Few-shot Joint Multimodal Aspect-Sentiment Analysis Based on Generative Multimodal Prompt

Xiaocui Yang[1,2], Shi Feng[1], Daling Wang[1], Qi Sun[2,3], Wenfang Wu[1], Yifei Zhang[1], Pengfei Hong[2], Soujanya Poria[2]

[1]Northeastern University, [2]Singapore University of Technology and Design,
[3]Nanjing University of Science and Technology

{yangxiaocui, wenfang}@stumail.neu.edu.cn,
{fengshi, wangdaling, zhangyifei}@cse.neu.edu.cn,
{pengfei_hong, sporia}@sutd.edu.sg, 319106003718@njust.edu.cn

Abstract

We have witnessed the rapid proliferation of multimodal data on numerous social media platforms. Conventional studies typically require massive labeled data to train models for Multimodal Aspect-Based Sentiment Analysis (MABSA). However, collecting and annotating fine-grained multimodal data for MABSA is tough. To alleviate the above issue, we perform three MABSA-related tasks with quite a small number of labeled multimodal samples. We first build diverse and comprehensive multimodal few-shot datasets according to the data distribution. To capture the specific prompt for each aspect term in a few-shot scenario, we propose a novel Generative Multimodal Prompt (GMP)[1] model for MABSA, which includes the Multimodal Encoder module and the N-Stream Decoders module. We further introduce a subtask to predict the number of aspect terms in each instance to construct the multimodal prompt. Extensive experiments on two datasets demonstrate that our approach outperforms strong baselines on two MABSA-related tasks in the few-shot setting.

1 Introduction

The Multimodal Aspect-Based Sentiment Analysis (MABSA) task has garnered significant attention in recent times, as evidenced by several recent studies (Chandrasekaran et al., 2021; Zhang et al., 2022a; Zhu et al., 2022; Gandhi et al., 2023). In the literature, MABSA is typically divided into three subtasks: Multimodal Aspect Term Extraction (MATE), Multimodal Aspect-oriented Sentiment Classification (MASC), and Joint Multimodal Aspect-Sentiment Analysis (JMASA) (Wu et al., 2020a; Zhang et al., 2021a; Yu and Jiang, 2019; Khan and Fu, 2021; Ju et al., 2021; Ling et al., 2022). Given a text-image pair, MATE aims to extract all the aspect terms mentioned in the text, MASC focuses on detecting the sentiment

corresponding to each extracted aspect term, and JMASA is designed to extract aspect terms and their corresponding sentiments jointly. Previous studies on Multimodal Aspect-Based Sentiment Analysis (MABSA) primarily focus on leveraging extensive training data (full training datasets), with some works resorting to additional data to improve performance (Ju et al., 2021; Ling et al., 2022). However, collecting and annotating such massive multimodal data for MABSA is time-intensive and laborious (Zhou et al., 2021). Moreover, in real-world applications, only a limited amount of labeled data is commonly available. To address this challenge, PVLM (Yu and Zhang, 2022) and UP-MPF (Yu et al., 2022) introduce prompt-based learning into Multimodal Aspect-oriented Sentiment Classification (MASC) in a few-shot scenario. Based on limited sentiment categories (three categories), PVLM and UP-MPF convert MASC to masked language modeling (MLM) tasks. However, the prerequisite of MASC is that the aspect terms are known, which requires aspect term extraction in advance, typically performed by Multimodal Aspect Term Extraction (MATE) or Joint Multimodal Aspect-Sentiment Analysis (JMASA). Both JMASA and MATE tasks are challenging due to the unknown and varying number of aspect items in each sample, as well as the distinct content of each aspect. Therefore, applying MLM in the few-shot setting is unsuitable for JMASA and MATE tasks, as depicted in Fig. 1. This paper addresses the challenges of JMASA, MASC, and MATE in a text-image few-shot setting, and to the best of our knowledge, there are no dedicated studies dealing with JMASA and MATE tasks in the multimodal few-shot scenario.

Prior few-shot text classification tasks with limited classification labels have manually designed general prompts for the entire dataset to mine knowledge from pre-trained language models (PLM) (Shin et al., 2020; Hosseini-Asl et al., 2022;

---

[1]https://github.com/YangXiaocui1215/GMP.

Zhang et al., 2022b). However, in the case of Joint Multimodal Aspect-Sentiment Analysis (JMASA) and Multimodal Aspect Term Extraction (MATE), where the content of each aspect term is unknown and assorted, manual prompts are infeasible for aspect extraction. To address this challenge, we propose a novel Generative Multimodal Prompt (GMP) model for few-shot Multimodal Aspect-Based Sentiment Analysis (MABSA), which includes the Multimodal Encoder (ME) module and the N-Stream Decoders (NSD) module. It is crucial to sample diverse and comprehensive data to build practical few-shot datasets in the multimodal few-shot setting. We construct few-shot training and development datasets by sampling data with combinations of different sentiments in instances, according to the data distribution, as shown in Table 1. Since the number of aspect terms in JMASA and MATE is unknown and vital, we leverage the Multimodal Encoder (ME) and Aspect-Num Decoder (AND) to predict the number of aspect terms as a subtask. The clues required for each aspect of an instance may vary. We generate aspect-oriented prompts for each aspect (aspect-level) using the ME and Aspect-oriented Prompt Decoder (APD). Similarly, we use the ME and Sentiment-oriented Prompt Decoder (SPD) to generate sentiment-oriented prompts. As the sentiment categories in all datasets are limited, we only reserve the instance-level sentiment prompts. The caption of the image modality is also captured as the image prompt. Lastly, specific multimodal prompts for different tasks are constructed based on the image caption, the predicted number of aspect terms, aspect prompts, and sentiment prompts. We feed the multimodal embedding with the multimodal prompt into the Multimodal Encoder-Decoder based BART model (Lewis et al., 2020) to generate triplet sequences. Our main contributions are summarized as follows:

- We propose a novel Generative Multimodal Prompt (GMP) model to handle Joint Multimodal Aspect-Sentiment Analysis (JMASA), Multimodal Aspect Sentiment Classification (MASC), and Multimodal Aspect Term Extraction (MATE) in the multimodal few-shot setting. To our knowledge, we are the first to focus on JMASA and MATE tasks in a multimodal few-shot scenario.

- To tackle the challenge of unknown number of multimodal aspect terms and construct effective multimodal prompts, we employ multitasking and build the few-shot dataset by taking into account the distribution of sentiment categories for each dataset.

- We conduct extensive experiments on the constructed few-shot datasets, and our results demonstrate that our proposed model outperforms strong baselines on JMASA and MASC in the few-shot setting.

2 Related Work

2.1 Multimodal Aspect Sentiment Analysis

In contrast to coarse-grained sentiment analysis (sentence-level) (Yang et al., 2021b; Li et al., 2022), MABSA requires not only extracting aspect terms, but also recognizing the corresponding sentiment associated with each aspect. Early research focuses on different subtasks, including Multimodal Aspect Term Extraction (MATE) (Sun et al., 2020; Yu et al., 2020; Wu et al., 2020b; Zhang et al., 2021b; Chen et al., 2022) and Multimodal Aspect Sentiment Classification (MASC) (Yang et al., 2021a; Yu and Jiang, 2019; Khan and Fu, 2021). More recently, Ju et al. (Ju et al., 2021) propose Joint Multimodal Aspect-Sentiment Analysis (JMASA), which jointly performs aspect term extraction and sentiment classification. Yang et al. (Yang et al., 2022b) introduce Cross-Modal Multitask Transformer (CMMT) for MABSA. VLP (Ling et al., 2022) further extends this by resorting to additional pre-training data and designing multiple pre-training tasks to enhance JMASA performance. However, few works specifically address MABSA in the few-shot scenario. Although VLP has conducted low-resource experiments, it includes over 17,000 pre-training data and utilizes the full development dataset, which violates our starting point of adopting few-shot data.

2.2 Few-shot Learning with Pre-trained Language Model

Prompt-based language modeling is applied to solve different few-shot tasks with PLM in Natural Language Process (NLP) due to its powerful representation (Liu et al., 2021), such as text classification (Shin et al., 2020; Hosseini-Asl et al., 2022), text regression (Gao et al., 2021), and text generation (Li and Liang, 2021). Existing works introduce Multimodal Prompt-based Fine-tuning (MPF) methods into multimodal settings by MLM, like Frozen (Tsimpoukelli et al., 2021), PVLM (Yu and Zhang, 2022), and UP-MPF (Yu et al., 2022).
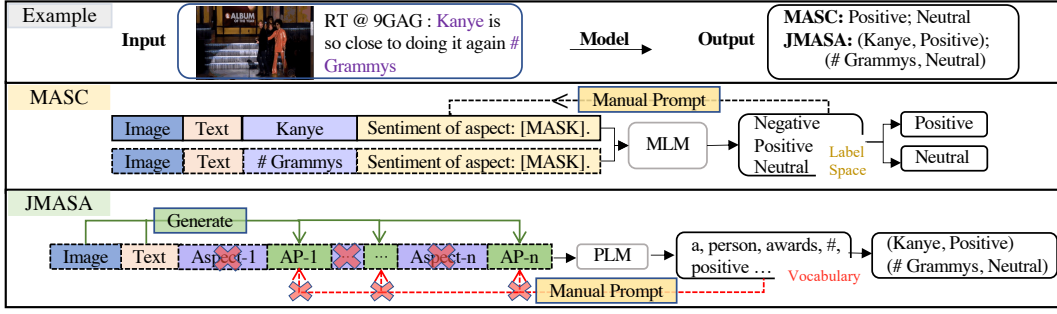
Figure 1: Examples of prompts for MASC and JMASA in a few-shot setting are shown. For MASC, the prompts consist of two triplets for the MABSA task, with aspect terms 'Kanye' and '#Grammy' shown in the upper part. The manual general prompt for the entire dataset can be efficient when the label space is limited, such as 'Negative', 'Positive', and 'Neutral', as shown in the middle part with the yellow box representing the manual prompt for few-shot MASC in UP-MPF (Yu et al., 2022). However, the manual prompt is infeasible for few-shot JMASA, as each aspect term is unknown and could be any phrase from the vocabulary, as shown by the red crosses. Hence, we generate a prompt for each aspect based on the multimodal context, as shown in the green box, where 'AP' represents Aspect-oriented Prompt and 'n' represents the number of aspect terms for an instance. 'Vocabulary' refers to the vocabulary of the Pre-trained Language Model (PLM).

Different from few-shot MASC (PLVM and UP-MPF), we simultaneously extract aspect terms and perform sentiment detection for each aspect in the multimodal few-shot scenario.

3 Our Proposed Model

In Joint Multimodal Aspect-Sentiment Analysis (JMASA), our goal is to extract aspect terms and classify sentiment corresponding to each aspect. However, due to the varying number of aspect terms in each instance and each diverse aspect term, a different prompt is needed for each aspect in the few-shot setting. To address this, we propose a Generative Multimodal Prompt (GMP) for few-shot JMASA, as illustrated in Fig. 2. Leveraging BART, we generate aspect-oriented prompts for each aspect based on the multimodal context, as well as instance-level sentiment-oriented prompts.

3.1 Task Formulation

In this paper, we assume access to a pre-trained language model $\mathcal{M}$, such as BART, that we wish to fine-tune for the aspect-sentiment sequence generation task using labeled data. For the few-shot multimodal training dataset $\mathcal{D}_{train}$, we select K training examples based on sentiment categories for each dataset, resulting in $\mathcal{D}train = (T^j, I^j, A^j, S^j, O^j)j = 1^K$, where $T = [t^1, t^2, ..., t^{l_t}]$ is the text modality with l_t as the text length; I is the image modality; $A = [a^1, ..., a^n]$ is the aspect list; $S = [s^1, ..., s^n]$ is the sentiment list corresponding to A; and $O = [(x_b^1, x_e^1, s^1), ..., (x_b^n, x_e^n, s^n)]$ is our output, which represents the index-sentiment list, e.g.,

$O = [(5, 5, POS), (13, 14, NEU)]$ for the instance in Fig. 3. Here, n denotes the number of aspects, x_b^k and x_e^k represent the beginning and end indices of the k_{th} aspect term, and $s^k \in \{POS, NEG, NEU\}$ denotes the sentiment label. For $\mathcal{D}_{dev}$, we select the same size of data as the few-shot training dataset, i.e., $|\mathcal{D}_{dev}| = |\mathcal{D}_{train}|$. Our task is to generate O in the few-shot multimodal setting. Following the formulation in (Yan et al., 2021; Ling et al., 2022), we define the outputs of the three subtasks as follows[2]:

- **JMASA**: $O = [(x_1^b, x_1^e, s_1), ..., (x_n^b, x_n^e, s_n)]$.

- **MASC**: $O = [(\underline{x_1^b}, \underline{x_1^e}, s_1), ..., (\underline{x_n^b}, \underline{x_n^e}, s_n)]$.

- **MATE**: $O = [(x_1^b, x_1^e), ..., (x_n^b, x_n^e)]$.

3.2 Generative Multimodal Prompt

GMP consists of two main modules: the Multimodal Encoder module and the N-Stream Decoders module.

3.2.1 Multimodal Encoder

In this section, we design the multimodal encoder to capture multimodal representations. We start by extracting image representations using NF-ResNet (Brock et al., 2021), and then project them to the text modality space for the image modality, I.

$$V = Reshape(W_i ResNet(I) + b_i)$$
$$= [v^1, ..., v^k, ..., v^{l_i}], v^k \in \mathbb{R}^{d_t}, \quad (1)$$

where V is reshaped image representation, $W_i \in \mathbb{R}^{d_v \times d_{nt}}$, $b_i \in \mathbb{R}^{d_{nt}}$, and $nt = l_i \times d_t$. l_i, which is

---

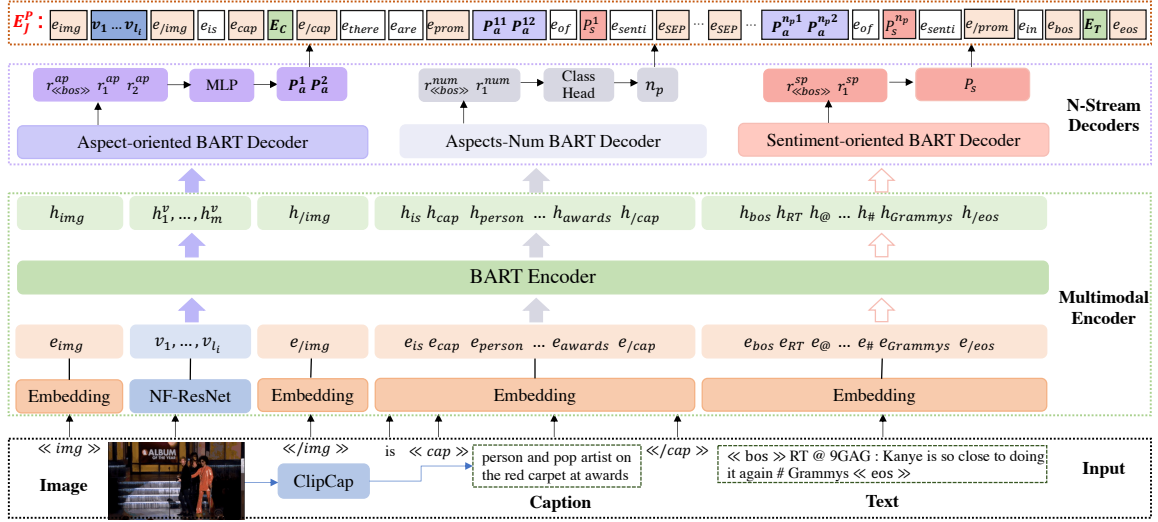[2]The underlined tokens are provided during inference.

Figure 2: The framework of our proposed Generative Multimodal Prompt (GMP) for Few-shot MABSA consists of two main modules: the Multimodal Encoder module (the green dashed box) and the N-Stream Decoders module (the purple dashed box). For JMASA, we apply the multimodal embedding with the generative multimodal prompt E_J^P using the Multimodal Encoder module. Similarly, for MASC and MATE, we design separate multimodal embeddings E_S^P and E_A^P respectively, as shown in Figure 5. The solid thick arrows (purple and gray) indicate the sharing of parameters between two multimodal encoders, while the thick hollow arrow (pink) does not share parameters with others. Special tokens such as img, $/img$, cap, $/cap$, bos, eos, $prom$, and $/prom$ are used in the multimodal embeddings. The embedding of "sentiment" is denoted as E_{senti}.

a hyperparameter, is the number of image slots that reserve initial image representation, and d_t represents the dimension of text embedding in BART.

Since BART is a pre-trained language model (PLM) that does not involve pre-training on image modality, we aim to alleviate the discrepancy issue of image representation in PLM. To achieve this, we further capture the image caption using ClipCap (Mokady et al., 2021), denoted as C, which can be regarded as the image prompt.

$$C = ClipCap(I). \quad (2)$$

We utilize the BART model to obtain text embeddings for both the text input and the image caption.

$$E_T = Embedding(T), E_T \in \mathbb{R}^{l_t \times d_t},$$
$$E_C = Embedding(C), E_C \in \mathbb{R}^{l_{cap} \times d_t}, \quad (3)$$

where l_{cap} is the length of image caption. The multimodal embedding E_M can be obtained, $E_M = [E_{img}, V, E_{/img}, E_{is}, E_{cap}, E_C, E_{/cap}, E_{bos}, E_T, E_{eos}]$.

Finally, we feed E_M into the BART Encoder to obtain the multimodal representation. We argue that subsequent decoders require specific information, so we leverage different multimodal BART Encoders for this purpose.

$$H_M^a = MBART_E^a(E_M), H_M^a \in \mathbb{R}^{l_m \times d},$$
$$H_M^s = MBART_E^s(E_M), H_M^s \in \mathbb{R}^{l_m \times d}, \quad (4)$$

where $l_m = l_i + l_{cap} + l_t + l_s$, l_s is the length of special tokens, and d is the hidden dimension.

3.2.2 N-Stream Decoders

IIn this section, we utilize the encoded multimodal representation from Eq. 4 to predict the number of aspect terms and generate aspect-oriented and sentiment-oriented prompts using different decoders for each instance. The 'N' in 'N-Stream' varies depending on the task, with values of 3, 2, and 1 for JMASA, MATE, and MASC, respectively.

Aspect-Num Decoder (AND). In the JMASA task, the number of aspects in each instance is significant but unknown, so we predict the number of aspects based on the multimodal context using the Aspect-Num BART Decoder as a subtask. Specifically, we input the multimodal encoder output H_M^a and the special token bos into the Aspect-Num Decoder, which then predicts the number of aspects $n_p \in \mathbb{R}^5$ as follows[3]:

$$h_n^{and} = AND(H_M^a; E_{bos}),$$
$$n_p = Softmax(MLP(h_n^{and})). \quad (5)$$

---

[3]Twitter 2017 dataset contains only 3 instances with more than 5 aspects. Therefore, we set "aspect-num" as 5 in the AND module to accommodate the maximum number of aspect terms in an instance.
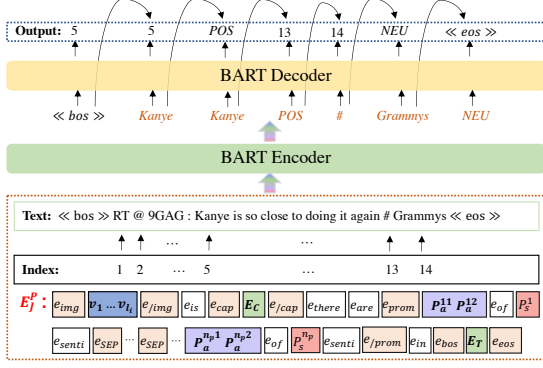
Figure 3: An example of triplet sequence generation for the JMASA task.

We leverage the cross-entropy loss for the subtask,

$$\mathcal{L}_c = -\sum_{j=1}^{K} n_g^j log(n_p^j), \qquad (6)$$

where n_g^j represents the label for the number of aspect terms. It's worth noting that in the MASC task, the gold number of aspect terms is provided to the model, and thus, this subtask is not required for MASC.

$$P_a^k = MLP^k([h_1^{apd}, h_2^{apd}]), \qquad (7)$$

where k is the k_{th} group of aspect of an instance, $P_a^k \in \mathbb{R}^{2 \times d}$. The generative aspect-oriented prompt $AP = [P_a^1, ..., P_a^{n_p}] \in \mathbb{R}^{2n_p \times d}$.

Aspect-oriented Prompt Decoder (APD). Prompts for few-shot multimodal classification tasks can be manually designed for specific datasets due to limited categories, as demonstrated in PVLM (Yu and Zhang, 2022) and UP-MPF (Yu et al., 2022). However, each text-image pair carries different context information, and the aspects of the text are diverse. Therefore, in the few-shot setting, we need to capture various cues for each aspect. Inspired by this, we design our model to generate aspect-oriented prompts based on the multimodal context. Specifically, we first generate an instance-level prompt based on the encoded multimodal representation. The final output of the JMASA task is a triplet sequence, where the first two positions of each triplet represent the beginning and ending indices for each aspect term. We set two aspect slots for each generated aspect-oriented prompt, resulting in an instance-level prompt length of $2n_p$. The decoder takes the encoder outputs H_m^a and previous decoder outputs $h^{apd} < (lap - 1)$ as inputs to compute the current

hidden state.

$$h_{l_{ap}}^{apd} = APD(H_M^a; (h_{<(l_{ap}-1)}^{apd})), \qquad (8)$$

where we feed the bos into APD as the beginning token and $l_{ap} = 2$.

Sentiment-oriented Prompt Decoder (SPD). The sentiment corresponding to each aspect is related to each instance. Similar to APD, we generate the sentiment-oriented prompt based on multimodal context. For JMASA, the last position in each triplet of the output sequence predicts the sentiment. We set one sentiment slot for each generated sentiment-oriented prompt, i.e., the length of the instance-level prompt is n_p.

$$P_s = h_1^{spd} = SPD(H_M^s; E_{bos}), \qquad (9)$$

where we feed the bos into SPD as the beginning token. As the sentiment categories are limited, they share a common label space. Therefore, we do not generate corresponding sentiment cues for each aspect. Instead, P_s is repeated n_p times to form the generative sentiment-oriented prompt $SP = [P_s^1, ..., P_s^{n_p}] \in \mathbb{R}^{n_p \times d}$, where d represents the dimensionality of the prompt.

3.3 Multimodal Embedding with Prompt

We construct the multimodal prompt for different tasks, including JMASA, MASC, and MATE, based on the text-image pair, aspect-oriented prompts, sentiment-oriented prompts, and prediction of the number of aspect terms. For JMASA, we design multimodal embedding with a generative multimodal prompt, denoted as E_J^P, as shown in Fig. 2. Similar to E_J^P, we separately design multimodal embedding with prompt for MASC and MATE, e.g, E_S^P and E_A^P as Fig. 5 shows in Appendix A.

3.4 Triplet Sequence Generation

We next feed the multimodal embedding with prompt into the Encoder-Decoder model to generate the triplet sequence. We take the JMASA task as an example, as Fig. 3 shows.

$$H_J^P = MBART_E^J(E_J^P), H_J^P \in \mathbb{R}^{l_J \times d}, \qquad (10)$$

where l_J is the length of E_J^P. Then, we use the BART decoder to get the last hidden state,

$$h_t^{dJ} = BART_D^J(H_J^P; \hat{O}_{<t}), \qquad (11)$$

where t is the t_{th} step and $\hat{O}_{<t}$ is the output of the previous t steps. Following (Yan et al., 2021), we

predict the token probability distribution P_t with $h_t^{dJ} \in \mathbb{R}^d$, as follows:

$$P_t = Predict([E_T; E_S]h_t^{dJ}), \qquad (12)$$

where $P_t \in \mathbb{R}^{l_t+l_c}$; E_S is the embedding of the sentiment label set, and its length is $l_c = 3$.

We employ cross-entropy loss for our sequence generation task.

$$\mathcal{L}_g = -\sum_{j=1}^{K} O^j log(P^j). \qquad (13)$$

3.5 Multitask Training

We optimize our main task and subtask.

$$\mathcal{L} = \mathcal{L}_g + \lambda \mathcal{L}_c, \qquad (14)$$

where λ is the hyperparameter to control the contribution of each task.

4 Experiments

We conduct experiments on two groups of few-shot multimodal datasets built according to the distribution of sentiment categories from Twitter-15 (15) and Twitter-17 (17) (Zhang et al., 2018; Lu et al., 2018). We compare our model with numerous approaches on three tasks, including Multimodal Aspect Term Extraction (MATE), Multimodal Aspect-oriented Sentiment Classification (MASC), and Joint Multimodal Aspect-Sentiment Analysis (JMASA).

4.1 Few-shot Datasets

To construct few-shot datasets for few-shot Multimodal Aspect-Based Sentiment Analysis (MABSA), it is important to select a few diverse samples that provide comprehensive coverage of the different sentiment categories. We sample data based on the distribution of sentiment categories in instances to create few-shot datasets. The statistics of the different datasets are presented in Table 1. For each dataset, we randomly sample three groups of few-shot training and development datasets based on three different seeds, such as [42, 87, 100], and each split is run 3 times. We report the average performance and standard deviation over 9 (3 × 3) times of training for a more robust evaluation.

4.2 Implementation Details

We utilize BART-Base with 140M parameters as our Pretrained Language Model (PLM), denoted as $\mathcal{M}$, and NF-ResNet-50 as our visual encoder. The number of epochs is set to 70, and the batch size is set to 4 for all tasks. The learning rates (lr) are set to 6.5e-5 for JMASA and MATE tasks, and for the MASC task, we set lr to 8e-5 and 7.5e-5 for Twitter-15 and Twitter-17, respectively. All models are implemented using PyTorch and the experiments are run on an A6000 GPU. Following (Ling et al., 2022), we evaluate our model on three subtasks of MABSA and use Micro-F1 score (F1), Precision (P), and Recall (R) as the evaluation metrics to measure the performance. For MASC, we also use Accuracy (Acc) to compare fairly with other approaches. GMP has 169.3M/155.6M/154.9M parameters for JMASA/MATE/MASC, respectively, and during training, all parameters are updated. The training time for GMP up to 70 epochs is 50/50/25 minutes for JMASA/MATE/MASC.

4.3 Baselines

To ensure a comprehensive comparison, we thoroughly evaluate our model against various approaches across different tasks.

Models for Joint Multimodal Aspect-Sentiment Analysis (JMASA). We first apply text-based approaches to perform Joint Aspect-Sentiment Analysis (JASA) with the following models: **BART** (Yan et al., 2021) adapts JASA to an Encoder-Decoder model. **D-GCN** (Chen et al., 2020) proposes directional graph convolutional networks for JASA. **SpanABSA** (Hu et al., 2019) applies an extraction-then-classification framework using a span-based labeling scheme. Next, we accomplish JMASA and MATE using multimodal approaches with the following models: **JML** (Ju et al., 2021) performs JMASA by introducing auxiliary cross-modal relation detection. **CMMT** (Yang et al., 2022b) proposes a multi-task learning framework that leverages two unimodal auxiliary tasks. **VLP** (Ling et al., 2022), which designs multiple Vision-Language pre-training tasks, is the state-of-the-art (SOTA) model for JMASA. However, since VLP introduces additional 17,000+ pre-training data, which violates our motivation to use few-shot data, we also present results for **NVLP**, which does not perform the pre-training task.

Models for Multimodal Aspect Sentiment Classification (MASC). We reproduce multimodal ap-

Datasets	POS	NEU	NEG	{POS, NEU}	{NEG, NEU}	{POS, NEG}	{POS, NEU, NEG}	All
15-Train	**32**/526	**64**/1084	**16**/214	**16**/178	**8**/79	**2**/13	**0**/7	**138**/2,101
15-Dev	**32**/162	**64**/375	**16**/71	**16**/69	**8**/44	**2**/6	**0**/0	**138**/727
15-Test	167	335	68	73	28	2	1	674
17-Train	**32**/534	**32**/328	**16**/150	**32**/535	**16**/153	**2**/26	**2**/20	**132**/1,746
17-Dev	**32**/177	**32**/109	**16**/49	**32**/180	**16**/50	**2**/7	**2**/5	**132**/577
17-Test	178	107	39	171	70	8	14	587

Table 1: Statistics on two datasets. POS: Positive, NEU: Neutral, NEG: Negative. For A/B, B represents the number of original data, and A represents the number of few-shot data. {Senti-1, Senti-2} means that both Senti-1 and Senti-2 simultaneously exist in the instance, and there can be more than one of each sentiment. For both datasets, the percentage of the constructed few-shot dataset accounts for about 7% of the overall training data.

Modality	Model	Twitter-15			Twitter-17		
		P	**R**	**F1**	**P**	**R**	**F1**
Text	BART	47.03 (±2.00)	41.90 (±3.80)	44.28 (±2.91)	48.59 (±1.90)	44.97 (±1.95)	46.70 (±1.81)
	D-GCN	42.02 (±2.71)	40.07 (±2.03)	40.95 (±2.18)	45.66 (±1.09)	45.81 (±1.41)	44.89 (±1.58)
	SpanABSA	48.52 (±0.84)	39.80 (±2.19)	43.71 (±1.60)	51.67 (±1.53)	48.44 (±0.75)	49.98 (±0.67)
Text -Image	JML	48.51 (±1.14)	41.59 (±2.56)	44.77 (±1.97)	50.13 (±0.41)	48.65 (±0.10)	49.38 (±0.25)
	CMMT	29.85 (±1.37)	36.23 (±2.05)	32.65 (±0.07)	39.64 (±0.51)	47.83 (±2.14)	43.34 (±1.12)
	NVLP	46.04 (±0.82)	42.40 (±0.25)	44.14 (±0.47)	50.66 (±2.09)	45.92 (±1.10)	48.16 (±1.28)
	VLP	46.56 (±0.94)	**49.08** (±1.64)	47.77 (±0.73)	51.32 (±0.19)	52.22 (±0.52)	51.76 (±0.21)
	GMP	**51.67** (±2.01)	47.19 (±1.46)	**49.33** (±1.71)	**54.28** (±1.08)	**53.31** (±1.71)	**53.79** (±1.31)

Table 2: Average results of different models in terms of **Precision** (P), **Recall** (R), and **F1** for MABSA.

proaches that are trained in full MSA datasets from published paper for MASC. **TomBERT** (Yu and Jiang, 2019) models the intra-modality and inter-modality dynamics to improve the performance of MASC. **CapTrBERT** (Khan and Fu, 2021) constructs an auxiliary sentence, which is the translation of the image, to provide multimodal information to a language model. **KEF** (Zhao et al., 2022) exploits adjective-noun pairs extracted from the image to improve the visual attention capability and sentiment prediction capability of the fine-grained MSA task **FITE** (Yang et al., 2022a), the state-of-the-art model for fine-grained MSA, leverages facial information from the image modality.

Additionally, we adapt and evaluate models originally designed for **few-shot text classification tasks** for multimodal aspect-based sentiment classification. **LM-BFF** (Gao et al., 2021) designs different text prompts based on each specific dataset and text demonstrations to solve few-shot text classification tasks. **LM-SC** (Jian et al., 2022) further introduces supervised contrastive learning based on LM-BFF to few-shot text tasks. **GFSC** (Hosseini-Asl et al., 2022) converts the classification task into a generation task and solves text classification tasks in the few-shot setting through a pre-trained generation model, namely GPT2 (Radford et al., 2018). Recently, **a few multimodal sentiment classification models** in the few-shot setting have emerged. **PVLM** (Yu and Zhang, 2022) proposes a prompt-

based vision-aware language modeling approach to MASC in a few-shot scenario. **UP-MPF** (Yu et al., 2022) applies a unified pre-training for multimodal prompt-based fine-tuning model, which is the state-of-the-art model for few-shot MASC.

4.4 Experimental Results and Analysis

4.4.1 Results of JMASA

Table 2 presents the results of JMASA on few-shot multimodal datasets, and several key observations can be made. We can make the following observations. First, multimodal models generally outperform unimodal models. Among the multimodal models, JML and VLP, which leverage additional data for relation detection and pre-training, respectively, achieve better performance compared to NVLP, which does not involve pre-training tasks, indicating the effectiveness of pre-training tasks in improving model performance. When considering the amount of data used by the models, it is more reasonable to compare our model with NVLP. Our model consistently outperforms NVLP across both datasets, indicating its superior performance. Notably, our model also outperforms the second-best model, VLP, by a significant margin, with 1.56 and 2.03 absolute percentage points in terms of F1 on Twitter-15 and Twitter-17, respectively. The superior performance of our model can be attributed to several factors. First, the generative multimodal prompt, which is based on the multi-

Modality	Model	Twitter-15	Twitter-17
Text	BART	65.57 (±3.07)	64.12 (±1.47)
	LM-BFF*	64.87 (±0.40)	52.08 (±0.54)
	LM-SC*	65.47 (±1.74)	57.51 (±2.95)
	GFSC*	60.75 (±1.07)	61.72 (±0.16)
Text -Image	TomBERT	61.78 (±3.27)	59.97 (±2.30)
	CapTrBERT	58.76 (±0.25)	56.48 (±1.61)
	JML-SC	60.36 (±0.90)	61.62 (±0.45)
	CMMT-SC	43.75 (±2.90)	51.94 (±2.11)
	KEF	55.81 (±3.74)	46.50 (±0.075)
	FITE	63.11 (±0.53)	60.89 (±1.40)
	NVLP	63.84 (±1.49)	62.72 (±2.95)
	VLP	59.34 (±1.35)	60.24 (±1.61)
	PVLM*	64.54 (±1.81)	61.45 (±2.31)
	UP-MPF*	63.71 (±3.62)	62.02 (±0.40)
	GMP	**67.06** (±0.55)	**66.20** (±1.12)

Table 3: Results of different models in terms of **Acc** for MASC on two datasets. "*" means that the model is proposed for the few-shot task.

modal context, enables the model to capture practical knowledge for each sample from the pre-trained language model. Second, the subtask information provides valuable clues for constructing the multimodal prompt, leading to improved performance in few-shot multimodal sentiment classification.

4.4.2 Results of the MASC

The results of the MASC task on few-shot multimodal datasets, in terms of accuracy (Acc), are presented in Table 3, while the corresponding F1 results are shown in Table 6 from Appendix B.1. The models with "*" are specifically introduced for few-shot scenarios. Several key observations can be made from the results. We can obtain the following observations. In the multimodal few-shot setting, 1) Our model demonstrates the best performance in the multimodal few-shot setting, indicating its superiority over other models in handling the challenges of limited labeled data. 2) Prompt-based methods outperform robust multimodal models, highlighting the effectiveness of prompt-based methods in low-resource scenarios. This suggests that leveraging prompt engineering techniques, such as our generative multimodal prompt, can lead to improved performance in few-shot MSA. 3) BART, which uses only the text modality, performs better than most multimodal models, indicating the strong performance of our base model. This suggests that the pre-trained language model, BART, provides a solid foundation for our multimodal model.

4.4.3 Results of MATE

Table 4 presents the results of the MATE task. Among the models, VLP achieves the best performance in MATE, although it deviates from our

Modality	Model	Twitter-15	Twitter-17
Text	BART	66.67 (±3.17)	70.12 (±1.73)
Text -Image	JML-MATE	71.95 (±4.30)	82.14 (±1.20)
	CMMT-MATE	73.19 (±2.50)	82.50 (±0.59)
	NVLP-MATE	65.95 (±1.83)	71.52 (±0.26)
	VLP-MATE	**77.61** (±0.25)	**83.35** (±0.53)
	GMP	73.65 (±1.35)	79.95 (±0.43)

Table 4: Results of different models in terms of **F1** for MATE on two datasets.

initial goal of applying low-resource data due to its reliance on additional data and multiple pretraining tasks on the MVSA-Multiple Dataset (Niu et al., 2016). Similarly, JML also leverages additional data to enhance its performance. An interesting observation is that MASC performs poorly in VLP when compared to NVLP, despite VLP showing better performance on the MATE and JMASA tasks compared to NVLP. We hypothesize that the pre-training task of VLP may be more aligned with the MATE task, which in turn may have an impact on the performance of MASC.

4.5 Ablation Experiments

We performed ablation experiments on the GMP model to assess the effectiveness of different modules. The results, as shown in Table 5, indicate that the complete GMP model consistently the best performance across all tasks. First, we remove the image modality (w/o Image) and built generative prompts based only on the text modality. The model's performance in all tasks is adversely affected, indicating that the image modality is crucial for achieving high performance in few-shot MSA tasks. Next, we only remove the image caption (w/o Caption) and retain the initial image features to evaluate the effectiveness of the image prompt. The results show that the image prompt contributes to the overall performance of the model, indicating its utility in capturing important information from the image modality. We also conduct experiments where we remove the multitask module (w/o Multitask) and set the number of aspect terms to 5 for each instance in the JMASA and MATE tasks. The performance of the models is affected, indicating that the subtask-specific modules are effective in capturing aspect-related information and improving performance. To verify the utility of the generative multimodal prompt, we remove the multimodal prompt (w/o Prompt) and use only the original textimage representation. The model's performance degraded, indicating that our proposed multimodal

Task	Model	Twitter-15	Twitter-17
JMASA	w / GSPrompt	47.11 (±3.12)	52.43 (±1.35)
	w/o Multitask	47.70 (±1.41)	49.77 (±1.69)
	w/o Image	44.71 (±1.64)	50.25 (±1.97)
	w/o Caption	47.31 (±1.12)	52.11 (±1.16)
	w/o Prompt	47.55 (±1.54)	51.20 (±1.71)
	w/o GAPrompt	48.05 (±1.38)	48.81 (±4.98)
	GMP	**49.33 (±1.71)**	**53.79 (±1.31)**
MATE	w/o Multitask	73.46 (±0.94)	79.02 (±1.16)
	w/o Image	68.54 (±0.99)	74.41 (±3.19)
	w/o Caption	72.06 (±1.52)	78.91 (±1.49)
	w/o Prompt	72.55 (±0.93)	79.01 (±0.90)
	w/o GAPrompt	71.62 (±0.71)	78.74 (±0.94)
	GMP	**73.65 (±1.35)**	**79.95 (±0.43)**
MASC	w/ DSPrompt	64.48 (±3.47)	64.17 (±1.31)
	w/o Image	65.09 (±1.66)	65.68 (±0.67)
	w/o Caption	64.81 (±3.60)	66.01 (±1.69)
	w/o Prompt	62.75 (±1.18)	64.34 (±1.76)
	w/o GSPrompt	65.03 (±1.49)	63.57 (±2.29)
	GMP	**67.06 (±0.55)**	**66.20 (±1.12)**

Table 5: Ablation experiment results in terms of **F1** on two tasks, including JMASA, MATE, and in terms of **Acc** on MASC. "w/" indicates "with" and "w/o" indicates "without".

prompt is beneficial in providing valuable cues for the sentiment analysis task. We further remove the generative aspect prompt (w/o GAP) to assess the importance of GAP. Interestingly, we observe that using generated sentiment prompts (GSP) resulted in better performance in the MASC task (w/o GSP), whereas we obtain the opposite result in the JMASA task (w/ GSP). This suggests that the generated aspect prompt provides sufficient information to the model, and GSP may introduce redundant information in the JMASA task. However, in the MASC task, GSP provides effective cues for sentiment classification. We further experiment with different generated sentiment prompts (w DSPrompt) and find that the performance significantly decrease. There are two possible reasons for this observation. First, the sentiment categories in our dataset are limited. When using generated sentiment prompts for each aspect, it may introduce noise and irrelevant information to MASC. Second, the generated prompts for each aspect provide sufficient information to guide the model in capturing aspect-related sentiment information.

4.6 Hyperparameters Setting

The hyperparameter experiments of JMASA are shown in Fig. 4. The hyperparameter experiments on other tasks are in Appendix B.2.

Hyperparameters l_i **and** λ **on JMASA**. In order to effectively utilize image information through NF-ResNet, we conduct experiments with different settings of the hyperparameter l_i in Eq. 1, and the results are shown in Fig. 4(a). We observe that our GMP model achieves the best performance on both datasets when the number of image slots, l_i, is set to 4. When l_i is smaller, the image information is not fully utilized, and the model's performance is compromised. On the other hand, retaining more image features by setting a larger value for l_i results in redundant information being provided to the model, which also leads to decreased performance. When l_i was set to 0, GMP only utilized the image prompt, i.e., the image caption C, and discarded the initial image representation V. We also employ the hyperparameter λ to balance the contribution of the subtask, as shown in Fig. 4(b). We find that the best value of λ varied across different datasets, with 0.1 being the optimal value for Twitter-15 and 0.15 for Twitter-17. When λ is set to a larger value, the model's performance dramatically drop. This is because a larger value of λ biases the model towards the subtasks, and we need to strike a balance among all tasks to achieve optimal performance.

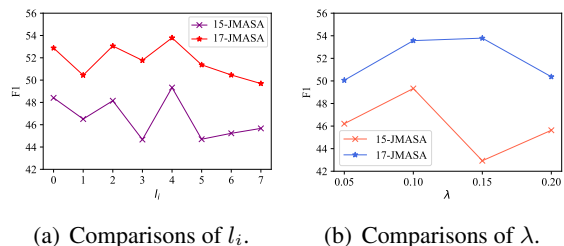(a) Comparisons of l_i. (b) Comparisons of λ.

Figure 4: **F1** comparisons of different Hyperparameters for JMASA.

5 Conclusion

We propose a novel Generative Multimodal Prompt (GMP) for Multimodal Aspect-Based Sentiment Analysis (MABSA) that includes JMASA, MASC, and MATE in the multimodal few-shot scenario. We further introduce a subtask to predict the number of aspect terms to form multitask training to improve the performance of GMP. Experimental results show that our proposed approach outperforms strong baselines on two subtasks of MABSA in the few-shot setting. We provide a new direction for related tasks of MABSA in the few-shot setting. In future work, we plan to exploit the fine-grained image features and achieve alignment between text and image modality to improve the performance of MABSA in the multimodal few-shot scenario.

Limitations

Although our model has shown superior performance, there are still a few limitations that could be improved in future work.

- We create few-shot datasets from the perspective of the combination of sentiment categories without considering the distribution of aspect items, such as the number of aspects in each sample. It may affect the performance of the model on the task of extracting aspects. We should create more efficient datasets for MABSA in the few-shot setting.

- As we put more emphasis on the performance of the main task, the performance of the subtask of predicting the number of aspect terms in each example may suffer. We will further improve the accuracy of the subtask in future work.

- We roughly exploit initial image features and do not perform alignment between text and image modalities. We plan to accomplish the alignment of multiple modalities further to improve the performance of MABSA in future work.

Acknowledgements

References

Andrew Brock, Soham De, and Samuel L. Smith. 2021. Characterizing signal propagation to close the performance gap in unnormalized resnets. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Ganesh Chandrasekaran, Tu N. Nguyen, and Jude Hemanth D. 2021. Multimodal sentimental analysis for social media applications: A comprehensive review. *WIREs Data Mining Knowl. Discov.*, 11(5).

Guimin Chen, Yuanhe Tian, and Yan Song. 2020. Joint aspect extraction and sentiment analysis with directional graph convolutional networks. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 272–279. International Committee on Computational Linguistics.

Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Good visual guidance makes A better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. *CoRR*, abs/2205.03521.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Inf. Fusion*, 91:424–444.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3816–3830. Association for Computational Linguistics.

Ehsan Hosseini-Asl, Wenhao Liu, and Caiming Xiong. 2022. A generative language model for few-shot aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 770–787. Association for Computational Linguistics.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 537–546. Association for Computational Linguistics.

Yiren Jian, Chongyang Gao, and Soroush Vosoughi. 2022. Contrastive learning for prompt-based few-shot language learners. In *NAACL*, pages 5577–5587.

Xincheng Ju, Dong Zhang, Rong Xiao, Junhui Li, Shoushan Li, Min Zhang, and Guodong Zhou. 2021. Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4395–4405. Association for Computational Linguistics.

Zaid Khan and Yun Fu. 2021. Exploiting BERT for multimodal target sentiment classification through

input space translation. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 3034–3042. ACM.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4582–4597. Association for Computational Linguistics.

Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao. 2022. CLMLF: A contrastive learning and multi-layer fusion method for multimodal sentiment detection. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2282–2294. Association for Computational Linguistics.

Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2149–2159. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1990–1999. Association for Computational Linguistics.

Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. Clipcap: CLIP prefix for image captioning. *CoRR*, abs/2111.09734.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El-Saddik. 2016. Sentiment analysis on multi-view social data. In *MultiMedia Modeling - 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II*, volume 9517 of *Lecture Notes in Computer Science*, pages 15–27. Springer.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4222–4235. Association for Computational Linguistics.

Lin Sun, Jiquan Wang, Yindu Su, Fangsheng Weng, Yuxuan Sun, Zengwei Zheng, and Yuanyi Chen. 2020. RIVA: A pre-trained tweet multimodal model based on text-image relation for multimodal NER. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1852–1862. International Committee on Computational Linguistics.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 200–212.

Hanqian Wu, Siliang Cheng, Jingjing Wang, Shoushan Li, and Lian Chi. 2020a. Multimodal aspect extraction with region-aware alignment network. In *Natural Language Processing and Chinese Computing - 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14-18, 2020, Proceedings, Part I*, volume 12430 of *Lecture Notes in Computer Science*, pages 145–156. Springer.

Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020b. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1038–1046. ACM.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2416–2429. Association for Computational Linguistics.

Hao Yang, Yanyan Zhao, and Bing Qin. 2022a. Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3324–3335. Association for Computational Linguistics.

Li Yang, Jin-Cheon Na, and Jianfei Yu. 2022b. Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis. *Inf. Process. Manag.*, 59(5):103038.

Li Yang, Jianfei Yu, Chengzhi Zhang, and Jin-Cheon Na. 2021a. Fine-grained sentiment analysis of political tweets with entity-aware multimodal network. In *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*, pages 411–420. Springer.

Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. 2021b. Multimodal sentiment detection based on multi-channel graph neural networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 328–339. Association for Computational Linguistics.

Jianfei Yu and Jing Jiang. 2019. Adapting BERT for target-oriented multimodal sentiment classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5408–5414. ijcai.org.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 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, ACL 2020, Online, July 5-10, 2020*, pages 3342–3352. Association for Computational Linguistics.

Yang Yu and Dong Zhang. 2022. Few-shot multi-modal sentiment analysis with prompt-based vision-aware language modeling. In *IEEE International Conference on Multimedia and Expo, ICME 2022, Taipei, Taiwan, July 18-22, 2022*, pages 1–6. IEEE.

Yang Yu, Dong Zhang, and Shoushan Li. 2022. Unified multi-modal pre-training for few-shot sentiment analysis with prompt-based learning. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 189–198. ACM.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021a. Multimodal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14347–14355.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021b. Multimodal graph fusion for named entity recognition with targeted visual guidance. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Arti-ficial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14347–14355. AAAI Press.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5674–5681. AAAI Press.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022a. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *CoRR*, abs/2203.01054.

Yue Zhang, Hongliang Fei, Dingcheng Li, and Ping Li. 2022b. Promptgen: Automatically generate prompts using generative models. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 30–37. Association for Computational Linguistics.

Fei Zhao, Zhen Wu, Siyu Long, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2022. Learning from adjective-noun pairs: A knowledge-enhanced framework for target-oriented multimodal sentiment classification. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 6784–6794. International Committee on Computational Linguistics.

Jie Zhou, Jiabao Zhao, Jimmy Xiangji Huang, Qinmin Vivian Hu, and Liang He. 2021. MASAD: A large-scale dataset for multimodal aspect-based sentiment analysis. *Neurocomputing*, 455:47–58.

Linan Zhu, Minhao Xu, Yinwei Bao, Yifei Xu, and Xiangjie Kong. 2022. Deep learning for aspect-based sentiment analysis: a review. *PeerJ Comput. Sci.*, 8:e1044.

A Multimodal Embedding with Prompt

For the MASC task, we design multimodal embedding with generative multimodal prompt, E_S^P, as Fig. 5(a) shows.

For the MATE task, we design multimodal embedding with generative multimodal prompt, E_A^P, as Fig. 5(b) shows.

B Experimental Results

B.1 F1 Results of MASC

The results of the MASC task in terms of F1 are shown in Table 6.

(a) The multimodal embedding with generative multimodal prompt for MASC.

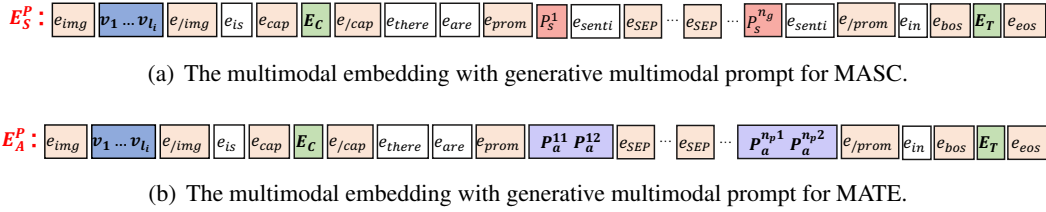(b) The multimodal embedding with generative multimodal prompt for MATE.

Figure 5: Multimodal embeddings with the generative multimodal prompt for MASC and MATE.

B.2 Hyperparameters Setting

Hyperparameters l_i on MASC: We use the gold number of aspect terms for the MASC task and don't use the subtask. Thus we only conduct experiments on the hyperparameter l_i. Similar to the JMASA task, our model achieves the best performance on two datasets when l_i is 4, as Fig. 6 shows.
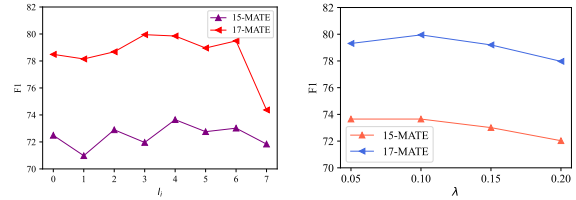

(a) Comparisons of l_i.   (b) Comparisons of λ.

Figure 7: **F1** comparisons of different Hyperparameters for MATE.

Modality	Model	Twitter-15	Twitter-17
Text	BART	57.21 ($\pm$4.62)	61.71 ($\pm$2.01)
	LM-BFF*	58.27 ($\pm$1.46)	49.04 ($\pm$3.40)
	LM-SC*	58.02 ($\pm$2.26)	55.97 ($\pm$2.54)
	GFSC*	29.3 ($\pm$1.97)	40.91 ($\pm$4.46)
Text -Image	TomBERT	43.16 ($\pm$8.08)	54.92 ($\pm$2.40)
	CapTrBERT	26.55 ($\pm$0.98)	49.59 ($\pm$3.69)
	JML-SC	44.77 ($\pm$2.10)	52.19 ($\pm$0.70)
	CMMT-SC	45.52 ($\pm$0.85)	51.92 ($\pm$1.00)
	KEF	43.54 ($\pm$0.24)	29.61 ($\pm$0.23)
	FITE	58.97 ($\pm$0.34)	59.16 ($\pm$2.15)
	NVLP	55.11 ($\pm$2.20)	59.37 ($\pm$4.09)
	VLP	44.56 ($\pm$3.83)	56.09 ($\pm$2.43)
	PVLM*	50.87 ($\pm$2.37)	59.62 ($\pm$1.81)
	UP-MPF*	55.15 ($\pm$1.33)	60.46 ($\pm$1.08)
	GMP	**60.31 ($\pm$1.83)**	**64.20 ($\pm$1.63)**

Table 6: Results of different models in terms of **F1** for MASC on two datasets.

Hyperparameters l_i and λ on MATE: Fig. 7 shows the hyperparameters of the MATE, including l_i and λ. On both datasets, our model has the best results when λ is 4. For the hyperparameter, l_i, our model achieves the best performance when l_i is 4 on the Twitter-15 dataset, and l_i is 3 on the Twitter-17 dataset.
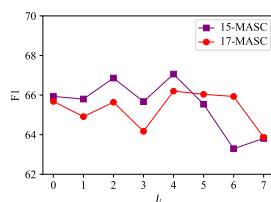

Figure 6: **Acc** comparisons of different Hyperparameters for MASC.

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☑ A2. Did you discuss any potential risks of your work?
Limitations

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract; 1; 5

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☐ Did you use or create scientific artifacts?

Not applicable. Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4.1

C ☑ Did you run computational experiments?

4.4; 4.5; 4.6

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. Left blank.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4.2; 4.6

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4.1; 4.4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4.2

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.