GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

Journal of Computer, Signal, and System Research

Review

Research on Recommendation Algorithms in Short Video Platforms: From Traditional Methods to Deep Learning and Multimodal Fusion

Yeqiang Zheng [1,2,*]

[1] Graduate School, University of the East, Manila, Philippines
[2] Yulin Normal University, Yulin, Guangxi, 537000, China
* Correspondence: Yeqiang Zheng, Graduate School, University of the East, Manila, Philippines; Yulin Normal University, Yulin, Guangxi, 537000, China

Abstract: With the rapid advancement of mobile internet technologies, short video platforms have swiftly risen to prominence, attracting hundreds of millions of users worldwide. At the heart of these platforms, recommendation algorithms serve as essential tools for mitigating information overload and enhancing user engagement and satisfaction. This paper provides a systematic and comprehensive review of the development of recommendation algorithms tailored for short video platforms, tracing their evolution from traditional recommendation methods to deep learning techniques and, more recently, to multimodal fusion strategies. A detailed analysis is conducted on the strengths, limitations, and application scenarios of different approaches. First, the fundamental concepts and evaluation metrics related to short video recommendation are introduced. Next, the applications and shortcomings of traditional recommendation algorithms in this domain are examined thoroughly. Subsequently, deep learning–based methods, including deep neural networks and sequence modeling, are explored extensively, followed by an in-depth investigation of the most recent advances in multimodal fusion for short video recommendation. Finally, the paper discusses current challenges and outlines potential future research directions. By presenting these in-depth discussions and critical comparisons, this work aims to provide researchers and practitioners with a panoramic perspective on short video recommendation algorithms and to foster further academic progress and practical innovation in this rapidly evolving field.

Keywords: short video platform; recommendation algorithm; deep learning; multimodal fusion; sequential behavior modeling; data sparsity

Received: 08 August 2025
Revised: 14 August 2025
Accepted: 28 August 2025
Published: 09 September 2025

1. Introduction

In recent years, with the rapid advancement of mobile internet technologies and the increasing fragmentation of user attention, short video platforms have swiftly emerged as a dominant form of online content consumption, attracting hundreds of millions of users worldwide. The exponential growth of short video content is largely driven by users' preference for instant gratification and their limited attention spans [1]. Platforms such as Douyin and Kuaishou rely heavily on advanced recommendation algorithms to effectively capture user interests and preferences, thereby consolidating their role as primary channels for online information and entertainment [2].

As the core technological backbone of short video platforms, recommendation systems are indispensable for alleviating information overload and enhancing user experience. With the explosive increase in video content, users often face intensified information overload and find it difficult to identify content aligned with their preferences [1]. At the same time, short video recommendation systems encounter several unique challenges. First, the production of content vastly exceeds its consumption, forcing recommendation

systems to handle massive and continuously expanding data streams. Second, user interactions are relatively limited, often restricted to likes, comments, and shares, which makes it challenging to capture fine-grained user preferences. Third, recommendation algorithms must meet extremely high demands for real-time responsiveness, matching users with suitable content within milliseconds. Finally, the widespread application of these algorithms can exacerbate the "information cocoon" effect, exposing users to homogeneous content and thereby narrowing their perspectives and reducing content diversity [2].

Unlike traditional long-form video, short videos are inherently multimodal, integrating visual, auditory, and textual modalities. Prior studies have shown that effectively exploiting multimodal information can substantially improve the accuracy of recommendations and overall user satisfaction [1]. For instance, by combining video frames, audio features, and text descriptions, recommendation systems are able to achieve a more comprehensive understanding of both video content and user preferences, thus delivering more precise and personalized recommendations.

This review systematically explores the evolution of recommendation algorithms for short video platforms, moving from traditional methods to deep learning approaches and, more recently, multimodal fusion strategies. A comparative analysis of the strengths, limitations, and application scenarios of each type of algorithm is provided. Specifically, we begin by introducing the basic concepts and evaluation metrics of short video recommendation. We then examine the applications and limitations of traditional recommendation algorithms in this domain, followed by an in-depth exploration of deep learning–based methods, including deep neural networks and sequence modeling. Subsequently, we analyze the latest advances in multimodal fusion strategies for short video recommendation. Finally, we discuss the major challenges and outline promising directions for future research. Through these comprehensive discussions, this paper seeks to provide researchers and practitioners with a panoramic overview of recommendation algorithms for short video platforms and to stimulate further academic progress and technological innovation in this rapidly evolving field.

2. Overview of Short Video Recommendation System

As a core component of modern social media platforms, short video recommendation systems are typically organized into a layered architecture consisting of a data collection layer, a feature extraction layer, a model training layer, and a recommendation generation layer. Unlike traditional recommendation systems, which usually operate on relatively static datasets, short video platforms must process vast amounts of real-time user behavior data, including viewing duration, likes, comments, shares, and other interactive signals [3]. To handle this dynamic environment, many systems adopt incremental multi-window scanning methods, which enable continuous feature extraction and facilitate the simultaneous modeling of both short-term user interests and long-term preference patterns.

A notable characteristic of short video recommendation is the sequential nature of user interactions. In contrast to conventional methods that rely on user–item rating matrices, modern systems take users' historical interaction sequences as input and apply sequence-aware recommendation algorithms to predict the next video a user is likely to engage with during the current session [4]. By incorporating temporal dynamics, these algorithms are better able to capture the rapid evolution of user interests, thereby enhancing both the accuracy and responsiveness of recommendations.

Short video recommendations differ considerably from traditional long-form video recommendations. First, short video content typically ranges from 15 seconds to several minutes, and users consume it at a much higher frequency, necessitating extremely high real-time responsiveness from recommendation systems. Second, the evaluation metrics used for short video platforms are more diverse. In addition to conventional click-through rates (CTR), they incorporate refined engagement indicators such as completion rates, like

rates, and comment rates [3]. Furthermore, such systems prioritize immediate user feedback, which requires rapid adaptation of recommendation strategies to accommodate fast-changing user interests.

Short video recommendation systems also face a number of distinctive challenges. A key difficulty lies in the sequential nature of user behavior: viewing patterns often display both strong short-term correlations and long-term evolutionary trends, which require algorithms to capture immediate interests while also modeling persistent preferences [4]. Another critical factor is the influence of social dynamics. On short video platforms, users not only interact with content but also form social connections with other users, and these relationships exert a substantial impact on recommendation effectiveness [5]. Consequently, the integration of social recommendation mechanisms has opened up new possibilities. By leveraging network connections and user similarities, such systems can enhance recommendation quality and mitigate the data sparsity problem prevalent in traditional methods. This integration accounts not only for user–content interactions but also for mutual influences between users, thereby producing recommendations that are more personalized and socially aware.

Short video recommendation systems also confront significant challenges related to data sparsity and the cold start problem. Due to the rapid pace of content creation and the continual emergence of new videos, accurately matching novel content with suitable audiences presents a major difficulty. Similarly, new users joining the platform lack historical behavioral data, making the provision of personalized recommendations particularly challenging.

To mitigate these issues, modern short video recommendation systems employ a variety of techniques. Hierarchical modeling of user interaction behaviors enables richer information capture while simultaneously reducing computational overhead. Transfer learning allows knowledge acquired from other datasets to be adapted to the current platform, enhancing generalization performance. In addition, ranking-based ensemble methods, especially when optimizing AUC metrics, have proven effective in improving recommendation accuracy [3].

In summary, as a sophisticated, multi-layered framework, short video recommendation systems must integrate multiple factors—including user behavior sequences, social relationships, and content characteristics—to deliver accurate and personalized recommendations in environments characterized by information overload. These distinctive characteristics and technical challenges lay the groundwork for a detailed exploration of algorithmic developments in subsequent chapters, highlighting both the complexity and the potential of modern short video recommendation technologies.

3. Recommendation Algorithms in Short Video Platforms

3.1. Traditional Recommendation Algorithms and Their Application in Short Video Platforms

Traditional recommendation algorithms form the foundation of short video recommendation systems, primarily comprising collaborative filtering (CF) and matrix factorization techniques. CF, one of the most widely adopted methods, identifies user interests and preferences based on historical interaction data, thereby enabling the recommendation of similar short video content [6]. Through analysis of user interactions, including views, likes, comments, and shares of micro-videos, a user–video interaction matrix is constructed, and CF algorithms are applied to generate personalized recommendations. CF methods are typically classified into user-based and item-based approaches: the former recommends items preferred by similar users, while the latter suggests items similar to those already liked by the target user [7].

Matrix factorization effectively addresses the challenges of data sparsity and large-scale interaction data by decomposing the high-dimensional user–item interaction matrix into low-dimensional latent feature vectors, thereby capturing latent patterns of user preferences and video characteristics [8]. This dimensionality reduction enables more efficient

and accurate recommendations on short video platforms, facilitating the modeling of both user and video representations in a compact latent space.

However, traditional recommendation algorithms face several limitations when applied to short video platforms. First, data sparsity is particularly severe: the large volume of short videos combined with limited user interactions results in an extremely sparse user–video interaction matrix, leading to suboptimal performance of traditional collaborative filtering methods. Second, such algorithms struggle to capture the multimodal nature of short videos, which contain rich visual, auditory, and textual information. Dependence solely on user–video interaction data constrains recommendation quality and limits the system's ability to fully understand video content. In addition, user behavior on short video platforms exhibits prominent sequential characteristics, which traditional collaborative filtering and matrix factorization methods find challenging to model. This makes it difficult to effectively capture users' dynamic interest changes and sequential behavior patterns[6]. Finally, the cold start problem is exacerbated in short video scenarios: new users or new videos often lack sufficient historical interaction data, hindering the system's ability to generate effective personalized recommendations [7].

To adapt to the unique needs of short video scenarios, traditional recommendation algorithms have undergone a series of improvements. On one hand, researchers have attempted to integrate content information into collaborative filtering frameworks by leveraging content features, such as video titles, tags, and cover images, to alleviate data sparsity [8]. For example, the Collaborative Embedding Regression (CER) model combines individual content features with user–video interactions, effectively recommending videos in both in-matrix and out-of-matrix scenarios. On the other hand, given the sequential nature of user behavior, traditional Markov chain models have been used to capture short-term changes in user interests. However, these methods often compress all historical records into a fixed hidden representation, making it difficult to model long-term dependencies [9].

As the data scale of short video platforms continues to expand and user demand grows, traditional recommendation algorithms are gradually evolving toward deep learning approaches. Deep learning, with its powerful feature-learning and nonlinear modeling capabilities, has effectively addressed many limitations of traditional methods. Autoencoders, as an unsupervised deep learning technique, excel at dimensionality reduction, feature extraction, and data reconstruction, enabling a more comprehensive understanding of user needs and item characteristics, thereby improving recommendation quality [10]. In the context of short video recommendations, sequential recommendation methods based on deep learning outperform traditional Markov chain models by more effectively capturing user sequential behavior. For instance, combining the Transformer architecture with collaborative filtering, employing self-attention mechanisms to capture the multimodal features of micro-videos, and learning user preferences from historical interactions through multi-head attention substantially enhances the accuracy of next-video predictions.

3.2. Short Video Recommendation Algorithm Based on Deep Learning

The application of deep learning technology in short video recommendation has achieved significant progress. With its powerful nonlinear feature learning and complex pattern recognition capabilities, deep learning has effectively enhanced both recommendation system performance and overall user experience. These methods can automatically learn effective feature representations from massive amounts of user behavior and content data, overcoming the limitations of traditional recommendation algorithms in handling high-dimensional and sparse data [7].

Deep neural networks, as the core architecture of deep learning, play a crucial role in short video recommendations. By performing multi-layered nonlinear transformations, deep neural networks can capture complex interaction patterns between users and videos.

Studies have shown that deep neural network models are capable of learning feature representations directly from raw data without relying on hand-crafted features, thus better adapting to the diverse and dynamic nature of short video content. When processing historical user behavior sequences, deep neural networks effectively capture both long-term and short-term user interests and preferences, providing a foundation for more accurate recommendations.

As an unsupervised deep learning model, autoencoders are primarily employed for feature dimensionality reduction and representation learning in short video recommendation. Autoencoders learn compact representations of data through an encoder–decoder architecture, effectively addressing the data sparsity problem in short video recommendation systems. In particular, during cold-start scenarios, autoencoders can generate meaningful user and video representations from limited behavior data, enabling reasonable recommendations for new users or newly uploaded videos [7]. Furthermore, variants of autoencoders, such as denoising autoencoders and variational autoencoders, have demonstrated promising performance in enhancing the robustness of recommendation systems and generating more diverse recommendations. In the context of short video recommendation, attention mechanisms can capture dynamic changes in user interests, identify key elements within video content, and model complex interactions between users and videos [11]. Research has indicated that deep recommendation models incorporating attention mechanisms substantially improve both the accuracy and interpretability of recommendations when processing user behavior sequences alongside video content features.

The application of deep learning technology in recommendation systems not only enhances performance but, more importantly, enables automatic learning of feature representations from raw data, thereby reducing reliance on manual feature engineering. By utilizing deep neural networks, recommendation systems can simultaneously process user–video interaction data and multimodal video content, allowing for more accurate personalized recommendations [8]. Furthermore, deep learning models exhibit strong generalization capabilities, making them more effective at addressing the cold-start problem commonly encountered on short video platforms [7].

Multimodal deep learning methods have demonstrated unique advantages in short video recommendation. Studies have shown that features extracted from video content using deep learning—including visual appearance, audio, and motion information—can significantly outperform traditional hand-crafted features. In particular, deep learning audio features and action-centric features have been shown to surpass conventional MFCC and iDT features in recommendation performance [12]. Multimodal representation learning methods can fully exploit the complementary information among different modalities by converting content from each modality into vectors within an embedding space and employing appropriate fusion strategies. For example, one study proposed concatenating multimodal content vectors as input to a multilayer perceptron and introducing a key–value memory mechanism to map dense real-valued vectors, achieving a more comprehensive nonlinear semantic representation. Experimental results indicate that this multimodal representation learning approach significantly enhances recommendation system performance, achieving superior outcomes in short video understanding and recommendation tasks [11].

Notably, combining various deep learning–extracted features with traditional hand-crafted features and text metadata can further improve recommendation system performance. Research has indicated that this hybrid feature fusion strategy fully leverages the complementary information among different feature types, resulting in substantial improvements in short video recommendation tasks [12]. Furthermore, to address the unique challenges of short video recommendation—such as the timeliness of content and the rapid evolution of user interests—researchers continue to explore new deep learning models and training strategies to enhance the real-time responsiveness and adaptability of recommendation systems.

3.3. Application of Multimodal Fusion in Short Video Recommendation

As an emerging content format, the multimodal nature of short videos presents both opportunities and challenges for recommendation systems. Multimodal fusion technology effectively compensates for the limitations of single-modal information by integrating multiple sources—including video content, audio, text, and user behavior—thereby significantly improving the accuracy of short video recommendations and enhancing user experience [13].

In short video recommendations, multimodal information primarily includes three categories: visual, auditory, and textual. Visual information encompasses video frame sequences, image features, and other visual elements; auditory information includes background music, voice, and sound effects; and textual information covers titles, descriptions, tags, and comments [11]. These multimodal elements form a complex and complementary relationship—for instance, visual information conveys video content intuitively, audio provides emotional context, and text offers explicit semantic descriptions [14]. Research has shown that effectively integrating these multimodal cues can bridge the semantic gap between items and provide users with more accurate recommendations [13].

Multimodal feature extraction is a fundamental step in short video recommendation. Current research mainly employs pre-trained models to obtain feature representations for each modality, such as convolutional neural networks (CNNs) for extracting visual features, recurrent neural networks (RNNs) for processing audio sequences, and natural language processing (NLP) models for analyzing textual content [15]. However, simple concatenation of these features cannot fully capture the complex relationships among modalities. To address this, researchers have proposed various multimodal representation learning methods, such as key–value memory networks, which map dense real-valued data into vectors to generate more comprehensive semantic representations in a nonlinear manner [11]. Additionally, the multi-level multimodal feature fusion (MLMF) method projects each modality into both shared and specific feature spaces, enhances feature representations based on inter-modality similarity, and produces fused features that preserve the unique characteristics of each modality while emphasizing their similarities [16].

Multimodal fusion plays a crucial role in enhancing the effectiveness of short video recommendation. By fully leveraging the complementary information across modalities, it enables a more comprehensive understanding of video content. Studies have shown that incorporating nonlinearly guided cross-modal signals and maintaining temporal consistency can significantly improve the performance of multimodal machine learning models, achieving excellent results on large-scale datasets such as YouTube-8M [14]. Moreover, multimodal fusion facilitates the discovery of latent video categories and better aligns recommendations with user interests. By interactively learning user–item representations, hidden video categories can be identified, and user preferences can be modeled at multiple levels [13].

Multimodal fusion technology also demonstrates great potential in practical applications, such as video advertising. By learning the multimodal similarities between advertisements and video content, it enables more natural in-video ad insertion, significantly improving user experience and advertising effectiveness. Experimental results indicate that ad matching methods based on multimodal modeling not only enhance objective evaluation metrics but also receive favorable recognition from user subjective evaluations [15].

In the future, the application of multimodal fusion in short video recommendation will face both challenges and opportunities. On one hand, efficiently processing large-scale multimodal data and reducing computational complexity remain important research directions. On the other hand, exploring more effective cross-modal interaction mechanisms, improving model interpretability, and combining multimodal fusion with techniques such as reinforcement learning and graph neural networks will further advance the development of short video recommendation systems.

4. Key Issues and Solutions in Short Video Recommendation

4.1. Cold Start Problem

The cold start problem is one of the core challenges facing short video recommendation systems. It can be categorized into two scenarios: new user cold start and new content cold start. When a new user joins the platform, the system lacks historical behavioral data, making it difficult to accurately capture their interests and preferences. Similarly, when new short video content is uploaded to the platform, the system struggles to assess its quality and potential audience due to insufficient user interaction data. Content cold start is a core issue in the recommendation field. By addressing the content cold start problem, service providers can tap into the potential value of content that most users have yet to discover and provide users with more accurate, personalized services [17]. The cold start problem is particularly prominent in short video recommendation scenarios. Short video platforms generate a large amount of new content daily, along with many new users joining the platform, posing significant challenges to recommendation systems. Short videos are characterized by diverse content formats, short duration, and rapid updates. This makes traditional collaborative filtering methods perform poorly in cold start scenarios due to their heavy reliance on user-item interaction data. Therefore, researchers have proposed a variety of solutions, including content-based, popularity-based, and transfer learning-based approaches, to address this challenge.

In recent years, deep learning and multimodal fusion techniques have made significant progress in addressing the cold start problem of short video recommendation. Research has shown that deep learning features excel at handling cold-start scenarios for new items. In particular, deep learning features, such as visual appearance, audio, and motion information extracted from video content, outperform traditional hand-crafted features [12]. For example, deep learning audio features and action-centric deep learning features achieve better recommendation performance than MFCC and state-of-the-art iDT features. In short video recommendation, content-based methods can quickly analyze the visual, audio, and textual features of newly uploaded videos to provide preliminary recommendations even without user interaction data. Multimodal meta-learning (MML) methods incorporate multimodal auxiliary information, such as text and images, into the meta-learning process, designing a set of multimodal meta-learners corresponding to each modality, along with an adaptive, learnable fusion layer to integrate predictions based on different modalities [18]. For short video recommendation, multimodal fusion can simultaneously leverage multiple information sources from a video, including visual, audio, text, and user comments, to comprehensively understand the video content and provide more accurate recommendations in cold-start situations.

In summary, the cold start problem in short video recommendation is a complex and significant challenge. Deep learning and multimodal fusion technologies provide new insights and approaches for addressing this issue. By leveraging multimodal information and transfer learning, recommendation performance in cold start scenarios can be significantly improved. With the further development of deep learning and multimodal techniques, it is expected that the cold start problem in short video recommendation will be more effectively mitigated in the near future.

4.2. Data Sparsity Problem

One of the core challenges facing short video recommendation systems is data sparsity. To address this challenge, researchers have proposed various solutions. Matrix completion techniques enhance data density by predicting missing user–item interactions. For example, one study proposed integrating linked open data (LOD) into a matrix factorization model (MF-LOD), leveraging external knowledge bases to supplement missing information and significantly improve recommendation accuracy [19]. Similarly, predictive network methods based on network science extract hidden structures between users

through link prediction and apply information diffusion techniques to enhance the rating matrix, effectively alleviating sparsity issues [20].

Deep learning technology provides new solutions for mitigating data sparsity. Deeper graph neural networks predict links on bipartite user–item graphs through information propagation and introduce attention mechanisms to handle variable-size inputs for each node, effectively extracting more information about interaction behaviors. However, these approaches also face challenges. For instance, stacking multiple network layers can lead to oversmoothing, causing all nodes to converge to similar values [21]. Furthermore, the high complexity of user-group networks poses difficulties. User interests within the same group can vary significantly, especially in large groups. Indiscriminate use of high-order neighbor information can introduce negative signals during embedding propagation [22].

Overall, different strategies offer advantages in alleviating data sparsity in short video recommendation, but they also have limitations. Matrix completion methods are simple to implement but depend on external data quality; feature augmentation effectively improves representation capabilities but entails high computational complexity; knowledge transfer can provide rich information but faces domain adaptation challenges; and deep learning methods offer superior performance but encounter oversmoothing and efficiency issues. In practical applications, it is essential to select an appropriate method or combine multiple strategies based on the specific scenario and data characteristics to achieve optimal recommendations under data sparsity conditions.

4.3. Sequential Behavior Modeling

Sequential behavior modeling is a core issue in short video recommendation systems, aiming to capture user behavior patterns and evolving interests on the platform. Traditional sequence modeling methods often struggle to effectively capture users' dynamically changing interests. To address this issue, researchers have proposed various improvements. Among these, the introduction of time-aware and content-aware controllers significantly enhances the performance of RNNs in user modeling, enabling them to better utilize contextual information for controlling state transitions [23].

In recent years, Transformer-based sequence modeling methods have shown significant advantages in short video recommendation. In particular, the concept of a "Behavior Pathway" has been proposed. Recognizing that only a few key actions in a user's behavior sequence influence their future actions, the Recommender Transformer (RETR) dynamically plans each user's specific behavior path through a path attention mechanism, effectively filtering out the interference of trivial behaviors [24].

When modeling user interests, distinguishing short-term interests from long-term preferences is crucial for improving recommendation effectiveness. Research has shown that combining collaborative filtering (CF) with user–video sequence interactions can more comprehensively capture user interests [6]. Moreover, attention-based frameworks can adaptively integrate users' long-term and short-term preferences to generate user representations in specific contexts [23]. Hierarchical memory networks, through a multi-scale feature memory mechanism, not only consider preferences at the item level but also capture preferences at the user feature level, further enhancing the accuracy of sequence modeling [25].

The advantages of deep learning in sequential behavior modeling are reflected in three main aspects. First, deep learning models can automatically learn complex sequential patterns without manual feature design; second, the attention mechanism enables the model to dynamically focus on key parts of historical behavior; and third, multimodal fusion capabilities allow the model to comprehensively leverage multiple information sources, including video content, user interactions, and social relationships. These strengths have enabled deep learning-based sequence models to achieve notable success

in short video recommendation, with experiments showing that they outperform traditional methods across multiple real-world datasets [6,23-25].

4.4. Real-Time Recommendations and Dynamic Updates

Short video recommendation systems face significant real-time and dynamic challenges. To address these issues, researchers have proposed a variety of online learning and incremental update techniques.

In the context of multimodal feature fusion, multi-view active learning methods provide new insights for real-time recommendation. By learning a mapping from visual views to textual views, this approach reduces reliance on manually annotated metadata. An active selection strategy based on prediction inconsistency and viewing frequency effectively identifies the most important videos for metadata queries, substantially lowering annotation costs [26]. This method is particularly well-suited for handling the large volume of newly uploaded content that lacks complete metadata on short video platforms, thereby enhancing the system's real-time adaptability.

4.5. Evaluation and Dataset of Short Video Recommendation System

The evaluation of short video recommendation systems is a crucial step in measuring the performance of recommendation algorithms. It not only determines the practical value of the algorithm but also provides researchers with guidance for improvement. The evaluation process requires comprehensive consideration of multiple dimensions, including accuracy, diversity, and novelty, as well as the representativeness and reliability of the evaluation dataset [27]. Evaluating short video recommendation systems presents unique challenges. First, short video content is multimodal, encompassing visual, audio, and textual information, which complicates the evaluation process [28]. Second, user preferences for short videos may be influenced by various factors, such as content timeliness and social influence, which are difficult to fully capture using traditional evaluation metrics. Furthermore, short video platforms typically rely on implicit feedback (such as viewing time, likes, and comments) rather than explicit ratings, which adds additional complexity to the evaluation [29].

To address these challenges, researchers have proposed the Framework for Evaluation of Recommender Systems (FEVR), which classifies the recommendation system evaluation space and emphasizes that a comprehensive evaluation of a recommendation system typically requires considering multiple aspects and perspectives across different dimensions. This multifaceted evaluation approach provides a structured and systematic foundation for the evaluation of short video recommendation systems and facilitates the adoption of appropriate evaluation configurations that comprehensively encompass the required multifacetedness. In terms of data processing, researchers have proposed a method for determining user preferences using the concept of resolution sets, which allows user preferences to be inferred even from very limited ratings. In addition, they have studied methods for providing recommendations when user ratings are imprecise, inconsistent, or incomplete, and for handling situations where users may not consider the values of certain item attributes [30]. These methods are of great reference value for effectively addressing the data sparsity, incompleteness, and uncertainty problems commonly encountered in short video recommendation systems.

In summary, evaluating short video recommendation systems is a complex and multifaceted process that requires thorough consideration of multiple evaluation metrics and dataset characteristics. Future research should focus on developing more comprehensive and reliable evaluation methods, along with datasets that more accurately represent real-world user behavior, in order to promote the effective development and practical application of short video recommendation systems.

5. Future Research Directions and Development Trends

5.1. Future Research Directions

The issue of fairness has become a major concern in recommender system research. As one of the most widespread applications of machine learning, recommender systems increasingly have a critical impact on society, as their use by a growing number of users shapes information acquisition and decision-making [31]. Therefore, it is essential to address potential unfairness, which may undermine user satisfaction, content provider interests, and overall platform effectiveness. Current research has proposed various fairness definitions and evaluation frameworks. However, balancing the fairness needs of different stakeholders in short video recommendation while maintaining recommendation accuracy remains an open challenge. In particular, in multimodal content environments, ensuring fair exposure opportunities for diverse content creators and preventing algorithmically induced information cocoons require further investigation.

The reproducibility and practical progress of recommender system research face significant challenges. In recent years, deep learning–based methods (neural networks) have dominated the literature on recommender systems, with many studies claiming to surpass the state of the art. However, a concerning analysis revealed that 11 of 12 reproducible neural recommendation methods published at prominent scientific conferences between 2015 and 2018 could be outperformed by conceptually simpler approaches, such as nearest neighbor heuristics or linear models [32]. This suggests that despite the proliferation of publications, common issues in current research practices may hinder progress. These challenges are particularly pronounced in short video recommendation systems, where complex deep learning models and multimodal fusion methods are often difficult to replicate and compare.

Scalability and data sparsity represent additional key challenges for short video recommendation systems. Recommender systems are widely applied across domains such as movies, news, and music, aiming to provide users with the most relevant recommendations from a diverse set of items. Although memory-based nearest neighbor methods are typical collaborative filtering approaches and are popular due to their high recommendation accuracy, their performance suffers in commercial applications with large user and item bases and limited ratings, a situation known as data sparsity [33]. Short video platforms feature large user bases and rapidly updating content. Consequently, designing recommendation algorithms capable of effectively handling high-dimensional sparse data while ensuring real-time responsiveness remains an urgent challenge.

Future research needs to achieve breakthroughs in several directions:

1) Explainable Recommendations: With the growing application of deep learning and multimodal fusion technologies in short video recommendations, model complexity has increased, making the results difficult to interpret. Developing recommendation algorithms that provide clear explanations not only enhances user trust and acceptance but also helps content creators understand recommendation mechanisms and optimize content creation. In particular, in multimodal fusion scenarios, explaining the contribution of different modal features to recommendation results is a research direction worth pursuing.
2) Privacy Protection: Short video platforms collect large amounts of user behavior data and multimedia content. Protecting user privacy while delivering personalized recommendations remains a major challenge. Future research should explore the application of privacy-preserving technologies, such as federated learning and differential privacy, in short video recommendation systems, as well as methods to achieve effective multimodal feature extraction and fusion without compromising user privacy.
3) Cross-Platform Recommendations: Users frequently engage across multiple short video and social media platforms. Integrating data and user behavior pat-

terns across these platforms to provide personalized cross-platform recommendations represents a promising research direction. This approach involves key technical issues, including data-sharing mechanisms, cross-domain user modeling, and multi-platform collaborative optimization.

5.2. Future Development Trends of Deep Learning and Multimodal Fusion Technology

First, the use of self-supervised learning in short video recommendations is expected to become more widespread. By leveraging large volumes of unlabeled data to learn effective representations, it is possible to alleviate data sparsity and improve recommendation performance. In multimodal scenarios, designing effective self-supervised tasks that capture correlations across different modalities represents a key research direction.

Second, the application of graph neural networks in recommendation systems will be further expanded. User–content interactions on short video platforms can naturally be represented as graph structures, and graph neural networks can effectively capture high-order connections and complex patterns. Future research will focus on incorporating multimodal information into graph neural networks and designing algorithms capable of handling large-scale, dynamic graphs.

Third, multi-task learning frameworks are expected to become more mature. Short video recommendation systems often need to optimize multiple objectives simultaneously, such as click-through rate, viewing time, and user retention. Designing effective multi-task learning frameworks that balance these objectives while fully leveraging multimodal information remains a critical research focus.

Finally, real-time recommendation technology will receive increasing attention. Short video content is highly time-sensitive and updates rapidly. Developing recommendation algorithms that can respond to changes in user interests and content updates in real time, while ensuring low latency and high throughput, remains a significant challenge.

In summary, short video recommendation systems face multiple challenges, including fairness, reproducibility, scalability, and data sparsity. Future research needs to achieve breakthroughs in explainable recommendations, privacy protection, and cross-platform recommendations, while also promoting innovative developments in deep learning and multimodal fusion technologies. Such research has both theoretical significance and broad practical value, ultimately enabling short video platforms to provide more intelligent, fair, and efficient recommendation services.

6. Conclusion

This paper reviews the current research on short video recommendation systems, focusing on deep learning, multimodal fusion, cold start, data sparsity, sequential behavior modeling, real-time adaptability, and evaluation methods. While significant progress has been made, challenges remain in fairness, reproducibility, scalability, and privacy protection. Future research directions include explainable recommendations, cross-platform personalization, self-supervised learning, graph neural networks, multi-task learning, and real-time recommendation technologies. Advancements in these areas will not only improve recommendation accuracy and user satisfaction but also provide theoretical guidance and practical value for the development of more intelligent, fair, and efficient short video recommendation systems.

References

1. D. Cao, L. Miao, H. Rong, Z. Qin, and L. Nie, "Hashtag our stories: Hashtag recommendation for micro-videos via harnessing multiple modalities," *Knowledge-Based Systems*, vol. 203, p. 106114, 2020, doi: 10.1016/j.knosys.2020.106114.
2. N. Li, et al., "An exploratory study of information cocoon on short-form video platform," in *Proc. 31st ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2022, pp. 4178–4182, doi: 10.1145/3511808.3557548.

3. Y. Liu, C. Lyu, Z. Liu, and D. Tao, "Building effective short video recommendation," in *2019 IEEE Int. Conf. Multimedia & Expo Workshops (ICMEW)*, Jul. 2019, pp. 651–656, doi: 10.1109/ICMEW.2019.00126.
4. M. Quadrana, D. Jannach, and P. Cremonesi, "Tutorial: Sequence-aware recommender systems," in *Companion Proc. 2019 World Wide Web Conf.*, May 2019, pp. 1316–1316, doi: 10.1145/3308560.3320091.
5. J. Shokeen and C. Rana, "A study on features of social recommender systems," *Artificial Intelligence Review*, vol. 53, no. 2, pp. 965–988, 2020, doi: 10.1007/s10462-019-09684-w.
6. S. Liu and Z. Chen, "Sequential behavior modeling for next micro-video recommendation with collaborative transformer," in *2019 IEEE Int. Conf. Multimedia and Expo (ICME)*, Jul. 2019, pp. 460–465, doi: 10.1109/ICME.2019.00086.
7. Z. Y. Khan, Z. Niu, S. Sandiwarno, and R. Prince, "Deep learning techniques for rating prediction: a survey of the state-of-the-art," *Artificial Intelligence Review*, vol. 54, no. 1, pp. 95–135, 2021, doi: 10.1007/s10462-020-09892-9.
8. X. Du, H. Yin, L. Chen, Y. Wang, Y. Yang, and X. Zhou, "Personalized video recommendation using rich contents from videos," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 3, pp. 492–505, 2018, doi: 10.1109/TKDE.2018.2885520.
9. H. Fang, D. Zhang, Y. Shu, and G. Guo, "Deep learning for sequential recommendation: Algorithms, influential factors, and evaluations," *ACM Trans. Inf. Syst. (TOIS)*, vol. 39, no. 1, pp. 1–42, 2020, doi: 10.1145/3426723.
10. G. Zhang, Y. Liu, and X. Jin, "A survey of autoencoder-based recommender systems," *Frontiers Comput. Sci.*, vol. 14, no. 2, pp. 430–450, 2020, doi: 10.1007/s11704-018-8052-6.
11. D. Guo, J. Hong, B. Luo, Q. Yan, and Z. Niu, "Multi-modal representation learning for short video understanding and recommendation," in *2019 IEEE Int. Conf. Multimedia & Expo Workshops (ICMEW)*, Jul. 2019, pp. 687–690, doi: 10.1109/ICMEW.2019.00134.
12. A. Almeida, J. P. de Villiers, A. De Freitas, and M. Velayudan, "The complementarity of a diverse range of deep learning features extracted from video content for video recommendation," *Expert Syst. Appl.*, vol. 192, p. 116335, 2022, doi: 10.1016/j.eswa.2021.116335.
13. J. Ma, J. Wen, M. Zhong, W. Chen, X. Zhou, and J. Indulska, "Multi-source multi-net micro-video recommendation with hidden item category discovery," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, Apr. 2019, pp. 384–400. Cham: Springer, doi: 10.1007/978-3-030-18579-4_23.
14. S. Ghosh, "Multimodal machine learning for video and image analysis," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery & Data Mining*, Aug. 2020, p. 3608, doi: 10.1145/3394486.3409558.
15. X. Song, B. Xu, and Y.-G. Jiang, "Predicting content similarity via multimodal modeling for video-in-video advertising," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 569–581, 2020, doi: 10.1109/TCSVT.2020.2979928.
16. X. Hu, Y. Ji, and G. A. Kumie, "Multi-level multi-modal feature fusion for action recognition in videos," in *Proc. 3rd Int. Workshop Human-Centric Multimedia Anal.*, Oct. 2022, pp. 25–33, doi: 10.1145/3552458.3556449.
17. P. Wang, Y. Jiang, C. Xu, and X. Xie, "Overview of content-based click-through rate prediction challenge for video recommendation," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2593–2596, doi: 10.1145/3343031.3356085.
18. X. Pan, Y. Chen, C. Tian, Z. Lin, J. Wang, H. Hu, and W. X. Zhao, "Multimodal meta-learning for cold-start sequential recommendation," in *Proc. 31st ACM Int. Conf. Inf. & Knowl. Manage.*, Oct. 2022, pp. 3421–3430, doi: 10.1145/3511808.3557101.
19. S. Natarajan, S. Vairavasundaram, S. Natarajan, and A. H. Gandomi, "Resolving data sparsity and cold start problem in collaborative filtering recommender system using linked open data," *Expert Syst. Appl.*, vol. 149, p. 113248, 2020, doi: 10.1016/j.eswa.2020.113248.
20. H. Zare, M. A. N. Pour, and P. Moradi, "Enhanced recommender system using predictive network approach," *Physica A: Stat. Mech. Appl.*, vol. 520, pp. 322–337, 2019, doi: 10.1016/j.physa.2019.01.053.
21. R. Yin, K. Li, G. Zhang, and J. Lu, "A deeper graph neural network for recommender systems," *Knowl.-Based Syst.*, vol. 185, p. 105020, 2019, doi: 10.1016/j.knosys.2019.105020.
22. Y. Chen, J. Wang, Z. Wu, and Y. Lin, "Integrating user-group relationships under interest similarity constraints for social recommendation," *Knowl.-Based Syst.*, vol. 249, p. 108921, 2022, doi: 10.1016/j.knosys.2022.108921.
23. Z. Yu, J. Lian, A. Mahmoody, G. Liu, and X. Xie, "Adaptive user modeling with long and short-term preferences for personalized recommendation," in Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI), Aug. 2019, vol. 7, pp. 4213–4219, doi: 10.24963/ijcai.2019/585.
24. Z. Yao, X. Chen, S. Wang, Q. Dai, Y. Li, T. Zhu, and M. Long, "Recommender transformers with behavior pathways," in *Proc. ACM Web Conf. (WWW)*, May 2024, pp. 3643–3654, doi: 10.1145/3589334.3645528.
25. B. Song, Y. Cao, W. Zhang, and C. Xu, "Session-based recommendation with hierarchical memory networks," in *Proc. 28th ACM Int. Conf. Inf. & Knowl. Manage. (CIKM)*, Nov. 2019, pp. 2181–2184, doi: 10.1145/3357384.3358120.
26. J. J. Cai, J. Tang, Q. G. Chen, Y. Hu, X. Wang, and S. J. Huang, "Multi-view active learning for video recommendation," in Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI), Aug. 2019, pp. 2053–2059, doi: 10.24963/ijcai.2019/284.
27. E. Zangerle and C. Bauer, "Evaluating recommender systems: Survey and framework," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 1–38, 2022, doi: 10.1145/3556536.
28. K. Abdalla, I. Menezes, and L. Oliveira, "Modelling perceptions on the evaluation of video summarization," *Expert Syst. Appl.*, vol. 131, pp. 254–265, 2019, doi: 10.1016/j.eswa.2019.04.065.

29. J. Y. Chin, Y. Chen, and G. Cong, "The datasets dilemma: How much do we really know about recommendation datasets?," in *Proc. 15th ACM Int. Conf. Web Search Data Mining (WSDM)*, Feb. 2022, pp. 141–149, doi: 10.1145/3488560.3498519.
30. A. Hertz, T. Kuflik, and N. Tuval, "Resolving sets and integer programs for recommender systems," *J. Global Optim.*, vol. 81, no. 1, pp. 153–178, 2021, doi: 10.1007/s10898-020-00982-0.
31. Y. Li, Y. Ge, and Y. Zhang, "Tutorial on fairness of machine learning in recommender systems," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval (SIGIR)*, Jul. 2021, pp. 2654–2657, doi: 10.1145/3404835.3462814.
32. M. Ferrari Dacrema, S. Boglio, P. Cremonesi, and D. Jannach, "A troubling analysis of reproducibility and progress in recommender systems research," *ACM Trans. Inf. Syst. (TOIS)*, vol. 39, no. 2, pp. 1–49, 2021, doi: 10.1145/3434185.
33. M. Singh, "Scalability and sparsity issues in recommender datasets: A survey," *Knowl. Inf. Syst.*, vol. 62, no. 1, pp. 1–43, 2020, doi: 10.1007/s10115-018-1254-2.