VERO: Verification and Zero-Shot Feedback Acquisition for Few-Shot Multimodal Aspect-Level Sentiment Classification

Kai Sun,[1,2] **Hao Wu,**[3] **Bin Shi,**[1,2*] **Samuel Mensah,**[4] **Peng Liu,**[3*] **Bo Dong**[1,5]

[1]Shaanxi Provincial Key Laboratory of Big Data Knowledge Engineering, Xi'an Jiaotong University, China
[2]School of Computer Science and Technology, Xi'an Jiaotong University, China
[3]School of computer science and engineering, Guangxi Normal University, Guilin, China
[4]Department of Computer Science, University of Sheffield, Sheffield, United Kingdom
[5]School of Continuing Education, Xi'an Jiaotong University, China
sunkai@xjtu.edu.cn, wuhao@stu.gxnu.edu.cn, shibin@xjtu.edu.cn, s.mensah@sheffield.ac.uk, liupeng@gxnu.edu.cn, dong.bo@xjtu.edu.cn

Abstract

Deep learning approaches for multimodal aspect-level sentiment classification (MALSC) often require extensive data, which is costly and time-consuming to obtain. To mitigate this, current methods typically fine-tune small-scale pretrained models like BERT and BART with few-shot examples. While these models have shown success, Large Vision-Language Models (LVLMs) offer significant advantages due to their greater capacity and ability to understand nuanced language in both zero-shot and few-shot settings. However, there is limited work on fine-tuning LVLMs for MALSC. A major challenge lies in selecting few-shot examples that effectively capture the underlying patterns in data for these LVLMs. To bridge this research gap, we propose an acquisition function designed to select challenging samples for the few-shot learning of LVLMs for MALSC. We compare our approach, **V**erification and Z**ERO**-shot feedback acquisition (VERO), with diverse acquisition functions for few-shot learning in MALSC. Our experiments show that VERO outperforms prior methods, achieving an F1 score improvement of up to 6.07% on MALSC benchmark datasets.

Code — https://github.com/absdog/vero

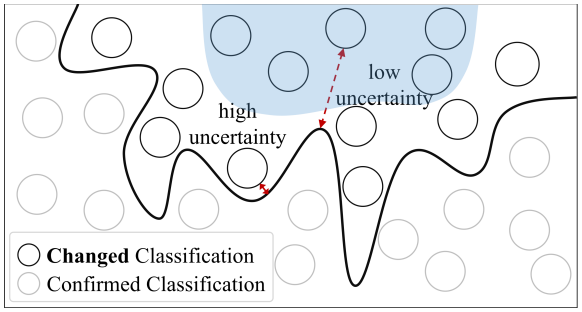

Figure 1: Illustrative example of VERO for LVLMs. The solid line represents the LVLM's self-verification decision boundary, separating samples into "changed" (black circles) and "confirmed" (gray circles) classifications based on the model's zero-shot and self-verification predictions on unlabeled data. The distance between circles and the boundary reflects the uncertainty level after self-verification. We focus on selecting samples within the low-uncertainty region of the "changed" class (highlighted in blue), as these are challenging and informative for fine-tuning the LVLM.

Introduction

With the growing abundance of multimodal data shared on social media, multimodal aspect-level sentiment classification (MALSC) has gained more attention. MALSC aims to determine the sentiment polarity of an aspect based on the given text-image pair. Previous approaches to MALSC focused on enhancing model performance by leveraging large amounts of training data or incorporating additional auxiliary data (Khan and Fu 2021; Ling, Yu, and Xia 2022; Yang, Xiao, and Du 2024). However, the collection and annotation of multimodal data for MALSC are inherently time-consuming and labor-intensive. To address this challenge, many studies have focused on Few-shot MALSC (Yu and Zhang 2022; Yu, Zhang, and Li 2022; Yang et al. 2023b), aiming is to identify and select a few highly informative samples from a large pool of unlabeled data to optimize the model performance on the task.

Recent advancements in Few-shot MALSC have primarily focused on fine-tuning small-scale pretrained language models (PLMs), such as BERT (Devlin et al. 2019) and BART (Lewis et al. 2020). Meanwhile, Large Vision-Language Models (LVLMs) demonstrate remarkable performance on the task using different prompt-based techniques, including zero-shot (Yang et al. 2024b), few-shot (Li et al. 2023; Ye et al. 2023), chain-of-thought (CoT) (Li et al. 2024) and most recently self-verification (Gero et al. 2023) - a better reasoning of CoT, where the LVLM performs a forward reasoning and backward verification to arrive at an answer. Then again, adapting LVLMs to unseen data through fine-tuning is crucial, yet this process on large-scale data is costly. Few-shot fine-tuning offers a promising alternative, allowing these models to quickly adapt to new domain-specific data while minimizing the resources and time typically required. However, limited studies have focused on few-shot fine-tuning of LVLMs for MALSC (Yang et al.

*Corresponding author

2024a).

To acquire samples to finetune LVLMs for the task, existing methods (Yu and Zhang 2022) rely on random sampling from a specific prior distribution (e.g., uniform or the distribution of the training dataset). The current state-of-the-art (SOTA) approach, MultiPoint (Yang et al. 2023b) proposed consistently distributed sampling (CDS), where the acquisition function ensured the consistency of the sentiment distribution between the full training dataset and the sampled dataset. Though these sampling methods (or acquisition functions) can be directly applied to LVLMs, they may not guarantee that the selected samples are informative to the LVLMs. This is because these techniques do not consider the capability of the LVLM itself. Considering that LVLMs already possess some zero-shot capabilities for MALSC, samples selected by these prior methods (Yu and Zhang 2022; Yang et al. 2023b) may lead to limited training benefits.

To this end, we propose an acquisition function that searches the pool of unlabelled data for challenging examples. Specifically, our method VERO, selects unlabeled data in the pool where there is a divergence between the LVLM's zero-shot predictions and subsequent self-verification predictions. Intuitively, the most informative samples are those that prompt the model to alter its classification during self-verification after an initial zero-shot prediction. Such samples likely contain complex features that cause the model to oscillate between different classes. As illustrated in Figure 1, we focus on selecting samples within the low-uncertainty region, shown in blue, because they present cases where self-verification reveals a confident shift in prediction. The fact that the LVLM becomes firmly convinced of a new class during re-evaluation indicates that these samples are challenging and crucial for refining the model's understanding and enhancing its overall performance.

We validated our approach using the LVLM-based models, LLaVA-7b and LLaVA-13b on the Twitter-2015 and Twitter-2017 datasets, demonstrating that our method outperforms previous techniques in few-shot learning scenarios, with an F1 score improvement of up to 6.07%. Our analyses demonstrate that the samples selected by our method challenge LVLMs more effectively than current methods (Yu and Zhang 2022; Yu, Zhang, and Li 2022; Yang et al. 2023b), leading to significant performance gains.

Our contributions are the following:

- To the best of our knowledge, we are the first to successfully fine-tune a Large Vision Language Model to solve the few-shot MALSC task.
- We propose VERO, a novel acquisition function that acquires challenging samples from a pool of unlabeled data for LVLM fine-tuning. VERO achieves this by leveraging the zero-shot and self-verification capability of LVLMs.
- We conduct extensive experiments on two benchmark datasets across 1% and 7% few-shot settings, demonstrating the remarkable superiority of our approach.

Related Work

Multimodal Aspect-Level Sentiment Classification

With the proliferation of multimodal data disseminated on social media, multimodal aspect-level sentiment classification (MALSC) began to receive increasing attention. Research in MALSC could be mainly categorized into three research lines: cross-modal attention methods, image translation methods and small-scale PLMs. Approaches based on cross-modal attention focus on utilizing attention mechanisms to implicitly align and fuse the semantic information and emotion information in the two modalities (Xu, Mao, and Chen 2019; Yu and Jiang 2019; Xiao et al. 2023; Yang, Xiao, and Du 2024). In image translation methods (Khan and Fu 2021; Yang, Zhao, and Qin 2022a), images will be translated into textual descriptions, and then these descriptions are used as additional context for the input text. Small-scale PLMs aim at improving representation learning (Lu et al. 2019; Nguyen, Vu, and Nguyen 2020; Ling, Yu, and Xia 2022). However, these methods require a large amount of annotated data for model fine-tuning, which is time-consuming and labour intensive.

Few-shot MALSC with PLMs

With the scaling of pre-trained language models (PLMs) from 110M parameters (Devlin et al. 2019) to over 500B parameters (Smith et al. 2022), the capabilities of these models have greatly improved. In recent studies, PLMs have been regarded as powerful tools for solving few-shot MALSC (Liu et al. 2023), which could be mainly categorized into two research lines: finetuning-based methods and in-context learning methods.

Finetuning-based Methods These models treat the classification task as a masked language modeling (MLM) task, where the model is fine-tuned with a set of prompts to guide its prediction by filling a special token, [MASK] (Gao, Fisch, and Chen 2021; Jian, Gao, and Vosoughi 2022; Hosseini-Asl, Liu, and Xiong 2022; Yu, Zhang, and Li 2022; Yu and Zhang 2022). Hosseini-Asl, Liu, and Xiong (2022) proposed a generative language model (GFSC) that reformulates the task as a language generation problem. Yang et al. (2023a) proposed a generative multimodal prompt (GMP) model for the joint multiple apsect-level sentiment analysis (JMALSA) task. Meanwhile, Yang et al. (2023b) proposed a unified multimodal prompt that allows for the joint processing of both text and image modalities in a coherent manner.

Despite the achievements of these methods, their limitations have become increasingly evident in the era of LVLMs. Firstly, they still persist in fine-tuning small-scale pre-trained models, without recognizing the powerful image-text understanding capabilities of LVLMs (Yang et al. 2023a,b, 2024a). Secondly, it is important for few-shot MALSC to select a small set of samples that can maximize the fine-tuning benefits. These methods (Yu and Zhang 2022; Yu, Zhang, and Li 2022; Yang et al. 2023b) typically sample from training and development set randomly according to a specific distribution (e.g., uniform, the distribution of original training set), which does not guarantee that the samples are informative with respect to the capability of LVLMs.

In-context learning methods These approaches enhance the performance of LVLMs by incorporating demonstrations into prompts without the need for parameter updating (Li et al. 2023; Ye et al. 2023; Yang et al. 2024b). For instance, Yang et al. (2024b) explored the potential of using ChatGPT for In-Context Learning (ICL) on the MALSA task and enchanced the ICL framework's performance in few-shot learning scenarios via an entity-aware contrastive learning method. Despite promising performances in extreme few-shot scenarios, the performance of the ICL method is often unstable, affected by factors such as instruction, the formatting of demonstrations, and the order of these examples.

Active Learning

Active Learning (AL) (Settles 2009) focuses on selecting the most "informative" training samples to finetune a model. AL falls into three main categories: uncertainty, representativeness, and performance-based methods. Uncertainty-based methods target the most uncertain instances (Zhu et al. 2009; Yang et al. 2015; Raj and Bach 2022), using criteria like entropy and margin. Representativeness-based methods select samples best representing the input distribution (Huang, Jin, and Zhou 2014; Xie et al. 2022). Performance-based methods directly optimize informativeness via surrogates, considering the impact of revealing an instance's label on future outcomes (Roy and McCallum 2001; Schein and Ungar 2007; Cai, Zhang, and Zhou 2013). Our method can be regarded as a an active learning approach, however it distinguishes itself from previous works by leveraging the inherent capability of the LVLM (e.g. zero-shot prediction, self-verification) for sample acquisition.

Uncertainty Estimation

Uncertainty estimation in language models is significant for analyzing the potential erroneous behaviors of these models. To aggregate uncertainty information obtained at the token level, Manakul, Liusie, and Gales (2023) propose four different metrics, including the maximum or average likelihood and the maximum or average entropy. Additionally, the prediction uncertainty of large language models can also be estimated by sample-based methods (Huang et al. 2023) and perturbation-based methods (Meister et al. 2023; Huang et al. 2023).

Methodology

We propose an acquisition function, called VERO, to select challenging samples from the pool of unlabeled data based on the LVLM's zero-shot predictions and subsequent self-verification. In the following sections, we begin by formally defining the problem of few-shot MALSC. Next, we introduce the proposed VERO for sample acquisition. The overview of our approach is presented in Figure 2.

Problem Statement

MALSC aims to classify the sentiment polarity y of a specified aspect a in a sentence-image pair, where a is a marked phrase in s and image v assists classification. Typically, MALSC is treated as a three-class classification task where the polarity y belongs to a predefined set of polarities, $y \in \{\text{Negative}, \text{Neutral}, \text{Positive}\}$. In this paper, we focus on the few-shot MALSC task. Given a training dataset $\mathcal{D}_{\text{train}} = \{(s^j, a^j, v^j, y^j)\}_{j=1}^{N}$, a subset of K samples is selected for training, where $K \ll N$. Specifically, the sentiment label y^j are not available during selection, i.e., we optimize model performance while saving annotation costs.

Verification and Zero-Shot Feedback Acquisition

Firstly, two multimodal prompts $\mathcal{P}_{\text{SC}}$ and $\mathcal{P}_{\text{SV}}$ are designed for sentiment classification (SC) and self-verification (SV) respectively. Each prompt consists of an *instruction*, an *input* and a *output format*. The *instruction* formally describes the task. The *input* presents the input sample for the task, while the *output format* formally describe the output structure. Examples of $\mathcal{P}_{\text{SC}}$ and $\mathcal{P}_{\text{SV}}$ are presented in Figure 2. Based on the constructed $\mathcal{P}_{\text{SC}}$ and $\mathcal{P}_{\text{SV}}$, we perform zero-shot prediction and self-verification for sample acquisition, which involves three steps: zero-shot prediction, self-verification guided sample division, ranking by uncertainty.

Zero-Shot Prediction $\mathcal{D}_{\text{unlabeled}} = \{(s^j, a^j, v^j)\}_{j=1}^{N}$ denotes the unlabeled training dataset. For the j-th sample in $\mathcal{D}_{\text{unlabeled}}$, we perform zero-shot prediction as follows:

$$\hat{y}^j = \text{LVLM}(\mathcal{P}_{\text{SC}}^j) \tag{1}$$

where $\hat{y}^j$ denotes the sentiment prediction.

Self-Verification Guided Sample Division Subsequently, we perform a self-verification and divide samples into $\mathcal{D}_{\text{no}}$ and $\mathcal{D}_{\text{yes}}$, as follows:

$$\hat{c}^j = \text{LVLM}(\mathcal{P}_{\text{SV}}^j) \tag{2}$$

$$\mathcal{D}_{\text{no}} = \{(s^j, a^j, v^j) | \hat{c}^j = \texttt{no}\}_{j=1}^{N} \tag{3}$$

$$\mathcal{D}_{\text{yes}} = \{(s^j, a^j, v^j) | \hat{c}^j = \texttt{yes}\}_{j=1}^{N} \tag{4}$$

Samples are assigned to $\mathcal{D}_{\text{no}}$ if there is a divergence between their zero-shot predictions and subsequent self-verification outputs ($\hat{c}^j = \texttt{no}$); otherwise, they are assigned to $\mathcal{D}_{\text{yes}}$ ($\hat{c}^j = \texttt{yes}$). Samples in $\mathcal{D}_{\text{no}}$ are considered challenging since they cause the model to shift its prediction after re-evaluation via self-verification. Fine-tuning the model on these samples is crucial for improving its understanding and can lead to greater training benefits.

Ranking by Uncertainty To identify the samples where the LVLM becomes firmly convinced of a new prediction during re-evaluation, we estimate the uncertainty of self-verification output and focus on selecting samples within the low-uncertainty. The uncertainty for a sample is computed as follows:

$$u^j = \frac{1}{\max(P(\hat{c}^j))} \tag{5}$$

where $P(\hat{c}^j)$ represents the probability distribution for the self-verification output $\hat{c}^j$, and $\max(P(\hat{c}^j))$ represents the highest probability of $P(\hat{c}^j)$. We rank the samples by their uncertainty scores and select the top-K samples with the lowest scores for fine-tuning.

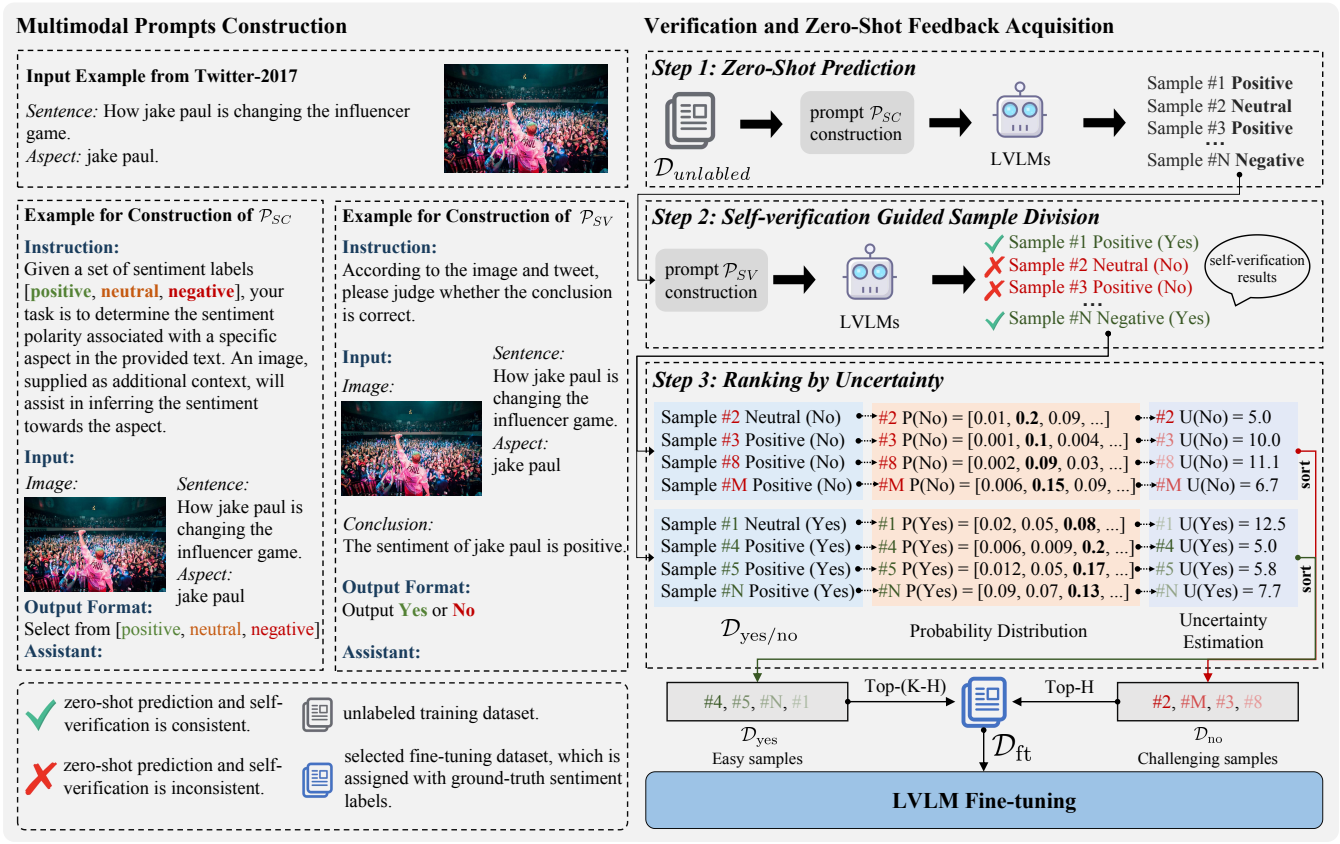

Figure 2: Overview of our approach.

Fine-tuning for Few-shot MALSC

The top-K samples with the lowest scores are considered challenging, as their predictions diverge significantly between zero-shot and self-verification. However, to help the model maintain stable performance across a range of examples rather than focusing too heavily on challenging cases during few-shot fine-tuning, we include a small set of relatively easy samples from $\mathcal{D}_{\text{yes}}$ to balance training. To achieve this, we introduce a hyperparameter $\lambda \in [0, 1]$ to quantify the proportion of samples selected from $\mathcal{D}_{\text{no}}$. Let $H = \lfloor \lambda K \rfloor$ be the number of samples selected from $\mathcal{D}_{\text{no}}$. The dataset for fine-tuning can be constructed as follows:

$$\begin{aligned} \mathcal{D}_{\text{ft}} &= \mathcal{D}^C \cup \mathcal{D}^E \\ \mathcal{D}_{\text{ft}} &= \{(x^j, y^j) | x^j \in \text{Sort}(\mathcal{D}_{\text{no}})\}_{j=1}^{H} \\ &\cup \{(x^j, y^j) | x^j \in \text{Sort}(\mathcal{D}_{\text{yes}})\}_{j=1}^{K-H} \end{aligned} \tag{6}$$

where $\mathcal{D}^C \subset \mathcal{D}_{\text{no}}$ (**C**hallenging samples) and $\mathcal{D}^E \subset \mathcal{D}_{\text{yes}}$ (**E**asy samples); Sort$(\cdot)$ function ranks samples by self-verification uncertainty scores in ascending order.

Loss Function Finally, we fine-tune the LVLM on $\mathcal{D}_{\text{ft}}$, where the sentiment label y is transformed to the target sequence $\bar{y}$ to fit the generative nature of LVLM. For example, the sentiment label *positive* is formatted as *The sentiment of* $\{aspect\}$ *is positive*. The target sequence is used to compute cross-entropy loss with the LVLM's output, as follows:

$$\mathcal{L} = \frac{1}{K \times M} \sum_{j=1}^{K} \sum_{i=1}^{M} \text{CE}(\hat{y}_i^j, \bar{y}_i^j) \tag{7}$$

where CE$(\cdot)$ denotes the cross-entropy loss function, K denotes the number of samples in $\mathcal{D}_{\text{ft}}$ and M denotes the length of target sequence.

Experiment

Setup Following the recent work (Yang et al. 2023b), we evaluate our approach on two benchmark datasets of MALSC: Twitter-2015 and Twitter-2017. To ensure a fair comparison, we directly utilize the preprocessed datasets provided by (Yang et al. 2023b). In the few-shot setting, 1% or 7% of samples are selected from the training dataset to fine-tune the models. The statistics of the two datasets are presented in Table 2. Implementation details and evaluation protocols are outlined in the accompanying code repository.

Main Result

The performance comparison on few-shot MALSC is presented in Table 1. Firstly, we find that some Text-Only models even achieve promising performances, compared to the Text-Image baselines. For example, LM-SC shows competitive performance with UP-MPF while surpassing all previ-

Modality	Model	Twitter-2015		Twitter-2017	
		Accuracy	**Weighted-F1**	**Accuracy**	**Weighted-F1**
Text-Only	RoBERTa (Liu et al. 2019)	55.58±4.13	52.32±2.28	48.22±2.95	46.37±3.17
	PT (Yang et al. 2023b)	61.97±3.15	60.11±3.38	58.77±3.70	57.85±3.63
	LM-BFF (Gao, Fisch, and Chen 2021)	60.87±3.38	59.63±3.04	56.84±3.51	55.96±3.48
	LM-SC (Jian, Gao, and Vosoughi 2022)	61.16±3.31	60.99±3.28	54.78±1.93	52.89±2.63
	GFSC (Hosseini-Asl, Liu, and Xiong 2022)	52.77±0.38	52.01±0.56	54.43±2.47	53.15±2.70
Text-Image	MFN (Yang et al. 2023b)	55.86±1.66	52.81±1.45	50.91±2.86	49.20±3.05
	CLMLF (Li et al. 2022)	56.97±2.08	52.04±2.35	49.63±2.40	45.72±2.17
	TomBERT (**?**)	55.95±5.17	43.25±0.06	47.47±2.26	36.93±5.89
	EF-CapTrBERT (Khan and Fu 2021)	57.81±1.45	42.72±1.00	47.41±1.01	33.58±3.58
	KEF (Ling, Yu, and Xia 2022)	57.58±2.04	43.09±0.25	45.74±0.78	31.29±2.39
	FITE (Yang, Zhao, and Qin 2022b)	58.42±0.18	43.29±0.11	46.20±0.52	29.97±0.70
	VLP-MABSA (Ling, Yu, and Xia 2022)	53.36±1.07	43.23±3.75	55.32±3.39	48.96±1.26
	PVLM (Yu and Zhang 2022)	59.25±2.02	54.45±3.33	54.28±3.17	51.02±5.24
	UP-MPF (Yu, Zhang, and Li 2022)	61.56±2.43	60.16±2.54	54.93±2.22	51.87±4.08
	MultiPoint (Yang et al. 2023b)	67.33±1.07	66.61±1.36	61.88±2.56	61.23±2.58
	LLaVA-7b	50.24	48.26	56.00	53.00
	LLaVA-13b	57.86	57.04	57.86	54.87
	$\text{VERO}_{\text{LLaVA}-7\text{b}}$ (ours)	67.30±0.41	66.49±0.18	64.56±0.68	64.22±0.86
	$\text{VERO}_{\text{LLaVA}-13\text{b}}$ (ours)	**69.17±0.46**	**69.42±0.46**	**67.46±0.28**	**67.30±0.29**

Table 1: Few-shot performance on the Twitter-2015 and Twitter-2017 datasets. For all datasets, the few-shot dataset represents 1% of the overall training data. The best performance is marked in bold. Our results are averaged over 10 runs with different random seeds, with standard deviation reported. Results for the compared models are retrieved from (Yang et al. 2023b).

	Twitter-2015			Twitter-2017		
	Train	Dev	Test	Train	Dev	Test
#POS	928	303	317	1508	515	493
#NEU	1883	679	607	1638	517	573
#NEG	368	149	113	416	144	168
Total	3179	1122	1037	3562	1176	1234
#ASP	1.34	1.33	1.35	1.41	1.45	1.45
#Len	16.72	16.74	17.05	16.21	16.37	16.38

Table 2: Dataset Statistics, including counts of positive **#POS**, neutral **#NEU**, and negative **#NEG** instances, as well as the average number of aspects per sentence **#ASP**, and the average sentence length **#LEN**

.

ous Text-Image baselines, suggesting the importance of developing a strong image understanding module to effectively use image information. Secondly, we find that MultiPoint, current state-of-the-art for this task, achieves a large performance margin over previous baselines by leveraging consistently distributed sampling (CDS) for few-shot fine-tuning.

To establish a clear baseline for LVLM-based models, we experimented with the vanilla LLaVA-7b/13b models and observed that MultiPoint outperforms them by large margins. This highlights the importance of fine-tuning LVLMs to achieve competitive performance for the task, as these models are designed as general-purpose solvers. VERO, our finetuned LLaVa-7b/13b, achieves gains of at least 8 Acc and 11 Weighted-F1 points over LLaVA-7b/13b, suggesting its gains stem from fine-tuning. Comparing to MultiPoint, $\text{VERO}_{\text{LLaVA}-7\text{b}}$ achieves comparable performance on Twitter-2015 and outperforms it on Twitter-2017. $\text{VERO}_{\text{LLaVA}-13\text{b}}$ further sets a new SOTA, surpassing MultiPoint by 3.71% in accuracy and 4.44% in weighted F1, while demonstrating stability with low standard deviations.

Some recent works (e.g. GMP (Yang et al. 2023a)), perform this task on the 7% few-shot setting. Accordingly, we evaluate our model on this setting as well, as presented in Table 3. It can be observed that the VERO surpasses the GMP with an accuracy margin of 3.22% and 3.18% on Twitter-2015 and Twitter-2017, further proving its superiority.

Model	Twitter-2015	Twitter-2017
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
TomBERT	61.78±3.27	59.97±2.30
CapTrBERT	58.76±0.25	56.48±1.61
KEF	55.81±3.74	46.50±0.08
FITE	63.11±0.53	60.89±1.40
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
VERO	**70.28±0.72**	**69.38±0.36**

Table 3: Performance comparison in terms of **Accuracy** on the Twitter-2015 and Twitter-2017 datasets. For all datasets, the few-shot dataset represents 7% of the overall training data. Results for the compared models are retrieved from (Yang et al. 2023a).

Performance on Ablated Acquisition Functions

To verify the contribution of each component in our method, we compare the performances of different acquisition functions ablated from our approach. We categorize these acquisition functions into uncertainty-based methods which select samples based on the uncertainty scores of the LVLM's zero-shot prediction or self-verification, and the diversity-based methods which select samples based on a prior distribution. The uncertainty-based methods include: (1) -$\mathcal{D}^E$ which forces the model to finetune on only challenging samples (2) "-$\mathcal{D}^E$,Ver" which additionally removes the self-verification from our model. Thus, the acquisition function selects samples based on the uncertainty scores of the zero-shot predictions. The diversity-based methods include: (3) "-$\mathcal{D}^E$,Unc" which reduces the model to randomly select challenging samples; (4) "-$\mathcal{D}^E$,Unc,Ver" which further reduces the model to follow the CDS method to randomly select samples from the training set. The results are presented in Table 4.

Base	Model	Twitter-2015	Twitter-2017
LLaVA-7b	Uncertainty-based Methods		
	VERO	**66.49±0.18**	**64.22±0.86**
	-$\mathcal{D}^E$	64.66±0.26	62.95±0.38
	-$\mathcal{D}^E$,Ver	64.44±0.24	62.80±0.59
	Diversity-based Methods		
	-$\mathcal{D}^E$,Unc	64.30±0.37	62.54±1.15
	-$\mathcal{D}^E$,Unc,Ver	63.97±0.40	61.35±0.98
LLaVA-13b	Uncertainty-based Methods		
	VERO	**69.42±0.46**	**67.30±0.29**
	-$\mathcal{D}^E$	69.23±0.45	66.21±0.54
	-$\mathcal{D}^E$,Ver	68.30±0.34	65.73±0.53
	Diversity-based Methods		
	-$\mathcal{D}^E$,Unc	68.21±0.67	65.84±1.28
	-$\mathcal{D}^E$,Unc,Ver	67.45±0.40	64.92±1.51

Table 4: **Weighted-F1** performance under different acquisition functions ablated from our approach.

Firstly, we find that the uncertainty-based methods generally outperform the diversity-based methods, indicating the importance of considering the capability of the LVLM itself to select challenging samples. Secondly, we also find that the performance of our model drops when removing the $\mathcal{D}^E$, demonstrating that easy samples improve the model's robustness. We also observe that the "-$\mathcal{D}^E$" model outperforms the "-$\mathcal{D}^E$,Ver" model, indicating the samples selected by self-verification provide greater training benefits than that by zero-shot prediction. The "-$\mathcal{D}^E$,Unc" model underperforms the "-$\mathcal{D}^E$" model, which is expected since ranking samples by their uncertainty scores results in more challenging samples than random selection. Meanwhile, the "-$\mathcal{D}^E$,Unc" model surpasses the "-$\mathcal{D}^E$,Unc,Ver" model, proving the effectiveness of selecting samples from the $\mathcal{D}_{\text{no}}$.

Further Analysis

Constrained by computational resources, additional analyses are performed using the LLaVA-7b backbone.

Error Analysis To examine the training benefits of our approach, we experiment with our acquisition function and its two ablated variants: "-$\mathcal{D}^E$" and "-$\mathcal{D}^E$,Unc,Ver" (denoted as "CDS"). The error analysis is demonstrated by the confusion matrices in Figure 3.

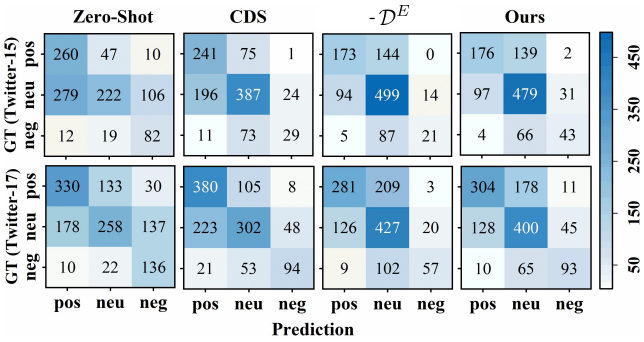

Figure 3: Confusion matrices for the test set of Twitter-2015 (left) and Twitter-2017 (right) under acquisition functions.

Firstly, let us take a look at the results on the Twitter-2015 dataset. The LVLM's zero-shot prediction shows strong performances on positive and negative sentiments while primarily failing on the neutral sentiment. We find the "CDS" model improves the performance on the neutral sentiment. However, it is still worse than the "-$\mathcal{D}^E$" model. This result proves the effectiveness of our acquisition function on enhancing the model's performance on challenging samples. Despite the large improvement on the neutral sentiment, the "-$\mathcal{D}^E$" model's performances on positive and negative sentiments drop compared to its zero-shot performances. This is due to focusing too heavily on challenging samples. By introducing a small set of samples in $\mathcal{D}_{\text{yes}}$ to balance training, VERO improves in performances on positive and negative sentiments, albeit sacrificing a bit of performance on neutral sentiment. Similar model behaviours can be observed in the Twitter-2017 dataset as well.

Impact of Hyperparameter λ The hyperparameter λ denotes the proportion of samples selected from $\mathcal{D}_{\text{no}}$. Now, we vary the λ from 1.0 to 0.0, to study its impact. $\lambda = 1.0$ means all samples are selected from $\mathcal{D}_{\text{no}}$, while $\lambda = 0.0$ means all samples are selected from $\mathcal{D}_{\text{yes}}$. The performance curves under different λ are shown in Figure 4 and the sentiment distributions of selected samples under different λ are shown in Figure 5.

Firstly, we find that the performance at $\lambda = 1.0$ significantly outperforms the performance at $\lambda = 0.0$, indicating that learning from samples where the LVLM firmly convinces shift in predictions via self-verification brings greater training benefits, which supports our motivation. Secondly, we find that the model achieves the best performance at $\lambda = 0.9$ on both datasets, indicating the importance of introducing a small proportion of most confident samples in

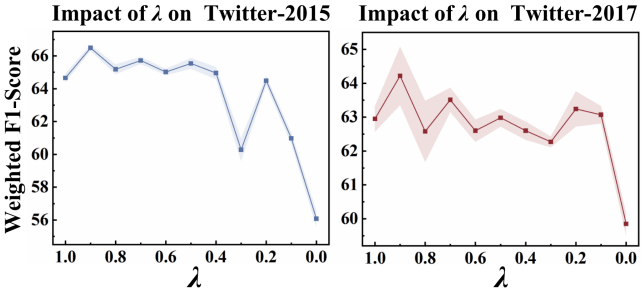

Figure 4: Weighted-F1 curves under different λ on the Twitter-2015 (left) and Twitter-2017 (right) datasets.

$\mathcal{D}_{\text{yes}}$. As observed in the "Error Analysis", these samples help balance the training process, reducing the model's tendency to overfit to challenging examples during fine-tuning.

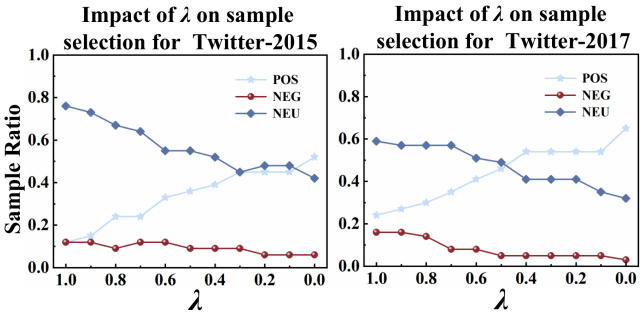

Figure 5: Sentiment distributions of the selected samples under different λ on the Twitter-2015 (left) and Twitter-2017 (right) datasets.

Turning to look at the sentiment distributions of selected samples under different λ, we find that at $\lambda = 1.0$ the neutral samples account for the highest proportion, about 0.75, while both positive and negative samples account for a very low proportion on the Twitter-2015 dataset. This result indicates that our acquisition function is clever to select the challenging samples against the zero-shot capability of the LVLM, which is corresponding to our observation in the "Error Analysis". As λ decreases, the proportion of neutral samples drops while the proportion of positive samples rises. This is expected since the LVLM's zero-shot performance on the positive sentiment is relatively higher. However, we find that the proportion of negative samples remains low as λ varies. This is due to the very low proportion of negative samples in the entire training set. A similar behaviour can also be observed on the Twitter-2017 dataset, where the LVLM's zero-shot performance on the positive sentiment is slightly lower than that on the Twitter-2015 dataset, resulting in a slightly higher proportion of positive samples at $\lambda = 1.0$.

Effectiveness of Self-Verification for Sample Acquisition We select challenging samples based on the uncertainty of the LVLM's self-verification, rather than its zero-shot prediction. However, selecting samples based on uncertainty of the model's predictions is a conventional approach in existing active learning studies. For a further analysis, we present the sentiment proportion and error rate of samples selected by the two strategies in Table 5, where selecting samples based on the uncertainty of the LVLM's self-verification and zero-shot prediction are denoted as "-$\mathcal{D}^E$" and "-$\mathcal{D}^E$,Ver" respectively.

Firstly, we find that the proportion of the neutral sentiment in selected samples by the "-$\mathcal{D}^E$" model is significantly higher than that by the "-$\mathcal{D}^E$,Ver" model on both datasets, while the proportions of neutral and positive sentiments for the "-$\mathcal{D}^E$,Ver" model are close on both datasets. It is evident that the distribution of samples selected by the "-$\mathcal{D}^E$" model aligns more closely with the model's zero-shot performance, with the neutral sentiment accounting for the majority of error cases. Furthermore, we observe that the error rate of the LVLM's zero-shot prediction on the samples selected by the "-$\mathcal{D}^E$" model is higher than that on the samples selected by the "-$\mathcal{D}^E$,Ver" model. This suggests that the self-verification mechanism is more effective at identifying challenging samples. We believe that self-verification represents a unique capability of LVLMs, distinct from zero-shot prediction. A majority of studies have also found that self-verification can effectively correct prediction errors. This ability offers more reliable cues for LVLMs to identify challenging samples, ultimately improving fine-tuning performance.

Dataset	Model	Sentiment Proportion			ER
		POS	NEG	NEU	
Twitter-15	-$\mathcal{D}^E$	0.12	0.12	0.76	51.52
	-$\mathcal{D}^E$,Ver	0.42	0.12	0.45	42.42
Twitter-17	-$\mathcal{D}^E$	0.24	0.16	0.59	54.05
	-$\mathcal{D}^E$,Ver	0.43	0.08	0.49	24.32

Table 5: The sentiment proportions of the selected samples for different models on the Twitter-2015 and Twitter-2017 datasets. "ER" denotes the error rate of the LVLM's zero-shot prediction on the selected samples.

Conclusion

In this paper, we solve the challenge of fine-tuning large vision-language models for few-shot multimodal aspect-level sentiment classification by introducing VERO, a novel acquisition function. VERO identifies challenging samples from a pool of unlabeled data by leveraging the zero-shot and self-verification capabilities of LVLMs, ensuring that the selected samples significantly contribute to the model's learning process. Our work also highlights the importance of strategically introducing easy samples, as a means to balance training and enhance the generalization of LVLMs. We experimented with VERO on two benchmark datasets and demonstrated that our approach consistently outperforms the recent competitive methods in both 1% and 7% few-shot settings. Further analyses demonstrate the advantage of VERO over existing sampling methods. The main limitation of our method is the underutilization of unlabeled data. Therefore, we plan to extend our approach to a semi-supervised framework in future work.

Acknowledgments

This research was partially supported by the National Key Research and Development Project of China No. 2021ZD0110700, the Key Research and Development Project in Shaanxi Province No. 2022GXLH01-03, the National Science Foundation of China No. (62250009, 62037001, 62406242, 62476215 and 62302380).

References

Cai, W.; Zhang, Y.; and Zhou, J. 2013. Maximizing expected model change for active learning in regression. In *2013 IEEE 13th international conference on data mining*, 51–60. IEEE.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *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.

Gao, T.; Fisch, A.; and Chen, D. 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 (Volume 1: Long Papers)*, 3816–3830.

Gero, Z.; Singh, C.; Cheng, H.; Naumann, T.; Galley, M.; Gao, J.; and Poon, H. 2023. Self-verification improves few-shot clinical information extraction. *arXiv preprint arXiv:2306.00024*.

Hosseini-Asl, E.; Liu, W.; and Xiong, C. 2022. A Generative Language Model for Few-shot Aspect-Based Sentiment Analysis. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 770–787.

Huang, S.-J.; Jin, R.; and Zhou, Z.-H. 2014. Active Learning by Querying Informative and Representative Examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(10): 1936–1949.

Huang, Y.; Song, J.; Wang, Z.; Zhao, S.; Chen, H.; Juefei-Xu, F.; and Ma, L. 2023. Look Before You Leap: An Exploratory Study of Uncertainty Measurement for Large Language Models.

Jian, Y.; Gao, C.; and Vosoughi, S. 2022. Contrastive Learning for Prompt-based Few-shot Language Learners. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5577–5587.

Khan, Z.; and Fu, Y. 2021. Exploiting BERT for multimodal target sentiment classification through input space translation. In *Proceedings of the 29th ACM international conference on multimedia*, 3034–3042.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 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*, 7871. Association for Computational Linguistics.

Li, X.; Lv, K.; Yan, H.; Lin, T.; Zhu, W.; Ni, Y.; Xie, G.; Wang, X.; and Qiu, X. 2023. Unified Demonstration Retriever for In-Context Learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4644–4668.

Li, Y.; Lan, X.; Chen, H.; Lu, K.; and Jiang, D. 2024. Multimodal PEAR Chain-of-Thought Reasoning for Multimodal Sentiment Analysis. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Li, Z.; Xu, B.; Zhu, C.; and Zhao, T. 2022. CLMLF:A Contrastive Learning and Multi-Layer Fusion Method for Multimodal Sentiment Detection. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Findings of the Association for Computational Linguistics: NAACL 2022*, 2282–2294. Seattle, United States: Association for Computational Linguistics.

Ling, Y.; Yu, J.; and Xia, R. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. *arXiv preprint arXiv:2204.07955*.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

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. *arXiv preprint arXiv:1907.11692*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Manakul, P.; Liusie, A.; and Gales, M. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9004–9017.

Meister, C.; Pimentel, T.; Wiher, G.; and Cotterell, R. 2023. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11: 102–121.

Nguyen, D. Q.; Vu, T.; and Nguyen, A. T. 2020. BERTweet: A pre-trained language model for English Tweets. *arXiv preprint arXiv:2005.10200*.

Raj, A.; and Bach, F. 2022. Convergence of uncertainty sampling for active learning (2021). *DOI: https://doi.org/10.48550/ARXIV*, 2110.

Roy, N.; and McCallum, A. 2001. Toward optimal active learning through sampling estimation of error reduction. int. conf. on machine learning.

Schein, A. I.; and Ungar, L. H. 2007. Active learning for logistic regression: an evaluation. *Machine Learning*, 68: 235–265.

Settles, B. 2009. Active learning literature survey.

Smith, S.; Patwary, M.; Norick, B.; LeGresley, P.; Rajbhandari, S.; Casper, J.; Liu, Z.; Prabhumoye, S.; Zerveas, G.;

Korthikanti, V.; et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Xiao, L.; Wu, X.; Yang, S.; Xu, J.; Zhou, J.; and He, L. 2023. Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis. *Information Processing & Management*, 60(6): 103508.

Xie, B.; Yuan, L.; Li, S.; Liu, C. H.; Cheng, X.; and Wang, G. 2022. Active learning for domain adaptation: An energy-based approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 8708–8716.

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

Yang, H.; Zhao, Y.; and Qin, B. 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*, 3324–3335.

Yang, H.; Zhao, Y.; and Qin, B. 2022b. Face-Sensitive Image-to-Emotional-Text Cross-modal Translation for Multimodal Aspect-based Sentiment Analysis. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3324–3335. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Yang, H.; Zhao, Y.; Wu, Y.; Wang, S.; Zheng, T.; Zhang, H.; Che, W.; and Qin, B. 2024a. Large Language Models Meet Text-Centric Multimodal Sentiment Analysis: A Survey. *arXiv preprint arXiv:2406.08068*.

Yang, J.; Xiao, Y.; and Du, X. 2024. Multi-grained fusion network with self-distillation for aspect-based multimodal sentiment analysis. *Knowledge-Based Systems*, 293: 111724.

Yang, L.; Wang, Z.; Li, Z.; Na, J.-C.; and Yu, J. 2024b. An empirical study of Multimodal Entity-Based Sentiment Analysis with ChatGPT: Improving in-context learning via entity-aware contrastive learning. *Information Processing & Management*, 61(4): 103724.

Yang, X.; Feng, S.; Wang, D.; Sun, Q.; Wu, W.; Zhang, Y.; Hong, P.; and Poria, S. 2023a. Few-shot Joint Multimodal Aspect-Sentiment Analysis Based on Generative Multimodal Prompt. In *Findings of the Association for Computational Linguistics: ACL 2023*, 11575–11589.

Yang, X.; Feng, S.; Wang, D.; Zhang, Y.; and Poria, S. 2023b. Few-shot multimodal sentiment analysis based on multimodal probabilistic fusion prompts. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6045–6053.

Yang, Y.; Ma, Z.; Nie, F.; Chang, X.; and Hauptmann, A. G. 2015. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113: 113–127.

Ye, J.; Wu, Z.; Feng, J.; Yu, T.; and Kong, L. 2023. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, 39818–39833. PMLR.

Yu, J.; and Jiang, J. 2019. Adapting BERT for Target-Oriented Multimodal Sentiment Classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 5408–5414. International Joint Conferences on Artificial Intelligence Organization.

Yu, Y.; and Zhang, D. 2022. Few-shot multi-modal sentiment analysis with prompt-based vision-aware language modeling. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Yu, Y.; Zhang, D.; and Li, S. 2022. Unified multi-modal pre-training for few-shot sentiment analysis with prompt-based learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, 189–198.

Zhu, J.; Wang, H.; Tsou, B. K.; and Ma, M. 2009. Active learning with sampling by uncertainty and density for data annotations. *IEEE Transactions on audio, speech, and language processing*, 18(6): 1323–1331.