Sign-Aware Multimodal Graph Recommendation

Yahong Lian[1], Haotian Tian[2], Chunyao Song[1*], Tingjian Ge [3]

[1]College of Computer Science, TJ Key Lab of NDST, DISSec, TMCC, TBI Center, Nankai University, Tianjin, China
[2]College of Software, Nankai University, Tianjin, China
[3]Department of Computer Science, University of Massachusetts Lowell, MA, USA
{yahong.lian, tedhao}@mail.nankai.edu.cn, chunyao.song@nankai.edu.cn, ge@cs.uml.edu

Abstract

A multimodal recommendation system (MRS), which leverages rich multimodal information to model user preferences, has recently attracted significant research interest. Most existing MRSs focus primarily on developing sophisticated encoders for feature extraction, typically relying on simple aggregation of interaction-based features for final predictions. However, this conventional paradigm fails to account for the critical semantic difference between high- and low-rating interactions: while high ratings indicate user preference, low ratings explicitly convey dissatisfaction. Such oversight of negative feedback semantics may significantly limit the system's recommendation performance. Recently, sign graphs—which model positive and negative feedback signals separately—have gained considerable attention. Inspired by this approach, we propose Sign-Aware Multimodal Graph Recommendation (SiMGR), a novel framework incorporating signed graphs into multimodal recommendation systems. SiMGR fuses multimodal features with signed interactions in a unified graph framework by integrating modality-specific representations and applying user-level thresholds to separate positive and negative subgraphs. A balanced pseudo-edge augmentation strategy is introduced to alleviate sparsity and enhance generalization. Experiments on three public multimodal recommendation datasets show that SiMGR outperforms state-of-the-art baselines, achieving an average 4.28% improvement in NDCG@20.

1 Introduction

Recommendation systems (RS) constitute a fundamental component of online ecosystems, enabling services to filter information and deliver personalized content based on user preferences and behavioral patterns (Aljunid, Manjaiah et al. 2025; Anand and Maurya 2025). The explosive growth of multimedia content on digital platforms has driven the development of advanced multimodal recommendation systems (MRSs) (Xu et al. 2025b; Malitesta et al. 2025), which utilize rich item-side information to build more comprehensive models, surpassing unimodal methods and overcoming their inherent limitations (Liu, Li, and Nie 2025; Liao et al. 2025). Current approaches (Guo et al. 2024; Xu et al. 2025a) capitalize on the expressive power of Graph Convolutional Networks (GCNs) to model user-item interactions and expand them into item-item graphs enriched with multimodal information.

However, existing MRSs often overlook negative interactions (e.g., post-purchase low ratings), mistakenly treating disliked items as preferred and reducing model accuracy (Heshmati et al. 2025; Liu et al. 2025a). To further substantiate that user-item interactions do not inherently represent positive feedback signals, we conducted two sets of controlled experiments. The first configuration followed conventional practices by treating all user ratings (spanning 1-5) as positive interactions (*green* bars in Figure 1), while the second configuration (*blue* bars in Figure 1) retained exclusively high-rating interactions (scores 4 and 5) for training. The experimental results demonstrate that low-rating data adversely affects recommendation precision.

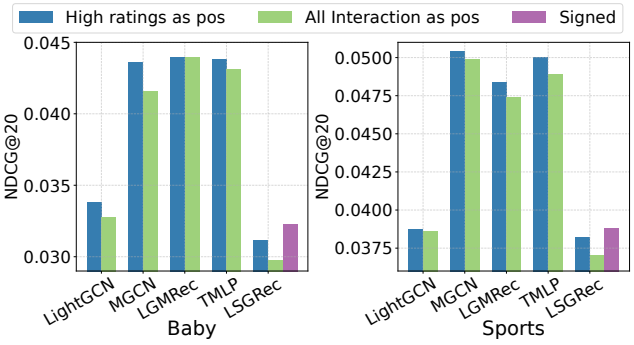

Figure 1: The effect of negative interaction data on Baby and Sports datasets

By distinguishing and encoding positive and negative interactions separately, sign-aware models (Derr et al. 2018; Zhang et al. 2024; Heshmati et al. 2025) better capture user preferences compared to unsigned approaches (Seo et al. 2022; Huang et al. 2023; Liu et al. 2025a). Figure 1 shows that LSGRec (Liu et al. 2025a)'s performance improves significantly when negative feedback is incorporated (*purple* bars) versus positive-only feedback variants (*blue*/*green* bars). Our results reveal significant performance gaps between unimodal approaches (e.g., signed method LSGRec and unsigned method LightGCN) and multimodal methods.

*Corresponding Author

This suggests that incorporating both positive and negative feedback into multimodal frameworks could substantially improve recommendation performance.

Nevertheless, integrating positive and negative feedback into MRSs remains non-trivial challenges. (i) The appropriate modeling of items' auxiliary information in conjunction with newly introduced signed graph structures is paramount—inadequate synergy between these components may result in detrimental effects on system performance. (ii) Current signed graph partitioning methods rely on rigid system-level thresholds (e.g., ratings<3 as negative), ignoring personalized rating behaviors. In e-commerce systems, generous raters may express dissatisfaction through relatively low scores within their typically high-rating patterns, while stringent raters might indicate satisfaction through relatively high scores within their usual low-rating tendencies. For example, a rating of 3 could signify satisfaction for a strict rater (User A) but dissatisfaction for a lenient one (User B). This limitation underscores the need for personalized threshold mechanisms that consider fine-grained user-specific rating patterns. (iii) Decomposing user-item graphs into signed positive and negative subgraphs leads to sparsity, requiring tailored data augmentation to enhance graph connectivity through additional edge generation.

To this end, we propose SiMGR, a novel Sign-aware Multimodal Graph Recommendation framework that captures user preferences more precisely than traditional implicit modeling. It unifies multimodal item features (e.g., textual and visual) with signed interactions in a structured graph-based architecture, enabling joint optimization of preference learning and signal alignment. The Signed Graph Augmentation (SGA) module explicitly integrates positive and negative user-item interactions into the multimodal recommendation process. To account for individual differences in feedback interpretation, we introduce the Personalized Sign Threshold (PST) module, which adaptively assigns thresholds based on historical behavior to distinguish genuine positive and negative signals at the user level, thereby enhancing personalized preference modeling. To address signed graph sparsity, we design the Adaptive Pseudo Feedback Injection (APFI) module, which selectively augments the positive and negative adjacency matrices to enrich the graph structure. Particularly: (1) For tail items with limited exposure opportunities, we introduce pseudo-positive edges when they exhibit strong semantic alignment with user preferences, compensating for their sparse interaction patterns; (2) For frequently exposed head items, we construct pseudo-negative edges when they demonstrate significant semantic divergence from user preferences, interpreting their interaction absence as potential dislike. This exposure-aware edge augmentation framework simultaneously mitigates the cold-start problem for tail items while calibrating recommendation relevance for head items.

The key contributions are summarized as follows:

- To the best of our knowledge, we are the first to propose SiMGR, a signed multimodal graph recommendation framework that explicitly models positive/negative feedback through signed graphs to fully leverage user rating semantics. It integrates these signals with multimodal data via a novel Signed Graph Augmentation (SGA) module.
- We introduce the Personalized Sign Threshold (PST) module to model user-specific preference boundaries and design the Adaptive Pseudo Feedback Injection (APFI) module to address graph sparsity by adding tailored pseudo-edges.
- Extensive experiments on three public datasets show SiMGR outperforms state-of-the-art baselines by 4.36% in $Recall@20$ and 4.28% in $NDCG@20$ on average.

2 Related Work

Multi-modal recommendation system has become a fundamental infrastructure of today's internet platforms. By modeling multimodal information and user-content interactions, it delivered accurate personalized recommendations. In graph-based recommendation methods, some studies constructed user-item bipartite graphs (Wei et al. 2019; Yi et al. 2022; Guo et al. 2024) to represent user preferences from interaction data. Another line of research focused on constructing item-item relationships (Lei et al. 2023; Xu et al. 2025a) to capture modality-aware collaborative relationships. In addition, some works improved models through higher-order modality processing (Shang et al. 2023; Yi et al. 2024a; Jiang et al. 2024; Liu and Lu 2025; Peng, Fu et al. 2025), effectively enhancing recommendation performance. Recently, researchers have delved deeper into multimodal information (Yi et al. 2024b; Xu et al. 2024; Wang et al. 2025; Wang, Liang et al. 2025), leveraging important semantic knowledge from each modality to enhance item representations. For instance, TMLP (Huang et al. 2025) uses a pure MLP to model user-item interactions and multimodal features jointly. Moreover, many GNN-based recommendation systems focus on high-rating interactions to improve accuracy, but effectively modeling low ratings remains challenging. Early works (Derr et al. 2018; Kim, Lee et al. 2023) introduced sign-aware graph recommendation models using balance theory for more accurate user-item signal propagation. Other approaches (Shen et al. 2018) focused on reconstructing sparse negative links. Subsequent methods (Huang et al. 2021; Zhang, Liu et al. 2023; Seo et al. 2022; Liu et al. 2025a) constructed signed bipartite graphs incorporating both positive and negative edges to enhance representation learning. Recent advances (Chen et al. 2024; Liu et al. 2025b) further improved performance through graph denoising and architectural optimization.

3 Preliminaries

Symbol and Notation Setup. In RS, the core input is a set of historical user-item interactions with associated ratings. Let $u \in \mathcal{U}$ and $i \in \mathcal{I}$ denote a user and an item, respectively. The capacity is $M = |\mathcal{U}|, N = |\mathcal{I}|$. We construct a user-item rating matrix $R \in \{0, w_{u,i}\}^{M \times N}$, where $w_{u,i}$ represents the rating given by user u to item i. Then, a bipartite graph $\mathcal{G} = (\mathcal{U} \cup \mathcal{I}, \mathcal{E})$ is formulated to model the interaction structure. $\mathcal{N}_u$ denotes items directly interacted with by user u, and $\mathcal{N}_i$ denotes users directly interacted with item i. In non-signed graph recommendation methods, $w_{u,i} = 1$, yielding R_b, a

binary interaction graph. For simplicity, we define $\mathcal{A} \Leftarrow R_b$ operation as: $\mathcal{A} = \begin{pmatrix} 0 & R_b \\ R_b^T & 0 \end{pmatrix} \in \mathbb{R}^{(M+N)\times(M+N)}$, where $\mathcal{A}$ is designated as the adjacency matrix.

Multimodal recommendation system (MRS). It jointly models interactions and multimodal item features, where $H^v \in \mathbb{R}^{N\times d^v}$ (visual) and $H^t \in \mathbb{R}^{N\times d^t}$ (textual) are associated with individual item representations $\mathbf{h}_i^v \in \mathbb{R}^{d^v}$ and $\mathbf{h}_i^t \in \mathbb{R}^{d^t}$, respectively. Given $\{\mathcal{G}, \mathcal{A}, H^v, H^t\}$, MRS generates user and item embeddings $E^* \in \mathbb{R}^{(M+N)\times d}$ by fusing interaction graph embeddings (IGE) and modality graph embeddings (MGE), with the Bayesian Personalized Ranking (BPR) (Rendle et al. 2009) loss. It is formulated as follows:

$$\mathcal{L}_{BPR} = -\sum_{(u,i,j)} ln\sigma(y_{ui} - y_{uj}), \tag{1}$$

where $y_{ui} = e_u^{*T} \cdot e_i^*$, and e_u^*, e_i^i denote individual user and item embeddings extracted from E^*. Also, σ is the sigmoid function. Here, $(u,i) \in \{R_b | w_{u,i} = 1\}$ is the historically interacted user-item pairs, and (u,j) is randomly sampled non-interacted pairs.

Interaction Graph Embedding. MRSs model high-order connectivity patterns via message propagation on the user-item graph, with ID embeddings serving as initial node representations ($E^0 = E^{id}$). Formally:

$$E^l = (D^{-\frac{1}{2}}\mathcal{A}D^{-\frac{1}{2}})E^{l-1}, E_{ig}^{id} = \frac{1}{L+1}\sum_{l=0}^{L} E^l \tag{2}$$

A lightweight Graph Convolutional Network (LGCN) (He et al. 2020a) is used, where D is the diagonal matrix of $\mathcal{A}$. Hidden layers are aggregated via mean pooling to obtain $E_{ig}^{id} \in \mathbb{R}^{(M+N)\times d}$, modeling interaction-based information.

Modality Graph Embedding. Given the semantic disparities among modalities and the heterogeneous dimensionalities of pre-trained feature extractors, MRS projects multimodal features h_i^v, h_i^t of items into a shared embedding space $\mathbb{R}^d$ to facilitate consistent modeling on the interaction graph, as: $\mathring{h}_i^v = W_v^T \cdot h_i^v, \mathring{h}_i^t = W_t^T \cdot h_i^t$, where $W_v \in \mathbb{R}^{d^v\times d}, W_t \in \mathbb{R}^{d^t\times d}$ are trainable parameters. For each user u, we aggregate their first-order neighbors to derive multimodal representation: $\mathring{h}_u^v = 1 \,/\, |\mathcal{N}_u| \sum_{i\in\mathcal{N}_u} \mathring{h}_i^v$, with $\mathring{h}_u^t$ obtained similarly. Let $m \in \{v,t\}$, we concatenate the user and item embedding $\mathring{h}_u^m, \mathring{h}_i^m$ to form the multimodal matrix $\mathring{H}^m \in \mathbb{R}^{(M+N)\times d}$. These are used as input to the LGCN (He et al. 2020a), producing modality-specific embeddings via: $\mathring{E}^{m,p} = (D^{-\frac{1}{2}}\mathcal{A}D^{-\frac{1}{2}})\mathring{E}^{m,p-1}$, $p \in \{1, ...P\}$, with $\mathring{E}^{m,0} = \mathring{H}^m$. The final layer $\mathring{E}^{m,P}$ captures modality-related information. The two modalities are fused using mechanisms such as weighted (Yu et al. 2023) or linear attention (Guo et al. 2024), resulting in $\mathring{E} = Fuse(\mathring{E}^{v,P}, \mathring{E}^{t,P})$. Finally, MRS combines interaction graph embedding and modality graph embedding to obtain final representations, as follows:

$$E^* = E_{ig}^{id} + \mathring{E} \tag{3}$$

With E^*, the overall loss of MRS is calculated as: $\mathcal{L} = \mathcal{L}_{BPR} + \lambda_r ||\Theta||$, where Θ denotes all learnable parameters, and λ_r is the regularization coefficient.

Sign-aware Recommendation System (SRS). In multimodal recommendation method, the interaction graph $\mathcal{G}$ is built with binary edges indicating user-item interaction. A signed graph, by contrast, retains actual ratings and splits into positive and negative subgraphs $\mathcal{G}^+ = \{\mathcal{U} \cup \mathcal{I}, \mathcal{E}^+\}$ and $\mathcal{G}^- = \{\mathcal{U} \cup \mathcal{I}, \mathcal{E}^-\}$ based on a predefined threshold δ, where $\mathcal{E}^+ = \{(u,i,1) | w_{ui} - \delta > 0, (u,i,w_{ui}) \in \mathcal{E}\}$, $\mathcal{E}^- = \{(u,i,1) | w_{ui} - \delta \le 0, (u,i,w_{ui}) \in \mathcal{E}\}$. The binary signed graphs R_b^+ and R_b^- generate feedback-specific embeddings, which are subsequently fused into the final representations e_u^*, e_i^*, using a sign cosine loss (Huang et al. 2023). The loss function is formulated as follows:

$$\mathcal{L}_{sign} = \begin{cases} 1 - cos(e_u^*, e_i^*), \ if \ (u,i) \in \mathcal{E}^+ \\ \gamma \cdot max(0, cos(e_u^*, e_i^*) - \psi), \ if \ (u,i) \in \mathcal{E}^- \end{cases}, \tag{4}$$

where γ and ψ are hyperparameters. Therefore, the overall loss function of the sign-aware recommendation system is formulated as: $\mathcal{L} = \mathcal{L}_{BPR} + \lambda_s \mathcal{L}_{sign} + \lambda_r ||\Theta||$, where λ_s denotes the coefficient of sign-related loss.

4 Methodology

In this section, we present SiMGR through a detailed description of its components. The overall framework of the SiMGR is shown in Figure 2, consisting of Signed Graph Augmentation, Personalized Sign Threshold, and Adaptive Pseudo Feedback Injection.

4.1 Signed Graph Augmentation (SGA)

This work introduces a novel approach by leveraging signed graphs to separate user-item interactions (rated from 1 to 5) into positive and negative feedback by setting $\delta = 3$ as in previous works (Liu et al. 2025a; Huang et al. 2023), thereby improving the modeling of users and items. We propose an augmented graph structure based on signed information and employ additional graph convolution operations to provide stronger support for multimodal recommendation tasks.

To encode the positive graph $\mathcal{G}^+$, we calculate following equation:

$$\dot{E}^{+(l+1)} = D^{+-\frac{1}{2}}\mathcal{A}^+D^{+-\frac{1}{2}}\dot{E}^{+(l)}, \tag{5}$$

where $\mathcal{A}^+ \Leftarrow R_b^+$, and $\mathcal{A}^+ \in \mathbb{R}^{(M+N)\times(M+N)}$ is the positive adjacency matrix. D^+ is the diagonal matrix with $D_{ii}^+ = \sum_j \mathcal{A}_{ij}^+$. Then, we get the final positive embedding as:

$$\ddot{E}^+ = \frac{1}{(L+1)}\sum_{l=0}^{L} \dot{E}^{+(l)}, \tag{6}$$

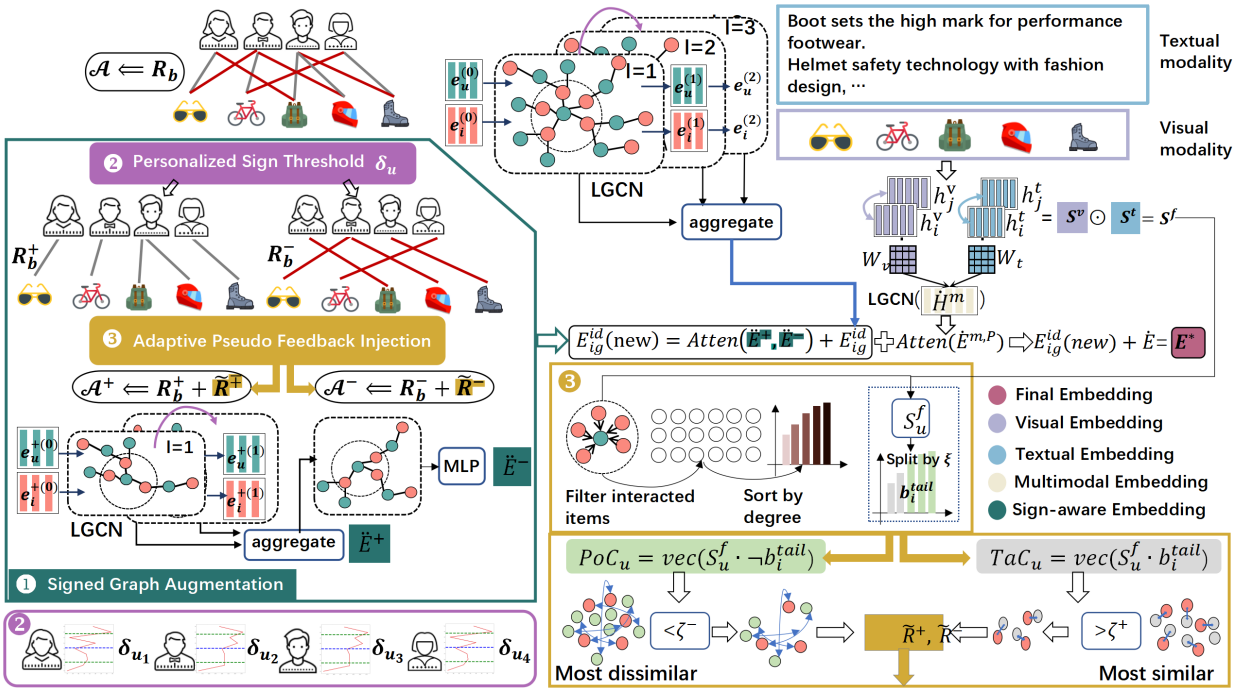

Figure 2: The framework of SiMGR

where $\dot{E}^{+(l)}$ is the trainable embeddings at the l-th layer. Equations (5) and (6) can be rewritten as:

$$\dot{e}_u^{+(l+1)} = \sum_{i \in \mathcal{N}_u^+} \frac{1}{\Gamma_{ui}} \dot{e}_i^{+(l)}, \dot{e}_i^{+(l+1)} = \sum_{u \in \mathcal{N}_i^+} \frac{1}{\Gamma_{ui}} \dot{e}_u^{+(l)}, \quad (7)$$

where $\Gamma_{ui} = \sqrt{|\mathcal{N}_u^+||\mathcal{N}_i^+|}$, $\mathcal{N}_u^+$ and $\mathcal{N}_i^+$ denote the sets of items connected to user u and users connected to item i in the positive graph $\mathcal{G}^+$, respectively.

And then, the final positive collaborative embeddings can be obtained by a mean pooling operation as follows:

$$\ddot{e}_u^+ = \frac{1}{(L+1)} \sum_{l=0}^{L} \dot{e}_u^{+(l)}, \ddot{e}_i^+ = \frac{1}{(L+1)} \sum_{l=0}^{L} \dot{e}_i^{+(l)}, \quad (8)$$

where $\ddot{e}_u^+$ and $\ddot{e}_i^+$ denote the final positive representations of user u and item i, respectively. For negative feedback, we propose to apply a new module to encode the negative edges, which consists of one GNN layer and an MLP layer, formally:

$$\ddot{E}^- = MLP(D^{-\frac{1}{2}} \mathcal{A}^- D^{-\frac{1}{2}} \ddot{E}^+), \quad (9)$$

where $\mathcal{A}^- \Leftarrow R_b^-$, and $\mathcal{A}^- \in \mathbb{R}^{(M+N)\times(M+N)}$ is the negative adjacency matrix, D^- is the diagonal matrix. In equation (9), we aggregate the final negative embeddings $\ddot{E}^-$ by passing the positive embedding $\ddot{E}^+$ through a GNN layer for structural aggregation, followed by an MLP to refine the representation. The MLP helps model heterophilic interactions, where linked nodes are more likely to be dissimilar (Zhu, Yan et al. 2020).

Then, we conduct a linear attention layer to combine both full user-item interaction representations E_{ig}^{id}, positive interaction representations $\ddot{E}^+$, and negative interaction representations $\ddot{E}^-$, as follows:

$$E_{ig}^{id}(new) = E_{ig}^{id} + ReLU(w^+ + w^-), \quad (10)$$

where w^+, w^- are calculated as follows:

$$w^+ = \mathbf{a}[0] \odot \ddot{E}^+, w^- = \mathbf{a}[1] \odot \ddot{E}^-, \mathbf{a} = \mathcal{F}_s([\mathbf{a}^+, \mathbf{a}^-]),$$
$$\mathbf{a}^+ = \frac{\ddot{E}^+ \ddot{W}^+}{\ddot{E}^+ \ddot{W}^+ + \ddot{E}^- \ddot{W}^-}, \mathbf{a}^- = \frac{\ddot{E}^- \ddot{W}^-}{\ddot{E}^+ \ddot{W}^+ + \ddot{E}^- \ddot{W}^-}, \quad (11)$$

where $\ddot{W}^+, \ddot{W}^- \in \mathbb{R}^d$ are learnable weights, and $\mathcal{F}_s$ denotes $Softmax$ function. Thereafter, we replace the first term in Equation (3) with $E_{ig}^{id}(new)$, thereby obtaining the final embedding representation.

4.2 Personalized Sign Threshold (PST)

Although SGA can effectively leverage user ratings (i.e., explicit feedback), individual users often exhibit distinct rating behaviors. To model such personalized preference expression, we propose a customized PST module, which adaptively determines user-specific thresholds to differentiate between positive and negative interactions.

Based on the original user-item rating matrix R, we then define a boolean mask matrix M, where:

$$M_{ui} = \mathbb{I}(R[u,i] > 0), \quad (12)$$

where $\mathbb{I}(\cdot)$ is an indicator function returning 1 if the condition holds, and 0 otherwise. The average score and the stan-

dard deviation are calculated as follows:

$$\mu_u = \frac{\sum_{i=1}^{n} R_{ui} M_{ui}}{\sum_{i=1}^{n} M_{ui}}, \varepsilon_u = \sqrt{\frac{\sum_{i=1}^{n} (R_{ui} - \mu_u)^2 M_{ui}}{\sum_{i=1}^{n} M_{ui}}} \tag{13}$$

Then the final threshold of user u is:

$$\delta_u = \nu * \mu_u - (1 - \nu) * \varepsilon_u, \tag{14}$$

where ν is a hyperparameter, controlling the influence of user mean ratings and standard deviation (see Section 5.4 for tuning). SiMGR substitutes the threshold δ with a user-specific threshold δ_u.

4.3 Adaptive Pseudo Feedback Injection (APFI)

We propose an APFI module to enhance sparse signed graphs using item multimodal knowledge. Pseudo-positive edges provide plausible positive feedback signals to uninteracted items that are semantically aligned with user preferences, enabling them to receive meaningful information propagation paths. Pseudo-negative edges, on the other hand, supplement reasonable "non-interest" signals where true negative samples are scarce, preventing bias caused by insufficient negative interactions. Together, they balance positive and negative signals, promoting stable and informative embedding representations.

Firstly, we construct a tail/popular mask to sort the items by their degrees in ascending order: $\mathcal{I}^{sorted} = sort(\mathcal{I}, d_i)$. Using a predefined threshold ratio ξ (a hyperparameter detailed in Section 5.4), we obtain the number of tail items as: $N_{tail} = \lfloor \xi \cdot |\mathcal{I}| \rfloor$. The item set is then partitioned into two subsets: $\mathcal{I}_{tail} = \{i_{idx} \in \mathcal{I}^{sorted} | idx \leq N_{tail}\}$, $\mathcal{I}_{pop} = \mathcal{I} \backslash \mathcal{I}_{tail}$. We then generate binary masks for popular and tail items as:

$$b_i^{tail} = \begin{cases} True, & if \quad i \in \mathcal{I}_{tail} \\ False, & otherwise \end{cases}, \tag{15}$$

where $b_i^{tail} \in \mathbb{B}^N$. Then, item similarities are computed from multimodal features as follows:

$$S^f = S^v \odot S^t, s_{i,j}^m = \frac{h_i^{m^T} \cdot h_j^m}{||h_i^m|| ||h_j^m||}, m \in \{v, t\}, \tag{16}$$

where $S^v, S^t \in \mathbb{R}^{N \times N}$ represent the item-item similarity from visual and textual modalities, respectively. Each element of the S^m matrix is calculated with a cosine similarity.

For individual user u, the interacted items are:

$$S_u^f = \{S_{[q,:]}, q \in \mathcal{N}_u\}, S_u^f \in \mathbb{R}^{(|\mathcal{N}_u| \times N)}, \tag{17}$$

multiply by the tail or popular mask $TaC_u = S_u^f \cdot b_i^{tail}, PoC_u = S_u^f \cdot \neg b_i^{tail}$. We flatten the 2D $|\mathcal{N}_u| \times N$ matrix into a 1D vector of shape $1 \times |\mathcal{N}_u| N$ using $vec(\cdot)$: $TaC_u = vec(TaC_u), PoC_u = vec(PoC_u)$.

For users linked to tail items, we select the most similar uninteracted items and inject them into the positive adjacency matrix $\mathcal{A}^+$. The positive pseudo-injection (PoPI) is defined as follows:

$$PoPI(u)_w, PoPI(u) = top_Q(TaC_u) \tag{18}$$

Here, Q is set to 20 for all settings. Then, we build the pseudo-positive interaction matrix $\widetilde{R}^+$ as:

$$\widetilde{R}^+ = \begin{cases} 1 & i \in PoPI(u) \ \& \ PoPI(u)_w > \zeta^+ \\ 0 & otherwise \end{cases}, \tag{19}$$

where ζ^+ is a hyperparameter discussed in Section 5.4. By leveraging semantic alignment in multimodal embedding space, we identify uninteracted tail items with strong user relevance. These items are incorporated as pseudo-positive connections into the positive adjacency matrix R_b^+, enhancing the model's ability to capture implicit user preference boundaries. For each user linked to popular items, we select the most dissimilar uninteracted items and add them to the negative adjacency matrix $\mathcal{A}^-$. The negative pseudo-injection (NePI) is formulated as follows:

$$NePI(u)_w, NePI(u) = top_Q(-PoC_u), \tag{20}$$

similarity, we build the pseudo-negative interaction matrix $\widetilde{R}^-$ as:

$$\widetilde{R}^- = \begin{cases} 1 & i \in NePI(u) \ \& \ NePI(u)_w < \zeta^- \\ 0 & otherwise \end{cases}, \tag{21}$$

where ζ^- is a hyperparameter discussed in Section 5.4. By adding pseudo-negative edges for items with substantial semantic divergence from user history, we reduce bias from highly visible items. This strategy decreases model attention on irrelevant candidates, enhancing focus. Finally, they are integrated as $R_b^+ \leftarrow R_b^+ + \widetilde{R}^+$ and $R_b^- \leftarrow R_b^- + \widetilde{R}^-$ to enrich the graph structure for collaborative learning.

4.4 Optimization

To mitigate the inconsistency between user-item and multimodal embeddings, we employ a contrastive learning (CL) objective using InfoNCE loss (Oord et al. 2018) to enhance preference modeling and enforce representation alignment. User side CL loss is formulated as:

$$\mathcal{L}_{cl}^u = \sum_{u \in \mathcal{U}} -log \frac{exp(e_u^{id} \cdot e_u^* / \tau)}{\sum_{z \in \mathcal{U}} exp(e_z^{id} \cdot e_z^* / \tau)}, \tag{22}$$

where e_u^* is from final representations E^* and e_u^{id} comes from the interaction graph embedding $E_{ig}^{id}(new)$. Temperature τ is set to 0.5. The item side CL loss $\mathcal{L}_{cl}^i$ follows Equation (22) similarly. The overall loss is as follows:

$$\mathcal{L} = \mathcal{L}_{BPR} + \lambda_s \mathcal{L}_{sign} + \beta_{cl}(\mathcal{L}_{cl}^u + \mathcal{L}_{cl}^i) + \lambda_r ||\Theta|| \tag{23}$$

Where β_{cl} controls the intensity of the contrastive loss.

5 Experiments and Analysis

We conducted experiments on three real-world datasets to address the following research questions: **RQ1:** How does SiMGR perform relative to state-of-the-art baselines? **RQ2:** How do SiMGR's key components influence recommendation accuracy? **RQ3:** How does hyperparameter variation affect SiMGR's performance?

5.1 Experiment Setup

Datasets. We evaluate different methods on Amazon review datasets (He and McAuley 2016), selecting three real-world domains: Baby, Sports, and Electronics. All datasets undergo 5-core filtering and are split into train/test/validation sets at an 8:1:1 ratio. Statistical summaries are presented in Table 1.

Dataset	#User	#Items	#Interactions	Density
Baby	19,445	7,050	160,792	0.117%
Sports	35,598	18,357	296,337	0.045%
Electronics	20,247	11,589	347,393	0.148%

Table 1: Dataset statistics

Baselines. We assess the effectiveness of our method by comparing it with top-performing approaches in three categories. (1) *Classical Recommendation Methods*: **BPR** (Rendle et al. 2009), and **LightGCN** (He et al. 2020b). (2) *Signed-Graph-Based Recommendation Methods*: **SiReN** (Seo et al. 2022), **SiGRec** (Huang et al. 2023), and **LSGRec** (Liu et al. 2025a). (3) *Multi-modal Recommendation Methods*: **MGCN** (Yu et al. 2023), **BM3** (Zhou et al. 2023), **Freedom** (Zhou and Shen 2023), **LGMRec** (Guo et al. 2024), **DiffMM** (Jiang et al. 2024), and **TMLP** (Huang et al. 2025).

Evaluation Protocol. We employ $Recall@K$ ($R@K$) and $NDCG@K$ ($N@K$) as evaluation metrics, where $K \in \{10, 20\}$. Early stopping is implemented when Recall@20 on the validation set fails to improve for 20 consecutive epochs.

Implementation Details. All baselines and SiMGR are implemented using the MMRec (Zhou 2023) . The learning rate is set as 0.001. The embedding size is set to 64, and the batch size is 2048. The GCN layers for the user-item graph are set to 2. Embedding parameters are initialized by the Xavier (Glorot and Bengio 2010) and optimized models with the Adam optimizer (Kingma and Ba 2015). Experiments are implemented on NVIDIA RTX 3090. Code is given here [1].

5.2 Overall Performance (RQ1)

Table 2 presents Recall and NDCG metrics for multiple models across three benchmark datasets to evaluate method effectiveness. The results demonstrate: (1) SiMGR outperforms existing multimodal baselines across all datasets, achieving 5.1%, 4.1%, and 3.8% improvements over the second-best model in $R@20$. This performance gain stems from sign-graph augmentation's capture of fine-grained positive/negative information in users' historical preferences, while PST and APFI modules enhance signed user-item interaction graph representations. (2) Certain multimodal baselines (e.g., LGMRec) perform well on Baby and Electronics datasets but underperform on Sports due to sparser user-item interactions. While LGMRec captures global user preferences through hypergraph construction, it yields suboptimal results by ignoring negative feedback information in user-item interaction graphs. Conversely, SiMGR achieves superior recommendation performance across datasets with varying interaction densities. (3) Comparison between traditional collaborative filtering methods (unsigned: LightGCN; signed: SiReN, SiGRec, LSGRec) and multimodal methods reveals that multimodal approaches substantially outperform traditional methods across all evaluation metrics. This confirms that incorporating multimodal information significantly enhances recommendation performance.

[1] https://github.com/DRec4AI/SiMGR2025

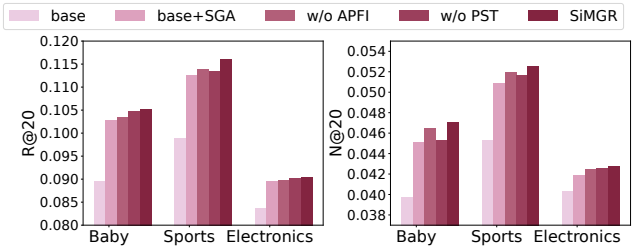

Figure 3: The different components of SiMGR

5.3 Ablation Study (RQ2)

This section presents a comprehensive experimental evaluation of SiMGR's effectiveness across multiple scenarios and configurations. SiMGR comprises three core modules: SGA, PST, and APFI. The ablation study systematically evaluates each component's contribution through four configurations: (1) **Base**: baseline model without proposed modules; (2) **Base + SGA**: baseline enhanced with SGA only; (3) **Base + SGA + PST (w/o APFI)**: baseline with SGA and PST modules, excluding APFI; (4) **Base + SGA + APFI (w/o PST)**: baseline with SGA and APFI modules, excluding PST. Figure 3 presents results that support the following conclusions: (1) SGA demonstrates substantial performance enhancement across all configurations. The significant improvement from Base to Base+SGA across all datasets underscores the critical importance of explicitly encoding both positive and negative collaborative signals from complex user-item interactions, thereby strengthening SiMGR's overall performance. (2) Building upon the SGA module, incorporating either PST or APFI yields further performance improvements, while the complete SiMGR model integrating all three modules achieves optimal results. This demonstrates that each proposed module contributes meaningful effectiveness, and their integration produces superior performance.

5.4 Hyperparameter Anaylsis (RQ3)

We analyze the impact of the hyperparameter ν, which controls the ratio of users' mean ratings to standard deviation in Equation (14), Section 4.2. Statistical analysis reveals that the average user actions for training datasets are 8.2, 16.1, and 17.1 for Baby, Sports, and Electronics datasets, respectively. Figure 4 presents $R@20$ (green bars) and $N@20$ (purple lines) performance on the Baby and Sports datasets. Based on these results, we set $\nu = 0.5$ for Baby and $\nu = 0.8$ for the Sports datasets. For the Electronics dataset, we set $\nu = 0.9$. From a scalability perspective, larger ν for dense

Datasets	Baby				Sports				Electronics			
Models	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
MF-BPR	0.0340	0.0557	0.0184	0.0240	0.0423	0.0634	0.0233	0.0288	0.0417	0.0662	0.0238	0.0306
LightGCN	0.0471	0.0750	0.0255	0.0322	0.0557	0.0846	0.0311	0.0386	0.0515	0.0791	0.0299	0.0374
SiReN	0.0429	0.0678	0.0232	0.0296	0.05	0.075	0.0271	0.0334	0.0399	0.0548	0.0278	0.0326
SiGRec	0.0489	0.0770	0.0257	0.0329	0.0524	0.0793	0.0281	0.035	0.035	0.0538	0.0247	0.0308
LSGRec	0.0480	0.0715	0.0264	0.0323	0.0569	0.0849	0.0316	0.0388	0.0287	0.0536	0.0234	0.0315
MGCN	0.0656	0.0984	0.0352	0.0436	0.0743	0.1114	0.0409	0.0504	0.0584	0.0879	0.0338	0.0421
BM3	0.0536	0.0845	0.0289	0.0369	0.0644	0.0968	0.0357	0.0441	0.0516	0.0792	0.03	0.0376
FREEDOM	0.0658	0.1001	0.0349	0.0438	0.0726	0.1096	0.0394	0.0490	0.0564	0.0881	0.0321	0.0409
LGMRec	0.0667	0.0967	0.0364	0.0441	0.0715	0.1064	0.0395	0.0484	0.0572	0.0866	0.0335	0.0416
DiffMM	0.0507	0.0796	0.0276	0.0349	0.0675	0.1014	0.0368	0.0454	0.057	0.0871	0.0314	0.0394
TMLP	0.0649	0.0973	0.0355	0.0438	0.0741	0.1111	0.0405	0.05	0.0551	0.0854	0.0319	0.0403
SiMGR	**0.0702**	**0.1052**	**0.0382**	**0.0471**	**0.0782**	**0.1160**	**0.0428**	**0.0526**	**0.0592**	**0.0904**	**0.0342**	**0.0428**
RelImp	5.247%	5.095%	4.945%	6.803%	5.249%	4.129%	4.645%	4.365%	1.370%	3.849%	1.183%	1.663%

Table 2: Results of baselines and SiMGR where the second best is underlined and the best is in **bold**.

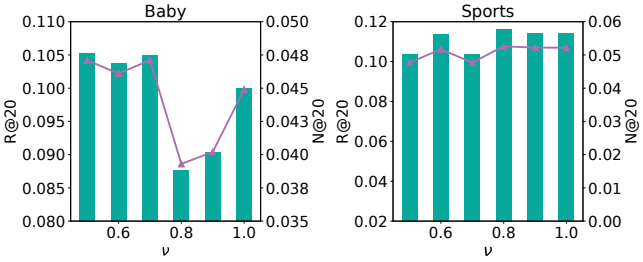

Figure 4: Effect of ν on Baby and Sports datasets

data, smaller ν for sparse — balancing mean and deviation emphasis.

The APFI module requires partitioning items into tail and popular categories. We employ the hyperparameter ξ (tail ratio) to control these divergent proportions. Figure 5 presents results across three datasets. We constrain $\xi \leq 0.2$, and conduct a grid search over $[0.05, 0.2]$ with step size of 0.05. Figure 5 demonstrates that model performance initially increases, then decreases as ξ increases. When ξ=0.05, performance remains suboptimal, likely because the augmented edges inadequately address data sparsity issues. Conversely, when $\xi > 0.1$, recommendation performance deteriorates, indicating that larger ξ values excessively classify items as tail items, introducing noise during pseudo-edge augmentation and yielding inferior results. Based on experiment results, we set $\xi = 0.1$ across all datasets.

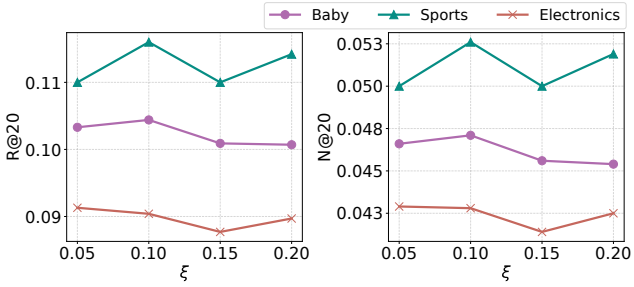

Figure 5: Impact of hyperparameter ξ

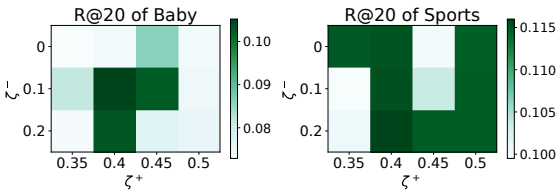

Figure 6: Impact of ζ^+, ζ^- on Baby and Sports datasets

The positive score ζ^+ in Eq. (19) determines the number of similar items selected for positive pseudo-graph augmentation, where higher values impose stricter selection criteria (grid search on [0.35, 0.5] with step size 0.05) . Conversely, the negative score ζ^- in Eq. (21) controls the number of dissimilar items added to negative pseudo-graphs, with lower values indicating more stringent requirements (grid search on [0, 0.1, 0.2]) . Figure 6 presents the results of $R@20$ on the Baby and Sports datasets. Regarding overall dataset sparsity, Sports exhibits the highest sparsity. Therefore, a larger value is assigned to ζ^- to add more pseudo-negative edges. The average user interaction count is 8.2 for Baby, receiving a smaller value, that is $\zeta^- = 0.1$. For optimal performance, we set $\zeta^+ = 0.4$ across three datasets, while ζ^- is configured as 0.1 for Baby, and 0.2 for Sports, respectively.

6 Conclusion

In this paper, we propose SiMGR, a sign-aware multimodal graph recommendation framework that addresses the limitation of neglecting negative feedback in existing MRS. Our approach contributes in three key aspects: (1) enabling collaborative multimodal feature fusion with positive/negative interaction signals within a unified graph architecture; (2) developing a personalized sign threshold strategy for positive/negative interaction subgraph generation; and (3) implementing a tailored pseudo-edge injection to address data sparsity in signed graphs. Experiments on three public datasets validate SiMGR's effectiveness and superiority over SOTA baselines. Our findings underscore the importance of modeling negative feedback and identifying new directions for signed graph integration in multimodal recommendation systems.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62172237). Tingjian Ge was supported in part by NSF (IIS-2124704, OAC-2106740) and U.S. Army CCDC (W911QY2020005). We also acknowledge support from Enowa Network Technology Co., Ltd., and thank Feipeng Dou and Hao Li for their valuable guidance.

References

Aljunid, M. F.; Manjaiah, D.; et al. 2025. A collaborative filtering recommender systems: Survey. *Neurocomputing*, 617: 128718.

Anand, V.; and Maurya, A. K. 2025. A survey on recommender systems using graph neural network. *ACM Transactions on Information Systems*, 43(1): 1–49.

Chen, S.; Chen, J.; Zhou, S.; et al. 2024. SIGformer: Sign-aware Graph Transformer for Recommendation. In *ACM SIGIR*, 1274–1284.

Derr, T.; et al. 2018. Signed graph convolutional networks. In *ICDM*, 929–934.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 249–256.

Guo, Z.; Li, J.; Li, G.; et al. 2024. LGMRec: Local and Global Graph Learning for Multimodal Recommendation. In *AAAI*, 8454–8462.

He, R.; and McAuley, J. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *Proceedings of the 25th International Conference on World Wide Web*, 507–517.

He, X.; Deng, K.; Wang, X.; et al. 2020a. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *ACM SIGIR*, 639–648.

He, X.; Deng, K.; Wang, X.; et al. 2020b. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *ACM SIGIR*, 639–648.

Heshmati, A.; Meghdadi, M.; Afsharchi, M.; et al. 2025. SiSRS: Signed social recommender system using deep neural network representation learning. *Expert Syst. Appl.*, 259: 125205.

Huang, J.; Qin, J.; Yu, Y.; et al. 2025. Beyond graph convolution: Multimodal recommendation with topology-aware mlps. In *AAAI*, volume 39, 11808–11816.

Huang, J.; Shen, H.; Cao, Q.; et al. 2021. Signed bipartite graph neural networks. In *ACM CIKM*, 740–749.

Huang, J.; Xie, R.; Cao, Q.; et al. 2023. Negative Can Be Positive: Signed Graph Neural Networks for Recommendation. *Inf. Process. Manag.*, 60(4): 103403.

Jiang, Y.; Xia, L.; Wei, W.; et al. 2024. Diffmm: Multimodal diffusion model for recommendation. In *ACM MM*, 7591–7599.

Kim, M.-J.; Lee, Y.-C.; et al. 2023. TrustSGCN: learning trustworthiness on edge signs for effective signed graph convolutional networks. In *ACM SIGIR*, 2451–2455.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Lei, F.; Cao, Z.; Yang, Y.; et al. 2023. Learning the user's deeper preferences for multi-modal recommendation systems. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(3s): 1–18.

Liao, H.; Wang, S.; Cheng, H.; et al. 2025. Aspect-enhanced explainable recommendation with multi-modal contrastive learning. *ACM Transactions on Intelligent Systems and Technology*, 16(1): 1–24.

Liu, F.; Li, Z.; and Nie, L. 2025. *Multimodal Learning Toward Recommendation*. Springer.

Liu, Y.; Dang, Y.; Liang, Y.; et al. 2025a. Towards Unified Modeling for Positive and Negative Preferences in Sign-Aware Recommendation. In *DASFAA*.

Liu, Z.; and Lu, W. 2025. MDN: Modality Decomposition Network for Multimodal Recommendation. In *ACM ICMR*, 871–879.

Liu, Z.; Wang, C.; Zheng, S.; et al. 2025b. Pone-GNN: Integrating Positive and Negative Feedback in Graph Neural Networks for Recommender Systems. *ACM Trans. Recomm. Syst.*

Malitesta, D.; Cornacchia, G.; Pomo, C.; et al. 2025. Formalizing multimedia recommendation through multimodal deep learning. *ACM Transactions on Recommender Systems*, 3(3): 1–33.

Oord, A. v. d.; et al. 2018. Representation learning with contrastive predictive coding. *arXiv preprint:1807.03748*.

Peng, H.; Fu, C.; et al. 2025. Indirect Interactions Discovering and True Negative Sampling for Multimodal Recommendation. *IEEE Transactions on Computational Social Systems*.

Rendle, S.; et al. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI*, 452–461.

Seo, C.; Jeong, K.; Lim, S.; et al. 2022. SiReN: Sign-Aware Recommendation Using Graph Neural Networks. *IEEE Trans. Neural Networks Learn. Syst.*, 35(4): 4729–4743.

Shang, Y.; Gao, C.; Chen, J.; et al. 2023. Enhancing adversarial robustness of multi-modal recommendation via modality balancing. In *ACM MM*, 6274–6282.

Shen, X.; et al. 2018. Deep network embedding for graph representation learning in signed networks. *IEEE transactions on cybernetics*, 50(4): 1556–1568.

Wang, J.; Liang, M.; et al. 2025. Multi-modal Negative Sampling for Recommendation with User Interest. In *Companion Proceedings of the ACM on Web Conference 2025*, 2187–2195.

Wang, Z.; Feng, Y.; Zhang, X.; et al. 2025. Multi-Modal Correction Network for Recommendation. *IEEE TKDE*, 37(2): 810–822.

Wei, Y.; Wang, X.; Nie, L.; et al. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *ACM MM*, 1437–1445.

Xu, F.; Zhu, Z.; Fu, Y.; et al. 2024. Collaborative denoised graph contrastive learning for multi-modal recommendation. *Information Sciences*, 679: 121017.

Xu, J.; Chen, Z.; Wang, W.; et al. 2025a. COHESION: Composite Graph Convolutional Network with Dual-Stage Fusion for Multimodal Recommendation. In *ACM SIGIR*, 1830–1839.

Xu, J.; Chen, Z.; Yang, S.; et al. 2025b. A Survey on Multimodal Recommender Systems: Recent Advances and Future Directions. *arXiv preprint:2502.15711*.

Yi, J.; et al. 2024a. Variational Mixture of Stochastic Experts Auto-encoder for Multi-modal Recommendation. *IEEE Transactions on Multimedia*.

Yi, Z.; Wang, X.; Ounis, I.; et al. 2022. Multi-modal graph contrastive learning for micro-video recommendation. In *ACM SIGIR*, 1807–1811.

Yi, Z.; et al. 2024b. A unified graph transformer for overcoming isolations in multi-modal recommendation. In *ACM RecSys*, 518–527.

Yu, P.; Tan, Z.; Lu, G.; et al. 2023. Multi-View Graph Convolutional Network for Multimedia Recommendation. In *ACM MM*, 6576–6585.

Zhang, Z.; Liu, J.; et al. 2023. Contrastive learning for signed bipartite graphs. In *ACM SIGIR*, 1629–1638.

Zhang, Z.; Zhao, P.; Li, X.; et al. 2024. Signed Graph Representation Learning: A Survey. *arXiv preprint: 2402.15980*.

Zhou, X. 2023. MMRec: Simplifying Multimodal Recommendation. In *ACM Multimedia Asia Workshops*, 6:1–6:2.

Zhou, X.; and Shen, Z. 2023. A Tale of Two Graphs: Freezing and Denoising Graph Structures for Multimodal Recommendation. In *ACM MM*, 935–943.

Zhou, X.; Zhou, H.; Liu, Y.; et al. 2023. Bootstrap Latent Representations for Multi-modal Recommendation. In *ACM WWW*, 845–854.

Zhu, J.; Yan, Y.; et al. 2020. Beyond homophily in graph neural networks: Current limitations and effective designs. *Advances in neural information processing systems*, 33: 7793–7804.