Hypercomplex Prompt-aware Multimodal Recommendation

Zheyu Chen*
The Hong Kong Polytechnic University
Hong Kong SAR, China
zheyu.chen@connect.polyu.hk

Jinfeng Xu*
The University of Hong Kong
Hong Kong SAR, China
jinfeng@connect.hku.hk

Hewei Wang
Carnegie Mellon University
Pittsburgh, PA, USA
heweiw@andrew.cmu.edu

Shuo Yang
The University of Hong Kong
Hong Kong SAR, China
shuoyang.ee@gmail.com

Zitong Wan
University College Dublin
Dublin, Ireland
zitong.wan@ucdconnect.ie

Haibo Hu†
The Hong Kong Polytechnic University
Hong Kong SAR, China
haibo.hu@polyu.edu.hk

Abstract

Modern recommender systems face critical challenges in handling information overload while addressing the inherent limitations of multimodal representation learning. Existing methods suffer from three fundamental limitations: (1) restricted ability to represent rich multimodal features through a single representation, (2) existing linear modality fusion strategies ignore the deep nonlinear correlations between modalities, and (3) static optimization methods failing to dynamically mitigate the over-smoothing problem in graph convolutional network (GCN). To overcome these limitations, we propose **HPMRec**, a novel **H**ypercomplex **P**rompt-aware **M**ultimodal **Rec**ommendation framework, which utilizes hypercomplex embeddings in the form of multi-components to enhance the representation diversity of multimodal features. HPMRec adopts the hypercomplex multiplication to naturally establish nonlinear cross-modality interactions to bridge semantic gaps, which is beneficial to explore the cross-modality features. HPMRec also introduces the prompt-aware compensation mechanism to aid the misalignment between components and modality-specific features loss, and this mechanism fundamentally alleviates the over-smoothing problem. It further designs self-supervised learning tasks that enhance representation diversity and align different modalities. Extensive experiments on four public datasets show that HPMRec achieves state-of-the-art recommendation performance.

CCS Concepts

• **Information systems → Recommender systems;**.

Keywords

Multimodal; Recommendation; Hypercomplex Algebra; Prompt-aware; Graph Learning

ACM Reference Format:
Zheyu Chen, Jinfeng Xu, Hewei Wang, Shuo Yang, Zitong Wan, and Haibo Hu. 2025. Hypercomplex Prompt-aware Multimodal Recommendation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3746252.3761174

1 Introduction

In the context of the exponential expansion of data volume, users encounter a significant challenge of information overload, thereby rendering recommender systems an effective method to mitigate this problem [15, 39, 41]. In multimodal recommendation scenarios, the synergy effect across modalities effectively mitigates the inherent data sparsity problem in traditional recommender systems[40, 45]. Since the data structure of the recommender system has a natural bipartite graph structure, graph convolutional network (GCN) technology is also widely used in the recommender system [7, 8, 42, 43]. Recently, the design of representation learning using hypercomplex algebra has garnered interest, and several works [5, 20, 31, 49] have started to investigate hypercomplex-based recommender systems in the conventional recommendation field. This design shares a similar motivation with multi-head mechanisms, as it enables parallel learning of diverse representations. Therefore, hypercomplex algebras offer a more expressive mathematical framework and enhance the capacity to encode multimodal information.

Despite these works exploring the multimodal recommendation and hypercomplex embedding ability, they still face three fundamental limitations: **Limitation 1:** Due to the richness of multimodal information, it is not sufficient for a single embedding to fully describe a user/item for each modality, and the traditional embedding structure restricts the representation diversity of multimodal features. **Limitation 2:** Existing linear modality fusion strategies (weighted sums or concatenations) ignore deep nonlinear correlations between modalities, which makes it difficult to fully explore the latent relationship between modalities, and leads to suboptimal exploration of cross-modality features. **Limitation 3:** For the over-smoothing problem, which indicates the representation tends to be indistinguishable from those of their neighbors during the message passing in GCN, existing methods [21, 24, 52] manually design static optimization strategies to mitigate the over-smoothing problem, without considering the dynamic mechanism.

*Equal contribution
†Corresponding author

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761174

To overcome these limitations, we propose the **H**ypercomplex **P**rompt-aware **M**ultimodal **Rec**ommendation **(HPMRec)** framework with the following tailored designs. Inspired by [19], we introduce the Cayley-Dickson algebra, an elegant structure of hypercomplex embedding that contains multiple components, as the structure of each modality's user/item representation. Based on this structure, we propose a hypercomplex graph convolution operator that learns these representations, enabling each component to capture diverse modality-specific features. Secondly, instead of regular dot products, the hypercomplex multiplication captures latent relations between two embeddings' components. We adopt this multiplication between different modalities' representations to capture nonlinear relations, which is beneficial to mine cross-modality features. In addition, we introduce a learnable prompt to dynamically compensate for the misalignment of components and the core modality-specific features loss. Moreover, the diversity of representations mitigates the over-smoothing problem by ensuring that the representations of users and items remain distinguishable from those of their neighbors. Furthermore, we design two self-supervised learning tasks. Specifically, we enhance user/item representation diversity by expanding the discrepancy between different components in hypercomplex embeddings, and we adopt cross-modality alignment, which also benefits modality fusion.

To summarize, our contributions are highlighted as follows:

- We propose the **HPMRec**, which utilizes hypercomplex embedding in the form of multi-components to enhance the representation diversity of modality-specific features. HPMRec leverages a novel nonlinear fusion strategy based on hypercomplex multiplication to bridge the semantic gap between modalities.
- We design the prompt-aware compensation mechanism to dynamically compensate for component misalignment and core modality-specific feature loss. This module also alleviates the over-smoothing problem.
- Our HPMRec integrates self-supervised learning tasks to enhance modal representation diversity by expanding the discrepancy between components to enhance the diversity of representation, and we also implement cross-modality alignment, which is beneficial for modality fusion.
- We conduct comprehensive experiments to show the effectiveness and robustness of HPMRec. These results show that our HPMRec outperforms state-of-the-art methods.

2 Related Work

In this section, we will introduce the latest works in multimodal recommendation methods, the application of hypercomplex algebra in recommendation systems, and the development of prompts in recommendation systems.

2.1 Multimodal Recommendation

To mitigate the data sparsity problem, recent multimodal recommendation models leverage visual and textual features through matrix factorization [6, 14] and graph-based architectures [32, 36, 53]. However, despite performance gains, three critical limitations remain. First, traditional embedding structures often force rich multimodal semantics into fixed-dimensional representations, hindering expressiveness. Second, modality fusion is typically handled via early [14, 48] or late [32, 36] strategies, both relying on linear operations that fail to capture latent cross-modality relations. Third, GCN-based methods such as NGCF [33] and LightGCN [15] suffer from over-smoothing, which recent multimodal extensions [24, 52] only mitigate through a static optimization strategy, lacking the ability of dynamic compensation. To this end, we propose the HPMRec, a novel framework that can overcome these limitations.

2.2 Hypercomplex-based Recommendation

Hypercomplex-based representation learning has proven effective in domains like computer vision [55] and natural language processing [30]. More recently, researchers have begun applying these techniques to recommender systems: previous works [4, 19] focused on pure collaborative filtering using interaction data, while subsequent studies [5, 20, 31] augmented that foundation by integrating auxiliary side information. Nevertheless, the inherent multi-component structure of the hypercomplex embedding makes it particularly suitable for encoding complex information such as multimodal features. To the best of my knowledge, no previous work has leveraged hypercomplex embeddings for multimodal recommendation. Our proposed HPMRec framework fills this gap and explores how hypercomplex embedding can benefit multimodal features through the representation capacity and structure.

2.3 Prompt-based Recommendation

Prompt learning has become an emerging research direction in the context of large pretrained models [3, 22], and some works explore the ability of prompt learning in the recommendation field. GraphPrompt [23] defines the paradigm of prompts on graphs. To transfer knowledge graph semantics into task data, KGTransformer [50] regards task data as a triple prompt for tuning. Additionally, prompt-based learning has also been introduced to enhance model fairness [37], sequence learning [38]. Recently, PromptMM [35] proposed a novel multimodal prompt learning method that can adaptively guide knowledge distillation. In our HPMRec, we consider using the prompt's capabilities to implement a dynamic compensation mechanism of the hypercomplex embedding, so that it can achieve diverse representations while retaining core modality-specific features. It also alleviates the inherent over-smoothing problem of graph convolutional networks through the diversity of representations. This design fully and reasonably utilizes the prompt's capabilities to improve recommendation performance.

3 Preliminary

3.1 Hypercomplex Algebra

A hypercomplex number $h_x \in \mathbb{H}_n$ in n-dimensional real vector space can be expressed as a representation in the form as follows:

$$h_x = x_1\mathbf{i}_1 + x_2\mathbf{i}_2 + \cdots + x_n\mathbf{i}_n = \sum_{k=1}^{n} x_k\mathbf{i}_k, \tag{1}$$

where $x_1, x_2, \cdots, x_n$ denote distinct components of the hypercomplex number. The elements $\mathbf{i}_1, \mathbf{i}_2, \cdots, \mathbf{i}_n$ are called hyperimaginary units, where $\mathbf{i}_1 = 1$ represents the vector identity element [1].

3.2 Cayley–Dickson Construction

The Cayley–Dickson algebra $\mathcal{A}$ is a sequence of hypercomplex algebras constructed from the real numbers using the Cayley–Dickson construction [2, 10]. Higher-dimensional Cayley–Dickson algebras can be obtained by doubling smaller algebras within the Cayley–Dickson construction [18]. Thus, the dimension of these algebras is a power of two. Specifically, such a construction procedure utilizes the n-th algebra $\mathcal{A}_n \in \mathbb{H}_{2^n}$ in the sequence to define the (n+1)-th algebra $\mathcal{A}_{n+1} \in \mathbb{H}_{2^{n+1}}$ as follows:

$$\mathcal{A}_{n+1} = \{h_a + h_b \mathbf{i}_{2^n+1}\}, \quad \text{and} \;\; h_a, h_b \in \mathcal{A}_n, \tag{2}$$

where $n \in \mathbb{N}$ and $\mathcal{A}_0 = \mathbb{R}$. Here $\mathbf{i}_{2^n+1} \notin \mathcal{A}_n$ is the additional hyperimaginary unit for doubling the dimension of $\mathcal{A}_n$, satisfying the following rules: $(\mathbf{i}_{2^n+1})^2 = -1$, $\mathbf{i}_1\mathbf{i}_{2^n+1} = \mathbf{i}_{2^n+1}\mathbf{i}_1$ and $\mathbf{i}_o\mathbf{i}_{2^n+1} = -\mathbf{i}_{2^n+1}\mathbf{i}_o = \mathbf{i}_{2^n}\mathbf{i}_o$ for all $o = 2, 3, \cdots, 2^n$.

The mathematical operations for Cayley-Dickson algebras are defined recursively [2, 9, 10]. For $h_x = h_a + h_b\mathbf{i}_{2^n+1} \in \mathcal{A}_{n+1}$, $h_y = h_c + h_d\mathbf{i}_{2^n+1} \in \mathcal{A}_{n+1}$; $h_a \in \mathcal{A}_n$, $h_b \in \mathcal{A}_n$, $h_c \in \mathcal{A}_n$, $h_d \in \mathcal{A}_n$; and $\gamma \in \mathbb{R}$, several widely utilized operations for Cayley-Dickson algebras are introduced as follows:

- The **addition** of h_x and h_y is defined as: $h_x \oplus_{\text{n+1}} h_y = (h_a \oplus_{\text{n}} h_c) + (h_b \oplus_{\text{n}} h_d)\,\mathbf{i}_{2^n+1}$. The *subtraction* follows the same principle analogously, but flipping $\oplus$ with $\ominus$.
- The **conjugate** of h_x is defined as: $\bar{h}_x = \bar{h}_a - h_b\mathbf{i}_{2^n+1}$. The conjugation for every $a \in \mathbb{R}$ is defined as: $\bar{a} = a$. $h_x \oplus_{\text{n+1}} h_y = (h_a \oplus_{\text{n}} h_c) + (h_b \oplus_{\text{n}} h_d)\,\mathbf{i}_{2^n+1}$.
- The **multiplication** of h_x and h_y is defined as: $h_x \otimes_{\text{n+1}} h_y = \left(h_a \otimes_{\text{n}} h_c \ominus_{\text{n}} \bar{h}_d \otimes_{\text{n}} h_b\right) + \left(h_a \otimes_{\text{n}} h_d \oplus_{\text{n}} h_b \otimes_{\text{n}} \bar{h}_c\right)\mathbf{i}_{2^n+1}$. When $n \geq 2$, the multiplication is asymmetric.
- The **scalar multiplication** of h_x by γ is defined as: $\gamma h_x = \gamma h_a + \gamma h_b\mathbf{i}_{2^n+1}$. Based on this form of mathematical operations, we can use the low-dimensional algebra operations to study the high-dimensional algebra operations recursively. We will employ these mathematical operations of Cayley-Dickson algebras to design our method HPMRec.

4 Methodology

In this section, we first formulate the problems. Then, we elaborate on the HPMRec framework. Finally, we discuss the optimization process of our HPMRec. The overall framework of HPMRec[1] is shown in Figure 1.

4.1 Task Definition

Let $\mathcal{U} = \{u_1, ..., u_{|\mathcal{U}|}\}$ denotes user set, $\mathcal{V} = \{v_1, ..., v_{|\mathcal{V}|}\}$ denotes item set, and $\mathcal{N} = \mathcal{U} \cup \mathcal{V}$ includes both user and item sets. We conceptualize the user-item graph $\mathcal{G} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$, where $\mathcal{U}, \mathcal{V}$ serve as the graph vertices, and $\mathcal{E}$ denotes the edge set. In the multimodal scenario, each item contains multiple modality features. We introduce modality-specific user/item embedding $\mathbf{h}^m_{u/v}$ for each u/v belonging to the set of modalities $\mathcal{M}$, and we let $\mathbf{H}^m_{u/v}$ denote the entire representation of user/item. Similarly, we let $p^m_{u/v}$ denote the learnable prompt of each user/item, and we let $\mathcal{P}^m_{u/v}$ denote the entire learnable prompt of all users/items. The historical interaction matrix is represented by $\mathcal{R} \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{V}|}$, in which $r_{uv} = 1$ indicates that user $u \in \mathcal{U}$ has engaged with item $v \in \mathcal{V}$ and zero otherwise. The goal of our HPMRec is to use the interaction matrix $\mathcal{R}$ and correlation features of each modality $m \in \mathcal{M}$ to predict user u's preference for item v that the user has never engaged with before.

4.2 Hypercomplex Multimodal Encoder

To enhance the representation ability, we introduce the hypercomplex algebra as the structure of representation. This structure makes modality-specific features no longer limited to a single vector, which improves the representation diversity of each node during the training process and helps to capture users, items, and their relationships in the user-item interaction graph at a more fine-grained level. Recent works [48, 53] find that jointly leveraging user-item heterogeneous graphs and item-item homogeneous graphs can substantially enhance recommendation performance. Building upon these findings, we develop a tailored hypercomplex multimodal encoder architecture to learn modality-specific features through user-item and item-item graphs. To be specific, we introduce the Cayley-Dickson algebra, a hypercomplex structure that contains multiple components, as the structure of the representation. Based on this structure, we propose a hypercomplex graph convolution operator that learns node representations, enabling each component to capture diverse modality-specific features.

4.2.1 Hypercomplex Embedding. We utilize the Cayley-Dickson construction to encode user and item features with modality m in the hypercomplex space $\mathcal{A}^m_{n+1}$. For user u and item v, their hypercomplex embeddings are defined as $\mathbf{h}^m_u$ and $\mathbf{h}^m_v$, respectively. We utilize items' embedding $\mathbf{h}^m_v$ as an example to illustrate this process:

$$\mathbf{h}^m_v = \mathbf{x}^m_v + \mathbf{y}^m_v\mathbf{i}_{2^n+1} = \sum_{k=1}^{2^{n+1}} \mathbf{v}^m_k \mathbf{i}_k, \tag{3}$$

where $n \in \mathbb{N}$; $\mathbf{x}^m_v = \sum_{k=1}^{2^n} \mathbf{v}^m_k \mathbf{i}_k \in \mathcal{A}^m_n$ and $\mathbf{y}^m_v = \sum_{k=2^n+1}^{2^{n+1}} \mathbf{v}^m_k \mathbf{i}_{k-2^n} \in \mathcal{A}^m_n$ are the subalgebras of $\mathbf{h}^m_v$; $\mathbf{v}^m_k \in \mathbb{R}^d$ is the real-valued representation for component k, and d denotes the feature dimension. Similar hypercomplex embedding is defined for user u.

4.2.2 Heterogeneous Graph. To capture high-order modality-specific features, we construct two **user-item graphs** $\mathcal{G} = \{\mathcal{G}^m \mid m \in \mathcal{M}\}$. Each graph $\mathcal{G}^m$ maintains the same graph structure and only retains the node features associated with each modality. Formally, the message propagation at l-th graph convolution layer can be formulated as:

$$\mathbf{h}^m_v(l) = \sum_{u \in \mathcal{N}_v} \frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_v|}} \mathbf{h}^m_u(l-1), \tag{4}$$

$$\mathbf{h}^m_u(l) = \sum_{v \in \mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_v|}} \mathbf{h}^m_v(l-1), \tag{5}$$

where $\mathbf{h}^m_{u/v}(l)$ represents the multi-component user/item representation in modality m at l-th graph convolution layer. $\mathcal{N}_{u/v}$ denotes the one-hop neighbors of u/v in $\mathcal{G}$. Then, we compute the final user/item embedding of each modality, $\bar{\mathbf{h}}^m_{u/v}$, and describe its aggregation process in detail in Section 4.3.

[1] The code is available at: https://github.com/Zheyu-Chen/HPMRec

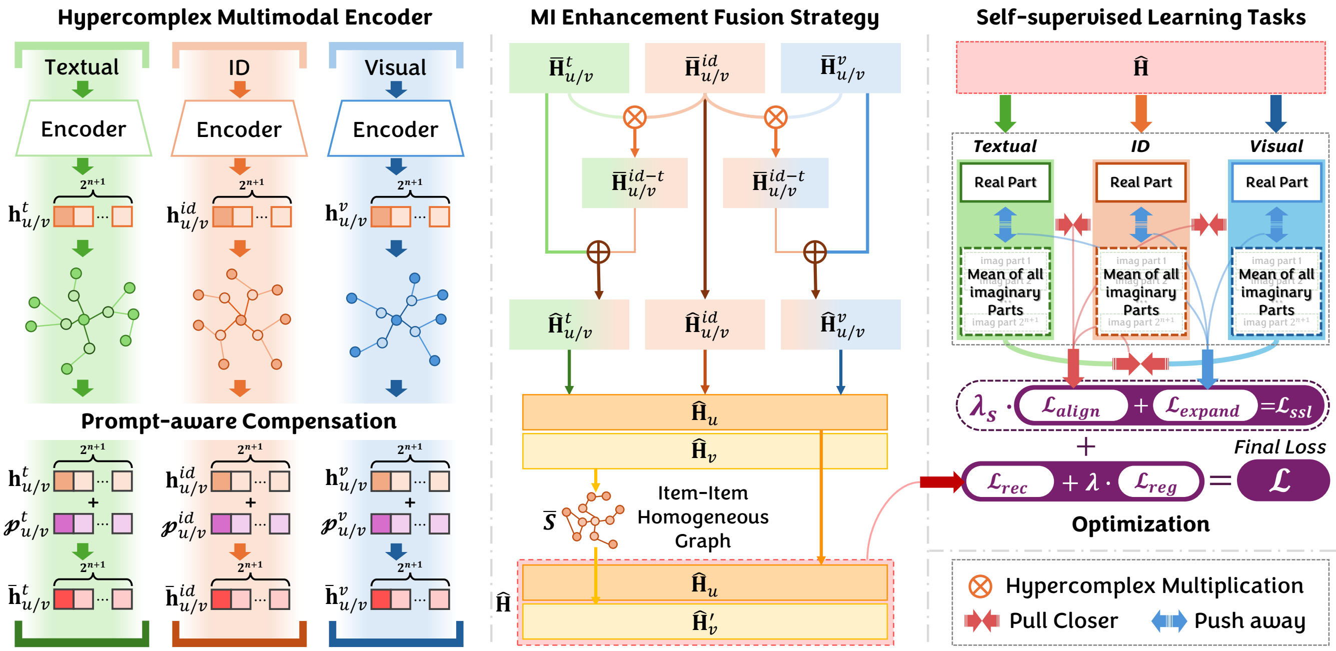
Hypercomplex Multimodal Encoder
MI Enhancement Fusion Strategy
Self-supervised Learning Tasks
Textual
ID
Visual
Encoder
2^{n+1}
$\mathbf{h}^t_{u/v}$
$\mathbf{h}^{id}_{u/v}$
$\mathbf{h}^v_{u/v}$
Prompt-aware Compensation
$p^t_{u/v}$
$p^{id}_{u/v}$
$p^v_{u/v}$
$\bar{\mathbf{h}}^t_{u/v}$
$\bar{\mathbf{h}}^{id}_{u/v}$
$\bar{\mathbf{h}}^v_{u/v}$
$\bar{\mathbf{H}}^t_{u/v}$
$\bar{\mathbf{H}}^{id}_{u/v}$
$\bar{\mathbf{H}}^v_{u/v}$
$\bar{\mathbf{H}}^{id-t}_{u/v}$
$\bar{\mathbf{H}}^{id-t}_{u/v}$
$\hat{\mathbf{H}}^t_{u/v}$
$\hat{\mathbf{H}}^{id}_{u/v}$
$\hat{\mathbf{H}}^v_{u/v}$
$\hat{\mathbf{H}}_u$
$\hat{\mathbf{H}}_v$
$\bar{S}$
Item-Item Homogeneous Graph
$\hat{\mathbf{H}}$
$\hat{\mathbf{H}}_u$
$\hat{\mathbf{H}}'_v$
$\hat{\mathbf{H}}$
Real Part
imag part 1
Mean of all imaginary Parts
imag part 2^{n+1}
$\lambda_S \cdot$ $\mathcal{L}_{align}$ + $\mathcal{L}_{expand}$ $=\mathcal{L}_{ssl}$
+
Final Loss
$\mathcal{L}_{rec}$ $+ \lambda \cdot$ $\mathcal{L}_{reg}$ = $\mathcal{L}$
Optimization
Hypercomplex Multiplication
Pull Closer
Push away

Figure 1: Overall Framework of HPMRec.

4.2.3 Homogeneous Graph. We use k-NN to establish the **item-item graph** based on the item features for each modality m to extract significant semantic relations between items. Particularly, we calculate the similarity score $S^m_{v,v'}$ between item pair $(v, v') \in \mathcal{V}$ by the cosine similarity $\mathrm{Sim}(\cdot)$ on their modality original features f^m_v and $f^m_{v'}$.

$$S^m_{v,v'} = \mathrm{Sim}(f^m_v, f^m_{v'}) = \frac{\left(f^m_v\right)^\top f^m_{v'}}{\|f^m_v\| \, \|f^m_{v'}\|}. \tag{6}$$

We only retain the top-k neighbors:

$$\bar{S}^m_{v,v'} = \begin{cases} S^m_{v,v'} & \text{if } S^m_{v,v'} \in \text{top-}k(S^m_{v,p} \mid p \in \mathcal{V}) \\ 0 & \text{otherwise} \end{cases}, \tag{7}$$

where $\bar{S}^m_{v,v'}$ represents the edge weight between item v and item v' within modality m. Thereafter, we further build a unified item-item graph $\bar{S}$ by aggregating all modality-specified graphs $\bar{S}^m$:

$$\bar{S} = \sum_{m \in \mathcal{M}} \alpha^m \bar{S}^m. \tag{8}$$

Inspired by [53], we freeze each item-item graph after initialization to eliminate the computational cost of the item-item graph during training. In addition, α_m is a trainable parameter with the same initial value for each modality.

4.3 Prompt-aware Compensation

Due to the multi-component structure of hypercomplex embedding, each component of each modality is able to learn different features, which poses a challenge to the efficient use of these representations. As these highly diverse components deviate from the initial semantic space, the representations of multiple components are misaligned. Directly concatenating[2] these components will result in a low-quality user/item representation and will even lose the core modality-specific features.

To this end, we designed the learnable prompt $p \in \mathbb{R}^{d \cdot 2^{n+1}}$ to independently compensate the features learned by each component. The final embedding for each modality m is as follows:

$$\bar{\mathbf{h}}^m_{u/v} = \sum_{l=0}^{L} \mathbf{h}^m_{u/v}(l) + p^m_{u/v}, \tag{9}$$

where L is the number of user-item graph layers. In addition, in the message passing process of GCN, user/item representation will inevitably tend to be the same as their neighbors'. HPMRec allows representations to learn diversity and then uses the learnable prompt for dynamic compensation, which fundamentally alleviates the over-smoothing problem. Specifically, the learnable prompt keeps the core modality-specific features in each component, and the diversity is retained. Therefore, the diversity ensures that the representations are not over-smoothing.

4.4 MI Enhancement Fusion Strategy

Previous works [21, 24] adopt linear strategies to fuse modalities, such as weighted sums or concatenations. However, linear fusion can not sufficiently mine the latent relation among modalities.

[2] If we accumulate or calculate the mean of the representations of these components, we can avoid this problem, but the diverse representations learned in each component will be lost. It goes against the purpose of adopting the multi-component structure.

Therefore, we apply hypercomplex algebraic multiplication to naturally build nonlinear relations among different modalities' components, enhancing the representation's ability to mine latent cross-modality features.

$$\bar{\mathbf{H}}^{id-v}_{u/v} = \bar{\mathbf{H}}^{id}_{u/v} \otimes_{n+1} \bar{\mathbf{H}}^{v}_{u/v}, \tag{10}$$

$$\bar{\mathbf{H}}^{id-t}_{u/v} = \bar{\mathbf{H}}^{id}_{u/v} \otimes_{n+1} \bar{\mathbf{H}}^{t}_{u/v}. \tag{11}$$

When two hypercomplex algebras are multiplied, the product incorporates the nonlinearities and higher-order dependencies between the original algebras [28]. Next, we add it back to the original modalities to enhance their cross-modality features, which is beneficial to modality fusion.

$$\hat{\mathbf{H}}^{v}_{u/v} = \bar{\mathbf{H}}^{v}_{u/v} + \epsilon_1 \cdot \bar{\mathbf{H}}^{id-v}_{u/v}, \tag{12}$$

$$\hat{\mathbf{H}}^{t}_{u/v} = \bar{\mathbf{H}}^{t}_{u/v} + \epsilon_2 \cdot \bar{\mathbf{H}}^{id-t}_{u/v}, \tag{13}$$

where ϵ_1, ϵ_2 are trainable parameters to control the MI enhancement strength, which are empirically initialized with 0.1. To simplify the formula expression, we let $\hat{\mathbf{H}}^{id}_{u/v} = \bar{\mathbf{H}}^{id}_{u/v}$. Then we calculate the final user/item representations:

$$\hat{\mathbf{H}}_{u/v} = \text{Con}(\beta^m \hat{\mathbf{H}}^m_{u/v} \mid m \in \mathcal{M}), \tag{14}$$

where the attention weight β^m is a trainable parameter, which is initialized with equal value for each modality. Then we enhance the item representations $\hat{\mathbf{H}}_v$ by item-item graphs $\bar{S}$. Ultimately, we fuse the final user representation and enhanced item representations to get the final representation:

$$\hat{\mathbf{H}} = \text{Con}(\hat{\mathbf{H}}_u, \hat{\mathbf{H}}_v^{'}), \quad \hat{\mathbf{H}}_v^{'} = \hat{\mathbf{H}}_v + \bar{S} \cdot \hat{\mathbf{H}}_v. \tag{15}$$

4.5 Self-Superised Learning Tasks

4.5.1 Cross-modality Alignment. We employ self-supervised learning, taking the mean of the Manhattan distance[3] to align ID-visual, ID-textual, and visual-textual modality pairs. Formally:

$$\mathcal{L}_{align} = -\frac{1}{|\mathcal{N}|} \sum_{n \in \mathcal{N}} \sum_{(a,b) \in C} \left[|\hat{\mathbf{h}}^a_n - \hat{\mathbf{h}}^b_n| \right], \tag{16}$$

where $C \in \{(id, v), (id, t), (v, t)\}$ denote set of modality pairs. This task brings the representations of each modality closer, which is beneficial to modality fusion and final rating prediction.

4.5.2 Real-Imag Discrepancy Expansion. We expand the discrepancy among different components to enhance the diversity of user/item representation. Specifically, we directly take the mean of the Manhattan distance between the real part and the mean of all imaginary parts of each modality.

$$\mathcal{L}_{expand} = -\frac{1}{|\mathcal{N}|} \sum_{n \in \mathcal{N}} \sum_{m \in \mathcal{M}} \left[|\hat{\mathbf{c}}^m_n - \mathbb{E}[\hat{\mathbf{d}}^m_n]| \right], \tag{17}$$

where $\mathcal{M} \in \{id, v, t\}$, $\mathcal{N} = \mathcal{U} \cup \mathcal{V}$, and $\mathbb{E}[\cdot]$ represents the mean calculation of the component-level. $\hat{\mathbf{c}}^m_n$ and $\hat{\mathbf{d}}^m_n$ denote the real part and imaginary parts of node n's representation $\hat{\mathbf{h}}^m_n$, respectively.

Algorithm 1 Learning Process of HPMRec

```
Input: $\mathcal{U}$, $\mathcal{V}$, $\mathcal{M}$, $\mathcal{G}$, node set $\mathcal{N} = \mathcal{U} \cup \mathcal{V}$, layer number $L$ of heterogeneous graph $\mathcal{G}$.
Output: Optimization loss $\mathcal{L}$
Initialize $\mathbf{H}^m_u, \mathbf{H}^m_v, \mathcal{P}^m_u, \mathcal{P}^m_v$;
for $l = 1...L$ do
    Conduct message passing in the heterogeneous graph $\mathbf{h}^m_v(l) \leftarrow \mathbf{h}^m_u(l-1)$ with Eq.4, or $\mathbf{h}^m_u(l) \leftarrow \mathbf{h}^m_v(l-1)$ with Eq.5;
end for
Get pormpt-aware compensated embedding $\bar{\mathbf{h}}^m_u$, $\bar{\mathbf{h}}^m_v$ for each modality with Eq.9;
Construct the unified item-item graph $\bar{S}$ with Eq.6-8;
Represent all node embeddings $\mathbf{h}$ as the entire node representation $\mathbf{H}$.
Apply hypercomplex multiplication with Eq.10-11;
Get enhanced representation $\hat{\mathbf{H}}^v_{u/v}$ and $\hat{\mathbf{H}}^t_{u/v}$ with Eq.12-13;
Attentively fuse all modality representations $\hat{\mathbf{H}}_{u/v} \leftarrow \hat{\mathbf{H}}^m_{u/v}$ with Eq.14;
Get final representation $\hat{\mathbf{H}}$ by item-item graphs $\bar{S}$ with Eq.15.
Calculate self-supervised learning loss $\mathcal{L}_{ssl}$ with Eq.16-18;
Calculate adaptive BPR loss $\mathcal{L}_{rec}$ with Eq.19;
Get final optimization loss $\mathcal{L}$ with Eq.20.
```

Here is the final self-supervised learning loss, formally:

$$\mathcal{L}_{ssl} = \mathcal{L}_{align} + \mathcal{L}_{expand}. \tag{18}$$

4.6 Optimization

We adopt LightGCN [15] as the backbone model and employ the Bayesian Personalized Ranking (BPR) loss [26] as the primary optimization objective. The BPR loss is specifically designed to improve the predicted preference distinction between positive and negative items for each triplet $(u, p, n) \in \mathcal{D}$, where $\mathcal{D}$ represents the training dataset. In this context, the positive item p is one with which user u has interacted, while the negative item n is randomly selected from the set of items that user u has not interacted with. Formally:

$$\mathcal{L}_{rec} = \sum_{(u,p,n) \in \mathcal{D}} -\log(\sigma(y_{u,p} - y_{u,n})) + \lambda \cdot \|\Theta\|^2_2, \tag{19}$$

where σ represents the sigmoid function, and λ controls the strength of L_2 regularization, and Θ denotes the parameters subject to regularization. The terms $y_{u,p}$ and $y_{u,n}$ correspond to the ratings of user u for the positive item p and the negative item n, respectively, computed as $\hat{\mathbf{h}}^\top_u \cdot \hat{\mathbf{h}}_p$ and $\hat{\mathbf{h}}^\top_u \cdot \hat{\mathbf{h}}_n$. The final loss function is given by:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_s \mathcal{L}_{ssl}, \tag{20}$$

where λ_s is the self-superised learning balancing hyper-parameter.

To provide a clearer overview of our HPMRec, we summarize the learning process of HPMRec in Algorithm 1.

5 Experiments

In this section, we conduct comprehensive experiments to evaluate the performance of our HPMRec framework on four widely used real-world datasets. The following five research questions can be well answered through experimental results: **RQ1:** Does

[3] We also considered adopting the Euclidean distance, but since the performance difference was almost the same and the performance consumption was higher, we chose an easy-to-use and effective method. In addition, compared with high-order distances, low-order metrics are more stable and less susceptible to extreme values, which is conducive to the stability of model training.

HPMRec outperform the state-of-the-art conventional and multimodal recommendation methods? **RQ2:** What impact do the key modules of our HPMRec framework have on its overall performance? **RQ3:** How does the representation scaling strategy affect the performance-efficiency trade-off? **RQ4:** How efficient is HPMRec compared with various state-of-the-art recommender systems? **RQ5:** How do different hyper-parameter settings impact the overall performance of HPMRec?

5.1 Datasets and Evaluation Metrics

To evaluate the performance of our proposed HPMRec in the recommendation task, we perform comprehensive experiments on four widely used Amazon datasets [25]: Office, Baby, Sports, and Clothing. These datasets offer both product descriptions and images. In line with previous works [44, 46], we preprocess the raw data with a 5-core setting for both items and users. Additionally, we utilize pre-extracted 4096-dimensional visual features and obtain 384-dimensional textual features using a pre-trained sentence transformer [51]. For a fair evaluation, we employ two widely recognized metrics: Recall@K (R@K) and NDCG@K (N@K). We present the average metrics for all users in the test dataset for both $K = 10$ and $K = 20$. We adhere to the standard procedure [53] with a random data split of 8:1:1 for training, validation, and testing.

Table 1: Statistics of datasets.

Datasets	#Users	#Items	#Interactions	Sparsity
Office	4,905	2,420	53,258	99.55%
Baby	19,445	7,050	160,792	99.88%
Sports	35,598	18,357	296,337	99.95%
Clothing	39,387	23,033	278,677	99.97%

5.2 Baselines

To comprehensively evaluate the effectiveness of HPMRec, we conduct a systematic comparison with state-of-the-art methods, categorized into traditional recommendation methods (focusing on collaborative filtering and graph-based learning) and multimodal recommendation methods (leveraging multiple modalities such as visual and textual features). Below, we provide a concise yet informative overview of each baseline method.

1) Conventional recommendation methods:

- **MF-BPR** [27]: optimized with Bayesian Personalized Ranking (BPR) loss, designed for learning user and item embeddings from implicit feedback.
- **LightGCN** [15]: removes unnecessary modules: nonlinear activations to improve recommendation performance.
- **SimGCL** [47]: enhances representation robustness by injecting controlled noise into embeddings.
- **LayerGCN** [52]: alleviating LightGCN's over-smoothing issue via residual connections, refining layer-wise aggregation for deeper GCNs.

2) Multimodal recommendation methods:

- **VBPR** [14]: extends matrix factorization by incorporating visual and textual features as side information for items.
- **MMGCN** [36]: employs separate GCNs per modality and fuses modality-specific predictions for final recommendations.
- **DualGNN** [32]: introduces a user-user graph to model latent preference patterns beyond user-item interactions.
- **LATTICE** [48]: constructs an item-item graph to capture high-order semantic relationships among items.
- **FREEDOM** [53]: enhances LATTICE by freezing the item-item graph and denoising the user-item graph.
- **SLMRec** [29]: employs node self-discrimination to uncover multimodal item patterns.
- **BM3** [54]: simplifies self-supervised learning via dropout-based representation perturbation.
- **MMSSL** [34]: combines modality-aware adversarial training with cross-modal contrastive learning to disentangle shared and modality-specific features.
- **LGMRec** [13]: unifies local (graph-based) and global (hypergraph-based) embeddings for multimodal recommendation.
- **DiffMM** [16]: leverages modality-aware graph diffusion to improve user representation learning.

5.3 Experimental Settings

Following the basic settings of previous works [53], we implement HPMRec in PyTorch and optimize with the Adam optimizer [17]. We apply Xavier initialization [12] for all initial random embeddings. As for hyper-parameter settings on HPMRec, we perform a grid search on the user-item heterogeneous graph $\mathcal{G}$'s GCN layer number L in {1, 2, 3}, regularization balancing hyper-parameter λ in $\{1e^{-2}, 1e^{-3}, 1e^{-4}\}$, self-supervised learning balancing hyper-parameter λ_s in $\{1e^{-2}, 1e^{-3}, 1e^{-4}\}$. We set n in hypercomplex algebra to {0, 1, 2, 3}, which indicate the components number 2^{n+1} in {2, 4, 8, 16}. We fix the learning rate as $1e^{-4}$, and adopt a single-layer GCN in the item-item homogeneous graph. The k of top-k in the item-item graph is set as 10. For convergence consideration, we fixed the early stopping at 20. Following the settings of [51], we update the best record by utilizing Recall@20 on the validation dataset as the indicator. All the experiments were conducted on the NVIDIA GeForce RTX 3090 GPU.

5.4 Overall Performance (RQ1)

Detailed experiment results are shown in Table 2. The optimal results are highlighted in bold, while the suboptimal ones are underlined. We have the following key observations:

Our framework consistently outperforms all baselines across all datasets and evaluation metrics, demonstrating both its effectiveness across datasets with varying scales and sparsity.

The multi-component structure of hypercomplex embeddings and the prompt-aware compensation mechanism effectively enhance the ability of representation. The hypercomplex embedding provides multiple components to capture the diverse modality-specific features, and the learnable prompt is able to dynamically compensate for the misalignment of multiple components and the loss of core modality-specific features. And due to the diversity of components, the user/item representation is not the same as the neighbors', so that it keeps the representation away from over-smoothing problems. Compared to previous works [21, 24, 52] that

Table 2: Performance comparison of baselines and HPMRec(our) in terms of Recall@K(R@K) and NDCG@K(N@K).

Model	Office				Baby				Sports				Clothing			
	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
MF-BPR	0.0572	0.0951	0.0331	0.0456	0.0357	0.0575	0.0192	0.0249	0.0432	0.0653	0.0241	0.0298	0.0187	0.0279	0.0103	0.0126
LightGCN	0.0791	0.1189	0.0459	0.0583	0.0479	0.0754	0.0257	0.0328	0.0569	0.0864	0.0311	0.0387	0.0340	0.0526	0.0188	0.0236
SimGCL	0.0799	0.1239	0.0470	0.0595	0.0513	0.0804	0.0273	0.0350	0.0601	0.0919	0.0327	0.0414	0.0356	0.0549	0.0195	0.0244
LayerGCN	0.0825	0.1213	0.0486	0.0593	0.0529	0.0820	0.0281	0.0355	0.0594	0.0916	0.0323	0.0406	0.0371	0.0566	0.0200	0.0247
VBPR	0.0692	0.1084	0.0422	0.0531	0.0423	0.0663	0.0223	0.0284	0.0558	0.0856	0.0307	0.0384	0.0281	0.0415	0.0158	0.0192
MMGCN	0.0558	0.0926	0.0312	0.0413	0.0378	0.0615	0.0200	0.0261	0.0370	0.0605	0.0193	0.0254	0.0218	0.0345	0.0110	0.0142
DualGNN	0.0887	0.1350	0.0505	0.0631	0.0448	0.0716	0.0240	0.0309	0.0568	0.0859	0.0310	0.0385	0.0454	0.0683	0.0241	0.0299
LATTICE	0.0969	0.1421	0.0562	0.0686	0.0547	0.0850	0.0292	0.0370	0.0620	0.0953	0.0335	0.0421	0.0492	0.0733	0.0268	0.0330
FREEDOM	0.0974	0.1445	0.0549	0.0669	0.0627	0.0992	0.0330	0.0424	0.0717	0.1089	0.0385	0.0481	0.0629	0.0941	0.0341	0.0420
SLMRec	0.0790	0.1252	0.0475	0.0599	0.0529	0.0775	0.0290	0.0353	0.0663	0.0990	0.0365	0.0450	0.0452	0.0675	0.0247	0.0303
BM3	0.0715	0.1155	0.0415	0.0533	0.0564	0.0883	0.0301	0.0383	0.0656	0.0980	0.0355	0.0438	0.0422	0.0621	0.0231	0.0281
MMSSL	0.0794	0.1273	0.0481	0.0610	0.0613	0.0971	0.0326	0.0420	0.0673	0.1013	0.0380	0.0474	0.0531	0.0797	0.0291	0.0359
LGMRec	0.0959	0.1402	0.0514	0.0663	0.0639	0.0989	0.0337	0.0430	0.0719	0.1068	0.0387	0.0477	0.0555	0.0828	0.0302	0.0371
DiffMM	0.0733	0.1183	0.0439	0.0560	0.0623	0.0975	0.0328	0.0411	0.0671	0.1017	0.0377	0.0458	0.0522	0.0791	0.0288	0.0354
HPMRec	**0.1092**	**0.1632**	**0.0632**	**0.0778**	**0.0667**	**0.1033**	**0.0357**	**0.0451**	**0.0751**	**0.1129**	**0.0410**	**0.0507**	**0.0658**	**0.0963**	**0.0351**	**0.0429**

utilize static optimization to alleviate the over-smoothing problem, we have superior performance.

Our MI enhancement fusion strategy and self-supervised learning tasks also positively impact overall performance, making the framework more robust. Notably, our nonlinear fusion strategy outperforms existing linear and attention-based strategies through deeper latent relation exploration. Comprehensive ablation studies in Section 5.5 will systematically dissect each module's contribution to overall performance.

Table 3: Performance Comparison on variants of HPMRec.

Variant	Baby		Sports	
	Recall@10	NDCG@10	Recall@10	NDCG@10
w/o Prompt	0.0592	0.0312	0.0541	0.0297
w/o MI	0.0657	0.0351	0.0739	0.0403
w/o SSL	0.0651	0.0343	0.0742	0.0401
w/o Enhance	0.0640	0.0346	0.0733	0.0398
w Explicit	0.0654	0.0348	0.0736	0.0403
HPMRec-Split	0.0654	0.0354	0.0708	0.0381
HPMRec-MLP	0.0582	0.0319	0.0682	0.0369
HPMRec	**0.0667**	**0.0357**	**0.0751**	**0.0410**

5.5 Ablation Study (RQ2 & RQ3)

In this section, we conduct extensive experiments to evaluate the effectiveness of each module in HPMRec. We also explored how the node representation's feature dimension scaling strategy affects the performance-efficiency trade-off in resource-constrained scenarios.

5.5.1 Effectiveness of key modules of HPMRec (RQ2).

- **HPMRec** *w/o* **Prompt:** Remove the learnable prompt from each node representation.
- **HPMRec** *w/o* **MI:** Remove the MI enhancement operation from the fusion stage.

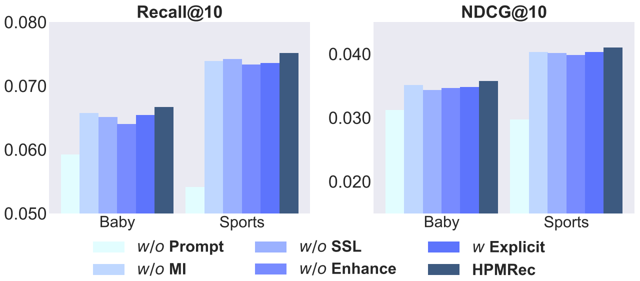

Figure 2: Effect of key modules in HPMRec.

- **HPMRec** *w/o* **SSL:** Remove the self-supervised learning tasks.
- **HPMRec** *w/o* **Enhance:** Remove both the MI enhancement operation and the self-supervised learning tasks.
- **HPMRec** *w* **Explicit:** Explicitly align the prompt with the initial layer node representation.

Detailed ablation study experiment results are shown in Table 2 and Figure 2. We have the following key observations:

Our multi-component hypercomplex embedding comprehensively explores the representational potential of each node, enabling rich and detailed feature extraction. However, introducing hypercomplex embeddings inevitably results in component-level misalignment, hindering effective representation learning. To this end, we propose a prompt-aware compensation mechanism that adaptively aligns the semantic spaces of different components. The performance of variant **HPMRec** *w/o* **Prompt** shows that the prompt-aware compensation mechanism is significant for our framework, and adopting hypercomplex embedding with multi-component structure alone is not feasible in multimodal scenarios. According to the result of the variant **HPMRec** *w/o* **Enhance**, with only the hypercomplex embedding and learnable prompt, we still surpass all baselines, indicating the significant effectiveness of these two modules.

The performance degradation observed in variant **HPMRec** *w/o* **MI** and variant **HPMRec** *w/o* **SSL** confirms the effectiveness of our MI enhancement fusion strategy and self-supervised learning tasks. These modules contribute not only to performance improvements but also to better framework robustness. In particular, the self-supervised learning module facilitates cross-modal alignment among ID, visual, and textual representations and enhances the diversity of different components for hypercomplex embedding. Moreover, our MI enhancement fusion strategy, which is based on hypercomplex multiplication, a naturally nonlinear calculation, outperforms existing linear and attention-based fusion strategies, demonstrating its superior capability in capturing the cross-modality features.

In the variant **HPMRec** *w* **Explicit**, we design the closest explicit guidance to the motivation of designing the learnable prompt, that is, aligning the learnable prompt with initial layer node representations. However, the performance degradation observed in variant **HPMRec** *w* **Explicit** demonstrates that the explicit guidance will harm the ability of the learnable prompt. Therefore, we employ no optimization task for the learnable prompt explicitly, only utilize the main recommendation task and self-supervised learning tasks to implicitly benefit its dynamic optimization. Our HPMRec's higher performance results shows the learnable prompt can perform better in implicit guidance than explicit guidance. We will further discuss how to maximize the ability of prompt in Section 6.2.

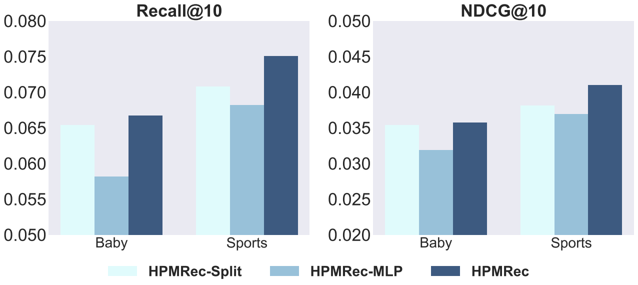

Figure 3: Effect of Dimension Scale Trade-off Strategy.

5.5.2 *Feature dimension scale trade-off strategy (RQ3).*

In Section 5.7, we found that the best performance of the framework does not gain advantages when the component number 2^{n+1} is very large, which makes it unnecessary to consider the computational resource consumption caused by the limited representation's feature dimension increasing, and we also conduct effieiency study in Section 5.6 to demonstrate the competitive efficiency of our framework. However, to consider a more comprehensive computation resources scenario, we design the following variant to explore the consumption-performance trade-off. Detailed ablation study experiment results are shown in Table 2 and Figure 3.

- **HPMRec-Split:** In the modality information encoding stage, each modality representation is partitioned into 2^{n+1} equal segments as components.
- **HPMRec-MLP:** In the modality information encoding stage, this variant employs an MLP to compress the feature dimension of each component from d to $d/2^{n+1}$, which will keep the feature dimension of each node representation to d, instand of $d \cdot 2^{n+1}$.

The dimensionality reduction of variant **HPMRec-Split** inevitably sacrifices some representation capacity, thereby capping the framework's potential performance. Furthermore, the simple dimension compression of variant **HPMRec-MLP** makes each component small and similar, meanwhile loses the diversity of user/item representation, which leads to a significant performance degradation.

The performance results of the two variants in Table 3 show that a sufficient feature dimension is crucial for exploring multimodal information. If the feature dimension scale is simply limited, the representation's capability will be reduced due to the lack of diverse modality-specific features. In addition, we found that variant **HPMRec-Split** has higher performance than variant **HPMRec-MLP**, which shows that although the representation is divided into multiple components, the origin modality-specific features are complete, ensuring a certain richness and diversity. It can still restore some representation capabilities under the compensation of the learnable prompt. The variant **HPMRec-MLP**, which directly compresses the feature dimension of the representation to a very small scale, not only fails to retain the core modality-specific features but also loses diversity. Although prompt has the ability to compensate the core modality-specific features, the user/item representation's diversity has been lost. This situation will be more obvious as the hypercomplex dimension grows, because the feature dimension of its single component will shrink as the component number grows.

In summary, variant **HPMRec-Split** reduces both computation and memory requirements in resource-constrained scenarios, which is suitable for scenarios with extremely limited computational resources, whereas variant **HPMRec-MLP** fails to achieve a favorable trade-off between efficiency and effectiveness due to the loss of component diversity. This ablation study experiment proves that multimodal information requires a sufficient feature dimension scale to explore diverse features, which is consistent with our motivation for using hypercomplex embedding. And the higher performance of variant **HPMRec-Split** also indirectly proves the effectiveness of our prompt-aware compensation mechanism.

5.6 Efficiency Study (RQ4)

We report the training time per epoch and memory usage of HPMRec and baselines in Table 4[4]. After analyzing the results of efficiency, we found that our framework maintains competitive efficiency in terms of training time per epoch and memory usage.

Thanks to the multi-component structure, the node representation contains rich and diverse modality-specific features. Thus, these powerful representations enable the framework to achieve the best performance with fewer convolutional layers on large but sparse datasets (e.g., Clothing), which means that we are not constrained by the high resource consumption of GCN, and have stable training time on all datasets.

5.7 Hyper-parameter Analysis (RQ5)

To evaluate the hyper-parameter sensitivity of HPMRec, we conduct comprehensive experiments on four datasets under varying hyperparameters settings: **Algebra Component Number** 2^{n+1}, **GCN**

[4] The results are under the best hyper-parameter settings on each dataset.

Table 4: Comparison of our HPMRec against state-of-the-art baselines on efficiency. (Time: s/Epoch; Memory: GB)

Dataset	**Baby**		**Sports**		**Clothing**	
Metrics	Time	Memory	Time	Memory	Time	Memory
DualGNN	5.63	2.05	11.59	2.81	14.19	3.02
MMGCN	4.09	2.69	14.93	3.91	17.48	4.24
LATTICE	3.20	4.53	11.07	19.93	16.53	28.22
FREEDOM	2.57	2.13	5.65	3.34	6.29	4.15
MMSSL	6.31	3.77	14.67	5.34	17.04	5.81
LGMRec	4.19	2.41	8.38	3.67	9.72	4.81
DiffMM	9.45	4.23	18.61	5.99	23.85	6.54
HPMRec	5.86	1.97	15.80	3.69	13.06	4.51

Layer Number L, **Regularization Balancing Hyper-parameter** λ, and **Self-supervised Learning Balancing Hyper-parameter** λ_s. The best result of each line is marked in Figure 4-7. According to these results, we have the following observations:

5.7.1 *Performance Comparison w.r.t 2^{n+1}.*

We analyze how different the component number 2^{n+1} influences the performance of the HPMRec. According to the result in Figure 4, we found that when component number 2^{n+1} equal to 4 ($n = 1$), HPMRec achieves optimal performance in terms of Recall@10 and NDCG@10 across Office and Baby datasets, and it achieves optimal performance in terms of Recall@10 and NDCG@10 when component number 2^{n+1} equal to 2 ($n = 0$) across Sports and Clothing datasets. When the component number is larger than 4 ($n > 1$), the performance does not improve, but rather has a negative influence. We attribute this situation to: four components are sufficient for the multimodal information encoder, a larger component number means higher diversity, which might introduce noise, resulting in suboptimal performance.

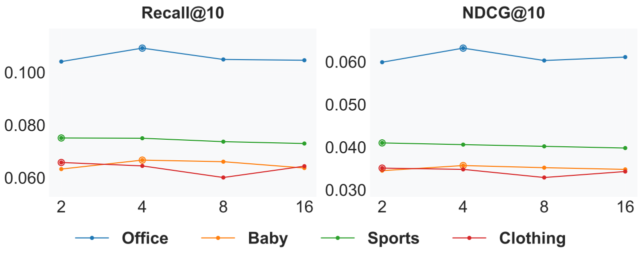

Figure 4: Effect of Algebra Component Number 2^{n+1}.

5.7.2 *Performance Comparison w.r.t L.*

As results show in Figure 5, we observe that the optimal value of layer number L is different across datasets: In terms of Recall@10 and NDCG@10, the framework achieves the best performance at 3 on the Baby, Sports, and Clothing datasets, and 2 on the Office dataset. Compared to other datasets, the Office dataset has a lower sparsity of user-item interaction, which means shallower GCNs are enough to extract the latent relationship, and the shallower message passing can avoid noise amplification.

5.7.3 *Performance Comparison w.r.t λ and λ_s.*

We analyze the effect of the regularization balancing hyper-parameter λ (shown in Figure 6). In terms of Recall@10 and NDCG@10, HPMRec achieves the best performance at $1e^{-2}$ on Baby, Sports, and Clothing datasets, and for Office, $1e^{-3}$ is best. As for the results of self-supervised learning regularizer λ_s shown in Figure 7, we find the same optimal setting ($1e^{-2}$) for all other three datasets except for the Clothing dataset. For the Clothing dataset, $1e^{-4}$ is best.

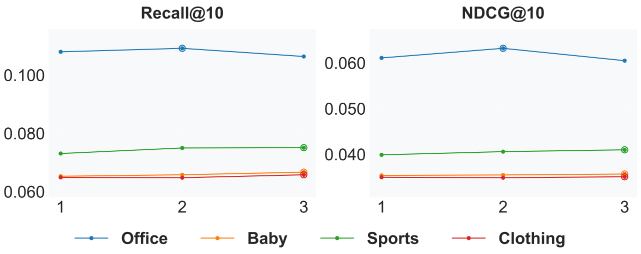

Figure 5: Effect of GCN Layer Number L.

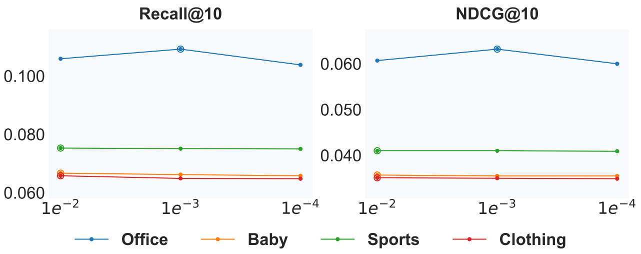

Figure 6: Effect of Balancing Hyper-parameter λ.

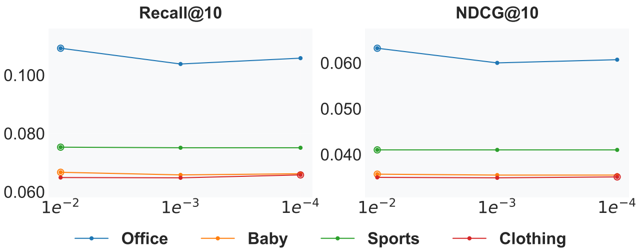

Figure 7: Effect of Balancing Hyper-parameter λ_s.

In summary, being flexible in choosing the hyper-parameter settings will allow us to adapt our model to multiple datasets. Although the optimal setting of these hyper-parameters varies, the performance differences are minimal, demonstrating the robustness and stability of HPMRec on different datasets.

6 Discussion

Based on the model implementation description in Section 4 and the results analysis of comprehensive ablation experiments in Section 5.5, we have the following discussion and interpretation of the effectiveness of each module of HPMRec and our design principles.

6.1 Model Joint Optimization

As shown in Section 4, our HPMRec adopts multiple modules to jointly optimize. The result analysis in Section 5.5 shows that each components have a positive effect on HPMRec. We will further discuss the crucial synergy and mutual constraints between modules that influence the model optimization. The prompt-aware compensation mechanism keeps the core modality-specific feature. Meanwhile, real-imag discrepancy expansion task enhances the ability of user/item representation to mine more modality-specific features, which can enhance the diversity of representation. These two modules' mutual constraints ensure that the representation will not lose the core modality-specific features in the pursuit of diversity, and deviate from the reasonable semantic space. However, the high diversity of each modality's representation will increase the gap between different modalities. To align different modalities and benefit the modality fusion, we design the cross-modality alignment task and MI enhancement fusion strategy. These two modules' synergy ensures that the gap between modalities does not affect modality fusion. In general, thanks to the prompt-aware compensation mechanism, the representation of each modality retains the core modality-specific features while mining more modality-specific features under the optimization of the real-imag discrepancy expansion task. With the joint optimization of all modules, HPMRec achieves state-of-the-art performance.

6.2 Maximize the Ability of Prompt

Through the analysis of variant **HPMRec** *w* **Explicit** in Section 5.5, we found that the learnable prompt can perform better in implicit guidance than explicit guidance. We attribute this phenomenon to the unsuitable optimization task and insufficient utilization of the powerful dynamic optimization capabilities of the learnable prompt. In our framework HPMRec, the main recommendation task and the self-supervised learning tasks implicitly optimize the prompt to facilitate the learning of core modality-specific features while avoiding the introduction of modality differences, and ensure a sufficiently flexible feature space (solution space) [11] to enhance user/item representations. Therefore, when there are no suitable explicit optimization task, utilizing implicit optimization tasks to ensure a sufficiently solution space of the learnable prompt can maximize its ability.

7 Conclusion

In this paper, we propose HPMRec, a hypercomplex, prompt-aware multimodal recommendation framework that enriches feature diversity and bridges semantic gaps across modalities. Specifically, HPMRec encodes each modality into a multi-component hypercomplex embedding, leveraging the multi-component representation ability of hypercomplex algebra to capture diverse modality-specific features. Secondly, HPMRec leverages the hypercomplex multiplication as naturally nonlinear fusion between modality pairs, thereby exploring more latent cross-modality features. Moreover, to mitigate component misalignment and keep core modality-specific features, we introduce a prompt-aware compensation mechanism that dynamically compensates each component, and this module also mitigates the over-smoothing problem. Finally, we design self-supervised learning tasks to further assist modality fusion and enhance the diversity of modality features. Extensive evaluations on four public datasets demonstrate that HPMRec outperforms state-of-the-art baselines in recommendation performance.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No: 62072390, and 92270123), and the Research Grants Council, Hong Kong SAR, China (Grant No: 15203120, 15226221, 15209922, and 15210023).

GenAI Usage Disclosure

No GenAI tools were used in any stage of the research, nor in the writing.

References

[1] Daniel Alfsmann. 2006. On families of 2 N-dimensional hypercomplex algebras suitable for digital signal processing. In *2006 14th European Signal Processing Conference*. IEEE, 1–4.

[2] John Baez. 2002. The octonions. *Bulletin of the american mathematical society* 39, 2 (2002), 145–205.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[4] Feiyu Chen, Junjie Wang, Yinwei Wei, Hai-Tao Zheng, and Jie Shao. 2022. Breaking Isolation: Multimodal Graph Fusion for Multimedia Recommendation by Edge-wise Modulation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 385–394.

[5] Tong Chen, Hongzhi Yin, Xiangliang Zhang, Zi Huang, Yang Wang, and Meng Wang. 2021. Quaternion factorization machines: A lightweight solution to intricate feature interaction modeling. *IEEE Transactions on Neural Networks and Learning Systems* 34, 8 (2021), 4345–4358.

[6] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 765–774.

[7] Zheyu Chen, Jinfeng Xu, and Haibo Hu. 2025. Don't Lose Yourself: Boosting Multimodal Recommendation via Reducing Node-neighbor Discrepancy in Graph Convolutional Network. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[8] Zheyu Chen, Jinfeng Xu, Yutong Wei, and Ziyue Peng. 2025. Squeeze and Excitation: A Weighted Graph Contrastive Learning for Collaborative Filtering. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2769–2773.

[9] Craig Culbert. 2007. Cayley-Dickson algebras and loops. *Journal of Forensic Biomechanics* 1, 1 (2007), 1–17.

[10] Leonard E Dickson. 1919. On quaternions and their generalization and the history of the eight square theorem. *Annals of Mathematics* 20, 3 (1919), 155–171.

[11] Taoran Fang, Yunchao Zhang, Yang Yang, Chunping Wang, and Lei Chen. 2023. Universal prompt tuning for graph neural networks. *Advances in Neural Information Processing Systems* 36 (2023), 52464–52489.

[12] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 249–256.

[13] Zhiqiang Guo, Jianjun Li, Guohui Li, Chaoyang Wang, Si Shi, and Bin Ruan. 2024. LGMRec: Local and Global Graph Learning for Multimodal Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8454–8462.

[14] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[15] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 639–648.

[16] Yangqin Jiang, Lianghao Xia, Wei Wei, Da Luo, Kangyi Lin, and Chao Huang. 2024. DiffMM: Multi-Modal Diffusion Model for Recommendation. *Proceedings of the 32ed ACM International Conference on Multimedia* (2024).

[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] Srdan Lazendic, Aleksandra Pizurica, and Hendrik De Bie. 2018. Hypercomplex algebras for dictionary learning. In *Early Proceedings of the AGACSE 2018 Conference*. 57–64.

[19] Anchen Li, Bo Yang, Huan Huo, and Farookh Hussain. 2022. Hypercomplex graph collaborative filtering. In *Proceedings of the ACM Web Conference 2022*. 1914–1922.

[20] Zhaopeng Li, Qianqian Xu, Yangbangyan Jiang, Xiaochun Cao, and Qingming Huang. 2020. Quaternion-based knowledge graph network for recommendation. In *Proceedings of the 28th ACM international conference on multimedia*. 880–888.

[21] Fan Liu, Zhiyong Cheng, Lei Zhu, Zan Gao, and Liqiang Nie. 2021. Interest-aware message-passing GCN for recommendation. In *Proceedings of the web conference 2021*. 1296–1305.

[22] Peng Liu, Lemei Zhang, and Jon Atle Gulla. 2023. Pre-train, prompt, and recommendation: A comprehensive survey of language modeling paradigm adaptations in recommender systems. *Transactions of the Association for Computational Linguistics* 11 (2023), 1553–1571.

[23] Zemin Liu, Xingtong Yu, Yuan Fang, and Xinming Zhang. 2023. Graphprompt: Unifying pre-training and downstream tasks for graph neural networks. In *Proceedings of the ACM web conference 2023*. 417–428.

[24] Kelong Mao, Jieming Zhu, Xi Xiao, Biao Lu, Zhaowei Wang, and Xiuqiang He. 2021. UltraGCN: ultra simplification of graph convolutional networks for recommendation. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 1253–1262.

[25] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 43–52.

[26] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. 452–461.

[27] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[28] Fernando E Rosas, Pedro AM Mediano, Michael Gastpar, and Henrik J Jensen. 2019. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Physical Review E* 100, 3 (2019), 032305.

[29] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2022. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia* (2022).

[30] Yi Tay, Aston Zhang, Luu Anh Tuan, Jinfeng Rao, Shuai Zhang, Shuohang Wang, Jie Fu, and Siu Cheung Hui. 2019. Lightweight and efficient neural natural language processing with quaternion networks. *arXiv preprint arXiv:1906.04393* (2019).

[31] Thanh Tran, Di You, and Kyumin Lee. 2020. Quaternion-based self-attentive long short-term user preference encoding for recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1455–1464.

[32] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. 2021. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia* (2021).

[33] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*. 165–174.

[34] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-Modal Self-Supervised Learning for Recommendation. In *Proceedings of the ACM Web Conference 2023*. 790–800.

[35] Wei Wei, Jiabin Tang, Lianghao Xia, Yangqin Jiang, and Chao Huang. 2024. Promptmm: Multi-modal knowledge distillation for recommendation with prompt-tuning. In *Proceedings of the ACM Web Conference 2024*. 3217–3228.

[36] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*. 1437–1445.

[37] Yiqing Wu, Ruobing Xie, Yongchun Zhu, Fuzhen Zhuang, Ao Xiang, Xu Zhang, Leyu Lin, and Qing He. 2022. Selective fairness in recommendation via prompts. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 2657–2662.

[38] Xin Xin, Tiago Pimentel, Alexandros Karatzoglou, Pengjie Ren, Konstantina Christakopoulou, and Zhaochun Ren. 2022. Rethinking reinforcement learning for recommendation: A prompt perspective. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 1347–1357.

[39] Jinfeng Xu, Zheyu Chen, Jinze Li, Shuo Yang, Hewei Wang, Xiping Hu, and Edith C-H Ngai. 2024. FourierKAN-GCF: Fourier Kolmogorov-Arnold Network – An Effective and Efficient Feature Transformation for Graph Collaborative Filtering. *arXiv preprint arXiv:2406.01034* (2024).

[40] Jinfeng Xu, Zheyu Chen, Jinze Li, Shuo Yang, Hewei Wang, Yijie Li, Mengran Li, Puzhen Wu, and Edith CH Ngai. 2025. Mdvt: Enhancing multimodal recommendation with model-agnostic multimodal-driven virtual triplets. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 3378–3389.

[41] Jinfeng Xu, Zheyu Chen, Jinze Li, Shuo Yang, Hewei Wang, and Edith C-H Ngai. 2024. AlignGroup: Learning and Aligning Group Consensus with Member Preferences for Group Recommendation. *arXiv preprint arXiv:2409.02580* (2024).

[42] Jinfeng Xu, Zheyu Chen, Wei Wang, Xiping Hu, Sang-Wook Kim, and Edith CH Ngai. 2025. COHESION: Composite Graph Convolutional Network with Dual-Stage Fusion for Multimodal Recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1830–1839.

[43] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, and Edith CH Ngai. 2025. The Best is Yet to Come: Graph Convolution in the Testing Phase for Multimodal Recommendation. *arXiv preprint arXiv:2507.18489* (2025).

[44] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Hewei Wang, and Edith CH Ngai. 2025. Mentor: multi-level self-supervised learning for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 12908–12917.

[45] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Hewei Wang, Wei Wang, Xiping Hu, and Edith Ngai. 2025. NLGCL: Naturally Existing Neighbor Layers Graph Contrastive Learning for Recommendation. *arXiv preprint arXiv:2507.07522* (2025).

[46] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Wei Wang, Xiping Hu, Steven Hoi, and Edith Ngai. 2025. A Survey on Multimodal Recommender Systems: Recent Advances and Future Directions. *arXiv preprint arXiv:2502.15711* (2025).

[47] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Lizhen Cui, and Quoc Viet Hung Nguyen. 2022. Are graph augmentations necessary? simple graph contrastive learning for recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 1294–1303.

[48] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3872–3880.

[49] Shuai Zhang, Lina Yao, Lucas Vinh Tran, Aston Zhang, and Yi Tay. 2019. Quaternion collaborative filtering for recommendation. *arXiv preprint arXiv:1906.02594* (2019).

[50] Wen Zhang, Yushan Zhu, Mingyang Chen, Yuxia Geng, Yufeng Huang, Yajing Xu, Wenting Song, and Huajun Chen. 2023. Structure pretraining and prompt tuning for knowledge graph transfer. In *Proceedings of the ACM web conference 2023*. 2581–2590.

[51] Xin Zhou. 2023. MMRec: Simplifying Multimodal Recommendation. *arXiv preprint arXiv:2302.03497* (2023).

[52] Xin Zhou, Donghui Lin, Yong Liu, and Chunyan Miao. 2023. Layer-refined graph convolutional networks for recommendation. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE, 1247–1259.

[53] Xin Zhou and Zhiqi Shen. 2023. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 935–943.

[54] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*. 845–854.

[55] Xuanyu Zhu, Yi Xu, Hongteng Xu, and Changjian Chen. 2018. Quaternion convolutional neural networks. In *Proceedings of the European conference on computer vision (ECCV)*. 631–647.