IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 4, August, 2025, pp. 2289-2299
Journal homepage: http://iieta.org/journals/ts

Cross-Modal Product Image Retrieval for E-Commerce Recommendation Systems via Deep Learning

Chao Zhang[1], Jing Feng[2*]

[1] School of Information Engineering, Suqian University, Suqian 223800, China
[2] Chengde College of Applied Technology, Chengde 067000, China

Corresponding Author Email: 19213080770@163.com

https://doi.org/10.18280/ts.420438

Received: 18 December 2024
Revised: 8 June 2025
Accepted: 20 June 2025
Available online: 14 August 2025

Keywords:
cross-modal product image retrieval, e-commerce recommendation, self-attention mechanisms, class-supervised learning, semantic masking, feature fusion

ABSTRACT

With the rapid evolution of e-commerce, increasing demands have been placed on the accuracy of product recommendation systems. Cross-modal product image retrieval, which serves as a critical bridge between visual content and textual descriptions, has a direct impact on user experience in such systems. However, existing retrieval approaches have been predominantly tailored for natural image scenarios, rendering them less effective in e-commerce contexts. Product images are required to highlight "purchasability features" but are often compromised by complex backgrounds, while textual descriptions—rich in both marketing language and functional attributes—frequently suffer from semantic dilution due to redundant content. Furthermore, generic models fail to optimize for the implicit alignment between "product attributes" and "user intent," resulting in suboptimal recommendation relevance. Although self-attention mechanisms have been introduced in cross-modal tasks, limitations persist in local visual feature extraction, core semantic focus in text, and deep alignment across modalities. To address these challenges, a cross-modal product image retrieval method tailored for e-commerce recommendation systems was proposed. A full-chain model based on self-attention mechanisms was constructed. A global-local visual feature extraction module was designed to enhance discriminative product regions using category labels. A textual feature extraction module was incorporated to suppress non-essential information and emphasize semantically decisive elements. Furthermore, a vision-language cross-extraction module was constructed to establish bidirectional mappings between sub-image patches and key textual tokens. This approach enables precise multimodal alignment, thereby providing a three-tiered matching rationale—image detail, textual demand, and purchase intent—for e-commerce recommendation systems.

1. INTRODUCTION

In the current digital era, e-commerce has become deeply integrated into daily life [1-4], with a continuous surge in the volume and diversity of available products. On leading e-commerce platforms, more than ten million new items are frequently added within a single quarter. Consumer shopping behavior has gradually shifted from a solely “text-based search” model to a hybrid pattern that combines “image reference with textual filtering” [5, 6]. For example, after uploading an image of a desired product, users may refine their selections further by adding descriptions such as “soft material” or “minimalist style.” However, current recommendation systems [7-9] often encounter challenges in handling such cross-modal demands. In particular, mismatches between fine-grained visual features and textual descriptions frequently result in recommendations that appear visually similar but deviate from the actual intent. Traditional cross-modal retrieval methods [10, 11] rely heavily on handcrafted feature extraction and shallow semantic mapping, rendering them insufficient for capturing the multi-dimensional relationships among product characteristics in e-commerce scenarios. In contrast, the strengths of deep learning in automatic feature learning and multimodal alignment modeling offer a promising pathway to overcome this core limitation.

The development of this research holds substantial practical and theoretical value. From an application perspective, accurate cross-modal product image retrieval enables recommendation systems to more effectively infer user intent. For instance, when a user uploads an image of a dress and adds the description “suitable for commuting,” the system should be capable of identifying associations between visual attributes—such as “knee-length” or “solid color design”—and the commuting context conveyed by the textual input, thereby delivering recommendations that are more aligned with the user’s needs. This capability not only reduces decision-making time but also significantly increases product click-through and conversion rates. From a technical perspective, optimization strategies designed specifically for e-commerce environments can bridge the gap left by general-purpose retrieval models in vertical domains. The proposed modality alignment approach is expected to serve as a valuable reference for cross-modal applications in other specialized fields. At the industry level, the adoption of this technology is expected to facilitate the transformation of e-commerce platforms from extensive, exposure-oriented recommendation models toward refined, precision-driven operational paradigms.

Although a range of methods has been developed within the field of cross-modal retrieval, significant limitations remain when applied to e-commerce recommendation scenarios. In the area of visual feature extraction, mainstream approaches based on Convolutional Neural Networks (CNNs) [12-14] primarily focus on global features, often overlooking discriminative local attributes. This deficiency leads to misidentification of products with similar overall styles but distinct local details, resulting in high matching scores for visually incongruent items. In textual feature extraction, methods based on word embeddings [15-17] frequently remain constrained to literal semantic matching and are unable to interpret implicit semantics—such as associating "slimming" with "waist-cinching design" or linking "breathable" with "mesh material." Furthermore, when abstract expressions such as "versatile" or "cost-effective" are present in product descriptions, feature representation tends to become semantically ambiguous. In the modality fusion stage, most existing approaches [18-20] simply concatenate or apply weighted summation to visual and textual features, failing to establish precise one-to-one mappings between local image features and key textual tokens. For example, the visual detail of an "embroidered collar" cannot be accurately aligned with the textual concept of "delicate embellishment," resulting in reduced robustness of cross-modal matching.

To address these challenges, a cross-modal product image retrieval method was proposed for e-commerce recommendation. This method comprises three core innovations. First, a global-local visual feature extraction module was designed, which incorporates product category labels as supervisory signals. This enables the model to learn global product contours while automatically attending to locally distinctive regions such as collars and cuffs, thereby enhancing the discriminative power of visual features. Second, a textual feature extraction module employing attention mechanisms was introduced to semantically filter product descriptions. General-purpose expressions such as "high quality" are downweighted, while core semantic elements such as "cotton fabric" and "wrinkle-resistant treatment" are reinforced, enhancing the specificity of the textual representation. Third, a bidirectional mapping mechanism was established through a cross-extraction module that integrates visual-text semantic information, enabling the dynamic association between visual local features and textual keywords, thereby allowing for precise alignment of deep semantic representations across modalities. This approach not only significantly improves the accuracy of cross-modal product retrieval and offers a practical solution for e-commerce recommendation systems but also contributes reusable design paradigms—such as global-local collaboration and semantic focus—to the broader field of cross-modal learning. As a result, it facilitates a shift from general-purpose adaptation to task-specific optimization in the application of deep learning within vertical domains.

2. CROSS-MODAL PRODUCT IMAGE RETRIEVAL METHOD FOR E-COMMERCE RECOMMENDATION

In the domain of natural images, cross-modal text-image retrieval has been advanced through techniques such as self-attention, enabling long-range semantic association and key information extraction. However, essential distinctions exist between product and natural images. Product images are required to emphasize "purchasability features," such as fabric texture in apparel or functional interfaces in home appliances, and are often embedded in complex backgrounds. Traditional approaches struggle to isolate core product features from such redundant visual information. Simultaneously, textual descriptions in e-commerce incorporate both marketing-driven and function-oriented expressions. For instance, phrases like "slimming dress" convey not only an effect-based descriptor but also implicit category information. Existing models are easily distracted by non-informative tokens such as "bestseller" or "hot item" and fail to focus on purchase-critical attributes such as "waist-cinching" or "chiffon." More critically, the primary objective of e-commerce recommendation lies in facilitating purchase decisions, which necessitates accurate mapping between latent user needs and product characteristics. Most existing cross-modal models are designed for general-purpose retrieval tasks and have not been optimized for the demand-attribute alignment required in recommendation scenarios. This underscores the necessity for task-specific approaches tailored to e-commerce recommendation systems.

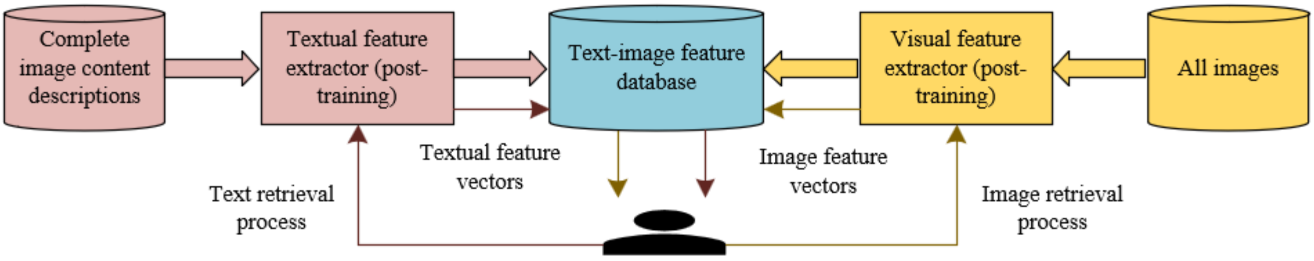
Complete image content descriptions
Textual feature extractor (post-training)
Text-image feature database
Visual feature extractor (post-training)
All images
Textual feature vectors
Image feature vectors
Text retrieval process
Image retrieval process

Figure 1. Conceptual framework of the cross-modal product image retrieval method for e-commerce recommendation

To address the specific demands of e-commerce recommendation, a dedicated cross-modal product image retrieval method was proposed. This approach is guided by the principle of "adapting to product-specific characteristics and serving recommendation objectives," and a full-chain optimization framework was constructed based on self-attention mechanisms, as illustrated in Figure 1. In the visual feature extraction stage, product images were divided into sub-image patches, and self-attention mechanisms were introduced. Attention weights were directed—under the guidance of class-supervised signals—toward discriminative regions such as collars and cuffs, while positional encoding was applied to preserve the structural relationships within the product. This design facilitates the differentiation of visually similar products that share the same style but differ in fine-grained details, a challenge that frequently arises in e-commerce scenarios.

In the textual feature extraction stage, semantic masking was combined with self-attention mechanisms. Non-essential expressions such as "recommended" or "bestseller" were first masked. Subsequently, semantic associations—such as those between "slim fit" and "tailoring"—were captured through self-attention, while representations of attributes that strongly influence purchasing decisions, such as "fabric" and "size," were reinforced. This process ensures that the extracted textual features remain closely aligned with the user's actual intent. During the cross-modal fusion stage, a bidirectional mapping between "visual sub-blocks" and "textual keywords" was constructed using self-attention-based mechanisms. For instance, when the query "durable running shoes" is entered, the model is designed to increase the attention weight linking the visual sub-block representing a "rubber sole" to the keyword "durable." This enables accurate alignment between product characteristics and user demands. As a result, the recommendation system is supported by a three-tiered matching rationale comprising image detail, textual demand, and purchase intent. This architecture ensures that the recommendations not only correspond to the product's appearance but also align with the user's practical usage scenarios. The overall model architecture for the proposed cross-modal product image retrieval method in e-commerce recommendation is presented in Figure 2.

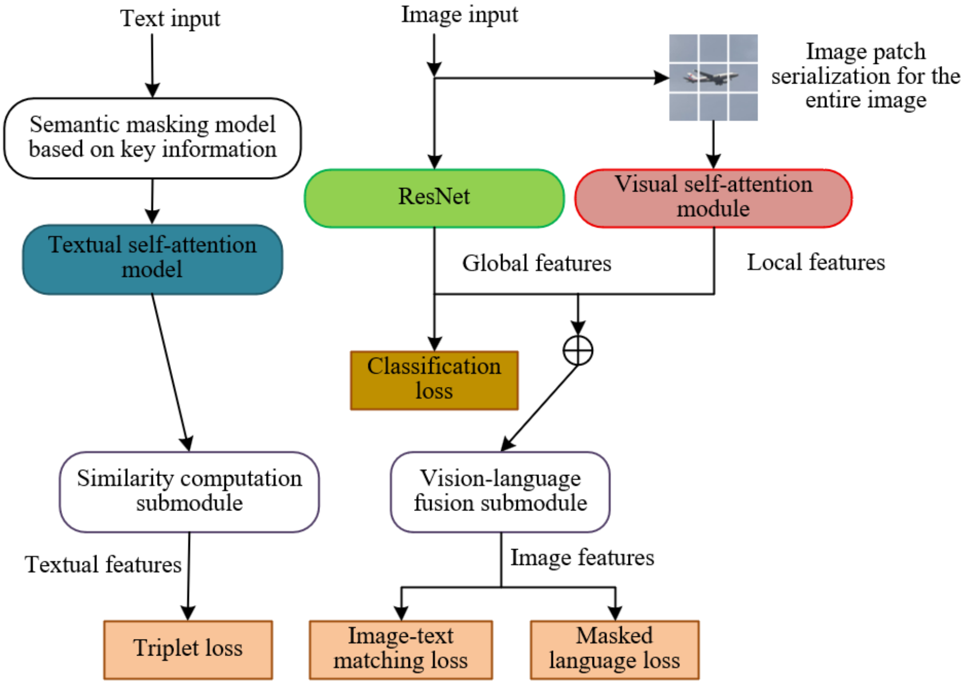
Text input
Image input
Image patch serialization for the entire image
Semantic masking model based on key information
ResNet
Visual self-attention module
Textual self-attention model
Global features
Local features
Classification loss
Similarity computation submodule
Vision-language fusion submodule
Textual features
Image features
Triplet loss
Image-text matching loss
Masked language loss

Figure 2. Model architecture of the cross-modal product image retrieval method for e-commerce recommendation

2.1 Global-local visual feature extraction module

The global-local visual feature extraction module, guided by class supervision, was designed to address the core requirement of accurately interpreting visual product attributes in e-commerce recommendation scenarios. In such contexts, visual features extracted from product images are directly associated with user purchasing decisions. A shirt may be selected based on a local detail such as "embroidered cuffs," while a pair of trousers may be dismissed due to a global characteristic like a "loose overall fit." However, due to the variability of shooting angles, small targets such as buttons and zippers, as well as densely distributed targets such as printed patterns, are often blurred in product images. Moreover, cluttered backgrounds introduced by accessories or props frequently interfere with the extraction of core product features. Product category information serves as a critical anchor for interpreting such features. For example, "dress" images should prompt attention to the neckline rather than the sole. Therefore, the proposed module was designed to jointly enable global orientation, local detail focus, and category-guided feature prioritization, providing a robust visual representation tailored to the requirements of cross-modal matching in e-commerce recommendation systems.

The global feature extraction module was built upon the Residual Neural Network (ResNet) and was optimized using class-supervised signals to enhance the detection of salient category-specific features, which is essential in scenarios where product categories strongly influence purchasing intent. The architecture incorporates interactive connections between shallow and deep features: shallow features preserve basic information such as edges and color, while deep features capture abstract semantics. This interaction mitigates the limitations of single features. Class supervision was introduced by converting product category labels into supervisory signals that guide the feature extraction process. This encourages the model to focus on category-relevant core attributes. For instance, in the case of the "sneakers" category, increased attention is allocated to the "sole pattern" and "lacing system," whereas unrelated elements such as "floor texture" are downweighted. As shown in the formulation below, the input image is transformed by the ResNet architecture to produce the global feature representation D_H. The class supervision signal adjusts the model parameters ϕ through a loss function, ensuring that D_H includes not only the global contour but also category-relevant core features, providing support for recognizing functional attributes of interest when users query terms like "sneakers." The feature extraction process using the ResNet model is denoted as L_{RES}, and the corresponding expression is:

$$D_H = L_{RES}\left(U;\varphi\right) \tag{1}$$

The local feature extraction module was constructed based on the Vision Transformer (ViT) architecture to address the challenge of representing small and densely packed visual targets in product images—an essential factor in e-commerce, where "fine details determine purchase decisions." In such images, local features often constitute critical decision-making elements. However, conventional convolutional methods are prone to detail loss due to the limited receptive field. ViT addresses this limitation by segmenting an image into $O\times O$ patches and transforming them into sequential inputs. Through the multi-head attention mechanism, long-range dependencies across image patches can be captured. For instance, the spatial relationship between the "left chest pocket" and "right sleeve logo" on a shirt can be simultaneously modeled. Additionally, multi-layer perceptrons (MLPs) and layer normalization enhance the stability of fine-grained representations. To mitigate the effects of variable camera angles, ViT's sequence-based modeling inherently ignores absolute position interference and instead emphasizes relative spatial relationships among patches. This enables the precise extraction of fine-grained features of small and densely distributed targets, thereby supporting the recommendation system in aligning with users' attention to fine detail. Denoting the multi-head attention layer as *MSA*, the MLP as *MLP*, and layer normalization as *LN*, the internal processing process of ViT is formulated as follows:

$$\begin{gathered} d_0 = \left[a_{CLS}; a_1 A; \ldots; a_v A\right] + A^{POS} \\ A \in \Re^{\left(O^2 \cdot Z\right)\times G}, A^{POS} \in \Re^{(V+1)\times G} \end{gathered} \tag{2}$$

$$d_f^{'} = MSA\left(LN\left(d_{f-1}\right)\right) + d_{f-1}, f = 1 \ldots F \tag{3}$$

$$d_f = MLP\left(LN\left(d_f^{'}\right)\right) + d_f^{'}, f = 1 \ldots F \tag{4}$$

$$D_M = LN\left(d_F^0\right) \tag{5}$$

The integration of global and local features represents a critical stage in adapting the module output to e-commerce recommendation tasks. This fusion enables both "macro-level category orientation" and "micro-level detail support." The global feature representation D_H carries salient, category-dominant features, while the local feature representation D_M encodes fine-grained visual details. During integration at the terminal stage of the visual pipeline, these features are not simply concatenated; rather, feature fusion is performed through weighted allocation, thereby emphasizing attributes that are most relevant to user intent. For instance, when the query is "warm down jacket," the weight of "loftiness" in the global feature is increased. Conversely, when durability is the focus—as in "durable down jacket"—the weight of "stitching quality" in the local feature is enhanced. The resulting fused representation retains both category-level attributes and detail-level distinctions, providing a rich and accurate visual semantic foundation for cross-modal alignment. This ultimately enables the recommendation system to infer not only the target product category but also the specific product details that users value. The integration of D_H and D_M is defined by the following expression:

$$D_N = D_H \oplus D_M \tag{6}$$

2.2 Textual feature extraction module

Textual descriptions in e-commerce are characterized by low semantic density and high informational focus. These descriptions typically consist of product category terms, key attribute phrases, and marketing-related vocabulary. Semantically relevant information—particularly that which strongly influences purchasing decisions—is concentrated in category and attribute terms, resulting in an overall semantic complexity significantly lower than that of natural language text. To accommodate these characteristics, a textual feature extraction module based on dominant semantic masking was designed. This module enables precise semantic parsing of product descriptions in e-commerce recommendation scenarios. The first six layers of the Bidirectional Encoder Representations from Transformers (BERT) model were employed to construct the core of the textual feature extraction module. These layers, comprising multi-head attention, MLPs, and layer normalization components, are sufficient for capturing shallow semantic associations such as "material-functionality" while avoiding the redundant computation associated with deeper network layers. This structure aligns with the computational efficiency requirements of real-time recommendation systems. The final six layers of BERT were reserved for subsequent cross-modal fusion, ensuring that textual features can interact deeply with visual representations. This design lays the foundation for accurate alignment between "product image details" and "textual attributes." A representative example of dominant semantic masking is illustrated in Figure 3. Specifically, a product description *S* is first tokenized and encoded into a sequence of word and positional vectors. Let the input sequence length be denoted by *v*, and let the dimensionality of each word vector be represented by *f*. The resulting representations can be formalized as:

$$S = \left[n_1, n_2, \ldots, n_v\right] \in E^{v\times f} \tag{7}$$

$$D_S = L_{BERT}\left(S\right) \tag{8}$$

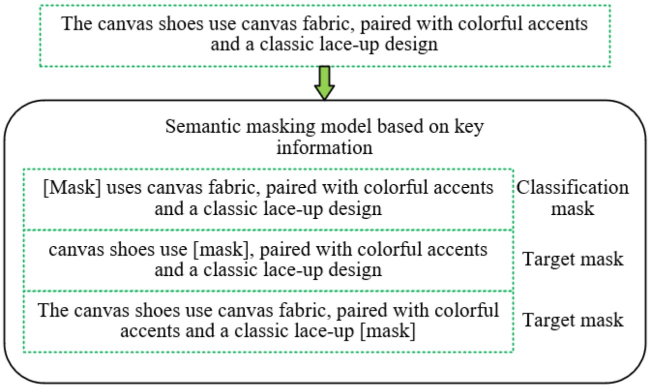
The canvas shoes use canvas fabric, paired with colorful accents and a classic lace-up design
Semantic masking model based on key information
[Mask] uses canvas fabric, paired with colorful accents and a classic lace-up design
Classification mask
canvas shoes use [mask], paired with colorful accents and a classic lace-up design
Target mask
The canvas shoes use canvas fabric, paired with colorful accents and a classic lace-up [mask]
Target mask

Figure 3. Example of dominant semantic masking

This module enhances the extraction of core semantics in e-commerce texts through a combined strategy of targeted masking and random masking, thereby addressing issues commonly observed in traditional textual feature extraction—namely, interference from marketing terms and the dilution of critical semantic content. In e-commerce descriptions, category terms such as "sneakers" and attribute terms such as "non-slip sole" or "breathable mesh" serve as essential semantic bridges connecting product features to user needs. In contrast, marketing expressions such as "hot sale" or "limited time" constitute non-essential information. The module first applies random masking to 15% of the input tokens, which preserves the model's ability to comprehend overall textual semantics. On this basis, category terms and key attribute terms are forcibly tagged and masked with higher probability or subjected to random replacement. For example, in the phrase "non-slip sneakers, limited-time discount," the terms "non-slip" and "sneakers" are preferentially masked to compel the model to strengthen its semantic encoding of these critical elements during training. This approach not only prevents the model from overfitting to non-informative marketing content but also leverages a "deletion-prediction" learning mechanism to deepen the representation of key semantics. As a result, the extracted textual features are more accurately aligned with genuine user intent.

Collaborative representation between the textual and visual feature extraction modules was achieved through the reuse of class supervision information, thereby enhancing consistency in cross-modal matching within e-commerce recommendation settings. The visual feature extraction module is guided to emphasize category-relevant visual features through class supervision. Correspondingly, the textual feature extraction module employs targeted masking of category terms to direct semantic alignment toward the same category space. For instance, when the visual module emphasizes neckline designs in the "dress" category, the masking of terms such as "dress" and "V-neck" in the textual module increases their semantic attention weights, facilitating alignment between textual and visual features at both the category and detail levels. This coordinated mechanism effectively mitigates the issue of cross-modal matching deviation among items within the same category—a common challenge in e-commerce environments. Meanwhile, the residual robustness introduced by random masking enables the model to tolerate irregularities in user-generated descriptions. Even when input text is irregular or imprecise, core semantic elements can still be accurately extracted, thus providing the recommendation system with a reliable link between textual intent and visual features.

2.3 Cross-extraction module

The cross-extraction module, which integrates visual and textual semantic information, was designed to address the core challenge of visual and textual semantic misalignment in e-commerce scenarios. A feature calibration architecture, constructed from the final six layers of the BERT model, was established and kept independent from the inference pipeline, bridging the correlation gap between visual and textual features in the context of e-commerce recommendation. Although visual and textual features may reference the same product attribute, independently extracted features are prone to a mismatch of "seemingly unrelated but semantically related" due to modal differences. For example, the visual representation of a "metallic zipper's gloss" and the textual description "durable hardware" may be intrinsically related but appear distant in the feature space, leading to false mismatches. This module focuses solely on semantic calibration and does not participate in the final retrieval inference. As such, it avoids interference with the primary inference path while leveraging the deep semantic modeling capability of BERT's final six layers to uncover latent correlations between "visual details" and "textual attributes," thereby providing foundational support for feature alignment in the recommendation system.

The cross-extraction module comprises two submodules. The similarity computation submodule L_{SIM} is responsible for precisely measuring the correlation between visual and textual features, thereby reinforcing the e-commerce-specific logic of "intra-category aggregation and inter-category separation" in the feature distribution. This submodule employs vector similarity measures, such as cosine distance, to directly compare visual features D_N with textual features D_S. For instance, for a query like "non-slip running shoes," the cosine similarity is computed between the visual feature representing the "rubber outsole texture" and the textual feature corresponding to "non-slip," ensuring that highly correlated features are located closer in the feature space. In addition, a triplet loss function was introduced to optimize the parameters. A triplet structure—consisting of a target product, a similar product from the same category, and a product from a different category—was constructed. The loss function minimizes the distance between the target and the similar product while maximizing the distance from the dissimilar product. This encourages the feature space to satisfy the dual demands of distinguishing categories and differentiating styles typical in e-commerce scenarios. As a result, the risk of erroneously recommending, for example, "leather shoes" when a user searches for "running shoes" is effectively minimized.

The vision-language fusion submodule L_{MUL} employs a dual-loss optimization strategy to enhance semantic interaction between modal features in e-commerce settings, addressing the challenge of achieving precise alignment between visual details and textual attributes. The masked language loss extends the dominant semantic masking strategy introduced in the textual module. Core attribute terms—such as "linen" or "non-slip"—are intentionally masked in the text, and the model is required to predict the masked tokens by leveraging visual features. For instance, when the textual token "linen" is masked, the model is prompted to infer this label based on visual cues such as the "loose fiber texture" extracted from the image, enhancing the mapping between visual fine-grained features and textual product attributes. In parallel, the image-text matching loss focuses on the compatibility judgment of product image-text pairs. For positive samples, such as a floral dress image paired with the phrase "floral long dress," the model is trained to produce a high probability of correspondence. For negative samples, such as pairing the same image with "solid short skirt," the model is optimized to yield a low correspondence score, ensuring that the system can distinguish whether an image and a text description refer to the same product. Through joint optimization of these two losses, visual features include not only what the product looks like but also which textual attributes it corresponds to. Simultaneously, textual features convey not only what attributes are described but also how they visually manifest. This ensures that the final recommendation aligns both with the product's physical appearance and with the user's expressed intent through textual queries. The cross-extraction module can be formally expressed as:

$$L_{CROSS} = \{L_{SIM}, L_{MUL}\} \tag{9}$$

2.4 Multi-objective loss function optimization

The multi-objective loss function optimization was designed with the core goal of achieving precise cross-modal feature alignment in e-commerce recommendation scenarios. Through the joint optimization of four distinct loss functions, model parameters were constrained from multiple dimensions, addressing complex optimization requirements that cannot be adequately resolved by a single loss function. In e-commerce settings, the alignment between product images and textual descriptions must extend beyond surface-level "appearance-to-description" correspondence to include deeper "function-to-demand" associations. A single loss function often optimizes only one dimension of model performance, whereas a multi-objective loss design enables comprehensive improvements through the combined application of the triplet loss $LOSS_{SY}$ (enhancing feature discriminability), the class-supervised cross-entropy loss $LOSS_{JCS}$ (improving category-specific feature salience), the masked language loss $LOSS_{YM}$ (strengthening semantic association), and the image-text matching loss $LOSS_{PP}$ (optimizing matching accuracy). Through their integration, a comprehensive objective can be achieved—ensuring retrieval accuracy, category clarity, semantic alignment, and robust matching performance—thereby meeting the multidimensional requirements of cross-modal matching in e-commerce recommendation systems.

The triplet loss serves as the core loss function, designed to improve cross-modal feature discriminability through bidirectional retrieval constraints. This approach supports the nuanced distinction of visually or semantically similar products—a frequent necessity in e-commerce environments. The optimization is carried out in two directions: When retrieving text based on image features, the feature distance between the query image and non-matching texts is maximized, while the distance to the corresponding text is minimized. Conversely, when retrieving images based on textual features, the distance to mismatched images is increased, and the distance to relevant images is reduced. This bidirectional constraint addresses the challenge of distinguishing between products with similar overall appearances but different attributes. For instance, when the user query is "loose-fit jeans," the model is guided by the triplet loss to strengthen the alignment between the visual representation of a "loose silhouette" and the corresponding textual description, while increasing the distance to the feature representation of "skinny jeans." This prevents the recommendation of visually similar but functionally mismatched products, thereby ensuring that retrieval results directly reflect user intent. Let the margin parameter that enlarges the separation between positive and negative pairs be denoted by β. Let $COS()$ represent the similarity score obtained from the cross-modal similarity computation module. The negative text corresponding to an image is represented as S', and the negative image corresponding to a text S is denoted as U'. The formal expression is:

$$\begin{aligned} LOSS_{SY} &= \sum_{S'}\left[\beta - COS(U,S) + COS(U,S')\right] + \\ &\sum_{U'}\left[\beta - COS(U,S) + COS(U',S)\right] \end{aligned} \tag{10}$$

The class-supervised cross-entropy loss $LOSS_{JCS}$ and the masked language loss $LOSS_{YM}$ were jointly designed to enhance the semantic association between category and attribute from the visual and textual modalities, respectively. This design aligns with user behavior patterns observed in e-commerce, wherein category guides decision-making and attribute determines selection. The class-supervised cross-entropy loss introduces category labels to guide the global visual feature extraction module, encouraging the generation of discriminative category-specific representations. For instance, in the case of the "sneakers" category, the loss function constrains the model to emphasize core visual features such as "sole pattern" and "collar height," ensuring that the resulting visual representation satisfies the foundational requirement of correct category identification for downstream recommendation. $LOSS_{YM}$ optimizes the prediction of masked core attribute terms within e-commerce product texts using image features. For example, if an image exhibits "linen-like texture," the loss function compels the model to accurately predict the masked token "linen," thereby minimizing prediction error and reinforcing the correspondence between local visual features and textual attribute terms. This mechanism ensures that textual representations more accurately reflect the actual characteristics of the product, addressing common disjunctions between image and text in product listings. Let b_u denote the ground-truth category of an image, and let o_u represent the predicted category probability distribution output by the model. Then the expression is as follows:

$$LOSS_{JCS} = -\sum_{u=1}^{V} b_u \log(o_u) + (1-b_u)\log(1-o_u) \tag{11}$$

Let the masked textual vector be denoted by S^, and let $o^{MA}(U,S\text{^})$ represent the predicted probability distribution of the masked text conditioned on the visual features. Then the expression is as follows:

$$LOSS_{YM} = R_{(U,S\wedge)\sim F}\, G\left(b^{MA}, o^{MA}, (U, S\wedge)\right) \tag{12}$$

The image-text matching loss $LOSS_{PP}$ was designed to optimize the overall compatibility of image-text pairs through a binary classification framework. This approach enhances the model's ability to discern the consistency between marketing descriptions and actual product visuals, a frequent concern in e-commerce platforms where image-text mismatches are common. Under this formulation, image-text matching is treated as a binary classification task. A fully connected layer was employed to output a matching probability, which was then constrained using cross-entropy loss. When an image and its corresponding text description are semantically aligned, the model is encouraged to assign a matching probability close to 1. Conversely, when the image-text pair is contradictory, the probability is pushed toward 0. This optimization enables the model to autonomously filter out "misleading promotional content" and ensures that recommendations are based solely on authentically aligned features. Moreover, $LOSS_{PP}$ complements $LOSS_{SY}$. While the former optimizes overall compatibility, the latter refines inter-feature distances. The combination of these two losses guarantees both the correctness of individual image-text matches and the discriminative power across different pairs, thereby achieving a retrieval performance that is both accurate and comprehensive—a critical requirement for effective e-commerce recommendation. The image-text matching loss generates the matching probability of an image-text pair, denoted by $o^{MT}(U,S)$, by introducing a fully connected layer. The overall optimization objective remains the minimization of cross-entropy loss, formally expressed as:

$$LOSS_{PP} = R_{(U,S)\sim F} G\left(b^{MA}, o^{MA}, (U,S)\right) \tag{13}$$

2.5 Algorithmic workflow

The algorithm training workflow was designed with the core objective of adapting to the unique characteristics of e-commerce data while strengthening cross-modal collaboration. Model performance was improved through a phased optimization process. The initial stage involved data preprocessing tailored to the properties of e-commerce data. For image data, background removal, angle normalization, and illumination enhancement were performed to ensure that visual input remains focused on the product itself. For textual data, non-essential marketing expressions such as “bestseller” or “hot sale” were filtered out. Only core elements—typically consisting of category and attribute terms—were retained. Synonym replacement was applied to enhance dataset diversity. Following preprocessing, the refined dataset entered the iterative training phase. Product images were processed through the visual feature extraction module to generate global features D_H and local features D_M, which were then fused into the comprehensive visual representation D_N. Simultaneously, the textual feature extraction module encoded the textual input into the representation D_S. During training, four loss functions operated in coordination to guide the model. The triplet loss ensured that products with the same style and attributes are positioned closer in the shared feature space. The class-supervised cross-entropy loss reinforced the salience of category-specific features. The masked language loss promoted fine-grained image-text alignment through the masked attribute terms. The image-text matching loss enabled the model to filter mismatched image-text pairs. The cross-extraction module indirectly optimized the single-modality modules. Through iterative learning, model parameters converged to an optimal configuration that enabled precise recognition of correlations between product characteristics and user intent in e-commerce settings. Figure 4 illustrates the conceptual architecture of the cross-modal product image retrieval model for e-commerce recommendation, while the complete training workflow is depicted in Figure 5.

The inference process was designed to satisfy the dual demands of real-time responsiveness and high-precision retrieval in e-commerce recommendation scenarios. To this end, a feature retrieval database tailored to e-commerce products was constructed in advance. Specifically, the pre-trained visual module was employed to generate *FV* using the product images, while the corresponding product description was processed by the textual module to generate D_S. All feature vectors were stored in association with their respective product IDs. Vector quantization techniques were applied to compress the feature dimensionality, thereby achieving a balance between storage cost and retrieval speed. This design was implemented to address the large-scale retrieval demands associated with e-commerce platforms containing tens of millions of products, while ensuring that feature comparison latency remains constrained within the millisecond range. Upon user query initiation, the inference pipeline was activated. The input image generated D_N using the visual module, while the textual input generated D_S via the textual module. The similarity computation module then compared the query features with the corresponding modal features stored in the database. The top *V* most similar candidates were selected based on the cosine distances. In parallel, e-commerce recommendation logic was incorporated to prioritize the retrieval of products with high sales volume and favorable user ratings, thereby ensuring that the final results not only satisfy feature-level matching but also align with user purchasing preferences.

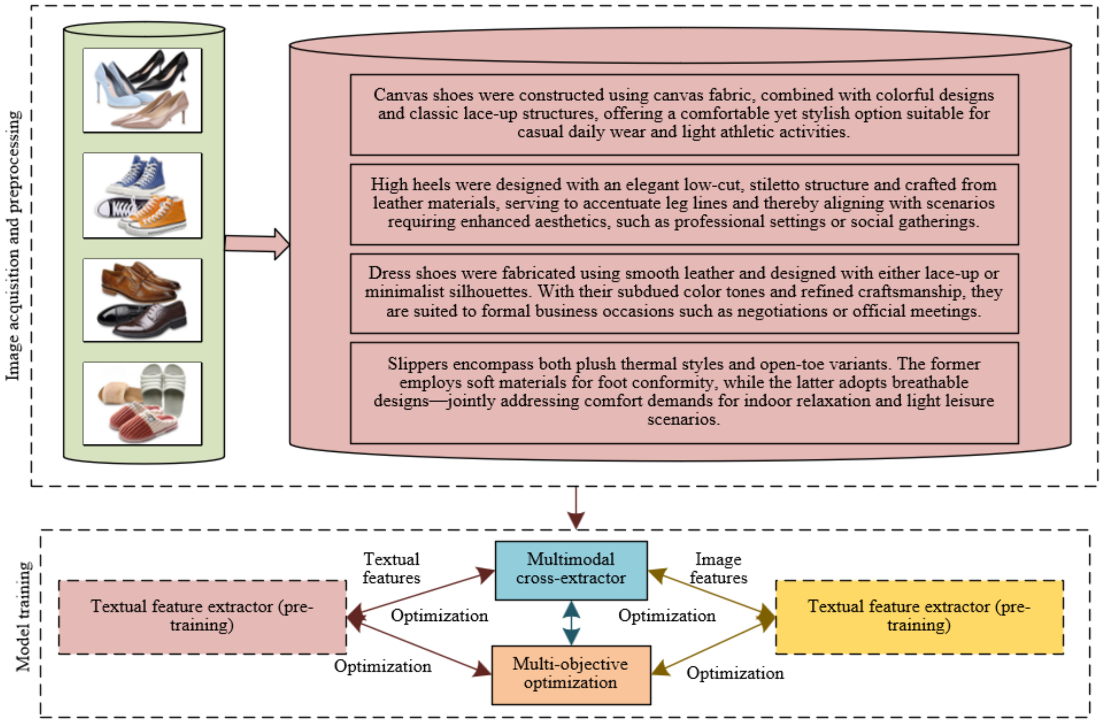
Image acquisition and preprocessing
Model training
Textual features
Multimodal cross-extractor
Image features
Textual feature extractor (pre-training)
Optimization
Optimization
Textual feature extractor (pre-training)
Optimization
Multi-objective optimization
Optimization

Canvas shoes were constructed using canvas fabric, combined with colorful designs and classic lace-up structures, offering a comfortable yet stylish option suitable for casual daily wear and light athletic activities.

High heels were designed with an elegant low-cut, stiletto structure and crafted from leather materials, serving to accentuate leg lines and thereby aligning with scenarios requiring enhanced aesthetics, such as professional settings or social gatherings.

Dress shoes were fabricated using smooth leather and designed with either lace-up or minimalist silhouettes. With their subdued color tones and refined craftsmanship, they are suited to formal business occasions such as negotiations or official meetings.

Slippers encompass both plush thermal styles and open-toe variants. The former employs soft materials for foot conformity, while the latter adopts breathable designs—jointly addressing comfort demands for indoor relaxation and light leisure scenarios.

Figure 4. Conceptual architecture of the proposed cross-modal product image retrieval model for e-commerce recommendation

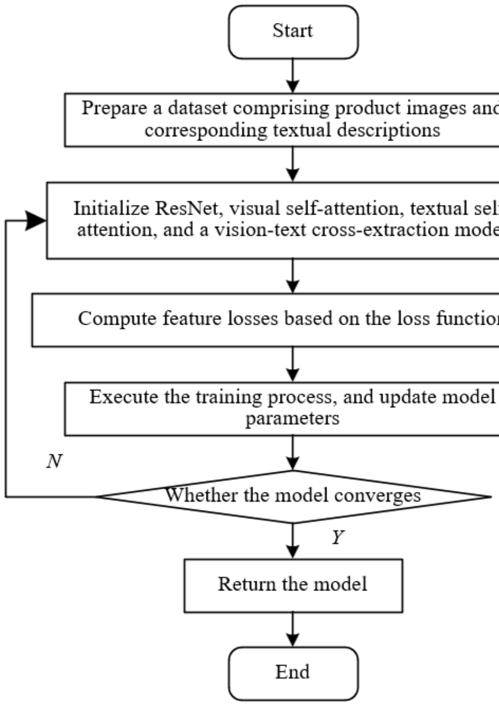
Start
Prepare a dataset comprising product images and corresponding textual descriptions
Initialize ResNet, visual self-attention, textual self-attention, and a vision-text cross-extraction model
Compute feature losses based on the loss function
Execute the training process, and update model parameters
N
Whether the model converges
Y
Return the model
End

Figure 5. Training workflow of the proposed cross-modal product image retrieval model for e-commerce recommendation

3. EXPERIMENTAL RESULTS AND ANALYSIS

As shown in Table 1, the proposed approach achieved a systematic breakthrough in bidirectional cross-modal retrieval. In the image-to-text retrieval task, Recall@1, Recall@5, and Recall@10 reached 12.69%, 32.69%, and 45.26%, respectively, corresponding to relative gains of 12.99%, 18.62%, and 20.51% over the baseline model Contrastive Language-Image Pretraining (CLIP) (11.23%, 27.56%, and 37.56%). In the text-to-image retrieval task, Recall@1, Recall@5, and Recall@10 reached 11.25%, 42.36%, and 61.59%, reflecting relative improvements of 13.98%, 48.58%, and 48.12% compared to CLIP's 9.87%, 28.51%, and 41.58%. The mean recall (mR) attained 33.34%, exceeding that of the Foundational Language And Vision Alignment (FLAVA) (25.62%) and the Visual Bidirectional Encoder Representations from Transformers (VisualBERT) (28.98%) by over four percentage points. These substantial performance gains were attributed to the synergistic integration of the global-local visual module and the textual module. The visual module, supervised by product category labels, captured not only the overall contours of items such as "shirts" and "dresses" but also the category-specific local features, including neckline cuts and cuff textures, thereby enhancing the fine-grained discriminative power of the visual features. Meanwhile, the textual module, guided by attention mechanisms, was able to filter out generic phrases such as "high quality" and emphasize core semantic descriptors such as "cotton fabric" and "lace stitching." This enabled the formation of precise mappings between textual and visual local features, resulting in a substantial bidirectional performance enhancement in fashion product cross-modal matching.

As shown in Table 2, further evaluation on the WAB dataset verified the model's adaptability to e-commerce scenarios. In the image-to-text retrieval task, Recall@1, Recall@5, and Recall@10 reached 5.24%, 14.59%, and 23.56%, respectively, representing improvements of 23.00%, 29.92%, and 44.10% over CLIP's 4.26%, 11.23%, and 16.35%. In the text-to-image retrieval task, Recall@1, Recall@5, and Recall@10 reached 4.65%, 21.36%, and 34.56%, outperforming CLIP's 3.89%, 15.23%, and 25.36% with relative gains of 19.54%, 40.25%, and 36.28%, respectively. The resulting mR of 16.35% exceeded that of the Align Before Fuse (ALBEF) (13.54%) and FLAVA (13.56%). These improvements were facilitated by the visual-textual cross-extraction module, which established a bidirectional mapping mechanism capable of dynamically aligning visual local features with textual keywords. This mechanism demonstrated robustness even under image blurriness or colloquial textual expressions, enabling precise cross-modal feature alignment. Experimental findings confirmed that the collaborative integration of three modules effectively addressed the fine-grained retrieval challenges in fashion product domains while adapting to the complex nature of livestreaming e-commerce, thereby offering enhanced cross-modal retrieval capabilities for recommendation systems in e-commerce environments.

Table 1. Comparative results on the Fashion-Gen dataset

Algorithm	**Image-to-Text Retrieval**			**Text-to-Image Retrieval**			***mR***
	***R@*1**	***R@*5**	***R@*10**	***R@*1**	***R@*5**	***R@*10**	
CLIP	11.23	27.56	37.56	11.23	27.56	42.56	25.36
ALBEF	12.36	24.59	38.52	9.78	28.51	41.58	25.42
FLAVA	12.35	25.36	37.51	8.12	26.35	43.23	25.62
ViLT	11.52	26.31	35.62	9.87	31.25	44.26	25.98
VisualBERT	11.54	23.54	42.65	12.36	33.56	53.62	28.32
Proposed approach	12.69	32.69	45.26	11.25	42.36	61.59	33.34

Table 2. Comparative results on the WAB dataset

Algorithm	**Image-to-Text Retrieval**			**Text-to-Image Retrieval**			***mR***
	***R@*1**	***R@*5**	***R@*10**	***R@*1**	***R@*5**	***R@*10**	
CLIP	4.26	11.25	16.35	3.89	15.23	25.36	12.36
ALBEF	5.78	11.23	18.52	3.65	15.62	25.64	13.54
FLAVA	5.23	11.58	22.34	4.23	14.23	26.68	13.56
ViLT	5.12	11.36	18.56	4.89	16.98	28.52	13.58
VisualBERT	5.36	14.23	22.36	4.87	17.56	32.31	15.32
Proposed approach	5.24	14.59	23.56	4.65	21.36	34.56	16.35

Table 3. Ablation study on the Fashion-Gen dataset

Algorithm	Image-to-Text Retrieval			Text-to-Image Retrieval			*mR*
	R@1	***R@5***	***R@10***	***R@1***	***R@5***	***R@10***	
Global-local visual feature extraction	12.36	28.56	42.56	11.23	35.62	57.56	32.65
Textual feature extraction	12.58	32.42	46.59	11.58	41.25	61.23	32.58
Cross-extraction	12.36	32.61	45.21	11.23	42.69	61.24	33.64

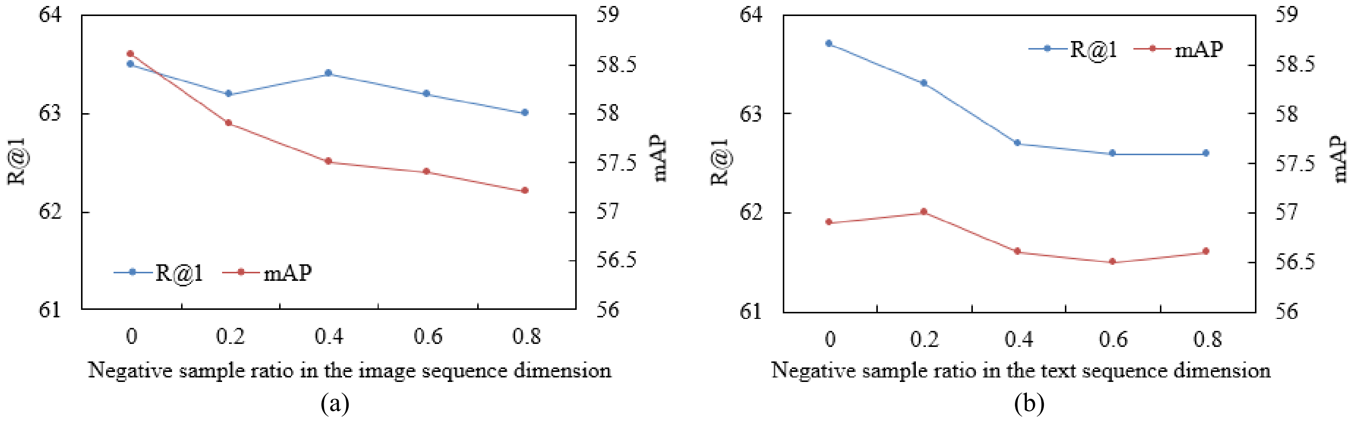

Figure 6. Impact of positive-to-negative sample ratio on retrieval performance on the Fashion-Gen dataset

As shown in Table 3, the results of the ablation study conducted on the Fashion-Gen dataset clearly delineated the individual contributions and synergistic potential of the three core modules. The global-local visual feature extraction module, guided by product category label supervision, enabled the model to effectively capture both the overall silhouette and fine-grained local attributes of products in the image-to-text retrieval task. Recall@1, Recall@5, and Recall@10 reached 12.36%, 28.56%, and 42.56%, respectively. These results demonstrated the module's capacity to focus on category-discriminative regions such as neckline contours and cuff textures, thereby enhancing the granularity and specificity of visual representations and supporting accurate text retrieval based on visual input. The textual feature extraction module filtered out generic expressions such as "high quality" through the application of attention mechanisms and emphasized semantically rich descriptors like "pure cotton fabric" and "lace stitching." This led to superior performance in the text-to-image retrieval task, where Recall@5 and Recall@10 reached 41.25% and 61.23%, representing improvements of 5.63 and 3.67 percentage points, respectively, over the visual module alone. These findings underscored the critical role of precise textual semantics in guiding visual retrieval and mitigating the semantic vagueness often encountered in e-commerce product descriptions. The cross-extraction module yielded the best overall performance, achieving Recall@5 of 42.69% and Recall@10 of 61.24% in the text-to-image retrieval task, along with the highest mR of 33.64%, because its bidirectional visual-textual mapping mechanism further broke through modal barriers. These results show that the module dynamically linked visual local features with key textual tokens, effectively addressing the issue of "misaligned representation" across modalities, thereby enhancing retrieval accuracy and robustness.

Figure 6 illustrates the model's robustness under complex noise conditions commonly encountered in e-commerce scenarios, examined along two dimensions: the negative sample ratios in the image and text sequences. In Figure 6(a), as the proportion of non-matching product images increased from 0 to 0.8, the Recall@1 metric remained stable within the 63-63.5% range. This resilience was attributed to the fine-grained discriminative capacity provided by the global-local visual feature extraction module. Supervised by product category labels, this module not only captured holistic contours of products but also precisely identified category-distinctive local features, including "neckline cut" and "cuff texture." As a result, the visual representations demonstrated strong robustness to interference. Even in the presence of a high volume of non-matching images in the retrieval pool, the target product could still be efficiently identified through its fine-grained visual features. Despite this, the mean Average Precision (mAP) exhibited a steady decline from 63.5% to 62.2%, indicating that an increased proportion of negative samples challenged the model's ranking precision. Nevertheless, by combining with the semantic filtering capabilities of the textual feature extraction module, generic expressions (e.g., "high quality") were suppressed and salient attribute terms (e.g., "pure cotton fabric") were amplified, enabling the model to remain stable in Recall@1, thereby demonstrating that a "coarse filtering-fine ranking" anti-noise mechanism had been established for the visual and textual features. In Figure 6(b), as the proportion of mismatched product descriptions increased from 0 to 0.8, Recall@1 initially dropped sharply from 64% to 63%, followed by a stabilizing trend, while mAP exhibited a "rise-then-fall" pattern. This behavior was primarily driven by the cross-extraction module integrating multimodal semantics. When the proportion of negative samples remained relatively low, strong correspondences were established between the "core semantics" emphasized by the text module and the "local features" extracted by the visual module through the bi-directional mappings, helping the model to precisely identify relevant image-text pairs under noisy conditions. However, once the negative sample ratio exceeded a critical threshold (e.g., >0.4), the inherent semantic generality of the text inputs led to a decline in mAP. Despite this, Recall@1 remained above 63%, further validating the discriminative power of the fine-grained features of the global-local visual module, which compensated for the limitations of the text module.

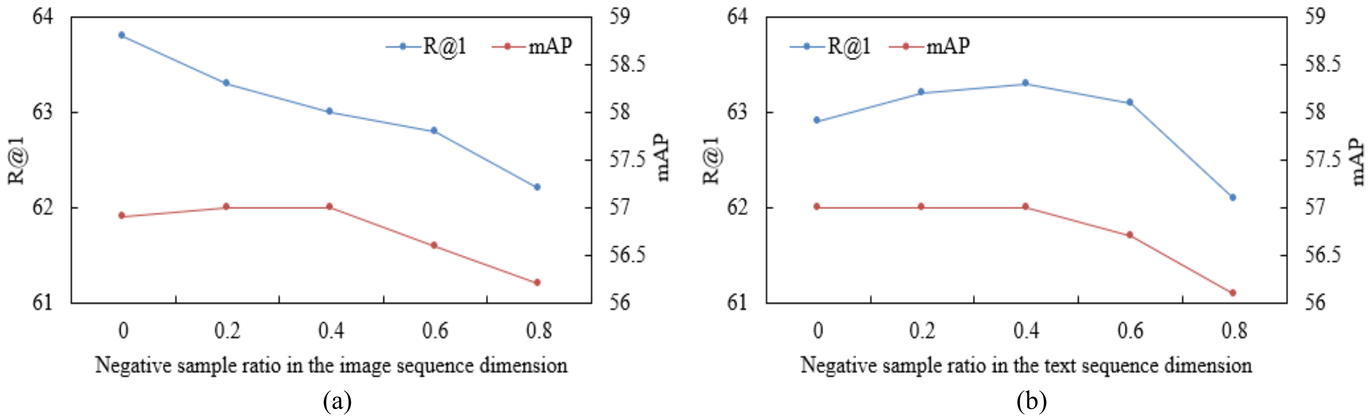

Figure 7. Impact of positive-to-negative sample ratio on retrieval performance under the WAB dataset

The retrieval robustness of the proposed approach under real-world livestreaming conditions was elucidated by examining the effect of varying negative sample ratios in the WAB dataset along two dimensions: the image sequence and the text sequence. As illustrated in Figure 7(a), when the proportion of mismatched product images increased from 0 to 0.8, Recall@1 declined only marginally from 63.8% to 62.2%, a reduction of merely 1.6 percentage points. Meanwhile, mAP dropped from 61.8% to 61.0%, reflecting a limited decrease of 0.8. This gradual degradation underscores the advantage of the global-local visual feature extraction module in terms of fine-grained local features. Under supervision from product category labels, the module was enabled not only to capture holistic product contours but also to extract category-discriminative local features with high precision, including neckline cut, cuff texture, and shoe curvature. Even when a large number of visually similar product images were introduced into the retrieval pool, these irreplaceable local features continued to support the rapid identification of the target product. However, as the number of negative samples increased, the ranking precision was mildly affected, reflecting a retrieval logic of "first-hit prioritization with progressive re-ranking," which aligns well with the practical requirements of e-commerce recommendation systems.

In Figure 7(b), when the mismatched product descriptions increased from 0 to 0.8, Recall@1 exhibited a rise-and-fall trend, peaking at 63.5% when the negative sample ratio reached 0.4, while mAP steadily decreased from 61.2% to 60.8%. This behavior was driven primarily by the synergy between the textual feature extraction module and the cross-extraction module. At lower negative sample ratios, the text module leveraged attention mechanisms to suppress generalized expressions such as “high quality” while enhancing domain-specific semantics such as “high-waisted design” or “genuine leather.” These enriched textual signals formed strong associations with local features of the visual module via the bidirectional mapping mechanism of the cross-module. As the negative sample ratio surpassed the threshold, semantic generalization in the textual modality began to surface, diminishing the mAP. Nonetheless, Recall@1 remained consistently above 62%, once again confirming that fine-grained features of the global-local visual module retained a decisive role in identifying relevant items, thereby establishing a closed loop against interference where semantic filtering in the textual modality is reinforced by visual fine-grained details.

4. CONCLUSION

To address core challenges in cross-modal product image retrieval within e-commerce recommendation systems, a retrieval method integrating global-local visual features, refined textual semantics, and bidirectional cross-modal mapping was proposed. Technical advancements were achieved through three innovative modules. The global-local visual feature extraction module, guided by product category label supervision, was shown to effectively resolve the limitation of missing fine-grained visual details by capturing category-distinctive local attributes such as neckline structure and sleeve texture. The textual feature extraction module, leveraging attention mechanisms, filtered redundant information and emphasized semantically critical descriptors such as “pure cotton fabric,” thereby enhancing the directional specificity of textual features. Furthermore, the vision-language cross-extraction module established a bidirectional mapping, enabling precise alignment between visual fine details and textual attributes. Experimental results demonstrated that Recall@1 exceeded 62% and mAP remained consistently above 60% on benchmark datasets, including Fashion-Gen and WAB. Compared to baseline models such as CLIP, performance improvements of 10%-15% were consistently observed. Notably, in complex e-commerce livestreaming scenarios, the proposed method exhibited strong resilience against challenges such as image blurring and colloquial textual input, providing retrieval support characterized by precise top-ranked results and ranking logic aligned with recommendation objectives, thereby substantially enhancing user decision-making efficiency and platform conversion rates.

Nevertheless, certain limitations remain. First, the datasets primarily comprised categories with distinct visual characteristics (e.g., fashion and apparel), leaving the adaptability to functional categories such as home appliances and consumer electronics insufficiently validated. Further investigation into modeling the relationship between local visual features and functional descriptions is warranted. Second, the current model possesses a large parameter count, posing challenges for real-time retrieval responsiveness in high-concurrency recommendation environments. Future research may be advanced in three directions: (a) expansion of the dataset to include a full range of e-commerce product categories, enabling the construction of a tri-layered annotation system linking visual features, functional attributes, and user needs; (b) application of knowledge distillation techniques to compress model size, supplemented by dynamic feature pruning strategies to improve real-time performance; and (c) incorporation of user behavioral data to integrate retrieval outputs with personalized preference weights, thereby achieving unified optimization of cross-modal matching and personalized recommendation, and further aligning system performance with practical e-commerce demands.

ACKNOWLEDGEMENTS

This work was supported by the Scientific Research Start-up Fund for Talent Introduction at Suqian University, the Top Undergraduate Major Construction Program in Jiangsu Province (Computer Science and Technology), the Key Discipline Program of the 14th Five-Year Plan in Jiangsu Province (Computer Science and Technology), and the Key Discipline of Intelligent Science and Technology at Suqian University.

REFERENCES

[1] Munshi, A., Alhindi, A., Qadah, T.M., Alqurashi, A. (2023). An electronic commerce big data analytics architecture and platform. Applied Sciences, 13(19): 10962. https://doi.org/10.3390/app131910962

[2] Amornkitvikai, Y., Tham, S.Y., Harvie, C., Buachoom, W.W. (2022). Barriers and factors affecting the e-commerce sustainability of Thai micro-, small-and medium-sized enterprises (MSMEs). Sustainability, 14(14): 8476. https://doi.org/10.3390/su14148476

[3] Mustafa, S., Hao, T., Qiao, Y., Kifayat Shah, S., Sun, R. (2022). How a successful implementation and sustainable growth of e-commerce can be achieved in developing countries; A pathway towards green economy. Frontiers in Environmental Science, 10: 940659. https://doi.org/10.3389/fenvs.2022.940659

[4] Ahi, A.A., Sinkovics, N., Sinkovics, R.R. (2023). E-commerce policy and the global economy: A path to more inclusive development? Management International Review, 63(1): 27-56. https://doi.org/10.1007/s11575-022-00490-1

[5] Tiwary, T., Mahapatra, R.P. (2023). Enhancement in web accessibility for visually impaired people using hybrid deep belief network-bald eagle search. Multimedia Tools and Applications, 82(16): 24347-24368. https://doi.org/10.1007/s11042-023-14494-y

[6] Zhang, G., Wei, S., Pang, H., Qiu, S., Zhao, Y. (2023). Enhance composed image retrieval via multi-level collaborative localization and semantic activeness perception. IEEE Transactions on Multimedia, 26: 916-928. https://doi.org/10.1109/TMM.2023.3273466

[7] Tiryaki, A.M., Yücebaş, S.C. (2023). An Ontology based product recommendation system for next generation e-retail. Journal of Organizational Computing and Electronic Commerce, 33(1-2): 1-21. https://doi.org/10.1080/10919392.2023.2226542

[8] Xiao, L.P., Lei, P.R., Peng, W.C. (2022). Hybrid embedding of multi-behavior network and product-content knowledge graph for tourism product recommendation. Journal of Information Science & Engineering, 38(3): 547-570. https://doi.org/10.6688/JISE.202205_38(3).0004

[9] Choudhary, V., Zhang, Z. (2023). Product recommendation and consumer search. Journal of Management Information Systems, 40(3): 752-777. https://doi.org/10.1080/07421222.2023.2229123

[10] Geigle, G., Pfeiffer, J., Reimers, N., Vulić, I., Gurevych, I. (2022). Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval. Transactions of the Association for Computational Linguistics, 10: 503-521. https://doi.org/10.1162/tacl_a_00473

[11] Kaur, P., Malhi, A.K., Pannu, H.S. (2022). Hybrid SOM based cross-modal retrieval exploiting Hebbian learning. Knowledge-Based Systems, 239: 108014. https://doi.org/10.1016/j.knosys.2021.108014

[12] Bhuvaneshwari, P., Rao, A.N., Robinson, Y.H. (2023). Top-N recommendation system using explicit feedback and outer product based residual CNN. Wireless Personal Communications, 128(2): 967-983. https://doi.org/10.1007/s11277-022-09984-5

[13] Latha, Y.M., Rao, B.S. (2024). Product recommendation using enhanced convolutional neural network for e-commerce platform. Cluster Computing, 27(2): 1639-1653. https://doi.org/10.1007/s10586-023-04053-3

[14] Devarajan, G.G., Nagarajan, S.M., Daniel, A., Vignesh, T., Kaluri, R. (2023). Consumer product recommendation system using adapted PSO with federated learning method. IEEE Transactions on Consumer Electronics, 70(1): 2708-2715. https://doi.org/10.1109/TCE.2023.3319374

[15] Liu, B., Guan, W., Yang, C., Fang, Z. (2023). Effective method for making Chinese word vector dynamic. Journal of Intelligent & Fuzzy Systems, 45(1): 941-952. https://doi.org/10.3233/JIFS-224052

[16] Jia, K., Meng, F., Liang, J., Gong, P. (2023). Text sentiment analysis based on BERT-CBLBGA. Computers and Electrical Engineering, 112: 109019. https://doi.org/10.1016/j.compeleceng.2023.109019

[17] Ghosh, R. (2024). Newspaper text recognition in Bengali script using support vector machine. Multimedia Tools and Applications, 83(11): 32973-32991. https://doi.org/10.1007/s11042-023-16862-0

[18] Khan, Q.W., Ahmad, R., Rizwan, A., Khan, A.N., Park, C.W., Kim, D. (2024). Multi-modal fusion approaches for tourism: A comprehensive survey of data-sets, fusion techniques, recent architectures, and future directions. Computers and Electrical Engineering, 116: 109220. https://doi.org/10.1016/j.compeleceng.2024.109220

[19] Hosseinpour, H., Samadzadegan, F., Javan, F.D. (2022). CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 184: 96-115. https://doi.org/10.1016/j.isprsjprs.2021.12.007

[20] Wan, B., Zhou, X., Sun, Y., Wang, T., et al. (2023). MFFNet: Multi-modal feature fusion network for VDT salient object detection. IEEE Transactions on Multimedia, 26: 2069-2081. https://doi.org/10.1109/TMM.2023.3291823