An Alternative to Human Annotation: Sentimental Image Generation for Aspect-based Sentiment Analysis

Anonymous ACL submission

Abstract

There have been multiple attempts to extract additional semantics from images in aspect-based sentiment analysis. However, relying on user-generated images limits their application scope to just a handful of text-image datasets and can not be grafted to massive text-only tasks, not to mention these images could be less informative and cannot fully match the semantics in user reviews. We argue that, compared with designing sophisticated model relying on the user-generated images, generating robust and comprehensive images from scratch could not only expand our application scope to unimodal textual tasks but also provide us more accurate information in multi-modal tasks. In this study, we propose novel Sentimental Image Generation and Sentimental Image Assessment, the former can precisely convert textual semantics into visual images as shown in Figure 1 while the latter can facilitate comprehensive image evaluation from multiple perspectives and select the most fit image to proceed. Finally, we adopt a unified vision-language model that is compatible with both uni- and multi-modal tasks for extraction. Extensive experiments build a new SOTA performance in several benchmarks, underscoring the effectiveness of our method and highlighting a promising direction for expanding our features.

1 Introduction

Aspect-based sentiment analysis (ABSA) as a topic of increasing interest in the research community, is comprised of four subtasks: aspect extraction, opinion extraction, aspect category classification, and aspect-level sentiment classification. The sentiment quadruple extraction task, which combines these four subtasks as shown in Figure 1, presents a significant challenge for traditional classification-based models. To address this challenge, recent research has adopted a unified generative approach that avoid the explicit modeling. These approaches

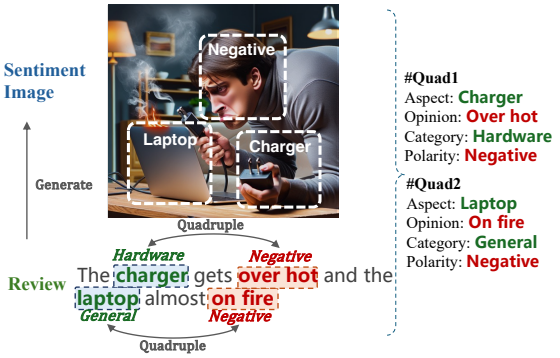


Figure 1: Example of sentimental image generation.

either craft complex features (Li et al., 2023; Xianlong et al., 2023; Bao et al., 2023b), or specify the desired target sequence (Yan et al., 2021; Zhang et al., 2021c,b; Bao et al., 2022; Hu et al., 2022).

Despite their effectiveness, most previous studies have concentrated on textual input (Bao et al., 2022; Hu et al., 2022), relying on various templates or linguistic features. As a result, these models are defective because they do not consider other data sources like images, which could be complementary as: (1) The aspect term in a sentence could be implicit (Cai et al., 2021) without specific words indicating it, but could be highlighted as a focused target in the image; (2) The polarity over a aspect could be hard to detect due to the short and informal nature of the sentence, but the associated image may help explicitly reflect it. However, it is challenging to incorporate images because there are no images associated with the reviews in the mainstream text-only datasets (Cai et al., 2021; Zhang et al., 2021b), making it inapplicable when considering any image data.

And for existing Multi-modal ABSA (MABSA) works (Ju et al., 2021; Sun et al., 2021), they rely on the user-generated images in the MABSA datasets such as Twitter2015/17 (Yu and Jiang, 2019), expose two obvious shortcomings: The first is the source of the images, which are defective for senti-

ment analysis as they are generated by users, could miss the significant expression of sentiment polarity or explicit focused target, making their images less informative. The second is their generalization. Previous works' (Sun et al., 2020; Zhang et al., 2021a) dependence on the user-generated images also shrinks their generalization scope, limiting their application only to the MABSA tasks and could not be grafted to other tasks.

We thus move our focus on generating the image from scratch as an alternative to the images posted by users, for not only providing precise and comprehensive semantics, but also making it better generalized to any situations where only text annotations are available. However, the quality of the generated images may could not reach the ideal level and wrongly provide miss-leading information. In this situation, two questions are raised: *1) how do we improve the quality of the generated image?* and the derived question: *2) how do we evaluate the quality of these images?* The former is to guarantee the quality of the generated images, which is the key for replacing the user-generated image and improving the overall performance. The later is to provide us an effective way for assessing the quality of images, and subsequently we can easily compare and pick out the image with highest quality to proceed.

To explore the first question, we introduce a novel sentimental image generation method to improve the overall quality for extraction. The key to crafting a high-quality image with a text-to-image model is ensuring it comprehends what elements to illustrate, especially when user reviews may be overly abstract. To tackle this challenge, we propose Sentimental Paraphrasing with Emphasis Prediction. This approach serves to convert abstract user reviews into vivid scene descriptions, thereby rendering them intelligible to the text-to-image model and facilitating the creation of high-quality images as shown in Figure 1. We further answer the second question with a novel Sentimental Image Assessment framework to measure the quality of generated images effectively. This system enables us to conduct robust assessment and contrast of images to pinpoint the most suitable image across different instances.

With the high-quality image generated, we adopt a Vision-Language Model (VLM) integrated with fusion instruction that adaptive to both uni- and multi-modal ABSA tasks to perform the extraction. The detailed evaluation shows that our model significantly advances the state-of-the-art performance on several uni- and multi-modal benchmark datasets. In addition, our Sentimental Image Generation method also reveals us a new direction for guiding large language models.

2 Related Work

2.1 Aspect-Based Sentiment Analysis

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).

Recently, there has been a growing interest in tackling the ABSA problem using generative approaches (Li et al., 2023; Xianlong et al., 2023; Bao et al., 2023b). These approaches involve treating the class index (Yan et al., 2021), the desired sentiment element sequence (Zhang et al., 2021c) or flatten structure (Bao et al., 2022) as the target of the generation model. Furthermore, Multi-view Prompting (MVP) (Gou et al., 2023) aggregates sentiment elements generated in different orders, mimicking different human views.

2.2 Multi-modal ABSA

The research on multi-modal ABSA(MABSA) generally follows the same routine as the uni-modal, starting at handling the aspect term (Zhang et al., 2018; Yu et al., 2020b; Wu et al., 2020a,b; Sun et al., 2020; Zhang et al., 2021a) and polarity classification (Xu et al., 2019; Yu et al., 2020a; Yang et al., 2021; Khan and Fu, 2021). The recent works tends to take the task a step further, they extract elements jointly (Ju et al., 2021). For example, JML (Ju et al., 2021) is proposed as multi-modal joint learning approach with cross-modal relation while Ling et al. (2022) proposed a VLP-MABSA that suits all the pre-training and downstream tasks.

Unlike prior studies that leaned on user-generated images, our research distinguishes itself by producing sentimental images that accurately capture the semantics of textual content from scratch. Further more, our work also facilitates superior generalization, enabling us to broaden our application across both uni- and multi-modal tasks.

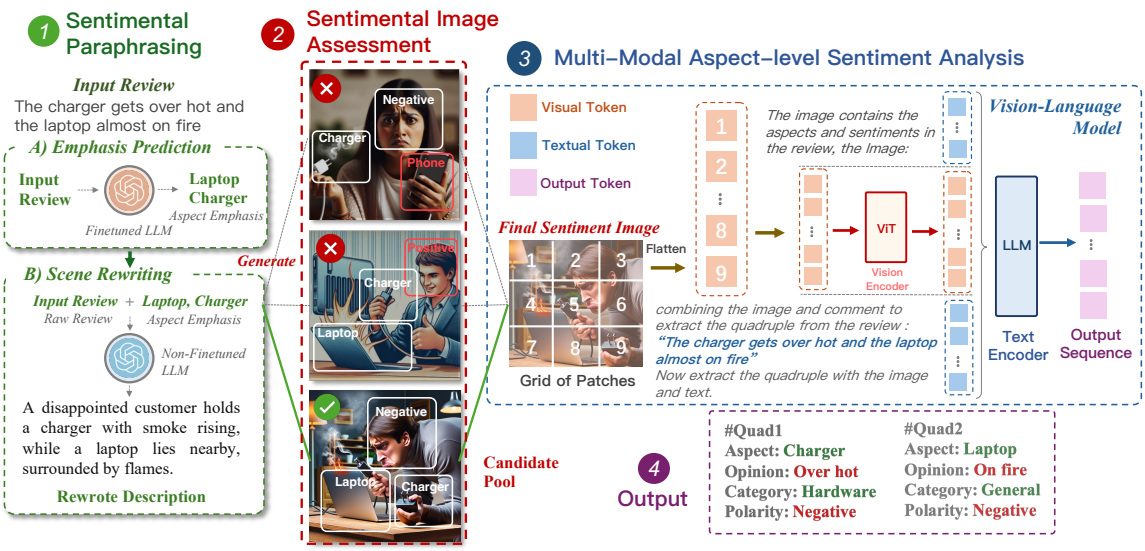


Figure 2: The illustration of our proposed method.

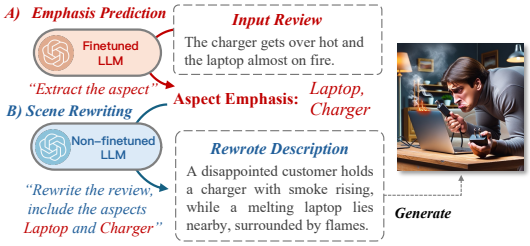


Figure 3: Process of the Sentimental Image Generation.

3 Sentimental Image Generation for Aspect-based Sentiment Analysis

In this study, we propose a novel sentimental image generation method. As shown in Figure 2 (1), we start by crafting scene descriptions with Sentimental Paraphrasing and generating a candidate pool of images based on them. Subsequently, we assess the images' quality using various metrics to identify the most fitting image that encapsulates the semantics of the original review, as illustrated in Figure 2 (2). We further bridge the textual and visual modality with a unified vision-language model showcased in Figure 2 (3). 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 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 3. Particularly, we first have Emphasis Prediction, 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 abstract user review to concrete scene description that can be understood by the text-to-image model with Scene Rewriting as shown in Figure 3 b). We feed them into a LLM to rewrite them into a scene description that meet 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

minimum the difficulty of model's understanding 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 assess procedure for best selection in next section.

3.2 Sentimental Image Assessment

However, even we can generate the high-quality images, we still need to pick out the one with the highest quality. In this case, an effective method is needed for not only choosing the image that could better reflect the content of the original review, but also checking 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 Relevant 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 4 (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:

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

  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.

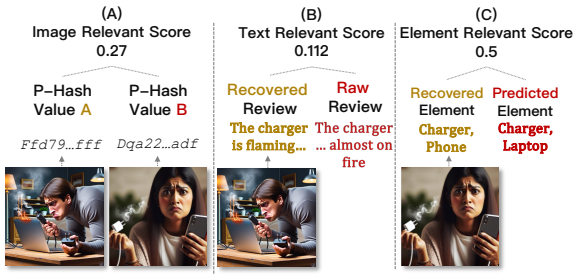


Figure 4: The illustration of proposed assessments.

- **Text Relevant Score** focuses on the origin, is designed to recover the generated image back to the review for building a cycle evaluation. Specifically, a recovered review will be generated on the basis of the image and BLUE score (Papineni et al., 2002) will be calculated between the recovered review and original review as shown in Figure 4 (b) to measure the semantic coverage of the image.

- **Element Relevant Score**: since we expect the vision-language model to extract the sentiment element from the sentimental image, we employ it to interpret the sentimental images in the candidates pool first, asking it to summary the aspects and the corresponding polarity appear in each image as shown in Figure 4 (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 down, we use Vision Transformer (ViT) as the image encoder to learn the visual representation. Specifically, as shown in Figure 2, the input image is divided into a grid of patches, and each patch is then embedded into a visual token. The grid is then flatten 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 the LLM as our text encoder also the modality fusioner. We specifically design the instructions for fuse the textual and visual input. The fusion instruction are designed as shown in Figure 2 (3), which include a guiding instruction at both

before and after the visual tokens, along with the specific text for extracting.

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, The decoder predicts the output sequence token-by-token. 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 functions is to maximize 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, then

$$\log p(X_T|X_O;\theta) = \sum_{i=1}^{n} \log p(x_T^i|x_T^1, x_T^2, ...x_T^{i-1}, X_O;\theta) \tag{8}$$

where $p(x_T^i|x_T^1, x_T^2, ...x_T^{i-1}, X_O;\theta)$ is calculated by decoder.

4 Experiment

In this section, we introduce the datasets used for evaluation and the baseline methods employed for comparison. We then report the experimental results and analysis conducted on uni-model task with different factors. The results of multi-model experiment will be introduced in next section.

4.1 Dataset and Experiment Setting

In this study, we use ABSA-ACOS (Cai et al., 2021) for uni-modal ABSA and Twitter2015/2017 (Yu and Jiang, 2019) dataset for multi-modal ABSA with their splitting.

For our VLM for finetuning and Sentimental Image Assessment, we employ the pre-trained weight 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). Stable-Diffusion-3 (Esser et al., 2024) is adopt for the text-to-image model for image generation.

In evaluation, a quadruple is viewed as correct if and only if the four elements, as well as their combination, are exactly the same as those in the gold quadruple (Cai et al., 2021; Zhang et al., 2021b).

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., 2021b), and Extract-Classify (Cai et al., 2021), typically rely on identifying relevant spans within the input text to extract sentiment quadruples. On the other hand, generative models, such as GAS (Zhang et al., 2021c), Paraphrase (Zhang et al., 2021b), 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. Besides, 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 large amount of methods without complex modeling, showing its great power in 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 when applied to LLM through extending its

[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 uni-modal datasets, we report the result with Pair Emphasis and Element Relevant Score.

semantic guiding. 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., 2021b) in Appendix A for holistic comparisons.

4.3 Contribution of Sentimental Image Generation

After analyzing the overall performance, we first check if our Sentimental Image Generation can answer the question *how do we improve the quality of the generated image?* by checking their contributions to the overall performance, where the higher performance means the higher image quality. Specifically, we gradually incorporate the image generate from the proposed rewrites into VLM. The non-rewriting images 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-level results in complex tasks like quadruple extraction. Significantly improved performance are observed when the sentimental images is included in the input, highlighting the superiority of 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 the comprehensive coverage of information within the review and facilitate the understanding of our VLM. Among these emphasises, Pair Emphasis outperforms Quadruple Emphasis, we

Method		Res	Lap	Phone
Text Only		0.6030	0.4106	0.5170
With 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.

believe the reason locates at 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.

4.4 Effectiveness of Sentimental Image Assessment

We subsequently answer the second question *how do we evaluate the quality of images?* To answer it, we need to check whether the scores produced by our proposed Sentimental Image Assessment can effectively pinpoint the image that be capable of enhancing the VLM performance, which indicates they have a superior quality in contrast to the disre-

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

Table 3: Results of different assessment scores.

Method	Twitter2015	Twitter2017
UMT+TomBERT	0.598	0.624
OSCGA+TomBERT	0.625	0.637
OSCGA-collapse	0.632	0.635
RpBERT-collapse	0.480	0.562
JML	0.641	0.660
VLP-MABSA	0.666	0.680
LLaVA	0.623	0.638
Ours (Original)	0.662	0.676
Ours (Generated)	0.674	0.687
Ours (Generated+Original)	**0.678**	**0.690**

Table 4: F1-score results in MABSA datasets Twitter2015/17. Baselines are from Ling et al. (2022)

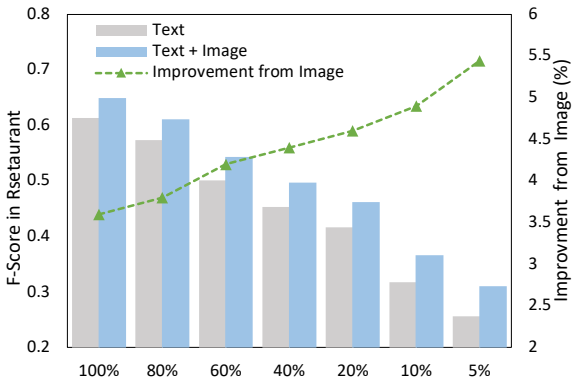


Figure 5: Improvement of data efficiency.

garded images. Specifically, we investigate this by making a comparison between our proposed assessments and the best single image generation method Pair Emphasis 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 3, 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 the assessments, the Element Relevant Score outperforms the other two, we believe this due to its congeniality with the fine-tuning: the target of both two tasks are extracting the elements instead of the reviews or images.

5 Analysis and Discussion

5.1 Results on Multi-modal Aspect-based Sentiment Analysis

Although our method is proposed for enhancing text-based models, it is also compatible with Multi-modal Aspect-based Sentiment Analysis (MABSA). We thus investigate whether our model can work in MABSA with Twitter2015/17 (Yu and Jiang, 2019) in Table 4, where we have three formations of input: 1) *Ours (Original)*, the inputted images will be the original user-generated images from the dataset; 2) *Ours (Generated)* represents inputting generated images only; 3) *Ours (Generated+Original)*, the original images will be feed before the generated image. Our baselines include 1) UMT/OSCGA+TomBERT (Yu and Jiang, 2019) 2) UMT-collapsed (Yu et al., 2020b); 3) OSCGA-collapsed (Wu et al., 2020b); 4) RpBERT-collapsed (Sun et al., 2021); 5) JML (Ju et al., 2021); 6) LoRA finetuned VLM LLaVA (Liu et al., 2023).

Referring to Table 4, it is evident that our *Ours (Generated)* approach surpasses all MABSA baselines by a considerable margin when incorporating the generated sentimental images. This result underscores the effectiveness of the sentimental image generation not only for uni-modal tasks but also for multi-modal tasks.

Furthermore, the generated image outperforms the original image. We attribute this superiority to the generated image's ability to offer a more explicit and focused target representation along with the sentiment polarity, while the original image's representation appears comparatively vague. This disparity highlights the capability of generated images to capture sentiment-related semantics with greater precision. It also hints at a novel avenue for exploration: substituting user-generated content with model-generated. We will have cases studies in next section for an more intuitive comparison.

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 sentimental images is the presence of a large number of shared portrayals, such as smiling faces, which can express polarities more explic-

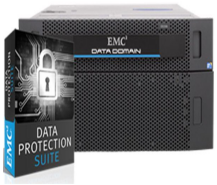
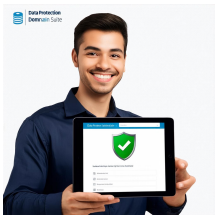
Review	Bus Selfie on the way to Harry Potter Studios @ WFCTrust @ NCSEast ShareYourSummer.	RT @ KTVU : UPDATE Protesters have blocked traffic on both sides of I - 80 at University in # Berkeley	BCMS students stoked to meet Clara ! ! clarasbigride.	We have you covered. See why the Data Protection Suite ampData Domain are even better together.
Original Image				
Output	(***Bus Selfie***, Positive) ✗	(***both sides of I - 80***, Netural) ✗	(BCMS, ***Negative***) ✗ (Clara, Positive)✓	(Data Protection Suite,Netural)✓ (Data Domain, ***Netural***)✗
Sentimental Image				
Output	(***Harry Potter Studios***, Positive) ✓	(***University in # Berkeley***, Netural) ✓	(BCMS, ***Positive***) ✓ (Clara, Positive)✓	(Data Protection Suite,Netural)✓ (Data Domain, ***Positive***)✓

Table 5: Cases studies for multi-modal Aspect-based Sentiment Analysis.

itly. This explicit representation makes it easier to establish semantic connections across samples, particularly when dealing with a limited amount of training data. We thus investigate how the sentimental image improves the data efficiency of our method 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 sentimental 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.

6 Cases Studies

We launch case studies on MABSA data to make a more intuitive comparison between the user-posted image and our generated image in Table 7.

We demonstrate that sentimental images can more effectively capture the intended target in the first two examples. The original image in the first two example misses "Harry Potter" and university campus respectively. while our sentimental image covers them, aiding the VLM in identifying the correct elements.

Furthermore, we illustrate that sentimental images can better convey sentiment polarity in the last two examples. The original image in the third example effectively captures the students but lacks discernible polarity. Similarly, the original image in the last example merely presents a product, whereas ours explicitly conveys a polarity attitude.

From the cases shown, we can find our generated sentimental image shows significant superiority and can better cover the semantics and in MABSA task. We also have uni-modal cases studies in Appendix B for holistic comparisons.

7 Conclusion

In this study, we move our sight outside the textual features, generating sentimental images from scratch and assessing them to provide extra visual semantics in aspect-level quadruple extraction. By extending the guiding for model to cross-modal scope, our proposed visual-language model achieves SOTA performance in both uni- and multi-modal benchmarks without the need for complex and costly annotation of additional features.

Furthermore, our results also validate that, besides from the conventional approaches of incorporating extra user-generated or linguistic features in downstream tasks, generating features from other modalities from scratch could also be considered as an efficiently way to provide us extra and robust semantic information.

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 two ABSA subtasks. While we have achieved promising results in them, it is important to acknowledge that the performance of our approach in other subtasks remains unknown. Extending our investigation to other ABSA subtasks and information extraction 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, 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.

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.

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.

Zaid Khan and Yun Fu. 2021. Exploiting bert for multimodal target sentiment classification through input space translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 3034–3042, New York, NY, USA. Association for Computing Machinery.

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.

Pan Li, Ping Li, and Kai Zhang. 2023. Dual-channel span for aspect sentiment triplet extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 248–261, Singapore. 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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. 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.

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.

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

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.

Hanqian Wu, Siliang Cheng, Jingjing Wang, Shoushan Li, and Lian Chi. 2020a. Multimodal aspect extraction with region-aware alignment network. In *Natural Language Processing and Chinese Computing*, pages 145–156, Cham. Springer International Publishing.

Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020b. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1038–1046, New York, NY, USA. Association for Computing Machinery.

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.

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.

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.

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. 2021a. Multi-modal graph fusion for named entity recognition with targeted visual guidance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14347–14355.

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).

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021b. 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. 2021c. 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.

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., 2021c). 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., 2021c), Paraphrase (Zhang et al., 2021b), 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., 2021c) dataset, demonstrating the effectiveness and generalization of our sentimental image generation method when applied to large language models, particularly through the extension of semantic guidance.

B Cases Studies of Uni-Modal Aspect-based Sentiment Analysis

We launch case studies from ABSA-ACOS datasets to make a more intuitive illustration of the effects of the sentimental image in uni-modal sentiment quadruple extraction. As demonstrated in Table 7, the correct prediction would be with a ✓ notation.

In the first example, without sentimental image, the model wrongly predicts the "*best*" as one of the opinion term, which obviously represents a positive polarity that conflict with the general sentiment polarity of the review. With the generated sentimental image that explicitly express a negative sentiment, our method easily excludes it.

In the second example, the review is missing leading for the model as it has the "*chefs*" that may cause confusing. In the generated sentimental image, the "*fish*" is presented as the focused target that

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 uni-modal datasets Rest15/16. The results are obtained from Hu et al. (2022) and Gou et al. (2023)

Text	this is sad for what once was one of the best places you could ever eat	good for that one night once in a blue moon when the chefs decideto use fish that's half-way decent	the backlit keyboard is awesome at night	however , mine came with an annoying trackpad
Sentimental Image				
w/o Image	(restaurant general, null, negative, best) ✗ (restaurant general, null, negative, sad) ✓	(food quality, chefs, negative, good) ✗	(laptop general, backlit, positive, awesome) ✗	(laptop general, trackpad, negative, annoying) ✗
w Image	(restaurant general, null, negative, sad) ✓	(food quality, fish, negative, decent) ✓	(keyboard design_features, backlit keyboard, positive, awesome) ✓	(keyboard general, trackpad, negative, annoying) ✓

Table 7: Cases studies in uni-modal Aspect-based Sentiment Analysis.

guide the model's attention to the correct aspect term and corresponding opinion term.

In the third example, the model without sentimental image wrongly classify the category into "*laptop general*" while in the image, there is a conspicuous backlit keyboard as the focused target and with the image, the model do the right classification.

In the fourth example, the model without sentimental image wrongly classify the category into "*laptop general*" while in the image, the "*trackpad*" of a laptop is printed as the focused target and with the image, the model do the right classification.

From the cases shown in Table 7, we can find that, with the extra semantic information from sentimental image, our method shows significant superiority in improving the performance of quadruple extraction.