Application of Multimodal Emotion Recognition Technology in Recommendation Systems

Wenhao Deng *

University of Electronic Science and Technology of China& Glasgow College, Chengdu, 330000, China

* Corresponding Author Email: 2960261D@student.gla.ac.uk

Abstract. Traditional recommender systems, relying mainly on users' historical behavioral data to predict preferences, often overlook the importance of real-time emotions in decision-making. As a result, they fail to meet the emotional needs of users. This study focuses on integrating multimodal emotion recognition technology with recommender systems to address this issue. It clarifies the necessity of such integration by analyzing the shift from correlation mining to causal understanding in recommendation systems and the value of multimodal emotion recognition. The research systematically analyzes five cutting-edge technologies: empathetic recommendation via large language models, robustness enhancement through causal inference, generative recommendation using diffusion models, emotion alignment via cross-modal contrastive learning, and privacy-preserving recommendation based on federated learning. Then, the research explores core challenges like efficiency, interpretability, and short-term homogenization, along with solutions such as knowledge distillation and neuro-symbolic methods. These technologies effectively tackle issues like emotional understanding, data bias, sparsity, modal alignment, and privacy protection. The study concludes that these technologies are converging, with future directions focusing on causal explanation, efficiency improvement, and long-term user well-being, driving recommender systems toward greater intelligence, robustness, reliability, and human-centricity.

Keywords: Multimodol emotion recognition technology, recommender system, large language model.

1. Introduction

In this era of information explosion, users are faced with a vast sea of data, making recommender systems a key bridge connecting them to relevant information. They are widely applied in applications such as e-commerce and streaming media platforms, influencing experience of users profoundly. However, there exist some inherent limitations in traditional recommender systems. They mainly rely on historical behavior data of users such as clicks, browsing records, purchase histories, or ratings to predict user preferences through methods like association rule mining or collaborative filtering. However, during this interaction, users' real-time emotion is ignored [1]. Human decision-making incorporates both rationality and sensibility, and emotion plays a decisive role in it [2]. It is fluctuations in human emotions that make traditional recommender systems fail sometimes.

To improve the accuracy of recommender systems, integrating emotional intelligence into recommending algorithms has become a new trend. Among various approaches, multimodal emotion recognition technology provides a brand-new and highly promising solution. This technology recognizes human emotions by comprehensively analyzing information from different modalities, such as semantic and emotional tones in text [3, 4].

This research aims to combine cutting-edge multimodal emotion recognition technology with recommender systems, constructing a new generation of intelligent recommendation systems that can perceive users' emotions and understand users' states. By dynamically adjusting strategies through real-time capture of emotional feedback, it will provide more personalized and humanized services, breaking through the limitations of traditional systems and contributing to the transformation of human-computer interaction.

This research helps promote recommendation systems from superficial association to in-depth empathy by integrating emotional theories from psychology with data modeling in computer science.

The research method of this review is to specifically and deeply prove the necessity of integrating emotional factors into the recommendation system. Summarize, review and compare key typical technologies in current multimodal emotion recognition research for recommender systems. Analyze the main challenges faced by current research at the theoretical and practical levels. Prospect the future development directions and cutting-edge trends in this field based on a profound understanding of existing research, providing valuable reference frameworks and research ideas for subsequent researchers. Organization of the Text.

2. Core Concepts and Technical Foundations

This chapter will first clarify the internal driving forces behind the development of recommender systems from traditional correlation mining to deeper understanding of cause and effect and explain why it is necessary to integrate multimodal emotion recognition technology with the recommender system.

2.1. Recommendation Systems: Challenges from Correlation to Causality

Recommendation systems are information filtering tools for online services, aiming to recommend items based on user preferences. Traditional recommendation systems, such as collaborative filtering or content-based filtering, mainly rely on two types of information. They are users' historical interaction data (ratings, reviews, clicks, purchase records, etc.) and item auxiliary information (identifiers and tabular features such as genre and year).

The core logic of this kind of system is correlation. They predict user preferences through mining "user-item" interaction patterns or content similarity. For example, collaborative filtering algorithms assume that "similar users like similar items", while content-based algorithms assume that "users are interested in items similar to their historical preferences". However, they only rely on correlation patterns instead of the fundamental motivation of user decisions.

There are significant limitations in such systems. Firstly, they lack semantic understanding. They only process behavioral symbols such as IDs and ratings, without understanding the deep reasons for user preferences. They know that user A "likes" movie B but cannot explain "why", completely ignoring emotion as a key factor in decision-making. This makes recommendations conform to historical correlations but violate users' real-time emotional needs. Secondly, data sparsity and cold start exist Systems rely on large amounts of interaction data, but user-item interactions are often sparse in reality. New users or items are difficult to get effective recommendations owing to the lack of a correlation basis.

To address these challenges, integrating users' emotional states as important contextual information into recommender models not only helps alleviate data sparsity but also enables systems to understand the deep reasons for user preferences. As a result, more accurate and humanized recommendations can be provided. This is the fundamental motivation for introducing emotion recognition technology into recommendation systems in this study.

2.2. Multimodal Emotion Recognition

Multimodal emotion recognition refers to the technical process of accurately identifying and understanding human emotional states by integrating and deeply analyzing information from multiple data modalities. These data modalities consist of speech, vision (e.g., facial expressions, body movements), text (e.g., semantic content, emotional vocabulary), and physiological signals (e.g., EEG, GSR, ECG). Its core idea is that information from a single modality is often ambiguous, incomplete, or even deceptive, while integrating multiple information sources can build a more robust and accurate emotional analysis system. This mechanism is rooted in biomimetics, aiming to imitate humans' ability to seamlessly integrate multisensory cues for emotional communication and understanding.

Compared with unimodal solutions depending on a single information source, multimodal methods have advantages in overcoming the inherent limitations of unimodality. For example, facial expression recognition systems might perform poorly under the circumstances with poor lighting, facial occlusion, or subtle expressions; speech systems are susceptible to background noise interference. By combining the rhythm of voice, facial micro-expressions and text semantics, multimodal systems can effectively filter noise and correct deviations in complex real-world scenarios (such as noisy public places and social interactions with emotional concealment). This helps obtain more reliable recognition results in the complex real world and provides solid technical support for emotion-driven intelligent systems [5].

This chapter sorts out the core concepts of recommendation systems and multimodal emotion recognition, clarifies the necessity of their integration, and sets a foundation for the subsequent analysis of typical technologies.

3. Analysis of Typical Multimodal Emotion Recognition Technologies for Recommender Systems

With the continuous deepening of the interdisciplinary field between recommender systems and multimodal emotion recognition, a series of typical technologies aimed at improving the system's empathy ability, robustness, generative ability, alignment accuracy, and privacy protection level have emerged. These technologies introduce the newest achievements in the field of artificial intelligence, fundamentally reshaping the construction of emotional recommender systems. This chapter will systematically analyze five representative cutting-edge technologies: empathy recommendation technology based on large language models, robustness enhancement technology based on causal inference, multimodal generative recommendation technology based on diffusion models, emotion alignment technology based on cross-modal contrastive learning, and privacy-preserving recommendation technology based on federated learning.

3.1. Empathetic Recommendation Technology Based on Large Language Models

In recent years, the rise of large language models (LLMs) has brought the potential for a paradigm revolution in the field of recommender systems. Especially when dealing with emotion-rich texts, they demonstrate the ability to capture subtle emotions that traditional models can hardly reach, providing a new technical path for constructing empathetic recommender systems. However, the high computational cost of LLMs and their inherent flaw of being prone to "hallucinations" pose severe challenges for them in recommendation scenarios requiring high efficiency and reliability.

The Sequential Emotion-Aware LLM-Based Personalized Recommendation System (SEALR) can effectively address the issues mentioned above while ensuring high performance. This framework divides the recommendation process into three stages. In the emotional label sequence generation stage, a RoBERTa model that has been pre-trained on Go Emotions is used to parse emotional label sequences (such as joy, admiration, sadness, etc.) from users' historical comments, providing structured emotional input for decision-making. To reduce computational burden and avoid hallucinations, the candidate item retrieval stage uses an efficient SASRec model to generate a high-quality candidate pool with controllable scale. In the instruction fine-tuning and recommendation generation stage, users' interaction history, emotional sequences, and the candidate item pool are fed into the instruction-fine-tuned LLaMA2-7B. In this core stage, the SEALR model comprehensively analyzes both behavioral and emotional clues, re-ranks the items in the candidate pool, and generates accurate recommendations [6].

The performance of the SEALR model was evaluated on the Amazon Product Reviews and Yelp datasets, using Hit Rate (HR@K) and Normalized Discounted Cumulative Gain (NDCG@K) as core evaluation metrics. On the Yelp dataset, SEALR achieved HR@10 of 0.2046 and NDCG@10 of 0.0930, significantly outperforming traditional strong baseline models. This confirms the great

potential of LLM frameworks that deeply integrate users' emotional sequences in accurately understanding user preferences.

3.2. Robustness Enhancement Technology Based on Causal Inference

Traditional multimodal recommender systems can integrate rich content information. However, they generally attribute users' interactions to their interest in item content, which oversimplifies the assumption. In fact, users' interactions are usually influenced by two independent factors: multimodal interest (the genuine preference for the content itself) and multimodal conformity (the willingness to follow others under the influence of popularity, trending reviews, etc.). This prevents models from identifying the fundamental causes of user behaviors, limiting their robustness.Causal inference theory, especially the analysis of collider structures, provides a solid theoretical foundation for solving this problem. User interaction refers to a "collider point", and its two independent causes (interest and conformity) become correlated after the interaction, thus requiring separation during modeling.

Multi-modal Graph Disentangled Causal Collaborative Filtering for Recommendation model (MGCE) realizes causal reasoning through a multimodal causal graph. It takes "multimodal interest" and "multimodal conformity" as two independent parent nodes, jointly pointing to the child node "interaction", forming a typical collider structure. Based on this causal graph, MGCE designs two parallel graph learning networks to learn the embedding representations of interest and conformity, respectively. Specifically, the interest learning network only uses the multimodal content features of users and items for message passing, aiming to capture users' pure preference. The conformity learning network additionally integrates the popularity features of items to explicitly model the conformity effect. To ensure that the embeddings learned by the two networks are semantically separated, the model introduces a differentiation task to enforce disentanglement by maximizing the distance between "interest embeddings" and "conformity embeddings" [7].

The final step of MGCE is the prediction and optimization process. The formula for calculating the final recommendation score is as follows:

$$\hat{y}_{u,i} = \hat{y}_{u,i}^{base} + \lambda_m^{\text{int}} \cdot \hat{y}_{u,i}^{\text{int}} + \lambda_m^{con} \cdot \hat{y}_{u,i}^{con} \tag{1}$$

Where $\hat{y}_{\text{u,i}}^{\text{base}}$ is the score obtained from a basic collaborative filtering model, $\hat{y}_{u,i}^{int}$ is the interest score calculated by the interest learning network. $\hat{y}_{u,\text{i}}^{con}$ is the conformity score calculated by the conformity learning network. λ_m^{int} and λ_m^{con} are hyperparameters that control the contribution of each component. This formula shows that the final recommendation decision is the result of comprehensively considering users' general preferences, genuine interests, and conformity psychology.

In terms of performance evaluation, the study adopts the standard Top-K recommendation paradigm and uses HR@k and NDCG@k as core evaluation metrics on three multimodal datasets: Beauty, Art, and Taobao. Experiment results show that MGCE is significantly better than strong baseline models such as DMRL and GRCN, verifying that causal inference can accurately capture the motivations behind user behaviors and enhance the robustness of the model.

3.3. Multimodal Generative Recommendation Technology Based on Diffusion Models

Another core challenge faced by recommender systems is data sparsity. Although existing self-supervised learning methods can alleviate this problem through data augmentation, simple augmentation strategies may introduce noise and damage the original semantic structure of the data. As the latest breakthrough in the field of generative models, diffusion models have generation capabilities with high-quality and accurate modeling of data distribution. Their application in recommender systems aims to address data sparsity. They do this through a generative augmentation paradigm that conforms to the inherent data distribution, specifically by generating high-quality, "modality-aware" user-item interaction graphs that integrate multimodal information. The technical

workflow of DiffMM, a representative work in this field, consists of two core stages. The forward diffusion process involves gradually adding Gaussian noise over T timesteps to a real "user-item" interaction graph until its structure is completely destroyed and it is transformed into pure noise [8].

Train a neural network to learn how to gradually recover the original graph structure from the noisy graph. In each denoising step, a modality-aware signal is injected. This signal is aggregated from the multimodal features of items (e.g., visual, textual features). In addition, it is used to guide the denoising network, ensuring that the generated interaction edges are consistent with the items' multimodal features. The model's loss function combines the standard ELBO loss and the modality signal injection loss (L_{msi}).

$$\mathcal{L}_{msi}^{m} = \left\| \hat{\alpha}_0 \cdot e_m^i - \alpha_0 \cdot e^i \right\|_2^2 \tag{2}$$

This is the reverse denoising process. Experiments on multimodal datasets such as TikTok and Amazon show that DiffMM achieves state-of-the-art (SOTA) performance in Top-K recommendation metrics including Recall@K and NDCG@K, significantly outperforming models such as LightGCN, MMGCN, and those based on other self-supervised learning methods. This fully shows the strong capability of diffusion models in generating interaction data with high quality and modality-aware to improve recommendation performance.

3.4. Emotion Alignment Technology Based on Cross-Modal Contrastive Learning

In multimodal scenarios, users' preferences are usually reflected by information from different modalities. For instance, a high numerical rating given by a user should be semantically consistent with their positive text review. However, data from these heterogeneous modalities resides in different representation spaces. It is hard for models to directly capture their consistency. It can be a tricky problem for traditional models if a user gives a high rating but a negative review. Cross-modal contrastive learning offers an efficient solution to this issue. Its core idea is to learn a unified, semantically aligned representation space by "pulling similar samples closer and pushing dissimilar samples farther apart" in the representation space [9].

The SERMON model is an innovative approach. Its workflow is as follows: Two modal representations are constructed for a single user-item interaction event. One is the interaction triple representation (h_r), which is composed of user ID, item ID, and key "aspect" embedding vectors; the other is the text representation (h_e) of rich text semantics encoded from user reviews. The (h_r h_e) pairs from the same interaction are considered as positive sample pairs, while those from different interactions are regarded as negative sample pairs. The model is optimized using the following contrastive loss function.

$$\mathcal{L}_{cl} = -\log \frac{e^{sim\left(h_{r,i}, h_{e,i}\right)/\tau}}{\sum_{j=1}^{N} e^{sim\left(h_{r,i}, h_{e,j}\right)/\tau}} \tag{3}$$

Here, "sim" stands for cosine similarity, and τ is the temperature hyperparameter. It forces the model to learn a shared representation space by maximizing the similarity of positive sample pairs and minimizing the similarity of negative sample pairs. In this space, structured interaction preferences and unstructured text sentiments achieve alignment.

On Amazon (cells), Yelp (restaurants), and TripAdvisor (hotels) datasets, SERMON demonstrates excellent performance. Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are used as evaluation metrics for rating prediction, while BLEU, ROUGE, and BERTScore serve as metrics for explanation generation. On the TripAdvisor dataset, SERMON achieves an RMSE of 0.791 and an MAE of 0.599 in rating prediction, which is better than strong baseline models such as NRT and PETER. In terms of explanation generation quality, its BLEU-4 score reaches 1.18 on the TripAdvisor dataset, improving 24.5% over baselines. This proves the effectiveness of cross-modal contrastive learning in enhancing model representation capabilities and performance on downstream tasks.

3.5. Privacy-Preserving Recommendation Technology Based on Federated Learning

With users taking data privacy more and more seriously, the traditional approach of centrally storing and training user data is facing significant challenges. This is especially critical in scenarios like intelligent healthcare, where user multimodal emotional data (e.g., physiological signals, facial expressions) is extremely sensitive.

Federated learning provides a solution to this problem. It is a distributed machine learning framework. It enables the joint training of a shared global model without transmitting users' raw data out of local devices, thereby achieving model optimization while providing powerful privacy protection.

In the PFL-MCL framework of federated learning, each client utilizes local, heterogeneous multimodal data to train a local model. This model typically contains a "personalization module" for capturing users' personality and a "federation module" for participating in global aggregation. Clients only encrypt and upload the model parameters of the updated "federation module" to the central server, while raw data is saved locally. The central server collects model updates from multiple clients and uses an aggregation algorithm to fuse them into a better global model. The server then distributes the aggregated new global model to each client as the beginning for the next round of local training. This process cycles repeatedly until the model converges [10].

Experiments were conducted on the Amazon Product Reviews dataset and the self-constructed Virtual Shopping Mall dataset, using Precision, Recall, and F-Measure as offline metrics, and click-through rate (CTR) as the online metrics. The results show that the PFL-MCL framework can not only protect privacy but also perform well. On the self-constructed Virtual Shopping Mall dataset, when the candidate set size M=5, the F-Measure of PFL-MCL in the recommendation task reaches 85.3%, which is significantly higher than the 63.8% of the non-federated centralized model and the 63.2% of the standard federated averaging algorithm. In online simulation tests, it also achieved the best performance in terms of click-through rate.

3.6. Technology Comparison

The above are five cutting-edge multimodal sentiment recognition technologies for recommender systems. These technologies promote the development of emotional recommendation from different dimensions respectively: LLMs equip the system with unprecedented empathetic understanding capabilities; causal inference works on improving the robustness and fairness of models; diffusion models provide a powerful generative solution to address data sparsity; cross-modal contrastive learning focuses on solving the challenges of alignment and fusion of heterogeneous data; and federated learning lays the foundation for privacy protection in the application of sensitive data.

The following table (Table 1.) compares the five key technologies for recommender systems mentioned above. These technologies can address core issues such as sentiment semantic understanding, data bias and robustness, data sparsity, multimodal alignment, and privacy protection respectively. They exert their advantages through their unique advantages, but also face different challenges. Among these challenges, efficiency and deployment (e.g., model computing costs, latency) and technical depth (e.g., explanation transformation, long-term modeling) are the main difficulties.

Table 1. Comparison of Five Key Technologies

Typical Technology	Problems Solved	Key Mechanisms	Advantages	Challenges
Large Language Model	Deep sentiment semantic understanding	Instruction tuning, in-context learning, generative reasoning	Handles complex sentiments, generates natural language explanations	Efficiency & Deployment: Enormous model parameters
Causal Inference	Data bias (popularity, conformity, etc.), model robustness and fairness	Causal graph construction, backdoor adjustment, counterfactual reasoning	Enhances model robustness & interpretability	Interpretability: Difficulty converting to user-friendly natural language
Diffusion Model	Data sparsity, high-quality data augmentation	Iterative forward noising and reverse denoising generation	High-quality generated data, accurate learning of data distribution,	Efficiency & Deployment:-
Cross-Modal Contrastive Learning	Heterogeneous multimodal data representation alignment & fusion	Pull positive pairs closer, push negative pairs farther in shared space	No strong supervised labels, learns unified cross-modal representations	Modeling Depth: Short-term interaction, no user long-term state evolution
Federated Learning	User data privacy protection	Distributed and decentralized model training, only parameter exchange	Raw data stays local, strong privacy protection	Efficiency & Deployment: High communication overhead, training latency

To conclude, these technologies are not isolated from each other but show a trend of integrated development. For instance, contrastive learning can be applied within the framework of federated learning to process local heterogeneous multimodal data. Additionally, causal inference can be used to guide the reasoning process of LLMs to reduce biases. Their combination will collectively drive recommender systems toward a more intelligent, more robust, more reliable, and more human-centric direction.

If you follow the “checklist” your paper will conform to the requirements of the publisher and facilitate a problem-free publication process.

4. Challenges and Cutting-Edge Solutions

Although these typical technologies have greatly improved the intelligence level of recommender systems, their adoption has also brought about profound challenges. These challenges not only relate to the computational efficiency and deployment feasibility of the technology but also to the fundamental issues such as users’ trust and long-term well-being. This chapter will focus on the three core current challenges in the field and discuss corresponding cutting-edge solutions.

4.1. Model Efficiency and Deployment Dilemmas

LLMs and diffusion models are widely applied, but recommender systems based on them face efficiency and deployment issues such as high latency and high overhead. For example, billions of parameters of the LLaMA2-7B large language model used by SEALR bring high training and inference costs. The generation process of the DiffMM diffusion model relies on computationally intensive iterative denoising, which has very strict requirements on latency. The frequent model parameter exchanges in the PFL-MCL federated learning framework also generate huge communication overhead.

To address the deployment difficulty of models, knowledge distillation provides an effective solution. Its core idea is to extract knowledge from a powerful "teacher model" for a lightweight

"student model" to learn. The student model not only learns from original data but also imitates the output of the teacher model. This methodology of knowledge distillation not only retains strong performance, but also maintains extremely low inference delay, which greatly improves the deployment feasibility [11].

4.2. From "Black-Box" to "Trustworthy" — Model Interpretability and User Trust Crisis

The "black-box" feature of deep learning models is a major hinder to attaining users' trust. That is, recommendation systems lack interpretability. Users are not accessible to the recommendation process, leading to a trust crisis. Although models like SERMON make an effort to provide interpretable recommendations by generating natural language, such explanations are still based on correlations in the data essentially. True interpretability needs to move from correlation to causality, that is, explanations should reveal "how decision factors (causes) influence user behaviors (results)". Correlational explanations lacking in-depth logic are difficult to resolve the user trust crisis fundamentally

Neuro-symbolic methods have built a new path for constructing trustworthy recommender systems. This mechanism aims to combine the powerful representational learning ability of neural networks with the clear reasoning ability of symbolic systems (such as logical rules and knowledge graphs). By using Logical Tensor Networks, domain knowledge can be encoded into first-order logical formulas and used as one of the objective functions for model training. In this way, recommendation systems can effectively solve the black-box problem, make the logical reasoning process transparent, and provide users with convincing explanations [12].

4.3. Short-Term Homogenization Limitations of Recommendation Systems

The current research paradigm of recommendation systems generally has a "myopia" problem: model design and evaluation over-concentrate on optimizing short-term interaction metrics. The evaluation metrics of SEALR, such as HR and NDCG, measure the accuracy of predicting users' next behavior. If users become tired of homogeneous items in the short term, recommender systems cannot effectively solve this problem, resulting in issues like information cocoons.

To extend user modeling from short-term prediction to long-term well-being, building high-fidelity user simulators is a highly promising direction. The RecUserSim model, which utilizes an agent-based user simulator built with LLMs, can generate diverse users like real humans that interact with recommender systems in multiple turns. With this method, the optimization goal of recommendation systems can shift from maximizing "the next click rate" to maximizing a long-term value function that comprehensively measures users' satisfaction, emotional health, etc., over a period [13]. This allows for evaluating and optimizing the long-term strategies of recommendation systems through simulation before actual deployment, making recommender systems more responsible.

5. Conclusion

This review systematically discusses a cutting-edge interdisciplinary field, the integrated application of multimodal emotion recognition technology and recommendation systems. It starts from a core argument: traditional recommender systems depend on users' historical interaction, leading to limitations in user experience and recommendation effectiveness. Multimodal emotion recognition technology, through the comprehensive analysis of multi-source information such as text, vision, and speech, can provide key technical support for building "empathetic" recommendation systems that can understand and respond to users' real-time emotions.

After introducing the basic concepts of recommendation systems from correlation to causality and multimodal emotion recognition, this review focuses on analyzing five typical technologies that drive the development of this field. Firstly, empathy recommendation technology based on large language models enhances the system's semantic understanding and interaction capabilities. Secondly, robustness enhancement technology based on causal inference is committed to solving data bias and

model credibility issues. Thirdly, multimodal generative recommendation technology based on diffusion models provides a new paradigm for addressing the challenge of data sparsity. Moreover, emotion alignment technology based on cross-modal contrastive learning effectively solves the problem of representation alignment between heterogeneous modalities. Finally, privacy-preserving recommendation technology based on federated learning provides security guarantees for processing sensitive emotional data.

However, while these cutting-edge technologies bring leaps in performance, they also trigger three severe challenges: model efficiency and deployment difficulties, model interpretability and user trust crisis, as well as the short-term homogenization limitation of recommendation systems. In conclusion, this review not only draws a panoramic view of the current development of technology, but also reveals the key bottlenecks in this field from technical implementation to application landing, pointing out the direction for future research.

Future research on multimodal emotion recognition for recommender systems will focus on three key directions. Integrate neuro-symbolic AI to realize causal explanation, solving the "black box" problem of deep learning, and building reliable recommendation systems. Improve recommendation efficiency to cope with deployment difficulties through model compression, lightweight basic models, and hybrid systems. Construct user simulators and combine the long-term prediction ability of recommendation systems to solve the problem of recommendation homogenization and develop more responsible recommendation systems.

Ultimately, research on multimodal emotion recognition for recommendation systems is moving towards a new crossroads, with the final goal of promoting a profound evolution of recommendation systems. Future recommender systems will no longer be cold data processors but "human-like" intelligent partners that can understand, respect, and actively guide users' long-term well-being.

References

[1] D. Wang, X. Zhao, "Affective video recommender systems: A survey," Frontiers in Neuroscience, vol. 16, pp. 984404, 2022.

[2] Q. Liu, J. Hu, Y. Xiao, X. Zhao, J. Gao, W. Wang, Q. Li, J. Tang, "Multimodal Recommender Systems: A Survey," arXiv preprint arXiv:2302.03883, 2024.

[3] J. Pan, Z. He, Z. Li, Y. Liang, and L. Qiu, "A review of multimodal emotion recognition," CAAI Transactions on Intelligent Systems, vol. 15, no. 4, pp. 633 - 645, 2020.

[4] Y. Wu, Q. Mi, T. Gao, "A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions," Biomimetics, vol. 10, no. 7, pp. 418, 2025.

[5] D. Mamieva, A. B. Abdusalomov, A. Kutlimuratov, B. Muminov, T. K. Whangbo, "Multimodal Emotion Detection via Attention-Based Fusion of Extracted Facial and Speech Features," Sensors, vol. 23, no. 12, pp. 5475, 2023.

[6] N. Lee, J. Kim, "SEALR: Sequential Emotion-Aware LLM-Based Personalized Recommendation System," 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), Padua, Italy, 2025, pp. 1 - 5.

[7] S. Li, F. Xue, K. Liu, D. Guo, R. Hong, "Multimodal Graph Causal Embedding for Multimedia-Based Recommendation," IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 12, pp. 8842 - 8858, 2024.

[8] Y. Jiang, L. Xia, W. Wei, D. Luo, K. Lin, C. Huang, "DiffMM: Multi-Modal Diffusion Model for Recommendation," arXiv preprint arXiv: 2406.11781, 2024.

[9] H. Liao, S. Wang, H. Cheng, W. Zhang, J. Zhang, M. Zhou, K. Lu, R. Mao, X. Xie, "Aspect-Enhanced Explainable Recommendation with Multi-modal Contrastive Learning," ACM Transactions on Intelligent Systems and Technology, vol. 16, no. 1, Article 8, 2025.

[10] X. Zhou, Q. Yang, X. Zheng, W. Liang, K. I-K. Wang, J. Ma, Y. Pan, Q. Jin, "Personalized Federated Learning with Model-Contrastive Learning for Multi-Modal User Modeling in Human-Centric Metaverse," IEEE Journal on Selected Areas in Communications, vol. 42, no. 4, pp. 817 - 832, 2024.

[11] Y. Cui, F. Liu, P. Wang, B. Wang, H. Tang, Y. Wan, J. Wang, J. Chen, "Distillation Matters: Empowering Sequential Recommenders to Match the Performance of Large Language Models," 18th ACM Conference on Recommender Systems (RecSys '24), Bari, Italy, 2024, pp. 507 - 517.

[12] T. Carraro, "Overcoming Recommendation Limitations with Neuro-Symbolic Integration," 17th ACM Conference on Recommender Systems (RecSys '23), Singapore, Singapore, 2023, pp. 1325 - 1331.

[13] L. Chen, Q. Dai, Z. Zhang, X. Feng, M. Zhang, P. Tang, X. Chen, Y. Zhu, Z. Dong, "RecUserSim: A Realistic and Diverse User Simulator for Evaluating Conversational Recommender Systems," ACM Web Conference 2025 (WWW '25 Companion), Sydney, Australia, 2025, pp. 133 - 142.