MGSPREC: MULTIMODAL SEQUENTIAL RECOMMENDATION WITH MULTI-GRANULARITY SHARED AND PRIVATE REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal sequential recommendation uses visual and textual content to enrich the modeling of users' dynamic preferences. However, raw modality features may contain information that is irrelevant to recommendation. Meanwhile, different modalities have distinct representation spaces and semantic structures. Direct parameter sharing or forced alignment across modalities may therefore introduce cross modal interference and weaken valuable modality specific semantics. To address these issues, we propose MGSPRec, a multimodal sequential recommendation method based on multi granularity shared and private representation learning. MGSPRec first uses behavior guided signals to refine visual and textual representations. This process suppresses recommendation irrelevant noise while preserving modality specific semantics. It then employs separate Transformer encoders for ID, visual, and textual sequences to reduce interference across modalities. Furthermore, ID based collaborative information is used as a semantic anchor. Shared and modality specific representations are learned at both the item and sequence levels. Experiments on three real world datasets, Baby, Pantry, and Office, show that MGSPRec consistently outperforms representative baselines. Compared with SGP4SR, MGSPRec achieves average relative improvements of 4.80%, 3.69%, and 2.29% over HR and NDCG at different cutoff values on Baby, Pantry, and Office, respectively. Ablation studies and parameter sensitivity analyses further validate the effectiveness of the proposed components.

1 INTRODUCTION

Sequential recommendation aims to model the evolution of user interests from historical interaction sequences and predict the next item a user is likely to interact with Dang et al. (2023); Zheng et al. (2024); Zhu et al. (2021); Xie et al. (2022). Traditional methods mainly rely on item IDs and user item interaction signals. When interaction data are sparse, limited collaborative information may be insufficient to characterize item properties and users' fine grained preferences Tang & Wang (2018); Hidasi et al. (2015); Kang & McAuley (2018); Sun et al. (2019). Multimodal sequential recommendation addresses this limitation by incorporating visual and textual content. Such information enriches item representations from complementary semantic perspectives and provides additional cues for modeling users' dynamic preferences Hou et al. (2022); Li et al. (2023); Ji et al. (2023).

Despite recent progress, learning reliable and complementary user preference representations from heterogeneous multimodal information remains challenging. (1) Raw multimodal features and their derived semantic relations may contain information that is irrelevant to users' actual preferences Li et al. (2026); Xu et al. (2025). Such information can be semantically discriminative but may not reflect real behavioral preferences. Constructing item relation graphs only from modality semantic similarity may connect items with similar content but weak behavioral relationships. These unreliable connections can propagate noise through the graph and affect subsequent preference modeling. (2) Different modalities have distinct representation spaces and sequential transition patterns Zhang et al. (2025). ID representations mainly capture collaborative relationships between users and items, while visual and textual representations describe visual attributes and linguistic semantics. These modalities differ in information distribution, semantic structure, and preference expression. Using shared parameter encoders may therefore introduce interference across modalities and limit the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

learning of modality specific sequential patterns. (3) Different modalities contain both task relevant shared information and complementary modality specific information Verma et al. (2019); Qian et al. (2026). Existing methods often directly fuse complete modality representations or apply global alignment constraints to reduce differences across modalities. Such strategies can improve consistency across modalities, but may also weaken useful modality specific information and spread irrelevant information between modalities.

To address the above challenges, we propose MGSPRec, a multimodal sequential recommendation method based on multi-granularity shared and private representation learning, with the overall framework illustrated in Figure 1. MGSPRec first uses collaborative behavioral signals to refine raw visual and textual information. Position aware item co-occurrence relations are used to improve modality features and item relation modeling. A modality preservation mechanism limits excessive behavioral intervention, which helps suppress recommendation irrelevant noise while preserving useful modality semantics. The model then employs separate Transformer encoders for ID, visual, and textual sequences. These encoders capture sequential patterns in different representation spaces and reduce interference across heterogeneous modalities. Modality interest prototypes are further introduced to enhance user interest representations in the visual and textual branches. Finally, MGSPRec uses the collaborative information encoded in ID representations as a task relevant semantic anchor. Visual and textual representations are decomposed into shared and private components at both the item and sequence levels. The shared representations capture common information across modalities, while the private representations preserve complementary information specific to each modality. The main components are summarized as follows:

- **Behavior Guided Modality Representation Optimization Mechanism.** We use collaborative behavioral signals to refine visual and textual representations. Position aware item co-occurrence information is used to improve modality features and item relation modeling. A modality preservation mechanism limits excessive behavioral intervention, which helps suppress recommendation irrelevant noise while preserving useful modality semantics.
- **Multimodal Sequential Modeling with Independent Parameters.** We employ separate Transformer encoders for ID, visual, and textual sequences. This design captures sequential patterns in different representation spaces and reduces interference across heterogeneous modalities. Modality interest prototypes are further introduced to enhance user interest representations in the visual and textual branches.
- **Multi-Granularity Shared and Private Representation Learning.** We use the collaborative information encoded in ID representations as a task related semantic anchor. Visual and textual representations are decomposed into shared and private components at both the item and sequence levels. The shared representations capture common information across modalities, while the private representations preserve complementary modality specific information.

2 Related Work

2.1 Multimodal Sequential Recommendation

Multimodal sequential recommendation introduces visual and textual content to complement the collaborative signals provided by item ID. This helps obtain richer representations of user interests. Existing studies mainly focus on multimodal fusion, pretrained semantic transfer, and noise reduction. MV-RNN Cui et al. (2018) introduced visual and textual features into sequential modeling. MMSR Hu et al. (2023), M3SRec Bian et al. (2023), and MMMLP Liang et al. (2023) further improved information interaction within and across modalities through multimodal sequential graphs, mixture of experts, and feature mixing mechanisms, respectively. With the development of pretrained vision and language models, MISSRec Wang et al. (2023) and IISAN Fu et al. (2024) further improved the use of content semantics through pretrained multimodal representations and parameter efficient adaptation.

Recent studies have paid more attention to extracting information related to user interests and reducing modality noise. HM4SR Zhang et al. (2025) uses a hierarchical mixture of experts to capture multimodal information related to user interests. SGP4SR Li et al. (2026) learns preferences for

each modality separately to reduce interference across modalities. It also incorporates behavioral information into modality relation modeling. However, the joint refinement of modality features and item relations remains insufficiently explored. Parameter interference among heterogeneous modalities also requires further investigation. In addition, existing methods provide limited exploration of the explicit separation between shared information and information specific to each modality.

2.2 Multimodal Shared and Private Representation Learning

A key challenge in multimodal representation learning is how to balance shared information across modalities with information specific to each modality. DeepCU Verma et al. (2019) jointly models common and modality specific information in multimodal data, showing that both are important for effective multimodal representation learning. MISA Hazarika et al. (2020) further projects different modalities into modality invariant and modality specific spaces to capture cross modal commonality and modality dependent characteristics. More recently, DecAlign Qian et al. (2026) decouples multimodal representations into modality common and modality specific components, while preserving modality specific information during cross modal alignment. These studies demonstrate the effectiveness of jointly modeling shared and modality specific information. However, they mainly focus on general multimodal representation learning or sentiment analysis, with limited use of collaborative behavioral information in recommendation. In addition, user preferences involve both item semantics and sequential interests. How to use task related collaborative information to guide shared and modality specific representation learning at multiple granularities remains insufficiently explored.

3 Methodology

3.1 Problem Definition

We study multimodal sequential recommendation with visual and textual content. Let $\mathcal{U} = \{u_1, \ldots, u_{|\mathcal{U}|}\}$ and $\mathcal{V} = \{v_1, \ldots, v_{|\mathcal{V}|}\}$ denote the user and item sets, respectively, and let $\mathcal{M} = \{I, T\}$ denote the visual and textual modalities. For each item $v_i \in \mathcal{V}$, e_i^{ID} denotes its learnable ID representation, while e_i^I and e_i^T denote its visual and textual features. For each user $u \in \mathcal{U}$, the historical interaction sequence is denoted by $S^u = \{v_1^u, \ldots, v_n^u\}$, where v_t^u is the item interacted with at time step t. According to the item indices in S^u, we obtain the corresponding ID, visual, and textual sequences, denoted by E_u^{ID}, E_u^I, and E_u^T. Given these inputs, the goal is to model the user's dynamic preference and predict the next interacted item. For a candidate item v, the prediction score is defined as $\hat{y}_{u,v} = f(S^u, E_u^{ID}, E_u^I, E_u^T, v)$, where $f(\cdot)$ denotes the multimodal sequential recommendation model. Candidate items are ranked by $\hat{y}_{u,v}$, and the top-K items are returned as recommendations.

3.2 Behavior-Guided Modality Representation Optimization

Item co-occurrence patterns in users' historical interaction sequences can reveal latent behavioral correlations. Moreover, items occurring at closer positions are more likely to correspond to similar user interest states. Based on this observation, for two items v_i and v_j, we define their behavioral correlation using a distance-decayed co-occurrence measure:

$$C_{ij} = \sum_{u \in \mathcal{U}} \sum_{(p,q) \in P_{ij}^u} \frac{1}{|q - p|}, \tag{1}$$

where P_{ij}^u denotes the set of all co-occurring position pairs of items v_i and v_j in the historical interaction sequence of user u, C_{ij} represents their behavioral correlation strength, and $|q - p|$ denotes the positional distance between the two interactions. Shorter positional distances contribute more to C_{ij}, emphasizing behaviorally closer item pairs. To reduce the scale differences caused by varying interaction frequencies, we normalize C_{ij} over the behavioral neighbors of each item and denote the resulting behavioral relation weight as T_{ij}. A larger T_{ij} indicates a stronger behavioral relation between items v_i and v_j. We then use the behavioral relation matrix T to refine the original

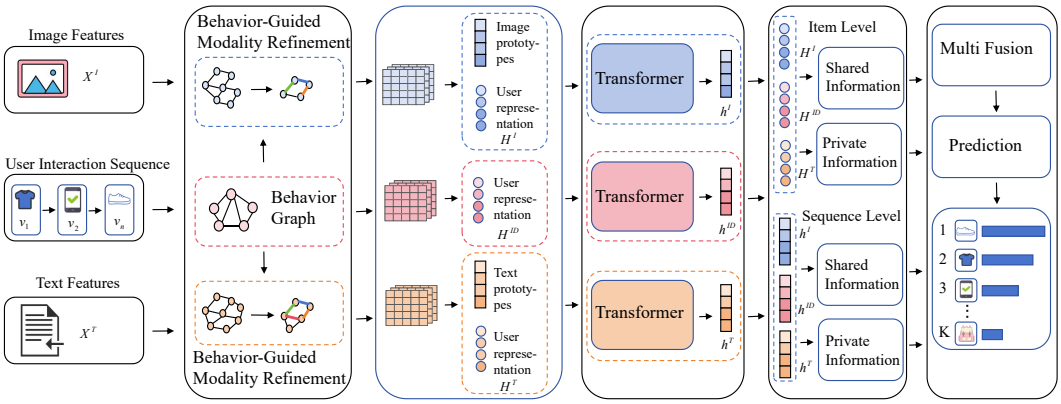


Figure 1: Overall framework of MGSPRec.

modality representations. For each modality $m \in \{I, T\}$, the behavioral neighborhood representation c_i^m is obtained by aggregating the modality representations of neighboring items according to their behavioral weights T_{ij}. In this way, c_i^m provides behavior-aware contextual information for item v_i. Since the amount of behavioral evidence varies across items, applying the same correction strength to all items may introduce unreliable signals, especially for items with sparse behavioral connections. Therefore, we estimate the reliability of behavioral evidence according to the number of effective behavioral neighbors d_i^b:

$$r_i = \frac{\log\left(1 + d_i^b\right)}{\max_k \log\left(1 + d_k^b\right) + \epsilon}, \tag{2}$$

where r_i denotes the behavioral reliability of item v_i, and ϵ is a small constant for numerical stability. A larger r_i indicates that the behavioral evidence associated with item v_i is relatively more sufficient and reliable. Finally, the modality representation refined by behavioral information is obtained through constrained residual injection:

$$E_i^m = e_i^m + \alpha r_i c_i^m, \quad m \in \{I, T\}, \tag{3}$$

where α controls the overall contribution of behavioral information. This design retains the original modality feature as the primary representation while introducing behavioral context through a residual connection. The correction strength is adaptively adjusted according to the reliability of behavioral evidence, thereby reducing recommendation irrelevant interference while preserving modality specific semantics.

After obtaining the behavior refined modality representations, we construct visual and textual item relation graphs separately. For each modality $m \in \{I, T\}$, the semantic similarity s_{ij}^m between items v_i and v_j is computed using the cosine similarity of their refined representations E_i^m and E_j^m. Based on s_{ij}^m, we first select a relatively large set of semantically similar items as candidate neighbors for each item. To further incorporate collaborative behavioral evidence into neighbor selection, we re-score each candidate neighbor using the behavioral relation T_{ij}:

$$q_{ij}^m = s_{ij}^m + \beta T_{ij}, \tag{4}$$

where β controls the contribution of collaborative behavioral information to neighbor selection. The top-K items with the highest q_{ij}^m values are retained as the final neighbor set $\mathcal{N}_i^m$. In this way, behaviorally supported item relations are prioritized, reducing the likelihood of introducing connections that are semantically similar but lack behavioral relevance. After determining the neighbor set, we further combine semantic similarity and behavioral information to assign edge weights:

$$A_{ij}^{m} = \begin{cases} s_{ij}^{m}\left(1+\eta T_{ij}\right), & v_j \in \mathcal{N}_i^{m}, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

where η controls the contribution of behavioral information to the edge weights. Through this process, we obtain the behavior-enhanced modality graph $A^m \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ for each modality $m \in \{I, T\}$.

3.3 Independent Multimodal Sequential Modeling

For the visual and textual branches, we propagate the learnable item ID embeddings over the modality relation graphs obtained in the previous section to derive item representations from different modality perspectives. Meanwhile, the behavioral relation graph is used to obtain collaborative item representations:

$$H^{m} = A^{m} E^{ID}, \quad m \in \{ID, I, T\}, \tag{6}$$

where A^{ID} denotes the behavioral relation graph, A^I and A^T denote the visual and textual relation graphs, respectively, and E^{ID} is the learnable item ID embedding matrix. Consequently, the three branches capture item representations from collaborative, visual, and textual perspectives, respectively. For a user's historical interaction sequence S^u, we retrieve the corresponding representations from the three item representation matrices and add positional embeddings. The resulting sequences are then fed into three Transformer encoders with separate parameters:

$$Z_u^{m} = \mathrm{Trm}_m\left(H^{m}[S^u] + P_{\mathrm{pos}}\right), \quad m \in \{ID, I, T\}, \tag{7}$$

where $\mathrm{Trm}_m(\cdot)$ denotes the Transformer encoder specific to branch m, and P_{pos} denotes the learnable positional embeddings. $H^m[S^u]$ denotes the sequence representation retrieved from H^m according to the item indices in S^u, while Z_u^m denotes the sequence of hidden representations produced by the Transformer. We use the hidden representation at the last valid position, denoted as z_u^m, as the final sequential representation.

Users' multimodal preferences may not only evolve dynamically with interaction sequences, but also concentrate around several representative interest patterns. To enhance latent interest modeling in the visual and textual branches, we introduce interest prototypes for the two content modalities. Specifically, for each modality $m \in \{I, T\}$, K-means clustering is performed on the graph enhanced item representations H^m to obtain K_c global interest prototypes, forming the prototype matrix W^m. Based on W^m, the personalized prototype representation P_u^m is further obtained through cross attention with the user's modality sequence, allowing the prototypes to adapt to individual historical behaviors. The matching relationships between the current modality sequence representation and the personalized prototypes are then computed, and Gumbel-Softmax is applied to obtain the prototype assignment weights R_u^m. The enhanced modality preference is formulated as

$$h^{m} = z_u^{m} + R_u^{m} P_u^{m}, \quad m \in \{I, T\}, \tag{8}$$

where h^I and h^T denote the user's visual and textual preferences, respectively, and z_u^m denotes the sequential representation output by the Transformer of modality m. The ID branch does not introduce additional interest prototypes, and its user preference h^{ID} is directly obtained from the representation at the last valid position of the ID sequence. To further integrate the basic sequential information from the ID, visual, and textual perspectives, we aggregate the representations of the three branches using learnable weights:

$$h^{g} = \rho_{ID} z_u^{ID} + \rho_I z_u^{I} + \rho_T z_u^{T}, \tag{9}$$

where h^g denotes the fused global user preference, and ρ_{ID}, ρ_I, and ρ_T control the contributions of the ID, visual, and textual branches, respectively.

3.4 Multi-Granularity Shared and Private Representation Learning

We perform shared and private representation learning at both the item and sequence levels. The item level captures common and complementary information in fine grained item semantics, while the sequence level captures shared and complementary patterns in users' dynamic interests across modalities. Specifically, the item level uses (H^{ID}, H^I, H^T), while the sequence level uses (h^{ID}, h^I, h^T). Accordingly, for each granularity $l \in \{\text{item}, \text{seq}\}$, we denote the corresponding representation of modality m as x_l^m, where $x_{\text{item}}^m = H^m$ and $x_{\text{seq}}^m = h^m$. For each modality $m \in \{I, T\}$, x_l^m is then projected into a shared space and a private space:

$$s_l^m = E_s^m\left(x_l^m\right), \qquad p_l^m = E_p^m\left(x_l^m\right), \quad m \in \{I, T\}, \tag{10}$$

where s_l^m denotes the task-relevant shared representation, while p_l^m captures modality-specific information. By performing the same shared-private decomposition at both granularities, the model jointly captures cross-modal commonality at the item-semantic level and complementarity at the user-interest level. Simply projecting visual and textual representations into a shared space does not guarantee that the retained information is truly relevant to the recommendation task. Since ID representations are mainly learned from user–item interactions and thus encode relatively direct collaborative preference signals, we use the ID representation at the same granularity as a semantic anchor to guide the learning of shared visual and textual representations. For each granularity $l \in \{\text{item}, \text{seq}\}$, the collaborative anchoring loss is defined as

$$\mathcal{L}_{anc}^l = \frac{1}{2} \sum_{m \in \{I,T\}} \left[1 - \cos\left(s_l^m, \text{sg}\left(z_l^{ID}\right)\right)\right], \tag{11}$$

where $\text{sg}(\cdot)$ denotes the stop gradient operation. This design allows the ID representation to serve only as a collaborative semantic reference, preventing the auxiliary objective from introducing reverse interference into the ID branch. To reduce redundancy between shared and private representations while avoiding representation collapse, we introduce orthogonality, variance, and reconstruction constraints. Their joint regularization objective is defined as

$$\mathcal{L}_{reg}^l = \lambda_o \mathcal{L}_{orth}^l + \lambda_v \mathcal{L}_{var}^l + \lambda_r \mathcal{L}_{rec}^l, \tag{12}$$

where $\mathcal{L}_{orth}^l$ encourages shared and private representations to be decorrelated, $\mathcal{L}_{var}^l$ prevents the private representations from collapsing, and $\mathcal{L}_{rec}^l$ preserves information from the original representations. Finally, the shared and private learning objectives at the item and sequence levels are jointly optimized as

$$\mathcal{L}_{sp} = \sum_{l \in \{\text{item,seq}\}} w_l \left(\mathcal{L}_{anc}^l + \mathcal{L}_{reg}^l\right), \tag{13}$$

where w_l controls the contribution of the corresponding granularity.

3.5 Prediction and Model Optimization

After the above modules, the model obtains the global user preference h^g, visual preference h^I, and textual preference h^T. For each candidate item $v_j \in \mathcal{V}$, the matching scores of the global and modality specific branches are computed as

$$\hat{y}_{u,j}^g = \left(h^g\right)^\top H_j^{ID}, \qquad \hat{y}_{u,j}^m = \left(h^m\right)^\top H_j^m, \quad m \in \{I, T\}, \tag{14}$$

where $\hat{y}_{u,j}^g$ denotes the prediction score produced by the global branch, while $\hat{y}_{u,j}^I$ and $\hat{y}_{u,j}^T$ denote the prediction scores produced by the visual and textual branches, respectively. The three branch scores are then combined to obtain the final prediction score:

$$\hat{y}_{u,j} = \hat{y}_{u,j}^g + \gamma \hat{y}_{u,j}^I + (1 - \gamma)\hat{y}_{u,j}^T, \tag{15}$$

Table 1: Overall performance comparison of all methods. The best results are highlighted in **bold**, and the second-best results are underlined.

Dataset	Metric	SASRec	UniSRec	SASRecF	MISSRec	HM4SR	SGP4SR	MGSPRec
Baby	HR@10	0.0419	0.0472	0.0408	0.0504	0.0422	0.0514	**0.0565**
	NDCG@10	0.0188	0.0207	0.0204	0.0246	0.0224	0.0261	**0.0283**
	HR@20	0.0667	0.0764	0.0647	0.0770	0.0660	0.0864	**0.0877**
	NDCG@20	0.0250	0.0280	0.0265	0.0311	0.0284	0.0348	**0.0361**
	HR@50	0.1191	0.1354	0.1160	0.1329	0.1153	0.1508	**0.1536**
	NDCG@50	0.0353	0.0396	0.0366	0.0420	0.0381	0.0475	**0.0491**
Pantry	HR@10	0.0477	0.0728	0.0389	0.0751	0.0511	0.0740	**0.0761**
	NDCG@10	0.0215	0.0323	0.0191	0.0349	0.0277	0.0360	**0.0383**
	HR@20	0.0771	0.1104	0.0661	0.1126	0.0813	0.1121	**0.1146**
	NDCG@20	0.0288	0.0418	0.0258	0.0442	0.0352	0.0455	**0.0480**
	HR@50	0.1355	0.1840	0.1203	0.1780	0.1387	0.1824	**0.1845**
	NDCG@50	0.0403	0.0563	0.0365	0.0569	0.0465	0.0594	**0.0618**
Office	HR@10	0.1129	0.1249	0.1052	0.1281	0.1118	0.1269	**0.1294**
	NDCG@10	0.0763	0.0848	0.0796	0.0853	0.0847	0.0861	**0.0885**
	HR@20	0.1355	0.1517	0.1259	0.1549	0.1325	0.1556	**0.1591**
	NDCG@20	0.0819	0.0916	0.0849	0.0919	0.0899	0.0934	**0.0959**
	HR@50	0.1742	0.1965	0.1632	0.2002	0.1701	0.2033	**0.2066**
	NDCG@50	0.0896	0.1005	0.0922	0.1008	0.0973	0.1028	**0.1053**

where γ controls the relative contributions of the visual and textual branches. All candidate items are ranked according to $\hat{y}_{u,j}$, and the top-K items are returned as the recommendation results.

During training, the three prediction branches are optimized using the full item cross entropy loss. We denote the corresponding losses as $\mathcal{L}_g$, $\mathcal{L}_I$, and $\mathcal{L}_T$ for the global, visual, and textual branches, respectively. The overall training objective is then formulated as

$$\mathcal{L} = \mathcal{L}_g + \mu\mathcal{L}_I + (1-\mu)\mathcal{L}_T + \lambda\mathcal{L}_{sp}, \tag{16}$$

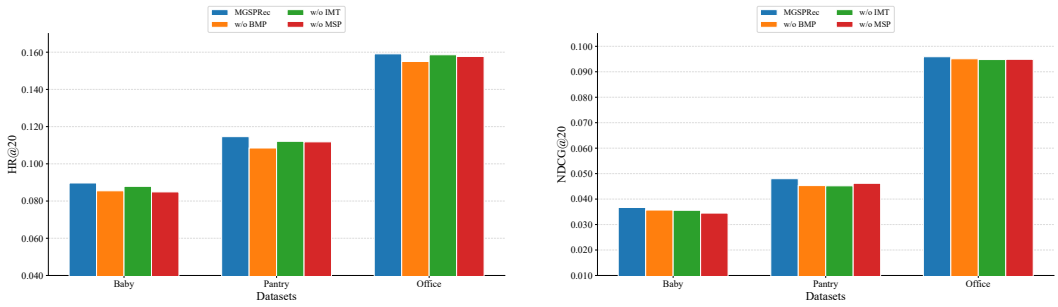
where $\mathcal{L}_g$, $\mathcal{L}_I$, and $\mathcal{L}_T$ denote the recommendation losses of the global, visual, and textual branches, respectively, and λ controls the contribution of the multi-granularity shared and private representation learning objective.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. We conduct experiments on three public subsets of the Amazon Review dataset: Baby, Prime Pantry, and Office Products. Following common preprocessing protocols for sequential recommendation, we remove duplicate interactions and apply a 5-core filtering strategy, retaining only users and items with at least five interactions. We adopt a chronological leave-one-out strategy, using the last interaction of each user for testing, the second-to-last interaction for validation, and the remaining interactions for training. We evaluate recommendation performance using HR and NDCG, and report HR@K and NDCG@K for $K \in \{10, 20, 50\}$, with higher values indicating better performance. Detailed dataset statistics are provided in the Appendix.

Baselines. To comprehensively evaluate the effectiveness of MGSPRec, we compare it with representative sequential recommendation and multimodal sequential recommendation methods. Specifically, SASRec Kang & McAuley (2018) relies solely on item IDs and models sequential dependencies through self-attention; UniSRec Hou et al. (2022) enhances sequential representations with pretrained textual semantics; SASRecF directly incorporates visual and textual features into SASRec; HM4SR Zhang et al. (2025) employs a hierarchical mixture-of-experts architecture to model dynamic multimodal user interests; MISSRec Wang et al. (2023) performs sequence pretraining through multimodal interest discovery and contrastive learning; and SGP4SR Li et al. (2026) alleviates multimodal noise from both cross-modal and intra-modal perspectives.



(a) HR@20 (b) NDCG@20

Figure 2: Ablation study of MGSPRec.

4.2 Performance Comparison

Table 1 reports the performance of all methods on Baby, Pantry, and Office. Overall, MGSPRec achieves the best results across all six metrics on the three datasets. Compared with SASRec, which relies only on item IDs, MGSPRec achieves average relative improvements of 38.22%, 57.08%, and 16.87% on Baby, Pantry, and Office, respectively. This result shows the benefit of incorporating visual and textual content as complementary information to collaborative signals. Compared with MISSRec, MGSPRec improves the average performance by 14.93%, 5.62%, and 3.25% on the three datasets. The corresponding improvements over HM4SR are 30.38%, 38.41%, and 12.78%. These results suggest that multimodal content alone may still contain information that is irrelevant to user preferences. Incorporating behavioral information into modality feature and item relation modeling can provide more reliable representations for recommendation.

SGP4SR also considers multimodal noise and incorporates behavioral information into modality relation modeling. MGSPRec consistently achieves better results across all evaluation metrics. The average relative improvements over SGP4SR are 4.80%, 3.69%, and 2.29% on Baby, Pantry, and Office, respectively. Different from SGP4SR, MGSPRec further combines behavior guided modality representation optimization, independent multimodal sequential modeling, and multi-granularity shared and private representation learning. These components reduce irrelevant interference and improve the modeling of shared and complementary information across modalities.

4.3 Ablation Study

To evaluate the effectiveness of the key components in MGSPRec, we construct three ablation variants. *w/o BMP* removes the behavior guided modality representation optimization module and directly uses the raw visual and textual features. *w/o IMT* replaces the separate Transformer encoders for the ID, visual, and textual branches with a shared parameter encoder. *w/o MSP* removes the multi-granularity shared and private representation learning module. Figure 2 presents the HR@20 and NDCG@20 results of MGSPRec and its variants on Baby, Pantry, and Office. Removing BMP leads to performance degradation on all three datasets, with a larger drop on Pantry. This result indicates that raw multimodal features may contain information that is irrelevant to user preferences. Refining modality features and item relations with collaborative behavioral information can improve the reliability of multimodal representations. Removing IMT also reduces the recommendation performance. Different modalities have distinct semantic structures and sequential patterns. Using shared parameters may limit the learning of information specific to each modality, while separate encoders can reduce interference across modalities. Removing MSP causes clear performance drops on Baby and Pantry, and also results in a consistent decline on Office. This result shows the benefit of explicitly modeling the relationships among different modalities. MGSPRec uses ID based collaborative representations as semantic anchors and learns shared and modality specific information at both the item and sequence levels. This allows the model to better capture common and complementary information across modalities. Overall, the ablation results demonstrate the effectiveness of BMP, IMT, and MSP.

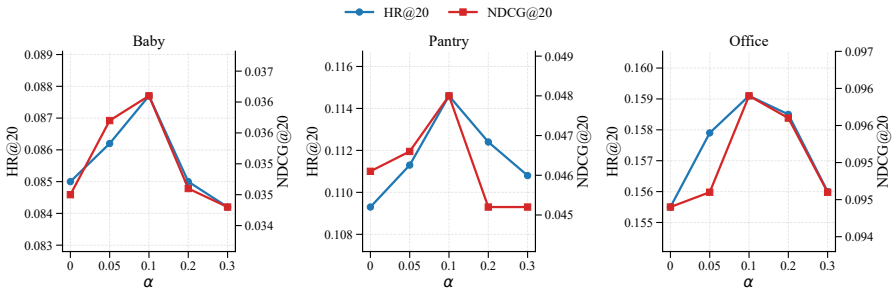


Figure 3: Influence of different α values on model performance.

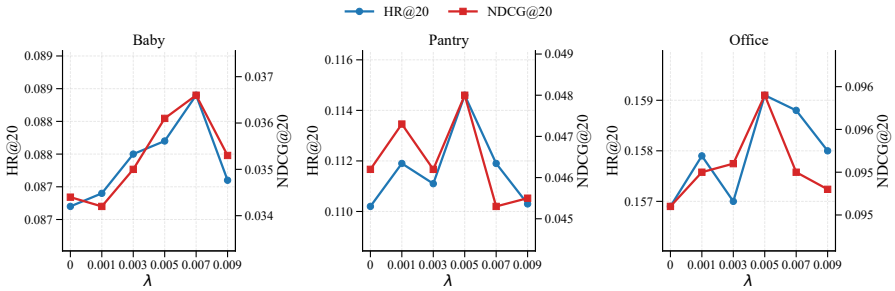


Figure 4: Influence of different λ values on model performance.

4.4 Effect of α in Behavior Guided Modality Representation Optimization

The parameter α controls the contribution of behavioral neighborhood information to the original modality features. We vary α over $\{0, 0.05, 0.10, 0.20, 0.30\}$. As shown in Figure 3, the performance on all three datasets generally increases first and then decreases. Baby, Pantry, and Office all achieve their best results at $\alpha = 0.1$. These results indicate that an appropriate amount of collaborative behavioral information can improve the reliability of modality representations. However, excessive behavioral calibration may weaken the original modality semantics. Therefore, balancing behavioral information and modality semantics is important for effective representation learning.

4.5 Influence of the Shared and Private Learning Weight λ

The parameter λ controls the contribution of the multi-granularity shared and private representation learning objective to the overall optimization. We vary λ over $\{0, 0.001, 0.003, 0.005, 0.007, 0.009\}$. As shown in Figure 4, the performance across the three datasets shows an overall increasing and then decreasing trend as λ increases, with better results obtained at moderate values. These results indicate that an appropriate auxiliary constraint helps learn task related shared information while preserving modality specific information. However, an excessively large λ may interfere with the main recommendation objective. Therefore, balancing the auxiliary objective and the recommendation objective is important for effective model optimization.

5 Conclusion

In this work, we propose MGSPRec, a multimodal sequential recommendation method based on multi-granularity shared and private representation learning. MGSPRec improves the reliability of multimodal information through behavior-guided modality representation optimization, reduces interference across heterogeneous modalities with parameter-independent sequential modeling, and captures shared and modality-specific information at both the item and sequence levels. Experiments on three real world datasets demonstrate the effectiveness of MGSPRec.

AI USE STATEMENT

Large Language Models (LLMs) were used to assist with language polishing, manuscript organization, and readability improvement. They were not involved in the development of the core research ideas, model design, experimental implementation, or result generation.

REFERENCES

Shuqing Bian, Xingyu Pan, Wayne Xin Zhao, Jinpeng Wang, Chuyuan Wang, and Ji-Rong Wen. Multi-modal mixture of experts represetation learning for sequential recommendation. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pp. 110–119, 2023.

Jianxin Chang, Chen Gao, Yu Zheng, Yiqun Hui, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. Sequential recommendation with graph neural networks. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 378–387, 2021.

Qiang Cui, Shu Wu, Qiang Liu, Wen Zhong, and Liang Wang. Mv-rnn: A multi-view recurrent neural network for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 32(2):317–331, 2018.

Yizhou Dang, Enneng Yang, Guibing Guo, Linying Jiang, Xingwei Wang, Xiaoxiao Xu, Qinghui Sun, and Hong Liu. Ticoserec: Augmenting data to uniform sequences by time intervals for effective recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 36(6):2686–2700, 2023.

Xinyan Fan, Zheng Liu, Jianxun Lian, Wayne Xin Zhao, Xing Xie, and Ji-Rong Wen. Lighter and better: low-rank decomposed self-attention networks for next-item recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 1733–1737, 2021.

Junchen Fu, Xuri Ge, Xin Xin, Alexandros Karatzoglou, Ioannis Arapakis, Jie Wang, and Joemon M Jose. Iisan: Efficiently adapting multimodal representation for sequential recommendation with decoupled peft. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 687–697, 2024.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 1122–1131, 2020.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*, 2015.

Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. Towards universal sequence representation learning for recommender systems. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 585–593, 2022.

Hengchang Hu, Wei Guo, Yong Liu, and Min-Yen Kan. Adaptive multi-modalities fusion in sequential recommendation systems. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pp. 843–853, 2023.

Wei Ji, Xiangyan Liu, An Zhang, Yinwei Wei, Yongxin Ni, and Xiang Wang. Online distillation-enhanced multi-modal transformer for sequential recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 955–965, 2023.

Zhixuan Jia, Yushun Fan, Jia Zhang, Chunyu Wei, Ruyu Yan, and Xing Wu. Improving next location recommendation services with spatial-temporal multi-group contrastive learning. *IEEE Transactions on Services Computing*, 16(5):3467–3478, 2023.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pp. 197–206. IEEE, 2018.

Changhong Li, Zhiqiang Guo, Guohui Li, Zhong Yang, and Chuhang Hong. Sgp4sr: Seperated-modality guided user perference learning for multimodal sequential reconmmendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15054–15062, 2026.

Jiacheng Li, Ming Wang, Jin Li, Jinmiao Fu, Xin Shen, Jingbo Shang, and Julian McAuley. Text is all you need: Learning language representations for sequential recommendation. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 1258–1267, 2023.

Jiahao Liang, Xiangyu Zhao, Muyang Li, Zijian Zhang, Wanyu Wang, Haochen Liu, and Zitao Liu. Mmmlp: Multi-modal multilayer perceptron for sequential recommendations. In *Proceedings of the ACM Web Conference 2023*, pp. 1109–1117, 2023.

Chengxuan Qian, Shuo Xing, Li Li, Yue Zhao, and Zhengzhong Tu. Decalign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *International Conference on Learning Representations*, volume 2026, pp. 152802–152824, 2026.

Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 811–820, 2010.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 1441–1450, 2019.

Jiaxi Tang and Ke Wang. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pp. 565–573, 2018.

Sunny Verma, Chen Wang, Liming Zhu, and Wei Liu. Deepcu: Integrating both common and unique latent information for multimodal sentiment analysis. In *International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 2019.

Jinpeng Wang, Ziyun Zeng, Yunxiao Wang, Yuting Wang, Xingyu Lu, Tianxiang Li, Jun Yuan, Rui Zhang, Hai-Tao Zheng, and Shu-Tao Xia. Missrec: Pre-training and transferring multi-modal interest-aware sequence representation for recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 6548–6557, 2023.

Yiqing Wu, Ruobing Xie, Yongchun Zhu, Fuzhen Zhuang, Xu Zhang, Leyu Lin, and Qing He. Personalized prompt for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3376–3389, 2024.

Yueqi Xie, Peilin Zhou, and Sunghun Kim. Decoupled side information fusion for sequential recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pp. 1611–1621, 2022.

Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Hewei Wang, and Edith CH Ngai. Mentor: multi-level self-supervised learning for multimodal recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 12908–12917, 2025.

Mengqi Zhang, Shu Wu, Xueli Yu, Qiang Liu, and Liang Wang. Dynamic graph neural networks for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(5): 4741–4753, 2022.

Shengzhe Zhang, Liyi Chen, Chao Wang, Shuangli Li, and Hui Xiong. Temporal graph contrastive learning for sequential recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 9359–9367, 2024.

Shengzhe Zhang, Liyi Chen, Dazhong Shen, Chao Wang, and Hui Xiong. Hierarchical time-aware mixture of experts for multi-modal sequential recommendation. In *Proceedings of the ACM on Web Conference 2025*, pp. 3672–3682, 2025.

Xiaoyao Zheng, Xingwang Li, Shengfei Jiang, Zhenghua Chen, Liping Sun, Qingying Yu, Liangmin Guo, and Yonglong Luo. Lighter sequential recommendation algorithm with time interval awareness augmentation. *IEEE Transactions on Services Computing*, 17(6):3857–3868, 2024.

Tianyu Zhu, Leilei Sun, and Guoqing Chen. Graph-based embedding smoothing for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):496–508, 2021.

A Additional Related Work

A.1 Sequential Recommendation

Sequential recommendation models temporal dependencies and item transition patterns in users' historical interactions to capture dynamic preferences and predict the next item of interest Jia et al. (2023); Wu et al. (2024). Early studies mainly relied on Markov chains and deep sequential models for preference modeling. FPMC Rendle et al. (2010) combined matrix factorization with Markov chains to capture short term item transitions while modeling users' long term preferences. Caser Tang & Wang (2018) and GRU4Rec Hidasi et al. (2015) employed convolutional and recurrent neural networks, respectively, to learn sequential patterns. With the development of self attention mechanisms, SASRec Kang & McAuley (2018) and BERT4Rec Sun et al. (2019) adopted Transformer architectures to model long range behavioral dependencies. LightSANs Fan et al. (2021) further improves efficiency through low rank self attention and decoupled positional encoding. Subsequent studies have also explored graph based sequential modeling Chang et al. (2021); Zhang et al. (2022; 2024). Despite their effectiveness, these methods mainly rely on item IDs and interaction signals. Limited collaborative information may be insufficient to characterize item content and fine grained user preferences, especially under sparse interactions. This limitation motivates the incorporation of visual and textual content into sequential recommendation.

B Additional Experimental Details

B.1 Dataset Statistics

Table 2: Statistics of the experimental datasets.

Statistics	**Baby**	**Pantry**	**Office**
# Users	19,446	13,102	87,436
# Items	7,051	4,899	25,986
# Interactions	141,347	113,861	684,837
Avg. Length	7.269	8.691	7.840
Sparsity	0.9990	0.9982	0.9997

B.2 Multimodal Feature Settings

For Pantry and Office, pretrained CLIP-B/32 is used to extract 512-dimensional visual and textual features. For Baby, we use the publicly available 4096-dimensional visual features, while textual features are extracted using a pretrained Sentence Transformer to obtain 384-dimensional embeddings.

B.3 Implementation Details

All experiments are implemented in PyTorch. The embedding dimension of MGSPRec is set to 300, and the maximum sequence length is set to 50. The training batch size is 256 for Baby and 512 for Prime Pantry and Office Products, while the evaluation batch size is fixed to 1024 for all datasets. We optimize the model using Adam with an initial learning rate of 0.001 and adopt early stopping when the validation performance does not improve for 20 consecutive epochs. For all baseline methods, we use their publicly available implementations and tune the hyperparameters following the recommended settings in the original papers or the released configurations.

C Training Procedure of MGSPRec

To provide a clear description of the overall execution process of MGSPRec, we summarize its training procedure in Algorithm 1.

Algorithm 1 Training Procedure of MGSPRec

Require: User interaction sequences S^u, visual features e^I, and textual features e^T
Ensure: Learned model parameters Θ

```
Initialize model parameters Θ
Compute the position aware item co-occurrence relation C from user interaction sequences, and construct the behavioral relation matrix T and behavioral graph A^ID
for m ∈ {I, T} do
    Refine the raw modality features using behavioral neighborhood information to obtain behavior enhanced representations E^m
    Construct the modality graph A^m using semantic similarity and behavioral relations
end for
Obtain graph enhanced item representations H^ID, H^I, and H^T from the behavioral and modality graphs
for m ∈ {I, T} do
    Perform K-means clustering on H^m to obtain the global interest prototype matrix W^m
end for
Feed the user sequences corresponding to H^ID, H^I, and H^T into separate Transformer encoders to obtain z_u^m, where m ∈ {ID, I, T}
for m ∈ {I, T} do
    Adapt the global interest prototypes using the user modality sequence to obtain personalized prototypes P_u^m
    Compute the prototype assignment weights R_u^m and obtain the enhanced modality preference h^m
end for
Set the ID preference as h^ID and fuse the three branches to obtain the global user preference h^g
Perform shared and private representation learning at the item level (H^ID, H^I, H^T) and the sequence level (h^ID, h^I, h^T) to obtain the auxiliary loss L_sp
Compute the prediction losses L_g, L_I, and L_T from h^g, h^I, and h^T
Compute the overall loss L = L_g + μL_I + (1 − μ)L_T + λL_sp
Update Θ according to L
return Θ
```

D Complexity Analysis

We analyze the computational complexity of MGSPRec. Let $|\mathcal{U}|$ and $|\mathcal{V}|$ denote the numbers of users and items, respectively. Let B denote the batch size, n the maximum sequence length, D the representation dimension, L the number of Transformer layers, and K_c the number of interest prototypes. In the preprocessing stage, constructing the position aware item co-occurrence relation requires $O(|\mathcal{U}|n^2)$ time in the worst case. Computing pairwise modality similarities for constructing the modality KNN graphs requires $O(|\mathcal{V}|^2 D)$. K-means clustering on the graph enhanced item representations H^m is used to initialize the global interest prototype matrix W^m, with a complexity of $O(I_k|\mathcal{V}|K_c D)$, where I_k denotes the number of clustering iterations.

During training, the main computational cost comes from the three separate Transformer encoders. Ignoring the constant factor introduced by the three branches, their complexity is $O(BL(n^2 D + nD^2))$, where $n^2 D$ corresponds to the self attention computation. The interaction between modality sequence representations and the K_c interest prototypes requires $O(BnK_c D)$. Since MGSPRec performs full item prediction using h^g, h^I, and h^T, the corresponding complexity is $O(B|\mathcal{V}|D)$. The multi-granularity shared and private representation learning module mainly involves linear projections and auxiliary constraints, and its additional cost does not change the dominant complexity with respect to the sequence length.

Therefore, the main training complexity of MGSPRec can be expressed as $O(BL(n^2 D + nD^2) + BnK_c D + B|\mathcal{V}|D)$. When L and K_c are small constants, MGSPRec retains the same quadratic complexity in the sequence length n as a standard Transformer.