

## RaMen

### Multi-Strategy Multi-Modal Learning for Bundle Construction

Nguyen, Huy Son; Nguyen, Quang Huy; Pham, Duc Hoang; Le, Duc Trong; Le, Hoang Quynh; Sitkrongwong, Padipat; Takasu, Atsuhiko; Mansoury, Masoud

#### DOI

[10.3233/FAIA251022](https://doi.org/10.3233/FAIA251022)

#### Publication date

2025

#### Document Version

Final published version

#### Published in

ECAI 2025 - 28th European Conference on Artificial Intelligence, including 14th Conference on Prestigious Applications of Intelligent Systems, PAIS 2025 - Proceedings

#### Citation (APA)

Nguyen, H. S., Nguyen, Q. H., Pham, D. H., Le, D. T., Le, H. Q., Sitkrongwong, P., Takasu, A., & Mansoury, M. (2025). RaMen: Multi-Strategy Multi-Modal Learning for Bundle Construction. In I. Lynce, N. Murano, M. Vallati, S. Villata, F. Chesani, M. Milano, A. Omicini, & M. Dastani (Eds.), *ECAI 2025 - 28th European Conference on Artificial Intelligence, including 14th Conference on Prestigious Applications of Intelligent Systems, PAIS 2025 - Proceedings* (pp. 1889-1896). (Frontiers in Artificial Intelligence and Applications; Vol. 413). IOS Press. <https://doi.org/10.3233/FAIA251022>

#### Important note

To cite this publication, please use the final published version (if applicable).  
Please check the document version above.

#### Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

#### Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.  
We will remove access to the work immediately and investigate your claim.

# RaMen: Multi-Strategy Multi-Modal Learning for Bundle Construction

Huy-Son Nguyen<sup>a,1</sup>, Quang-Huy Nguyen<sup>b,1</sup>, Duc-Hoang Pham<sup>b,1</sup>, Duc-Trong Le<sup>b</sup>, Hoang-Quynh Le<sup>b,\*</sup>,  
Padipat Sitkrongwong<sup>c</sup>, Atsuhiko Takasu<sup>c</sup> and Masoud Mansoury<sup>a</sup>

<sup>a</sup>Delft University of Technology, The Netherlands

<sup>b</sup>VNU University of Engineering and Technology, Hanoi, Vietnam

<sup>c</sup>National Institute of Informatics, Japan

**Abstract.** Existing studies on bundle construction have relied merely on user feedback via bipartite graphs or enhanced item representations using semantic information. These approaches fail to capture elaborate relations hidden in real-world bundle structures, resulting in suboptimal bundle representations. To overcome this limitation, we propose RaMen, a novel method that provides a holistic multi-strategy approach for bundle construction. RaMen utilizes both intrinsic (*characteristics*) and extrinsic (*collaborative signals*) information to model bundle structures through Explicit Strategy-aware Learning (ESL) and Implicit Strategy-aware Learning (ISL). ESL employs task-specific attention mechanisms to encode multi-modal data and direct collaborative relations between items, thereby explicitly capturing essential bundle features. Moreover, ISL computes hyperedge dependencies and hypergraph message passing to uncover shared latent intents among groups of items. Integrating diverse strategies enables RaMen to learn more comprehensive and robust bundle representations. Meanwhile, Multi-strategy Alignment & Discrimination module is employed to facilitate knowledge transfer between learning strategies and ensure discrimination between items/bundles. Extensive experiments demonstrate the effectiveness of RaMen over state-of-the-art models on various domains, justifying valuable insights into complex item set problems.

## 1 Introduction

Bundle construction, which focuses on grouping relevant items into appealing offers, is an increasingly valuable strategy in marketing for both physical stores and e-commerce platforms across various industries [43, 34, 28]. Beyond driving revenue growth, well-crafted bundles can enhance the customer experience by introducing variety and mitigating decision fatigue. Traditional bundle design has required manual effort from retailers, being not only time-consuming but also costly, making it difficult to scale across large datasets [34]. The emergence of automatic bundle manufacturers has garnered attention from researchers due to its scalability and efficiency [28, 34].

Most approaches oversimplify the complex strategies behind decision-making processes by relying on noisy datasets where bundles are defined using unreasonable heuristics [12, 34, 43]. For example, some studies equate collaborative items with bundles without considering the context-specific nature of such combinations [23, 8].

Others rely on user-generated lists in niche domains like *music* [17] or *gaming* [32], limiting their general applicability. These studies often overlook the underlying rationale for bundling decisions, assuming that historical bundles are readily available for recommendation purposes, which is unrealistic in many real marketing scenarios. State-of-the-art (SOTA) studies [27, 28, 11] on bundle-related tasks predominantly rely on user feedback, represented through bipartite graphs with LightGCN [16], or attempt to enhance item representations using semantic data [25]. Yet, such approaches often lead to suboptimal bundle representations, making it difficult to accurately capture the underlying structure of real-world bundling strategies.

In practice, successful bundle creation hinges on leveraging the inherent relationships between products, enabling businesses to develop combinations that meet specific customer needs [34]. To build potential bundles, it is vital to consider both intrinsic (*characteristic*) and extrinsic (*collaborative*) information of products, ensuring alignment with particular customer intents or preferences. Moreover, bundles are often tailored to target distinct customer groups or particular intentions, such as those curated by style, age, or price segmentation, etc. [34]. Thus, modeling shared latent attributes among items plays a significant role in determining optimal bundling tactics.

**Approaches and Contributions.** To address these limitations, we introduce RaMen, a novel framework that systematically models the bundle construction process through a multi-strategy multi-modal learning paradigm. RaMen leverages both intrinsic (*semantic*) and extrinsic (*collaborative*) information to effectively capture the latent structure of bundles. This is achieved by incorporating two key components: Explicit Strategy-aware Learning and Implicit Strategy-aware Learning. The Explicit Strategy focuses on encoding essential bundle characteristics by utilizing task-specific attention mechanisms, which highlight direct item relationships and relevant semantic information as *our first contribution*. Meanwhile, Implicit Strategy-aware Learning employs hypergraph message passing and hyperedge dependency matrices to uncover shared latent intents among item groups, capturing deeper implicit interactions that traditional models overlook as *our second contribution*. By integrating multi-strategy representations, RaMen constructs more comprehensive and generalizable bundle representations. Furthermore, Multi-strategy Alignment & Dispersion is designed to enhance knowledge transfer between learning strategies while maintaining discrimination between different object representations. As *our final contribution*, extensive experimental evaluations substantiate the efficacy of

\* Corresponding author: lhquynh@vnu.edu.vn

<sup>1</sup> Shared first-authors.

RaMen, revealing its ability to deliver novel insights and robust solutions to bundling problems. To the best of our knowledge, this study is the first to model the collaborative relationships and characteristics of items, combined with learning shared attributes between them, to identify the hidden intents of each constructed bundle.

## 2 Related Work

With the rapid growth of e-Commerce, studies related to complex sets of items such as next-basket recommendation [21, 29], bundle recommendation [33], and bundle construction [28] have garnered significant attention as a means of improving business revenue and reducing monotonous recommendations based on cross-selling concepts. Although research on bundle recommendation focuses on suggesting predefined bundles based on user interactions [26, 37, 2, 3], our objective in bundle construction is to predict collections of items to create bundles that fulfill specific needs to attract users [28, 27]. Traditional bundle recommendation approaches are based on predefined criteria [43] and matrix factorization [32] to capture and leverage user preferences. Recent advancements in deep learning have significantly improved the performance of bundle recommendation systems, including attention-based techniques [6], graph neural networks [4, 40], contrastive learning [26, 40], and generative methods [2, 3] but are still based purely on tripartite relations between user-bundle, bundle-item, and user-item pairs. Meanwhile, appropriately constructed bundles can enable the system to deliver more effective and targeted recommendations [34, 33].

Bundle construction tasks concentrate on completing partial bundles by identifying and selecting missing items from a pool of candidate products [28, 27, 34]. This process enables systems to automatically construct comprehensive and diverse bundles that better cater to a broad range of consumer preferences, ultimately enhancing product recommendations and improving user satisfaction. Common approaches in bundle construction leverage user-item interactions to uncover item-to-item relationships, thereby learning hidden bundle patterns to model the ultimate bundle representations [4, 8, 10]. Bundle representation learning consistently lies at the core of bundle-oriented challenges. Sequential models, such as Bi-LSTM [15], were employed to capture relations between consecutive items. However, as bundles are inherently unordered, conventional sequential models struggle to fully capture pairwise correlations. To tackle this issue, attention mechanisms [6], Transformers [37], and graph neural networks (GNNs) [30] have been utilized to model both pairwise and higher-order item relationships. Despite these advances in item correlation modeling, limited attention has been paid to multimodal information, leading to construct bundles that lack coherence in their item characteristics as well as meaningless intents [34].

The integration of multi-modal data proves effective in addressing key challenges such as data sparsity and cold-start issues [24, 28]. In the context of product bundling, several methods have leveraged multi-modalities to improve item representation learning. Recent SOTA model CLHE [28] leverages self-attention mechanisms [36] to fuse multi-modal features with user feedback, focusing on addressing data sparsity and cold-start issues. However, CLHE [28] improves the learning process by solely incorporating multi-modal features and a bipartite item-user graph with LightGCN [16] that can not model rigorous relationships between anchor items and accessories to determine primary intents of bundles. Another approach employs a multi-modal encoder along with cross-modal and cross-item contrastive loss to better capture item-to-item relationships [27]. CIRP thrives on employing cross-item relation to provide the pre-training

model of item representations [27]. Furthermore, some promising approaches integrate large language models (LLMs) into the bundle construction process, enhancing the model's understanding of relationships between different modalities semantically [25, 35]. The available results are remarkable, but authoritative studies have not yet been able to effectively address the modeling of corporate strategies based on both collaborative relationships and item characteristics.

SOTA models on related tasks mostly rely on bipartite graphs with LightGCN via user feedback [26, 28, 42, 41, 11], or attempt to enhance item representations merely using semantic data [25]. They often lead to suboptimal bundle representations, making it difficult to capture accurately the underlying structure of real-world bundling strategies. Different from previous works [28, 33, 34], our multi-strategy multi-modal learning paradigm aims to thoroughly model the collaborative relationships and characteristics of items, combined with learning shared attributes among items/bundles, to grasp the more comprehensive intents of each constructed bundle. Compared to the closest method, CLHE [28], we inherit the design of their evaluation protocol and input-output flows because CLHE is the pioneer and SOTA research on multimodal bundle construction. As mentioned above, CLHE solely incorporates multi-modal features and an item-user graph with LightGCN that can not model rigorous relationships between anchor items and accessories to determine the primary intents of bundles like RaMen. Meanwhile, ESL of RaMen not only encodes essential bundle characteristics by attention mechanisms, but also models association among items via our item-item graph. Notably, our refined attention-based propagation on item-item graph can learn more comprehensive associations among items, representing a significant improvement over the ubiquitous user-item LightGCN. Moreover, to tackle the issues mentioned about multi-strategy bundle construction, our ISL is designed to uncover shared latent intents among item groups. Besides the integration of two prime encoders, we devise MAD module to enhance knowledge transfer between learning strategies while maintaining discrimination between different items/bundles.

## 3 Methodology

Section 3 presents the overall architecture of RaMen as Fig. 1 for bundle construction tasks, consisting of four main modules: (i) Explicit Strategy-aware Learning, (ii) Implicit Strategy-aware Learning, (iii) Multi-strategy Alignment & Discrimination, and (iv) Retrieval & Joint Optimization.

### 3.1 Preliminaries

#### 3.1.1 Problem Formulation.

For the tasks of bundle construction, let  $\mathcal{I} = \{i_k\}_{k=1}^{|\mathcal{I}|}$ ,  $\mathcal{U} = \{u_k\}_{k=1}^{|\mathcal{U}|}$ ,  $\mathcal{B} = \{b_k\}_{k=1}^{|\mathcal{B}|}$  represents a set of items, users, and bundles, respectively. Relied on historical user behaviors, the user-item interaction is collected formally as a binary matrix  $X \in \{0, 1\}^{|\mathcal{U}| \times |\mathcal{I}|}$ , where  $X_{u,i} = 1$  indicates user  $u$  interacted with item  $i$ , and  $X_{u,i} = 0$  otherwise. Likewise, the bundle-item affiliation is defined in matrix  $Y \in \{0, 1\}^{|\mathcal{B}| \times |\mathcal{I}|}$ , where  $Y_{b,i} = 1$  if bundle  $b$  contains item  $i$ , and  $Y_{b,i} = 0$  otherwise. In particular, each bundle  $b = \{i_h\}_{h=1}^{|\mathcal{B}|} \in \mathcal{B}$  is a collection of pertinent items. We establish the training set of bundles as  $\tilde{\mathcal{B}} = \{b_k\}_{k=1}^{|\tilde{\mathcal{B}}|} \subset \mathcal{B}$ , and testing set of bundles as  $\hat{\mathcal{B}} = \{b_k\}_{k=|\tilde{\mathcal{B}}|+1}^{|\mathcal{B}|} \subset \mathcal{B}$ . Given partial bundle  $\hat{b}$  containing a few seed items from each unseen testing bundle  $b \in \hat{\mathcal{B}}$ , the objective of

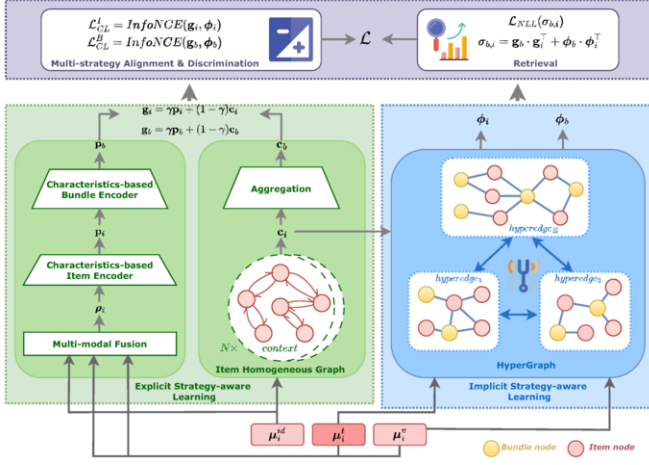


Figure 1. Overall architecture of our proposed model RaMen.

bundle construction is to efficiently predict the deficient items  $\{b \setminus \hat{b}\}$  to capture the comprehensive bundle. The training process complies with an auto-encoder approach [28], where the entire items within the bundle are considered as the input, and the same set should be predicted as the output.

### 3.1.2 Semantic Information Extraction.

Inherited from [24, 28], the textual and visual features of each item are derived from large-scale multi-modal feature extractors, represented as  $\{\mathbf{m}_i^t \in \mathbb{R}^{d_{mt}}, \mathbf{m}_i^v \in \mathbb{R}^{d_{mv}}\}$ , where  $d_{mt}$  and  $d_{mv}$  denote the dimensions of the textual and visual embeddings. Textual information, such as the item's title and description, is encoded into  $\mathbf{m}_i^t$ , while images are encoded into  $\mathbf{m}_i^v$ . The encoded visual and textual embeddings are respectively transformed into a unified latent space via specialized refinement MLP networks such as  $\text{MLP}^v, \text{MLP}^t$  to mitigate the misalignment caused by dimensional differences across modalities [22, 24]. This process is derived as follows:

$$\mu_i^v = \text{MLP}^v(\mathbf{m}_i^v), \quad \mu_i^t = \text{MLP}^t(\mathbf{m}_i^t), \quad (1)$$

where  $\mu_i^v, \mu_i^t \in \mathbb{R}^d$  are aligned embeddings after dimension adjustment, ensuring consistency across modalities.

### 3.1.3 Item-level Collaborative Relation Construction.

Learning collaborative signals based on user-item graphs [24, 28] may introduce noise when propagating higher-order collaborative signals. To tackle this limitation, an item homogeneous graph  $\mathcal{G}_{\mathcal{I}} = \{\mathcal{I}, \mathcal{E}_{\mathcal{I}}\}$  is designed to learn direct influences among items, where  $\mathcal{I}$  and  $\mathcal{E}_{\mathcal{I}} = \{e_{i,j} | i, j \in \mathcal{I}\}$  represent the set of vertices and edges. An item co-purchased matrix  $E \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{I}|}$  is first computed, where  $E = X^T \cdot X$ . The direct relation  $e_{i,j}$  between each pair of items  $i, j$  is established by discretizing the weighted matrix  $E$  with threshold  $\epsilon$  into an unweighted version to facilitate information propagation, derived as  $e_{i,j} = 1$  if  $E_{i,j} \geq \epsilon \wedge i \neq j$ , and 0 otherwise<sup>2</sup>.

## 3.2 Explicit Strategy-aware Learning

### 3.2.1 Characteristic Strategy Encoder.

Given the assumption that bundle construction is driven by enterprises leveraging distinctive characteristics [34, 43], this module

<sup>2</sup> This work can be developed more robustly by adaptively filtering noisy edges in the graph instead of empirical selection across diverse domains.

aims to synthesize relevant multi-modal features across items within each bundle, optimizing the bundling strategy from a characteristic-based perspective.

**Multi-modal Fusion.** The obtained semantic embeddings  $\mu_i^v, \mu_i^t$ , along with the initialized ID embedding  $\mu_i^{id}$  in the same latent space  $\mathbb{R}^d$  are synthesized into multi-modal item representation  $\rho_i \in \mathbb{R}^{2 \times d}$  through concatenation-based fusion, mathematically represented as:

$$\rho_i = \Xi(\Psi_\rho(\mu_i^v \parallel \mu_i^t), \mu_i^{id}), \quad (2)$$

where  $\Psi_\rho \in \mathbb{R}^{d \times 2d}$  is the linear transformation matrix, and  $\parallel$  denotes the vertical concatenation of the semantic features. Besides,  $\Xi$  performs horizontal concatenation, synthesizing these components into the multi-modal feature matrix of items.

**Characteristics-based Item Encoder.** Leveraging the proven effectiveness of attention mechanisms [36] and the diverse aspects of item features in recommender systems [9, 24], RaMen employs self-attention techniques to compute correlation scores between item-level multi-modal characteristics, formalized as:

$$\tilde{\rho}_i^{(l)} = \text{softmax} \left( \underbrace{\frac{1}{\sqrt{d}} \tilde{\rho}_i^{(l-1)} \Psi_I^K \left( \tilde{\rho}_i^{(l-1)} \Psi_I^Q \right)^T}_{\text{correlation score of item characteristics}} \right) \tilde{\rho}_i^{(l-1)}, \quad (3)$$

where  $\Psi_I^K$  and  $\Psi_I^Q \in \mathbb{R}^{d \times d}$  represent the key and query projection matrices. The feature matrix at layer  $l$ -th, denoted as  $\tilde{\rho}_i^{(l)} \in \mathbb{R}^{2 \times d}$ , evolves from the initial features  $\tilde{\rho}_i^{(0)} = \rho_i$ . The feature matrix  $\tilde{\rho}_i^{(L_1)} \in \mathbb{R}^{2 \times d}$  of item  $i$  is produced after  $L_1$  attention layers, and the corresponding characteristic vector  $\mathbf{p}_i \in \mathbb{R}^d$  is subsequently computed by mean pooling over the feature matrix  $\tilde{\rho}_i^{(L_1)}$ .

**Characteristics-based Bundle Encoder.** RaMen is capable of capturing the critical semantic features of items, enhancing the bundle construction process by focusing on the intricate correlations between multi-modal characteristics. With the obtained item embeddings, the bundle characteristics is formed by concatenating the embeddings of components, expressed as  $\rho_b = \Xi(\{\mathbf{p}_i\}_{i \in b})$ . The bundle representation is refined through  $L_2$  attention layers, defined as:

$$\tilde{\rho}_b^{(l)} = \text{softmax} \left( \underbrace{\frac{1}{\sqrt{d}} \tilde{\rho}_b^{(l-1)} \Psi_B^K \left( \tilde{\rho}_b^{(l-1)} \Psi_B^Q \right)^T}_{\text{correlation score of bundle characteristics}} \right) \tilde{\rho}_b^{(l-1)}, \quad (4)$$

where  $\tilde{\rho}_b^{(l)}$  denotes the feature matrix of bundle  $b$  in layer  $l$ -th with the initial value  $\tilde{\rho}_b^{(0)} = \rho_b$ ;  $\Psi_B^K, \Psi_B^Q \in \mathbb{R}^{d \times d}$  represent the learnable projection matrices. After refining through  $L_2$  attention layers, we adopt mean pooling to each bundle feature matrix  $\tilde{\rho}_b^{(L_2)}$ , aggregating the ultimate bundle characteristic  $\mathbf{p}_b \in \mathbb{R}^d$  correspondingly.

### 3.2.2 Collaborative Strategy Encoder.

Given the assumption that an effective bundle construction strategy should leverage item-level collaborative relations to align logically with user expectations, Collaborative Strategy Encoder employs advanced attention mechanisms [1, 30] to effectively propagate high-order collaborative signals with weighted causal influences among nodes of graph  $\mathcal{G}_{\mathcal{I}}$  underlying various contexts. The propagation of item neighborhood features is derived as:

$$\alpha_{i \leftarrow j}^{(n)} = \frac{\exp \left( \mathbf{q}_{(n)}^T \varphi(\Psi^{(n)} \mathbf{s}_i + \hat{\Psi}^{(n)} \mathbf{s}_j + \Delta) \right)}{\sum_{j' \in \mathcal{N}_i} \exp \left( \mathbf{q}_{(n)}^T \varphi(\Psi^{(n)} \mathbf{s}_i + \hat{\Psi}^{(n)} \mathbf{s}_{j'} + \Delta) \right)} \quad (5)$$

where  $\mathbf{q}_{(n)} \in \mathbb{R}^d$  is a learnable context vector;  $\mathbf{s}_i, \mathbf{s}_j \in \mathbb{R}^d$  represent the embeddings of item  $i$  and  $j$ ;  $\Delta$  denotes the bias weights; and  $\mathcal{N}_i$  denotes neighborhood set of item  $i$ . The symbol  $\varphi$  presents the activation function LeakyReLU. Compared to conventional attention techniques in GNN [13], we employ specialized transformation matrices  $\Psi^{(n)}, \hat{\Psi}^{(n)} \in \mathbb{R}^{d \times d}$  for the target-item node  $i$  and source-item node  $j$  at the  $n$ -th context to mitigate overfitting. The ID embeddings of items are adopted as input to Collaborative Strategy Encoder.

The final representation  $\mathbf{c}_i \in \mathbb{R}^d$  of item  $i$  is aggregated after  $N$  contexts, derived as follows:

$$\begin{aligned} \mathbf{s}_i^{(n)} &= \sum_{j \in \mathcal{N}_i} \alpha_{i \leftarrow j}^{(n)} \Psi_j^{(n)} \mathbf{s}_j, \\ \mathbf{c}_i &= \beta \frac{1}{N} \sum_{n=1}^N \mathbf{s}_i^{(n)} + (1 - \beta) \mathbf{s}_i \end{aligned} \quad (6)$$

where  $\mathbf{s}_i^{(n)} \in \mathbb{R}^d$  signifies the latent representation of item  $i$  at the  $n$ -th layer,  $\beta$  modulates the impact of the residual connection on the enhanced item embedding. Thereby, the Collaborative Strategy Encoder obtains the bundle embedding  $\mathbf{c}_b \in \mathbb{R}^d$  through the mean aggregation of the embeddings of items within bundle  $b$ .

The obtained representations of items/bundles from Characteristic Strategy Encoder and Collaborative Strategy Encoder are aggregated to compute Explicit Strategy-aware embeddings  $\mathbf{g}_b$  and  $\mathbf{g}_i$  for bundle  $b$  and item  $i$ , derived as follows:

$$\begin{aligned} \mathbf{g}_i &= \gamma \mathbf{p}_i + (1 - \gamma) \mathbf{c}_i, \\ \mathbf{g}_b &= \gamma \mathbf{p}_b + (1 - \gamma) \mathbf{c}_b, \end{aligned} \quad (7)$$

where  $\gamma$  controls the effect of embeddings from different encoders.

### 3.3 Implicit Strategy-aware Learning

The hypergraph architecture [14, 13, 38], which extends beyond pairwise relations, enables the latent representation of both intra-bundle and inter-bundle relations by modeling shared attributes among items as hyperedges. To effectively capture implicit strategies within groups of items, we introduce learnable hyperedge embeddings  $W_m \in \mathbb{R}^{H \times d}$ , designed to encode latent attributes specific to each modality  $m \in \{t, v\}$ , where  $H$  represents the number of hyperedges and  $t, v$  respectively denote textual/visual features. The dependency matrices for hyperedges and items/bundles are formally constructed as follows:

$$F_T^m = M_T^m (W_m)^\top, \quad F_B^m = Y (F_T^m)^\top, \quad (9)$$

where  $F_T^m \in \mathbb{R}^{|\mathcal{I}| \times H}$  and  $F_B^m \in \mathbb{R}^{|\mathcal{B}| \times H}$  are item-hyperedge and bundle-hyperedge dependency matrices, respectively;  $M_T^m = \{\mu_i^m\}_{i \in \mathcal{I}}$  is the feature matrix of modality  $m$ . The matrix  $F_T^m$  aims to capture the connections between items and hyperedges, grouping similar items under shared attributes. Besides, the matrix  $F_B^m$  reflects how bundles are indirectly associated with hyperedges via the items they contain. The stronger the affiliation between a bundle and items linked to latent attributes, the more likely the bundle's strategy is aligned with that attribute. Inspired by [38, 14], Gumbel-Softmax reparameterization technique [18] is adopted to mitigate the impact of noisy connections between items/bundles and hyperedges, defined as follows:

$$\hat{\mathbf{f}}_i^m = \text{softmax} \left( \frac{\log \boldsymbol{\theta} - \log(1 - \boldsymbol{\theta}) + \mathbf{f}_i^m}{\tau} \right), \quad (10)$$

where  $\hat{\mathbf{f}}_i^m \in \mathbb{R}^H$  represents the relation vector of item  $i$  with hyperedges in the fine-grained dependency matrix  $\hat{F}_T^m$ ; each value of the noise vector  $\boldsymbol{\theta}$  is sampled from a uniform distribution in range  $[0, 1]$ ; and temperature parameter  $\tau$  is empirically selected as 0.2. Likewise, we obtain the fine-grained bundle-hyperedge matrix  $\hat{F}_B^m$ . These fine-grained dependency matrices are then leveraged to propagate item and bundle attributes relied on each modality, derived as:

$$\begin{aligned} \phi_i^{m,(z+1)} &= \hat{F}_T^m \cdot (\hat{F}_T^m)^\top \cdot \phi_i^{m,(z)}, \\ \phi_b^{m,(z+1)} &= \hat{F}_B^m \cdot (\hat{F}_B^m)^\top \cdot \phi_b^{m,(z)}, \end{aligned} \quad (11)$$

where  $\phi_i^{m,(z)}$  is the embedding of item  $i$  corresponding to modality  $m$  at the  $z$ -th hypergraph layer, specifically  $\phi_i^{m,(0)} = \mathbf{c}_i$ . The final representations of Implicit Strategy-aware Learning are obtained after propagating  $Z$  hypergraph layers as follows:

$$\phi_i = \Psi \left( \sum_{m \in \{v, t\}} \phi_i^{m,(Z)} \right), \quad \phi_b = \Psi \left( \sum_{m \in \{v, t\}} \phi_b^{m,(Z)} \right) \quad (12)$$

where  $\Psi$  is  $L_p$  normalization function,  $\phi_i^{m,(Z)}$  and  $\phi_b^{m,(Z)} \in \mathbb{R}^d$  are embeddings of item  $i$  and bundle  $b$  corresponding to modality  $m$  obtained after  $Z$  hypergraph layers, respectively.

### 3.4 Multi-strategy Alignment & Discrimination

We apply the contrastive loss, specifically InfoNCE [31], to align the representations of the same item or bundle generated by different strategies and ensure the separation of embeddings corresponding to distinct items or bundles within the embedding space. This technique leads to more coherent and discriminative representation for each item or bundle. The item-level contrastive loss for multi-strategy learning is derived as:

$$\mathcal{L}_{CL}^I = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} -\log \frac{\exp(\cos(\mathbf{g}_i, \phi_i)/\tau)}{\sum_{j \in \mathcal{I}} \exp(\cos(\mathbf{g}_i, \phi_j)/\tau)}, \quad (13)$$

where  $\cos(\cdot)$  performs cosine similarity function. The bundle-level contrastive loss  $\mathcal{L}_{CL}^B$  is derived similarly. The objective loss of Multi-strategy Alignment & Discrimination module is to reconcile different strategy-based representations as:

$$\mathcal{L}_{MAD} = \mathcal{L}_{CL}^I + \mathcal{L}_{CL}^B, \quad (14)$$

### 3.5 Retrieval & Joint Optimization

#### 3.5.1 Retrieval.

To estimate the possibility that item  $i$  belongs to bundle  $b$ , we adopt the inner product to compute score  $\sigma_{b,i}$  from multi-strategy representations, as follows:

$$\sigma_{b,i} = \mathbf{g}_b \cdot \mathbf{g}_i^\top + \phi_b \cdot \phi_i^\top \quad (15)$$

Ground in the steering study of [28], the negative log-likelihood (NLL) is employed as the primary optimization objective after obtaining the score  $\sigma_{b,i}$ . By using NLL loss, the model learns to assign higher scores to items that are likely to belong to a bundle while minimizing scores for irrelevant items. Here, the NLL loss for optimizing

**Table 1.** The statistics of four benchmark datasets in diverse domains for bundle construction.

Dataset	#U	#I	#B	#B-I	#U-I	Avg.I/B	Avg.I/U	U-I Dens.
POG	17,449	48,676	20,000	72,224	237,519	3.61	13.61	0.0073%
Spotify	118,994	254,155	20,000	1,268,716	36,244,806	63.44	304.59	0.1198%
Electronic	888	3,499	1,750	6,165	6,165	3.52	6.94	0.1984%
Food	879	3,767	1,784	6,395	6,395	3.58	7.28	0.1931%

prediction is defined as:

$$\mathcal{L}_{NLL} = \frac{1}{|\tilde{\mathcal{B}}|} \sum_{b \in \tilde{\mathcal{B}}} \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} -\mathbb{1}_{i \in b} \log \left( \frac{\exp(\sigma_{b,i})}{\sum_{j \in \mathcal{I}} \exp(\sigma_{b,j})} \right), \quad (16)$$

where  $\mathbb{1}_{i \in b}$  represents an indicator function that equals 1 if the component item  $i$  belongs to the bundle  $b$ , and 0 otherwise.

### 3.5.2 Joint Optimization.

The overall objective function  $\mathcal{L}$  is composed of the defined loss functions combined with a regularization term, formulated as:

$$\mathcal{L} = \mathcal{L}_{NLL} + \lambda_1 \mathcal{L}_{MAD} + \lambda_2 \|\Theta\|_2^2, \quad (17)$$

where the hyperparameter  $\lambda_1$  controls the impact of contrastive-based loss, and  $\lambda_2$  denotes regularization weight with all the trainable parameters  $\Theta$  of model.

## 4 Experiments

We conduct extensive experiments to evaluate the effectiveness of RaMen, and analyze the significance of its main components. Moreover, some qualitative showcases accentuate the superior performance of RaMen compared to CLHE. Our repository is available on GitHub to facilitate reproducibility and extension.

### 4.1 Experimental Settings

#### 4.1.1 Datasets and Evaluation Protocols.

We utilize four datasets in diverse domains [5, 7, 34], as detailed in Table 1. POG [7] considers fashion outfits as bundles, and many music tracks in the same session of Spotify [5] are treated as bundles. Bundles of Food and Electronic [34] are constructed with high-quality metadata and meticulous intents. Pre-trained BLIP [20] is adopted to extract visual and textual embeddings across these datasets. To make fair comparisons with baselines, this work inherits the features extracted from baseline work [28] for Spotify and POG. We split all bundles into train:valid:test sets with a ratio of 7 : 1 : 2 for four datasets. Within the valid set and test sets, items in each bundle are randomly masked as the target items to be predicted, while the remaining items form the partial bundle [28]. The ubiquitous retrieval metrics [13, 24, 28, 34], such as Recall@K ( $R@K$ ) and NDCG@K ( $N@K$ ), are employed to evaluate the prediction of models.

#### 4.1.2 Comparative Baselines.

Based on the groundbreaking work in bundle construction task [28], we take into account the following baselines<sup>3</sup>: **Bi-LSTM** [15], **HyperGraph** [39], **Trans** [37], **TransCL** [28], **GAT** [1], **CLHE** [28]. In this study, GAT utilizes a graph attention mechanism to propagate high-order bundle-item affiliations, then computes the ultimate prediction as other comparative models. Meanwhile, the other baselines are followed to the experimental setups of Ma et al. [28].

<sup>3</sup> Due to space limitation, ‘Trans’ is an abbreviation for the ‘Transformer’-based model.

### 4.1.3 Implementation Details.

According to related works [26, 28, 30, 34], RaMen adopts Xavier initialization and Adam optimizer [19], setting the prevalent configuration including the embedding size as 64, the batch size as 1024, the learning rate as  $1e-3$  and regularization weight as  $1e-5$ . The hyperparameters are tuned by empirical studies, according to the related studies we inherit for each module. Inherited [41, 42],  $\epsilon$  is tuned in increments based on dataset size. The values of  $L_1, L_2, N, Z$  are empirically explored within range  $\{1, 2, 3, 4, 5\}$ , and  $\epsilon$  is set as 5 for POG, 450 for Spotify, 1 for Food/Electronic relied on its interaction distribution. The number of hyperedges  $H$  is chosen across  $\{4, 8, 16, 32, 64\}$ , and  $\beta, \gamma, \lambda_1$  are tuned in range  $\{0.1, 0.2, \dots, 0.8, 0.9\}$ . RaMen is implemented using PyTorch, and trained on NVIDIA A100 80GB GPUs & T4 15GB GPUs. Baselines are conducted in the same configuration and acknowledged available results in the steering work by Ma et al. [28]. Bi-LSTM results should be merely acknowledged according to [28], as we could not reproduce the same performance on Spotify. Our repository is available on Github via <https://github.com/Rec4Fun/RaMen>.

### 4.2 Performance Comparison

Table 2 demonstrates RaMen’s effectiveness and adaptability in various scenarios with different domains and distributions, proving its superior performance in bundle construction compared to SOTA approaches. The most significant improvements are indicated on benchmark datasets with small-sized bundle structures targeting specific intents, where RaMen achieves up to 77.31%, 66.61%, and 32.04% higher *w.r.t*  $R@20$  compared to the strongest baseline CLHE on Electronic, Food, and POG, respectively. Besides, RaMen significantly outperforms all competitive attention-based architectures, such as Trans, TransCL, GAT and CLHE, exemplifying its ability to comprehensively encode essential features hidden in both characteristic and collaborative strategies. Despite dealing with noise in learning target strategies for large bundles caused by numerous high-impact items within each bundle [30, 37], RaMen maintains modest yet steady gains on Spotify, especially in terms of ranking performance. We attribute this robustness to RaMen’s ability to dexterously model distinct decision-making strategies while integrating MAD module to enhance knowledge exchange between them. These observations prove RaMen adeptly exploits both explicit and implicit strategies, combining their potential to facilitate optimal decisions in this task.

### 4.3 Ablation Study

#### 4.3.1 Effect of different important components.

This study systematically removes the Characteristic Strategy Encoder (*w/o CrSE*), Collaborative Strategy Encoder (*w/o CbSE*), Implicit Strategy-aware Learning (*w/o ISL*), Multi-strategy Alignment & Discrimination (*w/o MAD*), textual (*w/o T*), and visual (*w/o V*) features to investigate the impact of RaMen’s core components in optimizing bundle construction strategies. As shown in Table 3, the findings underscore the significance of learning explicit strategies. Notably, *w/o CrSE* causes a more significant performance degradation than *w/o CbSE* on the sparse dataset POG, revealing the essence of capturing intrinsic information when extrinsic information is limited. Conversely, denser datasets like Electronics and Food show substantial drops in performance *w/o CbSE*, emphasizing the importance of understanding item relations in informed bundling decisions.

In addition, removing ISL leads to notable performance declines, particularly on POG, where detailed item descriptions promote targeting specific fashion segments. This highlights the effectiveness

Dataset	Metric	Bi-LSTM	HyperGraph	Trans	TransCL	GAT	CLHE	RaMen	% $\uparrow$
POG	R@10	0.0101	0.0113	0.0145	0.0160	0.0144	<u>0.0213</u>	<b>0.0264</b> <sup>‡</sup>	23.94
	N@10	0.0072	0.0074	0.0097	0.0109	0.0098	<u>0.0160</u>	<b>0.0191</b> <sup>‡</sup>	19.38
	R@20	0.0170	0.0207	0.0215	0.0202	0.0208	<u>0.0284</u>	<b>0.0375</b> <sup>‡</sup>	32.04
	N@20	0.0097	0.0111	0.0114	0.0134	0.0118	<u>0.0193</u>	<b>0.0226</b> <sup>‡</sup>	17.10
Spotify	R@10	-	0.0306	0.0552	0.0593	0.0506	<u>0.0689</u>	<b>0.0695</b>	0.87
	N@10	-	0.0923	0.1587	0.1698	0.1493	<u>0.1950</u>	<b>0.2060</b> <sup>‡</sup>	5.64
	R@20	0.0833	0.0572	0.0875	0.1014	0.0824	<u>0.1081</u>	<b>0.1091</b>	0.83
	N@20	0.1486	0.0941	0.1460	0.1696	0.1390	<u>0.1806</u>	<b>0.1882</b> <sup>‡</sup>	4.21
Electronic	R@10	0.0352	0.0616	0.1952	0.2355	0.3536	<u>0.4407</u>	<b>0.7410</b> <sup>‡</sup>	68.14
	N@10	0.0242	0.0344	0.1294	0.1562	0.2643	<u>0.3300</u>	<b>0.5104</b> <sup>‡</sup>	54.67
	R@20	0.0574	0.0928	0.2555	0.3050	0.3943	<u>0.4721</u>	<b>0.8371</b> <sup>‡</sup>	77.31
	N@20	0.0298	0.0430	0.1456	0.1757	0.2812	<u>0.3390</u>	<b>0.5373</b> <sup>‡</sup>	58.50
Food	R@10	0.0189	0.0712	0.2453	0.2346	0.3793	<u>0.4557</u>	<b>0.7575</b> <sup>‡</sup>	66.23
	N@10	0.0071	0.0379	0.1783	0.1769	0.2806	<u>0.3237</u>	<b>0.5028</b> <sup>‡</sup>	55.33
	R@20	0.0350	0.1055	0.3137	0.3088	0.4097	<u>0.5077</u>	<b>0.8459</b> <sup>‡</sup>	66.61
	N@20	0.0114	0.0478	0.1983	0.1985	0.2917	<u>0.3386</u>	<b>0.5242</b> <sup>‡</sup>	54.81

**Table 2.** Overall performances of RaMen compared with competitive baselines on four benchmark datasets from diverse domains. The best results are in **bold**, and the second best results are underlined. The symbol <sup>‡</sup> indicates statistically significant improvements over the second-best models with  $p - \text{value} < 0.05$  obtained through the average performance of five runs of each model.

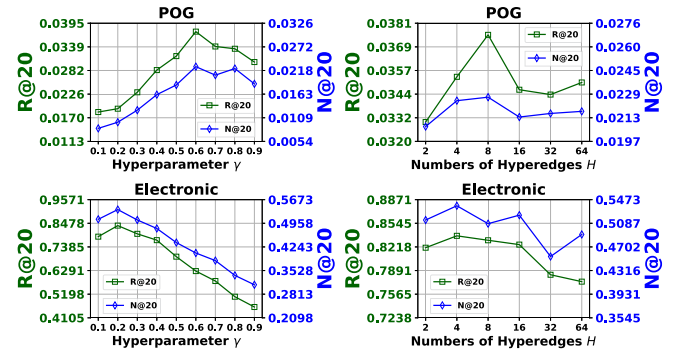
Dataset	Metric	w/o CrSE	w/o CbSE	w/o ISL	w/o MAD	w/o V	w/o T
POG	R@20	0.0218( $\downarrow 41.87$ )	0.0301( $\downarrow 19.73$ )	0.0335( $\downarrow 10.67$ )	0.0348( $\downarrow 7.20$ )	0.0332( $\downarrow 11.47$ )	0.0293( $\downarrow 21.87$ )
	N@20	0.0119( $\downarrow 47.35$ )	0.0209( $\downarrow 7.52$ )	0.0202( $\downarrow 10.62$ )	0.0210( $\downarrow 7.08$ )	0.0204( $\downarrow 9.73$ )	0.0171( $\downarrow 24.34$ )
Spotify	R@20	0.1028( $\downarrow 5.77$ )	0.0995( $\downarrow 8.80$ )	0.1071( $\downarrow 1.83$ )	0.1080( $\downarrow 1.01$ )	0.1041( $\downarrow 4.58$ )	0.1055( $\downarrow 3.30$ )
	N@20	0.1772( $\downarrow 5.84$ )	0.1662( $\downarrow 11.69$ )	0.1863( $\downarrow 1.01$ )	0.1838( $\downarrow 2.34$ )	0.1806( $\downarrow 4.04$ )	0.1830( $\downarrow 2.76$ )
Electronic	R@20	0.7376( $\downarrow 11.89$ )	0.3560( $\downarrow 57.47$ )	0.8081( $\downarrow 3.46$ )	0.7836( $\downarrow 6.39$ )	0.7383( $\downarrow 11.80$ )	0.7516( $\downarrow 10.21$ )
	N@20	0.4780( $\downarrow 11.04$ )	0.2595( $\downarrow 51.70$ )	0.5253( $\downarrow 2.23$ )	0.5035( $\downarrow 6.29$ )	0.4656( $\downarrow 13.34$ )	0.4936( $\downarrow 8.13$ )
Food	R@20	0.7871( $\downarrow 6.95$ )	0.3732( $\downarrow 55.88$ )	0.8116( $\downarrow 4.05$ )	0.8184( $\downarrow 3.25$ )	0.7951( $\downarrow 6.01$ )	0.8112( $\downarrow 4.10$ )
	N@20	0.4815( $\downarrow 8.15$ )	0.2788( $\downarrow 46.81$ )	0.5122( $\downarrow 2.29$ )	0.5104( $\downarrow 2.63$ )	0.4966( $\downarrow 5.27$ )	0.4978( $\downarrow 5.04$ )

**Table 3.** Ablation study on different vital components of RaMen. The figures in subscription with the symbol  $\downarrow$  denote the % reduction of performance when each component of the proposed model is omitted.

of our intuition in modeling latent shared attributes among items. These observations also explain the significant effect of textual/visual features on POG compared to other datasets. Fundamentally, the model derives most of its critical information from the two encoder mechanisms employed in Explicit Strategy-aware Learning (ESL). Meanwhile, ISL serves as a complementary module, which enables the system to construct more productive bundles by enriching item and bundle representations through alignment with latent shared attributes. Besides, the experiments of 'Only ISL' were also conducted and yielded extremely low results. This detection is expected, as this setting merely operates on primitive semantic embeddings of items combined with randomly initialized IDs. Completely omitting ESL prevents the model from capturing essential aspects of bundle construction strategies, such as distinctive characteristics and interdependence among items, which makes bundle representation almost meaningless. Thus, the inclusion of 'Only ISL' evaluation is deemed unnecessary, as its modest impact can be inferred from the performance drop observed when ISL is omitted. Finally, the considerable decreases in performance w/o MAD demonstrate that RaMen's multi-strategy learning architecture, enhanced by transferring supervision signals through MAD module, enables a more comprehensive and robust grasp of product bundling.

#### 4.3.2 Impacts of critical hyperparameters.

In practice, each domain (*music, fashion, etc.*) has different priorities in weighting strategies for producing bundles. In this work, empirical studies for hyperparameters are referred to in the related studies



**Figure 2.** Impact of  $\gamma, H$  on RaMen's performance.

where we modeled each component, and tuned it using grid search. These observations are clarified with the aim of providing the essential value range and insights for extending further work.

Figure 2 shows the impact of key hyperparameters on the model's performance, specifically the control parameter  $\gamma$  in Explicit Strategy-aware Learning and the number of hyperedges  $H$  in Implicit Strategy-aware Learning w.r.t R@20 and N@20. The findings demonstrate that RaMen's performance is highly sensitive to  $\gamma$  and  $H$ , requiring careful tuning to achieve optimal performance. RaMen can adapt the effect of collaborative strategy (*directly susceptible to data sparsity*) and characteristic strategy through  $\gamma$ . In essence, the sparser the data, the more crucial the characteristic encoder becomes.



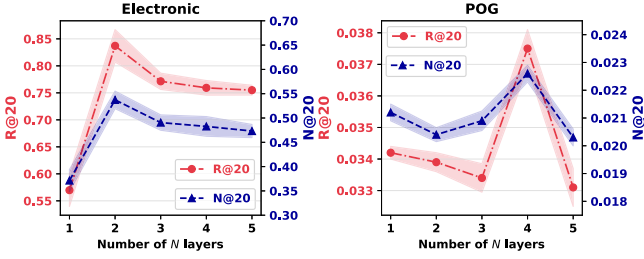


Figure 3. Impact of hyperparameter  $N$  on RaMen's performance.

In particular, the fluctuation of  $\gamma$  further reinforces the critical role of both CbSE and CrSE on different data domains, as discussed. As  $\gamma$  increases, the performance of RaMen initially improves, reaches an optimal value, and then declines across both datasets. This behavior can be attributed to the role of  $\gamma$  as a balancing parameter, which regulates the relative contributions of multi-modal features and collaborative signals in the Explicit Strategy-Aware Learning module. An improperly tuned value of  $\gamma$  may disrupt this balance, causing one encoder to dominate the other, ultimately leading to suboptimal performance. Regarding the hyperparameter  $H$ , both small and large-scale datasets benefit from a relatively low number of hyperedges (4 or 8), which conserves memory for computational resources while maintaining competitive performance.

The figure 3 illustrates the impact of varying the number of our refined attention layers ( $N$  layers) in the item-item graph component of ESL on RaMen's performance, evaluated on  $R@20$  and  $N@20$ . Two datasets (Electronic and POG) are shown to demonstrate the diversity in data size and sparsity. Besides, similar observations are obtained on other datasets. As shown in figure 3, the gain in the sparser dataset POG is more modest. The optimal number of attention layers varies: 2 for Electronic, 4 for POG. This indicates that the structure and density of the item-item graph significantly influence how deep the network should be. In both datasets, stacking too many layers (e.g., value of 5) degrades performance, due to oversmoothing - a common issue in GNNs where node representations become indistinguishably similar. More layers allow for the capture of higher-order relationships as well as multi-context interdependence among items in the bundle construction problem, but they also increase the risk of propagating irrelevant or redundant information, especially in large or noisy graphs. The oversmoothing problem is also easily obtained with hypergraphs via the observation of  $H$ .

#### 4.4 Qualitative Showcases

The results validate the effectiveness of multi-strategy multi-modal learning for automatic bundle construction, particularly in practical scenarios like bundles in POG (bundles with IDs as 2234 and 18762) and Electronic (bundles with IDs as 507 and 1090), which is aimlessly selected among sets of prominent predicted cases and depicted in Figure 4. In both scenarios, RaMen demonstrates superior capability in capturing bundling strategies compared to the state-of-the-art model CLHE, particularly in retrieving complementary items aligned with user intents. In contrast, CLHE primarily relies on semantic information and fails to exploit complex relations among items, resulting in not only biases towards substitute products (e.g., *memory cards*, *cameras* in Electronics; and *pants* in POG), but also the misconception of product segmentation (e.g., POG bundle aimed at *men's fashion*). Notably, in the first example, 5 out of the 10 items predicted by CLHE are shoes, which, although similar to the ground truth, reveal a tendency of CLHE to over-rely on multi-modal fea-

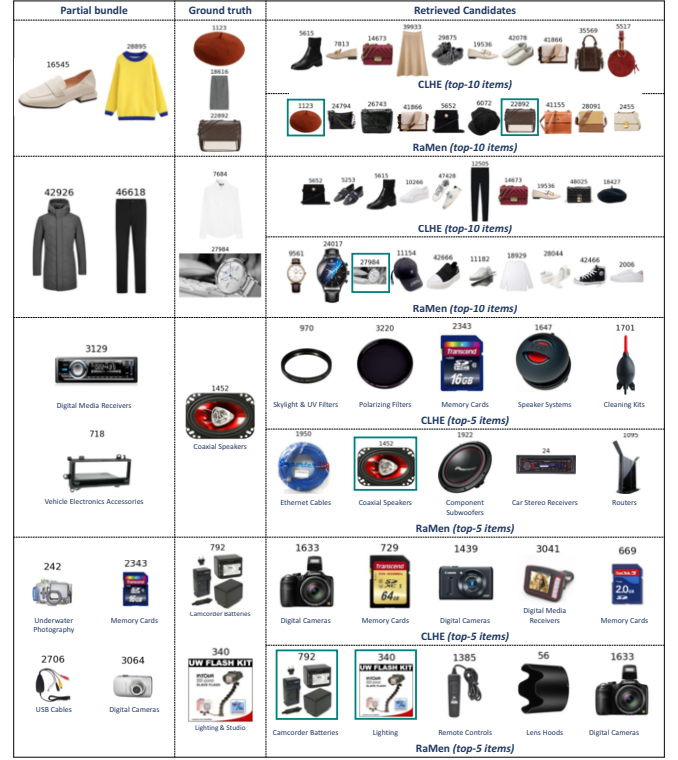


Figure 4. Practical studies of top- $K$  item candidates of RaMen compared to CLHE across the POG and Electronic dataset. The green box surrounding the item's image denotes the item that the model has correctly predicted.

ture similarity. This approach often results in the prediction of alternative rather than complementary items, a limitation in cases such as fashion bundles, where a balanced selection of items is crucial. In the third example, most of the items predicted by RaMen belong to the "Car & Vehicle Electronics" category, aligning well with the partial bundle's intent. In contrast, none of CLHE's predictions are relevant to the bundle's intended purpose. This highlights RaMen's superior ability to grasp bundle intents, a strength likely driven by the hypergraph structure in its Implicit Strategy-aware Learning module, where items with related information may be strongly connected with the same hyperedges. By explicitly leveraging item characteristics and collaborations and integrating implicit shared attributes, RaMen overcomes these shortcomings of SOTA models. Moreover, these observations provide valuable insights into how multi-strategy approaches can further improve performance in the recommendation of complex item sets [21, 34, 28, 33].

## 5 Conclusion

This study introduces RaMen, a novel bundle construction approach that effectively integrates intrinsic and extrinsic information using Explicit and Implicit Strategy-aware Learning. Through extensive experiments across multiple domains and datasets, RaMen consistently outperforms state-of-the-art methods. Moreover, RaMen demonstrates robustness in handling noise within large bundle structures by effectively modeling distinct decision-making strategies and facilitating knowledge transfer between them. Experiment analyses demystify RaMen's ability to model and integrate diverse decision-making strategies, providing a comprehensive and robust framework for solving bundle construction. Our insightful illustrations can further explore handling even larger bundle structures and refine strategy alignment techniques for complex item set recommendations.



## References

- [1] S. Brody, U. Alon, and E. Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022.
- [2] T.-N. Bui, H.-S. Nguyen, C.-V. N. Thi, H.-Q. Le, and D.-T. Le. Bridge: Bundle recommendation via instruction-driven generation. *arXiv preprint arXiv:2412.18092*, 2024.
- [3] T.-N. Bui, H.-S. Nguyen, C.-V. T. Nguyen, H.-Q. Le, and D.-T. Le. Personalized diffusion model reshapes cold-start bundle recommendation. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 3088–3091, 2025.
- [4] J. Chang, C. Gao, X. He, D. Jin, and Y. Li. Bundle recommendation and generation with graph neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 35(3):2326–2340, 2021.
- [5] C.-W. Chen, P. Lamere, M. Schedl, and H. Zamani. Recsys challenge 2018: Automatic music playlist continuation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 527–528, 2018.
- [6] L. Chen, Y. Liu, X. He, L. Gao, and Z. Zheng. Matching user with item set: Collaborative bundle recommendation with deep attention network. In *IJCAI*, pages 2095–2101, 2019.
- [7] W. Chen, P. Huang, J. Xu, X. Guo, C. Guo, F. Sun, C. Li, A. Pfadler, H. Zhao, and B. Zhao. Pog: personalized outfit generation for fashion recommendation at alibaba ifashion. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2662–2670, 2019.
- [8] Q. Deng, K. Wang, M. Zhao, R. Wu, Y. Ding, Z. Zou, Y. Shang, J. Tao, and C. Fan. Build your own bundle—a neural combinatorial optimization method. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2625–2633, 2021.
- [9] Z. Deng, J. Li, Z. Guo, W. Liu, L. Zou, and G. Li. Multi-view multi-aspect neural networks for next-basket recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1283–1292, 2023.
- [10] Y. Ding, P. Mok, Y. Ma, and Y. Bin. Personalized fashion outfit generation with user coordination preference learning. *Information Processing & Management*, 60(5):103434, 2023.
- [11] X. Du, K. Qian, Y. Ma, and X. Xiang. Enhancing item-level bundle representation for bundle recommendation. *ACM Transactions on Recommender Systems*, 2023.
- [12] Y. Fang, X. Xiao, X. Wang, and H. Lan. Customized bundle recommendation by association rules of product categories for online supermarkets. In *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, pages 472–475. IEEE, 2018.
- [13] C. Gao, Y. Zheng, N. Li, Y. Li, Y. Qin, J. Piao, Y. Quan, J. Chang, D. Jin, X. He, et al. A survey of graph neural networks for recommender systems: Challenges, methods, and directions. *ACM Transactions on Recommender Systems*, 1(1):1–51, 2023.
- [14] Z. Guo, J. Li, G. Li, C. Wang, S. Shi, and B. Ruan. Lgmrec: Local and global graph learning for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8454–8462, 2024.
- [15] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis. Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1078–1086, 2017.
- [16] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang. 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*, pages 639–648, 2020.
- [17] R. T. Irene, C. Borrelli, M. Zanon, M. Buccoli, and A. Sarti. Automatic playlist generation using convolutional neural networks and recurrent neural networks. In *2019 27th European signal processing conference (EUSIPCO)*, pages 1–5. IEEE, 2019.
- [18] E. Jang, S. Gu, and B. Poole. Categorical reparametrization with gumbel-softmax. In *International Conference on Learning Representations (ICLR 2017)*. OpenReview. net, 2017.
- [19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations*, 2015.
- [20] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.
- [21] M. Li, S. Jullien, M. Ariannezhad, and M. de Rijke. A next basket recommendation reality check. *ACM Transactions on Information Systems*, 41(4):1–29, 2023.
- [22] F. Liu, H. Chen, Z. Cheng, A. Liu, L. Nie, and M. Kankanhalli. Disentangled multimodal representation learning for recommendation. *IEEE Transactions on Multimedia*, 25:7149–7159, 2022.
- [23] G. Liu, Y. Fu, G. Chen, H. Xiong, and C. Chen. Modeling buying motives for personalized product bundle recommendation. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 11(3):1–26, 2017.
- [24] Q. Liu, J. Hu, Y. Xiao, X. Zhao, J. Gao, W. Wang, Q. Li, and J. Tang. Multimodal recommender systems: A survey. *ACM Computing Surveys*, 2023.
- [25] X. Liu, J. Wu, Z. Tao, Y. Ma, Y. Wei, and T.-s. Chua. Fine-tuning multimodal large language models for product bundling. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 848–858, 2025.
- [26] Y. Ma, Y. He, A. Zhang, X. Wang, and T.-S. Chua. Crosscbr: Cross-view contrastive learning for bundle recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1233–1241, 2022.
- [27] Y. Ma, Y. He, W. Zhong, X. Wang, R. Zimmermann, and T.-S. Chua. Crip: Cross-item relational pre-training for multimodal product bundling. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9641–9649, 2024.
- [28] Y. Ma, X. Liu, Y. Wei, Z. Tao, X. Wang, and T.-S. Chua. Leveraging multimodal features and item-level user feedback for bundle construction. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 510–519, 2024.
- [29] H.-S. Nguyen, T.-N. Bui, L.-H. Nguyen, D.-C. Can, C.-V. T. Nguyen, D.-T. Le, and H.-Q. Le. Hhmc: a heterogeneous x homogeneous graph-based network for multimodal cross-selling recommendation. In *the 15th International Conference on Knowledge and Systems Engineering*, pages 1–6. IEEE, 2023.
- [30] H.-S. Nguyen, T.-N. Bui, L.-H. Nguyen, H. Hoang, C.-V. Thi Nguyen, H.-Q. Le, and D.-T. Le. Bundle recommendation with item-level causation-enhanced multi-view learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 324–341. Springer, 2024.
- [31] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [32] A. Pathak, K. Gupta, and J. McAuley. Generating and personalizing bundle recommendations on steam. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1073–1076, 2017.
- [33] M. Sun, L. Li, M. Li, X. Tao, D. Zhang, P. Wang, and J. X. Huang. A survey on bundle recommendation: Methods, applications, and challenges. *arXiv preprint arXiv:2411.00341*, 2024.
- [34] Z. Sun, K. Feng, J. Yang, H. Fang, X. Qu, Y.-S. Ong, and W. Liu. Revisiting bundle recommendation for intent-aware product bundling. *ACM Transactions on Recommender Systems*, 2(3):1–34, 2024.
- [35] Z. Sun, K. Feng, J. Yang, X. Qu, H. Fang, Y.-S. Ong, and W. Liu. Adaptive in-context learning with large language models for bundle generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 966–976, 2024.
- [36] A. Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- [37] Y. Wei, X. Liu, Y. Ma, X. Wang, L. Nie, and T.-S. Chua. Strategy-aware bundle recommender system. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1198–1207, 2023.
- [38] L. Xia, C. Huang, Y. Xu, J. Zhao, D. Yin, and J. Huang. Hypergraph contrastive collaborative filtering. In *Proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval*, pages 70–79, 2022.
- [39] Z. Yu, J. Li, L. Chen, and Z. Zheng. Unifying multi-associations through hypergraph for bundle recommendation. *Knowledge-Based Systems*, 255:109755, 2022.
- [40] S. Zhao, W. Wei, D. Zou, and X. Mao. Multi-view intent disentangle graph networks for bundle recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4379–4387, 2022.
- [41] H. Zhou, X. Zhou, L. Zhang, and Z. Shen. Enhancing dyadic relations with homogeneous graphs for multimodal recommendation. In *ECAI 2023*, pages 3123–3130. IOS Press, 2023.
- [42] X. Zhou and Z. Shen. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of ACM International Conference on Multimedia*, pages 935–943, 2023.
- [43] T. Zhu, P. Harrington, J. Li, and L. Tang. Bundle recommendation in e-commerce. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 657–666, 2014.