A Novel Energy Based Model Mechanism for Multi-Modal Aspect-Based Sentiment Analysis

Tianshuo Peng[1,†], Zuchao Li[1,†,*], Ping Wang[3,4], Lefei Zhang[1,2], and Hai Zhao[5]

[1]School of Computer Science, Wuhan University, Wuhan, 430072, China,
[2]Hubei Luojia Laboratory, Wuhan 430072, P. R. China,
[3]Center for the Studies of Information Resources, Wuhan University, Wuhan 430072, China
[4]School of Information Management, Wuhan University, Wuhan 430072, China
[5]Department of Computer Science and Engineering, Shanghai Jiao Tong University
{pengts,zcli-charlie,wangping,zhanglefei}@whu.edu.cn, zhaohai@cs.sjtu.edu.cn

Abstract

Multi-modal aspect-based sentiment analysis (MABSA) has recently attracted increasing attention. The span-based extraction methods, such as FSUIE, demonstrate strong performance in sentiment analysis due to their joint modeling of input sequences and target labels. However, previous methods still have certain limitations: (i) They ignore the difference in the focus of visual information between different analysis targets (aspect or sentiment). (ii) Combining features from uni-modal encoders directly may not be sufficient to eliminate the modal gap and can cause difficulties in capturing the image-text pairwise relevance. (iii) Existing span-based methods for MABSA ignore the pairwise relevance of target span boundaries. To tackle these limitations, we propose a novel framework called DQPSA. Specifically, our model contains a Prompt as Dual Query (PDQ) module that uses the prompt as both a visual query and a language query to extract prompt-aware visual information and strengthen the pairwise relevance between visual information and the analysis target. Additionally, we introduce an Energy-based Pairwise Expert (EPE) module that models the boundaries pairing of the analysis target from the perspective of an Energy-based Model. This expert predicts aspect or sentiment span based on pairwise stability. Experiments on three widely used benchmarks demonstrate that DQPSA outperforms previous approaches and achieves a new state-of-the-art performance. The code will be released at https://github.com/pengts/DQPSA.

* Corresponding author. † Equal contribution. This work was supported by the National Natural Science Foundation of China (No. 62306216), the Natural Science Foundation of Hubei Province of China (No. 2023AFB816), the Fundamental Research Funds for the Central Universities (No. 2042023kf0133), the Special Fund of Hubei Luojia Laboratory (No. 220100014), National Natural Science Foundation of China [No. 72074171] [No. 72374161].

Introduction

As one of the most important tasks that examines a model's semantic comprehension and sentiment perception, Multi-modal Aspect-Based Sentiment Analysis (MABSA) is an challenging and fine-grained task in the Sentiment Analysis field and has attracted increasing attention. The MABSA task consists of three main tasks: given an image-text pair, Multi-modal Aspect Term Extraction (MATE) focuses on extracting all aspect terms with sentiment polarity in the sentence (Zhao et al. 2022; Lu et al. 2018; Wu et al. 2020a), Multi-modal Aspect-oriented Sentiment Classification (MASC) aims to determine the sentiment polarity of each given aspect (Xu, Mao, and Chen 2019; Yu and Jiang 2019; Yu, Jiang, and Xia 2020), Joint Multi-modal Aspect-Sentiment Analysis (JMASA), on the other hand, requires the model to extract aspect-sentiment pairs jointly (Ju et al. 2021; Ling, Yu, and Xia 2022; Zhou et al. 2023). Among all the methods of previous work, the span-based extraction methods, such as FSUIE (Peng et al. 2023), demonstrate strong performance in sentiment analysis due to their joint modeling of input sequences and target labels. Besides, it avoids complex structures for sequence labelling or sequence generation with a more concise structure.

In this scenario of fine-grained MABSA task, three main challenges are worth emphasizing: First, image contains a large amount of semantic information, and there is a difference in the focus of visual information between different analysis targets. Take figure 1 as an example: (1) The focus of visual information is different between MATE and MASC tasks: the MATE task should focus on all the potential entities across the image while MASC concentrates on the details of specific aspect which is fine-grained. (2) In the MASC task, different image regions may imply different sentimental tendencies, resulting the difference of focus among aspects. Previous work only focused on the general information in image features, while the visual information representing positive emotions, such as the person ponder-

ing with his chin resting on their hand, and the visual infor mation representing negative emotions, such as the perso with his head lowered would influence each other, introduc ing a significant amount of misleading information into thei sentiment analysis.

Second, most of the aforementioned studies focus on ex tracting modal features using pre-trained unimodal model and fusing them directly. However, separate training of im age feature and text feature ignores the semantic alignmen and modal gap between text and image, leading to the dif ficulty of the model in capturing the pairwise relevance be tween image and text. Therefore, it is crucial to design spe- cific structures to mitigate modal gap and strengthen image- text pairwise relevance

Besides, existing span-based models (Peng et al. 2023) consider independently the possibility of certain position as a start or end boundary while ignoring the pairwise relevance between span boundaries, i.e., the a priori knowledge that "the boundaries of spans should be semantically related".

Based on the challenges above, we proposed the DQPSA framework for Multi-modal Aspect-Based Sentiment Analysis, which unifies the MABSA tasks under one framework. Specifically, inspired by BLIP-2 (Li et al. 2023) that trains an adapter to filter features from visual encoder, we designed the *Prompt as Dual Query* module to address the issue that different analysis targets pay different attention to visual information and strengthen the pairwise relevance between image and text. *Prompt as Dual Query* uses prompt as both visual query and language query. The visual query interacts with image features from frozen pre-trained image encoder in alternating self-attention layers and cross-attention layers, from which the prompt-aware visual information with the highest semantic relevance to the analysis target is extracted. The language query will act as one of the input of text encoder, guiding model to output prediction based on the current analysis target. Considering the pairwise relevance between target span boundaries, we introduce the idea of Energy Based Model (LeCun et al. 2006) to give better span scores and proposed the novel *Energy based Pairwise Expert* that predicts span based on pairing stability. Experiments on three widely used benchmarks verify that DQPSA outperforms previous approaches and achieves the state-of-the-art performance. A large amount of complementary experiments and ablation studies further demonstrate the effectiveness of components we proposed.

In summary, our contributions are as follows:

• We proposed *Prompt as Dual Query* module that satisfy diverse focus of different analysis targets on visual information, and strengthen the pairwise relevance between visual information and analysis target.

• Inspired by the Energy Based Model (EBM) that quantifying compatibility between variables using an energy scalar, we proposed a novel *Energy based Pairwise Expert* that models the boundaries pairwise stability of target span.

• Experiments on three widely used benchmarks Twitter2015, Twitter2017 and Political Twitter shows that DQPSA outperforms previous approaches and achieves SOTA performance. DQPSA also significantly outperforms Multi-modal Large Language Model (LLM) VisualGLM-6B and Uni-modal LLM ChatGPT-3.5 under fair comparison.

Text: *Chess World Championship: contemplating Liren Ding and frustrated Nebo just now decided the winner!*

Aspect	*contemplating Liren Ding*	*frustrated Nebo*
Sentiment	*Positive*	*Negative*

Figure 1: Example of the variability in the focus of different analysis tasks

Related Work

Multi-modal Aspect-Based Sentiment Analysis

With the proliferation of multi-modal data disseminated on social media, images are considered to be an important complementary information for sentiment analysis. Thus, MABSA began to receive increasing attention.

Wang (2022) injects knowledge-aware information through multi-modal retrieval. Cai (2019) treats image attribute features as supplementary modalities to bridge the gap between texts and images. Ju (2021) first proposes JMASA task that jointly extract aspects and corresponding sentiment polarity to better satisfy the practical applications, Ling (2022) performs various vision-language pre-training tasks to capture crossmodal alignment, and Zhou (2023) designs an aspect-aware attention module to select textual tokens and image blocks that are semantically related to the aspects. The above approaches for JMASA either treat MATE and MASC as sequence labelling and binary classification, or require an additional decoding module for sequence generating, and do not emphasize the variability of image focus across analysis targets. Unlike previous works, our proposed DQPSA address MATE and MASC under a unified framework as span recognition, dispensing with the complex sequence generation structure. Meanwhile, we design the *Prompt as Dual Query* module that uses prompt as both visual query and language query, in order to differentially extract prompt-aware visual information and strengthen image-text pairwise relevance.

Energy Based Model

The concept of Energy Based Model was first proposed by (LeCun et al. 2006). The core idea is to establish a mapping between different variable configurations and a scalar energy that will be able to measure compatibility, thus capturing the dependence between different variable configurations. The target of learning is to find an efficient energy function that maps correct variable configurations to low energy values while mapping incorrect variable configurations to high energy values. The goal of inference is to find variable configurations that minimize the energy. The loss function selected during training can be used to measure the effectiveness of the energy function. Zhou (2021) introduce the concept of EBM to text adversarial domain adaptation

and take text sentiment classification as one of benchmarks. However, due to the unreleased code and data and significant differences in dataset and task, it's hard for us to make a fair comparison with it.

Inspired by the idea of Energy Based Model, we quantify boundary pairing stability of potential span with scalar energies and design the *Energy based Pairwise Expert* to predict span based on pairwise stability. To the best of our knowledge, this is the first attempt that adapting Energy Based Model into MABSA task.

Method

In this section, we first introduce the general framework of our proposed DQPSA. Then we introduce the specific structure of *Prompt as Dual Query*, followed by a detailed description of *Energy based Pairwise Expert* . Figure 2 demonstrates the framework of our proposed DQPSA, which consists of four main components: a frozen image encoder, a *Prompt as Dual Query* module, a text encoder and an *Energy based Pairwise Expert*. To strengthen the correlation between visual information and analysis target, we design the *Prompt as Dual Query* that allows prompt to interacting with both image and text. Besides, in order to consider the boundary positions of the target span more comprehensively, we design the *Energy based Pairwise Expert* to extract span considering both start and end boundaries simultaneously. In the following subsections, we will illustrate the details of our proposed model.

Prompt as Dual Query (PDQ)

We propose the *Prompt as Dual Query* (PDQ) module, which leverage prompt as both visual query and language query, guiding model to focus on different perspectives of visual information and text information according to concrete analysis targets.

Figure 3 shows the specific structure of PDQ, which is mainly composed of two kinds of blocks stacked alternately: cross-attention block and self-attention block. we assume that the image feature obtained from the frozen image encoder is $F_{\mathrm{I}} = \{I_0, I_1, I_2 \cdots I_{L_{\mathrm{I}}}\}$. For each image, we constructed a description for its content, the features of the constructed description and the prompt are $F_{\mathrm{D}}^i = \left\{I_0^i, I_1^i, I_2^i \cdots I_{L_{\mathrm{D}}}^i\right\}$ and $F_{\mathrm{P}}^i = \left\{I_0^i, I_1^i, I_2^i \cdots I_{L_{\mathrm{P}}}^i\right\}$, where L_{I}, L_{D}, L_{P} are the lengths of the corresponding sequences. $i \in [\text{-}1, \mathrm{N}]$ represents the index of the block in which F_{D} and F_{P} are located and $i = -1$ represents the original word embedding. The PDQ receives three types of text sequences belonging to $S = \{F_{\mathrm{D}}^{-1}, F_{\mathrm{P}}^{-1}, [F_{\mathrm{P}}^{-1} : F_{\mathrm{D}}^{-1}]\}$ as inputs, with only sequences belonging to $\hat{S} = \{F_{\mathrm{P}}^{-1}, [F_{\mathrm{P}}^{-1} : F_{\mathrm{D}}^{-1}]\}$ interacting with image features in cross-attention layer.

The basic formula for attention can be expressed as:

$$\textsc{Attn}(Q, K, V) = \mathrm{Softmax}\left(\frac{Q\,\mathrm{K}^\top}{\sqrt{\mathrm{d_k}}}\right)V. \tag{1}$$

Suppose the hidden state H^{i-1} of the previous layer, for input $\in S$, the result of self-attention layer can be represented as follows, where $W_{QS}, W_{KS}, W_{VS}, W_{QC}, W_{KC}, W_{VC}$ are the parameters to be optimized::

$$\begin{aligned}&\textsc{Self-Attn}(H^{i-1})\\ &= \textsc{Attn}(W_{\mathrm{QS}}H^{i-1}, W_{\mathrm{KS}}H^{i-1}, W_{\mathrm{VS}}H^{i-1}).\end{aligned} \tag{2}$$

As for input $\in \hat{S}$, it will go through an additional cross-attention layer in cross-attention block that can be represented as:

$$\begin{aligned}&\textsc{I2T-Attn}(H^{i-1})\\ &= \textsc{Attn}(W_{\mathrm{QC}}H^{i-1}[:L_{\mathrm{P}}], W_{\mathrm{KC}}F_{\mathrm{I}}, W_{\mathrm{VC}}F_{\mathrm{I}}),\\ &\textsc{Cross-Attn}(H^{i-1})\\ &= \textsc{cat}[\textsc{I2T-Attn}(H^{i-1}) : H^{i-1}[L_{\mathrm{P}} :]].\end{aligned} \tag{3}$$

where $H^{i-1}[: L_{\mathrm{P}}]$ represents the sub-sequence of H^{i-1} up to the L_{P}-th token and $H^{i-1}[L_{\mathrm{P}} :]$ represents the sub-sequence of H^{i-1} from and include the L_{P}-th token. In the inference process, we use F_{P}^{-1} as input. Through multiple fusion with F_{I}, prompt guides the model to filter the prompt-aware visual information that semantically related to the analysis target.

To further strength the pairwise relevance between visual information and analysis targets, we introduce image-text matching task and in-batch image-text contrastive learning, the specific algorithmic process is as follows:

For image-text matching task, we have

$$\begin{aligned}&\textsc{Loss}_{\mathrm{ITM}} = -\frac{1}{2}\sum_{i=0}^{1} p_i log(q_i),\\ &p = \textsc{mean}((W_{\mathrm{ITM}}[F_{\mathrm{VQ}}^{\mathrm{N}} : F_{\mathrm{D}}^{\mathrm{N}}])[: L_{\mathrm{VQ}}]).\end{aligned} \tag{4}$$

where the hidden state in the last block of PDQ will go though an linear projection and we select the average of tokens corresponding to visual query as model prediction p. q is the label that identifies whether the image and description match or not. W_{ITM} is the parameter to be optimized.

For in-batch image-text comparative learning, we firstly use the first token of $F_{\mathrm{VQ}}^{\mathrm{N}}$ and $F_{\mathrm{D}}^{\mathrm{N}}$ as [CLS] token to construct the similarity matrix between different $F_{\mathrm{VQ}}^{\mathrm{N}}$ and $F_{\mathrm{D}}^{\mathrm{N}}$ within the same batch:

$$\begin{aligned}I^{\mathrm{ITC}} &= [I_1^{\mathrm{ITC}}, I_2^{\mathrm{ITC}} \cdots I_{\mathrm{B}}^{\mathrm{ITC}}],\\ T^{\mathrm{ITC}} &= [T_1^{\mathrm{ITC}}, T_2^{\mathrm{ITC}} \cdots T_{\mathrm{B}}^{\mathrm{ITC}}].\end{aligned} \tag{5}$$

where B is the batch size, I_i^{ITC} and T_i^{ITC} is the [CLS] token of the i-th $F_{\mathrm{VQ}}^{\mathrm{N}}$ and $F_{\mathrm{D}}^{\mathrm{N}}$ in batch. And then we construct the similarity vector p_{i2d} p_{d2i}, and corresponding label q_{i2d} q_{d2i}, for the i-th $F_{\mathrm{VQ}}^{\mathrm{N}}$ and $F_{\mathrm{D}}^{\mathrm{N}}$ in the batch:

$$\begin{aligned}&p^{\mathrm{i2d}} = {T^{\mathrm{ITC}}}^\top I_i^{\mathrm{ITC}}, p^{\mathrm{d2i}} = {I^{\mathrm{ITC}}}^\top T_i^{\mathrm{ITC}},\\ &q^{\mathrm{i2d}}, q^{\mathrm{d2i}} \in R^{\mathrm{B}\times 1}, q_j^{\mathrm{i2d}}, q_j^{d2i} = 1\ if\ j == i\ else\ 0.\end{aligned} \tag{6}$$

finally, the $\textsc{Loss}_{\mathrm{ITC}}$ can be represented as:

$$\textsc{Loss}_{\mathrm{ITC}} = -\frac{1}{\mathrm{B}}(\sum_{j=1}^{\mathrm{B}} p_j^{\mathrm{i2d}} log(q_i^{\mathrm{i2d}}) + \sum_{j=1}^{\mathrm{B}} p_j^{\mathrm{d2i}} log(q_i^{\mathrm{d2i}})). \tag{7}$$

Optimizing $\textsc{Loss}_{\mathrm{ITM}}$ and $\textsc{Loss}_{\mathrm{ITC}}$ enables model to capture the pairwise relevance of image-text pairs thus obtain the capability of prompt-aware visual information extraction.

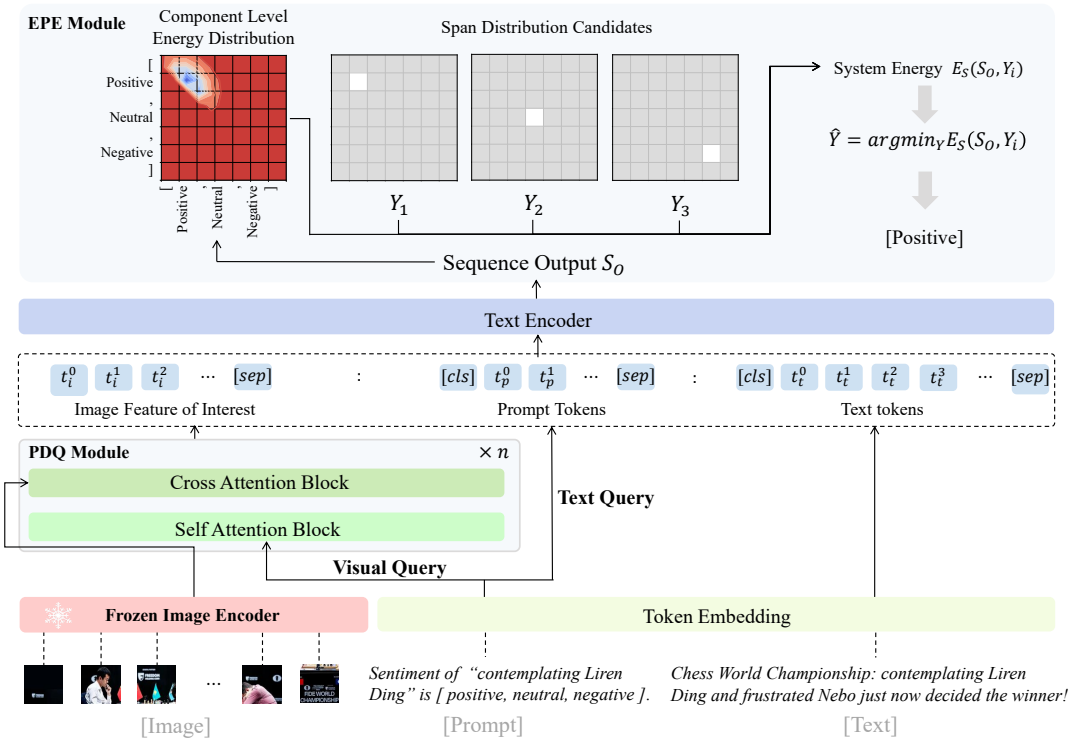


Figure 2: The overview of our proposed DQPSA.

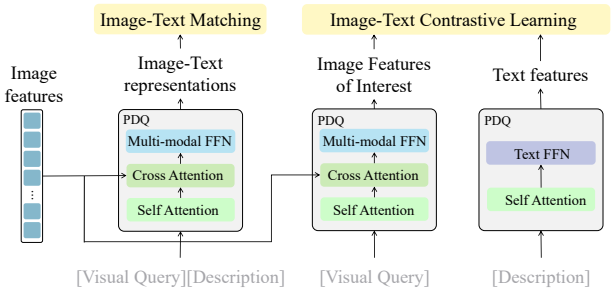


Figure 3: Demonstration of Prompt as Dual Query

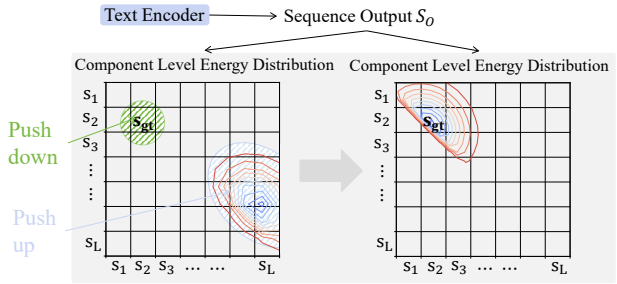


Figure 4: Demonstration of Energy based Pairwise Expert

Energy Based Pairwise Expert (EPE)

To capture the semantical relevance of start and end boundary within a span, we introduce the idea of Energy Based Model (LeCun et al. 2006) to give better span scores and proposed the novel *Energy based Pairwise Expert* that predicts span based on pairwise stability.

EBM captures the correlation between variables by creating a mapping of variable combinations to a scalar energy. In the inference phase, the model aims to find the combination of variables that minimizes the energy of the system, in the training phase, the model is forced to find an energy function that assigns lower energy values to paired combinations of variables while assigning higher energy values to unpaired combinations of variables.

The core theory of EBM is learning an energy function that assigns lower energy values to paired combinations of variables while assigning higher energy values to unpaired combinations of variables.

Following the theory of EBM, we try to map the boundary pairing relations of a target span to a scalar energy, which in turn quantitatively describe the intensity of the pairing relations at the boundary of a given span. Figure 4 demonstrates the learning process of EPE, in which we lower the energy of positions with pairwise relevance while raising the energy of others.Specifically, we design the *Component-level energy function* E_C and *System-level energy function* E_S. Let the output sequence of the text encoder be $S_O = \{s_1, s_2 \cdots s_L\}$, L is the length of S_O. We use E_C to denote the pairing energy of the span whose boundaries are the i-th and j-th to-

ken, denoted as

$$E_{\mathrm{C}}(W_{\mathrm{S}}, W_{\mathrm{E}}, s_i, s_j) = -(s_i^{\top} W_{\mathrm{S}_i}^{\top}(W_{\mathrm{E}_j} s_j)). \quad (8)$$

where W_{S} and W_{E} are the parameters to be optimized. The smaller E_{C} represents the more stable pairing relation, i.e., the higher the probability that $S_{\mathrm{O}}[i:j+1]$ is the target span.

For the distribution of predicted span over the entire sequence, we use E_{S} to measure the energy of the entire system. Let the span distribution matrix $Y \in R^{\mathrm{L}\times\mathrm{L}}$ be the label of the target span, where $y_{ij} \in \{0,1\}$ indicates whether $S_{\mathrm{O}}[i:j+1]$ is the target span or not. The system level energy can be expressed as

$$E_{\mathrm{S}}(S_{\mathrm{O}}, Y) = \sum_{i=0}^{\mathrm{L}} \sum_{j=i}^{\mathrm{L}} (E_{\mathrm{C}}(s_i, s_j) y_{ij}). \quad (9)$$

In the inference phase, we select the span distribution matrix that minimizes the energy of the system. In the training phase, we construct the loss function directly using E_{C} as

$$\begin{aligned} \textsc{Loss}_{\mathrm{EPE}} = -\frac{1}{\frac{\mathrm{L}(\mathrm{L}+1)}{2}} \sum_{i=0}^{\mathrm{L}} \sum_{j=i}^{\mathrm{L}} \Big(& y_{ij} log(x_{ij}) \\ & + (1-y_{ij}) log(1-x_{ij}) \Big), \\ x_{ij} = \mathrm{Sigmoid}(-E_{\mathrm{C}}(s_i, s_j)). & \end{aligned} \quad (10)$$

by optimizing $\textsc{Loss}_{\mathrm{EPE}}$, the model will learn the energy function E_{C} that allocates lower energy to the paired systems (S_{O}, Y) while allocating higher energy to the unpaired ones.

Based on all the above narratives, the final training loss of the model is represented as

$$\textsc{Loss} = \lambda_1 \textsc{Loss}_{\mathrm{ITM}} + \lambda_2 \textsc{Loss}_{\mathrm{ITC}} + \lambda_3 \textsc{Loss}_{\mathrm{EPE}}. \quad (11)$$

where $\lambda_1, \lambda_2, \lambda_3$ are predefined hyper-parameters

Experiments

Experimental Settings

Dataset Following previous works, we adopt two widely used benchmarks: Twitter2015 and Twitter2017 (Yu and Jiang 2019) to evaluate our proposed DQPSA. Besides, we employ another Political Twitter dataset[1] from (Yang et al. 2021) for JMASA task. In pre-training stage, we use COCO2017 dataset and ImageNet dataset.

Implementation Details In our proposed model, we choose CLIP-ViT-bigG-14-laion2B-39B-b160k (Radford et al. 2021) as the frozen image encoder, FSUIE-base (Peng et al. 2023) as the text encoder, *Prompt as Dual Query* module is based on BERT-base (Devlin et al. 2019) architecture and pre-training parameter and randomly initialized cross-attention layers. Refer to Appendix A for detailed information of our backbone and selections of hyper-parameters.

We first employ a two-stage pre-training with 5 epochs for each stage. Then We trained model for 50 epochs with an AdamW optimizer on the datasets of each task, and selected the final model based on the performance on the development set. Associated code and pre-trained models will be made publicly available upon receipt.

Evaluation Metrics Following previous work, we evaluate the performance of our model on the MATE and JMASA tasks with Precision (P), Recall (R) and Micro-F1 (F1) score, while on MASC task, we report Accuracy (Acc) and Micro-F1 (F1) score for comparison.

Pre-training In order to equip the PDQ with initial capability of prompt-controlled image comprehending, we employed a two-stage pre-training before adapting to the specific MABSA task. For ImageNet data, we train model to predict the entity class contained in the image under the guidance of prompt, which helps the model to capture the word level pairwise relevance between images and entities. For COCO data, the model is trained to predict the descriptions that are relevant to the content of image under the guidance of prompt, from which the model will learn the sentence level pairwise relevance between image and text. See Appendix B for specific constructs of prompt, description and text.

During phase 1, to prevent the initialized PDQ from detracting the semantic comprehension of text encoder, we freeze all parameters except PDQ and EPE. While for subsequent training, we train all model parameters except image encoder.

Main Results

Results of JMASA Task Table 1 and 2 show the results of JMASA task. It can be seen that, by introducing the *Prompt as Dual Query* module and the *Energy based Pairwise Expert* module, DQPSA significantly outperforms the sub-optimal models (3.3 on Twitter2015 and 0.9 on Twitter2017) and achieves SOTA results. This demonstrates the effectiveness of differentially leveraging visual information according to different analytical targets and focusing on the pairwise relevance of the target span.

Compared to the text-based models, the multi-modal models perform better in general and DQPSA far exceeds all text-based models. This verifies that image modal introduced by our method does provide important supplementary information for sentiment analysis.

Compare to the span based methods, EPE improves the span boundary recognition process. Instead of separately predicting the start and end boundaries of span, EPE capture the pairwise relevance of span boundaries, which provides a better understanding of the distribution of target spans. Unlike directly using visual features as a prefix to textual inputs, PDQ module filters out visual noise and selects visual information beneficial to the current task. It also bridge the modality gap between visual and textual features through stacked attention operation. Furthermore, our method focuses on task-specific visual information rather than token-specific visual information, allowing for a more macro-level utilization of visual information. Compare to the span based methods, EPE improves the span boundary recognition process. Instead of separately predicting the start

[1]Political Twitter for evaluaton is consistence with dataset released by (Yang, Na, and Yu 2022)

	Methods	Twitter2015			Twitter2017		
		P	R	F1	P	R	F1
Text-based	SPAN (Hu et al. 2019)	53.7	53.9	53.8	59.6	61.7	60.6
	D-GCN (Chen, Tian, and Song 2020)	58.3	58.8	59.4	64.2	64.1	64.1
	BART (Yan et al. 2021)	62.9	65.0	63.9	65.2	65.6	65.4
Multi-modal	UMT+TomBERT (Yu et al. 2020; Yu and Jiang 2019)	58.4	61.3	59.8	62.3	62.4	62.4
	OSCGA+TomBERT (Wu et al. 2020b; Yu and Jiang 2019)	61.7	63.4	62.5	63.4	64.0	63.7
	OSCGA-collapse (Wu et al. 2020b)	63.1	63.7	63.2	63.5	63.5	63.5
	RpBERT-collapse (Sun et al. 2021)	49.3	46.9	48.0	57.0	55.4	56.2
	UMT-collapse (Yu et al. 2020)	61.0	60.4	61.6	60.8	60.0	61.7
	JML (Ju et al. 2021)	65.0	63.2	64.1	66.5	65.5	66.0
	VLP-MABSA (Ling, Yu, and Xia 2022)	65.1	68.3	66.6	66.9	69.2	68.0
	CMMT (Yang, Na, and Yu 2022)	64.6	68.7	66.5	67.6	69.4	68.5
	AoM (Zhou et al. 2023)	67.9	69.3	68.6	68.4	**71.0**	69.7
	DQPSA (ours)	**71.7**	**72.0**	**71.9**	**71.1**	70.2	**70.6**

Table 1: Results of Twitter2015 and Twitter2017, JMASA task. The best results are bold-typed.

Methods	Political-Twitter		
	P	R	F1
RoBERTa (Liu et al. 2019)	63.1	62.1	62.6
UMT+collapse (Yu et al. 2020)	54.9	54.7	54.8
JML (Ju et al. 2021)	63.6	59.4	61.4
UMT-RoBERTa (Yu et al. 2020; Liu et al. 2019)	63.8	63.4	63.6
JML-PoBERTa (Ju et al. 2021; Liu et al. 2019)	63.0	60.2	61.6
CMMT (Yang, Na, and Yu 2022)	65.3	**65.7**	65.5
DQPSA (ours)	**68.3**	65.5	**66.9**

Table 2: Results of Political Twitter, JMASA task. The best results are bold-typed.

Methods	Twitter2015			Twitter2017		
	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	90.2	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
DQPSA (ours)	**88.3**	87.1	**87.7**	**95.1**	93.5	**94.3**

Table 3: Results of Twitter2015 and Twitter2017, MATE task. The best results are bold-typed.

Methods	Twitter2015		Twitter2017	
	ACC	F1	ACC	F1
ESAFN (Yu, Jiang, and Xia 2020)	73.4	67.4	67.8	64.2
TomBERT (Yu and Jiang 2019)	77.2	71.8	70.5	68.0
CapTrBERT (Khan and Fu 2021)	78.0	73.2	72.3	70.2
JML (Ju et al. 2021)	78.7	-	72.7	-
VLP-MABSA (Ling, Yu, and Xia 2022)	78.6	73.8	73.8	71.8
CMMT (Yang, Na, and Yu 2022)	77.9	-	73.8	-
AoM (Zhou et al. 2023)	80.2	75.9	**76.4**	**75.0**
DQPSA (ours)	**81.1**	**81.1**	75.0	**75.0**

Table 4: Results of different methods for MASC.

and end boundaries of span, EPE capture the pairwise relevance of span boundaries, which provides a better understanding of the distribution of target spans. Unlike directly using visual features as a prefix to textual inputs, PDQ module filters out visual noise and selects visual information beneficial to the current task. It also bridge the modality gap between visual and textual features through stacked attention operation. Furthermore, our method focuses on task-specific visual information rather than token-specific visual information, allowing for a more macro-level utilization of visual information.

In contrast to approaches that use collapse labels or work with pipelines using different models, our approach uses a unified framework for the MASC and MATE tasks, completely eliminating the formal differences between two tasks, helping the model to learn the interactions between the two sub-tasks in terms of semantic information and sentiment with a more concise structure, resulting in better performance. Compared with methods that focus on token-various visual information, our approach focuses on target-various visual information to filter and leverage image features from a more macroscopic perspective, achieving better performance while reducing computational effort.

Results of MATE and MASC Task Table 3 and 4 show the results of the MATE and MASC task. Consistent with the results of JMASA task, DQPSA also achieves a significant improvement or competitive performance in the two sub-tasks performance.

Specifically, DQPSA outperforms the sub-optimal method in the MATE task by 1.5 on Twitter2015 and 2.0 on Twitter2017, which suggests that our method helps model to focus on the image information related to the aspects and filter out irrelevant information.

As for MASC, we note that the model made a huge improvement on Twitter2015 (5.2 over previous SOTA), but only achieved competitive results on Twitter2017. On the one hand, we think it is because the sentiment analysis data in Twitter2017 is more challenging for it contains a significant number of unresolvable and unidentifiable symbols, including emojis commonly used on Twitter, which posed a relative hard challenge for DQPSA, on the other hand, model with large scale may better capture the correlation between aspect and sentiment in difficult cases. However,

Methods	Twitter2015			Twitter2017		
	MATE	MASC	JMASA	MATE	MASC	JMASA
DQPSA	87.7	81.1	71.9	94.3	75.0	70.6
w/o EPE	86.3	80.9	69.1	92.5	73.1	66.8
w/o PDQ	87.4	78.4	69.9	93.8	69.7	65.6
w/o PDQ&EPE	84.4	76.4	63.3	90.8	68.6	64.1

Table 5: Results of Ablation Study

considering the difference in the number of trainable parameters, the results of MASC are still convincing in proving the effectiveness of our approach.

Ablation Study

To further investigate the contribution of each component to model performance improvement, we conducted ablation studies on Twitter2015 and Twitter2017 for MATE, MASC, and JMASA tasks, and the results of F1 scores are illustrated in table 5. To examine the effect of *Energy based Pairwise Expert* module, we follow the original setting of text encoder to predict positions of start and end boundaries independently. To examine the effect of *Prompt as Dual Query* module, we replace the visual query with a set of optimizable tokens.

It can be seen that removing both the PDQ and the EPE detracts somewhat from the model's performance. Specifically, the model without PDQ achieves competitive performance on the MATE task, but performance on the MASC and JMASA tasks drops significantly. This is because the prompt we constructed for the MASC task is different among potential aspects, and optimizable tokens are not sufficient to capture this target-variability. However, the prompt used by MATE is relatively fixed, so optimizable tokens can fit the target requirements of MATE to a certain extent and acts as soft prompt. The JMASA task, as a prolongation of the two subtasks, naturally shows a decrease in performance due to decrease on MATE and MASC. And this verifies that our *Prompt as Dual Query* can satisfy the differential focus on visual information of different targets in the MABSA topic.

For model without EPE, we can see that the model shows a significant performance degradation on all three tasks. This is due to the fact that EPE, as a module for the model to make span decisions, has an auxiliary effect on all the tasks performed by the model. By introducing EPE, the model does not consider the boundaries of the target span in isolation, but makes full use of the pairwise relevance between the boundaries of the spans, which further enriches the knowledge learnt by the model, and thus improves the effect on multiple tasks.

Considering the whole table, the introduction of PDQ and EPE alone can significantly improve the model performance, in which EPE leads to comprehensive performance improvement by capture the span pairwise relevance, while PDQ focuses more on guiding the model to filter visual information according to different target requirements thus is more effective in enhancing fine-grained MASC tasks. Simultaneous application of the two will further boost the model performance.

Methods	14lap	14res	15res	16res
FSUIE-base	65.6	74.1	70.6	75.8
PSA (ours)	69.8	78.5	78.3	79.2

Table 6: Results on ASTE-DATA-V2 datasets (14lap, 14res, 15res, and 16res)

Methods	Twitter2015			Twitter2017		
	MATE	MASC	JMASA	MATE	MASC	JMASA
DQPSA	87.7	81.1	71.9	94.3	75.0	70.6
PSA	80.8	77.7	62.6	91.2	68.6	61.6

Table 7: Results of models w & w/o image modal

Case Study

For in-depth analysis, we also trained a PSA model based on FSUIE-base without using *Prompt as Dual Query* module and only receiving text as input. We apply PSA model to different case study as follows:

EPE on Sentiment Analysis of Plain Text To verify the robustness of *Energy based Pairwise Expert*, we apply PSA to the ASTE-Data-V2 (Xu et al. 2020) of Aspect Sentiment Triplet Extraction (ASTE) task, and compare it with the best result of the existing work FSUIE-base (Peng et al. 2023). Table 6 shows the F1 scores of PSA on the ASTE task, it shows that EPE delivers a huge performance improvement on all the four datasets and achieves the up-to-date SOTA results. This demonstrates the extensibility and robustness of our proposed EPE, as well as its strong performance under span recognition especially in cases involving sentiment analysis.

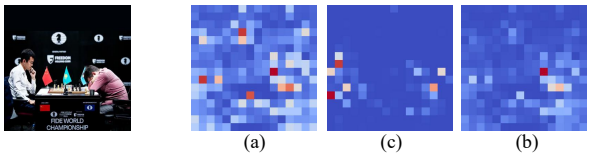
W/o Information from Image Modal To investigate whether our approach helps the model utilize visual information more efficiently, we apply PSA comparing to DQPSA on Twitter2015 and Twitter2017 to examine the effects of the introduction of image modal on the model performance. Table 7 reports F1 scores for PSA and DQPSA on Twitter2015 and Twitter2017 for the three tasks.

It can be seen from table 7 that: with *Prompt as Dual Query* module , our DQPSA significantly outperformed PSA that without image modal on a variety of tasks. This verifies that our proposed method is efficient in helping models to exploit the rich information from images.

Sentiment Analysis Compared to LLMs In order to verify whether our model has an advantage over large models on the MABSA task, we compare the model's performance with that of common Large Language Models (LLMs) include VisualGLM-6B and ChatGPT-3.5. Since LLMs are not designed for identifying aspects in text and it is more difficult to unify the output structure, we only test them on the MASC task for a fair comparison. Table 8 shows the results of DQPSA and other LLMs on the MASC task. The results show that our DQOSA using fewer parameters obtains significantly better performance than LLMs, which also verifies the superiority and effectiveness of our proposed DQPSA framework. It should be noted that due to the lack of a visual

Models	Twitter2015			Twitter2017		
	P	R	F1	P	R	F1
DQPSA (ours)	81.1	81.1	81.1	75.0	75.0	75.0
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

Table 8: Results of comparison with LLMs on MASC task



(a) "Extract the aspects in the sentence"
(b) "Sentiment of "contemplating Liren Ding (left)" is [positive, neutral, negative]"
(c) "Sentiment of "frustrated Nebo (right)" is [positive, neutral, negative]"

Figure 5: Visualization of Prompt as Dual Query

module, ChatGPT 3.5 only receives text as input. Therefore, despite its impressive generation capability, the performance of ChatGPT3.5 is not very outstanding. This also verifies that there is a great potential and importance for the integration and utilization of information from different modalities in multi-modal tasks.

Visualization

To further validate the effectiveness of our proposed method, we visualize the attention matrices in the last cross-attention layer to verify if *Prompt as Dual Query* module guide model to focus on different visual information based on various analysis targets. Figure 5 presents the visualization results along with the corresponding prompts. It can be observed that for the MATE task, model exhibits a relatively uniform distribution of attention over the image, indicating that model analyzes potential aspects by incorporating a larger receptive field. However, when analyzing different aspects of sentiment, model demonstrates distinct focus towards different regions of visual information. This result once again confirms that our proposed PDQ effectively captures the variations in the focus of visual information across different analysis targets.

Conclusion

In this paper, we proposed a novel framework, named DQPSA, for Multi-modal Aspect-Based Sentiment Analysis (MABSA). We use a well-designed *Prompt as Dual Query* module that leveraging prompt as both visual query and language query to extract the prompt-aware visual information, thus satisfying various focus of different analysis targets on the visual information. Besides, we capture the boundaries pairing of analysis target with the perspective of Energy based Model and predict span based on pairwise stability with an *Energy base Pairwise Expert* module. Performance on three widely used benchmark datasets verifies that our method outperforms previous methods.

Appendix

A. Backbone Inrtoduction&Hyper-parameters Selection

Specifically, the frozen image encoder contains 48 layers of 104-head Transformer layers with hidden size of 1664, the *Prompt as Dual Query* module and the text encoder have 12 layers of 12-head Transformer layers and a hidden size of 768. Hyper-parameters used during training are shown in table 9.

Training Stage	λ_1	λ_2	λ_3	learing rate
Pretraing-Stage1	2.0	2.0	1.0	5e-5
Pretraing-Stage2	1.0	1.0	1.0	3e-5
Finetuning	0.1	0.1	1.0	2e-5

Table 9: Hyper-parameters used during training

B. Data Construction for Pre-training

Prompt	"Provide a description for image."
Description	the label description of image
Text	label description, Irrelevant description1, Irrelevant description2, Irrelevant description3.

Table 10: COCO pre-training data construction

COCO Dataset For COCO data, each image has corresponding five descriptions. We randomly select one relevant description and three irrelevant descriptions to splice as the input text, the model needs to predict the descriptions that are relevant to the content of the picture under the guidance of prompt. Training data is constructed as table 10.

Prompt	"What does this image contains."
Description	"It's an image of {image label}."
Text	Correct label, fake label1, fake label2 $\cdots$ fake label9.

Table 11: ImageNet pre-training data construction

ImageNet Dataset For ImageNet data, which contains 1000 classes of images and corresponding labels, we select 100 of these classes to construct the pre-training data, and divide them into 10 groups equally. For the same group of data, the input text is constructed as a splice of ten class names, and the model needs to predict the entity class contained in the current image under the guidance of prompt. Training data is constructed as table 11.

References

Cai, Y.; Cai, H.; and Wan, X. 2019. Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. In

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2506–2515. Florence, Italy: Association for Computational Linguistics.

Chen, G.; Tian, Y.; and Song, Y. 2020. Joint Aspect Extraction and Sentiment Analysis with Directional Graph Convolutional Networks. In *Proceedings of the 28th International Conference on Computational Linguistics*, 272–279. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Hu, M.; Peng, Y.; Huang, Z.; Li, D.; and Lv, Y. 2019. Open-Domain Targeted Sentiment Analysis via Span-Based Extraction and Classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 537–546. Florence, Italy: Association for Computational Linguistics.

Ju, X.; Zhang, D.; Xiao, R.; Li, J.; Li, S.; Zhang, M.; and Zhou, G. 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*, 4395–4405. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Khan, Z.; and Fu, Y. 2021. Exploiting BERT for Multimodal Target Sentiment Classification through Input Space Translation. In Shen, H. T.; Zhuang, Y.; Smith, J. R.; Yang, Y.; César, P.; Metze, F.; and Prabhakaran, B., eds., *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, 3034–3042. ACM.

LeCun, Y.; Chopra, S.; Hadsell, R.; Ranzato, M.; and Huang, F. 2006. A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *CoRR*, abs/2301.12597.

Ling, Y.; Yu, J.; and Xia, R. 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)*, 2149–2159. Dublin, Ireland: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Lu, D.; Neves, L.; Carvalho, V.; Zhang, N.; and Ji, H. 2018. Visual Attention Model for Name Tagging in Multimodal Social Media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1990–1999. Melbourne, Australia: Association for Computational Linguistics.

Peng, T.; Li, Z.; Zhang, L.; Du, B.; and Zhao, H. 2023. FSUIE: A Novel Fuzzy Span Mechanism for Universal Information Extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16318–16333. Toronto, Canada: Association for Computational Linguistics.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Sun, L.; Wang, J.; Zhang, K.; Su, Y.; and Weng, F. 2021. RpBERT: A Text-image Relation Propagation-based BERT Model for Multimodal NER. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 13860–13868. AAAI Press.

Wang, X.; Cai, J.; Jiang, Y.; Xie, P.; Tu, K.; and Lu, W. 2022. Named Entity and Relation Extraction with Multi-Modal Retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 5925–5936. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Wu, H.; Cheng, S.; Wang, J.; Li, S.; and Chi, L. 2020a. Multimodal Aspect Extraction with Region-Aware Alignment Network. In Zhu, X.; Zhang, M.; Hong, Y.; and He, R., eds., *Natural Language Processing and Chinese Computing - 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14-18, 2020, Proceedings, Part I*, volume 12430 of *Lecture Notes in Computer Science*, 145–156. Springer.

Wu, Z.; Zheng, C.; Cai, Y.; Chen, J.; Leung, H.; and Li, Q. 2020b. Multimodal Representation with Embedded Visual Guiding Objects for Named Entity Recognition in Social Media Posts. In Chen, C. W.; Cucchiara, R.; Hua, X.; Qi, G.; Ricci, E.; Zhang, Z.; and Zimmermann, R., eds., *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, 1038–1046. ACM.

Xu, L.; Li, H.; Lu, W.; and Bing, L. 2020. Position-Aware Tagging for Aspect Sentiment Triplet Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2339–2349. Online: Association for Computational Linguistics.

Xu, N.; Mao, W.; and Chen, G. 2019. Multi-Interactive Memory Network for Aspect Based Multimodal Sentiment Analysis. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The*

Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, 371–378. AAAI Press.

Yan, H.; Dai, J.; Ji, T.; Qiu, X.; and Zhang, Z. 2021. A Unified Generative Framework for Aspect-based Sentiment Analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2416–2429. Online: Association for Computational Linguistics.

Yang, L.; Na, J.; and Yu, J. 2022. Cross-Modal Multitask Transformer for End-to-End Multimodal Aspect-Based Sentiment Analysis. *Inf. Process. Manag.*, 59(5): 103038.

Yang, L.; Yu, J.; Zhang, C.; and Na, J. 2021. Fine-Grained Sentiment Analysis of Political Tweets with Entity-Aware Multimodal Network. In Toeppe, K.; Yan, H.; and Chu, S. K., eds., *Diversity, Divergence, Dialogue - 16th International Conference, iConference 2021, Beijing, China, March 17-31, 2021, Proceedings, Part I*, volume 12645 of *Lecture Notes in Computer Science*, 411–420. Springer.

Yu, J.; and Jiang, J. 2019. Adapting BERT for Target-Oriented Multimodal Sentiment Classification. In Kraus, S., ed., *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, 5408–5414. ijcai.org.

Yu, J.; Jiang, J.; and Xia, R. 2020. Entity-Sensitive Attention and Fusion Network for Entity-Level Multimodal Sentiment Classification. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28: 429–439.

Yu, J.; Jiang, J.; Yang, L.; and Xia, R. 2020. 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*, 3342–3352. Online: Association for Computational Linguistics.

Zhao, G.; Dong, G.; Shi, Y.; Yan, H.; Xu, W.; and Li, S. 2022. Entity-level Interaction via Heterogeneous Graph for Multimodal Named Entity Recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 6345–6350. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Zhou, R.; Guo, W.; Liu, X.; Yu, S.; Zhang, Y.; and Yuan, X. 2023. AoM: Detecting Aspect-oriented Information for Multimodal Aspect-Based Sentiment Analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, 8184–8196. Toronto, Canada: Association for Computational Linguistics.

Zou, H.; Yang, J.; and Wu, X. 2021. Unsupervised Energy-based Adversarial Domain Adaptation for Cross-domain Text Classification. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1208–1218. Online: Association for Computational Linguistics.