Beyond Polarity: Continuous Affect-Enhanced Multimodal Aspect-Based Sentiment Classification

Ling-Ang Meng[1], Tianyu Zhao[1], Dawei Song[1,2]*, Jingxu Cao[1], Youhui Zuo[1]
[1]School of Computer Science & Technology, Beijing Institute of Technology, Beijing, China
[2]School of Computing and Communications, The Open University, UK
ling.ang.meng@bit.edu.cn , dwsong@bit.edu.cn

Abstract

Multimodal aspect-based sentiment classification (MABSC) requires aspect-level sentiment inference from textual-image data that jointly convey opinions. Yet most existing approaches primarily exploit discrete polarity patterns and generic visual embeddings, making them less effective when the affect is subtle, implicit, or expressed through imagery. In this work, we propose ***VADE***, a Valence–Arousal–Dominance (***VAD***)-***E***nhanced MABSC framework that brings continuous VAD signals into multimodal sentiment reasoning and learns emotion-sensitive image representations. Specifically, we design a VAD encoder to extract continuous affect cues from text for aspect-level sentiment reasoning. Furthermore, we fine-tune a CLIP-based image encoder on affect-enriched image–text pairs to obtain visual representations that are more sensitive to sentiment cues. To support the fine-tuning process, we construct an affect-enriched image–text dataset ***Senti-COCO*** by rewriting MSCOCO captions with a multimodal large language model, which yields large-scale image-text pairs with richer affective expressions. Experiments on two mainstream datasets, Twitter-15 and Twitter-17, show that VADE achieves state-of-the-art results, demonstrating the effectiveness of incorporating VAD signals for MABSC. Our code and dataset are publicly available at https://github.com/Maydayflower/VADE.

1 Introduction

Multimodal aspect-based sentiment classification (MABSC) aims to predict the sentiment polarity toward a given aspect term in user-generated posts, such as tweets, where text and images jointly convey opinions (Zhao et al., 2024). Compared with sentence-level multimodal sentiment analysis, MABSC requires finer-grained reasoning because multiple aspects can co-exist in the same tweet and each aspect may associate with a different sentiment. As shown in Figure 1, each sample consists of an image–text pair with aspect terms marked in the text, and the model predicts a polarity label for each aspect term.

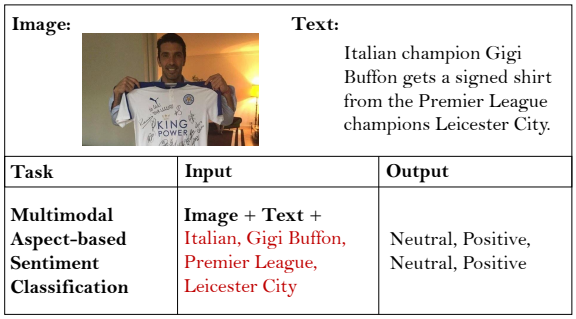

Figure 1: An example of multimodal aspect-based sentiment classification (MABSC): given an image, its accompanying tweet, and the aspect terms (highlighted in the text), the goal is to predict the sentiment polarity (POSITIVE/NEUTRAL/NEGATIVE) for each aspect term.

Existing MABSC methods have mainly focused on improving how textual and visual evidence is aligned and fused, aiming to reduce modality noise while highlighting aspect-relevant sentiment cues (Yu and Jiang, 2019; Yu et al., 2019; Khan and Fu, 2021; Xu et al., 2019; Yang et al., 2022). As a result, current approaches typically treat the image as a source of auxiliary evidence and rely on certain fusion mechanisms to discover whatever cues the visual modality may contain. This reflects a common research paradigm in sentiment analysis, where pretrained models are used to extract features and a classifier is applied to predict discrete sentiment labels.

However, affect is inherently continuous: decades of psychology describe core affect along continuous dimensions (Russell, 1980, 2003). To address this gap, recent research has introduced emotion space theory into sentiment analysis, advo-

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 39717–39727

cating for Dimensional Sentiment Analysis (DSA) that models continuous affective variation rather than relying solely on discrete categories (Park et al., 2021; Ghosh et al., 2023). While dimensional sentiment analysis has proven effective in capturing nuanced human emotions through continuous affective dimensions such as valence and arousal (Buechel and Hahn, 2017; Yu et al., 2016), its application has been largely confined to document-level or sentence-level analysis in unimodal settings. Although the dimABSA shared task (Lee et al., 2024) represents progress toward dimensional aspect-level analysis, the participating systems (Zhang et al., 2024; Zhu et al., 2024) remain text-centric, overlooking the complementary visual cues that current MABSC research has shown to be valuable. Therefore, a critical question remains unresolved: *how can we effectively incorporate continuous dimensional affect representations into MABSC to capture both the nuanced intensity of emotions and the complementary information from visual and textual modalities?*

In this work, we propose ***VADE***, a Valence–Arousal–Dominance (***VAD***)-***E***nhanced framework for MABSC that explicitly integrates continuous affect into aspect-level sentiment reasoning. To facilitate affect-aware visual representation learning, we first construct Senti-COCO, an affect-enriched image–text dataset derived from MSCOCO (Chen et al., 2015) by rewriting caption sets with a multimodal large language model, Qwen2.5-VL-7B (Bai et al., 2025). Specifically, we craft prompts that ask Qwen2.5-VL-7B to keep the visual content intact while rephrasing the captions to express stronger and more nuanced emotion, so that each image is paired with a caption that conveys richer affective signals. Built on this dataset, we develop an Affect-Aware CLIP visual encoder by fine-tuning CLIP on Senti-COCO so that the resulting image representations are more sensitive to affective cues. Together with an aspect-aware text encoder and a BERT-based VAD encoder for extracting Valence–Arousal–Dominance signals from text, these components form our full VADE framework for multimodal aspect-level sentiment classification.

Our contributions are summarized as follows:

- We propose VADE, a VAD-Enhanced MABSC framework that incorporates continuous Valence–Arousal–Dominance signals into aspect-level sentiment reasoning.
- We design a VAD encoder to extract affect cues from text and integrate the resulting VAD features into the VADE model, complementing the traditional discrete polarity patterns.
- We create Senti-COCO, an affect-enriched image-text dataset by rewriting the MSCOCO captions with Qwen2.5-VL-7B, and use it to fine-tune a CLIP-based image encoder, yielding emotion-sensitive visual representations that better capture affective cues.

2 Related Work

2.1 Multimodal Aspect-based Sentiment Classification (MABSC)

MABSC extends traditional aspect-based sentiment analysis by incorporating multiple modalities, such as images, to enhance sentiment inference at the aspect level. Early studies demonstrate that visual information can provide complementary sentiment cues when textual expressions are implicit or ambiguous, motivating the integration of visual and textual signals for fine-grained sentiment understanding.

Prior MABSC research has mainly improved aspect-level prediction by enhancing cross-modal alignment and fusion while reducing modality noise. Early target-aware models introduce target-sensitive text representations and target–image matching (Yu and Jiang, 2019), and subsequent work develops richer interaction and fusion mechanisms (Yu et al., 2019; Zhang et al., 2021). To better bridge the modality gap, several methods adopt alignment/translation or structured reasoning designs, including coarse-to-fine image–target matching (Yu et al., 2022b), hierarchical interaction with reconstruction (Yu et al., 2022a), and caption/graph-assisted grounding or filtration (Xiao et al., 2023; Huang et al., 2023; Wang et al., 2023). More recently, vision–language pretraining has been explored to encode aspect/sentiment signals (Ling et al., 2022; Zhou et al., 2023), with additional efforts incorporating sentiment-aware pretraining or affective region modeling (Ye et al., 2022; Jia et al., 2023).

2.2 Dimensional Sentiment Analysis

Dimensional Sentiment Analysis models affective states in a continuous space rather than assigning

discrete sentiment labels. Early studies in psychology represent emotions using low-dimensional continuous variables, such as Valence-Arousal (VA) or Valence-Arousal-Dominance (VAD), which describe emotional polarity, intensity, and control, respectively. Among them, the circumplex model proposed by Russell characterizes emotions as points in a two-dimensional valence-arousal space, providing a principled foundation for continuous affect modeling.

Building upon these theories, computational approaches have explored dimensional sentiment analysis primarily in text-based settings. Early methods rely on lexicon-based mappings or regression models to predict affective dimensions from linguistic features (Mohammad, 2018, 2025). Mohammad obtained reliable human ratings of valence, arousal, and dominance for over 20,000 English words using Best-Worst Scaling, creating the NRC VAD Lexicon. With the advent of deep learning, neural models employing recurrent or transformer-based architectures have been proposed to capture contextual cues and map text representations to continuous emotion scores (Park et al., 2021; Alahmadi et al., 2025). Such approaches demonstrate improved flexibility in modeling subtle emotional variations compared to discrete sentiment classification.

However, dimensional sentiment analysis remains largely absent in multimodal and fine-grained domains. Existing approaches predominantly focus on holistic emotion prediction at the document or sentence level (Buechel and Hahn, 2017; Yu et al., 2016), without explicitly modeling affective dimensions for specific aspects or entities within the content. Similarly, while recent work has begun exploring dimensional representations in aspect-based sentiment analysis (Lee et al., 2024), these approaches have primarily remained in the unimodal text domain, neglecting the rich semantic information available in accompanying images (Zhang et al., 2024).

3 Method

Task Definition. Multimodal Aspect-Based Sentiment Classification (MABSC) aims to identify the sentiment polarity expressed toward a specific aspect by jointly exploiting textual and visual information. Formally, each data instance is represented as a triplet (I, T, a), where I denotes an associated image, $T = \{w_1, \ldots, w_n\}$ is a sentence, and a indicates the target aspect term. The objective of MABSC is to predict a sentiment label $y \in \{\textit{positive}, \textit{neutral}, \textit{negative}\}$ that reflects the sentiment toward aspect a conditioned on both modalities, which can be formulated as:

$$y = \arg\max_{c \in \mathcal{Y}} \; p(c \mid I, T, a). \tag{1}$$

Model Overview. As shown in Figure 2, our proposed framework consists of (i) a feature extractor that produces textual features, (ii) a VAD encoder that maps an input sentence to a 3-dimensional Valence–Arousal–Dominance vector, and (iii) an Affect-Aware CLIP module that adapts the CLIP image encoder using affect-enriched image–text pairs constructed via an MLLM. For each instance, we obtain the text representation t, image representation v, and VAD vector e. We concatenate these features and use an MLP classifier to predict the sentiment polarity.

3.1 VAD Encoder

To incorporate continuous affect, we train a VAD encoder that maps a sentence to a 3-dimensional affect vector $e = [\mathrm{Val}, \mathrm{Aro}, \mathrm{Dom}] \in \mathbb{R}^3$. Although the NRC-VAD lexicon provides word-level VAD ratings, we train the model to predict sentence-level VAD via weak labeling.

3.1.1 Weak Sentence-level VAD Targets from NRC-VAD

Given an input sentence $S = \{w_1, \ldots, w_m\}$, we match tokens that appear in the NRC-VAD lexicon. Let $\mathcal{M}(S)$ be the set of matched tokens and $\mathbf{v}(w) \in \mathbb{R}^3$ be the lexicon VAD vector for token w. We construct a pseudo sentence-level target by averaging matched token scores:

$$\mathbf{e}_S^* = \frac{1}{|\mathcal{M}(S)|} \sum_{w \in \mathcal{M}(S)} \mathbf{v}(w). \tag{2}$$

3.1.2 Training and Inference

We use a BERT encoder (Devlin et al., 2019) followed by a regression head $g(\cdot)$ to predict VAD:

$$\mathbf{h}_S = \mathrm{BERT}(S), \qquad \hat{\mathbf{e}}_S = g(\mathbf{h}_S) \in \mathbb{R}^3. \tag{3}$$

The training objective is mean squared error:

$$\mathcal{L}_{\mathrm{VAD}} = \|\hat{\mathbf{e}}_S - \mathbf{e}_S^*\|_2^2 . \tag{4}$$

During MABSC inference, we feed the tweet sentence (we use the same sequence $\tilde{X}$ for consistency) into the VAD encoder to obtain:

$$e = f_{\mathrm{VAD}}(\tilde{X}) \in \mathbb{R}^3, \tag{5}$$

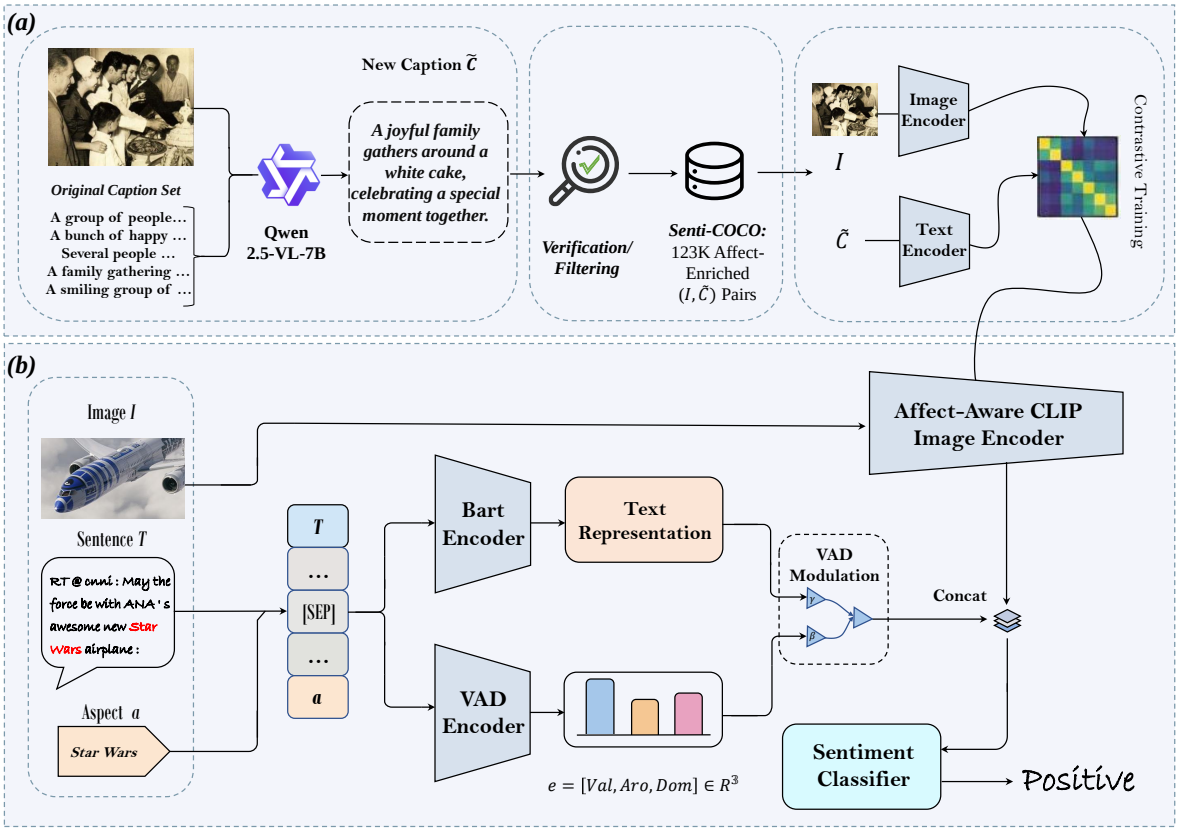

Figure 2: Overview of the proposed VADE framework. (a) Construction of Senti-COCO: for each MSCOCO image, the original caption set is rewritten by Qwen2.5-VL-7B into an affect-enriched caption; the resulting 123K image–caption pairs are used to fine-tune CLIP with a contrastive objective. (b) VADE for MABSC: the tweet sentence and aspect term are concatenated as $T[SEP]a$ and encoded by a BART encoder to obtain textual features, while a VAD encoder predicts a continuous affect vector that modulates the text representation; in parallel, the image is encoded by the affect-aware CLIP image encoder. The modulated text feature and image feature are concatenated and fed into a sentiment classifier to predict the aspect-level polarity.

which is then fused with t and v for polarity prediction.

3.1.3 VAD-Conditioned Modulation of Text Representations

Beyond concatenating VAD as an auxiliary feature, we further inject affect into the textual stream via a VAD-conditioned feature modulation layer. Let $H \in \mathbb{R}^{B\times L\times d_t}$ denote the token-level hidden states output by the text encoder (BART) for a mini-batch of size B and sequence length L. Given the VAD vector $e \in \mathbb{R}^3$, we generate a sample-specific scaling vector and bias vector through two linear mappings:

$$\gamma = W_\gamma e + b_\gamma \in \mathbb{R}^{d_t},$$
$$\beta = W_\beta e + b_\beta \in \mathbb{R}^{d_t}. \tag{6}$$

We then apply a channel-wise affine transformation (FiLM-style modulation) to all tokens:

$$\tilde{H}_{b,l,:} = \gamma_b \odot H_{b,l,:} + \beta_b \tag{7}$$

$$\forall\, b \in \{1,\dots,B\},\ l \in \{1,\dots,L\}, \tag{8}$$

where $\odot$ denotes element-wise multiplication. Finally, we obtain a VAD-modulated sentence representation by mean pooling over the sequence:

$$t_{\text{vad}} = \frac{1}{L}\sum_{l=1}^{L} \tilde{H}_{:,l,:} \in \mathbb{R}^{B\times d_t}. \tag{9}$$

This module enables the continuous VAD signal to explicitly control the magnitude and bias of textual feature channels, thereby producing affect-aware text representations for downstream fusion and classification.

3.2 Affect-Aware CLIP

Standard CLIP features primarily emphasize general semantic alignment between images and text. To obtain affect-sensitive image representations, we fine-tune CLIP on an affect-enriched dataset constructed by rewriting captions to contain richer emotional expressions.

3.2.1 Dataset Construction

We start from MSCOCO, where each image is associated with multiple human-written captions. As shown in Figure 2(a), for each image I, we collect its caption set $\mathcal{C} = \{C_1, \ldots, C_m\}$ and input $(I, \mathcal{C})$ into the multimodal large language model **Qwen2.5-VL-7B**. The model is prompted to produce an affect-enriched caption $\tilde{C}$ that (i) preserves the visual content described by $\mathcal{C}$ and (ii) is more emotionally expressive. We then conduct a manual verification step to filter out low-quality generations (e.g., content drift or inconsistent descriptions), ensuring the rewritten captions remain faithful to the image content. This yields an affect-enriched training pair $(I, \tilde{C})$. Aggregating over images, we construct a dataset of 123K affect-enriched image–text pairs, denoted as **Senti-COCO**. More details about the construction of Senti-COCO are provided in Appendix A.

3.2.2 CLIP Fine-tuning Objective

Let $\tilde{\mathcal{D}} = \{(I_i, T_i)\}_{i=1}^{N}$ denote the constructed affect-enriched dataset, where $T_i = \tilde{C}_i$ is the affect-enriched caption rewritten for image I_i. We adapt CLIP by optimizing a bidirectional contrastive learning objective that pulls matched image–text pairs together while pushing mismatched pairs apart within each mini-batch.

Encoders and normalized embeddings. CLIP consists of an image encoder $f_{\text{img}}(\cdot)$ (ViT-B/32) and a text encoder $f_{\text{txt}}(\cdot)$, followed by linear projections into a shared 512-dimensional embedding space. For a mini-batch of size B, we compute ℓ_2-normalized embeddings:

$$\begin{aligned} \mathbf{v}_i &= \frac{W_{\text{img}} f_{\text{img}}(I_i)}{\|W_{\text{img}} f_{\text{img}}(I_i)\|_2}, \\ \mathbf{u}_i &= \frac{W_{\text{txt}} f_{\text{txt}}(T_i)}{\|W_{\text{txt}} f_{\text{txt}}(T_i)\|_2}, \end{aligned} \tag{10}$$

where $\mathbf{v}_i, \mathbf{u}_i \in \mathbb{R}^{512}$.

Similarity with temperature. We compute the pairwise cosine similarities for all image–text pairs in the batch and scale them by a learnable temperature $\tau > 0$:

$$S_{ij} = \frac{\mathbf{v}_i^\top \mathbf{u}_j}{\tau}, \qquad i, j \in \{1, \ldots, B\}. \tag{11}$$

Here, (I_i, T_i) is treated as the only positive pair for image I_i (and text T_i), while $\{T_j\}_{j \neq i}$ and $\{I_j\}_{j \neq i}$ serve as in-batch negatives.

Image-to-text and text-to-image contrastive learning. We define a categorical distribution over candidate texts given an image:

$$p(j \mid i) = \frac{\exp(S_{ij})}{\sum_{k=1}^{B} \exp(S_{ik})}, \tag{12}$$

and similarly a distribution over candidate images given a text:

$$q(j \mid i) = \frac{\exp(S_{ji})}{\sum_{k=1}^{B} \exp(S_{ki})}. \tag{13}$$

The image-to-text loss encourages the matched caption T_i to be the most probable text for image I_i:

$$\begin{aligned} \mathcal{L}_{i2t} &= -\frac{1}{B} \sum_{i=1}^{B} \log p(i \mid i) \\ &= -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(S_{ii})}{\sum_{j=1}^{B} \exp(S_{ij})}. \end{aligned} \tag{14}$$

The text-to-image loss is defined analogously:

$$\begin{aligned} \mathcal{L}_{t2i} &= -\frac{1}{B} \sum_{i=1}^{B} \log q(i \mid i) \\ &= -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(S_{ii})}{\sum_{j=1}^{B} \exp(S_{ji})}. \end{aligned} \tag{15}$$

Finally, we optimize the symmetric CLIP objective:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2} \left(\mathcal{L}_{i2t} + \mathcal{L}_{t2i} \right), \tag{16}$$

which enforces bidirectional alignment between affect-enriched captions and their corresponding images. In practice, we treat τ as a learnable scalar updated jointly with the encoder and projection parameters.

3.2.3 Fine-tuning Configuration

We perform full-parameter fine-tuning (vision encoder, text encoder, and projections; $\sim$ 150M trainable parameters). We use AdamW with learning rate 1×10^{-6}, weight decay 0.01, and $\beta =$

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

Table 1: Statistics on two datasets of MABSC.

$(0.9, 0.98)$, and apply gradient clipping with max norm 1.0. The learning rate follows cosine annealing with 10% warmup and a minimum learning rate of 1×10^{-7}. We train for 10 epochs with FP16 mixed precision and an effective batch size of 128 (batch size 32 with 4-step gradient accumulation).

4 Experiments

4.1 Experimental Settings

Datasets and Evaluation Metrics. We evaluate our approach on two widely used datasets for MABSC: Twitter-15 and Twitter-17, where each instance consists of an image-tweet pair and a set of aspect terms annotated in the tweet text. The statistics for these datasets are presented in Table 1. Following the standard protocol in prior work, we use the official train/dev/test splits provided by the benchmarks. We report **Accuracy** (Acc) and **Macro-F1** as evaluation metrics.

Implementation Details. Our model contains three main components: an aspect-aware text encoder, a VAD encoder, and a CLIP image encoder adapted by affect-enriched image–text pairs. For the textual input, we follow an aspect-conditioned formulation and construct the sequence by appending the aspect term to the tweet with a separator:

$$\tilde{X} = X \texttt{ [SEP] } a, \tag{17}$$

which is fed into BART to obtain hidden states. In parallel, the VAD encoder takes the same sequence $\tilde{X}$ as input and outputs a 3-dimensional VAD vector $e \in \mathbb{R}^3$.

To inject affect into the textual stream, we implement a VAD-conditioned modulation layer (FiLM-style): two linear mappings transform e into a per-sample scaling vector and bias vector, which are applied channel-wise to the text hidden states, followed by mean pooling to form the affect-aware text representation. We then concatenate the affect-aware text feature and the image feature, and feed the resulting representation into a lightweight MLP classifier for polarity prediction.

We set the maximum input length to 128 and train the MABSC model for 20 epochs with a batch size of 16. We optimize the model using AdamW with learning rate 2×10^{-5}, warmup steps 0, and gradient clipping with $\texttt{max\_grad\_norm} = 1.0$. We report the average performance over multiple random seeds. All the models are trained and implemented based on PyTorch with one NVIDIA RTX A6000 GPU.

Compared Baselines. We compare VADE with representative baselines under three settings: *image-only*, *text-only*, and *text+image* multimodal models. For *image-only* methods, we include ResTarget (Yu and Jiang, 2019), which extracts visual features for target-aware sentiment prediction. For *text-only* methods, we consider MGAN (Fan et al., 2018) and BERT (Devlin et al., 2019), representing strong pretrained language models. For *text+image* baselines, we compare against a broad set of multimodal MABSC approaches, including MIMN (Xu et al., 2019), TomBERT (Yu and Jiang, 2019), EF-CapTrBERT (Khan and Fu, 2021), FITE (Yang et al., 2022), ITM (Yu et al., 2022b), and VLP-MABSA (Ling et al., 2022), which cover target-aware fusion, caption-augmented modeling, and task-specific vision–language pretraining. We further include AMIFN (Yang et al., 2024), DPCI (Liu et al., 2025), and TCMT (Zou et al., 2025) as recent state-of-art methods, and compare large multimodal or foundation-model baselines, including DeepSeek-V3 (Liu et al., 2024a), LLaMA (Touvron et al., 2023), and LLaVA-v1.5 (Liu et al., 2024b). For fair comparison, we follow the evaluation protocol and data splits reported in prior work and use the officially released implementations or re-implement the models when necessary.

4.2 Main Results

Table 2 reports the performance comparison on Twitter-15 and Twitter-17. VADE achieves the best results on both benchmarks, reaching 81.87 Acc / 77.40 Macro-F1 on Twitter-15 and 77.45 Acc / 76.38 Macro-F1 on Twitter-17. Compared with the strongest non-ours MABSC baseline, TCMT, VADE yields consistent improvements: +0.47 Acc / +0.70 Macro-F1 on Twitter-15 and +0.15 Acc / +0.58 Macro-F1 on Twitter-17. VADE also surpasses the recent DPCI model by a clear margin (+1.45 Acc / +1.01 Macro-F1 on Twitter-15;

	Models	Twitter-15		Twitter-17	
		Acc	Macro-F1	Acc	Macro-F1
Image Only	Res-Target (Yu and Jiang, 2019)	59.88	46.48	58.59	53.98
Text Only	MGAN (Fan et al., 2018)	71.17	64.21	64.75	61.46
	BERT (Devlin et al., 2019)	74.15	68.86	68.15	65.23
Text and Image	MIMN (Xu et al., 2019)	71.84	65.69	65.88	62.99
	TomBERT (Yu and Jiang, 2019)	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., 2022b)	78.27	74.19	72.61	71.97
	VLP-MABSA (Ling et al., 2022)	78.60	73.80	73.80	71.80
	AMIFN (Yang et al., 2024)	78.68	75.50	72.29	70.21
	DPCI (Liu et al., 2025)	80.42	76.39	75.20	74.73
	TCMT (Zou et al., 2025)	81.4	76.7	77.3	75.8
	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
	VADE (ours)	**81.87**†	**77.40**†	**77.45**†	**76.38**†

Table 2: Performance comparison on Twitter-2015 and Twitter-2017 datasets. ♣ denotes the results from DPCI (Liu et al., 2025). † indicates that our model is significantly better than all the compared methods with p-value < 0.05.

	Twitter-15		Twitter-17	
	Acc	Macro-F1	Acc	Macro-F1
VADE (full)	**81.87**	**77.40**	**77.45**	**76.38**
w/o VAD Encoder	80.03	76.00	74.47	73.69
w/o VAD modulation	81.25	76.41	75.73	73.51
w/o Affect-Aware CLIP	80.89	77.15	74.68	73.45
w/o Valence	79.52	74.78	75.21	73.91
w/o Arousal	80.73	76.05	76.39	75.03
w/o Dominance	81.18	76.63	76.82	75.71

Table 3: Ablation study of VADE on Twitter-15 and Twitter-17. We evaluate the impact of removing key components, including the VAD encoder, VAD modulation, and Affect-Aware CLIP, as well as individual VAD dimensions (Valence, Arousal, Dominance).

+2.25 Acc / +1.65 Macro-F1 on Twitter-17). Notably, VADE improves over strong general-purpose VL/LLM baselines (e.g., LLaMA and LLaVA-v1.5), indicating that explicitly modeling continuous affect and learning affect-sensitive visual representations provides complementary benefits beyond generic multimodal semantics. Overall, these results validate the effectiveness of VADE for robust aspect-level sentiment prediction in multimodal tweets.

4.3 Ablation Study

To better understand the contribution of each component in VADE, we conduct ablation experiments on Twitter-15 and Twitter-17, as shown in Table 3. First, removing the VAD encoder leads to a clear performance drop on both datasets (e.g., Macro-F1 drops from 77.40 to 76.00 on Twitter-15), demonstrating that explicit modeling of continuous affect provides useful complementary signals for aspect-level sentiment prediction. Second, eliminating the VAD modulation mechanism also degrades performance (76.41 Macro-F1 on Twitter-15), indicating that simply extracting VAD features is insufficient; effectively injecting them into textual representations is crucial for leveraging affect information. Third, replacing the affect-aware CLIP with the original CLIP results in consistent declines (e.g., 2.25 Acc on Twitter-15), highlighting the importance of adapting visual representations to capture affective cues rather than relying on general semantic features. Finally, we analyze the impact of individual VAD dimensions. Removing Valence causes the most significant degradation, confirming that it is the primary signal aligned with sentiment polarity. Meanwhile, removing Arousal or Dominance also leads to noticeable drops, suggesting that these dimensions provide complementary information (e.g., emotional intensity and control) that cannot be captured by valence alone. Overall, these results verify that each component in VADE contributes to the final performance, and that jointly modeling continuous affect across modalities is essential for robust MABSC.

Figure 3: Case study on multimodal aspect-based sentiment classification. We compare predictions from TomBERT and VLP-MABSA with our VADE on three tweet–image examples. Each instance contains an image, tweet text, and multiple aspect terms with gold polarities. VADE produces more accurate aspect-level predictions, especially when sentiment is subtle, implicit, or requires affective interpretation beyond surface semantics. Correct and incorrect predictions are marked with ✓ and ×, respectively.

4.4 Case Study

Figure 3 presents qualitative comparisons between VADE and two representative baselines (TomBERT and VLP-MABSA) on three samples. In the first sample, the tweet reports an injury-related news event ("Kyrie Irving will miss the rest of NBA Finals with a fractured left kneecap"), where the gold labels indicate negative sentiment toward Cavs but neutral sentiment toward NBA Finals. TomBERT incorrectly predicts Cavs as neutral, suggesting limited sensitivity to the affective implication of the event, while VLP-MABSA correctly identifies the negative polarity for Cavs but fails on NBA Finals. In contrast, VADE correctly distinguishes the different sentiments associated with the two aspects, reflecting improved affect-aware reasoning.

In the second sample, the tweet contains a metaphorical expression ("Lionel Messi could nutmeg a Mermaid"), where the gold annotations assign positive sentiment to Lionel Messi and neutral sentiment to Mermaid. TomBERT misclassifies Lionel Messi as neutral, and VLP-MABSA overpredicts negativity for Mermaid. VADE correctly predicts the polarity for both aspects, consistent with its ability to incorporate continuous affect cues into aspect-level inference.

In the third sample, the tweet describes weather conditions for baseball (Another warm evening for baseball in Flushing"), with a neutral label for **Flushing**. While TomBERT yields the correct neutral prediction, VLP-MABSA predicts positive, likely influenced by the upbeat phrasing (warm evening") that can correlate spuriously with positive polarity. VADE remains neutral, suggesting that VAD-informed modeling helps avoid overcommitting to discrete polarity cues when the overall affect is mild or informational. Overall, these cases illustrate that VADE better captures implicit affect and reduces spurious polarity shifts, leading to more reliable aspect-level sentiment prediction in multimodal tweets.

5 Conclusion

In this paper, we presented ***VADE***, a VAD-Enhanced framework for multimodal aspect-based sentiment classification that explicitly incorporates continuous affect cues into aspect-level multimodal sentiment reasoning. VADE combines a BART text encoder, a BERT-based VAD encoder trained with weak supervision from the NRC-VAD lexicon, and an affect-aware CLIP image encoder adapted on **Senti-COCO**, a 123K affect-enriched image–text

dataset constructed by rewriting MSCOCO caption sets with Qwen2.5-VL-7B. Moreover, we inject VAD into the textual stream via a lightweight VAD-conditioned modulation mechanism, enabling continuous affect to directly shape textual representations before fusion. Extensive experiments on Twitter-15 and Twitter-17 demonstrate that VADE achieves strong performance and that continuous VAD signals, together with affect-sensitive visual representations, provide complementary benefits for MABSC. In future work, we plan to explore finer-grained affect modeling (e.g., aspect-specific affect estimation) and more general affect-aware multimodal pretraining strategies beyond caption rewriting.

Limitations

Although VADE demonstrates strong effectiveness for multimodal aspect-based sentiment classification, it has several limitations. First, our study focuses exclusively on MABSC and does not investigate the other two closely related tasks in the MABSA family, namely multimodal aspect term extraction (MATE) and joint multimodal aspect-sentiment analysis (JMABSA). Extending VAD-Enhanced affect modeling to jointly handle aspect discovery and sentiment prediction remains an important direction for future work. Second, our affect-aware visual adaptation relies on caption rewriting to construct Senti-COCO; while this strategy is scalable, the rewritten captions may still introduce stylistic bias or occasional content drift, which could affect the learned affective alignment. Third, our VAD encoder is trained with weak sentence-level targets aggregated from a word-level lexicon, which may be less accurate for compositional, ironic, or context-dependent affect.

Acknowledgments

This work is funded in part by the Natural Science Foundation of China (grant number: 62376027).

References

Khaled Alahmadi, Sultan Alharbi, Juan Chen, and Xianzhi Wang. 2025. Generalizing sentiment analysis: a review of progress, challenges, and emerging directions. *Social Network Analysis and Mining*, 15(1):1–28.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *ACL-HLT 2019*, pages 4171–4186.

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3433–3442, Brussels, Belgium. Association for Computational Linguistics.

Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2023. Vad-assisted multitask transformer framework for emotion recognition and intensity prediction on suicide notes. *Information Processing Management*, 60(2):103234.

Yufeng Huang, Zhuo Chen, Jiaoyan Chen, Jeff Z Pan, Zhen Yao, and Wen Zhang. 2023. Target-oriented sentiment classification with sequential cross-modal semantic graph. In *International Conference on Artificial Neural Networks*, pages 587–599. Springer.

Li Jia, Tinghua Ma, Huan Rong, and Najla Al-Nabhan. 2023. Affective region recognition and fusion network for target-level multimodal sentiment classification. *IEEE Transactions on Emerging Topics in Computing*.

Zaid Khan and Yun Fu. 2021. Exploiting bert for multimodal target sentiment classification through input space translation. In *ACM Multimedia 2021*, pages 3034–3042.

Lung-Hao Lee, Liang-Chih Yu, Suge Wang, and Jian Liao. 2024. Overview of the SIGHAN 2024 shared task for Chinese dimensional aspect-based sentiment analysis. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 165–174, Bangkok, Thailand. Association for Computational Linguistics.

Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. In *ACL 2022*, pages 2149–2159.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 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, Jiaqian Ren, Xu Bai, and Yanan Cao. 2025. Dual-path counterfactual integration for multimodal aspect-based sentiment classification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22759–22769.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Saif M Mohammad. 2025. Nrc vad lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms. *arXiv e-prints*, pages arXiv–2503.

Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, and Alice Oh. 2021. Dimensional emotion detection from categorical emotion. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 4367–4380.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Qianlong Wang, Hongling Xu, Zhiyuan Wen, Bin Liang, Min Yang, Bing Qin, and Ruifeng Xu. 2023. Image-to-text conversion and aspect-oriented filtration for multimodal aspect-based sentiment analysis. *IEEE Transactions on Affective Computing*.

Luwei Xiao, Xingjiao Wu, Shuwen Yang, Junjie Xu, Jie Zhou, and Liang He. 2023. Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis. *Information Processing & Management*, 60(6):103508.

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

Hao Yang, Yanyan Zhao, and Bing Qin. 2022. Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In *EMNLP 2022*, pages 3324–3335.

Juan Yang, Mengya Xu, Yali Xiao, and Xu Du. 2024. Amifn: Aspect-guided multi-view interactions and fusion network for multimodal aspect-based sentiment analysis. *Neurocomputing*, 573:127222.

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.

Jianfei Yu, Kai Chen, and Rui Xia. 2022a. Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. *IEEE Transactions on Affective Computing*.

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

Jianfei Yu, Jing Jiang, and Rui Xia. 2019. 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, Jieming Wang, Rui Xia, and Junjie Li. 2022b. Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching. In *IJCAI 2022*, pages 4482–4488.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016. Building Chinese affective resources in valence-arousal dimensions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540–545, San Diego, California. Association for Computational Linguistics.

Yice Zhang, Hongling Xu, Delong Zhang, and Ruifeng Xu. 2024. A hybrid approach to dimensional aspect-based sentiment analysis using bert and large language models. *Electronics*, 13(18).

Zhe Zhang, Zhu Wang, Xiaona Li, Nannan Liu, Bin Guo, and Zhiwen Yu. 2021. Modalnet: an aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network. *World Wide Web*, 24:1957–1974.

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 *ACL 2023 Findings*, pages 8184–8196.

Senbin Zhu, Hanjie Zhao, Xingren Wang, Shanhong Liu, Yuxiang Jia, and Hongying Zan. 2024. ZZU-NLP at SIGHAN-2024 dimABSA task: Aspect-based sentiment analysis with coarse-to-fine in-context learning. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 112–120, Bangkok, Thailand. Association for Computational Linguistics.

Wang Zou, Xia Sun, Wenhuan Wu, Qiang Lu, Xiaodi Zhao, Qirong Bo, and Jianqiang Yan. 2025. Tcmt: Target-oriented cross modal transformer for multimodal aspect-based sentiment analysis. *Expert Systems with Applications*, 264:125818.

A Senti-COCO Construction Details

In this section, we provide detailed information about the construction of **Senti-COCO**, our affect-enriched image–text dataset used for training the affect-aware CLIP encoder.

A.1 Data Source

We build Senti-COCO based on the MSCOCO dataset (Chen et al., 2015), where each image is associated with multiple human-written captions. For each image I, we collect its caption set $\mathcal{C} = \{C_1, \ldots, C_m\}$ as the semantic basis for subsequent rewriting.

A.2 Affect-Enriched Caption Generation

To introduce affective information, we employ a multimodal large language model, Qwen2.5-VL-7B (Bai et al., 2025), to rewrite the original captions. Given an image-caption set pair $(I, \mathcal{C})$, where $\mathcal{C} = \{C_1, \ldots, C_m\}$, we construct a prompt that explicitly instructs the model to integrate the information from multiple captions while enhancing affective expression. The prompt is formulated as follows:

> *Here are several different descriptions of the image: {captions}. Please use them as reference to generate a single, more emotional caption for the image. Add suitable adjectives or verbs, make it concise and less than 20 words.*

This prompt encourages the model to (i) preserve the core visual content described in the original captions, (ii) enrich the expression with stronger and more nuanced affective cues, and (iii) maintain conciseness for stable training. The model then generates a rewritten caption $\tilde{C}$, forming an affect-enriched pair $(I, \tilde{C})$ for downstream use.

Metric	Train	Validation
Dataset Size		
# Samples	118,348	5,000
MSCOCO Raw Captions		
# Captions	592,058	25,014
Avg. Captions / Image	5.00	5.00
Caption Length (avg)	52.34	52.28
Senti-COCO Captions		
# Emotional Captions	118,348	5,000
Caption Length (avg)	103.56	103.54

Table 4: Statistics of MSCOCO captions and the constructed Senti-COCO dataset. Caption length is reported in characters.

A.3 Dataset Statistics

After filtering, we obtain a total of **123K** high-quality affect-enriched image–text pairs, forming the final Senti-COCO dataset. Compared to the original MSCOCO captions, Senti-COCO provides richer affective descriptions, making it more suitable for learning emotion-sensitive visual representations. Detailed statistics are summarized in Table 4.

A.4 Discussion

By combining large-scale caption rewriting with human verification, Senti-COCO balances scalability and quality. While automatic generation enables efficient data expansion, the manual checking process reduces noise such as content drift or exaggerated affect, thereby improving the reliability of the dataset for affect-aware representation learning.