Large Language Model Empowered Spatio-Visual Queries for Extended Reality Environments

Mohammadmasoud Shabanijou[1], Vidit Sharma[1], Suprio Ray[1], Rongxing Lu[1], Pulei Xiong[2]
[1]University of New Brunswick, Fredericton, Canada
Email: {shabani.m, vidit.sharma, sray, rlu1}@unb.ca
[2]National Research Council (NRC), Canada
Email: pulei.xiong@nrc-cnrc.gc.ca

***Abstract*—With the technological advances in creation and capture of 3D spatial data, new emerging applications are being developed. Digital Twins, metaverse and extended reality (XR) based immersive environments can be enriched by leveraging geocoded 3D spatial data. Unlike 2D spatial queries, queries involving 3D immersive environments need to take the query user's viewpoint into account. Spatio-visual queries return objects that are visible from the user's perspective.**

In this paper, we propose enhancing 3D spatio-visual queries with large language models (LLM). These kinds of queries allow a user to interact with the visible objects using a natural language interface. We have implemented a proof-of-concept prototype and conducted preliminary evaluation. Our results demonstrate the potential of truly interactive immersive environments.

I. Introduction

The volume of three dimensional (3D) spatial data is experiencing a rapid growth due to recent technological advances. This is driven by several factors, including advances in 3D scanning techniques, such as LiDAR or photogrammetry. The fusion of Building Information Modeling (BIM) and three-dimensional Geographic Information Systems (3D GIS) offers new spatial analysis capabilities in a number of areas such as urban planning, transportation, energy and environmental engineering.

Advances in 3D game engine software and the advent of devices such as Microsoft Hololens and Meta Quest have enabled the development of Extended Reality (XR) applications. XR is a broad term that encompasses the experiences of Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR). Whereas, in Virtual Reality (VR) 3D representations of physical objects are used to mimic a real-world environment, in AR a physical environment is enhanced by overlaying computer-generated visual elements. MR enhances AR by providing the ability to interact with the virtual objects that are overlaid onto a physical environment. The integration of 3D GIS and XR ushers in new opportunities for developing immersive 3D spatial environments that will allow the users feel more present in the spatial representation [1]. Such spatial immersion experience can enable realistic simulation in geo-spatial applications, such as performing 'what-if' scenarios by urban planners for city planning, disaster response and greener city management.

XR-enhanced geo-spatial applications also entail spatial queries that are different from traditional spatial queries. The incorporation of *visibility* with spatial queries along with user trajectories opens the door to new types of queries, called *spatio-visual* queries. In such cases, the fidelity of the virtual environments has to considered, where some background objects may be occluded. Depending on the viewpoint of the user, only objects that are visible need to be accessed and objects further away from the viewpoint may be approximated by coarser representations than those that are nearer. An example spatio-visual query is *viewpoint query* [2] that returns all objects that are visible from the query viewpoint. Based on the movement of a viewpoint, a walkthrough application will continuously refresh the set of visible objects as the viewpoint moves. Another such query is *visible reverse nearest neighbor*r (VRNN) search [3], which considers the impact of obstacles on the visibility of objects. The *spatio-visual keyword* (SVK) query [4] is designed to support retrieving spatial Web objects that are both visually conspicuous and relevant to the user.

Spatio-visual queries offer new ways to retrieve information from XR-enhanced 3D spatial environments. However, their applicability is still somewhat limited, since the result-sets retrieved by these queries are static spatial objects. These queries cannot deal with complex questions posed in a natural language, which could be quite powerful. For instance, research in digital civics and urban planning [5] advocates for citizens participation in urban planning by empowering them to provide feedback.

Recently, pre-trained large language models (LLMs) [6] have attracted a lot of attention from the research community. One of the areas that is yet to be explored by the researchers is the full potential of LLMs in geo-spatial tasks. There have been a few studies that investigated geo-spatial capabilities of LLMs. For instance, a recent work [7] tried to determine to what extent geospatial knowledge, awareness, and reasoning abilities are enbedded within pretrained LLMs.

Due to the growing interest in XR environments and LLMs, our research focuses on an intersection of these two

Note: This is the authors' version of the work. Not for redistribution.

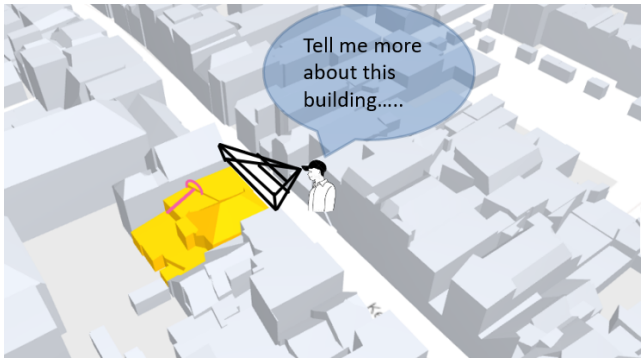

Fig. 1: A use case scenario

areas. To that end, we propose to enhance 3D spatio-visual queries with natural language processing capabilities based on LLMs. In this project, we have developed a prototype to demonstrate a XR environment in which a user can walk through a city and pose spatio-visual queries about some city buildings. Then the user can ask a series of prompt queries in natural language about any building. To efficiently process spatio-visual queries and calculate visibility of city objects in particular, we utilize HDoV-tree [8] index structure. We use Gemma2b [9] as our large language model. To our knowledge, ours in the first attempt to enhance spatio-visual queries with LLM-based natural language query capabilities. We believe that our system can be the basis of many smart cities applications.

The remainder of this paper is organized as follows. In Section II we describe our methodology. Specifically, we illustrate a use case scenario and then describe the components of our system. Then we outline the evaluation of our system in Section III, where we describe the dataset, query set, LLM model and fine-tuning process, and evaluation results. Finally, we conclude the paper in Section IV

II. Methodology

A. Use case scenario and problems addressed

A use case is illustrated in Figure 1, in which an urban tourist is visiting a new city in-person or virtually. The figure shows an example bounding box that pertains to a sample viewpoint query that a pedestrian might have and the visible building (highlighted) that that user might see from that specific point of view. We assume that the user is equipped with either a head-mounted display (HMD) device and participating in a 3D VR environment based application. Or the user may be equipped with a mixed-reality wearable, such as Meta Smart Glasses and actively involved in a MR environment. In this use case, the user is moving along a street (in a VR or MR environment) and directing her vision forward to a part of the urban landscape. Based on her interest in a particular area of her viewpoint, she can trigger a spatio-visual range query, which will return all objects that are not occluded from her perspective. Then, she can initiate an interactive conversation in natural language regarding any of the objects returned by the spatio-visual range query. For instance,

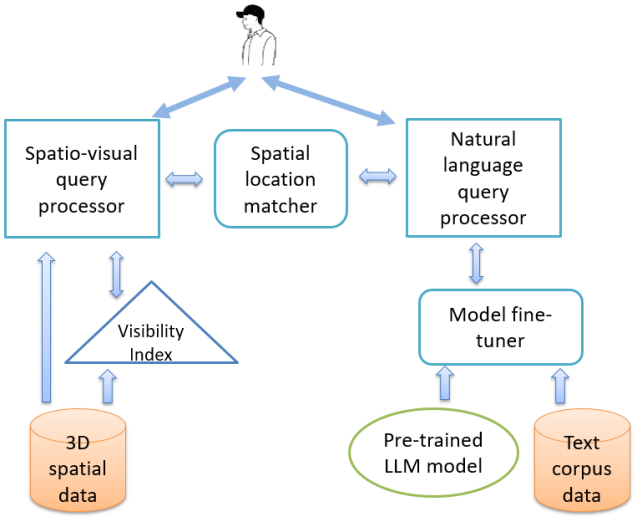

Fig. 2: Our system architecture

she can ask a series of questions such as, “Tell me more about this building”, “Is this a hotel?”, “Can I book a hotel room?”, “Tell me the phone number of the hotel reservation” and so on.

B. Our system

We have implemented a system to support the aforementioned use case. The high-level architecture of our system is shown in Figure 2. We describe the system components next.

1) Spatio-visual query processor: The spatio-visual query processor handles spatial viewpoint queries, particularly, spatio-visual range queries. A spatial visual range query takes as input a spatial viewpoint, a query rectangle, and a direction. In response, it returns all spatial objects that are visible from the given viewpoint. The query dataset can be any 3D spatial dataset, but typically urban 3D building datasets are common for our application. Processing a spatio-visual range query can be computationally expensive. To efficiently process such a query we build a visibility index described next.

2) Visibility Index: To support efficient retrieval of objects based on visibility, an efficient technique is needed that can avoid full scan of all the objects during query time. One option is to partition the user viewpoint space *a priori* into disjoint cells, and maintain visible objects for each such cells. However, this approach may entail a large number of objects to be loaded. To address this issue, an index structure called HDoV-tree [8] was proposed. HDoV-tree is an extension of R-tree that maintains visibility information of the objects along with their internal level-of-details (LoDs).

For our system, we leverage the HDoV-tree to index the 3D spatial objects. When a user issues a spatio-visual query, the index is utilized to obtain objects that are visible from the query’s viewpoint. Any object that is occluded by another object from a user’s perspective is excluded.

3) Spatial location matcher: A user can ask natural lanugage based prompt queries on any object returned by the spatio-visual query processor. The spatial location matcher enhances spatial context-awareness of the natural language query processor.

4) Natural language query processor: Natural language query processor enables users to have interaction with viewpoint query objects in a natural language. It utilizes advanced natural language processing (NLP) techniques to parse the queries from the user and process them using pre-trained LLMs.

5) Model fine-tuner: Model fine-tuning has emerged as a powerful technique that allows a pre-trained model to be adapted to a specific dataset or domain. This process involves of further training a pre-trained model on a specific data or domain [10]. Recently there have been attempts towards exploring different strategies for this technique, such as layer-wise learning rate decay [11], and prompt-based fine-tuning [12], which have shown to be capable of improving the model performance and accuracy. In addition to these techniques, approaches such as few-shot learning and meta-learning have been introduced for situations with limited labeled data. In our system, the model fine-tuner is used to adapt a LLM model with the domain knowledge, which in our case, is the text corpus concerning the objects in an urban environment.

III. Evaluation

A. Experimental setup

We used a Windows machine, having an Intel(R) Core(TM) i9-9880H CPU clocked at 2.30GHz and 64 GB of main memory. We used PostgreSQL with PostGIS to store the 3D spatial dataset. The spatio-visual query processor and HDoV-tree index were implemented in Java. The LLM-based prompt query answering component was implemented in Python.

B. Dataset

For the 3D spatial data, we used the DenHaag data from the Topography-of-the-netherlands-in-3d dataset [13]. It contains of 3D roof shapes and a 2.5D terrain model of the Municipality of The Hague in the Netherlands. This dataset includes a total of 100,000 building models within the Municipality of The Hague and approximately 12,500 models in neighboring municipalities. The dataset includes Level of Detail 2 (LOD2) data, that is used for the calculation of the visibility by the HDoV-tree index.

For the text corpus data, we used a dataset containing information about buildings in The Hague [14]. This dataset contains information about different building properties in The Hague and additional details about their homeowners. We used this dataset to fine-tune our LLM model based on that.

C. Queries

We used a random bounding rectangle generator to generate a number of spatio-visual range queries as example viewpoint queries that a hypothetical user might ask in a real-world scenario. These queries are processed by the spatio-visual query processor. Each of these queries returns all buildings that are inside its query rectangle. Then, based on the visibility score assigned to each building by the visibility index, we remove the buildings that we consider occluded and pass the information about the buildings visible from the user's point of view into the natural language query processor component.

We generate a sequence of prompt queries as example natural language queries that a user may ask about the visible buildings. These prompt queries are then processed by the natural language query processor.

D. LLM model and fine-tuning

We chose Gemma2b [9] as our large language model component to conduct our experiments. The main reason for this choice was that this model was lighter and less computationally intensive compared to their 7B parameter model counterparts. In order to finish our experiments in time, we prioritized speed over model performance. However, we plan to use a better model (such as Mistral7B [15] or other more powerful models) in the future. Some of the fine-tuning parameters included the number of epochs= 1, batch size = 4, and maximum steps = 100.

E. Prototype evaluation and results

We executed the spatio-visual queries and their corresponding prompt queries (described in Section III-C) several times. In Figure 3 we show a series of prompt queries about a building in The Hague and the best responses received.

For some of the prompt queries, the model returned relevant and straightforward answers. Whereas for some of the queries, it did not return relevant answers, as it seemed that the model was hallucinating at some points. Given the fact that we leaned towards the speed in fine-tuning the model (in the interest of time), this seemed reasonable. As a future work, we plan to conduct experiments with more powerful models and dedicate more time and effort for fine-tuning, which would result in better performance and less hallucinating.

Prompt1

"What is the address of Ocean House Scheveningen?"

Response1

"Scheveningslaan 49, 2587GT, The Hague (The Netherlands)."

Prompt2

"Can you tell me the phone number of this?"

Response2

"+31 7034207067."

Prompt3

"Is Ocean House Scheveningen a hotel"

Response3

"Hotel yes, because you may stay in one of the 51 rooms of this impressive building. Ocean House is a luxury hotel that consists of 12 floors and is located close to one of the most impressive buildings in The Hague, the Parliament of the House of Representatives."

Prompt4

"Can you tell me some more information about Ocean House Scheveningen?"

Response4

"This magnificent property was built on the dunes for Lord Stanley, head of the Royal Household, and was extensively renovated to create a grand ocean-front hotel that would become the first of its kind in the Netherlands, serving a diverse and exclusive clientele that included European Royalty."

Prompt5

"Can you provide some user reviews about Ocean House Scheveningen?"

Response5

"We were just in Scheveningen and have to say we were disappointed. We read the previous reviews and are confused about the mix of reviews. Here's what we thought: The location: We stayed in room 646, which we thought was a good location: close to the beach, close enough to be able to walk to a bunch of really excellent restaurants, close"

Fig. 3: Prompt queries and responses

IV. Conclusion

With the proliferation of 3D spatial data, and recent advances in extended reality (XR) and large language model (LLM) technologies, opportunities arise for developing new kinds of spatio-visual queries with natural language capabilities. We have proposed empowering spatio-visual queries with large language model based reasoning capabilities. We have developed a prototype system as a proof-o-concept.

One of the goals of this project is to develop a state-of-the-art system that enables users to interact with smart city environments using natural language and XR interfaces. Our initial results show that our system is capable of providing valuable information and can assist users in a smart city environment, As a future work, we plan to further explore and expand the scope of this system.

Acknowledgment

This project was supported in part by collaborative research funding from the National Research Council of Canada's Digital Health and Geospatial Analytics Program.

We thank Minh Duc Nguyen. Ngoc Phuong Anh Nguyen, Viet Anh Nguyen and Ronnit Peter for their contributions.

References

[1] J. Keil, D. Edler, T. Schmitt, and F. Dickmann, "Creating Immersive Virtual Environments Based on Open Geospatial Data and Game Engines," *Journal of Cartography and Geographic Information*, pp. 53–65, 2021.

[2] L. Shou, Z. Huang, and K.-L. Tan, "The hierarchical degree-of-visibility tree," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1357–1369, 2004.

[3] Y. Gao, B. Zheng, G. Chen, W. Lee, K. C. K. Lee, and Q. Li, "Visible reverse k-nearest neighbor queries," in *ICDE*, 2009, pp. 1203–1206.

[4] C. Zhang, L. Shou, K. Chen, and G. Chen, "See-to-retrieve: efficient processing of spatio-visual keyword queries," in *SIGIR*, 2012, p. 681–690.

[5] Z. Zhou, Y. Lin, D. Jin, and Y. Li, "Large language model for participatory urban planning," 2024. [Online]. Available: https://arxiv.org/abs/2402.17161

[6] W. X. Zhao *et al.*, "A survey of large language models," 2023. [Online]. Available: https://arxiv.org/abs/2303.18223

[7] P. Bhandari, A. Anastasopoulos, and D. Pfoser, "Are large language models geospatially knowledgeable?" ser. SIGSPATIAL.

[8] L. Shou, Z. Huang, and K.-L. Tan, "Hdov-tree: the structure, the storage, the speed," in *ICDE*, 2003, pp. 557–568.

[9] "Gemma2b," 2024. [Online]. Available: https://huggingface.co/google/gemma-2b

[10] M. E. Peters, S. Ruder, and N. A. Smith, "To tune or not to tune? adapting pretrained representations to diverse tasks," *arXiv preprint arXiv:1903.05987*, 2019.

[11] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," *arXiv preprint arXiv:1908.03265*, 2019.

[12] T. B. Brown, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[13] "Topography-of-the-netherlands-in-3d," 2024. [Online]. Available: https://www.cityjson.org/datasets/the-topography-of-the-netherlands-in-3d

[14] "Dataplatform portal," https://denhaag.dataplatform.nl//home, accessed: 2024-9-7.

[15] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.