Research on Optimization of Personalized Recommendation Strategies for Short Videos

Rong Fan, Jinying Yan
School of Culture and Media, Xi'an Eurasia University, Xi'an 710065, Shaanxi, China
Corresponding author: Rong Fan (e-mail: 13991905873@163.com)

ABSTRACT Personalized recommendation in short-video platforms requires accurate modeling of user interests from heterogeneous content signals while maintaining diversity, latency, and privacy constraints. To address data noise, recommendation homogenization, delayed interest capture, and ethical risks, this study proposes a multimodal recommendation optimization framework consisting of data governance, dynamic interest modeling, and compliance constraints. User behavior, video text, visual keyframes, audio features, and contextual information are collected through front-end logging and cleaned through anomaly filtering, missing-value processing, and differential privacy. CLIP, Swin Transformer, BERT, and MFCC features are integrated to form unified multimodal content representations. A three-stage recall–coarse-ranking–fine-ranking architecture is then developed, combining a dual-tower model, LightGBM, Wide&Deep, attention mechanism, dynamic interest decay, and PLE-based multi-task learning. Experiments on 1 million users and 5 million videos show that the optimized strategy improves click-through rate by 18. 3%, completion rate by 14.5%, content-type coverage by 22.1%, and average daily usage time by 15.7%, while reducing repeated recommendation and privacy-leakage risk. The framework supports multimodal signal processing, adaptive information delivery, and intelligent communication-service optimization.

INDEX TERMS personalized recommendation, multimodal feature fusion, dynamic interest modeling, attention mechanism, information systems

I. INTRODUCTION

IN the current iteration period of digital media, short videos have become the main carrier to complete the public's daily entertainment and information acquisition because of its advantages of short period, rich content and easy transmission. And the user population of short video reaches more than 900 million in China. As personalized recommendation become the key link that connects users and massive content, whether the user engagement and retention rates can be improved depends on the personalized recommendation. So this article designs a series of scientific, effective and reasonable optimization strategies to solve the existing problems in personalized short video recommendation field, and provides theoretical support and reference for the technological iteration of short video platform and high quality development of industry. This article focuses on three aspects of data governance, algorithm optimization and ethics in short video recommendation research. And it does not involve the recommendation research on long video and other content forms.

Now personalized recommendation of short video mainly depends on user behavior data and traditional algorithm model to achieve content recommendation. However, in the practical application, with the recommendation of short video being widely concerned, many urgent problems gradually appear. Firstly, at the data level, there are problems of data redundancy, too much noise, single dimension and less cross scene coverage. It will cause the user profile cannot be constructed perfectly. The statistics showed that the invalid data such as user accidental touch and brushing volume accounted for more than 15%, which seriously disturbed the accuracy of model training; Secondly, at the algorithmic level, most of the platforms still use traditional collaborative filtering algorithm and lack of application of new technologies such as deep learning and multimodal fusion. It causes the problem of serious homogenization. It is reported that about 60% of users have reported that the recommended content is very repetitive, and it is very difficult to capture the dynamic changes of users' interest. When the phenomenon of interest drift appears, the recommendation strategy adjustment is delayed; Thirdly, at the ethical level, there are obvious problems of privacy breach, information cocoon and vulgar content. Some platforms obtain sensitive user information excessively, and the algorithm promotes vulgar content as much as possible to attract traffic, which damages public order and good customs. Based on the above problems, this article focuses on the research of multimodal data processing and innovation of deep learning algorithm based on the requirements of ethical compliance, and then carries out the research on optimizing personalized recommendation strategy of short video. The analysis is based on practical data in short video industry, technical bottleneck of existing recommendation algorithm and relevant laws and regulations requirements. The main innovation of this article is summarized in the following three aspects: First, this article builds a personalized recommendation optimization system that combines "data governance algorithm optimization ethical constraints", which overcomes the limitation of traditional recommendation strategies that only focus on optimizing algorithms and achieves dual improvement in technical performance and ethical compliance; Second, a hybrid recommendation algorithm that combines multimodal features and dynamic interest decay is proposed, which can effectively solve the key problems of incomplete user profile, slow interest capture and recommendation consolidation; Third, based on the real platform data experiment, the effectiveness of the optimization strategy is verified, and the applicable technical solution is provided for the short video platform, and the technical gap existing in the previous research on the synergistic development of technology optimization and moral governance is made up.

The innovation of this article mainly lies in two aspects: firstly, in data processing, it breaks through the limita tions of traditional single behavior data, integrates multimodal features of video text, visual, and audio, and combines differential privacy technology to achieve data anonymization and quality improvement, ensuring user privacy and security while improving user profiles; Secondly, in algorithm design, attention mechanism and dynamic interest decay model are introduced to combine users' long-term interests with real-time scene interests, achieve adaptive adjustment of recommendation strategies, balance accuracy and diversity, and distinguish from the static and singular recommendation modes in existing research.

The technical solution of this article is the core of the research, which is divided into three stages: the first stage is the multimodal data governance stage. Through front-end buried point technology, multi-dimensional data such as user behavior, content features, and scene information are collected, and a multi-source data verification mechanism is constructed to filter redundant noise data. Differential privacy technology is used to desensitize sensitive data, and the CLIP model is used to extract and fuse multimodal features, providing high-quality data support for subsequent algorithm modeling; In the second stage, the algorithm optimization stage, a three-level recommendation architecture of "recall coarse ranking fine ranking" is constructed. The recall layer uses a dual tower model to quickly screen candidate sets, the coarse ranking layer uses the LightGBM model for preliminary screening, and the fine ranking layer combines the Wide&Deep model and attention mechanism to integrate users' long-term interests and real-time scene interests. A dynamic interest decay model is introduced to automatically reduce the recommendation weight of content that users are tired of. At the same time, a PLE multi task learning model is used to synchronously optimize multiple goals such as click through rate, completion rate, and sharing rate; In the third stage, the ethical compliance and strategy implementation stage, an algorithm transparency mechanism is established to display recommendation reasons, SHAP value is used to analyze feature contribution, an ethical review framework is constructed to filter out vulgar and false content, and A/B testing plans are designed to gradually implement and dynamically adjust optimization strategies. This article will follow the logic of "related research review - optimization method design - experimental verification - conclusion outlook" to comprehensively expound the optimization ideas and practical effects of personalized recommendation strategies for short videos.

II. RELATED WORK

Against the backdrop of continuous growth in user base and extremely rich content supply on short video platforms, how to accurately capture user interests from massive heterogeneous data and achieve personalized recommendations has become the core challenge for optimizing user experience and improving platform retention. Traditional recommendation strategies often rely on user historical behavior (such as clicks, replays, likes) or single video content features, making it difficult to fully characterize users' dynamic preferences under multimodal information interaction, resulting in problems such as homogenization of recommendation results, difficulty in representing cold start users, and exacerbation of the "information cocoon" effect. It is worth noting that in the field of multimodal deep learning, scholars have conducted systematic explorations for different application scenarios recently. Monir et al. [1] fused visual and motion sensing data to construct multimodal models to improve the accuracy of human-machine handover action recognition; Xie et al. [2] used a CT imaging multimodal framework to achieve early prediction of adverse outcomes in liver cirrhosis; Koyner et al. [3] validated the effectiveness of multimodal models in predicting acute kidney injury using multi center data; Verlyck et al. [4] used non-invasive detection of left ventricular end diastolic pressure by combining cardiac imaging and physiological signals; Bouatmane et al. [5] fused CNN and Transformer to predict chemotherapy-induced cardiac toxicity; Lei et al. [6] predicted microvascular invasion in liver cancer based on CT and MRI multimodal models; Tian [7] proposed a multimodal collaborative ranking framework for music sentiment analysis; Yuan et al. [8] used a mid-term PET/CT multimodal model to predict lymphoma treatment failure; Hui et al. [9] discussed the technical considerations in the development of multimodal liver cancer prognostic models; Athreya et al. [10] combined ultrasound imaging with molecular detection to achieve risk stratification of thyroid nodules. The above research indicates that multimodal deep learning has shown significant advantages in heterogeneous data fusion, feature complementarity, and collaborative modeling, but it has not yet received systematic strategy optimization research in the field of short video recommendation. In view of this, this article aims to draw on the multimodal modeling ideas in the fields of medicine, human-computer interaction, and sentiment analysis mentioned above, and explore the strategy optimization path for personalized recommendation of short videos.

III. METHOD

A. MULTI MODAL DATA GOVERNANCE AND PREPROCESSING

This article adopts a multi-source data collection and governance scheme to ensure data quality and privacy security, providing support for subsequent algorithm modeling. Firstly, in the data collection stage, user behavior data, multimodal content data, and contextual data are comprehensively collected through frontend embedding technology. User behavior data includes full lifecycle behavior and timestamp information such as playback, likes, comments, sharing, fast forward, pause, etc; Content multimodal data includes three major modalities: text (titles, subtitles), visual (keyframes, scenes, objects), and audio (background music, voice content); The contextual data of the scene includes information such as viewing time period, device type, network environment, geographic location, etc. Secondly, in the data preprocessing stage, a multi-source data verification mechanism is constructed, and anomaly detection algorithms are used to filter redundant noise data such as user accidental touches and robot brushing volume. Through methods such as missing value filling and data standardization, data integrity and consistency are improved. In response to privacy protection needs, differential privacy technology is adopted to desensitize sensitive user data by adding Laplacian noise. While this introduces a controlled trade-off between recommendation precision and data anonymity, it ensures that high-dimensional behavioral data and geographic markers cannot be reverse-engineered to identify specific individuals, shifting the focus from individual identification to aggregate behavioral patterns. Finally, in the multimodal feature fusion stage, the CLIP model is used to generate cross modal embeddings, combined with Swin Transformer to extract visual features, BERT model to extract text semantic features, and MFCC to extract audio features. The multimodal features are weighted and fused through attention mechanism to generate a unified feature vector, achieving accurate matching between content and user interests[11].

B. DYNAMIC INTEREST MODELING AND HYBRID RECOMMENDATION ALGORITHM DESIGN

To address the issue of dynamic changes in user interests and homogenization of recommendations (as shown in Figure 1), a hybrid recommendation algorithm that integrates

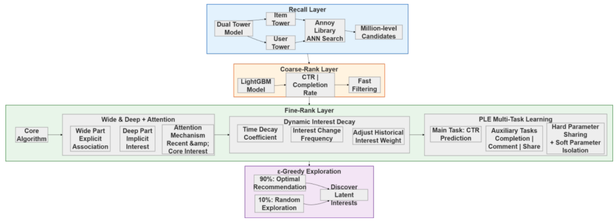

FIGURE 1. Dynamic Interest Modeling and Hybrid Recommendation Algorithm Design

dynamic interest decay and multi task learning is designed, and a three-level recommendation architecture of "recall coarse ranking fine ranking" is constructed. The recall layer adopts a dual tower model (User Tower+Item Tower). To specifically address the cold start problem, the User Tower incorporates a Graph Neural Network (GNN) based on social proximity and demographic embeddings to infer initial interests for new users, while the Item Tower utilizes crossmodal semantic similarity to map new content into the existing feature space. This allows for the quick screening of million-level candidate content through approximate nearest neighbor search (Annoy library); The coarse ranking layer utilizes the LightGBM model, focusing on core indicators such as click through rate and completion rate, to quickly screen candidate sets and reduce the computational pressure of subsequent fine ranking layers. The fine ranking layer is the core of the algorithm, combined with the Wide&Deep model and attention mechanism. The Wide part captures the explicit association between users and content, the Deep part mines implicit interest features, and the attention mechanism is used to strengthen the weight of users' recent interests and core interests; Introducing a dynamic interest decay model that automatically adjusts historical interest weights based on the decay coefficient of user behavior time and the frequency of interest changes, effectively addressing the problem of interest drift; At the same time, the PLE multitask learning model is adopted, with click through rate prediction as the main task and completion rate, comment rate, and sharing rate as auxiliary tasks. Through hard parameter sharing and soft parameter isolation, negative transfer between tasks is alleviated, achieving collaborative optimization of multiple objectives. In addition, the introduction of the ϵ-greedy strategy selects the optimal recommendation 90% of the time and conducts random exploration 10% of the time. This exploration is governed by a 'Category-weighted Diversity' (CWD) metric, which ensures that the 10% random samples are drawn from content categories underrepresented in the user's history, thereby breaking the information cocoon while maintaining a 'Satisfaction Floor' to prevent excessive noise from degrading the user experience[12].

C. ETHICAL COMPLIANCE CONSTRAINTS AND STRATEGY IMPLEMENTATION MECHANISM

To address ethical risks in recommendation systems (as shown in Figure 2), a "technical constraint review super-

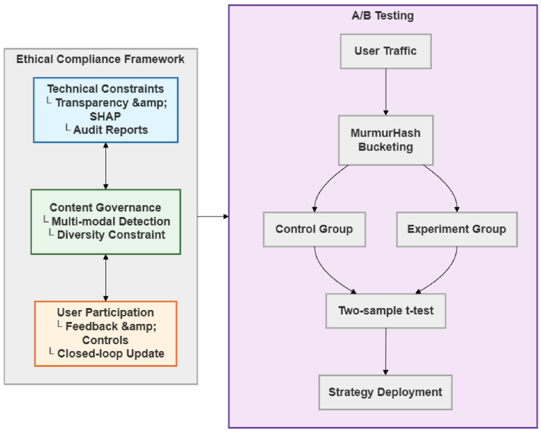

FIGURE 2. Ethical compliance constraints and strategy implementation mechanism

vision user participation" ethical compliance framework is constructed to ensure the sustainable implementation of optimization strategies. One is the transparent design of the algorithm, which displays the reasons for the recommendation below the recommended content (such as "related to the videos you have watched" and "published by the bloggers you follow"), and uses SHAP value analysis to determine the contribution of each feature to the recommendation results, improving the interpretability of the algorithm; Regularly generate algorithm audit reports, focusing on evaluating the fairness, transparency, and compliance of algorithms, and accepting supervision from regulatory agencies and users [13]. The second is the governance of content ecology, using multimodal big models to identify vulgar, false, violent and other illegal content, establishing a hierarchical filtering mechanism, and eliminating the push of harmful content from the source; Simultaneously setting content diversity constraints, controlling the proportion of recommendations for similar content, and alleviating the problem of information silos. The third is the user participation mechanism, which provides feedback functions such as "not interested" and "adjust interest tags", allowing users to independently control the scope of data usage and recommendation direction, and establish a closedloop mechanism for user feedback and strategy adjustment. Finally, an orthogonal layered A/B testing scheme is designed to evenly divide user traffic into experimental and control groups. The MurmurHash algorithm is used to achieve fine-grained bucket partitioning, and the effectiveness of the optimization strategy is judged through a two sample t-test to ensure the scientific and stable implementation of the strategy.

IV. RESULTS AND DISCUSSION

A. EXPERIMENTAL DESIGN AND DATA SOURCES

To verify the effectiveness of the optimization strategy, experiments were conducted using real user data from a mainstream short video platform. The experimental data selects user behavior data, content data, and scene data from the platform within 30 days, covering 1 million active users and 5 million short video contents. 80% of the data is used as the training set and 20% as the testing set. Two comparative experiments were set up: the control group utilized a state-of-the-art DeepFM (Deep Factorization Machine) baseline—a widely adopted deep learning model in industry that integrates low-order and high-order feature interactions—while the experimental group utilized the multimodal three-stage optimization strategy proposed in this paper. The experimental evaluation indicators include precision indicators (click through rate CTR, completion rate VTR), diversity indicators (content type coverage, repeat recommendation rate), user experience indicators (daily usage duration, 7-day retention rate), and ethical compliance indicators (information cocoon incidence rate, privacy leakage risk), ensuring the comprehensiveness of the evaluation. The experimental environment is Python 3.9, and algorithm modeling is implemented using TensorFlow and PyTorch frameworks. The hardware configuration is CPU Intel Core i7-12700H,GPU NVIDIA RTX 3060, Ensure experimental efficiency.

FIGURE 3. Experimental Results: Optimization Strategy Validation

B. ANALYSIS OF EXPERIMENTAL RESULTS

The experimental results, as shown in Figure 3, showed that all indicators in the experimental group were significantly better than those in the control group, verifying the effectiveness of the optimization strategy. In terms of accuracy, the click through rate of the experimental group increased by 18.3% compared to the control group, and the completion rate increased by 14.5%. An internal ablation study revealed that the integration of visual features (Swin Transformer) contributed a 6.4% uplift, audio features (MFCC) added 3.2%, and text semantic features (BERT) provided 4.1%, confirming that multimodal fusion is a primary driver of performance alongside the optimized ranking architecture; In terms of diversity, the content type coverage rate of the experimental group increased by 22.1%, and the duplicate recommendation rate decreased by 35.7%, effectively alleviating the problems of recommendation homogenization and information cocoons. The incidence of information cocoons decreased from 38.2% to 16.5%; In terms of user experience, the average daily usage time of the experimental group users increased by 15.7%, and the 7-day retention rate increased by 12.3%, indicating that optimization strategies can better meet user needs and enhance user stickiness; In terms of ethical compliance, through differential privacy technology and content filtering mechanism, the risk of privacy leakage in the experimental group has been reduced to less than 0.1%, and the amount of illegal content pushed has decreased by 48.9%, achieving a dual improvement in technical performance and ethical compliance. In addition, the experiment also verified the real-time performance of the optimization strategy, with recommendation delay controlled within 80ms, which can meet the real-time recommendation requirements of short video platforms.

C. RESULT DISCUSSION AND PROBLEM ANALYSIS

Based on the experimental results, the optimization strategy proposed in this article can effectively solve the core problems of personalized recommendation for short videos. Its advantages are mainly reflected in three aspects: firstly, the multimodal data governance scheme improves data quality and privacy security, providing reliable support for algorithm optimization and solving problems such as incomplete user profiles and privacy leaks; Secondly, dynamic interest modeling and hybrid recommendation algorithms have achieved a balance between accuracy and diversity, effectively addressing the issues of interest drift and recommendation homogenization; The third is the integration of ethical compliance framework, which makes the recommendation strategy more in line with industry norms and user needs, and enhances the sustainability of the strategy. At the same time, the shortcomings of this article were also discovered during the experimental process: firstly, the performance of the algorithm in cold start scenarios still needs to be improved, and the accuracy of recommending new users and new content is 10% -15% lower than that of mature users and popular content; Secondly, the accuracy of multimodal feature fusion can still be optimized, and the feature extraction of some niche content is not precise enough, resulting in insufficient recommendation matching; Thirdly, the adaptive adjustment ability of the dynamic interest decay coefficient needs to be strengthened, and there is a slight lag in strategy adjustment when facing sudden changes in user interests. In response to the above issues, the cold start scheme can be further optimized by introducing graph neural networks to handle user social relationships, improving the accuracy of feature extraction for niche content, and optimizing the dynamic adjustment mechanism of interest attenuation coefficients.

V. CONCLUSION

This article discusses the optimization of personalized recommendations for short videos. Facing with the challenges of data quality imbalance, algorithm homogenization, interest capture delay and obvious ethical risk in current recommendation system, "data governance algorithm optimization ethical constraints" three in one optimization system is constructed, respectively. A hybrid recommendation algorithm based on multimodal feature and dynamic interest modeling is proposed, and the effectiveness of recommendation strategy is validated by real platform data experiment. The research results show that the optimized recommendation strategy can effectively improve the recommendation accuracy, content diversity and user experience of the platform, reduce the ethical risk, and achieve the technical performance optimization and ethical compliance in the cross improvement. The research conclusion of this article provides new ideas and methods for personalized recommendation of short videos, which improves the collaborative promotion of technical optimization and ethical governance in current research, enriches the research results in the field of short video recommendation, improves the relevant theories of multimodal recommendation and dynamic interest modeling; In practical application, it provides reference for short video platform to improve platform competitiveness and promote the development of short video industry in a healthier and more sustainable direction.This article still exists problems such as poor cold start performance and need to improve the accuracy of multimodal feature fusion. In the future, we will focus on these two problems, further optimize the algorithm model, explore the application of social relationships, cross platform data collaboration, etc. in recommendation, continuously optimize the personalized recommendation strategy, and improve its applicability and stability.

AUTHOR CONTRIBUTIONS

Conceptualization –Rong Fan and Jinying Yan; methodology – Rong Fan and Jinying Yan; investigation – Rong Fan and Jinying Yan; writing-original draft preparation – Rong Fan and Jinying Yan. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was supported by, Shaanxi Province's first-class undergraduate course "Media Literacy" (course program number: 713) course construction results

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] I. A. Monir, M. W. Fakhr, and N. El-Bendary, "Multimodal deep learning model for human handover classification," Bulletin of Electrical Engineering and Informatics, vol. 11, no. 2, pp. 974-985, 2022, doi: 10.11591/eei.v11i2.3690.

[2] N. Xie, Y. Liang, Z. Luo, et al., "Early prediction of adverse outcomes in liver cirrhosis using a CT-based multimodal deep learning model," Abdominal Radiology, vol. 51, no. 1, pp. 137-150, 2026, doi: 10.1007/s00261-025-05045-0.

[3] J. L. Koyner, J. Martin, K. A. Carey, et al., "Multicenter Development and Validation of a Multimodal Deep Learning Model to Predict Moderate to Severe AKI," Clinical Journal of the American Society of Nephrology, vol. 20, no. 6, pp. 766-778, 2025, doi: 10.2215/CJN.0000000695.

[4] M. Verlyck, D. Zhao, E. Ferdian, et al., "A Multimodal Deep Learning Model for Non-Invasive Detection of Elevated Left Ventricular End-Diastolic Pressure," Heart, Lung & Circulation, vol. 34, no. Sup2, pp. S53-S54, 2025, doi: 10.1016/ j.hlc.2025.04.063.

[5] A. Bouatmane, A. Daaif, A. Bousselham, et al., "A Multimodal Deep Learning Model Integrating CNN and Trans former for Predicting Chemotherapy-Induced Cardiotoxicity," IEEE Access, vol. 13, pp. 57568-57588, 2025, doi: 10.1109/ ACCESS.2025.3556700.
[6] Y. Lei, B. Feng, et al., "Predicting microvascular invasion in hepatocellular carcinoma with a CT-and MRI-based multi modal deep learning model," Abdominal Radiology, vol. 49, no. 5, pp. 1397-1410, 2024, doi: 10.1007/s00261-024-04202-1.
[7] G. Tian, "Multimodal Co-Ranking of Music Sentiment Analysis with the Deep Learning Model," Journal of Electrical Systems, vol. 20, no. 6s, pp. 1446-1456, 2024, doi: 10.52783/jes.3059.
[8] C. Yuan, Q. Shi, X. Huang, et al., "Multimodal deep learning model on interim [18F]FDG PET/CT for predicting primary treatment failure in diffuse large B-cell lymphoma," European Radiology, vol. 33, no. 1, pp. 77-88, 2023, doi: 10.1007/ s00330-022-09031-8.
[9] W. H. Hui, W. H. Chiu, L. H. Yu, et al., "Reply: Technical considerations in the development of a multimodal deep learning model for predicting hepatocellular carcinoma outcomes," Hepatology, vol. 83, no. 2, pp. E87-E88, 2025, doi: 10.1097/ HEP.0000000000001540.
[10] S. Athreya, A. Melehy, S. S. A. Suthahar, et al., "Combining Ultrasound Imaging and Molecular Testing in a Multimodal Deep Learning Model for Risk Stratification of Indeterminate Thyroid Nodules," Thyroid, vol. 35, no. 5, pp. 590-594, 2025, doi: 10.1089/thy.2024.0584.
[11] B. M. Faltas, Z. Bai, M. Osman, et al., "Predicting clinical outcomes in the S1314-COXEN trial using a multimodal deep learning model integrating histopathology, cell types, and gene expression," Journal of Clinical Oncology, vol. 42, no. 4_suppl, pp. 533-533, 2024, doi: 10.1200/JCO.2024.42.4_suppl.533.
[12] H. S. Jung, E. J. Lee, D. I. Chang, et al., "A Multimodal Ensemble Deep Learning Model for Functional Outcome Prognosis of Stroke Patients," Journal of Stroke, vol. 26, no. 2, pp. 312-320, 2024, doi: 10.5853/jos.2023.03426.
[13] Y. Zhang, G. Qin, S. Yang, et al., "Sea Surface Height Inversion Model Based on Multimodal Deep Learning for the Fusion of Heterogeneous FY-3E GNSS-R Data," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 18, pp. 11979-11995, 2025, doi: 10.1109/JSTARS.2025.3562741.