Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/154/2025.TJ23119

The Evolution of Multi-modal Recommendation Algorithms for Short Videos

Wenxin Xiao

Faculty of Science and Technology, Beijing Normal-Hong Kong Baptist University, Zhuhai, China
2267167975@qq.com

Abstract: The rapid proliferation of short video platforms has necessitated the evolution of recommendation systems beyond traditional unimodal approaches. This survey comprehensively analyzes the advancements, challenges, and future directions of multi-modal recommendation algorithms tailored for short videos. Unlike conventional methods reliant on singular data sources (e.g., user logs or text), multi-modal systems integrate visual, audio, textual, and behavioral signals to address critical limitations such as data sparsity, cold starts, and dynamic user intent. We systematically categorize multi-modal fusion strategies—data-level, feature-level, and decision-level—and highlight their applications in short video scenarios, including cross-modal alignment (e.g., contrastive learning), audio-visual synchronization, and behavior-driven personalization. Key challenges span data noise, computational inefficiency, interest drift, and privacy risks while emerging trends emphasize lightweight fusion, generative AI-driven content synthesis, and explainable recommendation mechanisms. By synthesizing cutting-edge frameworks like M3CSR and SVARM, this review underscores how multi-modal techniques enhance recommendation accuracy, user engagement, and platform competitiveness. Future research must prioritize efficient edge-compatible architectures, ethical AI governance, and interdisciplinary innovations to sustain the growth of short video ecosystems.

Keywords: Multimodal recommendation systems, short video recommendations, cross-modal alignment, user behavior modeling

1. Introduction

During the evolution of the digital media ecosystem, short video platforms such as TikTok, Kwai, and YouTube are leading to the restructuring of global content consumption. Through fragmented and high-density information presentation, these platforms not only reshape the user attention allocation mechanism but also push the recommendation system to the strategic core of platform competition. Recommendation algorithms have become the core assets of tech giants. For instance, TikTok's recommendation engine is based on user interest signals rather than social connections, enabling it to capture users' preferences within seconds and dynamically adjust the content flow [1]. This mechanism significantly enhances user stickiness and platform retention rate and has become a key support for ByteDance's global strategy. While traditional recommendation methodologies—including collaborative filtering and content-based filtering — have demonstrated efficacy in domains like e-commerce and long-form video platforms, their application to short-video scenarios reveal systemic limitations under the recommendation standard of AUC and other methods [2-4]. These

DOI: 10.54254/2755-2721/154/2025.TJ23119

traditional recommendation methods have problems such as sparse data and cold boot, primarily are their reliance on unimodal data processing frameworks, where algorithms predominantly leverage singular information dimensions such as user interaction logs or textual metadata [5-7]. Such approaches prove inadequate in addressing the inherent multimodality of short-video content, which necessitates the synergistic analysis of dynamic visual compositions, affective audio signals, real-time textual interactions, and microsecond-level behavioral sequences. Therefore, a breakthrough in this field urgently requires the establishment of a new generation of recommendation framework that integrates multimodal feature learning, dynamic interest modeling and low-latency inference. Moreover, with the explosive growth of user-generated content (UGC) and the increasing complexity of user behaviors, recommendation systems are facing higher requirements for real-time performance and personalization. Under such circumstances, this paper aims to explore the shift from traditional to multimodal recommendation approaches in short-video platforms, highlighting the limitations of existing methods in processing rich multimodal content. It further discusses key advancements such as multimodal feature learning and real-time inference, while emphasizing the significance of building more personalized and efficient recommendation systems.

2. Multimodal recommendation system

2.1. Definition and characteristics of Multi-modal data

Multi-modal data refers to information derived from multiple sources, such as text, images, audio, and video. Unlike traditional single-modal data, which relies on a single type of input, multi-modal data integrates diverse modalities to provide a more comprehensive representation of content.

In short video recommendation, multi-modal data includes visual features (e.g., objects, scenes, and colors), audio signals (e.g., background music and speech), and textual metadata (e.g., captions and hashtags). These modalities complement each other, enhancing content understanding and personalization. Thus, effectively fusing and interpreting these modalities is crucial for improving recommendation accuracy and user engagement.

2.2. Main approaches in Multi-modal recommender systems

Multimodal recommender systems leverage various fusion strategies to effectively integrate heterogeneous data sources, and these methods are commonly categorized into data-level fusion, feature-level fusion, and decision-level fusion [8], as shown in Figure 1.

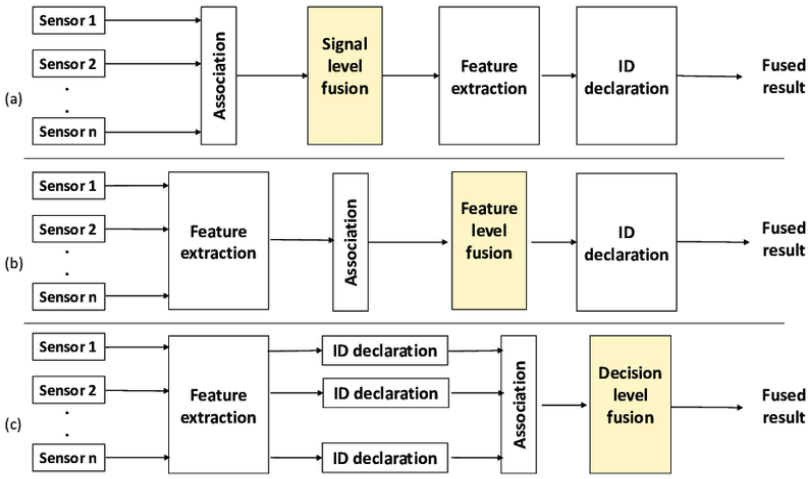

Figure 1: Data fusion at three different levels: (a) signal-level fusion, (b) feature-level fusion, and (c) decision-level fusion [9]

DOI: 10.54254/2755-2721/154/2025.TJ23119

Data-level fusion operates at the earliest stage by directly merging raw data from different modalities—such as audio, video, and text—into a unified representation before any feature extraction takes place [10]. This method preserves complete information and tends to yield accurate results due to minimal data loss. However, it demands high synchronization across modalities and is thus best suited to structured environments with stable system communications, such as integrated e-commerce platforms or curated content applications like Netflix, where video, subtitles, and user metadata can be tightly coupled [11].

Feature-level fusion, on the other hand, extracts feature independently from each modality—e.g., visual CNN embeddings, audio MFCCs, and textual embeddings—and then transforms and integrates them into a common feature space [12]. Its advantage is that only one classifier is needed to complete the classification task, thus reducing the computational complexity [13]. In addition, it combines features from different modes (such as images, sounds, etc.) It can improve the stability and reliability of the system [14]. It is widely adopted in platforms like TikTok and Kwai, where the interaction of dynamic visual elements, background music, and on-screen text is key to understanding content relevance and user interest. The flexibility of feature-level fusion also allows it to adapt to noisy or asynchronous data, making it a popular choice in real-world, high-throughput recommendation settings.

Decision-level fusion is when each classifier makes decisions independently and then combines these decisions in some way to form a final unified decision [15]. This method is often used to improve the accuracy and robustness of classification These individual outputs—such as content relevance scores or user engagement probabilities—are then combined, often via ensemble strategies or weighted voting, to arrive at a final recommendation. This method excels in systems where modality-specific signals are strong and relatively independent, such as news aggregators or hybrid recommendation engines like YouTube's multi-source model, which combines search behavior, watch history, and thumbnail preferences [16]. While decision-level fusion offers high interpretability and robustness, it may struggle to capture subtle cross-modal dependencies unless further optimized.

In practice, these fusion strategies are not mutually exclusive. State-of-the-art systems frequently utilize hybrid architectures that combine feature-level and decision-level techniques, thereby leveraging both deep integration and modular interpretability. Understanding the trade-offs among these methods is crucial for selecting or designing an architecture that balances performance, efficiency, and scalability in a specific application context.

3. Multi-modal applications in short video recommendation systems

3.1. Characteristics of short video platforms and recommendation needs

Short video recommendation systems differ significantly from traditional ones due to their unique content characteristics and user interaction patterns. Platforms like TikTok and Kwai operate on highly dynamic timelines, where the content lifecycle is extremely short, and user attention spans are limited. Unlike conventional video platforms that primarily rely on explicit user preferences (e.g., ratings or subscriptions), short video platforms emphasize implicit feedback, including watch time, swipe behavior, pauses, comment frequency, and sharing rates [17]. These interactions provide rich but noisy signals that require robust algorithms capable of capturing fine-grained user intent and preferences.

Additionally, short video content is inherently multimodal—integrating text (titles, hashtags), audio (music, speech), and visuals (frames, motion, objects)—making the use of unimodal models inadequate. Users might be attracted to a dance clip because of a catchy beat, a trending hashtag, or vibrant visuals, all of which must be interpreted in tandem. This complexity necessitates the

DOI: 10.54254/2755-2721/154/2025.TJ23119

deployment of advanced multi-modal recommendation algorithms that can handle the heterogeneous nature of input data and provide real-time personalization at scale.

3.2. Multi-modal technology applications in short video recommendations

Multimodal recommendation technologies enhance user satisfaction and engagement by incorporating various modalities into the recommendation pipeline. Techniques such as cross-modal feature fusion, user behavior modeling, and contrastive representation learning have been widely adopted in recent models.

Textual metadata such as titles, captions, and hashtags convey semantic cues, which, when combined with visual analysis (e.g., scene recognition, object detection, color palette extraction), improve content understanding. The M3CSR framework, for instance, addresses the cold-start problem by jointly encoding textual descriptions and video frame sequences. This allows the system to make informed recommendations even for newly uploaded content with limited interaction data. Experimental results from M3CSR show an average 12.3% improvement in click-through rate (CTR) and a 9.6% increase in watch duration compared to single-modality models [18].

Audio + Video Synchronization relies on background music, sound effects, and speech not only to enhance video appeal but also to influence user retention. The SVARM model synchronizes audio spectral analysis with scene action recognition to detect synergistic patterns—such as rhythmic dancing to pop music—which correlates strongly with higher engagement [19]. The researchers propose a multimodal graph Convolutional network (MMGCN) framework that aims to enable personalized micro-video recommendations by integrating information from users' multimodal interactions with items, including visual, audio, and text. The model was tested on TikTok, Kwai, and MovieLens data sets, and showed better performance than other multi-modal recommendation methods [20].

Analyzing user comments—especially through sentiment analysis using NLP techniques—helps infer viewer satisfaction. Text + User Behavior Modeling collects emojis, slang, and sarcasm, providing a holistic understanding of user intent when paired with CTR and sharing metrics [21]. For example, a video receiving overwhelmingly negative feedback may be downweighed in ranking, while content with high positive sentiment and shareability can be prioritized for exposure. Beyond algorithmic optimization, recent research has explored emotion-aware design for social media recommendation systems, aiming to support users' emotional well-being. A case study by Deng on TikTok introduced an emotion reminder prototype that combines a Naive Bayes text classifier with an SVM-based facial expression analyzer. Although the sentiment classification accuracy was moderate (0.51 for text and 0.69 for facial expression), user feedback revealed a higher sense of emotional awareness and control during browsing [22]. Unlike automatic content filtering, the reminder served as a gentle emotional cue, which users preferred for maintaining privacy and enabling a balanced emotional experience that includes both positive and negative content. These findings highlight that emotion-sensitive recommendation interfaces can enhance user trust and long-term engagement when designed with psychological nuance in mind.

Cross-modal Representation Alignment benefits significantly from advanced techniques such as contrastive learning [23], which aim to align the embedding spaces of different modalities and resolve semantic gaps. For instance, associating a "rock music" label with energetic visual styles or synchronizing emotional speech with scene tension are examples of semantic alignment. Netflix has pioneered similar techniques in long-form content recommendation—using multi-head self-attention and cross-modal transformers to align visual, audio, and subtitle embeddings within the same latent space [24]. Although targeted at episodic content, these methods provide a technical blueprint for short-video platforms, where the challenge of sparse interaction data and rapidly shifting trends demands equally robust alignment strategies. The successful adaptation of such architectures in short

DOI: 10.54254/2755-2721/154/2025.TJ23119

video systems has led to improved AUC and NDCG scores, enabling finer-grained personalization and thematic continuity.

Overall, the integration of multimodal technologies into short-video recommendation systems not only improves the precision and relevance of recommendations but also yields measurable commercial benefits—such as increased advertising revenue, longer user engagement times, and improved content discoverability. As short video platforms continue to evolve, the adoption and optimization of multimodal recommendation strategies will be critical for sustaining competitive advantage.

4. Challenges and future trends

4.1. Data and algorithm challenges

The problem of data sparsity is particularly prominent. User-item interaction behaviors typically follow a long-tail distribution, making it difficult for unpopular items to obtain effective representations. Additionally, multimodal data is often accompanied by noise interference, such as irrelevant backgrounds in images or redundant information in text, which may mislead model learning [25]. The cross-modal alignment problem further complicates the situation. Semantic differences between different modalities (such as the matching deviation between product images and descriptions) need to be addressed through refined mapping. Specifically, Meta points out that manually designed sparse features in traditional DLRM systems (e.g., page ids clicked, etc.) lose sequence information and fine-grained context [26].

Feature selection and fusion are crucial for multi-modal learning, yet balancing information richness and computational efficiency remains difficult [27]. Transformer-based architectures improve representation learning but demand high computational resources, limiting real-time recommendations. Moreover, the lack of interpretability in deep learning-based multi-modal models hinders their practical deployment, necessitating methods for better transparency. The extensive use of multi-modal data—including visual embeddings, audio signals, and comment text—raises growing privacy concerns [28]. These data types often contain sensitive personal cues, such as facial expressions, voice characteristics, or personal opinions, which could be exploited if not properly protected. To address this, advanced privacy-preserving techniques like federated learning (which keeps user data localized on devices while training global models) and differential privacy (which injects noise to mask individual-level information) are considered as potential approaches.

4.2. Future development trends

Building upon the challenges identified in data quality, algorithm design, and user privacy, future research in multimodal recommender systems is shifting toward solutions that are not only more efficient but also contextually adaptive and ethically aligned. With the exponential growth of short video content, multimodal recommendation algorithms are evolving to address existing bottlenecks through innovative approaches [29].

Techniques such as knowledge distillation and dynamic modality pruning help reduce the computational burden introduced by Transformer-based architectures, making it feasible to deploy high-performing models on mobile devices. These methods can directly mitigate the algorithmic challenge of resource consumption, while cross-modal pretraining boosts feature robustness and generalization—offering partial solutions to data sparsity and noise interference.

Context-aware recommendation addresses user interest drift by integrating spatiotemporal data and sensor inputs (e.g., GPS, and device usage patterns). These networks can capture user context in real-time, enhancing interest modeling and reducing cold-start limitations for underrepresented content.

DOI: 10.54254/2755-2721/154/2025.TJ23119

4.3. Integration with generative AI

The emergence of generative AI offers promising solutions for both user engagement and feature enrichment. Diffusion models can synthesize personalized content (e.g., tailored video previews), potentially alleviating the semantic gap across modalities. Meanwhile, multimodal conversational agents such as ChatGPT-4 with vision allow users to interactively refine preferences, improving interpretability. However, as these techniques increasingly access sensitive data, federated learning and differential privacy remain essential for ethical compliance—ensuring that performance improvements do not compromise user trust. Moreover, Meta is driving research into transformer, state-space models, linear attention, and KV-cache optimization, all of which are future recommended architecture trends [26].

5. Conclusion

This paper systematically reviews the evolution process of multimodal recommendation algorithms for short videos and summarizes their core contributions in data modeling, feature fusion, and application scenarios. In response to the high dynamics, multimodal heterogeneity, and complex user behaviors of short video content, existing research has significantly improved recommendation accuracy and user immersion experience through cross-modal alignment, lightweight fusion frameworks, and deep representation learning. Multimodal recommendation systems not only overcome the data sparsity and cold start problems of traditional single-modal methods but also capture user intentions in a refined manner by integrating multi-source signals such as visual, audio, text, and behavioral sequences, becoming the core technical engine for content distribution on short video platforms.

Future research should further explore the following directions: Firstly, develop a multimodal fusion architecture that balances efficiency and accuracy, addressing the high computational cost of pre-trained models and the problem of edge-device compatibility; Secondly, deepen the collaborative innovation between generative AI and recommendation systems, achieving dynamic personalized content generation and interaction while ensuring content quality and ethical security; Thirdly, build a transparent and interpretable recommendation mechanism, integrating visualization technology and privacy computing framework, to balance algorithm performance and user trust. Moreover, interdisciplinary cross-pollination (such as cognitive science, and human-computer interaction) will drive the development of multimodal recommendation systems towards a more humanized and scenario-oriented direction, contributing to the sustainable development of the short-video ecosystem.

References

[1] *Ye, Josh. "Explainer: What Is so Special about TikTok's Technology." Reuters, 26 Apr. 2024, www.reuters.com/technology/what-is-so-special-about-tiktoks-technology-2024-04-26/.*

[2] *Liu, H., & Liu, C. (2024). Research on short video recommendation algorithms based on user preferences under the background of big data. Broadcasting and Television Networks, 31(06), 104-106. https://doi.org/10.16045/j.cnki.catvtec.2024.06.027*

[3] *Schröder, G., Thiele, M., & Lehner, W. (2011, October). Setting goals and choosing metrics for recommender system evaluations. In UCERSTI2 workshop at the 5th ACM conference on recommender systems, Chicago, USA (Vol. 23, p. 53).*

[4] *Chen, M., & Liu, P. (2017). Performance evaluation of recommender systems. International Journal of Performability Engineering, 13(8), 1246.*

[5] *Lin, T. (2023). Research Progress on Collaborative Filtering Recommendation Algorithm. Information Record Materials, 24 (11), 16-18. doi:10.16009/j.cnki.cn13-1295/tq.2023.11.069.*

[6] *Wu, Z., Zhang, D., & Li, G. (2025). A multimodal fusion recommendation algorithm based on joint self-supervised learning. Computer Applications, 1-14. http://kns.cnki.net/kcms/detail/51.1307.TP.20241008.1103.002.html*

DOI: 10.54254/2755-2721/154/2025.TJ23119

[7] Huo, Y., Jin, B., & Liao, Z. (2024). Short video recommendation model enhanced by multi-modal information. Journal of Zhejiang University (Engineering Edition), 58(06), 1142-1152

[8] Liu, P. R. (2022). Research on multimodal data fusion applications in the field of intelligent education. Intelligent City, 8(11), 13-15. https://doi.org/10.19301/j.cnki.zncs.2022.11.005

[9] Khan, Md Nazmuzzaman, and Sohel Anwar. "Paradox Elimination in Dempster–Shafer Combination Rule with Novel Entropy Function: Application in Decision-Level Multi-Sensor Fusion." Sensors, vol. 19, no. 21, 5 Nov. 2019, p. 4810, https://doi.org/10.3390/s19214810

[10] T, Haylat. "DATA FUSION." Haileleol Tibebu, 3 Feb. 2020, medium.com/haileleol-tibebu/data-fusion-78e68e65b2d1.

[11] Zhang, Pengfei, et al. "A Data-Level Fusion Model for Unsupervised Attribute Selection in Multi-Source Homogeneous Data." Information Fusion, vol. 80, 1 Apr. 2022, pp. 87–103, www.sciencedirect.com/science/article/pii/S1566253521002256, https://doi.org/10.1016/j.inffus.2021.10.017. Accessed 17 May 2023.

[12] Ross, Arun. "Fusion, Feature-Level." Encyclopedia of Biometrics, 2009, pp. 597–602, https://doi.org/10.1007/978-0-387-73003-5_157.

[13] Kalpeshkumar Ranipa, et al. "A Novel Feature-Level Fusion Scheme with Multimodal Attention CNN for Heart Sound Classification." Computer Methods and Programs in Biomedicine, vol. 248, 15 Mar. 2024, pp. 108122–108122, https://doi.org/10.1016/j.cmpb.2024.108122. Accessed 17 Nov. 2024.

[14] Rasool, Rabab A. "Feature-Level vs. Score-Level Fusion in the Human Identification System." Applied Computational Intelligence and Soft Computing, vol. 2021, 21 June 2021, pp. 1–10, https://doi.org/10.1155/2021/6621772. Accessed 19 Aug. 2021.

[15] Yashvi Chandola, et al. "Lightweight End-To-End Pre-Trained CNN-Based Computer-Aided Classification System Design for Chest Radiographs." Elsevier EBooks, 1 Jan. 2021, pp. 167–183, https://doi.org/10.1016/b978-0-323-90184-0.00001-1.

[16] Osadciw, Lisa, and Kalyan Veeramachaneni. "Fusion, Decision-Level." Encyclopedia of Biometrics, 2009, pp. 593–597, https://doi.org/10.1007/978-0-387-73003-5_160. Accessed 12 Apr. 2025.

[17] Vargas, S., & Castells, P. (2011, October). Rank and relevance in novelty and diversity metrics for recommender systems. In Proceedings of the fifth ACM conference on Recommender systems (pp. 109-116).

[18] Chen, G., Sun, R., Jiang, Y., Cao, J., Zhang, Q., Lin, J., ... & Zhang, X. (2024, October). A Multi-modal Modeling Framework for Cold-start Short-video Recommendation. In Proceedings of the 18th ACM Conference on Recommender Systems (pp. 391-400).

[19] Jiang, X. Q. (2024). Short video advertisement recommendation model based on large model text feature enhancement and fusion (Master's thesis). Jilin Institute of Chemical Technology. https://doi.org/10.27911/d.cnki.ghjgx.2024.000242

[20] Wei, Yinwei, et al. "MMGCN." Proceedings of the 27th ACM International Conference on Multimedia, 15 Oct. 2019, https://doi.org/10.1145/3343031.3351034.

[21] Kumar, Sumit. "A Guide to User Behavior Modeling." Sumit's Diary, xxxx, 7 Jan. 2024, blog.reachsumit.com/posts/2024/01/user-behavior-modeling-recsys/.

[22] Deng, Yawen. "Design an Emotionally Positive Experience via Sentiment Classification for Social Media Recommendation Systems: A Case Study in TikTok." DIVA, 2023, www.diva-portal.org/smash/record.jsf?dswid=-8863&pid=diva2%3A1814509. Accessed 12 Apr. 2025.

[23] Liu, Zhiyuan, et al. Representation Learning for Natural Language Processing. Springer Nature, 2 Nov. 2023.

[24] Netflix Technology Blog. "Foundation Model for Personalized Recommendation - Netflix TechBlog." Medium, Netflix TechBlog, 21 Mar. 2025, netflixtechblog.com/foundation-model-for-personalized-recommendation-1a0bd8e02d39.

[25] He, Y. (2023). Research on short video recommendation methods based on graph contrastive representation learning (Master's thesis). Hefei University of Technology. https://doi.org/10.27101/d.cnki.ghfgu.2023.000017

[26] Reddy, Sri. "Sequence Learning: A Paradigm Shift for Personalized Ads Recommendations." Engineering at Meta, 19 Nov. 2024, engineering.fb.com/2024/11/19/data-infrastructure/sequence-learning-personalized-ads-recommendations/.

[27] Vargas, S., & Castells, P. (2011, October). Rank and relevance in novelty and diversity metrics for recommender systems. In Proceedings of the fifth ACM conference on Recommender systems (pp. 109-116)

[28] Li, A. X. (2021). Design and implementation of a short video recommendation system based on multimodal information and differential privacy (Master's thesis). Beijing University of Posts and Telecommunications. https://doi.org/10.26969/d.cnki.gbydu.2021.002187

[29] Lu, Y., Huang, Y., Zhang, S., Han, W., Chen, H., Fan, W., ... & Wu, F. (2023, July). Multi-trends enhanced dynamic micro-video recommendation. In CAAI International Conference on Artificial Intelligence (pp. 430-441). Singapore: Springer Nature Singapore