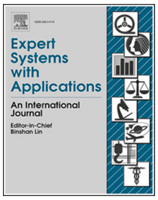
Expert Systems With Applications 320 (2026) 132124

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Conditional graph diffusion with granular-ball representation for multimodal recommendation

Xiaofei Zhu [*], Ling Tan

School of Computer Science and Engineering, Chongqing University of Technology, 400054, Chongqing, China

ARTICLE INFO

Keywords:
Multimodal recommendation
Granular-ball computing
Conditional diffusion model
Graph neural network

ABSTRACT

Multimodal Recommendation Systems (MRSs) aim at enhancing recommendation performance by incorporating multimodal information. Despite recent advances, two critical challenges remain. First, existing graph-based methods mainly focus on local propagation based on the fine-grained interaction graph, overlooking latent associations among distant nodes from a coarse-grained global perspective. Second, existing methods often struggle to mitigate the inherent noise in user historical records, particularly in multimodal scenarios. To tackle these issues, we propose a novel approach named Conditional Graph Diffusion with Granular-Ball Representation (GBDiff). Specifically, we propose to capture coarse-grained global user interest information based on the Granular-Ball Computing technique, which is utilized to complement the fine-grained user-item interaction modeling. In addition, we develop a novel conditional diffusion model which leverage multimodal signals to guide the denoising process. Extensive experiments on two widely used datasets demonstrate that our proposed approach GBDiff consistently outperforms all state-of-the-art baseline methods. The source code of the proposed GBDiff model is available at the following link: https://github.com/dddorren/GBDiff.

1. Introduction

With the rapid proliferation of online multimedia platforms, multimodal recommender systems have become essential for delivering personalized services (Xu et al., 2025). By leveraging visual, acoustic, and textual features (Luo et al., 2025; Wang et al., 2020), these systems model user preferences. However, effectively integrating diverse multimodal information to capture complex user interests remains a significant challenge.

To address multimodal integration in recommender systems, various approaches have been proposed. For instance, VBPR (He & McAuley, 2016) extends matrix factorization by incorporating visual features alongside item IDs to enhance preference modeling. ACF (Chen et al., 2017) employs a hierarchical attention network to capture component-level user preferences. More recently, graph-based methods like MMGCN (Wei et al., 2019) construct modality-aware graphs and utilize graph convolutional networks (GCNs) to optimize user and item representations. LATTICE (Zhang et al., 2021) further explores inter-item relationships through multimodal features, integrating them into GNN message-passing processes for enhanced semantic aggregation. These advancements have inspired efficient models like MICRO (Zhang et al., 2023a) which extends the LATTICE (Zhang et al., 2021) to fuse multimodal features by introducing contrastive learning to capture modality-shared and modality-specific information. Additionally, self-supervised learning (SSL) techniques, such as SGL (Wu et al., 2021) and NCL (Lin et al., 2022), incorporate data augmentation into collaborative filtering.

Despite significant advancements in multimodal recommender systems, several critical challenges persist. Traditional graph-based recommendation models like LightGCN (He et al., 2020) learn user and item representations through multiple graph convolutions based on fine-grained graph structures. While graph convolutions effectively aggregate local information, they often fail to capture latent coarse-grained global associations among distant nodes in the graph structure, thus preventing the model from capturing user interests from a comprehensive perspective. From the human cognition perspective, a “global-first” (Xia et al., 2023a) perception mechanism aligns more closely with cognitive reasoning processes, where macroscopic information is first apprehended before fine-grained judgments are made (Xie et al., 2025; Yang et al., 2025c). Therefore, relying solely on fine-grained modeling is insufficient to fully capture true interests of users, requiring coarse-grained modeling to provide global semantic support.

Moreover, real-world recommendation scenarios frequently encounter noisy user-item interaction data, such as misclicks behaviors or exploratory actions. Such noise interactions weaken true preference signals, particularly in sparse interaction settings, significantly degrading

[*] Corresponding author.
E-mail addresses: zxf@cqut.edu.cn (X. Zhu), tanling@stu.cqut.edu.cn (L. Tan).

https://doi.org/10.1016/j.eswa.2026.132124
Received 21 September 2025; Received in revised form 6 March 2026; Accepted 16 March 2026
Available online 22 March 2026

the quality of learned representations. Recent studies have explored diffusion models to mitigate noise in interactions, leveraging their robust generative capabilities for effective denoising (Jiang et al., 2024a,b). However, these methods often fail to fully utilize multimodal information to generate precise user-item interaction structures, thus limiting denoising effectiveness.

To address these challenges, we propose a unified framework named Conditional Graph Diffusion with Granular-Ball Representation (GBDiff) for multimodal recommendation. Inspired by the human "coarse-to-fine" cognitive mechanism (Li et al., 2025), we introduce granular-ball computing to semantically partition user and item nodes into stable, coarse-grained granular-balls and design rules to construct a bipartite interaction matrix between user and item granular-balls. Through convolutional aggregation, we obtain global user interest information to complement fine-grained user-item interaction modeling. Furthermore, to mitigate noisy interactions between users and items, we incorporate a novel conditional diffusion model that utilizes multimodal signals as condition to guide the reverse diffusion process, generating denoised user-item interaction matrices. Experiments on two publicly available datasets demonstrate the effectiveness of our proposed approach. We further conduct ablation studies to validate the contribution of each main component. In summary, our main contributions are as follows:

- We design a granular-gall based representation learning module to enhance model performance by further modeling coarse-grained global user interest information.
- We propose a multimodal guided conditional graph diffusion module, which leverages multimodal information as conditional guidance for diffusion-based reconstruction of user-item interaction structures.
- Experiments on two widely used datasets verify GBDiff's effectiveness and the contribution of each component.

2. Related work

In this section, we briefly review the latest advancements in related topics, including Multimodal Recommendation, Diffusion Models in Recommendation and Granular-Ball Computing.

2.1. Multimodal recommendation

Multimodal recommender systems utilize diverse data modalities including images, product descriptions, videos, and audio to model preference of users. VBPR (He & McAuley, 2016) enhances traditional matrix factorization by incorporating visual features with product ID embeddings. Recent self-supervised learning (SSL)-based approaches, such as SGL (Wu et al., 2021) and NCL (Lin et al., 2022), employ distinct augmentation strategies: SGL utilizes random node and edge dropping, while NCL explores global user-item interactions. SLMRec (Tao et al., 2022) further applies random perturbations to modality features for contrastive learning. Graph neural networks (GNNs) have become prevalent for capturing user-item relationships. LightGCN (He et al., 2020) simplifies message passing with a parameter-free linear mechanism, while MMGCN (Wei et al., 2019) constructs modality-specific graphs for representation learning. GRCN (Wei et al., 2020) optimizes graph structure by identifying and pruning noisy edges. LATTICE (Zhang et al., 2021) emerges as a leading framework through its efficient dual aggregation of user-item and item-item graphs, proving that simplified graph structures enhance recommendation performance. FREEDOM (Zhou & Shen, 2023) extends the LATTICE framework by introducing a structure denoising module and a graph freezing strategy, which effectively filters out noisy latent relationships and enhances recommendation performance. Then, LATTICE (Zhang et al., 2021)inspires methods like MICRO (Zhang et al., 2023a) and MGCN (Wei et al., 2019) that model co-occurrence patterns. MICRO (Zhang et al., 2023a) employs contrastive learning for modality fusion, and BM3 (Zhou et al., 2023) uses embedding dropout for self-supervised learning. Although LGMRec (Guo et al., 2024) models global-local features via hypergraphs, its over-reliance on co-occurrence patterns hinders differentiation between similar users, limiting preference representation accuracy. MCDRec (Ma et al., 2024) injects modality-aware uncertainty into item embeddings via diffusion to align multimodal features with collaborative signals, then leverages these refined representations to denoise user-item graphs. MIG-GT (Hu et al., 2025) designs different GNN receptive field for each modality and integrates a sampling-based global Transformer to model long-range dependencies across the user-item graph.

2.2. Diffusion models in recommendation

Diffusion models have become essential for generative tasks since the introduction of DDPM (Ho et al., 2020). Advances in sampling efficiency and conditional diffusion now enable performance rivaling VAEs (Kingma & Welling, 2014) and GANs (Goodfellow et al., 2020), without their stability issues. These developments inspired two recommendation approaches: generating user-item interactions in graph space or producing embeddings in latent space (Zhao et al., 2024). Seminal works demonstrate effectiveness of DM for recommendation. SGMs (Song et al., 2021a; Song & Ermon, 2020; Song et al., 2021b) adapt score-based generation for collaborative filtering via perturbation-recovery, while recommendation DDPM (Ho et al., 2020)denoises user and item embeddings to handle implicit feedback noise. Graph-based methods like GiffCF (Zhu et al., 2024) with heat equation smoothing and CF-Diff (Hou et al., 2024) with multi-hop attention leverage DMs to enhance collaborative filtering. Social recommendation like RecDiff (Li et al., 2024) denoises latent social space, though noisy edges persist. Multimodal approaches (DiffMM Jiang et al., 2024a, MCDRec Ma et al., 2024) align visual, textual and audio features with interactions, significantly boosting accuracy.

2.3. Granular-ball computing

In the evolution of granular computing theory, Wang (Wang, 2017) innovatively integrated traditional granular computing principles with the human cognitive characteristic of "macro-to-micro perspective" (Chen, 1982), establishing a theoretical framework for multi-granularity cognitive computing. This groundbreaking research laid a crucial foundation for the subsequent development of granular-ball computing by the Xia team (Xia et al., 2019). This novel approach represents data through encapsulated granular balls, overcoming the limitations of traditional point-wise granular computing while demonstrating significant advantages in computational efficiency, methodological robustness, and result interpretability. Currently, granular-ball computing has achieved important breakthroughs across multiple machine learning domains, with notable applications including: granular-ball clustering algorithms (Cheng et al., 2024; Xie et al., 2024a,b,c), granular-ball classification models (Xia et al., 2024; Yang et al., 2025a, 2026), granular-ball sampling techniques (Xia et al., 2023b), granular-ball rough set theory (Xia et al., 2022; Zhang et al., 2023b), granular-ball three-way decision methods (Xia et-al., 2024; Yang et al., 2024, 2025b), and granular-ball reinforcement learning frameworks (Liu et al., 2024). Furthermore, multiple application studies have validated the effectiveness of granular-ball representation, such as in textual adversarial defense (Wang et al., 2024a), label noise resistance (Wang et al., 2024b) and feature selection (Cao et al., 2024). MGBCC (Su et al., 2025) proposes a granular-ball contrastive learning method for multi-view clustering, which groups data into coarse-grained balls to capture local structures and cross-view relationships, avoiding the limitations of instance-level and cluster-level approaches.

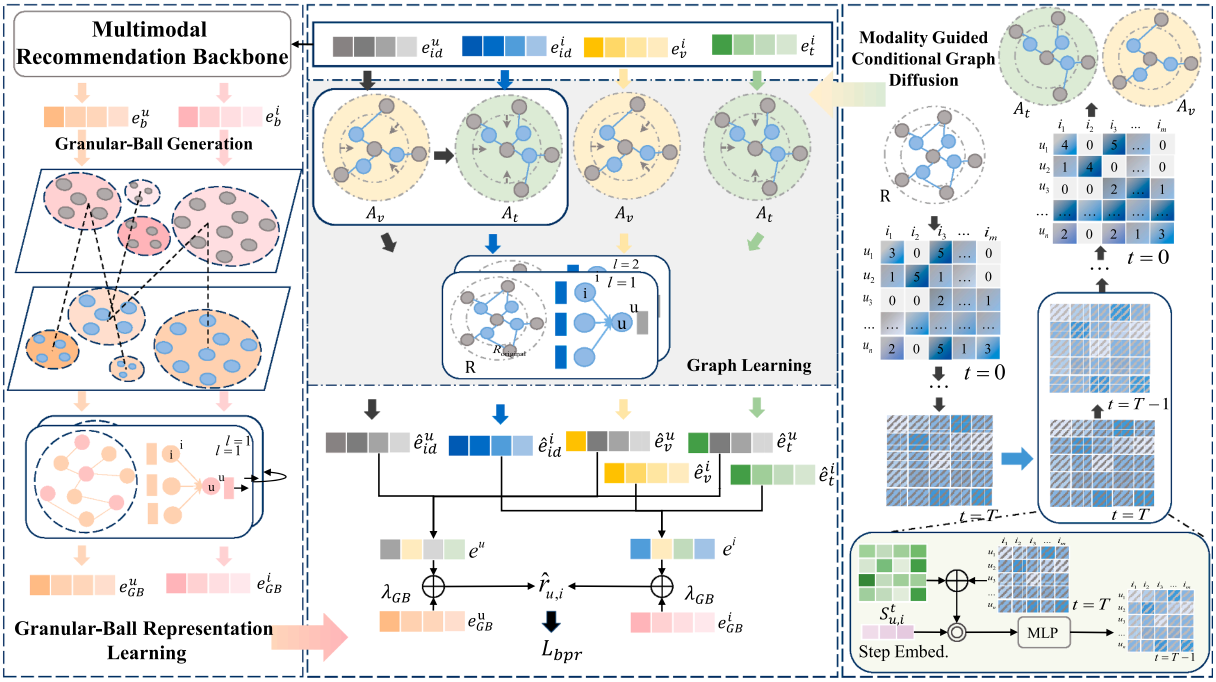


Fig. 1. The framework of the proposed GBDiff.

3. Proposed method

In this section, we introduce GBDiff in details. The overall framework is shown in Fig. 1. GBDiff comprises three main components: (a) Granular-Ball Representation Learning Module, (b) Modality-guided Conditional Graph Diffusion Module, (c) Graph Learning Module.

4. Preliminaries

Let $\mathcal{U}$, $\mathcal{I}$ denote the collections of users and items, respectively. Following mainstream recommendation methods (Guo et al., 2024), we randomly initialize user and item ID embeddings, denoted as $e^u_{id}, e^i_{id} \in \mathbb{R}^d$, where d is the embedding dimension. Furthermore, each item is associated with multimodal features $f^m \in \mathbb{R}^{d_m}$, where d_m is the original modality embedding dimension, and $m \in \mathcal{M} = \{V, T\}$ indicates the modality type (V for visual and T for textual). These original modality features are mapped into low-dimensional embeddings $\mathbf{e}^i_m \in \mathbb{R}^d$ via modality-specific MLPs. For users, multimodal features $\mathbf{e}^u_m$ are initialized by aggregating the modality-specific embeddings of the items they have interacted with in the user-item interaction graph. User historical behavior data can be represented as a sparse binary interaction matrix $R \in \{0,1\}^{|U|\times|I|}$, where $r_{ui} = 1$ if user u interacted with item i, and 0 otherwise. These interaction data naturally form a bipartite graph structure $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with the vertex set $\mathcal{V} = U \cup I$ containing all user and item nodes, and the edge set $\mathcal{E} = \{(u,i)|r_{ui} = 1\}$ corresponding to actual user-item interactions. Finally, given the interaction data and the multimodal features of items, our goal is to predict the preference scores of a specific user on each item.

4.1. Granular-ball representation learning module

The Granular-Ball Representation Learning module attempts to model the coarse-grained semantic patterns to complement instance-level graph learning. Specifically, we partition user and item embeddings into coarse-grained representative units, i.e., granular-balls, and establish cross-view relationships among granular-balls. Compared with instance-level modeling, our approach forms coarse-grained understanding before making fine-grained judgments according to the cognitive theory which suggests that humans tend to process information in a “global-first” manner (Xia et al., 2023a). In addition, granular-balls enable the discovery of latent associations among distant nodes in the graph, thereby enriching global information.

4.1.1. Granular-ball generation

Given the user feature matrix $\mathbf{E}^u_b \in \mathbb{R}^{N_u\times d}$ and the item feature matrix $\mathbf{E}^i_b \in \mathbb{R}^{N_i\times d}$ learned from a multimodal recommendation backbone (Ong & Khong, 2025), where N_u and N_i denote the numbers of users and items respectively and d is the embedding dimension. we denote $\mathbf{e}^u_b \in \mathbb{R}^d$ and $\mathbf{e}^i_b \in \mathbb{R}^d$ as the embedding vectors of an individual user and item, corresponding to the rows of $\mathbf{E}^u_b$ and $\mathbf{E}^i_b$, we apply the granular-ball generation process independently to the user and item views. To be specific, for each view, we initialize all data points (e.g., users or items) as a single granular-ball and iteratively partition it using a granular-ball generation method (Su et al., 2025). To control the granularity, we introduce a parameter p which indicates the scale at which granular-balls are generated. The number of granular-balls k_v in view $v \in \{u, i\}$ is computed as:

$$k_v = \max\left(\left\lfloor \frac{N_v}{p} \right\rfloor, 1\right) \tag{1}$$

For simplicity, we set the same granularity parameter p for each view and adopt this setting in the subsequent analysis and experiments. We adopt the same granular-ball generation strategy as Su et al. (2025) to partition embeddings into k_u (or k_i) granular-balls. For the mth granular-ball GB^v_m in view $v \in \{u, i\}$, its center $\mathbf{c}^v_m$ and radius r^v_m are computed as:

$$\mathbf{c}^v_m = \frac{1}{N_m} \sum_{e^v_b \in GB^v_m} e^v_b, \quad r^v_m = \frac{1}{N_m} \sum_{e^v_b \in GB^v_m} \|\mathbf{c}^v_m - e^v_b\|_2, \tag{2}$$

where N_m is the number of samples in GB^v_m.

To address the issue of overly small or outlier-heavy granular-balls, we introduce a merging process. Specifically, granular-balls with fewer samples than a threshold η are merged with their nearest neighbor, based on the Euclidean distance between centers of balls. After merging, the new center and radius are recalculated, ensuring that the resulting balls remain the integrity of the local structure. Overall, granular-balls transform fine-grained embeddings into coarse-grained semantics that

Algorithm 1 User and item granular-ball generation.

```
Input: User feature matrix $\mathbf{E}_b^u \in \mathbb{R}^{N_u \times d}$; Item feature matrix $\mathbf{E}_b^i \in \mathbb{R}^{N_i \times d}$;
    Granularity parameter $p$; Minimum threshold $\eta$.
Output: User granular-ball set $\mathcal{S}^u = \{GB_n^u\}_{n=1}^{k_u}$; Item granular-ball set
    $\mathcal{S}^i = \{GB_m^i\}_{m=1}^{k_i}$.
for each view $v \in \{u, i\}$ do
    Initialize all embeddings in view $v$ as a single granular-ball
    Compute the number of granular-balls by Eq. (1);
    Apply granular-ball generation method to partition embeddings into $k_v$ granular-balls
    for each granular-ball $GB_m^v$ do
        Compute center $\mathbf{c}_m^v$ and radius $r_m^v$ by Eq. (2);
    end for
    while exists granular-ball $GB_m^v$ with $|GB_m^v| < \eta$ do
        for each granular-ball $GB_l^v \in \mathcal{S}^v$, $l \neq m$ do
            Compute center distance: $d_{m,l} = \|\mathbf{c}_m^v - \mathbf{c}_l^v\|_2$
        end for
        Identify the nearest granular-ball: $n = \arg\min_{l \neq m}(d_{m,l})$
        Merge $GB_m^v$ into $GB_n^v$
        Remove $GB_m^v$ from granular-ball set
        Update center $\mathbf{c}_n^v$ and radius $r_n^v$
    end while
    Obtain granular-ball set $\mathcal{S}^v$
end for
return $\mathcal{S}^u, \mathcal{S}^i$
```

are both noise-resistant and structurally representative, making them well-suited for recommendation tasks. The detailed procedure of the granular-ball generation is shown in Algorithm 1.

4.1.2. Granular-ball association and learning

We further establish associations between user and item granular-balls to capture coarse-grained collaborative patterns. To this end, we construct a cross-view association matrix $\mathbf{P}^{(u,i)} \in \{0,1\}^{k_u \times k_i}$ based on the original user-item interaction matrix. For the nth user granular-ball GB_n^u and the mth item granular-ball GB_m^i, the set of interaction pairs is defined as:

$$\mathbf{Id}_{\text{pair}} = \{(u,i) \mid u \in \mathbf{Id}(GB_n^u), i \in \mathbf{Id}(GB_m^i), R_{u,i} = 1\}, \tag{3}$$

where $\mathbf{Id}(GB_n^u)$ and $\mathbf{Id}(GB_m^i)$ denote the user and item indices in their corresponding granular-balls. $\mathbf{R} \in \{0,1\}^{N_u \times N_i}$ is the original interaction matrix, and $t_{\text{pair}} = |\mathbf{Id}_{\text{pair}}|$ represents the number of interaction pairs. The association between granular-balls is determined by the normalized interaction ratio:

$$\mathbf{P}_{nm}^{(u,i)} = \begin{cases} 1, & \text{if } \frac{t_{\text{pair}}}{\sqrt{t_n^u \cdot t_m^i + \epsilon}} \geq \tau, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where t_n^u and t_m^i are the sizes of in GB_n^u and GB_m^i, respectively, ϵ is introduced to avoid division by zero, and τ is a threshold. This normalization ensures robustness against varying granular-ball sizes. Then, we build a bipartite graph with user- and item-balls as nodes and edges defined by $\mathbf{P}^{(u,i)}$. The adjacency matrix is $\bar{\mathbf{P}} = \begin{bmatrix} \mathbf{0} & \mathbf{P}^{(u,i)} \\ \mathbf{P}^{(u,i)\top} & \mathbf{0} \end{bmatrix}$. For each granular-ball, the initial embedding is computed as the average of its member embeddings. For example, for a user granular-ball GB_n^u with user indices $\mathbf{Id}(GB_n^u)$, its initial embedding is:

$$e_{GB_n}^u = \frac{1}{|\mathbf{Id}(GB_n^u)|} \sum_{u \in \mathbf{Id}(GB_n^u)} e_b^u, \tag{5}$$

where $\mathbf{e}_b^u \in \mathbb{R}^d$ is the embedding of user u. Similarly, item granular-ball embeddings are obtained. These embeddings are then concatenated as input for graph convolution. Finally, we adopt LightGCN to encode the granular-ball interaction graph:

$$\mathbf{E}_{GB}^{(l)} = \left(\mathbf{D}^{-1/2}\bar{\mathbf{P}}\mathbf{D}^{-1/2}\right)\mathbf{E}_{GB}^{(l-1)}, \tag{6}$$

where D is the degree matrix of $\bar{\mathbf{P}}$. The final layer output is used as the representation of granular-balls.

4.2. Modality-guided conditional graph diffusion model

To mitigate the adverse effects of irrelevant or noisy interactions, we propose a modality-guided conditional graph diffusion module which reconstructs user-item interaction graphs guided by modality-specific information, thereby improving the modeling of user preferences. In particular, our graph diffusion paradigm over the original user-item interactions includes two crucial processes. In the forward process, the original user-item graph is corrupted by incrementally introducing Gaussian noise. In the reverse reconstruction phase, modality-specific denoising networks leverage similarity matrices between users and items in the respective modality as guiding signals to gradually restore the noise-disrupted interaction relationships. By iteratively refining the corrupted graph, our approach integrates collaborative user-item signals with modality-specific guidance to obtain more robust user-item connections.

4.2.1. Forward graph diffusion process

Given user $u \in \mathcal{U}$ with interaction vector $r_u = [r_u^0, r_u^1, \ldots, r_u^{|\mathcal{I}|-1}]$ over items $\mathcal{I}$, where $r_u^i \in \{0,1\}$ whether user u interacted with item i. The diffusion process starts with $x_0 = r_u$, where r_u represents the observed interactions of user u. The forward process constructs $x_{1:T}$ through a Markov chain that incrementally introduces Gaussian noise over T time steps, indexed by t. Each transition from x_{t-1} to x_t follows the standard diffusion parameterization as follows:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}), \tag{7}$$

where $\{\beta_t\}_{t=1}^T$ controls the noise injection rate, $t \in \{1,2,\ldots,T\}$ is the current step, $\mathbf{I}$ is the identity matrix, and $\mathcal{N}$ is the Gaussian distribution which means x_t is sampled from this distribution. According to the additivity of independent Gaussian noises and reparameterization trick, x_t can be directly obtained from x_0 in the calculation:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{8}$$

where $\bar{\alpha}_t = \prod_{t'=1}^t \alpha_{t'}$, $\alpha_t = 1-\beta_t$. We reparameterize $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

4.2.2. Reverse graph diffusion process

The reverse process aims to iteratively denoise x_t for t steps to ultimately approximate the initial user-item interactions:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \boldsymbol{\mu}_\theta(x_t, t, \hat{\mathbf{s}}_{u,*}^m), \boldsymbol{\Sigma}_\theta(x_t, t)), \tag{9}$$

where $\boldsymbol{\Sigma}_\theta(x_t,t) = \sigma_t^2\mathbf{I} = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\mathbf{I}$ denotes the variance. $\hat{\mathbf{s}}_{u,*}^m$ quantifies the semantic similarity between a given user u and all items under modality m, with its definition provided in the subsequent section **Conditional Estimator**. In the recommendation domain, it is common to train a suitable neural estimator $f_\theta(\cdot)$ directly to approximate x_0 for performing the reverse steps. The mean $\boldsymbol{\mu}_\theta(x_t, t, \hat{\mathbf{s}}_{u,*}^m)$ is computed as:

$$\boldsymbol{\mu}_\theta(x_t, t, \hat{\mathbf{s}}_{u,*}^m) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} f_\theta(x_t, t, \hat{\mathbf{s}}_{u,*}^m)\right) \tag{10}$$

4.2.3. Conditional estimator

The vanilla denoising diffusion probabilistic model (Ho et al., 2020) is not suitable to generate reasonable propagation paths, as it lacks controllable conditions derived from modality information and historical interactions, and thus fails to effectively mitigate the noisy connections. To address this issue, we propose to utilize the modality-specific semantic signals to guide the generation of propagation paths.

For each modality $m \in \mathcal{M}$, we generate high-quality modality-specific embeddings $\mathbf{e}_m^u$ and $\mathbf{e}_m^i$ for users and items by employing LightGCN to perform information aggregation over the original user-item interaction graph $\mathbf{R}$:

$$\begin{aligned}\mathbf{e}_m^{u,l+1} &= \sum_{i \in \mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_i|}} \mathbf{e}_m^{i,l}, \\ \mathbf{e}_m^{i,l+1} &= \sum_{u \in \mathcal{N}_i} \frac{1}{\sqrt{|\mathcal{N}_i|}\sqrt{|\mathcal{N}_u|}} \mathbf{e}_m^{u,l},\end{aligned} \tag{11}$$

where $\mathcal{N}_u$ and $\mathcal{N}_i$ denote the neighbor sets of user u and item i, respectively, l denotes the layer. The final user and item embeddings are obtained from the Lth layer, denoted as $\tilde{\mathbf{e}}_{m,L}^u$ and $\tilde{\mathbf{e}}_{m,L}^i$. Subsequently, we construct a user-item bipartite graph using a k-nearest neighbors (KNN) approach based on these embeddings and compute the cosine similarity matrix:

$$S_{u,i}^m = \frac{(\tilde{\mathbf{e}}_{m,L}^u)^\top \tilde{\mathbf{e}}_{m,L}^i}{\|\tilde{\mathbf{e}}_{m,L}^u\| \|\tilde{\mathbf{e}}_{m,L}^i\|} \tag{12}$$

To focus on the most semantically relevant connections, we sparsify the graph by retaining only the top-k edges for each user node:

$$\hat{S}_{u,i}^m = \begin{cases} s_{u,i}^m & \text{if } S_{u,i}^m \in \text{top-}k(S_u^m) \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

This resulting top-k similarity matrix $\hat{\mathbf{s}}_{u,*}^m$ provides semantic guidance for the reverse diffusion process. To integrate this condition into the denoising step, we fuse the similarity vector $\hat{\mathbf{s}}_{u,*}^m$ with x_t at time step t:

$$x_t' = x_t + \hat{\mathbf{s}}_{u,*}^m, \tag{14}$$

where $\hat{\mathbf{s}}_{u,*}^m \in \mathbb{R}^{1 \times N}$ denotes the similarity vector corresponding to user u in modality m, extracted from the top-k similarity matrix $\hat{\mathbf{S}}_{u,i}^m$. The augmented representation x_t' is concatenated with the temporal embedding e_t, i.e., $h_t = [x_t' \parallel e_t]$, where $\parallel$ denotes the concatenation operation. The combined feature h_t undergoes a nonlinear transformation through a two-layer multilayer perceptron (MLP) to obtain the denoised representation x_{t-1}:

$$x_{t-1} = \mathbf{W}_2 \sigma(\mathbf{W}_1 h_t + \mathbf{b}_1) + \mathbf{b}_2, \tag{15}$$

where $\sigma(\cdot)$ represents the ReLU activation function, $\mathbf{W}_1, \mathbf{W}_2$ are trained parameters, and $\mathbf{b}_1, \mathbf{b}_2$ are bias terms. This process can be formally expressed as:

$$x_{t-1} = f_\theta(x_t, \hat{\mathbf{S}}_{u,*}^m, e_t), \tag{16}$$

where f_θ denotes the denoising function parameterized by a neural network. The resulting x_{t-1} represents the denoised user-item interactions which effectively captures both semantic relationships from the conditional top-k similarity matrix and temporal dynamics from the diffusion process, ensuring that the generated interactions are semantically consistent with user preferences and modality-specific informations.

4.2.4. Optimization

To optimize the diffusion model for generating modality-aware user-item interaction graph, we employ a composite loss function that combines reconstruction loss and modality-aware signal injection (MSI) loss (Jiang et al., 2024a). The reconstruction loss aligns the learned prior distribution $p_\theta(x_{t-1}|x_t)$ with the posterior $q(x_{t-1}|x_t, x_0)$, formalized as the Kullback-Leibler (KL) divergence:

$$L_{\text{vlb}} = D_{\text{KL}}\big(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)\big) \tag{17}$$

Following the Denoising Diffusion Probabilistic Model (DDPM) framework (Ho et al., 2020), this KL divergence is simplified to a Mean-Squared Error (MSE) loss, focusing on reconstructing the original data x_0 from the noisy input $\mathbf{x}_t$, and ensuring the generated interactions converge to the true data distribution:

$$L_{dm} = \left\| x_0 - f_\theta(x_t, \hat{\mathbf{S}}_{u,*}^m, e_t) \right\|_2^2 \tag{18}$$

To incorporate modality-specific semantics, we introduce the MSI loss. For each modality $m \in \mathcal{M}$, we aggregate aligned item modal features $\mathbf{e}_m^i$ with predicted interaction probabilities $\hat{x}_0$, and item ID embeddings $\mathbf{e}_{id}^i$ with observed interactions x_0. The discrepancy is minimized via an MSE loss:

$$L_{msi}^m = \left\| \hat{\mathbf{x}}_0 \cdot \mathbf{e}_m^i - \mathbf{x}_0 \cdot \mathbf{e}_{id}^i \right\|_2^2 \tag{19}$$

This joint optimization guarantees that the learned user-item interactions not only preserve fidelity to the observed data but also incorporate semantic signals from multiple modalities, ultimately improving recommendation accuracy. In particular, λ_m serves as a hyperparameter to control the relative contributions of different modality-specific losses.

$$L_{diff} = L_{dm} + \sum_{m \in \mathcal{M}} \lambda_m L_{msi}^m \tag{20}$$

4.2.5. Inference

Our multimodal graph diffusion model predicts user-item interactions by corrupting the observed interaction probabilities x_0 over T forward diffusion steps to obtain $\hat{x}_T$, followed by a deterministic reverse denoising process guided by conditional information derived from modality-specific semantic similarity $S_{u,i}^m$ to reconstruct $\hat{x}_0$. For each user u, we select the top-k items predicted probabilities from $\hat{x}_0 = [\hat{x}_{u,0}, \hat{x}_{u,1}, \dots, \hat{x}_{u,|I|-1}]$ to construct the modality-specific user-item graph $\mathbf{A}_m$, capturing semantic relationships for modality m.

4.3. Graph learning module

4.3.1. Collaborative graph learning

To enhance the modeling of fine-grained collaborative patterns, we design a collaborative graph learning module that first performs graph convolutions on modality-specific interaction graphs generated through the diffusion process, and then applies multi-layer convolution on the original interaction graph. The modality-specific interaction graphs $\mathbf{A}_m$ are derived from the conditional graph diffusion process described in Section 4.2, where $m \in \{v, t\}$ denotes visual and textual modality. We apply symmetric normalization to $\mathbf{A}_v$, $\mathbf{A}_t$, and the original user-item interaction graph $\mathbf{R}$, as follows:

$$\tilde{\mathbf{A}}_m = \mathbf{D}_m^{-1/2} \mathbf{A}_m \mathbf{D}_m^{-1/2}, \quad \tilde{\mathbf{A}} = \mathbf{D}^{-1/2} \mathbf{R} \mathbf{D}^{-1/2}, \tag{21}$$

where $\mathbf{D}_m$ and $\mathbf{D}$ are the corresponding degree matrices. Such normalization ensures numerical stability and improves propagation efficiency on large-scale graphs. We initialize the embeddings as $\mathbf{E}_{id} \in \mathbb{R}^{(N_u+N_i)\times d}$, which are constructed by concatenating user embeddings $\mathbf{e}_{id}^u \in \mathbb{R}^d$ and item embeddings $\mathbf{e}_{id}^i \in \mathbb{R}^d$. To capture local modality-specific signals, the embeddings are first propagated on the normalized modality-specific graphs:

$$\hat{\mathbf{E}}_{id} = \tilde{\mathbf{A}}_t (\tilde{\mathbf{A}}_v \mathbf{E}_{id}) \tag{22}$$

This progress integrates localized modality-aware interaction patterns which enriches the representations with signals from neighbors that are particularly relevant in visual or textual contexts. At the same time, it mitigates the influence of noisy interactions by reinforcing structurally coherent patterns within the modality graph. The refined embeddings $\hat{\mathbf{E}}_{id}$ are then propagated through a multi-layer LightGCN on the normalized original interaction graph $\tilde{\mathbf{A}}$ to capture high-order collaborative signals:

$$\hat{\mathbf{E}}_{id}^{(l)} = \tilde{\mathbf{A}} \hat{\mathbf{E}}_{id}^{(l-1)}, \quad l = 1, \dots, L \tag{23}$$

Finally, to focus on modeling fine-grained collaborative signals between users and items, we obtain the output representations by averaging the embeddings across all layers:

$$\hat{\mathbf{E}}_{id} = \frac{1}{L+1} \sum_{l=0}^{L} \hat{\mathbf{E}}_{id}^{(l)} \tag{24}$$

4.3.2. *Multimodal graph learning*

The multimodal graph learning module is designed to preserve modality-specific semantics while capturing fine-grained local interactions. For each modality, we concatenate $\mathbf{e}^u_m$ and $\mathbf{e}^i_m$ to form initial modality embeddings $\mathbf{E}_m$. We then perform message passing on the normalized graph $\tilde{\mathbf{A}}_m$ to obtain refined embeddings $\hat{\mathbf{E}}_m = \tilde{\mathbf{A}}_m\mathbf{E}_m$, which enhances local modality-aware features. These refined embeddings are subsequently propagated through multiple LightGCN layers over the normalized original interaction graph $\tilde{\mathbf{A}}$ to capture higher-order local structures:

$$\hat{\mathbf{E}}^{(l)}_m = \tilde{\mathbf{A}}\hat{\mathbf{E}}^{(l-1)}_m, \quad l = 1, \dots, L \tag{25}$$

4.4. *Optimization objectives*

By integrating signals from different perspectives, we obtain the final expressive and robust representations of users and items, as follows:

$$\mathbf{E} = \hat{\mathbf{E}}_{id} + \hat{\mathbf{E}}_v + \hat{\mathbf{E}}_t + \lambda_{GB}\mathbf{E}_{GB}, \tag{26}$$

where λ_{GB} is the weight of granular-ball representation. The predicted interaction score is computed as $\hat{y}_{u,i} = (\mathbf{E}^u)^\top \mathbf{E}^i$. To optimize the model, we adopt the Bayesian Personalized Ranking (BPR) loss, which encourages higher prediction scores for positive items compared to negative ones:

$$\mathcal{L}_{\text{BPR}} = \sum_{(u,i,j)\in\mathcal{D}} -\ln\sigma\big(\hat{y}_{u,i} - \hat{y}_{u,j}\big), \tag{27}$$

where $\mathcal{D} = \{(u,i,j) \mid (u,i) \in \mathcal{D}^+, (u,j) \in \mathcal{D}^-\}$ denotes the interaction set, including positive interactions $\mathcal{D}^+$ and negative interactions $\mathcal{D}^-$, with σ representing the sigmoid function. The overall model is optimized by integrating the BPR loss with a diffusion loss and an L2 regularization term, resulting in the following total objective:

$$\mathcal{L} = \mathcal{L}_{\text{BPR}} + \lambda_1\mathcal{L}_{\text{dm}} + \lambda_2\|\Theta\|_2^2, \tag{28}$$

where λ_1 and λ_2 control the contributions of the diffusion loss and L2 regularization, respectively.

5. Experiments

We conducted extensive experiments to address the following research questions:

- **RQ1**: How does the proposed model GBDiff perform compared to state-of-the-art general and multimodal recommendation systems?
- **RQ2**: What are the contributions of the key components in GBDiff to its overall effectiveness?
- **RQ3**: How do variations in hyperparameters affect the performance of the proposed model?
- **RQ4**: How does GBDiff perform under different levels of data sparsity?
- **RQ5**: How does GBDiff impact the distribution of user and item representations?
- **RQ6**: How does the conditional fusion strategies affect recommendation performance?
- **RQ7**: What is the impact of the merging threshold η on recommendation performance?
- **RQ8**: How do different granular-ball similarity metrics affect the model performance?
- **RQ9**: How efficient is GBDiff compared to other multimodal recommendation models?

5.1. *Experiment configurations*

5.1.1. *Datasets*

To assess the performance of our proposed GBDiff model, we perform extensive experiments on two publicly available Amazon datasets (McAuley et al., 2015), including “Baby”, and “Sports and Outdoors”. For ease of reference, we label them as Baby, and Sports, respectively. Each dataset is processed using a 5-core filtering approach to ensure sufficient user and item interactions. These datasets incorporate both visual and textual modality features. Following prior work (Zhou et al., 2023), we utilize pre-extracted features: 4096-dimensional visual features and 384-dimensional textual features. A summary of the dataset statistics is provided in Table 1.

Table 1
Statistics of the two datasets.

Dataset	#Users	#Items	#Interactions	Sparsity
Baby	19,445	7050	160,792	99.883%
Sports	35,598	18,357	296,337	99.955%

5.1.2. *Baselines*

To evaluate the performance of our proposed model GBDiff, we compare it with 16 baseline methods which can be roughly grouped into two categories:

- **General Recommenders**:
  - **MF-BPR** (Rendle et al., 2012) optimizes recommendation systems based on matrix factorization techniques by incorporating Bayesian Personalized Ranking (BPR) loss.
  - **LightGCN** (He et al., 2020) simplifies the GNN architecture by adopting parameter-free linear message passing for efficient user-item relationship modeling.
  - **SGL** (Wu et al., 2021) employs self-supervised learning through random node or edge dropping augmentations on the interaction graph to enhance graph representations.
  - **NCL** (Lin et al., 2022) enhances contrastive learning by capturing global user-item interaction patterns beyond local graph structures.
- **Multimodal Recommenders**:
  - **VBPR** (He & McAuley, 2016) extends matrix factorization by incorporating visual features with item ID embeddings.
  - **MMGCN** (Wei et al., 2019) learns modality-specific user preferences by performing message passing on the user-item bipartite graph of each modality.
  - **GRCN** (Wei et al., 2020) optimizes graph topology by dynamically identifying and pruning noisy edges.
  - **LATTICE** (Zhang et al., 2021) leverages multimodal features to discover latent semantic structures between items and aggregates information derived from item-item graphs.
  - **SLMRec** (Tao et al., 2022) generates contrastive views through three types of data augmentation at different granularity levels on modality features to enhance multimodal recommendation.
  - **FREEDOM** (Zhou & Shen, 2023) improves LATTICE by incorporating a structure denoising module and freezing item-item graphs, allowing for more efficient and robust multi-modal recommendation through refined latent relation mining.
  - **MICRO** (Zhang et al., 2023a) adopts modality-aware contrastive learning to improve the quality of feature fusion.
  - **BM3** (Zhou et al., 2023) utilizes an embedding dropout mechanism for robust self-supervised representation learning.
  - **DiffMM** (Jiang et al., 2024a) generates modality-aware user-item interaction graphs through diffusion and improves recommendation performance via cross-modal contrastive learning on these graphs.
  - **LGMRec** (Guo et al., 2024) models global-local dependencies via hypergraphs but may face challenges in distinguishing preferences of similar users.
  - **MCDRec** (Ma et al., 2024) boosts item embeddings by introducing modality-specific uncertainty, aiming to alleviate biases between collaborative features and multimodal features.

Table 2
Overall performance comparison of GBDiff and baseline models on two datasets in terms of Recall@K (R@K) and NDCG@K (N@K). The best results are bolded, and the second-best results are underlined.

		Baby				Sports			
Category	Model	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
General Recommenders	MF-BPR	0.0379	0.0607	0.0202	0.0261	0.0452	0.0690	0.0252	0.0314
	LightGCN	0.0464	0.0732	0.0251	0.0320	0.0553	0.0829	0.0307	0.0379
	SGL	0.0532	0.0820	0.0289	0.0363	0.0620	0.0945	0.0339	0.0423
	NCL	0.0538	0.0836	0.0292	0.0369	0.0616	0.0940	0.0339	0.0421
Multimodal Recommenders	VBPR	0.0424	0.0663	0.0223	0.0284	0.0556	0.0854	0.0301	0.0378
	GRCN	0.0531	0.0835	0.0291	0.0370	0.0600	0.0921	0.0324	0.0407
	MMGCN	0.0498	0.0749	0.0261	0.0315	0.0582	0.0825	0.0305	0.0382
	LATTICE	0.0536	0.0858	0.0287	0.0370	0.0618	0.0950	0.0337	0.0423
	SLMRec	0.0540	0.0810	0.0296	0.0361	0.0676	0.1007	0.0374	0.0462
	FREEDOM	0.0627	0.0992	0.0361	0.0452	0.0717	0.1089	0.0385	0.0481
	MICRO	0.0570	0.0905	0.0310	0.0406	0.0675	0.1026	0.0365	0.0463
	BM3	0.0538	0.0857	0.0301	0.0378	0.0659	0.0979	0.0354	0.0437
	DiffMM	0.0625	0.0975	0.0323	0.0411	0.0681	0.1017	0.0370	0.0458
	LGMRec	0.0644	0.1002	0.0349	0.0440	0.0720	0.1068	0.0390	0.0480
	MCDRec	0.0644	0.1013	0.0343	0.0438	0.0737	0.1100	0.0392	0.0488
	MIG-GT	0.0665	0.1021	0.0361	0.0452	0.0753	0.1130	0.0414	0.0511
	GBDiff	**0.0674**	**0.1034**	**0.0368**	**0.0461**	**0.0780**	**0.1154**	**0.0426**	**0.0520**
	Improv.	1.35%	1.27%	1.94%	1.99%	3.59%	2.12%	2.90%	1.76%

Table 3
Ablation study results on Amazon Baby and Sports datasets.

	Baby				Sports			
Method	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
Ours	**0.0674**	**0.1034**	**0.0368**	**0.0461**	**0.0780**	**0.1154**	**0.0426**	**0.0520**
w/o GB	0.0660	0.1011	0.0361	0.0452	0.0753	0.1123	0.0411	0.0507
w/o Cond	0.0648	0.0992	0.0353	0.0442	0.0755	0.1128	0.0410	0.0506
w/o Diff	0.0642	0.0983	0.0349	0.0437	0.0698	0.1048	0.0379	0.0469

- **MIG-GT** (Hu et al., 2025) enhances multimodal recommendation by employing modality-independent receptive fields in GNNs and a sampling-based global Transformer to capture both fine-grained local structures and broad global dependencies.

5.1.3. Evaluation standards

To assess the performance of our top-K recommendation results, we employ two standard metrics: *Recall@K (R@K)* and *Normalized Discounted Cumulative Gain (NDCG@K)*. Following Guo et al. (2024), we split the interaction data into 80% for training, 10% for validation, and 10% for testing. We evaluate the performance of various methods by reporting the average metrics in test set under top-K condition, with K empirically set at 10 and 20.

5.1.4. Implementation details

We utilize the open-source LGMRec framework (Guo et al., 2024) to develop the proposed model. To ensure a fair comparison, we strictly adhere to the operational settings of existing research. All baseline models are trained with a default batch size of 2048, a learning rate of 0.001, and an embedding dimension of $d = 64$. For all graph-based methods, the number of collaborative graph propagation layers L is set to 2. We initialize all trainable parameters using the Xavier initialization method (Glorot & Bengio, 2010) and optimize models using the Adam optimizer. Early stopping is triggered if the Recall@20 metric on the validation set does not improve for 20 consecutive steps. For our GBDiff model, the number of modality graph embedding layers and collaborative graph embedding layers are tuned within the set $\{1, 2, 3, 4\}$; the weighting factor λ_{GB} is searched in the range $\{0.4, 0.6, \dots, 1.0\}$; and the regularization coefficient λ_1 is optimized over the range $\{10^{-5}, 10^{-4}, \dots, 0.1\}$. We tuned the hyperparameters extensively through experiments to achieve the best model performance across different datasets.

5.2. Overall performance comparison (RQ1)

Table 2 reports the performance comparison of state-of-the-art baseline methods and our method on two datasets. We can observe that: First, multimodal recommendation methods mostly demonstrate superior performance compared with general recommendation methods (e.g., matrix factorization or graph based methods). The results verify the effectiveness of multimodal recommenders which introduce auxiliary information to enrich the representations and improve performance. Second, among all multimodal recommendation methods, diffusion-based approaches DiffMM (Jiang et al., 2024a) and MCDRec (Ma et al., 2024) achieve relatively strong performance in most cases by introducing a diffusion mechanism to mitigate noisy interactions and refine item ID features. In addition, graph-based multimodal recommender such as MIG-GT (Hu et al., 2025) yields the most significant performance among all baselines. This is due to the fact that it addresses inconsistent modality sensitivity via independent receptive fields and captures global dependencies through a Transformer to overcome limitation of GNNs.

Finally, comparing with all state-of-the-art baseline methods, our proposed approach GBDiff achieves the best performance on all two datasets. For instance, GBDiff outperforms the best baseline MIG-GT (Hu et al., 2025) by 1.94% and 2.90% on the Baby and Sports datasets in terms of NDCG@10, respectively. The results demonstrate that the effectiveness of our method GBDiff by integrating diffusion with the conditions of multimodal information and the granular-ball mechanism as it can effectively suppress noisy interactions and model user preferences in a multi-grained perspective.

5.3. Ablation studies (RQ2)

To evaluate the contribution of each component in our proposed model, we compared the performance of our model against three

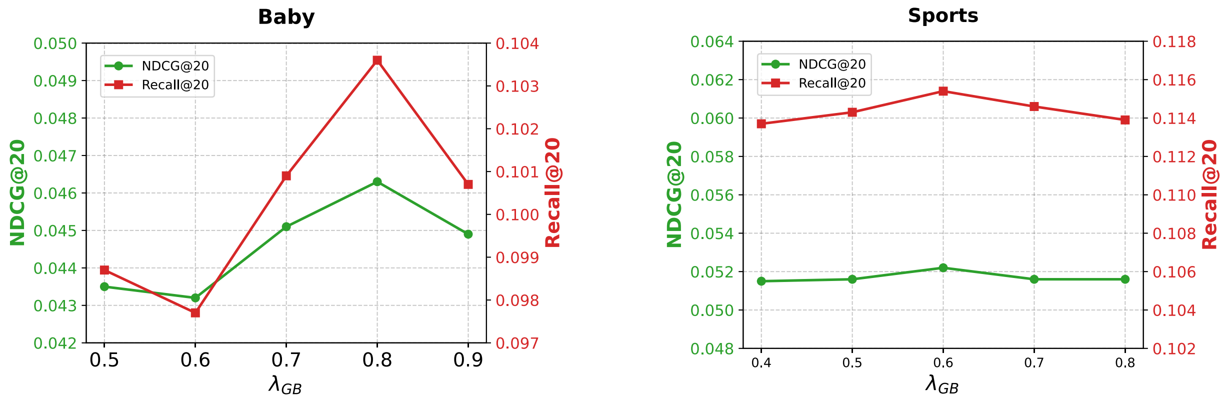


Fig. 2. The effect of λ_{GB} on model performance.

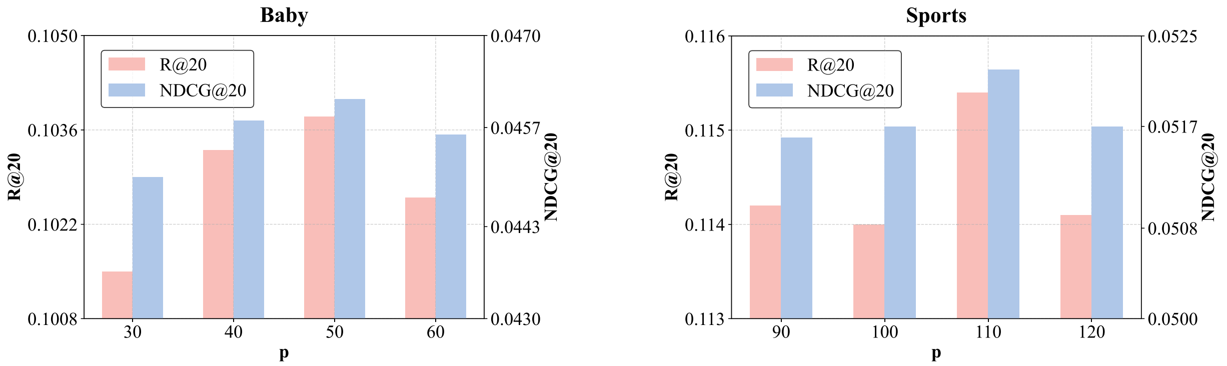


Fig. 3. The effect of p on model performance.

variants, including: (1) **w/o GB**, which removes the granular-ball representation learning module and relies solely on modality information of items and user-item interaction data for recommendation; (2) **w/o Cond**, which disregards conditional information in the diffusion process, employing only a classical diffusion model; and (3) **w/o Diff**, which neglects the modality guided conditional graph diffusion module, solely performing convolution on the original user-item interaction graph. The results are presented in Table 3, and we have the following observations:

- Removing the granular-ball representation learning moduele (**w/o GB**) leads to a moderate performance drop across all datasets, suggesting that the incorporation of coarse-grained patterns and global information in user-item interactions based on granular-ball representation learning can lead to the superior performance of our proposed model. This is because the introduction of user and item balls enables the model to capture coarse-grained user preference from a global perspective, and facilitates the modeling of potential associations between distant nodes.
- Disregarding the conditional guidance (i.e., **w/o Diff**) results in a significant decline in performance across all datasets. It indicates that introducing the specific modality information as condition to guide the reverse process of diffusion is crucial for filtering noise and focusing on meaningful latent relationship propagation. This stems from the fact that denoising on the original interaction graph alone usually suffers from insufficient denoising due to limited information. By introducing specific modality information as diffusion condition, the denoising process can be guided with richer information from multiple perspectives.
- The modality guided conditional graph diffusion module plays a critical role in eliminating the adverse effects of irrelevant or noisy interactions in original user-item graphs. we can observe that the absence of the modality guided conditional graph diffusion module (i.e., **w/o Diff**) leads to a considerable performance decline across all datasets.

5.4. Hyperparameter sensitivity analysis (RQ3)

The weight of granular-ball λ_{GB}: λ_{GB} represents the weight of granular-ball embedding representation in the aggregated representation. To validate the influence of λ_{GB} to our model performance, we tune it in $\{0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ on the two datasets, i.e., Sports and Baby. The results are presented in Fig. 2. We observe that the performance improves with increasing value of λ_{GB} and reaches a peak when λ_{GB} equals to 0.6 on the Sports dataset and 0.8 on the Baby dataset, respectively. This is because the granular-ball representation contains rich global user information. If the value of λ_{GB} continues to increase, the performance will drop. The reason is that when λ_{GB} becomes too large, more coarse-grained information will overwhelm fine-grained local preference of user.

The granularity control parameter p: p is used to control the granularity of granular-ball generation. To investigate the effect of p, we test $p \in \{30, 40, 50, 60\}$ on the Baby and $p \in \{90, 100, 110, 120\}$ on the Sports. The results are reported in Fig. 3. It can be observed that the performance raises with the increase of the value of p and reaches a peak when p equals to 50 on the Baby dataset or 110 on the Sports dataset. If we continue to raise the value of p, the performance will start to drop quickly. This phenomenon occurs because it generates more granular-balls with each ball contains only few samples when p is too small, which can

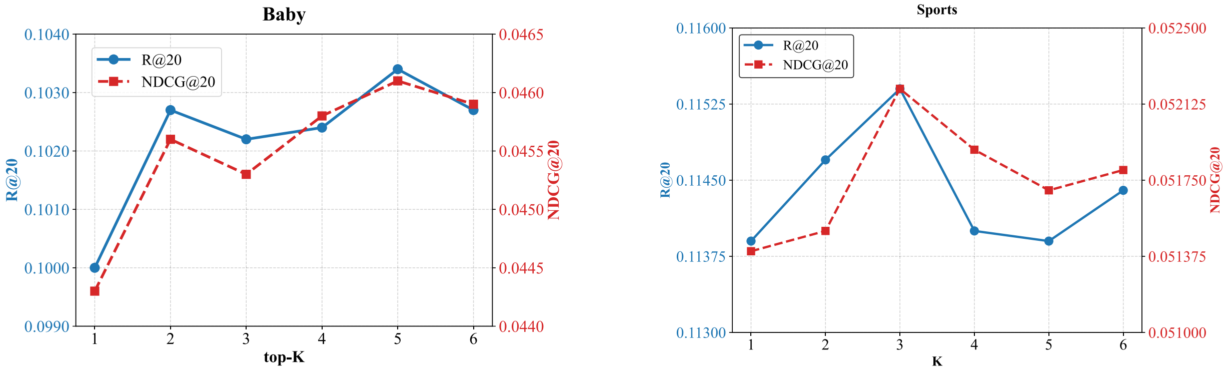


Fig. 4. The effect of k on model performance.

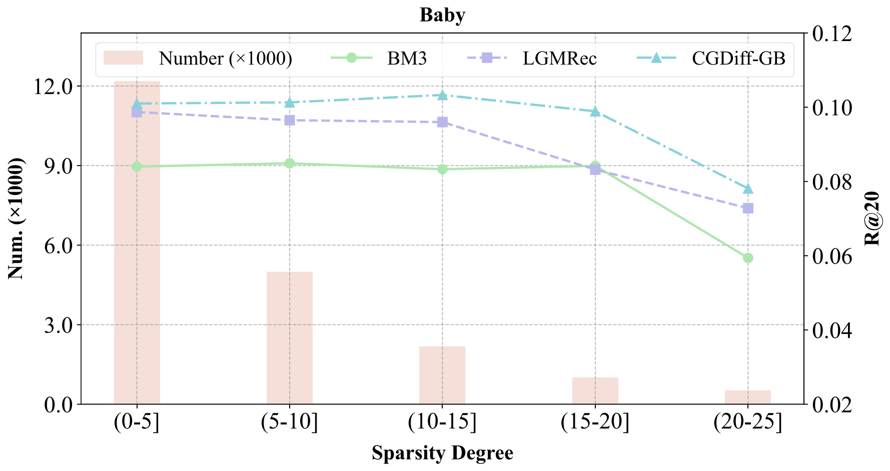


Fig. 5. Performances under different user interaction sparsity in terms of R@20 on the Baby dataset.

not effectively capture the coarse-grained information among users or items. However, when p becomes too large, the generated granular-balls contain too many samples and fail to capture intricate relationships.

Top-k for similarity matrix in the reverse diffusion k**:** k is the number of edges retained in the similarity matrix in the Eq. (13), and we vary it from 1 to 6 with a step size of 1. The results are reported in Fig. 4. We can observe that on the Baby and Sports datasets, model performance shows an increase with the increasing value of k and reaches a peak when k equals to 3 on the Baby dataset or 5 on the Sports dataset. If the value of k continues to raise, the performance will start to decline. This is because k becomes too large and may introduce additional noise during the denoising step, resulting in inadequate denoising and suboptimal performance. The result demonstrates that our model can effectively capture semantically significant user-item interactions with a proper setting of k, which will provide informative guidance for the reverse diffusion process.

5.5. *Performance with different data sparsity (RQ4)*

To investigate the robustness of the GBDiff model under a varying interaction sparsity, we compared it with two multimodal recommendation baselines (LGMRec Guo et al., 2024 and BM3 Zhou et al., 2023) on the Baby dataset. Specifically, we define multiple user groups based on the number of interactions, ranging from 0 to 25, grouped into 5 intervals of 5 interactions each (1-5, 6-10, 11-15, 16-20, 20-25). The experimental results are illustrated in Fig. 5. First, we can find that both baselines and our model exhibit robustness under different levels of sparsity. When the sparsity degree decreases, the performance of all comparing methods show a trend of decline. We can further observe that our proposed approach GBDiff consistently outperforms the two baselines across all user groups, demonstrating its ability to enhance robustness by constructing granular-balls to capture coarse-grained information and effective denoising in the diffusion process under the guidance of condition.

5.6. *Visualization of user and item embedding distributions (RQ5)*

To verify the quality of fusion features learned by the GBDiff model, we visualized user and item representations from the Baby dataset by performing dimensionality reduction using t-SNE (Ong & Khong, 2025), as shown in Figs. 6 and 7. The embedding distributions are visualized via Gaussian kernel density estimation, and the unit hypersphere S^1 depicts the estimation of arctan(y, x) at the bottom of each figure. The LGMRec model (Guo et al., 2024) is used as a baseline for comparison. In Fig. 6, the user representations learned by the GBDiff model exhibit a more uniformly distributed structure, demonstrating its ability to effectively capture distinct semantic features of each item. In contrast, the distribution of the LGMRec model (Guo et al., 2024) is partially condensed, reflecting high semantic similarity among certain users and indicating representation degradation caused by noise amplification. Fig. 7 demonstrates a similar pattern with previous observations. These results suggest that the GBDiff model excels in modeling user preferences and integrating item features by introducing global coarse behavioral patterns and enhancing individual embeddings through denoised structural propagation.

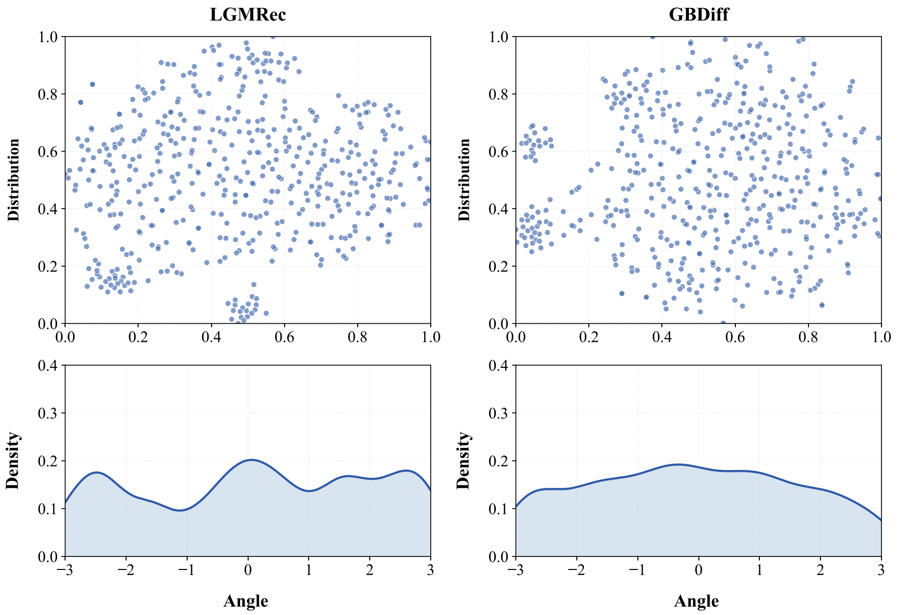


Fig. 6. Distribution of user representations learned from the Baby dataset.

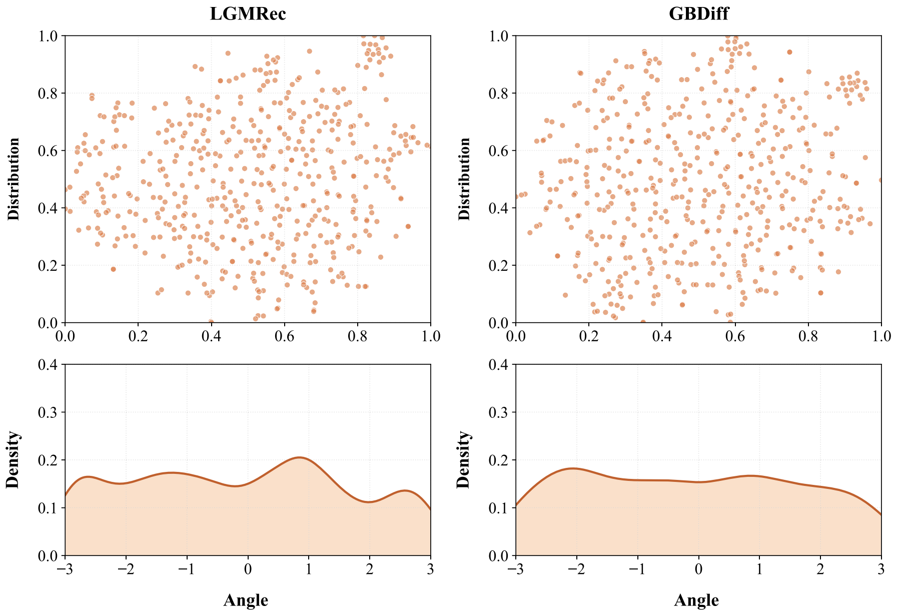


Fig. 7. Distribution of item representations learned from the Baby dataset.

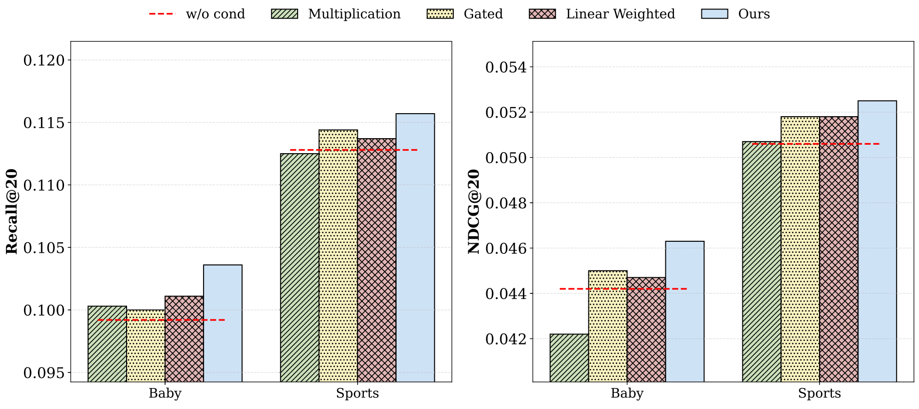


Fig. 8. Impact of conditional fusion strategies.

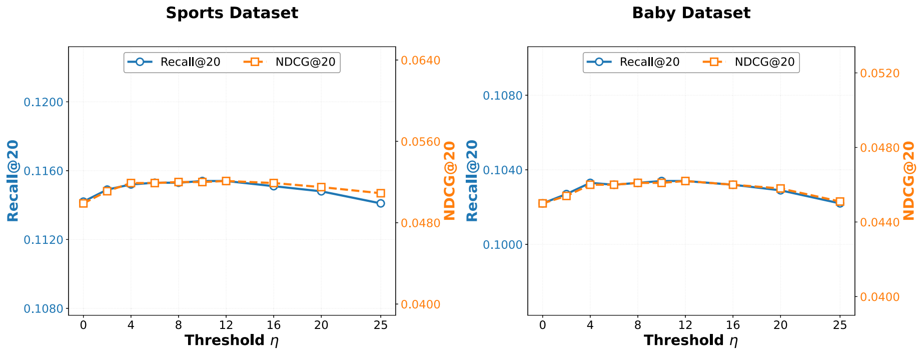


Fig. 9. Impact of threshold η.

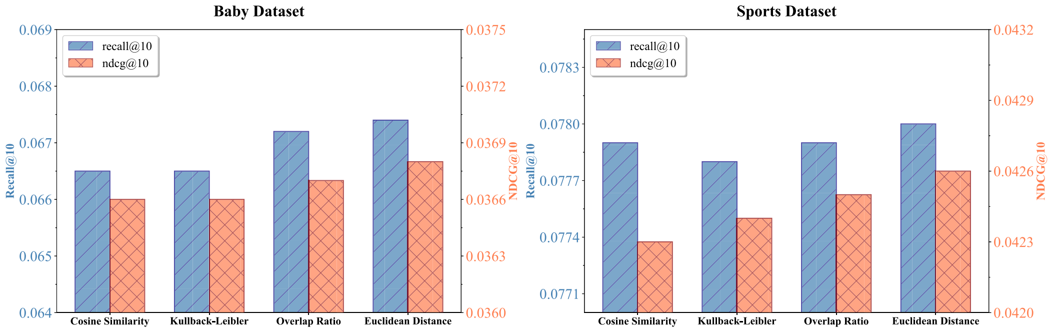


Fig. 10. Impact of granular-ball merging strategies.

5.7. Performance of conditional fusion strategies (RQ6)

To investigate whether the simple addition strategy is optimal for injecting conditional information, we further explore several fusion mechanisms. Specifically, we compare the proposed simple addition strategy with three representative fusion methods, including element-wise multiplication, gated fusion, and linear weighted fusion. Fig. 8 presents the performance comparison of different conditional fusion strategies on two datasets in terms of Recall@20 and NDCG@20. We first observe that all four fusion strategies lead to noticeable performance improvements, indicating that incorporating conditional information is generally beneficial for enhancing diffusion quality and recommendation performance. Among the compared methods, the simple addition strategy consistently achieves the best results across all datasets. In contrast, more complex fusion mechanisms, such as gated fusion and linear weighted fusion, exhibit slight performance degradation on certain datasets. This observation indicates that diffusion-based denoising benefits more from lightweight and stable conditional injection. The simple addition strategy provides an effective way to incorporate multimodal conditional signals.

5.8. Impact of merging threshold η (RQ7)

To investigate the sensitivity of our proposed model to the threshold η, we vary η within the range $\{0, 2, 4, 6, 8, 10, 12, 16, 20, 25\}$ on the Baby and Sports datasets and report the results in Fig. 9. We can observe that GBDiff exhibits an inferior performance when η is very small (e.g., $\eta = 0$ or 2), which could be attributed to excessive fragmentation of granular-balls and improper semantic aggregation. The model achieves a stable peak performance when η in set to values in $\{4, 6, 8, 10, 12\}$ in terms of both Recall@20 and NDCG@20. After that, the performance degrades gradually if we continue to increase η. A similar stability trend is also observed on the Baby dataset, indicating that our proposed model exhibits relatively promising performance across a broad range of η and the proposed granular-ball generation strategy is relatively less sensitive to the setting of η.

5.9. Performance of granular-ball similarity metrics (RQ8)

To investigate the impact of different similarity metrics on granular-ball construction quality, we compare the proposed distance-based strategy (Algorithm 1) with other three similarity metrics:

- **Cosine Similarity**: It computes the cosine similarity between two granular-ball center vectors, and the most similar pair is merged.
- **KL Divergence (Kullback-Leibler Divergence)**: This strategy quantifies the difference between the data distributions represented by two granular balls. Assuming Gaussian distributions within each ball, the pair with the smallest distribution divergence is merged.
- **Overlap Ratio**: It jointly considers the center distance and radius of two granular-balls by measuring their overlap in the embedding space. The overlap between GB_i and GB_j is defined as: $\text{Overlap}(GB_i, GB_j) = \max\left(0, \frac{r_i + r_j - \|\mathbf{c}_i - \mathbf{c}_j\|_2}{r_i + r_j + \epsilon}\right)$, where $\mathbf{c}_i$ and $\mathbf{c}_j$ denote the centers of the two granular-balls, r_i and r_j are their corresponding radii, and ϵ is a small constant for numerical stability. The pair with the highest overlap ratio is selected for merging.
- **Euclidean Distance**: This is the strategy utilized in our model. It employs Euclidean distance to measure geometric proximity between granular-ball center embeddings. A smaller distance indicates closer proximity in the embedding space, and the nearest pair is merged.

The results, presented in Fig. 10, show that cosine similarity and KL divergence yield comparatively weaker performance, particularly on the Baby dataset. The overlap ratio outperforms both of these metrics. Among all methods compared, our Euclidean distance achieves the best performance, consistently surpassing all other similarity metrics on both datasets. These results indicate that Euclidean distance-based merging is more suitable for granular-ball construction. The rationale behind is

Table 4
Efficiency comparison of different recommendation models.

Dataset	Metric	Model				
		LATTICE	FREEDOM	LGMRec	DiffMM	GBDiff
Baby	Memory (GB)	4.53	2.13	2.41	1.90	3.6
	Time (s/epoch)	3.63	1.67	2.63	8.21	12.32
Sports	Memory (GB)	19.93	3.34	3.67	3.58	3.97
	Time (s/epoch)	11.54	3.66	7.12	28.70	24.77

that it directly measures how close the ball centers are in the embedding space, which helps preserve the original local structural information.

5.10. *Efficiency analysis (RQ9)*

To evaluate computational efficiency, we compare memory usage and per-epoch training time of different models on both Baby and Sports datasets. The results are shown in Table 4. Specifically, the memory consumption of GBDiff is comparable to other comparing methods. It costs 3.6 GB on the Baby dataset, which is slightly higher than FREEDOM, LGMRec and DiffMM but much lower than LATTICE. Similar results can be observed on the Sports dataset. Regarding training time, the diffusion-based models, such as DiffMM and GBDiff, is relatively higher than other models. This increase is attributed to the additional computational expense required for the diffusion process. Among both diffusion-based models, the training time of our proposed model GBDiff is higher than that of DiffMM on the larger Sports dataset while lower than it on the small Baby dataset. On the Sports dataset, GBDiff takes 24.77 s/epoch, faster than DiffMM (28.70 s/epoch). The results suggest that the training time of GBDiff remains controlled, especially on larger datasets. Overall, GBDiff keeps the computational cost within an acceptable range while achieving performance improvements, demonstrating its potential to be applied in practical large-scale recommendation scenarios.

6. Conclusion

In this paper, we propose a unified framework named Conditional Graph Diffusion with Granular-Ball Representation for multimodal recommendation (GBDiff). In particular, GBDiff consists of two key modules, including the Modality Guided Conditional Graph Diffusion module and the Granular-Ball Representation Learning module. To strengthen the denoising capability, we propose to apply the modality-specific semantic signals to guide the reverse graph diffusion process. In addition, we introduce the Granular-Ball Computing technique and establish associations between user and item granular-balls to capture coarse-grained collaborative patterns, which will be utilized to complement traditional fine-grained user-item interaction modeling.Extensive experiments on two widely used datasets demonstrate that our proposed approach GBDiff is consistently superior to all state-of-the-art baseline methods.

Despite its effectiveness, GBDiff still has several potential limitations. First, the quality of modality-specific features may affect the reliability of modality-guided diffusion, making the model sensitive to noisy or incomplete modality information. Second, similar to many graph-based recommender systems, in cold-start scenarios with extremely sparse interactions, single structural signals in both the interaction and granular-ball graphs may cause over-smoothing with deeper propagation. This results in embedding homogenization and limits the effective integration of fine-grained and coarse-grained information.

In the future, we plan to explore cross-modality signals rather than modality-specific signals to guide the denoising process. In addition, we will further investigate modality-level denoising strategies to alleviate the influence of noisy or low-quality modality features, thereby improving the robustness and stability of model. Moreover, we will introduce coarse-grained user and item granular-balls into the conventional user-item interaction graph, and propose a novel granular-ball enhanced hypergraph where we can enhance structural diversity and aggregate information from multiple granularities.

CRediT authorship contribution statement

Xiaofei Zhu: Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing; **Ling Tan:** Conceptualization, Methodology, Validation, Investigation, Writing – original draft, Visualization.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (62472059), the Science and Technology Innovation Key R&D Program of Chongqing (CSTB2024TIAD-STX0027), the Chongqing Talent Plan Project, China (CSTC2024YCJH-BGZXM0022), the Open Research Fund of Key Laboratory of Cyberspace Big Data Intelligent Security, Ministry of Education (CBDIS202403).

References

Cao, X., Yang, X., Xia, S., Wang, G., & Li, T. (2024). Open continual feature selection via granular-ball knowledge transfer. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, *36* (12), 8967–8980.

Chen, J., Zhang, H., He, X., Nie, L., Liu, W., & Chua, T. S. (2017). Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention. In *Proceedings of the international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 335–344).

Chen, L. (1982). Topological structure in visual perception. *Science*, *218* (4573), 699–700.

Cheng, D., Li, Y., Xia, S., Wang, G., Huang, J., & Zhang, S. (2024). A fast granular-ball-based density peaks clustering algorithm for large-scale data. *IEEE Transactions on Neural Networks and Learning Systems*, *35* (12), 17202–17215.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the international conference on artificial intelligence and statistics* (pp. 249–256).

Goodfellow, I., Pouget-Abadie, J., Mirza, M. et al. (2020). Generative adversarial networks. *Communications of the ACM*, *63* (11), 139–144.

Guo, Z., Li, J., Li, G. et al. (2024). LGMRec: Local and global graph learning for multimodal recommendation. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)* (pp. 8454–8462). (*vol. 38*).

He, R., & McAuley, J. (2016). VBPR: Visual Bayesian personalized ranking from implicit feedback. In *Proceedings of the thirtieth AAAI conference on artificial intelligence (AAAI)* (pp. 144–150).

He, X., Deng, K., Wang, X. et al. (2020). LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 639–648).

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, *33*, 6840–6851.

Hou, Y., Park, J. D., & Shin, W. Y. (2024). Collaborative filtering based on diffusion models: Unveiling the potential of high-order connectivity. In *Proceedings of the international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 1360–1369).

Hu, J., Hooi, B., He, B., & Wei, Y. (2025). Modality-independent graph neural networks with global transformers for multimodal recommendation. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 11790–11798). (*vol. 39*).

Jiang, Y., Xia, L., Wei, W. et al. (2024a). DiffMM: Multi-modal diffusion model for recommendation. In *Proceedings of the ACM international conference on multimedia* (pp. 7591–7599).

Jiang, Y., Yang, Y., Xia, L. et al. (2024b). DiffKG: Knowledge graph diffusion model for recommendation. In *Proceedings of the ACM international conference on web search and data mining (WSDM)* (pp. 313–321).

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In *Proceedings of 2nd International Conference on Learning Representations (ICLR)*. arXiv: 1312.6114.

Li, Y., Ouyang, X., Pan, C. et al. (2025). Multi-granularity open intent classification via adaptive granular-ball decision boundary. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)* (pp. 24512–24520). (*vol. 39*).

Li, Z., Xia, L., & Huang, C. (2024). RecDiff: Diffusion model for social recommendation. In *Proceedings of the ACM international conference on information and knowledge management (CIKM)* (pp. 1346–1355).

Lin, Z., Tian, C., Hou, Y., & Zhao, W. X. (2022). Improving graph collaborative filtering with neighborhood-enriched contrastive learning. In *Proceedings of the international conference on world wide web (WWW)* (pp. 2320–2329).

Liu, J., Hao, J., Ma, Y., & Xia, S. (2024). Unlock the cognitive generalization of deep reinforcement learning via granular ball representation. In *Proceedings of the 41st international conference on machine learning (ICML)* (pp. 31062–31079). (*vol. 235*).

Luo, X., Cao, J., Sun, T., Yu, J., Huang, R., Yuan, W. et al. (2025). Qarm: Quantitative alignment multi-modal recommendation at kuaishou. In *Proceedings of the 34th ACM international conference on information and knowledge management (CIKM)* (pp. 5915–5922).

Ma, H., Yang, Y., Meng, L. et al. (2024). Multimodal conditioned diffusion model for recommendation. In *Companion proceedings of the ACM web conference (WWW)* (pp. 1733–1740).

McAuley, J., Targett, C., Shi, Q., & Van Den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 43–52).

Ong, R. K., & Khong, A. W. H. (2025). Spectrum-based modality representation fusion graph convolutional network for multimodal recommendation. In *Proceedings of the eighteenth ACM international conference on web search and data mining (WSDM)* (pp. 773–781).

Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2012). BPR: Bayesian personalized ranking from implicit feedback. arXiv preprint, .

Song, Y., Durkan, C., Murray, I., & Ermon, S. (2021a). Maximum likelihood training of score-based diffusion models. In *Advances in neural information processing systems (neurIPS)* (pp. 1415–1428).

Song, Y., & Ermon, S. (2020). Improved techniques for training score-based generative models. In *Advances in neural information processing systems (neurIPS)* (pp. 12438–12448).

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021b). Score-based generative modeling through stochastic differential equations. In *International conference on learning representations (ICLR)* arXiv:2011.13456

Su, P., Huang, S., Ma, W. et al. (2025). Multi-view granular-ball contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)* (pp. 20637–20645). (*vol. 39*).

Tao, Z., Liu, X., Xia, Y., Wang, X., Yang, L., Huang, X., & Chua, T. S. (2022). Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia*, *25*, 5107–5116.

Wang, D., Wang, Q., An, Y., Gao, X., & Tian, Y. (2020). Online collective matrix factorization hashing for large-scale cross-media retrieval. In *Proceedings of the international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 1409–1418). ACM.

Wang, G. (2017). DGCC: Data-driven granular cognitive computing. *Granular Computing*, *2* (4), 343–355.

Wang, Z., Li, J., Xia, S., Lin, L., & Wang, G. (2024a). Text adversarial defense via granular-ball sample enhancement. In *Proceedings of the international conference on multimedia retrieval (ICMR)* (pp. 348–356). 2024a.

Wang, Z. L., Zhang, T., Xia, S. Y., Lin, L. L., & Wang, G. Y. (2024b). GBRAIN: Combating textual label noise by granular-ball based robust training. In *Proceedings of the international conference on multimedia retrieval (ICMR)* (pp. 357–365). 2024b.

Wei, Y., Wang, X., Nie, L., He, X., & Chua, T. S. (2020). Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM international conference on multimedia (ACM MM)* (pp. 3541–3549).

Wei, Y., Wang, X., Nie, L., He, X., Hong, R., & Chua, T. S. (2019). MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the ACM international conference on multimedia* (pp. 1437–1445).

Wu, J., Wang, X., Feng, F., He, X., Chen, L., Lian, J., & Xie, X. (2021). Self-supervised graph learning for recommendation. In *Proceedings of the international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 726–735).

Xia, D., Wang, G., Zhang, Q., Yang, J., & Xia, S. (2024). Three-way approximations fusion with granular-ball computing to guide multigranularity fuzzy entropy for feature selection. *IEEE Transactions on Fuzzy Systems*, *32* (10), 5963–5977.

Xia, S., Lian, X., Wang, G., Gao, X., Chen, J., & Peng, X. (2024). GBSVM: An efficient and robust support vector machine framework via granular-ball computing. *IEEE Transactions on Neural Networks and Learning Systems*, (pp. 1–15). 2024b.

Xia, S., Liu, Y., Ding, X., Wang, G., Yu, H., & Luo, Y. (2019). Granular ball computing classifiers for efficient, scalable and robust learning. *Information Sciences*, *483*, 136–152.

Xia, S., Wang, G., Gao, X. et al. (2023a). Granular-ball computing: An efficient, robust, and interpretable adaptive multi-granularity representation and computation method.

Xia, S., Zhang, H., Li, W., Wang, G., Giem, E., & Chen, Z. (2022). GBNRS: A novel rough set algorithm for fast adaptive attribute reduction in classification. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, *34* (3), 1231–1242.

Xia, S., Zheng, S., Wang, G., Gao, X., & Wang, B. (2023b). Granular ball sampling for noisy label classification or imbalanced classification. *IEEE Transactions on Neural Networks and Learning Systems*, *34* (4), 2144–2155. 2023b.

Xie, J., Cheng, Y., Xia, S., Hua, C., Wang, G., & Gao, X. (2025). AW-GBGAE: An adaptive weighted graph autoencoder based on granular-balls for general data clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (pp. 1–17).

Xie, J., Dai, M., Xia, S., Zhang, J., Wang, G., & Gao, X. (2024a). An efficient fuzzy stream clustering method based on granular-ball structure. In *Proceedings of the international conference on data engineering (ICDE)* IEEE: 901–913. 2024a.

Xie, J., Hua, C., Xia, S., Cheng, Y., Wang, G., & Gao, X. (2024b). W-GBC: An adaptive weighted clustering method based on granular-ball structure. In *Proceedings of the international conference on data engineering (ICDE)* (pp. 914–925). 2024b.

Xie, J., Xiang, X., Xia, S., Jiang, L., Wang, G., & Gao, X. (2024c). MGNR: A multi-granularity neighbor relationship and its application in KNN classification and clustering methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46* (12), 7956–7972. 2024c.

Xu, J., Chen, Z., Yang, S., Li, J., Wang, W., Hu, X., Hoi, S., & Ngai, E. (2026). A survey on multimodal recommender systems: Recent advances and future directions, arXiv: 2502.15711.

Yang, J., Liu, X., Wang, G. et al. (2025a). A robust three-way classifier with shadowed granular balls based on justifiable granularity. *IEEE Transactions on Neural Networks and Learning Systems*, *36* (9), 16534–16548.

Yang, J., Liu, X., Wang, G. et al. (2026). A three-way incremental granular-ball classifier using shadowed set. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Yang, J., Liu, Z., Xia, S., Wang, G., Zhang, Q., Li, S., & Xu, T. (2024). 3WC-GBNRS++: A novel three-way classifier with granular-ball neighborhood rough sets based on uncertainty. *IEEE Transactions on Fuzzy Systems*, *32* (8), 4376–4387.

Yang, J., Lu, F., Wang, G. et al. (2025b). Three-way outlier detection based on shadowed granular-balls. *IEEE Transactions on Fuzzy Systems*, *34* (1), 101–113.

Yang, J., Zhao, F., Wang, G., Pedrycz, W., Xia, S., Liu, Y., & Zhang, Q. (2025c). CS3W-GBG: A cost-sensitive three-way granular-ball generation method. *IEEE Transactions on Fuzzy Systems*, *33* (10), 3681–3694.

Zhang, J., Zhu, Y., Liu, Q., Wu, S., Wang, S., & Wang, L. (2021). Mining latent structures for multimedia recommendation. In *Proceedings of the ACM international conference on multimedia* (pp. 3872–3880).

Zhang, J., Zhu, Y., Liu, Q., Zhang, M., Wu, S., & Wang, L. (2023a). Latent structure mining with contrastive modality fusion for multimedia recommendation. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, *35* (9), 9154–9167.

Zhang, Q., Wu, C., Xia, S., Zhao, F., Gao, M., Cheng, Y., & Wang, G. (2023b). Incremental learning based on granular ball rough sets for classification in dynamic mixed-type decision system. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, *35* (9), 9319–9332.

Zhao, J., Wenjie, W., Xu, Y. et al. (2024). Denoising diffusion recommender model. In *Proceedings of the international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 1370–1379).

Zhou, X., & Shen, Z. (2023). A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM international conference on multimedia* (pp. 935–943).

Zhou, X., Zhou, H., Liu, Y., Zeng, Z., Miao, C., Wang, P., You, Y., & Jiang, F. (2023). Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the web conference (WWW)* (pp. 845–854).

Zhu, Y., Wang, C., Zhang, Q. et al. (2024). Graph signal diffusion model for collaborative filtering. In *Proceedings of the international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 1380–1390).