Narrative-Driven Travel Planning: Geocultural-Grounded Script Generation with Evolutionary Itinerary Optimization

Anonymous ACL submission

Abstract

To enhance tourists' experiences and immersion, this paper proposes a narrative-driven travel planning framework called NarrativeGuide, which generates a geocultural-grounded narrative script for travelers, offering a novel, role-playing experience for their journey. In the initial stage, NarrativeGuide constructs a knowledge graph for attractions within a city, then configures the worldview, character setting, and exposition based on the knowledge graph. Using this foundation, the knowledge graph is combined to generate an independent scene unit for each attraction. During the itinerary planning stage, NarrativeGuide models narrative-driven travel planning as an optimization problem, utilizing a genetic algorithm (GA) to refine the itinerary. Before evaluating the candidate itinerary, transition scripts are generated for each pair of adjacent attractions, which, along with the scene units, form a complete script. The weighted sum of script coherence, travel time, and attraction scores is then used as the fitness value to update the candidate solution set. Experimental results across four cities, i.e., Nanjing and Yangzhou in China, Paris in France, and Berlin in Germany, demonstrate significant improvements in narrative coherence and cultural fit, alongside a notable reduction in travel time and an increase in the quality of visited attractions. Our study highlights that incorporating external evolutionary optimization effectively addresses the limitations of large language models in travel planning.

1 Introduction

Large language models (LLMs) have demonstrated significant success in various generation tasks, such as role-playing (Wang et al., 2023b). These applications not only offer a convenient alternative to human labor but also enhance the user's narrative immersion (Ahn et al., 2024; Lu et al., 2024), such as the educational chatbot (Wang et al., 2024) and the sales agent (Chang and Chen, 2024). Moreover, in the tourism domain, some studies (Wei et al., 2024; Vasic et al., 2024; Helmy et al., 2024) have explored employing LLMs as virtual tour guides. Although these systems offer increased convenience, they do not necessarily improve the overall user experience. This is because tourists' modes of travel remain unchanged, limiting the potential for deeper immersion, and LLMs often lack robust itinerary planning capabilities.

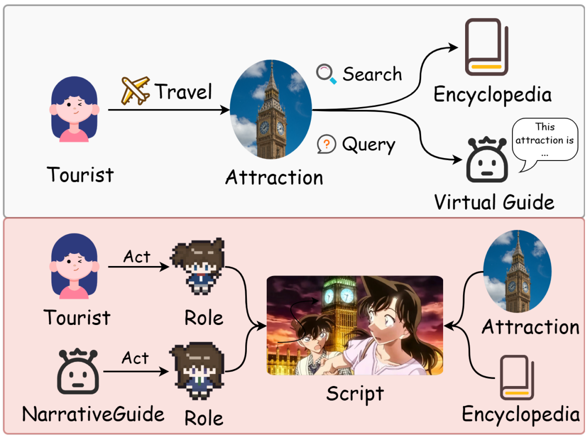

Figure 1: Comparison between narrative-driven travel and traditional tourism. In traditional tourism (top figure), tourists typically search encyclopedias or consult virtual guides to obtain information about attractions. Narrative-driven travel (bottom figure) immerses tourists in a personalized storyline, where they assume roles within a script based on the geocultural background of the attractions. Guided by the *NarrativeGuide*.

Indeed, the powerful story creation capabilities of LLMs have the potential to transform the tourism industry (Wang et al., 2023a; Mirowski et al., 2023). By combining LLM-driven story generation with agent-based role-playing, narratives can be effortlessly brought to life (Han et al., 2024; Wu et al., 2024). Accordingly, we propose the concept of narrative-driven travel planning, as illustrated in

Figure 1. By generating a geocultural-grounded script, tourists can assume the roles of characters within the narrative, thereby enhancing their immersive experience.

Unlike existing tasks in script generation and virtual tour guiding, narrative-driven travel planning faces two primary challenges. First, for the task of generating a travel guide script, LLMs should incorporate geocultural references from authentic tourist attractions to ensure an immersive experience. Second, itineraries must satisfy tourists' constraints, such as travel duration, while optimizing narrative coherence. However, recent research highlights the planning limitations of LLMs. For instance, the TravelPlanner benchmark (Xie et al.) reveals that LLMs struggle to meet user requirements, achieving a success rate of only 0.6%.

Contributions. We model narrative-driven travel planning as an optimization problem. The objective is to select a subset of attractions within a city and determine an itinerary that traverses them, thereby optimizing the narrative coherence, travel time, and attraction score. To address this, we propose *NarrativeGuide*, a framework that integrates geocultural knowledge graphs with genetic algorithms (GA). First, we construct a knowledge graph incorporating historical, cultural, and geographical information for each attraction and generate an independent narrative script for each attraction. Then, we apply a GA-based optimization approach. In each iteration, a new sequence of attraction visits is generated, transition scripts are added to ensure narrative coherence, and their narrative coherence is evaluated. Finally, the itinerary with the optimal weighted sum of script quality, travel time, and attraction satisfaction is selected along with its corresponding travel script. We evaluate our approach using different LLMs across four cities, Nanjing, Yangzhou, Paris, and Berlin. Experimental results demonstrate that *NarrativeGuide* significantly improves script quality compared to baseline methods and enhances itinerary planning by reducing travel time and selecting more popular attractions.

2 Related Work

2.1 Long-Form Script Generation

Long-form narrative generation is a key research area in natural language processing, aiming to produce coherent and creative stories. Guo et al. (2018) introduced LeakGAN, which combines generative adversarial networks (GANs) with policy gradients to guide long-text generation. Yao et al. (2019) introduced the "Plan-and-Write" framework, which divides the story generation into two stages: planning and writing. You et al. (2023) proposed the "EIPE-text" method, which refines plans iteratively using an evaluation mechanism to produce more coherent narratives. In the domain of scriptwriting, Mirowski et al. (2023) developed the Dramatron system, which leverages large language models (LLMs) to co-write movie and theatre scripts. Dramatron generates coherent scripts by hierarchically creating titles, characters, story beats, location descriptions, and dialogues.

2.2 Automatic Itinerary Planning

Numerous studies have addressed automated travel itinerary planning, employing various methods to tackle the problem. Some studies use exact algorithms, such as Verbeeck et al. (2014), which applies a branch-and-cut approach to solve self-guided tour planning. Since travel itinerary planning is NP-hard (Liao and Zheng, 2018; Castro et al., 2015; Gavalas et al., 2013), approximation methods are often employed to enhance solution efficiency. Consequently, metaheuristic algorithms are commonly used. For instance, Abbaspour and Samadzadegan (2009) employed a genetic algorithm to address itinerary planning, focusing on time and multimodal transport constraints. Zhang et al. (2024) use a cooperative co-evolutionary algorithm for cross-city itinerary planning, while Chen et al. (2023) apply an improved ant colony algorithm, considering restaurant and hotel selections. More recently, some researchers have explored the use of LLMs for itinerary planning. For example, Singh et al. (2024) leverage LLMs for personalized travel route planning, and Li (2023) utilize the ChatGPT model to enable users to generate travel plans and suggestions based on keywords.

3 Method

Given a travel itinerary for a tourist, LLMs can directly generate a narrative script for the journey. However, this approach encounters challenges such as attention sink and difficulties in maintaining global consistency. To address these issues, this paper adopts a segmented planning approach, dividing the complete narrative script into scene units based on individual attractions. By pre-configuring the worldview and character settings, the logical consistency of each attraction's independent nar-

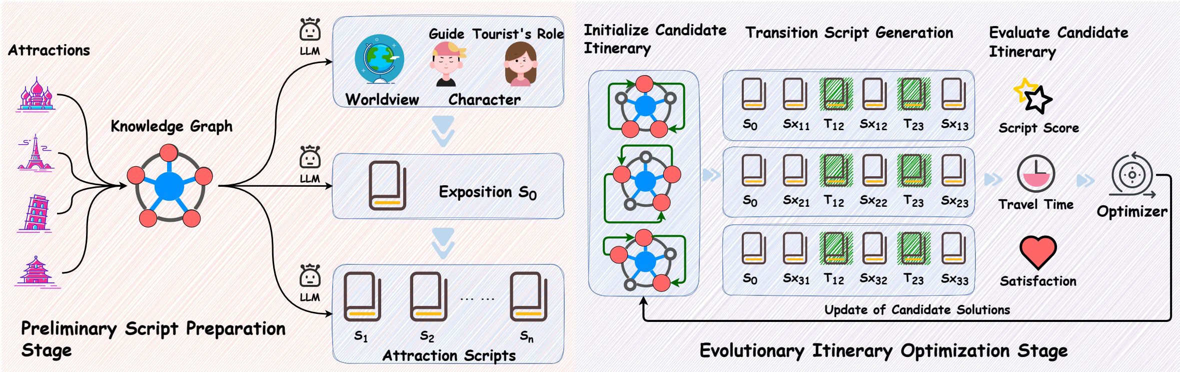

Figure 2: The pipeline of the proposed NarrativeGuide. This framework consists of two stages. The first stage, preliminary script preparation, involves constructing a knowledge graph based on the historical, cultural, and geographical background of various attractions in the city. Using this foundation, NarrativeGuide generates a worldview and character settings, followed by the exposition and independent sub-scripts for each attraction. The second stage, evolutionary itinerary optimization, begins by generating multiple candidate itineraries and their corresponding transition scripts. Each itinerary is then evaluated based on script coherence, travel time, and attraction satisfaction. Finally, GA is employed to optimize the itinerary.

rative script is ensured. After the tourist selects a segment of the itinerary, transition scripts are generated for pairs of adjacent attractions, ultimately forming a complete travel narrative script. Moreover, the challenge of narrative-driven travel planning lies not only in the quality of the script but also in the ability to plan the itinerary. To this end, we model the problem as an optimization task and use the GA to determine the final itinerary, optimizing the script score, travel time, and attraction satisfaction. Figure 2 illustrates the pipeline of the proposed NarrativeGuide framework.

3.1 The Optimization Model

We model narrative-driven travel planning as an optimization problem. To formalize this, we define an undirected, connected, and weighted graph $G = (V, E)$, where the vertex set $V = \{v_1, v_2, \ldots, v_n\}$ represents the scenic spots, and the edge set $E = \{e_1, e_2, \ldots, e_m\}$ represents the relationships between the scenic spots. Each edge $e_k \in E$ connects two distinct vertices v_i and v_j $(i \neq j)$ and is associated with an attribute vector $\mathbf{w}(v_i, v_j)$, which encodes the historical or cultural connections between these two spots, along with geographical attributes such as travel time. Each vertex $v_i \in V$ is also associated with an attribute vector $\mathbf{w}(v_i)$, which encapsulates information about the scenic spot, including its historical background, cultural significance, main attractions, geographical location, ticket price, and other relevant details.

The objective is to select a subset $S \subset V$ and determine an optimal visiting sequence $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ for the selected subset. This arrangement is designed to maximize the tourist's experience, such as the coherence of the corresponding narrative script, the quality of the attractions, and the travel time. The objective function can be expressed as follows:

$$\begin{aligned} \max_{\mathbf{x}} F(\mathbf{x}) =& w_1 f_1(\mathbf{x}) + w_2 \sum_{i=1}^{k-1} f_2(x_i, x_{i+1})^{-1} \\ &+ w_3 \sum_{i=1}^{k} f_3(x_i) \end{aligned} \tag{1}$$

where $f_1(\mathbf{x})$ represents the smoothness score, $f_2(x_i, x_{i+1})$ represents the travel time between attractions x_i and x_{i+1}, and $f_3(x_i)$ represents the popularity of attraction x_i, w_1, w_2, and w_3 are weighting factors that control the relative importance of each component in the optimization. To transform the problem into a maximization problem, we take the reciprocal of the travel time f_2 as the second term in the objective function $F(\mathbf{x})$.

3.2 Geoculturally-Grounded Narrative Script Generation

To create an immersive experience for tourists, the narrative script must be grounded in the geocultural context of the attractions. Therefore, we initially construct a knowledge graph by extracting information about attractions from Wikipedia and inputting it into the LLM. The LLM is responsible for summarizing this information into five key attributes, i.e., historical background, cultural significance, historical stories, main attractions, and geographical location. In the knowledge graph, each attraction is represented as a node, and each node is associated with an attribute vector that includes the aforementioned five attributes of the attraction, collectively referred to as the attraction information. Subsequently, we input these attributes into the LLM to extract historical or cultural connections between the attractions. These connections are used as edges to connect the nodes, and each edge is associated with an attribute vector that includes historical or cultural relevance. In this manner, we construct a weighted and connected knowledge graph that enables the generation of geocultural-grounded narrative scripts, as depicted in Fig. 3.

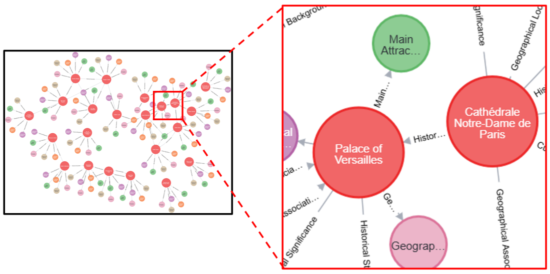

Figure 3: Knowledge graph of attraction information.

Algorithm 1 Narrative Script Generation for Tourist Itinerary

Initialize knowledge graph G, attractions V, and itinerary $\mathbf{x}$
Generate world view $\mathcal{W}$ and character settings $\mathcal{C}$;
Generate exposition S_0 based on $\mathcal{W}$, $\mathcal{C}$;
for each attraction v_i in V **do**
  Generate scene unit S_i for v_i based on $\mathcal{W}$ and $\mathcal{C}$;
end for
for $i \leftarrow 1$ **to** $k-1$ **do**
  Generate transition script T_{ij} between scripts S_{x_i} and $S_{x_{i+1}}$;
end for
Return Narrative script $\{S_0, S_{x_1}, T_{12}, S_{x_2}, \dots, T_{k-1,k}, S_{x_k}\}$

Consider an itinerary $\mathbf{x} = (x_1, x_2, \dots, x_k)$, we generate a narrative script for the tourist in a multi-level manner, as outlined in Algorithm 1. First, we create a personalized worldview and character settings for the tourist. Then, we generate the exposition, which immerses the tourist in their role. Next, based on the geocultural information of each attraction, we generate an independent sub-script for each attraction, treating them as scene units. Finally, for the itinerary $\mathbf{x}$, we create a transition script for each pair x_i, x_{i+1}, considering their cultural, historical, and geographical relationships, ensuring smooth scene transitions and maintaining the tourist's immersion.

Worldview and Character Setting. We instruct the LLM to generate a worldview, denoted as $\mathcal{W}$, by integrating the storylines and cultural backgrounds of attractions. This process follows a predefined format and an example worldview provided as reference. The LLM is tasked with producing a foundational description of the fictional world, encompassing its history, culture, and geographical features, while ensuring consistency with the background of the attractions. Additionally, it defines world rules that align with these elements.

The generated worldview $\mathcal{W}$ serves as the basis for creating two characters, i.e., the user character and the guide character. Using $\mathcal{W}$, a predefined character setting format, and example character profiles, the LLM determines the names, identities, personality traits, background stories, and relationships with the user or travel purposes for both characters. The complete character settings are represented as $\mathcal{C} = \{C_u, C_g\}$, where C_u and C_g correspond to the user character and guide character, respectively.

Exposition. The LLM synthesizes the worldview $\mathcal{W}$ and character settings $\mathcal{C}$ to generate an engaging exposition, denoted as S_0. This introduction establishes the narrative framework by presenting the initial encounter between the user and the guide, defining the journey's starting point and purpose, and offering a glimpse into the forthcoming adventure. The LLM is tasked with crafting S_0 to ensure coherence with the predefined elements, effectively setting the stage for the unfolding storyline.

Geocultural-Grounded Attraction Script. We begin by extracting detailed attraction information v_i from the knowledge graph. Using this data, we construct a comprehensive prompt that integrates v_i, the overarching worldview $\mathcal{W}$, and character settings $\mathcal{C}$. The prompt specifies the script structure, divided into "Intro," "Development," "Climax," and "Ending", along with the desired narrative style, character interactions, and key plot elements. This structured prompt guides the LLM in generating complete and coherent attraction scripts. Following these guidelines, the LLM produces multiple scripts $\mathcal{S}_i$ that align with the specified criteria.

Transition Script. After designing a travel route $S = \{S_0, S_{x_1}, S_{x_2}, \ldots, S_{x_k}\}$ consisting of multiple attraction scripts, we focus on generating the transitional script T_{ij} for adjacent attraction scripts S_{x_i} and S_{x_j} using the LLM. To guide the LLM, we establish several requirements: a common cultural theme, a time-space portal triggered by historical events, clear reasoning for the scene transitions, and consistency in character goals. Based on these instructions, the LLM generates T_{ij}.

Next, we combine S_{x_i}, T_{ij}, and S_{x_j} and input them back into the LLM for evaluation. The LLM assesses the transitional script T_{ij} according to a predefined questionnaire, considering four aspects, i.e., narrative coherence, character interaction, spatiotemporal consistency, and immersion (each dimension contains three sub-questions). The average evaluation score serves as the smoothness score for T_{ij}, providing valuable data for travel planning. Upon completion, we obtain the full travel script $T(\mathbf{x}, S) = \{S_0, S_{x_1}, T_{12}, S_{x_2}, \ldots, T_{k-1,k}, S_{x_k}\}$.

3.3 Genetic Algorithm for Narrative-Driven Travel Planning

NarrativeGuide utilizes the GA to determine the final itinerary. The following sections introduce the algorithm from two aspects: the encoding scheme and the update of candidate solutions.

Encoding Scheme and Population Initialization. In the GA, each candidate solution represents a candidate itinerary, and the dimension of the solution indicates the upper bound of the number of attractions that can be visited. Since a single attraction may correspond to multiple different scripts, once the sequence of attractions is determined, it is necessary to further specify the script number for each attraction. To achieve this, the encoding consists of two main pieces of information: the attraction number and the script choice for each attraction. Therefore, a two-dimensional encoding scheme is used, where the first row contains the attraction numbers or 0 (indicating a placeholder that does not correspond to any attraction). The second row contains the script choices for the corresponding attractions. Fig. 4 provides an encoding example, representing an itinerary from attraction 1 $\rightarrow$ attraction 2 $\rightarrow$ attraction 3, with corresponding scripts 3, 2, and 1, respectively.

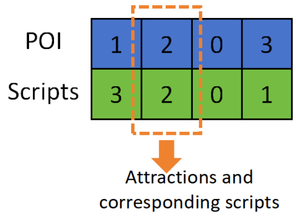

Figure 4: The encoding example of a candidate solution in GA.

Each candidate solution in the population is initialized as follows. For the first row, each position is randomly selected from the set of attractions V, with a certain probability of being set to 0. For the second row, if a position corresponds to a visited attraction, a random script is selected from the available scripts for that attraction; otherwise, the value is set to 0.

Update of Candidate Solutions The population update process consists of two core steps, i.e., crossover and mutation. To prevent duplicate attractions during these operations, adjustments are made to the crossover and mutation strategies. Although the encoding is two-dimensional, with attraction numbers and their corresponding scripts located in the same column, only the attraction numbers need to be checked for duplication. Therefore, the encoding is treated as one-dimensional for both crossover and mutation operations.

In the crossover strategy, two crossover points, n_1 and n_2 ($n_1 < n_2$), are selected. The encoding between columns n_1 and n_2 is exchanged between two individuals. During this exchange, we check for duplicate attraction numbers within the segment, ensuring no duplication occurs with other parts of the encoding. If any duplication is found, the exchange is aborted.

In the mutation strategy, a mutation point m is selected, and a random attraction number, not already present in the encoding, is chosen. The script choice for this attraction is then selected randomly.

The pseudocode of the algorithm is shown in 2.

4 Experiment

In this section, we conduct tests in four cities, i.e., Nanjing and Yangzhou in China, Paris in France, and Berlin in Germany. We evaluate both the quality of the scripts generated by the proposed algorithm and its ability to plan travel itineraries. The

Algorithm 2 Genetic Algorithm for Narrative-Driven Travel Planning

Initialize population $P = \{\mathbf{x}_1, \ldots, \mathbf{x}_\lambda\}$
while Termination condition not satisfied **do**
  for $i \leftarrow 1$ **to** λ **do**
    Perform Crossover operation;
    Perform Mutation operation;
    Generate new $\mathbf{x}'$;
    $\{S_0, S_{x'_1}, T_{12}, \ldots, T_{k-1,k}, S_{x'_k}\} = T(\mathbf{x}', S)$;
    Evaluate the fitness of $\mathbf{x}'$, $f(\mathbf{x}')$;
    If $\mathbf{x}'$ is better than $\mathbf{x}$, replace $\mathbf{x}$;
  end for
end while
Return $\mathbf{x}^* \leftarrow \arg\min(f(\mathbf{x}))$

section begins with a description of the experimental setup, followed by an analysis of the results.

4.1 Experimental Design

The experiment consists of two parts. The first part aims to evaluate the quality of the generated travel scripts based on predefined criteria. For this, we use OpenAI's GPT-4 model as the evaluation model. The model receives as input the algorithm-generated narrative scripts, which represent travel itineraries for various destinations, including descriptions of attractions, historical and cultural context, character interactions, and attraction information. First, the model is paired with relevant attraction information, followed by the use of custom evaluation prompts. The evaluation focuses on four aspects: plot coherence, character interaction, time and space coherence, and cultural fit. Weight adjustment rules are applied based on script length: a factor of 0.7 for scripts under 1500 characters, no adjustment for scripts between 1500 and 7000 characters, and a factor of 1.2 for scripts exceeding 7000 characters. The detailed evaluation criteria and weights are included in Appendix A.

The second part of the experiment aims to evaluate the quality of the generated travel itineraries. The comparison focuses on two metrics, i.e., travel time and attraction score. The attraction score is calculated as the product of the number of reviews and the rating for each attraction on the Ctrip website. Each model is tested ten times, and the average values are used for comparison.

4.2 Experimental Results

4.2.1 Script Quality

Table 1 presents the experimental results of the proposed algorithm across different LLMs, including Deepseek-v3, GPT-4o, GPT-4o-mini, GPT-4, and Qwen2.5-max. We use pure GPT-4 as the baseline algorithm for comparison. The evaluation dimensions include narrative coherence (NC), character interaction (CI), spatiotemporal consistency (SC), cultural fit (CF), and the overall score. Here, we generate Chinese scripts for Nanjing and Yangzhou in China, while English scripts are generated for Paris and Berlin in France and Germany, respectively.

City	Model	NC	CI	SC	CF	Overall
Berlin	Baseline	7.40	2.40	8.00	7.50	25.30
	Deepseek-v3	7.92	2.52	9.12	9.00	28.56
	GPT-4o-mini	10.08	2.52	8.88	10.20	**31.68**
	GPT-4o	9.36	2.52	9.12	9.60	30.60
	GPT-4	8.88	2.88	8.40	9.00	29.16
	Qwen2.5-max	8.88	2.52	9.12	9.00	29.52
NanJing	Baseline	5.18	1.68	5.18	5.25	17.29
	Deepseek-v3	7.40	2.10	7.00	7.50	24.00
	GPT-4o-mini	7.40	2.10	7.00	7.00	23.50
	GPT-4o	7.40	2.10	7.00	7.50	24.00
	GPT-4	7.20	2.10	7.00	7.50	23.80
	Qwen2.5-max	7.40	2.10	7.60	7.50	**24.60**
Paris	Baseline	7.40	2.40	7.00	7.50	24.30
	Deepseek-v3	8.88	2.88	9.12	9.00	29.88
	GPT-4o-mini	8.64	2.52	8.40	9.00	28.56
	GPT-4o	7.40	2.40	7.40	7.50	24.70
	GPT-4	9.84	2.88	9.12	9.00	30.84
	Qwen2.5-max	9.84	2.88	9.60	10.80	**33.12**
YangZhou	Baseline	5.18	1.68	5.18	5.25	17.29
	Deepseek-v3	7.40	2.10	7.40	7.50	24.40
	GPT-4o-mini	8.16	2.52	8.40	8.40	**27.48**
	GPT-4o	7.40	2.10	7.00	7.50	24.00
	GPT-4	7.00	2.10	7.00	7.00	23.10
	Qwen2.5-max	7.40	2.10	7.00	7.50	24.00

Table 1: The experimental results of the proposed algorithm across Deepseek-v3, GPT-4o, GPT-4o-mini, GPT-4, and Qwen2.5-max are compared with the baseline GPT-4.

From the results in Table 1, it can be observed that the proposed algorithm outperforms the baseline method across all four cities and various LLMs. Among the LLMs tested, Qwen2.5-max performed the best in Nanjing (China) and Paris (France), while GPT-4o-mini showed the best results in Yangzhou (China) and Berlin (Germany). Compared to the baseline algorithm, the scores improved by 28%–59%, demonstrating that the proposed algorithm significantly enhances LLM-driven narrative-based travel planning tasks. Furthermore, the experimental results indicate that better LLM performance did not lead to a significant improvement in script generation tasks. In fact, our approach decomposes a large itinerary script into several scene units (representing individual attractions) and constructs the complete narrative script through transition scripts. This reduces the demand for LLMs' long-text generation capabilities.

By comparing the scores across the four detailed metrics, we can observe significant improvements in narrative coherence (NC), cultural fit (CF), and spatiotemporal consistency (SC), while the improvement in character interaction (CI) was relatively modest. The improvements in narrative coherence and spatiotemporal consistency can be attributed to the evolutionary algorithm's optimization of the itinerary, which considers both the geographical proximity of attractions and their cultural and historical relevance. The enhancement in cultural fit arises from the algorithm's approach of assigning an independent narrative to each attraction, ensuring that the attraction's script aligns with its cultural background. However, when generating the overall script, the occurrence of hallucinations may reduce the cultural fit score.

Moreover, the experiments also indicate that the language of the script (Chinese versus English) has a significant impact on the quality of the generated scripts. As shown in Table 1, the overall scores for Chinese scripts (Nanjing and Yangzhou) were consistently lower than those for English scripts (Berlin and Paris). This suggests inherent differences in how models process historical content across various linguistic and cultural contexts.

4.2.2 Travel Itinerary Planning

This section compares the proposed NarrativeGuide with several representative LLMs, including GPT-4o, GPT-4o-mini, and Qwen2.5-max, focusing on their travel itinerary planning capabilities. The comparison primarily examines travel time and the quality of the planned attractions, as shown in Tables 2 and 3. Note that the base LLM used in NarrativeGuide is GPT-4o.

From the results in Table 2, it is evident that with the introduction of GA optimization, the algorithm tends to recommend attractions that are clustered together, significantly reducing travel time. For example, for the city of Berlin, the travel time for the itinerary recommended by GPT-4o is 27 times greater than that of the NarrativeGuide. These results once again highlight that pure LLMs lack the capability for itinerary planning and are unable to suggest a reasonable travel plan. Additionally, the results in Table 3 further support this conclusion. The itineraries generated by NarrativeGuide have higher attraction scores, indicating that they feature popular destinations. However, this advantage is not as significant as the travel time reduction, as LLMs possess enough internal knowledge to recommend popular attractions. Yet, due to the inability to collect real-time data from the real world, this outcome is based on prior knowledge rather than updated, accurate data. Overall, the use of GA as an external planner in NarrativeGuide proves to be beneficial, significantly enhancing the ability of LLMs to address real-world problems and meet practical demands.

	NarrativeGuide	GPT-4o-mini	GPT-4o	Qwen2.5-max
Nanjing	**0.384**	1.659	1.935	1.854
Yangzhou	**0.786**	2.656	1.645	1.288
Paris	**0.249**	3.283	3.143	3.566
Berlin	**0.212**	3.836	5.886	3.595

Table 2: The travel time (h) of NarrativeGuide with GPT-4o are compared with the baseline GPT-4o, GPT-4o-mini, and Qwen2.5-max.

	NarrativeGuide	GPT-4o-mini	GPT-4o	Qwen2.5-max
Nanjing	**3.79E+05**	1.65E+05	8.87E+04	9.63E+04
Yangzhou	**4.24E+05**	9.40E+04	1.28E+05	1.71E+05
Paris	**5.17E+04**	1.08E+04	9.55E+03	9.55E+03
Berlin	**1.30E+04**	4.11E+03	3.25E+03	3.67E+03

Table 3: The attraction score of NarrativeGuide with GPT-4o are compared with the baseline GPT-4o, GPT-4o-mini, and Qwen2.5-max.

5 Conclusion

This study introduces *NarrativeGuide*, a novel framework that combines geocultural knowledge graphs with evolutionary algorithms to improve narrative-driven travel planning. By grounding script generation in real-world attractions and optimizing itineraries using GA, our approach addresses the dual challenges of narrative coherence and practical travel constraints. Experimental evaluations across four cities, i.e., Nanjing, Yangzhou, Paris, and Berlin, show significant improvements: script quality metrics, including narrative coherence, cultural fit, and spatiotemporal consistency, increased by 28%–59% compared to baseline methods, while travel time was reduced by up to 27-fold in cities such as Berlin. The framework's integration of LLM-generated scene units with GA-driven itinerary optimization ensures both immersive storytelling and efficient route planning, overcoming the limitations of traditional LLMs in handling real-world constraints.

Limitations

Data Dependency The quality of generated scripts heavily relies on the completeness and accuracy of the knowledge graph, which may limit scalability to regions with sparse cultural or historical data.

Character Interaction While narrative coherence and cultural fit were strengths, character interaction scores remained suboptimal (e.g., 1.68–2.88), indicating a need for deeper modeling of dynamic character behaviors.

Language and Cultural Gaps A performance disparity (23%) was observed between English and Chinese scripts, suggesting potential biases in LLMs' handling of non-Western cultural contexts.

Algorithm Scalability The genetic algorithm's efficiency may degrade for large-scale cities or highly complex constraints (e.g., multi-day itineraries).

User Personalization The current framework prioritizes narrative fluency over individualized preferences, such as varying travel interests or activity types. Future work could incorporate adaptive user profiling to address this gap.

References

RA Abbaspour and F Samadzadegan. 2009. Itinerary planning in multimodal urban transportation network. *Journal of Applied Sciences*, 9(10):1898–1906.

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. 2024. Timechara: Evaluating point-in-time character hallucination of role-playing large language models. *arXiv preprint arXiv:2405.18027*.

Marco Castro, Kenneth Sörensen, Pieter Vansteenwegen, and Peter Goos. 2015. A fast metaheuristic for the travelling salesperson problem with hotel selection. *4or*, 13:15–34.

Wen-Yu Chang and Yun-Nung Chen. 2024. Injecting salesperson's dialogue strategies in large language models with chain-of-thought reasoning. *arXiv e-prints*, pages arXiv–2404.

Shuo-Tsung Chen, Tsung-Hsien Wu, Ren-Jie Ye, Liang-Ching Lee, Wen-Yu Huang, Yi-Hong Lin, and Bo-Yao Wang. 2023. Application of ant colony optimization computing to a recommended travel itinerary planning system with repeatedly used nodes. *Applied Sciences*, 13(24):13221.

Damianos Gavalas, Charalampos Konstantopoulos, Konstantinos Mastakas, Grammati Pantziou, and Yiannis Tasoulas. 2013. Cluster-based heuristics for the team orienteering problem with time windows. In *Experimental Algorithms: 12th International Symposium, SEA 2013, Rome, Italy, June 5-7, 2013. Proceedings 12*, pages 390–401. Springer.

Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. 2018. Long text generation via adversarial training with leaked information. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Senyu Han, Lu Chen, Li-Min Lin, Zhengshan Xu, and Kai Yu. 2024. IBSEN: Director-actor agent collaboration for controllable and interactive drama script generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1607–1619.

Mona Helmy, Yomna Safaa El-Din, Omar Tarek Mohamed, Omar Salah Abdel Kader, Shehab Adel Ramadan, Amr Essam Kamal, and Mohamed Reda Mohamed Selim. 2024. Navigating the world with an intelligent tourist guide using generative ai. In *2024 International Telecommunications Conference (ITC-Egypt)*, pages 1–6. IEEE.

Bohao Li. 2023. Everywheregpt: An ai travel planning assistant based on chatgpt. In *Proceedings of the 4th International Conference on Artificial Intelligence and Computer Engineering*, pages 995–1003.

Zhixue Liao and Weimin Zheng. 2018. Using a heuristic algorithm to design a personalized day tour route in a time-dependent stochastic environment. *Tourism Management*, 68:284–300.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. *arXiv e-prints*, pages arXiv–2401.

Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Harmanpreet Singh, Nikhil Verma, Yixiao Wang, Manasa Bharadwaj, Homa Fashandi, Kevin Ferreira, and Chul Lee. 2024. Personal large language model agents: A case study on tailored travel planning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 486–514.

Iva Vasic, Hans-Georg Fill, Ramona Quattrini, and Roberto Pierdicca. 2024. LLM-aided museum guide: Personalized tours based on user preferences. In *International Conference on Extended Reality*, pages 249–262. Springer.

Cédric Verbeeck, Pieter Vansteenwegen, and E-H Aghezzaf. 2014. An extension of the arc orienteering problem and its application to cycle trip planning. *Transportation research part E: logistics and transportation review*, 68:64–78.

Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024. Book2dial: Generating teacher-student interactions from textbooks for cost-effective development of educational chatbots. *arXiv e-prints*, pages arXiv–2403.

Yuxin Wang, Jieru Lin, Zhiwei Yu, Wei Hu, and Börje F Karlsson. 2023a. Open-world story generation with structured knowledge enhancement: A comprehensive survey. *Neurocomputing*, page 126792.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. 2023b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Qikai Wei, Mingzhi Yang, Jinqiang Wang, Wenwei Mao, Jiabo Xu, and Huansheng Ning. 2024. Tourllm: Enhancing llms with tourism knowledge. *arXiv e-prints*, pages arXiv–2407.

Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Jiale Hong, Hai Zhao, and Min Zhang. 2024. From role-play to drama-interaction: An llm solution. *arXiv e-prints*, pages arXiv–2405.

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. Travelplanner: A benchmark for real-world planning with language agents. In *Forty-first International Conference on Machine Learning*.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

Wang You, Wenshan Wu, Yaobo Liang, Shaoguang Mao, Chenfei Wu, Maosong Cao, Yuzhe Cai, Yiduo Guo, Yan Xia, Furu Wei, et al. 2023. Eipe-text: Evaluation-guided iterative plan extraction for long-form narrative text generation. *arXiv e-prints*, pages arXiv–2310.

Ziyu Zhang, Peilan Xu, Zhaoguo Wang, and Wenjian Luo. 2024. Cooperative coevolution for cross-city itinerary planning. In *International Conference on Intelligent Information Processing*, pages 382–391. Springer.

Appendix A: Evaluation Criteria and Weights

Dimension	Criteria	Score Range	Weight	Description
1. Plot Coherence	Event Logic	0-10	0.4	Logic of event connections and cause-effect relationships
	Attraction Relevance	0-10	0.4	Connection of attractions to the overall plot
	Transition Smoothness	0-10	0.2	Smoothness and naturalness of transitions between events and locations
2. Character Interaction	Dialogue Authenticity	0-10	0.3	Authenticity of dialogue in relation to character identities and historical/cultural context
	Cultural-Driven Actions	0-10	0.4	Actions of characters based on cultural/historical context
	Metaphorical Dialogue	0-10	0.3	Use of dialogue that adds deeper, symbolic meanings related to the attractions or historical context
3. Time and Space Coherence	Spatiotemporal Corridor	0-10	0.6	Logic of time/space transitions and their relevance to the storyline and attractions
	Route Rationality	0-10	0.4	Historical and geographical logic in selecting travel paths
4. Cultural Fit	Cultural Depth	0-10	0.5	Depth of cultural integration in the narrative (impact on decisions, symbolism, etc.)
	Multi-Dimensional Linkage	0-10	0.5	Complexity of connections between historical, cultural, and geographical elements across attractions

Table 4: Evaluation Criteria and Weights

Appendix B: Prompt

1. Generating Worldview

<table>
<tr><th>Description</th><th>Content</th></tr>
<tr><td>Table Input</td><td><table><tr><td>Location</td><td>Features/Culture/History/Legends</td></tr><tr><td>{item['location']}</td><td>{item['features']}</td></tr></table></td></tr>
<tr><td>Requirement</td><td>Construct a travel script worldview based on the table. Connect various attractions' storylines and describe basic information about this world.</td></tr>
<tr><td>Example</td><td>Travel Script Worldview Setting
Name: Time Journey: Dream Hunting in Jinling
Background:
At the intersection of modern technology and ancient wisdom exists a secret organization - 'Time Guardians'. This group consists of individuals who can travel through historical periods to protect cultural heritage. They can access a parallel world called 'Historical Realm' that preserves the most glorious cultural legacies and captivating stories from history.
In this realm, Nanjing (known as 'Jinling') is a mysterious place full of historical charm. Each attraction represents a temporal node containing rich history and hidden passages to other eras. These passages only appear during specific historical moments, and the Time Guardians' mission is to guide travelers through these nodes while protecting cultural heritage from temporal erosion.</td></tr>
</table>

Table 5: Worldview Generation Template

2. Generating Characters

Description	Content	
Worldview	{worldview}	
Character 1: User's Role	Name	*{Example: Lin Yi}*
	Identity	*{Example: Traveler}*
	Personality	*{Example: Curious, observant}*
Character 2: Guide	Name	*{Example: Murong Yun}*
	Identity	*{Example: Time Guardian}*
	Expertise	*{Example: Temporal navigation}*
Example	**Character Settings:** – Traveler: Lin Yi (Modern history enthusiast) – Guide: Murong Yun (Time Guardian) ▷ Goal: Protect cultural heritage nodes ▷ Key traits: Temporal navigation abilities	

Table 6: Character Generation Template

3. Generating Opening Script

Description	Content
Worldview	{worldview}
Characters	{characters}
Requirements	• Brief introduction of worldview • First meeting between user and guide • Journey starting point and purpose • Preview of upcoming travels • Make the opening engaging and intriguing
Example	**Travel Script Opening** Worldview: {worldview} Characters: {characters} Key Elements: • Introduced the "Historical Realm" parallel world • Established mentor-protégé relationship in first encounter • Set journey goal: Protect cultural heritage nodes • Foreshadowed conflicts with temporal erosion

Table 7: Opening Script Generation Template

4. Generating Script for Attraction

Description	Content
Background and Worldview Setting	{worldview}
Character Setting	{characters}
Attraction Setting	**Location**: {attraction['name']} **Historical Context**: {attraction['history']} **Cultural Features**: {attraction['culture']} **Legends**: {attraction['legends']}
Script Requirements	• Four-act structure: Intro/Development/Climax/Ending • Historical accuracy with emotional character arcs • Adventure elements with environmental interactions • Action-driven climax with tangible conflicts • Self-contained narrative resolution
Special Requirements	• Temporal transition effects between eras • Cultural symbolism in dialogue/actions • Consistent character voices • Skip redundant introductions • Explicit section markers

Table 8: Attraction Script Generation Template

5. Generate Transition Scripts

Description	Content
Current Scenic Spot Script	{script1}
Next Scenic Spot Script	{script2}
Transition Script	**Example Transition**: • Used shared cultural motif (e.g., "Dragon Gate" legend) • Introduced time portal triggered by historical event • Added guide's explanation linking both locations • Maintained character goals across transition

Table 9: Transition Script Generation Template

6. Score with Transition Script

Description	Content
Previous Script	{previous_script}
Transition Script	{transition_script}
Next Script	{next_script}
Combined Script	{combined_script}
Survey Questions	{survey_text}
Scoring Requirement	**Example Scoring**: 4,5,3,2,3,4,1,2,3,1,3,3 **Interpretation**: • First 3 scores: Plot Coherence (Q1-Q3) • Next 3: Character Interaction (Q4-Q6) • Next 3: Spatiotemporal Coherence (Q7-Q9) • Last 3: Immersion (Q10-Q12)

Table 10: Script Scoring Template

Appendix C: Fluency Survey

Category	Question	Rating (1-5)
Plot Coherence	1. Plot continuity across transition scripts	1: Fragmented – 5: Seamless
	2. Logical story progression	1: Forced – 5: Natural
	3. Utilization of prior plot elements	1: Incoherent – 5: Coherent
Character Interaction	4. Consistency with character profiles	1: Inconsistent – 5: Faithful
	5. Dialogue/action naturalness	1: Artificial – 5: Organic
	6. Engagement level	1: Disconnected – 5: Immersive
Spatiotemporal Coherence	7. Historical/geographical accuracy	1: Anachronistic – 5: Authentic
	8. Environmental consistency	1: Jarring – 5: Continuous
	9. Transition clarity	1: Confusing – 5: Intuitive
Immersion	10. Narrative depth	1: Superficial – 5: Layered
	11. Multimedia support	1: Distracting – 5: Enhancing
	12. Innovative storytelling	1: Generic – 5: Original

Table 11: Transition Script Evaluation Criteria