

Research on Multimodal Recommendation Algorithms for Short Video Platforms

Zhenhang Liu

*School of Artificial Intelligence, Chongqing Technology and Business University, Chongqing, China
liuzhenhang@ctbu.edu.cn*

Abstract. Short video platforms have become important channels for online information distribution, entertainment, and digital consumption. Because short videos usually contain visual frames, audio signals, captions, covers, and textual descriptions, traditional recommendation algorithms based only on user-item interactions are insufficient for fully understanding user preferences and item characteristics. Multimodal recommendation has therefore become an important research direction for improving short video recommendation quality, especially under data sparsity, cold start, and noisy feedback conditions. This paper reviews recent research on multimodal recommendation algorithms for short video platforms. It first introduces the industrial recommendation pipeline and sequential user behavior modeling, and then summarizes multimodal feature representation, graph-based multimodal recommendation, self-supervised learning, cold-start recommendation, fine-grained interest modeling, and dataset evaluation. Representative models are compared according to their tasks, modalities, methods, contributions, and limitations. The results find that multimodal information improves short video recommendation by enhancing content understanding and user preference modeling, but future research still needs to address modality noise, semantic gaps, exposure bias, computational cost, and explainable recommendation.

Keywords: Short Video Recommendation, Multimodal Recommendation, Graph Neural Network, Self-supervised Learning, Cold-start Recommendation

1. Introduction

Short video platforms such as TikTok, Kwai, and YouTube Shorts have changed the way users consume online information. Unlike traditional item recommendation, short video recommendation must process fast-changing user behavior, newly uploaded videos, short interaction sessions, and rich multimodal content. A recent survey shows that multimodal features are useful for alleviating data sparsity and improving semantic understanding in recommendation systems [1]. Therefore, multimodal recommendation has become an important research direction for short video platforms.

Industrial video recommendation systems usually follow a multi-stage architecture. The YouTube deep neural recommendation system uses candidate generation to retrieve a small set of relevant videos and then using a ranking model to order them according to user engagement objectives [2]. In ranking tasks, Deep Factorization-Machine-based Neural Network (DeepFM) models low-order and

high-order feature interactions in an end-to-end way for click-through rate (CTR) prediction [3]. They provide the basic industrial background for short video ranking systems.

Sequential modeling is another important foundation. Self-attention to sequential recommendation (SASRec) applied self-attention to sequential recommendation and captured both short-term and long-term user behavior patterns [4]. Sequential Recommendation with Bidirectional Encoder Representations from Transformer (BERT4Rec) further used bidirectional self-attention and a Cloze objective to model user behavior sequences [5]. These methods are important because short video users often show dynamic and session-based interests.

With the development of multimodal representation learning, visual-language models demonstrate that image-text contrastive pre-training can learn transferable visual representations [6]. Multi-modal Graph Convolution Network (MMGCN) first emphasized modality-specific user preferences in micro-video recommendation, while Light Graph Convolution Network (LightGCN) simplified graph convolution for collaborative filtering [7, 8]. Latent structure mining method for multimodal recommendation (LATTICE) mined latent item-item structures from multimodal features [9]. Later studies such as Multi-Modal Graph Contrastive Learning (MMGCL), Meta Multimodal Fusion (MetaMMF), Multi-Modal Self-Supervised Learning (MMSSL), Bootstrapped Multi-Modal Model (BM3), and Freezing and Denoising Graph Structures for Multimodal Recommendation (FREEDOM) further explored contrastive learning, dynamic fusion, self-supervision, bootstrapping, graph freezing, and graph denoising [10-14]. For short video-specific problems, Multi-modal Modeling Framework for Cold-start Short-video Recommendation (M3CSR) focuses on cold-start recommendation, Fine-grained Preference-modeling for Micro-video Recommendation (FRAME) models fine-grained user interests, KuaiRand and KuaiRec provide important datasets for unbiased and fully observed evaluation, and the improved matrix decomposition recommendation algorithm (DASVD) introduces duration preference into short video recommendation [15-19]. Knowledge graph-based recommendation is also relevant for future explainable recommendation because it can use structured semantic relations to improve interpretability [20].

This paper reviews the problems of how multimodal recommendation algorithms can improve short video recommendation under practical challenges such as fast-changing user interests, sparse interactions, cold-start videos, noisy multimodal content, and biased evaluation data. This review provides a systematic reference for future research on multimodal short video recommendation.

2. Foundations of short video recommendation

2.1. Industrial recommendation pipeline

Short video recommendation systems usually contain several stages, as shown in Figure 1. A platform may contain millions or billions of videos, allowing a limited number to be displayed to each user. Candidate generation improves efficiency by narrowing options, and ranking models then estimate engagement indicators such as clicks, watch time, likes, comments, shares, or completion rate [2]. This multi-stage process forms a typical recommendation pipeline, in which user behavior and multimodal content features are gradually transformed into the final recommendation list.

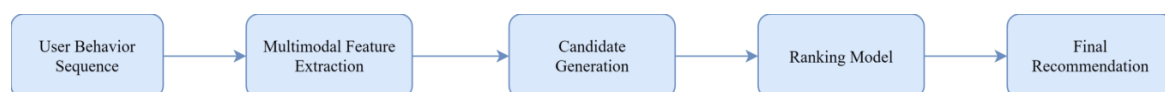


Figure 1. General pipeline of multimodal short video recommendation

Following this pipeline, user behavior sequences provide information about recent interests. Multimodal feature extraction converts frames, audio, captions, covers, and textual descriptions into machine-readable embeddings. Candidate generation retrieves potentially relevant videos, while the ranking model combines user features, item features, context features, and multimodal representations to produce the final recommendation list. DeepFM is a useful background model because it shows how industrial ranking systems combine sparse categorical features, dense embeddings, and feature interactions [3].

2.2. User behavior and sequential modeling

Short video recommendation depends heavily on sequential behavior because user interests can change quickly within a short session. For example, a user may watch several food videos in a row and then suddenly shift to sports or music videos. Traditional collaborative filtering cannot fully capture this temporal pattern. Self-attention-based models provide a solution by allowing the model to assign different weights to historical behaviors. SASRec uses self-attention to identify relevant items from a user's historical sequence and can adapt to both sparse and dense datasets [4]. BERT4Rec extends this idea by using bidirectional self-attention and predicting masked items through a Cloze objective [5]. These models' logic is highly relevant because short video platforms rely on behavior sequences, session interests, and implicit feedback.

In short video scenarios, user behavior is not limited to clicks or ratings but includes fine-grained signals such as watch time, skip position, and completion rate. FRAME models fine-grained user interests through user-clip relations, showing that video-level feedback may be too coarse for micro-video recommendation [16]. DASVD further shows that incorporating user duration preference and duration sensitivity can improve personalized short video recommendation [19].

3. Multimodal recommendation algorithms

3.1. Multimodal feature representation

Short videos contain visual, acoustic, and textual information. Visual frames and covers describe objects, scenes, human actions, and aesthetic styles. Audio provides music, speech, tone, and environmental signals. Captions and textual descriptions provide semantic topics and hashtags. Multimodal representations can help model cold-start videos and long-tail videos because content features are available even when user interactions are limited.

Contrastive Language-Image Pre-training (CLIP) is an important background model for multimodal representation because it learns aligned image-text representations through contrastive pre-training [6]. Although CLIP is not a recommendation model, its idea of aligning visual and textual semantics is useful for short video content understanding. In short video recommendation, multimodal representation is usually combined with user-item interaction signals. A survey divides multimodal recommender systems into several technical components, including modality encoding, feature interaction, feature enhancement, and model optimization [1]. This taxonomy is useful for organizing current studies.

This content-based perspective is also important for cold-start short video recommendation, where newly uploaded videos have limited user-item interactions. For example, M3CSR uses modality-specific cluster ID embeddings and original modality features to model user multimodal preferences for cold-start videos [15].

3.2. Multimodal graph-based recommendation

Graph neural networks are widely used in recommendation because user-item interactions can naturally form a bipartite graph. MMGCN is a core work in micro-video recommendation. It constructs modality-specific user-item graphs and learns representations for users and micro-videos through graph convolution, capturing user preferences for different modalities [7]. This is important because users may prefer different aspects of a video: one user may be attracted by visual style, while another may focus on audio or text.

LightGCN is not a multimodal model, but it is an important graph recommendation baseline. It argues that feature transformation and nonlinear activation are not always necessary for collaborative filtering, and that neighborhood aggregation is the essential component [8]. This simplified idea influences later multimodal graph recommendation models.

LATTICE further improves multimodal recommendation by mining latent item-item structures from multimodal features [9]. It constructs item-item graphs based on semantic similarity across modalities. This helps discover videos that are similar in visual or textual content even if they do not share many users. However, learning item-item graph structures can be computationally expensive. FREEDOM addresses this problem by freezing item-item graphs and denoising user-item interaction graphs, reducing memory cost while improving recommendation performance [14]. Table 1 summarizes and compares the selected models from these five aspects.

Table 1. Comparison of representative multimodal recommendation models

Model	Main Task	Modalities	Core Method	Main Contribution	Limitation
MMGCN	Micro-video recommendation	Visual, acoustic, textual	Modality-specific graph convolutional network	Models user preference for different modalities	Fusion is relatively static
LATTICE	Multimedia recommendation	Mainly visual and textual	Latent item-item graph learning	Mines semantic item-item structures	High graph construction cost
MMGCL	Micro-video recommendation	Visual, acoustic, textual	Graph contrastive learning	Uses modality edge dropout and modality masking	Depends on contrastive sampling design
MetaMMF	Micro-video recommendation	Visual, acoustic, textual	Meta-learning-based fusion	Learns item-specific fusion functions	More complex training process

Table 1. (continued)

MMSSL	Multimedia recommendation	Visual, textual, acoustic	Generative and contrastive self-supervised learning	Captures modality-aware dependencies	Additional SSL modules increase complexity
BM3	Multimodal recommendation	Visual and textual features	Bootstrap latent representation	Avoids negative sampling and auxiliary graphs	Mainly focuses on efficiency and representation alignment
FREEDOM	Multimodal recommendation	Image and textual features	Graph freezing and denoising	Reduces memory cost and removes noisy edges	Relies on precomputed graph quality
M3CSR	Cold-start short-video recommendation	Visual, textual, acoustic	Multimodal clustering and interest decoupling	Targets industrial cold-start scenarios	Designed mainly for cold-start setting
FRAME	Fine-grained micro-video recommendation	Frame-level visual features	User-clip graph and hybrid supervision	Models clip-level user interests	Requires fine-grained feedback and visual data

3.3. Self-supervised multimodal recommendation

Self-supervised learning is valuable for short video recommendation because labeled user feedback is sparse and noisy. MMGCL applies graph contrastive learning to micro-video recommendation by generating multiple views through modality edge dropout and modality masking [10]. This design is suitable for multimodal data because different modalities should not always be treated with equal importance.

MMSSL enhances self-supervised multimodal recommendation by combining modality-aware structure learning and cross-modal contrastive learning [12]. It attempts to preserve common semantic information and diverse user preferences across modalities. BM3 proposes a more efficient self-supervised framework by using bootstrapped latent representations without auxiliary graphs or negative sampling [13]. This matters for large-scale platforms, since negative sampling and auxiliary graph construction can increase training cost. Overall, self-supervised learning can reduce dependence on explicit labels and improve robustness under sparse interaction data.

4. Challenges and future directions

Current multimodal recommendation algorithms still face several challenges. First, modality noise is still a serious issue. Visual frames, audio, and captions may contain irrelevant or misleading information. For example, a video cover may be attractive but not representative of the actual video content. Second, semantic gaps exist among different modalities. MMGCN and MetaMMF show that users may have different preferences toward visual, acoustic, and textual modalities, and static fusion may not be sufficient [7, 11]. Third, data sparsity and cold-start problems remain challenging. Self-supervised methods such as MMSSL and BM3 reduce label dependence, while (M3CSR) improves cold-start recommendation through multimodal interest modeling [12, 13, 15]. However, practical deployment still requires efficient training and real-time inference. Fourth, graph-based models like LATTICE improve semantic item representation but introduce item-item graph learning cost [9]. FREEDOM provides a useful direction by freezing and denoising graphs [14]. Fifth, evaluation remains a critical issue because many short video recommendation models are evaluated on historical logs with exposure bias. KuaiRand provides randomly exposed videos for unbiased sequential recommendation evaluation, while KuaiRec provides fully observed user-item interactions to analyze how data density and exposure bias affect offline evaluation [17, 18]. Future research should combine unbiased datasets, counterfactual evaluation, and online A/B testing when possible. Finally, explainable recommendation is becoming increasingly important. Knowledge graph-based recommendation can provide structured semantic paths and improve interpretability [20]. Combining multimodal features with knowledge graphs may help explain why a certain short video is recommended, such as because of its topic, visual style, music type, or relation to previous user interests.

5. Conclusion

This paper reviews multimodal recommendation algorithms for short video platforms. The reviewed studies show that short video recommendation has developed from simple user-item matching into a more complex task involving user behavior sequences, multimodal video content, platform exposure mechanisms, and multiple implicit feedback signals. Overall, multimodal information is valuable because it can alleviate data sparsity, improve cold-start recommendation, enrich item representation, and capture more detailed user preferences.

However, this review still has some limitations. First, it relies on literature review and comparative analysis, so it doesn't run new experiments to verify the performance of different algorithms under the same dataset and evaluation setting. Second, the selected literature mainly focuses on representative studies in multimodal and short video recommendation, so some related works may not be fully covered. Third, some technical details, such as model architectures, loss functions, and industrial deployment issues, are only discussed at a general level.

Future research may focus on more efficient multimodal fusion, robust representation learning, unbiased evaluation, fine-grained feedback modeling, and explainable recommendation. Combining multimodal learning with knowledge graphs and large language models is also promising, as it can improve both recommendation accuracy and interpretability. Overall, multimodal recommendation offers a key technical path for short video platforms, but further work is still needed to balance accuracy, efficiency, fairness, and explainability.

References

- [1] Liu, Q., Hu, J., Xiao, Y., Zhao, X., Gao, J., Wang, W., Li, Q., & Tang, J. (2024). Multimodal recommender systems: A survey. *ACM Computing Surveys*, 57(2), Article 26.
- [2] Covington, P., Adams, J., & Sargin, E. (2016). Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 191-198).
- [3] Guo, H., Tang, R., Ye, Y., Li, Z., & He, X. (2017). DeepFM: A factorization-machine based neural network for CTR prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (pp. 1725-1731).
- [4] Kang, W.-C., & McAuley, J. (2018). Self-attentive sequential recommendation. In *Proceedings of the IEEE International Conference on Data Mining* (pp. 197-206).
- [5] Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019). BERT4Rec: Sequential recommendation with bidirectional encoder representations from Transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 1441-1450).
- [6] Radford, A., et al. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8748-8763).
- [7] Wei, Y., Wang, X., Nie, L., He, X., Hong, R., & Chua, T.-S. (2019). MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia* (pp. 1437-1445).
- [8] He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., & Wang, M. (2020). LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 639-648).
- [9] Zhang, J., Zhu, Y., Liu, Q., Wu, S., Wang, S., & Wang, L. (2021). Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia* (pp. 3872-3880).
- [10] Yi, Z., Wang, X., Ounis, I., & Macdonald, C. (2022). Multi-modal graph contrastive learning for micro-video recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1807-1811).
- [11] Liu, H., Wei, Y., Liu, F., Wang, W., Nie, L., & Chua, T.-S. (2023). Dynamic multimodal fusion via meta-learning towards micro-video recommendation. *ACM Transactions on Information Systems*, 42(2), Article 47.
- [12] Wei, W., Huang, C., Xia, L., & Zhang, C. (2023). Multi-modal self-supervised learning for recommendation. In *Proceedings of the ACM Web Conference* (pp. 790-800).
- [13] Zhou, X., Zhou, H., Liu, Y., Zeng, Z., Miao, C., Wang, P., You, Y., & Jiang, F. (2023). Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference* (pp. 845-854).
- [14] Zhou, X., & Shen, Z. (2023). A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 935-943).
- [15] Chen, G., et al. (2024). A multi-modal modeling framework for cold-start short-video recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems* (pp. 391-400).
- [16] Shang, Y., Gao, C., Chen, J., Jin, D., Wang, M., & Li, Y. (2023). Learning fine-grained user interests for micro-video recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 433-442).
- [17] Gao, C., Li, S., Zhang, Y., Chen, J., Li, B., Lei, W., Jiang, P., & He, X. (2022). KuaiRand: An unbiased sequential recommendation dataset with randomly exposed videos. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management* (pp. 3953-3957).
- [18] Gao, C., Li, S., Lei, W., Chen, J., Li, B., Jiang, P., He, X., Mao, J., & Chua, T.-S. (2022). KuaiRec: A fully-observed dataset and insights for evaluating recommender systems. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management* (pp. 540-550).
- [19] Li, J., Lu, Z., Zhang, Y., Zhu, M., Wang, J., & Sun, Z. (2026). An improved matrix factorization model for personalized short video recommendation. *Journal of Modern Information*. Advance online publication.
- [20] Han, Y., & Wei, G. (2025). Overview of task optimization and domain application development of knowledge graph and recommendation algorithm. *Computer Engineering and Applications*. Advance online publication.