Sentimental Image Generation for Aspect-based Sentiment Analysis

Anonymous ACL submission

Abstract

Recent research work on textual Aspect-Based Sentiment Analysis (ABSA) have achieved promising performance. However, a persistent challenge lies in the limited semantics derived from the raw data. To address this issue, researchers have explored enhancing textual ABSA with additional augmentations, they either craft audio (Guo et al., 2024), text (Seo et al., 2024) and linguistic features (Bao et al., 2022) based on the input, or rely on user-posted images (Yu and Jiang, 2019). Yet these approaches have their limitations: the former three formations are heavily overlap with the original data, which undermines their ability to be supplementary while the user-posted images are extremely dependent on human annotation, which not only limits its application scope to just a handful of text-image datasets, but also propagates the errors derived from human mistakes to the entire downstream loop. In this study, we explore the way of generating the sentimental image that no one has ever ventured before. We propose a novel Sentimental Image Generation method that can precisely provide ancillary visual semantics to reinforce the textual extraction as shown in Figure 1. Extensive experiments build a new SOTA performance in ACOS, ASQP and en-Phone datasets, underscoring the effectiveness of our method and highlighting a promising direction for expanding our features.

1 Introduction

Aspect-based sentiment analysis (ABSA) is a topic of increasing interest in the research community, it is comprised of four subtasks: aspect term extraction, opinion term extraction, aspect category classification, and aspect-level sentiment classification. The Aspect-Category-Opinion-Sentiment (ACOS) Quadruple Extraction task, which combines these four subtasks as shown in Figure 1, presents a significant challenge for traditional classification-based models.

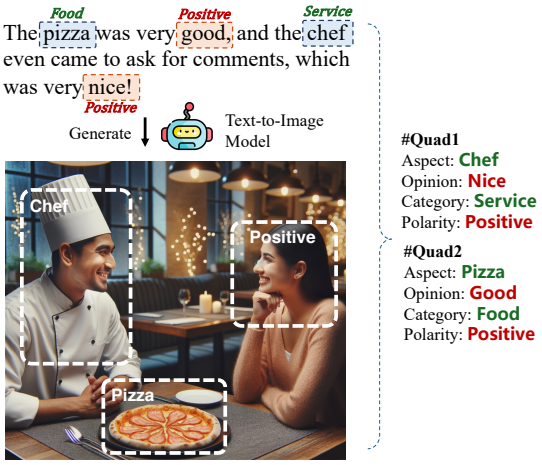


Figure 1: Example of sentimental image generation.

In response, recent research has adopted a unified generative approach to avoid the complex modeling. These approaches either design complex training or inference patterns (Kim et al., 2024; Gou et al., 2023; Xianlong et al., 2023; Bao et al., 2023b), or specify the desired target sequence (Yan et al., 2021; Zhang et al., 2021b,a; Bao et al., 2022; Hu et al., 2022) to simplify the overall task and improve performance. Despite their effectiveness, most previous studies are restricted to raw input data (Zhang et al., 2021b; Hu et al., 2022), fail to consider other data sources that could be supplementary to textual ABSA systems.

To alleviate this problem, recent research tend to introduce external knowledge from data augmentations to enhance the textual ABSA performance. They either craft audio (Guo et al., 2024), text (Seo et al., 2024) and linguistic features (Bao et al., 2022) on the basis of the textual samples, or rely on user-posted images (Yu and Jiang, 2019). Nevertheless, these approaches have notable limitations: most of the knowledge introduced in the first three formations either heavily overlap with the raw data (such as audio and text) or not stranger for the language models that are pre-trained on a mas-

sive corpus (such as linguistic features), thereby hindering their ability to enrich the knowledge of ABSA models. And for the images posted by users, they completely rely on human annotation, which not only restricts their application scope to a few labeled text-image datasets but also risks propagating the vague sentiment expression and weak text-image association caused by human mistakes into downstream extraction.

We hence shift our focus to generating sentimental images from scratch as an alternative to the images posted by users. Such generated images can be generalized to any ABSA dataset where only text annotations are available instead of sticking to the text-image dataset. More importantly, unlike the user-posted images whose flaws are from humans and are not revisable, this approach can grant us control over the association between the input text and the generated image, enabling us to iteratively adjust the images towards the positive reinforcement of the extraction.

However, it is challenging to tailor the generated image for better reinforcing the ABSA task, which requires the text-to-image model to comprehend the aspect-level information in the sample, especially when user reviews may be overly abstract and vague. Only by this can the content of the generated image be reflective and strongly associated with the aspect-level elements appeared, thereby facilitating the surpassing of the user-posted image in both application scope and downstream extraction performance.

In this study, we introduce a novel sentimental image generation method for aspect-level quadruple extraction. To craft effective images, we first propose Sentimental Paraphrasing with Emphasis Prediction. This approach serves to convert abstract user reviews into vivid scene descriptions that covers all the aspect-level elements, thereby rendering them intelligible to the text-to-image model and facilitating its creation of effective images as shown in Figure 1. Furthermore, to ensure that the generated images contribute to model performance, we subsequently introduce a Sentimental Image Assessment framework to conduct a robust assessment and contrast of images generated, it measures the text-image relevance of generated images and finally pinpoints the most suitable image across different instances.

With the sentimental image generated, we adopt a Vision-Language Model (VLM) integrated with fusion instruction to perform the extraction. The detailed evaluation shows that our model significantly advances the state-of-the-art performance on several benchmark datasets. To the best of our knowledge, our Sentimental Image Generation method stands out as the first to augment textual data with generated images, revealing a new direction for guiding large language models.

2 Related Work

Research on ABSA typically progresses from addressing individual sub-tasks to tackling their intricate combinations. Initially, the focus is often on predicting a single sentiment element (Tang et al., 2016; Chen et al., 2022; Liu et al., 2021; Seoh et al., 2021; Zhang et al., 2022). Many studies also explore the joint extractions, aiming to capture more complex sentiment information (Xu et al., 2020; Li et al., 2022; Bao et al., 2023a,b).

Recently, there has been a growing interest in tackling the ABSA problem using generative approaches (Zhang et al., 2021a). These approaches involve treating the class index (Yan et al., 2021) or the desired sentiment element sequence (Zhang et al., 2021b) as the target of the generation model, feeding a prompt to generate the sequence of aspect terms and opinion words (Yan et al., 2021; Zhang et al., 2021a; Bao et al., 2022). Furthermore, Multi-view Prompting (MVP) (Gou et al., 2023) aggregates sentiment elements generated in different orders, mimicking human-like problem-solving processes from different views.

Some works subsequently explore augmenting features from extra modalities to provide additional semantics. There are initial attempts on linguistic features, such as syntactic (Bao et al., 2022) and dependence trees (Chen et al., 2022), are combined into downstream models with linearization or graph networks. Some works further explore audio, leveraging the pitches and tones in the speech (Zhang et al., 2023a; Guo et al., 2024; Zhang et al., 2023b) to dig the implicit sentiment information behind the samples. Recently, the rise of LLMs have introduced text generation-based approaches: ATOSS (Seo et al., 2024) propose a plug-and-play module that splits input sentence into multiple aspect-oriented sub-sentences; UniGen (Choi et al., 2024) producing zero-shot dataset based on the knowledge from LLMs and SCRAP (Kim et al., 2024) distills chain-of-thought reasoning text and performs a vote over them.

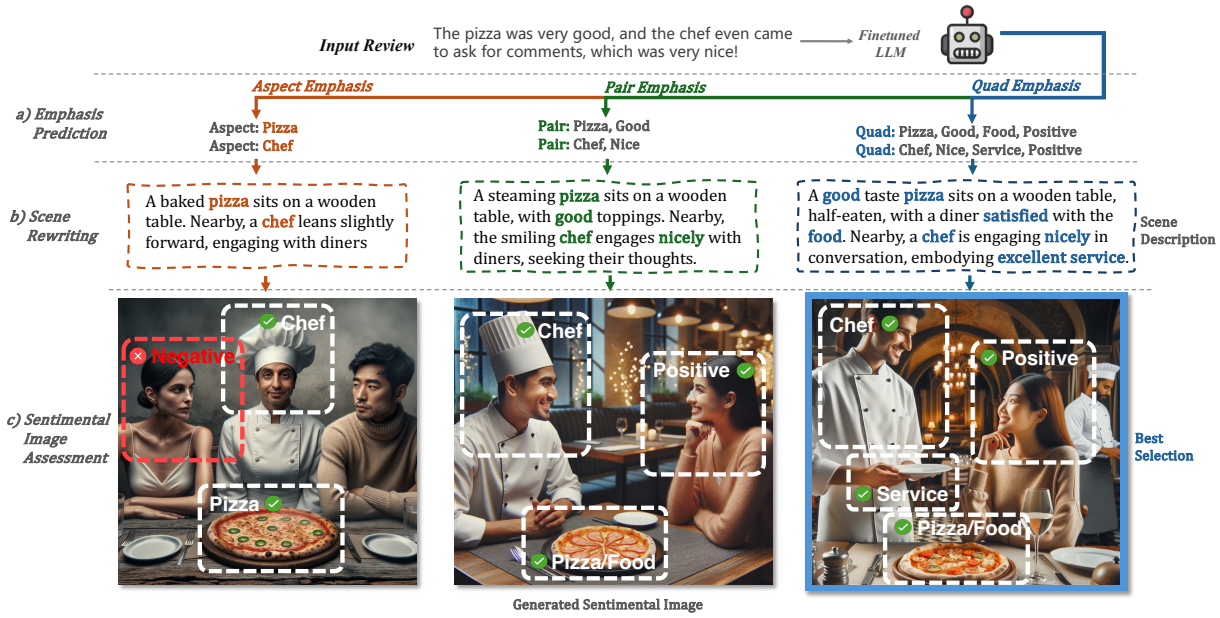


Figure 2: The illustration of our proposed Sentimental Image Generation.

In contrast to previous studies, our research stands out by first introducing generated visual content to the textual ABSA task. This novel approach significantly surpasses prior augmentation methods in enhancing the extraction of aspect-level quadruples, and more importantly, extends the applicability of visual augmentations to scenarios where only text data is accessible.

3 Sentimental Image Generation for Aspect-based Sentiment Analysis

In this study, we propose a novel sentimental image generation method which generally includes two part: Sentimental Image Generation and Sentimental Image Assessment. As shown in Figure 2, we start by crafting scene descriptions with Sentimental Paraphrasing and generating a candidate pool of images based on them. Subsequently, we assess the images' text-image relevance with Sentimental Image Assessment to identify the most fitting image as illustrated in Figure 2 (c). We further bridge the textual and visual modality with a unified vision-language model showcased in Figure 4. We will discuss these steps one by one.

3.1 Sentimental Image Generation

We first illustrate the process for generating the sentimental image. Given a customer review, it could be too abstract and missing a focused target, making it hard to be understood by text-to-image models. Besides, since the text-to-image models are not pre-trained on aspect-level tasks, they may not be able to cover the elements involved.

To solve that, we propose Sentimental Paraphrasing, the workflow of which is shown in Figure 2. Particularly, we first have Emphasis Prediction in Figure 2 a), employing a finetuned LLM to predict the silver label of sentiment elements for a given review as the semantic emphasis, making up for the relative low performance of the text-to-image model's semantic understanding. The target of Emphasis Prediction could be different combinations:

- **Aspect Emphasis** is an intuitive injection, providing the pre-predicted aspect terms as the hint since they are the core elements of the aspect-level information.
- **Pair Emphasis** provides one more element of polarity compared with the previous one to better help the model generate the explicit expression in the image.
- **Quadruple Emphasis** is similar to the Pair Emphasis, but the pre-predict and emphasis target is the quadruples to provide the comprehensive aspect-level information.

We further rewrite the original review together with the emphasis from the abstract user review to concrete scene description that can be understood by the text-to-image model with Scene Rewriting as shown in Figure 2 b). We feed them into a LLM to rewrite them into a scene description that meets

the following requirements: 1) having a customer involved with the explicit expression of sentiment polarity. 2) covering the aspect-level information emphasized. The two requirements are designed to minimize the difficulty the LLM might have in understanding the text and ensure its coverage of the semantics.

Finally, we feed the scene descriptions rewritten with different emphasises into a text-to-image model to generate a candidate pool of images. We also expand our pool with two more non-rewriting images generated based on either the original review or the predicted silver quadruple solely. Subsequently, this pool will undergo an assessment procedure for best selection in next section.

3.2 Sentimental Image Assessment

Once we finish generating the images, we need an effective method to adjust the generation result by choosing the image that could better reflect the content of the original aspect-level contents, and also check the effectiveness of our proposed Sentimental Paraphrasing.

Specifically, each image in the pool will be evaluated by Sentimental Image Assessment to choose the image that best matches the semantics of the original review, from the following perspectives:

- **Image Relevance Score** is the most intuitive one, where a similarity score will be calculated based on *Perceptual Hash Algorithm(P-Hash)*(Fei et al., 2017) between any two candidate images as shown in Figure 3 (a). Particularly, the image will be divided into $M \times M$ non-overlapping blocks and a 2D Discrete Cosine Transform (DCT) will be applied the to each block to obtain the DCT coefficients:

$$\mathrm{F(u, v)} = \sum_{x=0}^{M-1} \sum_{y=0}^{M-1} f(x,y) \cos\left(\frac{(2x+1)u\pi}{2M}\right) \cos\left(\frac{(2y+1)v\pi}{2M}\right) \quad (1)$$

  where the $f(x, y)$ is the pixel intensity at position (x, y). $F(u, v)$ is the DCT coefficient at frequency (u, v). The DCT coefficients will be quantised to obtain a 64-bit binary hash for each image, and the Hamming Distance $H(A, B)$ between two hashed image A and B will be employed as the measurement of similarity by:

$$\mathrm{H(A, B)} = \left(\sum_{i=1}^{64} |A_i - B_i| \right) / 64 \quad (2)$$

  The Hamming Distance will be calculated between any two images in the pool. The image with the lowest average Hamming Distance will be regarded as the most representative one and picked out as the final image.

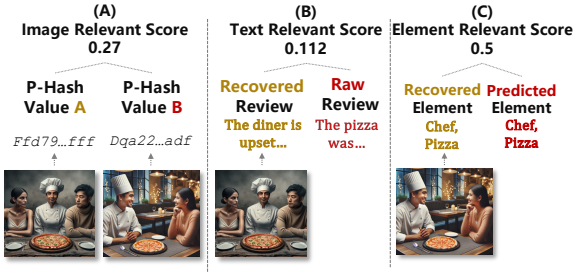


Figure 3: The illustration of proposed assessments.

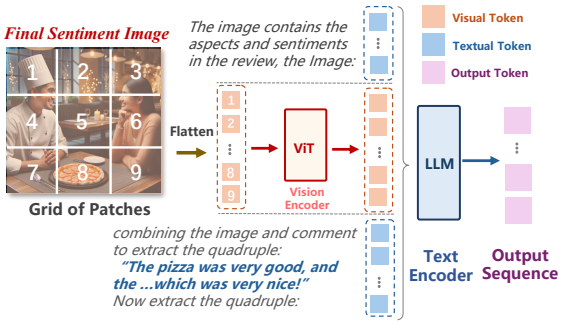


Figure 4: The illustration of our vision-language model.

- **Text Relevance Score** focuses on the origin, is designed to recover the original review based on the generated image for building a cycle evaluation. Specifically, a recovered review will be generated on the basis of the image and BLEU score (Papineni et al., 2002) will be calculated to measure the overlap between the recovered review and original review as shown in Figure 3 (b) to evaluate the image's semantic coverage.

- **Element Relevance Score**: since we expect the vision-language model to extract the sentiment element from the generated sentimental image, we employ it to interpret the generated images in the candidates pool first, asking it to summarize the aspect-level elements in each image as shown in Figure 3 (c). The final assessment score S^i of a particular image i will be calculated by the overlap rate between the summarized pairs P^i_{image} and the predicted pairs $P_{predict}$ produced during the Emphasis Prediction in Section 3.1:

$$S^i = (P^i_{image} \cap P_{predict}) / P_{predict} \quad (3)$$

3.3 Vision Encoder

Once the final sentimental image is settled, we use a Vision Transformer (ViT) as the image encoder

to learn the visual representation. Specifically, as shown in Figure 4, the input image is divided into a grid of patches, and each patch is then embedded into a visual token. The grid is then flattened into a sequence. Then the encoded image representations x_v can be obtained from image I.

3.4 Text Encoder with Fusion Instruction

We employ a LLM as our text encoder and the modality fusioner. We specifically design the instructions to fuse the textual and visual input. The fusion instruction are shown in Figure 4, which includes guiding instructions both before and after the visual tokens.

When provided with a image and text, the LLM processes the vision encoder's output as visual tokens x_v and the tokenized text as textual tokens x_{t_before} and , x_{t_after}. These tokens are subsequently merged to create the input x:

$$x = [x_{t_before},\ x_v,\ x_{t_after}] \tag{4}$$

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 in the linearized form, and decoder state h_i^d as:

$$y_i, h_i^d = ([h_1^d, ..., h_{i-1}^d], y_{i-1}) \tag{5}$$

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

$$p(y|x) = \prod_{i=1}^{|y|} p(y_i|y_{<i}, x) \tag{6}$$

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 X_T probability given the review sentence X_O. Therefore, we optimize the negative log-likelihood loss function:

$$\mathcal{L} = \frac{-1}{|\tau|} \sum_{(X_O, X_T) \in \tau} \log p(X_T|X_O; \theta) \tag{7}$$

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

4 Experiment

4.1 Dataset and Experiment Setting

In this study, we use the ABSA-ACOS (Cai et al., 2021) and en-Phone (Zhou et al., 2023) dataset and their splitting for textual ABSA experiments.

For our VLM for finetuning and Sentimental Image Assessment, we employ the pre-trained InternLM-XComposer2-VL(Dong et al., 2024) and LoRA finetune the LLM adapter parameters. In terms of the LLMs for Sentimental Paraphrasing, we employ LLaMA-3-8B (AI@Meta, 2024) as our sliver label annotator, the accuracy of which can be found in Table 8. Stable-Diffusion-3 (Esser et al., 2024) is adopt for the text-to-image model.

In evaluation, a quadruple is viewed as correct if and only if the four elements, as well as their combination (Cai et al., 2021; Zhang et al., 2021a).

4.2 Main Results

In Table 1, we present a comprehensive comparison of our model with various state-of-the-art baselines. These baselines include both classification-based and generative models, as well as LLMs.

Classification-based methods, such as TAS-BERT (Wan et al., 2020; Zhang et al., 2021a), and Extract-Classify (Cai et al., 2021), typically relies on identifying relevant spans within the input text to extract sentiment quadruples. On the other hand, generative models, such as GAS (Zhang et al., 2021b), Paraphrase (Zhang et al., 2021a), DLO (Hu et al., 2022), Seq2Path (Mao et al., 2022), OTG (Bao et al., 2022)[1], One-ASQP (Zhou et al., 2023) and MvP (Gou et al., 2023), aim to generate sentiment quadruples in target templates, potentially allowing for more flexibility and creativity in their outputs. Additionally, we also have LLMs include closed-source zero-shot ChatGPT (Ouyang et al., 2022) and LoRA fine-tuned LLaMA-3-8B (AI@Meta, 2024) as our baselines.

From Table 1 we observe that generative models easily surpass previous classification-based approaches. Furthermore, the LLM (Touvron et al., 2023) outperforms 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 Sentimental Image Generation method for quadruple extraction which enhances LLMs with semantic guid-

[1]We adopt the OTG performance without external resource for fair comparison.

Method	Restaurant			Laptop			Phone		
	P	R	F1	P	R	F1	P	R	F1
TAS-BERT	0.2629	0.4629	0.3353	0.4715	0.1922	0.2731	0.3453	0.2207	0.2693
Extract-Classify	0.3854	0.5296	0.4461	0.4556	0.2948	0.3580	0.3128	0.3323	0.3223
Seq2Path	0.6029	0.5961	0.5995	0.4448	0.4375	0.4411	0.5263	0.4994	0.5125
OTG	0.6191	0.6085	0.6164	0.4395	0.4383	0.4394	0.5302	0.5659	0.5474
One-ASQP	**0.6591**	0.5624	0.6069	0.4380	0.3954	0.4156	**0.5742**	0.5096	0.5400
GAS	0.6069	0.5852	0.5959	0.4160	0.4275	0.4217	0.5072	0.4815	0.4940
Paraphrase	0.5898	0.5911	0.5904	0.4177	0.4504	0.4334	0.4672	0.4984	0.4832
DLO	0.5904	0.6029	0.5966	0.4359	0.4367	0.4363	0.5451	0.5173	0.5308
MvP	-	-	0.6154	-	-	0.4392	-	-	-
ChatGPT	0.5014	0.3625	0.4207	0.4492	0.3123	0.3541	0.4514	0.4627	0.4569
LLaMA	0.6213	0.6024	0.6117	0.4334	0.4201	0.4266	0.5314	0.5478	0.5394
Ours	0.6544	**0.6443**	**0.6493**	**0.4543**	**0.4524**	**0.4534**	0.5312	**0.5809**	**0.5549**

Table 1: Results of textual ABSA datasets, we report the result with Pair Emphasis and Element Relevance Score.

Method		Res	Lap	Phone
Text Only		0.6030	0.4106	0.5170
With Generated Sentimental Image	*Original Review*	0.6244	0.4428	0.5323
	Silver Quadruple	0.6323	0.4464	0.5414
	Aspect Emphasis	0.6283	0.4489	0.5433
	Pair Emphasis	0.6368	0.4497	0.5468
	Quadruple Emphasis	0.6256	0.4482	0.5399
	All (Ours)	**0.6493**	**0.4534**	**0.5549**

Table 2: F1-score results of different emphasises.

ance. To the best of our knowledge, this is the first attempt to generate semantic representation in the form of images and leverage them to enhance the text-based model in ABSA task.

We also have Rest15/16 datasets(Zhang et al., 2021a) in Appendix A and the analysis of time cost in Appedix B for a holistic comparison.

4.3 Contribution of Sentimental Image Generation

After the overall performance, we first check the contribution of our Sentimental Image Generation to the overall performance. Specifically, we gradually incorporate the images generated from the proposed rewrites into the VLM. The non-rewriting images generated from the original review and the predicted silver quadruple are also included.

As depicted in Table 2, when using only textual features, the performance of VLM is notably low, underscoring the necessity of enriched features to achieve SOTA results in complex tasks like quadruple extraction. Significantly improvement are observed when the generated sentimental images is included in the input, highlighting the superiority of the visual modality in capturing semantics.

Furthermore, all the proposed emphasises contribute positively to quadruple extraction and surpass the two non-rewriting images, demonstrating the effectiveness of our proposed Sentimental Paraphrasing. This technique is designed for ensuring image's comprehensive coverage of the review's semantics and making it easy to be understood by our VLM. Among these emphases, Pair Emphasis outperforms Quadruple Emphasis, we believe the reason is due to the intricate and voluminous information encapsulated in Quadruple Emphasis, which may potentially overwhelm the text-to-image model because of its relative low performance in semantic understanding since it is not trained or finetuned on this task.

Additionally, our proposed model, which combines all the image enhancement methods to incorporate visual guiding, achieves the best performance and showcases the value of visual sentiment semantics in sentiment analysis. We also show our sentimental image generation is generalize towards various ABSA subtasks in Appendix C.

4.4 Effectiveness of Sentimental Image Assessment

We subsequently check whether the relevant scores produced by our proposed Sentimental Image Assessment can effectively pinpoint the image that is capable of enhancing the VLM performance, which indicates they have a superior text-image relevance. Specifically, we investigate this by making a comparison between our proposed assessments and the best single image generation method Pair Emphasis

Augmentation Type	Method	Twitter2015	Twitter2017
Text Baseline		0.598	0.613
Textual	MvP+ATOSS	0.653	0.654
	SCRAP	0.648	0.659
Linguistic	OTG	0.631	0.633
Original Visual	OSCGA+TomBERT	0.632	0.635
	JML	0.641	0.660
	VLP-MABSA	0.666	0.680
	Ours (Original)	0.662	0.676
Generated Visual	Ours (Generated)	0.674	0.687
	Ours (Generated+Original)	**0.678**	**0.690**

Table 3: Results of different augmentations in Twitter2015/17 datasets.

Method	Res	Lap	Phone
Single Generation	0.6368	0.4497	0.5468
Image Relevance Score	0.6395	0.4476	0.5487
Text Relevance Score	0.6426	0.4512	0.5525
Element Relevance Score(Ours)	**0.6493**	**0.4534**	**0.5549**

Table 4: Results of different assessment scores.

found in previous section.

We show that our image assessment can effectively pick out the high-quality image and improve the overall performance in Table 4, where all of our assessments can surpass the best single generation baseline, giving us a conclusion that combining different generation paths can provide us more comprehensive semantics. Among them, the Element Relevance Score outperforms the other two, we believe this due to its congeniality with the finetuning: the target of both two tasks are extracting the elements instead of the reviews or images.

5 Analysis and Discussion

5.1 Comparison of Augmentations

We subsequently make a comparison of different augmentation methods. We switch our benchmark to the multi-modal ABSA (MABSA) dataset Twitter2015/17 (Yu and Jiang, 2019) to facilitate the comparison with user-posted images. The comparison include: **Textual Augmentations**: 1) MvP+ATOSS (Seo et al., 2024). 2) SCRAP (Kim et al., 2024). **Linguistic Augmentations**: OTG Bao et al. (2022). **Original Visual Augmentations** that the augmenting images are user-posted: 1) OSCGA+TomBERT (Yu and Jiang, 2019); 2) JML (Ju et al., 2021); 3) VLP-MABSA (Ling et al., 2022); 3) Ours (Original). **Generated Visual Augmentations** where the augmenting images are generated sentimental images: 1) Ours (Generated) represents generated images; 2) Ours (Generated+Original) represents the original image will be fed together with the generated image. We also have the baseline that solely rely on the original sentences, named "Text Baseline".

Referring to Table 3, it is evident that the methods based on visual augmentations surpass the textual and linguistic augmentations by a considerable margin when incorporating either generated or posted images, showing the superiority of visual augmentations in supplementing textual tasks. On the other side, as most of the knowledge introduced in the text and linguistic-based augmentations have heavy overlap with original sample, their lower performance is expected.

Furthermore, inside the visual augmentations, the generated image outperforms the original image. We attribute this superiority to the generated image's ability to offer a more explicit text-image association, while the original image's representation appears comparatively vague, could miss the significant expression of sentiment polarity or aspect terms, making their images less informative. It also hints at a novel avenue for exploration: substituting user-posted content with model-generated.

In addition, the combination of the two types of the images achieves the SOTA performance in MABSA task. This can be attributed to the enriched semantic information provided by the combination, and also reinforces the significance of visual sentiment semantics in sentiment analysis.

5.2 Analysis of Data Efficiency

When compared to textual content, one of the advantages of generated sentimental images is the presence of a large number of shared portrayals, such as smiling faces, which can express polarities more explicitly. This explicit representation makes it easier to establish semantic connections across samples. We thus investigate how the generated

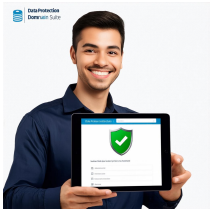
Review	Bus Selfie on the way to Harry Potter Studios @ WFCTrust @ NCSEast ShareYourSummer.	RT @ KTVU : UPDATE Protesters have blocked traffic on both sides of I - 80 at University in # Berkeley	BCMS students stoked to meet Clara ! ! clarasbigride.	We have you covered. See why the Data Protection Suite ampData Domain are even better together.
without Sentimental Image	(***Bus Selfie***, Positive) ✗	(***both sides of I - 80***, Neutral) ✗	(BCMS, ***Negative***) ✗ (Clara, Positive)✓	(Data Protection Suite,Neutral)✓ (Data Domain, ***Neutral***)✗
Generated Sentimental Image				
with Sentimental Image	(***Harry Potter Studios***, Positive) ✓	(***University in # Berkeley***, Neutral) ✓	(BCMS, ***Positive***) ✓ (Clara, Positive)✓	(Data Protection Suite,Neutral)✓ (Data Domain, ***Positive***)✓

Table 5: Cases studies for our generated sentimental image.

sentimental image improves the data efficiency by comparing with using textual modality solely under limited training data in Figure 5.

From the figure, we find that the more training data, the higher performance our proposed model can reach. Moreover, the improvement brought by the generated image information increases under limited data size, showing the superiority of visual sentiment semantics in low resource situation, where a pool of shared features can be easily built compared with relying on textual modality solely.

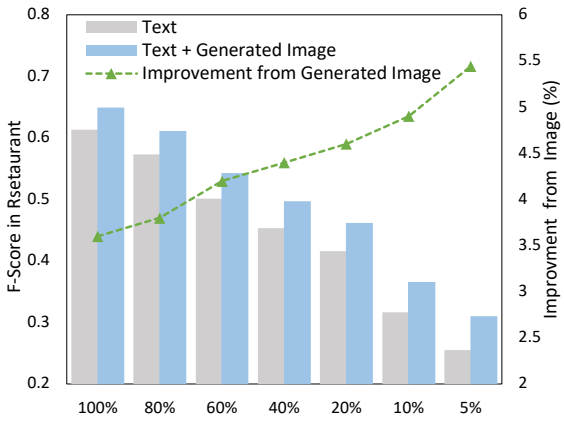


Figure 5: Improvement of data efficiency.

6 Cases Studies

We launch case studies to make a more intuitive comparison between the extraction result with and without our generated image in Table 5.

We show that generated sentimental images can effectively capture the intended target in the first two examples. The extraction without generated images in the first two example misses "Harry Potter Studios" and "University in # Berkeley" respectively, while our generated sentimental images successfully cover them, aiding the VLM in identifying the correct elements.

Furthermore, we illustrate that generated sentimental images can better convey sentiment polarity in the last two examples. The extraction without generated images in the third example successfully captures the aspect target but lacks discernible polarity. It performs similarly in the last example, wrongly classifies into Neutral polarity, whereas our generated image explicitly conveys a correct Positive polarity and helps the final classification.

From the cases shown in Table 5, we can find that, with the enhancement of the generated sentimental image, our method shows significant superiority in improving aspect-level extraction.

7 Conclusion

In this study, we address the long-overlooked limitations of existing data augmentation methods of textual ABSA and shift our focus toward generating sentimental images from scratch as a promising alternative. With proposed Sentimental Image Generation and Assessment, we generate effective images to assist textual ABSA, achieving SOTA performance in multiple benchmarks.

Our results also validate that, in addition to the conventional approaches of incorporating extra user-posted features, leaning on machines-generated features generated from scratch could also be considered as an efficiently way to provide us with supplementary semantic insights.

Limitations

The limitations of our work can be stated from two perspectives. Firstly, besides the image, there is another feature whose effect on downstream tasks is not yet known such speech. In future research, further exploration of the impact of text-to-speech could provide valuable insights.

Secondly, our focus has been primarily on utilizing image generation in ABSA. While we have achieved promising results in them, it is important to acknowledge that the performance of our approach in other field such as event extraction remains unknown. Extending our investigation to other tasks would allow us to gain a more comprehensive understanding of the generalizability and effectiveness of our methodology.

References

AI@Meta. 2024. Llama 3 model card.

Xiaoyi Bao, Xiaotong Jiang, Zhongqing Wang, Yue Zhang, and Guodong Zhou. 2023a. Opinion tree parsing for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7971–7984. 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. 2023b. 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.

Hongjie Cai, Yaofeng Tu, Xiangsheng Zhou, Jianfei Yu, and Rui Xia. 2020. Aspect-category based sentiment analysis with hierarchical graph convolutional network. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 833–843, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350, Online. 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.

Shaowei Chen, Yu Wang, Jie Liu, and Yuelin Wang. 2021. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12666–12674.

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, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis.

Mengjuan Fei, Zhaojie Ju, Xiantong Zhen, and Jing Li. 2017. Real-time visual tracking based on improved perceptual hashing. *Multimedia Tools Appl.*, 76(3):4617–4634.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. MvP: Multi-view prompting improves aspect sentiment tuple prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4380–4397, Toronto, Canada. Association for Computational Linguistics.

Zirun Guo, Tao Jin, and Zhou Zhao. 2024. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1726–1736, Bangkok, Thailand. Association for Computational Linguistics.

Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai, and Shiwan Zhao. 2022. Improving aspect sentiment quad prediction via template-order data augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*,

pages 7889–7900, Abu Dhabi, United Arab Emirates. 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.

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.

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.

Yue Mao, Yi Shen, Jingchao Yang, Xiaoying Zhu, and Longjun Cai. 2022. Seq2Path: Generating sentiment tuples as paths of a tree. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2215–2225, Dublin, Ireland. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion Word Expansion and Target Extraction through Double Propagation. *Computational Linguistics*, 37(1):9–27.

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.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z. Pan. 2020. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In *AAAI 2020*, pages 9122–9129.

Luo Xianlong, Meng Yang, and Yihao Wang. 2023. Tagging-assisted generation model with encoder and decoder supervision for aspect sentiment triplet extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2078–2093, Singapore. 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.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2416–2429, Online. Association for Computational Linguistics.

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

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023a. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 756–767, Singapore. Association for Computational Linguistics.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510, Online. Association for Computational Linguistics.

Xinlang Zhang, Zhongqing Wang, and Peifeng Li. 2023b. Multimodal chinese event extraction on text and audio. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

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.

Junxian Zhou, Haiqin Yang, Yuxuan He, Hao Mou, and Junbo Yang. 2023. A unified one-step solution for aspect sentiment quad prediction. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12249–12265. Association for Computational Linguistics.

A Results in ASQP Datasets

In Table 6, we show our proposed model can also achieve the state-of-the-art performance on ASQP (Zhang et al., 2021b). These baselines include both classification-based methods and generative models, include classification-based methods, such as TASO-BERT-CRF (Cai et al., 2021) and generative models, such as GAS (Zhang et al., 2021b), Paraphrase (Zhang et al., 2021a), DLO (Hu et al., 2022) and MvP (Gou et al., 2023).

Our proposed model achieves statistically significant improvements over all previous studies ($p < 0.05$) on the ASQP (Zhang et al., 2021b) dataset, demonstrating the effectiveness and generalization of our Sentimental Image Generation method when applied to quadruple extraction.

B Analysis of Inference Cost

In this section, we compare our method with other augmentation methods including MvP and SCRAP we mentioned before as in the following Table 7, We implement all of them with LLaMA-3-8B for the right. The speed is measured with seconds of generating 100 samples. The first two strategies are the textual argumentation baselines MvP+ATOSS and SCRAP. The other one is the linguistic argumentation baseline OTG.

As evident from the results, SCRAP is the slowest as it needs to pre-generate even 20 times to get the augmentations, on the other hand, OTG achieves the fastest speed. However, this comes at the cost of reduced performance as it only generates once, having very narrow searching space. If we take both aspects into consideration, our method emerges as the clear winner. It outperforms all other strategies while maintaining an acceptable inference speed.

C Analysis of Generalization

As a complex task, ABSA contains multiple subtasks that focus on analysing different combination of targets. To fully explore the generalization of our sentimental image, we analyze it with the LLaMA-3-8B in different ABSA subtask, which also serves as our powerful silver label annotator in Section 3.1. In particular, there are seven popular subtask:

- **AE** is the most basic subtask, it means the single extraction of the aspect term.
- **AO/Pair** means that we only extract aspect term and opinion term from review text (Qiu et al., 2011; Xu et al., 2020; Li et al., 2022).
- **OS** means that we extract opinion term and polarity from review text.
- **AC** means that we extract aspect term and category from review text.

Method	Rest15			Rest16		
	P	R	F1	P	R	F1
HGCN-BERT+BERT-TFM*	0.2555	0.2201	0.2365	0.2740	0.2641	0.2690
TASO-BERT-CRF*	0.4424	0.2866	0.3478	0.4865	0.3968	0.4371
GAS	0.4531	0.4670	0.4598	0.5454	0.5762	0.5604
Paraphrase	0.4616	0.4772	0.4693	0.5663	0.5930	0.5793
DLO	0.4708	0.4933	0.4818	0.5792	0.6180	0.5979
MvP	-	-	0.5221	-	-	0.6039
Ours	**0.5359**	**0.5433**	**0.5396**	**0.6447**	**0.6245**	**0.6344**

Table 6: Results of textual datasets Rest15/16. The results are obtained from Hu et al. (2022) and Gou et al. (2023)

Method	Time(s)	Res	Lap	Phone
MvP+ATOSS	423.13	0.6154	0.4392	-
SCRAP	1767.49	0.6095	0.4313	0.5237
OTG	283.12	0.6164	0.4394	0.5474
Ours	876.97	**0.6493**	**0.4534**	**0.5549**

Table 7: Results of inference time. The speed is measured with seconds of generating 100 samples.

Subtask	Domain	LLaMA	Ours
AE	Restaurant	0.7739	0.7984
	Laptop	0.7684	0.7910
	Phone	0.7957	0.8132
AO	Restaurant	0.6906	0.7381
	Laptop	0.7201	0.7602
	Phone	0.7123	0.7294
OS	Restaurant	0.7306	0.7629
	Laptop	0.7412	0.7511
	Phone	0.7233	0.7484
AC	Restaurant	0.6703	0.6922
	Laptop	0.6828	0.7199
	Phone	0.7356	0.7462
AOS	Restaurant	0.6582	0.6881
	Laptop	0.6461	0.6843
	Phone	0.7038	0.7323
ACS	Restaurant	0.6323	0.6624
	Laptop	0.4412	0.4690
	Phone	0.5643	0.5702
ACOS	Restaurant	0.6117	0.6493
	Laptop	0.4266	0.4534
	Phone	0.5394	0.5549

Table 8: Results our method's generalization towards different ABSA subtask, LLaMA also serves as our sliver label annotator, the results are measured by F1-score.

- **AOS/Triple** means that we extract aspect term, opinion term, and polarity from review text (Zhang et al., 2021b; Chen et al., 2021).
- **ACS** means that we extract aspect term, category, and polarity from review text.
- **ACOS/Quad** is the quadruple schema that extracts all four sentiment elements to form the opinion quadruple (Cai et al., 2020; Zhang et al., 2021a; Bao et al., 2022).

Note that, we make minor modifications to our workflow, and let it suitable for the corresponding subtask (i.e., delete the prediction of silver opinion words in ACS subtask). From Table 8, we can find that our model outperforms LLaMA in all the schemas. It indicates that our sentimential image is generalized and can be used to handle different subtasks in aspect-based sentiment analysis.