Utilizing Large Language Models for Indoor Tour Guidance

Zhiqiang Jiang
zhiqiang.jiang@ucalgary.ca
University of Calgary
Calgary, Canada

Isaac Huang
isaac.huang@uwaterloo.ca
University of Waterloo
Waterloo, Canada

Xin Wang*
xcwang@ucalgary.ca
University of Calgary
Calgary, Canada

ABSTRACT

Most indoor venues depend heavily on human tour guides or information desk staff for indoor tour guiding services due to the lack of efficient and effective representation of indoor layouts, support for dynamic tour requests and natural language interaction. This paper addresses the challenges of indoor tour guiding by using large language models (LLMs) to automate three core tasks: pathfinding, tour planning, and tour Q&A through the introduction of IndoorRoaming, an LLM-based indoor tour guidance system. This system leverages an Indoor Entity Graph to bolster LLMs' precise comprehension of indoor environments. It integrates a State-Driven Dynamic Planning method for interactive and adaptive tour planning. Additionally, IndoorRoaming employs an Object Labelling Workflow to enhance the multi-modal capabilities of LLMs, enabling them to handle tour visual question-answering tasks more effectively. Instruction tuning and fine-tuning of LLMs are designed to align with the system functions, producing indoor tour guidance comparable to that of human tour guides. The performance of IndoorRoaming has been evaluated through extensive testing in a large shopping mall (CF Market Mall) and a museum (Studio Bell Museum). A quantitative user study demonstrates the effectiveness of IndoorRoaming in terms of accuracy, generation quality, and user satisfaction across all implemented tasks.

CCS CONCEPTS

• **Computing methodologies → Spatial and physical reasoning**; **Natural language generation**.

KEYWORDS

Large Language Models, Indoor Spatial Data Modeling, Qualitative Spatial Representation, Personalized Geospatial Recommendation Systems

ACM Reference Format:
Zhiqiang Jiang, Isaac Huang, and Xin Wang. 2024. Utilizing Large Language Models for Indoor Tour Guidance. In *Proceedings of 4th SIGKDD Workshop on Deep Learning for Spatio-temporal Data, Applications, and Systems (DEEPSPATIAL '24).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Most indoor venues rely on human tour guides or information desk staff to support indoor tour guiding services due to the limitations of indoor GPS positioning [2] and the absence of standardized indoor mapping system [7]. A tour guide or information desk staff is able to adequately fulfill the indoor guiding service without depending on any indoor positioning or digital map because of familiarity with the venue spatial environment and guiding information.

Large language models (LLMs), such as ChatGPT [11], Gemini [6] and LLaMA [10], with their profound capacity for natural language understanding, context-awareness, generative abilities, and multilingual support, have the potential to provide a user-friendly interactive tour experience for visitors. The goal of this paper is to develop an LLM-based indoor tour guidance system that employs large language models to handle three core tasks – pathfinding, tour planning, and tour question-answering – within the context of indoor tourism environments such as museums and shopping malls. However, there are a few challenges:

Firstly, the pathfinding task presents a significant hurdle for an LLM-based tour guide system, as it requires a comprehensive understanding of the indoor venue's topology, the ability to identify the shortest path, and the generation of accurate and coherent route guidance language. While LLMs possess remarkable natural language processing capabilities, they require additional support to fully comprehend the intricate spatial relationships and topological structures of indoor environments.

Secondly, the tour planning task involves recording and maintaining a recommended tour plan, including both routes and accompanying commentary. Specifically, the task needs to track the progress of the tour plan, generating relevant commentary at each step, and dynamically modify subsequent portions of the plan based on visitor inputs or preferences. Developing a computational model that can support an LLM-based system in implementing these functions coherently and responsively remains a a formidable undertaking.

Thirdly, the tour Q&A task necessitates the ability to learn domain-specific knowledge and engage in interactive question-answering sessions with visitors. While multimodal large language models (MLLMs), such as LLaVA[3][4], excel in acquiring both textual and visual domain knowledge, they require additional support to handle visual tour Q&A interactions effectively, particularly when dealing with clear spatial references. One key challenge in this task is the accurate detection and labeling of objects to address spatial reference issues in visual Q&A scenarios. Despite remarkable progress in computer vision and natural language processing, enabling MLLMs to seamlessly integrate spatial information from visual inputs and provide visitors with precise responses remains a significant challenge.

In this paper, an LLM-based tour guidance system called IndoorRoaming, is proposed to support the implementation of the three

---

[*] Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

DEEPSPATIAL '24, August 25-26, 2024, Barcelona, Spain
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

core tasks outlined above. Our main contributions are summarized as follows:

- To enable LLMs to accurately comprehend indoor topology and generate precise pathfinding instructions, we introduce the Indoor Entity Graph. This graph serves as the foundation for LLMs to infer the shortest path and generate natural language route guidance for the pathfinding task. The Indoor Entity Graph comprises detailed node representations for various indoor entities, as well as edge representations that capture both visible corridors and strategic connections within the indoor space. Furthermore, the graph incorporates turn representations that explicitly denote the directions required at intersections, enhancing LLMs' ability to provide clear and accurate pathfinding instructions to visitors.
- We introduce the State-Driven Dynamic Planning method to enable LLMs to accurately perform the indoor tour planning task. This method represents each Point of Interest (POI) as an independent state, treating visitor movements as transitions between these states. By maintaining a state-based representation of the tour plan, the method allows LLMs to generate contextually relevant commentary and modify the subsequent portions of the tour route in response to visitor inputs or preferences, facilitating a seamless and adaptive indoor tour experience.
- To address the spatial reference challenges in visual tour Q&A interactions with visitors, we have designed an Object Labeling Workflow for MLLMs. The workflow integrates an additional object detection model and dynamically adjusts object labels based on visitor inquiries, ensuring the accuracy and relevance of object identifications. This adaptive labeling process is crucial for answering queries that require a comprehensive understanding of the spatial arrangement of elements within a visual input. It enables MLLMs to provide precise and contextually relevant responses during the tour Q&A sessions.
- We evaluated IndoorRoaming to assess its performance in providing indoor tour guidance services for a large shopping mall and a museum, demonstrating the great promise of IndoorRoaming for application in real-world scenarios. Our results show that IndoorRoaming effectively enhances visitor experiences through accurate guidance and tailored information dissemination

The remainder of this paper is organized as follows: Section 2 discusses the related work. Section 3 introduces the system overview and our methodology on each task. Section 4 provides our case studies of the IndoorRoaming implementation. Section 5 evaluates the effectiveness of these initiatives. Section 6 closes with a conclusion and future work.

2 RELATED WORK

Limited research works have been conducted using LLMs in spatial systems. MapGPT [12] supports existing outdoor digital maps service to generalize across complex query intents and invoke suitable chains of digital maps services. It focuses on how LLMs support outdoor digital map services in understanding user query intents. However, MapGPT does not target the indoor environment where GPS positioning and digital map services are unavailable. There is one notable research reference using LLMs to understand map spatial information: the standout AI experiment, Stanford Smallville [8]. This paper clearly elaborates that Smallville sandbox environment is represented and operates as a tree data structure with the Node and Edge relationship. Although the Stanford Smallville research focuses on the interactive simulation of human behavior with LLMs, the experiment proves that LLMs is able to learn map spatial relationship if its topology, distance, and direction are all clearly defined with a computational model. However, the experiment is performed in a game environment without any real-world implementation. Another important spatial-related LLMs research field is robotic navigation and autopilot. The work by Dorbala et al. [9] explores how LLMs generate wayfinding instructions based on the egocentric images taken by embodied robot along indoor paths, which inspires us to extend our IndoorRoaming to support indoor embodied robot hardware as future work.

3 METHODOLOGY

3.1 System Overview

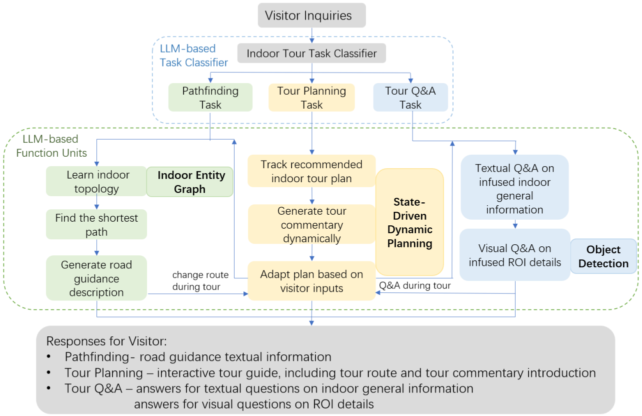

Visitor Inquiries
LLM-based Task Classifier
Indoor Tour Task Classifier
Pathfinding Task
Tour Planning Task
Tour Q&A Task
LLM-based Function Units
Learn indoor topology
Indoor Entity Graph
Find the shortest path
Generate road guidance description
change route during tour
Track recommended indoor tour plan
Generate tour commentary dynamically
Adapt plan based on visitor inputs
State-Driven Dynamic Planning
Q&A during tour
Textual Q&A on infused indoor general information
Visual Q&A on infused ROI details
Object Detection
Responses for Visitor:
Pathfinding- road guidance textual information
Tour Planning – interactive tour guide, including tour route and tour commentary introduction
Tour Q&A – answers for textual questions on indoor general information
answers for visual questions on ROI details

Figure 1: IndoorRoaming System Overview

Figure 1 shows the overview of the proposed IndoorRoaming system. The system comprises two primary LLM-based components: task classifier and function units. The indoor tour task classifier leverages LLMs to categorize visitor inquiries into one of three core tasks: pathfinding, tour planning, or tour Q&A. By understanding the visitor's intent, this module routes the inquiry to the appropriate function units. The LLM-based function units handle the specific requirements of each core task.

- **Pathfinding Task:** The proposed Indoor Entity Graph enables LLMs to learn the indoor topology of a venue. Then IndoorRoaming can find the shortest path and generate detailed textual road guidance for visitors' pathfinding request.
- **Tour Planning Task:** The task tracks a recommended indoor tour plan and dynamically generates commentary tailored to the visitor's interests and interactions. The proposed State-Driven Dynamic Planning (SDDP) method enables IndoorRoaming to adjust the tour plan on the fly, providing a

responsive and engaging visitor experience. When a visitor requests a route change or poses a question during the tour, IndoorRoaming uses the function units of Pathfinding or Tour Q&A to meet the visitor's needs and preferences.

- **Tour Q&A Task:** This task enriches the visitor experience by providing detailed responses to any visitor questions about the indoor tour. The Textual Q&A function unit supports inquiries about indoor general information. Detailed visual information about Region of Interest (ROI) learned by MLLMs is handled by the Visual Q&A function unit and an integrated object detection model, ensuring clear spatial referencing.

In the following, we will discuss each task in detail.

3.2 Pathfinding Task

The pathfinding task requires accurately describing the indoor path from point A to point B. To assist LLMs in understanding indoor layout, we propose an Indoor Entity Graph including Nodes, Edges and Turns to represent the spatial entities, their topological and distance relationships.

Definition 1 (Indoor Entity Graph): An indoor entity graph G is an ordered triple $G = (N, E, T)$, where

- N represents a set of units represent indoor location entities such as rooms, stores, exhibitions or general any indoor object of interest. Besides, intersections of corridors are regarded as landmark type of nodes in an indoor entity graph. Each node has attributes including its ID, name, type, located corridor, distances to both start and end intersection nodes of the located corridor, and the side on its Corridor according to the direction from start to end node. Specifically, N is represented as:

$$N = \{n_i \mid n_i = (id, name, type, corridor, side, distanceFromStart, distanceToEnd)\} \quad (1)$$

- E represents a set of connections between nodes. Note that every edge is bidirectional. Moreover, edges include not only all connections between intersection nodes, which are visible corridors on the indoor map, but also two connections from each node to the start and end intersection nodes of its located corridor, which are invisible on the indoor map. Specifically, E is represented as:

$$E = \{e_k \mid e_k = (id, startNode, endNode, length)\} \quad (2)$$

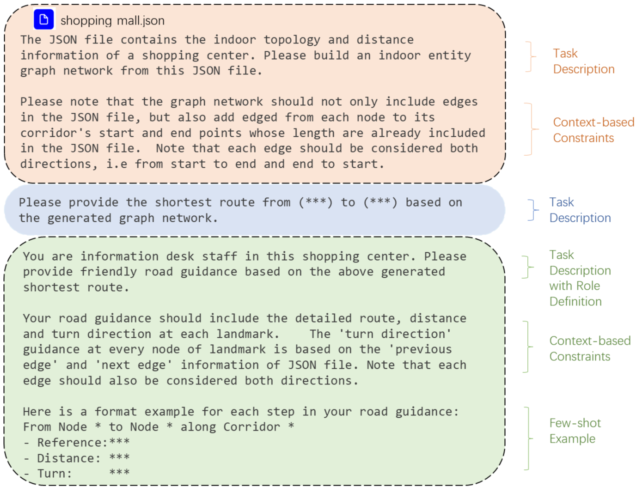
- T represents as set of turning directions at each intersection node. For a directional path, each turn item includes the previous edge, the next Edge, and the turning direction from the previous Edge to the next Edge. Specifically, T is represented as:

$$T = \{t_j \mid t_j = (id, previousEdge, nextEdge, turnDirection)\} \quad (3)$$

Each edge links two nodes, and each turn is associated with transitions between edges, essential for defining movement paths. Given a source node s and a destination node t, the goal of the pathfinding task is to find the shortest path P^* from s to t.

$$P^* = \arg\min_{P \in \mathcal{P}(s,t)} (D(P) = \sum_{(u,v) \in P} d(u,v)) \quad (4)$$

, where $\mathcal{P}(s,t)$ represents all possible paths from source s to target t. $D(P)$ represents the total distance of a path P. $d(u,v)$represents the distance of each segment along the path from start to finish.


shopping mall.json
Task Description
Context-based Constraints
Task Description
Task Description with Role Definition
Context-based Constraints
Few-shot Example

```
The JSON file contains the indoor topology and distance
information of a shopping center. Please build an indoor entity
graph network from this JSON file.

Please note that the graph network should not only include edges
in the JSON file, but also add edged from each node to its
corridor's start and end points whose length are already included
in the JSON file.  Note that each edge should be considered both
directions, i.e from start to end and end to start.
```

```
Please provide the shortest route from (***) to (***) based on
the generated graph network.
```

```
You are information desk staff in this shopping center. Please
provide friendly road guidance based on the above generated
shortest route.

Your road guidance should include the detailed route, distance
and turn direction at each landmark.    The 'turn direction'
guidance at every node of landmark is based on the 'previous
edge' and 'next edge' information of JSON file. Note that each
edge should also be considered both directions.

Here is a format example for each step in your road guidance:
From Node * to Node * along Corridor *
- Reference:***
- Distance: ***
- Turn:     ***
```

Figure 2: Prompt Example for Pathfinding Task

To ensure IndoorRoaming provides pathfinding guidance akin to a human tour guide, instruction tuning prompts are designed. The prompt templates and techniques are presented in Figure 2. The prompt typically includes a task description with role definition, context-based constraints, followed by a few examples for few-shot learning. This structure helps the model understand the context and apply its learned capabilities to similar new tasks. Chain-Of-Thought technique is applied to enhance the interpretability and logical reasoning by breaking down the whole task into three steps:

- The first prompt directs the LLM to extract the indoor topology information from an uploaded JSON file and build an indoor entity graph for a venue, such as a shopping center as shown in Figure 4. It specifically instructs the addition of extra edges to construct a complete indoor entity graph.
- The second prompt requests the calculation of the shortest route within the created indoor entity graph, demonstrating the practical application of the shortest path searching algorithms such as Dijkstra.
- The third prompt starts with task description with the role definition for LLMs, such as placing the LLM in the role of an information desk staff member at a shopping center, and asking to provide friendly road guidance based on the generated shortest route. The context-based constraints guides the LLM to interprete the graph data to provide actionable instructions, simulating a real-world customer service scenario. The few-shot example format provided emphasizes precision and attention to detail, which are crucial for effective communication in pathfinding tasks.

3.3 Tour Planning Task

A typical tour plan includes a designated route through multiple Points of Interest (POIs), accompanied by specific commentary for each POI. Indoor venues, such as museums, usually offer a recommended tour plan, which the indoor tour guide can adjust based on visitor requests during the tour. To enable LLMs to adapt the tour route and corresponding commentary changes, we propose the State-Driven Dynamic Planning (SDDP) method. This method helps LLMs track tour progress and make adjustments to the route or commentary, while maintaining coherence with the existing plan. In the indoor scenario of a museum, nodes of indoor entity graph represent POIs. In the SDDP method, each POI is considered an independent state, and the visitor's movements from one POI to another are perceived as transitions between states. Any route changes requested by the visitor can be implemented by adding the necessary transitions and removing obsolete ones on the fly. Additionally, the SDDP method supports entry-state action definition, allowing the tour commentary for each POI to be dynamically generated based on the visitor's feedback regarding their interests and preferred commentary content.

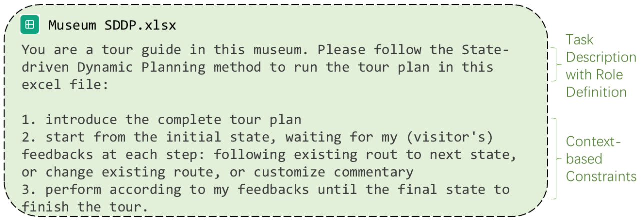
The SDDP method process is defined with a tuple (S, V, δ, C) and Algorithm 1:

- S is the set of states such as POIs, with s_0 being the initial state such as the museum entrance or lobby, and s_f as the final state such as the museum exit.
- V is the set of visitor inputs, with V_0 including initial requirements such as tour duration, language, and commentary content, which can be changed during the tour.
- $\delta : S \times V \longrightarrow S$ is the state transition function dependent on visitor input.
- $C : S \times V \longrightarrow$ Text is the dynamic commentary function.

Algorithm 1 Dynamic Tour and Commentary Adjustment (SDDP)

```
Require: Tour Plan initial state $s_0$, final state $s_f$, Visitor's initial
    input $V_0$, model components $\langle S, V, \delta, C \rangle$
  $V \longleftarrow V_0$
  Counter $\longleftarrow$ 0
  $s \longleftarrow s_0$
  while $s \notin F$ do
    Commentary $\longleftarrow C(s, V)$ {Generate commentary based on current state and visitor input}
    Present Commentary to Visitor and Collect Feedback
    $V \longleftarrow$ Update visitor inputs
    if there is a request for change in route then
      updateTransition$(s, s')$
      $s' \longleftarrow \delta(s, V)$ {Determine next state based on updated route}
    end if
    Counter $\longleftarrow$ Counter + 1
  end while
  return $s, V$
```

The algorithm dynamically customizes both the tour path and the commentary based on real-time visitor feedback. It starts with initial visitor preferences and the tour's starting point, and through continuous feedback loops, it adjusts the visitor's experience by updating the tour route and commentary in response to changes in visitor preferences. The process continues until the tour reaches its planned end, ensuring a personalized and engaging experience throughout the visit.



```
Museum SDDP.xlsx
You are a tour guide in this museum. Please follow the State-
driven Dynamic Planning method to run the tour plan in this
excel file:

1. introduce the complete tour plan
2. start from the initial state, waiting for my (visitor's)
feedbacks at each step: following existing rout to next state,
or change existing route, or customize commentary
3. perform according to my feedbacks until the final state to
finish the tour.
```

Task Description with Role Definition
Context-based Constraints

Figure 3: Prompt Example for Running Tour Plan Using SDDP

The LLM is instruct-tuned using specially designed prompts, as illustrated in Figure 3. The prompt begins by establishing the role of the LLMs as a tour guide, and is followed by three key instructions:

Introduction of the Plan The LLM-based guide is requested to provide an introduction to the tour, offering visitors a clear overview of the recommended tour plan.

Interactive Engagement The LLM-based guide is instructed to operate according to Algorithm 1, starting from the initial state and awaiting the visitor's feedback at each step. This interactive element is essential for making the tour adaptable and personalized.

Continuous Adaptation The prompt specifies that the guide should adapt the tour based on the visitor's feedback. Possible actions include: following the existing route to the next state, changing the route, or customizing the commentary. The guide is instructed to continue until reaching the final state as defined in Algorithm 1.

3.4 Tour Q&A Task

This task involves addressing visitor inquiries about the indoor venue, including information on buildings, floors, and venue management rules. Fine-tuning LLMs is essential to equip them with the necessary domain-specific knowledge to respond to visitors' questions effectively. After this knowledge infusion, LLMs excel at providing satisfactory Textual Q&A service concerning general indoor information, such as floor introductions, opening hours and events. Besides the explanatory general information, narratives of the details of the individual exhibits constitute a crucial component of indoor tour guiding service; however, textual descriptions often do not convey the richness that visual representation can offer. Thus, Multimodal Large Language Models (MLLMs) are adopted for visual Q&A about exhibit details. Open-source MLLMs, such as LLaVA, are recommended for executing these visual Q&A tasks. This approach aligns with the requirements of indoor venues, which necessitate processing domain-specific image data using private MLLMs on local servers to ensure data safety.

During an indoor tour, current MLLMs often face a spatial-reference issue with visual Q&A inquiries. For example, when visitors ask MLLMs a common question like 'Who is this person?' at an image, they cannot point directly at the person as they would with a human tour guide. Such spatial cues are crucial for accurate interpretation but are absent in digital exchanges, rendering MLLMs incapable of effectively resolving these interactive queries. An Object Labeling Workflow is proposed to address the spatial reference issue as follows:

(1) Identify visual questions that involve spatial reference ambiguity.
(2) Employs Object Detection model to detect and label multiple objects within different Regions of Interest (ROIs). In the context of computer vision, "Region of Interest" (ROI) refers to a specific portion of an image that is targeted for a particular analysis or operation. When visitors ask questions related to specific visual elements, identifying and analyzing the appropriate ROI ensures that the responses are precise and relevant to the queried subject matter. This is particularly useful in dense or detailed images where multiple subjects might be present.
(3) Upon receiving the supplementary ROI label from the visitor, the MLLM's inference method is activated, with inputs including the labeled image and a consolidated question with the ROI information.
(4) Outputs the MLLM's inference answer for visual question concerning the specified object.

4 CASE STUDY

In this section, we evaluate the IndoorRoaming system in two typical indoor environments: shopping mall and museum.

4.1 Pathfinding in CF Market Mall

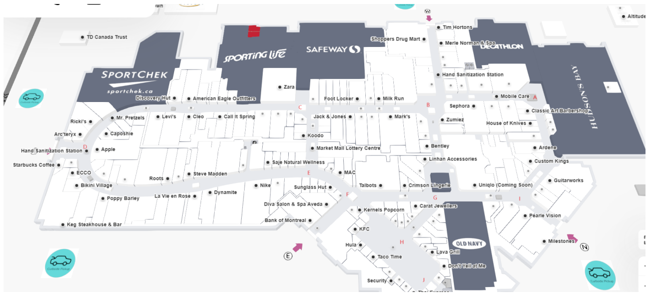


Figure 4: CF Market Mall Indoor Map

CF Market Mall, located in Calgary, Alberta, is one of the largest malls in the area with different stores, restaurants, and entertainment options. It serves as a shopping mall example to illustrate how to instruct LLMs to learn an indoor layout through the Indoor Entity Graph. Based on its indoor map shown in Figure 4, we first extract all stores and corridors information from the map. Then the information is represented as the 71 nodes, 17 edges and 66 turns as shown in Figure 5. Note that the nodes of the intersection type are labeled as Letters A to J on the indoor map.

With these spatial data, the LLM is instructed to build the indoor entity graph, apply Dijkstra's algorithm to calculate the shortest route according to visitor pathfinding requests, and to generate accurate road guidance descriptions. The instruction tuning and execution of the CF Market Mall pathfinding task are detailed in Table 1. It shows that the final generated road guidance includes accurate references along the route, based on the prompt's few-shot example, demonstrating the LLM's comprehensive understanding of indoor spatial relationships at CF Market Mall.

Node ID	Name	Type	Corridor	Side	Distance From Start	Distance To End
1	Tim Hortons	Store	9	Left	5	95
2	Shoppers Drug Mart	Store	9	Right	10	90
...	...	...	...	...	...	...
71	Intersection J	Intersection	–	–	–	–

Edge ID	StartNode	EndNode	Length
1	70	63	200
2	63	64	100
...	...	...	...
17	69	71	20

Turn ID	Previous Edge	Next Edge	Turn Direction
1	1	9	Right
2	9	2	Right
...	...	...	...
66	7	1	Left

Figure 5: Spatial Data of CF Market Mall

4.2 Tour Planning in Studio Bell Museum

The Studio Bell Museum, known as the Canadian National Music Centre, is a museum featuring five floors of exhibitions, interactive instrument installations, and a vast collection of musical artifacts. It serves as an example to illustrate the LLM-based tour planning method.

Figure 6 presents the state transitions diagram of an SDDP tour example at the Studio Bell Museum, detailing the following steps:

(1) Music Square Entrance-Level 1: The tour enters the initial state by starting with a welcome and tour overview commentary. If the visitor follows the current plan, the system displays the route to the next showroom.
(2) Trailblazer Showroom-Level 2: This area generates commentary for State 2. If a visitor requests more interesting commentary with stories, the system adapts the commentary accordingly. Following the plan, the system shows the route to the next showroom.
(3) Music and Wellness Showroom-Level 3: Here, the system adapts the commentary style and generates commentary for State 3. If the visitor wishes to bypass this level, the system changes the route and shows the way directly to the final state.
(4) Country Music Showroom-Level 4: Commentary for State 4 is generated. If a visitor requests country music recommendations, the system provides these recommendations based on an additional online search. If the visitor follows the plan, it leads to the next showroom.
(5) Songwriters Showroom-Level 4: This showroom generates commentary for State 5. If a visitor requests a summary of

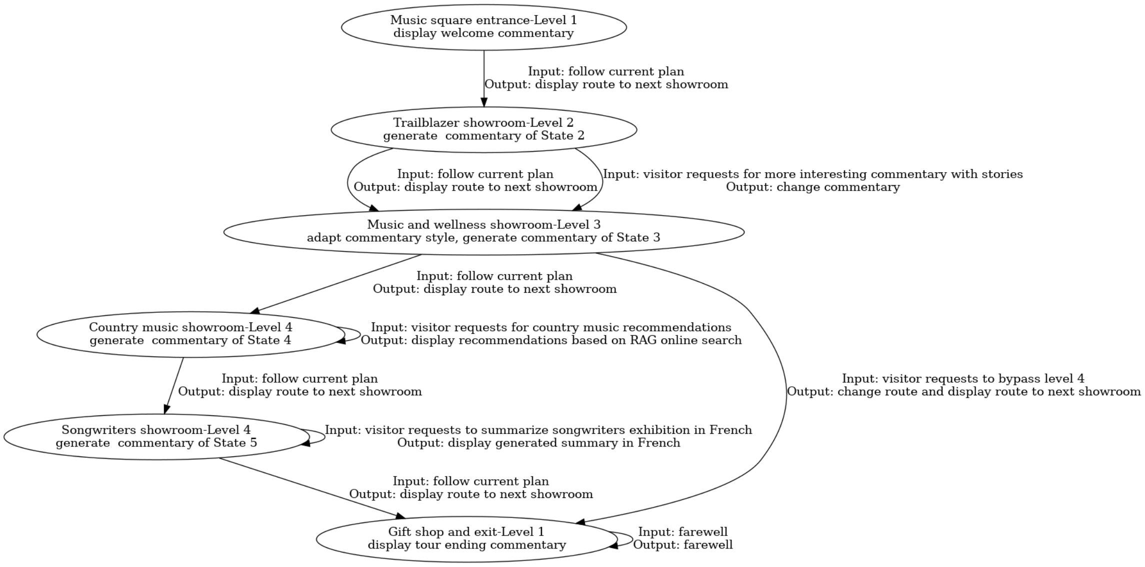

Music square entrance-Level 1
display welcome commentary
Input: follow current plan
Output: display route to next showroom
Trailblazer showroom-Level 2
generate commentary of State 2
Input: follow current plan
Output: display route to next showroom
Input: visitor requests for more interesting commentary with stories
Output: change commentary
Music and wellness showroom-Level 3
adapt commentary style, generate commentary of State 3
Input: follow current plan
Output: display route to next showroom
Country music showroom-Level 4
generate commentary of State 4
Input: visitor requests for country music recommendations
Output: display recommendations based on RAG online search
Input: visitor requests to bypass level 4
Output: change route and display route to next showroom
Input: follow current plan
Output: display route to next showroom
Songwriters showroom-Level 4
generate commentary of State 5
Input: visitor requests to summarize songwriters exhibition in French
Output: display generated summary in French
Input: follow current plan
Output: display route to next showroom
Gift shop and exit-Level 1
display tour ending commentary
Input: farewell
Output: farewell

Figure 6: Example of SDDP Tour at Studio Bell Museum

the songwriters exhibition in French, the system provides a generated summary in French.

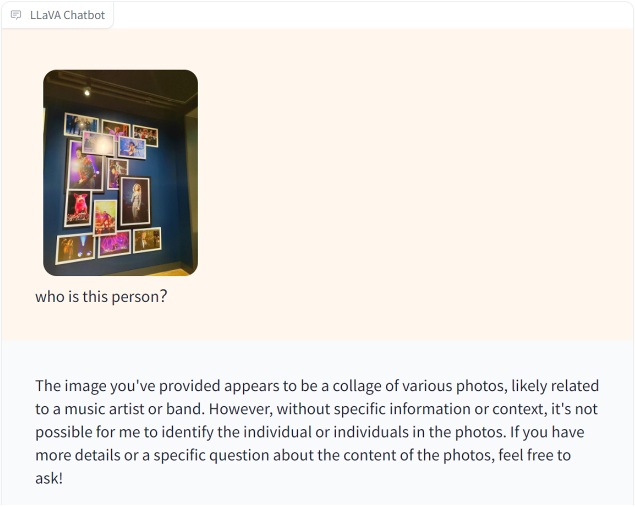
(6) Gift Shop and Exit-Level 1: At the end of the tour, visitors reach the final state where the tour ending commentary is displayed. If the visitor says farewell, the output is also farewell.

The instruction tuning and execution of a Studio Bell Museum tour planning example are detailed in Table 2. It shows that SDDP provides a reliable framework for LLMs to perform the tour planning task with full control over visitor progress and experience.

4.3 Tour Q&A in Studio Bell Museum

For the interactive tour Q&A, we still use the Studio Bell Museum for our case study. As mentioned, LLaVA faced the challenge of spatial reference in the Visual Q&A task. An example is shown in Figure 7. To facilitate easy spatial referencing, we implemented the proposed Object Labeling Workflow using Gradio UI [1] as the visitor interface and Grounding DINO [5] as the object detection model. The workflow is implemented as the following steps:

(1) Identify that the visual question in Figure 7 involves spatial reference ambiguity.
(2) Given that the object in question is person, Grounding DINO is applied to detect and label persons within different ROIs, as demonstrated in Figure 8. Gradio UI displays the labeled image and prompts the visitor to specify the ROI label as supplementary input. Note that the actual implementation utilizes the APIs of Grounding DINO and Gradio UI as a backend process instead of a frontend demo as shown here.
(3) When the visitor provides supplementary ROI label input — "red bounding box labeled with the score of 0.77", LLaVA accurately identifies the specified individual from among multiple persons.



who is this person?

The image you've provided appears to be a collage of various photos, likely related to a music artist or band. However, without specific information or context, it's not possible for me to identify the individual or individuals in the photos. If you have more details or a specific question about the content of the photos, feel free to ask!

Figure 7: Spatial Reference Problem in Visual Q&A

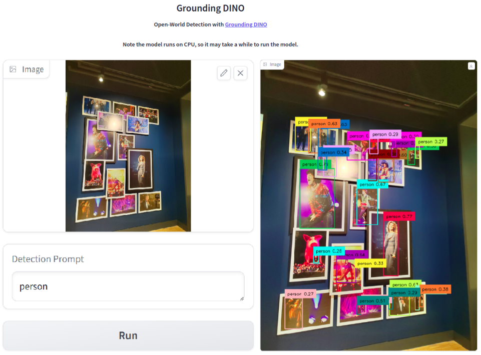
(4) With infused visual knowledge about the specified individual, LLaVA gives a correct response that the person is Celine Dion, a Canadian singer, as demonstrated in Figure 9.

5 EVALUATION

5.1 User Study

We evaluate the IndoorRoaming system using five common LLMs system evaluation metrics:

- **Accuracy** - the number of correct responses out of 5 tests.
- **Generation Quality** - coherence and logical consistency of the LLMs' outputs.
- **Response Time** - speed of LLMs' inference.



Grounding DINO

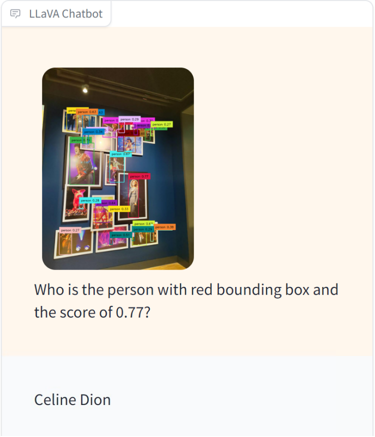
Open-World Detection with Grounding DINO

Note the model runs on CPU, so it may take a while to run the model.

Detection Prompt

person

Figure 8: Labeling Persons using Object Detection Model



Who is the person with red bounding box and the score of 0.77?

Celine Dion

Figure 9: Spatial Reference Solution in Visual Q&A

- **Robustness** - ability of the LLMs-based system to handle erroneous inputs.
- **User Satisfaction** - overall interaction experience and utility value of the LLMs-based system.

A user study was conducted where each tour guidance core task was tested by nine individuals and scored from 1 (low) to 5 (high) on the above five metrics.

5.2 Results & Findings

Quantitative user study results are presented in Figure 10. The findings indicate:

- **Pathfinding Task:** Accuracy, generation quality, response time, and user satisfaction all scored 4 or above. This demonstrates that the Indoor Entity Graph effectively supports LLMs in learning spatial knowledge and assisting with pathfinding. However, the robustness average score is around 3, which indicates that careful handling of visitor error inputs is necessary to improve the interaction experience.
- **Tour Planning Task:** Accuracy, generation quality, and response time all scored 4 or above, suggesting that the State-Driven Dynamic Planning method significantly enhances LLM-based interactive tour planning. Despite this, one user scored only 2 on user satisfaction, noting, "It would be better if LLMs could automatically introduce the exhibit I was looking at instead of requiring me to take a photo and input it into the chatbot."
- **Visual Q&A Task:** Generation quality and user satisfaction both scored 4 or above, reflecting satisfactory answers provided by LLaVA to questions about the museum. This score also demonstrates that the Object Labelling Workflow is helpful for users in identifying specific objects for the visual Q&A task.

These results provide valuable insights into the strengths and areas for improvement in the LLM-based IndoorRoaming system.

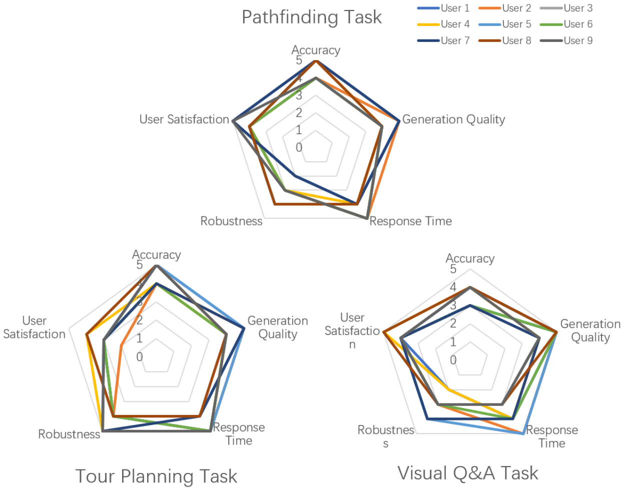


Figure 10: Evaluation Results of the User Studies

6 CONCLUSION AND FUTURE WORKS

This paper introduces IndoorRoaming, an LLM-based system to support indoor tour guidance. The system has demonstrated its effectiveness to perform the core indoor guidance tasks of pathfinding, tour planning and tour Q&A, with designed Indoor Entity Graph, State-driven Dynamic Planning method, and Object Labelling Workflow. IndoorRoaming holds great promise for integration into real-world applications, marking a solid advancement in the research domain of indoor spatial modeling and reasoning with LLMs.

There is broad scope to refine and escalate our work across various dimensions. First, we will enhance the multimodal capabilities of LLMs for the indoor tour guidance tasks. This includes providing landmark visual guidance with directional signs and exploring MLLMs' supports for Audio Q&A in audio exhibitions. Second, we plan to integrate with suitable hardware, such as AR device or embodied tour guide robot, to build a comprehensive LLMs-based tour guidance solution for indoor venues. This aims to address the feedback from User Survey which suggested that the LLMs embody should be able to automatically introduce the exhibit a visitor is viewing, rather than requiring the visitor to take a photo and input it into the chatbot to enquiry.

Table 1: Instruction Tuning for Shopping Mall Pathfinding

Step 1: Build indoor entity graph	**Input:** shopping center.json The JSON file contains the indoor topology and distance information of a shopping center. Please build an Indoor Entity Graph network from this JSON file. Please note that the graph network should not only include edges in the JSON file, but also add edged from each node to its corridor's start and end points whose length are already included in the JSON file. Note that each edge should be considered both directions, i.e from start to end and end to start. **Output:** The indoor entity graph network has been successfully built from the provided JSON data. If you need any specific operations to be performed on this graph, such as finding the shortest path between two points, please let me know!
Step 2: Based on the graph network, calculate the shortest route.	**Input:** Please provide the shortest route from Tim Horton to Nike based on the generated graph network. **Output:** The shortest route from Tim Hortons to Nike based on the graph network involves traversing the following nodes: 1 (Tim Hortons), 63, 64, 66, to 39 (Nike). The total distance for this route is 285.0 meters.
Step 3: instruct LLMs to provide friendly road guidance based on the shortest path and indoor entity graph	**Input:** You are information desk staff in this shopping center. Please provide friendly road guidance based on the above generated shortest route. Your road guidance should include the detailed route, distance and turn direction at each landmark. The 'turn direction' guidance at every node of landmark is based on the 'previous edge' and 'next edge' information of JSON file. Note that each edge should also be considered both directions. Here is a format example for each step in your road guidance: From Node * to Node * - Reference:*** - Distance:*** - Turn:*** **Output:** Here's the detailed route from Tim Hortons to Nike, including the closest store references at each turning landmark for better navigation: Starting from Tim Hortons 1. Exit Tim Hortons and head towards the junction with Intersection B - Reference: You will see Tim Hortons itself as you start. - Distance: Approximately 95 meters - Turn: Turn right onto Corridor 2. 2. Continue from node Intersection B to Intersection C - Reference: Safeway is nearby, helping you recognize the turning point. - Distance: Approximately 100 meter - Turn: Turn left onto Corridor 3. 3. Proceed from node Intersection C to Intersection E - Reference: Lottery Centre is the closest store you will see before the next turn. - Distance: Approximately 60 meters - Turn: Turn right onto Corridor 4. 4. Follow Corridor 4 straight to Nike. - Reference: Nike is the destination, directly on this path. - Distance: Approximately 30 meters - Turn: Continue straight; Nike will be on your right. This route totals approximately 285 meters and features clear store landmarks to guide you effectively through the shopping center. Follow these directions to reach Nike easily.

Table 2: Instruction Tuning for Museum Tour Planning

Step 1: Instruct the LLM to start running the recommended tour plan in SDDP method	**Input:** Museum Tour Plan SDDP.xlsx You are a tour guide in this museum. Please follow the State-driven Dynamic Planning method to run the tour plan in this excel file: 1. introduce the complete tour plan 2. start from the initial state, waiting for my (visitor's) feedbacks at each step: following existing rout to next state, or change existing route, or customize commentary 3. perform according to my feedback until the final state to finish the tour. **Output:** Tour Plan Introduction The tour is structured around a sequence of showrooms, each designed to provide an immersive experience through curated commentary and guided transitions. Starting the Tour: We are now at the Music Square Entrance - Level 1. I'll begin by providing a welcome commentary. Please let me know how you would like to proceed: Continue following the current plan to the next state. Change the existing route. Customize the commentary. Your feedback will guide the next steps!
Step 2: The visitor gives feedback to proceed with the current plan to the next step.	**Input:** continue **Output:** Great, let's proceed! Current State: Trailblazer Showroom - Level 2 Action: I'm generating commentary about the Trailblazer showroom, which focuses on pioneers in various music genres and their contributions to the evolution of music. Next Step: After the commentary, we'll head to the Music and Wellness Showroom on Level 3. Please indicate how you would like to proceed: Follow the current plan to the next state. Change the route. Customize the commentary for the next state or any specific requests for this showroom.
Step 3: The visitor gives feedback to change tour commentary content	**Input:** change the commentary more fun with more stories **Output:** Understood! I'll make the commentary more engaging with stories about the trailblazers in music. Enhanced Commentary for the Trailblazer Showroom - Level 2 In this showroom, we celebrate the bold spirits who dared to redefine genres and inspire generations. Let me tell you about Ella Fitzgerald, a pioneer of jazz and bebop. Ella, often known as the "First Lady of Song," had a remarkable ability to mimic instrumental sounds, which she showcased in her scat singing. Her improvisational skill was so profound that many bandleaders felt she could out-swing any instrument in the orchestra! Next Steps We are now ready to move on to the Music and Wellness Showroom on Level 3. Would you like to: Continue following the current plan to the next state? Change the route? Further customize the commentary?
Step 4: The visitor gives feedback to change tour route	**Input:** bypass all stops in Level 4 **Output:** Understood! We'll bypass all stops at Level 4 and proceed directly to the final stage of our tour. Updated Next State: Gift shop and exit-Level 1 Action: Proceed directly to the gift shop and the exit. Welcome back to Level 1! As we conclude our tour, the gift shop offers a chance to take a piece of musical history home with you. Let me know if there's anything else you would like to do or explore before we finish the tour.

REFERENCES

[1] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild. arXiv:1906.02569 [cs.LG]

[2] Anahid Basiri, Elena Simona Lohan, Terry Moore, Adam Winstanley, Pekka Peltola, Chris Hill, Pouria Amirian, and Pedro Figueiredo e Silva. 2017. Indoor location based services challenges, requirements and usability of current solutions. *Computer Science Review* 24 (2017), 1–12. https://doi.org/10.1016/j.cosrev.2017.03.002

[3] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744 [cs.CV]

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 34892–34916. https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf

[5] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023).

[6] Timothy R. McIntosh, Teo Susnjak, Tong Liu, Paul Watters, and Malka N. Halgamuge. 2023. From Google Gemini to OpenAI Q* (Q-Star): A Survey of Reshaping the Generative Artificial Intelligence (AI) Research Landscape. arXiv:2312.10868 [cs.AI]

[7] Roi Otero, Susana Lagüela, Iván Garrido, and Pedro Arias. 2020. Mobile indoor mapping technologies: A review. *Automation in Construction* 120 (2020), 103399. https://doi.org/10.1016/j.autcon.2020.103399

[8] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 2, 22 pages. https://doi.org/10.1145/3586183.3606763

[9] Vishnu Sashank Dorbala, Sanjoy Chowdhury, and Dinesh Manocha. 2024. Can LLMs Generate Human-Like Wayfinding Instructions? Towards Platform-Agnostic Embodied Instruction Synthesis. *arXiv e-prints*, Article arXiv:2403.11487 (March 2024), arXiv:2403.11487 pages. https://doi.org/10.48550/arXiv.2403.11487 arXiv:2403.11487 [cs.RO]

[10] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]

[11] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. *IEEE/CAA Journal of Automatica Sinica* 10, 5 (2023), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

[12] Chiqun Zhang, Antonios Karatzoglou, Helen Craig, and Dragomir Yankov. 2023. Map GPT Playground: Smart Locations and Routes with GPT. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23)*. Association for Computing Machinery, New York, NY, USA, Article 52, 4 pages. https://doi.org/10.1145/3589132.3625595