Exploring Fine-Grained Text-Image Alignment for Multimodal Aspect-based Sentiment Analysis

Anonymous ACL submission

Abstract

As a multimodal task, there have been increasing works focused on the text-image alignment in Multimodal Aspect-Based Sentiment Analysis (MABSA). Yet they trend to rely on various black box modules to self-learn the text-image alignment during training, either in attention-based (Ling et al., 2022), graph-based (Zhou et al., 2023), or vision-language model (Liu et al., 2025a), such implicit alignment cannot clearly state that the important textual phrases reflect on which part of the image, thereby can significantly support the cross-modal interaction. To this end, we are motivated to explore building explicit fine-grained alignment. We thus propose **I**nterpretation-based **E**xplicit **A**lignment (**IEA**) framework, it includes a Sentimental Image Interpreter and a Fine-grained Aligner, the former can effectively segment and interpret the region-level semantics while the latter can build the explicit fine-grained alignment between regions and textual phrases. Extensive experiments build a new SOTA performance in Twitter 2015/17 datasets, indicating the effectiveness of our alignment and revealing a new direction in cross-modal interaction.

1 Introduction

Multimodal Aspect-based Sentiment Analysis (MABSA) is a topic of increasing interest in the community, with the input pair of textual sentence and image, it is comprised of two subtasks: aspect term extraction (Zhang et al., 2021), aspect-oriented sentiment classification (Liu et al., 2025a). The Joint Multimodal Aspect-Sentiment Analysis (JMASA) (Yu and Jiang, 2019), which combines these two subtasks to jointly extract the aspect terms and their corresponding sentiments as shown in Figure 1, presents a significant challenge for traditional classification-based models.

Existing research (Yu et al., 2020b; Xu et al., 2019; Yu et al., 2020a; Yang et al., 2021; Yang and Li, 2023) on MABSA have primarily focused on designing various strategies to model the cross-modal alignment between textual context and the image. Their methods typically rely on certain fusion modules, include attention-based (Feng et al., 2024; Liu et al., 2023a), graph-based (Zhou et al., 2023), and joint pre-training (Ling et al., 2022). More recently, vision-language models (VLMs) have also been introduced into the MABSA task (Liu et al., 2025a; Bao et al., 2025) to utilize the large scale pre-trained knowledge of them, have shown superior performance over previous entries.

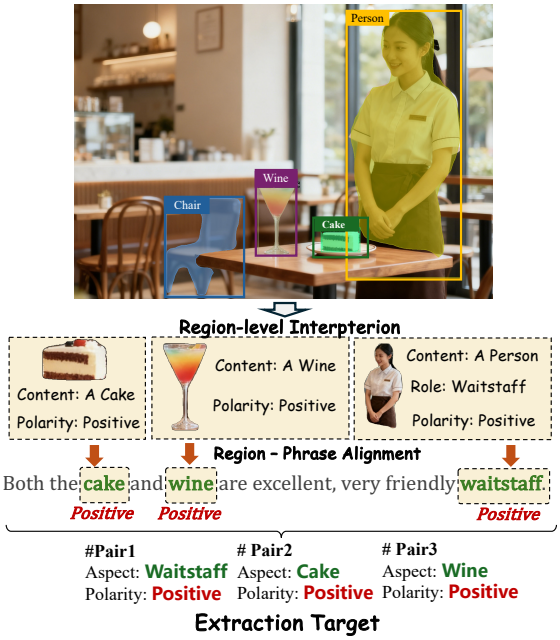


Figure 1: Our explicit fine-grained alignment.

Despite their effectiveness, previous works' strategies of alignment can be summarized as relying on a black box module to self-learn the text-image alignment implicitly during training. Such a manner results in two significant drawbacks: 1) Forcing them to rely on massive model parameters and training data to get the module fully fitted, involves high computational cost and poor adap-

tation scope. 2) Making their alignment nontransparent and difficult to be captured by the model, lacking an explicit association that the important textual phrases reflect on which part of the image as shown in Figure 1. To this end, we are motivated to explore building an explicit fine-grained alignment that can be directly injected into the model, providing clear alignment guidance that can be easily captured without sophisticated training, thereby reducing computational demands and remaining competitive in small model and data size situations.

However, it is difficult to tailor the explicit alignment, the challenges are threefold: 1) On the visual slide, how do we deconstruct the image into semantic regions and determine the sentiment meaning of them? Especially there may exist irrelevant regions that can disturb the analysis. 2) Once we obtain the regions and their semantics, how can we map them with the corresponding textual span of the role phrase to be aligned? 3) Finally, upon the alignments are settled, how can we model them with current powerful large language models?

In this study, we propose **I**nterpretation-based **E**xplicit **A**lignment (**IEA**) for MABSA. As shown in Figure 2, our approach answers the first two questions with our Sentimental Image Interpreter and Fine-grained Aligner. The interpreter works on the visual end to address the first challenge, it dynamically segments the image into non-overlapping semantic regions and interpret them by perceiving descriptive content and region-level sentiment evaluation. We further propose our Fine-grained Aligner for the second question. Our aligner first filters the regions that could be too small or vague to be meaningful with three conditions. Then, for each region, the aligner maps it with the fine-grained textual phrase to form the explicit text-image alignment. To train our aligner, we build the first **F**ine-**G**rained **Cross-modal A**lignment (**FGCA**) dataset that modified based on the existing semantic segmentation dataset PASCAL-Context (Mottaghi et al., 2014).

We finally model the obtained alignment with our Nearest-to-Match Instruction that directly injects visual regions and interpretations next to their textual counterparts, which is generalize and can fit any VLMs that support multi-image input. The detailed evaluation shows that our proposed model significantly advances the SOTA performance on several benchmarks and remains highly competitive with smaller models and data. To the best of our knowledge, we are the first to explore explicit cross-modal alignment in MABSA works.

2 Related Work

Unimodal ABSA: Research on ABSA generally starts from predicting a single sentiment element (Wang et al., 2021; Hu et al., 2019; Tang et al., 2016; Chen et al., 2022; Liu et al., 2021; Seoh et al., 2021; Zhang et al., 2022). Many studies further delve into exploring the joint extraction of sentiment elements (Xu et al., 2020; Li et al., 2022; Bao et al., 2023; Zhang and Qian, 2020).

Recently, generative paradigms for ABSA have gained notable traction: OTG (Bao et al., 2022) and dotGCN (Chen et al., 2022) relied on syntactic cues and dependency trees; ATOSS (Seo et al., 2024) introduced a plug-and-play module that decomposes a sentence into sub-sentences; UniGen (Choi et al., 2024) constructed zero-shot datasets by drawing on LLM knowledge; SCRAP (Kim et al., 2024) distilled chain-of-thought rationales and applied self-consistent voting to enhance stability and accuracy; MUL (Hu et al., 2023) controlled the token-level generation with template-agnostic methods.

Multimodal ABSA: The research on MABSA generally share a similar routine with the unimodal, from aspect term extraction (Zhang et al., 2018; Yu et al., 2020b) and polarity classification (Xu et al., 2019; Yu et al., 2020a; Yang et al., 2021) to Joint Multimodal Aspect-Sentiment Analysis (JMASA) task: Tom-BERT (Yu et al., 2020b) designed a target-oriented template; VLP-MABSA (Ling et al., 2022) designed a joint pre-training pattern; AoM (Zhou et al., 2023) relied on GCN to capture the token-level cross-modal graph; UnifiedTMSC (Liu et al., 2023a) adopted descriptive prompt paraphrasing; A^2II (Feng et al., 2024) adopt a cross attention module to learn the alignment; MVAT (Li et al., 2026) incorporated GCN with a Differential Transformer. Some works further utilize the powerful LLMs: DPCI (Liu et al., 2025a) relied on a LLAVA (Liu et al., 2023b) to integrate the image information; SIG (Bao et al., 2025) LoRA finetuned a VLM with both the original and the their generated images.

Nevertheless, the alignments in previous works can be summarized as implicit alignments, which rely on black box module to self-learn the relations, making them hard to be captured under computational conditions that are not particularly demanding. Different from previous studies, our method stands out as the first to construct explicit alignment at fine-grained level in MABSA, highlighting a new direction of constructing cross-modal connection.

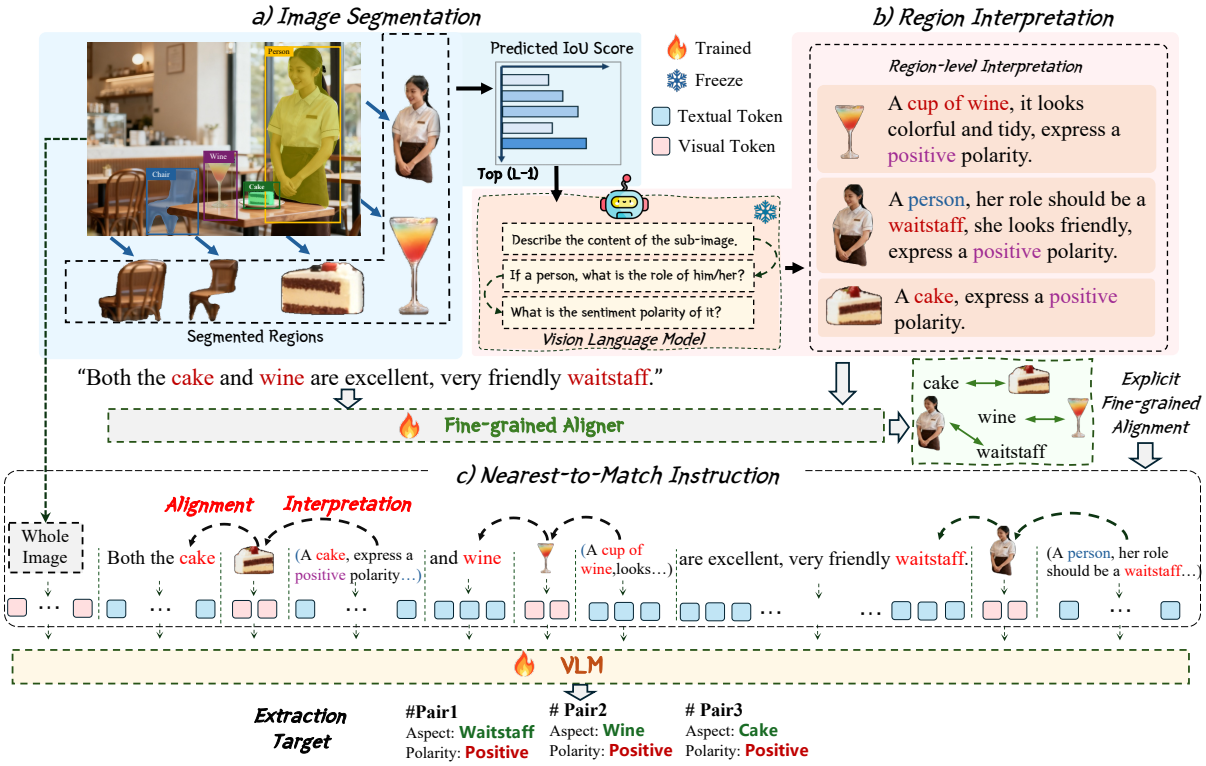


Figure 2: Illustration of our Interpretation-based Explicit Alignment.

3 Interpretation-based Explicit Alignment for MABSA

As shown in Figure 2, our IEA generally include three parts: 1) we first decompose an image into semantically coherent regions and extract region-level interpretation with our Sentimental Image Interpreter as shown in Figure 2 a) and b). 2) We then precisely ground each visual region to its corresponding phrase span in the text with our Fine-Grained Aligner, trained on our FGCA dataset that novelly built through a reverse synthetical procedure as shown in Figure 3. 3) We finally design a Nearest-to-Match Instruction which directly injects visual regions next to their textual counterparts, distilling the explicit alignments and adapting them to arbitrary VLMs as shown in Figure 2 c).

3.1 Sentimental Image Interpreter

The very first condition before building the fine-grained alignment is obtaining the regions to be aligned. We thus propose our Sentimental Image Interpreter, its overall pipeline proceeds as follows: 1) Segment Anything Model (SAM) (Kirillov et al., 2023) is first employed to automatically partition each original image into semantically coherent regions. 2) Subsequently, a VLM produces concise object and sentiment interpretation for each region. We will go though them one by one as follows.

Image Segmentation

To enable fine-grained distortion perception, the input image I is segmented into semantic regions in a parallel branch to preserve structural continuity while suppressing the influence of irrelevant regions. As shown in Figure 2 a), semantic region segmentation is based on a powerful large-scale foundation model, SAM (Kirillov et al., 2023), which can accurately segment any object in any image without additional training. The segmentation mask generated can be formulated as:

$$M = \text{Postprocess}(\text{SAM}(I)) \quad (1)$$

where $\text{Postprocess}(\cdot)$ ensures that the segmentation masks $M \in \mathbb{R}^{H\times W\times L}$ divide the image into L non-overlapping regions. It ranks the masks produced by $\text{SAM}(\cdot)$ in descending order by their Predicted IoU Scores, where a higher score represents the model's confidence for a more semantically meaningful region. We retain the top $(L-1)$ masks with the highest scores. Pixels belonging to multiple masks are assigned to the one with the highest score, while those not covered by the top $(L-1)$ masks are designated as the background region, resulting in a total of L semantic regions.

Region Interpretation

Following the above segmentation, each region is provided to a frozen VLM together with task-specific instruct prompts to elicit an interpretation of the semantic content and associated sentiment ratings, as depicted in Figure 2 b). For each region R in the image I, the per-region evaluations and ratings are subsequently aggregated into a holistic region-level textual description E.

By doing so, we can effectively obtain the visual regions and their interpretations. We will subsequently introduce our fine-grained aligner to align the regions with their textual counterparts.

3.2 Fine-Grained Aligner

How to find the corresponding textual span for each region become our next target. Ideally, the input of which should include the whole image I, segmented region R with its interpretation E, and the texts T that contains the phrase to be aligned, the output label would be the span $[P_{start}, P_{end}]$ of the aligned phrase P. The pattern of our fine-grained alignment task can then be formulated as:

$$[I, R, E, T] \rightarrow [P_{start}, P_{end}] \tag{2}$$

To handle situations where the split regions are too small or vague to be meaningful or there do not exist corresponding phrases, we further introduce a filtering mechanism with three judging conditions. Specifically, we first filter the regions with the size:

$$\min\{\text{width}(R), \text{height}(R)\} < t_{\text{size}}, \tag{3}$$

We also condition the clarity by computing the variance of the region with a Laplacian operator (Pech-Pacheco et al., 2000), where lower values indicating blurrier regions, specifically:

$$\text{Var}\left(\nabla^2 R\right) < t_{\text{clarity}}, \tag{4}$$

and confidence measured by Predicted IoU Score:

$$\text{Predicted\_IoU}(R) < t_{\text{conf}}. \tag{5}$$

If a region does not meet any of the above conditions, we subsequently filter it out and keep the rest to form the candidates pool pending alignment.

We then solve the lack of training data. To avoid the expensive labor cost of manual annotation, we novelly propose an automatic method that can build the dataset from existing semantic segmentation dataset reversely. We build the first fine-gained alignment dataset **F**ine-**G**rained **C**ross-modal

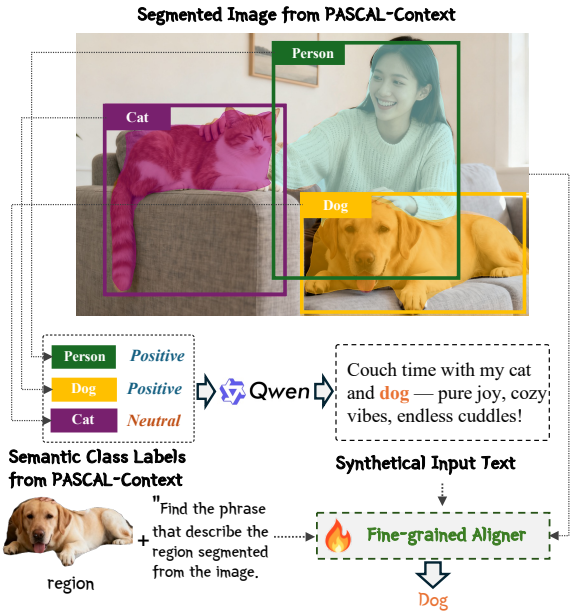


Figure 3: Our process of building FGCA from semantic segmentation dataset reversely and the training pattern of our Fine-Grained Aligner.

Alignment (**FGCA**) modified based on semantic segmentation dataset PASCAL-Context (Mottaghi et al., 2014) as shown in Figure 3. In PASCAL-Context we already have the label of image I, segmented region R, predicted interpretation E from our interpreter, and the region's corresponding phrase P (semantic class of the region), the only target that we need to build is the input text T. We thus adopt Qwen3-8B (Yang et al., 2025) with a clear prompt to instruct it to construct the synthetical input text T' based on all of the phrases (semantic classes) P appear in image I, and record the span $[P'_{start}, P'_{end}]$ as training output.

3.3 Nearest-to-Match Instruction

We then design the instructions to model our fine-grained alignment. We follow the principle of nearest-to-match, directly inject the regions next to their textual counterparts. As shown in Figure 2 c), given the whole image I,the texts T, segmented regions $R_1, R_2, ..., R_n$ with their aligned phrases $P_1, P_2, ..., P_n$ and interpretations $E_1, E_2, ..., E_n$, we first identify the aligned triplet unit U_n as:

$$U_n = [P_n, R_n, E_n] \tag{6}$$

we then organize the input x with the format:

$$X = [I, c_1, U_1, c_2, U_2, ..., U_n, c_m] \tag{7}$$

where the $c_1, c_2, ..., c_m$ refers to the rest content in the text T that does not belong to a phrase in

the aligned triplets, been organized in their original relative order in the raw context. The output target would be the pair of aspect term and corresponding polarity directly. Our instruction does need any complex modeling modules, can adapt to any powerful VLMs that accept multi-image input.

We finally finetune a VLM for the JMABSA extraction with our nearest-to-match instruction. Given the fused sequence $X = x_1, ..., x_{|x|}$ as input. At the i-th step of generation, the decoder predicts the i-th token y_i, and decoder state h_i^d as:

$$y_i, h_i^d = ([h_1^d, ..., h_{i-1}^d], y_{i-1}) \tag{8}$$

The conditional probability of the whole output sequence $p(y|X)$ is progressively combined by the probability of each step $p(y_i|y_{<i}, X)$:

$$p(y|x) = \prod_{i=1}^{|y|} p(y_i|y_{<i}, X) \tag{9}$$

where $y_{<i} = y_1...y_{i-1}$, and $p(y_i|y_{<i}, X)$ are the probabilities over target vocabulary V.

The objective function maximizes the output target sequence Y probability given the sentence X. Therefore, we optimize the negative log-likelihood loss function:

$$\mathcal{L} = \frac{-1}{|\tau|} \sum_{(X,Y)\in\tau} \log p(Y|X;\theta) \tag{10}$$

where θ is the model parameters, and (X, Y) is a (*sentence, target*) pair in training set τ.

4 Experiment

4.1 Dataset and Experiment Setting

In this study, we use the Twitter2015/2017 (Yu and Jiang, 2019) dataset and their splitting for multimodal ABSA experiments. For our VLM for finetuning, we employ the pre-trained Qwen3-VL-8B (Yang et al., 2025) and LoRA finetune the LLM adapter parameters with 30 epochs. In terms of the VLM for Region Interpretation, we employ the smaller Qwen3-VL-2B (Yang et al., 2025) for saving computing cost. Segment Anything Model (SAM) (Kirillov et al., 2023) is adopted and the number L of retained regions is set to 5. The threshold t_{size}, t_{clarity} and t_{conf} for filtering are set to 25, 100 and 0.7 respectively. More detailed training settings can be found in Appendix A.

For our FGCA, we finally filter 2,000 samples for the train set, 500 samples for the dev and test set. As shown in Figure 3, the training input include the whole image I, a specific region R pending alignment and the synthetical text T', the output target is the span of the semantic class $[P'_{start}, P'_{end}]$ appeared in T'. We adopt Qwen3VL-2B [1] and LoRA finetune it on our FGCA as our aligner, the detailed distribution of our FGCA can be found in Table 1.

Split	#Sample	#Class	#Region/Image
Train	2,000	74	4.65
Val	500	48	5.52
Test	500	56	4.21

Table 1: Distribution of our FGCA for training aligner.

4.2 Main Results

In Table 2, we present a comprehensive comparison of our model with various state-of-the-art baselines. These baselines include both pure-textual and multimodal models, as well as VLMs and LLMs, specifically: **Pure-Textual**: 1) MvP+ATOSS (Seo et al., 2024). 2) SCRAP (Kim et al., 2024). 3) OTG Bao et al. (2022). **Multimodal**: 1) UMT+TomBERT (Yu et al., 2020b) 2) OSCGA+TomBERT (Yu and Jiang, 2019) 3) RpBERT-collapse (Sun et al., 2021) 4) SIG (Bao et al., 2025) 5) JML (Ju et al., 2021) 6) VLP-MABSA (Ling et al., 2022) 7) CMMT (Yang et al., 2022) 8) CORSA (Liu et al., 2025b); **LoRA finetuned LLMs**: 1) LLaMA3-8B (AI@Meta, 2024); 2) Qwen3-8B (Yang et al., 2025); **LoRA finetuned VLMs**: 1) InternLM-XComposer2-VL (Dong et al., 2024); 2) Qwen3-VL-8B (Yang et al., 2025);

From Table 2 we observe that multimodal models easily surpass previous pure-textual approaches. Furthermore, the VLMs outperform a large number of approaches without complex modeling, showing its efficacy for the complex extraction task. The results also highlight the effectiveness of the unified generation architecture, which can fully utilize the rich label semantics by encoding the natural language label into the target output for extraction.

Moreover, our proposed model exhibits significant improvements over all prior studies ($p < 0.05$), demonstrating the efficacy of our explicit alignment, where the key cross-modal relations are fed more easier to be captured by the model. We further investigate its performance under limited parameters and data in Section 5.3 and Appendix C.

4.3 Ablation Study

After presenting the overall performance, we then move to analyze how each component contributes to the observed gains in Table 2. To this end, we

[1] https://github.com/QwenLM/Qwen3-VL

Manner	Method	Twitter2015			Twitter2017		
		P	R	F1	P	R	F1
Pure-Textual	MvP+ATOSS	0.667	0.641	0.653	0.663	0.644	0.654
	SCRAP	0.657	0.638	0.648	0.647	0.671	0.659
	OTG	0.641	0.623	0.631	0.639	0.628	0.633
LLM	LLaMA3-8B	0.633	0.621	0.626	0.628	0.637	0.632
	Qwen3-8B	0.633	0.652	0.642	0.635	0.641	0.637
VLM	InternLM-XComposer2-VL	0.646	0.662	0.653	0.649	0.662	0.655
	Qwen3-VL-8B	0.684	0.703	0.693	0.703	0.689	0.695
Multimodal	UMT+TomBERT	0.584	0.613	0.598	0.623	0.624	0.624
	OSCGA+TomBERT	0.617	0.634	0.625	0.634	0.640	0.637
	RpBERT-collapse	0.493	0.469	0.480	0.570	0.554	0.562
	JML	0.650	0.632	0.641	0.665	0.655	0.660
	VLP-MABSA	0.651	0.683	0.666	0.669	0.692	0.680
	CMMT	0.646	0.687	0.665	0.676	0.694	0.685
	SIG	0.662	0.695	0.678	0.683	0.698	0.690
	CORSA	0.690	0.708	0.699	0.701	0.710	0.706
	IEA(Ours)	**0.714**	**0.722**	**0.717**	**0.716**	**0.727**	**0.721**

Table 2: Results of JMABSA task on Twitter2015 and Twitter2017 datasets.

Method	Twitter15	Twitter17
Ours	0.717	0.721
Image Interpreter		
- Image Segmentation	0.702	0.705
- Region Interpretation	0.696	0.698
- Nearest-to-Match Instruction	0.691	0.699
- Fine-Grained Aligner & Nearest-to-Match Instruction	0.684	0.690

Table 3: The results of ablation study.

	Sentiment		Aspect	
	Twitter15	Twitter17	Twitter15	Twitter17
w/o IEA	0.756	0.773	0.853	0.917
with IEA	0.783 (+2.70%)	0.784 (+1.10%)	0.885 (+3.20%)	0.940 (+2.30%)
Cover Rate	0.5942	0.5375	0.4438	0.39419

Table 4: Results of our IEA on different elements, we adopt Qwen3-VL-8B as our base model.

progressively ablate the interpreter, the aligner, and the instruction module, removing them one by one to isolate their impact and quantify the effect of each module on the final improvements.

As depicted in Table 3, the full model performs best. Significantly decrease are observed when our components are excluded in the model, among them our Nearest-to-Match Instruction contributes the most, underscoring that precise and direct visual-text alignment is essential for mapping opinion elements to localized visual evidence. Region Interpretation contributes the second, indicating it is supplementary for model's visual understanding.

4.4 Results on Sentiment Elements

We further examine the performance of our method on different sentiment elements to check if our explicit alignment can efficiently help building connection between sentiment elements and their visual reflections. In particular, we ablate our alignment from our model, solely rely on Qwen3-VL-8B to isolate and measure our alignment's impact on each specific element type. In addition, we further measure our aligner's performance by computing its alignment coverage rate over sentiment elements, therefore we can analyze how does our aligner perform, and how does its performance affect the extraction results across different elements?

As shown in Table 4, IEA yields substantial performance improvement on both the aspect and sentiment elements, indicating that our explicit alignment effectively associates each fine-grained element with its corresponding visual region. Notably, the cover rates exhibit a positive correlation with the magnitude of improvement, and even relatively modest coverage still produces sizable gains, suggesting that partial alignment alone can provide strong and robust guidance for element extraction.

We also have the comparison with pervious works in Multimodal Aspect Term Extraction (MATE) and Multimodal Aspect-oriented Sentiment Classification (MASC) in Appendix B for a more comprehensive evaluation.

Manner	Method	Twitter2015			Twitter2017		
		P	R	F1	P	R	F1
Attention-Based	TMFN	0.690	0.684	0.696	0.709	0.706	0.712
	MVAT	0.688	0.700	0.694	0.700	0.709	0.705
Graph-Based	AoM	0.679	0.693	0.686	0.684	0.710	0.697
VLM-Based	SIG	0.662	0.695	0.678	0.683	0.698	0.690
	Qwen3-VL-8B	0.689	0.697	0.693	0.688	0.702	0.695
Explicit	**IEA(Ours)**	**0.714**	**0.722**	**0.717**	**0.716**	**0.727**	**0.721**

Table 5: Results of different alignment methods.

VLM	NMI	Twitter2015	Twitter2017
InternVL2.5-2B	with	0.663	0.681
	w/o	0.624	0.632
Qwen3-VL-4B	with	0.709	0.713
	w/o	0.679	0.684
InternLM-2-VL-8B	with	0.677	0.673
	w/o	0.653	0.655
LLaVA-NeXT-8B	with	0.643	0.656
	w/o	0.628	0.631

Table 6: Results on different VLMs, NMI refers to our Nearest-to-Match Instruction.

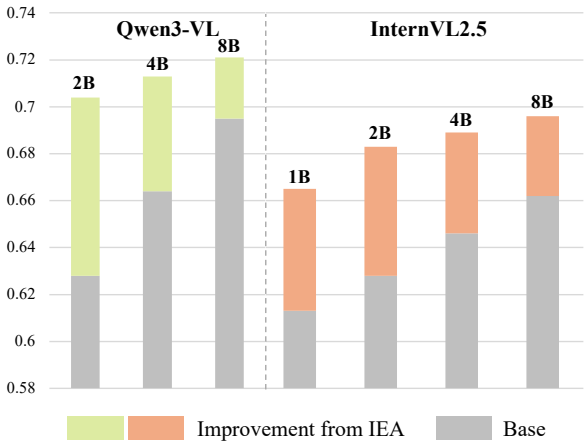


Figure 4: Comparison of parameter efficiency.

5 Analysis and Discussion

5.1 Comparison of Different Alignments

We subsequently make a comparison of different alignment manners. The comparison include: **Attention-Based**: 1) TMFN (Wang et al., 2024). 2) MVAT (Li et al., 2026). **Graph-Based**: AoM Zhou et al. (2023). **VLM-Based** : 1) SIG (Bao et al., 2025); 2) Qwen3-VL-8B (Yang et al., 2025);

Referring to Table 5, it is evident that the methods based on attention mechanism surpass the graph-based and VLM-based alignment by a considerable margin, showing the superiority of attention. Furthermore, our explicit alignment surpasses all above, we attribute this superiority to our alignment's ability of offering an explicit association, where the region-phrase alignment can be easily captured by the model, yet the previous implicit alignment appears comparatively vague and needs to be iteratively learned during training. It also hints us a new avenue for exploring multimodal understanding: substituting previous self-learned blackbox alignment with explicit alignment.

5.2 Generalization of Nearest-to-Match Instruction

We then analyze the effect of our Nearest-to-Match Instruction on different VLMs to check the generalization of it. In particular, we check the effectiveness of our instruction on popular VLMs that support multi-image input, include Qwen3-VL-4B (Yang et al., 2025), InternLM-2-VL-8B (Dong et al., 2024), InternVL2.5-2B (Chen et al., 2024) and LLaVA-NeXT-8B (Liu et al., 2023b). We use "without" in Table 6 refers the remove of our Nearest-to-Match Instruction, solely inputs the whole image the text and relies on the VLM to construct the alignment internally.

As shown in Table 6, NMI generally improves performance across diverse VLMs, among them InternLM and LLaVA-NeXT benefit notably from NMI, indicating good model-agnostic generalization. This indicates NMI supplies explicit, fine-grained grounding that complements each model's internal alignment, reducing spurious matches and improving robustness with different base models.

5.3 Analysis of Parameter Efficiency

Different from previous implicit alignments that needs massive parameters in the downstream model to learn the relation during finetuning, our explicit alignment can be built externally, does not rely on the model itself. This bring us a huge advantage in parameter efficiency, where a smaller model can

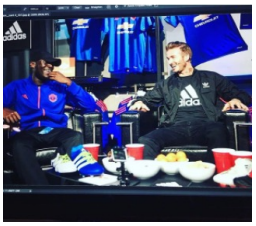
Text	Spent the morning chilling on a sofa with David Beckham.	Chicago cops play impromptu football game with kids in Lawndale	At the mets game with BBQ , beer and baseball. I ' m all set .	Throwback Thursday of me taking a ride with my best mate docile pet sheep Kate.
Image				
Qwen3-VL-8B	(***sofa***, Positive) ✗	(***football game***, Positive) ✗	(***baseball***, ***Neutral***) ✗	(***docile pet***, Positive) ✗
Explicit Alignment	 David Beckham	 Kids  Chicago Cops	 Baseball   Beer BBQ	 Me  Sheep Kate
IEA(Ours)	(***David Beckham***, Positive) ✓	(***Chicago***, Positive) ✓	(***mets***, Positive) ✓	(***Kate***, Positive) ✓

Table 7: Cases studies.

easily surpass the larger models with our explicit alignment. We thus investigate this by comparing the performance of small VLMs with our alignment verse large VLMs on Twitter 2017 to check if our method can improve the parameter efficiency.

As shown in Figure 4, without our alignment, the advantages of larger VLMs are highly evident, It also reveals that modeling complex relationships like the text-image alignment truly requires powerful models with more parameters. On the other side, with our explicit alignment, smaller VLMs (e.g., 1B or 2B) can achieve higher performance even when their size is only one-fifth that of larger VLMs (e.g., 8B or 13B), showing our alignment can narrow the capacity gap without inflating parameter counts. This highlights a practical path toward cost-effective deployment: pairing compact VLMs with our explicit alignment to unlock performance traditionally reserved for larger models.

We also have the analysis of data efficiency in Appendix C for showing the superiority of our IEA can be built upon limited data resources.

6 Case Study

We launch case studies to make a more intuitive comparison between our method and one of the strongest baselines Qwen3-VL-8B in Table 7.

We show that our method can effectively identify the person appear in the image and align them with the text in the first two examples. In the first one the baseline mistakenly focuses on the wrong target such as "sofa" while our IEA successfully between the region and phrase of "David Beckham". In the second example, our model successfully distinguishes the two groups of people and help the model identify the "Chicago" as aspect term.

In the last two examples, we illustrate that our method also performs better in recognizing multiple common items such as the "Baseball","Beer" and "BBQ" in the third example, effectively helping the model building text-image connection. The fourth example is similar, where our explicit alignment on the region of the sheep significantly helps the extraction while the baseline falsely recognizes the adjective of the correct target as output.

7 Conclusion

In this study, we highlight that prior MABSA approaches largely rely on implicit black-box alignment, which limits their transparency and hindered their efficiency in both the model parameter and training data sizes. We therefore propose an explicit, fine-grained alignment with **I**nterpretation-based **E**xplicit **A**lignment (**IEA**) for MABSA. Combining with Sentimental Image Interpreter and Fine-grained Aligner, our IEA not only surpasses the prior works in overall performance, but also shows great advantage in parameter and data efficiency, revealing us a new direction for the future exploration of cross-modal interactions.

Limitations

The limitations of our work can be articulated from two perspectives. First, beyond images, there are additional modalities such as video whose influence on downstream tasks is still uncertain. A more systematic examination of how text and video align, including factors like timing, prosody, and noise robustness, could clarify their contributions and provide actionable guidance for model design.

Second, our investigation has primarily centered on applying explicit alignment within MABSA. While the results are encouraging, the behavior of our approach in other problem settings such as event extraction remains untested. Evaluating the method on a broader set of tasks with different structures and data distributions would help determine its generalization and limits.

References

AI@Meta. 2024. Llama 3 model card.

Xiaoyi Bao, Jinghang Gu, Zhongqing Wang, and Chu-Ren Huang. 2025. Sentimental image generation for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4070–4081, Vienna, Austria. Association for Computational Linguistics.

Xiaoyi Bao, Zhongqing Wang, Xiaotong Jiang, Rong Xiao, and Shoushan Li. 2022. Aspect-based sentiment analysis with opinion tree generation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4044–4050. ijcai.org.

Xiaoyi Bao, Zhongqing Wang, and Guodong Zhou. 2023. Exploring graph pre-training for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3623–3634, Singapore. Association for Computational Linguistics.

Chenhua Chen, Zhiyang Teng, Zhongqing Wang, and Yue Zhang. 2022. Discrete opinion tree induction for aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2064, Dublin, Ireland. Association for Computational Linguistics.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Juhwan Choi, Yeonghwa Kim, Seunguk Yu, JungMin Yun, and YoungBin Kim. 2024. UniGen: Universal domain generalization for sentiment classification via zero-shot dataset generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1–14, Miami, Florida, USA. Association for Computational Linguistics.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, and 4 others. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *Preprint*, arXiv:2401.16420.

Junjia Feng, Mingqian Lin, Lin Shang, and Xiaoying Gao. 2024. Autonomous aspect-image instruction a2II: Q-former guided multimodal sentiment classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1996–2005, Torino, Italia. ELRA and ICCL.

Mengting Hu, Yinhao Bai, Yike Wu, Zhen Zhang, Liqi Zhang, Hang Gao, Shiwan Zhao, and Minlie Huang. 2023. Uncertainty-aware unlikelihood learning improves generative aspect sentiment quad prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13481–13494, Toronto, Canada. Association for Computational Linguistics.

Mengting Hu, Shiwan Zhao, Li Zhang, Keke Cai, Zhong Su, Renhong Cheng, and Xiaowei Shen. 2019. CAN: Constrained attention networks for multi-aspect sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4601–4610, Hong Kong, China. Association for Computational Linguistics.

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

Jieyong Kim, Ryang Heo, Yongsik Seo, SeongKu Kang, Jinyoung Yeo, and Dongha Lee. 2024. Self-consistent reasoning-based aspect-sentiment quad prediction with extract-then-assign strategy. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7295–7303, Bangkok, Thailand. Association for Computational Linguistics.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *arXiv:2304.02643*.

Hongxin Li, Bin Gao, Linlin Li, Yutong Li, Shutian Liu, and Zhengjun Liu. 2026. Multi-level visual-textual alignment transformer for multimodal aspect-based sentiment analysis. *Expert Systems with Applications*, 299:130148.

Junjie Li, Jianfei Yu, and Rui Xia. 2022. Generative cross-domain data augmentation for aspect and opinion co-extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4219–4229, Seattle, United States. Association for Computational Linguistics.

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

Dan Liu, Lin Li, Xiaohui Tao, Jian Cui, and Qing Xie. 2023a. Descriptive prompt paraphrasing for target-oriented multimodal sentiment classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4174–4186, Singapore. Association for Computational Linguistics.

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

Jian Liu, Zhiyang Teng, Leyang Cui, Hanmeng Liu, and Yue Zhang. 2021. Solving aspect category sentiment analysis as a text generation task. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4406–4416, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rui Liu, Jiahao Cao, Jiaqian Ren, Xu Bai, and Yanan Cao. 2025a. Dual-path counterfactual integration for multimodal aspect-based sentiment classification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22759–22769, Suzhou, China. Association for Computational Linguistics.

Xinjing Liu, Ruifan Li, Shuqin Ye, Guangwei Zhang, and Xiaojie Wang. 2025b. Multimodal aspect-based sentiment analysis under conditional relation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 313–323, Abu Dhabi, UAE. Association for Computational Linguistics.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. 2014. The role of context for object detection and semantic segmentation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

J.L. Pech-Pacheco, G. Cristobal, J. Chamorro-Martinez, and J. Fernandez-Valdivia. 2000. Diatom autofocusing in brightfield microscopy: a comparative study. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 3, pages 314–317 vol.3.

Yongsik Seo, Sungwon Song, Ryang Heo, Jieyong Kim, and Dongha Lee. 2024. Make compound sentences simple to analyze: Learning to split sentences for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11171–11184, Miami, Florida, USA. Association for Computational Linguistics.

Ronald Seoh, Ian Birle, Mrinal Tak, Haw-Shiuan Chang, Brian Pinette, and Alfred Hough. 2021. Open aspect target sentiment classification with natural language prompts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6311–6322, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: A text-image relation propagation-based bert model for multimodal ner. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13860–13868.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective lstms for target-dependent sentiment classification. In *COLING 2016*, pages 3298–3307.

Di Wang, Yuzheng He, Xiao Liang, Yumin Tian, Shaofeng Li, and Lin Zhao. 2024. TMFN: A target-oriented multi-grained fusion network for end-to-end aspect-based multimodal sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16187–16197, Torino, Italia. ELRA and ICCL.

Qianlong Wang, Zhiyuan Wen, Qin Zhao, Min Yang, and Ruifeng Xu. 2021. Progressive self-training with discriminator for aspect term extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 257–268, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349, Online. Association for Computational Linguistics.

Nan Xu, Wenji Mao, and Guandan Chen. 2019. Multi-interactive memory network for aspect based multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):371–378.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao,

Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Bin Yang and Jinlong Li. 2023. Visual elements mining as prompts for instruction learning for target-oriented multimodal sentiment classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6062–6075, Singapore. Association for Computational Linguistics.

Hao Yang, Yanyan Zhao, and Bing Qin. 2022. 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*, pages 3324–3335, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Li Yang, Jianfei Yu, Chengzhi Zhang, and Jin-Cheon Na. 2021. Fine-grained sentiment analysis of political tweets with entity-aware multimodal network. In *Diversity, Divergence, Dialogue: 16th International Conference, IConference 2021, Beijing, China, March 17–31, 2021, Proceedings, Part I*, page 411–420, Berlin, Heidelberg. Springer-Verlag.

Jianfei Yu and Jing Jiang. 2019. Adapting bert for target-oriented multimodal sentiment classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5408–5414. International Joint Conferences on Artificial Intelligence Organization.

Jianfei Yu, Jing Jiang, and Rui Xia. 2020a. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020b. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3352, Online. Association for Computational Linguistics.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal graph fusion for named entity recognition with targeted visual guidance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14347–14355.

Mi Zhang and Tieyun Qian. 2020. Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3540–3549, Online. Association for Computational Linguistics.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Zheng Zhang, Zili Zhou, and Yanna Wang. 2022. SSEGCN: Syntactic and semantic enhanced graph convolutional network for aspect-based sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4916–4925, Seattle, United States. Association for Computational Linguistics.

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 the Association for Computational Linguistics: ACL 2023*, pages 8184–8196, Toronto, Canada. Association for Computational Linguistics.

A Parameter Setting

For our VLM for finetuning, we employ the pre-trained Qwen3-VL-8B (Yang et al., 2025) and LoRA finetune the LLM adapter parameters. Training was conducted for 30 epochs with a global effective batch size of 4 (per-device train batch size = 1, gradient accumulation steps = 4). Inputs were truncated to a maximum length of 4096 tokens, and images were resized to 490 pixels on the longer side. Mixed-precision was enabled with bfloat16. The vision encoder and the sampler were frozen throughout training. We applied fine-tuning via LoRA (the LoRA alpha is 128 and the LoRA rank is 64.), while other components were frozen. Optimization used AdamW with learning rate 5e-5, weight decay 0.1, 2 = 0.95, and a cosine learning-rate schedule with 1% warm-up.

For our Fine-Grained Aligner, we finetune the smaller Qwen3-VL-2B(Yang et al., 2025) in a similar setting, with a shorter 10 epochs training. The filtering mechanism in our aligner finally filter out 37.12% of the segmented regions.

B Results of MATE and MASC task

We further examine our method on two extra MABSA subtask: Multimodal Aspect Term Extraction (MATE) and Multimodal Aspect-oriented Sentiment Classification (MASC). The former extracts all aspect terms in the sentence at the prompt of the image, and the latter predicts the sentiment polarities for the given aspect.

As shown in Table 8 and 9, IEA achieves leading or highly competitive results on both MATE

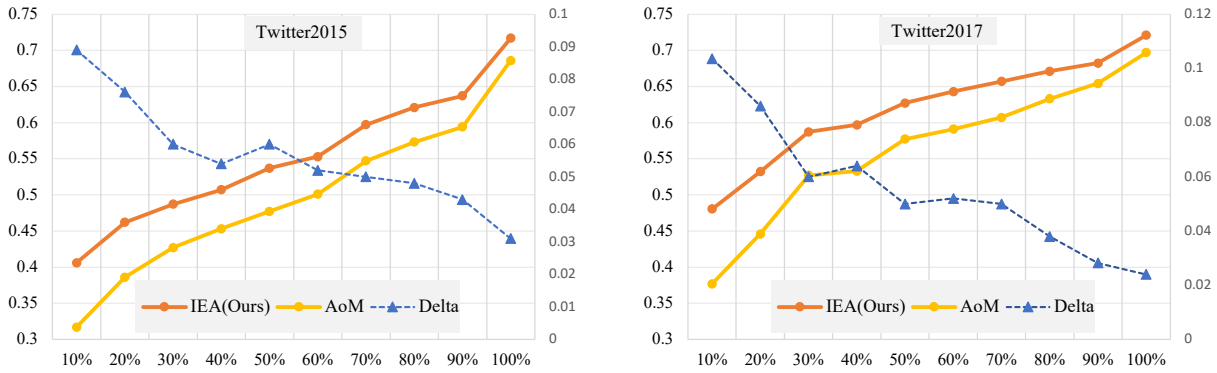


Figure 5: Improvement of data efficiency.

Method	Twitter2015			Twitter2017		
	P	R	F1	P	R	F1
UMT	0.778	0.817	0.797	0.867	0.868	0.867
OSCGA	0.817	0.821	0.819	0.902	0.907	0.904
JML	0.836	0.812	0.824	0.920	0.907	0.914
VLP-MABSA	0.836	0.879	0.857	0.908	0.926	0.917
CMMT	0.839	0.881	0.859	0.922	0.939	0.931
AoM	0.846	0.879	0.862	0.918	0.928	0.923
IEA(Ours)	**0.883**	**0.889**	**0.885**	**0.934**	**0.947**	**0.940**

Table 8: Results of MATE task on Twitter2015/17.

Method	Twitter2015		Twitter2017	
	ACC	F1	ACC	F1
TomBERT	0.772	0.718	0.705	0.680
JML	0.787	-	0.727	-
VLP-MABSA	0.786	0.738	0.738	0.718
CMMT	0.779	-	0.738	-
AoM	0.802	0.759	0.764	0.750
DPCI	0.804	0.763	0.752	0.747
IEA(Ours)	**0.834**	**0.783**	**0.778**	**0.784**

Table 9: Results of MASC task on Twitter2015/17.

and MASC, demonstrating strong cross-task transferability and robustness. It is worth noting that our method can well adapt to them, only a few modifications on our instruction are needed.

C Analysis of Data Efficiency

When compared to implicit alignment that relies on various module to self-learn the alignment relation during training, one of the advantages of our fine-grained alignment is the construction of the alignment is done externally with our aligner, does not rely on the massive training set, particularly when dealing with a limited amount of training data. We thus investigate how the fine-grained alignment improves the data efficiency by comparing with one of the strongest baselines of AoM (Zhou et al., 2023).

From Figure 5, we find that the more training data, the higher performance our proposed model can reach. Moreover, the improvement brought by our fine-grained alignment increases under limited data size, showing the superiority of fine-grained alignment in low resource situation, where the alignment information can be built easily from external when compared with relying on implicit alignment learned during training.