and Image Query

For the RAG database construction phase, we combine the user-centric EMP-16 dataset [56], comprising 4.6 million geo-tagged images captured by ordinary individuals, with the OSV-5M dataset [13], consisting of 5.2 million road street-view geo-tagged images. By merging everyday user-generated photos with structured street-view imagery, we build a hybrid vector database that offers broader coverage of real-world environments and richer location information for evaluating geolocation performance. This approach ensured the inclusion of a greater variety of images from different regions in the vector database, providing the image encoder with a broader range of options to establish close or distant relationships with incoming images. Here, the image encoder is utilized both for creating a vector database from 10 million previously obtained images during the RAG database construction phase as shown in Figure 1 and image query phase for encoding newly received images to perform searches within the retrieval vector database.

For our approach, the image encoder plays a pivotal role and serves as a cornerstone of the methodology. The SigLIP model (siglip-so400m-patch14-224), an open-weights release by Google [26], has been selected as the image encoder.

SigLIP was selected as the image encoder not only because it achieved the best results on benchmarks [26] such as ImageNet-1k and COCO R@1, but also because it outperformed alternatives like OpenAI CLIP [25], LAION [68], and StreetCLIP [69] in our own tests conducted. Additionally, once all embeddings are generated through using an image encoder, we utilise FAISS [57], a tool and library designed to enhance efficiency in vector-based databases, to facilitate searches and manage storage within the database.

After constructing the RAG database, we use the SigLIP image encoder to identify the most similar and most dissimilar neighbors for the incoming image. By "most similar" and "most dissimilar," we refer to finding the nearest and farthest embeddings within the vector space generated by the SigLIP image encoder. In our case, the similarity and dissimilarity are determined based on the L2 distance (Euclidean distance) between vector spaces. This approach allows us to retrieve the location information for as many images as desired, corresponding to the closest and farthest embeddings.

The location information from the most similar and most dissimilar images is then used to augment the prompt, which serves as input to the MLLMs system. The augmented prompt, comprising the given image and the retrieved geolocations (x, y coordinates), is fed into the MLLMs for geolocation estimation. These steps are illustrated in Figure 2.

B. Model Selection

In this study, extensive testing was conducted on numerous open-source MLLMs, including but not limited to the Qwen2-VL [70], InternVL2 [36, 71], Pixtral [72], Llama 3.2 Vision [39], and Aria [73]. The primary reason for selecting these models was their accessibility as open-weight resources on HuggingFace [74].

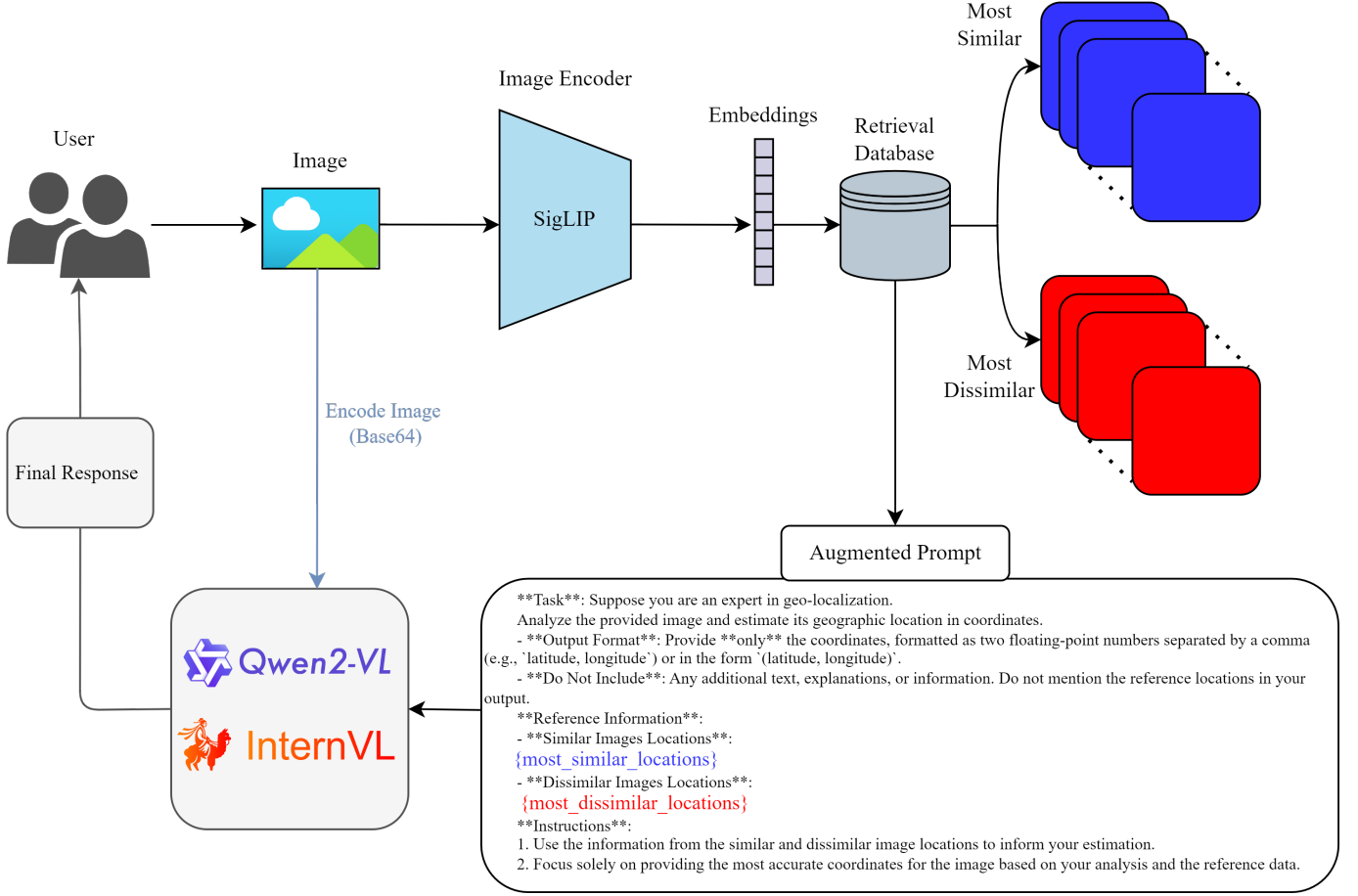


Fig. 2. Overall geolocation estimation framework of the proposed method.

The preliminary test results are consistent with previous research [75, 18]. This can promise a higher performance for MLLMs with a higher number of parameters, in geolocation estimation tasks, similar to other tasks [76, 77, 78]. Consequently, certain models mentioned above were excluded from the final analysis to ensure fairness in the comparison. However, these high parameter models have significant graphics processing unit (GPU) resource requirements. To address this issue, the quantized versions of these models, published by their developers using AWQ [79] and GPTQ [80], were selected. Ultimately, two of the most successful of our tasks MLLMs —Qwen2-VL-72B-Instruct and InternVL2-Llama3-76B— are utilized.

C. Implementation

All experiments are conducted utilizing the high-performance computing (HPC) resources of the university affiliated with the authors. Considering the scale of the models employed, we utilized 2 x NVIDIA RTX 6000 Ada GPUs with 48GB GDDR6 memory (18,176 CUDA Cores) on the HPC to facilitate parallel processing and enable efficient model execution.

The construction of the RAG system was supported by libraries such as PyTorch [81], Faiss [57], Pillow [82], Transformers [83], and pandas [84]. Additionally, to run and per-

form inference with MLLMs, we leveraged the architectures and tools provided by the vLLM [85] and LMDeploy [86] libraries.

The selected MLLMs used specific hyperparameter settings to optimize their performance for the geolocation estimation task. For the RAG phase, the prompt generation incorporated both the 16 most similar and the 16 most dissimilar location information. We empirically determined that retrieving the 16 most similar and 16 most dissimilar embeddings yielded the best performance. Fewer neighbors (1, 5, or 10) provided less contrastive information, while larger sets did not yield additional improvements and increased computational overhead. Furthermore, the following hyperparameters were applied for MLLMs: a temperature [87] of "0.1", a top-p value [87] of "0.1", a maximum model length [88] of "6,000", and a maximum token limit [88] of "512".

To evaluate the obtained outputs for each image included in the benchmark datasets, the geodetic distance between the location estimated by the MLLMs and the coordinates is calculated using the GeoPy library [89]. The computed distance error is then categorised according to thresholds established in previous studies [24, 46, 1]: 1 km for street-level accuracy, 25 km for city-level accuracy, 200 km for region-level accuracy, 750 km for country-level accuracy, and 2,500 km for continent-level accuracy. We used these thresholds to

TABLE I
GEOLOCATION ESTIMATION ACCURACY OF THE PROPOSED METHOD COMPARED TO PREVIOUS METHODS, ACROSS BENCHMARK DATASETS.

Benchmark	Method	Distance (% @ km)				
		Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2,500 km
IM2GPS [23]	PlaNet [24]	8.4	24.5	37.6	53.6	71.3
	CPlaNet [49]	16.5	37.1	46.4	62.0	78.5
	ISNs(M, f^*, S_3) [48]	16.9	43.0	51.9	66.7	80.2
	Translocator [50]	19.9	48.1	64.6	75.6	86.7
	GeoDecoder [51]	22.1	50.2	69.0	80.0	89.1
	PIGEOTTO [66]	14.8	40.9	63.3	82.3	91.1
	Ours (InternVL2-76B)	22.1	49.7	62.8	76.3	89.8
	Ours (Qwen2-VL-72B-Instruct)	23.2	50.2	62.8	78.0	90.7
	Δ (% points)	+1.1	0	-6.2	-4.3	-0.4
IM2GPS3k [1]	PlaNet [24]	8.5	24.8	34.3	48.4	64.6
	CPlaNet [49]	10.2	26.5	34.6	48.6	64.6
	ISNs(M, f^*, S_3) [48]	10.1	27.2	36.2	49.3	65.6
	Translocator [50]	11.8	31.1	46.7	58.9	80.1
	GeoDecoder [51]	12.8	33.5	45.9	61.0	76.1
	PIGEOTTO [66]	11.3	36.7	53.8	72.4	85.3
	GeoCLIP [67]	14.1	34.5	50.7	69.7	83.8
	Img2Loc(GPT4V) [45]	17.1	45.1	57.9	72.9	84.7
	Ours (InternVL2-76B)	15.3	37.0	49.4	65.6	81.1
	Ours (Qwen2-VL-72B-Instruct)	17.1	38.7	51.4	66.6	85.6
	Δ (% points)	0	-6.4	-6.5	-6.3	+0.9
YFCC4k* [1]	PlaNet [24]	5.6	14.3	22.2	36.4	55.8
	CPlaNet [49]	7.9	14.8	21.9	36.4	55.5
	ISNs(M, f^*, S_3) [48]	6.7	16.5	24.2	37.5	54.9
	Translocator [50]	8.4	18.6	27.0	41.1	60.4
	GeoDecoder [51]	10.3	24.4	33.9	50.0	68.7
	PIGEOTTO [66]	10.4	23.7	40.6	62.2	77.7
	GeoCLIP [67]	9.5	19.3	32.6	55.0	74.6
	Img2Loc(GPT4V) [45]	14.1	29.6	41.4	59.3	76.9
	Ours (InternVL2-76B)	20.8	30.0	39.0	54.6	70.7
	Ours (Qwen2-VL-72B-Instruct)	24.3	35.1	44.5	59.5	75.2
	Δ (% points)	+9.9	+5.5	+3.1	-2.7	-2.5

* During the benchmarking process, we encountered an issue where 169 images out of the expected 4,536 were unavailable and returned the message “This photo is no longer available.” Thus, the benchmark was conducted using 4,367 images, which corresponds to 96.3% of the dataset. This implies that 3.7% of the images could not be analysed using our method.

ensure a fair comparison with previous studies conducted on this task and to demonstrate how well our proposed method performs compared to other models.

IV. RESULTS

To assess the efficacy of our proposed method, we conducted comprehensive experiments on three benchmark datasets: IM2GPS [23], IM2GPS3k [1], and YFCC4k [59]. We compared our approach against several methods, including PlaNet [24], CPlaNet [49], ISNs [48], TransLocator [50], GeoDecoder [51], PIGEOTTO [66], GeoCLIP [67], and Img2Loc(GPT4V) [45]. The evaluation metrics are based on the percentage of images localized within specific distance thresholds: 1 km (Street level), 25 km (City level), 200 km (Region level), 750 km (Country level), and 2,500 km (Continent level). Table I summarizes the geolocation estimation accuracy across these methods and datasets.

In the IM2GPS benchmark dataset, our method, utilising the Qwen2-VL-72B-Instruct model, achieved a street-level (1 km) accuracy of 23.2%, surpassing all prior approaches and establishing a new state-of-the-art at this level. At street-level accuracy, our method also ranked second, with the InternVL2-76B model and the GeoDecoder method achieving 22.1%. Compared to the best-performing model in this category, our method represents an improvement of 1.1%. At the city-level

(25 km), our approach demonstrated exceptional performance, matching the GeoDecoder method to secure the best accuracy. On this benchmark dataset, our models achieved the highest overall accuracy of 50.2%, correctly estimating the location for nearly one in two images. At the 200 km and 750 km distance error thresholds, our models exhibited competitive performance, although trailing the best results by 6.2% and 4.3%, respectively. Furthermore, at the continent level accuracy, our approach was among the two methods achieving over 90% precision.

For the IM2GPS3k benchmark dataset, our approach achieved a street-level accuracy of 17.1%, matching the state-of-the-art performance of Img2Loc (GPT4V), which also reached 17.1%. While our method fell short of the highest accuracy at the city, region, and country levels by 6.4%, 6.5%, and 6.3%, respectively, it excelled at the continent level, achieving an accuracy of 85.6%. This was the highest among all compared methods, marking an overall improvement of 0.9%.

In the YFCC4k benchmark dataset, our method demonstrated state-of-the-art performance at street, city, and region levels. Using the Qwen2-VL model, our approach surpassed the nearest competitor at the street level by 9.9%, achieving an accuracy of 24.3%, thereby setting a new state-of-the-art benchmark. Similarly, at the city level, our method achieved

an accuracy of 35.1%, outperforming the closest approach by 5.5%. At the region level, it maintained state-of-the-art accuracy with a score of 44.5%, exceeding the nearest competitor by 3.1%. While the results at the country and continent levels were highly competitive, it is important to note that 169 images in the dataset remained inaccessible despite attempts from various sources. These images were excluded solely from our tests, accounting for 3.7% of the dataset not being used in the analysis.

Across all benchmarks, our method demonstrates superior performance for geolocation estimation at the street level, which is the most ambitious level of granularity. The integration of MLLMs and RAG databases enables our method to leverage detailed visual and textual cues, enhancing its ability to make precise geolocation predictions. The delta values in Table I represent the difference in percentage points between our method and the best methods. Positive delta values indicate an improvement over existing methods, and negative delta values at higher distance thresholds suggest areas for potential enhancement. Our method, in the IM2GPS benchmark dataset, there has been a +1.1% improvement at the street level; in the IM2GPS3k benchmark dataset, a +0.1% improvement at the continent level; and in the YFCC4k benchmark dataset, improvements of +9.9% at the street level, +5.5% at the city level, and +3.1% at the region level.

V. DISCUSSION

In artificial intelligence research, MLLMs, which excel in addressing generalization problems through training on extensive datasets, appear less proficient in geographic location estimation [90, 91] without the application of RAG. However, as evidenced by our results, the integration of RAG significantly enhances their performance. Despite utilizing quantized weights rather than the originally trained weights of open-weight models, our approach achieved state-of-the-art accuracy at the street level — a critical metric for applications requiring precise geolocation — highlighting the method’s effectiveness.

Our findings demonstrate that this refined approach achieves results that outperform existing methods across all metrics, including street-level and city-level granularity, without relying on costly pretraining or fine-tuning processes. This has been delivered by integrating a larger number of street-level photographs into the RAG database, employing a different image encoder, while prioritising open-source models. The consistency of our method’s performance across diverse datasets and evaluation metrics underscores its robustness and generalizability.

It is worth noting, however, that the benchmark datasets used to compare the performance by this study, are subject to temporal inconsistency too. Therefore, some images remained inaccessible in a few cases and it is important to consider the potential impacts. Nevertheless, this limitation did not significantly affect overall results, and our method demonstrated notable performance and resilience against these challenges, reaffirming its robustness when applied to such benchmark datasets.

VI. CONCLUSION

This paper harnessed the power of MLLMs and the capabilities of RAG databases to achieve significantly improved performance in geolocation estimation accuracy at street level and other scales across multiple benchmark datasets. The superior results of our approach, compared to its predecessors following similar methodologies, can be attributed to three key factors: the inclusion of a substantially larger number of street-level images in the RAG database, the selection of SigLIP as the image encoder due to its superior performance metrics, and the preference for open-weight models, chosen for their reproducibility and open-source nature, despite not being the largest models available.

This demonstrates that solving a complex problem like geolocation estimation can be achieved with high precision, without the need for extensive model fine-tuning and retraining, thereby saving considerable time and resources.

Our findings contribute valuable insights to the field, paving the way for the development of more efficient and accurate geolocation estimation tasks.

Future research, conducted with a focus on openly sharing resources such as datasets, open-weight models, and code, will undoubtedly enhance the reproducibility of these approaches. Furthermore, while fine-tuning such large models is perceived to be highly resource-intensive and costly, attempting it could yield intriguing results, given the strong language and vision capabilities of MLLMs.

REFERENCES

- [1] N. Vo, N. Jacobs, and J. Hays, “Revisiting im2gps in the deep learning era,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2621–2630.
- [2] A. Hernandez-Suarez, G. Sanchez-Perez, K. Toscano-Medina, H. Perez-Meana, J. Portillo-Portillo, V. Sanchez, and L. J. García Villalba, “Using twitter data to monitor natural disaster social dynamics: A recurrent neural network approach with word embeddings and kernel density estimation,” *Sensors*, vol. 19, no. 7, p. 1746, 2019.
- [3] R. Suwaileh, T. Elsayed, and M. Imran, *Role of Geolocation Prediction in Disaster Management*. Singapore: Springer Nature Singapore, 2022, pp. 1–31. [Online]. Available: https://doi.org/10.1007/978-981-16-8800-3_176-2
- [4] B. Zhao and D. Z. Sui, “True lies in geospatial big data: detecting location spoofing in social media,” *Annals of GIS*, vol. 23, no. 1, pp. 1–14, 2017.
- [5] G. Kordopatis-Zilos, S. Papadopoulos, and I. Kompatsiaris, “Geotagging text content with language models and feature mining,” *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1971–1986, 2017.
- [6] P. Louro, M. Vieira, and M. A. Vieira, “Geolocalization and navigation by visible light communication to address automated logistics control,” *Optical Engineering*, vol. 61, no. 1, pp. 016 104–016 104, 2022.
- [7] A. Gupta and A. Yilmaz, “Ubiquitous real-time geospatial localization,” in *Proceedings of the Eighth ACM*

- SIGSPATIAL International Workshop on Indoor Spatial Awareness*, 2016, pp. 1–10.
- [8] D. Acharya, R. Tennakoon, S. Muthu, K. Khoshelham, R. Hoseinnezhad, and A. Bab-Hadiashar, “Single-image localisation using 3d models: Combining hierarchical edge maps and semantic segmentation for domain adaptation,” *Automation in Construction*, vol. 136, p. 104152, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0926580522000255>
 - [9] Q. Cui, Y. Zhang, G. Yang, Y. Huang, and Y. Chen, “Analysing gender differences in the perceived safety from street view imagery,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 124, p. 103537, 2023.
 - [10] Y. Zhang, Y. Li, and F. Zhang, “Multi-level urban street representation with street-view imagery and hybrid semantic graph,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 218, pp. 19–32, 2024.
 - [11] F. Biljecki and K. Ito, “Street view imagery in urban analytics and gis: A review,” *Landscape and Urban Planning*, vol. 215, p. 104217, 2021.
 - [12] Y. Hou, M. Quintana, M. Khomiakov, W. Yap, J. Ouyang, K. Ito, Z. Wang, T. Zhao, and F. Biljecki, “Global streetscapes – a comprehensive dataset of 10 million street-level images across 688 cities for urban science and analytics,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 215, pp. 216–238, 2024.
 - [13] G. Astruc, N. Dufour, I. Siglidis, C. Aronsson, N. Bouia, S. Fu, R. Loiseau, V. N. Nguyen, C. Raude, E. Vincent, L. XU, H. Zhou, and L. Landrieu, “Openstreetview-5m: The many roads to global visual geolocation,” 2024. [Online]. Available: <https://arxiv.org/abs/2404.18873>
 - [14] R. Rodrigues and M. Tani, “Are these from the same place? seeing the unseen in cross-view image geolocalization,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3753–3761.
 - [15] S. Lisovski, C. M. Hewson, R. H. Klaassen, F. Korner-Nievergelt, M. W. Kristensen, and S. Hahn, “Geolocation by light: accuracy and precision affected by environmental factors,” *Methods in Ecology and Evolution*, vol. 3, no. 3, pp. 603–612, 2012.
 - [16] J. Kinnari, F. Verdoja, and V. Kyrki, “Season-invariant gnss-denied visual localization for uavs,” *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 232–10 239, 2022.
 - [17] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” 2020. [Online]. Available: <https://arxiv.org/abs/2005.14165>
 - [18] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, “A survey on multimodal large language models,” *National Science Review*, vol. 11, no. 12, Nov. 2024. [Online]. Available: <http://dx.doi.org/10.1093/nsr/nwae403>
 - [19] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” 2021. [Online]. Available: <https://arxiv.org/abs/2005.11401>
 - [20] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, “Retrieval-augmented generation for large language models: A survey,” 2024. [Online]. Available: <https://arxiv.org/abs/2312.10997>
 - [21] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam, “A review on large language models: Architectures, applications, taxonomies, open issues and challenges,” *IEEE Access*, vol. 12, pp. 26 839–26 874, 2024.
 - [22] H. Xu, J. Yuan, A. Zhou, G. Xu, W. Li, X. Ban, and X. Ye, “Genai-powered multi-agent paradigm for smart urban mobility: Opportunities and challenges for integrating large language models (llms) and retrieval-augmented generation (rag) with intelligent transportation systems,” 2024. [Online]. Available: <https://arxiv.org/abs/2409.00494>
 - [23] J. Hays and A. A. Efros, “Im2gps: estimating geographic information from a single image,” in *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 2008, pp. 1–8.
 - [24] T. Weyand, I. Kostrikov, and J. 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*. Springer, 2016, pp. 37–55.
 - [25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
 - [26] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, “Sigmoid loss for language image pre-training,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 975–11 986.
 - [27] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.
 - [28] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, “Dinov2: Learning robust visual features without supervision,” *arXiv preprint arXiv:2304.07193*, 2023.
 - [29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: <http://arxiv.org/abs/1706.03762>

- [30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2019. [Online]. Available: <https://arxiv.org/abs/1810.04805>
- [31] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” 2019. [Online]. Available: <https://arxiv.org/abs/1907.11692>
- [32] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, “Palm: Scaling language modeling with pathways,” 2022. [Online]. Available: <https://arxiv.org/abs/2204.02311>
- [33] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [34] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Łukasz Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Łukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O’Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, and B. Zoph, “Gpt-4 technical report,” 2024. [Online]. Available: <https://arxiv.org/abs/2303.08774>
- [35] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” 2023. [Online]. Available: <https://arxiv.org/abs/2304.08485>
- [36] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, B. Li, P. Luo, T. Lu, Y. Qiao, and J. Dai, “Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks,” 2024. [Online]. Available: <https://arxiv.org/abs/2312.14238>
- [37] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He, Q. Chen, H. Zhou, Z. Zou, H. Zhang, S. Hu, Z. Zheng, J. Zhou, J. Cai, X. Han, G. Zeng, D. Li, Z. Liu, and M. Sun, “Minicpm-v: A gpt-4v level mllm on your phone,” 2024. [Online]. Available: <https://arxiv.org/abs/2408.01800>
- [38] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, “Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond,” 2023. [Online]. Available: <https://arxiv.org/abs/2308.12966>
- [39] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang,

- A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, L. Rantala-Yearly, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenheide, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. E. Tan, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Grattafiori, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Vaughan, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Franco, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Wyatt, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Ozgenel, F. Caggioni, F. Guzmán, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Thattai, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, I. Damla, I. Molybog, I. Tufanov, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Prasad, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Huang, K. Chawla, K. Lakhotia, K. Huang, L. Chen, L. Garg, L. A. L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabza, M. Avalani, M. Bhatt, M. Tsimpoukelli, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. P. Laptev, N. Dong, N. Zhang, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Li, R. Hogan, R. Battey, R. Wang, R. Maheswari, R. Howes, R. Rinott, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Kohler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Albiero, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wang, X. Wu, X. Wang, X. Xia, X. Wu, X. Gao, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Hao, Y. Qian, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, and Z. Zhao, “The llama 3 herd of models,” 2024. [Online]. Available: <https://arxiv.org/abs/2407.21783>
- [40] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show

- and tell: A neural image caption generator,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.
- [41] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.
- [42] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “Vqa: Visual question answering,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.
- [43] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, “Vse++: Improving visual-semantic embeddings with hard negatives,” *arXiv preprint arXiv:1707.05612*, 2017.
- [44] J. Lu, D. Batra, D. Parikh, and S. Lee, “Vilbert: Pre-training task-agnostic visiolinguistic representations for vision-and-language tasks,” *Advances in neural information processing systems*, vol. 32, 2019.
- [45] Z. Zhou, J. Zhang, Z. Guan, M. Hu, N. Lao, L. Mu, S. Li, and 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*, 2024, pp. 2749–2754.
- [46] J. Hays and A. A. Efros, “Large-scale image geolocalization,” *Multimodal location estimation of videos and images*, pp. 41–62, 2015.
- [47] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros, “What makes paris look like paris?” *Communications of the ACM*, vol. 58, no. 12, pp. 103–110, 2015.
- [48] E. Müller-Budack, K. Pustularen, and R. Ewerth, “Geolocation estimation of photos using a hierarchical model and scene classification,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 563–579.
- [49] P. H. Seo, T. Weyand, J. Sim, and B. Han, “Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 536–551.
- [50] S. Pramanick, E. M. Nowara, J. Gleason, C. D. Castillo, and R. Chellappa, “Where in the world is this image? transformer-based geo-localization in the wild,” in *European Conference on Computer Vision*. Springer, 2022, pp. 196–215.
- [51] B. Clark, A. Kerrigan, P. P. Kulkarni, V. V. Cepeda, and M. 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*, 2023, pp. 23 182–23 190.
- [52] S. Zhu, M. Shah, and C. 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*, 2022, pp. 1162–1171.
- [53] J. Lin, Z. Zheng, Z. Zhong, Z. Luo, S. Li, Y. Yang, and N. Sebe, “Joint representation learning and keypoint detection for cross-view geo-localization,” *IEEE Transactions on Image Processing*, vol. 31, pp. 3780–3792, 2022.
- [54] X. Zhang, X. Li, W. Sultani, Y. Zhou, and S. Wshah, “Cross-view geo-localization via learning disentangled geometric layout correspondence,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 3480–3488.
- [55] M. Larson, M. Soleymani, G. Gravier, B. Ionescu, and G. J. Jones, “The benchmarking initiative for multimedia evaluation: Mediaeval 2016,” *IEEE MultiMedia*, vol. 24, no. 1, pp. 93–96, 2017.
- [56] J. Theiner, E. Müller-Budack, and R. Ewerth, “Interpretable semantic photo geolocation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2022, pp. 750–760.
- [57] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, “The faiss library,” 2024.
- [58] J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with gpus,” *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.
- [59] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, “Yfcc100m: The new data in multimedia research,” *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.
- [60] M. A. Aguilar, F. J. Aguilar, M. d. Mar Saldaña, I. Fernández *et al.*, “Geopositioning accuracy assessment of geoeye-1 panchromatic and multispectral imagery,” *Photogrammetric Engineering & Remote Sensing*, vol. 78, no. 3, pp. 247–257, 2012.
- [61] Y. Zhuang, Z. Syed, J. Georgy, and N. El-Sheimy, “Autonomous smartphone-based wifi positioning system by using access points localization and crowdsourcing,” *Pervasive and mobile computing*, vol. 18, pp. 118–136, 2015.
- [62] J. Zhuang, M. Dai, X. Chen, and E. Zheng, “A faster and more effective cross-view matching method of uav and satellite images for uav geolocalization,” *Remote Sensing*, vol. 13, no. 19, 2021. [Online]. Available: <https://www.mdpi.com/2072-4292/13/19/3979>
- [63] M. Zhou, X. Chen, N. Samano, C. Stachniss, and A. Calway, “Efficient localisation using images and open-streetmaps,” in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 5507–5513.
- [64] V. A. Krylov, E. Kenny, and R. Dahyot, “Automatic discovery and geotagging of objects from street view imagery,” *Remote Sensing*, vol. 10, no. 5, 2018. [Online]. Available: <https://www.mdpi.com/2072-4292/10/5/661>
- [65] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.
- [66] L. Haas, M. Skreta, S. Alberti, and C. Finn, “Pigeon: Predicting image geolocations,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

- Recognition (CVPR)*, June 2024, pp. 12 893–12 902.
- [67] V. Vivanco Cepeda, G. K. Nayak, and M. Shah, “Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [68] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, “Openclip,” Jul. 2021, if you use this software, please cite it as below. [Online]. Available: <https://doi.org/10.5281/zenodo.5143773>
- [69] L. Haas, S. Alberti, and M. Skreta, “Learning generalized zero-shot learners for open-domain image geolocalization,” 2023.
- [70] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, “Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution,” 2024. [Online]. Available: <https://arxiv.org/abs/2409.12191>
- [71] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma *et al.*, “How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites,” *arXiv preprint arXiv:2404.16821*, 2024.
- [72] P. Agrawal, S. Antoniak, E. B. Hanna, B. Bout, D. Chaplot, J. Chudnovsky, D. Costa, B. D. Monicault, S. Garg, T. Gervet, S. Ghosh, A. Héliou, P. Jacob, A. Q. Jiang, K. Khandelwal, T. Lacroix, G. Lample, D. L. Casas, T. Lavril, T. L. Scao, A. Lo, W. Marshall, L. Martin, A. Mensch, P. Muddireddy, V. Nemychnikova, M. Pellat, P. V. Platen, N. Raguraman, B. Rozière, A. Sablayrolles, L. Saulnier, R. Sauvestre, W. Shang, R. Soletskyi, L. Stewart, P. Stock, J. Studnia, S. Subramanian, S. Vaze, T. Wang, and S. Yang, “Pixtral 12b,” 2024. [Online]. Available: <https://arxiv.org/abs/2410.07073>
- [73] D. Li, Y. Liu, H. Wu, Y. Wang, Z. Shen, B. Qu, X. Niu, G. Wang, B. Chen, and J. Li, “Aria: An open multimodal native mixture-of-experts model,” 2024. [Online]. Available: <https://arxiv.org/abs/2410.05993>
- [74] “Hugging Face,” <https://huggingface.co/>, accessed: December 8, 2024.
- [75] J. Huang and J. Zhang, “A survey on evaluation of multimodal large language models,” 2024. [Online]. Available: <https://arxiv.org/abs/2408.15769>
- [76] W. Zhang, M. Aljunied, C. Gao, Y. K. Chia, and L. Bing, “M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 5484–5505, 2023.
- [77] P. Xu, W. Shao, K. Zhang, P. Gao, S. Liu, M. Lei, F. Meng, S. Huang, Y. Qiao, and P. Luo, “Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2024.
- [78] B. Li, Y. Ge, Y. Ge, G. Wang, R. Wang, R. Zhang, and Y. Shan, “Seed-bench: Benchmarking multimodal large language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 13 299–13 308.
- [79] J. Lin, J. Tang, H. Tang, S. Yang, W.-M. Chen, W.-C. Wang, G. Xiao, X. Dang, C. Gan, and S. Han, “Awq: Activation-aware weight quantization for llm compression and acceleration,” 2024. [Online]. Available: <https://arxiv.org/abs/2306.00978>
- [80] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, “Gptq: Accurate post-training quantization for generative pre-trained transformers,” 2023. [Online]. Available: <https://arxiv.org/abs/2210.17323>
- [81] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” 2019. [Online]. Available: <https://arxiv.org/abs/1912.01703>
- [82] A. Murray, H. van Kemenade, wiredfool, J. A. Clark, A. Karpinsky, O. Baranovič, C. Gohlke, Yay295, J. Dufresne, M. Brett, DWesl, D. Schmidt, K. Kopachev, A. Houghton, REDxEYE, S. Mani, S. Landey, A. Koskela, J. Ware, vashek, Piolie, J. Douglas, S. T., D. Caro, U. Martinez, S. Kossouho, and R. Lahd, “python-pillow/pillow: 11.0.0,” Oct. 2024. [Online]. Available: <https://doi.org/10.5281/zenodo.13935429>
- [83] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “Huggingface’s transformers: State-of-the-art natural language processing,” 2020. [Online]. Available: <https://arxiv.org/abs/1910.03771>
- [84] T. pandas development team, “pandas-dev/pandas: Pandas,” Feb. 2020. [Online]. Available: <https://doi.org/10.5281/zenodo.3509134>
- [85] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, “Efficient memory management for large language model serving with pagedattention,” in *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.
- [86] L. Contributors, “Lmdeploy: A toolkit for compressing, deploying, and serving llm,” <https://github.com/InternLM/lmdeploy>, 2023.
- [87] OpenAI, “Openai api reference - create chat completion,” <https://platform.openai.com/docs/api-reference/chat/create>, accessed on 23-12-2024.
- [88] vLLM, “Api documentation - sampling parameters,” https://docs.vllm.ai/en/latest/dev/sampling_params.html, accessed on 23-12-2024.
- [89] K. Esmukov and the GeoPy Contributors, “Geopy - geocoding library for python,” <https://github.com/geopy/geopy>, 2014–2024, accessed: November 21, 2024. Contributors include Adam Tygart, Adrián López, Afonso Queiros, Albina, Alessandro Pasotti, Álvaro Mondéjar, Andrea Tosatto, Ann Paul, and many others (see full

contributor list on GitHub).

- [90] L. Li, Y. Ye, B. Jiang, and W. Zeng, “Georeasoner: Geo-localization with reasoning in street views using a large vision-language model,” in *Forty-first International Conference on Machine Learning*.
- [91] M. Wu, Q. Huang, S. Gao, and Z. Zhang, “Mixed land use measurement and mapping with street view images and spatial context-aware prompts via zero-shot multi-modal learning,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 125, p. 103591, 2023.