OX-MABSR: A Benchmark for Open-domain Explainable Multimodal Aspect-Based Sentiment Reasoning

Xinjing Liu, ZiXin Xue, Pengyue Lin, Xinyu Tu, Siwei Xu, Ruifan Li*

School of Artificial Intelligence, Beijing University of Posts and Telecommunications
{liuxj_ai, xuezixin, linpengyue, tuxinyu, xusw, rfli}@bupt.edu.cn

Abstract

Multimodal Aspect-Based Sentiment Analysis (MABSA) involves extracting aspect terms from text-image pairs and identifying their sentiments. Most existing tasks consider one fixed sentiment category with explicitly mentioned aspects. However, these tasks seldom consider expressive sentiment categories, implicit aspects, and explainability. To this end, we introduce a novel task of Open-domain Explainable Multimodal Aspect-Based Sentiment Reasoning (OX-MABSR). This task enables the prediction of open-vocabulary aspect-sentiment pairs, together with the generation of sentiment explanations and reasoning paths. To benchmark OX-MABSR task, we construct OX-MABSR-Bench, a dataset annotated with explicit and implicit aspects, expressive sentiment categories, as well as perceptual and cognitive two-level explanations. The explanations capture visual and textual cues, including aesthetics, facial expressions, scenes, and textual semantics, together with background and situational knowledge. In addition, we annotate the reasoning paths that trace how the sentiment evolves from surface cues to a deeper contextual understanding. To address OX-MABSR task, we propose MABSR-LLM. Extensive experimental results show our MABSR-LLM outperforms strong baselines. To the best of our knowledge, we are the first to provide a unified framework for open-domain and explainable MABSR.

Code and Data: —
https://github.com/Liuxj-Anya/OX-MABSR

1 Introduction

Textual Sentiment Analysis (TSA) (Liu 2020) has long served as a foundational task in natural language processing, aiming to enable AI systems to understand human sentiment. Previous works have moved from assigning a coarse-grained sentiment polarity (Cambria et al. 2020; Zhang et al. 2023) to determining fine-grained sentiment for specific aspects, i.e., Aspect-Based task (ABSA) (Li et al. 2021; Ouyang et al. 2024). Recently, the rise of multimodal content has driven the evolution of multimodal sentiment analysis (MSA) (Wu et al. 2024a; Hossain et al. 2025). Early efforts in MSA remained at the coarse level, predicting the overall sentiment. Furthermore, Multimodal ABSA (MABSA) (Peng et al. 2024; Zhu et al. 2025c; Liu et al. 2025b) enables aspect sentiment detection across modalities.

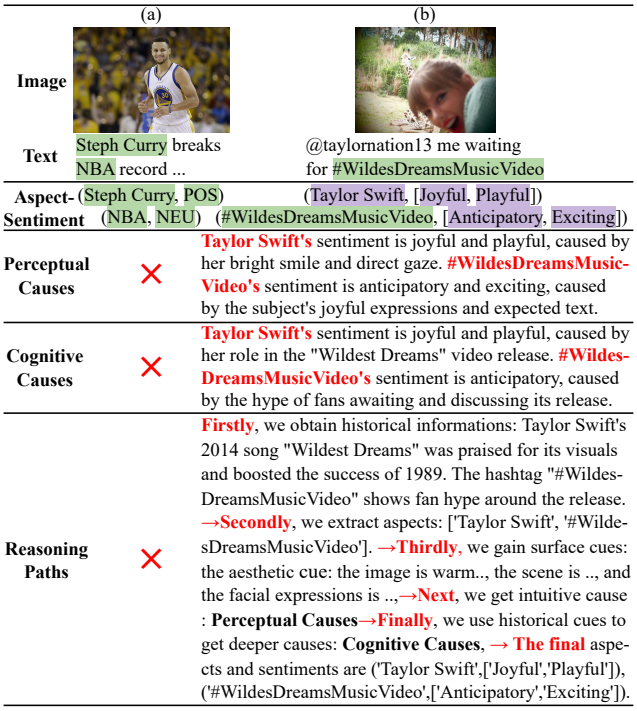

Figure 1: Comparison of the existing MABSA (Col. (a)) with our proposed OX-MABSR (Col. (b)). Our task includes aspect-sentiment, perceptual and cognitive causes, and reasoning paths. *Open* aspects and the sentiments are shown in purple. *Explainability* is shown at the bottom three rows.

As shown in Figure 1 (a), however, existing MABSA tasks (Han et al. 2025) mainly focus on classifying sentiment polarity. Thus, they could provide limited insight into the underlying causes behind sentiment. Moreover, these tasks assume that aspects are explicitly mentioned in the text and rely on one predefined sentiment category (e.g., Pos/Neu/Neg). Thus, they restrict the ability to handle implicit aspects and express multi-nuanced sentiments. Yet, sentiments are rarely arbitrary; they are influenced by surface cues and latent contextual factors, such as aesthet-

*Corresponding author

ics, experiences, social interactions, and historical context. Moreover, aspects are sometimes implicitly mentioned, and sentiment can be nuanced. Therefore, we suppose that without considering these factors, sentiment analysis AI systems struggle to provide a deep and meaningful understanding.

To bridge this gap, in this paper we introduce a novel OX-MABSR task. This task enables the prediction of open-vocabulary aspect-sentiment pairs, together with the generation of two-level sentiment explanations and reasoning paths. For example, as shown in Figure 1 (b), the task first requires inferring the implicit aspect "Taylor Swift" by jointly interpreting visual cues and textual context. Then, the task requires predicting the expressive sentiment and the corresponding causes. By comparison, the OX-MABSR task extends the conventional MABSA, enabling *Open-domain* and *Explainability* on image-text modalities.

To benchmark the OX-MABSR task, we construct a novel dataset, OX-MABSR-Bench. It is built based on the previous MVSA-Multiple dataset (Niu et al. 2016), which only provides coarse sentiment for image-text pairs. For our purpose, we propose an automatic annotation approach based on Chain-of-Thought (CoT) prompting with MLLM. It simulates human-like reasoning steps to enrich each sample with open-vocabulary aspect-sentiment pairs and two-level sentiment explanations. Specifically, the annotation approach works in multiple steps. First, the MLLM is prompted to extract events relevant to the given sample and to identify aspects likely to be sentiment targets from them. Next, it constructs the two-level causes. The MLLM first analyzes observable surface cues such as visual aesthetics, facial expressions, image scenes, and textual semantics to determine perceptual causes. Subsequently, to construct cognitive causes, it analyzes latent contextual knowledge. This knowledge includes historical background and significant events, enabling it to infer more implicit motivations. Finally, the MLLM merges two-level causes for multi-expressive sentiments. In addition, we extract intermediate reasoning paths from the CoT process, capturing the inference trajectory from surface cues to deeper understanding.

To ensure the quality of annotations, we propose a validation approach. It combines MLLM reasoning capabilities with human verification. For each annotated sample, the MLLM first assumes the annotations are correct and provides supportive reasons. Simultaneously, it critiques the annotations are incorrect and provides oppositive reasons. Then, given both supportive and oppositive reasons, it makes a final judgment on its validity. This scheme guides the MLLM to reflect on both sides of reasoning, reducing hallucinations and enhancing factual consistency. Finally, we conduct a round of manual review to ensure overall quality. Thus, we obtain high-quality OX-MABSR-Bench that supports not only the prediction of open-vocabulary aspect-sentiment but also the explainability analysis of sentiment.

To address the OX-MABSR task, we design MABSR-LLM. The model adopts a visual encoder (SigLIP-2) and a language backbone (Qwen3) for image-text modalities. Moreover, our model is trained not only with aspect-sentiment labels and two-level explanations but also with the intermediate reasoning paths. This enables the model to perform sentiment inference in multiple steps. It first recognizes aspects and surface-level cues, then incorporates contextual knowledge for deeper interpretation. Thus, our model advances aspect-sentiment prediction toward open-domain sentiment understanding and enhanced explainability.

To summarize, our major contributions are threefold. **1**) We introduce OX-MABSR task, which extends MABSA task by enabling open-vocabulary aspect-sentiment pairs prediction, explanations and reasoning paths generation. **2**) We build OX-MABSR-Bench with open-vocabulary labels to capture expressive sentiments and explicit/implicit aspects. Moreover, it offers explanations at perceptual and cognitive levels, along with reasoning paths. **3**) We propose MABSR-LLM, a strong baseline that uses step-wise reasoning to achieve explainable and open-domain understanding of aspect-based sentiment.

2 OX-MABSR: Dataset Construction

Existing sentiment analysis datasets are limited in both open-domain generality and explainability. As shown in Table 1, previous datasets rely on one fixed category and assume aspects are explicitly mentioned in text. This limits the reasoning on implicit aspects and multi-nuanced sentiments. Furthermore, they focus solely on polarity classification, seldom touching the explanation on sentiment causes.

To this end, we design a data construction pipeline primarily driven by MLLM and assisted by human, shown in Figure 2. The pipeline consists of three stages: annotation generation, annotation verification, and reasoning data construction. Thereby, we introduce an open-domain explainable dataset, OX-MABSR-Bench.

2.1 Annotation Generation

We extend MVSA-Multiple dataset by generating additional annotations. While MVSA contains image-text pairs labeled with coarse sentiments, we aim to enrich each sample with open-vocabulary aspect-sentiment pairs labels, together with two-level explanations across perceptual and cognitive levels. To this end, we design an automatic annotation approach based on CoT prompting with Qwen2.5-VL-72B.

The annotation steps are as follows. **1)** Event Extraction. Given an image-text pair, MLLM is prompted to identify salient events described in the input. These events provide the narrative structure for subsequent sentiment inference, anchoring the reasoning in specific contexts. **2)** Aspect Detection. From extracted events, MLLM is prompted to detect aspects (explicit or implicit) that are likely to be sentiment targets. **3)** Perceptual-level Reasoning. MLLM is guided to analyze surface-level cues of visual aesthetics, facial expressions, scene composition, and textual semantics. With the observable cues, it generates corresponding perceptual causes for each aspect. This level reflects intuitive and first-impression explanations. **4)** Cognitive-level Reasoning. MLLM is further prompted to analyze latent contextual knowledge, including previous events, historical background, and implicit textual references. Thus, it infers deeper cognitive causes for each aspect. **5)** Sentiments Generation. MLLM combines both levels of causes to derive multiple open-vocabulary sentiments for each aspect.

Dataset	Modality	#Sample	Task Output	Granularity	Open/ Fixed	Explainability
CR (Hu and Liu 2004)	Text	6,180	Sentiment	Coarse	Fixed	✗
Yelp (Zhang and LeCun 2015)	Text	59,800	Sentiment	Coarse	Fixed	✗
SemEval (Pontiki et al. 2016)	Text	8,886	Target, Aspect, Sentiment	Fine	Fixed	✗
ACOS (Cai, Xia, and Yu 2021)	Text	5,045	Target, Aspect, Opinion, Sentiment	Fine	Fixed	✗
MVSA-Multiple (Niu et al. 2016)	Image+Text	19,600	Sentiment	Coarse	Fixed	✗
Twitter-15/17 (Yu and Jiang 2019)	Image+Text	6,412	Aspect, Sentiment	Fine	Fixed	✗
OX-MABSR-Bench (Ours)	Image+Text	14,705	Aspect, Sentiment, Perceptual Causes, Cognitive Causes, Reasoning Paths	Fine	Open	✓

Table 1: Comparison of our dataset OX-MABSR-Bench with existing sentiment datasets, including text-only and text-image multimodal datasets. Only our dataset provides open-vocabulary aspect-sentiment pairs and sentiment explanations.

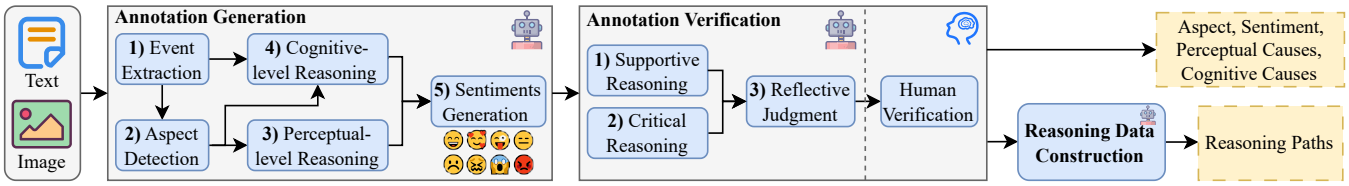

Figure 2: The pipeline of data construction, including annotation generation, verification, and reasoning data construction.

2.2 Annotation Verification

To ensure the annotation quality, we introduce a validation approach. It first leverages the reflective reasoning capability of MLLMs to verify annotated samples in a principled and self-consistent manner. Then it adopts a round of human verification for final quality assurance.

Inspired by adversarial prompting (Paulus et al. 2024) and argumentation-based evaluation (Liu et al. 2024a), our MLLM-based validation approach involves three steps. **1)** Supportive Reasoning. For a generated annotation, MLLM is prompted to assume the annotation is correct and provide supportive reasons by retrieving contextual and perceptual evidence from the image-text pairs. **2)** Critical Reasoning. The same annotation is then assumed incorrect, and MLLM critiques it to provide oppositive reasons by uncovering inconsistencies or implausible evidence. This step provides a counter-perspective to expose potential hallucinations. **3)** Reflective Judgment. Finally, MLLM considers the annotation along with the supportive and oppositive reasons, reflecting on both perspectives to decide whether the annotation could be accepted. This reflection step helps MLLM make a more robust and self-aware decision. The MLLM utilized in this validation approach is Qwen2.5-VL-72B.

Our verification approach encourages MLLM to reason from both supportive and critical views, reducing hallucinations and improving annotation quality. At last, we employed three human annotators to perform the final review, where each annotation was accepted by majority voting.

2.3 Reasoning Data Construction

With the aforementioned two stages, we obtain the final aspect-sentiment pairs and two-level explanations. Furthermore, we extract intermediate reasoning results produced during the CoT prompting process. The reasoning results capture the human-like sentiment inference trajectory from surface cues to contextual understanding. Specifically, we first extract the intermediate outputs from CoT prompting process (Sec. 2.1). These outputs are then concatenated using predefined templates, forming a complete reasoning trajectory of each aspect. The obtained reasoning paths data takes the form of a sequence, i.e., events $\rightarrow$ aspect terms $\rightarrow$ perceptual causes $\rightarrow$ cognitive causes $\rightarrow$ final open-vocabulary aspect-sentiment pairs.

2.4 OX-MABSR-Bench Overview and Insight

To summarize, our OX-MABSR-Bench includes five elements. **1)** The *aspect* is the sentiment target, which may be explicitly mentioned in the text or implicitly inferred from visual or contextual cues. **2)** The *sentiment* is a set of open-vocabulary labels describing expressive sentiments, free from predefined sentiment categories. **3)** The *perceptual causes* is a surface explanation based on observable cues such as facial expressions, visual aesthetics, scene context, and textual semantics. **4)** The *cognitive causes* is a deeper explanation based on contextual knowledge such as social factors and historical events. **5)** The *reasoning paths* is a complete sequence of intermediate inference steps from raw inputs to the final sentiment interpretation.

We highlight the characteristics of our proposed OX-MABSR-Bench. Table 2 reports primary statistics. It comprises a total of 14,705 image-text pairs, and includes over 70,000 aspect-sentiment pairs. Notably, over two-thirds of the aspects are implicit, which poses a significant challenge for models to infer the implicit aspects. Another property is its rich sentiment expressiveness. The dataset contains 872 unique sentiment labels, tremendously exceeding conventional predefined three sentiment categories. Each aspect contains an average of 1.86 sentiment labels, suggesting that

Statistic	Train	Dev	Test
# Image-Text Pairs	10,295	2,940	1,470
# Aspect-Sentiments Pairs	49,101	14,153	7,044
# Implicit Aspects	34,227	9,977	4,910
# Unique Sentiment Labels	872	566	413
Avg. Sentiments per Aspect	1.86	1.86	1.86
Avg. Reasoning Paths Steps	5.00 k	5.02 k	5.02 k

Table 2: Statistics of the OX-MABSR-Bench.

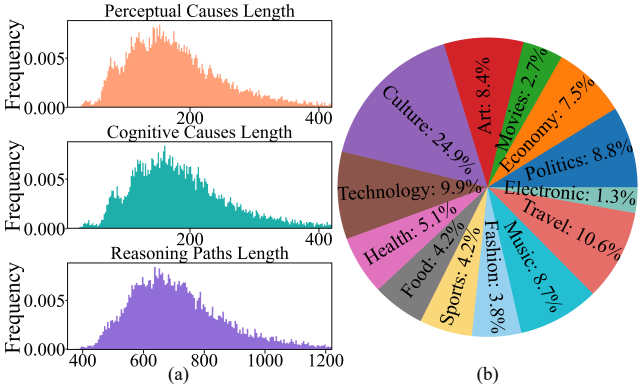

Figure 3: (a) Length distributions of Perceptual-cognitive causes and reasoning paths; (b) Domain topic distribution.

sentiments are multifaceted. Moreover, the average reasoning paths steps exceed 5,000 characters, showing the complexity involved in tracing sentiment causes.

Furthermore, we perform statistics on causes and reasoning paths, together with topics. As shown in Figure 3 (a), we present the length distribution for perceptual causes, cognitive causes, and reasoning paths. The length is measured by the number of words. The three elements exhibit distributions spanning a relatively long range. This indicates that explanations in our OX-MABSR-Bench are not shallow but instead involve complex reasoning processes. Finally, our dataset demonstrates broad topical coverage across real-world domains, such as culture, technology, travel, music, as shown in Figure 3 (b). Such diversity supports robust generalization across sociocultural and thematic variations.

3 OX-MABSR: Evaluation Benchmark

We first define our OX-MABSR task and then design a benchmark framework with appropriate evaluation metrics.

3.1 Task Definition

We define subtasks involved in OX-MABSR. The prediction task of open-vocabulary aspect-sentiment pairs requires jointly generating aspect terms and their sentiments in free form. In contrast, MABSA relies on span extraction and predefined sentiment categories. The generation task of two-level sentiment explanations aims to generate both perceptual and cognitive levels causes. The reasoning paths generation task aims to generate the reasoning trajectory.

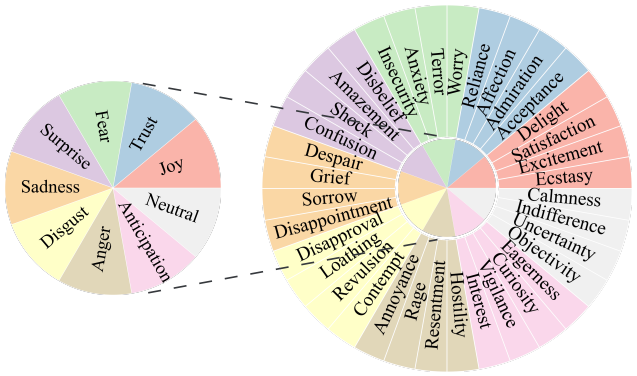

Figure 4: An illustration of Plutchik's sentiment wheel. Nine inner core sentiments are shown on the left and the corresponding outer fine-grained sentiments are on the right.

3.2 Aspect-Sentiment Joint Evaluation

The open-vocabulary sentiment labels in OX-MABSR pose challenges for the conventional evaluation. To this end, we propose a hierarchical sentiment mapping scheme based on Plutchik's sentiment wheel (Plutchik 1980), as shown in Figure 4. Our scheme consists of two successive normalization steps. First, a lemmatization function $\mathrm{lem}(\cdot)$ is applied to standardize the sentiment. Next, the standardized sentiment is mapped onto a hierarchical space using two functions. The mapping $S^{\text{in}}(\cdot)$ gives one of the nine inner sentiments, while the mapping $S^{\text{out}}(\cdot)$ produces outer fine-grained sentiments,

$$G^* = S^*(\mathrm{lem}(sentiment)). \tag{1}$$

Next, let $\mathcal{G} = \{(A_i, \mathcal{Y}_i)\}_{i=1}^{N}$ be the true aspect-sentiment pairs and $\mathcal{P} = \{(\hat{A}_j, \hat{\mathcal{Y}}_j)\}_{j=1}^{M}$ denote the predicted ones. We compute precision, recall, and F1 as follows,

$$\begin{cases} \mathrm{P}^* = \dfrac{1}{M}\displaystyle\sum_{j=1}^{M} \begin{cases} \frac{|G^*(\mathcal{Y}_i)\cap G^*(\hat{\mathcal{Y}}_j)|}{|G^*(\hat{\mathcal{Y}}_j)|}, & \text{if } \exists A_i = \hat{A}_j \\ 0, & \text{otherwise} \end{cases} \\ \mathrm{R}^* = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \begin{cases} \frac{|G^*(\mathcal{Y}_i)\cap G^*(\hat{\mathcal{Y}}_j)|}{|G^*(\mathcal{Y}_i)|}, & \text{if } \exists \hat{A}_j = A_i \\ 0, & \text{otherwise} \end{cases} \\ \mathrm{F1}^* = 2 \times \dfrac{\mathrm{P}^* \times \mathrm{R}^*}{\mathrm{P}^* + \mathrm{R}^*}, \end{cases} \tag{2}$$

where the star * represents two cases, in or out. In addition, M and N are the numbers of predicted and true pairs.

3.3 Explanations and Reasoning Path Evaluation

We evaluate the model's ability to generate accurate and meaningful causes and reasoning paths using two groups of metrics. **1)** The lexical-level metrics involve BLEU-1 and ROUGE-L. These metrics measure the 1-gram overlap (Papineni et al. 2002) and the longest common subsequence (Lin 2004) between the predicted output $\hat{g}$ and the true output g. **2)** The semantic-level metric is Semantic Coherence (SC). We define a coherence scoring function f_{sc} as follows,

$$\text{SemanticCoherence} = f_{\text{sc}}(\hat{g}, g). \tag{3}$$

The function f_{sc} uses Qwen3-based prompt evaluation (See Appendix B). The coherence score ranges from 0 to 1. A larger score of SC means better consistency.

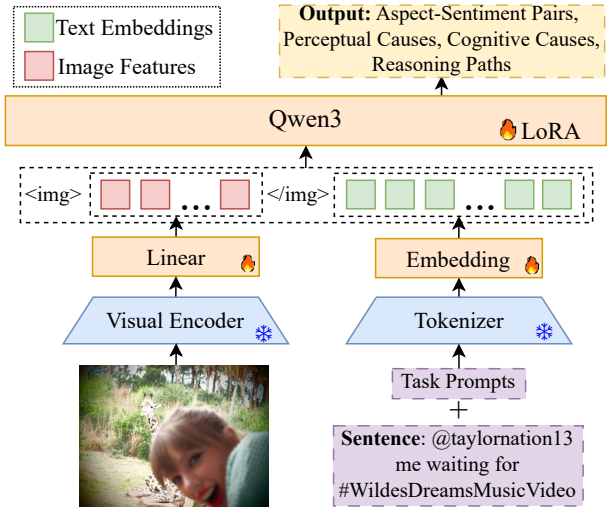

Figure 5: Overall architecture of our MABSR-LLM model.

4 Our Proposed MABSR-LLM

We present Multimodal Aspect-Based Sentiment Reasoning LLM (MABSR-LLM) to tackle OX-MABSR challenges.

4.1 Model Architecture

Figure 5 shows the architecture of our proposed MABSR-LLM. The model is designed to process both visual and textual inputs in a unified generative framework, enabling open-domain and explainable sentiment reasoning. Specifically, we use SigLIP-2 (Tschannen et al. 2025) to extract image features, which are projected into the text space via a linear projection layer. Textual inputs, including task prompts and sentence, are encoded using the embedding layer of Qwen3-8B (Qwen Team 2025). To distinguish visual inputs from textual ones, we enclose the projected image features with two tokens $\langle \text{img} \rangle$ and $\langle /\text{img} \rangle$, and concatenate them with text embeddings. The obtained sequence is fed into Qwen3, which jointly attends to image-text tokens via self-attention. Guided by task prompts, the model generates structured output elements, including aspect–sentiment pairs, perceptual and cognitive causes, and reasoning paths.

4.2 Training Objective

We train our MABSR-LLM in two stages with the corresponding objectives. Among them, LoRA-based (Hu et al. 2022) fine-tuning is applied to Qwen3 for efficient adaptation, while the image projection layer is fully tuned to ensure effective cross-modal alignment. Firstly, we perform Supervised Fine-Tuning (SFT) using OX-MABSR-Bench. All subtasks are trained jointly using multi-task learning. The objective is to minimize the standard auto-regressive language modeling loss, i.e.,

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, x), \tag{4}$$

where x denotes the multimodal input and $y_{1:T}$ is the target output sequence. This stage equips MABSR-LLM with the ability of foundational sentiment reasoning.

Method	Aspect-Sentiment Joint Evaluation					
	P^{in}	R^{in}	$F1^{in}$	P^{out}	R^{out}	$F1^{out}$
mPLUG-Owl-7B	11.36	10.91	11.13	10.29	9.64	9.95
DeepSeek-VL2	10.46	13.29	11.70	10.11	11.62	10.81
LLaVA-NeXT-7B	12.71	12.02	12.35	13.26	12.94	13.09
Qwen2.5-VL-7B	14.52	11.33	12.72	13.51	10.91	12.07
Janus-Pro-7B	13.59	15.85	14.63	11.82	13.07	12.41
InternVL3-8B	15.73	14.49	15.08	16.49	15.11	15.76
MABSR-LLM	**44.94**	**44.37**	**44.65**	**40.79**	**40.53**	**40.66**

Table 3: Performance comparison on Aspect-Sentiment.

Secondly, we apply GRPO-based reinforcement learning (RL) (Shao et al. 2024b) to improve the model's reasoning ability. Specifically, we define a hybrid reward function. A lexical-level reward f_{rouge} is first applied, which evaluates lexical overlap based on ROUGE-L. Although this reward captures the literal correctness, it could penalize the semantically valid output. Therefore, a semantic-level reward f_{sc} that measures semantic coherence is applied. The final reward is a weighted sum of these two terms, i.e.,

$$r(\hat{y}, y) = \lambda f_{\text{rouge}}(\hat{y}, y) + (1-\lambda) f_{\text{sc}}(\hat{y}, y), \tag{5}$$

where y is the true output and $\hat{y}$ is the predicted one. For better semantic consistency, we set $\lambda = 0.3$ to downweight ROUGE-L for lexical diversity and upweight SC. Finally, our GRPO loss encourages the model to generate outputs with higher rewards via preference-based optimization,

$$\mathcal{L}_{\text{GRPO}} = \mathbb{E}_{\hat{y}\sim\pi_\theta}\left[r(\hat{y}, y) \log \pi_\theta(\hat{y})\right]. \tag{6}$$

Here, the model π_θ is parameterized by θ.

5 Experiments and Analysis

5.1 Experimental settings

Baselines. We benchmark our MABSR-LLM against existing MLLMs on OX-MABSR-Bench. The baseline models include mPLUG-Owl (Ye et al. 2024), DeepSeek-VL2 (Wu et al. 2024b), LLaVA-NeXT (Liu et al. 2024b), Qwen2.5-VL (Bai et al. 2025), InternVL3 (Zhu et al. 2025b) and Janus-Pro (Chen et al. 2025). Each model receives identical multimodal inputs and unified task prompts (See Appendix C). They are required to generate aspect–sentiment pairs, perceptual and cognitive causes, and reasoning paths.

Implementation Details. We train our model on four A6000 GPUs using AdamW optimizer (Loshchilov and Hutter 2017) and the LoRA rank is set to 8. For SFT, we use a batch size of 4 and a learning rate of $5e^{-5}$ for 5 epochs. For GRPO, we train for $10e^3$ steps with a learning rate of $5e^{-5}$.

5.2 Main Results

Table 3 and Table 4 report the performance of MABSR-LLM against other MLLMs on OX-MABSR-Bench. **1)** Aspect–Sentiment Pairs Prediction. Our model achieves the best performance, outperforming the second-best model InternVL3-8B by 29.57 and 24.90 F1-score, respectively.

Method	Perceptual Causes			Cognitive Causes			Reasoning Paths		
	BLEU-1	ROUGE-L	SC	BLEU-1	ROUGE-L	SC	BLEU-1	ROUGE-L	SC
mPLUG-Owl-7B (Ye et al. 2024)	15.62	16.73	31.83	22.43	18.14	50.82	8.09	12.17	49.33
DeepSeek-VL2 (Wu et al. 2024b)	12.40	15.86	44.04	19.59	16.05	47.42	7.18	9.34	45.13
LLaVA-NeXT-7B (Liu et al. 2024b)	16.54	18.36	46.70	21.05	18.27	55.65	7.54	11.73	47.28
Janus-Pro-7B (Chen et al. 2025)	19.11	18.41	34.21	23.68	18.66	51.42	9.13	12.31	52.40
Qwen2.5-VL-7B (Bai et al. 2025)	29.91	19.77	59.40	26.01	17.90	57.96	14.80	15.38	51.94
InternVL3-8B (Zhu et al. 2025b)	26.63	18.35	54.80	27.94	18.93	58.47	15.89	14.61	53.76
MABSR-LLM (Ours)	**45.46**	**40.79**	**71.21**	**45.55**	**40.60**	**70.53**	**61.73**	**47.82**	**72.17**

Table 4: Performance comparison on perceptual causes, cognitive causes, and reasoning paths generation.

Method	Aspect-Sentiment		Perceptual Causes			Cognitive Causes			Reasoning Paths		
	$F1^{in}$	$F1^{out}$	BLEU-1	ROUGE-L	SC	BLEU-1	ROUGE-L	SC	BLEU-1	ROUGE-L	SC
w/o Multimodal	31.15	27.67	31.28	29.97	48.34	32.07	30.42	50.39	45.91	34.25	52.78
w/o RL	38.61	35.22	40.05	34.68	62.87	39.97	33.71	60.88	57.28	41.33	65.18
w/o RP-Task	41.72	38.69	42.85	37.94	67.26	41.61	36.52	66.78	51.38	38.16	60.77
Full MABSR-LLM	**44.65**	**40.66**	**45.46**	**40.79**	**71.21**	**45.55**	**40.60**	**70.53**	**61.73**	**47.82**	**72.17**

Table 5: Ablation results on MABSR-LLM, showing the effect of various components on performance across all subtasks.

While general-purpose MLLMs can partially identify aspect–sentiment pairs via prompts, they struggle with open-vocabulary settings. In contrast, our model demonstrates remarkable performance through two-stage learning. **2)** Sentiment Explanations Generation & Reasoning Paths Generation. Our model achieves the highest performance on both subtasks. For example, in the first task, it surpasses the second-best model by 11.81 and 12.03 in SC score, respectively. In the second task, it surpasses the second-best model by 18.41 in SC score. Baseline MLLMs tend to produce generic explanations and lack step-by-step sentiment reasoning, whereas our model conducts multi-step reasoning to generate coherent causes and sentiment trajectories based on perceptual and contextual cues.

5.3 Ablation Study

On Model Component. The result is given in Table 5. *w/o* Multimodal Input denotes that we train text-only MABSR-LLM without images. The performance degrades greatly, showing the importance of images in reasoning implicit aspects, expressive sentiments and causes. *w/o* RL denotes the model trained only with SFT. The corresponding performance drops, showing RL's effectiveness in improving accuracy and semantic consistency. *w/o* Reasoning Path Task denotes the model is trained with only aspect–sentiment pairs and two-level explanations. The big drop in reasoning path performance shows that the model struggles with step-wise reasoning. Moreover, the aspect–sentiment prediction and explanations generation also decline, showing that learning reasoning paths helps improve the model's sentiment ability.

On LLM Backbone. We replace the Qwen3 backbone in MABSR-LLM with other LLMs, including LLaMA3-8B (Llama Team 2024) and ChatGLM3-6B (GLM et al. 2024). As shown in Figure 6 (a), Qwen3 outperforms other backbones in semantic coherence, validating its effectiveness.

On LoRA Rank. As shown in Figure 6 (b), we find that

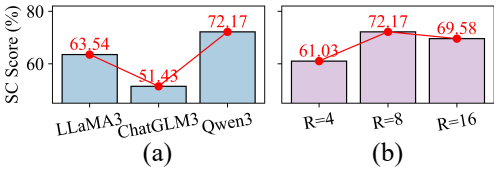

Figure 6: Ablation study on (a) LLMs and (b) LoRA rank.

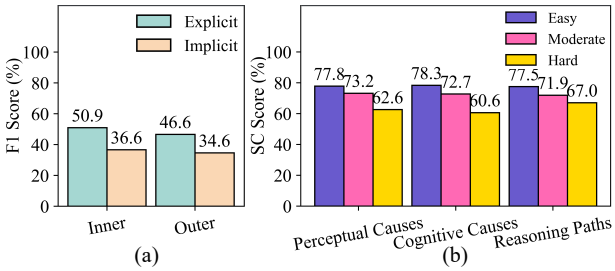

Figure 7: (a) F1 for two types of aspects prediction; (b) SC for explanation generation under three-level difficulty.

a rank of 8 yields the best reasoning quality. In contrast, a lower rank of 4 results in degraded performance, due to a limited learning capacity. A slight performance drop with the rank increased to 16, possibly due to the model's overfitting. The rank of 8 is our best choice in LoRA.

On Open-vocabulary Aspects. As shown in Figure 7 (a), we analyze the performance gap between explicit and implicit aspects in the aspect-sentiment pairs prediction task. The results clearly show a significant drop in F1 scores for implicit aspects. These implicit aspects can only be inferred from visual cues and textual context, indicating that recognizing them remains a challenge due to lack of direct cues.

On Explanation Difficulty. We divide explanation difficulty levels based on the explanation length. The top one-

third with the shortest length is considered Easy. The middle and the bottom are taken as Moderate and Hard, respectively. In Figure 7 (b), we report the sentiment explanations and reasoning paths generation tasks under three difficulty levels. Performance consistently drops with increasing difficulty, suggesting longer, more compositional explanations are harder to generate due to higher reasoning complexity.

The above two studies show that our dataset remains challenging and is a valuable benchmark for OX-MABSR task.

5.4 Case Study

We present a case study comparing our MABSR-LLM with the second-best MLLM InternVL3-8B, as shown in Figure 8. The aspect "The Giving Pledge" is not explicitly mentioned in the textual input. The competitor InternVL3-8B fails to recognize it. In contrast, our model identifies this aspect by leveraging visual cues and background knowledge, specifically the blue-box region, which includes "99% of his wealth will go to charity" and shows a strong connection to this aspect. This demonstrates our model's prediction capacity for multimodal implicit aspects. In sentiment prediction, the competitor produces only a single rather simplistic sentiment for this aspect. Conversely, our model is capable of generating multiple and more expressive sentiments. Furthermore, regarding explanations generation, the competitor produces shallow and generic causes (e.g., "his achievements" or "public persona") without considering observable details and background knowledge. In contrast, our MABSR-LLM successfully identifies visual causes, such as "friendly expression", "calm demeanor", and "long-standing success". In addition, it captures more abstract background attributions, including "a principled approach to life" and "philanthropy that inspired millions globally".

	Image	Text
Input		Wizard of Omaha #Warren Buffett turns 85 @WarrenBuffett

	Ground Truth	MABSR-LLM (Ours)	InternVL3-8B
Aspect-Sentiment	**Warren Buffett**,[Confident, Admirable],**The Giving Pledge,**[Approval,Appreciative]	**Warren Buffett**,[Confident, Admirable],**The Giving Pledge,**[Approval,Appreciative]	[**Warren Buffett**, Positive]
Perceptual Causes	**Warren Buffett's** sentiment is approachable and confident, conveyed through his friendly expression, relaxed posture, and the emphasis on his generosity and long-standing success. **The Giving Pledge's** sentiment is approval and support, caused by its alignment with Buffett's dedication to donating the majority of his wealth to charity.	**Warren Buffett's** sentiment is content, confident,and generous, conveyed through his friendly expression,calm demeanor, philanthropic values, and long-standing business success. **The Giving Pledge's** sentiment is approval and support, caused by Buffett's commitment to giving away the majority of his wealth to charitable causes.	**Warren Buffett**, known as the Oracle of Omaha, is highlighted for his age,wealth, generosity, intelligence, and philanthropic efforts.
Cognitive Causes	**Warren Buffett's** sentiment is admirable and respectable, caused by his financial acumen, philanthropy, humble lifestyle, and generous commitment to giving away 99% of his wealth. **The Giving Pledge's** sentiment is compassionate and generous, caused by its mission to encourage the wealthy to donate most of their wealth, reflecting Buffett and Gates's commitment to social good.	**Warren Buffett's** sentiment is admirable and respectable, caused by his exceptional financial acumen, generosity, principled approach to life, which have inspired millions globally. **The Giving Pledge's** sentiment is appreciative and generous, caused by it reflects Buffett's commitment to philanthropy and resonates with those who value social responsibility and wealth redistribution.	**Warren Buffett**, the famed Oracle of Omaha, turned 85, a major milestone . The deeper cause stems from his success, charity, and profile.

Figure 8: A case result by MABSR-LLM and a competitor.

6 Related Work

6.1 Aspect-Based Sentiment Analysis

TSA has evolved from coarse-grained classification (Cambria et al. 2020; Zhang et al. 2023) to fine-grained ABSA (Yin and Zhong 2024), which identifies sentiments of specific aspects and has been extended to extract sentiment tuples (Nie et al. 2024; Jian et al. 2025). With the popularity of multimodal content, researchers have explored MSA (Gao et al. 2024; Xia et al. 2025) using visual cues to enhance sentiment analysis. Furthermore, MABSA integrates textual and visual inputs to detect aspect-level sentiment (Zhu et al. 2025a; Liu et al. 2025a; Sun et al. 2025), emphasizing cross-modal alignment. Although MABSA has made significant progress, most existing works still focus on sentiment classification, without uncovering deeper causes, and assume aspects are explicit and sentiment are predefined.

In contrast, we propose OX-MABSR task, which extends MABSA to predict open-vocabulary aspect-sentiment pairs and generate explanations and reasoning paths.

6.2 LLM-Guided Data Annotation

Recently, LLMs are widely used for scalable data annotation, providing richer context and commonsense than traditional methods. CoT prompting (Wei et al. 2022; Sahoo et al. 2024) improves LLMs' complex reasoning by guiding step-by-step inference. In data construction, several studies leverage CoT prompting (Shao et al. 2024a; Chen et al. 2024) to generate explanations and structured datasets, such as visual explanations (Mitra et al. 2024) and explainable QA datasets (Chu et al. 2025). These works show that LLMs can serve as annotators and reasoners, generating more detailed, explanation-oriented supervision signals. We construct OX-MABSR-Bench, driven by MLLMs and assisted by humans. We further design MABSR-LLM to address this task.

7 Conclusion and Future Work

In this paper, we introduce a novel task, OX-MABSR. It significantly extends traditional MABSA tasks to open-vocabulary aspect-sentiment prediction and explainable sentiment reasoning. To benchmark this task, we construct OX-MABSR-Bench, a challenging dataset annotated with explicit/implicit aspects, expressive sentiment categories, and two-level (perceptual/cognitive) explanations along with reasoning paths. To address this task, we propose MABSR-LLM, a step-wise multimodal reasoning model for open-domain explainable aspect-based sentiment understanding. Experiments show it outperforms strong baselines.

In the future, we will enhance contextual reasoning with commonsense knowledge and causal inference. We plan to extend this benchmark with additional modalities (e.g., audio and video) and to the multimodal dialogue domain.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFC3305902, by the Special Project for Industrial Foundation Reconstruction and High-Quality Development of Manufacturing under Grant Agreement No. ZC25T320057/100, by the National Natural Science Foundation of China No. 62076032, by the China Computer Federation of Zhipu Foundation No. CCF-Zhipu202407, by Industry-University-Research Innovation Fund for Chinese Universities No. 2024MZ028, and by BUPT Kunpeng and Ascend Center of Cultivation. The authors thank the anonymous editors and reviewers for their valuable comments on improving the final version of this paper.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Cai, H.; Xia, R.; and Yu, J. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, 340–350.

Cambria, E.; Li, Y.; Xing, F. Z.; Poria, S.; and Kwok, K. 2020. SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In *Proceedings of the 29th ACM international conference on information & knowledge management*, 105–114.

Chen, Q.; Qin, L.; Zhang, J.; Chen, Z.; Xu, X.; and Che, W. 2024. M^3CoT: A Novel Benchmark for Multi-Domain Multi-step Multi-modal Chain-of-Thought. arXiv:2405.16473.

Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; and Ruan, C. 2025. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling. arXiv:2501.17811.

Chu, X.; Tan, Z.; Xue, H.; Wang, G.; Mo, T.; and Li, W. 2025. Domaino1s: Guiding LLM Reasoning for Explainable Answers in High-Stakes Domains. *arXiv preprint arXiv:2501.14431*.

Gao, Z.; Jiang, X.; Xu, X.; Shen, F.; Li, Y.; and Shen, H. T. 2024. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26876–26885.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Han, Z.; Hu, M.; Bai, Y.; Wang, X.; and Luo, B. 2025. DEQA: Descriptions Enhanced Question-Answering Framework for Multimodal Aspect-Based Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22): 23987–23995.

Hossain, M. S.; Hossain, M. M.; Chaki, S.; Mridha, M.; Rahman, M. S.; and Moni, M. A. 2025. Dimension-Wise Gated Cross-Attention for Multimodal Sentiment Analysis. In *Companion Proceedings of the ACM on Web Conference 2025*, 1979–1987.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Hu, M.; and Liu, B. 2004. Mining opinion features in customer reviews. In *AAAI*, volume 4, 755–760.

Jian, Z.; Chen, Y.; Li, J.; Wang, S.; Zeng, X.; Yao, J.; An, X.; and Wu, Q. 2025. SimRP: Syntactic and Semantic Similarity Retrieval Prompting Enhances Aspect Sentiment Quad Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 24248–24256.

Li, R.; Chen, H.; Feng, F.; Ma, Z.; Wang, X.; and Hovy, E. 2021. Dual graph convolutional networks for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 6319–6329.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Liu, B. 2020. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Studies in Natural Language Processing. Cambridge University Press. ISBN 9781108486378.

Liu, C.; Wang, Y.; Flanigan, J.; and Liu, Y. 2024a. Large language model unlearning via embedding-corrupted prompts. *Advances in Neural Information Processing Systems*, 37: 118198–118266.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, X.; Li, R.; Ye, S.; Zhang, G.; and Wang, X. 2025a. Multimodal Aspect-Based Sentiment Analysis under Conditional Relation. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 313–323. Abu Dhabi, UAE: Association for Computational Linguistics.

Liu, X.; Lin, P.; Tu, X.; Jia, W.; Jiang, C.; and Li, R. 2025b. SDG-MLLM: Injecting Structured Dialogue Graphs into MLLM for Multimodal Conversational Aspect-Based Sentiment Analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 14114–14121.

Llama Team. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mitra, C.; Huang, B.; Darrell, T.; and Herzig, R. 2024. Compositional Chain-of-Thought Prompting for Large Multimodal Models. arXiv:2311.17076.

Nie, Y.; Fu, J.; Zhang, Y.; and Li, C. 2024. Modeling implicit variable and latent structure for aspect-based sentiment quadruple extraction. *Neurocomputing*, 586: 127642.

Niu, T.; Zhu, S.; Pang, L.; and El-Saddik, A. 2016. Sentiment Analysis on Multi-View Social Data. In *MultiMedia Modeling*, 15–27.

Ouyang, J.; Yang, Z.; Liang, S.; Wang, B.; Wang, Y.; and Li, X. 2024. Aspect-based sentiment analysis with explicit sentiment augmentations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 18842–18850.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Paulus, A.; Zharmagambetov, A.; Guo, C.; Amos, B.; and Tian, Y. 2024. Advprompter: Fast adaptive adversarial prompting for llms. *arXiv preprint arXiv:2404.16873*.

Peng, T.; Li, Z.; Wang, P.; Zhang, L.; and Zhao, H. 2024. A novel energy based model mechanism for multi-modal aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18869–18878.

Plutchik, R. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, 3–33. Elsevier.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; Al-Smadi, M.; Al-Ayyoub, M.; Zhao, Y.; Qin, B.; De Clercq, O.; et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, 19–30.

Qwen Team. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Sahoo, P.; Singh, A. K.; Saha, S.; Jain, V.; Mondal, S.; and Chadha, A. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Shao, H.; Qian, S.; Xiao, H.; Song, G.; Zong, Z.; Wang, L.; Liu, Y.; and Li, H. 2024a. Visual CoT: Advancing Multi-Modal Language Models with a Comprehensive Dataset and Benchmark for Chain-of-Thought Reasoning. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 8612–8642. Curran Associates, Inc.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Sun, K.; Wu, H.; Shi, B.; Mensah, S.; Liu, P.; and Dong, B. 2025. VERO: Verification and Zero-Shot Feedback Acquisition for Few-Shot Multimodal Aspect-Level Sentiment Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25210–25218.

Tschannen, M.; Gritsenko, A.; Wang, X.; Naeem, M. F.; Alabdulmohsin, I.; Parthasarathy, N.; Evans, T.; Beyer, L.; Xia, Y.; Mustafa, B.; Hénaff, O.; Harmsen, J.; Steiner, A.; and Zhai, X. 2025. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. arXiv:2502.14786.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, D.; Yang, D.; Zhou, Y.; and Ma, C. 2024a. Robust Multimodal Sentiment Analysis of Image-Text Pairs by Distribution-Based Feature Recovery and Fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 5780–5789. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706868.

Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; Xie, Z.; Wu, Y.; Hu, K.; Wang, J.; Sun, Y.; Li, Y.; Piao, Y.; Guan, K.; Liu, A.; Xie, X.; You, Y.; Dong, K.; Yu, X.; Zhang, H.; Zhao, L.; Wang, Y.; and Ruan, C. 2024b. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. arXiv:2412.10302.

Xia, W.; Jia, G.; Zhao, S.; and Yang, J. 2025. Seek Common Ground While Reserving Differences: Semi-Supervised Image-Text Sentiment Recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29601–29611.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; Li, C.; Xu, Y.; Chen, H.; Tian, J.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2024. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. arXiv:2304.14178.

Yin, S.; and Zhong, G. 2024. Textgt: A double-view graph transformer on text for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 19404–19412.

Yu, J.; and Jiang, J. 2019. Adapting BERT for target-oriented multimodal sentiment classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. IJCAI.

Zhang, W.; Deng, Y.; Liu, B.; Pan, S. J.; and Bing, L. 2023. Sentiment Analysis in the Era of Large Language Models: A Reality Check. arXiv:2305.15005.

Zhang, X.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Zhu, A.; Hu, M.; Wang, X.; Yang, J.; Tang, Y.; and An, N. 2025a. DaNet: Dual-Aware Enhanced Alignment Network for Multimodal Aspect-Based Sentiment Analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, 14369–14381.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025b. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Zhu, L.; Sun, H.; Gao, Q.; Liu, Y.; and He, L. 2025c. Aspect Enhancement and Text Simplification in Multimodal Aspect-Based Sentiment Analysis for Multi-Aspect and Multi-Sentiment Scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1683–1691.