

DaNet: Dual-Aware Enhanced Alignment Network for Multimodal Aspect-Based Sentiment Analysis

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, Ning An
School of Computer Science and Information Engineering, Hefei University of Technology
zhuaqiang@mail.hfut.edu.cn, {jsjxhumin, xh_wang, jiaoyun, ymtang}@hfut.edu.cn, ning.g.an@acm.org

Abstract

Multimodal Aspect-Based Sentiment Analysis (MABSA) aims to extract aspect-sentiment pairs from text and image data. While significant progress has been made in aspect-image alignment, due to the subtlety and complexity of language expressions, there are not always explicit aspect words in the language to align with images. Existing methods typically assume a direct alignment between images and aspects, matching the entire image with a corresponding aspect. This rough alignment of images and aspects introduces noise. To address the above issues, this paper proposes a Dual-Aware Enhanced Alignment Network (DaNet) designed for fine-grained multimodal aspect-image alignment and denoising. Specifically, we first introduce a Multimodal Denoising Encoder (MDE) that jointly image and text to guide the compression and denoising of visual sequences. And then, aspect-aware and sentiment-aware networks are constructed to jointly enhance fine-grained alignment and denoising of text-image information. To better align implicit aspects, an Implicit Aspect Opinion Generation (IAOG) pretraining is designed under the guidance of large language model. Extensive experiments across three MABSA subtasks demonstrate that DaNet outperforms existing methods.

1 Introduction

Multimodal Aspect-Based Sentiment Analysis (MABSA) aims to jointly extract aspect terms from text-image pairs and predict their sentiment polarity (Zhao et al., 2024). As one of the significant and complex tasks in sentiment analysis, MABSA has received increasing attention in recent years (Zhang et al., 2023; Ghorbanali and Sohrabi, 2023). Depending on the task, MABSA typically also contains two subtasks: Multimodal Aspect Term Extraction (MATE) and Multimodal Aspect-oriented Sentiment Classification (MASC). MATE focuses

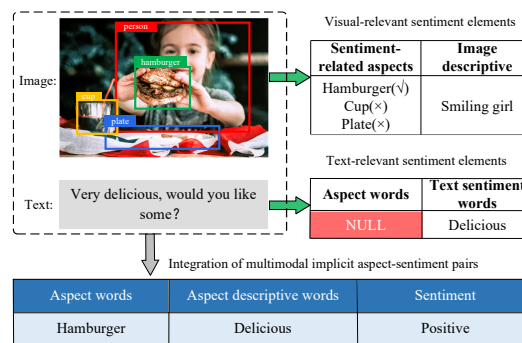


Figure 1: An example of an implicit aspect in MABSA. The implicit aspect means that the aspect corresponding to the sentiment is not explicitly in the text (not missing data).

on extracting all aspect terms (Zhao et al., 2022), while MASC aims to classify the sentiment of each given aspect term (Khan and Fu, 2021). There are significant differences in information density between image and text data. This undeniable data heterogeneity and semantic gap make the learning of fine-grained multimodal remains challenging.

Previous research on MABSA primarily focused on exploring effective alignment between image and text aspects (Zhao et al., 2024). Addressing the alignment issue, some methods (Ling et al., 2022; Yang et al., 2023; Yu et al., 2023) use trained object detectors to identify potential objects in images, achieving alignment between text words and visual objects. However, these approaches are limited by the categories and performance of the object detectors, and the identified potential objects may be unrelated to the text description. Other approaches, based on visual coders such as ViT (Dosovitskiy et al., 2021) or CLIP (Radford et al., 2021), focus on aligning images and text globally semantically (Peng et al., 2024; Wang et al., 2024a; Zhu et al., 2024b). While overcoming the limitations of the target detector, these methods associate entire image with the text, neglecting that different

regions of the image may correspond to different aspects of the text. Additionally, the sentiment of specific aspects may be influenced by other aspects, and global coarse-grained alignment introduces noise (Bao et al., 2022). Moreover, due to the subtlety and complexity of language expression, aspect terms may be omitted. For example, as shown in Fig. 1, the term "hamburger" is not explicitly mentioned in the language but is simplified and implied. Existing methods typically assume a clear alignment between images and aspects, making it difficult to address such implicit alignment issues.

To resolve the above issues, two key challenges need to be addressed: 1)How to efficiently focus on regions in text-image pairs that are relevant to opinions and aspects, and filter out irrelevant regions to reduce noise? 2)How to recognize implicit aspects (e.g., "hamburger") from multimodal information and align them with textual opinions (e.g., "delicious")?

Based on the above thinking, this paper proposes a Dual-Aware Enhanced Alignment Network (DaNet) designed for fine-grained multimodal aspect-image alignment and denoising. The motivation is that feature regions in images and text that are related to aspect and sentiment should receive higher attention, while irrelevant regions should receive less attention. Specifically, we first introduce a Multimodal Denoising Encoder(MDE) that segments the global visual input into strongly correlated and weakly correlated regions. Guided by both image and text, the weakly correlated regions are then injected into the corresponding strongly correlated regions, reducing interference from unrelated areas while enhancing the representation of the strongly correlated regions. And then, this paper enhances fine-grained alignment in the text-image pairs from an aspect perspective, filtering out irrelevant regions to reduce noise. However, aspect-to-aspect alignment lacks the interaction of sentiment semantics. Additionally, relying solely on aspects does not achieve alignment of implicit aspects related to opinions. Therefore, we further propose to perceive the alignment of emotions and aspects from the perspective of emotional semantics, aiming to achieve implicit aspect alignment and reduce interference between different aspects through the understanding of sentiment-semantic relations of opinion words. To better align implicit aspects, we propose a pre-training task for Implicit Aspect Opinion Generation (IAOG) guided by large language models. Additionally, introduc-

ing Visual Aspect Opinion Generation (VAOG) pre-training task enhances the model's ability to learn the semantic relationships between common aspect-sentiment pairs.

In summary, our contributions are as follows:

- This paper proposes a Dual-Aware Enhanced Alignment Network (DaNet), which reduces the introduction of noisy regions and jointly aligns the cross-modal information of text and images from both aspect and affective semantic perspectives, reducing mutual interference between different aspects.
- This paper proposes a Multimodal Denoising Encoder (MDE), which enhances the representation of strongly correlated regions by injecting weakly correlated regions of the task, thereby reducing interference from irrelevant regions.
- This paper proposes two specific pre-training tasks which combine multimodal aspects (explicit and implicit) with opinion alignment and awareness, aiming to capture common sentiment patterns and aspect-sentiment semantic within text-image information.

2 Related Work

Early MABSA studies usually straightforwardly combined MATE and MASC subtasks or adapted MATE to perform MABSA tasks (Ling et al., 2022). UMT-collapse (Yu et al., 2020b), OSCGA-collapse (Wu et al., 2020b), and RpBERT-Collapse (Yu and Jiang, 2019) were adapted from models for MATE by using collapsed labels to represent aspect and sentiment pairs. These simple combination methods ignore the alignment and interaction between the semantic information and sentiment of the two subtasks, and the results are usually not satisfactory. Recently, a growing number of research efforts have been devoted to the alignment of joint images and corresponding aspects. JML (Ju et al., 2021) first proposed jointly extracting aspects and classifying the corresponding sentiments to better meet practical applications. VLP-MABSA (Ling et al., 2022) and CMMT (Wu et al., 2022) design multiple visual-language pre-training tasks to achieve a unified multimodal architecture for specific tasks. Although specific pre-training helps improve model performance, the lack of alignment between aspects and emotions hinders further performance enhancement. DPFN (Wang et al.,

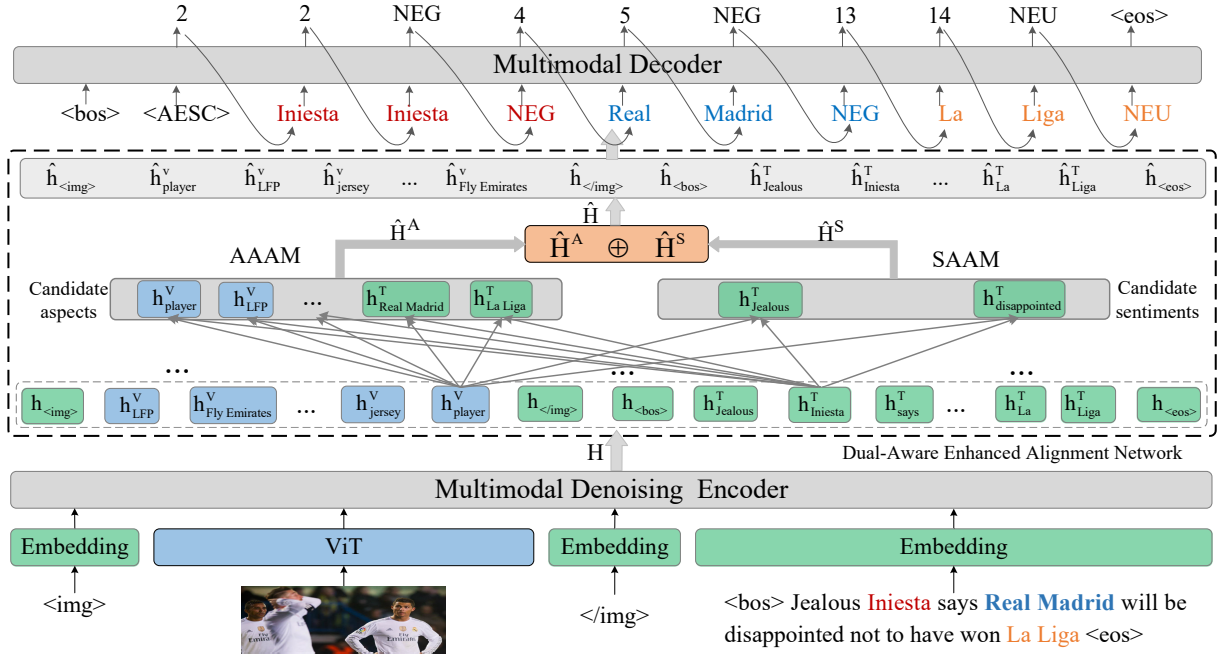


Figure 2: The overall framework of DaNet. The depicted multimodal decoder is designed for the MABSA task, and it differs in its structure from the decoders used in the MATE and MASC tasks.

2024b) and Atlantis (Xiao et al., 2024) achieve multi-modal fusion from different granularity features, but do not address the mutual interference between aspects and between sentiments. Further research has focused on more fine-grained alignment, TMFN (Wang et al., 2024a) and AoM (Zhou et al., 2023) focus on the alignment of text tokens and image regions to reduce the interference of irrelevant regions, but overlook the complementarity of sentiment and aspect alignment. Recent studies have further improved the performance of MABSA through specific prompt (Peng et al., 2024) and aspect-enhanced pre-training (Zhu et al., 2024b). Although these methods have made great progress on the MABSA task, these methods associate entire image with the text, neglecting that different regions of the image may correspond to different aspects of the text. Additionally, the sentiment of specific aspects may be influenced by other aspects, and global coarse-grained alignment introduces noise.

3 Methodology

In Fig. 2, DaNet mainly consists of a multimodal denoising encoder, a dual-aware enhanced alignment network, and multimodal decoder. In addition, two specific pretraining tasks (i.e., IAOG and VAOG, as shown in Fig. 4) were introduced to further enhance DaNet’s ability to learn joint multi-

modal representation and the relationships between aspect-sentiment in multimodal data.

3.1 Multimodal Denoising Encoder

We employ ViT (Dosovitskiy et al., 2021) to extract image representations. To be consistent with the text representation, we adopt a linear transformation layer to project the image features F^I to d -dimensional vectors.

$$V = \text{Reshape}(W_I \text{ViT}(F^I) + b_i), \quad (1)$$

where V denotes the projected image features, $W_I \in \mathbb{R}^{d_v \times r \times d}$ and r denotes the number of visual blocks. We consider every feature of a visual block as an atomic feature. We use the BART (Lewis et al., 2020) to obtain the embedding of the text T :

$$E_T = \text{Embedding}(T), E_T \in \mathbb{R}^{l_t \times d}, \quad (2)$$

where l_t denotes the length of the text sequence. The multimodal embedding can be obtained:

$$E_M = [E_{\langle \text{img} \rangle}, V, E_{\langle / \text{img} \rangle}, E_{\langle \text{bos} \rangle}, E_T, E_{\langle \text{eos} \rangle}], \quad (3)$$

where $E_M \in \mathbb{R}^{l_m \times d}$, $l_m = r + l_t + l_s$ is the length of multimodal sequence and l_s is the length of special tokens. As shown in Fig. 2, following Liu et al. (2021), we use $\langle \text{img} \rangle$ and $\langle / \text{img} \rangle$ to mark the beginning and end of visual features, and $\langle \text{bos} \rangle$ and $\langle \text{eos} \rangle$ for text boundaries.

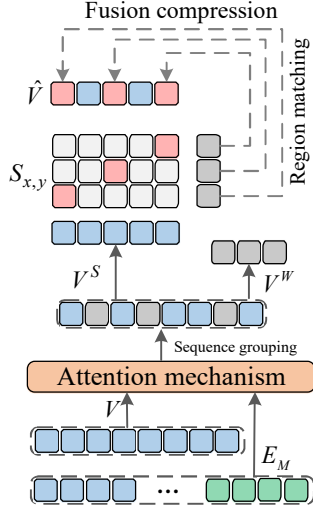


Figure 3: Joint image and text guidance for visual compression and denoising.

In data, not all the image content has the matching text description. It is necessary to prune the visual sequence to reduce the redundant information unrelated to the text. Therefore, we propose an image-text jointly guided visual sequence compression and noise reduction method, as shown in Fig. 3. We take E_M as the Q (query) and V as the K (key) to obtain the attention weights.

$$a = \text{softmax}\left(\frac{E_M \cdot V^T}{\sqrt{d}}\right), \quad (4)$$

where a is the attention distribution of E_M to V . Based on the attention weights and the specified retention ratio α (Appendix C provides a detailed analysis of the α), the visual regions are classified into strongly correlated regions V^S and weakly correlated regions V^W . We consider that neighboring visual regions express similar features to some extent, so for each visual marker in a weakly correlated region v_x , a nearest-neighbor visual region found in a strongly correlated region v_y corresponds to it. By cosine similarity, the nearest neighbor region is calculated as:

$$S_{x,y} = \frac{v_x^T v_y}{\|v_x\| \|v_y\|}, \quad (5)$$

$$v_{near} = \arg \max_{v_y \in V^S} S_{x,y}, \quad (6)$$

where v_{near} denotes the nearest neighbor region of weakly correlated region v_x . Then, the fusion weights of v_x and v_{near} are computed, and v_x is weighted and merged into v_{near} to obtain the up-

dated v_{new} .

$$\theta_x = \frac{\exp(S_{x,near})}{\exp(S_{x,near}) + e}, \quad (7)$$

$$v_{new} = (1 - \theta_x)v_{near} + \theta_x v_x, \quad (8)$$

where θ_x is the fusion weight of the weakly correlated region v_x , e is Euler's number. Multimodal representation E_M is reconstructed using the weighted fused visual sequence $\hat{V} \in \mathbb{R}^{\alpha r \times d}$ to obtain the noise-reduced representation \hat{E}_M :

$$\hat{E}_M = [E_{}, \hat{V}, E_{}, E_{<bos>}, E_T, E_{<eos>}], \quad (9)$$

where $\hat{E}_M \in \mathbb{R}^{\alpha r + l_t + l_s}$. We feed \hat{E}_M into the Multimodal Encoder to obtain the multimodal representation.

3.2 Dual-Aware Enhanced Alignment Network

To achieve fine-grained cross-modal text-image aspect alignment and sentiment-aspect semantic alignment, we propose a dual-aware enhanced alignment network. As shown in Fig. 2.

1) Aspect-Aware Alignment Module (AAAM):

Given a text-image pair, we first use the NLP tool SpaCy¹ to extract noun phrases from the sentence and YOLOv8² to identify target entities in the image. After merging and deduplicating the extracted noun phrases and target entities, we obtain k candidate aspects (CA). Then, the hidden layer states $H^{CA} = \{h_1^{CA}, \dots, h_t^{CA}, \dots, h_k^{CA}\}$, $H^{CA} \in \mathbb{R}^{k \times d}$ corresponding to the candidate aspects are obtained through the multimodal encoder. Given the t -th hidden feature h_t of the multimodal hidden state H as the query, the attention distribution $\alpha_t^A \in \mathbb{R}^{l_m \times k}$ guided by the candidate aspects as keys can be represented as:

$$Z_t(h_t, H^{CA}) = \tanh(\text{cat}(W_H h_t + b_H, W_{CA} H^{CA} + b_{CA}; \text{dim} = -1)), \quad (10)$$

$$\alpha_t^A(h_t, H^{CA}) = \text{softmax}(W_\alpha Z_t + b_\alpha), \quad (11)$$

where $Z_t \in \mathbb{R}^{\alpha r \times k \times 2d}$ represents the extracted composite features, $W_H \in \mathbb{R}^{d \times d}$, $W_{CA} \in \mathbb{R}^{d \times d}$, $W_\alpha \in \mathbb{R}^{1 \times 2d}$, $b_H \in \mathbb{R}^d$, $b_{CA} \in \mathbb{R}^d$, and $b_\alpha \in \mathbb{R}$ are learnable parameters. Then, by calculating the weighted sum of α_t^A and all candidate aspects, we obtain the aspect-related hidden feature h_t^A .

$$h_t^A = \sum_i^k \alpha_t^A \cdot h_i^{CA} \quad (12)$$

¹<https://spacy.io>

²<https://yolov8.com>

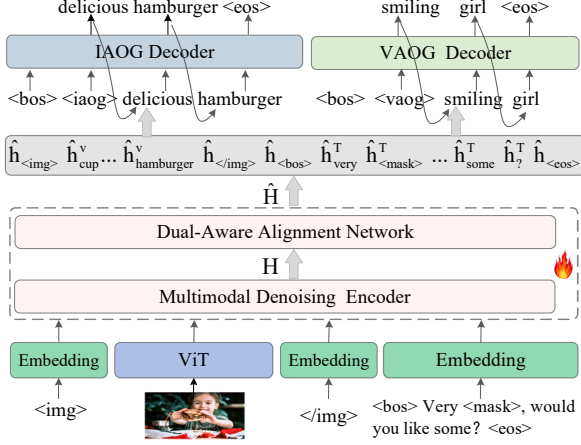


Figure 4: The framework of the pre-training tasks.

Since not all image regions are aspect-related, we introduce a weight factor β_t to learn the additive weight of the original hidden feature h_t and the aspect-related hidden feature h_t^A , as detailed below:

$$\beta_t = \text{sigmoid}(W_\beta \text{cat}(W_1 h_t, W_2 h_t^A) + b_\beta), \quad (13)$$

$$\hat{h}_t^A = \beta_t h_t + (1 - \beta_t) h_t^A, \quad (14)$$

where $W_\beta \in \mathbb{R}^{1 \times 2d}$, $W_1 \in \mathbb{R}^{d \times d}$, $W_2 \in \mathbb{R}^{d \times d}$, and $b_\beta \in \mathbb{R}$ are learnable parameters of the fully connected layers. $\hat{h}_t^A \in \hat{H}^A$ is the final output of AAAM.

2) Sentiment-Aware Alignment Module (SAAM): Firstly, according to the VADER sentiment dictionary (Hutto and Gilbert, 2014), the words with sentiment significance in the text sequence are searched as candidate sentiments (CS). Then, the hidden states of these candidate sentiments are obtained through the multimodal encoder, represented as $H^{CS} \in \mathbb{R}^{w \times d}$, where w denotes the number of candidate sentiment words in the sequence. As shown in Fig. 2, SAAM has a similar network structure to AAAM. By comparing the final output Eq. (14) of AAAM, the final output \hat{h}_t^S of SAAM after sentiment-aware alignment can be obtained:

$$\hat{h}_t^S = \delta_t h_t + (1 - \delta_t) h_t^S \quad (15)$$

Similarly, δ_t is the weight factor, and $h_t^S = \sum_i^e \alpha_t^S \cdot h_i^{CS}$ represents the hidden features obtained by weighting the sentiment-guided attention distribution $\alpha_t^S \in \mathbb{R}^{l_m \times e}$ and all candidate sentiments H^{CS} . Unlike AAAM, which focuses on the alignment between aspects, SAAM emphasizes the semantic relationship alignment between sentiments and aspects.

Finally, we add the \hat{h}_t^A and \hat{h}_t^S to obtain the final dual-aware enhanced alignment hidden layer feature $\hat{h}_t \in \hat{H}$.

$$\hat{h}_t = \hat{h}_t^A + \hat{h}_t^S \quad (16)$$

3.3 Multimodal Decoder

The multimodal decoder takes \hat{H} and the previous decoder output $y_{<t}$ as inputs to generate y_t . We take the MABSA task as an example, as Fig. 2 shows.

$$h_t^d = \text{Decoder}(\hat{H}, y_{<t}), \quad (17)$$

where t is the t_{th} step. We predict the token probability distribution P_t with h_t^d , as follows:

$$P_{y_t} = \text{softmax}([E_T; E_S] h_t^d), \quad (18)$$

where E_S is the embedding of the sentiment label set. The loss function is as follows:

$$\mathcal{L} = - \sum_{t=1}^N \log P(y_t | y_{<t}, \hat{H}), \quad (19)$$

where N is the length of the target sequence.

3.4 Pre-training Tasks

To further enhance DaNet’s ability to learn joint multimodal representation and the relationships between aspect-sentiment, we designed two specific pre-training tasks on MVSA-Multi (Niu et al., 2016), as shown in Fig 4.

1) Implicit Aspect-Opinion Generation (IAOG): First, using the method mentioned in Section 3.2, we obtain k candidate aspects (CA) and an sentiment word. Next, we create the following prompt: “Please choose a word from ‘candidate aspects’ that is most suitable to be modified by ‘sentiment word’, and output the corresponding sentiment word and aspect. If none, output ‘none’.” as the input for ChatGPT-3.5 (Brown et al., 2020), leveraging ChatGPT to generate the matched sentiment-aspect pairs.

The opinion-aspect pairs generated by ChatGPT-3.5 serve as the target supervision for $G^I = \{g_1, \dots, g_t, \dots, g_{|G^I|}\}$. Then, IAOG Decoder takes the multimodal encoder output \hat{H} and the previous decoder $G_{<t}^I$ as inputs to predict the probability distribution $P(g_t)$ of the token.

$$h_t^d = \text{IAOGD}(\hat{H}; G_{<t}^I), \quad (20)$$

$$P(g_t) = \text{Softmax}(E^T h_t^d), \quad (21)$$

Methods	Venues	Twitter-2015			Twitter-2017		
		P	R	F1	P	R	F1
UMT+TomBERT*	ACL 2020; IJCAI 2019	58.4	61.3	59.8	62.3	62.4	62.4
OSCGA+TomBERT*	MM 2020; IJCAI 2019	61.7	63.4	62.5	63.4	64.0	63.7
OSCGA-collapse*	MM 2020	63.1	63.7	63.2	63.5	63.5	63.5
UMT-collapse*	ACL 2020	61.0	60.4	61.6	60.8	60.0	61.7
RpBERT-collapse*	AAAI 2021	49.3	46.9	48.0	57.0	55.4	56.2
JML*	EMNLP 2021	65.0	63.2	64.1	66.5	66.5	66.0
VLP-MABSA*	ACL 2022	65.1	68.3	66.6	66.9	69.2	68.0
CMMT*	IPM 2022	64.6	68.7	66.5	67.6	69.4	68.5
AoM*	ACL 2023	67.9	69.3	68.6	68.4	71.0	69.7
AESAL*	IJCAI 2024	68.7	70.4	69.5	69.4	74.8	72.0
DaNet(ours)	-	70.8	71.5	71.2	71.3	72.9	72.1

Table 1: Results of different methods for MABSA. "*" denotes the results from AESAL (Zhu et al., 2024b).

where $h_t^d \in \mathbb{R}^d$ is the output of the decoder and E denotes the embedding matrix of all tokens in the vocabulary.

$$\mathcal{L}_{IAOG} = - \sum_{t=1}^{|G^I|} \log P(g_t | g_{<t}, \hat{H}) \quad (22)$$

2) Visual Aspect-Opinion Generation (VAOG):

We use the pre-trained ANP (Adject-noun Pair) detector DeepSentiBank (Chen et al., 2014) to predict the ANP in the image, and select the ANP with the highest probability as the supervision signal for the VAOG task. For example, the ANP predicted for the image shown in Fig. 1 is a "Smiling girl". Compared to IAOG, the loss function for VAOG can be expressed as:

$$\mathcal{L}_{VAOG} = - \sum_{t=1}^{|G^V|} \log P(g_t | g_{<t}, \hat{H}) \quad (23)$$

To optimize all model parameters, we adopt an alternating optimization strategy to iteratively optimize our two pre-training tasks. The objective function is as follows:

$$\mathcal{L}_{pre} = \lambda_1 \mathcal{L}_{IAOG} + \lambda_2 \mathcal{L}_{VAOG}, \quad (24)$$

where λ_1 and λ_2 are the trade-off hyperparameters to control the contribution of each task.

4 Experiments

4.1 Experimental Settings

Datasets. Following prior studies (Ju et al., 2021; Zhou et al., 2023; Ling et al., 2022), we evaluate the performance of DaNet using two widely

used benchmark datasets: Twitter-2015 and Twitter-2017 (Yu and Jiang, 2019). The details of the datasets are shown in Appendix A.1.

Training Details. All models are built on the PyTorch (Paszke et al., 2019) with RTX A40 GPU. λ_1 and λ_2 are all set to 1. Appendix A.2 describes the details of hyper-parameter setting.

Evaluation Metrics. Following previous work (Ju et al., 2021; Zhou et al., 2023; Ling et al., 2022), we evaluate the performance of our model on the MABSA task and MATE task using F1 score (F1), Precision (P), and Recall (R). For the MASC task, we evaluate performance using Accuracy (Acc) and F1.

Baselines. Please refer to Appendix B for more details on baselines.

4.2 Quantitative Results and Analysis

We compare DaNet with state-of-the-art methods on MABSA’s three subtasks, where it achieves superior or competitive results across both datasets.

Results of MABSA: As shown in Table 1, DaNet outperforms all other methods on the MABSA task, except for slightly lower than AESAL on metric R. Specifically, although the R-metric of DaNet on Twitter-2017 is slightly lower than AESAL, all other metrics outperform AESAL. The average performance of DaNet on Twitter-2015 is 2.36% higher than that of AESAL, which further proves the effectiveness of DaNet.

Results of MATE: As shown in Table 2, in the MATE task, while DaNet significantly outperforms all other models except AESAL across all metrics, it only outperforms AESAL in R-value on Twitter-

Methods	Twitter-2015			Twitter-2017		
	P	R	F1	P	R	F1
RAN*	80.5	81.5	81.0	90.7	90.7	90.0
UMT*	77.8	81.7	79.7	86.7	86.8	86.7
OSCGA *	81.7	82.1	81.9	90.0	90.7	90.4
JML*	83.6	81.2	82.4	92.0	90.7	91.4
VLP-MABSA*	83.6	87.9	85.7	90.8	92.6	91.7
CMMT*	83.9	88.1	85.9	92.2	93.9	93.1
AoM*	84.6	87.9	86.2	91.8	92.8	92.3
AESAL*	90.2	90.6	90.4	93.1	96.4	94.7
DaNet (ours)	87.6	90.8	89.2	94.6	93.9	94.2

Table 2: Results of MATE. "*" denotes the results from AESAL (Zhu et al., 2024b).

Methods	Twitter-2015		Twitter-2017	
	Acc	F1	Acc	F1
ESAFN*	73.4	67.7	67.8	64.2
TomBERT*	77.2	71.8	70.5	68.0
CapTrBERT*	78.0	73.2	72.3	70.2
JML*	78.7	-	72.7	-
VLP-MABSA*	78.6	73.8	73.8	71.8
CMMT*	77.9	-	73.8	-
AoM*	80.2	75.9	76.4	75.0
AESAL*	80.1	75.2	78.8	75.9
DaNet (ours)	81.3	78.5	79.0	76.4

Table 3: Results of MASC. "*" denotes the results from AESAL (Zhu et al., 2024b).

2015 and P-value on Twitter-2017, and the overall performance is still lower than AESAL. The main reason is that AESAL performs supervised aspect-related pre-training on the Twitter-2015 and Twitter-2017 datasets, which helps to extract aspect information in the MATE task. In contrast, the IAOG and VAOG pre-training performed by DaNet on MVSA-Mult dataset relies on ChatGPT and DeepSentiBank for supervision labels due to the inability to obtain actual opinion-aspect pairs, which may lead to errors and affect the reliability of the results. However, considering the performance improvement of the main task MSABA through IAOG and VAOG pre-training, as well as the exploration of implicit aspect alignment, this research remains significant. We also believe that providing more precise opinion-aspect pairs will yield better results.

Results of MASC: Table 3 shows that DaNet outperforms all other models on the MASC task. Specifically, compared to the currently publicly

Methods	Twitter-2015			Twitter-2017		
	P	R	F1	P	R	F1
VisualGLM-6B*	69.2	64.6	66.8	57.2	52.0	54.5
ChatGPT-3.5*	66.3	66.3	66.3	58.9	58.9	58.9
DQPSA*	81.1	81.1	81.1	75.0	75.0	75.0
DaNet (ours)	79.1	77.9	78.5	76.8	76.0	76.4

Table 4: Results of comparison with LLMs on MASC task. "*" denotes the results from DQPSA (Peng et al., 2024).

available best model AESAL, DaNet improved the F1 scores on the Twitter-2015 and Twitter-2017 datasets by 4.39% and 0.66%, respectively. This indicates that DaNet more effectively integrates text and image information for the sentiment classification of each aspect term.

4.3 Sentiment Analysis Compared to LLMs

Considering the excellent performance of large language models (Touvron et al., 2023; Du et al., 2022) in NLP tasks, we compare the performance of DaNet with VisualGLM-6B and ChatGPT-3.5. Since LLMs are not specifically designed to recognize aspects in text, the output structure is difficult to unify. Therefore, following previous research (Peng et al., 2024), we only tested on the MASC task to ensure a fair comparison. Table 4 shows that DaNet outperforms LLMs (e.g., VisualGLM-6B and ChatGPT-3.5) with fewer parameters, validating its effectiveness. However, on the Twitter-2015 dataset (61.65% one-aspect data), DaNet underperforms compared to DQPSA, primarily owing to DQPSA’s use of prompts such as "Sentiment of 'aspect term' is [positive, neutral, negative]", which enhance one-aspect sentiment classification. In contrast, on the multi-aspect Twitter-2017 dataset (66.46%), DaNet surpasses DQPSA across all metrics. This indicates that as the number of aspect terms increases, DQPSA’s performance deteriorates due to interference between aspects, whereas DaNet’s dual-aware alignment effectively reduces such interference, maintaining stable performance.

4.4 Ablation Studies

Table 5 shows that without the various components of the DaNet model, the overall performance decreases. Further analysis reveals that the performance is most significantly affected without AAAM, with the average F1 score of the three subtasks decreasing by 3.82%. This indicates that the aspect-aware alignment of text and images in

Methods	Twitter-2015			-
	MABSA	MATE	MASC	Avg
DaNet (full)	71.2	89.2	78.5	-
w/o MDE	69.9	87.3	77.5	1.74% ↓
w/o AAAM	69.1	85.9	74.7	3.82% ↓
w/o SAAM	70.8	88.6	75.9	1.66% ↓
w/o Pretraining	69.3	88.9	77.6	1.38% ↓
w/o IAOG	70.5	89.8	77.8	0.40% ↓
w/o VAOG	70.7	88.2	78.4	0.65% ↓

Table 5: Results of ablation studies. We evaluate all tasks in terms of F1. w/o denotes without. “Avg” denotes the average decrease of performance. “↓” means decrease.

the MABSA task is crucial. As a complementary module to AAAM, SAAM also plays a significant role in the MABSA task. Furthermore, the performance drop without pre-training suggests that task-specific pre-training is beneficial for enhancing model performance. This is consistent with that pre-training language models improve model performance (Zhu et al., 2024a).

4.5 Evaluation of Implicit Aspect Influence

To analyze the impact of implicit aspect scenarios on model performance, we compared DaNet with high performing models that provide reproducible code (i.e., VLP-MABSA, CMMT, and AoM). On the Twitter-2015 dataset, we simulated implicit aspect scenarios by randomly replacing 50% of the aspect terms in the test and validation sets with “<mask>”. To further explore the influence of the IAOG, AAAM, and SAAM modules on DaNet in this scenarios, we conducted ablation analysis.

From Table 6, it is evident that even the current better performing models experience a significant performance drop in scenarios simulating implicit aspects. For example, the F1 scores of VLP-MABSA and CMMT for three tasks in implicit scenarios decreased by 42.89% and 37.67%, respectively, compared to the average in general scenarios. The main reason is that both VLP-MABSA and MMTT, as multi-task pre-trained and joint learning models, lack alignment specifically for sentiment and aspects compared to AoM and DaNet. This indicates that the ability to learn such alignment enables AoM and DaNet to maintain a certain level of performance in scenarios involving implicit aspects. DaNet further introduces the aware alignment of sentiment to aspects and the pre-training task of implicit aspect generation, which allows DaNet to

Methods	Twitter-2015			-
	MABSA	MATE	MASC	Avg
DaNet (ours)	65.1 (71.2)	87.9 (89.2)	71.6 (78.5)	6.27% ↓
VLP-MABSA	25.1 (66.5)	40.2 (85.0)	63.6 (73.7)	42.89% ↓
CMMT	29.7 (66.6)	49.6 (85.4)	61.8 (73.2)	38.3% ↓
AoM	60.4 (66.7)	80.2 (85.7)	65.9 (74.5)	9.13% ↓
w/o IAOG	64.8 (71.2)	87.1 (89.2)	68.4 (78.5)	8.07% ↓
w/o AAAM	61.4 (71.2)	80.8 (89.2)	70.5 (78.5)	11.12% ↓
w/o SAAM	64.3 (71.2)	84.5 (89.2)	69.4 (78.5)	8.85% ↓

Table 6: Results of implicit aspect analysis. We evaluate all tasks in terms of F1. The values in “()” denote the results of the general scenario in the experimental setting of this paper. “Avg” denotes the average decrease of performance. “↓” means decrease.

perform better in implicit aspect scenarios while improving overall performance. In addition, we observe that the performance of DaNet decreases significantly without SAAM and AAAM, which further demonstrates the importance of aspect and sentiment awareness and alignment.

To investigate the effectiveness of the model in detecting aspect-opinion related information, Appendix C provides the impact of image retention rates α , Appendix D provides a visualization, and Appendix E provides the case studies.

5 Conclusion

In this paper, we propose DaNet, a dual-aware enhanced alignment network designed for fine-grained multimodal aspect-image alignment and denoising. Specifically, we first introduce a Multimodal Denoising Encoder (MDE) that jointly image and text to guide the compression and denoising of visual sequences. And then, DaNet aligns aspects with image regions based on aspect awareness to filter out irrelevant regions and reduce noise. Additionally, DaNet aligns sentiment with aspects through sentiment-aware semantic relationships. This not only enhances the sentiment-semantic relationship and facilitates implicit aspect alignment between text and image, but also reduces mutual interference between different aspects. MABSA-related pre-training tasks further enhance DaNet’s ability to learn common sentiment patterns and semantic relationships in text-image information. Extensive experiments and ablation studies on the three sub-tasks of MABSA, validating the effectiveness of the DaNet and the necessity of each module.

Limitations

This paper explores the fine-grained multimodal aspect-image alignment and denoising, thereby further enhancing the performance of MABSA. However, there are still some limitations that need to be addressed in future work. Firstly, given that we do not have access to the actual opinion pairs from the pre-trained dataset MVSA-Multi (Niu et al., 2016), it is not possible to compute the exact error rates for ChatGPT and DeepSentiBank, which could potentially lead to flawed results. Secondly, in the experimental analysis of implicit aspects, we masked aspect words in text on the available dataset to simulate the implicit experimental setting. Implicit aspects are not missing data and are scenarios with higher requirements for alignment. To make the dataset usable, we retained the position of the masked words in the sequences, which reduced the testing difficulty to some extent. In the future, we will collect more samples with implicit aspects from social media to further enhance the training of our model in realistic scenarios.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62176084, and Grant 62176083, and in part by the National Key Research and Development Program of China under Grant 2023YFC3604704.

References

- Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. [Vlmo: Unified vision-language pre-training with mixture-of-modality-experts](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 32897–32912.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Tao Chen, Damian Borth, Trevor Darrell, and Shih Fu Chang. 2014. DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks. *Computing Research Repository*.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, and Sylvain Gelly. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
- Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. [GLM: General language model pretraining with autoregressive blank infilling](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.
- Alireza Ghorbanali and Mohammad Karim Sohrabi. 2023. [A comprehensive survey on deep learning-based approaches for multimodal sentiment analysis](#). *Artificial Intelligence Review*, 56(Suppl 1):1479–1512.
- C. Hutto and Eric Gilbert. 2014. [Vader: A parsimonious rule-based model for sentiment analysis of social media text](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.
- Xincheng Ju, Dong Zhang, Rong Xiao, Junhui Li, Shoushan Li, Min Zhang, and Guodong Zhou. 2021. [Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4395–4405.
- 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*, MM ’21, page 3034–3042.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.
- 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.
- Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multi-view social data. In *MultiMedia Modeling: 22nd International Conference, MMM 2016, 2016, Proceedings, Part II 22*, pages 15–27. Springer.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- Tianshuo Peng, Zuchao Li, Ping Wang, Lefei Zhang, and Hai Zhao. 2024. [A novel energy based model mechanism for multi-modal aspect-based sentiment](#)

- analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18869–18878.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763.
- Xuefeng Shi, Min Hu, Fuji Ren, Piao Shi, and Satoshi Nakagawa. 2024. [Aspect based sentiment analysis with instruction tuning and external knowledge enhanced dependency graph](#). *Applied Intelligence*, pages 1–18.
- Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. [Rpbert: A text-image relation propagation-based bert model for multimodal ner](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13860–13868.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. [Llama: Open and efficient foundation language models](#). *arXiv preprint arXiv:2302.13971*.
- Di Wang, Yuzheng He, Xiao Liang, Yumin Tian, Shaofeng Li, and Lin Zhao. 2024a. [TMFN: A target-oriented multi-grained fusion network for end-to-end aspect-based multimodal sentiment analysis](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16187–16197.
- Di Wang, Changning Tian, Xiao Liang, Lin Zhao, Lihuo He, and Quan Wang. 2024b. [Dual-perspective fusion network for aspect-based multimodal sentiment analysis](#). *IEEE Transactions on Multimedia*, 26:4028–4038.
- Xiaohua Wang, Wenlong Fei, Min Hu, Qingyu Zhang, and Aoqiang Zhu. 2024c. [MEVTR: A multilingual model enhanced with visual text representations](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11247–11261.
- Hanqian Wu, Siliang Cheng, Jingjing Wang, Shoushan Li, and Lian Chi. 2020a. [Multimodal aspect extraction with region-aware alignment network](#). In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, 14–18, 2020, Proceedings, Part I 9*, pages 145–156. Springer.
- Yang Wu, Yanyan Zhao, Hao Yang, Song Chen, Bing Qin, Xiaohuan Cao, and Wenting Zhao. 2022. [Sentiment word aware multimodal refinement for multimodal sentiment analysis with ASR errors](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1397–1406, Dublin, Ireland.
- Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020b. [Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts](#). In *Proceedings of the 28th ACM International Conference on Multimedia, MM '20*, page 1038–1046.
- Luwei Xiao, Xingjiao Wu, Junjie Xu, Weijie Li, Cheng Jin, and Liang He. 2024. [Atlantis: Aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis](#). *Information Fusion*, 106:102304.
- Xiaocui Yang, Shi Feng, Daling Wang, Qi Sun, Wenfang Wu, Yifei Zhang, Pengfei Hong, and Soujanya Poria. 2023. [Few-shot joint multimodal aspect-sentiment analysis based on generative multimodal prompt](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11575–11589.
- Jianfei Yu, Kai Chen, and Rui Xia. 2023. [Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis](#). *IEEE Transactions on Affective Computing*, 14(3):1966–1978.
- Jianfei Yu and Jing Jiang. 2019. [Adapting bert for target-oriented multimodal sentiment classification](#). In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5408–5414.
- Jianfei Yu, Jing Jiang, and Rui Xia. 2020a. [Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439.
- Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020b. [Improving multimodal named entity recognition via entity span detection with unified multimodal transformer](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3352.
- Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2023. [A survey on aspect-based sentiment analysis: Tasks, methods, and challenges](#). *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11019–11038.
- Gang Zhao, Guanting Dong, Yidong Shi, Haolong Yan, Weiran Xu, and Si Li. 2022. [Entity-level interaction via heterogeneous graph for multimodal named entity recognition](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6345–6350.
- 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.

Ru Zhou, Wenya Guo, Xumeng Liu, Shenglong Yu, Ying Zhang, and Xiaojie Yuan. 2023. **AoM: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8184–8196.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Fuji Ren. 2024a. **KEBR: Knowledge enhanced self-supervised balanced representation for multimodal sentiment analysis**. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM '24*, page 5732–5741.

Linlin Zhu, Heli Sun, Qunshu Gao, Tingzhou Yi, and Liang He. 2024b. **Joint multimodal aspect sentiment analysis with aspect enhancement and syntactic adaptive learning**. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6678–6686.

Labels	Twitter -2015			Twitter -2017		
	Train	Dev	Test	Train	Dev	Test
Negative	368	149	113	416	144	168
Neutral	1883	670	607	1638	517	573
Positive	928	303	317	1508	515	493
One aspect	2,159 (61.65%)			976 (33.54%)		
Mult. aspects	1,343 (38.35%)			1,934 (66.46%)		
Total Aspects	3,502			2,910		

Table 7: Dataset statistics."Mult. aspects" denotes "Multiple aspects".

A Experimental settings

A.1 Datasets

Following prior research (Ju et al., 2021; Zhou et al., 2023; Ling et al., 2022), we evaluate the performance of our proposed DaNet model using two widely used benchmark datasets: Twitter-2015 and Twitter-2017 (Yu and Jiang, 2019). The statistics of these two datasets are summarized in Table 7.

A.2 Implementation Details

All models are built on the PyTorch with the NVIDIA RTX A40 GPU. The Adam optimizer is used for both pre-training and fine-tuning. Our model is based on BART (Wang et al., 2024c), trained for 60 epochs with a batch size of 32 on the pretraining task and 35 epochs with a batch size of 16 on the downstream task of MABSA. The learning rate is set to $7e-5$. The trade-off hyperparameters λ_1 and λ_2 are all set to 1. The hidden layer size is 768. SpaCy, YOLOv8, DeepSentiBank, and ChatGPT all use the default settings provided on

Hyper-parameter	Value
d	768
α	0.7
λ_1, λ_2	1
Hidden size	768
Optimizer	Adam
Learning rate	$7e-5$
Epoch for training	35
Epoch for pretraining	60
Batch size for training	16
Batch size for pretraining	35

Table 8: Hyper-parameters setting.

their official websites. The experimental results are taken as the average of three consecutive experiments. The details of the relevant parameters are given in Table 8.

A.3 Computational Overhead

Table 9 provides the computational overhead for model pre-training (on MVSA-Multi (Niu et al., 2016)) and training (on Twitter-2015 (Yu and Jiang, 2019)).

B Baselines

This paper compares the performance of DaNet with other baseline models on the MATE, MASC, and MABSA tasks.

Baselines for MATE: RAN (Wu et al., 2020a) proposes a region-aware alignment network to align text with object regions in images. **UMT** (Yu et al., 2020b) uses a cross-modal transformer to fuse text and image representations to mitigate visual biases. **OSCGA** (Wu et al., 2020b) bridges visual and linguistic information by utilizing object labels as embeddings for multimodal interaction.

Baselines for MASC: ESAFN (Yu et al., 2020a) learns entity-sensitive visual representations and integrates them with LSTM to mitigate visual noise. **TomBERT** (Yu and Jiang, 2019) applies BERT to model intra-modality dynamics for obtaining aspect-sensitive text representations. **CapTrBERT** (Khan and Fu, 2021) constructs an auxiliary sentence for the translation of an image, to provide multimodal information for the language model.

Baselines for MABSA: UMT-collapse (Shi et al., 2024), **OSCGA-collapse** (Wu et al., 2020b), and **RpBERT-collapse** (Sun et al., 2021) are adapted from models for MATE

Model	Params	Time / Epoch
Pre-training	153 M	27m 34s
Training	158 M	1m 38s

Table 9: Computational overhead.

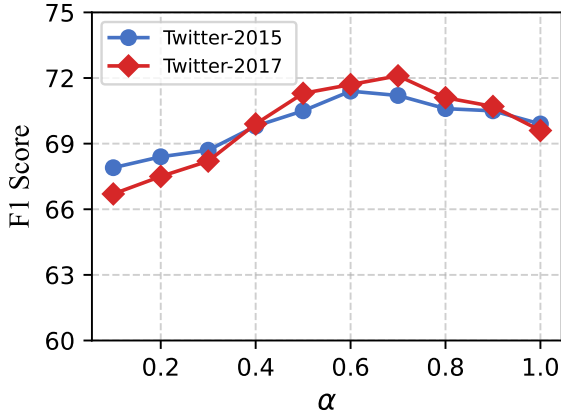


Figure 5: The impact of different image retention rates α .

by using collapsed labels to represent aspect-sentiment pairs. **UMT+TomBERT** and **OSCGA+TomBERT** achieve joint multimodal aspect-based sentiment analysis by combining UMT (Yu et al., 2020b) and OSCGA (Wu et al., 2020b) with TomBERT (Yu and Jiang, 2019). **JML** (Ju et al., 2021) performs the MABSA task by introducing auxiliary cross-modal relation detection. **CMMT** (Wu et al., 2022) introduced a gating mechanism to control the contribution of multimodal information during the interaction process between modalities. **VLP-MABSA** (Ling et al., 2022) designs multiple visual language pre-training tasks to perform task-specific unified multimodal architectures. **AoM** (Zhou et al., 2023) focuses on the alignment of text tokens and image blocks to reduce the interference of irrelevant regions. **AESAL** (Zhu et al., 2024b) proposes an aspect-enhanced and syntax-adaptive learning approach to capture differences in the importance of various words within the text.

C Impact of Image Retention Rate.

To evaluate the impact of image retention rates α on performance, we tested MABSA under varying α values and measured F1 scores, as shown in Fig. 5. Results indicate that both excessively high and low α degrade performance across datasets. For Twitter-2015, optimal performance occurs at $\alpha \approx 0.6$, while for Twitter-2017, it peaks at $\alpha \approx 0.7$. This difference arises because Twitter-2015

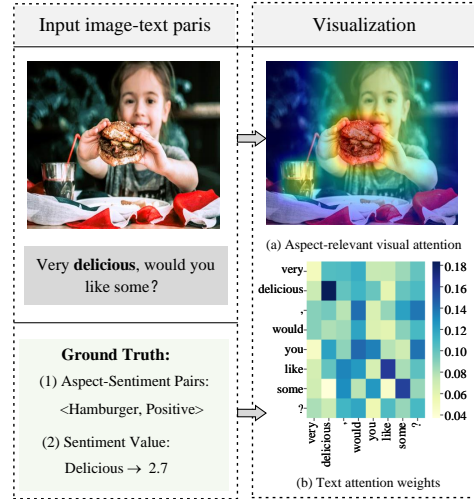


Figure 6: Visualization aspects related to attention.

primarily contains single-aspect data (61.65%), requiring fewer retained image regions and exhibiting greater stability. In contrast, Twitter-2017 includes more multi-aspect data (66.46%), where excessive regions introduce noise, and insufficient regions lose critical information. Notably, this effect is manageable due to our approach of weighting weakly relevant regions and fusing them with strongly relevant ones, preserving key information. Since Twitter-2017 achieves a satisfactory F1 score of 71.2 at $\alpha = 0.7$, we standardized $\alpha = 0.7$ for all datasets in subsequent experiments.

D Implicit Case Study and Visualization

To evaluate DaNet’s ability to detect implicit aspect-opinion relationships, we visualize the attention mechanism using the example in Fig. 1, shown in Fig. 6.

For visual attention: Fig. 6(a) illustrates the proportion of visual information retained in the last step of the dual-aware enhanced alignment network, where we weighted add the representation of visual patches and the corresponding aspects. It can be observed from Fig. 6(a) that semantically relevant image regions are successfully located. For example, in the case of Fig. 6, the text “Very delicious, would you like some?” semantically refers to food. Combined with the visual information of a girl happily offering a hamburger, it can be known that the semantically related area in the image is a hamburger. The heatmap of Fig. 6(a) shows that the regions associated with “hamburger” are prominently preserved, while other irrelevant regions are ignored.



Image			
Text	(a) Team @ Equality_MI is back in Detroit today for day two of @ MotorCityPride and the pride parade .	(b) Houston ‘has emerged’ as serious threat to sign Chris Paul (per @ ESPNSteinLine)	(c) Jealous Iniesta says Real Madrid will be disappointed not to have won La Liga
Ground Truth	(Equality_MI, POS) (Detroit, NEU) (MotorCityPride, NEU)	(Houston, NEG) (Chris Paul, NEU)	(Iniesta , NEG) (Real Madrid, NEG) (La Liga, NEU)
VLP-MABSA	(Equality_M, POS) (✓, ✓) (Detroit, NEU) (✓, ✓) (MotorCityPride, POS) (✓, ✗)	(Houston, NEG) (✓, ✓) (Chris Paul, NEG) (✓, ✗)	(Iniesta , NEG) (✓, ✓) (Real Madrid, NEG) (✓, ✓) (La Liga, NEU) (✓, ✓)
CMMT	(Equality_M, POS) (✓, ✓) (Detroit, NEU) (✓, ✓) (MotorCity, POS) (✗, ✗)	(Houston, NEG) (✓, ✓) (Chris Paul, NEU) (✓, ✓)	(Iniesta , NEG) (✓, ✓) (Real Madrid, NEG) (✓, ✓) (La Liga, NEG) (✓, ✗)
AoM	(Equality_M, POS) (✓, ✓) (Detroit, NEU) (✓, ✓) (MotorCityPride, POS) (✓, ✗)	(Houston, NEG) (✓, ✓) (Chris Paul, NEU) (✓, ✓)	(Iniesta , NEG)(✓, ✓) (Real Madrid, NEG)(✓, ✓) (La Liga, NEG)(✓, ✗)
DaNet	(Equality_M, POS) (✓, ✓) (Detroit, NEU) (✓, ✓) (MotorCityPride, NEU) (✓, ✓)	(Houston, NEG) (✓, ✓) (Chris Paul, NEU) (✓, ✓)	(Iniesta , NEG)(✓, ✓) (Real Madrid, NEG)(✓, ✓) (La Liga, NEU)(✓, ✓)

Figure 7: Case studies. NEU, POS, and NEG denote Neutral, Positive, and Negative sentiments.

For text attention: Fig. 6(b) shows the attention given to the input text by the last hidden layer state of the encoder after the dual-aware enhanced alignment network. From Fig. 6(b), it is evident that text related to aspects and sentiments receives more attention. For example, from the text "Very delicious, would you like some?", candidate aspects ("same") and candidate sentiments ("delicious", "like") are extracted. These candidate words receive relatively high attention, as shown in Fig. 6(b). When analyzed together with the image, "delicious hamburger" fits the context better, and "delicious" receives more attention compared to "like". This highlights DaNet’s semantic understanding, enhanced by guidance from LLM.

E Case Studies

To further demonstrate the effectiveness of our model DaNet, we compared DaNet with high performing models that provide reproducible code (i.e., VLP-MABSA, CMMT, and AoM). Fig. 7 presents three examples with predictions from VLP-MABSA, CMMT, AoM, and our DaNet.

Fig. 7 clearly illustrates that, in example (a), both VLP-MABSA and AoM incorrectly classified the aspect term "MotorCityPride". It is possible that "MotorCityPride" is a compound word and that both the word "Pride" and the content of the im-

age express positive sentiment. However, CMMT only identified the aspect terms "Equality_M" and "Detroit", failing to predict "MotorCityPride". In example (b), all baseline models made correct predictions except for VLP-MABSA, which incorrectly predicted the sentiment for "Chris Pau". It is worth noting that all models make the correct entity prediction in case (c), which may be affected by the fact that case (c) as a whole expresses negative sentiment, and AoM and CMMT put the neutral "La Liga" prediction as negative sentiment. Benefiting from fine-grained alignment with dual awareness of aspects and sentiments, as well as enhancement from specific pre-training tasks, our model, DaNet, accurately extracted all aspect terms and classified the sentiments in all three cases.