Dual-Path Counterfactual Integration for Multimodal Aspect-Based Sentiment Classification

Rui Liu[1,2], Jiahao Cao[3], Jiaqian Ren[4], Xu Bai[1,5]*, Yanan Cao[1,5]
[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] China Mobile Communications Corporation, Jiutian Team, [3]ByteDance Ltd.
[4]China Mobile (Hangzhou) Information Technology Co., Ltd.
[5] School of Cyber Security, University of Chinese Academy of Sciences
liuruiyjy@chinamobile.com, caojiahao.98@bytedance.com,
renjiaqian@cmhi.chinamobile.com, {baixu, caoyanan}@iie.ac.cn

Abstract

Multimodal aspect-based sentiment classification (MABSC) requires fine-grained reasoning over both textual and visual content to infer sentiments toward specific aspects. However, existing methods often rely on superficial correlations—particularly between aspect terms and sentiment labels—leading to poor generalization and vulnerability to spurious cues. To address this limitation, we propose DPCI, a novel Dual-Path Counterfactual Integration framework that enhances model robustness by explicitly modeling counterfactual reasoning in multimodal contexts. Specifically, we design a dual counterfactual generation module that simulates two types of interventions: replacing aspect terms and rewriting descriptive content, thereby disentangling the spurious dependencies from causal sentiment cues. We further introduce a sample-aware counterfactual selection strategy to retain high-quality, diverse counterfactuals tailored to each generation path. Finally, a confidence-guided integration mechanism adaptively fuses counterfactual signals into the main prediction stream. Extensive experiments on standard MABSC benchmarks demonstrate that DPCI not only achieves state-of-the-art performance but also significantly improves model robustness.

1 Introduction

Multimodal aspect-based sentiment classification (MABSC) has attracted increasing attention in recent years. This task aims to determine the sentiment polarity of specific aspect terms by jointly leveraging textual and visual information. As illustrated in Figure 1, the image-text pair contains two aspects: *"man"* and *"Elijah Jordan Wood"*. By integrating the textual description with the visual cue of a smiling face, the sentiment polarity toward *"man"* is identified as positive. In contrast, the sentiment polarity for *"Elijah Jordan Wood"* is assessed as neutral. This example highlights the importance of effectively capturing aspect-related contextual information within image-text pairs.

Figure 1: An example of multimodal aspect-based sentiment classification (MABSC).

Existing research (Xu et al., 2019; Yu and Jiang, 2019a; Khan and Fu, 2021; Ling et al., 2022; Yang and Li, 2023; Feng et al., 2024) on multimodal aspect sentiment classification have primarily focused on designing various strategies to model the semantic relationship between target aspects and their context. These include employing attention mechanisms to capture cross-modal aspect features and leveraging multi-task learning frameworks. More recently, large language models (LLMs) have also been introduced into the MABSC task. While these approaches have led to performance improvements, they largely overlook dataset bias issues, which may hinder further advances in model generalization and robustness.

Although MABSC has made significant progress in recent years, current research rarely focuses on the causal relationship between aspect terms and sentiment labels, especially ignoring the spurious correlation between them. The spurious correlation between aspect words and sentiment labels refers to the fact that due to data imbalance and label bias in the training corpus, the model tends to learn biased statistical information rather than intrinsic contextual semantic information when predicting the sentiment of aspect words, thereby making incorrect sentiment predictions.

*The corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22748–22758
November 4-9, 2025

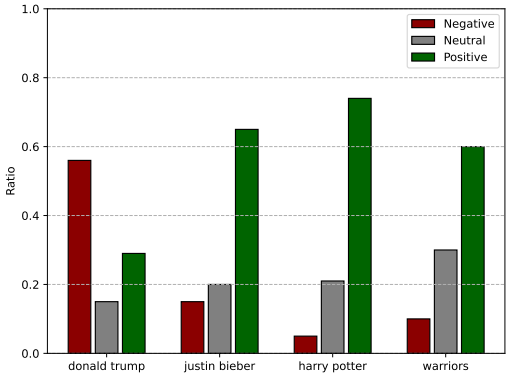

Figure 2: The sentiment distribution of several aspects.

Figure 2 shows that certain aspects in Twitter-17 exhibit skewed sentiment distributions. For example, *"harry potter"* is predominantly positive in training dataset. In the test set, the advanced MABSC models misclassified most neutral *"harry potter"* samples as positive, consistent with the dominant training distribution. Similar trends are observed for other aspects, confirming that aspect–sentiment imbalance translates into systematic prediction errors.

To mitigate the impact of spurious correlations between aspect terms and sentiment labels in multimodal aspect-based sentiment classification (MABSC), we propose a novel Dual-Path Counterfactual Integration (DPCI) framework, as illustrated in Figure 4. Unlike previous methods that passively model semantic relationships, our approach is grounded in a causal perspective. Specifically, we decompose the causal structure of MABSC and identify the major sources of spurious correlations: biased co-occurrence between aspect terms and sentiment labels. To address this issue, we design a dual-path counterfactual generation module that constructs both aspect and description counterfactuals. Aspect counterfactuals are created by replacing the target aspect while preserving contextual semantics, whereas description counterfactuals alter the surrounding descriptions (including image captions generated by LLaVA) while holding the aspect fixed.

To fully exploit the knowledge from the dual-path counterfactuals, we propose a novel integration-based fusion mechanism. During training, we construct augmented samples that blend both types of counterfactual knowledge into a unified representation and fine-tune LLaMA with LoRA. During inference, we introduce a confidence-aware, training-free fusion strategy that adaptively integrates predictions from the original and counterfactual views based on their predictive confidence. This approach enables flexible and robust decision-making without additional training overhead. We evaluate DPCI on two widely used MABSC benchmarks, Twitter-15 and Twitter-17. Experimental results show that our framework consistently outperforms existing state-of-the-art methods.

Our contributions are as follows:

- We propose a novel Dual-Path Counterfactual Integration (DPCI) framework that explicitly models the causal relationships, effectively mitigating spurious correlations overlooked by traditional MABSC models in multimodal aspect-based sentiment classification.
- We develop dual-path counterfactual generation and selection strategy that constructs both aspect and context counterfactual samples, combined with a confidence-based sample fusion mechanism to enhance the counterfactual learning ability and reduce noise.
- We conduct extensive experiments on two multimodal sentiment datasets, demonstrating that our DPCI framework consistently outperforms existing competitive methods.

2 Related Work

2.1 Multimodal Aspect-based Sentiment Classification

Multimodal aspect-based sentiment classification (MABSC) (Ling et al., 2022; Yang and Li, 2023; Wang et al., 2024; Liu et al., 2025) is a new task that has emerged in recent years. Compared with multimodal sentiment analysis (Xu et al., 2018; Truong and Lauw, 2019; Hazarika et al., 2020; Yang et al., 2023; Li et al., 2024) and aspect-based sentiment classification (Xu et al., 2019; Liu et al., 2022; Chang et al., 2024; Ouyang et al., 2024), this task focuses on determining the sentiment polarity of fine-grained aspect terms within text–image pairs, which poses greater challenges. (Xu et al., 2019) first proposed a multi-interactive memory network to capture the relationships between modalities. (Yu and Jiang, 2019a) proposed two publicly annotated multimodal Twitter datasets and designed a target attention mechanism to learn the alignment. (Khan and Fu, 2021) translated the image into text and then sent it into BERT as an auxiliary sentence

for MABSC. (Yu et al., 2022) proposed a new multi-task learning architecture to achieve fine-grained image-target matching. (Feng et al., 2024) used the instruction tuning paradigm and leveraged the ability of large vision-language models to alleviate the limitation in the fusion stage. Previous works have achieved impressive results, but they have ignored the impact of aspect bias. In multimodal data analysis, especially when processing text and images from social media, sentiment analysis models are easily disturbed by this bias, which may affect their performance.

2.2 Spurious Correlation Mitigation

In recent years, research on mitigating spurious correlations, such as shortcuts, dataset biases, and group robustness, has been widely used in various fields such as computer vision (Wang et al., 2021) and natural language processing (Chang et al., 2024). Spurious correlation refers to a situation where two variables appear to be correlated, but this relationship is accidental or confounded with an external variable (Ye et al., 2024). Such correlations can mislead the model and may affect the generalization and robustness of the model. To address this, some studies (Srivastava et al., 2020; Wu et al., 2023) modify the input of the model to enhance the generality and diversity of data distribution. Other studies focus on improving the representation within the model through methods like causal intervention (Wang et al., 2021; Agarwal et al., 2020; Yang et al., 2024; Xu et al., 2025; Yang et al., 2024) and invariant learning (Krueger et al., 2021; Eastwood et al., 2023), to help the model better capture the potential relationship between variables.

3 Preliminary

3.1 Task Definition

Given a set of multimodal samples M, each sample $m_i \in M$ consists of sentence s_i, image v_i and aspect words a_i. Multimodal aspect-based sentiment analysis aims to identify the sentiment polarity $y_i \in \{positive, negative, neutral\}$ of the given aspect a_i, where the sentence $s_i = [\omega_1, \ldots, \omega_{a+1}, \ldots, \omega_{a+m}, \ldots, \omega_n]$ is a sequence consisting of n words. $a_i = [\omega_{a+1}, \ \ldots, \ \omega_{a+m}]$ is the aspect consisting of m words.

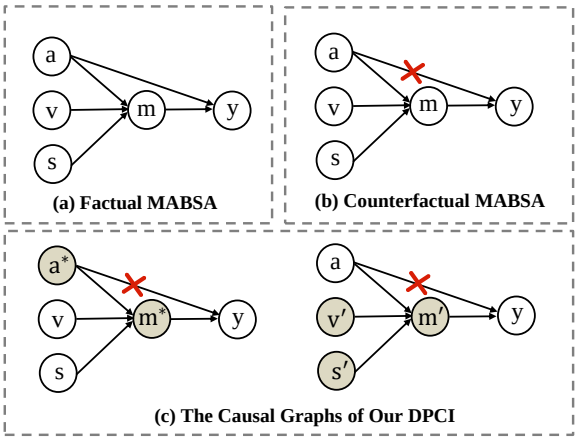

Figure 3: The causal graphs for factual MABSC, counterfactual MABSC and our DPCI. a and a^*: original and replaced aspect term. v and v': original and counterfactual image information. s and s': original and counterfactual sentence. m, m^* and m': multimodal fusion knowledge of (a,v,s), (a^*,v,s) and (a,v',s'). y: sentiment label.

3.2 Cause-Effect

We use causal graphs to illustrate the traditional MABSC method and our DPCI framework as shown in Figure 3. The causal graph reflects the causal relationship between variables, where $\rightarrow$ represents the direct effect between variables. The traditional MABSC model indirectly acts on the sentiment label through the joint action of images, sentences, and aspects. This usually involves the intermediate role of the multimodal information mediator M, that is, $(A, V, S) \rightarrow M \rightarrow Y$. However, due to the bias of the aspect term, it also affects the sentiment label through a direct path, that is, $A \rightarrow Y$. Therefore, we can rewrite $Y_{a,v,s}$ as a function $Z(\cdot)$ of A, V, S and M:

$$Y_{a,v,s} = Z(A = a, V = v, S = s, M = m) \quad (1)$$

where $m = M_{a,v,s}$ denotes the multimodal fusion knowledge of a, v and s. However, traditional MABSC methods ignore the causal impact of aspect on the results, leading the model to learn spurious correlations between aspect and sentiment labels, which negatively impacts its performance.

We want to remove the influence of aspect bias, that is, remove the direct path $A \rightarrow Y$, as shown in Figure 3(b). In our DPCI framework, we propose two complementary strategies to eliminate aspect bias. We first replace the value of A to a^*, the V and S remain unchanged, and M will reach a value

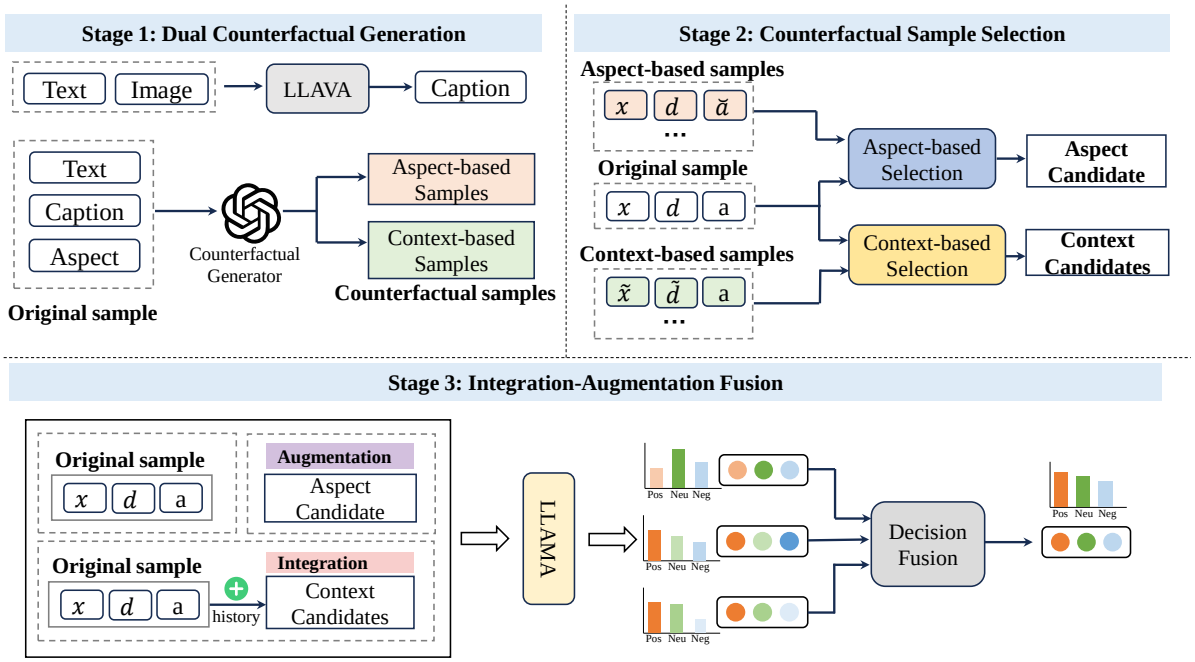

Figure 4: The proposed dual-path counterfactual integration (DPCI) framework.

of m^*. In this case, $Y_{a*,v,s}$ is represented as:

$$Y_{a^*,v,s} = Z(A = a^*, V = v, S = s, M = m^*) \tag{2}$$

Another strategy is to keep A unchanged and replace the values of V and S with v' and s', respectively. In this case, the value of M becomes m'. Then $Y_{a,v',s'}$ is expressed as:

$$Y_{a,v',s'} = Z(A = a, V = v', S = s', M = m') \tag{3}$$

This will eliminate the direct impact of aspect bias on the prediction results, help the model better learn the intrinsic semantic connection.

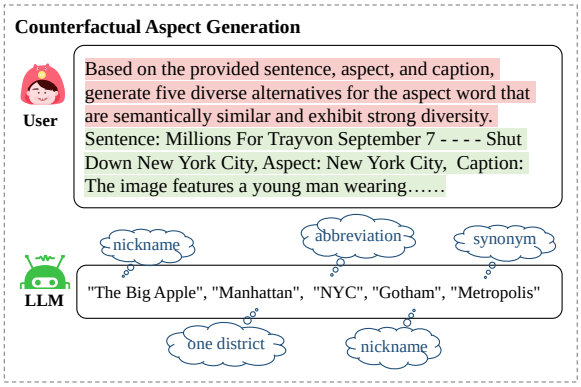

Figure 5: A counterfactual aspect generation example of our DPCI on Twitter-15 dataset.

4 Methodology

4.1 Overview

Figure 4 shows an overview of our DPCI framework. It consists of three stages. In the dual counterfactual generation stage, we first convert the image into text descriptions, and then build two types of prompt templates to feed into a large language model, particularly GPT-4o, to generate two types of counterfactual samples: aspect-based and context-based samples. In the counterfactual sample selection stage, we design two sample selection mechanisms based on aspect words and context to obtain high-quality candidate samples. In the integration-augmentation fusion stage, we integrate and enhance the two types of counterfactual candidates based on the original samples, and perform multi-decision fusion on the model inference results to obtain the final sentiment prediction.

4.2 Dual Counterfactual Generation

To mitigate the spurious correlation between the input features and output sentiment labels in the original dataset, we design dual counterfactual generation to generate counterfactual samples for the original training samples. We first input images into LLaVA based on sentence and aspect composition prompts to generate corresponding image descriptions. Afterwards, we design dual-path counterfactual generation.

Counterfactual Aspect Generation. To alleviate the spurious correlation between aspect terms

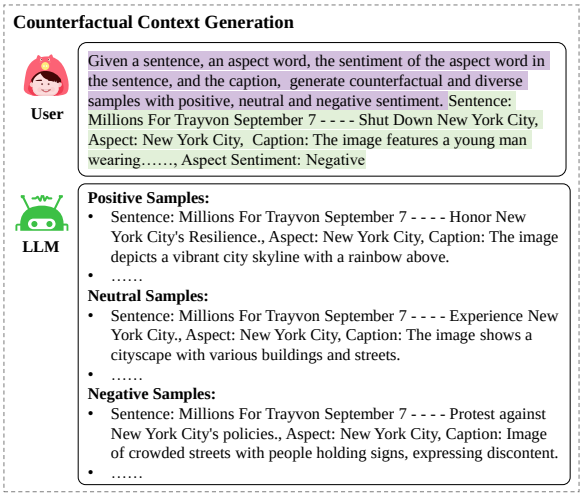

Figure 6: A counterfactual context generation example of our DPCI on Twitter-15 dataset.

and sentiment labels in the original dataset, we use aspect counterfactual generation. By constructing prompts, we let LLM generate five semantically similar and diverse aspect replacement words while keeping the original sample sentiment unchanged, which can be expressed in Eq.(2). As shown in Figure 5, we generate five aspect replacement words (*"The Big Apple"*, a well-known nickname for New York) (*"Manhattan"*, one district of New York) (*"NYC"*, the abbreviation of New York) (*"Gotham"*, another nickname of New York) (*"Metropolis"*, the synonym of New York) for the original aspect *"NEW York City"*.

Counterfactual Context Generation. As shown in Figure 3(c), in addition to replacing aspect words to obtain counterfactual samples with the same sentiment to alleviate the spurious correlation, we can also obtain counterfactual samples of context by replacing sentences and captions, which can be expressed in Eq.(3). As shown in Figure 6, we retain the aspect words of the original sample and use LLM to generate aspect-level sentiment samples with positive, neutral, and negative sentiments.

4.3 Counterfactual Sample Selection

To filter out counterfactual samples with diversity and high quality, we first fine-tune BERT using the training set D_{train}. Specifically, for sample m_i, we can get its original sentence s_i, the description information d_i generated by the image v_i, and the aspect words a_i. We construct the form of "[CLS] s_i. d_i [SEP] a_i [SEP]" as the input of BERT, and estimate the sentiment label of the current sample by obtaining the embedding of the [CLS] position. To better adapt our aspect-based and context-based counterfactual samples, we design two counterfactual sample selection schemes.

Aspect-based Counterfactual Selection. One work (Cao et al., 2022) found that the replacement of aspect words has minimal impact on the sentiment of the overall ABSA sample. Meanwhile, when generating aspect word counterfactual samples, we require that the generated replacement words are semantically similar to the original aspect words, so as to ensure the accuracy of the sentiment of the aspect words in the counterfactual samples to a large extent. Therefore, we hope to select samples with strong diversity from these counterfactual samples to further improve the model effect. Specifically, for the set of five counterfactual aspect replacement words $A_i^* = \left\{a_{i,1}^*, a_{i,2}^*, a_{i,3}^*, a_{i,4}^*, a_{i,5}^*\right\}$, we define a minimum estimated probability threshold p. We select the sample with the smallest estimated correct probability greater than p as the expansion sample:

$$a_i^* = min\left\{a_{i,j}^* \in A_i^* \mid BERT(a_{i,j}^*) > p\right\} \quad (4)$$

where a_i^* is the selected high-quality aspect counterfactual candidate, $j \in [1, 5]$. This can ensure diversity and improve the performance of the model.

Context-based Counterfactual Selection. Unlike aspect-based counterfactual samples, the accuracy of sentiment labels for aspect words in context counterfactual samples can not be guaranteed. Directly using MLLMs to predict on MABSC datasets does not yield satisfactory results (Zhao et al., 2024). Therefore, for the set of context-based counterfactual samples C_i, we aim to select the most accurate samples possible. To achieve this, we set two thresholds, q_1 and q_2, where q_1 represents the minimum estimated probability and q_2 represents the maximum difference between the predicted probability of the current sample e and context-based counterfactual samples C_i. We select the samples with the highest predicted probability that exceeds both thresholds as our context counterfactual candidates, which can be represented as:

$$C_i^* = max\left\{c_{i,j} \in C_i | P(c_{i,j}) > q_1, P(e) - P(c_{i,j}) < q_2\right\} \quad (5)$$

where C_i^* is the selected context counterfactual candidates. This ensures that the selected samples have high prediction accuracy and meet the diversity requirements.

4.4 Integration-Augmentation Fusion

Integration-Augmentation. The aspect counterfactual candidate serves primarily as an augmentation. By being exposed to diverse and semantically similar aspect candidate, the model can more accurately capture actual sentiment information rather than rely solely on superficial aspect patterns. Besides, we integrate the contextual counterfactual candidates into the historical context of the original samples. This integration enhances the ability of our DPCI to handle complex relationships between aspects and their surrounding context.

Decision Fusion. We propose an uncertainty-based decision fusion mechanism to combine original samples with counterfactual information. Specifically, we input three types of samples into the trained LLaMA model: the original sample e, the new sample e^* formed by combining the original sample e with the history of the contextual counterfactual candidates C_i^*, and the sample e' of the aspect word counterfactual candidate a_i^* to obtain three output logits l_i^1, l_i^2, l_i^3. We first calculate the uncertainty of each sample. Specifically, we introduce an uncertainty (Taha et al., 2022), which is the normalized difference between the logits of the winner and second winner classes:

$$\varphi(x) = \frac{Max1(x) - Max2(x)}{|(Max1(x) + Max2(x))|} \quad (6)$$

If the confidence of e^* is smaller than e and e', we define it as a hard samples:

$$\varphi(e^*) < min(\varphi(e), \varphi(e')) \quad (7)$$

where $Max1$ is the largest and $Max2$ is the second largest logit value. To better handle these hard samples, we perform weighted fusion by confidence:

$$\hat{p_i} = \alpha\varphi(e) \cdot l_i^1 + \beta\varphi(e^*) \cdot l_i^2 + \gamma\varphi(e') \cdot l_i^3 \quad (8)$$

where α, β and γ are hyperparameters. A higher confidence makes the sample more influential in the final decision. This decision fusion can effectively handle samples with counterfactual information and improve the ability to predict sentiments in complex and uncertain scenarios.

5 Experiments

5.1 Datasets

We conduct experiments on two publicly available standard datasets for our MABSC task: Twitter-15 and Twitter-17, both of which are multimodal datasets from (Yu and Jiang, 2019a). The Twitter-15 dataset contains user tweets posted during 2014-2015, while the Twitter-17 dataset includes tweets from 2016-2017. The statistics for these datasets are presented in Table 1.

Table 1: Statistics on two datasets of MABSC.

Label	Twitter-15			Twitter-17		
	Train	**Dev**	**Test**	**Train**	**Dev**	**Test**
Positive	928	303	317	1508	515	493
Neutral	1883	670	607	1638	517	573
Negative	368	149	113	416	144	168
Total	3179	1122	1037	3562	1176	1234

5.2 Implementation Details

During the dual counterfactual generation stage, each original instance is augmented with five aspect-based counterfactual samples and nine context-based counterfactual samples, with three generated for each sentiment category (positive, negative, and neutral). In the counterfactual sample selection stage, we use BERT-base-uncased English version as the filter, with sample thresholds set to p=0.6, q_1=0.6 and q_2=0.1 in Twitter-15. In Twitter-17, p=0.7, q_1=0.7 and q_2=0.2. We obtain counterfactual candidate sets using the same generation and selection mechanism on both the training and test sets. In the decision fusion module, the weights of the samples are controlled with α=0.7, β=0.1 and γ=0.2 on Twitter-15. In Twitter-17, α=0.3, β=0.4 and γ=0.1. During the training phase, the LLaMA3 (8B) model is fine-tuned with LoRA for four epochs to ensure efficient adaptation.

5.3 Baseline Methods

To comprehensively evaluate our DPCI, we compare it with existing competitive methods. The image-only methods include Res-Target (Yu and Jiang, 2019b), Faster R-CNN-Aspect (Ye et al., 2022) and CLIP (Ye et al., 2022). The text-only methods include BERT (Devlin et al., 2019), BERT-Pair-QA (Sun et al., 2019) and Tk-Instruct (Wang et al., 2022). The text and image methods include MIMN (Xu et al., 2019), TomBERT (Yu and Jiang, 2019b), EF-CapTrBERT (Khan and Fu, 2021), FITE (Yang et al., 2022), ITM (Yu et al., 2022), VLP-MABSA (Ling et al., 2022), VEMP (Yang and Li, 2023), InstructBLIP (Dai et al., 2023), A^2II (Feng et al., 2024), DeepSeek-V3 (Liu et al.,

Table 2: Experimental results comparison on two multimodal datasets.

	Models	Twitter-15		Twitter-17	
		Acc	Macro-F1	Acc	Macro-F1
Image Only	Res-Target (Yu and Jiang, 2019b)	59.88	46.48	58.59	53.98
	Faster RCNN-Aspect (Ye et al., 2022)	59.98	37.71	57.94	54.71
	CLIP (Ye et al., 2022)	61.23	48.77	52.03	47.83
Text Only	BERT (Devlin et al., 2019)	74.15	68.86	68.15	65.23
	BERT-Pair-QA (Sun et al., 2019)	74.35	67.70	63.12	59.66
	Tk-Instruct (Wang et al., 2022)	77.35	71.88	71.07	69.66
Text and Image	MIMN (Xu et al., 2019)	71.84	65.69	65.88	62.99
	TomBERT (Yu and Jiang, 2019b)	77.15	71.75	70.34	68.03
	EF-CapTrBERT (Khan and Fu, 2021)	78.01	73.25	69.77	68.42
	FITE (Yang et al., 2022)	78.49	73.90	70.90	68.70
	ITM (Yu et al., 2022)	78.27	74.19	72.61	71.97
	VLP-MABSA (Ling et al., 2022)	78.60	73.80	73.80	71.80
	VEMP (Yang and Li, 2023)	78.88	75.09	73.01	72.42
	A^2II (Feng et al., 2024)	79.46	75.16	74.39	72.35
	InstructBLIP (Dai et al., 2023)	57.57	59.63	60.37	35.96
	DeepSeek-V3 (Liu et al., 2024a)	62.49	62.28	63.29	61.83
	LLaMA (Touvron et al., 2023)	78.30	74.10	73.58	73.44
	LLaVA-v1.5 (Liu et al., 2024b)	77.90	74.30	74.60	74.30
	BERT+DPCI	77.63	73.53	72.04	70.59
	LLaMA+DPCI (ours)	**80.42**	**76.39**	**75.20**	**74.73**

2024a), LLaMA(8B) (Touvron et al., 2023) and LLaVA-v1.5(13B) (Liu et al., 2024b).

5.4 Comparison Results

We use accuracy and macro-F1 as evaluation metrics. We compare our method with the advanced MABSC methods and the results are shown in Table 2. It can be seen that our DPCI achieves the best performance on Twitter-15 and Twitter-17. Compared to the competitive baselines VEMP and A^2II, where both designed instruction prompts for instruction-tuning and utilize the rich knowledge of LLMs, our DPCI still achieves better performance. This is due to the fact that we generate counterfactual augmentation information and incorporate it into model learning through integration and augmentation. Besides, we add our method to the relatively small model BERT and see significant improvements, demonstrating the effectiveness and universality of our method. Our method can reduce the false correlation in the aspect sentiment prediction process and help the model learn the real sentiment information.

5.5 Ablation Study

To better understand the role of each module, we conduct ablation study and the results are shown in Table 3. "only LLaMA" indicates the result of fine-tuning the original dataset using only LLaMA3 (8B), where the same image descriptions are used as the image information. "w/o AG" and "w/o CG" represent the generation of aspect counterfactual samples removed and context counterfactual samples removed respectively. The results show a decrease in model performance when only one type of counterfactual data is generated. The performance drop is more significant when the context counterfactual samples are removed. “w/o fusion-original”, “w/o fusion-aspect” and “w/o fusion-context” represent the effects of the model on original samples, augmented aspect counterfactual candidate and integrated context counterfactual candidates respectively, without decision fusion in the integration-augmentation fusion stage. The ablation results confirm the effectiveness of each proposed part.

5.6 Analysis of Sample Selection

The selection of counterfactual samples plays a critical role in the performance. To further investigate its impact, we conduct a detailed analysis on the effects of q_1 and q_2 (as defined in Eq.(3) in the context counterfactual sample selection. The results, as shown in the Figure 7, illustrate the per-

Table 3: Ablation experiments of each module of DPCI.

Models	Twitter-15		Twitter-17	
	Acc	F1	Acc	F1
only LLaMA	78.30	74.10	73.58	73.44
w/o AG	80.33	76.20	74.96	74.54
w/o CG	79.94	75.97	74.72	74.37
w/o fusion-original	79.75	75.76	74.47	74.42
w/o fusion-aspect	78.01	73.60	73.50	73.07
w/o fusion-context	79.84	75.31	74.15	73.98
our DPCI	**80.42**	**76.39**	**75.20**	**74.73**

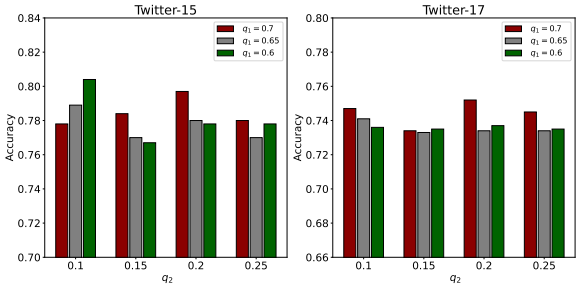

Figure 7: Effect of different q_1 and q_2 on two datasets.

formance of our DPCI under various combinations of q_1 and q_2 for both the Twitter-15 and Twitter-17. From the analysis, we observe that the performance fluctuates to some extent depending on the selected threshold values for q_1 and q_2. Specifically, in Twitter-15 dataset, the best performance is achieved when q_1=0.6 and q_2=0.1, indicating that this combination of thresholds enables the model to better capture high-quality context counterfactual samples. For Twitter-17 dataset, the optimal combination occurs when q_1= 0.7 and q_2=0.2. This suggests that a slightly higher threshold for q_1 and a more moderate value for q_2 lead to the best model performance in this context.

In addition, we conduct experiments with different numbers of augmented samples, which can be shown in Table 4. Specifically, on the Twitter-15 dataset, the model attain the highest accuracy of 80.42% when 6290 augmented samples (total 9469 samples). Similarly, on the Twitter-17 dataset, the accuracy peak at 75.20% with 6969 augmented samples (total 10531 samples). These results suggest that introducing an appropriate amount of augmentation can enhance model robustness and generalization, whereas excessive augmentation may introduce redundant or noisy information that offsets these benefits. Hence, striking a balance between the quantity and quality of augmented data is crucial for achieving optimal performance.

Table 4: The performance of DPCI with different numbers of augmented samples.

Dataset	Augmented	Accuracy (%)
Twitter-15	1258	79.75
	3145	80.13
	4403	80.23
	6290	**80.42**
	7548	80.33
Twitter-17	1394	74.72
	3485	74.95
	4878	75.04
	6969	**75.20**
	8363	75.12

5.7 Case Study

To better analyze the effectiveness of our DPCI, we further conduct case analysis as shown in Figure 8. TomBERT is a representative relatively small model, and LLaMA is a fine-tuned model with the same parameter scale as our DPCI to ensure fair comparison. Neg, Neu and Pos represent negative, neutral and positive sentiments respectively. The aspect in the sentence have been bolded. "Label" represents the true sentiment of aspect. We can see that aspect words such as "Justin Bieber" and "Donald Trump" (which have been counted in Figure 2), TomBERT and LLaMA models are easily affected by statistical bias and lack sufficient counterfactual modeling capabilities, resulting in sentiment prediction errors. In contrast, our DPCI effectively alleviates the spurious correlation through effective counterfactual learning, thereby more accurately identifying their sentiment polarities.

5.8 Evaluation under Imbalanced Data

We further evaluate the performance of the DPCI model under imbalanced data distribution followed by (Chang et al., 2024). As shown in Table 5, we conduct experiments on Twitter-15 and Twitter-17 to compare the performance of BERT, LLaMA and our DPCI on different sentiment labels. On the Twitter-15 dataset, the DPCI model achieves performance improvements over BERT and LLaMA in all three categories, especially in the Positive and Negative categories, reaching 75.39% and 69.03%, respectively, which are 6.0% and 1.77% higher than the LLaMA model. On the Twitter-17 dataset, the DPCI model also shows excellent performance

Image				
Text	RT @ HitDaBoogieZ : [***Nigga***] said I look like I bark at people .	RT @ RandiLawson : [***Oscar***] fact : [***JK Simmons***] is the voice of the yellow m amp m . [***MIND BLOWN***]	Meta Theory : [***Donald Trump***] will win the election WITH THE POWER OF THE CHAOS EMERALDS ? !	Tell me why the guy who sings [***Despacito***] with [***Justin Bieber***] looks like [***Rufus Humphrey***] 😂
Label	(1-**Neu**)	(1-**Neu** , 2-**Pos** , 3-**Neu**)	(1-**Neu**)	(1-**Neu**, 2-**Neu**, 3-**Pos**)
TomBERT	(1-**Neg**$_\times$)	(1-**Pos**$_\times$, 2-**Pos**$_\surd$, 3-**Pos**$_\times$)	(1-**Neg**$_\times$)	(1-**Neu**$_\surd$, 2-**Pos**$_\times$, 3-**Neu**$_\times$)
LLaMA	(1-**Neg**$_\times$)	(1-**Pos**$_\times$, 2-**Pos**$_\surd$, 3-**Pos**$_\times$)	(1-**Neg**$_\times$)	(1-**Neu**$_\surd$, 2-**Pos**$_\times$, 3-**Neu**$_\times$)
DPCI (ours)	(1-**Neu**$_\surd$)	(1-**Neu**$_\surd$, 2-**Pos**$_\surd$, 3-**Pos**$_\times$)	(1-**Neu**$_\surd$)	(1-**Neu**$_\surd$, 2-**Neu**$_\surd$, 3-**Neu**$_\times$)

Figure 8: Case studies of our DPCI and other baselines.

Table 5: The performance of the MABSC models on two datasets with imbalanced data distributions.

Models	Positive	Neutral	Negative
Twitter-15			
BERT	70.98	83.20	66.37
LLaMA	69.40	85.01	67.26
our DPCI	**75.39**	**85.17**	**69.03**
Twitter-17			
BERT	70.99	74.69	66.07
LLaMA	70.39	78.71	65.48
our DPCI	**72.82**	**79.41**	**67.86**

in all categories. These results verify the effectiveness of DPCI under imbalanced data distribution. By reducing the spurious correlation, DPCI not only outperforms on majority classes (positive and neutral), but also has more obvious improvements on the minority class (negative), proving the robustness and generalization ability of our DPCI in scenarios with imbalanced data distribution.

6 Conclusion

In this paper, we present DPCI, a dual-path counterfactual integration framework that introduces counterfactual reasoning into MABSC. Our DPCI introduces dual-path counterfactual generation with sample-aware selection, and integrates the resulting samples through confidence-guided fusion to enhance its robustness. Experimental results show its effectiveness. In future work, we aim to refine counterfactual generation and selection with stronger semantic constraints and causal priors, further enhancing the interpretability and generalization of our framework.

Limitations

While the proposed approach shows promising results in MABSC through counterfactual learning, it still has some limitations:

- Dependence on quality of counterfactual samples: The effectiveness of the model heavily relies on the generation of high-quality counterfactual samples. Although aspect and context counterfactual generation are designed to capture diverse expressions, their generation process can still be prone to inaccuracies or inconsistencies, especially when the model's underlying knowledge is limited.
- Limited dataset diversity: Our experiments are conducted on two publicly available datasets (Twitter-15 and Twitter-17). While widely used, they may not fully capture the diversity of real-world sentiment expressions across different domains. Expanding the evaluation to include more diverse and complex datasets would provide a more comprehensive assessment of the generalization capabilities of models.

Acknowledgements

This research is supported by the National Key R&D Program of China (No.2023YFC3303800).

References

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic

editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698.

Jiahao Cao, Rui Liu, Huailiang Peng, Lei Jiang, and Xu Bai. 2022. Aspect is not you need: No-aspect differential sentiment framework for aspect-based sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1599–1609.

Mingshan Chang, Min Yang, Qingshan Jiang, and Ruifeng Xu. 2024. Counterfactual-enhanced information bottleneck for aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17736–17744.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Cian Eastwood, Shashank Singh, Andrei L Nicolicioiu, Marin Vlastelica Pogančić, Julius von Kügelgen, and Bernhard Schölkopf. 2023. Spuriosity didn't kill the classifier: Using invariant predictions to harness spurious features. *Advances in Neural Information Processing Systems*, 36:18291–18324.

Junjia Feng, Mingqian Lin, Lin Shang, and Xiaoying Gao. 2024. Autonomous aspect-image instruction a2ii: Q-former guided multimodal sentiment classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1996–2005.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131.

Zaid Khan and Yun Fu. 2021. Exploiting bert for multimodal target sentiment classification through input space translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3034–3042.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. 2024. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12458–12468.

Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2149–2159.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Rui Liu, Jiahao Cao, Lei Jiang, Chaodong Tong, Haimei Qin, and Yanan Cao. 2025. Bottleneck-constrained contrastive decoupled network for multimodal aspect-based sentiment classification. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Rui Liu, Jiahao Cao, Nannan Sun, and Lei Jiang. 2022. Aspect feature distillation and enhancement network for aspect-based sentiment analysis. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1577–1587.

Jihong Ouyang, Zhiyao Yang, Silong Liang, Bing Wang, Yimeng Wang, and Ximing Li. 2024. Aspect-based sentiment analysis with explicit sentiment augmentations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18842–18850.

Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. 2020. Robustness to spurious correlations via human annotations. In *International Conference on Machine Learning*, pages 9109–9119. PMLR.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 380–385.

Abdel Aziz Taha, Leonhard Hennig, and Petr Knoth. 2022. Confidence estimation of classification based on the distribution of the neural network output layer. *arXiv preprint arXiv:2210.07745*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Quoc-Tuan Truong and Hady W Lauw. 2019. Vistanet: Visual aspect attention network for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 305–312.

Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. 2021. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100.

Wenbin Wang, Liang Ding, Li Shen, Yong Luo, Han Hu, and Dacheng Tao. 2024. Wisdom: Improving multimodal sentiment analysis by fusing contextual world knowledge. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2282–2291.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109.

Shirley Wu, Mert Yuksekgonul, Linjun Zhang, and James Zou. 2023. Discover and cure: Concept-aware mitigation of spurious correlation. In *International Conference on Machine Learning*, pages 37765–37786. PMLR.

Nan Xu, Wenji Mao, and Guandan Chen. 2018. A co-memory network for multimodal sentiment analysis. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 929–932.

Nan Xu, Wenji Mao, and Guandan Chen. 2019. Multi-interactive memory network for aspect based multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 371–378.

Zhi Xu, Dingkang Yang, Mingcheng Li, Yuzheng Wang, Zhaoyu Chen, Jiawei Chen, Jinjie Wei, and Lihua Zhang. 2025. Debiased multimodal understanding for human language sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14450–14458.

Bin Yang and Jinlong Li. 2023. Visual elements mining as prompts for instruction learning for target-oriented multimodal sentiment classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6062–6075.

Dingkang Yang, Mingcheng Li, Dongling Xiao, Yang Liu, Kun Yang, Zhaoyu Chen, Yuzheng Wang, Peng Zhai, Ke Li, and Lihua Zhang. 2024. Towards multimodal sentiment analysis debiasing via bias purification. In *European Conference on Computer Vision*, pages 464–481. Springer.

Hao Yang, Yanyan Zhao, and Bing Qin. 2022. Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3324–3335.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7617–7630.

Junjie Ye, Jie Zhou, Junfeng Tian, Rui Wang, Jingyi Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022. Sentiment-aware multimodal pre-training for multimodal sentiment analysis. *Knowledge-Based Systems*, 258:110021.

Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. 2024. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*.

Jianfei Yu and Jing Jiang. 2019a. Adapting bert for target-oriented multimodal sentiment classification. IJCAI.

Jianfei Yu and Jing Jiang. 2019b. Adapting bert for target-oriented multimodal sentiment classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 5408–5414.

Jianfei Yu, Jieming Wang, Rui Xia, and Junjie Li. 2022. Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching. In *IJCAI*, pages 4482–4488.

Tianyu Zhao, Ling-ang Meng, and Dawei Song. 2024. Multimodal aspect-based sentiment analysis: a survey of tasks, methods, challenges and future directions. *Information Fusion*, 112:102552.