WisdoM: Improving Multimodal Sentiment Analysis by Fusing Contextual World Knowledge

Anonymous ACL submission

Abstract

Sentiment analysis is rapidly advancing by utilizing various data modalities (*e.g.,* text, image). However, most previous works relied on superficial information, neglecting the incorporation of contextual world knowledge (*e.g.,* background information derived from but beyond the given image and text pairs) and thereby restricting their ability to achieve better multimodal sentiment analysis (MSA). In this paper, we proposed a plug-in framework named WisdoM, to leverage the contextual world knowledge induced from the large vision-language models (LVLMs) for enhanced MSA. WisdoM utilizes LVLMs to comprehensively analyze both images and corresponding texts, simultaneously generating pertinent *context*. To reduce the noise in the context, we also introduce a training-free contextual fusion mechanism. Experiments across diverse granularities of MSA tasks consistently demonstrate that our approach has substantial improvements (brings an average +1.96% F1 score among five advanced methods) over several state-of-the-art methods. The code will be released.

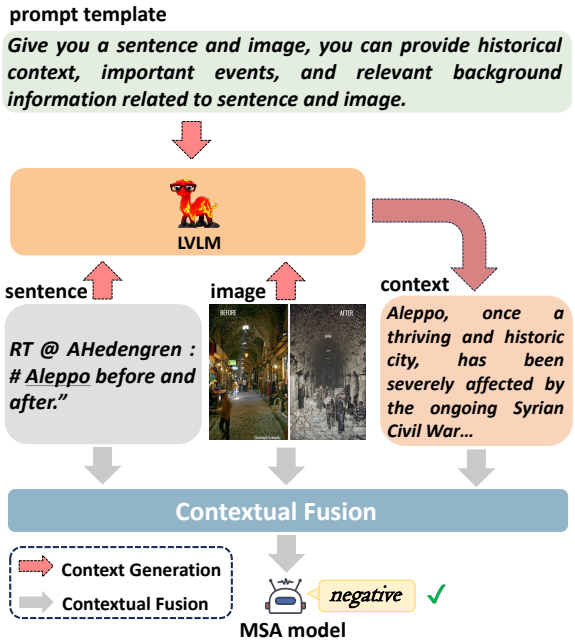


Figure 1: **The simple schematic of our method**. The sentiment polarity of Aleppo is negative, which is hard to directly predict by existing methods while our WisdoM predicts correctly via incorporating *context* generated by the world knowledge-rich LVLMs.

1 Introduction

Sentiment analysis (SA, Medhat et al., 2014; Wankhade et al., 2022) task focuses on identifying human sentiment polarity (Das and Singh, 2023). With the development of the Internet, the way people express their sentiments is not limited to text, but also includes multimodal data (*e.g.,* images). Detecting the sentiments of such information is challenging because of their short, informal nature, but utilizing the paired images can offer valuable insights. Therefore, how to make accurate sentiment classification by effectively marring modalities is at the core of multimodal sentiment analysis (MSA, Majumder et al., 2018; Wang et al., 2020).

Recent studies that improve MSA by carefully designing fusion strategies can be categorized into two types: 1) for aspect-level MSA, existing methods (Ju et al., 2021; Ling et al., 2022; Yang et al., 2022; Zhou et al., 2023) estimate whether the image contains conducive information by learning the relationship between images and text, and locally fuse the aspect-aware images; 2) for sentence-level MSA, researchers primarily focus on the global fusion at feature-level and decision-level (Hazarika et al., 2020; Yu et al., 2021; Han et al., 2021; Guo et al., 2022). Despite their empirical success, the above studies only consider the *superficial information*[1] between image and text (See the case "*Aleppo*" in Fig. 1), and sometimes it is difficult to predict the true polarity without their background world

[1]only reflects the surface or literal information without considering their deep (*e.g.,* historical and cultural) meaning.

knowledge ("*Aleppo has been severely affected by the ongoing Syrian Civil War...*" induced from the large vision-language models.). This raises the following question:

Could world knowledge boost MSA?

Take the test case in Fig. 1 as an instance, given a comparative image at different periods alongside a sentence, it is required to answer the question: *What's the sentiment polarity of "Aleppo"?*. We employ the current state-of-the-art (SOTA) MSA model (Zhou et al., 2023) as the backbone. As expected, even the SOTA model gives the wrong prediction (*i.e.,* "neutral" rather than the groundtruth–"negative") because of lacking the deeper knowledge of "Aleppo" (which is a city in Syria, in conjunction with the image, we might infer that the difference between the before and after of this city is caused by the Syrian war). Therefore, employing world knowledge is essential.

Correspondingly, we propose a plug-and-play framework to utilize the contextual world knowledge (simply induced from the large vision-language models) to complement the existing text-image pair with only superficial information, namely WisdoM. In particular, our WisdoM follows a three stages process: ❶ Prompt Templates Generation, ❷ Context Generation, and ❸ Contextual Fusion. In **stage 1**, we ask language models (*e.g.,* ChatGPT[2]) to generate prompt templates (See the Prompt Template "*Give you a sentence and image, you can ...*" in Fig. 1) which are used to construct instructions for **stage 2**. Then, we employ the advanced large vision-language model (LVLM, *e.g.,* LLaVA Liu et al., 2023b in **stage 2** to generate the contextual information (See the Context "*Aleppo, ... ongoing Syrian Civil War...*" in Fig. 1) based on the provided image and sentence. Note that, we refer to this contextual information as *context*. Due to the noisy nature of the derived *context*, we further introduce a training-free Contextual Fusion mechanism, in **stage 3**, to wisely incorporate the *context* for hard samples.

We validated our WisdoM on several benchmarks including Twitter2015, Twitter2017 (Yu and Jiang, 2019) and MSED (Jia et al., 2022) over several models: LLaVA-v1.5 (Liu et al., 2023a), MMICL (Zhao et al., 2023), mPLUG-Owl2 (Ye et al., 2023), AoM (Zhou et al., 2023), ALMT (Zhang et al., 2023b). Experiments demonstrate the effectiveness and universality of our approach, and extensive analyses provide insights into when and how our method works. Our main **contributions** are:

- We propose a plug-in framework WisdoM, leveraging the LVLM to generate explicit contextual world knowledge, to enhance the multimodal sentiment analysis ability.
- To achieve wise knowledge fusion, we introduce a novel contextual fusion mechanism to mitigate the impact of noise in the *context*.
- Experiments on three MSA benchmarks upon several advanced LVLMs, show that WisdoM brings consistent and significant improvements (up to +6.3% F1 score).

2 Related Work

2.1 Multimodal Sentiment Analysis

Multimodal Sentiment Analysis (MSA), diverging from conventional text-based approaches (Hussein, 2018), incorporates diverse modalities (*e.g.,* image, speech) to enhance sentiment classification accuracy (Soleymani et al., 2017). Numerous advanced models have been proposed, covering different levels of granularity, such as sentence and aspect:

Sentence-Level MSA. Wang et al. (2014) integrate images and text for microblog analysis. You et al. (2015) employ deep neural networks for textual and visual sentiment analysis. Zhao et al. (2019) explore image-text correlations in movie reviews. Li et al. (2020) propose a ConvTransformer, blending Transformer (Vaswani et al., 2017) and CNN technologies for sentiment analysis. Das and Singh (2022) propose a multi-stage multimodal method for the Assamese language, leveraging both text and images. Zhang et al. (2023b) present an advanced model, ALMT, enhancing multimodal analysis with a focus on language-guided features to handle irrelevant or conflicting data across different modalities.

Aspect-Level MSA. Yu and Jiang (2019) introduce TomBERT, leveraging two multimodal tweet datasets with target annotations, and proposed an aspect-oriented multimodal BERT model. Khan and Fu (2021) propose a two-stream model employing an object-aware transformer for image translation in the input space, followed by a single-pass

[2] https://chat.openai.com

non-autoregressive generation approach (Wu et al., 2020; Ding et al., 2021). Zhou et al. (2023) introduce an aspect-oriented network, AoM, designed to reduce visual and textual distractions arising from intricate image-text interactions.

Although existing approaches relying on sophisticated fusion techniques have achieved remarkable performance in MSA, their limitation lies in relying on superficial information for fusion, without incorporating contextual world knowledge.

2.2 Large Vision-Language Models

Large Vision-Language Models (LVLMs) are becoming a fundamental tool for solving general tasks (Li et al., 2023; Fu et al., 2023; Liu et al., 2023b; Zhao et al., 2023; Ye et al., 2023; Dai et al., 2023; Zhu et al., 2023; Alayrac et al., 2022; Chen et al., 2023). Liu et al. (2023a) introduce LLaVA, which connects the CLIP ViT-L/14 visual encoder (Dosovitskiy et al., 2020) with the large language model Vicuna (Chiang et al., 2023b) through a simple projection matrix, utilizing a two-stage instruction-tuning method. Zhao et al. (2023) propose MMICL to tackle the complexity of multi-modal prompts, focusing on improving LVLMs from both model and data viewpoints. Ye et al. (2023) propose mPUG-Owl2, a multi-modal language model designed for enhanced collaboration between modalities. It features a modular network where the language decoder acts as a universal interface for different modalities.

Here, we leverage LVLMs to enhance multi-modal sentiment analysis by generating relevant world knowledge. Additionally, we introduce a new training-free module named Contextual Fusion, designed to minimize noise in the context.

2.3 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) enhances Language Models by incorporating retrieved text, significantly improving performance in knowledge-based tasks, applicable in both fine-tuned and off-the-shelf scenarios (Gao et al., 2023; Gupta et al., 2024). Traditional RAG (Lewis et al., 2020), also known as Naive RAG, incorporates retrieval content to aid generation but faces key challenges: 1) varying retrieval quality, 2) generation of responses prone to inaccuracies, and 3) difficulties in coherently integrating retrieved-context with current tasks. To overcome the limitations of Naive RAG, advanced methods introduce more contextually rich information during inference. The DSP framework (Khattab et al., 2022) facilitates an intricate exchange between frozen LMs and retrieval models, improving context richness, while PKG (Luo et al., 2023) allows LLMs to retrieve relevant information for complex tasks without altering their parameters. The working mechanism of WisdoM is similar to RAG, but **different** at the following aspects: ① WisdoM utilizes LVLM to generate world knowledge to provide coherent and accurate context rather than retrieval, ② WisdoM incorporates a contextual fusion mechanism to diminish noise within the context. For additional experimental analysis and discussion, please refer to § 5.3.

Aspect-level Task Instruction I_{aspect}

Sentence: [sentence] Use the image as a visual aids to help you answer the question. What is the sentiment polarity of the aspect [aspect] in this sentence?
A). positive
B). neutral
C). negative
Answer with the option's letter from the given choices directly.

Sentence-level Task Instruction $I_{sentence}$

Sentence: [sentence] Use the image as a visual aids to help you answer the question. Given the sentence and image, what is the sentiment conveyed?
A). positive
B). neutral
C). negative
Answer with the option's letter from the given choices directly.

Figure 2: **Template of task instruction**.

3 Preliminary

We first describe the notation of the MSA, then review two typical frameworks for modelling the MSA tasks, where we experiment with our schema upon them: task-specific framework (Zhou et al., 2023; Zhang et al., 2023b) and general-purpose framework (Liu et al., 2023a; Ye et al., 2023).

Notation. Let $\mathcal{M}$ be a set of multimodal samples. Each sample $m_i \in \mathcal{M}$ consists of a sentence s_i and image v_i. For *aspect*-level MSA tasks, there are several aspects a_i which is a subsequence of s_i, *i.e.,* $a_i \in s_i$. We denote $f(\cdot)$ as the sentiment classifier. The output of $f(\cdot)$ is the sentiment polarity $y_i \in$ {negative, neutral, positive}, with corresponding predicted probability denoted as $P_i = \{p_i^{neg}, p_i^{neu}, p_i^{pos}\}$.

Task-Specific Framework. For *aspect*-level MSA tasks, the goal is to predict the sentiment polarity y_i and probability P_i for the specific aspect a_i conditioned on the (v_i, s_i), *i.e.,* $(y_i, P_i) =$

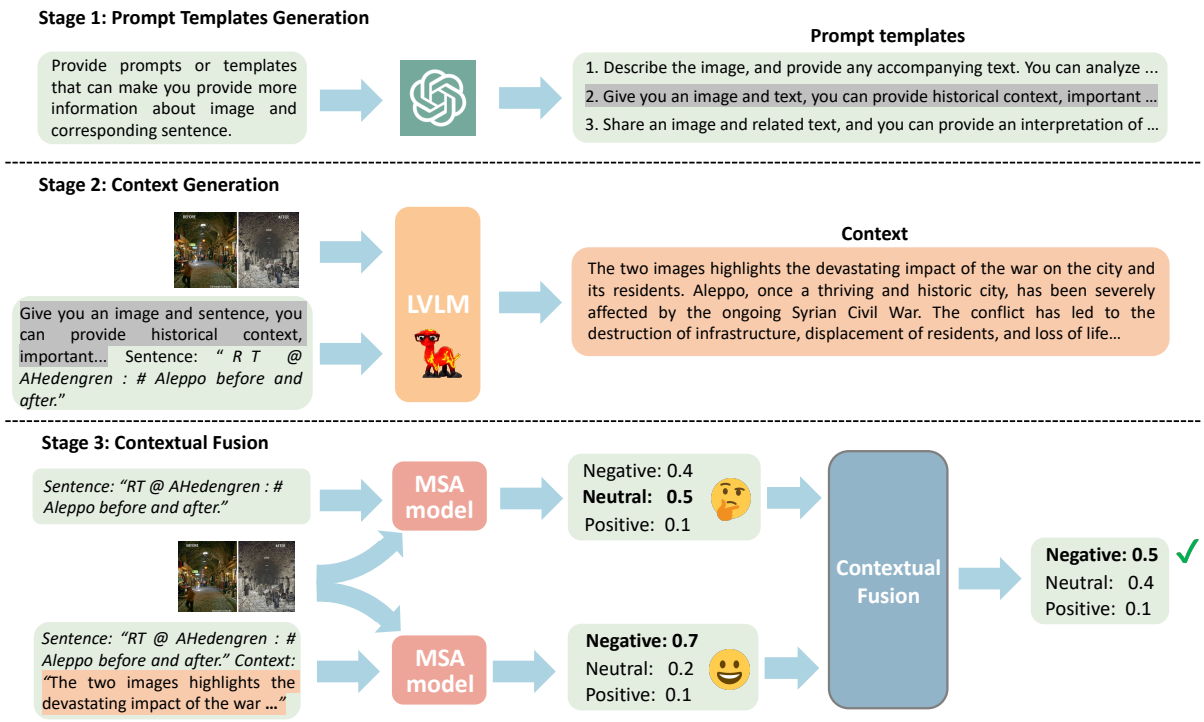


Figure 3: **Detailed illustration of our proposed schema** WisdoM with a running example. (1) Using ChatGPT to provide prompt templates. (2) We then prompt LVLMs to generate *context* using the prompt templates with image and sentence. (3) A training-free mechanism Contextual Fusion mitigates the noise in the *context*.

$f(v_i, s_i, a_i)$. For *sentence*-level MSA tasks, the y_i and P_i are predicted by sentence s_i alongside the image v_i, *i.e.*, $(y_i, P_i) = f(v_i, s_i)$.

General-Purpose Framework. To verify that our WisdoM works well on arbitrary architectures, we also apply WisdoM to general-purpose LVLMs. We follow Wang et al. (2023b) to construct the task instructions I_{aspect} and $I_{sentence}$ for each task to elicit its ability to the corresponding task. The task instructions are presented as single-choice questions with well-formatted options (Shown in Fig. 2). The LVLMs can be seen as a sentiment classifier $f(\cdot)$. Therefore, the *aspect*-level task and *sentence*-level task can be formulated as $(y_i, P_i) = f(I_{aspect}, v_i, s_i, a_i)$ and $(y_i, P_i) = f(I_{sentence}, v_i, s_i)$ respectively.

4 Methodology

Overview. Fig. 3 illustrates the overview of our method following three stages. In the (1) Prompt Templates Generation, we use large language models, particularly ChatGPT, to provide prompt templates. These prompt templates are fed into the LVLM with sentence s_i and image v_i to generate *context*, also called the (2) Context Generation. During (3) Contextual Fusion, we first compute the confidence, determining if the sample is uncertain (referred to as a hard sample). For hard samples, we fuse the predicted probability P_i with $\hat{P}_i$ which is obtained by incorporating *context*. Otherwise, we use P_i as the final prediction.

4.1 Stage 1: Prompt Templates Generation

The main purpose of this stage is to design the prompt templates used to generate the *context*, so that the LVLM can better understand our intention and thus provide a more comprehensive contextual world knowledge. Inspired by (Jiao et al., 2023; Zhong et al., 2023), we ask ChatGPT to provide the appropriate prompt templates. The prompt templates provided by ChatGPT consider world knowledge of different perspectives, including history, social and cultural, *etc.* We insert a "Sentence: [x]" at the end of the prompt template to place the input sentence s_i. The example of prompt templates are shown in Appendix A.5.

4.2 Stage 2: Context Generation

In the context generation stage, prompt templates in (1) are used to generate *context* that explicitly incorporates world knowledge based on given the image v_i and sentence s_i by LVLMs (Liu et al., 2023b; Ye et al., 2023). Specifically, we construct

Method	Twitter2015		Twitter2017	
	Acc.	***F1***	***Acc.***	***F1***
ESAFN (Yu et al., 2019)	73.4	67.4	67.8	64.2
TomBERT (Yu and Jiang, 2019)	77.2	71.8	70.3	68.0
CapTrBERT (Khan and Fu, 2021)	77.9	73.9	72.3	70.2
JML (Ju et al., 2021)	78.7	-	72.7	-
VLP-MABSA (Ling et al., 2022)	78.6	73.8	73.8	71.8
CMMT (Yang et al., 2022)	77.9	-	73.8	-
MMICL (Zhao et al., 2023)*	76.0	72.7	74.1	74.0
-w/ WisdoM	77.3 (+1.3)	74.2 (+1.5)	75.7 (+1.6)	75.7 (+1.1)
LLaVA-v1.5 (Liu et al., 2023a)*	77.9	74.3	74.6	74.3
-w/ WisdoM	78.9 (+1.0)	75.6 (+1.3)	75.6 (+1.0)	75.3 (+1.0)
mPLUG-Owl2 (Ye et al., 2023)*	76.8	72.3	74.2	73.0
-w/ WisdoM	77.3 (+0.5)	73.4 (+1.1)	74.5 (+0.3)	73.7 (+0.7)
AoM (Zhou et al., 2023)*	80.0	75.2	75.9	74.5
-w/ WisdoM	**81.5** (+1.5)	**78.1** (+2.9)	**77.6** (+1.7)	**76.8** (+2.3)

Table 1: **Comparison of our method (upon several advanced models) with existing works** on Twitter2015 and 2017 benchmarks. The highest results are highlighted in bold, and * indicates the reproduced results.

instruction by replacing the "[x]" in the prompt template with the sentence s_i. In addition, different LVLMs require a special token to indicate where the image v_i is inserted. Taking LLaVA (Liu et al., 2023b) as an example, we insert a special token "<image>" at the beginning of the instruction.

4.3 Stage 3: Contextual Fusion

We use the *sentence*-level task as an example. After obtaining the *context*, we can intuitively use predicted sentiment polarity $\hat{y}_i$ obtained by incorporating *context*, *i.e.,* $(\hat{y}_i, \hat{P}_i) = f(v_i, s_i, context)$. However, the *context* may contain irrelevant information that could disturb performance. Therefore, we first determine the hard samples and then fuse the P_i and $\hat{P}_i$ in the hard sample.

Determining the Hard Samples. Inspired by Zhang et al. (2021), we found that the ambiguous hard sample is commonly found around boundary areas of the sentiment polarity (*e.g.,* the boundary areas of negative and neutral). Therefore, given a sample m_i, we only consider the difference δ_i between the highest and the second highest probabilities to determine whether it is a hard sample:

$$\delta_i = 2 \times max(P_i) + min(P_i) - 1. \quad (1)$$

Then, we denote uncertain threshold α to select samples that not exceed α as hard, *i.e.,* $\mathcal{V}_{hard} = \{m_i | \delta_i \leq \alpha\}$, where we leave $\alpha = 0.3$ as default.

Fusion with Context. Inspired by (Li et al., 2022; O'Brien and Lewis, 2023), we take the convex combinations of P_i and $\hat{P}_i$ to obtain the final prediction $\tilde{P}_i$ for hard sample $m_i \in \mathcal{V}_{hard}$:

$$\tilde{P}_i = P_i + \beta \cdot (\hat{P}_i - P_i), \quad (2)$$

where β is an interpolation coefficient. Intuitively, $(\hat{P}_i - P_i)$ represents the information incorporating by *context*. β is used to control the proportion of information introduced in *context*. When $\beta \to 0$, the effect brought by *context* is completely ignored and vice versa. Note that, we use (y_i, P_i) as the final prediction when $m_i \notin \mathcal{V}_{hard}$. We study the impact of α and β in Appendix B.1.

5 Experiments

In this section, we apply WisdoM to *aspect*-level and *sentence*-level MSA tasks to verify its effectiveness and conduct extensive analysis to better understand the proposed method.

5.1 Experimental Settings

Datasets. For aspect-level tasks, our two benchmark datasets are Twitter2015 and Twitter2017 (Yu and Jiang, 2019). Twitter2015 and Twitter2017 are comprised of multimodal tweets, where each tweet incorporates textual content, an accompanying image, aspects contained within the tweet, and the sentiment associated with each aspect. Each aspect is assigned a label from the predefined set {negative, neutral, positive}. For sentence-level tasks, we

Method	MSED		
	Precision	***Recall***	***F1***
DCNN+AlexNet (Jia et al., 2022)	71.02	70.09	70.31
DCNN+ResNet (Jia et al., 2022)	74.73	74.73	74.64
BiLSTM+AlexNet (Jia et al., 2022)	78.73	79.22	78.89
BERT+AlexNet (Jia et al., 2022)	83.22	83.11	83.16
Multimodal Transformer (Jia et al., 2022)	83.56	83.45	83.50
ALMT (Zhang et al., 2023b)*	83.73	83.98	83.73
-w/ WisdoM	89.92 (+6.19)	90.14 (+6.16)	90.01 (+6.28)
MMICL (Zhao et al., 2023)*	86.24	86.55	86.17
-w/ WisdoM	89.92 (+3.68)	88.24 (+1.69)	88.85 (+2.68)
mPLUG-Owl2 (Ye et al., 2023)*	87.75	88.28	87.98
-w/ WisdoM	88.72 (+0.97)	89.35 (+1.07)	88.95 (+0.97)
LLaVA-v1.5 (Liu et al., 2023a)*	88.98	88.77	88.75
-w/ WisdoM	**90.58** (+1.60)	**90.41** (+1.64)	**90.48** (+1.73)

Table 2: **Performance of applying our WisdoM to advanced models** on MSED benchmark, with reference results from existing works. The best results are bolded, and the * denotes the reproduced results.

evaluate our WisdoM on a multimodal and multi-task dataset MSED (Jia et al., 2022), containing 9,190 text-image pairs.

Models. To demonstrate WisdoM generalizes across architectures and sizes, we experimented on MMICL (14B) (Zhao et al., 2023), LLaVA-v1.5 (13B) (Liu et al., 2023a), mPLUG-Owl2 (8.2B) (Ye et al., 2023), AoM (105M) (Zhou et al., 2023), ALMT (112.5M) (Zhang et al., 2023b).

Evaluation Metrics. For aspect-level tasks, we use Accuracy (***Acc.***) and macro-F1 (***F1***) following previous studies (Khan and Fu, 2021; Ju et al., 2021; Ling et al., 2022; Zhou et al., 2023). For the sentence-level task, we adopt precision, recall, and macro-F1 (***F1***) as evaluation metrics.

5.2 Main Results

Results of Aspect-Level MSA Task. We compare against advanced aspect-level MSA methods on Twitter2015 and Twitter2017, and report the results on Table 1. We show that our WisdoM achieves consistent and significant improvement on four models across two datasets. The WisdoM brings max 2.9% and 2.3% F1-gains on Twitter2015 and Twitter2017 respectively, showing that our method has a clear advantage.

Results of Sentence-Level MSA Task. As shown in Table 2, notably, our WisdoM upon LLaVA-v1.5 achieves the **new SOTA** F1 score: 90.48%, outperforming LLaVA-v1.5 (88.75%), mPLUG-Owl2 (87.98%), MMICL (86.17%) and ALMT (83.73%), consistently. The most significant improvement is achieved on ALMT (112.5M), where we bring an encouragingly 6.28% F1 gain, suggesting that contextual world knowledge is particularly crucial for small models in MSA.

Method	Twitter2017		MSED
	Acc.	***F1***	***F1***
MMICL			
Baseline	74.1	74.1	86.2
+ *context*	72.6 (-1.5)	74.4 (+0.3)	86.8 (+0.6)
+ CF	**75.7** (+1.6)	**75.7** (+1.6)	**88.9** (+2.7)
LLaVA-v1.5			
Baseline	74.6	74.3	88.8
+ *context*	73.7 (-0.9)	73.5 (-0.8)	87.8 (-1.0)
+ CF	**75.6**(+1.0)	**75.3** (+1.0)	**90.5** (+1.7)

Table 3: **Ablation study of *context* and its wise fusion module.** "CF" denotes our Contextual Fusion. We first only incorporate "*context*" and subsequently introduce the "contextual fusion" module.

5.3 Ablation Study

To better understand the role of each module in our method WisdoM, Table 3 presents the ablation results of the gradual addition of different components. Compared with the baselines (MMICL and LLaVA-v1.5), only adding *context* results in a slight performance degradation (-0.23% average F1 score), while with the help of our proposed Context Fusion mechanism, we achieve a consistent and significant improvement (+1.75% average F1 score).

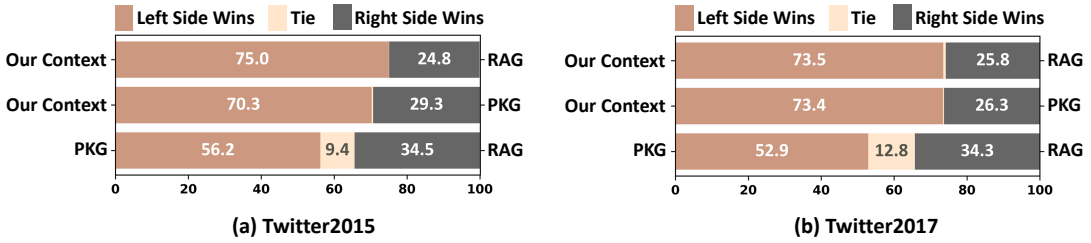


Figure 4: **Comparative winning rates of Our Context *v.s.* RAG-based methods on Twitter2015 and Twitter2017 benchmarks.** We can see that our contexts are better than the knowledge provided by RAG and PKG.

Method	Twitter2015		Twitter2017	
	Acc.	*F1*	*Acc.*	*F1*
RAG	75.1	71.0	71.9	70.7
PKG	76.2	72.4	72.8	71.5
Our Context	**76.3**	**72.7**	**73.7**	**73.5**

Table 4: **Comparative results of *context* generated by RAG, PKG, and our stage 2**. We incorporate the *contexts* on LLaVA-v1.5 directly.

Through (error) case studies in Appendix C.2, we found that containing irrelevant information in the original context leads to bad performance, showing the necessity of further context fusion mechanism.

Analyzing Effects of *Context*. To further analyse the effect of *context*, we compare our *context* generated in **stage 2** with the document retrieved by RAG (Lewis et al., 2020) and knowledge generated by PKG (Luo et al., 2023), collectively termed as "context" for simplicity. The assessment focuses on the context's pertinence to a given image v and sentence s, alongside its applicability in MSA tasks. We employ the LLM-based metric, *i.e.,* **LLM-as-a-Judge** (Chiang et al., 2023a) to quantify the quality of *context*. Specifically, we craft a prompt for GPT-4V (OpenAI, 2023) to compare our context with that provided by RAG and PKG. The detailed experimental settings can be found in Appendix A.7. As shown in Fig. 4, our context "Our Context" significantly beats the RAG and PKG counterparts, demonstrating its superiority. We provide the examples of context provided by different methods in Appendix C.1. Besides analyzing the contexts' pertinence of different methods, we report their downstreaming performance on MSA tasks in Table 4. Clearly, the MSA performance with our context is the best. The results above illustrate that ***our method can provide more precise context, thus bringing better MSA performance.***

Analyzing Effects of Contextual Fusion. In Table 5, we experimentally explore different fusion

Method	Twitter2015	
	Acc.	*F1*
AoM	80.00	75.20
-w/ Average	80.33 (+0.33)	70.69 (-4.51)
-w/ Max	80.56 (+0.56)	76.81 (+1.61)
-w/ JS	80.78 (+0.78)	77.39 (+2.19)
-w/ CXMI	79.66 (-0.34)	75.74 (+0.54)
-w/ **CF**	**81.45** (+1.45)	**78.12** (+2.92)
MMICL	75.98	72.71
-w/ Average	75.53 (-0.45)	72.33 (-0.38)
-w/ Max	75.42 (-0.56)	72.16 (-0.55)
-w/ JS	77.09 (+1.11)	73.92 (+1.21)
-w/ CXMI	76.31 (+0.33)	73.37 (+0.66)
-w/ **CF**	**77.32** (+1.34)	**74.18** (+1.47)

Table 5: **Ablation study of different fusion strategies.** "JS" denotes Jensen-Shannon divergence. "CF" denotes our Contextual Fusion.

strategies, including $mean(P_i, \hat{P}_i)$ ("**Average**"), $max(P_i, \hat{P}_i)$ ("**Max**"), Jensen-Shannon divergence (Menéndez et al., 1997) ("**JS**"), conditional cross-mutual information score f_{cxmi} (Wang et al., 2023c) ("**CXMI**"), and our Context Fusion ("**CF**"). For JS, we calculate the JS divergence of P_i with the uniform distribution to serve as the fusion weight, *i.e.,* β. As for CXMI, if $f_{cxmi} > 1.1$[3], we adopt (y_i, P_i) as our ultimate prediction, otherwise, we use $(\hat{y}_i, \hat{P}_i)$ as the final prediction. The results show that ***our Contextual Fusion module performs the best among all competitive alternatives, confirming its effectiveness.***

5.4 When and Why Does Our Method Work?

To better understand when and why our method works? we conduct extensive analysis to provide the following insights:

[3]In preliminary study, we grid-searched values ranging from 0.5 to 2.0, and 1.1 performs best on the dev set, thus leaving as our default setting.

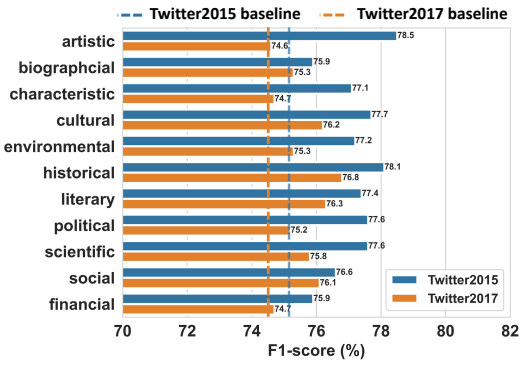


Figure 5: **Effects of different types of world knowledge.** We analyse the effect of different types of world knowledge by applying WisdoM to AoM. The orange dash line and blue dash line represent the F1-score of vanilla AoM on Twitter2015&2017 respectively.

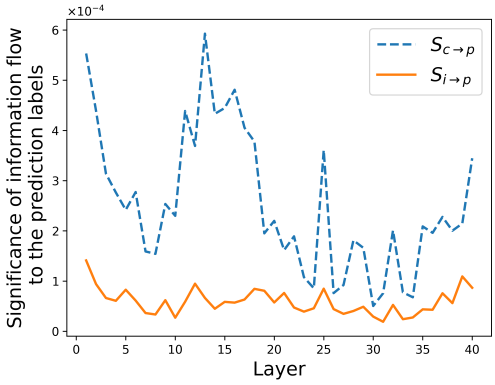


Figure 6: **Comparison of context ($S_{c\to p}$) and input's ($S_{i\to p}$) correlation to the final prediction across layers** in LLaVA-v1.5 on Twitter2015. High score means a strong correlation with final decision-making.

World Knowledge Enhances MSA, while Domain-related Knowledge is more Helpful. We explore the effects of different types of world knowledge from 11 perspectives (artistic, biographical, etc.), upon AoM. Fig. 5 shows that 1) nearly all types of extra knowledge enhance the MSA performance, 2) historical knowledge significantly enhances MSA on Twitter datasets. We conjecture that this is because tweets often convey sentiment through historical references, and domain-related world knowledge is more helpful. To verify our hypothesis, we conduct experiments on Financial PhraseBank (FPB, Malo et al., 2014) dataset and Twitter financial sentiment validation dataset (Magic, 2022), and find that financial and historical knowledge are the two types of world knowledge with the greatest gain (bringing 4.2% and 2.1% improvements on average F1 scores, respectively), confirming our conjunction. Details can be refer to Appendix B.5.

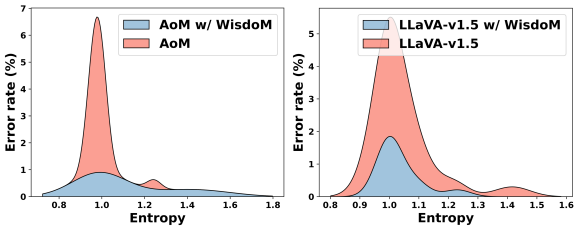


Figure 7: **Visualizing of error rate for hard samples** ($\delta \leq 0.3$) on Twitter2015 benchmark.

Context is Dominant for Prediction. To draw a clearer picture of the information flow for MSA, we calculate $S_{c\to p}$ and $S_{i\to p}$ to represent the mean significance of information flow from *context* (c) and original input (i) to the prediction labels (p) respectively (Simonyan et al. (2014); Wang et al. (2023a), detailed in Appendix A.8). Fig. 6 reveals that the significance of the information flow from *context* to the prediction label is remarkably high than its *input* counterpart, suggesting that *context* outweighs image and sentence when making the final prediction. Results of other datasets show identical trends and can be found in Appendix B.6.

WisdoM Effectively Reduces the Uncertainty of Hard Samples. To further explore how our WisdoM affects the hard samples, we visualize the error rate within high entropy in Fig. 7. After integrating WisdoM, the error rate is significantly decreased compared with the baseline (AoM and LLaVA-v1.5), demonstrating that our WisdoM effectively reduces the uncertainty of hard samples and improves performance.

6 Conclusion

In this paper, we propose a simple and effective plug-in framework WisdoM to enhance the ability of MSA. WisdoM follows three stages: Prompt Templates Generation, Context Generation, and Contextual Fusion. Firstly, we employ ChatGPT to generate prompt templates, enabling the large vision-language model to produce pertinent contextual world knowledge (referred to as *context*) derived from both the image and sentence. Subsequently, we incorporate this *context* using Contextual Fusion, minimizing the introduction of noise in the process. We empirically demonstrated the effectiveness and universality of the WisdoM on a series of widely used benchmarks.

Limitation

Our work has several potential limitations. Although our experiments revealed the enhanced performance of the multimodal sentiment analysis model with the introduction of context, the adaptive incorporation of context requires further exploration. Additionally, there is a need for further research into introducing models capable of handling additional modalities (*e.g.,* speech).

Ethics Statement

We take ethical considerations very seriously. This paper focuses on improving multimodal sentiment analysis by making the most of large vision-language models. All experiments are conducted on open datasets and the findings and conclusions of this paper are reported accurately and objectively. Thus, we believe that this research will not pose ethical issues.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023a. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023b. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint*.

Ringki Das and Thoudam Doren Singh. 2022. A multi-stage multimodal framework for sentiment analysis of assamese in low resource setting. *ESWA*.

Ringki Das and Thoudam Doren Singh. 2023. Multimodal sentiment analysis: A survey of methods, trends and challenges. *CSUR*.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F Wong, Dacheng Tao, and Zhaopeng Tu. 2021. Understanding and improving lexical choice in non-autoregressive translation. In *ICLR*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint*.

Jiwei Guo, Jiajia Tang, Weichen Dai, Yu Ding, and Wanzeng Kong. 2022. Dynamically adjust word representations using unaligned multimodal information. In *ACM MM*.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. 2020. Accelerating large-scale inference with anisotropic vector quantization. In *ICML*.

Aman Gupta, Anup Shirgaonkar, Angels de Luis Balaguer, Bruno Silva, Daniel Holstein, Dawei Li, Jennifer Marsman, Leonardo O Nunes, Mahsa Rouzbahman, Morris Sharp, et al. 2024. Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. *arXiv preprint*.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint*.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *ACM MM*.

Doaa Mohey El-Din Mohamed Hussein. 2018. A survey on sentiment analysis challenges. *JKSUES*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *TMLR*.

Ao Jia, Yu He, Yazhou Zhang, Sagar Uprety, Dawei Song, and Christina Lioma. 2022. Beyond emotion: A multi-modal dataset for human desire understanding. In *NAACL*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint*.

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 *EMNLP*.

Zaid Khan and Yun Fu. 2021. Exploiting bert for multimodal target sentiment classification through input space translation. In *ACM MM*.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*, 33.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint*.

Pengfei Li, Peixiang Zhong, Jiaheng Zhang, and Kezhi Mao. 2020. Convolutional transformer with sentiment-aware attention for sentiment analysis. In *IJCNN*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint*.

Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. *arXiv preprint*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *ICLR*.

Ziyang Luo, Can Xu, Pu Zhao, Xiubo Geng, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Augmented large language models with parametric knowledge guiding. *arXiv preprint*.

Neural Magic. 2022. Twitter financial news sentiment.

Macedo Maia, Siegfried Handschuh, Andre Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra. Balahur. 2018. Financial opinion mining and question answering. In *WWW*.

Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, Erik Cambria, and Soujanya Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-based systems*.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *JASIST*.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*.

ML Menéndez, JA Pardo, L Pardo, and MC Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *NeurIPS*.

Sean O'Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models. *arXiv preprint*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint*.

K Simonyan, A Vedaldi, and A Zisserman. 2014. Deep inside convolutional networks: visualising image classification models and saliency maps. In *ICLR*.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *IVC*.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *SIGIR*, SIGIR '21, page 2443–2449, New York, NY, USA. Association for Computing Machinery.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS*.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023a. Label words are anchors: An information flow perspective for understanding in-context learning. In *EMNLP*.

Min Wang, Donglin Cao, Lingxiao Li, Shaozi Li, and Rongrong Ji. 2014. Microblog sentiment analysis based on cross-media bag-of-words model. In *ICIMCS*.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023b. Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint*.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023c. Learning to filter context for retrieval-augmented generation. *arXiv preprint*.

Zilong Wang, Zhaohong Wan, and Xiaojun Wan. 2020. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *WWW*.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*.

Di Wu, Liang Ding, Fan Lu, and Jian Xie. 2020. Slotrefine: A fast non-autoregressive model for joint intent detection and slot filling. In *EMNLP*.

Li Yang, Jin-Cheon Na, and Jianfei Yu. 2022. Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis. *Information Processing & Management*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint*.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Joint visual-textual sentiment analysis with deep neural networks. In *ACM MM*.

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

Jianfei Yu, Jing Jiang, and Rui Xia. 2019. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *TASLP*.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *AAAI*.

Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023a. Enhancing financial sentiment analysis via retrieval augmented large language models. In *ICAIF*.

Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. 2021. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *CVPR*.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023b. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *EMNLP*.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint*.

Ziyuan Zhao, Huiying Zhu, Zehao Xue, Zhao Liu, Jing Tian, Matthew Chin Heng Chua, and Maofu Liu. 2019. An image-text consistency driven multimodal sentiment analysis approach for social media. *Information Processing & Management*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint*.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint*.

A Implementation Details

A.1 Details of Datasets

In this work, we conduct experiments covering aspect-level (Twitter2015 and Twitter2017) and sentence-level (MSED) MSA benchmarks. We present the statistics of all datasets in Table 6. Then, each dataset is described as:

Twitter2015 & 2017. Twitter2015 and Twitter2017 datasets encompass multimodal tweets, wherein each tweet comprises textual content, an accompanying image, embedded aspects, and corresponding sentiment annotations for each aspect. These aspects are categorized into labels from the set {negative, neutral, positive}.

MSED. MSED comprises 9190 pairs of text and images sourced from diverse social media platforms, including but not limited to Twitter, Getty Images, and Flickr.

A.2 Model Details

To verify the effectiveness of our WisdoM, we apply it within two standard frameworks for modeling the MSA tasks. For task-specific framework, we conduct experiment using the AoM (105M, Zhou et al., 2023) and ALMT (112.5M, Zhang et al., 2023b). For general-purpose framework, we experiment on mPLUG-Owl2 (8.2B, Ye et al., 2023), LLaVA-v1.5 (13B, Liu et al., 2023a) and MMICL (14B, Zhao et al., 2023). The detailed model information is listed in Table 7.

A.3 Training Details

For a fair comparison, we fine-tune models introduced in § 5.1 on target datasets. We optimize our model with AdamW (Loshchilov and Hutter, 2018). The learning rate is grid searched in {1e-5, 2e-5, 7.5e-5, 1e-4}, and batch size is in {16, 32, 256}. It should be noted that *contexts* are not incorporated during the training stage.

A.4 Inference Details

For task-specific methods (Zhou et al., 2023; Zhang et al., 2023b), we employ the same inference procedures as their original works. For general-purpose methods (Liu et al., 2023a; Zhao et al., 2023; Ye et al., 2023), we compute the likelihood that an LVLM generates the content of this choice given the question. We select the choice with the highest likelihood as the model's prediction.

A.5 Prompt Templates of World Knowledge

In Table 8, we list the prompt templates which are provided by ChatGPT in prompt templates generation stage. In the context generation stage, we replace "[x]" with sentence s_i and a special image token (*e.g.,* "<image>") is inserted at the beginning of the prompt template.

A.6 Pseudo Code for Contextual Fusion

Algorithm 1 provides the pseudo-code of Contextual Fusion. The simplicity of our method requires only a few lines of code.

Algorithm 1: Python-style pseudo-code for Contextual Fusion.

```
# p_o: numpy array, represents P_i
# p_c: numpy array, represents P̂_i
# alpha: float, the threshold of choosing hard sample
# beta: float, interpolation coefficient
def contextual_fusion(p_o, p_c, alpha, beta):
    # step1 : calculate delta
    delta = 2 * np.max(p_o) + np.min(p_o) - 1
    if delta > alpha:
        return p_o
    # step2 : calculate the final prediction
    p_f = p_o + beta * (p_c - p_o)
    return p_f
```

A.7 RAG Experiment Setup

In § 5.3, we compare our WisdoM with two RAG-based methods: (1) a naive RAG (Lewis et al., 2020), which initially searches for relevant documents related to a given question and then employs a generator to predict an answer; (2) PKG (Luo et al., 2023), an advanced RAG method that incorporates a knowledge-guided module, allowing for information retrieval without modifying the parameters of language models. The experimental setting is described as below.

Knowledge Sources. The source of knowledge for our experiments is the Wikipedia-Image-Text (WIT) dataset (Srinivasan et al., 2021), which has published in 2021. This dataset comprises images from Wikipedia, along with their alt-text captions and contextualized text passages.

Methods. For naive RAG, we use off-the-shelf Contriever-MSMARCO (Izacard et al., 2022) as the textual retriever and CLIP-ViT as the visual retriever. Specifically, we utilize CLIP-ViT and Contriever-MSMARCO for encoding query q and knowledge source, and employ Maximum Inner Product Search (MIPS, Guo et al., 2020) to find the five nearest neighbors (*knowledge*) to q within the knowledge source. For PKG, we use LLaVA-v1.5

	Twitter2015			Twitter2017			MSED		
	#Train	***#Dev***	***#Test***	***#Train***	***#Dev***	***#Test***	***#Train***	***#Dev***	***#Test***
Negative	368	149	113	416	144	168	1939	308	613
Neutral	1883	679	607	1638	517	573	1664	294	569
Positive	928	303	317	1508	515	493	2524	419	860
Total	3179	1122	1037	3562	1176	1234	6127	1021	2042

Table 6: Dataset statistics.

Model	Model Type	Source
AoM	Task-specific	https://github.com/SilyRab/AoM
ALMT	Task-specific	https://github.com/Haoyu-ha/ALMT
LLaVA-v1.5	General-purpose	https://huggingface.co/liuhaotian/llava-v1.5-13b
mPLUG-Owl2	General-purpose	https://huggingface.co/MAGAer13/mplug-owl2-llama2-7b
MMICL	General-purpose	https://huggingface.co/BleachNick/MMICL-Instructblip-T5-xxl

Table 7: Information of all models used in this study.

(13B) as knowledge-guided module fine-tuning on WIT and then generate the *knowledge* according to the image v and sentence s. Subsequently, we directly predict sentiment polarity by incorporating *knowledge*.

Evaluation Setting. To evaluate the relevance of the context to a specific image v and sentence s, we employ LLM-based metric, *i.e.,* LLM-as-a-Judge (Chiang et al., 2023a). Specifically, a prompt is crafted for GPT-4V (OpenAI, 2023) to assess the winning rates of our context in comparison to those derived from RAG-based methods. The detailed prompt can be found in Table 9. We also evaluate performance on MSA tasks.

A.8 Calculation of $S_{c\to p}$ and $S_{i\to p}$

To measure the significance of *context* and original input (*i.e.,* image and sentence), we use $S_{c\to p}$ and $S_{i\to p}$ for highlighting critical token interactions. Following previous work (Simonyan et al., 2014; Wang et al., 2023a), we use the Taylor expansion (Michel et al., 2019) to calculate the score for each element of the attention matrix:

$$I_l = \sum_h |A_{h,l}^{\mathsf{T}} \frac{\partial \mathcal{L}(x)}{\partial A_{h,l}}|. \tag{3}$$

Here, $A_{h,l}$ represents the value of the attention matrix of the h-th attention head in the l-th layer, x is the input, and $\mathcal{L}(x)$ is the loss function. We calculate the saliency matrix I_l for the l-th layer by averaging across all attention heads. $I_l(k,j)$ denotes the importance of the information flow from the j-th word to the k-th word. We propose two quantitative metrics based on I_l. The definitions of the two quantitative metrics are below.

$S_{c\to p}$, the mean significance of information flow from *context*(c) to the prediction label (p).

$$S_{c\to p} = \frac{\sum_{(k,j)\in C_{cp}} I_l(k,j)}{|C_{cp}|}, \quad C_{c,p} = \{(c,p) : c \in \textit{context}\}. \tag{4}$$

$S_{i\to p}$, the mean significance of information flow from image (v) and sentence (t) to the prediction label (p).

$$S_{i\to p} = \frac{\sum_{(k,j)\in C_{ip}} I_l(k,j)}{|C_{ip}|}, \quad C_{i,p} = \{(i,p) : i \in [v,t]\}. \tag{5}$$

$S_{c\to p}$ and $S_{i\to p}$ indicate the intensity of information aggregation onto the prediction label. A high S demonstrates strong information for final decision-making.

B Additional Experimental Results

B.1 Hyperparameter Selection

In stage 3, Contextual fusion has two major hyperparameters interpolation coefficient β and uncertain threshold α. To systematically study the impact of β and α, we first fix $\alpha = 0.3$ and search different configurations for the β. Then we fix β to the optimal value in the Twitter2015 dev set and search for α.

Type name	Prompt template
Artistic	Identify and discuss any artistic movements or styles that influenced the creation of the image. Explore how the artist's choice of style aligns with or deviates from prevalent artistic trends of the time. Sentence: [x].
Biographcial	Delve into the backgrounds of individuals associated with the image and text. Explore the biographies of artists, authors, or other relevant figures, and discuss how their life experiences shaped the creation and interpretation of the work. Sentence: [x]
Character	Focus on characters within the image and sente. Analyze their personalities, relationships, and potential character development. Discuss how the visual and textual elements contribute to character portrayal. Sentence: [x]
Cultural	Explore how the image and sentence reflect or represent aspects of a particular culture. Discuss the cultural significance, traditions, or values implied by the elements in the image and sentence. Sentence: [x]
Environmental	Examine the environmental elements within the image and sentence, discussing ecological factors, environmental changes, or the relationship between human activities and the depicted setting. Sentence: [x]
Historical	Give you an image and sentence, you can provide historical context, important events, and relevant background information related to the image and sentence. Sentence: [x]
Literary	Conduct a literary analysis of the sentence, exploring themes, symbolism, and narrative techniques. Discuss how the words complement or contrast with the visual elements in the image. Sentence: [x]
Political	Examine the political during the time the image and text were created. Discuss any political events, movements, or ideologies that may have influenced the content and tone of the work. Sentence: [x]
Scientific	Investigate the scientific elements within the image, delving into discoveries, advancements, or breakthroughs related to the subject matter mentioned in the sentence. Sentence: [x]
Social	Investigate the image and text as a form of social commentary. Analyze how the work reflects or critiques social issues, norms, or inequalities prevalent at the time of creation. Sentence: [x]
Financial	Give you a sentence and image, you should provide related financial knowledge. Sentence: [x]

Table 8: **Example of the prompt template** generated in stage 1 and used in stage 2.

Parameter Analysis of β. As shown in Fig. 8, we find that: 1) different models are sensitive to the values of β. 2) β values within the range of $[0.4, 0.5]$ demonstrate strong performance among various models. 3) Excessive values of β result in performance degradation. We conjecture that this decline may be caused by introducing excessive noise within the *context*. For a fair comparison, we select the optimal value of β on the dev set for evaluation.

Parameter Analysis of α. Fig. 9 shows that: 1) models exhibit insensitivity to the value of α. 2) α values within the range of $[0.3, 0.4]$ exhibit strong performance across diverse models. 3) There is no significant decrease in trends with increasing α. Thus, we set $\alpha = 0.3$ as our default setting.

B.2 Scalability of WisdoM

Our plug-in method is data- and model-agnostic, therefore, it is expected to be highly scalable. Here we scale our WisdoM up to different model sizes and data volumes.

Performance on Different Model Sizes. We experiment with scaling the model size to see if there are ramifications when operating at a larger scale. Fig. 10 (a) reveals that the performance increases as the LVLM size increases. In addition, we found that as the size of the model increased, the performance gains became more pronounced.

Evaluation Prompt
System: In this task, you will be asked to compare the relevance of two paragraphs to determine which one is more pertinent to the provided source sentence and image and benefits the sentiment analysis task the most. There are three options for you to choose from: 1. Context1 is better. If you think Context 1 is more relevant to the source sentence and image and benefits the sentiment analysis task. 2. Context2 is better. If you think Context 2 is more relevant to the source sentence and image and benefits the sentiment analysis task. 3. Context1, Context2 are the same: If you think Context1, Context2 have the same relevance to the source sentence and image, then choose this option. **Your answer is a JSON DICT that has one key: answer. For example: {"answer": "x. Context x is better."}** **INPUT** Source Sentence: "[s]" Context1: "[x1]" Context2: "[x2]" **OUTPUT**

Table 9: **The Evaluation Prompt** we used for GPT-4V. [s] represents the input sentence. [x1] and [x2] represent the context generated by different methods.

Performance on different Data Volumes. We conduct experiments on different ratios of training data to verify the robustness of WisdoM. As shown in Fig. 10 (b), we find that even when only 25% of the training data was used, our method WisdoM resulted in a 0.29% improvement.

B.3 Exploring Contexts Derived from Various LVLMs

To explore the relationship between *context* and LVLMs capability, we conduct experiments on AoM using *contexts* derived from mPLUG-Owl2 (8.2B) and LLaVA-v1.5 (13B). As depicted in the Table 10, the results show that **the stronger the capability of LVLMs, the more accurate and helpful the generated context is for MSA.**

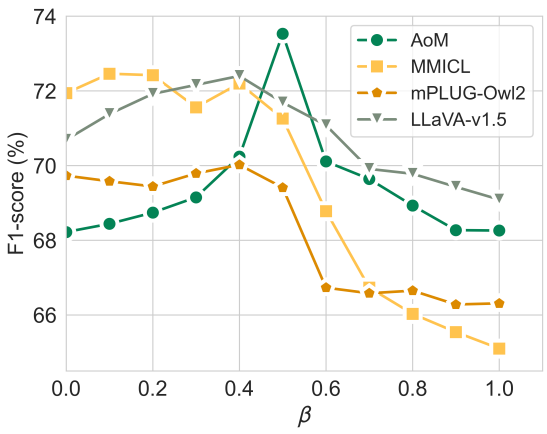


Figure 8: **Effect of interpolation coefficient β.** We show how F1-score changes when varying the β values.

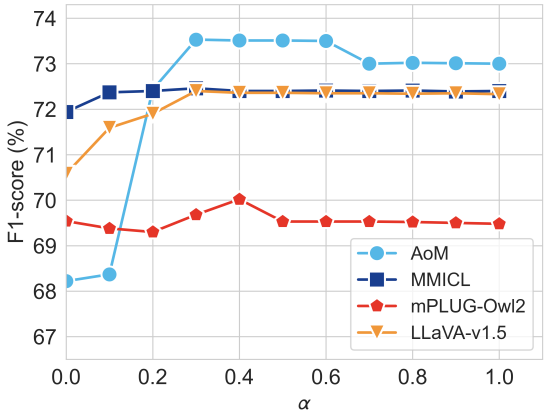


Figure 9: **Impact of uncertain threshold α**, illustrating how the F1-score changes when varying α.

B.4 Training with Context

To investigate the impact of context incorporation during training, we conduct experiments with three different setups: 1) vanilla, which without incorporation of context during training and inference; 2) training with context, which incorporation of context into training; 3) our WisdoM, to leverage world knowledge during inference phrase. The results, depicted in Fig. 11, demonstrate that our WisdoM consistently outperforms the vanilla across all models. However, the performance of training with context, specifically AoM and mPLUG-Owl2, experienced a significant decline. The possible reason is that the noisy nature of the context used during training. We conjecture that developing a method to effectively filter out this noise could potentially ameliorate the performance of models trained with context.

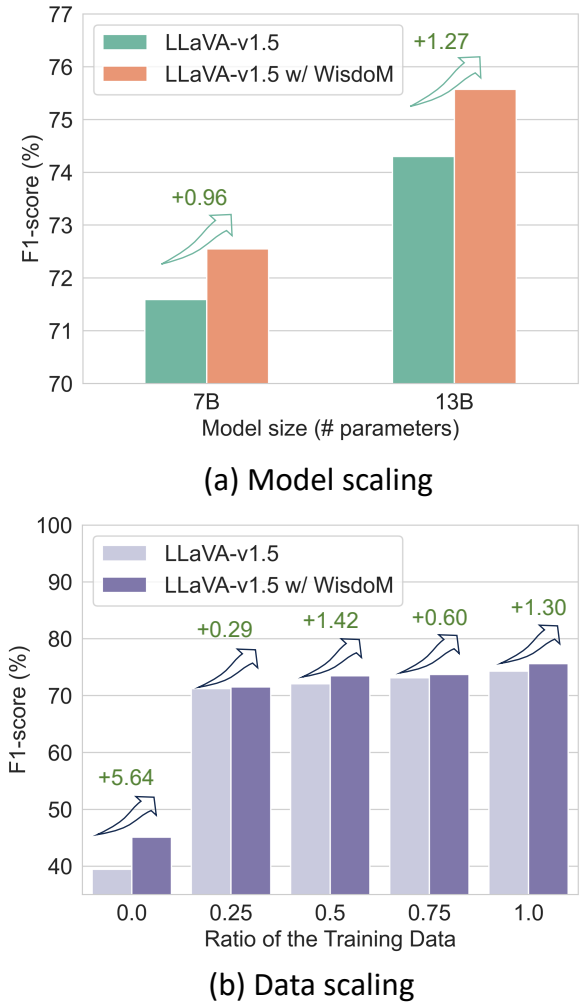


Figure 10: **Performance of scaling WisdoM** on Twitter2015 with different a) model and b) data scales.

Method	Twitter2015		Twitter2017	
	Acc.	*F1*	*Acc.*	*F1*
mPLUG-Owl2	76.8	72.3	74.2	73.0
LLaVA-v1.5	77.9	74.3	74.6	74.3
AoM	80.0	75.2	75.9	74.5
-w/ Context$_m$	81.2	77.8	76.4	75.2
-w/ Context$_L$	**81.5**	**78.1**	**77.6**	**76.8**

Table 10: **Comparison of contexts derived from different LVLMs.** "Context$_m$" represents the context derived from mPlUG-Owl2. "Context$_L$" represents the context derived from LLaVA-v1.5.

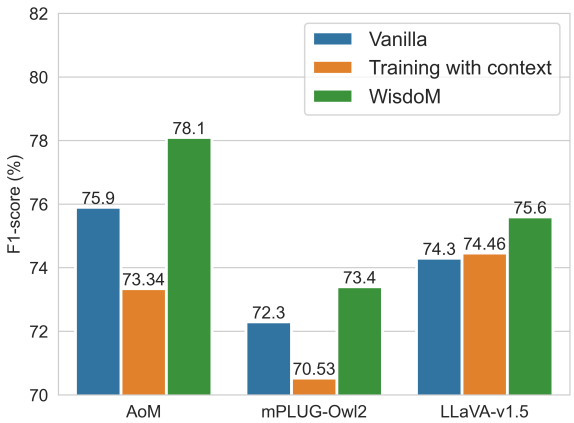


Figure 11: **Comparison of WisdoM with vanilla and training with context on Twitter2015.** We utilize the historical world knowledge generated by LLaVA-v1.5 for training and observed significant decreases in AoM and mPLUG-Owl2.

B.5 Additional World Knowledge Analysis

In § 5.4, we analyzed the effect of different types of world knowledge in MSA tasks and conjecture that domain-related world knowledge is more helpful. To verify our hypothesis, we perform experiment on financial sentiment analysis. The experimental setting is described as below.

Datasets. Following (Zhang et al., 2023a), we perform instruction tuning on Twitter Financial News dataset (Magic, 2022) and FiQA (Maia et al., 2018). We conduct evaluation on Financial PhraseBank dataset (FPB, Malo et al., 2014) and Twitter financial news sentiment validation dataset (Twitter Val, Magic, 2022). Twitter Financial News Sentiment Dataset focuses on financial sector tweets, while FiQA consists of 961 annotated samples. The Financial PhraseBank dataset includes 4840 randomly selected samples from financial news articles in the LexisNexis database.

Baselines. The experiment includes three different types of baselines: 1) Direct generation without knowledge, without providing any knowledge for a given task and ask the MSA model to response directly; 2) Generation with retrieval financial knowledge, retrieveing related knowledge from external knowledge sources (*e.g., News Source*, *Research Publication Platforms* and *Social Media Platforms*), following the approach of prior works (Zhang et al., 2023a); 3) our WisdoM with non-financial *context*, employing gpt-3.5-turbo to generate historical knowledge in our stage 2.

Implementation Details. Since FPB and Twitter Val are textual sentiment analysis benchmarks, we adopt the same experimental setup as (Zhang et al., 2023a), using Llama-7B (Touvron et al., 2023) as our MSA model. In our WisdoM, we employ gpt-3.5-turbo to generate historical and financial knowledge respectively and select the hyperparameter on

FiQA validation dataset. The performance metric is F1 score.

Result. Table 11 shows that: 1) our WisdoM consistently brings improvement across all benchmarks among various types of knowledge, 2) financial knowledge brings more improvement than historical knowledge. The results prove that ***domain-related knowledge is more helpful.***

Method	FPB	Twitter Val
Llama-7B	72.1	78.7
-w/ RAG	75.0	81.2
-w/ WisdoM(Historical)	74.8	80.2
-w/ **WisdoM(Financial)**	**76.0**	**83.1**

Table 11: **Experimental results of Llama-7B with different methods on FPB and Twitter Val.** History and Financial represent the incorporation of history and financial knowledge respectively.

B.6 $S_{c\to p}$ and $S_{i\to p}$ of Other datasets

Fig. 12 illustrates the $S_{c\to p}$ and $S_{i\to p}$ on Twitter2017 and MSED. $S_{c\to p}$ is prominent, while $S_{i\to p}$ is less significant.

C Case Study

C.1 Example of Contexts

Examples of context from RAG-based methods and our WisdoM are presented in Table 12. It is evident that our WisdoM offers more precise context, providing detailed descriptions of image elements. In contrast, RAG-based methods exhibit a lack of specificity in image details and demonstrate weak relevance to the associated sentences.

C.2 Qualitative Examples of Aspect-Level MSA

We present two qualitative examples from Table 13 showcasing historical knowledge. In the first, adding background information about Aleppo aided in accurate decision-making. Conversely, in the second, the inclusion of satirical content led to misclassification. However, with Contextual Fusion, LLaVA-v1.5 ultimately make the correct decision.

C.3 Qualitative Examples of Sentence-Level MSA

Table 14 illustrates two instances of incorporating scientific world knowledge into the MSED dataset.

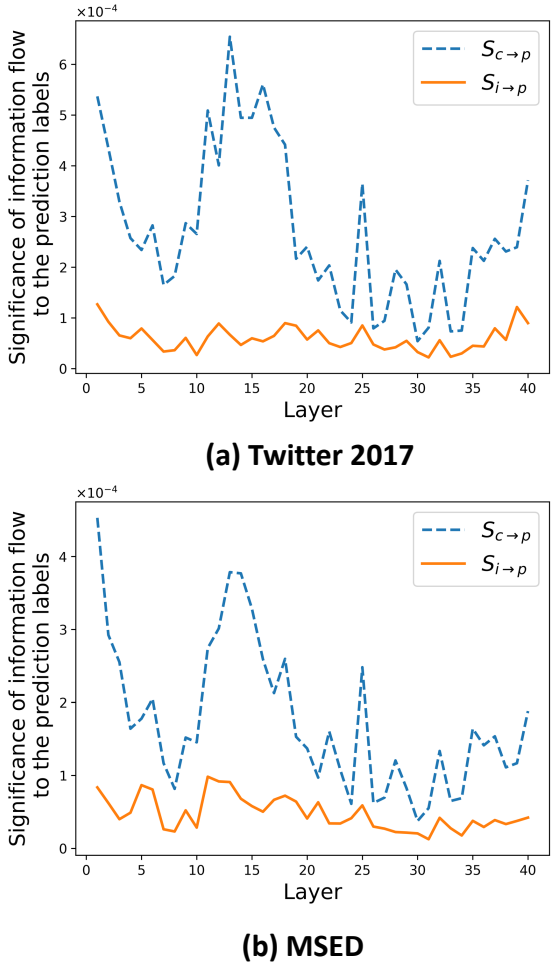


Figure 12: **$S_{c\to p}$ and $S_{i\to p}$ on Twitter2017 and MSED.** $S_{c\to p}$ represents the information flow from *context* to the prediction label and $S_{i\to p}$ represents information flow from image and sentence to prediction label.

In the first example, ALMT initially predicts negative, but incorporating information on masks' positive role in epidemic prevention leads to a positive sentiment prediction. In the second example, excessive focus on coffee benefits causes misclassification by LLaVA-v1.5. However, with Contextual Fusion, LLaVA-v1.5 ultimately predicts correctly.

C.4 Qualitative Examples of Financial Sentiment Analysis

In addition, we also provide examples in Financial Sentiment Analysis in Table 15. It can be observed that for samples with higher expertise, Llama-7B tends to predict incorrectly, but by introducing explanations for these samples, Llama-7B can predict correctly.

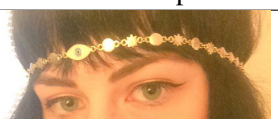
Sentence	RT @ shanilpanara : Bus Selfie on the way to Harry Potter Studios @ WFCTrust @ NCSEast # ShareYourSummer
Image	
RAG	Harry Potter and the Forbidden Journey is a motion-based dark ride located in The Wizarding World of Harry Potter-themed areas of Islands of Adventure in Orlando, Florida; Universal Studios Hollywood in Universal City, California, Universal Studios Japan in Osaka, Japan and is upcoming to Universal Studios Beijing in Beijing, China. The ride takes guests through scenes and environments in and around Hogwarts Castle from the Harry Potter series of books and films. Mark Woodbury, president of Universal Creative, described the ride as an in-depth look at the world of Harry Potter, which utilizes never-before-seen technology which transforms "the theme park experience as you know it". The ride first opened at Islands of Adventure with The Wizarding World of Harry Potter on June 18, 2010; subsequent versions opened at Universal Studios Japan on July 15, 2014, and at Universal Studios Hollywood on April 7, 2016.
PKG	A picture of a person with a camera and a big lens in front of them, taken by a tourist.
WisdoM	The image shows a large group of people, including men and women, sitting on a bus. They are all wearing black and white clothing, which suggests that the photo might have been taken during a specific time period or event when black and white clothing was more common. The group appears to be diverse, with people of different ages and ethnicities, which indicates that the bus might be a public transportation vehicle or a special event bus. In the context of historical events, the black and white clothing could imply that the photo was taken during the early to mid-20th century when colour photography was not widely available. The group of people might be attending a social event, a gathering, or a public transportation service during that time. The presence of a diverse group of people on the bus also reflects the social dynamics and cultural diversity of the period.
Sentence	RT @ wemssecnirp : Now I really am like Cleopatra # hypnotizing # findom # goddess
Image	
RAG	Cleopatra has been the subject of literature, films, plays, television programs, and art. Only those with Wikipedia articles are cited.
PKG	A portrait of Cleopatra, the Queen of Egypt, wearing a golden headdress and a necklace. The portrait was drawn in the style of Raphael and engraved on copper.
WisdoM	The image shows a woman wearing a gold headband with a large, ornate design featuring eyes. This headband is likely inspired by ancient Egyptian or African culture, as it resembles the traditional headpieces worn by women in these regions. The eyes on the headband may symbolize protection, wisdom, or the ability to see beyond the physical realm. In ancient Egypt, headbands were often worn by both men and women as a symbol of status, power, and religious beliefs. They were made from various materials, such as gold, silver, or precious stones, and often featured intricate designs and symbols. Similarly, in African cultures, headbands have been used as a form of adornment and expression of identity, with designs and patterns often representing specific tribes, clans, or regions. The woman in the image is wearing a modern interpretation of these traditional headbands, which demonstrates the enduring influence of ancient cultures on contemporary fashion and design.

Table 12: **Examples of contexts generated by RAG, PKG and our WisdoM.** Our contexts provide a detailed description that matches the elements present in images.

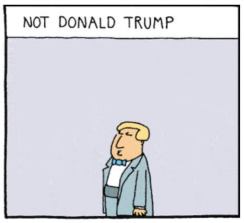
Sentence	RT @ AHedengren : # Aleppo before and after . # Syria .
Image	
Aspect	Aleppo
Context	The image shows a comparison of the same street in Aleppo, Syria, before and after the devastating war. The before picture depicts a bustling and lively street with people walking and cars driving, while the after picture displays a desolate and destroyed street with rubble and ruins. The stark contrast between the two images highlights the devastating impact of the war on the city and its residents. Aleppo, once a thriving and historic city, has been severely affected by the ongoing Syrian Civil War. The conflict has led to the destruction of infrastructure, displacement of residents, and loss of life. The image serves as a powerful reminder of the human cost of war and the need for peaceful resolution to conflicts.
Label	Negative
AoM	Neutral
+ *Context*	Negative
+ Contextual Fusion	Negative
Sentence	Donald Trump will win in a landslide. * The cartoonist behind ‘ Dilbert ’ explains why.
Image	
Aspect	Donald Trump
Context	The image features a cartoon of a man in a suit and tie, standing in front of a wall with a sign that says "Not Donald Trump." This cartoon is likely a political satire or a commentary on the 2016 United States presidential election, as Donald Trump was one of the main candidates running for office at that time. The cartoon could be a representation of the political climate, opinions, or reactions towards Donald Trump and his campaign. It might also be a way to express the artist’s or the creator’s perspective on the election, the political landscape, or the candidates involved.
Label	Positive
LLaVA-v1.5	Positive
+ *Context*	Negative
+ Contextual Fusion	Positive

Table 13: **Qualitative examples of aspect-level MSA**. Two cases with predictions by SOTA methods and our method on Twitter2015 and Twitter2017. We initially integrate only the *context* and then introduce Contextual Fusion. Clues to answering the input are highlighted in green within the *Context*. The *Context* that will disrupt MSA is highlighted in red.

Sentence	Group of teenagers posing showing their protective face masks during COVID-19 coronavirus epidemic spread.
Image	
Context	The image features a group of teenagers posing with their protective face masks, which are an essential part of personal protective equipment (PPE) during the COVID-19 pandemic. The use of face masks has been a crucial measure to reduce the transmission of the virus, as they help prevent respiratory droplets from spreading into the air. Scientifically, the effectiveness of face masks in preventing the spread of COVID-19 has been extensively studied. Research has shown that wearing a face mask can significantly reduce the transmission of respiratory viruses, including COVID-19. The masks work by filtering out respiratory droplets that may contain the virus, preventing them from reaching other people's noses and mouths. In conclusion, the image of the group of teenagers posing with their face masks highlights the importance of wearing PPE during the COVID-19 pandemic. The scientific evidence supporting the effectiveness of face masks in preventing the spread of respiratory viruses, including COVID-19, has been instrumental in promoting their use as a crucial public health measure.
Label	Positive
ALMT	Negative
+ *Context*	Positive
+ Contextual Fusion	Positive
Sentence	Handsome young man standing in the kitchen with a cup of coffee at home.
Image	
Context	The image features a handsome young man standing in a kitchen with a cup of coffee. Recent studies have shown that moderate coffee consumption can have positive effects on cognitive function, memory, and mood. Additionally, coffee has been found to have potential health benefits, such as reducing the risk of certain diseases, including type 2 diabetes and certain types of cancer. In terms of advancements in coffee production and consumption, there has been a growing interest in sustainable and eco-friendly coffee practices. This includes the use of organic farming methods, fair trade practices, and the development of new technologies to reduce waste and improve the efficiency of coffee production.
Label	Neutral
LLaVA-v1.5	Neutral
+ *Context*	Positive
+ Contextual Fusion	Neutral

Table 14: **Qualitative examples of MSED.** We initially integrate only the *context* and then introduce Contextual Fusion. Clues to answering the input are highlighted in green within the *Context*. Red represents incorrect prediction.

Sentence	Last year the company raised its turnover to approximately 7 million litas EUR 2 mln, from 6.1 million litas in 2004.
Context	The text you provided discusses the company's turnover, which refers to the total amount of money generated from sales of goods or services. It's important to note that turnover is different from profit, as it represents the company's total sales before deducting any expenses. In this case, the company's turnover increased from approximately 6.1 million litas in 2004 to 7 million litas (equivalent to EUR 2 million) last year. This indicates a positive trend in the company's sales performance. Increasing turnover can be a sign of business growth and improved market demand for the company's products or services.
Label	Positive
Llama-7B	Neutral
+ *Context*	Positive
+ Contextual Fusion	Positive
Sentence	$XOM (+5.8% pre) Exxon cuts full-year capex forecast by 30%, maintains long-term outlook - SA
Context	Exxon Mobil Corporation (XOM) has increased by 5.8% in pre-market trading. The company has announced a 30% reduction in its full-year capital expenditure forecast while maintaining its long-term outlook. This move reflects the company's response to changing market conditions and its commitment to long-term sustainability. Capital expenditure forecasts are important indicators of a company's investment plans and financial health. Investors often monitor these forecasts closely to assess a company's growth prospects and financial management. The market's positive reaction to this news suggests that investors may view Exxon's decision as a prudent and forward-thinking strategy.
Label	Positive
Llama-7B	Negative
+ *Context*	Positive
+ Contextual Fusion	Positive
Sentence	Stock Market Update: Stock market drifts in record territory
Context	The stock market drifting in record territory typically indicates a period of stability and potential growth. Investors may interpret this as a positive sign for the economy and the companies listed on the stock exchange. However, it's important to consider various factors such as interest rates, inflation, and geopolitical events that could impact market movements. Investors should also diversify their portfolios and stay informed about market trends to make well-informed decisions. It's advisable to consult with a financial advisor for personalized guidance based on individual financial goals and risk tolerance.
Label	Positive
Llama-7B	Negative
+ *Context*	Positive
+ Contextual Fusion	Positive

Table 15: **Qualitative examples of Financial PhraseBank (FPB) and Twitter financial news sentiment validation (Twitter Val).** Clues to answering the input are highlighted in green within the *Context*. Red represents incorrect prediction.