Multimodality In Recommender Systems: Does It Help, and Should We Expect An Answer? ⋆

Aixin Sun

Nanyang Technological University, Singapore

Abstract

Multimodal recommender systems integrate diverse information sources, such as text and visual data, into predictive models for personalization. While multimodality promises richer representations and potentially improved relevance, it remains an open question whether multimodal inputs genuinely enhance recommender systems. In this talk, I present findings from an evaluation of reproducible multimodal RecSys models, designed to address this question. The results show that multimodality does not consistently lead to better performance. I argue that while multimodal inputs can yield incremental gains, their effectiveness must be considered in relation to user interaction dynamics, task objectives, and system context — their value is inherently contextual.

Keywords

Multimodality Recommender System, Evaluation, User-Decision Process

1. Does Multimodality Really Help?

Recommender systems (RecSys) are foundational technologies for digital platforms ranging from e-commerce to media streaming. Traditional systems primarily rely on user-item interactions, leveraging collaborative filtering or content features extracted from text. However, the proliferation of rich multimedia content has spurred a growing body of research on multimodal recommendation, where heterogeneous signals — such as text descriptions, product images, and short videos — are fused into the recommendation pipeline.

As a trending topic in RecSys, multimodal recommendation has attracted significant attention. Yet, a central question remains: Does multimodality truly improve recommender systems? Interestingly, three groups have conducted independent studies around the same time, each aiming to answer this question from different perspectives and experimental setups [1, 2, 3]. In this extended abstract, I briefly review key findings from [3]. The authors collected 41 papers published between 2019 and 2024 in top-tier venues such as SIGIR, WWW, TOIS, and TKDE. A paper qualifies as a study on multimodal RecSys if it introduces a novel technique and addresses issues specific to multimodal recommendation.

While the community increasingly emphasizes reproducibility, not all papers release their source code or datasets for various reasons. Among the 41 papers, 12 were considered reproducible — meaning both *code reproducible* (source code publicly available and functioning correctly) and *dataset available* (publicly accessible datasets or raw data with preprocessing scripts). For benchmarking, the team used three datasets: two e-commerce datasets (Amazon and Taobao) and one short-video dataset (DY). Largely following the experimental settings of the reproducible papers, each dataset was randomly partitioned into training, validation, and test sets with an 8:1:1 ratio. Although this random split can introduce data leakage [4, 5], it remains a common practice in academic research, as it avoids modifying models to accommodate temporally aware splits. The evaluation metrics used were Recall and NDCG.

The team conducted multiple evaluations focusing on the role of multimodality. To assess its benefits, two classic baselines — ItemKNN and UserKNN — were employed, both relying solely on interaction

DaQuaMRec'25: First International Workshop on Data Quality-Aware Multimodal Recommendation, September 22, 2025, Prague, Czech Republic

⋆ This paper is an extended abstract of my keynote talk at the DaQuaMRec workshop @ RecSys 2025.

axsun@ntu.edu.sg (A. Sun)

https://personal.ntu.edu.sg/axsun/ (A. Sun)

0000-0003-0764-4258 (A. Sun)

data. Experimental results show that these traditional KNN-based methods perform comparably to, or even better than, several sophisticated multimodal recommendation models. Moreover, when comparing results across the three datasets, no consistent advantage of multimodality emerges. In particular, very few methods outperform the KNN baselines on the DY dataset.

The team then conducted experiments comparing the recommendation accuracy of multimodal and single-modality models. Interestingly, multimodal systems do not always achieve the best performance compared to their single-modality counterparts. Furthermore, different types of modal information contribute differently depending on the recommendation scenario. Specifically, in e-commerce settings, textual features often play a more important role, whereas visual information tends to be marginally more useful for short-video recommendations.

While many more interesting findings are presented in [3], the results highlighted here do not always provide a positive answer to the question posed in this section's title. In fact, negative or inconclusive results are not uncommon in RecSys research. For example, through a large-scale evaluation comparing 18 algorithms across 85 datasets, McElfresh et al. [6] observed that classic ItemKNN remains a competitive method outperforming many others. More recently, two papers have discussed broader issues in RecSys research, and their titles are themselves quite indicative [7, 8]. These negative findings also relate to the second question in the title of this paper.

2. Should We Expect an Answer?

Before addressing the question of multimodality specifically, it is helpful to view recommender systems from a broader perspective. ACM RecSys is a conference under SIGCHI, emphasizing human-computer interaction. Yet, many academic papers in the field focus primarily on algorithmic or methodological advances, often overlooking the HCI dimension.

Recently, we conducted a survey of real-world recommender systems, including only those studies that reported online A/B testing results on production systems [9]. The survey revealed that real-world recommendation scenarios are highly diverse and often differ substantially in their interaction settings, which in turn shape distinct recommendation logics.

In particular, we categorize commonly observed RecSys tasks into two broad types: *Transaction-Oriented RecSys* and *Content-Oriented RecSys*. The goal of the former is to drive transactional actions — optimizing for conversion rates, revenue, or purchase likelihood. E-commerce platforms are typical examples. In contrast, the latter focuses on promoting user consumption and engagement, optimizing for metrics such as dwell time, clicks, or user satisfaction to encourage continued interaction, such as watching videos, listening to music, or reading news articles.

Multimodal RecSys must be situated within the larger ecosystem of recommendation. Research often narrows to algorithmic accuracy, but real-world systems involve broader considerations: interfaces, feedback loops, optimization objectives, and operational constraints like cost and latency. Evaluation differences between offline metrics and online A/B testing further complicate insights.

The effectiveness of different modalities depends heavily on context, visibility, and the extent to which these modalities influence user decision-making [10]. For instance, in e-commerce, product images and textual descriptions can be critical factors affecting whether users click on a product among many recommended items and ultimately make a purchase. In contrast, within short-video platforms, users primarily engage with visual content, while textual cues are often ignored. In audio streaming, songs are delivered continuously unless the user actively intervenes. In this case, visual or textual attributes may influence only the selection of the first song that initiates playback, whereas subsequent recommendations depend less on multimodal signals. A more detailed discussion on the different stages of user-item interaction can be found in [10].

Mixed findings suggest no universal superiority of multimodality. Domain specificity, task dependence, evaluation design, and user visibility all shape outcomes. While multimodality enriches representation, its value is inherently contextual.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT-5 for spelling check and sentence polishing. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] C. Pomo, M. Attimonelli, D. Danese, F. Narducci, T. Di Noia, Do recommender systems really leverage multimodal content? a comprehensive analysis on multimodal representations for recommendation, in: Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 2377–2387. URL: https://doi.org/10.1145/3746252.3761398. doi:10.1145/3746252.3761398.

[2] Y. Ye, J. Fu, Y. Song, K. Zheng, J. M. Jose, Are multimodal embeddings truly beneficial for recommendation? a deep dive into whole vs. individual modalities, ArXiv abs/2508.07399 (2025). URL: https://arxiv.org/abs/2508.07399, to appear in ECIR 2026.

[3] H. Zhou, Y. Zhang, A. Sun, Z. Shen, Does multimodality improve recommender systems as expected? a critical analysis and future directions, ArXiv abs/2508.05377 (2025). URL: https://arxiv.org/abs/2508.05377.

[4] Y. Ji, A. Sun, J. Zhang, C. Li, A critical study on data leakage in recommender system offline evaluation, ACM Trans. Inf. Syst. 41 (2023). doi:10.1145/3569930.

[5] A. Sun, Take a fresh look at recommender systems from an evaluation standpoint, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23, ACM, New York, NY, USA, 2023, p. 2629–2638. doi:10.1145/3539618.3591931.

[6] D. McElfresh, S. Khandagale, J. Valverde, J. P. Dickerson, C. White, On the generalizability and predictability of recommender systems, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NeurIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022.

[7] A. Said, M. S. Pera, M. D. Ekstrand, We're still doing it (all) wrong: Recommender systems, fifteen years later, in: Beyond Algorithms: Reclaiming the Interdisciplinary Roots of Recommender Systems Workshop (BEYOND 2025), co-located with the ACM RecSys 2025, Prague, Czech Republic, 2025.

[8] K. Higley, R. Burke, M. D. Ekstrand, B. P. Knijnenburg, What news recommendation research did (but mostly didn't) teach us about building a news recommender, in: Beyond Algorithms: Reclaiming the Interdisciplinary Roots of Recommender Systems Workshop (BEYOND 2025), co-located with the ACM RecSys 2025, Prague, Czech Republic, 2025.

[9] K. Zou, A. Sun, A survey of real-world recommender systems: Challenges, constraints, and industrial perspectives, ArXiv abs/2509.06002 (2025). URL: https://arxiv.org/abs/2509.06002.

[10] A. Sun, A task-centric perspective on recommendation tasks, ArXiv abs/2503.21188 (2025). URL: https://arxiv.org/abs/2503.21188, to appear in Communications of the ACM (CACM).