Data Quality Challenges in Multimodal Tourism Recommender Systems

Zehui Wang[1,2,*], Wolfram Höpken[1] and Dietmar Jannach[2]

[1]University of Applied Sciences Ravensburg-Weingarten, Doggenriedstrasse, Weingarten, Weingarten, 88250, Germany
[2]University of Klagenfurt, Universitätsstraße 65–67, Klagenfurt am Wörthersee, 9020, Austria

Abstract

Modern recommender systems increasingly consider diverse data modalities to enhance model performance. However, the quality of the underlying multimodal data (e.g. data from social media platforms) has received limited attention so far. Focusing on tourism recommendation, we examine real-world challenges encountered when using datasets from Yelp and Instagram. Issues include noise from fake reviews, hallucinated content from LLMs, modality contradictions, and demographic bias. We outline future directions such as quality-aware fusion, structured and controlled LLM-based content generation to reduce hallucinations, and user group differentiation to improve the robustness and reliability of multimodal recommendation systems.

Keywords

Multimodal Recommendation, Data Quality, Tourism, Large Language Models, Modality Fusion

1. Motivation

As a key application domain of recommender systems, tourism recommendation involves tasks such as providing personalized suggestions for transportation, accommodation, and points-of-interest (POIs) [1, 2, 3]. Recently, researchers have increasingly explored the use of diverse information types to improve model performance [4, 5]. Beyond check-in records that incorporate timestamps and geographic coordinates, additional sources such as POI profiles (e.g., textual attributes), social relationships (e.g., friendship networks), user feedback (e.g., reviews), and visual content (e.g., photos) have been integrated through representation learning, attention-based neural architectures, or late fusion schemes [4].

Despite the growing reliance on such heterogeneous data sources, the quality of multimodal information remains largely underexplored. This oversight can introduce noise, bias, and contradictions into the recommendation process. In the following, we draw on findings from real-world datasets to illustrate the challenges that arise when working with imperfect, incomplete, or biased multimodal data in tourism recommender systems.

2. Challenges in Multimodal Data Quality

Recent investigations in tourism-related user behavior analysis and next-POI recommendation have revealed several fundamental issues regarding the quality of multimodal data. Based on two real-world datasets from [6, 7], namely the Multimodal Yelp Dataset[1] and an Instagram dataset collected from the Lake Constance region in Germany[2], we summarize the key challenges as follows.

Noise in User-Generated and Model-Generated Content. Real-world datasets often contain noisy information that can significantly degrade the effectiveness of recommendation models. For example, the multimodal Yelp dataset is constructed from user-generated reviews and check-in records

DaQuaMRec Workshop at RecSys '25, September 22–26, 2025, Prague, Czech Republic
*Corresponding author.
Zehui.Wang@rwu.de (Z. Wang); Wolfram.Hoepken@rwu.de (W. Höpken); Dietmar.Jannach@aau.at (D. Jannach)
0000-0001-5401-8773 (Z. Wang); 0000-0002-4175-1295 (W. Höpken); 0000-0002-4698-8507 (D. Jannach)

[1]https://huggingface.co/datasets/wzehui/Yelp-Multimodal-Recommendation
[2]This dataset cannot be publicly shared due to platform restrictions and privacy considerations.

on the Yelp platform. However, some of these reviews originate from fake or automated accounts that produce fabricated interactions, leading to distorted representations of businesses. Prior work [7] has tackled this issue by analyzing behavioral patterns to identify non-human users. Suspicious accounts are first flagged based on unusually high posting frequency compared to normal user activity. These flagged accounts are then further examined for implausible movement speeds between consecutive check-ins and eventually removed. In the Instagram dataset, burst postings (multiple posts within a few minutes at the same location) are also observed. Retaining such data does not contribute to the analysis of meaningful sequential behavior patterns, but instead dilutes the signal of genuine user trajectories and biases the statistical distribution of activities.

In addition, with the increasing use of LLMs to process textual and visual modalities, a new form of semantic noise has emerged. Hallucinated or irrelevant content generated by LLMs can significantly degrade the quality of downstream embeddings and harm recommendation performance. One approach to address this issue is to use structured summarizations instead of open-ended generation. By enforcing predefined fields in the summaries, the process becomes more controllable and less susceptible to hallucination. In particular, for visual content, summaries are generated selectively, only for images that contain clear and interpretable visual cues [7]. Ambiguous images are filtered out to avoid introducing additional noise.

Demographic Bias in User Behavior and Content Distribution. Demographic bias occurs when there is a mismatch between the distribution of users in training data and the characteristics of the target population [8]. A typical case in tourism is the behavioral divergence between tourists and locals. For example, Instagram posts collected from Lake Constance region includes contributions from both travelers and nearby residents. If such data are used indiscriminately, the recommendation model may overfit to local preferences that are irrelevant for tourist-oriented next-POI suggestions.

Demographic bias can also arise from population imbalances in content creation. For instance, Instagram content may primarily come from younger users, whereas recommendations targeting regions like Lake Constance are often aimed at an older audience. This mismatch can reduce the relevance and effectiveness of recommendations. Moreover, the majority of media originates from the DACH region (Germany, Switzerland, Austria), reflecting cultural and behavioral patterns specific to this geographic area, which limits the generalizability of the dataset to broader international contexts. One possible approach to mitigate this issue is to further analyze the user profiles to improve audience segmentation and better align the system output with the preferences of the target user group.

Semantic Contradictions Across Modalities. Contradictions refer to semantic inconsistencies between content from different modalities. In the current datasets, such inconsistencies mainly stem from user-generated content (e.g., misaligned captions, incorrect location tags, irrelevant hashtags) or from business-provided information that is outdated or inaccurate. A common example arises in the multimodal Yelp dataset, where an LLM-generated summary based on the majority of user reviews might describe a restaurant as “clean and quiet”, while user-uploaded photos reveal a cluttered or noisy environment. Similar inconsistencies can also be found in Instagram posts, where images, captions, and location tags do not always align. A photo may be taken in one city but tagged to another, or hashtags may be added that are unrelated to the actual content. Such contradictions reduce the coherence and reliability of the fused multimodal representation, posing challenges for both modality alignment and inference stability in downstream models.

3. Possible Solutions & Future Directions

Given such challenges, we propose several directions for future research: (1) Future models should adopt data-aware multimodal learning, where the quality of each modality is explicitly modeled during training. This way, low-quality modalities can be dynamically downweighted in the modality fusion process [9, 10]; (2) Current fusion strategies assume consistent and complete input across modalities,

which rarely holds in real-world scenarios. Future research should develop robust fusion mechanisms to tolerate missing, noisy, or redundant inputs [11, 12, 13]. This includes support for preserving modality-specific uniqueness, reducing cross-modal redundancy, and enhancing synergistic signals to better adapt to imperfect data; (3) Future studies should also focus on modeling user intent divergence across different groups, such as tourists versus locals. Identifying and representing such behavioral differences during preprocessing or representation learning can mitigate distributional mismatches and improve recommendation relevance [14, 15]; (4) In addition to these directions, we observe that videos (e.g., ones posted by social media users or 360° immersive marketing videos) [16, 17] have not been leveraged much in the literature for improved tourism recommendations. We see the incorporation of these additional visual signals as a promising area for future work; (5) Given the increasing role of LLMs in multimodal recommendation, we advocate for more research on structured and controlled LLM-based generation to reduce hallucinations, thereby enhancing the reliability of generated content for downstream use [18].

In summary, integrating data quality considerations into both upstream preprocessing and downstream modeling is essential for building trustworthy, personalized, and data-aligned multimodal recommender systems.

References

[1] F. Ricci, M. Fuchs, U. Gretzel, W. Höpken, Recommender Systems in Tourism, Springer International Publishing, Cham, 2020, pp. 1–18. doi:10.1007/978-3-030-05324-6_26-1.

[2] J. Borràs, A. Moreno, A. Valls, Intelligent tourism recommender systems: A survey, Expert Systems with Applications 41 (2014) 7370–7389. doi:https://doi.org/10.1016/j.eswa.2014.06.007.

[3] J. L. Sarkar, A. Majumder, C. R. Panigrahi, S. Roy, B. Pati, Tourism recommendation system: a survey and future research directions, Multimedia Tools and Applications 82 (2023) 8983–9027. doi:10.1007/s11042-022-12167-w.

[4] Z. Wang, W. Höpken, D. Jannach, A survey on point-of-interest recommendations leveraging heterogeneous data, Information Technology & Tourism 27 (2025) 29–73. doi:10.1007/s40558-024-00301-3.

[5] Q. Zhang, P. Yang, J. Yu, H. Wang, X. He, S.-M. Yiu, H. Yin, A survey on point-of-interest recommendation: Models, architectures, and security, IEEE Transactions on Knowledge and Data Engineering 37 (2025) 3153–3172. doi:10.1109/TKDE.2025.3551292.

[6] Z. Wang, S. Schwarzenbacher, W. Höpken, M. Fuchs, Do travel destinations meet my expectations? a comparison of tourists' perceptions and destinations' self-presentation through instagram posts by a convolutional neural network, in: L. Nixon, A. Tuomi, P. O'Connor (Eds.), Information and Communication Technologies in Tourism 2025, Springer Nature Switzerland, Cham, 2025, pp. 289–299. doi:10.1007/978-3-031-83705-0_24.

[7] Z. Wang, W. Höpken, D. Jannach, Beyond visit trajectories: Enhancing poi recommendation via llm-augmented text and image representations, in: Proceedings of the Nineteenth ACM Conference on Recommender Systems, RecSys '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 521–526. URL: https://doi.org/10.1145/3705328.3748014. doi:10.1145/3705328.3748014.

[8] N. Neophytou, B. Mitra, C. Stinson, Revisiting popularity and demographic biases in recommender evaluation and effectiveness, in: Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I, Springer-Verlag, Berlin, Heidelberg, 2022, p. 641–654. doi:10.1007/978-3-030-99736-6_43.

[9] Q. Zhang, H. Wu, C. Zhang, Q. Hu, H. Fu, J. T. Zhou, X. Peng, Provable dynamic fusion for low-quality multimodal data, in: Proceedings of the 40th International Conference on Machine Learning, ICML'23, JMLR.org, 2023.

[10] S. Wei, Y. Luo, Y. Wang, C. Luo, Robust multimodal learning via representation decoupling, in: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4,

2024, Proceedings, Part XLII, Springer-Verlag, Berlin, Heidelberg, 2024, p. 38–54. doi:10.1007/978-3-031-72946-1_3.

[11] R. Lin, H. Hu, Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis, Transactions of the Association for Computational Linguistics 11 (2023) 1686–1702. doi:10.1162/tacl_a_00628.

[12] Y.-H. H. Tsai, M. Ma, M. Yang, R. Salakhutdinov, L.-P. Morency, Multimodal routing: Improving local and global interpretability of multimodal language analysis, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 1823–1833. URL: https://aclanthology.org/2020.emnlp-main.143/. doi:10.18653/v1/2020.emnlp-main.143.

[13] C. Xu, Y. Zhang, Z. Guan, W. Zhao, Trusted multi-view learning with label noise, in: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI '24, 2024. doi:10.24963/ijcai.2024/582.

[14] P. Sanchez, L. W. Dietz, Travelers vs. locals: The effect of cluster analysis in point-of-interest recommendation, in: Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 132–142. doi:10.1145/3503252.3531320.

[15] A. Derdouri, T. Osaragi, A machine learning-based approach for classifying tourists and locals using geotagged photos: the case of tokyo, Information Technology & Tourism 23 (2021) 575–609. doi:10.1007/s40558-021-00208-3.

[16] M. Casillo, F. Colace, A. Lorusso, D. Santaniello, C. Valentino, Integrating physical and virtual experiences in cultural tourism: An adaptive multimodal recommender system, IEEE Access 13 (2025) 28353–28368. doi:10.1109/ACCESS.2025.3539205.

[17] L. Argyriou, D. Economou, V. Bouki, Design methodology for 360° immersive video applications: the case study of a cultural heritage virtual tour, Personal and Ubiquitous Computing 24 (2020) 843–859. doi:10.1007/s00779-020-01373-8.

[18] D. King, Z. Shen, N. Subramani, D. S. Weld, I. Beltagy, D. Downey, Don't say what you don't know: Improving the consistency of abstractive summarization by constraining beam search, in: A. Bosselut, K. Chandu, K. Dhole, V. Gangal, S. Gehrmann, Y. Jernite, J. Novikova, L. Perez-Beltrachini (Eds.), Proceedings of the Second Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 555–571. doi:10.18653/v1/2022.gem-1.51.