Proceedings of CONF-CDS 2026 Symposium: Machine Learning and Neural Network Applications in Engineering
DOI: 10.54254/2755-2721/2026.AST36916

Multimodal Recommendation Systems for Cultural and Tourism Applications: A Survey and Future Directions

Xize Zhang

College of Information Science and Technology, Northeast Normal University, Changchun, China
15104656219@163.com

Abstract. Traditional cultural tourism interest point recommendation systems rely solely on single-modal data modeling, they fail to accurately uncover tourists' deeper preferences and dynamic travel needs, making it difficult to meet the evolving demands of contemporary cultural tourism consumption. This paper reviews the multimodal data architecture in the cultural tourism domain, categorizes cultural tourism data into five major types, and clarifies the applicable scenarios and application advantages of various domestic and international datasets. Furthermore, leveraging cutting-edge technologies, this paper constructs a comprehensive multimodal recommendation framework for cultural tourism. Subsequently, from the two core dimensions of data and algorithms, the paper conducts an in-depth analysis of the key bottlenecks and practical implementation challenges in current research and, in alignment with industry development trends, proposes several forward-looking research directions. This study enhances the theoretical and technical framework for multimodal recommendation in cultural tourism, providing a solid theoretical basis and engineering reference for numerous smart cultural tourism scenarios.

Keywords: Digital Cultural Tourism, Multimodal Recommendation, POI Recommendation, Heterogeneous Data Fusion, Cross-scenario Adaptation

1. Introduction

The deep integration of the digital economy and the cultural tourism industry has made smart cultural tourism a core driver of sector transformation. Platforms such as Ctrip, TikTok, and Dianping have accumulated vast multimodal data—user travel trajectories, attraction reviews, and audio-visual materials—breaking away from the traditional single-information-service model. Tourists increasingly seek customized, immersive, and differentiated services, imposing higher demands on cultural tourism recommendations [1]. Mainstream cultural tourism POI recommendation previously centered on single-modal data modeling, relying on check-ins and short-text reviews to mine preferences [2]. Over-reliant on shallow structured data, this approach characterized user preferences simplistically, yielding homogeneous recommendations that failed to distinguish user groups or travel purposes, limiting the enhancement of smart cultural tourism services [3]. Maturing artificial intelligence technologies have overcome the limitations of single-modal systems, providing a technical foundation for multi-source data modeling and deep preference mining [4, 5]. However, industry applications still face data silos, noisy data, cultural

DOI: 10.54254/2755-2721/2026.AST36916

distortion in AI-generated content, and privacy breaches [6]. A multimodal recommendation framework balancing precision, security, and cultural integrity is therefore urgently needed [7].

This paper reviews multimodal recommendation systems for cultural and tourism scenarios. First, it categorizes the multimodal datasets, distinguishing international public benchmarks from domestic proprietary ones [8]. Second, it outlines five mainstream technical approaches—foundational deep learning, graph neural networks, cross-modal fusion, large model enhancement, and generative and federated recommendation [5, 9]. Third, it deconstructs key implementation bottlenecks from data and algorithm perspectives [7]. Fourth, it proposes feasible future directions aligned with industry demands for privacy, cultural preservation, and lightweight deployment [10, 11].

This study has both theoretical and applied value. Theoretically, it integrates fragmented research into a systematic framework, filling a gap in systematic studies and providing a standardized reference for subsequent work [4, 8]. Practically, the summarized challenges and directions guide the development of personalized recommendation, intelligent itinerary planning, and cultural heritage promotion systems, helping the industry achieve an intelligent, personalized, culturally enriched upgrade that balances privacy, security, and cultural preservation [12].

2. Multimodal dataset for cultural and tourism scenarios

Multimodal datasets are the core foundation for training cultural and tourism recommendation models, and their adaptability, completeness, and quality directly determine recommendation performance [8]. Cultural and tourism data feature diverse modalities, scattered sources, heterogeneous structures, and scenario specificity, differing significantly from general datasets [1]. This paper categorizes them into five major types and reviews their sources, features, and application scenarios.

2.1. Text-based cultural tourism dataset

The textual modality is the most fundamental and mature data type, offering clear semantics, low annotation costs, and strong interpretability, and serving as the core vehicle for mining emotional preferences, travel intentions, and cultural interests [1]. Product review data covers evaluations of scenic spots, hotels, and dining; the international TripAdvisor dataset is a global benchmark, while domestic data from Ctrip and Dianping aligns with local consumption characteristics. User behavior datasets use the Foursquare and Gowalla check-in datasets from Stanford's SNAP database as international benchmarks, while domestic datasets rely on anonymized OTA logs, addressing cold-start modeling [2].

The travel guide text dataset centers on TravelQA, a question-answering dataset released by AAAI 2026 and tailored for itinerary planning, while anonymized travelogues from Xiaohongshu and Mafengwo align with local scenarios [8]. The cultural heritage text dataset draws on public digital resources from museums and intangible cultural heritage platforms, covering authoritative texts on ancient architecture, heritage narratives, and cultural relics, serving culture-driven recommendation and niche cultural tourism promotion [12].

2.2. Image-based cultural tourism dataset

Image datasets intuitively represent scenic landscapes and ambiance, capture aesthetic preferences, and compensate for the textual modality's limits in conveying visual experiences. They fall into

DOI: 10.54254/2755-2721/2026.AST36916

general-purpose, domain-specific, and local datasets [8]. General-purpose POI image datasets, centered on Yelp Image and Foursquare attraction images, feature comprehensive annotations and diverse coverage, and are widely used for visual model pre-training and text-image alignment [2]. T-ECBM, a tourism-specific text-image dataset released in 2025, provides extensive paired attraction data, addressing the scenario adaptability gap of general-purpose datasets [13].

Domestic researchers have developed localized datasets, including the Guilin Landscape Cultural Tourism Visual Dataset and the Yunnan Ethnic Cultural Tourism Image Collection, filling the gap in international datasets regarding China's regional characteristics and providing essential data support for precise local recommendations [8].

2.3. Video and audio modality cultural and tourism dataset

Short videos and audio are core carriers of cultural and tourism communication in the new media era, containing dynamic landscapes, voiceover explanations, and scene atmospheres. CM-DCTD, China's first large-scale official cultural and tourism short video dataset, integrates nationwide short videos with user interaction data, suiting dynamic feature extraction and new media recommendation modeling. The Migu Cultural Tourism Audio Dataset provides vast scenic voiceovers and folk audio, filling the data gap in the audio modality and supporting audio-text fusion and intelligent speech recommendation [4].

2.4. Other auxiliary data types

Auxiliary heterogeneous data enrich cultural and tourism scenarios and enhance models' dynamic adaptation [1]. Spatio-temporal geographic datasets—scenic coordinates, meteorological and seasonal time-series data—characterize travel patterns and support dynamic time-series recommendations [2]. IoT context-aware data from smart scenic area sensors captures real-time visitor flow, congestion, and environmental conditions, enabling real-time recommendation adjustment. The Multimodal Knowledge Graph (MMKG) dataset, exemplified by the 'Visit Yunnan' dataset, enables knowledge-driven, culturally constrained recommendation modeling [12].

2.5. Standardized multimodal fusion benchmark dataset

Standardized integrated benchmark datasets are the core basis for algorithm comparison and generalization testing [8]. Major international benchmarks—Stanford's SNAP universal cultural and tourism dataset, dedicated route planning datasets, and global urban cultural and tourism datasets—support cross-scenario and cross-city research [2]. Domestic benchmarks focus on the Forbidden City, intangible cultural heritage sites, and distinctive scenic spots, serving human-machine collaboration and cultural heritage recommendation [12]. General recommendation benchmarks also enable standardized performance validation of such models.

3. Multimodal recommendation methods for cultural and tourism scenarios

Building on these datasets, academia has established a mature framework for scenario-based recommendation systems. This paper categorizes cultural and tourism multimodal recommendation methods into five core approaches and outlines their principles and application scenarios.

DOI: 10.54254/2755-2721/2026.AST36916

3.1. Foundational deep learning-based multimodal modeling

Foundational deep learning underpins cultural tourism multimodal recommendation, enabling modality-specific feature extraction and basic fusion modeling [2]. Time-series networks capture temporal travel trajectories, suiting personalized itinerary sequence recommendation. CNNs extract basic visual features, ViT captures global image details, and CLIP achieves cross-modal image-text alignment, uncovering aesthetic preferences [13]. Pre-trained language models perform sentiment analysis, intent classification, and semantic matching to extract subjective preferences [9].

Modal fusion has formed four fundamental paradigms: early, late, hybrid, and hierarchical fusion [8]. Early fusion is efficient and suits simple scenarios; late fusion is robust against interference and ideal for noisy data. Hybrid and hierarchical fusion combine these advantages, enabling precise modeling of heterogeneous multimodal data, and represent the current mainstream foundational approach [4].

3.2. Graph neural network (GNN)-based recommendation methods

Cultural tourism scenarios involve numerous users, attractions, cultural elements, and geographic entities with inherent graph structures, making graph neural networks the core technology for cultural tourism POI recommendation [14]. GCN and GAT model heterogeneous graphs by mining entity-related features and adaptively assigning interaction weights to distinguish user preferences [3]. Integrating GNNs with multimodal knowledge graphs combines structured knowledge with multimodal features, enhancing the cultural relevance and accuracy of recommendations [12].

Furthermore, the Shuangliu Graph Attention Framework enables deep text-image integration, spatiotemporal graph generation models suit dynamic scenario recommendation, and multi-type entity heterogeneous graph models integrate multi-dimensional relationships, fully meeting the modeling requirements of complex cultural tourism scenarios [4].

3.3. Cross-modal fusion and contrastive learning

The cross-modal semantic gap is a core challenge in cultural tourism multimodal modeling, which contrastive learning and dynamic fusion can effectively address [4]. The Cross-Attention Dynamic Weighting Mechanism adaptively adjusts modal weights according to users and scenarios, overcoming the limitations of fixed fusion weights. Bilateral collaborative contrastive learning narrows the semantic distance between homogeneous modalities, precisely aligning text-image and audio-text data and reducing modal bias errors [13].

To handle the high noise and inconsistent quality of cultural tourism UGC data, robust fusion strategies for low-quality data enhance utilization through noise filtering, feature correction, and modal completion [6]. Context-aware fusion frameworks dynamically optimize fusion using contextual information such as seasonality and visitor flow, improving stability in complex scenarios.

3.4. Large language model-enhanced recommendation

Multimodal large language models, with powerful semantic understanding and reasoning, have revolutionized traditional cultural tourism recommendation [5, 11]. Alignment techniques precisely distinguish local residents' preferences from those of non-local tourists, resolving ambiguous user identity. Retrieval-enhanced models integrated with cultural tourism knowledge bases mitigate

DOI: 10.54254/2755-2721/2026.AST36916

hallucination [3]. Culture-and-tourism-specific thinking chains enable logical reasoning and intelligent planning for long-sequence itineraries, suiting customized recommendation [10].

Intelligent agents and memory-enhanced frameworks enable long-term retention of user preferences and end-to-end interactive recommendation, suiting human-machine collaborative services [5]. Parameter-efficient fine-tuning enables low-cost vertical adaptation of large models; combined with visual retrieval architectures, it achieves precise image-text matching and personalized recommendation, enhancing practical applicability [11].

3.5. Generative, federated and conversational recommendation

Generative, federated, and conversational recommendation technologies address the industry's challenges in content generation, privacy security, and user interaction [7]. Generative models automatically produce personalized itineraries and attraction combinations, overcoming the limits of retrieval-based systems [4]. Federated learning trains models through cross-platform collaboration without sharing raw private data, mitigating data silos and privacy breaches and meeting compliance requirements [14].

Conversational recommendation systems adjust strategies dynamically through real-time user dialogues, suiting itinerary modification and personalized consultation. Multimodal pre-training frameworks unify feature representation across domains, enhancing generalization [9]. Human-AI collaborative models combine AI modeling with manual cultural verification, mitigating cultural distortion in AI-generated content [12].

4. Challenges in current research on recommendation systems for cultural and tourism scenarios

Data-level challenges: Although multimodal recommendation technologies for cultural tourism have formed a comprehensive framework, bottlenecks in data infrastructure, algorithmic modeling, and engineering implementation hinder large-scale adoption. This paper reviews the core challenges from the data and algorithm dimensions [7]. First, the industry suffers from severe data silos and lacks a unified annotation framework [14]. Data are scattered across platforms with inconsistent formats and standards, preventing collaborative modeling of multi-source data; China has yet to establish an exclusive multimodal annotation standard, and large variations across datasets hinder cross-domain comparison and generalization. Second, UGC data quality is highly uneven, with pronounced noise [6], undermining cross-modal alignment and preference mining accuracy [8]. Third, modal distribution is imbalanced, with pronounced cold-start and small-sample challenges. While textual and visual data are abundant and well-labeled, audio, short video, and IoT data are scarce and poorly annotated; data for niche scenic spots and emerging destinations are extremely limited, exacerbating cold-start problems and weakening small-sample generalization [7]. Finally, structured data (spatiotemporal, meteorological, passenger flow) differ greatly in semantics from unstructured text, image, and audio-visual data, making joint modeling difficult [4].

Algorithmic challenges: First, the cross-modal semantic gap cannot be completely eliminated; existing algorithms achieve only shallow alignment, and persistent modal bias makes user preference characterization incomplete. Second, high-precision models incur excessive computational costs, hindering deployment [10]; GNNs and multimodal large language models have vast parameters and compute demands that small and medium-sized platforms can hardly afford. Third, such algorithms may exacerbate information cocoons by amplifying historical preferences,

DOI: 10.54254/2755-2721/2026.AST36916

causing homogeneous recommendations and insufficient exposure of niche and intangible cultural heritage resources [7].

Fourth, large-scale models suffer from hallucinations and logical inconsistencies in long texts, so general-purpose models cannot fully meet the refined requirements of specific cultural tourism scenarios [5, 11]. Fifth, their dynamic adaptability is weak—static modeling fails to accommodate seasonal, holiday, and visitor-flow changes, reducing accuracy during special periods.

5. Conclusion

This paper systematically investigates multimodal recommendation systems for cultural and tourism scenarios, reviewing the domain-specific data architecture, technical landscape, and bottlenecks to construct a complete research framework. It classifies multimodal data, clarifying the sources, characteristics, and applicable scenarios of each dataset and distinguishing domestic and international application advantages. It also analyzes the five mainstream technical approaches and their principles, scenarios, and trade-offs, offering references for algorithm selection. Finally, it deconstructs the core bottlenecks from data and algorithm dimensions—data quality, modal fusion, computational cost, privacy and security, and cultural fidelity—charting the course for future research.

Addressing current bottlenecks and emerging trends, this paper proposes seven research directions: (1) lightweight culture-and-tourism-specific multimodal large models cutting parameters and costs for small and medium-sized platforms; (2) multimodal federated privacy-preserving recommendation enabling cross-platform collaborative modeling while balancing performance and privacy; (3) memory-enhanced agent-based human-machine collaboration for dynamic interaction and personalized itinerary planning; (4) robust fusion algorithms for low-quality data raising UGC utilization in real-world scenarios; (5) a knowledge-graph-constrained culture-fidelity generation framework preventing cultural distortion in AI-generated content; (6) a standardized domestic multimodal annotated dataset filling China's benchmark gap; and (7) a spatiotemporal adaptive recommendation framework integrating real-time data for dynamic strategy adjustment and higher accuracy in complex scenarios.

This systematic framework still has certain limitations. First, most existing algorithms remain at the laboratory validation stage without large-scale deployment or real-world traffic testing, so their stability and effectiveness require further verification. Second, the field lacks a unified evaluation framework and quantitative metrics for interpretability, fairness, and cultural suitability, hindering standardized comparison. Third, modal research is imbalanced: static text and image studies are mature, while fusion techniques for dynamic modalities such as short videos and audio remain scarce.

References

[1] Wang, Z., Höpken, W., & Jannach, D. (2025). A survey on point-of-interest recommendations leveraging heterogeneous data. *Information Technology & Tourism, 27*(1), 29–73.

[2] Islam, M. A., Mohammad, M. M., et al. (2022). Deep learning in point-of-interest recommendation: A comprehensive survey. *Neurocomputing, 472*, 306–325.

[3] Zhang, F., Chen, J., et al. (2026). DiMA: Distinguishing resident and tourist preferences via multi-modal LLM alignment for out-of-town cross-domain recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence, 40*(19), 16262–16270.

[4] Pan, L., Pan, Z., Cai, F., & Chen, H. (2026). Multimodal recommender systems: A survey of representation, modeling, and optimization. *Information Fusion, 128*, 103991.

DOI: 10.54254/2755-2721/2026.AST36916

[5] Lopez-Avila, A., & Du, J. (2025). A survey on large language models in multimodal recommender systems. *arXiv preprint*, arXiv:2505.09777.
[6] Zhang, Q., Wei, Y., et al. (2026). Multimodal fusion on low-quality data: A comprehensive survey. *Information Fusion, 135*, 104437.
[7] Bellogin, A., Dietz, L., Ricci, F., & Sánchez, P. (2026). Point of interest recommendation: Pitfalls and research directions. *ACM Transactions on Recommender Systems*. https://doi.org/10.1145/3816430.
[8] Khan, Q. W., Ahmad, R., et al. (2024). Multi-modal fusion approaches for tourism: A comprehensive survey of datasets, fusion techniques, recent architectures, and future directions. *Computers & Electrical Engineering, 116*, 109220.
[9] Liu, Q., Hu, J., et al. (2025). Multimodal recommender systems: A survey. *ACM Computing Surveys, 57*(12), 1–17.
[10] Zhao, Z., Fan, W., et al. (2024). Recommender systems in the era of large language models (LLMs). *IEEE Transactions on Knowledge and Data Engineering, 36*(11), 6889–6907.
[11] Nawara, D., & Kashef, R. (2025). A comprehensive survey on LLM-powered recommender systems: From discriminative, generative to multi-modal paradigms. *IEEE Access, 13*, 145772–145798.
[12] Zhang, G., Li, H., Li, S., Wang, B., & Ding, Z. (2024). MMKG-PAR: Multi-modal knowledge graphs-based personalized attraction recommendation. *Sustainability, 16*(5), 1–22.
[13] Chen, J., Cong, J., Li, M., Sun, Y., & Zhang, J. (2025). T-ECBM: A deep learning-based text-image multimodal model for tourist attraction recommendation. *Scientific Reports, 15*(1), 41638.
[14] Zhang, Q., Yang, P., et al. (2025). A survey on point-of-interest recommendation: Models, architectures, and security. *IEEE Transactions on Knowledge and Data Engineering, 37*(6), 3153–3172.