Multimodal Diffusion With Attention Self-Supervised Learning for Recommendation

Shuhui Han, Dan Yang, Xi Gong

Abstract—**Effectively integrating heterogeneous modal information and alleviating performance bottlenecks caused by sparse interactions remain core challenges in multimodal recommendation systems. To address the limitations of existing methods in modal fusion, feature redundancy, and computational overhead, this paper proposes a novel Diffusion-based Attention Self-Supervised Learning Recommendation algorithm (DASRec). The method introduces a dynamic sparse generation process through diffusion models, which enhances representation quality under sparse data conditions via forward perturbation and reverse denoising. Simultaneously, it incorporates a modal attention mechanism to learn the importance of different modalities dynamically and designs a modal-aware signal injection strategy to guide the diffusion process in generating semantically consistent interaction graphs. To further enhance modal consistency and cross-modal collaborative representation, DASRec introduces a cross-modal contrastive learning strategy that jointly optimizes alignment constraints between primary and modal perspectives, thereby improving model generalization and robustness. Extensive experiments on two real-world multimodal datasets demonstrate that DASRec significantly outperforms various existing recommendation methods across evaluation metrics, mainly exhibiting superior personalized recommendation performance in high-sparsity scenarios. These results validate its broad applicability and superior performance in multimodal sparse recommendation tasks.**

Index Terms—**Attention Mechanism, Multimodal Fusion, Diffusion Modal, Contrastive Learning, Recommendation**

I. Introduction

With the rapid development of internet technologies, personalized recommendation systems are widely applied across various domains, including e-commerce, short video platforms, music streaming, and social media. These systems recommend suitable content or products based on their historical behaviors, preference patterns, and social relationships, enhancing user experience and commercial benefits. However, as data grows and user demands become increasingly diverse, existing recommendation methods face several challenges, such as data sparsity[1], difficulty in multimodal information fusion[2], and limited interpretability of recommendation results.

Manuscript received May 8, 2025; revised July 10.

Shuhui Han is a postgraduate student at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (e-mail: hsh_yeying@163.com).

Dan Yang is a professor at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (corresponding author to provide email: asyangdan@163.com).

Xi Gong is a lecturer at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (e-mail: askdjy05gx@ 163. com).

In recent years, multimodal recommendation systems[3] have emerged as a research focus. Compared with unimodal methods, multimodal systems integrate multiple sources of information, such as text, images, and audio, to comprehensively understand user preferences and improve recommendation accuracy. For instance, in short, video recommendation systems[4], textual descriptions, cover images, and background music of the videos are used to model user interests. Nonetheless, multimodal recommendation methods still encounter several challenges, including: (1) semantic discrepancies and heterogeneity across different modalities, which hinder effective information fusion; (2) missing or noisy modality data, which degrades recommendation performance; (3) the high dimensionality of multimodal features, which results in sizeable computational overhead and limits scalability. Therefore, effectively leveraging multimodal information while improving computational efficiency and recommendation performance is a critical research direction in current recommender systems.

In multimodal recommendation systems, traditional approaches typically fuse features from different modalities through weighted summation or simple concatenation. However, such approaches fail to adequately consider the relative importance of different modalities, which may lead to interference from irrelevant or redundant information. Moreover, current multimodal recommendation methods usually adopt fixed fusion strategies[5], without adaptively adjusting modality weights according to different users or recommendation scenarios[6], thus limiting model flexibility and generalization capability.

Meanwhile, the data sparsity issue severely limits recommendation systems' performance. In real-world applications, user-item interaction data is often minimal, as most users interact with only a few items. This sparsity makes it difficult for collaborative filtering methods[7] to accurately learn user preferences. To alleviate the problem of data sparsity, self-supervised learning (SSL)[8] is widely applied in recommendation systems in recent years. SSL extracts potential supervisory signals from unlabeled data and generates auxiliary training objectives to improve the generalization ability of models. However, existing self-supervised learning methods still exhibit limitations when applied to multimodal recommendation tasks. For example, some approaches generate self-supervised signals through random data augmentation[9], such as node or edge dropout. However, these augmentation strategies are often designed based on heuristic rules[10], which fail to account for the characteristics of different modalities fully and may introduce irrelevant noise.

This paper proposes a Multimodal Diffusion Attention

With Self-Supervised Recommendation algorithm (DASRec) to address the aforementioned challenges. This method enhances the performance of recommendation systems by incorporating cross-modal attention mechanisms[11] and a dynamic sparse attention mechanism based on diffusion processes[12]. The main contributions of this work are summarized as follows:

- A dynamic sparse attention mechanism based on diffusion models is proposed. A diffusion process is introduced on the user-item interaction graph and combined with a modality attention mechanism to adaptively adjust the weights of multimodal features, thereby improving the accuracy of personalized recommendations.
- A self-supervised multimodal fusion strategy is explored. A cross-modal contrastive learning framework enhances the model's capability to learn from different modalities of features.
- Extensive experiments are conducted to validate the effectiveness of the proposed model. The results show that the model consistently performs better in recommendation tasks, especially in sparse data scenarios, significantly improving recommendation accuracy and personalization.

II. Related Work

This section reviews several important research directions in the field of recommendation algorithms, with a focus on the integration of multimodal diffusion and attention mechanisms, as well as the application of self-supervised learning in recommendation systems.

A. Recommendation Algorithms Based on Multimodal

Multimodal recommendation algorithms have emerged as one of the key research hotspots in recent years. Unlike traditional unimodal recommendation methods, multimodal recommendation systems incorporate various types of information—such as text, images, audio, and video—to enhance the modeling of user interests. Reference[13] proposes a self-supervised multimodal recommendation model that does not require auxiliary graph augmentation or negative sampling. It generates contrastive views via dropout and jointly optimizes user and item representations through interaction reconstruction and modality alignment, thereby improving recommendation performance while reducing computational costs. Reference[14] introduces a self-supervised multimodal graph contrastive learning model. It constructs multiple views by dropping modality-specific edges and applying modality masking, and incorporates a novel negative sampling strategy to enhance multimodal representation learning. This approach better captures user preferences across modalities and improves micro-video recommendation performance and convergence speed. Reference[15] proposes a multimodal variational graph autoencoder model. It uses modality-specific variational encoders to learn Gaussian variables for users and items. It applies a product-of-experts strategy to fuse embeddings from different modalities, thereby balancing semantic informativeness and uncertainty.

B. Recommendation Algorithms Based on Diffusion

Diffusion models represent a powerful class of generative modeling techniques that have achieved significant progress in computer vision, natural language processing, and related fields in recent years. The core idea of diffusion models is to learn the underlying data distribution through forward noise injection and reverse denoising reconstruction, enabling the generation of high-quality samples. Reference[16] proposes a diffusion-based recommendation system that learns the generation process of user interactions through denoising. It introduces two task-specific extensions: L-DiffRec performs diffusion in the latent space to reduce the computational cost, while T-DiffRec captures evolving user preferences via time reweighting, enhancing recommendation performance. Related work[17] proposes a deep influence diffusion recommendation model, which simulates the recursive diffusion process of users in social networks through a hierarchical influence propagation structure. This model dynamically updates user embeddings to alleviate data sparsity and improve social recommendation effectiveness.

C. Recommendation Algorithms Based on Self-Supervised Learning

Self-supervised learning (SSL) has recently been widely applied in recommendation systems. The core idea of SSL is to extract supervision signals from the data itself, thereby reducing dependence on manual annotations and improving the robustness and generalization ability of recommendation models. Related work[18] is a commonly used SSL approach. It constructs positive and negative samples to make the representations of similar users/items closer while pushing apart dissimilar representations. SLMRec[19] is a multimedia recommendation model based on SSL. It enhances item representation quality by leveraging multimodal data augmentation and contrastive learning to explore potential relationships across modalities. RGCL[20] is a review-aware recommendation model based on graph contrastive learning. It constructs a user-item graph with enhanced edge features incorporating user-item ratings and review semantics. By performing node- and edge-level contrastive tasks, RGCL provides self-supervised signals to improve the representation learning for users and items.

D. Recommendation Algorithms Based on Attention Mechanism

The attention mechanism is a dynamic weighting strategy that estimates input features' importance, thereby enhancing models' expressive power. Modality attention mechanisms dynamically adjust the weights of different modalities, allowing the model to integrate multimodal information more effectively. CRMMAN[21] is a collaborative recommendation model based on a multimodal multi-view attention network. It simultaneously models user preferences and aversions and enriches item representations using semantic and structural information, improving recommendation comprehensiveness and accuracy. MMKDGAT[22] is a remote sensing image recommendation model based on a deep graph attention network aware of multimodal knowledge graphs. It constructs a multimodal knowledge graph to integrate various attributes and visual information of remote sensing images and performs information aggregation through a deep relational attention mechanism. This model achieves strong performance in cold-start scenarios.

III. Preliminaries

This section introduces the key notations and definitions used throughout the paper and briefly overviews the problem. Table I lists the specific meanings of the symbols:

TABLE I Symbol Descriptions

Symbol	Description
$U=\{u_1,u_2,...,u_{\|U\|}\}$	The set of users
$I=\{i_1,i_2,...,i_{\|I\|}\}$	The set of items
$(u,i)\in E$	Indicates an interaction between user u and item i
$G=(U,I,E)$	User-item interaction graph
M	The set of modalities
d_m	Feature dimension of modality m
$\hat{f}_i^m \in \mathbb{R}^{d_m}$	Feature vector of item i under modality m
$G^M=(G,\{F_i \mid i\in I\})$	Multimodal interaction graph
$\hat{y}_{ui}$	Predicted interaction probability between user u and item i

Definition 1. Let U denote the set of users $U=\{u_1,u_2,...,u_{|U|}\}$, and I denote the set of items $I=\{i_1,i_2,...,i_{|I|}\}$. Let $|U|$ and $|I|$ denote the total number of users and items, respectively.

Definition 2. The user-item interaction graph is denoted as $G=(U,I,E)$, where U is the user set, I is the item set, and E is the set of edges. An edge $(u,i)\in E$ indicates that user u has interacted with item i.

Definition 3. For each item i, a multimodal feature vector $\hat{F}_i$ is introduced to incorporate information from different modalities $\hat{F}_i=(\hat{f}_i^1,...,\hat{f}_i^m,...,\hat{f}_i^{|M|})$. Let M represent the set of modalities (e.g., textual, visual, acoustic). For each modality m, $\hat{f}_i^m \in \mathbb{R}^{d_m}$ denotes the feature vector of item i, and d_m is the corresponding feature dimension.

In DASRec, the recommendation algorithm predicts potential interactions between users and items based on the graph structure once the graph is constructed. The prediction function is defined as follows:

$$\hat{y}_{u,i}=f(G^M) \tag{1}$$

where $\hat{y}_{u,i}$ denotes the predicted interaction score between user u and item i.

IV. Algorithm Framework

This section presents a detailed description of the proposed recommendation algorithm, DASRec, which consists of four key components: 1) Dynamic Sparse Diffusion. A diffusion model is employed to simulate the denoising generation process of user-item interactions, enhancing representation learning quality. The diffusion process adopts a dynamic sparsity strategy to denoise high-noise data, improving the recommendation model's robustness. 2) Modality-Attentive Multimodal Aggregation. A modality attention mechanism is utilized to dynamically learn the importance of different modalities, thereby optimizing user and item representations. 3) Cross-Modal Contrastive Learning. Positive and negative sample pairs are constructed to pull similar data closer and push dissimilar data apart, thus enhancing the representation quality of users and items across modalities. 4) Recommendation Prediction. The final prediction is optimized using a loss function incorporating Bayesian Personalized Ranking (BPR)[23], enabling the model to generate personalized recommendation results. The overall framework of DASRec is illustrated in Figure 2.

A. Dynamic Sparse Diffusion

Inspired by the successful application of diffusion models in preserving essential patterns during data generation, this work designs a multimodal graph diffusion module that generates user-item interaction graphs enriched with modality information, enabling more accurate modeling of user preferences.

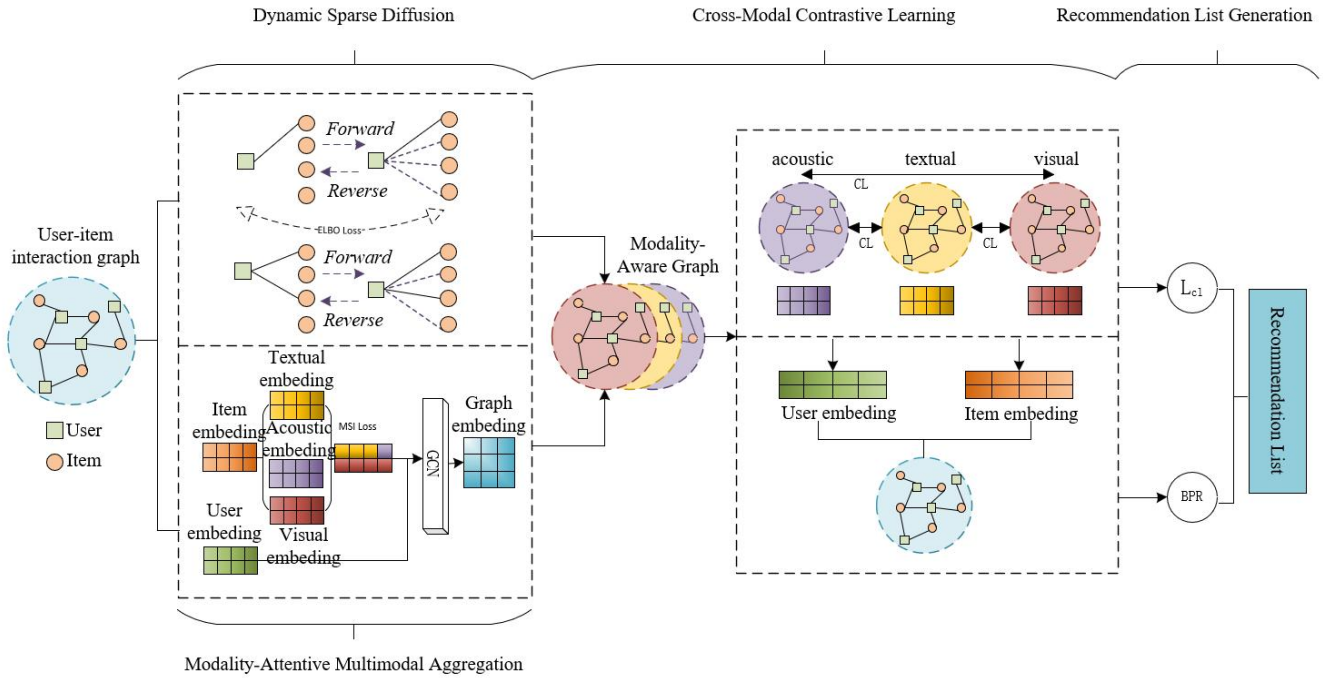

Fig. 1. Overall Framework of the DASRec.

In multimodal recommendation tasks, modality information may contain irrelevant or noisy features, negatively affecting recommendation performance. To address this issue, a modality-aware denoising diffusion probabilistic model is proposed. This model fuses collaborative signals from user-item interactions with multimodal information. Specifically, the user-item interaction graph is progressively perturbed, and iterative denoising learning is performed to recover the interaction structure. This denoising process effectively incorporates multimodal knowledge while suppressing irrelevant or noisy signals.

In addition, a modality-aware signal injection mechanism is introduced to guide the recovery of interaction relationships. This mechanism provides adequate modality-level support during the reconstruction of user-item interactions, allowing the final generated graph to reflect users' actual preferences better. DASRec presents an efficient and robust solution for multimodal recommendation by combining diffusion modeling with modality-aware signal guidance.

1) Dynamic Forward Diffusion Process

Let the interaction behavior of user u over the item set I represented as $\alpha_u = \left[a_0^u, a_1^u, ..., a_{|I|-1}^u \right]$, where each element $a_i^u \in \{0,1\}$ indicates whether user u has interacted with item i. The forward diffusion process gradually adds Gaussian noise over T time steps, transforming the interaction distribution toward a standard Gaussian form. The transition equation of the diffusion process is defined as:

$$q(\alpha_t \mid \alpha_{t-1}) = N(\alpha_t; \sqrt{1-\beta}\alpha_{t-1}, \beta_t I) \tag{2}$$

Where $\beta_t \in (0,1)$ controls the noise intensity added at each time step.

In this work, user sparsity information is introduced to dynamically regulate the diffusion process to better align with real-world data distribution. The sparsity of user u is defined as:

$$s_u = 1 - \frac{\| a_0^u \|_1}{M} \tag{3}$$

Where $\| a_0^u \|_1$ denotes the number of interactions for user u in the interaction matrix, and M is the maximum possible number of interactions (i.e., the total number of items). The sparsity score s_u ranges from 0 to 1 and reflects the degree of sparsity in user behavior.

This metric quantifies the sparsity level of a user's interactions across all items. A higher value s_u indicates that the user exhibits more sparse interaction behavior. During the forward diffusion process, this sparsity measure is incorporated to adaptively adjust the noise intensity, enabling the model to fit users with varying sparsity levels better. The adjusted noise scheduling function is defined as:

$$q(\alpha_t \mid \alpha_0, s_u) = N(\alpha_t; \sqrt{\bar{\gamma}_t}\alpha_0, (1-\bar{\gamma}_t)s_u I) \tag{4}$$

To regulate the amount of noise added during the diffusion process, two parameters are introduced, $\gamma_t = 1 - \beta_t$ which represent the proportion of original information retained at time step t. $\bar{\gamma}_t = \prod_{t'=1}^{t} \gamma_{t'}$, which accumulates the overall information retention rate from t = 1 to time t.

Based on this, the diffusion data α_t can be reparameterized as:

$$\alpha_t = \sqrt{\bar{\gamma}_t}\alpha_0 + \sqrt{1-\bar{\gamma}_t}\varepsilon, \varepsilon \sim N(0, I) \tag{5}$$

where α_0 denotes the original (clean) data and ε is random noise sampled from a standard normal distribution.

To control the magnitude of noise introduced during diffusion, a linear noise scheduling strategy is adopted, defined as:

$$1 - \bar{\gamma}_t = s_u \cdot (\gamma_{\min} + \frac{t-1}{T-1}(\gamma_{\max} - \gamma_{\min})) \tag{6}$$

where $\gamma_{\min}$ and $\gamma_{\max}$ (both within the range (0, 1)) define the lower and upper noise bounds, respectively.

This scheduling strategy enables effective control over the level of noise added at different diffusion steps, ensuring that the model receives appropriate perturbation throughout the process. As a result, the quality and stability of the generated data are significantly improved.

2) Dynamic Reverse Diffusion Process

DASRec aims to progressively remove the noise introduced α_t during the reverse process and recover the original clean data α_{t-1}, enabling the multimodal diffusion to effectively capture subtle variations in the generative process. The reverse diffusion process starts from the final noisy representation α_T and iteratively reconstructs the user-item interaction information through denoising transformations.

The transition of the reverse process is defined as:

$$p_\theta(\alpha_{t-1} \mid \alpha_t, s_u) = N(\alpha_{t-1}; \mu_\theta(\alpha_t, t, s_u), \sum\nolimits_\theta \alpha_t, t) \tag{7}$$

Where $\mu_\theta(\alpha_t, t)$ and $\sum\nolimits_\theta(\alpha_t, t)$ denote the predicted mean and covariance of the Gaussian distribution, respectively. These parameters are generated by a neural network with learnable parameters θ.

This approach ensures that the diffusion model gradually removes noise during the reverse generation while preserving the critical information in the data, resulting in a more accurate and representative reconstruction of user-item interactions.

3) Diffusion Model Training

The training objective of the diffusion model is to optimize the evidence lower bound (ELBO) to maximize the log-likelihood of the observed user-item interaction α_0, formulated as:

$$L_{elbo} = \mathrm{E}_{q(\alpha_0)}\left[-\log p_\theta(\alpha_0) \right] \le \sum_{t=0}^{T} \mathrm{E}_q \left[L_t \right] \tag{8}$$

The term L_t at different time steps is defined as:

$$L_t = \begin{cases} -\log p_\theta(\alpha_0 \mid \alpha_1), & t = 0 \\ D_{KL}(q(\alpha_T \mid \alpha_0) \| p(\alpha_T)), & t = T \\ D_{KL}(q(\alpha_{t-1} \mid \alpha_t, \alpha_0) \| p(\alpha_{t-1} \mid \alpha_t)), & t \in \{1,2,...,T-1\} \end{cases} \tag{9}$$

Here, L_0 represents the negative reconstruction error α_0. The terms L_t for $t \in \{1,...,T-1\}$ constrain the model to approximate the reverse transition distribution $p_\theta(\alpha_{t-1} \mid \alpha_t)$ to the true posterior $q(\alpha_{t-1} \mid \alpha_t, \alpha_0)$.

To optimize the diffusion process, a neural network is designed for denoising. According to Bayes' theorem, the

posterior $q(\alpha_{t-1}|\alpha_t,\alpha_0)$ has the following closed-form expression:

$$q(\alpha_{t-1}|\alpha_t,\alpha_0)\alpha N(\alpha_{t-1};\widetilde{\mu}(\alpha_t,\alpha_0,t),\sigma^2(t)I) \quad (10)$$

Where:

$$\overline{\mu}(\alpha_t,\alpha_0,t)=\frac{\sqrt{\gamma_t}(1-\overline{\gamma}_{t-1})}{1-\overline{\gamma}_t}\alpha_t+\frac{\sqrt{\overline{\gamma}_{t-1}}(1-\overline{\gamma}_t)}{1-\overline{\gamma}_t}\alpha_0 \quad (11)$$

$$\sigma^2(t)=\frac{(1-\gamma_t)(1-\overline{\gamma}_{t-1})}{1-\overline{\gamma}_t} \quad (12)$$

To simplify computation and enhance training stability, the variance term is set as $\sum_\theta(\alpha_t,t)=\sigma^2(t)I$. The final loss function at time step t is defined as:

$$L_t=\frac{1}{2\sigma^2(t)}\|\mu_\theta(\alpha_t,t)-\widetilde{\mu}(\alpha_t,\alpha_0,t)\|_2^2 \quad (13)$$

The neural network learns the mean $\mu_\theta(\alpha_t,t)$, which is defined as:

$$\mu_\theta(\alpha_t,t)=\frac{\sqrt{\gamma_t}(1-\overline{\gamma}_{t-1})}{1-\overline{\gamma}_t}\alpha_t+\frac{\sqrt{\overline{\gamma}_{t-1}}(1-\gamma_t)}{1-\overline{\gamma}_t}\widehat{\alpha}_\theta(\alpha_t,t) \quad (14)$$

Here, $\widehat{\alpha}_\theta(\alpha_t,t)$ is the network's prediction of α_0 , implemented using a multilayer perceptron (MLP) with inputs α_t and the embedding of time step t. For the initial step L_0 , the loss is computed as:

$$L_0=\|\widehat{\alpha}_\theta(\alpha_1,1)-\alpha_0\|_2^2 \quad (15)$$

In practice, the time step t is uniformly sampled to reduce the computational cost $\{1,2,...,T\}$. The final training loss is expressed as:

$$L_{elbo}=\mathrm{E}_{t\sim U(1,T)}\mathrm{E}_{q(\alpha_0)}\left[\|\widehat{\alpha}_\theta(\alpha_t,t)-\alpha_0\|_2^2\right] \quad (16)$$

B. Modality-Attentive Multimodal Aggregation

A modality attention mechanism is designed to perform multimodal aggregation to effectively integrate multimodal semantic features and guide the diffusion process in generating modality-aware user-item graph structures. This mechanism injects modality-aware signals into the aggregated information to extract user preferences under different modalities and enhance the semantic alignment of the generated graph structure.

1) Modality Attention Aggregation

For each modality $m\in\{1,...,M\}$, the predicted user-item interaction probability $\widehat{\alpha}_0\in\mathbb{R}^{|U|\times|I|}$ is first used to perform weighted aggregation on the modality-specific item features $e_i^m\in\mathbb{R}^{d_m}$, yielding the latent preference representation of user u under modality m:

$$z_u^m=\sum_{i\in I}\widehat{\alpha}_{0,ui}\cdot e_i^m \quad (17)$$

Next, each modality-specific representation is fed into a modality attention network to learn the contribution weight of each modality to the user's preferences:

$$a_u^m=Soft\max(w_2^T\cdot\sigma(W_1z_u^m+b_1)) \quad (18)$$

Finally, a weighted fusion strategy is adopted to aggregate the user preferences from all modalities and obtain the final modality-aware user representation z_u :

$$z_u=\sum_{m=1}^{M}a_u^m\cdot z_u^m \quad (19)$$

2) Modality-Aware Signal Injection Mechanism (MSI)

To enhance the diffusion module's ability to model multimodal semantics in the construction of user-item graph structures, a Modality-aware Signal Injection (MSI) mechanism is proposed. MSI aims to guide the model in generating user-item interaction graphs that are semantically aligned across modalities. After modality attention aggregation, a semantic path is constructed based on the aggregated modality-aware representations. Simultaneously, based on the observed binary interaction matrix α_0 , an aggregation of item ID embeddings e_i is performed to construct a structure-aware path. By minimizing the mean squared error (MSE) between these two types of paths, the model is encouraged to generate modality-aware interaction graphs semantically consistent with the proper interaction structure. The loss function is defined as:

$$L_{msi}^m=\|\widehat{\alpha}_0\cdot e_i^m-\alpha_0\cdot e_i\|_2^2 \quad (20)$$

where e_i^m denotes the item feature under modality m, $\widehat{\alpha}_0$ is the predicted interaction probability from the diffusion model, and α_0 is the observed binary interaction matrix.

C. Cross-Modal Contrastive Learning

A contrastive learning-based cross-modal alignment mechanism is proposed to fully exploit the commonalities in user behavior across different modality feature spaces and improve cross-modal representations' consistency and generalization. This mechanism performs dual-directional alignment from both the modality and main view perspectives by constructing positive and negative sample pairs and maximizes semantic consistency between different modalities.

Based on the modality-aware user-item graph G^m constructed in the previous module, a graph neural network (GNN) is applied to perform structured feature modeling. For each modality $m\in M$, the original feature vectors $\widehat{f}^m\in\mathbb{R}^{d_m}$ are aligned in dimensionality, mapped to a shared embedding space via a single-layer multilayer perceptron (MLP), and normalized to obtain modality-aligned item feature representations:

$$e_i^m=Norm(Trans(\widehat{f}^m)),\quad m\in M \quad (21)$$

$Trans(\cdot)$ denotes a nonlinear transformation that maps inputs from $\mathbb{R}^{d_m}$ to $\mathbb{R}^d$, and $Norm(\cdot)$ represents feature normalization. Subsequently, one-step graph neighborhood aggregation is performed on the modality-aware graph $G^m\in\mathbb{R}^{|U|\times|I|}$ to derive structural representations for users and items under modality m:

$$z_u^m=\overline{G}_{u,*}^m\cdot E_u,\quad z_i^m=\overline{G}_{*,i}^m\cdot E_i^m \quad (22)$$

Here, $E_i^m\in\mathbb{R}^{|I|\times d}$ denotes user embeddings, $E_i^m\in\mathbb{R}^{|I|\times d}$ represents modality-aligned item features, and $\overline{G}^m$ is the symmetrically normalized adjacency matrix of G^m . The normalization is defined as:

$$\overline{G}_{u,i}^{m} = \frac{G_{u,i}^{m}}{\sqrt{|N_u^m||N_i^m|}} \tag{23}$$

Where N_u^m, N_i^m denote the neighbor sets of user u and item i in the modality-specific graph G^m, respectively. To further exploit cross-modal structural synergies in the original interaction graph, multilayer message propagation is applied to the original user-item interaction graph G, yielding higher-order semantic-enhanced representations:

$$Z_m^{(l+1)} = \overline{G} \cdot Z_m^{(l)}, Z_m^{(0)} = Z_m \tag{24}$$

Here, $\overline{G}$ is the normalized adjacency matrix of the interaction graph. Finally, the modality-aware representations for contrastive learning are obtained by aggregating all layer-wise outputs via summation:

$$\overline{Z}_m = \sum_{l=0}^{L} Z_m^{(l)} \tag{25}$$

1) Modality Perspective as Anchor

Given any two distinct modalities $m_1, m_2 \in M$ (e.g., text, image, audio), the feature embeddings from these modalities are treated as anchors for contrastive learning. For a given user, the representations are derived from the two modalities $z_u^{m_1}$ and $z_u^{m_2}$ are considered positive sample pairs, while representations of other users under either modality form antagonistic sample pairs. By maximizing the similarity between positive samples and minimizing the similarity between negative samples, the model is guided to learn the latent consistency of user representations across different modalities. Specifically, the InfoNCE loss is adopted and defined as:

$$L_{cl}^{user} = \sum_{\substack{m_1, m_2 \in M \\ m_1 \neq m_2}} \sum_{u \in U} -\log \frac{\exp(s(z_u^{m_1}, z_u^{m_2}) / \tau)}{\sum_{v \in U} \exp(s(\overline{z}_u^{m_1}, \overline{z}_v^{m_2}) / \tau)} \tag{26}$$

where $s(\cdot,\cdot)$ denotes a similarity function (e.g., cosine similarity), and τ is a temperature coefficient that controls the smoothness of the distribution.

2) Main View as Anchor

Beyond enforcing consistency across modalities, the final user/item representations $\hat{h}_u$ generated by the main recommendation task are also used as anchors to align the modality-specific feature views. This alignment enhances the consistency between the main task and the multimodal features. Specifically, the InfoNCE loss on the user side is defined as:

$$L_{cl}^{user} = \sum_{m \in M} \sum_{u \in U} -\log \frac{\exp(s(\hat{h}_u, z_u^m) / \tau)}{\sum_{v \in U} \exp(s(\hat{h}_u, z_u^m) / \tau)} \tag{27}$$

Similarly, the item representation $\hat{h}_i$ from the main view is also used as an anchor to compute the consistency between the primary representation and the modality-specific item embeddings z_i^m, resulting in a corresponding contrastive loss L_{cl}^{item}.

To comprehensively enhance the cooperative learning of cross-modal representations, the user-side and item-side contrastive losses are combined to form the final cross-modal contrastive learning objective, defined as:

$$L_{cl} = L_{cl}^{user} + L_{cl}^{item} \tag{28}$$

3) Cross-Modal Aggregated Representation

To further enhance the contribution of multimodal information to recommendation performance, the final user and item embeddings $\overline{h}_u, \overline{h}_i \in \mathbb{R}^d$ for prediction are generated based on their structure-aware representations from all modalities.

First, the raw features $\hat{f}^m$ of each modality are aligned in dimensionality using an MLP mapping function, yielding modality-aligned feature representations e_i^m. Subsequently, multilayer graph aggregation is performed on both the original user-item interaction graph $\overline{G}$ and the modality-aware graph $\overline{G}^m$, generating structural representations $\hat{z}_u^m, \hat{z}_i^m$ for users and items under each modality:

$$\begin{aligned} \hat{z}_u^m &= \overline{A}_{u,*} \cdot E_u + \overline{A}_{u,*} \cdot (\overline{A}_{u,*} \cdot E_u) + \overline{A}_{*,u} \cdot E_u \\ \hat{z}_i^m &= \overline{A}_{i,*} \cdot E_i + \overline{A}_{i,*} \cdot (\overline{A}_{i,*} \cdot E_i) + \overline{A}_{*,i} \cdot E_i \end{aligned} \tag{29}$$

This design integrates first- and second-order connectivity from the original interaction graph while incorporating neighbor information from the modality-aware graph, thereby preserving cross-modal collaborative signals.

After obtaining individual structural representations for all modalities, a weighted aggregation is applied to generate the final fused multimodal representations. To account for the varying importance of modalities, a learnable modality weight vector k_m is introduced as a regulator, performing a weighted summation:

$$h_u = \sum_{m \in M} k_m \hat{z}_u^m, \quad h_i = \sum_{m \in M} k_m \hat{z}_i^m \tag{30}$$

To further enhance the modeling of high-order collaborative relationships, multilayer graph neural network (GNN) propagation is applied to the original interaction graph $\overline{G}$:

$$H^{(l+1)} = \overline{G} \cdot H^{(l)}, \quad H^{(0)} = h_u \text{ or } h_i \tag{31}$$

Where $H^{(l)}$ denotes the embeddings at the l-th layer, and the GNN is stacked for L layers. Normalized initial embeddings are incorporated as residual terms to mitigate over-smoothing caused by excessive message passing. The final user or item representations are generated by summing all layer-wise embeddings with weighted residuals:

$$\overline{H} = \sum_{l=0}^{L} H^{(l)} + \omega \cdot Norm(H^{(0)}) \tag{32}$$

ω is a hyperparameter controlling the residual ratio.

D. Recommendation List Generation

After learning the final user representation h_u and item representation h_i, the predicted interaction score between user uu and item i is computed via their inner product:

$$\hat{y}_{ui} = h_u^T \cdot h_i \tag{33}$$

The Bayesian Personalized Ranking (BPR) loss is adopted as the primary optimization objective for recommendation:

$$L_{bpr} = \sum_{(u,i,j) \in O} -\log \sigma(\hat{y}_{ui} - \hat{y}_{uj}) \tag{34}$$

To integrate cross-modal contrastive learning objectives, the total loss function is jointly optimized as follows:

$$L_{rec} = L_{bpr} + \lambda_1 L_{cl} + \lambda_2 \| \Theta \|_2^2 \quad (35)$$

Here, Θ represents the set of trainable parameters, λ_1, λ_2 adjust the relative importance between the contrastive loss and the regularization term.

V. Experiments

This section describes the datasets, evaluation metrics, baseline comparisons, and ablation studies conducted to validate the proposed method. Detailed analyses of experimental results are provided.

A. Experimental Setup

1) Datasets

The model is evaluated on two widely used real-world datasets, TikTok and Amazon-Baby, to assess recommendation algorithms, covering user rating behaviors across diverse items. These datasets represent two common multimodal domains: short-video recommendations and e-commerce. TikTok Dataset， derived from the popular short-video platform TikTok, includes user interaction data such as video views, likes, and shares. Amazon-Baby Dataset focuses on maternal and infant products (e.g., bottles, strollers, toys). It contains user reviews, ratings, and multimodal features (e.g., textual, visual, and acoustic) of related products. Detailed statistics for both datasets are summarized in Table II.

Specifically, the letter T denotes the textual modality, V represents the visual modality, and A stands for the acoustic modality. The TikTok dataset includes all three modalities (T, V, A), whereas the Amazon-baby dataset contains only two modalities (T and V).

The sparsity of each dataset is calculated as:

$$Sparsity = 1 - \frac{\text{int}eractions}{users \times items} \quad (36)$$

TABLE II Statistical Information Of The Datasets

Datasets	TikTok	Amazon-baby
Modality Embed Dim	T V A 768 128 128	T V 1024 4096
Users	9319	19445
Items	6710	7050
Interaction	59541	139110
Sparsity	99.904%	99.899%

2) Evaluation Metrics

Three widely-used metrics are adopted to evaluate recommendation performance: Recall@k, Normalized Discounted Cumulative Gain (NDCG@k), and Precision@k, where @k denotes the ranking position, typically evaluating the top-k recommended items (in this work, k=20).

- Recall@k. Measures the overlap between recommended items and users' truly relevant items, reflecting recommendation comprehensiveness. It is computed as:

$$\text{Re}call@k = \frac{\left| \left\{ i \in \tau_u^{rel} : i \in \hat{\tau}_u^{(k)} \right\} \right|}{\left| \tau_u^{rel} \right|} \quad (37)$$

Where τ_u^{rel} represents the set of all relevant items for user u, and $\hat{\tau}_u^{(k)}$ denotes the top-k items recommended to u.

- NDCG@k. Evaluate ranking quality by considering the recommended items' relevance and positional importance. The formula is:

$$NDCG@k = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{\log_2 (p_i + 1)} \quad (38)$$

where i_j indicates the position of the i-th relevant item in the recommendation list. If the i-th relevant item is not in the top-k list, the term $\frac{1}{\log_2(i_j + 1)}$ defaults to 0.

- Precision@k. A standard metric in information retrieval and recommendation systems, measuring the accuracy of the top-k recommendations. It calculates the proportion of truly relevant items among the top-k results:

$$\text{Pr}ecision@K = \frac{\sum_{i=1}^{N} rel(i)}{K} \quad (39)$$

N is the total number of users and $rel(i)$ is an indicator function that equals one if the i-th recommended item is relevant to the user, and zero otherwise.

Three metrics evaluate model performance, targeting the top 20 items in the recommendation list. This paper employs a full-ranking evaluation strategy to calculate the recommendation effectiveness, and the average scores across all users in the test set are aggregated as the final evaluation metrics. Higher metric values indicate superior recommendation performance of the DASRec model.

3) Baselines

The DASRec model is compared with the following 12 baseline methods:

- NGCF[24]: Utilizes multilayer graph convolutional networks (GCNs) to propagate information on user-item interaction graphs, learning latent user and item representations. It emphasizes neighbor information propagation and aggregation to enhance recommendation effectiveness.
- LightGCN[25]: Simplifies GCN propagation by removing nonlinear activations and feature transformations, improving computational efficiency and model performance.
- SGL[26]: Introduces contrastive learning signals through stochastic data augmentation (e.g., node/edge dropout) to enhance graph collaborative filtering.
- NCL[27]: Employs Expectation-Maximization (EM) clustering to identify neighboring nodes and construct contrastive views for generating high-quality positive pairs.
- HCCF[28]. Captures local and global collaborative relationships via hypergraph neural networks and optimizes recommendations through cross-view contrastive learning.
- MMGCN[29]. Propagates modality-specific embeddings using GNNs and models user preferences for different modalities tailored for micro-video recommendation.
- GRCN[30]. Refines interaction data using structurally optimized GCNs, reducing false-positive feedback and noise via pruning.

TABLE III PERFORMANCE COMPARISON OF DASREC WITH BASELINES ON TWO DATASETS

Datasets	Metric	NGCF	LightGCN	SGL	NCL	HCCF	MMGCN	GRCN	LATTICE	CLCRec	SLMRec	BM3	DiffMM	DASRec
LastFM	Precision@20	0.0030	0.0033	0.0030	0.0034	0.0030	0.0037	0.0036	0.0042	0.0035	0.0042	0.0048	0.0056	**0.0057**
	Recall@20	0.0604	0.0653	0.0603	0.0658	0.0670	0.0730	0.0806	0.0842	0.0627	0.0845	0.0957	0.1129	**0.1148**
	NDCG@20	0.0238	0.0282	0.0282	0.0269	0.0267	0.0307	0.0350	0.0369	0.0265	0.0353	0.0403	0.0456	**0.0483**
Amazon-baby	Precision@20	0.0032	0.0037	0.0036	0.0038	0.0037	0.0032	0.0041	0.0044	0.0034	0.0043	0.0044	0.0051	**0.0052**
	Recall@20	0.0591	0.0601	0.0678	0.0705	0.0705	0.0640	0.0754	0.0829	00613	0.0765	0.0839	0.0975	**0.0990**
	NDCG@20	0.0261	0.0261	0.0296	0.0311	0.0308	0.0284	0.0336	0.0367	0.0286	0.0352	0.0361	0.0411	**0.0424**

- BM3[31]. Leveraging self-supervised learning for user-item interaction modeling avoids reliance on randomly sampled negative samples.
- LATTICE[32]. Discovers latent item-item relationships via modality-aware homogeneous item graphs to improve recommendation performance.
- SLMRec[33]. Enhances recommendations through multimodal data augmentation, including feature perturbation and modality-aware pattern recognition.
- DiffMM[34]. Augments user-item interaction representations with diffusion processes and optimizes multimodal fusion via self-supervised learning.
- CLCRec[35]. CLCRec is a contrastive learning-based framework designed for cold-start recommendation. It enhances recommendation accuracy by maximizing the mutual information between item content features and collaborative representations, enabling effective predictions even for items without historical interactions.

4) Parameter Setting

The DASRec model is implemented under the PyTorch framework and updated using the Adam optimizer with Xavier initialization (default parameters) to ensure fair comparisons. During training, the batch size is set to 1024, the embedding dimension is set to 128, and the diffusion steps are set to 5. In the modality attention mechanism, the number of attention heads is set to 4 to dynamically adjust the weights of different modality features, with a default value of 1 (enabled). The dynamic noise intensity coefficients are searched within the range $\{0.1, 0.3, 0.5, 0.7\}$, while the modality similarity threshold is adjusted within the range $\{0.3, 0.4, 0.5, 0.6, 0.7\}$. The same optimization algorithms, parameter initialization methods, and batch sizes as DASRec are adopted for baseline methods.

B. Analysis of Experimental Results

Extensive experiments are conducted on the TikTok and Amazon-Baby datasets, and the results are compared with 11 baseline methods, as shown in Table 3. The experimental results demonstrate the following:

Models such as NGCF and LightGCN adopt graph neural network-based collaborative filtering methods. Specifically, BiasMF optimizes matrix factorization by introducing bias scores, while AutoRec learns embeddings through autoencoder-based interaction reconstruction, validating the effectiveness of these methods in recommendation tasks. Models including SGL, NCL, and HCCF are self-supervised learning-enhanced recommendation frameworks that effectively capture high-order collaborative signals between users and items via distinct message-passing mechanisms. The multimodal recommendation frameworks are GRCN, LATTICE, BM3, SLMRec, DiffMM, CLCRec, and MMGCN.

The results show that the DASRec model consistently outperforms all baselines, significantly improving evaluation metrics. This indicates that cross-modal attention enhancement and the dynamic sparse attention mechanism based on diffusion processes substantially improve model performance. Notably, the sparsity levels of the two datasets are 99.904% and 99.899%, respectively, which benefit more from the dynamic sparse attention mechanism.

C. Model Analysis

This section investigates the effectiveness of individual modules in DASRec and analyzes the impact of key parameters through ablation studies and parameter sensitivity analyses.

1) Ablation Experiments

To validate the contributions of the data augmentation method and attention mechanism, two variants of DASRec are generated by removing specific modules. DASRec-n excludes the dynamic sparse attention mechanism based on the diffusion process. DASRec-a turns off cross-modal attention enhancement to adjust attention weights. Experimental results on the TikTok and Amazon-Baby datasets are summarized in Table IV.

2) Analysis of Hyperparameters

This section discusses the impact of hyperparameters γ and the similarity sim-threshold on the performance of DASRec. The detailed analysis is as follows:

- Impact of γ.

The parameter γ controls the noise intensity in the diffusion model, which directly affects the augmentation strategy for interaction data.

TABLE IV PERFORMANCE COMPARISON OF DASREC WITH OTHER ABLATION METHODS

Datasets	Metrics	DASRec-n	DASRec-a	DASRec
TikTok	Recall@20	0.1131	0.1130	**0.1147**
	NDCG@20	0.0471	0.0473	**0.0479**
	Precision@20	0.0056	**0.0057**	**0.0057**
Amazon-baby	Recall@20	0.0051	**0.0052**	**0.0052**
	NDCG@20	0.0981	**0.0990**	**0.0990**
	Precision@20	0.0415	**0.0426**	0.0424

A moderate noise level enhances the robustness of interaction representations. A low value of γ (e.g., 0.1) may lead to insufficient augmentation, while a high value (e.g., 0.7) may introduce excessive noise, degrading recommendation accuracy. Taking the TikTok dataset as an example, all other hyperparameters are fixed while varying γ within a defined range $\{0.1, 0.3, 0.5, 0.7\}$. The results are shown in Figure 2:

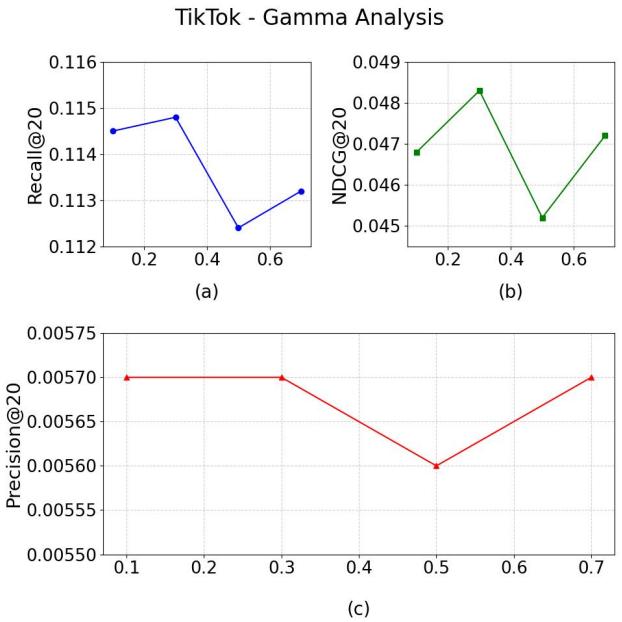

Fig. 2. Performance Comparison of different γ

As illustrated in the figure, the model achieves optimal performance when γ =0.3.

- Impact of the sim-threshold.

The sim-threshold filters the similarity between modalities, ensuring the quality of information fusion. The default value is set to 0.5, meaning that only modality pairs with similarity above 0.5 contribute to the final representation. A lower threshold (e.g., 0.3) may introduce irrelevant information, while a higher threshold (e.g., 0.7) may exclude potentially helpful information. Again, using the TikTok dataset, the sim threshold is tested with values 0.3, 0.4, 0.5, 0.6, and 0.7 while keeping other hyperparameters fixed. The results are shown in Figure 3:

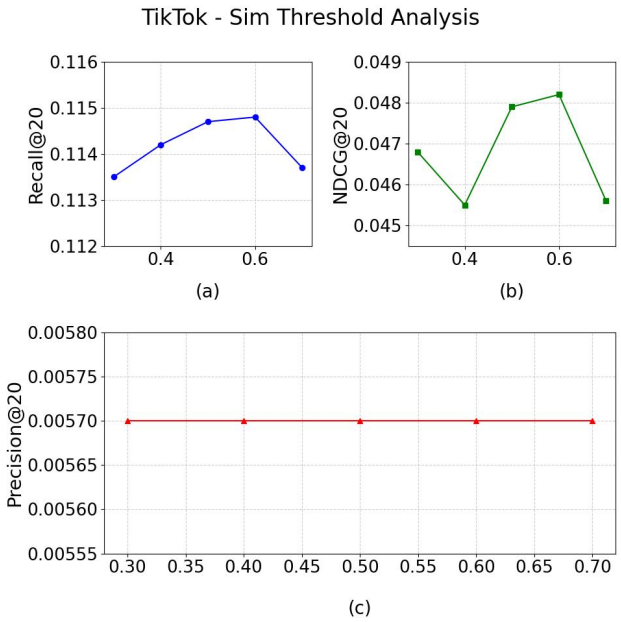

Fig. 3. Performance Comparison of different sim-thresholds

The results show that recommendation performance varies significantly with different sim thresholds. The model performs best when sim-threshold = 0.6, while higher and lower values lead to noticeable performance degradation. Therefore, choosing an appropriate sim-threshold value to achieve optimal results is essential.

VI. CONCLUSION

This paper addresses key challenges in multimodal recommendation systems, including difficulties in modality fusion, severe data sparsity, and lack of representation consistency. A novel recommendation algorithm, DASRec, is proposed, a self-supervised learning framework with attention-enhanced multimodal diffusion. From a generative modeling perspective, DASRec introduces a diffusion process to construct a dynamically sparse, attention-enhanced denoising interaction graph, significantly improving the model's ability to handle sparse interaction data. A modality attention mechanism is designed to model the importance of different modalities, and a modality-aware signal injection mechanism is incorporated to guide the diffusion-based reconstruction, ensuring semantic consistency in the generated multimodal structure. Experimental results demonstrate that DASRec consistently outperforms existing methods on the TikTok and Amazon-baby datasets, confirming its effectiveness and practical value. Future work may extend this framework to real-time recommendations, cold-start scenarios, and more complex multimodal fusion settings, enhancing its adaptability in dynamic environments.

REFERENCES

[1] Yin H, Wang Q, Zheng K, et al., "Overcoming data sparsity in group recommendation," IEEE Trans. Knowl. Data Eng., vol. 34, no. 7, pp. 3447–3460, 2020.
[2] Xiao F, Deng L, Chen J, et al., "From abstract to details: A generative multimodal fusion framework for recommendation," in Proc. 30th ACM Int. Conf. Multimedia, pp. 258–267, 2022.
[3] Mu Y., Wu Y., "Multimodal movie recommendation system using deep learning," Mathematics, vol. 11, no. 4, pp. 1–12, 2023.

[4] Liu Y, Lyu C, Liu Z, et al., "Building effective short video recommendation," in Proc. IEEE Int. Conf. Multimedia & Expo Workshops (ICMEW), pp. 651–656, 2019.
[5] Wei W, Huang C, Xia L, et al., "Multimodal self-supervised learning for recommendation," in Proc. ACM Web Conf., pp. 790–800, 2023.
[6] Xu Y, Zhu L, Cheng Z, et al., "Multimodal discrete collaborative filtering for efficient cold-start recommendation," IEEE Trans. Knowl. Data Eng., vol. 35, no. 1, pp. 741 – 755, 2021.
[7] Wu L, He X, Wang X, et al., "A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation," IEEE Trans. Knowl. Data Eng., vol. 35, no. 5, pp. 4425 – 4445, 2022.
[8] Yu J, Yin H, Xia X, et al., "Self-supervised learning for recommender systems: A survey," IEEE Trans. Knowl. Data Eng., vol. 36, no. 1, pp. 335–355, 2023.
[9] Han Y, Wu L, Wang H, et al., "Guesr: A global unsupervised data-enhancement with bucket-cluster sampling for sequential recommendation," in Proc. Int. Conf. Database Syst. Adv. Appl., Springer, pp. 286 – 296, 2023.
[10] García-Peñalvo F. J., Vázquez-Ingelmo A., García-Holgado A., "Explainable rules and heuristics in AI algorithm recommendation approaches—A systematic literature review and mapping study," CMES-Computer Modeling in Engineering and Sciences, vol. 136, no. 2, pp. 1023–1051, 2023.
[11] Wang R, Li C, Zhao Z, "Towards user-specific multimodal recommendation via cross-modal attention-enhanced graph convolution network," Appl. Intell., vol. 55, no. 1, Art. no. 2, pp:2853-2865, 2025.
[12] Wang S, Sui Y, Wu J, et al., "Dynamic sparse learning: A novel paradigm for efficient recommendation," in Proc. 17th ACM Int. Conf. Web Search Data Min., pp. 740 – 749, 2024.
[13] Zhou X, Zhou H, Liu Y, et al., "Bootstrap latent representations for multimodal recommendation," in Proc. ACM Web Conf., pp. 845–854, 2023.
[14] Yi Z, Wang X, Ounis I, et al., "Multi-modal graph contrastive learning for micro-video recommendation," in Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 1807–1811, 2022.
[15] Wang J, Zeng Z, Wang Y, et al., "Missrec: Pre-training and transferring multi-modal interest-aware sequence representation for recommendation," in Proc. 31st ACM Int. Conf. Multimedia, pp. 6548–6557, 2023.
[16] Wang W, Xu Y, Feng F, et al., "Diffusion recommender model," in Proc. 46th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 832–841, 2023.
[17] Wu L, Sun P, Fu Y, et al., "A neural influence diffusion model for social recommendation," in Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 235–244, 2019.
[18] Jiang Y, Huang C, Huang L, "Adaptive graph contrastive learning for recommendation," in Proc. 29th ACM SIGKDD Conf. Knowl. Discov. Data Min., pp. 4252–4261, 2023.
[19] Tao Z, Liu X, Xia Y, et al., "Self-supervised learning for multimedia recommendation," IEEE Trans. Multimedia, vol. 25, pp. 5107–5116, 2022.
[20] Shuai J, Zhang K, Wu L, et al., "A review-aware graph contrastive learning framework for recommendation," in Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 1283–1293, 2022.
[21] Z. Hu, S. M. Cai, J. Wang, and T. Zhou, "Collaborative recommendation model based on multi-modal multi-view attention network: Movie and literature cases," Applied Soft Computing, vol. 144, Art. no. 110518, 2023.
[22] Wang F, Zhu X, Cheng X, et al., "MMKDGAT: Multi-modal knowledge graph-aware deep graph attention network for remote sensing image recommendation," Expert Syst. Appl., vol. 235, Art. no. 121278, 2024.
[23] Fei Wang, Xianzhang Zhu, Xin Cheng, Yongjun Zhang, and Yansheng Li, "MMKDGAT: Multi-modal knowledge graph-aware deep graph attention network for remote sensing image recommendation," Expert Systems with Applications, vol. 235, Art. no. 121278, 2024.
[24] Wang X, He X, Wang M, Feng F, Chua T-S, "Neural graph collaborative filtering," in Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 165–174, 2019.
[25] He X, Deng K, Wang X, Li Y, Zhang Y, Wang M, "LightGCN: Simplifying and powering graph convolution network for recommendation," in Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 639 – 648, 2020.
[26] Wu J, Wang X, Feng F, et al., "Self-supervised graph learning for recommendation," in Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 726–735, 2021.
[27] Lin Z, Tian C, Hou Y, Zhao W X, "Improving graph collaborative filtering with neighborhood-enriched contrastive learning," in Proc. Web Conf. (WWW), pp. 2320–2329, 2022.
[28] Xia L, Huang C, Xu Y, et al., "Hypergraph contrastive collaborative filtering," in Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 70–79, 2022.
[29] Wei Y, Wang X, Nie L, He X, Hong R, Chua T-S, "MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video," in Proc. ACM Multimedia Conf., pp. 1437–1445, 2019.
[30] Wei Y, Wang X, Nie L, He X, Chua T-S, "Graph-refined convolutional network for multimedia recommendation with implicit feedback," in Proc. ACM Multimedia Conf., pp. 3541–3549, 2020.
[31] Zhou X, Zhou H, Liu Y, et al., "Bootstrap latent representations for multi-modal recommendation," in Proc. Web Conf. (WWW), pp. 845–854, 2023.
[32] Zhang J, Zhu Y, Liu Q, Wu S, Wang S, Wang L, "Mining latent structures for multimedia recommendation," in Proc. ACM Multimedia Conf., pp. 3872 – 3880, 2021.
[33] Z. Tao, X. Liu, Y. Xia, et al., "Self-supervised learning for multimedia recommendation," IEEE Transactions on Multimedia, vol. 25, pp. 5107–5116, 2022.
[34] Jiang Y, Xia L, Wei W, et al., "DiffMM: Multi-modal diffusion model for recommendation," in Proc. 32nd ACM Int. Conf. Multimedia, pp. 7591–7599, 2024.
[35] Y. Wei, X. Wang, Q. Li, L. Nie, Y. Li, X. Li, and T.-S. Chua, "Contrastive learning for cold-start recommendation," in Proc. 29th ACM Int. Conf. Multimedia (MM), pp. 5382 – 5390, 2021.