Dual-Enhanced Item Representation for Bundle Construction via Category-Wise and Cross-Modality Learning

Nguyen, Long Hai; Nguyen, Huy Son; Thi Nguyen, Cam Van; Le, Duc Trong; Takasu, Atsuhiro; Le, Hoang Quynh

DOI
10.1145/3767695.3769501

Publication date
2025

Document Version
Final published version

Published in
SIGIR-AP 2025 - Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region

RESEARCH-ARTICLE

Dual-Enhanced Item Representation for Bundle Construction via Category-Wise and Cross-Modality Learning

LONG HAI NGUYEN, Vietnam National University, Hanoi, Hanoi, Vietnam

HUY SON NGUYEN, Delft University of Technology, Delft, Zuid-Holland, Netherlands

CAM-VAN THI NGUYEN, Vietnam National University, Hanoi, Hanoi, Vietnam

DUC TRONG LE, Vietnam National University, Hanoi, Hanoi, Vietnam

ATSUHIRO TAKASU TAKASU, Research Organization of Information and Systems National Institute of Informatics, Tokyo, Japan

HOANGQUYNH LE, Vietnam National University, Hanoi, Hanoi, Vietnam

Published: 07 December 2025

SIGIR-AP 2025:Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region
December 7 - 10, 2025
Xi'an, China

Conference Sponsors:
SIGIR

https://doi.org/10.1145/3767695.3769501
ISBN: 9798400722189

Dual-Enhanced Item Representation for Bundle Construction via Category-Wise and Cross-Modality Learning

Long-Hai Nguyen
VNU University of Engineering and Technology
Hanoi, VietNam
longhai230303@gmail.com

Huy-Son Nguyen
Delft University of Technology
Delft, The Netherlands
h.s.nguyen@tudelft.nl

Cam-Van Thi Nguyen
VNU University of Engineering and Technology
Hanoi, VietNam
vanntc@vnu.edu.vn

Duc-Trong Le
VNU University of Engineering and Technology
Hanoi, VietNam
trongld@vnu.edu.vn

Atsuhiro Takasu
National Institute of Informatics
Tokyo, Japan
takasu@nii.ac.jp

Hoang-Quynh Le*
VNU University of Engineering and Technology
Hanoi, VietNam
lhquynh@vnu.edu.vn

Abstract

Bundle recommender systems merely learn from existing bundles, but obtaining large-scale, high-quality bundle datasets remains a challenge, especially for platforms newly adopting bundle services. Bundle construction is the task of automatically selecting a set of compatible items to form a coherent bundle, a vital step before making recommendations on bundle-aware platforms. Groundbreaking work on bundle construction, like CLHE, has been designed solely on user-item interaction and self-attention modules to learn item/bundle representations. These techniques fall short of the standards for coherent bundles in real-world applications, where the relation among the semantic information of items should be considered more thoroughly. To address these challenges, we explicitly leverage category-wise information and employ cross-modal fusion to enhance item representations. By doing so, we propose **Caro**: Dual-Enhanced Item Representation for Bundle Construction via Category-Wise and Cross-Modality Learning. **Caro** captures the inherent relationships between items within analogous categories, improving bundle coherence. It comprises three main components: (1) cross-modality enhanced item representation, (2) category-enhanced item representation, and (3) bundle contrastive learning. Extensive experiments and detailed analyzes using multiple real-world datasets demonstrate that our method outperforms existing state-of-the-art techniques and provides valuable insight into the bundle construction problem. Notably, **Caro** achieves a $5-8\%$ higher *Recall*@20 than the strongest baseline, underscoring its performance gains through dual category-wise and cross-modal enhancements. Our repository is available at https://github.com/L2R-UET/CaRo.

*Corresponding author.

SIGIR-AP 2025, Xi'an, China

ACM ISBN 979-8-4007-2218-9/2025/12
https://doi.org/10.1145/3767695.3769501

CCS Concepts

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

Keywords

Bundle construction, Category-wise, Cross-attention, Multimodal recommendation

ACM Reference Format:
Long-Hai Nguyen, Huy-Son Nguyen, Cam-Van Thi Nguyen, Duc-Trong Le, Atsuhiro Takasu, and Hoang-Quynh Le. 2025. Dual-Enhanced Item Representation for Bundle Construction via Category-Wise and Cross-Modality Learning. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP 2025), December 7–10, 2025, Xi'an, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3767695.3769501

1 Introduction

Recently, bundle recommendation [5, 30] has emerged as a practical paradigm in e-commerce, fashion, and social media platforms, offering users a set of items that serve as a cohesive solution rather than isolated products. Compared with traditional recommendation systems that merely focus on individual items, bundle recommendation improves user interest and increase revenue through cross-selling and reduced decision fatigue of users [18, 21]. Bundle construction refers to the automated process of generating a group of items that are not only relevant but also semantically and visually compatible [17, 21]. However, constructing high-quality bundles is still a challenge, especially from a large dataset of items where user interaction data is sparse or new items lack historical engagement.

Bundle construction offers several merits for enterprises, including reducing manual labor and time costs, streamlining inventory and logistics management, and facilitating large-scale manufacturing or deployment. For example, in the electronics industry, companies often bundle smartphones with accessories like chargers, cases, and earbuds, which simplifies packaging and reduces the need for separate quality checks and handling. Similarly, in e-commerce, automated bundle generation helps platforms like Amazon or Shopee reduce the effort required to manually curate related products, accelerating promotions and improving customer

satisfaction. By grouping complementary items, bundle construction not only enhances the shopping experience but also promotes efficient product delivery and inventory turnover.

Recent advances in bundle construction have focused on leveraging user-item interactions and multimodal information to generate candidate bundles. Despite their effectiveness, existing methods such as CLHE [17] often suffer from two significant limitations. First, they typically treat multimodal features (e.g, text, image, user feedback) as flat concatenations without modeling the cross-modal relationships. Second, they overlook category-level semantics, which are crucial for ensuring coherence among items within a bundle. As a result, generated bundles may lack contextual relevance and be less coherent.

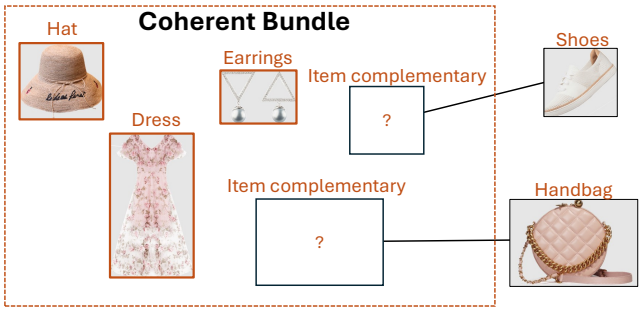

Figure 1: An illustrative example of bundle coherence. Given a partial bundle (hat, earrings, dress), our task aims to recommend complementary items (e.g., shoes and handbag) that align in style, category, and function.

Therefore, it is important to distinguish between bundle recommendation and bundle construction. Bundle recommendation assumes a predefined catalog of bundles and mainly focus on recommending the most suitable ones to users, whereas bundle construction aims to dynamically generate new bundles, or complete bundles by selecting complementary items for a large candidate pool. Our work addresses the latter, which is inherently more challenging since it requires modeling item complementarity and semantic coherence rather than relying on pre-existed bundles.

Approach and Contribution In this paper, we propose `Caro` (Category-aware Cross Modal Attention Enhancing bundle construction). This novel framework incorporates category-wise information and multimodal fusion to address key limitations in existing bundle construction. Specifically, `Caro` is designed to capture both inter-modal relationships and structured item-category dependencies, thereby improving the semantic coherence and contextual compatibility of generated bundles. `Caro` consists of three main components:

- Cross-modal fusion of textual, visual, and user-item interaction features that aims to improve the semantic quality of item embeddings;
- A category enhanced item representation learning module that propagates semantic signals from item categories using a graph-based approach;
- A bundle contrastive learning module with contrastive learning objectives that obtain coherent item relationships and enhance generalization.

2 Related Work

The research on product bundling spans two well-known tasks [21]: **bundle recommendation**, which assumes a predefined bundle catalog and ranks them for users [18], and **bundle construction**, which dynamically assembles new bundles by selecting complementary items [17]. While both aim to improve user experience, recommendation focuses on matching existing bundles, whereas construction must explicitly model item complementarity and coherence. The latter has recently been treated as an independent task, yet it is crucial for controllability and scalability in practice.

Graph-based approaches. Graph-based models represent users, items, or bundles as nodes and learn embeddings enriched by structural context. Early techniques such as LightGCN [10] and KGAT [26] capture user–item relations, while bundle-specific methods refine these strategies. For bundle recommendation, CrossCBR [16] and MultiCBR [15] use contrastive learning (adopting InfoNCE [8]) to align user and item representations across modalities, which can address inconsistencies between individual and bundle-level preferences and ensure semantic consistency. For bundle construction, RaMen [19] further emphasizes the combination of collaborative signals, semantic features, and latent intent to better capture bundle-level structures. Besides, CLHE [17] enhances the representation learning process by utilizing multi-modal features and a item-user graph with LightGCN [10], which lacks the capability to capture the intricate condition among items based on its categories.

Generative-based approaches. Generative models learn the distribution of valid bundles from historical data and synthesize new ones accordingly [3, 4]. Recently, Liu et al. [14] propose a fine-tuning LLM-based model (Bundle-MLLM) for bundle construction by integrating textual, media, and interaction data into a coalesced tokenization. In terms of bundle recommendation, Bui et al. [4], it first groups correlated items, then generates pseudo-"ideal" bundles by combining these clusters with user interaction data. This generative process enriches sparse interaction matrices and improves recommendation performance over baselines.

Our approach. Unlike previous work, our framework emphasizes *controllable and semantically coherent bundle construction* by integrating category-enhanced item representations with multi-modal cross-attention fusion. This structured design captures dependencies more explicitly while avoiding reliance on large-scale generative sampling.

3 Methodology

In this section, we formally define the bundle construction problem and outline its details. Subsequently, we demonstrate the four important modules of `Caro` illustrated in Figure 2 including: (1) Cross-Modality Enhanced Item Representation; (2) Category Enhanced Item Representation; (3) Bundle Contrastive Learning; (4) Prediction and Joint Optimization.

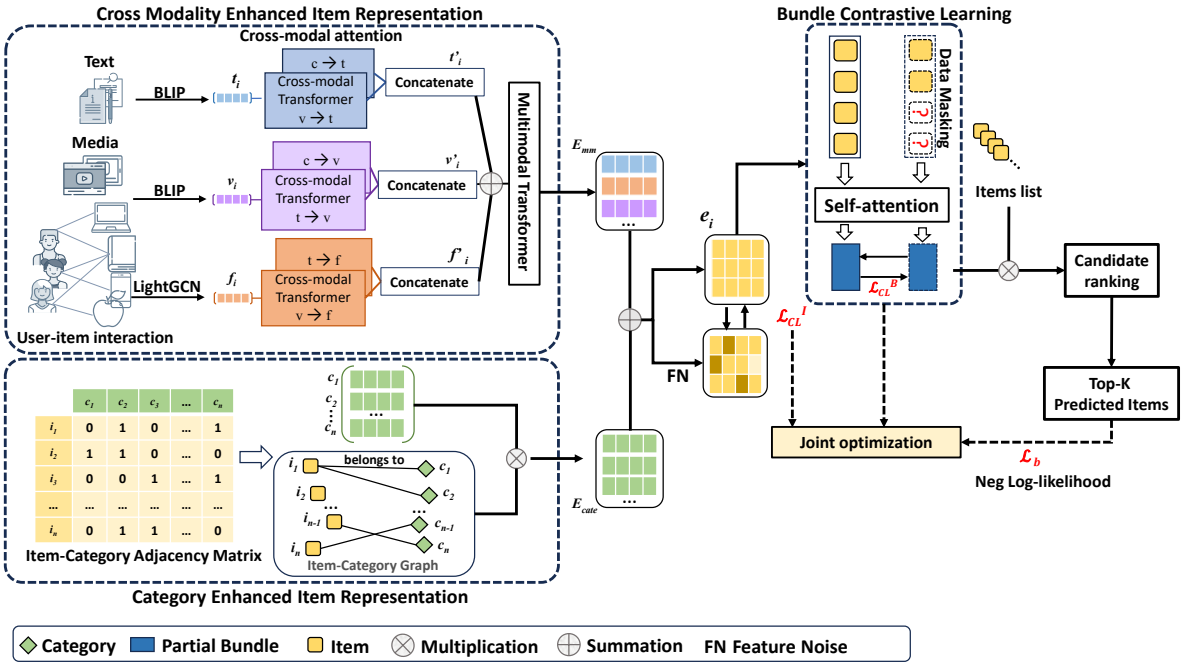

Figure 2: The schematic illustration of our proposed model Caro.

3.1 Preliminaries

Given an item set of $\mathcal{I} = \{i_1, i_2, ..., i_{|\mathcal{I}|}\}$, each item has three modalities as input: textual input t_i can be its titles, descriptions; visual input v_i can be image or video of the item; and representation from item-level user feedback f_i. Furthermore, item-user feedback data, which is denoted as user-item interaction matrix $\mathbf{X}^{|\mathcal{U}|\times|\mathcal{I}|} = \{x_{ui} | u \in \mathcal{U}, i \in \mathcal{I}\}$, where $\mathcal{U} = \{u_1, u_2, ..., u_{|\mathcal{U}|}\}$ is the user set, with entries ($x_{ui} = 1$) if user u interacted with item i. We define the set of categories as $C = \{c_1, c_2, ..., c_{|C|}\}$, and build the sparse binary adjacency matrix for the category-wise item information, as follows:

$$A \in \{0,1\}^{|C|\times|\mathcal{I}|}, \quad A_{c,i} = \begin{cases} 1, & \text{if item } i \text{ belongs to category } c, \\ 0, & \text{otherwise.} \end{cases}$$

Each bundle, as a set of items, is denoted as $b = \{i_1, i_2, ..., i_{|b|}\}$, where $|b|$ is the size of the bundle. Given a partial bundle $b_s \subset b$, where the model aims to predict missing items $i \in \{b \backslash b_s\}$. In addition, we also define a set of known bundles for training, denoted as $\mathcal{B}_t = \{b_1, b_2, ..., b_T\}$ and a set of unseen bundles for testing, denoted as $\overline{\mathcal{B}}_t = \{b_{T+1}, b_{T+2}, ..., b_{T+\overline{T}}\}$, where T is the number of training bundles and $\overline{T}$ is the number of testing bundles. The objective of this work is to predict missing items from an incomplete item bundle. We define **complementarity** as the property that items play different but compatible roles when combined within a bundle (e.g., a shirt and a pair of pants forming a complete outfit). In a successful bundle, items are chosen for their complementarity, ensuring the resulting set is coherent, diverse, and non-redundant.

3.2 Cross-Modal Enhanced Item Representation

Multimodal feature extraction. Building upon the work of [17], the textual and visual features of items are extracted using large-scale multi-modal foundation models. Specifically, we employ the BLIP (Bootstrapped Language Image Pretraining) model [13], a state-of-the-art vision language model to obtain both textual and visual representations. Textual information refers to the item's title and its product description. On the other hand, visual features are extracted from the image of the item, capturing its characteristics, such as color, shape, and material. After the feature extraction process, this research acquires the corresponding textual $t_i \in \mathbb{R}^{768}$ and visual features $v_i \in \mathbb{R}^{768}$ for each item. These embeddings serve as inputs to the multimodal learning module, enabling the model to learn the characteristics of the multimodal item effectively.

Item-level user feedback feature extraction In the context of **item-level user feedback**, *feedback* refers to individual items that have user interactions (e.g, clicks, views, purchases, or ratings). In this work, those interactions are mainly treated as binary implicit feedback: 1 if the user has interacted with the item; 0 if there is no recorded interaction or indicates low interest through lows ratings. Our research captures user preferences toward individual items, not just bundle level interactions by using user-item interaction graphs based on the user-item interaction matrix. Proposed method constructs a bipartite user-item interaction graph $\mathcal{G} = (\mathcal{U} \cup \mathcal{I}, \mathcal{E})$, where $\mathcal{U}$ is the set of users, $\mathcal{I}$ is the set of items, $\mathcal{E}$ represents interactions between users and items. By applying LightGCN [10] for representation learning due to its performance and optimal

complexity:

$$\begin{cases} \mathbf{p}_u^{(k+1)} = \sum_{i \in \mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}} \mathbf{p}_i^{(k)}, \\ \mathbf{p}_i^{(k+1)} = \sum_{u \in \mathcal{N}_i} \frac{1}{\sqrt{|\mathcal{N}_i||\mathcal{N}_u|}} \mathbf{p}_u^{(k)}, \end{cases} \tag{1}$$

where $\mathbf{p}_u^{(k)}$ and $\mathbf{p}_i^{(k)}$ denote the embeddings of user u and item i at the k-th propagation layer, respectively. $\mathcal{N}_u$ represents the set of items that user u has interacted with, and $\mathcal{N}_i$ represents the set of users who have interacted with item i. The normalization $\frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}}$ balances message passing by adjusting for the degrees of users and items. After that, we use $\mathbf{p}_i^{(k)}$ which had captured item-level user feedback information by aggregating the item representation through K layers' propagation, as follows:

$$\mathbf{p}_i = \frac{1}{K} \sum_{k=0}^{K} \mathbf{p}_i^{(k)}. \tag{2}$$

resulting representation of item-level user feedback is f_i.

Multimodal fusion. In this work, three modalities are used: text (t), visual (v), collaborative filtering information (f). We denote with $X_{\{t,v,f\}} \in \mathbb{R}^{F_{\{t,v,f\}} \times d}$. For the rest of this work, $F(\cdot)$ and $d(\cdot)$ are used to represent sequence length and feature dimension. These notations are used in the multimodal fusion module, which consists of four steps.

1) 1D convolutions layer: To enhance local contextual representation and reduce noise in the input embeddings, we first apply a 1D convolution layer. These steps help refine raw modality, enabling them to be effective for cross-modal fusion later. All modality inputs were processed using a one-dimensional convolutional layer:

$$\hat{X}_{\{t,v,f\}} = \text{Conv1D}(X_{\{t,v,i\}}, k_{\{t,v,f\}}) \in \mathbb{R}^{F_{\{t,v,f\}} \times d} \tag{3}$$

where k denote the convolutional kernel sizes for each modality $\{t, v, f\}$, d is the shared embedding dimension. Because the convolutions layers project the different modalities to the same dimension d, enabling the dot products to be computed in the subsequent cross-modal attention stage.

2) Positional embedding: Although input features $\hat{X}_{\{t,v,f\}}$ are extracted from pretrained encoder - BLIP, the output embedding loses explicit sequence awareness when used in downstream models. To encode multimodal information and provide structural order, positional embeddings (PE) are applied to extracted features $\hat{X}_{\{t,v,f\}}$. Since transformers do not recognize the order of input tokens, PE provides essential ordering, allowing the model to attend meaningfully across positions. We adopt fixed sinusoidal embeddings, which were proposed by Vaswani et al. [25]:

$$H^{[0]}_{\{t,v,f\}} = \hat{X}_{\{t,v,f\}} + \text{PE}(F_{\{t,v,f\}}, d) \tag{4}$$

where $H^{[0]}$ denotes the low level positionally augmented input representations for different modalities, $\text{PE}(F_{\{t,v,f\}}, d) \in \mathbb{R}^{F_{\{t,v,f\}} \times d}$ generates fixed embeddings for each position.

3) Cross-modal transformers: Our work considers combine modalities as a pair, in this research γ and φ are denoted as two modalities with sequences from each as $X_\gamma \in \mathbb{R}^{F_\gamma \times d}$ and $X_\varphi \in \mathbb{R}^{F_\varphi \times d}$, where F indicates the represent sequence length and d is feature embedding size which was set to the same size. Inspired by Tsai et al. [23], our work adopts a **cross-modal attention mechanism** which provides latent adoptaion across modalities (e.g, γ to φ or vice versa). This step is mandatory because unimodal representations often lack contextual information that only exists in other modalities. For instance, the expression in an image can change the meaning of a description from neutral to sarcastic, whereas user-item interaction patterns can provide important background that further explains the context of text or image. Therefore, by allowing each modality to attend to the others, the proposed model can build more context-aware representations. In our research, three modalities $\{t, v, f\}$ were used. To make each modality receive information from the other two, a pair of cross-modal transformer modules is applied to each modality. As a result, with three different modalities in use, we have a total of six cross-modal transformer modules (see Figure 2).

We defines the Querys, Keys, and Values as below:

$$Q_\varphi = X_\varphi W_{Q_\varphi}, \quad K_\gamma = X_\gamma W_{K_\gamma}, \quad V_\gamma = X_\gamma W_{V_\gamma} \tag{5}$$

where $W_{Q_\varphi} \in \mathbb{R}^{d_\varphi \times d_k}$, $W_{K_\gamma} \in \mathbb{R}^{d_\gamma \times d_k}$, and $W_{V_\gamma} \in \mathbb{R}^{d_\gamma \times d_v}$. The latent adaptation from γ to φ is presented as cross-modal attention $Y_\varphi = \text{CrossAtt}_{\gamma \to \varphi}(X_\varphi, X_\gamma)$

$$Y_\varphi = \text{softmax}\left(\frac{Q_\varphi K_\beta^\top}{\sqrt{d_k}}\right) V_\gamma \tag{6}$$

$$Y_\varphi = \text{softmax}\left(\frac{X_\varphi W_{Q_\varphi} W_{K_\gamma}^\top X_\gamma^\top}{\sqrt{d_k}}\right) X_\gamma W_{V_\gamma} \tag{7}$$

Building on the cross-modal attention mechanism, our work applies the Transformer framework that allows one modality to integrate information from another modality. This research fixes all dimensions ($d\{\gamma, \varphi, k, v\}$ for each cross-modal attention block as d In Figure 2, passing textual (t) information to visual (v) is denoted by ($t \to v$). Each cross-modal transformer consists of D layers of cross-modal attention blocks. A cross-modal transformer computes feed-forward for $i = 1, \cdots, D$ layers:

$$\begin{aligned} H^{[0]}_{\gamma \to \varphi} &= H^{[0]}_\varphi \\ \hat{H}^{[i]}_{\gamma \to \varphi} &= \text{CrossAtt}^{[i],\text{mul}}_{\gamma \to \varphi}\left(\text{LN}(H^{[i-1]}_{\gamma \to \varphi}),\ \text{LN}(H^{[0]}_\gamma)\right) + \text{LN}(H^{[i-1]}_{\gamma \to \varphi}) \\ H^{[i]}_{\gamma \to \varphi} &= f^{[i]}_{\gamma \to \varphi}\left(\text{LN}(\hat{H}^{[i]}_{\gamma \to \varphi})\right) + \text{LN}(\hat{H}^{[i]}_{\gamma \to \varphi}) \end{aligned} \tag{8}$$

where f is a position wise feed-forward sublayer, and $\text{CrossAtt}^{[i],\text{mul}}_{\gamma \to \varphi}$ is a multihead version of $\text{CrossAtt}_{\gamma \to \varphi}$ at layer i. LN means layer normalization [1] At this stage, each modality continues to update itself through the multi-head cross-modal attention.

4) Self-Attention Transformers: At the final step of this module, we concatenate the outputs from previous stage, that correspond to the same target modality, resulting in $H_{\{t,v,f\}} \in \mathbb{R}^{T_{\{t,v,f\}} \times 2d}$. Each concatenated representation is then processed through a sequence modeling component to get the final item representations, which are subsequently used in the bundle learning phase. The final representation of item is denoted as E^{mm}_{item}

3.3 Category Enhanced Item Representation

Using item category-wise information provides high-level semantics that enhance representation learning. To capture item complementarity within bundles, we propose a category-wise item representation learning module which propagates semantic information from category embeddings to item representation via a graph-based approach.

We initialize a learnable embedding matrix for item categories:

$$E_{\text{cate}} = \text{Embed}(C) \in \mathbb{R}^{C \times d}, \tag{9}$$

where C is the set of C categories, and d is the embedding dimension. Each row $E^{(0)}_{\text{cate}}[c]$ represents the embedding of category c, initialized randomly and updated during training.

To model the relationships between categories and items, we construct a bipartite graph in the form of a sparse binary adjacency matrix $A \in \mathbb{R}^{C \times I}$, where $A_{c,i} = 1$ if item i belongs to category c, and $A_{c,i} = 0$ otherwise. Since most item belongs to a limited number of categories, the resulting matrix is sparse.

To perform embedding propagation from categories to items, we normalize the adjacency matrix to balance the influence from multiple categories. Specifically, we define a normalized matrix $\hat{A} \in \mathbb{R}^{C \times I}$ as follows:

$$\hat{A}_{c,i} = \frac{A_{c,i}}{\sum_{c'=1}^{C} A_{c',i}}. \tag{10}$$

Instead of an unweighted sum, this normalization step ensures every item receives information from its associated categories. Using this normalized graph, we compute the dense category-aware item embeddings as follows:

$$E^{\text{cate}}_{\text{item}} = \hat{A}^{\top} E_{\text{cate}} \in \mathbb{R}^{N \times d}. \tag{11}$$

Each row of $E^{\text{cate}}_{\text{item}}$ represents a dense embedding for item i. Optionally, we apply a linear transformation and non-linearity to project these features into a more expressive space:

$$E^{\text{cate}}_{\text{item}} = \text{ReLU}(\hat{A}^{\top} E_{\text{cate}} W_c), \quad W_c \in \mathbb{R}^{d \times d}. \tag{12}$$

This transformation allows the model to refine the category features before integrating with other item information.

Finally, we fuse the category-aware item embedding $E^{\text{cate}}_{\text{item}}$ with other item-level features, multimodal embeddings $E^{\text{mm}}_{\text{item}}$ produced from previous section. The final item representation is computed via concatenation followed by an MLP, as follows:

$$\mathbf{e}_i = \text{MLP}\left([E^{\text{mm}}_{\text{item}} \,\|\, E^{\text{cate}}_{\text{item}}] \right), \tag{13}$$

This fused representation captures both low-level (visual, textual) and high-level (category) semantics, enabling the model to make more coherent and context-aware predictions during bundle construction.

3.4 Bundle Contrastive Learning

After obtaining the item representation including both multimodal features and category-aware information, we employ a self-attention module to capture representation of the given partial bundle. The representation of a partial bundle is computed as follows:

$$\begin{aligned} \mathbf{A}^{(z)}_b &= \frac{1}{\sqrt{d}} \hat{\mathbf{E}}^{(z-1)}_b \mathbf{W}^K_B \left(\hat{\mathbf{E}}^{(z-1)}_b \mathbf{W}^Q_B \right)^{\top}, \\ \tilde{\mathbf{E}}^{(z)}_b &= \text{softmax}(\mathbf{A}^{(z)}_b) \hat{\mathbf{E}}^{(z-1)}_b, \end{aligned} \tag{14}$$

where $\mathbf{W}^K_B, \mathbf{W}^Q_B \in \mathbb{R}^{d \times d}$ are trainable weight matrices used to project item embeddings into key and query spaces. The notation $\hat{\mathbf{E}}^{(z-1)}_b \in \mathbb{R}^{|b| \times d}$ denotes the hidden representations of items in the partial bundle b at the $(z-1)$-th attention layer. The result $\tilde{\mathbf{E}}^{(z)}_b$ is the attention-weighted output after the z-th layer. By stacking Z self-attention layers, this research obtains the final set of item-level representations for the bundle $\tilde{\mathbf{E}}^{(Z)}_b = \text{concat}(\{\mathbf{e}_i\}_{i \in b})$. To aggregate this into a single, fixed-length bundle representation $\mathbf{e}_b$, our research applies average pooling over all item embeddings within the bundle, as follows:

$$\mathbf{g}_b = \text{average}(\tilde{\mathbf{E}}^{(Z)}_b). \tag{15}$$

3.5 Prediction and Joint Optimization

Prediction. To estimate the probability that item i belongs to partial bundle g_{b_p}, we adopt the inner product to compute score $\hat{y}_{b_p,i}$, which indicates the possibility of item i being included into bundle b to make bundle complementary, defined as:

$$\hat{y}_{b_p,i} = g_{b_p} e_i^T \tag{16}$$

Inspired by [20], the negative log-likelihood (NLL) is employed as the primary optimization objective after obtaining the prediction, therefore the loss for bundle b is denoted as:

$$\mathcal{L}_b = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} -y_{b_p,i} \log \delta_b(\hat{y}_{b_p,i}), \tag{17}$$

where $\delta(\cdot)$ denotes the softmax function applied over the entire item space, and $\hat{y}_{b_s,i}$ represents the predicted score for item i in bundle b_s. The ground-truth label $y_{b_s,i}$ indicates whether item i belongs to the target bundle.

Contrastive learning. For each item, this research gets its representation $\mathbf{e}_i$ in Equation 13. We use data augmentation method the generate the augmented view $\mathbf{e}'_i$ by adding a small-scaled random noise vector to the item's features, this technique is called Feature Noise (FN), inspired by [28]. We employ the InfoNCE technique [8] to align knowledge from various strategy aware perspectives and ensure the discrimination of embeddings for individual objects, leading to more distinct comprehensive representations for items/bundles. The item-level contrastive loss is derived as:

$$\mathcal{L}^I_{CL} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} -\log \frac{\exp\left(\cos(\mathbf{e}_i, \mathbf{e'}_i)/\tau\right)}{\sum_{j \in \mathcal{I}} \exp\left(\cos(\mathbf{e}_i, \mathbf{e'}_j)/\tau\right)}, \tag{18}$$

where $\cos(\cdot)$ is the cosine similarity function, τ is the temperature parameter, and $\mathcal{I}$, $\mathcal{B}$ is the set of item indices, bundles indices respectively.

$$\mathcal{L}^B_{CL} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} -\log \frac{\exp\left(\cos(\mathbf{g}_b, \mathbf{g'}_b)/\tau\right)}{\sum_{j \in \mathcal{I}} \exp\left(\cos(\mathbf{g}_b, \mathbf{g'}_j)/\tau\right)} \tag{19}$$

At bundle level, we apply category popularity masking strategy for data augmentation producing $\mathbf{g'}_b$, which aim to preserve the

Datasets	$\lvert\mathcal{U}\rvert$	$\lvert\mathcal{I}\rvert$	$\lvert\mathcal{B}\rvert$	$\lvert C\rvert$	$\mathcal{E}_{\mathcal{BI}}$	$\mathcal{E}_{\mathcal{UI}}$	Avg.$\mathcal{I}/\mathcal{B}$	Avg.$\mathcal{B}/\mathcal{I}$
POG	17,449	48,676	20,000	72	72,224	237,519	3.61	1.48
Electronic	888	3,499	1,750	513	6,165	6,165	3.52	1.76
Food	879	3,767	1,784	735	6,395	6,395	3.58	1.70

Table 1: Dataset statistics of benchmark datasets in the bundle construction task.

semantic coherence of the bundle. In contrast, random masking may remove the anchor items within a bundle, potentially degrading the efficiency of contrastive learning.

Joint optimization. To train the model effectively, proposed method combines this reconstruction loss with both item-level and bundle-level contrastive losses, along with an L2 regularization term. The total objective function is formulated as:

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \mathcal{L}_b + \lambda_1 \mathcal{L}_{CL}^I + \lambda_2 \mathcal{L}_{CL}^B + \beta \|\Theta\|_2^2, \tag{20}$$

where λ_1, λ_2, and β are hyperparameters that balance the influence of the contrastive and regularization components. Here, $\|\Theta\|_2^2$ denotes the L2 norm of all trainable parameters in the model.

4 Experiments

4.1 Experimental Setup

Component		Parameters
Embeddings	Item Embedding	`n_item × d`
	Bundle Embedding	`n_bundle × d`
	Category Embedding	`n_category × d`
Multimodal Attention	Number of Heads	`num_heads = 4`
	Transformer Layers	`layers = 2`
	Fusion Type	`fusion = 'cross-attention'`
Bundle Masking Strategy	Bundle Ratio	`bundle_ratio = 0.5`
Contrastive Learning	Item-Level Loss	`item_contrastive_loss = 2.0`
	Bundle-Level Loss	`bundle_contrastive_loss = 0.5`
Optimization	Optimizer	`optimizer = 'Adam'`
	Learning Rate	`lr = 1e-3`
	Batch Size	`batch_size = 1024`

Table 2: Hyper-parameters of the proposed model.

Datasets. Our work uses three datasets in various domains. POG [6] is a fashion dataset collected from TaoBao which is the largest publicly available dataset of fashion items with rich information compared to existing datasets Fashion-136K [11]. Polyvore Outfits [24], which contains all the metadata required for the multimodal usage in this work. Two additional Electronic and Food datasets are used from the famous Amazon product corpus, which provides user-item interactions, along with item metadata, such as category labels, images, and textual descriptions. The structure of the dataset is based on Sun et al. (2022). Following [17], we randomly split all bundles into train:valid:test set with a ratio of 7:1:2. For the valid and test sets, items in each bundle are masked by strategy-based as the targeted items to be predicted, while the remaining items form the partial bundle.

The statistics are shown in Table 1. Where $|\mathcal{U}|$, $|\mathcal{I}|$, and $|B|$ represent the number of users, items, and bundles, respectively. $\mathcal{E}_{\mathcal{UI}}$ and $\mathcal{E}_{\mathcal{UB}}$ denote the number of connections among users-items and users-bundles, respectively. As shown in 1, Avg.$\mathcal{I}/\mathcal{B}$ indicates the mean number of items within each bundle across the entire dataset. A unique characteristic of the POG dataset is that each item is associated with only one category. This limits the diversity of category information for learning, which partly explains the smaller performance obtains compared to Electronic and Food datasets, where items typically belong to multiple categories. The richer category associations in those domains provide more informative supervision, which allow model to better leverage category information during training. The data also highlights the strong correlation between bundle construction and the performance of our proposed model, which will be discussed in more detail.

Evaluation Metrics. Recall (*R@K*) and Normalized Discounted Cumulative Gain (*N@K*) are two commonly employed metrics for evaluating the performance of a method in a bundle recommendation task. *R@K* measures the proportion of test bundles within the top-K ranking list. *N@K* manifests normalized discounted cumulative gain scores aimed at obtaining relevant items at higher positions on the ranking list. Both metrics are employed with $K \in \{10, 20\}$ for performance validation. The smaller the top-K, the more clearly it demonstrates the model's practical recommendation performance. Average performances in 5 runs with various random initializations are reported. Comparisons are evaluated by two tailed paired sample Student's *t-test* with *p-value* of 0.05.

Baselines. Following [17], we consider the following baselines, as this recommendation task remains relatively underexplored within the recommendation field, all of the dataset settings are applied to all baselines with the same ratio:

- **Bi-LSTM**[9]: This methods capture sequential and relational dynamics between items within a bundle. The key idea is to treat bundle components as sequences and use Bi-LSTM to get the contextual compatibility between items.
- **HyperGraph** [29]: This research applies hypergraphs to bundle recommendation, indicating that hypergraph can effectively unify the modeling of multiple user-bundle association by representing them as hyper-edges It has been proved that hypergraph modeling technique simplifies and extends the descriptions of numerous associations
- **Transformer** (Trans) [27] : a framework for bundle recommendation that models strategy-aware representation of both users and bundles to enhance user-bundle prediction. It includes 3-component token embedding layers for input representation, a Hierarchical Graph Transformer layer for capturing multi-level structure, and a prediction layer.
- **Transformer Contrastive Learning** (TransCL)[17] : is the version that adds bundle level contrastive loss to the above Transformer [27] model.
- **GAT**[2] : GAT utilizes a graph attention mechanism to propagate high-order bundle-item affiliations

Dataset	POG				Electronic				Food			
Metric	*R@10*	*R@20*	*N@10*	*N@20*	*R@10*	*R@20*	*N@10*	*N@20*	*R@10*	*R@20*	*N@10*	*N@20*
Bi-LSTM [9]	0.0101	0.0170	0.0072	0.0097	0.0352	0.0574	0.0242	0.0298	0.0189	0.0350	0.0071	0.0114
HyperGraph[29]	0.013	0.0207	0.0074	0.0111	0.0616	0.0928	0.0344	0.0430	0.0712	0.1055	0.0379	0.0478
Transformer[27]	0.0145	0.0215	0.0097	0.0114	0.1952	0.2555	0.1294	0.1456	0.2453	0.3137	0.1783	0.1983
TransformerCL[17]	0.0160	0.0202	0.0109	0.0134	0.2355	0.3050	0.1562	0.1757	0.2346	0.3088	0.1769	0.1985
GAT[2]	0.0144	0.0208	0.0098	0.0118	0.3536	0.3943	0.2643	0.2812	0.3793	0.4097	0.2806	0.2917
CLHE [17]	0.0205	0.0284	0.0159	0.0193	0.4407	0.4721	0.3300	0.3390	0.4557	0.5077	0.3237	0.3386
Caro	**0.0215**	**0.0303**	**0.0169**	**0.0195**	**0.7719**	**0.7943**	**0.6030**	**0.6092**	**0.7432**	**0.7904**	**0.5714**	**0.5848**
Imp (%)	4.9	8.0	6.3	0.8	75.2	68.2	82.7	79.7	63.1	55.7	76.5	72.7

Table 3: Overall performance of Caro compared with competitive baselines on three benchmark datasets from diverse domains. The best results are in bold, and the second-best results are underlined.

- **CLHE** [17]: CLHE utilizes multimodal feature of items and leverage both item, bundle contrastive learning to perform the bundle construction task.

Implementation Details. Caro adopts Xavier initialization [7] and Adam optimizers [12], setting the embedding size to 64, the batch size to 1024, the learning rate to $1e-3$ and regulazrization weight to $1e-5$. Hyperparameters $\lambda_1, \lambda_2, \beta$ are tuned in range $\{0.1, 0.2..., 0.8, 0.9\}$. We considered results in $K \in \{10, 20\}$ as this range is typically taken into consideration in other recommendation research. Caro is conducted using PyTorch, and trained on A100 GPU & T4 15GB GPus. All baselines are conducted in the same configuration, settings, and acknowledge the available results in a prestigious publication [17].

Model configuration. The detailed fine-calibrated parameters of all components in the optimal proposed model are shown in Table 2.

4.2 Performance Comparison to Baselines

Table 3 shows the performance of Caro against the above mentioned baselines across three benchmark datasets: POG [6], Electronic [22], and Food [22]. Across all three evaluation datasets—POG, Electronic, and Food,Caro consistently outperforms existing baselines in terms of both *Recall@K* and *NDCG@K*, demonstrating its robust generalization across domains with varying sparsity and semantic structures. On the **POG dataset**, which is the largest but also the sparsest in terms of interactions, Caro achieves a *Recall*@20 of 0.0303, surpassing the strongest baseline CLHE [17] by 8.0%, respectively. In addition, these gains underscore the model's capacity to convey semantic category signals and utilize contrastive learning, which is particularly beneficial in low-signal environments where item co-occurrence is limited.

In the **Electronic** and **Food** datasets, which are more compact but semantically dense, Caro delivers even more significant improvements. It achieves a *Recall*@20 of 0.7943 and *NDCG*@20 of 0.6092 on the Electronic dataset, outperforming CLHE by 68.2% and 79.7%, respectively. On the Food dataset, Caro reaches *Recall*@20 of 0.7904 and *NDCG*@20 of 0.5848, marking relative improvements of 55.7% and 72.7% over the best baseline. These results affirm the model's effectiveness in capturing both compatibility and semantic coherence, even in domains with high redundancy. The observed improvements can be primarily attributed to two components of Caro: (1) the cross-modal attention mechanism effectively aligns and integrates multimodal features, overcoming limitations seen in prior works that use naive concatenation; and (2) Item-category Enhancement Module: By using a graph-based item-category learning, Caro manage to get semantic information from category embeddings to items. Together, these components enable Caro to model both item-level and high-level contextual signals, leading to substantial gains across diverse settings.

4.3 Ablation Study

Impact of modality. To assess the impact of integrating knowledge from different modalities, our work divides into five ablated settings to examine its effective: removing multimodal fusion module (w/o MM Fusion), removing multimodal feature by only using item embedding for training (w/o MM), using only textual features (only Textual), using only media (only Visual), and using only Item-level User Feedback information (only Item-level User). The percentage decrease from the full Caro performance is also reported to further explore the importance of each modality in Table 4.

Effectiveness of Multimodal Information. We first examine the impact of removing multimodal features entirely by training the model with only ID-based item embeddings. This variant, denoted as *w/o MM*, results in a substantial performance drop: Recall@20 decreases from 0.7943 to 0.425 on the Electronic dataset and from 0.7904 to 0.4585 on the Food dataset—corresponding to a relative decline of over 46%. These results confirm that multimodal features provide rich semantic signals beyond what can be captured by basic embeddings alone, thereby enhancing item understanding and compatibility modeling.

Effectiveness of Multimodal Fusion. While prior work such as CLHE [17] also incorporates multimodal signals, it does so via simple concatenation, which weakly models modality interactions. To validate the importance of structured fusion, we replace our cross-modal attention mechanism with a simple pooling-based fusion strategy (*w/o MM Fusion*). This leads to a significant degradation in performance across all metrics and datasets, demonstrating that the manner in which modalities are combined plays a crucial role. Without appropriate fusion, semantic conflicts and modality misalignment can hinder representation learning.

Effectiveness of Individual Modalities. We further evaluate single-modality variants of our model, using only one input type:

Dataset	Metric	*w/o MM Fusion*	*w/o MM*	*only Textual*	*only Visual*	*only Item-level User CF*	**Caro**
Electronic	*R@10*	$0.4357_{(\downarrow 43.55)}$	$0.3960_{(\downarrow 48.70)}$	$0.3631_{(\downarrow 52.96)}$	$0.3669_{(\downarrow 52.47)}$	$0.7238_{(\downarrow 6.23)}$	**0.7719**
	R@20	$0.4614_{(\downarrow 41.91)}$	$0.4245_{(\downarrow 46.55)}$	$0.3960_{(\downarrow 50.15)}$	$0.4017_{(\downarrow 49.43)}$	$0.7660_{(\downarrow 3.57)}$	**0.7943**
	N@10	$0.3339_{(\downarrow 44.63)}$	$0.3058_{(\downarrow 49.29)}$	$0.2674_{(\downarrow 55.66)}$	$0.2721_{(\downarrow 54.87)}$	$0.5322_{(\downarrow 11.74)}$	**0.6030**
	N@20	$0.3411_{(\downarrow 44.01)}$	$0.3132_{(\downarrow 48.59)}$	$0.2763_{(\downarrow 54.65)}$	$0.2821_{(\downarrow 53.69)}$	$0.5442_{(\downarrow 10.67)}$	**0.6092**
Food	*R@10*	$0.4283_{(\downarrow 42.37)}$	$0.4073_{(\downarrow 45.19)}$	$0.2806_{(\downarrow 62.25)}$	$0.3947_{(\downarrow 46.89)}$	$0.7138_{(\downarrow 3.96)}$	**0.7432**
	R@20	$0.4680_{(\downarrow 40.79)}$	$0.4585_{(\downarrow 42.00)}$	$0.3205_{(\downarrow 59.45)}$	$0.4286_{(\downarrow 45.78)}$	$0.7738_{(\downarrow 2.10)}$	**0.7904**
	N@10	$0.3184_{(\downarrow 46.80)}$	$0.3040_{(\downarrow 44.29)}$	$0.2247_{(\downarrow 60.67)}$	$0.3054_{(\downarrow 46.56)}$	$0.5236_{(\downarrow 8.37)}$	**0.5714**
	N@20	$0.3299_{(\downarrow 45.69)}$	$0.3176_{(\downarrow 43.59)}$	$0.2370_{(\downarrow 59.47)}$	$0.3148_{(\downarrow 46.17)}$	$0.5401_{(\downarrow 7.63)}$	**0.5848**

Table 4: The performance of Caro once omitting different modalities. The subscription ↓ denotes the percentage decrease of the overall performance when a specific modality is ablated from Caro.

user feedback at the text, visual, or item level, respectively. Among these, models using only item-level user feedback perform the best across the entire model, suggesting that user-item interaction data capture valuable latent preference signals and play a crucial role in bundle construction. In contrast, using only textual or visual features results in significant performance drops, reinforcing the importance of multimodal feature richness. However, these results also caution that when modalities are not fused effectively, they can degrade performance, highlighting the need for rich input and effective integration.

Effectiveness of Category-wise Item Learning. Figure 3 shows that item category-aware learning consistently improves recommendation performance across both the Electronic and Food datasets. In the Electronic dataset, excluding the category-aware learning module, model performance drops from 0.7943 to 0.7562 in terms of *Recall*@20, representing a decline of approximately 5.25% compared to Caro. A similar pattern is observed in the Food domain, where removing the category-aware learning step causes *Recall*@20 drop from 0.7904 to 0.7631. These findings show the importance of incorporating category-level semantics to enhance item representation and reinforce the model's ability to deliver more accurate and category-aware recommendations.

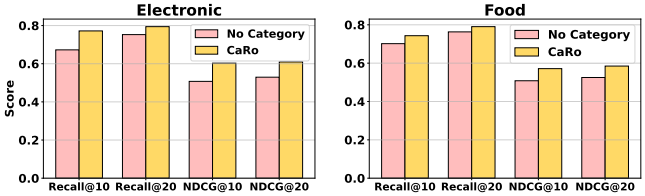

Figure 3: Performance Comparison With and Without Category Information across two datasets: Electronic and Food.

4.4 Qualitative Showcase

To qualitatively verify the effectiveness of our proposed approach in this research, we examine a real-world example from the Electronic dataset. This analysis shows how Caro captures item compatibility and semantic coherence more effectively than the current SOTA baseline CLHE model.

Example overview: Given a partial bundle containing items, the task is to retrieve additional items that semantically complement the bundle. These seed items indicate bundle intent focused on photography related equipment.

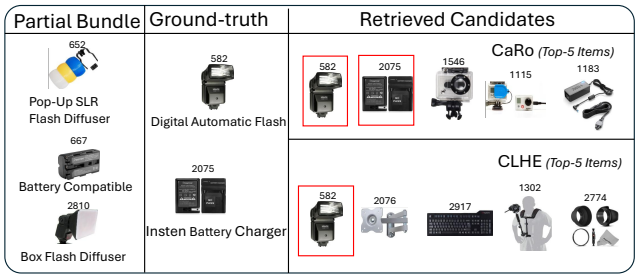

Figure 4: Practical cases of retrieved candidates from CLHE and Caro from Electronic dataset. The correct predicted items are surrounded by red boxes.

Category-aware Learning Effect. Our proposed model leverages category-wise item information to get the high-level semantic structure into the item retrieval process. By explicitly modeling category-item relationships, our model learns that items such as in Figure 4 *battery*, *flash*, and *charger* belong to a shared category, it is photography equipment in this case. These category embeddings act as semantic anchors that guide the model toward retrieving items within relevant domains. Thereby, output predictions become more coherent, even when items have low visual similarity.

Contribution of Multimodal Cross-attention. In addition, our model incorporates a multimodal cross attention mechanism to capture the interaction between three modalities. This multimodal alignment allows Caro to make more accurate and reasonable predictions when constructing bundles.

5 Conclusion

In conclusion, this research presents Caro, a novel framework for bundle construction that effectively integrates category-aware semantics and cross-modality learning item features. By jointly learning item representations through cross-modal fusion and category-item graph propagation, Caro demonstrates a superior ability to capture both semantic coherence and user preferences. Extensive evaluations across three benchmark datasets validate the effectiveness of our approach, significantly outperforming strong baselines on multiple metrics. Our work highlights the importance of incorporating structured category knowledge and cross-modality alignment in bundle construction, which offers new perspectives for future research in recommendation systems.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. arXiv:1607.06450 [stat.ML] https://arxiv.org/abs/1607.06450

[2] Shaked Brody, Uri Alon, and Eran Yahav. 2022. How Attentive are Graph Attention Networks?. In *International Conference on Learning Representations*.

[3] Tuan-Nghia Bui, Huy-Son Nguyen, Cam-Van Thi Nguyen, Hoang-Quynh Le, and Duc-Trong Le. 2025. Personalized Diffusion Model Reshapes Cold-Start Bundle Recommendation. In *Companion Proceedings of the ACM on Web Conference 2025*. 3088–3091.

[4] Tuan-Nghia Bui, Huy-Son Nguyen, Cam-Van Nguyen Thi, Hoang-Quynh Le, and Duc-Trong Le. 2024. BRIDGE: Bundle Recommendation via Instruction-Driven Generation. *arXiv preprint arXiv:2412.18092* (2024).

[5] Jianxin Chang, Chen Gao, Xiangnan He, Depeng Jin, and Yong Li. 2020. Bundle recommendation with graph convolutional networks. In *Proceedings of the 43rd international ACM SIGIR conference on Research and development in Information Retrieval*. 1673–1676.

[6] Wen Chen, Pipei Huang, Jiaming Xu, Xin Guo, Cheng Guo, Fei Sun, Chao Li, Andreas Pfadler, Huan Zhao, and Binqiang Zhao. 2019. POG: personalized outfit generation for fashion recommendation at Alibaba iFashion. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2662–2670.

[7] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*.

[8] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 297–304.

[9] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. 2017. Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 25th ACM international conference on Multimedia*. 1078–1086.

[10] 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.

[11] Vignesh Jagadeesh, Robinson Piramuthu, Anurag Bhardwaj, Wei Di, and Neel Sundaresan. 2014. Large scale visual recommendations from street fashion images. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1925–1934.

[12] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization.. In *International Conference on Learning Representations (Poster)*.

[13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[14] Xiaohao Liu, Jie Wu, Zhulin Tao, Yunshan Ma, Yinwei Wei, and Tat-seng Chua. 2025. 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*. 848–858.

[15] Yunshan Ma, Yingzhi He, Xiang Wang, Yinwei Wei, Xiaoyu Du, Yuyangzi Fu, and Tat-Seng Chua. 2024. Multicbr: Multi-view contrastive learning for bundle recommendation. *ACM Transactions on Information Systems* 42, 4 (2024), 1–23.

[16] Yunshan Ma, Yingzhi He, An Zhang, Xiang Wang, and Tat-Seng Chua. 2022. CrossCBR: cross-view contrastive learning for bundle recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1233–1241.

[17] Yunshan Ma, Xiaohao Liu, Yinwei Wei, Zhulin Tao, Xiang Wang, and Tat-Seng Chua. 2024. 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*. 510–519.

[18] Huy-Son Nguyen, Tuan-Nghia Bui, Long-Hai Nguyen, Hung Hoang, Cam-Van Thi Nguyen, Hoang-Quynh Le, and Duc-Trong Le. 2024. Bundle Recommendation with Item-Level Causation-Enhanced Multi-view Learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 324–341.

[19] Huy-Son Nguyen, Quang-Huy Nguyen, Duc-Hoang Pham, Duc-Trong Le, Hoang-Quynh Le, Padipat Sitkrongwong, Atsuhiro Takasu, and Masoud Mansoury. 2025. RaMen: Multi-Strategy Multi-Modal Learning for Bundle Construction. *arXiv preprint arXiv:2507.14361* (2025).

[20] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[21] Zhu Sun, Kaidong Feng, Jie Yang, Hui Fang, Xinghua Qu, Yew-Soon Ong, and Wenyuan Liu. 2024. Revisiting bundle recommendation for intent-aware product bundling. *ACM Transactions on Recommender Systems* 2, 3 (2024), 1–34.

[22] Zhu Sun, Jie Yang, Kaidong Feng, Hui Fang, Xinghua Qu, and Yew Soon Ong. 2022. Revisiting Bundle Recommendation: Datasets, Tasks, Challenges and Opportunities for Intent-aware Product Bundling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2900–2911.

[23] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, Vol. 2019. 6558.

[24] Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. 2018. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European conference on computer vision (ECCV)*. 390–405.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[26] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 950–958.

[27] Yinwei Wei, Xiaohao Liu, Yunshan Ma, Xiang Wang, Liqiang Nie, and Tat-Seng Chua. 2023. Strategy-aware bundle recommender system. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1198–1207.

[28] Junliang Yu, Xin Xia, Tong Chen, Lizhen Cui, Nguyen Quoc Viet Hung, and Hongzhi Yin. 2023. XSimGCL: Towards extremely simple graph contrastive learning for recommendation. *IEEE Transactions on Knowledge and Data Engineering* 36, 2 (2023), 913–926.

[29] Zhouxin Yu, Jintang Li, Liang Chen, and Zibin Zheng. 2022. Unifying multi-associations through hypergraph for bundle recommendation. *Knowledge-Based Systems* 255 (2022), 109755.

[30] Tao Zhu, Patrick Harrington, Junjun Li, and Lei Tang. 2014. Bundle recommendation in ecommerce. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 657–666.