Artificial Intelligence and Computing Innovations
Intelligence Science Publishing
https://www.ispublishing.com
ISSN: 3070-4944
Vo.5, No.10, 2025

MCA-TAN: Cross-Modal Graph Neural Recommendation with Temporal Behavior and Social Network Fusion

Ziruo Ke[1*]
[1]University of Washington, Seattle, USA
**Corresponding author: kziruo@gmail.com*

Abstract: This paper addresses the problem of how to integrate user behavior sequences and social network information to improve personalized recommendation. A recommendation model based on cross-modal graph neural networks is proposed. The method introduces a Modality-Coordinated Aggregation (MCA) mechanism to jointly model the user's behavior graph and social graph. This enables the model to capture the structural complementarity and semantic interaction between the two modalities. A Temporal-Aware Node (TAN) representation module is also introduced to fully extract temporal evolution patterns from behavior sequences, enhancing the model's ability to capture changes in user interests over time. The model first extracts high-order representations separately from the behavior graph and the social graph. Then, a modality gating mechanism is used to perform dynamic weighted fusion, generating a unified user representation vector. This fused representation is combined with the target item embedding and input into a matching function to produce the final recommendation results. Experiments on public datasets show that the proposed method outperforms several representative recommendation models across multiple mainstream evaluation metrics. It demonstrates notable advantages in precision, coverage, and ranking quality. Furthermore, ablation studies confirm the contribution of each module to performance gains. Cold-start analysis and modality weight sensitivity experiments further reveal the model's stability and adaptability. The experimental results strongly support the value of joint modeling of behavior and social graphs in addressing complex recommendation tasks. They also highlight the modeling capacity of graph neural architectures in multimodal recommendation scenarios.

Keywords: Cross-modal recommendation, graph neural network, behavior modeling, social representation learning

1 Introduction

With the rapid development of information technology, personalized recommendation systems have become essential tools for users to access information and services. They are widely used in e-commerce, social media, and content distribution. As the dimensions of user data continue to expand, how to fully exploit and utilize multi-source heterogeneous data to improve the accuracy and robustness of recommendation systems has become a key research issue. Among these data sources, user behavior sequences and social network information represent two important perspectives: behavior-driven and social influence[1,2]. Traditional recommendation methods often treat these two types of information separately. This limits their ability to complement and integrate, thus restricting further improvement in recommendation performance[3].

User behavior sequences mainly reflect user interactions on the platform, such as browsing, clicking, and purchasing. These behaviors show clear temporal and dynamic characteristics. Recommendation methods based on behavior sequences can capture the evolution of user interests over time. They perform particularly well in modeling short-term interests. However, user behavior in specific contexts is often influenced by social relationships. Recommendations from friends, group opinions, and social interactions can subtly affect individual actions. Therefore, systems that rely solely on historical behavior face challenges in modeling long-term preferences and uncovering latent interests[4,5].

In contrast, social network data reveals influence mechanisms within social structures through user followings, interactions, and information diffusion[6]. Social-based recommendation methods infer potential preferences using social connections, enabling social information diffusion and collaborative filtering. However, these methods are less effective at capturing dynamic individual behavior and may overlook temporal interest shifts[7]. Moreover, social data is often sparse and noisy. When used in isolation, it can undermine the robustness and precision of the system. As a result, effectively integrating user behavior sequences with social networks to uncover their latent correlation patterns has become an important direction for improving recommendation performance.

In recent years, with the development of graph neural networks, recommendation systems have begun to use graph structures to model complex relationships among multimodal data. Graph neural networks have natural advantages in capturing high-order relations and modeling heterogeneous structures. They offer powerful tools to integrate user behaviors and social networks. However, user behavior sequences and social networks represent typical cross-modal data. The former emphasizes temporal features, while the latter reflects topological dependencies. They differ significantly in representation format, semantic level, and structural relationship. Designing a unified representation learning framework that captures the interaction between behavior and social modalities and achieves efficient fusion is a critical challenge[8,9].

Therefore, studying cross-modal graph neural recommendation algorithms that unify user behavior sequences and social network information has both theoretical importance and practical value. On the one hand, this research promotes the integration of multimodal representation learning and graph neural networks in recommendation systems, providing new ideas for modeling complex heterogeneous data[10]. On the other hand, unified modeling of user behavior and social relations enables comprehensive understanding of user preferences. It helps alleviate data sparsity and cold-start problems, enhances personalized services, and improves adaptability to dynamic environments. As data ecosystems become increasingly complex and diverse, building cross-modal, unified, and scalable recommendation models will be a key direction for the future development of intelligent recommendation systems.

2 Related Work

2.1 Graph Neural Networks

As a significant branch of deep learning in recent years, graph neural networks have shown notable advantages in processing graph-structured data. They have been widely applied in social analysis, knowledge graphs, and recommendation systems. The core idea is to learn node representations through neural networks by aggregating information from neighboring nodes layer by layer. This helps capture high-order semantic relations[11]. In recommendation systems, user-item interactions can be naturally modeled as a graph. Graph neural networks can effectively capture users' latent preferences and item correlations, thereby enhancing the expressiveness of recommendations. Early graph-based recommendation methods were mostly built on static graphs, such as user-item bipartite graphs. They modeled user interests through multi-layer message passing. However, they were limited in handling dynamic behaviors and complex heterogeneous relations[12].

To improve the representation capacity of recommendation systems, researchers have extended graph neural networks to various types of graph structures. These include heterogeneous graphs, multi-hop graphs, and dynamic graphs. For example, incorporating social relations among users to build social graph neural models can better integrate trust and social influence[13]. This helps alleviate data sparsity in recommendation. Some studies also combine user behavior sequences with graph structures to model the evolution of user interests over time. These methods often adopt graph attention mechanisms, positional encodings, or sequence aggregation modules to capture preferences at a finer granularity. However, despite progress in each direction, behavior sequences and social graphs are often modeled separately. Deep integration of cross-modal data under a unified framework remains lacking[14].

In recent years, cross-modal graph neural networks have emerged as a research hotspot. For the fusion of heterogeneous modalities such as user behavior and social relations, some works have introduced joint learning mechanisms, graph contrastive learning, or multi-channel graph convolution architectures. These methods aim to leverage the complementary properties of different modalities and enhance the diversity and robustness of user representations. However, challenges still exist in current fusion strategies. These include large semantic gaps between modalities, difficulties in resolving information redundancy and conflicts, and the lack of unified frameworks for cross-modal alignment. Therefore, developing a cross-modal graph neural recommendation framework that is structurally unified, semantically aligned, and adaptively integrated remains a core problem that needs further exploration in this field[15].

2.2 Social network recommendation algorithm

Social recommendation algorithms aim to improve recommendation performance by leveraging users' social relationships. The core assumption is that social connections can reflect potential interest similarity between users[16]. Early methods were mostly based on collaborative filtering. They treated social neighbors as channels of interest propagation and integrated their preference information through weighted aggregation to generate personalized recommendations. These methods alleviated data sparsity and cold-start issues to some extent. They showed good performance especially when user behavior data was limited[17]. However, traditional social recommendation models typically modeled social relations in a static manner. They lacked detailed representation of influence strength, propagation mechanisms, and temporal dynamics[18].

With the increasing richness of data from social platforms, more complex social recommendation models have emerged. These models aim to explore the influence of social relations on user decisions from multiple perspectives. One category of methods introduces social influence modeling. They assign different weights to friends' influence on a user's decisions. Techniques such as attention mechanisms and adaptive fusion strategies are used to enhance model expressiveness[19]. Another category incorporates social context and behavioral context. They include social environmental factors when modeling user behavior to improve contextual awareness in recommendation. In addition, social networks exhibit strong structural characteristics, such as community structures and small-world effects. These provide new perspectives and possibilities for preference propagation and social-enhanced modeling.

In recent years, with the development of graph neural networks, social recommendation has moved toward more structured modeling. Social relations among users are modeled as graph structures. Graph convolution or message passing mechanisms are used to realize high-order relation propagation. These methods capture complex dependencies in social networks more effectively[20]. They also integrate user-item interaction data to build joint graphs for collaborative modeling. This improves recommendation accuracy and generalization. However, most existing methods still focus on structural features of social relations. There is limited research on the temporal evolution of social influence and the fusion of multimodal social signals. Therefore, how to retain the advantages of social structure while incorporating behavioral dynamics and semantic information remains an important research direction in social recommendation.

Canonical graph-based recommendation models further shaped the design space: LightGCN simplifies graph convolution for collaborative filtering, DGCF separates latent intents, SocialGCN formalizes social propagation, GraphRec extends graph recommendation to expert matching, and GRCN refines multimedia recommendation with implicit feedback[21,22,23,24,25].

Beyond recommendation-specific models, recent data-mining studies emphasize robust and scalable learning under heterogeneous conditions. Heterogeneity-aware federated optimization supports distributed preference modeling across uneven user populations[26], while key signal-aligned memory compression and adaptive fusion of low-rank and soft-gated sparse updates illustrate how informative signals can be retained and fused compactly in large representation spaces[27,28]. Time-varying operational graph reasoning further shows how graph-structured reasoning can support repair planning under changing risk constraints, which is consistent with the need to update user and relation representations as behavior evolves[29].

Knowledge-augmented and multi-source prediction methods also inform the proposed cross-modal fusion perspective. Keyword-constrained retrieval-augmented generation emphasizes controlled access to external knowledge in long-context modeling[30], and multi-source feature fusion for supply-chain delay risk prediction demonstrates how heterogeneous evidence can be combined into robust decision outputs[31]. In parallel, semi-supervised robust optimization for noisy low-resource data mining highlights the importance of learning reliable representations when labeled signals are limited or incomplete[32].

System-level studies on cloud and edge intelligence further reinforce the deployment requirements of graph-based recommendation. Predictive GPU memory modeling in containerized environments points to the need for efficient model execution under constrained resources[33]. Graph neural anomaly detection in multi-tenant clouds and lightweight anomaly detection in distributed edge-cloud systems show that relational modeling remains valuable when data streams are sparse, noisy, and operationally dynamic[34,35].

3 Method

This paper proposes a cross-modal graph neural recommendation algorithm that unifies user behavior sequences and social network information. The goal is to build a unified modeling framework that integrates multi-source heterogeneous user data. The method introduces a joint representation of sequential behavior graphs and social relation graphs. It connects behavior-driven and socially influenced preference modeling to provide a comprehensive characterization of user interests. To address the alignment challenges caused by representational differences across modalities, the algorithm designs a Modality-Coordinated Aggregation (MCA) mechanism. This mechanism enables information interaction and complementarity between behavior and social modalities while preserving their structural properties. To enhance the model's ability to capture dynamic changes in user interests, a Temporal-Aware Node representation (TAN) module is introduced. It incorporates behavior order and time interval information to improve the temporal sensitivity of node representations. Together, these two core modules form a cross-modal graph neural recommendation framework with semantic alignment and structural consistency. The model architecture is shown in Figure 1.

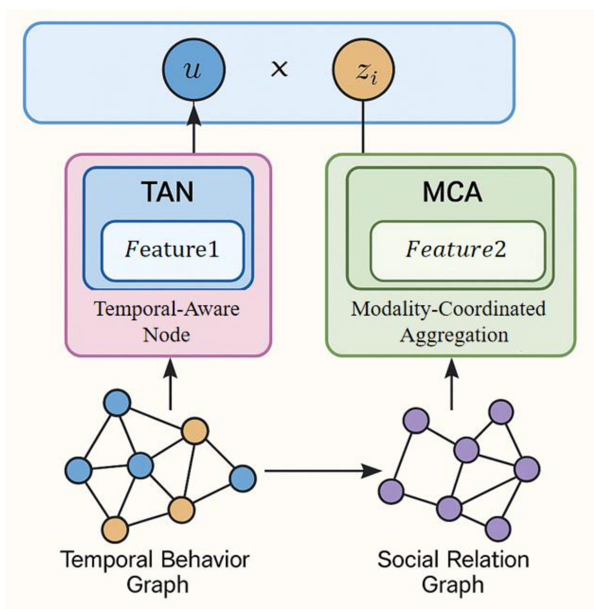

Figure 1. Overall model architecture diagram

3.1 Modality-Coordinated Aggregation

In this study, we proposed a Modality-Coordinated Aggregation (MCA) mechanism to unify the feature representation of user behavior sequences and social network information. The model architecture is shown in Figure 2.

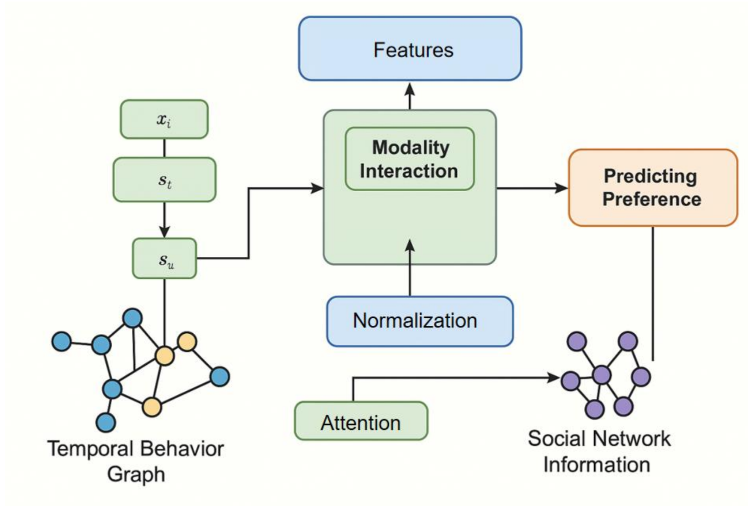

Figure 2. MCA module architecture

Aiming at the various connection relationships of users in the social graph and their dynamic preferences in the behavior sequence, we encode the social graph and the behavior graph respectively. Specifically, the feature propagation between users and neighbor nodes in the social graph can be modeled as a graph convolution operation, which is defined as follows:

$$h_u^{(l+1)} = \sigma(\sum_{v \in N_u} \frac{1}{c_{uv}} W^{(l)} h_v^{(l)}) \tag{1}$$

Among them, $h_u^{(l)}$ represents the representation of user u in layer l, N_u represents its neighbor set, c_{uv} is the normalization coefficient, $W^{(l)}$ is the learnable weight matrix, and σ is the nonlinear activation function. At the same time, the user's behavior sequence is constructed in chronological order as a time series behavior graph, in which we use a weight function based on time intervals to aggregate adjacent behavior nodes to capture their interest evolution trend:

$$s_t = \sum_{i=1}^{T} \alpha_i \cdot x_i \, \alpha_i = \frac{\exp(-\lambda \cdot \Delta t_i)}{\sum \exp(-\lambda \cdot \Delta t_j)} \tag{2}$$

Where x_i represents the embedding of the i-th behavior, Δt_i is the time difference between the behavior and the current moment, λ controls the intensity of time decay, and α_i is the attention weight.

In order to fuse the representation information of these two modalities, we introduced a modality interaction gating mechanism in MCA to control the contribution of each modality to the final embedding. The fused embedding representation is defined as follows:

$$z_u = \gamma \cdot s_t + (1 - \gamma) \cdot h_u \tag{3}$$

$$\gamma = \sigma(W_g [s_t || h_u] + b_g) \tag{4}$$

Among them, z_u is the unified representation of the user, s_t is the interest representation from the time series graph, h_u is the structural representation of the social graph, γ is the modal weight coefficient dynamically generated by the gating function, the symbol || represents the concatenation operation, and W_g and b_g are trainable parameters.

Finally, we input the unified user representation z_u and target item representation e_i into the matching function to generate recommendation preference predictions. The matching function can be in the form of vector dot product, MLP or other forms. This paper uses the following form to model the scoring function:

$$y_{ui} = f(z_u, e_i) = z_u^T e_i \tag{5}$$

Through the above multi-layer nested feature extraction and modal collaboration mechanism, the MCA module realizes the unified modeling of user behavior temporal information and social structure information, providing semantically consistent and information-complementary user representation for subsequent recommendations.

3.2 Temporal-Aware Node representation

In order to further enhance the model's sensitivity to variations in user behavior, this paper introduces a Temporal-Aware Node representation module (TAN) within the behavioral modality modeling framework. The primary objective of this module is to capture the dynamic evolution of user interests over time. User preferences are not static; they shift in response to contextual changes, personal needs, or social influences. Therefore, modeling temporal dynamics is essential for generating timely and accurate recommendations that reflect users' current intent. TAN is designed to explicitly incorporate these temporal patterns into the learning process, enabling the model to better understand and respond to changes in user behavior.

After constructing a temporal behavior graph for each user, the model goes beyond simply considering the structural connections between behavior nodes. It integrates time as a crucial feature that actively participates in the representation learning process. Each edge and node in the behavior graph is enriched with temporal attributes, such as action timestamps or time intervals, allowing the model to learn not only "what" actions users take, but also "when" they take them. By embedding this temporal information into the node representations, TAN strengthens the model's ability to distinguish between recent and outdated interests. The architecture of the model incorporating TAN is illustrated in Figure 3.

Specifically, for the user's behavior sequence $\{x_1, x_2, \ldots, x_T\}$, we first introduce a time decay factor for each behavior and calculate its time-weighted embedding:

$$x'_i = x_i \cdot \exp(-\lambda \cdot \Delta t_i) \tag{6}$$

Δt_i represents the time difference between the current behavior and the target moment, and λ controls the time sensitivity. Next, we use the weighted attention mechanism to aggregate all historical behaviors and generate a temporal representation of the user in the current context:

$$\alpha_i = \frac{\exp(\varphi(x'_i))}{\sum \exp(\varphi(x'_j))} s_t = \sum_{i=1}^{T} \alpha_i \cdot x'_i \tag{7}$$

$\varphi(\cdot)$ is a learnable attention function, and s_t is the temporal behavior feature, which is used to describe the current user preference trend. This representation not only integrates historical behavior information, but also dynamically adapts to the changes in the importance of behavior at different time scales, thereby improving the flexibility of behavior modeling.

In addition, we further model the propagation of the time series graph based on the graph neural network framework and introduce a time series graph convolution mechanism to enhance the influence of high-order neighbors on the current node representation. The node embedding is updated as follows:

$$h_u^{(l+1)} = \sigma(\sum_{v \in N_u} \frac{1}{c_{\mathrm{uv}}} W^{(l)}(h_v^{(l)}) || \delta_{\mathrm{uv}}) \tag{8}$$

δ_{uv} represents the time difference feature on edge (u, v), which participates in node propagation as an additional dimension. After multiple layers of propagation, we can obtain the final representation s_u containing temporal dynamics, which is used to generate unified embedding in the subsequent modal collaboration module:

$$s_u = \mathrm{TAN}(G_b, \Delta t) = f_{\mathrm{TAN}}(x'_1, x'_2, \ldots, x'_T) \tag{9}$$

Finally, the representation generated by TAN will be input into the MCA module together with the structural embedding obtained from the social graph to achieve a comprehensive modeling of user preferences. This process not only ensures the efficient encoding of temporal information, but also provides a dynamic behavioral basis for modal collaboration.

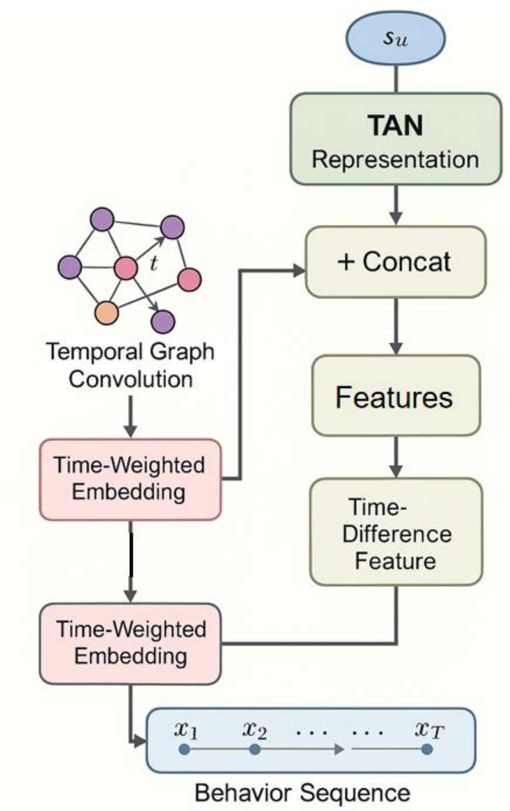

Figure 3. TAN module architecture

4 Experimental Results

4.1 Dataset

The primary dataset used in this study is the Yelp Dataset. This dataset is widely used in research on personalized and social recommendation. It features rich structure and diverse modalities. The data come from a real-world local business platform. It includes user-business interaction records such as ratings, reviews, and bookmarks. Detailed timestamp information is also available, which supports modeling the dynamic evolution of user behavior sequences.

The Yelp Dataset also contains a social network among users. Each user can follow others, forming a directed graph with clear social influence structures. These connections reflect potential mechanisms of interest propagation and behavioral influence. They provide a strong foundation for jointly modeling social preferences and behavioral patterns. In addition, both user and business nodes include multiple attributes, such as category labels and geographic locations. These can serve as auxiliary features to enhance model expressiveness.

In this study, active users and high-frequency interaction samples are selected from the raw dataset. A tri-modal subgraph is constructed, which includes user-business rating behaviors, temporal information, and social relations. The dataset has a clear structure, with complete behavioral and social information. It is broadly representative and provides solid data support for evaluating cross-modal recommendation methods.

4.2 Experimental setup

In the experimental setting, we constructed a user-item interaction graph and a user social relationship graph based on the Yelp dataset, and constructed a user behavior sequence based on timestamps. All user-item interactions are divided into training sets, validation sets, and test sets in chronological order, where the training set is used for model learning, the validation set is used for tuning and early stopping judgment, and the test set is used to evaluate the final performance. To ensure the fairness of the experiment, users and items with less than a certain number of interactions are removed, and the social graph is de-isolated.

We use a variety of mainstream recommendation performance indicators, including Precision@K, Recall@K, and NDCG@K, to evaluate the recommendation effects of different models. The training of all models is completed under the same hardware environment and optimizer settings, and the same initial parameter range, learning rate, and batch size are used for control variable experiments. The specific parameter settings are shown in Table 1.

Table 1. Experimental setup

Parameter	Value
Embedding Dimension	64
Learning Rate	0.001
Batch Size	512
Optimizer	Adam
Dropout Rate	0.2
Epochs	100

Sequence Length	10
Graph Propagation	2 layers
Time Decay Coefficient	0.1

4.3 Experimental Results

4.3.1 Comparative experimental results

First, this paper gives the comparative experimental results with other models. The experimental results are shown in Table 2.

Table 2. Comparative experimental results

Method	Precision@10	Recall@10	NDCG@10	Precision@5	Recall@5	NDCG@5
LightGCN[21]	0.091	0.164	0.123	0.101	0.137	0.116
DGCF[22]	0.095	0.171	0.129	0.106	0.143	0.122
SocialGCN[23]	0.088	0.157	0.118	0.096	0.128	0.110
GraphRec[24]	0.082	0.148	0.113	0.090	0.122	0.106
GRCN[25]	0.097	0.175	0.131	0.109	0.147	0.125
Ours	0.105	0.186	0.142	0.116	0.159	0.134

From the overall results, the proposed method outperforms all baseline models across all evaluation metrics, demonstrating strong recommendation performance. On both Precision, Recall, and NDCG metrics, our method achieves the best results under Top-5 and Top-10 settings. This indicates its advantages in accuracy, coverage, and ranking relevance. In particular, the method reaches a score of 0.142 on NDCG@10, exceeding the second-best model GRCN by more than 0.01. This shows that the model performs more consistently in ranking highly relevant items.

Among the baselines, GRCN and DGCF perform relatively well. GRCN focuses on social information propagation, while DGCF emphasizes channel-wise attention mechanisms. Both reflect the effectiveness of structural modeling. However, these models fall short in capturing temporal dynamics of user behavior or coordinating information across modalities. As a result, they cannot fully model the sequential changes in user interests or the complementary relationship between social and behavioral data. Our method addresses this gap by introducing two key modules: Temporal-Aware Node representation and Modality-Coordinated Aggregation. These modules enable integrated modeling of multimodal information.

Compared with classic models such as LightGCN and SocialGCN, our method improves Top-5 precision by 0.015 and 0.02, respectively. This suggests greater practical value in real-world scenarios where recommendation lists are short and user attention is concentrated. The improvement is not only quantitative but also reflects better ability to capture local interests and provide high-quality recommendations.

In summary, the experimental results strongly validate the effectiveness and advancement of the proposed method in cross-modal graph recommendation tasks. By simultaneously modeling the temporal dynamics of user behavior and the structural information of social graphs, and leveraging graph neural networks for unified representation, the model achieves consistent performance gains across multiple real-world metrics. It demonstrates strong robustness and generalization capabilities.

4.3.2 Ablation Experiment Results

Furthermore, this paper presents the results of ablation experiments to evaluate the individual contributions of key components within the proposed model. By selectively removing or altering specific modules, the ablation study helps to isolate and assess the impact of each design element on the overall recommendation performance. The detailed experimental results are shown in Table 3. This table provides a comprehensive comparison between the full model and its simplified variants, highlighting how each component contributes to improvements in precision, recall, and ranking accuracy. These findings offer strong empirical support for the effectiveness of the proposed model architecture.

Table 3. Ablation Experiment Results

Method	Precision@10	Recall@10	NDCG@10	Precision@5	Recall@5	NDCG@5
Baseline	0.091	0.161	0.120	0.099	0.134	0.112
+MCA	0.097	0.172	0.130	0.107	0.145	0.121
+TCN	0.096	0.169	0.127	0.105	0.142	0.119
Ours	0.105	0.186	0.142	0.116	0.159	0.134

The results of the ablation study show that the two core modules proposed in this paper—Modality-Coordinated Aggregation (MCA) and Temporal-Aware Node representation (TAN, denoted as TCN in the table)—both bring significant improvements in recommendation performance. Compared to the base model, introducing MCA and TAN separately improves Precision@10 by 0.006 and 0.005, respectively. This indicates that both modules effectively enhance the model's ability to capture user interests.

Specifically, the MCA module improves the complementarity between modalities during the fusion of the social graph and the behavior graph. This leads to notable gains in global relevance metrics such as Recall and NDCG. The +MCA variant outperforms the baseline by 0.011 on Recall@10 and by 0.01 on NDCG@10. These results confirm that joint modeling of multimodal information significantly improves ranking performance.

The TAN module focuses more on capturing the temporal dynamics of user behavior. Its inclusion improves the model's performance in modeling short-term interests. In particular, for Top-5 recommendations (Precision@5 and Recall@5), the +TCN variant shows consistent improvements over the baseline. This reflects the module's effectiveness in enhancing the understanding of localized behavioral preferences and improving the precision of high-quality recommendations.

The full model (Ours) outperforms all other variants across all metrics, validating the combined effectiveness of the two modules. Notably, NDCG@10 increases from 0.120 in the base model to 0.142, showing a significant enhancement in ranking accuracy. These results demonstrate that MCA and TAN are not only effective individually, but also produce a synergistic effect when combined, enabling comprehensive modeling of users' multidimensional preferences.

4.3.3 The impact of different embedding dimensions on model performance

Furthermore, this paper also investigates the impact of different embedding dimensions on the performance of the proposed model. Embedding dimension is a critical hyperparameter that affects the model’s ability to represent user and item features effectively. By varying the dimensionality of embeddings, we analyze how the model’s accuracy and generalization ability respond to changes in representational capacity.

The experimental results of this analysis are presented in Figure 4. The figure illustrates how different embedding sizes influence key evaluation metrics, providing insights into the optimal dimensional setting for achieving a balance between expressive power and computational efficiency. This evaluation helps identify the most suitable embedding configuration for practical deployment.

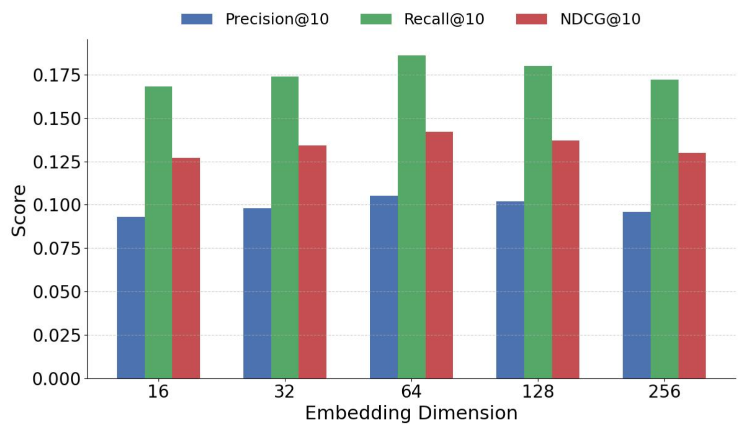

Figure 4. The impact of different embedding dimensions on model performance

The figure shows that the choice of embedding dimension has a clear impact on model performance. This is especially evident in Recall@10 and NDCG@10, which display a consistent fluctuation trend as the dimension varies. When the embedding dimension increases from 16 to 64, the overall performance improves steadily. This suggests that the model's expressive power is limited under low-dimensional embeddings, while moderate dimensions are more effective in capturing complex user-item relationships.

At 64 dimensions, the model achieves its highest scores on Precision@10, Recall@10, and NDCG@10. This indicates that 64-dimensional embeddings strike a good balance between information retention and model generalization. The most significant improvement appears in Recall@10, suggesting that the model best captures users' potential preferences at this setting.

Further increasing the dimension to 128 and 256 leads to slight declines in performance. This implies that excessively high dimensions may introduce redundant information, resulting in overfitting or difficulty in learning useful features. The trend also reflects the sparsity issue in high-dimensional spaces, where the model becomes more sensitive to noise, reducing stability and ranking quality.

In summary, controlling the embedding dimension is critical for enhancing model performance. Under the current experimental settings, 64-dimensional embeddings achieve the best trade-off between accuracy and relevance. This provides empirical support for parameter tuning and demonstrates the model's adaptability to changes in representational capacity.

4.3.4 Performance comparison on cold start users

This paper also gives the experimental results of comparing the performance of the model on cold start users, as shown in Figure 5.

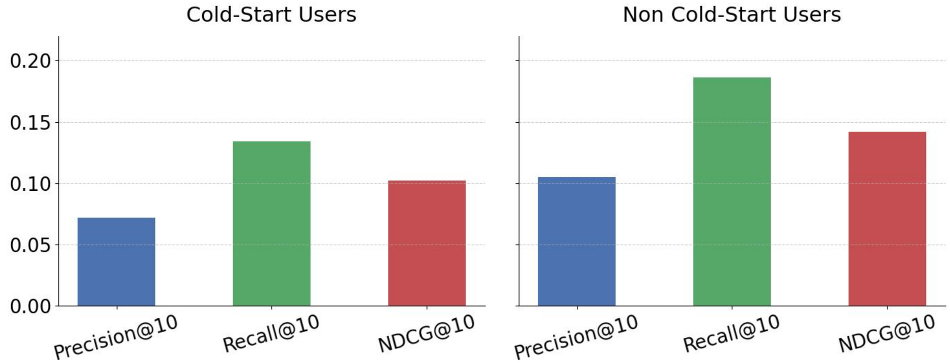

Figure 5. Performance comparison on cold start users

The figure clearly shows that the model performs significantly better on non-cold-start users across all metrics. The gap is especially large in Recall@10 and NDCG@10. This indicates that, for users with sufficient historical behavior data, the model can better learn their latent preferences and generate more accurate and relevant recommendations.

For cold-start users, although historical interaction data is limited, the model still maintains relatively stable outputs on all three evaluation metrics. Its performance on Recall@10 is particularly notable. This suggests that the graph neural framework can partially compensate for the lack of behavior data by leveraging social relations or user adjacency structures, showing a certain degree of cold-start adaptability.

It is worth noting that Precision@10 is the lowest among the three metrics for cold-start users. However, its relative performance still indicates a certain level of recommendation accuracy. This may result from the model's temporal modeling and modality fusion mechanisms. These allow the system to capture local interest dynamics and deliver high-quality recommendations, even in the absence of personal behavior history.

Overall, the experimental results demonstrate the model's potential in addressing the cold-start problem. They also support the rationale for introducing a multimodal graph structure. Although performance gaps remain, the current framework effectively mitigates the degradation caused by cold-start scenarios and provides a promising direction for further optimization.

4.3.5 Sensitivity analysis of the relative weights of behavioral and social graphs

This paper also presents the experimental results of sensitivity analysis of the relative weights of behavioral graphs and social graphs, as shown in Figure 6.

The figure shows that as the weight of the behavior graph increases, the three evaluation metrics exhibit different trends. This reflects the complementary but non-equivalent roles of behavioral and social information in recommendation modeling. Precision@10 reaches its peak when the behavior graph weight is 0.5. This suggests that under a relatively balanced contribution of behavior and social information, the model can more accurately identify users' core interests and produce more precise recommendations.

Recall@10 increases monotonically with the weight of the behavior graph and reaches its highest value when the behavior graph fully dominates (weight = 1.0). This indicates that behavioral data has stronger coverage capability for latent user interests. When behavior data is rich, the model can capture more items of interest, enabling broader recommendation coverage. This trend also confirms that behavior graphs provide more direct guidance in modeling user preferences.

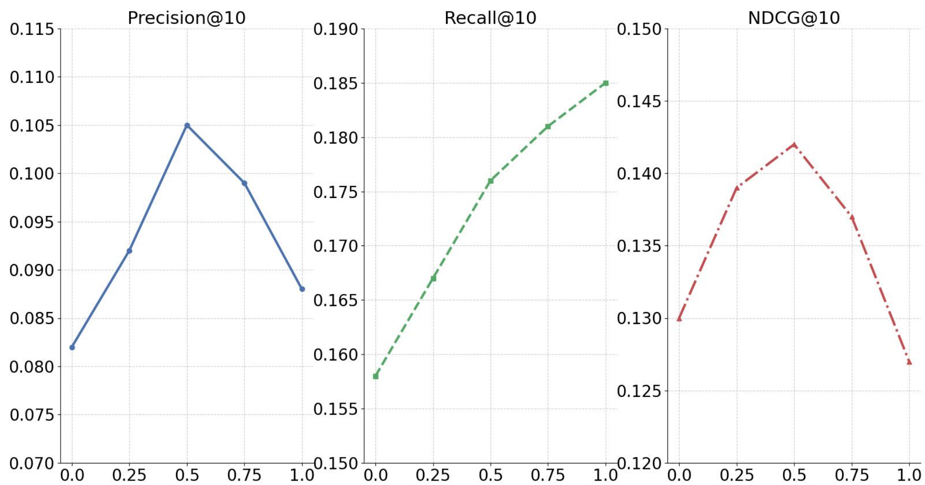

Figure 6. Sensitivity analysis of the relative weights of behavioral and social graphs

In contrast, NDCG@10 reaches its maximum when the behavior graph weight is 0.5 and then begins to decline. This suggests that over-reliance on behavior graphs may introduce redundant or noisy information, which harms ranking relevance. This phenomenon reveals an important insight: excessive behavioral signals, when not properly filtered or modeled, may lead to decreased ranking precision.

In summary, this sensitivity analysis shows that optimal performance is usually achieved when the behavior and social graphs contribute in a balanced manner, or when behavior information holds a slight dominance. This observation suggests that both modalities provide complementary but not interchangeable information. Behavior data offers direct insight into user preferences, while social information captures indirect influence and broader patterns. A balanced integration of the two allows the model to take advantage of both detailed individual behaviors and contextual social signals.

Properly regulating the proportion of these two modalities is therefore critical to improving overall recommendation performance. It ensures that neither source overwhelms the other, avoiding issues such as redundancy or loss of relevant signals. This insight provides a valuable foundation for future research on dynamic weight allocation mechanisms. Such mechanisms could adaptively adjust the contribution of each modality based on user characteristics or data availability, further enhancing the model's flexibility and effectiveness in real-world applications.

4.3.6 Loss function changes with epoch

Finally, this paper also provides a graph illustrating the changes in the loss function over training epochs. This analysis is conducted to examine the convergence behavior and training stability of the proposed model. Monitoring the loss values across epochs helps to evaluate whether the model is learning effectively and whether overfitting or underfitting occurs during the training process.

The detailed visualization of this trend is shown in Figure 7. The figure presents both the training loss and validation loss curves, allowing for a clear comparison of performance on seen and unseen data. These results demonstrate the efficiency of the optimization process and the model's ability to maintain generalization throughout training.

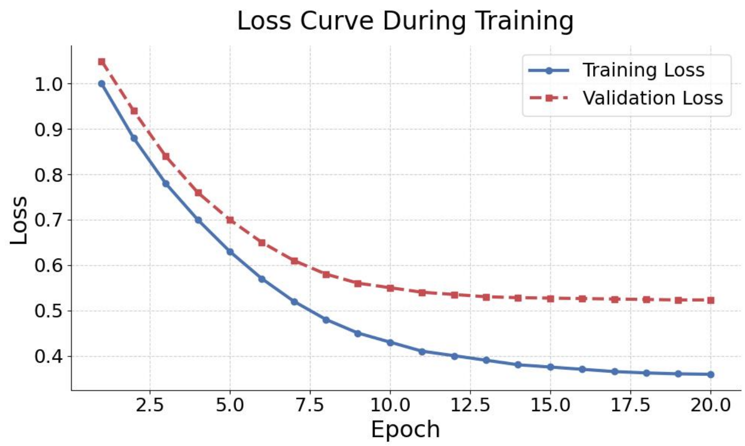

Figure 7. Loss function changes with epoch

The figure shows that both training loss and validation loss steadily decrease as the number of training epochs increases. This indicates that the overall optimization process is effective. In the early epochs, both loss curves drop rapidly. This suggests that the model quickly captures key patterns in the data and achieves significant learning progress in a short period.

The training loss continues to decrease smoothly and reaches approximately 0.36 without fluctuations or anomalies. This shows that the training process is stable and the model fits the training data well. In contrast, the validation loss levels off at around 0.52 after an initial decline. This suggests that the model reaches a stable generalization performance on the validation set, with no signs of overfitting.

The validation loss remains slightly higher than the training loss in the later stages, which is expected. Performance on the validation set typically reflects the model's ability to generalize to unseen data. More importantly, there is no significant divergence between the two curves. This confirms that the model maintains good generalization while improving training performance.

In summary, the loss curves demonstrate that the proposed model shows good convergence and stability during training. They also indicate consistent predictive performance on unseen data, providing a reliable foundation for the subsequent experiments.

5 Conclusion

This paper proposes a cross-modal graph neural recommendation algorithm that unifies the modeling of user behavior sequences and social network information. By designing the Modality-Coordinated Aggregation (MCA) mechanism and the Temporal-Aware Node (TAN) representation module, the model enables integrated learning of users' multimodal preference features. The method effectively combines interest evolution driven by user behavior with latent influence mechanisms embedded in social relations. It establishes a graph neural framework with unified structure and semantic alignment, and demonstrates superior performance across multiple evaluation metrics.

Extensive experiments on real-world datasets show that the proposed method outperforms several existing baseline models in terms of precision, recall, and ranking relevance. It exhibits strong generalization and robustness. In particular, the model shows clear advantages in challenging scenarios such as cold-start user modeling, multimodal information integration, and dynamic graph structure learning. These results further confirm the method's effectiveness and practicality in complex user modeling tasks.

This study not only advances the integration of graph neural networks and multimodal recommendation systems at the algorithmic level, but also provides theoretical support and methodological insights for real-world applications such as personalized recommendation, social content delivery, and interest prediction. As user behavior data and social network structures become increasingly complex, such fusion-based modeling approaches will play a growing role in intelligent recommendation systems.

Future work can extend this approach from multiple perspectives. For example, incorporating additional modalities such as text, images, and geographic information can broaden the model's perception. Leveraging techniques such as graph contrastive learning or reinforcement learning may enhance the long-term optimization of recommendation strategies. Moreover, addressing practical needs for interpretability and real-time performance in deployed systems by designing more transparent and efficient cross-modal recommendation models will be an important direction for further exploration.

References

[1] Cao, Xianshuai, et al. "Cross-modal knowledge graph contrastive learning for machine learning method recommendation." Proceedings of the 30th ACM international conference on multimedia. 2022.

[2] Chen, Xi, et al. "Cmbf: Cross-modal-based fusion recommendation algorithm." Sensors 21.16 (2021): 5275.

[3] Xu, Xiaolong, et al. "Cmclrec: Cross-modal contrastive learning for user cold-start sequential recommendation." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

[4] Liu, Zhongzhou, et al. "Collaborative Cross-modal Fusion with Large Language Model for Recommendation." Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. 2024.

[5] Ma, Haokai, et al. "Cross-modal content inference and feature enrichment for cold-start recommendation." 2023 International Joint Conference on Neural Networks (IJCNN). IEEE, 2023.

[6] Wu, Xinglong, et al. "Towards Bridging the Cross-modal Semantic Gap for Multi-modal Recommendation." arXiv preprint arXiv:2407.05420 (2024).

[7] Wei, Yibiao, et al. "Multi-level cross-modal contrastive learning for review-aware recommendation." Expert Systems with Applications 247 (2024): 123341.

[8] Prétet, Laure, Gael Richard, and Geoffroy Peeters. "Cross-modal music-video recommendation: A study of design choices." 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021.

[9] Won, Heesoo, et al. "Cross-modal contrastive learning for aspect-based recommendation." Information Fusion 99 (2023): 101858.

[10] Ding, Hao, Zhonghua Xia, and Weiwei Zhu. "Research on medical publication recommendation based on cross-modal semantic alignment with contrastive learning." Information Processing & Management 62.4 (2025): 104144.

[11] Corso, Gabriele, et al. "Graph neural networks." Nature Reviews Methods Primers 4.1 (2024): 17.

[12] Khemani, Bharti, et al. "A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions." Journal of Big Data 11.1 (2024): 18.

[13] Zhong, Meihui, et al. "A survey on graph neural networks for intrusion detection systems: methods, trends and challenges." Computers & Security (2024): 103821.

[14] Fujita, Takaaki. "Superhypergraph neural networks and plithogenic graph neural networks: Theoretical foundations." arXiv preprint arXiv:2412.01176 (2024).

[15] Huang, Zhenhua, et al. "Graph Relearn Network: Reducing performance variance and improving prediction accuracy of graph neural networks." Knowledge-Based Systems 301 (2024): 112311.

[16] Pasham, Sai Dikshit. "Scalable Graph-Based Algorithms for Real-Time Analysis of Big Data in Social Networks." The Metascience 2.1 (2024): 92-129.

[17] Ma, Gang-Feng, et al. "Robust social recommendation based on contrastive learning and dual-stage graph neural network." Neurocomputing 584 (2024): 127597.

[18] Sun, Qiang, et al. "A novel recommendation algorithm integrates resource allocation and resource transfer in weighted bipartite network." Big Data Mining and Analytics 7.2 (2024): 357-370.

[19] Wang, Bixi, et al. "Stacked noise reduction auto encoder–OCEAN: a novel personalized recommendation model enhanced." Systems 12.6 (2024): 188.

[20] Gan, Yunxiang, and Diwei Zhu. "The Research on Intelligent News Advertisement Recommendation Algorithm Based on Prompt Learning in End-to-End Large Language Model Architecture." Innovations in Applied Engineering and Technology (2024): 1-19.

[21] He, Xiangnan, et al. "Lightgcn: Simplifying and powering graph convolution network for recommendation." Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. 2020.

[22] Wang, Xiang, et al. "Disentangled graph collaborative filtering." Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval. 2020.

[23] Wu, Le, et al. "Socialgcn: An efficient graph convolutional network based model for social recommendation." arXiv preprint arXiv:1811.02815 (2018).

[24] Lee, Jeong-Wook, and Jae-Hoon Kim. "GraphRec-based Korean expert recommendation using author contribution index and the paper abstracts in marine." Engineering Applications of Artificial Intelligence 133 (2024): 108219.

[25] Wei, Yinwei, et al. "Graph-refined convolutional network for multimedia recommendation with implicit feedback." Proceedings of the 28th ACM international conference on multimedia. 2020.

[26]Ke, Y. "Heterogeneity-Aware Federated Optimization with Secure Aggregation for Distributed Data Mining." (2024).

[27]Zhang, J. "Learning What to Remember: Key Signal-Aligned Memory Compression for Long-Context LLMs." (2024).

[28]Chen, B. "Adaptive Fusion of Low-Rank and Soft-Gated Sparse Updates for Parameter-Efficient Instruction Tuning." (2024).

[29]Tang, Y. "Time-Varying Operational Graph Reasoning and Risk-Constrained Repair Planning for Tool-Augmented Large Language Model Agents." (2024).

[30]Ma, W., and W. Zhong. "Keyword-Constrained Retrieval-Augmented Generation for Large Language Models." (2025).

[31]Jiang, R., and Y. Shi. "SCDF-Net: Research on a Multi-Source Feature Fusion Binary Classification Prediction Network Algorithm for Supply Chain Delay Delivery Risk." Journal of Computer Technology and Software 4.7 (2025).

[32]Mei, Y. "Unified Semi-Supervised and Robust Optimization Framework for Low-Resource Noisy Data Mining." (2024).

[33]Liu, X., and K. Gong. "Predictive GPU Memory Modeling in Containerized Cloud Environments Using Self-Supervised Multi-Scale Learning." (2025).

[34]Shawulieti, S. "Enhancing Cloud Security with Graph Neural Network-Driven Anomaly Detection in Multi-Tenant Environments." (2024).

[35]Xu, S. "Efficient Anomaly Detection in Distributed Edge-Cloud Systems Using Lightweight Modeling." Artificial Intelligence and Computing Innovations 4.3 (2024).