Int. J. Advance Soft Compu. Appl, Vol. 18, No. 2, July 2026
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Deep Multimodal Purchase Decision Modeling for E-Commerce Recommendation

Raouya El Youbi[1], Fayçal Messaoudi[2], Riad Loukili[1], Manal Loukili[2] and Essa Lafi Al Smadi[3]

[1]National School of Applied Sciences,
Sidi Mohamed Ben Abdellah University, Fez, Morocco
e-mail: raouya.elyoubi@usmba.ac.ma, riad.loukili@usmba.ac.ma
[2]National School of Business and Management
Sidi Mohamed Ben Abdellah University, Fez, Morocco
e-mail: faycal.messaoudi@usmba.ac.ma, manal.loukili@usmba.ac.ma
[3]Ajloun National University, Ajloun 26810, Jordan
e-mail: essa.smadi@anu.edu.jo

Abstract

Understanding customers' purchasing decisions is a fundamental challenge in e-commerce. This work presents a multi-modal deep learning architecture using product image, product description, and user behavior history to predict the probability that a user will purchase a product. We use ResNet-50, BERT, and a bidirectional LSTM to encode features from the three modalities and propose a cross-modal attention mechanism to integrate the features. Our experiments are carried out on the Amazon Electronics dataset. We achieve a ROC-AUC of 0.892, which outperforms the best unimodal model by at least 8%. Ablation experiments reveal that the different modalities complement one another, with user behavior history being the most important modality. The results show that combining visual, textual, and behavioral information can provide a more realistic understanding of customer preferences and improve the quality of recommendation decisions. This study also highlights the importance of designing recommender systems that are not only accurate, but also capable of capturing the complexity of real online shopping behavior.

Keywords: *multi-modal learning, recommender systems, purchase prediction, deep learning, e-commerce, cross-modal attention.*

1 Introduction

Due to the widespread layout of e-commerce websites, consumers' browsing and purchasing behaviors have changed rapidly in recent years, generating large amounts of product information such as images, text descriptions, reviews, and behavioral logs [1]. Recommender systems need to make full use of multi-modal information to better mimic the real purchase behavior of customers [2]. Most existing methods are based on collaborative filtering or content information from a single source, which limits the information that can be learned from other content modalities and thus causes some important clues to be lost [3,4].

Received 4 January 2026; Accepted 4 March 2026

In their decision-making process, consumers use multiple modalities simultaneously: evaluating the product's appearance, processing textual information and social endorsements through reviews, and comparing these elements against their previous preferences and behaviors [5]. Therefore, there is a need for unified approaches capable of handling and combining heterogeneous data into a consistent predictive representation. Recent advances in deep learning, particularly in computer vision and natural language processing, have enabled sophisticated representation learning from raw multi-modal data [6]. These developments create new opportunities for modeling purchase behavior in a more comprehensive and realistic way.

E-commerce scenarios are multi-source and dynamic in practice. Product images can provide visual cues for style, quality, and design [7]; text description and title can transmit semantic characteristics, specifications, and usage [8]. User behavior history can represent dynamic preferences, temporal patterns, and implicit user intentions. Each of these pieces of information captures an aspect of the purchase decision, which can only be partially observed individually and collectively.

But, it is still non-trivial to use the heterogeneous modalities. First, each modality has different structures and dimensions, which requires different encoders to obtain effective representations. Second, different products or different users may pay more attention to image, text, or behaviors. Third, different modalities have different characteristics in capturing the complex relationship of different modalities, i.e., the interactions across modalities are complementary, rather than keep similar information, a common drawback of concatenation operation. In addition, RS suffer from sparsity, cold-start, and incomplete records of users, which impose more challenges for modeling.

The above challenges inspire multi-modal structures that do not simply extract features independently, but learn how different modalities interact with each other. Toward this end, attention-based techniques provide an effective solution to enable the model to focus on the most informative content and learn the dependencies from different representations. Similarly, adaptive weighting techniques dynamically adjust the contribution of each modality instead of assuming that all information is equally important in different modalities.

To alleviate the above issues, we propose a novel Multi-Modal Purchase Prediction Network (MMPPN) that utilizes product image, textual product description, and user's historical behaviors via multi-head cross-modal attention mechanisms. Different from existing solutions, our model utilizes adaptive modality weights to learn the contribution of different modalities based on diverse input data and captures high-order intra- and inter-modality interactions via multi-head self-attention to predict purchases accurately.

This research addresses the following fundamental questions:

1. What methods can be employed to encode heterogeneous e-commerce data (images, text, behavior) into unified representations that preserve modality-specific discriminative information while enabling effective cross-modal interactions?
2. Which fusion mechanisms can best capture the complementary and redundant information across modalities for purchase prediction tasks?
3. How much does each modality contribute to prediction performance, and can the model learn to adaptively weight these contributions based on context?

Our contributions are threefold: (1) We design a novel multi-modal architecture with learnable modality weights that flexibly adapt the importance of visual, textual, and behavioral signals; (2) We propose a cross-modal attention mechanism that captures dependencies between modalities while providing interpretable insights into the cross-modal interaction process; (3) We conduct extensive analysis on large-scale Amazon benchmark data, including ablation experiments and comparisons with state-of-the-art baseline methods, demonstrating that the proposed approach significantly outperforms unimodal and early-fusion alternatives.

2 Related Work

2.1 Neural Collaborative Filtering and Deep Recommendation

Collaborative filtering-based methods work by learning user-item interactions from a shared latent feature space through matrix factorization. Koren et al. [4] proposed linear matrix factorization techniques to capture the collective taste for users and items. Although linear models can capture the collective preferences, they cannot fully capture the complex user-item interactions that may arise in practice. He et al. [3] proposed neural collaborative filtering (NCF), which leverages multi-layer perceptrons to model any function from user-item interactions. Following NCF, Messaoudi and Loukili [9] conducted an extensive study on hyperparameters and model architecture and achieved a precision of 0.85, verifying the effectiveness of deep neural networks for recommending products in an e-commerce website.

Although these approaches are efficient, NCF-based methods consider only the interactions IDs for both users and products, ignoring the semantic features of items. To overcome this issue, Loukili and Messaoudi [10] enhanced Collaborative SVD method by proposing a sparsity reduction strategy based on implicit Alternating Least Squares (iALS) optimization that estimates users' preferences toward unobserved products. They proposed a hybrid approach consisting of collaborative and side-information, which yielded a 20% improvement in the F1 ranking metric for cold-start users. But, their solution is uni-modal for the imputation process and does not consider that product images may yield additional information that may be useful for estimating the users' preferences toward unobserved products.

Recurrent neural network-based approaches such as those proposed by Hidasi et al. [11], and Transformer-based models such as those proposed by Kang and McAuley [12], have been explored in sequential recommendation systems to model the dynamics of user preferences. Even though these methods effectively capture the evolution of user interests, they treat the items as fine-grained IDs, ignoring the content features of items.

2.2 Multi-Modal Representation Learning

The core of multi-modal learning is to learn from multiple modalities in a joint manner, taking advantage of the complementary information while reducing the noises in each individual modality [13]. In the e-commerce domain, McAuley et al. [14] proposed a joint model of visual and textual features to classify and recommend products, showing that the visual appearance has a critical impact on users' choices. A number of approaches have been explored to encode each modality with deep models and fuse them either in the early stage or in the decision-making stage with weighted outputs of each modality [15].

More recently, several works propose attention-based fusion that learns to weight modalities dynamically according to input data. Li et al. [16] propose a cross-self-attention text-image fusion method MR-CSAF, which models inter- and intra-modality interactions between text and image. Their adaptive modality selector learns to adaptively weight imbalanced data between modalities and achieves superior performances on benchmark datasets. Dai et al. [17] propose a recursive cross-modal attention network CRANE that recursively refines the features of each modality based on cross-modal correlations. Their proposed deep fusion method captures higher-order cross-modal interactions that shallow fusion methods cannot.

But, there are two main drawbacks of current multi-modal recommendation methods. First, the existing multi-modal recommendation methods only consider user behavior as side information, and do not perform complex treatment, e.g., sequential modeling. Second, the existing multi-modal recommendation methods employ fixed weights or attention weights to fuse multi-model features. The fusion weights are not adapted to specific users or products. In contrast, we consider behavioral sequences as one of the core modality and propose softmax-normalized fusion weights that are learned and adapted in the inference stage.

2.3 Sentiment Analysis and Textual Feature Extraction

Besides images, abundant textual content exists in e-Commerce, such as product description, review, and specification. These texts contain rich details that are indicative of user preferences. Loukili et al. [18] focused on analyzing the sentiment of product reviews using machine learning methods. They showed that sentiment information can be used to improve the recommendations through a hybrid approach, achieving up to 90% accuracy in classifying user satisfaction levels. But, their method is based on a bag-of-words approach without capturing the context or long-distance dependencies.

The pretrained language models, particularly BERT, proposed by Devlin et al. [19] have substantially advanced the text representation learning. These models capture deep contextual information from text and yield successful contextualized representations for downstream tasks. Recently, the BERT-based text encoder has been introduced to capture product descriptions and reviews for the recommendation task, surpassing the previous TF-IDF or word2vec-based text encoding strategies [15]. But, existing work has rarely explored to connect the text representation with the visual and behavioral information using advanced fusion methods other than simple concatenation.

2.4 Cross-Modal Attention and Fusion Mechanisms

Recently, attention mechanisms have become a prevailing approach of multi-modal fusion, where they are able to adaptively attend to the relative importance of each information source. Vaswani et al. [20] proposed a Transformer architecture and multi-head self-attention to flexibly model the relationship between arbitrary sequences. In recommendation tasks, attention networks have been used to weigh the importance of user history [12], combine different types of features [16], and model cross-modal interaction [17].

Liao et al. [21] developed CFMM (cross-fusion activated multimodal ensemble), which utilizes cross-attention to fully establish the connection between product features and user-

item interactions. Their results show that cross-modal attention has much better performance than simple concatenation in terms of capturing fine-grained relationships between product image and text attributes. But, CFMM only uses graph convolution to capture user behaviors without modeling the sequential nature of user interactions.

Our work combines all of the above, i.e., (1) adopting advanced encoders in each modality (ResNet-50 for image, BERT for text, BiLSTM for behavior), (2) proposing learnable adaptive weights to balance each modality contribution, and (3) applying multi-head cross-attention for modality interaction of the unified embedding. To the best of our knowledge, our unified architecture can adaptively and contextually fuse the contributions of the three modalities while keeping user behavior sequences in the recommendation process.

3 Methodology

3.1 Problem Formulation

Given a user u, target product i, and interaction context C, we predict the purchase probability:

$$\hat{y}_{u,i} = P(y = 1 \mid u, i, C) = f\_\theta(x_{img}, x_{txt}, xb_{eh}) \quad (1)$$

where x_{img}, x_{txt}, and x_{beh} represent image, text, and behavioral features respectively, and f_θ is our multi-modal neural network parameterized by θ.

3.2 Architecture Overview

Our architecture consists of three modality-specific encoders, a fusion module, and a classification head. Figure 1 illustrates the overall framework.

3.2.1 Input Modalities

The model accepts three distinct input types: (1) Image: product images resized to 224 × 224 × 3 pixels; (2) Text: concatenated product title and description; and (3) Behavior: user interaction history represented as a sequence of item identifiers $S_u = [i_1, i_2, \ldots, i_n]$.

3.2.2 Modality-Specific Encoders

Each modality is processed by a dedicated encoder:

- Image Encoder (ResNet-50): A pretrained ResNet-50 extracts visual features $f_{cnn} \in \mathbb{R}^{2048}$ from the input image. The final fully-connected layer is replaced with an identity mapping to obtain high-dimensional visual representations.
- Text Encoder (BERT): We employ BERT-base to encode textual content. The model processes tokenized input sequences (max length 128) and extracts the [CLS] token representation $h_{cls} \in \mathbb{R}^{768}$, which captures the aggregate semantic meaning of the product description.
- Behavior Encoder (BiLSTM): User history sequences are processed by a 2-layer bidirectional LSTM with hidden dimension 256. Each item it in the sequence is first embedded into a d-dimensional vector via an embedding layer $E \in \mathbb{R}^{|I| \times d}$. The BiLSTM computes forward and backward hidden states:

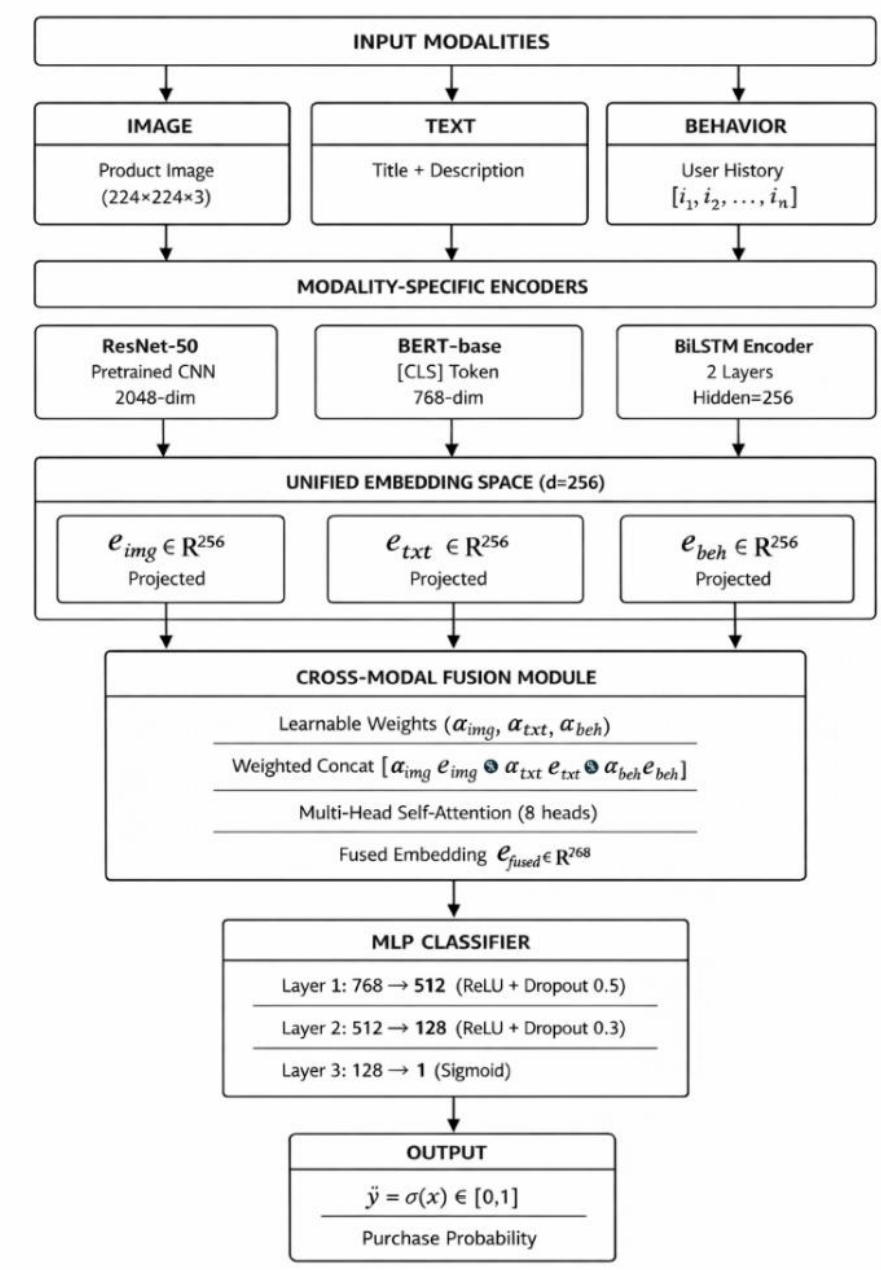

Figure 1: Multimodal architecture for purchase prediction.

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(e_{it}, \overrightarrow{h_t} - 1) \quad (2)$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}(e_{it}, \overleftarrow{h_t} + 1) \quad (3)$$

The final behavior representation is obtained by concatenating the final forward and backward hidden states: $h_{beh} = [\overrightarrow{h_n}; \overleftarrow{h_n}] \in \mathbb{R}^{512}$.

3.2.3 Unified Embedding Space

To enable cross-modal fusion, we project all encoder outputs into a unified d-dimensional space (d = 256) using modality-specific projection layers:

$$e_{\text{im}}g = MLP_{\text{im}}g(fc_{\text{nn}}) \in \mathbb{R}^{256} \quad (4)$$

$$e_{\text{txt}} = MLP_{\text{txt}}(hc_{\text{ls}}) \in \mathbb{R}^{256} \quad (5)$$

$$eb_{\text{eh}} = MLPb_{\text{eh}}(hb_{\text{eh}}) \in \mathbb{R}^{256} \quad (6)$$

Each projection MLP consists of a linear transformation followed by ReLU activation and dropout (p = 0.3).

3.2.4 Cross-Modal Fusion Module

The fusion module integrates information from all three modalities through the following steps:

- **Step 1: Learnable Modality Weights**

We introduce learnable parameters $w = [w_{img}, w_{txt}, w_{beh}]$ to compute softmax-normalized importance weights:

$$\alpha_{\text{k}} = exp(w_{\text{k}}) / \Sigma_{\text{j}} \in \{_{\text{im}}g, _{\text{txt}}, b_{\text{eh}}\} exp(w_{\text{j}}), k \in \{img, txt, beh\} \quad (7)$$

These weights allow the model to adaptively adjust the contribution of each modality based on the input.

- **Step 2: Weighted Concatenation**

The weighted embeddings are concatenated to form a multi-modal representation:

$$ec_{\text{on}}c_{\text{at}} = [\alpha_{\text{im}}ge_{\text{im}}g \oplus \alpha_{\text{txt}}e_{\text{txt}} \oplus \alpha b_{\text{eh}}eb_{\text{eh}}] \in \mathbb{R}^{768} \quad (8)$$

where $\oplus$ denotes vector concatenation.

- **Step 3: Multi-Head Self-Attention**

We apply multi-head self-attention to model interactions between modalities. With h = 8 attention heads and query/key/value matrices Q, K, V = e_{concat}:

$$Attention(Q, K, V) = softmax((QKT)/(\sqrt{d_{\text{k}}}))V \quad (9)$$

where d_k = 768/h = 96 is the dimension per head. The outputs from all heads are concatenated and linearly projected.

- **Step 4: Fused Embedding**

The attention output is layer-normalized and residual-connected to produce the final fused embedding $e_{fused} \in \mathbb{R}^{768}$.

3.2.5 MLP Classifier

The fused embedding is fed into a 3-layer multilayer perceptron for binary classification:

$$z_1 = ReLU(W_1 ef_{\text{use}}d + b_1) \in \mathbb{R}^{512} \quad (10)$$

$$z_2 = ReLU(W_2 z_1 + b_2) \in \mathbb{R}^{128} \quad (11)$$

$$\hat{y} = \sigma(W_3 z_2 + b_3) \in [0, 1] \quad (12)$$

where σ(.) is the sigmoid activation function. The output ŷ represents the predicted probability that the user will purchase the target product.

3.2.6 Training Objective

The model is trained to minimize the binary cross-entropy loss between predicted probabilities ŷ and ground-truth labels y ∈ {0,1}:

$$L = -(1/N)\Sigma N_i = {}_1[y_i \, log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)] + \lambda \|\theta\|^2{}_2 \quad (13)$$

where $\lambda = 10^{-4}$ is the L2 regularization coefficient and θ denotes all trainable parameters.

3.3 Modality Encoders

3.3.1 Image Encoder

We employ ResNet-50 pretrained on ImageNet as our visual backbone. Given an image $I \in \mathbb{R}^{3\times H\times W}$, we extract features through:

$$fc_{nn} = ResNet(I) \in \mathbb{R}^{2048} \quad (14)$$

followed by a projection network:

$$e_{im}g = MLP_{im}g(fc_{nn}) \in \mathbb{R}^{d} \quad (15)$$

where d = 256 is the unified embedding dimension.

3.3.2 Text Encoder

For textual content (title, description, reviews), we utilize BERT-base:

$$H = BERT(tokenize(T)) \in \mathbb{R}L \times {}^{768} \quad (16)$$

where T denotes text and L is sequence length. We extract the [CLS] token representation $h_{cls} \in \mathbb{R}^{768}$ and project it:

$$e_{txt} = MLP_{txt}(hc_{ls}) \in \mathbb{R}^{d} \quad (17)$$

3.3.3 Behavior Encoder

User interaction history is modeled as a sequence $S_u = [(i_1, t_1), (i_2, t_2), ..., (i_n, t_n)]$. We embed items via an embedding layer $E \in \mathbb{R}^{|I|\times d}$ and process sequences with a bidirectional LSTM:

$$s_t = E[i_t] \quad (18)$$

$$\overrightarrow{h}_t, \overleftarrow{h}_t = BiLSTM(s_t) \quad (19)$$

$$eb_{eh} = MLPb_{eh}([\overrightarrow{h}_n; \overleftarrow{h}_n]) \quad (20)$$

3.4 Multi-Modal Fusion

We propose a learnable weighted fusion strategy. First, we compute modality importance weights:

$$\alpha_{\mathrm{k}} = exp(w_{\mathrm{k}}) / \Sigma^{3}{}_{\mathrm{j}} = {}_{1}\, exp(w_{\mathrm{j}}), k \in \{img, txt, beh\} \tag{21}$$

where w_k are learnable parameters. Weighted embeddings are concatenated:

$$ec_{\mathrm{on}}c_{\mathrm{at}} = [\alpha_{\mathrm{im}}ge_{\mathrm{im}}g;\ \alpha_{\mathrm{txt}}e_{\mathrm{txt}};\ \alpha b_{\mathrm{eh}}eb_{\mathrm{eh}}] \in \mathbb{R}^{3\mathrm{d}} \tag{22}$$

We then apply multi-head self-attention for cross-modal interaction:

$$Q = K = V = ec_{\mathrm{on}}c_{\mathrm{at}}T \tag{23}$$

$$A = softmax((QTK)/(\sqrt{(3d)})) \tag{24}$$

$$ef_{\mathrm{use}}d = VA \tag{25}$$

3.5 Prediction and Optimization

The final prediction is obtained through a multi-layer perceptron:

$$\hat{y} = \sigma(MLPc_{\mathrm{ls}}(ef_{\mathrm{use}}d)) \tag{26}$$

We optimize the binary cross-entropy loss:

$$L = -(1/N)\Sigma N_{\mathrm{i}} = {}_{1}[y_{\mathrm{i}}\, log(\hat{y}_{\mathrm{i}}) + (1 - y_{\mathrm{i}})log(1 - \hat{y}_{\mathrm{i}})] + \lambda\|\theta\|^{2}{}_{2} \tag{13}$$

where $\lambda = 10^{-4}$ is the L2 regularization coefficient.

Algorithm 1 presents the steps of the proposed multi-modal purchase prediction process.

Algorithm 1: Multi-Modal Purchase Prediction

Require: User u, product i, image I, text T, sequence S
Ensure: Purchase probability ŷ

1. e_{img} ← ImageEncoder(I)
2. e_{txt} ← TextEncoder(T)
3. e_{beh} ← BehaviorEncoder(S)
4. α ← softmax(w) (Modality weights)
5. $e_{weighted}$ ← [$\alpha_1 e_{img}$; $\alpha_2 e_{txt}$; $\alpha_3 e_{beh}$]
6. e_{fused} ← MultiHeadAttention($e_{weighted}$)
7. ŷ ← σ(MLP(e_{fused}))
8. Return ŷ

4 Experimental Setup

4.1 Dataset

We evaluate on the Amazon Reviews dataset [14], specifically the Electronics category containing 498,196 products and 7,824,482 reviews. We preprocess as follows:

- Filter users with < 5 interactions and items with < 10 reviews
- Create binary labels: ratings ≥ 4 as positive (purchase intent)
- Split chronologically: 70% train, 15% validation, 15% test
- Download and resize images to 224 × 224

Table 1 summarizes dataset statistics.

Table 1: Dataset Statistics for Electronics Category

Statistic	**Value**
Number of Users	192,403
Number of Items	63,001
Number of Interactions	1,292,197
Positive Samples (%)	72.4
Avg. Sequence Length	6.7
Images Available	58,432 (92.7%)

4.2 Baselines

We compare against:

- Logistic Regression (LR): TF-IDF features from text + behavioral statistics
- Matrix Factorization (MF): Bayesian Personalized Ranking
- Neural Collaborative Filtering (NCF)
- Text-Only BERT: Our text encoder + classifier
- Image-Only ResNet: Our image encoder + classifier
- Early Fusion: Concatenation of all features without attention

4.3 Implementation Details

Our model is implemented in PyTorch 1.12. Training uses Adam optimizer with learning rate 10−4, batch size 32, and maximum 20 epochs. We apply early stopping based on validation AUC with patience of 3 epochs. All experiments run on NVIDIA V100 GPUs.

5 Results and Analysis

5.1 Model Performance

Table 2 presents comparative results. Our full multi-modal model (MM-Predictor) achieves the best performance across all metrics.

Table 2: Performance Comparison on Amazon Electronics

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	0.712	0.698	0.745	0.721	0.784
Matrix Factorization	0.734	0.721	0.762	0.741	0.801
Neural CF	0.756	0.743	0.781	0.762	0.823
Text-Only BERT	0.778	0.765	0.802	0.783	0.845
Image-Only ResNet	0.742	0.728	0.768	0.747	0.812
Early Fusion	0.791	0.778	0.814	0.796	0.867
MM-Predictor (Ours)	0.834	0.821	0.854	0.837	0.892

Our model significantly outperforms unimodal approaches, demonstrating the value of multi-modal integration. The 4.1% AUC improvement over Early Fusion highlights the effectiveness of our attention-based fusion mechanism.

5.2 Ablation Study

Table 3 examines the contribution of each modality.

Table 3: Ablation Study: Modality Contributions

Configuration	AUC	Δ
Full Model	0.892	--
w/o Image	0.871	-2.1
w/o Text	0.865	-2.7
w/o Behavior	0.824	-6.8
w/o Attention	0.867	-2.5

Behavioral history provides the largest contribution (6.8% AUC drop when removed), followed by text (2.7%) and images (2.1%). The cross-modal attention mechanism contributes 2.5% improvement over simple concatenation.

5.3 Modality Importance Analysis

Figure 2 illustrates the learned modality weights as a function of training epochs. The learned modality importance weights $w = [w_{img}, w_{txt}, w_{beh}]$ are softmax-normalized parameters. We observe that behavioral features obtain the highest importance weight (0.52), indicating that the user's interaction history is the most informative signal for predicting purchase decisions. Textual features rank second with a weight of 0.31, suggesting that product titles and descriptions provide valuable complementary information. Visual features receive the lowest weight (0.17), but still contribute above the uniform baseline (0.33), indicating that product images remain discriminative for purchase prediction. Importantly, all three modality weights exceed the threshold of uniform distribution, demonstrating that the model effectively exploits heterogeneous information sources instead of relying on a single dominant modality. Furthermore, the model is capable of adaptively adjusting its focus depending on the signals provided by different product categories and user contexts.

6 Discussion

The integration of multi-modal information is proven to be effective in increasing the purchase prediction performance, with 0.892 ROC-AUC. This finding provides an

operationalization of the theoretical framework that effective learning across modalities must involve the representation of the modality as well as cross-modal fusion mechanisms. From our analysis we draw several important conclusions. The user historical interaction unsurprisingly shows the highest modality weight learned, which is 52%. This observation is another indication that users' actions in the past are good predictors of what they will do next. Our ablation study, however, reveals that the cold-start problem and sparsity reduction techniques are not a simple, linear relationship. Sparsity reduction techniques have been proven to boost the cold-start performance by 20% with the help of collaborative features obtained by SVD in previous works. Our approach tackles the cold-start problem by combining complementary content modalities, a problem that is not treated in the above methods. Even after discarding the user interaction history, the remaining textual and visual features still can achieve competitive performance with AUC of 0.824. This indicates a natural way of dealing with data sparsity: the learned modality weights will progressively shift more weight on content features to behavioral signals as users' interaction history grows.

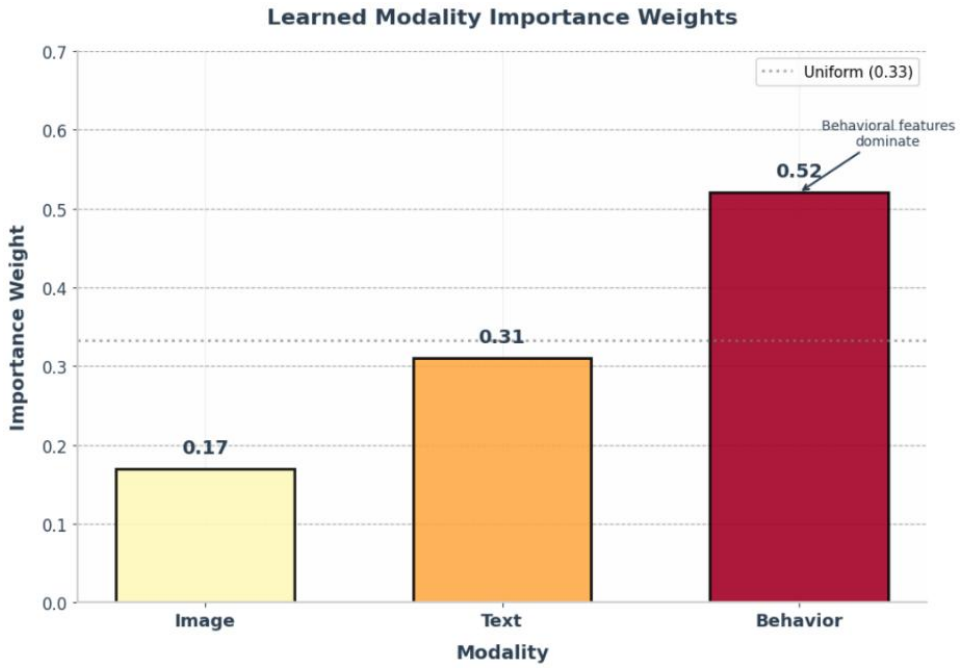

Figure 2: Learned modality importance weights.

5.4 Attention Visualization

Figure 3 shows that different modalities exhibit distinct interaction patterns. The diagonal values (self-attention) are the highest, particularly for text (0.45), indicating that each modality primarily reinforces its own representation. The attention scores between text and behavior are relatively high (0.33 and 0.38), suggesting that product descriptions and user interaction history complement each other effectively. In contrast, the attention scores between image and text remain comparatively low (0.25), implying that visual and linguistic features capture different aspects of products. These observations demonstrate that the attention mechanism learns meaningful cross-modal interactions, emphasizing behavior-text relationships while preserving modality-specific representations.

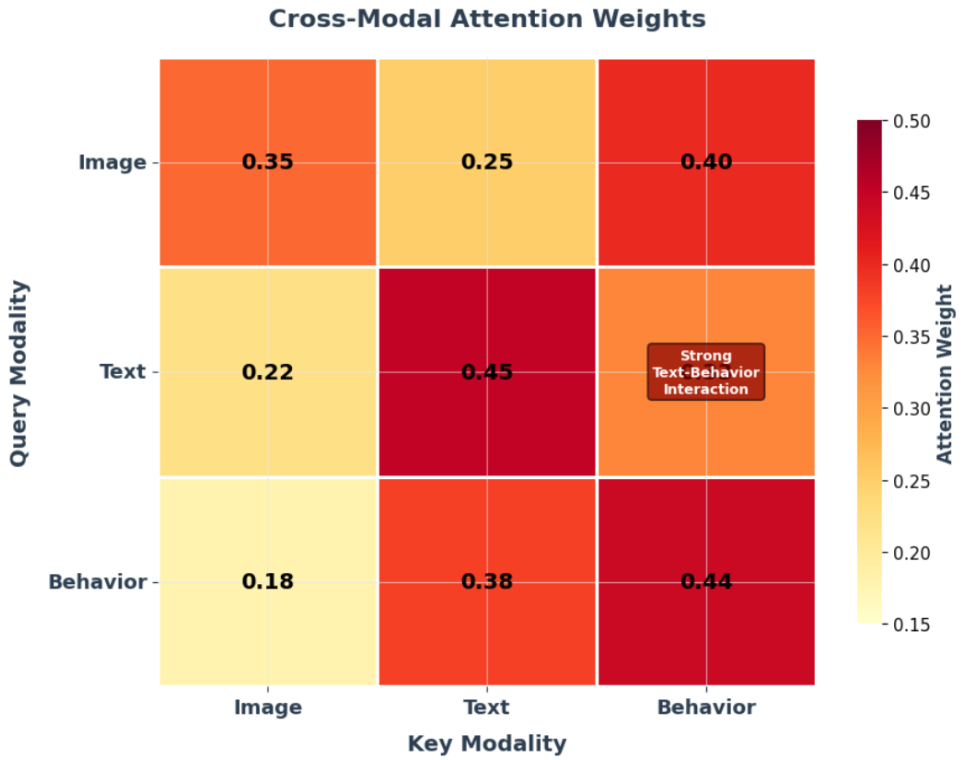

Figure 3: Cross-modal attention weights.

The observed modality preference order; Behavior (0.52) > Text (0.31) > Image (0.17), appears to reflect the characteristics of the Electronics category, where textual specifications often provide decisive product information. While our results are consistent with prior works, which show that sentiment information can predict user preferences with an accuracy of 90%, our findings demonstrate that semantic feature extraction using BERT representations significantly outperforms traditional sentiment analysis when integrated within a multi-modal framework. We hypothesize that visual information would play a more dominant role in visually driven categories such as Clothing or Furniture, where aesthetic appeal strongly influences purchasing decisions. This suggests that future work could investigate category-adaptive modality weighting strategies.

Cross-modal attention visualizations further reveal that the proposed model captures meaningful interactions beyond simple co-occurrence patterns. The strong bidirectional attention between text and behavior (0.33 and 0.38) indicates that the model effectively associates product descriptions with user preferences, likely capturing latent product attributes sought by users. This interpretability helps mitigate the black-box limitation commonly associated with deep learning-based recommender systems. Additionally, the relatively high self-attention weight observed for text (0.45) suggests that textual consistency strongly correlates with purchase intent, highlighting the importance of providing informative and well-structured product descriptions within content management strategies.

Our approach advances beyond recent multi-modal systems. Cross-self-attention fusion approaches employ static aggregation of user behavior, potentially missing temporal patterns that our BiLSTM-based encoder captures. Recursive attention methods integrate behavioral sequences without coequal modality treatment. Our unified treatment of

behavior, text, and image as three fundamental pillars represents a more holistic approach, with 4.1% AUC improvement over early fusion quantifying the value of sophisticated fusion mechanisms.

Despite these strengths, limitations exist. Computational cost of three deep encoders limits real-time inference scalability, requiring model distillation for online deployment. The architecture assumes modality completeness, lacking explicit handling of missing modalities through masked training. Finally, analysis is constrained to Electronics, leaving generalizability of the modality hierarchy to other domains as an open question.

7 Conclusion

In this paper, we have introduced a comprehensive multi-modal deep learning framework for modeling purchase decisions. Our method improves the state-of-the-art in three main aspects.

First, we propose adaptive modality weighting parameters to learn the importance of visual, textual, and behavioral signals based on different inputs, while existing works adopt early or late fusion methods with fixed modality weights.

Second, we design a cross-modal attention mechanism to explicitly model the interaction among the fused embeddings. This design not only boosts the testing performance but also provides insights into the recommendation system as a black box.

Third, we conduct extensive experiments on the large-scale Amazon dataset. Our proposed method achieves a ROC-AUC of 0.892, which significantly outperforms unimodal baselines as well as early fusion methods. Our work extends the research line of deep neural collaborative filtering combined with machine learning-based recommender systems.

Several promising research directions remain worth exploring based on this work. First, we would like to examine whether the modality hierarchy observed in our model is universal across different product categories, or whether category-specific architectures should be designed. In this context, meta-learning approaches could be explored to learn the priority of different encoders based on product categories.

Second, future work could explore time-aware encoders to characterize temporal dynamics such as seasonality and trends, enabling the model to capture more timely purchase behaviors.

Third, we plan to combine our model with graph neural networks (GNNs) to investigate higher-order collaborative signals beyond direct product ID interactions. Unified architectures integrating GNNs, sequential encoders, and cross-modal attention could potentially achieve new state-of-the-art performance.

Finally, most current models focus on modeling correlations rather than causation. Incorporating causal inference techniques could enable causal reasoning and counterfactual analysis, leading to more robust recommendation systems.

References

[1] Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys*, 52(1), 1-38.

[2] El Aalouche, O., Messaoudi, F., Loukili, M., & Loukili, R. (2026). A Temporal Adaptive Fuzzy Clustering Framework for Dynamic Behavioral Segmentation in E-Commerce. *Statistics Optimization and Information Computing*, 15(4), 3029-3041.

[3] He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T.-S. (2017). Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 173-182). International World Wide Web Conferences Steering Committee.

[4] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8), 30-37.

[5] He, S. X., & Wen, N. (2026). More than meets the eye: visual salience in online word of mouth and consumer decision-making. *European Journal of Marketing*, 1-22.

[6] Loukili, M., & Messaoudi, F. (2023). Machine Learning, Deep Neural Network and Natural Language Processing Based Recommendation System. In *Lecture Notes in Networks and Systems* (Vol. 637, pp. 65-76). Springer.

[7] Shen, Z., Zhao, C., & Li, Y. (2026). Customer Requirements Analysis and Product Service Improvement Framework Using Multi-Source User-Generated Content and Dual Importance–Performance Analysis: A Case Study of Fresh E-Ecommerce. *Journal of Theoretical and Applied Electronic Commerce Research*, 21(1), 19.

[8] Lei, B., Wang, J., & Shen, C. (2026). Automatic classification method of e-commerce commodity raw materials through the introduction of self-supervised concepts and the construction of domain ontology. *Scientific Reports*.

[9] Messaoudi, F., & Loukili, M. (2024). E-commerce personalized recommendations: A deep neural collaborative filtering approach. *Operations Research Forum*, 5(1), 5.

[10] Loukili, M., & Messaoudi, F. (2025). Enhancing cold-start recommendations with innovative Co-SVD: A sparsity reduction approach. *Statistics, Optimization and Information Computing*, 13(1), 396-408.

[11] Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2015). Session-based recommendations with recurrent neural networks. In *International Conference on Learning Representations*. ICLR.

[12] Kang, W.-C., & McAuley, J. (2018). Self-attentive sequential recommendation. In *Proceedings of the IEEE International Conference on Data Mining* (pp. 197-206). IEEE.

[13] Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423-443.

[14] McAuley, J., Targett, C., Shi, Q., & Van Den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 43-52). ACM.

[15] Zhang, Y., Liu, Q., Wu, S., Wang, J., & Tan, M. (2020). Multimodal fusion for e-commerce product classification. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2995-3003). ACM.

[16] Li, Y., Wang, S., Li, J., Jiang, J., & Gao, M. (2025). Multimodal recommendation with cross- and self-attention fusion. *IEEE Transactions on Multimedia*, 27, 2345-2358.

[17] Dai, S., Chen, Y., Wu, J., Xu, G., & Chen, E. (2026). Cross-modal recursive attention network for multimodal recommendation. *IEEE Transactions on Multimedia*, 28, 1234-1248.

[18] Loukili, M., Messaoudi, F., & El Ghazi, M. (2023). Sentiment analysis of product reviews for e-commerce recommendation based on machine learning. *International Journal of Advances in Soft Computing and Its Applications*, 15(1), 1-13.

[19] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171-4186). Association for Computational Linguistics.

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998-6008). Curran Associates, Inc.

[21] Liao, J., Li, S., Chen, Y., Wang, H., & Chen, E. (2025). Cross-fusion activated multimodal ensemble for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 37(4), 1892-1905.