MSRec: Multimodal Semantic Learning for Cross-modal Recommendation via Multi-curvature Geometric Spaces

1st Yuxin Dong
Beijing University of Posts and Telecommunications,
Beijing 100-876, China
yuxindong@bupt.edu.cn

2nd Zheming Yang*
Institute of Computing Technology,
Chinese Academy of Sciences
yangzheming@ict.ac.cn
*Corresponding author

***Abstract*—The rapid growth of multimedia content has further stimulated research on multimodal recommendation. However, existing methods, when integrating modality embeddings with ID embeddings, often implicitly assume a unified latent geometry, lack rigorous theoretical support, and tend to dilute item representations. We argue that heterogeneous signals like text, vision, and behavioral IDs—exhibit distinct geometric regularities, including local proximity, hierarchical structure, and cyclic constraints, forcing them into a single Euclidean manifold entangles shared, modality-specific, and spurious components, thereby degrading ranking performance. To address this issue, we propose a Multimodal Semantic Learning for Cross-modal Recommendation (MSRec) that embeds each modality and the ID representation into parallel Euclidean, hyperbolic, and spherical manifolds and performs geometry-aware inference. Specifically, MSRec employs per-space modality weighting and ID-conditioned gating to extract cross-modality commonalities and preserve discriminative modality-specific cues in parallel manifolds, the fused representations are written back to the ID anchor via native manifold algebra, with cross-space consensus and geometry-tailored branches jointly forming the final item embedding for ranking. We further introduce a graph-neighborhood contrastive objective that integrates top-K neighborhood sampling, prototype construction, and near-boundary hard example generation, jointly optimized with a standard BPR loss. Experiments on three public datasets demonstrate that MSRec consistently outperforms strong baselines, achieving up to 9.62% relative improvement in Recall@10 and consistent gains of 2–9% across Recall@K and NDCG@K.**

***Keywords*—Multimodal recommendation, multi-curvature space, semantic-invariant, semantic-specific**

I. Introduction

Recommender systems have become essential infrastructure for information access, with collaborative filtering long serving as the dominant paradigm [1]–[3]. By modeling user–item ID interactions, ID-centric methods scale effectively but treat items as opaque identifiers, underutilizing semantic content and yielding an incomplete view of user preference formation. To address this limitation, multimodal recommendation incorporates content signals such as images, text, and video. The central challenge is not feature inclusion but integration: coherently reconciling interaction-driven ID information with modality-specific evidence within unified embeddings [4], [5], which is critical for reliable preference modeling. Existing multimodal recommendation (MR) methods largely follow two lines, namely matrix factorization–based and graph–based approaches, which we review in detail in Section II.

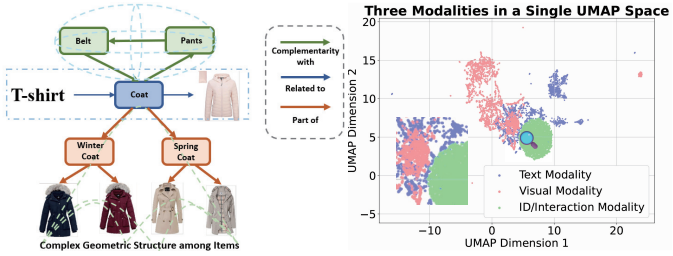


Fig. 1: The illustration showcases the complex geometric structures among items. The left panel centers on Coat (blue). Complementarity edges (green) link Coat with Belt and Pants; Related-to edges (blue) connect Coat to textual and visual evidence; Has-part / is-a edges (orange) expand the taxonomy. The right panel visualizes the embeddings of ID, text, and vision modalities after dimensionality reduction, illustrating how different geometric structures capture distinct semantic relations and motivating the proposed multi-curvature modeling.

Despite substantial progress, modeling modality diversity remains challenging. Real multimodal descriptors entangle (i) semantic-invariant information shared across modalities, (ii) modality-specific cues, and (iii) noise from redundancy or background clutter. As illustrated in Fig.1, these components also differ in latent geometry: local similarity neighborhoods are well approximated by Euclidean space, hierarchical relations exhibit lower distortion in hyperbolic space, and stylistic or complementary patterns are more naturally captured on hyperspherical manifolds. Empirical evidence [6], such as tree-like category graphs and circular concentration of seasonal or style signals, supports this mixed-curvature view. Forcing all evidence into a single Euclidean embedding entangles heterogeneous relations, increases distortion, and degrades ranking performance.

We propose MSRec, a multi-geometry recommender that aligns multimodal content with interaction IDs without trans-

port solvers. By embedding representations into Euclidean, hyperbolic, and hyperspherical spaces, MSRec performs ID-conditioned aggregation, geometry-native write-back, and cross-space fusion to obtain robust item embeddings. A graph-neighborhood contrastive objective with prototype-based hard samples further enforces semantic-ID consistency. Our code will be released on https://github.com/dong-yuxin/MSRec. In summary, our contributions are threefold:

- We propose **MSRec**, which embeds items and modalities into three explicit spaces: hyperbolic for hierarchical relations, spherical for complementarity and style cycles, and Euclidean for local metric similarity. Geometry-native aggregation and fusion yield unified item and user representations. To our knowledge, this is the first attempt to introduce multi-geometry interaction into multimodal recommendation.
- We design a semantic-guided contrastive scheme that selects nearest positives and farthest negatives across spaces, and sharpens decision boundaries via prototype interpolation and near-boundary hard mixing.
- Extensive experiments on three public benchmarks show that MSRec achieves new SOTA with statistically significant improvements on Recall and NDCG. Ablation studies further validate the contributions of each geometry and the contrastive module.

II. Related Work

A. Traditional Recommendation Methods

Traditional recommender systems are primarily grounded in collaborative filtering (CF), which infers user preferences from historical interactions [7]. Matrix Factorization (MF) and its variants represent users and items with ID-based embeddings and predict preferences via latent similarity [8]. Although effective and scalable, these ID-centric approaches rely heavily on dense interaction data and often generalize poorly in sparse or cold-start settings.

To address these limitations, subsequent research extends CF along several directions. Feature-aware models incorporate side information to enrich representations [9], while knowledge-aware methods leverage external knowledge graphs to inject structured semantics [10]. In parallel, graph-based recommenders explicitly model the user–item interaction graph and propagate higher-order relational information via graph convolution [11], with simplified variants improving efficiency and stability [12]. More recently, multimodal graph models further integrate content signals into graph-based frameworks [13], [14].

Despite these advances, most traditional methods either operate in unimodal settings or rely on explicit relational structures, and thus struggle to coherently integrate heterogeneous semantic signals from modern multimedia content. In contrast, our work departs from these paradigms by explicitly modeling joint multimodal semantics through geometry-aware representations.

B. Multimodal Recommendation

Multimodal recommendation aims to integrate heterogeneous content modalities with user–item interactions. Existing methods are commonly categorized into a separated framework (SF) and an end-to-end framework (EF) [15].

Separated Framework (SF): SF-based approaches extract modality features using external encoders and inject them into downstream recommenders. Representative examples include VBPR [16] and CKE [17], which incorporate visual and textual signals into ID-centric models. These methods correspond to the matrix factorization–based line mentioned in the introduction. Subsequent work improves modality utilization via attention or graph-based aggregation [18], [19].

End-to-End Framework (EF): EF-based methods jointly optimize content encoders and interaction models within a unified architecture. Early neural approaches [20], [21] demonstrate the effectiveness of end-to-end learning, while more recent studies leverage self-supervised objectives to alleviate sparsity and enhance robustness [22], [23]. Within this category, graph-based models exploit the user–item bipartite structure to propagate higher-order relations via graph convolution [24], [25], with MMGCN [13] constructing modality-specific graphs and GRCN [25] refining edge reliability using multimodal cues. A complementary line of work augments MR with explicit relational structures to model inter-item and inter-user dependencies [26], [27]. LATTICE [28] dynamically constructs an item–item similarity graph from multimodal features and propagates the induced relations via graph convolution, while DualGNN [14] further incorporates user–user affinities to guide personalized modality fusion. Despite steady progress, most multimodal methods ultimately combine modality features into item embeddings without explicitly addressing the distributional alignment between ID-based signals and multimodal content, and theoretical guarantees remain limited.

C. Graph Contrastive Learning

Graph contrastive learning (GCL) learns node or graph representations by contrasting semantically related pairs against dissimilar ones, and has shown strong effectiveness in self-supervised graph representation learning [29]. Representative methods construct positive and negative pairs via graph augmentations or structural views, such as subgraph sampling in GraphCL [30] or low-rank structural refinement in LightGCL [31]. Extensions further incorporate external knowledge to reduce noise and improve interpretability [32].

Despite their success, most GCL methods rely on heuristic or stochastic augmentation strategies to generate contrastive pairs, largely ignoring the semantic and geometric structure of the data, which can introduce noisy supervision and limit representation quality.

In contrast, our proposed MSRec adopts a semantic-guided contrastive scheme that constructs prototypes from neighborhood structure and generates near-boundary hard examples via controlled mixing, enabling geometry-aware contrastive learning without hand-crafted sampling rules.

III. Preliminaries

A. Riemannian Manifolds

A D-dimensional Riemannian manifold $(\mathcal{M}, g)$ is a smooth space endowed with a position-dependent inner product on

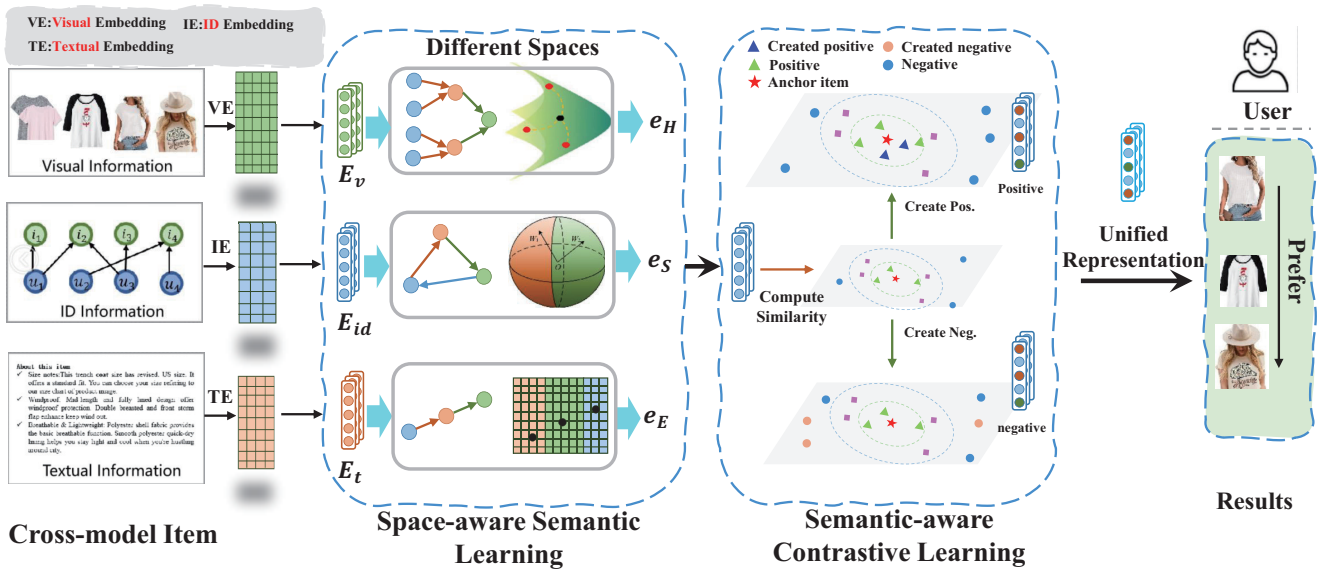


Fig. 2: The overall framework of MSRec. a) We first obtain the embeddings for visual and textual information by corresponding encoders. b) Then we align the multimodal embeddings to ID information by jointly utilizing modal-invariant and modal-specific multi-curvature space. c) Subsequently, we introduce semantic-enhanced contrastive learning to explore the semantic relevance among ID information and multimodal information.

each tangent space, which induces curvature-aware geometry. We consider constant-curvature model spaces

$$\mathcal{M}_c^D \in \{\mathbb{E}^D\ (c=0),\ \mathbb{H}_c^D\ (c<0),\ \mathbb{S}_c^D\ (c>0)\}, \quad (1)$$

corresponding to Euclidean, hyperbolic, and hyperspherical geometries. These spaces provide complementary inductive biases for modeling local similarity, hierarchy, and cyclic or directional patterns, respectively.

Learning on manifolds is enabled by the exponential and logarithmic maps, which transfer representations between the manifold and its tangent space:

$$\exp_x^c : \mathcal{T}_x\mathcal{M} \to \mathcal{M}, \qquad \log_x^c : \mathcal{M} \to \mathcal{T}_x\mathcal{M}. \quad (2)$$

They allow geometry-consistent aggregation by performing vector operations in tangent spaces and mapping the results back to the manifold.

In practice, we instantiate hyperbolic geometry using the Poincaré ball model $\mathbb{H}_c^D\ (c<0)$, where the geodesic distance admits the closed form

$$\mathrm{dist}_c(\boldsymbol{x},\boldsymbol{y}) = \frac{2}{\sqrt{-c}}\,\mathrm{atanh}\big(\sqrt{-c}\,\|-\boldsymbol{x}\oplus_c \boldsymbol{y}\|\big), \quad (3)$$

with $\oplus_c$ denoting Möbius addition. Analogous formulations hold for hyperspherical space. These operators form the mathematical basis for geometry-native aggregation and alignment in MSRec.

B. Contrastive Learning

Traditional contrastive loss [33] plays a critical role in aligning different views. Denoting B as a batch of training data containing pairs from two views, we have $B = \{(u_i, v_i)\}_{i=1}^N$, where u_i and v_i belong to views U and V, respectively. After passing through view-specific or shared encoders, the embeddings of u_i and v_i are denoted as $\boldsymbol{p}_i$ and $\boldsymbol{q}_i$ respectively. The contrastive loss function for aligning the views U with V is given by:

$$\mathcal{L}_{U\to V} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\langle \boldsymbol{p}_i, \boldsymbol{q}_i\rangle\right)}{\sum_{j\in[N]}\exp\left(\langle \boldsymbol{p}_i, \boldsymbol{q}_j\rangle\right)} \quad (4)$$

where $[N]$ represents the index set $\{1, 2, \ldots, N\}$. A corresponding loss function, $\mathcal{L}_{V\to U}$ can also be defined, with the total loss being $\mathcal{L}_{U\to V} + \mathcal{L}_{V\to U}$. Specifically, each pair (u_i, v_i) is considered a positive example, while all other combinations $\{v_j\}_{j\neq i}$ are treated as negative examples. This training approach encourages the model to bring similar instances closer together while pushing dissimilar instances apart, ultimately leading to the effective alignment of the different views.

IV. Multi-geometry Semantic Alignment Recommendation Framework

Problem Statement. Let $\mathcal{U}$ and $\mathcal{I}$ denote the sets of users and items. We maintain an ID-centric embedding table $\boldsymbol{E}_{\mathrm{id}} \in \mathbb{R}^{(|\mathcal{U}|+|\mathcal{I}|)\times d}$, whose rows represent users and items in a d-dimensional latent space. Each item is associated with modality-specific content. For modality $m \in \mathcal{M}$ (typically $\mathcal{M} = \{t, v\}$ for text and vision), let $\boldsymbol{E}_m \in \mathbb{R}^{|\mathcal{I}|\times d_m}$ denote the corresponding feature matrix, with $\boldsymbol{e}_{i,m} \in \mathbb{R}^{d_m}$ the feature of item i. User–item interactions are given by a sparse implicit-feedback matrix $\boldsymbol{R} \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{I}|}$, where $\boldsymbol{R}_{u,i} = 1$ indicates that user u has interacted with item i. These interactions define a bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V} = \mathcal{U} \cup \mathcal{I}$ and $\mathcal{E} = \{(u, i) : \boldsymbol{R}_{u,i} = 1\}$.

Given $\mathcal{G}$ and modality features $\{\boldsymbol{E}_m\}_{m\in\mathcal{M}}$, the goal of multimodal recommendation is to learn a scoring function $f : \mathcal{U} \times \mathcal{I} \to \mathbb{R}$ that predicts $\hat{y}_{u,i} = f(u, i)$ for ranking items per user. Our framework, **MSRec**, addresses this task

by aligning ID and content representations across multiple geometric spaces (Fig. 2).

ID Embedding Backbone. To capture collaborative signals from $\mathcal{G}$, we adopt a LightGCN-style propagation scheme. Let $\boldsymbol{E}_{\text{id}}^{(0)} \in \mathbb{R}^{(|\mathcal{U}|+|\mathcal{I}|)\times d}$ denote the initial ID embeddings. We define the bipartite adjacency and degree matrices as

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{R} \\ \boldsymbol{R}^\top & \boldsymbol{0} \end{bmatrix}, \qquad \boldsymbol{D}_{vv} = \sum_{v'} \boldsymbol{A}_{vv'}, \tag{5}$$

and the symmetrically normalized propagation operator

$$\boldsymbol{S} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}. \tag{6}$$

At layer $l \geq 1$, ID embeddings are updated via linear neighborhood diffusion,

$$\boldsymbol{E}_{\text{id}}^{(l)} = \boldsymbol{S}\boldsymbol{E}_{\text{id}}^{(l-1)}, \tag{7}$$

where normalization by $\boldsymbol{S}_{ui} = 1/\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}$ stabilizes aggregation. Higher layers incorporate increasingly global connectivity. Following standard practice, the final ID representation is obtained by averaging all layers:

$$\boldsymbol{E}_{\text{id}} = \frac{1}{L+1}\sum_{l=0}^{L} \boldsymbol{E}_{\text{id}}^{(l)}. \tag{8}$$

This $\boldsymbol{E}_{\text{id}}$ serves as the collaborative backbone that MSRec subsequently aligns and fuses with multimodal content representations.

A. Semantic Alignment by Multi-Space Interaction

ID embeddings from the backbone and modality embeddings from pretrained encoders reside in heterogeneous latent geometries, making naive fusion prone to noise and distortion. MSRec performs implicit semantic alignment by projecting both ID and modality features into three geometry channels, namely Euclidean, hyperbolic, and hyperspherical, and aligning them through space-aware aggregation, ID-anchored distribution, and cross-space fusion.

Space-Aware Multimodal Encoding. Let $\mathcal{M} = \{t, v\}$ denote textual and visual modalities with embeddings $\{\boldsymbol{E}_m\}$, and let $\boldsymbol{E}_{id}$ be the item ID embedding. We embed both modality and ID representations into three geometric spaces—Euclidean, hyperbolic, and hyperspherical—via

$$\begin{aligned} \boldsymbol{E}_{m,s} &= \begin{cases} \boldsymbol{E}_m, & s = E, \\ \exp_0^{\kappa_s}(\boldsymbol{E}_m), & s \in \{H, S\}, \end{cases} \\ \boldsymbol{E}_{id,s} &= \begin{cases} \boldsymbol{E}_{id}, & s = E, \\ \exp_0^{\kappa_s}(\boldsymbol{E}_{id}), & s \in \{H, S\}. \end{cases} \end{aligned} \tag{9}$$

where $s \in \{E, H, S\}$ with curvature $\kappa_s \in \{0, c, u\}$ ($c < 0$, $u > 0$). An ID-conditioned preference vector is computed in Euclidean space and lifted to non-Euclidean spaces accordingly.

(1) Space-Wise Aggregation and Write-Back. Within each space s, modality information is aggregated in the tangent space using ID-conditioned attention and gating:

$$\begin{aligned} \boldsymbol{h}_s = \sum_{m\in\mathcal{M}} \alpha_{m,s} \Big(& \sigma\big(\boldsymbol{G}_s[\log_s(\boldsymbol{E}_{m,s}); \log_s(\boldsymbol{E}_{id,s})]\big) \\ & \odot \log_s(\boldsymbol{E}_{m,s}) \Big). \end{aligned} \tag{10}$$

where $\alpha_{m,s}$ denotes attention weights over modalities. The aggregated signal is written back to the space via a geometry-native update:

$$\boldsymbol{z}_s = \boldsymbol{E}_{id,s} \oplus_{\kappa_s} \exp_s(\boldsymbol{h}_s), \tag{11}$$

yielding a geometry-aligned item embedding $\boldsymbol{z}_s$ in space s.

(2) Cross-Space Fusion. Let $\boldsymbol{Z} = [\boldsymbol{z}_E, \boldsymbol{z}_H, \boldsymbol{z}_S]$. We decompose the representation into a space-invariant component and geometry-specific residuals:

$$\boldsymbol{h}_{\text{unified}} = \boldsymbol{W}_U \Big[\boldsymbol{W}_I \boldsymbol{Z},\ \boldsymbol{Z} \odot \sigma(\boldsymbol{W}_{\text{Spe}}\boldsymbol{Z})\Big]. \tag{12}$$

Optionally, the unified embedding is stabilized by mixing with the Euclidean ID anchor:

$$\boldsymbol{E}_{i,\text{ms}} = (1-\beta)\boldsymbol{h}_{\text{unified}} + \beta \boldsymbol{E}_{id}. \tag{13}$$

(3) User Embeddings. User representations are obtained by degree-normalized aggregation over interacted items:

$$\boldsymbol{E}_{u,\text{ms}} = \sum_{i\in\mathcal{N}_u} \frac{\boldsymbol{W}_2\boldsymbol{E}_{i,\text{ms}} + \boldsymbol{b}_2}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}}. \tag{14}$$

Why it works. ID-conditioned attention suppresses modality noise before geometry-aware write-back; constant-curvature spaces capture complementary structural patterns; cross-space fusion balances geometry-invariant semantics with space-specific traits.

B. Contrastive Learning

In recommendation, we construct positives and negatives for each anchor item from its top-K similarity neighborhood. Let $\boldsymbol{E}_i \in \mathbb{R}^d$ be the fused embedding of item i and obtain an auxiliary semantic representation via a lightweight MLP:

$$\boldsymbol{e}_i^{\text{att}} = \text{MLP}(\boldsymbol{E}_i) \tag{15}$$

We compute cosine similarity

$$\text{sim}_{ij} = \frac{\boldsymbol{e}_i^{\text{att}\top}\boldsymbol{e}_j^{\text{att}}}{\|\boldsymbol{e}_i^{\text{att}}\|\,\|\boldsymbol{e}_j^{\text{att}}\|}, \tag{16}$$

and define candidate sets by the top-K nearest and farthest neighbors:

$$\begin{aligned} \mathcal{P}[i] &= \text{TopK}\big(\{\, j \neq i \,\};\, \text{sim}_{ij}\big), \\ \mathcal{N}[i] &= \text{TopK}\big(\{\, j \neq i \,\};\, -\text{sim}_{ij}\big). \end{aligned} \tag{17}$$

The neighbor sets $\mathcal{P}[i]$ and $\mathcal{N}[i]$, which contain the K most similar items and the K least similar (hard negatives) respectively,[1] are recomputed per epoch using the latest item embeddings, ensuring dynamic adaptation of contrastive supervision.

To diversify supervision, we synthesize prototype positives/negatives via random convex combinations of neighbors. At each epoch, draw $w_j \sim \mathcal{U}[0,1]$ i.i.d. and form

$$\boldsymbol{e}^+ = \frac{\sum_{j=1}^{|\mathcal{P}[i]|} w_j\, \boldsymbol{e}_j^{\text{att}}}{\sum_{j=1}^{|\mathcal{P}[i]|} w_j}, \qquad \boldsymbol{e}^- = \frac{\sum_{j=1}^{|\mathcal{N}[i]|} w_j\, \boldsymbol{e}_j^{\text{att}}}{\sum_{j=1}^{|\mathcal{N}[i]|} w_j}. \tag{18}$$

[1] In large catalogs, we sample $\mathcal{N}[i]$ from a low-similarity pool (e.g., bottom $q\%$ by sim) before taking TopK to reduce computation while keeping hardness.

a) Head Mixing.: We generate near-boundary hard samples by mixing the prototypes with the anchor. Sample $\alpha, \beta \sim \mathcal{U}[0,1]$ and set

$$\boldsymbol{e}^{+}_{\text{mix}} = \alpha\, \boldsymbol{e}^{+} + (1-\alpha)\, \boldsymbol{e}^{\text{att}}_{i}, \qquad \boldsymbol{e}^{-}_{\text{mix}} = \beta\, \boldsymbol{e}^{-} + (1-\beta)\, \boldsymbol{e}^{\text{att}}_{i}. \tag{19}$$

Because $\boldsymbol{e}^{+}$ and $\boldsymbol{e}^{-}$ are derived from nearest and farthest neighborhoods, respectively, head mixing preserves their relative similarity while increasing difficulty.[2]

b) Contrastive Objective.: Let $\text{score}(\boldsymbol{x}, \boldsymbol{y})$ be an inner-product scoring function and $\mathcal{G}$ the set of anchor items. We optimize

$$\mathcal{L}_{\text{CL}} = -\frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} \log \sigma\Big(\text{score}(\boldsymbol{e}^{\text{att}}_{i},\, \boldsymbol{e}^{+}_{\text{mix}}) - \text{score}(\boldsymbol{e}^{\text{att}}_{i},\, \boldsymbol{e}^{-}_{\text{mix}})\Big), \tag{20}$$

which yields continual, semantically grounded near-boundary contrasts without requiring explicit examples in the data.

c) Integration with Recommendation.: Let $\boldsymbol{E}_u, \boldsymbol{E}_i$ be the final user/item representations and $\hat{y}_{ui} = \boldsymbol{E}_u^{\top} \boldsymbol{E}_i$ the interaction score. We optimize the BPR loss [13]

$$\mathcal{L}_{\text{rec}} = \sum_{(u,i,j) \in \mathcal{D}} -\log \sigma\big(\hat{y}_{ui} - \hat{y}_{uj}\big), \tag{21}$$

and jointly train with

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \gamma\, \mathcal{L}_{\text{CL}}, \tag{22}$$

where $\gamma > 0$ balances ranking and contrastive learning.

V. Experiments

In what follows, we subject **MSRec** to a rigorous empirical assessment on three standard, publicly available benchmarks. The study is deliberately organized to address three reviewer-oriented concerns, i.e., overall utility, soundness of the design, and stability under tuning, articulated as the following research questions:

- **Q1.** Does **MSRec** deliver superior accuracy relative to leading multimodal recommenders and strong collaborative-filtering baselines?
- **Q2.** What incremental value does each constituent module contribute, and how does altering or removing a component change end-to-end performance?
- **Q3.** How resilient are the results to variations in hyperparameters, and to what extent do different configurations affect the final outcomes?

A. Experimental Settings

Datasets. We follow prior work [41] and reuse the released multimodal features to ensure strict comparability, including 4,096-dimensional visual descriptors and 384-dimensional textual embeddings. Experiments are conducted on three widely used Amazon categories, Baby, Sports, and Clothing, as in [28], [41]. Dataset statistics are summarized in Table II, and the raw data are publicly available.[3]

[2]Optionally enforce $\text{sim}(\boldsymbol{e}^{\text{att}}_{i}, \boldsymbol{e}^{+}_{\text{mix}}) \geq \min_{j \in \mathcal{P}[i]} \text{sim}_{ij}$ and $\text{sim}(\boldsymbol{e}^{\text{att}}_{i}, \boldsymbol{e}^{-}_{\text{mix}}) \leq \max_{j \in \mathcal{N}[i]} \text{sim}_{ij}$ by re-sampling α, β to strictly preserve the ordering.

[3]http://jmcauley.ucsd.edu/data/amazon/links.html

Compared Methods. We compare MSRec with representative baselines covering general CF, graph-based, hypergraph-based, and multimodal recommendation models, including BPR [34], LightGCN [12], SGL [26], HCCF [35], VBPR [16], MMGCN [13], LATTICE [37], BM3 [41], SLMRec [40], and LGMRec [44], among others. This set covers strong collaborative baselines and state-of-the-art multimodal methods.

Evaluation Protocol. We adopt the standard protocol used in [39], [41]. User interactions are split into training/validation/test sets with an 8:1:1 ratio. At test time, we perform all-item ranking with training interactions masked, and evaluate performance using $\text{Recall@}K$ and $\text{NDCG@}K$.

Implementation Details. All models are implemented using the public codebase of [45]. We use Adam optimization and follow the original hyperparameter settings of each baseline. Unless stated otherwise, embeddings are 64-dimensional, the batch size is 2048, and training uses early stopping based on validation Recall@20.

B. Overall Performance (RQ1)

Table I reports the comparative results on three benchmarks. MSRec consistently achieves the best performance across all metrics, demonstrating its superior ability to exploit multimodal content. The improvements are substantial rather than incremental: on the Baby dataset, for instance, MSRec improves Recall@10 by 9.60% over the strongest baseline. Similar gains are observed for both $\text{Recall@}K$ and $\text{NDCG@}K$.

These results stem from two key factors. First, instead of naive feature fusion, MSRec jointly aligns modality-invariant and modality-specific signals within a multi-curvature geometric framework, reducing cross-modal semantic distortion while preserving complementary information. Second, the semantic-aware contrastive objective explicitly couples multimodal representations with ID-based semantics by pulling together related items and separating unrelated ones, effectively enhancing mutual information between behavioral and content signals. Together, these mechanisms translate into state-of-the-art accuracy across all evaluated datasets.

The behavior of classical baselines further contextualizes these gains. Content-augmented CF models (e.g., VBPR) consistently outperform ID-only methods, confirming the value of multimodal signals. However, graph-based multimodal approaches may suffer from modality preference divergence, where message passing amplifies dominant-modal biases and degrades representations. While self-supervised methods partially mitigate this issue via auxiliary objectives, they remain sensitive to cross-modal preference conflicts. By explicitly disentangling and aligning shared and modality-specific semantics in multi-curvature spaces, MSRec effectively curbs such divergence, leading to its consistent superiority.

Ablation Study (RQ2)

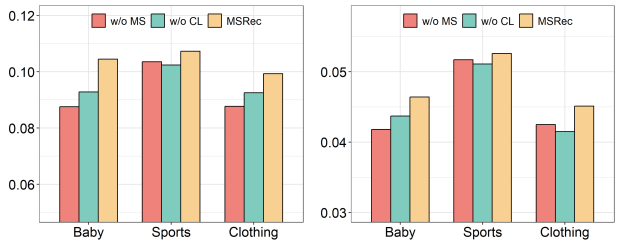
To identify the sources of MSRec's gains, we conduct ablation studies by selectively disabling key components. We consider two variants: (i) **w/o MS**, which removes the multi-curvature semantic alignment and replaces it with naive elementwise fusion of pre-trained visual/text features; and (ii) **w/o**

TABLE I: Overall performances of MSRec and other baselines on three datasets. The best result is in boldface and the second best is underlined. The t-tests validate the significance of performance improvements with the p-value < 0.01.

Dataset	Baby				Sports				Clothing			
Metrics	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
BPR [34]	0.0379	0.0607	0.0202	0.0261	0.0452	0.0690	0.0252	0.0314	0.0211	0.0315	0.0118	0.0144
LightGCN [12]	0.0464	0.0732	0.0251	0.0320	0.0553	0.0829	0.0307	0.0379	0.0331	0.0514	0.0181	0.0227
SGL [26]	0.0532	0.0820	0.0289	0.0363	0.0620	0.0945	0.0339	0.0423	0.0392	0.0586	0.0216	0.0266
NCL [27]	0.0538	0.0836	0.0292	0.0369	0.0616	0.0940	0.0339	0.0421	0.0410	0.0607	0.0228	0.0275
HCCF [35]	0.0480	0.0756	0.0259	0.0332	0.0573	0.0857	0.0317	0.0394	0.0342	0.0533	0.0187	0.0235
SHT [36]	0.0470	0.0748	0.0256	0.0329	0.0564	0.0838	0.0306	0.0384	0.0345	0.0541	0.0192	0.0243
VBPR [16]	0.0424	0.0663	0.0223	0.0284	0.0556	0.0854	0.0301	0.0378	0.0281	0.0412	0.0158	0.0191
MMGCN [13]	0.0498	0.0749	0.0261	0.0315	0.0582	0.0825	0.0305	0.0382	0.0329	0.0564	0.0219	0.0253
GRCN [25]	0.0531	0.0835	0.0291	0.0370	0.0600	0.0921	0.0324	0.0407	0.0431	0.0664	0.0230	0.0289
LATTICE [37]	0.0536	0.0858	0.0287	0.0370	0.0618	0.0950	0.0337	0.0423	0.0459	0.0702	0.0253	0.0306
MMGCL [38]	0.0522	0.0778	0.0289	0.0355	0.0660	0.0994	0.0362	0.0448	0.0438	0.0669	0.0239	0.0297
MICRO [39]	0.0570	0.0905	0.0310	0.0406	0.0675	0.1026	0.0365	0.0463	0.0496	0.0743	0.0264	0.0332
SLMRec [40]	0.0540	0.0810	0.0296	0.0361	0.0676	0.1007	0.0374	0.0462	0.0452	0.0675	0.0247	0.0303
BM3 [41]	0.0538	0.0857	0.0301	0.0378	0.0659	0.0979	0.0354	0.0437	0.0450	0.0669	0.0243	0.0295
DGVAE [42]	0.0636	0.1009	0.0340	0.0436	0.0753	0.1127	0.0410	0.0506	0.0619	0.0917	0.0336	0.0412
MGCN [43]	0.0620	0.0964	0.0339	0.0427	0.0729	0.1106	0.0397	0.0496	0.0641	0.0945	0.0347	0.0428
LGMRec [44]	0.0644	0.1002	0.0349	0.0440	0.0720	0.1068	0.0390	0.0480	0.0555	0.0828	0.0302	0.0371
MSRec	**0.0681**	**0.1072**	**0.0372**	**0.0484**	**0.0770**	**0.1191**	**0.0431**	**0.0525**	**0.0686**	**0.0992**	**0.0371**	**0.0450**
p-value	$7.4e^{-5}$	$5.5e^{-5}$	$4.4e^{-5}$	$7.7e^{-5}$	$3.3e^{-4}$	$6.2e^{-5}$	$5.0e^{-5}$	$8.7e^{-4}$	$4.3e^{-6}$	$5.0e^{-6}$	$7.4e^{-5}$	$2.3e^{-5}$
Improv.	9.62%	6.52%	9.34%	8.65%	2.37%	5.57%	5.35%	3.97%	6.84%	5.06%	7.22%	5.39%

TABLE II: Statistics of the datasets

Dataset	# Users	# Items	# Interactions	Density
Clothing	39,387	23,033	237,488	0.00026
Sports	35,598	18,357	256,308	0.00039
Baby	19,445	7,050	139,110	0.00101



(a) Recall@20 (b) NDCG@20

Fig. 3: Performance Comparison between different variants of MSRec.

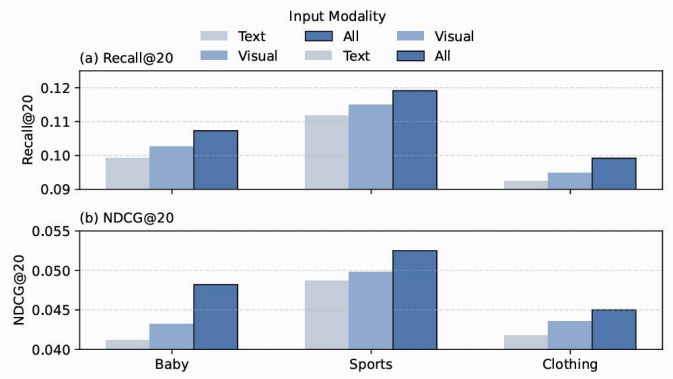


Fig. 4: Performance Comparison for MSRec under different modalities. This figure evaluates the impact of different modality configurations on recommendation performance. The results indicate that joint utilization of textual and visual modalities consistently improves top-K accuracy, validating the effectiveness of MSRec's multimodal representation modeling.

CL, which preserves the architecture but drops the semantic-aware contrastive objective. As shown in Fig. 3, both variants incur consistent performance degradation in Recall@20 across all datasets.

The larger drop of **w/o MS** indicates that raw multimodal features contain preference-irrelevant noise; multi-curvature alignment is crucial for disentangling modality-invariant and modality-specific semantics and for bridging the gap between ID and content representations. The decline of **w/o CL** further underscores the importance of semantic neighborhoods: contrastive learning guided by similarity weighting exposes complementary cross-item cues that are otherwise weakly expressed. Together, these results confirm that both MS and CL are essential to MSRec's effectiveness.

We further assess the contribution of individual modalities under three settings: Text (ID + text), Visual (ID + image), and All (ID + text + image). Fig 4 shows that augmenting ID embeddings with either modality consistently improves accuracy, with the Visual setting typically yielding larger gains. This reflects domain characteristics, where appearance plays a dominant role in user decisions, while textual descriptions often include redundant details. Integrating all modalities combines complementary cues and delivers the most robust performance.

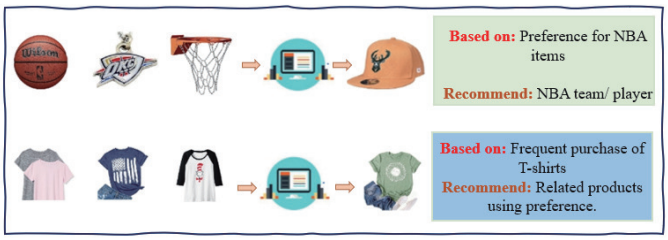


Fig. 5: Case study on Sports and Clothing datasets for MSRec.

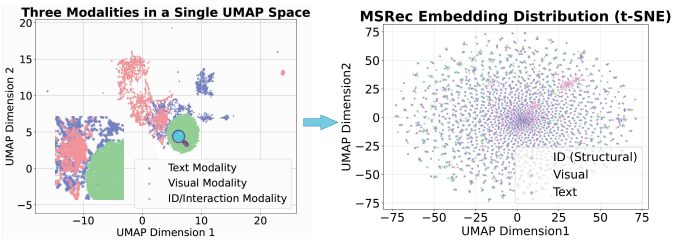


Fig. 6: Modality Representation Visualization. Left: before Right: after.

C. Qualitative Analysis and Hyperparameter Study (RQ3)

Case Study. We present two qualitative examples from *Sports* and *Clothing*. Fig. 5 shows, for each case, the user's historical interactions, MSRec's recommendations, and the corresponding personalized explanations. In the *Sports* case, interactions with a basketball, a logo necklace, and a basketball net indicate an NBA-oriented interest, leading MSRec to recommend NBA hats with explanations aligned to this theme. In the *Clothing* case, the user's history centers on T-shirts; MSRec both reinforces this preference and captures complementary semantics by recommending Shorts as a natural co-purchase. These examples illustrate that MSRec not only infers user intent from interaction histories, but also exploits latent complementarity to generate coherent and persuasive recommendations.

Effect of the contrastive objective. We examine the influence of the contrastive loss weight γ on the Baby dataset via grid search (Fig 7). Performance exhibits a clear U-shaped trend. When γ is too small, contrastive supervision is insufficient to shape the representation space, limiting the benefit of semantic neighborhoods. Conversely, overly large γ overemphasizes semantic alignment and weakens the primary ranking objective, degrading Recall@K and NDCG@K. An intermediate value achieves the best trade-off, effectively regularizing multimodal representations while preserving ranking fidelity.

Modality Representation Visualization. Fig 6 visualizes the learned item representations across modalities before and after applying MSRec. From left to right, the plots reveal a clear structural evolution. Initially, textual, visual, and ID-based embeddings exhibit pronounced heterogeneity, forming loosely related and weakly aligned clusters. After MSRec training, representations from different modalities become locally well-aligned within semantic neighborhoods, while still preserving modality-specific dispersion at a global scale. This pattern indicates that MSRec does not collapse heterogeneous modalities into a single homogeneous space; instead, it promotes

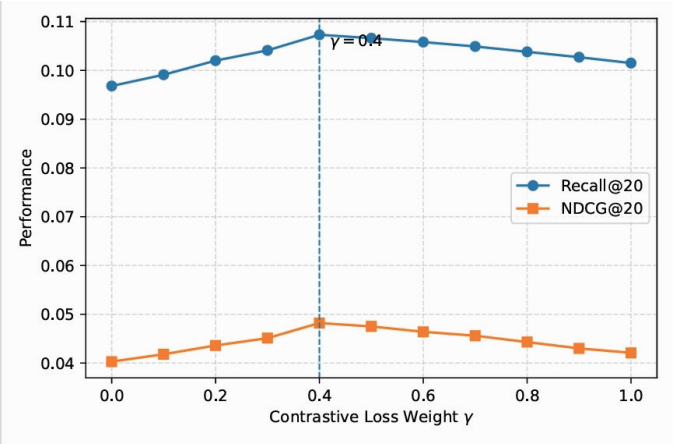


Fig. 7: Hyperparameter Analysis of Contrastive Learning Coefficients.

local semantic coherence across modalities while maintaining their intrinsic structural differences. **Such behavior suggests that the proposed multi-geometry alignment effectively mitigates adverse effects induced by modality heterogeneity, enabling consistent semantic organization without sacrificing modality diversity.**

VI. Conclusion and Future Work

We presented MSRec, a multimodal recommender that addresses modality heterogeneity and cross-item semantic relations through learning in multiple geometric spaces with distinct curvature. Instead of naive content fusion, MSRec jointly aligns modality-invariant and modality-specific representations with ID semantics, preserving shared meaning while retaining complementary cues unique to each modality. A semantic-aware contrastive objective further strengthens this alignment by exploiting neighborhoods of semantically related items, uncovering structure that is inaccessible to ID-only training. Theoretical analysis supports the stability and effectiveness of both the multi-geometry alignment and the contrastive learning mechanism, while extensive experiments on three real-world benchmarks demonstrate consistent improvements over strong collaborative filtering and multimodal baselines. Looking ahead, integrating MSRec with large language models to leverage their world knowledge and generalization capability offers a promising direction for enhancing robustness, adaptability, and cross-domain transfer.

VII. Acknowledgement

We thank the reviewers for their comments and we thank our mentor for his patient guidance and unconditional support.

References

[1] J. Li, X. Wang, G. Lv, and Z. Zeng, "Graphcfc: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition," *IEEE Trans. Multim.*, vol. 26, pp. 77–89, 2024.

[2] W. Nie, X. Wen, J. Liu, J. Chen, J. Wu, G. Jin, J. Lu, and A. Liu, "Knowledge-enhanced causal reinforcement learning model for interactive recommendation," *IEEE Trans. Multim.*, vol. 26, pp. 1129–1142, 2024.

[3] W. You, J. Ji, L. Sun, C. Yang, M. Yu, S. Chen, and J. Yao, "Automatic generation of interactive nonlinear video for online apparel shopping navigation," *IEEE Trans. Multim.*, vol. 26, pp. 474–486, 2024.

[4] H. Tang, G. Zhao, J. Gao, and X. Qian, "Personalized representation with contrastive loss for recommendation systems," *IEEE Trans. Multim.*, vol. 26, pp. 2419–2429, 2024.

[5] Y. Wu, G. Zhao, M. Li, Z. Zhang, and X. Qian, "Reason generation for point of interest recommendation via a hierarchical attention-based transformer model," *IEEE Trans. Multim.*, vol. 26, pp. 5511–5522, 2024.

[6] Z. Qiao, P. Wang, Y. Fu, Y. Du, P. Wang, and Y. Zhou, "Tree structure-aware graph representation learning via integrated hierarchical aggregation and relational metric learning," in *2020 IEEE International Conference on Data Mining (ICDM)*, 2020, pp. 432–441.

[7] D. Zhang, "An item-based collaborative filtering recommendation algorithm using slope one scheme smoothing," in *2009 Second International Symposium on Electronic Commerce and Security*, vol. 2, 2009, pp. 215–217.

[8] X. Xu, "Matrix factorization recommendation algorithm based on deep neural network," in *2019 2nd International Conference on Information Systems and Computer Aided Education (ICISCAE)*, 2019, pp. 320–323.

[9] T. Chen, W. Zhang, Q. Lu, K. Chen, Z. Zheng, and Y. Yu, "Svdfeature: a toolkit for feature-based collaborative filtering," *J. Mach. Learn. Res.*, vol. 13, pp. 3619–3622, 2012.

[10] H. Wang, M. Zhao, X. Xie, W. Li, and M. Guo, "Knowledge graph convolutional networks for recommender systems," in *WWW*, 2019, pp. 3307–3313.

[11] R. v. d. Berg, T. N. Kipf, and M. Welling, "Graph convolutional matrix completion," *arXiv preprint arXiv:1706.02263*, 2017.

[12] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "Lightgcn: Simplifying and powering graph convolution network for recommendation," in *SIGIR*, 2020, pp. 639–648.

[13] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T. Chua, "MMGCN: multi-modal graph convolution network for personalized recommendation of micro-video," in *ACM MM,*, 2019, pp. 1437–1445.

[14] Q. Wang, Y. Wei, J. Yin, J. Wu, X. Song, and L. Nie, "Dualgnn: Dual graph neural network for multimedia recommendation," *IEEE Trans. Multim.*, vol. 25, pp. 1074–1084, 2023.

[15] J. M. George, E. Joy, and S. Evangeline, "End-to-end personalized medical recommendation framework for disease prediction," in *2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL)*, 2025, pp. 1429–1433.

[16] R. He and J. J. McAuley, "VBPR: visual bayesian personalized ranking from implicit feedback," in *AAAI*, 2016, pp. 144–150.

[17] F. Zhang, N. J. Yuan, D. Lian, X. Xie, and W. Ma, "Collaborative knowledge base embedding for recommender systems," in *KDD*. ACM, 2016, pp. 353–362.

[18] Y. Zhou, G. Wang, and A. Guo, "Enhancing kgcn-based recommendation algorithms via attention mechanism integration," in *2025 6th International Conference on Computer Engineering and Application (ICCEA)*. IEEE, 2025, pp. 1756–1760.

[19] F. Meng, H. Ma, D. Zhao, and Q. Deng, "Knowledge graph-based dual feature aggregation recommendation model," in *2024 5th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA)*, 2024, pp. 371–375.

[20] T. Sun, J. Wen, and J. Gong, "Personalized learning resource recommendation using differential evolution-based graph neural network: A graphsage approach," in *2023 4th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC)*, 2023, pp. 636–639.

[21] R. Wang, Y. Luo, X. Li, Z. Zhang, J. Hu, and W. Liu, "A hybrid recommendation approach integrating matrix decomposition and deep neural networks for enhanced accuracy and generalization," in *2025 5th International Conference on Neural Networks, Information and Communication Engineering (NNICE)*, 2025, pp. 1778–1782.

[22] Y. She and C. Ouyang, "Self-supervised learning recommendation algorithm based on meta-paths," in *2023 IEEE 5th International Conference on Advanced Information and Communication Technologies (AICT)*, 2023, pp. 1–4.

[23] Z. Bao, X. Sun, and W. Zhang, "A pre-game item recommendation method based on self-supervised learning," in *2023 IEEE 6th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, 2023, pp. 961–966.

[24] K. Liu, F. Xue, D. Guo, P. Sun, S. Qian, and R. Hong, "Multimodal graph contrastive learning for multimedia-based recommendation," *IEEE Trans. Multim.*, vol. 25, pp. 9343–9355, 2023.

[25] Y. Wei, X. Wang, L. Nie, X. He, and T. Chua, "Graph-refined convolutional network for multimedia recommendation with implicit feedback," in *ACM MM*. ACM, 2020, pp. 3541–3549.

[26] J. Wu, X. Wang, F. Feng, X. He, L. Chen, J. Lian, and X. Xie, "Self-supervised graph learning for recommendation," in *SIGIR*, 2021, pp. 726–735.

[27] Z. Lin, C. Tian, Y. Hou, and W. X. Zhao, "Improving graph collaborative filtering with neighborhood-enriched contrastive learning," in *WWW*, 2022, pp. 2320–2329.

[28] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, and L. Wang, "Mining latent structures for multimedia recommendation," in *ACM MM*, 2021, pp. 3872–3880.

[29] G. V. Ramteke, V. Gupta, and P. Singla, "Enhancing recommendation systems with graph contrastive learning: A comprehensive study," in *2025 8th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2025, pp. 903–908.

[30] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," *Advances in neural information processing systems*, vol. 33, pp. 5812–5823, 2020.

[31] X. Cai, C. Huang, L. Xia, and X. Ren, "Lightgcl: Simple yet effective graph contrastive learning for recommendation," *arXiv preprint arXiv:2302.08191*, 2023.

[32] C. Wei, B. Yuan, C. Hu, J. Li, C.-D. Wang, and M. Guizani, "Knowledge-aware intent-guided contrastive learning for next-basket recommendation," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

[33] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *NeurIPS*, vol. 33, pp. 18 661–18 673, 2020.

[34] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: bayesian personalized ranking from implicit feedback," in *UAI*, 2009, pp. 452–461.

[35] L. Xia, C. Huang, Y. Xu, J. Zhao, D. Yin, and J. Huang, "Hypergraph contrastive collaborative filtering," in *SIGIR*, 2022, pp. 70–79.

[36] L. Xia, C. Huang, and C. Zhang, "Self-supervised hypergraph transformer for recommender systems," in *KDD*, 2022, pp. 2100–2109.

[37] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, and L. Wang, "Mining latent structures for multimedia recommendation," in *ACM MM*, 2021, pp. 3872–3880.

[38] Z. Yi, X. Wang, I. Ounis, and C. Macdonald, "Multi-modal graph contrastive learning for micro-video recommendation," in *SIGIR*, 2022, pp. 1807–1811.

[39] J. Zhang, Y. Zhu, Q. Liu, M. Zhang, S. Wu, and L. Wang, "Latent structure mining with contrastive modality fusion for multimedia recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 9, pp. 9154–9167, 2023.

[40] Z. Tao, X. Liu, Y. Xia, X. Wang, L. Yang, X. Huang, and T. Chua, "Self-supervised learning for multimedia recommendation," *IEEE Trans. Multim.*, vol. 25, pp. 5107–5116, 2023.

[41] X. Zhou, H. Zhou, Y. Liu, Z. Zeng, C. Miao, P. Wang, Y. You, and F. Jiang, "Bootstrap latent representations for multi-modal recommendation," in *WWW*, 2023, pp. 845–854.

[42] C. Li, X. Qin, D. Yang, and G. Wei, "Dgvae: An end-to-end model for link prediction in directed graphs," in *Proceedings of the 2021 5th International Conference on Innovation in Artificial Intelligence*, 2021, pp. 202–207.

[43] P. Yu, Z. Tan, G. Lu, and B.-K. Bao, "Multi-view graph convolutional network for multimedia recommendation," in *ACM MM*, 2023, pp. 6576–6585.

[44] Z. Guo, J. Li, G. Li, C. Wang, S. Shi, and B. Ruan, "Lgmrec: Local and global graph learning for multimodal recommendation," in *AAAI*, 2024, pp. 8454–8462.

[45] X. Zhou, "Mmrec: Simplifying multimodal recommendation," in *ACM Multimedia Asia Workshops*. ACM, 2023, pp. 6:1–6:2.