Knowledge and Information Systems (2026) 68:226
https://doi.org/10.1007/s10115-026-02844-9

RESEARCH

Joint aspect-based sentiment and overall rating prediction via a cross-modal transformer in user reviews

Xiaoning Wang[1] · Rui Li[1] · Yuxuan Guo[2] · Xiaoling Lu[3] · Jiawei Ren[1] · Libo Sun[4]

Received: 7 October 2025 / Revised: 30 March 2026 / Accepted: 21 June 2026

Abstract
Rating prediction is an important task in review mining. Aspect category sentiment analysis (ACSA) and overall rating prediction (ORP) are closely related tasks, yet existing joint models are typically limited to text-only inputs, while existing multimodal review models usually address only a single task. To bridge this gap, we propose MMRP, an end-to-end multimodal multitask framework for jointly modeling ACSA and ORP from image-text reviews. The model integrates textual and visual information through a cross-modal Transformer and further introduces modality-aware learnable positional encoding and a gated multi-head cross-modal attention mechanism to better handle the structure of multimodal reviews, where text forms token sequences and images appear as variable-sized sets. We evaluate MMRP on the ZOL mobile review dataset and an additional cross-domain multimodal benchmark, ViMACSA. Extensive experiments, including comparative evaluation, ablation analysis, sensitivity analysis, and case studies, show that MMRP consistently achieves strong performance and exhibits good robustness, cross-domain applicability, and practical deployment potential. These results demonstrate the effectiveness of jointly leveraging multimodal information for aspect-level and overall rating prediction in user reviews.

Keywords Review mining · Multimodal data analysis · Multitask rating prediction

1 Introduction

Sharing user-generated comments has become a prevalent social practice, facilitating the exchange of experiences and opinions. Particularly on e-commerce platforms, user reviews hold substantial importance for documenting and disseminating consumption experiences [1, 2]. The user review data holds significant value for both businesses and customers. For businesses, review data aids in understanding user preferences, monitoring perceptions of

✉ Xiaoling Lu
xiaolinglu@ruc.edu.cn

1 School of Data Science and Media Intelligence, Communication University of China, Dingfuzhuang, Chaoyang, Beijing 100024, China

2 Maxwealth Fund Management Company Limited, Pudong District, ShangHai 200120, China

3 Center for Applied Statistics, School of Statistics, Innovation Platform, Renmin University of China, Zhongguancun, Haidian, Beijing 100872, China

4 School of Data Science, Fudan University, Handan, Yangpu, Shanghai 200433, China

Published online: 22 July 2026

products, and driving improvements. For consumers, review data assists in identifying product strengths and weaknesses, thereby aiding better purchasing decisions [3–5]. Consequently, in-depth review mining has become an important research topic.

Rating prediction is an important supervised task in the field of review mining. The Rating Prediction (RP) task entails predicting specific ratings based on the sentiment expressed in reviews. RP is a central focus of this study due to its practical significance, as many platforms employ 5-star scale ratings for user feedback. Aspect Category Sentiment Analysis (ACSA) on online platforms aims to analyze and understand users' sentiment expressions at a fine-grained level, predicting the sentiment expressions on various aspect categories. Specifically, for a given review content like "The fish is delicious, but the waiter is terrible!", the ACSA task aims to infer the sentiment polarity of the aspect category food as positive and the opinion on the aspect category service as negative. The ACSA task focused on in this study requires predicting sentiment scores on given aspect categories, so the essence of ACSA is rating prediction.

The ACSA task and the overall rating prediction problem (ORP) often occur simultaneously. ACSA focuses on predicting sentiment scores for different aspect categories, which is actually a fine-grained rating task. ORP, on the other hand, focuses on predicting the overall sentiment from review content, which is a coarse-grained overall rating task. These two tasks are highly correlated, covering sentiment polarity detection from fine-grained to coarse-grained levels. Most studies on these two tasks have treated them as single-task problems [6–9]. These studies focus on building advanced model structures to achieve more accurate rating prediction results and integrate various auxiliary information to improve prediction accuracy. Joint modeling of ACSA and ORP can leverage information at different granularities to improve prediction accuracy [10].

With the widespread usage of taking pictures on mobile phones, reviews include more images than ever before. As a result, many e-commerce reviews are multimodal. Review mining for text content is a mature research field, but the studies on multimodal review mining with image information are still limited. Studies have shown that images are often more visually impactful than text, and that images in reviews contain strong user opinions [11, 12]. Text and images in reviews are usually complementary rather than independent, and visual content often supplements or reinforces the textual message. In other words, images do not tell the whole story on their own. Compared to textual analysis alone, multimodal reviews containing images provide a rich source of information. The inclusion of visual content enhances our comprehension of review context, consequently elevating the accuracy of tasks such as rating prediction. Intuitive visual information can reduce consumers' purchase risk and improve shopping satisfaction [13].

In sum, current research in rating prediction faces several limitations. First, although ACSA and ORP are highly related tasks, they are typically studied in isolation, with most models focusing on single-task learning. Existing models for review rating prediction often address only one task (either ACSA or ORP), missing the opportunity to leverage the strong correlation between these tasks to enhance prediction accuracy. Second, most existing models rely on single-modality approaches. These models predominantly depend on textual data, failing to fully utilize the rich information provided by images in reviews. Third, traditional multimodal fusion methods, such as feature concatenation, are insufficient for rating prediction tasks. These methods fail to fully exploit the interactions between modalities, resulting in suboptimal performance.

To address these gaps, we propose the Multimodal Multitask Rating Prediction (MMRP) model. MMRP leverages both textual and visual data, overcoming the limitations of single-modality approaches. By incorporating image data, the model captures visual cues that may

be overlooked in text, enabling a more comprehensive understanding of user sentiment. Furthermore, MMRP simultaneously addresses ACSA and ORP tasks, allowing the model to exploit the relationship between these tasks. This joint modeling approach improves the efficiency and accuracy of both tasks by sharing information across them. To enhance multimodal fusion, MMRP employs a cross-modal Transformer to integrate textual and visual information. It excels at matching highly related content across modalities. This aligns closely with our goal of addressing the ACSA problem. Therefore, we adopt the cross-modal Transformer as the primary structure for data fusion, aiming to achieve further improvements in both the ACSA and ORP tasks compared to existing approaches.

We note that neither multitask learning nor cross-modal attention/Transformer is new by itself. The novelty of this work lies in instantiating an end-to-end multimodal multitask framework for review mining that jointly solves ACSA and ORP from image–text reviews, a setting that is not addressed by representative prior work. In particular, ASAP (Aspect Category Sentiment Analysis and Rating Prediction) [10] is a joint ACSA–ORP model but it is text-only, while existing multimodal review models typically focus on a single task (either ACSA or ORP). Our MMRP integrates visual evidence into a shared representation that benefits both tasks and is specifically designed for the structure of e-commerce reviews, where text describes fine-grained aspects and images provide complementary evidence. Technically, rather than directly reusing a standard cross-modal Transformer, we introduce two task-tailored adaptations for multimodal reviews: modality-aware learnable positional encoding to distinguish textual token sequences from variable-sized image sets, and a gated multi-head cross-modal attention mechanism to suppress noisy visual interference. These designs make the fusion process better aligned with aspect-bearing text and sentiment-relevant visual evidence in image–text reviews.

This study makes the following three contributions:

- We introduce a novel multimodal, multitask model that can simultaneously address both ACSA and ORP tasks by jointly considering both textual and visual review content. This modeling approach leads to significant performance gains.
- The model integrates textual and visual information through a cross-modal Transformer, and further introduces modality-aware learnable positional encoding and a gated multi-head cross-modal attention mechanism to better handle the structure of multimodal reviews, where text forms token sequences and images appear as variable-sized sets.
- We conduct extensive experiments on both the ZOL mobile review dataset and an additional cross-domain benchmark, ViMACSA, to evaluate not only the effectiveness of MMRP in the source domain but also its transferability under realistic distribution shifts. In addition, we compare MMRP with a recent instruction-based multimodal baseline and provide an efficiency-oriented discussion on model complexity and deployment feasibility.

The remainder of the paper is organized as follows. Section 2 presents the related work. Section 3 illustrates in detail the proposed MMRP model. Section 4 presents the experimental settings. The results of comparison experiments, ablation experiments, sensitivity analysis, and case study are presented in Sect. 5. Finally, Sect. 6 concludes the paper.

2 Related work

2.1 Rating prediction

Rating prediction for reviews is an important research direction in the fields of data science and natural language processing, aiming to predict the ratings or sentiment tendencies corresponding to reviews by analyzing the user-generated review content [2, 14]. The research on the rating prediction for e-commerce platforms is of great value to both consumers and merchants [15–17]. Most of the existing rating prediction studies use only text information and focus only on overall rating prediction (ORP). Although text-only models have achieved steady progress, they still face two fundamental limitations in review mining. First, overall ratings are often a coarse-grained synthesis of multiple aspect-level opinions, and purely modeling ORP may overlook the fine-grained signals required for reliable interpretation and error analysis. Second, many existing approaches treat ACSA and ORP as separate problems, which prevents them from leveraging their strong correlation to improve both fine-grained aspect prediction and coarse-grained overall rating estimation. This motivates a unified framework that can jointly learn aspect-level sentiment and overall rating from shared representations. Feng et al. [18] introduce the InterSentiment model, which utilizes a deep neural network model to analyze user-product interactions and review sentiments for rating prediction. Wang et al. [19] propose a rating prediction model MRMRP based on multiple comments, which collects supplementary comments from similar users, extracts features from them, integrates them into different neural models, and improves prediction accuracy. Bauman and Tuzhilin [20] propose a new context parsing method to systematically extract context information from user-generated comments, predict ratings, and improve recommendation performance.

The goal of Aspect Category Sentiment Analysis (ACSA) is to predict users' sentiment expressions at each aspect category, so it is also essentially a rating prediction task. Moreover, [21] provided a comprehensive survey of aspect-based sentiment analysis, highlighting the strengths and limitations of existing techniques and pointing out promising directions for future multimodal sentiment research. For the ACSA task, traditional tokenization methods may fail to accurately capture the semantic content in review mining. The ARP model [22] aims to address this issue. The core idea of ARP is to adopt a multi-view learning framework to capture semantic information from different perspectives (i.e., words and characters), thereby predicting review ratings more accurately. Att-BiLSTM [23] model applies Bidirectional Long Short-Term Memory (BiLSTM) networks, combined with attention mechanisms, effectively processing and understanding sentence structures without relying on external lexical resources or complex natural language processing scenarios. Long et al. [24] propose a novel attention model trained by cognition grounded eye-tracking data, which firstly uses eye-tracking data to establish a reading prediction model, and then uses the predicted reading time to construct a cognitive-based attention layer for neural sentiment analysis.

To solve multiple tasks simultaneously, Bu et al. [10] propose a joint model which focuses on both sentiment analysis (ACSA) and rating prediction (ORP) tasks in e-commerce. This structure effectively integrates the two tasks, achieving the prediction of sentiment polarity of aspect categories and overall sentiment in user reviews, thus improving prediction accuracy and efficiency. The research indicates a high correlation between ACSA and ORP, often jointly used in practical e-commerce scenarios. Moreover, several multitask learning frameworks, such as MoE [25] and MMoE [26], have also been proposed. It is more appropriate and

concise to directly add the two designed loss functions together, because the two tasks are strongly related.

There are several multimodal studies that focus on rating prediction, using both picture and text information, but they are single-task models. For the ACSA problem, the MIMN [27] employed interactive learning methods to enhance analysis accuracy. The Co-Memory+Aspect [28] model uses MultiSentiNet module as a key part, aiming to enhance the accuracy of sentiment analysis by integrating textual and visual data. For the overall rating prediction problem, BiGRU-aVGG model [29] and HAN-aVGG [30] model are two examples. The BiGRU-aVGG model consists of a BiGRU model for processing textual data and a VGG model for processing images. For HAN-aVGG, the core idea of the Hierarchical Attention Network (HAN) model is to utilize the hierarchical structure of documents for document classification. Lin et al. [31] proposes a MSA hierarchical graph contrastive learning framework, which constructs a unimodal graph for each modal representation and employs graph contrastive learning strategy to explore potential relationships based on unimodal graph enhancement. Lu et al. [32] proposes a coordinated-joint translation fusion framework with affective interaction graph. Existing multimodal review models are typically task-specific: some models are designed for ACSA (e.g., MIMN and Co-Memory+Aspect) while others focus on ORP (e.g., BiGRU-aVGG and HAN-aVGG). As a result, they cannot exploit the mutual benefits between aspect-level and overall-level supervision. Moreover, many ORP-oriented multimodal methods adopt a relatively late or coarse fusion paradigm, where global visual features (e.g., VGG embeddings) are merged with textual representations without enforcing fine-grained alignment between aspect-related tokens and image evidence, which is crucial for aspect-aware reasoning in e-commerce reviews.

In contrast, MMRP is explicitly designed for multimodal and multitask review mining: it jointly solves ACSA and ORP in one framework and adopts a cross-modal Transformer to model interaction-aware fusion, enabling the model to align and integrate complementary cues between review text and images for both tasks.

2.2 Multimodal sentiment analysis

Multimodal sentiment analysis aims to infer sentiment signals by jointly modeling heterogeneous modalities (e.g., text and images), and its performance heavily depends on how cross-modal representations are aligned and fused. In multimodal review mining, the core of multimodal feature fusion is to unify features from different modalities into a consistent form. The choice of fusion strategy significantly affects the quality of feature extraction and downstream prediction performance. Existing multimodal sentiment models can be broadly grouped into three categories, while recent advances increasingly emphasize fine-grained cross-modal alignment and multi-granularity fusion. For example, Yu et al. [33] further developed a multi-grained feature gating fusion network to better integrate heterogeneous modal features for multimodal sentiment analysis.

A straightforward fusion strategy is to concatenate unimodal representations, as used in early multimodal review analysis by Ma et al. [34]. Hu and Flaxman [35] developed the Deep Sentiment model, which combines GloVe word vectors and pre-trained Inception networks; with LSTM networks and feature concatenation, it obtains multimodal features for sentiment prediction. Such methods are simple and intuitive, but they often fail to explicitly model cross-modal interactions, because concatenation itself does not enforce meaningful alignment between modalities and relies on subsequent neural layers to implicitly learn how to exploit multimodal cues.

To capture richer multiplicative interactions, bilinear pooling methods (e.g., low-rank bilinear pooling and separable bilinear pooling) have been widely explored, especially in the Visual Question Answering (VQA) domain [36, 37]. Representative approaches include Multimodal Compact Bilinear (MCB) pooling Fukui et al. [38] and Multimodal Low-rank Bilinear (MLB) pooling Kim et al. [39]. MCB approximates the outer product between image and text features in a low-dimensional space, while MLB uses Hadamard product and linear projections to approximate bilinear fusion with reduced computational complexity. Although bilinear pooling can strengthen cross-modal feature interactions, it may still be limited in capturing long-range or aspect-aware alignment required by fine-grained multimodal sentiment understanding.

More recent multimodal sentiment models adopt attention mechanisms to achieve adaptive and interpretable fusion. For example, Truong and Lauw [40] treat images as auxiliary signals and learn attention scores between image and text features to enhance textual sentiment representations. Cross-modal Transformer-based fusion further advances this line of work by explicitly using one modality as Query and the other as Key/Value to perform targeted multi-head attention, enabling more precise cross-modal interaction modeling [41, 42]. Beyond generic attention, the latest multimodal sentiment research has increasingly focused on multi-grained alignment, knowledge-guided dynamic modality weighting, and multi-level fusion. Specifically, Lin et al. [43] propose a semantic-guided multi-grained cross-modal alignment and fusion network for multimodal aspect-based sentiment analysis, highlighting the importance of aligning semantics across modalities at different granularities. Feng et al. [44] introduce a knowledge-guided dynamic modality attention fusion framework, where external knowledge is leveraged to guide modality attention and improve robustness under modality ambiguity or noise. Li et al. [45] further emphasize multi-level textual-visual alignment and fusion for multimodal ABSA, suggesting that alignment should be performed at multiple representation layers rather than only at a single fusion stage. Yu et al. [46] further proposed a dual syntactic graph network with joint contrastive learning for multimodal aspect-based sentiment analysis, showing that explicitly modeling both textual and visual dependencies can improve fine-grained sentiment prediction. In addition to sentiment polarity prediction, multimodal affective understanding tasks such as sarcasm detection also benefit from dynamic and multi-granularity fusion, as shown by He et al. [47], who propose a dual dynamic multi-granularity fusion method to capture subtle incongruity across modalities.

2.3 Large language models (LLMs) and vision–language models (VLMs)

Recent LLMs have shown strong general-purpose performance and have been increasingly explored for sentiment-related tasks via prompting or fine-tuning [48, 49]. However, many widely used LLM backbones remain primarily text-only (e.g., encoder-decoder LMs) and therefore cannot exploit visual evidence in image–text reviews without additional vision encoders or multimodal interfaces [50]. In multimodal settings, VLMs integrate visual representations with language modeling and can potentially support multimodal sentiment reasoning; nevertheless, they typically require task-specific adaptation for aspect category–level prediction and rating regression, especially under missing-aspect annotations [51–53].

In this work, we focus on a lightweight multimodal multitask framework for joint ACSA and ORP. We view LLM/VLM approaches as complementary: they can be used as strong teachers (e.g., pseudo-labeling or prompt-based reasoning), while our model offers a deployable alternative with lower inference cost for large-scale review mining.

Overall, these studies indicate that effective multimodal sentiment modeling requires not only combining heterogeneous features, but also ensuring where and how modalities interact (e.g., alignment granularity, dynamic modality importance, and layered fusion). Motivated by these insights, our model adopts a cross-modal Transformer style fusion to explicitly model directional cross-modal interactions, which is particularly suitable for multimodal reviews where text and images often describe correlated product aspects and sentiment cues.

3 Methodology

3.1 Problem definition

In the context of multimodal product reviews, let the textual review of user u on product p be denoted as $R_{u\to p} = \{w_1, w_2, \ldots, w_L\}$, where w_j represents the j-th word, and L is the length of the review. Let the set of images corresponding to this review be denoted as $I_{u\to p} = \{i_1, i_2, \ldots, i_C\}$, where C represents the number of images attached to the review, and i_k denotes the k-th image.

This study focuses on two rating tasks in the domain of review mining: Aspect Category Sentiment Analysis (ACSA) and Overall Rating Prediction (ORP). The ACSA task aims to predict the score y_i of the review on aspect category a_i, based on the textual review R and accompanying images I. Here, $i \in \{1, 2, \ldots, N\}$, where N is the number of aspect categories in the set. To address the issue of some aspect categories not being discussed in the review, a masking vector $P = [p_1, p_2, \ldots, p_N]$ is introduced to indicate the presence of aspect categories in the set. When aspect category a_i is mentioned in the review, $p_i = 1$; otherwise, $p_i = 0$. Assuming that K aspect categories are mentioned in the review, we have $\sum_{i=1}^{N} p_i = K$. The ORP task aims to predict the overall rating g of the review based on the textual review R and accompanying images I.

For both ACSA and ORP, the ratings are usually given by the reviewers, according to the design of the platform. In different practical settings, y_i and g can take various forms, such as binary classification into positive or negative sentiment or a numerical score ranging from 1 to 10.

3.2 MMRP model

This study proposes a comprehensive model, namely the Multimodal Multitask Rating Prediction (MMRP), consisting of three key modules to achieve a more comprehensive understanding of multimodal data and accomplish the multitask rating prediction tasks. The structure of the MMRP model is illustrated on the left side of Fig. 1.

Firstly, this study introduces a feature extraction module aimed at fully exploiting the rich information from text and images. For the input text data, BERT is used for feature extraction to obtain text representations. For image data, VGG model is employed to extract image feature representations. The objective of this stage is to capture rich information from multimodal data and convert text and images into feature representations for subsequent processing.

Secondly, a cross-modal fusion module is introduced, which serves as the core of the entire model. In this module, our study adopts a cross-modal Transformer model to fuse text features with image features. Transformer encoders with S layers are utilized to interactively learn the relationships between text features and image features, enabling the model to better

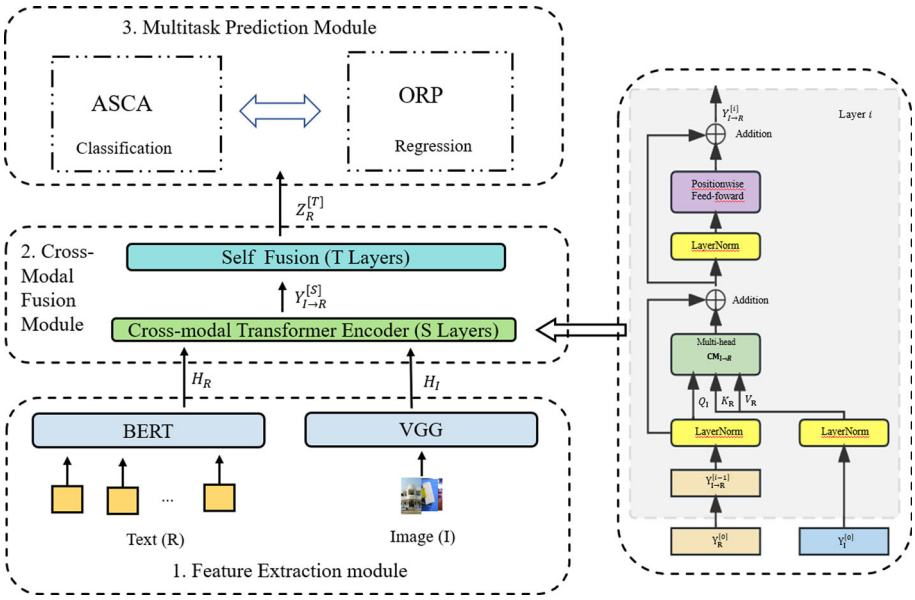


Fig. 1 The structure of the MMRP

understand the correlation between text and images, followed by a T-layer self-fusion module to further process features. The goal of this step is to preserve and enhance the semantic correlations between text and images after fusion, providing more informative inputs for subsequent tasks.

Finally, a multitask prediction module is designed. This module is responsible for predicting ACSA and ORP labels. Leveraging the features processed by the cross-modal fusion module, an appropriate classifier and regression model are utilized for label prediction. The aim of this step is to translate the deep cross-modal information learned by the model into meaningful outputs for specific tasks, achieving accurate prediction of ratings.

In the following, we will give detailed descriptions for these modules.

3.2.1 Feature extraction module

For the input textual review $R = \{w_1, w_2, ..., w_L\}$, tokenization processing is applied to convert it into the input format required by the BERT model[1] The architecture of the BERT model consists primarily of encoders, which are designed as a multi-layer stacked Transformer structure. Each layer consists of a multi-head self-attention mechanism and a feedforward neural network, which allows the model to focus on different parts of the input sequence at the same time. In each Transformer encoder layer, the BERT model encodes bidirectional context information through multi-head self-attention mechanisms. This mechanism allows the model to focus on information in different locations simultaneously when processing input sequences, rather than just being limited to one-way contexts. Specifically, the multi-head self-attention mechanism allows the model to compute multiple different attention weights, enabling each position to fuse information from the entire input sequence. In addition, each

[1] BERT pre-trained model: https://huggingface.co/google-bert/bert-base-chinese.

encoder layer also includes a feedforward neural network layer to further process the output of the self-attention mechanism. The feedforward neural network can extract higher level semantic information and enhance the representation ability of the model for the input sequence. With the multi-layer stacked Transformer encoder, BERT model can gradually learn and extract complex semantic structures from input sequences to form more abstract and rich representations. Through processing by the BERT model, the final hidden layer matrix of the textual review $H_R : \{h_{R1}, h_{R2}, ..., h_{RL}\}$ is obtained, where $H_R \in \mathbb{R}^{L \times d_R}$, d_R is the dimension of the BERT hidden layer, and L represents the length of the review. This processing step retains rich semantic information for each word in the review, providing strong textual features for subsequent multimodal fusion.

For the image data, feature extraction is performed using the VGG16 model.[2] Compared with traditional CNN networks, VGG uses deeper network layers, and the size of the convolutional kernel is unified to 3×3. This design makes the model more expressive, and can better capture the features of the input data. The VGG model suggests that using multiple smaller convolution kernels is preferable to using a single larger convolution kernel. By continuously using the small convolution kernel of 3 × 3, the VGG model can improve the representation ability of the network. This design has made the VGG model a classic model in deep learning. The VGG16 network structure used in this study is as follows, which is shown in Fig. 2.

- Input Layer: 224 × 224 colored images
- Convolutional Layer: 13 convolutional layers, each using 3 × 3 small convolutional kernels with a stride of 1
- Pooling Layer: 4 2 × 2 max-pooling layers
- Fully Connected Layer: 3 fully connected layers, with the final layer outputting the probability corresponding to each category

For the image set $I : \{i_1, i_2, ..., i_C\}$, after processing by the VGG model, the final hidden layer of the image is extracted as the image feature, resulting in the image feature matrix $H_I : \{h_{I1}, h_{I2}, ..., h_{IC}\}$, where $H_I \in \mathbb{R}^{C \times d_I}$, d_I is the dimension of the image feature, and C represents the number of images. This step helps retain advanced semantics and features of the images, providing strong support for the subsequent processing.

According to these two pre-trained models, BERT and VGG, high-dimensional feature matrices H_R and H_I are obtained for review texts and images, respectively. This provides rich data representations for subsequent modal fusion, laying a solid foundation for improving the performance of deep learning models.

3.2.2 Cross-modal fusion module

MulT [54] targets unaligned multimodal language sequences (e.g., audio/video/text streams) and applies cross-modal attention blocks across sequential modalities. Our scenario differs in that e-commerce reviews consist of text and a set (or weakly ordered list) of images, where text is the primary carrier of aspect categories and images provide auxiliary evidence.

To fit this structure, we employ a directional cross-modal Transformer encoder ($I \rightarrow R$) that injects image information into the text representation (text as the target modality for aspect reasoning), and then apply a T-layer unimodal Transformer self-fusion to refine the fused features before multitask prediction. This two-stage design is intended to (i) align visual cues with aspect-relevant tokens and (ii) consolidate cross-modal evidence into a compact representation shared by both ACSA and ORP.

[2] VGG16 pre-trained model: https://download.pytorch.org/models/vgg16-397923af.pth.

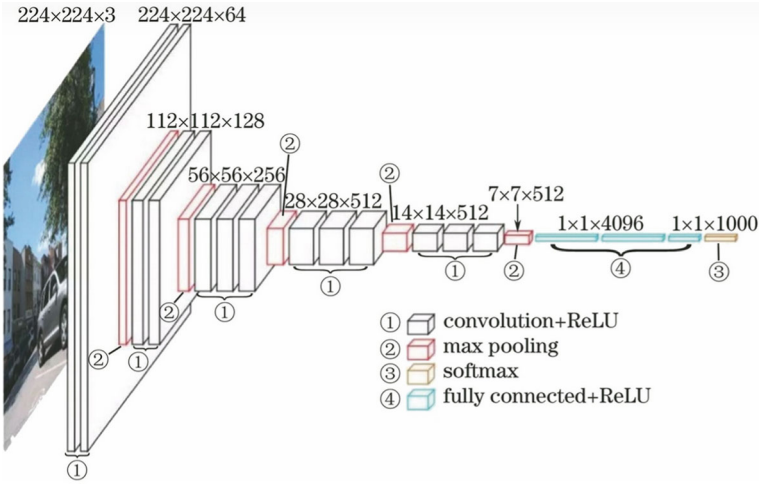


Fig. 2 VGG16 network structure

After obtaining the textual and image features, a cross-modal Transformer is used to fuse these two modalities, as shown on the right side of Fig. 1.

This choice is well-suited for our task because textual and visual product reviews are inherently correlated: both modalities describe the same product aspects and reflect consistent sentiment tendencies. Cross-modal Transformers are effective at aligning and matching highly related content across modalities, which aligns closely with our goal of addressing ACSA. Therefore, we adopt a cross-modal Transformer as the primary fusion backbone, aiming to further improve ORP performance while jointly optimizing ACSA.

A standard Transformer does not encode ordering information by itself, so positional encoding is typically introduced. Although positional encoding is a standard component, our multimodal review setting has two notable properties: (i) the text modality is a token sequence, while (ii) the image modality is a set (or weakly ordered list) of images with a variable number of elements per review. To better fit this structure, we adopt a modality-aware learnable positional encoding with separate learnable positional embeddings and modality-type embeddings for the two modalities. Let the text features be $H_R \in \mathbb{R}^{L\times d_R}$ and the image features be $H_I \in \mathbb{R}^{C\times d_I}$, where L is the text length and C is the number of images. We project them into a shared dimension d_k and add modality-aware embeddings:

$$Y_R^{[0]} = H_R W_R + E_R^{pos} + \mathbf{1}_L (e_R^{type})^\top, \tag{1}$$

$$Y_I^{[0]} = H_I W_I + E_I^{pos} + \mathbf{1}_C (e_I^{type})^\top, \tag{2}$$

where $W_R \in \mathbb{R}^{d_R\times d_k}$ and $W_I \in \mathbb{R}^{d_I\times d_k}$ are linear projections, $E_R^{pos} \in \mathbb{R}^{L\times d_k}$ and $E_I^{pos} \in \mathbb{R}^{C\times d_k}$ are learnable positional embeddings for text token indices and image indices, $e_R^{type}, e_I^{type} \in \mathbb{R}^{d_k}$ are modality-type embeddings to explicitly distinguish the two modalities, and $\mathbf{1}_L, \mathbf{1}_C$ are all-one vectors for broadcasting.

This change keeps the Transformer backbone intact while making the positional modeling review-structure–aware and compatible with variable C.

To achieve cross-modal feature fusion from image to text, we introduce a multi-head attention mechanism. Following the $(I \rightarrow R)$ direction, we treat text as the target modality (Query) and *images as the source modality* (Key/Value), so that the fused representation

preserves the same length L as the text tokens. For the m-th head ($m = 1, \ldots, M$), we define:

$$Q_R^m = \mathrm{LN}(Y_{(I\to R)}^{[i-1]})W_Q^m, \tag{3}$$

$$K_I^m = \mathrm{LN}(Y_I^{[0]})W_K^m, \tag{4}$$

$$V_I^m = \mathrm{LN}(Y_I^{[0]})W_V^m, \tag{5}$$

where $W_Q^m, W_K^m, W_V^m \in \mathbb{R}^{d_k\times d_k}$ and $\mathrm{LN}(\cdot)$ denotes layer normalization. The cross-modal attention output for head m is:

$$\mathrm{CM}_{(I\to R)}^m(Y_{(I\to R)}^{[i-1]}, Y_I^{[0]}) = \mathrm{softmax}\left(\frac{Q_R^m(K_I^m)^\top}{\sqrt{d_k}}\right)V_I^m \in \mathbb{R}^{L\times d_k}. \tag{6}$$

To reduce the impact of irrelevant images and avoid over-reliance on a vanilla formulation, we introduce a lightweight head-wise gate to adaptively re-weight each head:

$$g_m^{[i]} = \sigma\left((w_g^m)^\top\left[\mathrm{pool}(Y_{(I\to R)}^{[i-1]}); \mathrm{pool}(Y_I^{[0]})\right] + b_g^m\right), \quad g_m^{[i]} \in (0, 1), \tag{7}$$

where $\mathrm{pool}(\cdot)$ denotes mean pooling, $\sigma(\cdot)$ is the sigmoid function, and w_g^m, b_g^m are learnable parameters. The gated multi-head cross-modal attention is then given by:

$$\mathrm{HGCM}^{[i]}(Y_{(I\to R)}^{[i-1]}, Y_I^{[0]}) = \mathrm{Concat}_{m=1}^M\left(g_m^{[i]}\cdot \mathrm{CM}_{(I\to R)}^m(Y_{(I\to R)}^{[i-1]}, Y_I^{[0]})\right)W_O, \tag{8}$$

where $W_O \in \mathbb{R}^{(Md_k)\times d_k}$. This gating is a minor adjustment but helps suppress noisy cross-modal interactions (e.g., irrelevant review images) and emphasizes sentiment-relevant heads.

Stacked cross-modal Transformer encoder.

Based on the above head-gated cross-modal attention, we further define a cross-modal Transformer module consisting of S layers. Each layer integrates image information into the fused multimodal representation from the previous layer:

$$\hat{Y}_{(I\to R)}^{[i]} = \mathrm{HGCM}^{[i]}\left(\mathrm{LN}(Y_{(I\to R)}^{[i-1]}), \mathrm{LN}(Y_I^{[0]})\right) + \mathrm{LN}(Y_{(I\to R)}^{[i-1]}), \quad i = 1, \ldots, S, \tag{9}$$

$$Y_{(I\to R)}^{[i]} = f_{\theta_i}\left(\mathrm{LN}(\hat{Y}_{(I\to R)}^{[i]})\right) + \mathrm{LN}(\hat{Y}_{(I\to R)}^{[i]}), \quad i = 1, \ldots, S, \tag{10}$$

with initialization

$$Y_{(I\to R)}^{[0]} = Y_R^{[0]}. \tag{11}$$

Here f_{θ_i} represents the position-wise feedforward sublayer in the i-th layer of the cross-modal encoder.

After S layers, we obtain the fused feature $Y_{(I\to R)}^{[S]} \in \mathbb{R}^{L\times d_k}$.

Next, we apply a unimodal Transformer with T layers to further refine the fused representation. Multiple rounds of self-fusion allow the information from both modalities to be thoroughly consolidated and the most salient features to be captured. For the i-th layer ($i = 1, \ldots, T$), we compute:

$$\hat{Z}_R^{[i]} = \mathrm{SA}^{[i]}\left(\mathrm{LN}(Z_R^{[i-1]})\right) + \mathrm{LN}(Z_R^{[i-1]}), \tag{12}$$

$$Z_R^{[i]} = g_{\theta_i}\left(\mathrm{LN}(\hat{Z}_R^{[i]})\right) + \mathrm{LN}(\hat{Z}_R^{[i]}), \tag{13}$$

with

$$Z_R^{[0]} = Y_{(I\to R)}^{[S]}. \tag{14}$$

Here $\mathrm{SA}^{[i]}(\cdot)$ denotes the standard multi-head self-attention sublayer in the i-th unimodal encoder layer, and g_{θ_i} represents the position-wise feedforward sublayer. After T layers, the output feature $Z_R^{[T]} \in \mathbb{R}^{L\times d_k}$ is obtained and fed into the subsequent multitask prediction module.

3.2.3 Multitask prediction module

In the ACSA task, for each aspect category a, the following formula is used for calculating the score M_i^a at each hidden layer i.

$$M_i^a = \tanh\left(Z_R^{[T]} W_i^a\right), \tag{15}$$

where M_i^a represents the characteristic representation of the i-th aspect class throughout the process, which is the intermediate quantity further used for the softmax operation, $W_i^a \in \mathbb{R}^{(d_k\times d_k)}$ is the weight matrix for attention of the aspect category. Then, the softmax operation is used to compute attention weights α_i:

$$\alpha_i = \mathrm{softmax}(M_i^a \omega_i), \tag{16}$$

where $\omega_i \in \mathbb{R}^{d_k}$ is the weight vector. The attention representation r_i for the i-th aspect category a_i of the comment is calculated using the following formula:

$$r_i = \tanh\left(\alpha_i^T Z_R^{[T]} W_i^p\right), \tag{17}$$

where $W_i^p \in \mathbb{R}^{(d_k\times d_k)}$ is the weight matrix for comment attention.

Finally, the score prediction for the ACSA task is performed using the following formula:

$$\hat{y}_i = \mathrm{softmax}(W_i^q r_i + b_i^q), \tag{18}$$

where $W_i^q \in \mathbb{R}^{(N\times d_k)}$ is the weight matrix for the ACSA task, and $b_i^q \in \mathbb{R}^q$ is the bias vector. $\hat{y}_i$ represents the predicted probability vector for the i-th aspect a_i, where each component represents the probability of a_i being classified into the corresponding category, and the true label y_i is a one-hot vector.

The loss function for the ACSA task, taking a classification problem with K categories as an example, is defined as entropy loss:

$$\mathrm{loss}_{\mathrm{ACSA}} = -\frac{1}{T}\sum_{i=1}^{N} p_i \sum_{k=1}^{K} y_{ik}\log \hat{y}_{ik}, \tag{19}$$

where N is the total number of aspects, $p_i \in \{0, 1\}$ indicates whether the i-th aspect is mentioned, $K = \sum_{i=1}^{N} p_i$ is the number of mentioned aspects for the sample.

In the ORP task, we focus on overall comment ratings, the first embedding after self-fusion $h_R = Z_R^{[i]}[0; :] \in \mathbb{R}^{d_k}$ is used to predict g. Specifically, the calculation is performed using the following formula:

$$\hat{g} = \beta^T \tanh(W^r h_R + b^r), \tag{20}$$

where $\beta \in \mathbb{R}^{d_k}$, $W^r \in \mathbb{R}^{(d_k\times d_k)}$, and $b^r \in \mathbb{R}^{d_k}$.

The loss function for the ORP task, taking a regression problem with numeric values as an example, is defined as absolute loss:

$$\mathrm{loss}_{\mathrm{ORP}} = |\hat{g} - g| \tag{21}$$

Finally, the losses for the ACSA and ORP tasks are combined to improve the model performance by jointly considering these two tasks to more comprehensively understand user reviews and predict overall attitudes:

$$\text{loss} = \text{loss}_{\text{ACSA}} + \text{loss}_{\text{ORP}} \tag{22}$$

Differences from multitask baselines (e.g., ASAP[10]). Unlike ASAP, which performs joint learning on textual reviews only, our multitask heads are built upon the shared multimodal representation produced by cross-modal alignment and self-fusion. Moreover, the ACSA head is aspect-category-aware and uses a masking mechanism to handle the fact that only a subset of aspect categories is mentioned in each review, while the ORP head predicts an overall rating from the fused representation. This design enables cross-task regularization: fine-grained aspect signals can inform overall rating prediction and vice versa, while visual cues can support both tasks through the shared-bottom layers.

4 Experiment settings

4.1 ZOL dataset

This study evaluates the models on the ZOL Mobile Phone Reviews dataset from Zhongguancun Online (ZOL) [27][3] ZOL is a leading IT information and commercial portal website in China. It consists of 40 major channels covering news, shopping centers, hardware, downloads, games, mobile phones, etc. The dataset contains 12587 reviews, of which 7359 are single-modal reviews containing only text information, and the rest are multimodal reviews with both texts and images, covering a total of 114 brands and 1318 types of mobile phones.

In this dataset, each multimodal review contains a piece of text review, an image set, and at least one but no more than six aspect categories. These six aspect categories are value for money, performance & configuration, battery life, design & feel, camera quality, and display quality. Each aspect category corresponds to an integer sentiment score ranging from 1 to 10, used as labels for the ACSA classification task. Additionally, the overall review also has a numerical sentiment score, used as the label for the ORP regression task.

4.2 Baseline models

We selected eight baseline models to validate the performance of our proposed model. These eight models can be divided into two groups according to whether they use multimodal information. Att-BiLSTM [23], ARP [22] ASAP [10] and Chimera [55] are single-modal models using only textual information, while MIMN [27], Co-Memory+Aspect [28], HAN-aVGG [30], BiGRU-aVGG [29] models are multimodal models using both textual and visual information. The specific details of the eight baseline models are shown in Table 1. The proposed MMRP model in this study is the only multimodal model that can simultaneously perform both tasks.

Among the four single-modal pure text models, Att-BiLSTM, ARP and Chimera are single-task models, and they were used separately for the ACSA task or the ORP task in the experimental comparison process. The ASAP model is a multitask model that can simultaneously output results for both the ACSA task and the ORP task for comparison. Among the four multimodal models, MIMN and Co-Memory+Aspect were proposed to handle the

[3] The dataset can be downloaded from https://github.com/xunan0812/MIMN.

Table 1 Details of the baseline models

Model	ACSA task	ORP task	Multitask	Multimodal
Att-BiLSTM	✓	✓	✗	✗
ARP	✓	✓	✗	✗
ASAP	✓	✓	✓	✗
MIMN	✓	✗	✗	✓
Co-Memory+Aspect	✓	✗	✗	✓
HAN-aVGG	✗	✓	✗	✓
BiGRU-aVGG	✗	✓	✗	✓
Chimera	✓	✗	✗	✗
MMRP	✓	✓	✓	✓

ACSA task, while HAN-aVGG and BiGRU-aVGG were used for the ORP task. They were used to compare the ACSA and ORP results output by MMRP, respectively.

The baseline models are selected based on three considerations: task compatibility, representative coverage, and recency. We include representative text-only, multimodal, single-task, and multitask methods that are comparable under the unified ACSA/ORP evaluation setting. In particular, given the strong performance of recent large language models across multiple tasks, we additionally include Chimera, a large-model-based baseline, to compare with our proposed model.

4.3 Evaluation metrics

For the ACSA classification task, this study employs accuracy (Acc.) and Macro-F1 score as evaluation metrics. The accuracy is defined as follows:

$$\text{Acc.} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Observations}} \tag{23}$$

Macro-F1 can better evaluate balanced performance across rating categories and is less affected by class imbalance than Accuracy. The Macro-F1 score is defined as the arithmetic mean of the class-wise F1-scores:

$$\text{Macro-F1} = \frac{1}{K}\sum_{k=1}^{K}\frac{2P_kR_k}{P_k + R_k}, \tag{24}$$

where K denotes the number of classes. For the k-th class, precision P_k and recall R_k are defined as

$$P_k = \frac{TP_k}{TP_k + FP_k}, \qquad R_k = \frac{TP_k}{TP_k + FN_k},$$

where TP_k, FP_k, and FN_k denote the numbers of true positives, false positives, and false negatives for the k-th class, respectively. Macro-F1 computes the F1-score for each class independently and then averages them across all classes, thereby assigning equal importance to each class regardless of its sample size.

For the ORP regression task, Mean Absolute Error (MAE) is used for evaluation.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{25}$$

The original ORP scores ranging from 1 to 10 are further treated as discrete categories. Specifically, the predicted ORP scores are rounded to the nearest integer and then evaluated as a classification problem. Based on this setting, Macro-F1 and Accuracy are additionally adopted as evaluation metrics.

Besides predictive performance, the practical applicability of a multimodal multitask framework should also be evaluated in terms of computational cost and real-world feasibility. We therefore further examine MMRP from four aspects: computational complexity, time efficiency, scalability under different data scales, and deployment feasibility. The results are shown in Sect. 5.4.

4.4 Environment and hyperparameter settings

For the proposed MMRP model, Adam optimizer is used for training. The model is implemented in the PyTorch framework, and all experiments are conducted on an NVIDIA TITAN V 12 G GPU. Specific hyperparameter settings are described as follows: learning rate is set to 0.001, batch size is set to 32, maximum sequence length is set to 512, attention dimension is set to 80, the number of layers in the cross-modal Transformer encoder S and the transformer encoder T are both set to 6, and the number of attention heads is set to 8. Other baseline models adopt hyperparameters specified in their respective papers or source code. All experiments are repeated 10 times.

5 Experiment results

5.1 Results of model comparison

5.1.1 ACSA task

In order to comprehensively evaluate the results of the proposed MMRP, this study not only considers the performance indicators but also takes into account the stability of the models. The mean of the indicators for 10 repeated experiments is used as the evaluation of model performance, and the standard deviation of the 10 experiments is used as the evaluation of model stability. The better the mean indicator, the higher the accuracy of the model; the smaller the standard deviation of the indicator, the better the stability of the model.

Table 2 reports the comparison results for the ACSA task. The best model results are indicated in bold. The significance marks are added to the baseline models to indicate that their results are used as reference values in subsequent significance comparison, whereas no significance marks are attached to MMRP and MMRP (w/o ORP), since they serve as the proposed model and its ablated variant rather than comparison targets.

Among the four single-modal models, Att-BiLSTM achieves the highest Macro-F1, reaching 62.69%. Chimera obtains the highest Accuracy, at 75.07%, but its Macro-F1 is the lowest among these models, only 44.09%. This suggests that although Chimera performs well in terms of overall accuracy, its performance across different aspect sentiment categories is highly imbalanced. In contrast, the proposed MMRP model achieves a Macro-F1 of 63.27%, outperforming all the single-modal models. Its Accuracy is 66.83%, which is lower than that of Chimera but higher than that of ARP and ASAP, and slightly higher than Att-BiLSTM. Considering that Macro-F1 is a more comprehensive metric for the ACSA task. In particular, Accuracy tends to mislead in imbalanced classification settings, because a model can obtain

Table 2 Model comparison results for ACSA task

Model type	Model name	Macro-F1	Acc
Single modal	Att-BiLSTM	62.69%***	66.33%***
		(0.0027)	(0.0051)
	ARP	44.43%***	56.84%***
		(0.0037)	(0.0049)
	ASAP	59.78%***	63.57%***
		(0.0029)	(0.0033)
	Chimera	44.09%***	**75.07%****
		(0.0008)	(0.0058)
Multimodal	MIMN	61.27%***	59.48%***
		(0.0038)	(0.0037)
	Co-Memory+Aspect	60.57%***	58.37%***
		(0.0034)	(0.0026)
	MMRP Model	62.37%	66.64%
	(w/o ORP)	(0.0022)	(0.0027)
	MMRP Model	**63.27%**	66.83%
		(0.0030)	(0.0025)

The numbers in brackets are standard deviations.
***significant at 0.001 level
Bold values indicate the best performance for each metric

a high Acc value by over-predicting the majority classes while performing poorly on minority classes. Therefore, we argue that the proposed MMRP model provides a better balance between classification effectiveness and robustness, and is thus more suitable for the ACSA task.

In the multimodal models, the MIMN model and the Co-Memory+Aspect model exhibit higher accuracy compared to others, with MIMN's Macro-F1 and Acc at 61.27% and 59.48%, and Co-Memory+Aspect's Macro-F1 and Acc at 60.57% and 58.37%, respectively. This suggests that different model structures and modal fusion methods for multimodal models have a significant impact on downstream tasks. At the same time, comparing the experimental results of single-modal and multimodal models, it is observed that the optimal single-modal model Att-BiLSTM outperforms MIMN in experimental accuracy, indicating that simply adding multimodal information does not necessarily improve model accuracy; rather, selecting appropriate modal fusion methods is essential for improving model accuracy.

Additionally, we further examine the impact of the multitask structure on model accuracy, where MMRP (w/o ORP) denotes the experimental accuracy obtained by exclusively handling the ACSA task without considering the ORP task loss function. It was found that simultaneously handling ACSA and ORP tasks results in higher experimental accuracy than dealing with the ACSA task alone, indicating that information between the two tasks complements each other, and the information regarding the entire comment rating in the ORP task can help improve the judgment of attitude toward aspect categories.

Table 3 Model comparison results for ORP task

Model type	Model name	Macro-F1	Acc	MAE
Single-modal	Att-BiLSTM	51.57%***	53.59%***	0.6531***
		(0.0049)	(0.0136)	(0.0086)
	ARP	42.62%***	51.28%***	0.7707***
		(0.0016)	(0.0151)	(0.01326)
	ASAP	32.77%***	42.58%***	0.9445***
		(0.0017)	(0.0030)	(0.0049)
Multimodal	HAN-aVGG	48.51%***	52.40%***	0.7380***
		(0.0054)	(0.0068)	(0.0198)
	BiGRU-aVGG	52.04%***	53.71%***	0.6765***
		(0.0061)	(0.0022)	(0.0024)
	MMRP	53.68%	52.99%	0.6575
	(w/o ACSA)	(0.0017)	(0.0098)	(0.0125)
	MMRP	**57.33%**	**60.26%**	**0.5663**
		(0.0043)	(0.0031)	(0.0079)

The numbers in brackets are standard deviations
***significant at 0.001 level
Bold values indicate the best performance for each metric

5.1.2 ORP task

The evaluation metrics for the ORP task are similar to the ACSA task, considering both the mean metrics and the standard deviations of 10 repeated experiments. The improvement ratio of the proposed models over the baseline models is computed, and t-tests are performed. The mean results for ORP task metrics are shown in Table 3, with the corresponding standard deviations in the brackets below.

As in Table 2, significance marks in Table 3 are attached only to the baseline models, since they serve as the reference models in subsequent statistical comparisons, whereas MMRP and MMRP (w/o ACSA) are the proposed model and its ablated variant.

Among the single-modal baselines, Att-BiLSTM achieves the best overall performance, with a Macro-F1 of 51.57%, an Accuracy of 53.59%, and an MAE of 0.6531. Although its Macro-F1 standard deviation is 0.0049, which is the largest among the single-modal models, its overall results are still consistently better than those of ARP and ASAP. In contrast, the proposed MMRP model achieves the best ORP performance, with a Macro-F1 of 57.33%, an Accuracy of 60.26%, and an MAE of 0.5663. Compared with ASAP, MMRP improves Macro-F1 and Accuracy by 74.95% and 41.52%, respectively, while reducing MAE by 40.05%. Even when compared with Att-BiLSTM, the strongest single-modal baseline, MMRP still achieves clear gains of 11.16% in Macro-F1 and 12.46% in Accuracy, together with a 13.30% reduction in MAE. These results demonstrate that incorporating visual information can substantially improve the performance of overall rating prediction.

Among the multimodal baselines, HAN-aVGG and BiGRU-aVGG show relatively comparable performance, while BiGRU-aVGG performs slightly better overall, achieving higher Macro-F1 and lower MAE than HAN-aVGG, as well as smaller standard deviations in Accuracy and MAE. Notably, BiGRU-aVGG also outperforms all the single-modal baselines in terms of Accuracy. Nevertheless, the proposed MMRP model still delivers the best overall

Table 4 Ablation results on ZOL

Configuration	ORP			ACSA	
	Macro-F1 (%)	Acc (%)	MAE (%)	Macro-F1 (%)	Acc (%)
BERT only	40.06	46.41	0.7688	44.30	54.07
+ Cross-modal	44.59	47.81	0.7085	61.91	66.18
Full MMRP	**57.33**	**60.26**	**0.5663**	**63.27**	**66.83**

Bold values indicate the best performance for each metric

results among all multimodal models, while maintaining relatively small experimental standard deviations. Compared with BiGRU-aVGG, MMRP improves Macro-F1 by 10.16%, Accuracy by 12.20%, and reduces MAE by 16.29%. These findings further verify that the proposed multimodal interaction mechanism is more effective for the ORP task than existing visual-text fusion methods.

Furthermore, for the ORP task, we further examine the impact of the multitask structure, where MMRP (w/o ACSA) denotes the results obtained by exclusively handling the ORP task without considering the ACSA task. Similar to the conclusions drawn from the ACSA task, the experiments reveal that simultaneously handling ACSA and ORP tasks results in higher experimental accuracy than handling single tasks. According to the experimental results, the contribution of ACSA supervision to ORP is more pronounced, indicating that integrating fine-grained aspect category scores complements overall score prediction, thereby achieving better prediction results.

5.2 Ablation experiment

The design of the model structure has a significant impact on performance. To validate the effectiveness of the proposed model architecture, we investigate the role of core modules of the proposed model, which are cross-modal Transformer and the self-attention mechanism. The model that only utilizes the BERT model to extract text encoding for rating prediction serves as the baseline for comparison.

As summarized in Table 4, each stage in the proposed pipeline contributes measurable gains on ZOL. Starting from the text-only BERT baseline, introducing the cross-modal Transformer yields the most substantial improvement for ACSA, boosting Macro-F1 from 44.30 to 61.91% (+17.61 points) and Accuracy from 54.07 to 66.18% (+12.11 points), which highlights the importance of explicit image–text fusion for aspect-level sentiment understanding. For ORP, cross-modal fusion brings a modest but consistent benefit (Accuracy: 46.41% $\rightarrow$ 47.81%, +1.40 points; MAE: 0.7688 $\rightarrow$ 0.7085), indicating that multimodal alignment alone improves rating prediction but is not sufficient.

After further adding the refinement stage (self-fusion/self-attention) and the full multitask learning objective, the complete MMRP achieves the best overall performance, reaching 63.27%/66.83% for ACSA Macro-F1/Accuracy and 57.33%/60.26% for ORP Macro-F1/Accuracy, while reducing MAE to 0.5663. Compared with the cross-modal-only variant, this final step produces a relatively small additional gain for ACSA (+1.36 Macro-F1 and +0.65 Accuracy) but a much larger improvement for ORP (+12.45 Accuracy and a further MAE reduction of 0.1422), suggesting that the refinement module is particularly effective at strengthening the global semantic representation required by overall rating prediction. Overall, the ablation results validate that cross-modal alignment and subsequent refinement

Table 5 Results for different image and text encoder combinations on the ZOL

Image encoder	Text encoder	ACSA		ORP		
		Macro-F1 (%)	Acc (%)	Macro-F1 (%)	Acc (%)	MAE
VGG	BERT	**63.27**	66.83	**57.33**	**60.26**	**0.5663**
	RoBERTa	56.41	**67.25**	53.73	59.23	0.6035
	Tencent	51.07	57.04	47.00	47.21	0.9356
ResNet	BERT	61.51	64.25	55.27	57.69	0.6213
	RoBERTa	60.23	64.02	52.91	56.36	0.6580
	Tencent	50.26	57.58	46.53	48.15	0.9822

Bold values indicate the best performance for each metric

are complementary: the former mainly drives aspect-level sentiment improvements, while the latter is critical for enhancing rating prediction and error reduction, supporting the proposed architecture as an integrated design rather than a simple combination of off-the-shelf components.

5.3 Sensitivity and robustness analysis

To further evaluate the robustness of MMRP, we conduct a series of sensitivity and robustness analyses on model components, multitask frameworks, input settings, and structural hyperparameters.

5.3.1 Feature extraction encoder

In the field of multimodal data analysis, the choice of image and text encoders has an important impact on the final performance of the model. In this section, we compare the impact of selecting different image and text encoders on the accuracy of the ACSA and ORP tasks. Regarding image encoders, in addition to the VGG model used by MMRP, the commonly used ResNet model [56] is also compared. For text encoders, besides the BERT model used by MMRP, the RoBERTa model [54] and Tencent word vectors[4] are also compared. The experimental results are shown in Table 5.

Regarding text encoders, compared to using RoBERTa or Tencent word vectors, the BERT model has a significant advantage in all metrics. In the ACSA task, the combination of VGG as the image encoder and BERT as the text encoder achieved the highest Macro-F1 at 63.27%, among all experimental combinations. The Acc reached 66.83%, second only to the combination of VGG and RoBERTa. In the ORP task, the combination of VGG and BERT also performed the best in terms of Macro-F1, reaching 57.33%, with an Acc of 60.26% and an MAE of 0.5663, outperforming any other combination in all metrics. These results highlight the advantage of BERT in text feature extraction. In terms of image encoder selection, the VGG model has a significant advantage over ResNet. Regardless of whether BERT, RoBERTa, or Tencent word vectors are selected as text encoders, the VGG model achieved good accuracy in both ACSA and ORP tasks. In conclusion, using VGG and BERT as image and text encoders, respectively, demonstrates advantages in the multitask rating prediction task.

[4] Tencent word vectors: https://modelscope.cn/models/lili666/text2vec-word2vec-tencent-chinese/files.

Table 6 Experiment results for multitask framework selection

Multitask framework	ACSA		ORP		
	Macro-F1	Acc	Macro-F1	Acc	MAE
Shared-bottom	63.27%	**66.83%**	**57.33%**	**60.26%**	**0.5663**
	(0.0030)	(0.0025)	(0.0043)	(0.0031)	(0.0079)
MMoE	**63.32%**	66.66%	51.38%	54.76%	0.6189
	(0.0072)	(0.0055)	(0.0122)	(0.0135)	(0.0139)
AdaTT	61.86%	66.58%	51.18%	54.80%	0.6091
	(0.0391)	(0.0138)	(0.0501)	(0.0267)	(0.0326)

Bold values indicate the best performance for each metric

5.3.2 Multitask framework

Since this study focuses on a multitask learning problem, selecting an appropriate multitask framework is crucial in balancing training cost and model accuracy. The choice of different frameworks significantly impacts the final performance of the model. In this section, we compare the proposed MMRP model with alternative multitask frameworks to investigate their influence on the results of the ACSA and ORP tasks.

The MMRP model adopts a simple shared-bottom architecture, where lower layers are shared for representation learning, while task-specific layers are placed on top. In this work, the cross-modal Transformer Encoder and Transformer Encoder serve as the shared lower layers. To further explore the impact of different multitask frameworks, we compare MMRP against two alternative architectures: MMoE and AdaTT.

MMoE [57], an earlier and widely used multitask framework, employs a mixture of experts shared across all tasks, with gating networks dynamically fusing them for each task. These gating networks process input features and generate softmax-based weights to selectively utilize different experts. AdaTT [58], a more recent framework designed for recommendation tasks, introduces task-specific experts, shared experts (optional), and gating modules to explicitly model task-to-task interactions at both the task-pair and global levels.

In our experiments, all experts are implemented as single-layer MLPs, with hidden dimensions matching the attention dimension (set to 80 in this study). The MMoE framework shares three experts across the two tasks, with each task assigning weights through its respective gating network. In contrast, AdaTT assigns two dedicated experts per task, while also utilizing a gating network to allocate weights across all experts to capture inter-task dependencies. Both MMoE and AdaTT are integrated into the architecture shown in Fig. 1, positioned after the Transformer Encoder.

To evaluate the effect of different multitask frameworks, we trained models under each setting ten times, and the results are summarized in Table 6, with the corresponding standard deviations in the brackets below.

In the following discussion, we use MMRP-Sb, MMRP-MMoE, and MMRP-AdaTT to refer to the models under the three different multitask frameworks. For the ACSA task, the performance of all three frameworks is highly similar. The results of the t-test indicate that there are no significant differences among the outcomes of these three methods. MMRP-Sb achieves the highest accuracy (66.83%), while MMRP-MMoE attains the highest Macro-F1 (63.32%). For the ORP task, MMRP-Sb outperforms the other frameworks across all three

Table 7 Improvement and t-test results of different multitask frameworks for ORP task

Multitask framework	Macro-F1	Acc	MAE
MMoE	+11.57%***	+10.05%***	+8.51%***
AdaTT	+12.02%***	+9.96%***	+7.02%***

***significant at 0.001 level

Table 8 Experiment results for the influence of image number

Model	ACSA		ORP		
	Macro-F1 (%)	Acc (%)	Macro-F1 (%)	Acc (%)	MAE
img+3	**63.27**	66.83	**57.33**	**60.26**	**0.5663**
img+2	63.11	**67.09**	56.38	58.10	0.5985
img+1	63.09	66.91	56.05	57.57	0.5912
MEAN	63.14	66.71	55.48	55.46	0.6078

Bold values indicate the best performance for each metric

evaluation metrics, achieving a Macro-F1 of 57.33%, accuracy of 60.26%, and MAE of 0.5663.

We conducted t-tests on the ORP task results, as shown in Table 7. The hypothesis testing results indicate that MMRP-Sb demonstrates a statistically significant improvement over MMRP-MMoE and MMRP-AdaTT.

5.3.3 Number of images

In the above analysis, for each review, we utilize at most $B = 3$ images. For reviews with more than B images, B of the images are randomly selected; if the number of images is less than B, all available images are used. The experiments in this section mainly investigated whether the number of images has a significant impact on model performance.

The experiment examined the performance of the MMRP model on ACSA and ORP tasks when integrating different numbers of image information (1 or 2). The notation "MEAN" represents averaging all images' pixels to obtain one image content for all reviews for comparison. The experimental results are shown in Table 8, with the best results highlighted in bold.

Specifically, in the ACSA task, the img+3 model achieved a Macro-F1 of 63.27% and an accuracy of 66.83%. Except for the slightly lower accuracy compared to img+2 with an accuracy of 67.09%, it outperformed all indicators of img+1 and MEAN models. In the ORP task, the img+3 model led with a Macro-F1 of 57.33%, an accuracy of 60.26%, and an MAE of 0.5663, outperforming all other models. The model incorporating three images (img+3) demonstrated superiority in multiple evaluation metrics, indicating that the img+3 model had the strongest comprehensive performance in sentiment analysis tasks.

5.3.4 Structural hyperparameters

To further evaluate the practical feasibility of the proposed MMRP framework, we conducted additional sensitivity experiments on several key architectural hyperparameters, including

Table 9 Sensitivity analysis of structural hyperparameters

Parameter	Value	ACSA		ORP		
		Macro-F1 (%)	Acc (%)	Macro-F1 (%)	Acc (%)	MAE
S	2	62.90	66.51	53.13	56.37	0.5897
d	120	61.22	65.04	52.07	55.18	0.6602
h	4	61.39	65.19	48.96	52.19	0.6833
Default	–	**63.27**	**66.83**	**57.33**	**60.26**	**0.5663**

Bold values indicate the best performance for each metric

the number of cross-modal fusion layers (S), the hidden dimension (d), and the number of attention heads (h). The default configuration of the model uses $S = 6$, $d = 80$, and $h = 8$. In the experiments, we varied one hyperparameter at a time while keeping the others unchanged.

Table 9 shows that the model remains relatively robust under moderate hyperparameter variations. When adjusting the fusion depth to $S = 2$, the performance is only slightly lower than the default setting, indicating limited sensitivity to this parameter. In contrast, increasing the hidden dimension to $d = 120$ leads to noticeable degradation, suggesting that the default configuration provides a better balance between representation capacity and stability. Similarly, reducing the number of attention heads to $h = 4$ significantly harms performance on both ACSA and ORP tasks. Overall, the default setting consistently achieves the best results across all metrics.

These results indicate that the proposed MMRP framework is reasonably robust under moderate architectural variations. Small changes in structural hyperparameters do not lead to drastic performance fluctuations, suggesting that the model can maintain stable performance in practical applications.

5.4 Computational complexity

5.4.1 Time efficiency and scalability

In large-scale review mining, the primary computational bottleneck of MMRP arises from the attention operations within the cross-modal Transformer fusion module. Given a text token sequence length L and an image count C, the dominant computational complexity per layer is approximately $O((L + C)^2 d)$, where d denotes the hidden dimension. This scaling behavior is consistent with standard Transformer architectures. In practice, the number of images per review is typically small (e.g., 1–3) and can be capped if necessary, while long textual inputs can be truncated or filtered via sentence selection. Therefore, the runtime grows predictably with $(L + C)$, making the model suitable for high-throughput batch processing on large corpora.

While theoretical complexity provides a useful upper-bound view of the computational cost, practical applicability also depends on how the model behaves in actual training and deployment settings. Therefore, beyond the complexity analysis above, we further examine the efficiency and scalability of MMRP from an empirical perspective, focusing on its time cost under different data scales and its feasibility for real-world deployment.

First, the original training and validation sets are merged, reshuffled with a fixed seed, and then re-split into a new training pool and validation set at an 8:2 ratio, while the test

Table 10 Performance and training time under different training data ratios

Ratio (%)	Train size	Time (h)	ACSA		ORP		
			Macro-F1 (%)	Acc (%)	Macro-F1 (%)	Acc (%)	MAE
30	1134	7.72	45.23	52.60	35.66	42.83	0.8852
50	1890	10.24	51.15	57.08	37.76	44.82	0.7825
80	3024	15.86	60.32	62.84	52.32	54.78	0.6405
100	3780	17.12	**60.82**	**64.93**	**52.72**	54.78	**0.6310**

Bold values indicate the best result, and underlined values indicate the second-best result in each column. For Macro-F1 and Acc, higher values are better; for Time and MAE, lower values are better

set remains unchanged. Next, we train the model using 30%, 50%, 80%, and 100% of the training pool. As reported in Table 10, the training time grows from 7.72 to 17.12 h as the number of training samples increases from 1134 to 3780, indicating a near-linear growth pattern.

Meanwhile, the performance on both ACSA and ORP improves consistently as more training data are used. Notably, the improvement from 80 to 100% is relatively small, despite the additional training time. These results suggest that the proposed model has good scalability and maintains a practical trade-off between performance and computational cost.

5.4.2 Deployment

To complement the predictive evaluation and cross-domain experiments, we further analyze the practical deployment feasibility of MMRP.

For online scenarios, such as real-time monitoring of user feedback or dynamic product-quality dashboards, MMRP can be deployed using a two-stage pipeline. First, visual features are extracted offline using a vision encoder and cached as embeddings. Second, the online system performs lightweight cross-modal fusion and prediction by consuming cached image embeddings together with incoming text tokens. This design significantly reduces online latency and enables the system to satisfy near real-time requirements under moderate traffic conditions. Additional acceleration can be achieved through batching, mixed-precision inference, and restricting the maximum text length and image count per request.

From a memory perspective, the additional parameters introduced by the multitask prediction heads are marginal compared with the fusion backbone. Moreover, the overall architecture remains considerably lighter than large end-to-end vision–language pre-trained models, resulting in lower GPU memory consumption and improved fine-tuning efficiency. These properties make MMRP more practical for deployment in resource-constrained environments.

Overall, MMRP supports both offline large-scale analytics and near real-time inference, providing a favorable trade-off between effectiveness and computational efficiency for multimodal review mining applications.

5.5 Cross-domain evaluation

5.5.1 ViMACSA dataset

To evaluate the cross-domain generalization ability of the proposed model beyond the electronics domain (ZOL), we additionally using ViMACSA[59][5] a multimodal hotel-review dataset containing user-written review texts, corresponding image sets, and multi-aspect ratings. Each multimodal instance in ViMACSA consists of (i) a piece of review text, (ii) an associated image set (with a varying number of images), and (iii) aspect-level rating annotations together with an overall rating, which allows us to construct the same multitask learning setting as in ZOL (i.e., ACSA + ORP). Following the same preprocessing protocol used for ZOL, we convert ViMACSA into the unified ZOL data format to ensure consistent input representation, label processing, and evaluation.

ViMACSA contains 4,876 samples, split into 2,876/1,000/1,000 for Train/Valid/Test, respectively. Compared with ZOL (mobile phones), ViMACSA exhibits a clear domain shift in both linguistic expressions and visual content: the reviews are written in the hospitality context, and images typically depict room environments, facilities, and service-related scenes rather than product appearance. Most reviews contain 1–2 images, and the overall rating distribution is skewed toward medium-to-high scores, reflecting real-world rating behaviors in hotel platforms. These characteristics make ViMACSA a suitable benchmark to assess the model's robustness and transferability across domains under the same multitask framework.

5.5.2 Comparison results

Importantly, we use the same multitask formulation (ACSA + ORP), preprocessing pipeline, and evaluation protocol for both ZOL and ViMACSA. This unified setting allows us to attribute performance differences primarily to domain shift rather than to inconsistent task definitions or metric choices.

We first report the main comparative results on ViMACSA. For the sake of space, we only report the ACSA results. As shown in Table 11, for the ACSA task, MMRP achieves the best Macro-F1 among all compared methods, reaching 22.32%, and also obtains competitive accuracy at 26.94%. Although Chimera yields a much higher accuracy (62.34%), its Macro-F1 is only 20.28%, indicating that its predictions are less balanced across aspect categories. In contrast, the proposed MMRP maintains better category-level balance, which is particularly important in cross-domain aspect prediction. Compared with the single-task variant MMRP (w/o ORP), the full multitask model further improves ACSA Macro-F1 from 21.86 to 22.32% and accuracy from 26.84 to 26.94%, suggesting that the ORP supervision still provides useful auxiliary information even under domain shift.

5.5.3 Ablation study

Beyond the direct comparison with baselines, we further examine whether the internal architectural design of MMRP remains effective on ViMACSA. Table 12 reports the ablation results. Removing the self-encoder decreases ORP Macro-F1 from 33.31 to 29.72% and ACSA Macro-F1 from 22.32 to 21.30%. Removing the cross-encoder leads to an even larger degradation, with ORP Macro-F1 dropping to 26.99%, ORP Accuracy to 34.40%, ORP

[5] The dataset can be downloaded from https://github.com/hoangquy18/Multimodal-Aspect-Category-Sentiment-Analysis.

Table 11 ACSA comparison results on ViMACSA

Model type	Model name	Macro-F1	Acc
Single Modal	Att-BiLSTM	13.77%***	29.95%***
		(0.0080)	(0.0033)
	ARP	12.88%***	29.75%***
		(0.0023)	(0.0002)
	ASAP	21.54%***	26.88%***
		(0.0015)	(0.0025)
	Chimera	20.28%***	**62.34%****
		(0.0012)	(0.0014)
Multi-Modal	MIMN	19.82%***	26.27%***
		(0.0063)	(0.0076)
	Co-Memory+Aspect	19.53%***	25.48%***
		(0.0024)	(0.0079)
	MMRP	21.86%	26.84%
	(w/o ORP)	(0.0040)	(0.0040)
	MMRP	**22.32%**	26.94%
		(0.0014)	(0.0016)

The numbers in brackets are standard deviations, *** $p < 0.001$
Bold values indicate the best performance for each metric

Table 12 Ablation results on ViMACSA

Configuration	ACSA		ORP		
	Macro-F1 (%)	Acc (%)	Macro-F1 (%)	Acc (%)	MAE
w/o self-encoder	21.30	26.31	29.72	39.10	0.8042
w/o cross-encoder	20.98	25.68	26.99	34.40	0.9445
Full MMRP	**22.32**	**26.94**	**33.31**	**39.53**	**0.7615**

MAE rising to 0.9445, and ACSA Macro-F1 decreasing to 20.98%. These results show that both modules contribute to performance on the hotel-review domain, while the cross-modal encoder plays the more foundational role. In other words, even under cross-domain shift, the explicit fusion of textual and visual evidence remains essential, and the subsequent self-fusion stage further refines the shared representation for downstream prediction.

5.5.4 Sensitivity analysis

We also analyze the influence of encoder choices on cross-domain performance. As shown in Table 13, encoder selection has a visible effect on the final results. Under the VGG image encoder, PhoBERT gives the strongest ORP result among the three text encoders, whereas under the ResNet image encoder, XLM-RoBERTa achieves the best overall balance, reaching 23.99% / 28.47% on ACSA and 36.04% / 40.80% / 0.7407 on ORP. Interestingly, this ResNet + RoBERTa combination slightly surpasses the default ViMACSA setting in several metrics, suggesting that the optimal encoder choice may depend on the target domain rather than being universally fixed. At the same time, FastText-based settings perform consistently worse,

Table 13 Results for different encoder combinations on ViMACSA

Image encoder	Text encoder	ACSA		ORP		
		Macro-F1 (%)	Acc (%)	Macro-F1 (%)	Acc (%)	MAE
VGG	PhoBERT	22.32	26.94	33.31	39.53	0.7615
	RoBERTa	20.94	27.07	30.55	37.70	0.8017
	FastText	18.36	24.97	22.34	32.50	1.0566
ResNet	PhoBERT	22.50	26.69	30.10	40.40	0.7464
	RoBERTa	**23.99**	**28.47**	**36.04**	**40.80**	**0.7407**
	FastText	19.42	25.46	19.65	30.00	1.0255

The best results are shown in bold, and the second-best results are underlined. For the ViMACSA dataset, PhoBERT refers to `vinai/phobert-base-v2`, RoBERTa refers to `facebookai/xlm-roberta-base`, and FastText refers to FastText word embeddings

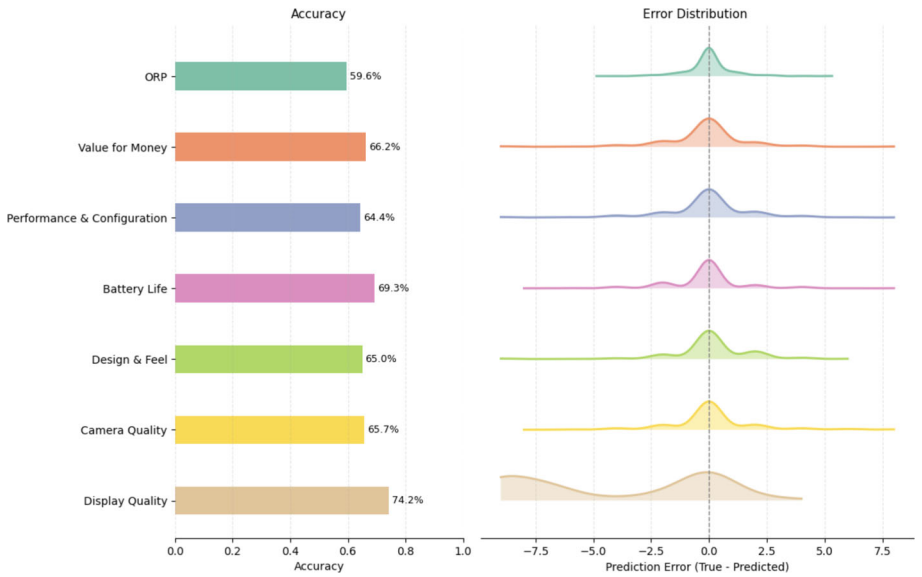


Fig. 3 Accuracy and prediction error of ORP and aspect categories

which indicates that stronger pre-trained contextual language encoders remain important for handling the semantic variability of hotel reviews.

5.6 Case study

5.6.1 Objective versus subjective targets in ZOL

Figure 3 presents the case-study results of MMRP on the ZOL dataset, including prediction accuracy and error distribution for overall rating (ORP) and six aspect categories (value for money, performance configuration, battery life, design feel, camera quality, and display quality).

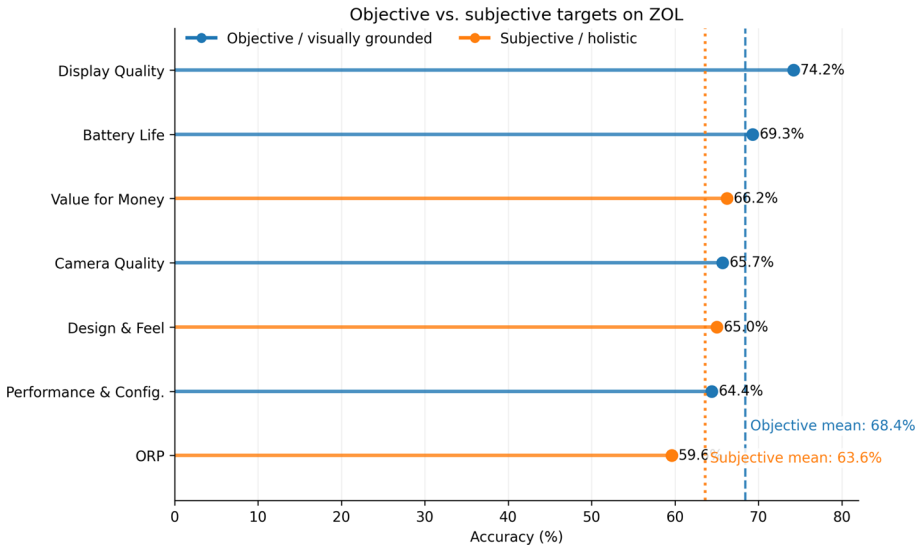


Fig. 4 Objective versus subjective targets in the ZOL case study

The MMRP model achieves its highest accuracy on concrete aspects such as display quality (74.2%) and battery life (69.3%), while it obtains the lowest accuracy for more subjective targets such as ORP (59.6%). This suggests that clear visual cues translate into more reliable predictions, whereas holistic judgments remain challenging.

Error distributions indicate that tight, peaked curves for ORP and camera quality indicate more consistent predictions, while slight under-prediction bias on value for money and performance & configuration suggests that a small calibration step could improve balance on those aspects.

To further clarify the difference between subjective and objective targets, Figure 4 reorganizes the ZOL case-study results into two groups: more objective / visually grounded targets and more subjective / holistic targets. As shown in Fig. 4, the model generally achieves higher accuracy on the objective group, with the best results on display quality (74.2%) and battery life (69.3%), whereas more subjective targets remain relatively harder, especially ORP (59.6%).

As expected, performance decreases on ViMACSA due to substantial domain shift, including differences in aspect schema, rating calibration behaviors, writing style, and visual content. Specifically, ACSA Macro-F1 drops from 62.37% on ZOL to 23.00% on ViMACSA, while ORP MAE increases from 0.5663 to 0.7677. These findings confirm that cross-domain multimodal multitask learning remains challenging and highlight the sensitivity of aspect-level prediction to domain-specific semantics and annotation structures.

5.6.2 Joint textual semantics and visual evidence in ViMACSA

Unlike product-centric reviews, hotel feedback often contains more subjective language and relies heavily on contextual visual cues. In the example shown in Fig. 5, the reviewer praises room cleanliness and interior design while implicitly expressing dissatisfaction with the surrounding environment.

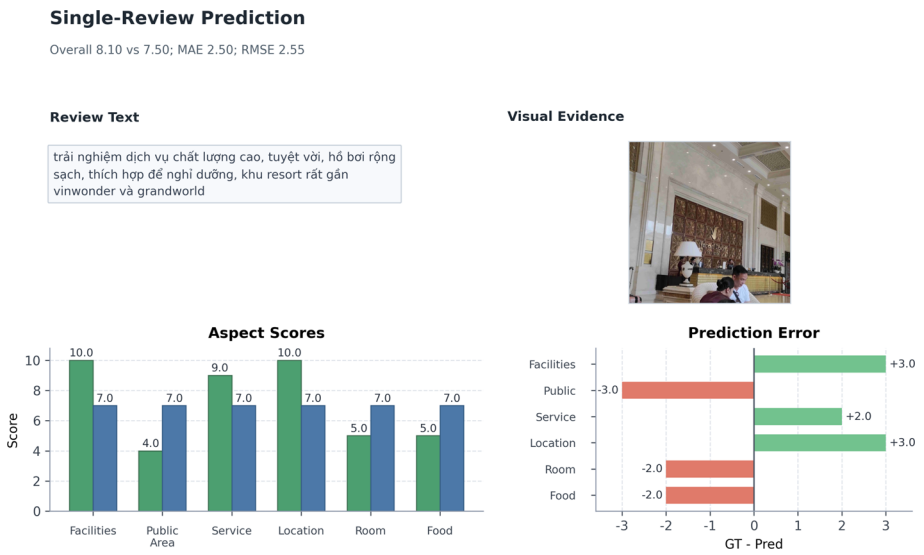


Fig. 5 The cross-domain reviews with text and image

By jointly modeling textual semantics and visual evidence, MMRP assigns positive sentiment to cleanliness-related aspects while detecting weaker sentiment toward environmental factors, resulting in a balanced overall rating prediction. In contrast, text-only approaches tend to depend primarily on explicit sentiment expressions and may overlook complementary visual signals.

From an application perspective, such multimodal reasoning supports more reliable automated analysis of customer feedback. For instance, hospitality providers could leverage aspect-level insights to identify recurring service issues, prioritize operational improvements, and monitor satisfaction at scale. More broadly, these results suggest that MMRP generalizes beyond electronic product reviews and retains practical utility under realistic distribution shifts.

Nevertheless, the observed performance gap also motivates future research directions, including domain-adaptive pretraining, improved aspect alignment across domains, and more robust fusion strategies for scenarios with sparse or missing annotations.

6 Discussion

With the development of social media and e-commerce platforms, user reviews have become crucial references for analyzing brands and products. These reviews not only contain textual information but also include images, offering richer and more intuitive expressions of sentiment. Consequently, multimodal review mining has emerged as a challenge in the field of data mining. Aspect Category Sentiment Analysis (ACSA) and Overall Rating Prediction (ORP) are two key tasks in review mining. This study proposes an innovative multimodal and multitask rating prediction network structure, MMRP, which leverages text and image information to enhance prediction accuracy by jointly addressing the ACSA and ORP tasks. Compared to traditional methods that rely on textual information, MMRP improves the prediction accuracy of user sentiment polarity and rating inclination by integrating image information. This

indicates that images, as an intuitive emotional expression, can transcend language barriers and provide a new perspective for understanding users' sentiments. The model is capable of simultaneously handling both ACSA and ORP tasks. This joint modeling not only enhances the understanding of fine-grained sentiment but also improves the overall rating prediction accuracy. Compared to other multimodal models, such as those using attention mechanisms or bilinear pooling methods, MMRP adopts a more effective cross-modal Transformer fusion strategy to achieve higher prediction accuracy. Through experimental comparisons, this study validates the superior performance of the MMRP model.

These observations suggest that the current model is better suited for offline or near-offline multimodal sentiment analysis scenarios, such as large-scale review mining, periodic brand monitoring, and batch-level opinion analytics, where improved predictive performance is more critical than strict real-time latency. By contrast, for strongly latency-sensitive settings, additional optimization would still be desirable, such as parameter reduction, lightweight fusion, pruning, or distillation. Although the present study supplements the practical discussion with empirical training-time scaling analysis, a systematic comparison of parameter count, inference latency, and memory footprint across baselines is still absent and will be included in future work.

From a practical perspective, the above results indicate that MMRP maintains stable predictive performance under moderate architectural variations and increasing data scales. This suggests that the framework is suitable for offline or near-offline applications such as large-scale review mining, batch sentiment analysis, and periodic product evaluation. In these scenarios, predictive accuracy is usually prioritized over strict latency constraints.

For real-time applications, additional engineering techniques such as model compression, pruning, distillation, or lightweight multimodal encoders may further improve deployment efficiency. We leave this direction for future work. There is vast research space in this field for the future. The semantic alignment and fusion strategy of textual and visual data remain an open research issue. Designing more effective algorithms to automatically identify intrinsic connections between text and images, as well as constructing efficient models to handle such highly complex data relationships, will be the focus of future research. Meanwhile, with the increasing awareness of privacy protection and the enactment of relevant regulations, conducting multimodal data mining under the premise of protecting user privacy is also an important issue that future research needs to address.

In summary, through innovative model design and in-depth analysis, this study has made significant progress in the field of multimodal review mining and rating prediction, improving prediction accuracy and providing new perspectives and tools for understanding and utilizing user-generated content.

Beyond predictive performance and computational efficiency, responsible deployment is a critical consideration for modern AI systems.

As multimodal sentiment analysis systems are increasingly applied to real-world review mining, it is important to consider their potential societal and ethical implications. User-generated reviews often contain subjective opinions, cultural nuances, and platform-specific rating behaviors, which may introduce unintended biases into predictive models.

To reduce these risks, this study adopts several precautionary strategies. First, the supervision signals used in our experiments originate from platform-provided aspect ratings and overall scores rather than synthetic annotations, helping maintain consistency with real user feedback. Second, the multimodal design of MMRP jointly leverages textual and visual evidence, mitigating the risk of over-reliance on a single modality that could encode biased or incomplete information. Third, the introduction of a cross-domain dataset (ViMACSA)

enables evaluation under distribution shift, providing an additional perspective on model robustness and transferability.

Future work will explore fairness-aware training strategies, bias diagnostics across aspect categories, and domain-adaptive techniques to improve model reliability. We also encourage cautious deployment of automated sentiment systems in decision-making pipelines, emphasizing their role as supportive analytic tools rather than replacements for human judgment.

Acknowledgements This work is supported by the National Natural Science Foundation of China (No. 72171229), and the MOE Project of Key Research Institute of Humanities and Social Sciences (No. 22JJD110001) and Big Data and Responsible Artificial Intelligence for National Governance, Renmin University of China and Fundamental Research Funds for the Central Universities (CUC25GT23).

Author Contributions The contributions of the authors to this study are as follows: XW: Conceptualization, Methodology, Formal analysis, Writing—original draft and revised RL: Methodology, Formal analysis. YG: Conceptualization, Methodology, Data Curation, Formal analysis, Writing—original draft. XL: Conceptualization, Supervision, Writing—review & editing. JR: Methodology, Formal analysis. LS: Formal analysis, Writing—original draft

Data Availability Data are provided within the manuscript.

Declarations

Conflict of interest The authors declare no conflict of interest.

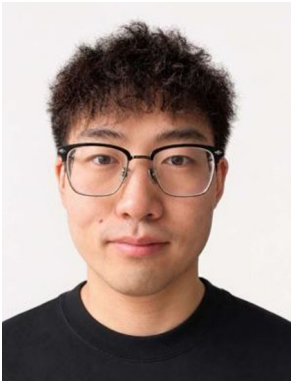
References

1. Pang B, Lee L et al (2008) Opinion mining and sentiment analysis. Found Trends® Inform 2(1–2):1–135
2. Liu B (2022) Sentiment analysis and opinion mining. Springer, Cham
3. Liu B (2020) Sentiment analysis: mining opinions, sentiments, and emotions. Cambridge University Press, Cambridge
4. Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, pp. 168–177
5. Wu P, Tang T, Zhou L, Martínez L (2024) A decision-support model through online reviews: consumer preference analysis and product ranking. Inf Process Manag 61(4):103728
6. Seo S, Huang J, Yang H, Liu Y (2017) Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In: Proceedings of the eleventh ACM conference on recommender systems, pp. 297–305
7. Li F, Liu NN, Jin H, Zhao K, Yang Q, Zhu X (2011) Incorporating reviewer and product information for review rating prediction. In: Twenty-second international joint conference on artificial intelligence
8. Tang D, Qin B, Liu T, Yang Y (2015) User modeling with neural network for review rating prediction. In: Twenty-fourth international joint conference on artificial intelligence
9. Zhang W, Yuan Q, Han J, Wang J (2016) Collaborative multi-level embedding learning from reviews for rating prediction. IJCAI 16:2986–2992
10. Bu J, Ren L, Zheng S, Yang Y, Wang J, Zhang F, Wu W (2021) Asap: a chinese review dataset towards aspect category sentiment analysis and rating prediction. arXiv preprint arXiv:2103.06605
11. You Q, Luo J, Jin H, Yang J (2016) Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In: Proceedings of the ninth ACM international conference on web search and data mining, pp. 13–22
12. Luo C, Luo XR, Xu Y, Warkentin M, Sia CL (2015) Examining the moderating role of sense of membership in online review evaluations. Inf Manag 52(3):305–316
13. Truong Q-T, Lauw HW (2017) Visual sentiment analysis for review images with item-oriented and user-oriented CNN. In: Proceedings of the 25th ACM international conference on multimedia, pp. 1274–1282
14. Khan ZY, Niu Z, Sandiwarno S, Prince R (2021) Deep learning techniques for rating prediction: a survey of the state-of-the-art. Artif Intell Rev 54:95–135
15. Xiong Y, Liu Y, Qian Y, Jiang Y, Chai Y, Ling H (2024) Review-based recommendation under preference uncertainty: an asymmetric deep learning framework. Eur J Oper Res 316:1044–1057

16. Dhillon PS, Aral S (2021) Modeling dynamic user interests: a neural matrix factorization approach. Mark Sci 40(6):1059–1080
17. Zhou F, Jiang Y, Qian Y, Liu Y, Chai Y (2024) Product consumptions meet reviews: inferring consumer preferences by an explainable machine learning approach. Decis Support Syst 177:114088
18. Feng S, Song K, Wang D, Gao W, Zhang Y (2021) Intersentiment: combining deep neural models on interaction and sentiment for review rating prediction. Int J Mach Learn Cybern 12:477–488
19. Wang X, Xiao T, Tan J, Ouyang D, Shao J (2020) Mrmrp: multi-source review-based model for rating prediction. In: Database systems for advanced applications: 25th international conference (DASFAA 2020). Springer, Cham
20. Bauman K, Tuzhilin A (2022) Know thy context: parsing contextual information from user reviews for recommendation purposes. Inf Syst Res 33(1):179–202
21. Kandhro IA, Ali F, Uddin M, Kehar A, Manickam S (2024) Exploring aspect-based sentiment analysis: an in-depth review of current methods and prospects for advancement. Knowl Inf Syst. https://doi.org/10.1007/s10115-024-02104-8
22. Wu C, Wu F, Liu J, Huang Y, Xie X (2019) Arp: Aspect-aware neural review rating prediction. In: Proceedings of the 28th ACM international conference on information and knowledge management, pp. 2169–2172
23. Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, Xu B (2016) Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short Papers), pp. 207–212
24. Long Y, Lu Q, Xiang R, Li M, Huang CR (2017) A cognition based attention model for sentiment analysis. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp. 462–471. Association for Computational Linguistics, Copenhagen, Denmark. https://doi.org/10.18653/v1/D17-1048, https://aclanthology.org/D17-1048/
25. Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE (1991) Adaptive mixtures of local experts. Neural Comput 3(1):79–87
26. Ma J, Zhao Z, Yi X, Chen J, Hong L, Chi EH (2018) Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 1930–1939
27. Xu N, Mao W, Chen G (2019) Multi-interactive memory network for aspect based multimodal sentiment analysis. In: Proceedings of the AAAI conference on artificial intelligence, vol. 33, pp. 371–378
28. Xu N, Mao W (2017) Multisentinet: a deep semantic network for multimodal sentiment analysis. In: Proceedings of the 2017 ACM on conference on information and knowledge management, pp 2399–2402
29. Tang D, Qin B, Liu T (2015) Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp. 1422–1432
30. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp. 1480–1489
31. Lin Z, Liang B, Long Y, Dang Y, Yang M, Zhang M, Xu R (2022) Modeling intra-and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis. In: Proceedings of the 29th international conference on computational linguistics, vol 29. Association for Computational Linguistics, Berkeley
32. Lu Q, Sun X, Gao Z, Long Y, Feng J, Zhang H (2024) Coordinated-joint translation fusion framework with sentiment-interactive graph convolutional networks for multimodal sentiment analysis. Inf Process Manag 61:103538
33. Yu B, Li C, Shi Z (2025) Multi-grained feature gating fusion network for multimodal sentiment analysis. Knowl Inf Syst 67:6879–6905
34. Ma Y, Xiang Z, Du Q, Fan W (2018) Effects of user-provided photos on hotel review helpfulness: an analytical approach with deep leaning. Int J Hosp Manag 71:120–131
35. Hu A, Flaxman S (2018) Multimodal sentiment analysis to explore the structure of emotions. In: Proceedings of the 24th ACM sigkdd international conference on knowledge discovery & data mining, pp. 350–358
36. Yu Z, Yu J, Fan J, Tao D (2017) Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp. 1821–1830
37. Yu Z, Yu J, Xiang C, Fan J, Tao D (2018) Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans Neural Netw Learn Syst 29(12):5947–5959
38. Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M (2016) Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847

39. Kim J-H, On K-W, Lim W, Kim J, Ha J-W, Zhang B-T (2016) Hadamard product for low-rank bilinear pooling. arXiv preprint arXiv:1610.04325
40. Truong Q-T, Lauw HW (2019) Vistanet: visual aspect attention network for multimodal sentiment analysis. In: Proceedings of the AAAI conference on artificial intelligence, vol. 33, pp. 305–312
41. Tsai Y-HH, Bai S, Liang PP, Kolter JZ, Morency L-P, Salakhutdinov R (2019) Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the conference. Association for computational linguistics. Meeting, vol. 2019, p. 6558. NIH Public Access
42. Yu J, Wang J, Xia R, Li J (2022) Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, pp. 4482–4488
43. Lin Y, Wang Z, Qin G, Wu L, Li Y (2025) Semantic-guided multi-grained cross-modal alignment and fusion network for multimodal aspect-based sentiment analysis. Inf Fusion 127:103878
44. Feng X, Lin Y, He L, Li Y, Chang L, Zhou Y (2024) Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In: Findings of the association for computational linguistics: EMNLP 2024, pp. 14755–14766
45. Li Y, Ding H, Lin Y, Feng X, Chang L (2024) Multi-level textual-visual alignment and fusion network for multimodal aspect-based sentiment analysis. Artif Intell Rev 57(4):78
46. Yu B, Xing Y, Yang Y, Cao C, Shi Z (2025) Multimodal aspect-based sentiment analysis based on a dual syntactic graph network and joint contrastive learning. Knowl Inf Syst 67:5125–5149
47. He L, Lin Y, Qin G, Liu J, Feng X, Zhou Y (2025) Dual dynamic multi-granularity fusion method for multimodal sarcasm detection. Neurocomputing. https://doi.org/10.1016/j.neucom.2025.131985
48. Zhang W, et al (2024) Sentiment analysis in the era of large language models. In: Findings of the association for computational linguistics: NAACL 2024. https://aclanthology.org/2024.findings-naacl.246/
49. Simmering PF, et al (2023) Large language models for aspect-based sentiment analysis. arXiv preprint arXiv:2310.18025 [cs.CL]
50. OpenAI: Gpt-4 technical report. arXiv preprint (2023) arXiv:2303.08774 [cs.CL]
51. Li J, Li D, Savarese S, Hoi S (2023) Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th international conference on machine learning (ICML). Proceedings of Machine Learning Research, vol. 202
52. Dai W, Li J, Li D, Tiong AMH, Zhao J, Wang W, Li B, Fung P, Hoi S (2023) Instructblip: towards general-purpose vision-language models with instruction tuning. In: Advances in neural information processing systems (NeurIPS)
53. Liu H, Li C, Wu Q, Lee YJ (2023) Advances in neural information processing systems (NeurIPS) (Visual instruction tuning)
54. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692
55. Xiao L, Mao R, Zhao S, Lin Q, Jia Y, He L, Cambria E (2025) Exploring cognitive and aesthetic causality for multimodal aspect-based sentiment analysis. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2025.3565506
56. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778
57. Ma J, Zhao Z, Yi X, Chen J, Hong L, Chi EH (2018) Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. KDD '18, pp. 1930–1939. Association for Computing Machinery, New York
58. Li D, Zhang Z, Yuan S, Gao M, Zhang W, Yang C, Liu X, Yang J (2023) Adatt: adaptive task-to-task fusion network for multitask learning in recommendations. In: Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining. KDD '23, pp. 4370–4379. Association for Computing Machinery, New York
59. Nguyen QH, Nguyen M-VT, Van Nguyen K (2025) New benchmark dataset and fine-grained cross-modal fusion framework for Vietnamese multimodal aspect-category sentiment analysis. Multimed Syst 31(1):4

Xiaoning Wang is an associate professor at School of Data Science and Media Intelligence, Communication University of China. His research interests include data mining, sampling survey, LLM and RAG.



Rui Li is a graduate student at School of Data Science and Media Intelligence, Communication University of China. His research interests include data mining, LLM, multimodal learning, and recommender systems.

Yuxuan Guo is a researcher of Maxwealth Fund Management Company Limited.

Xiaoling Lu is a professor at School of Statistics, Renmin University of China. Her research interest includes graph neural network, spatiotemporal data mining, data science, and LLM.

Jiawei Ren is a postgraduate student at School of Data Science and Media Intelligence, Communication University of China. His research interests include data science, LLM, and RAG.

Libo Sun is a PhD student at School of Data Science, Fudan University. His research interests include multimodal large language models and autonomous agents.