Aspect-based Sentiment Analysis via Synthetic Image Generation

Ge Chen, Zhongqing Wang, and Guodong Zhou
Natural Language Processing Lab, Soochow University, Suzhou, China
20245227045@stu.suda.edu.cn, {wangzq, gdzhou}@suda.edu.cn

Abstract

Recent advancements in Aspect-Based Sentiment Analysis (ABSA) have shown promising results, yet the semantics derived solely from textual data remain limited. To overcome this challenge, we propose a novel approach by venturing into the unexplored territory of generating sentimental images. Our method introduce a synthetic image generation framework tailored to produce images that are highly congruent with both textual and sentimental information for aspect-based sentiment analysis. Specifically, we firstly develop a supervised image generation model to generate synthetic images with alignment to both text and sentiment information. Furthermore, we employ a visual refinement technique to substantially enhance the quality and pertinence of the generated images. After that, we propose a multimodal model to integrate both the original text and the synthetic images for aspect-based sentiment analysis. Extensive evaluations on multiple benchmark datasets demonstrate that our model significantly outperforms state-of-the-art methods. These results highlight the effectiveness of our supervised image generation approach in enhancing ABSA.

1 Introduction

Aspect-based sentiment analysis (ABSA) represents a fine-grained approach to text sentiment analysis, which pinpoints sentiment information pertinent to specific aspects and offers businesses and organizations deeper market insights. (Pontiki et al., 2014) ABSA is commonly divided into two independent subtasks: Aspect Term Extraction (ATE) and Aspect Sentiment Classification (ASC). ATE aims to extract all aspect terms from a given text, while ASC is concerned with determining the sentiment polarity of each aspect (Ju et al., 2021).

Previous studies have demonstrated significant advancements in ABSA by utilizing pre-trained encoder-decoder language models (Zhang et al., 2021a). These studies have adopted various strategies, treating the class index (Yan et al., 2021), natural language descriptions (Zhang et al., 2021a), or the desired sentiment element sequence (Zhang et al., 2021b; Bao et al., 2022; Gou et al., 2023) as the target for the generation model. Despite their effectiveness, most of these studies have been limited to raw input textual data, neglecting other additional data sources that could enrich and complement textual ABSA systems.

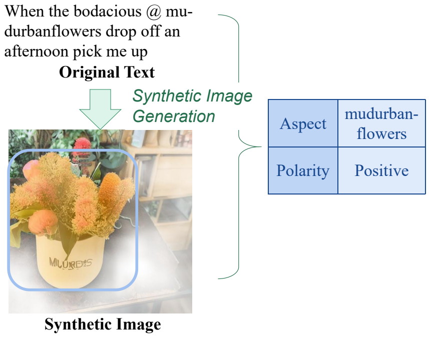
When the bodacious @ mudurbanflowers drop off an afternoon pick me up
Original Text
Synthetic Image Generation
Aspect
mudurban-flowers
Polarity
Positive
Synthetic Image

Figure 1: An example of the synthetic image generation for aspect-based sentiment analysis.

Therefore, an increasing number of methods have begun to incorporate image information to enhance the performance of the ABSA task (Ling et al., 2022; Zhou et al., 2023; Yang et al., 2024b). Integrating image information can offer several advantages. For one, visual elements can provide a more intuitive and vivid understanding of the text content, especially when dealing with complex sentiment expressions. Images can also capture subtle emotional cues that might be overlooked in text alone, thereby enriching the sentiment analysis process. However, for original texts, it is a difficult endeavor to collect and acquire corresponding image information. Furthermore, numerous original texts lack directly corresponding images. Thus, it is essential to generate images based on the original texts.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21818–21829
November 4-9, 2025 ©2025 Association for Computational Linguistics

Nevertheless, directly generating images that correspond to the original text presents multiple challenges. Firstly, the generated images must be relevant to the original text not only in a literal sense but also in terms of sentiment. Secondly, it is desirable for the generated images to incorporate additional information that can aid in sentiment analysis.

To tackle these challenges, we introduce *synthetic image generation framework* tailored to produce images that are highly congruent with both textual and sentimental information for aspect-based sentiment analysis, as shown in 1. Specifically, our framework is divided into two stages: In the first stage, we develop a supervised image generation model to generate synthetic images with alignment to both text and sentiment information. Moreover, we employ visual refinement techniques to enhance the synthetic images by highlighting regions relevant to sentiment information and masking irrelevant areas, thereby enabling the images to focus more closely on sentiment-related visual cues.

Subsequently, we propose a multi-modal model that combines the original text with the synthetic images for aspect-based sentiment analysis. Extensive experimental evaluations demonstrate that our model significantly outperforms existing state-of-the-art methods on multiple benchmark datasets, opening a new avenue for multimodal sentiment analysis and enhancing large language models' capabilities in visual–text understanding.

2 Related Works

In this section, we introduce two related topics of this study: aspect-based sentiment analysis, and multi-modal sentiment analysis.

2.1 Aspect-based Sentiment Analysis

Recent studies using pre-trained encoder-decoder language models show great improvements in ABSA (Zhang et al., 2021a). They either treated the class index (Yan et al., 2021), natural language (Zhang et al., 2021a), or the desired sentiment element sequence (Zhang et al., 2021b; Peper and Wang, 2022; Lee and Kim, 2023) as the target for the generation model.

With the development of pre-trained models, researchers are designing more complex architectures. Some works combine graph neural networks with semantic knowledge (Chen et al., 2024; Bao et al., 2023b; Song et al., 2024), while others replace syntactic trees with abstract meaning representations and enhance self-attention with semantic relations (Ma et al., 2023). In addition, reasoning chains based on syntax-opinion-sentiment (Bai et al., 2024) and soft prompt methods (Fan et al., 2025) have been proposed to better capture sentiment information and improve few-shot performance.

2.2 Multi-modal Sentiment Analysis

Multi-modal Aspect-Based Sentiment Analysis (MABSA) integrates multiple modalities to enhance sentiment analysis by leveraging complementary information (Yang et al., 2022; Ju et al., 2021). Existing studies combine text with images (Zhou et al., 2023; Ling et al., 2022), audio (Yao et al., 2021), or video (Zhang et al., 2023; Yang et al., 2023), enabling richer emotional understanding beyond single-modality features.

Among these, text-image fusion has attracted increasing attention for its broad applications. Researchers have improved sentiment fusion through attention-based matching (Zhao et al., 2022; Xiao et al., 2023), cross-modal contrastive learning (Yang et al., 2024b), fine-grained joint learning (Ju et al., 2021), dual encoders with alignment (Yu et al., 2022), and multi-granularity denoising (Zhao et al., 2023).

Different from previous studies, we pioneer the use of generated visual content to enhance textual ABSA, achieving superior extraction performance and expanding visual augmentation to pure-text scenarios.

3 Synthetic Image Generation

As illustrated in Figure 2, we introduce a synthetic image generation framework tailored to produce images that are highly congruent with both textual and sentiment information for aspect-based sentiment analysis.

Specifically, our framework is divided into two stages: In the first stage, we develop a *supervised image generation model* to generate synthetic images aligned with both text and sentiment. Furthermore, we employ a *visual refinement technique* to substantially enhance the quality and pertinence of the generated images, ensuring they more accurately reflect the aspects and sentiments conveyed in the original text.

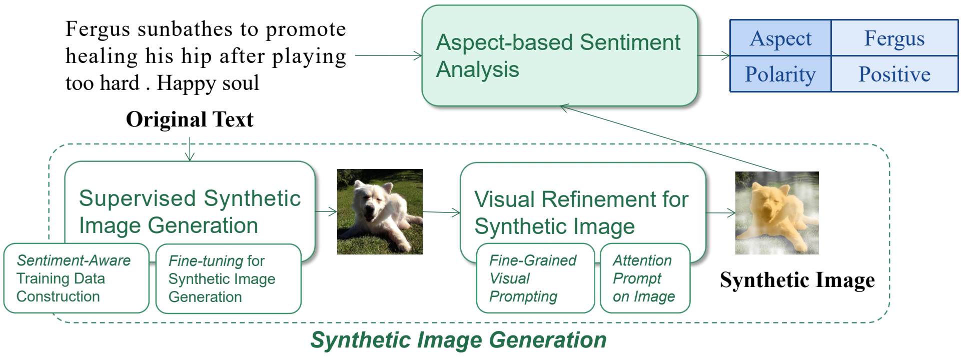
Fergus sunbathes to promote healing his hip after playing too hard . Happy soul
Original Text
Aspect-based Sentiment Analysis
Aspect
Fergus
Polarity
Positive
Supervised Synthetic Image Generation
Sentiment-Aware Training Data Construction
Fine-tuning for Synthetic Image Generation
Visual Refinement for Synthetic Image
Fine-Grained Visual Prompting
Attention Prompt on Image
Synthetic Image
Synthetic Image Generation

Figure 2: Overview of the proposed model.

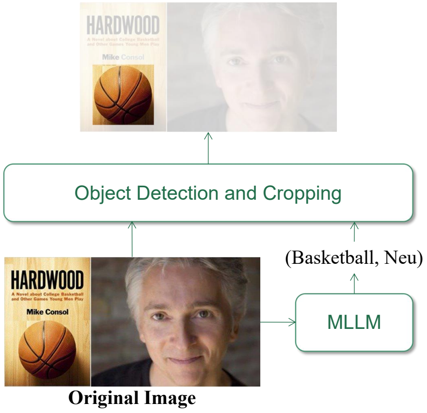

Figure 3: An example of sentiment-aware training data construction.

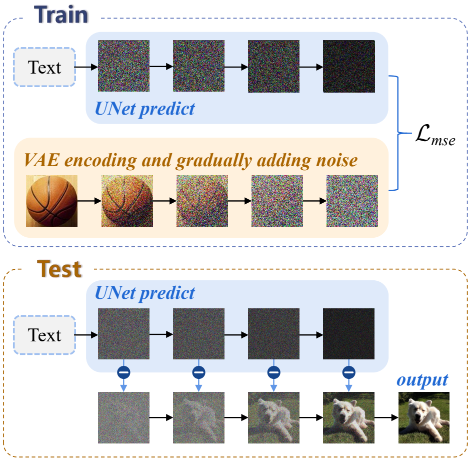

Figure 4: An example of supervised image generation.

3.1 Supervised Synthetic Image Generation

In this subsection, we introduce a novel supervised image generation framework that aims to generate synthetic images with alignment to both text and sentiment information. Specifically, our framework encompasses the construction of a high-quality training dataset and the fine-tuning of a pre-trained image generation model. The fine-tuned model produces images closely corresponding to the input text and sentiment.

Sentiment-Aware Training Data Construction

To effectively fine-tune the image generation model, it is essential to construct a high-quality sentiment-aware training dataset comprising text-image pairs. As illustrated in Figure 3, we begin by leveraging a multi-modal learning model[1] (Dong et al., 2024) to analyze raw images and texts, generating pseudo-labels that encapsulate various aspects and their corresponding sentiment polarities.

Subsequently, we input the raw images, along with their associated aspect terms, into an object detection model[2] (Minderer et al., 2023). This model is utilized to detect and crop the target region images (I_c) that are pertinent to the aspect terms. These cropped images are then paired with their corresponding texts, forming the training dataset necessary for fine-tuning the image generation model.

To further enhance the model's generalization capabilities, we augment the training data through techniques such as random cropping and rotation of the images. This augmentation helps to prevent overfitting and bolsters the model's robustness.

Fine-tuning for Synthetic Image Generation

Next, we fine-tune the Stable Diffusion model[3] (Rombach et al., 2022) to generate the synthetic images based on the above sentiment-

[1] https://huggingface.co/internlm/internlm-xcomposer2-vl-7b

[2] https://huggingface.co/google/owlv2-base-patch16-ensemble

[3] https://huggingface.co/CompVis/stable-diffusion-v1-4

aware training dataset. The fine-tuning process is shown in Figure 4.

In particular, we employ the Low-Rank Adaptation (LoRA) (Hu et al., 2021) method for parameter-efficient fine-tuning of the pre-trained Stable Diffusion model. The specific process is as follows: first, we load the pre-trained weights of the base model, including the text encoder (CLIP), the variational autoencoder (VAE), and the UNet diffusion model, while freezing all parameters of the base model. Next, we insert trainable LoRA layers only in the attention module of the UNet, setting the rank of the low-rank matrices to 4 and initializing them using a Gaussian distribution.

The training images x processed in the previous step are resized to a resolution of 512×512, followed by center cropping and random horizontal flipping for augmentation, and then compressed into latent variables using the VAE encoder:

$$z = \text{VAE}_{\text{enc}}(x) \tag{1}$$

The input text T is converted into an ID sequence using the CLIP tokenizer:

$$t = \text{CLIP}_{\text{tokenizer}}(T) \tag{2}$$

We use the AdamW (Loshchilov and Hutter, 2017) optimizer with an initial learning rate of 1e-5 and a batch size of 4. The loss is calculated using the mean squared error (MSE) of the noise prediction:

$$\mathcal{L}_{mse} = ||\hat{\mathcal{E}} - \mathcal{E}||_2^2 \tag{3}$$

where $\hat{\mathcal{E}}$ represents the predicted noise and $\mathcal{E}$ represents the true noise.

Through this fine-tuning process, we iteratively optimize the image generation model, enabling it to effectively capture the semantic relationships between text and images.

3.2 Visual Refinement for Synthetic Image

After generating the synthetic images, we introduce a visual refinement technique to significantly improve the quality and relevance of the generated images, making them better aligned with the aspects and opinions expressed in the original text.

As illustrated in Figure 5, we develop a visual cue module that integrates an advanced form of fine-grained visual prompting with attention heatmaps. This module is guided by textual queries, ensuring that the visual enhancements are tightly coupled with the textual content.

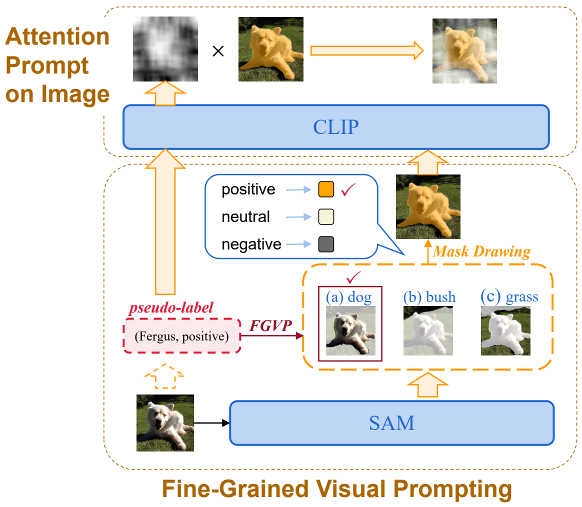

Figure 5: An example of visual refinement for synthetic image.

Fine-Grained Visual Prompting

We firstly introduce the fine-grained visual prompting module designed to improve the quality of generated synthetic images. This module employs color-based prompting techniques to differentiate entities associated with various sentiment polarity within the images.

Given the synthetic image I_s, we use the Segment Anything Model (SAM) (Kirillov et al., 2023) to generate semantic segmentation masks for the image:

$$M = \text{SAM}(I_s) \tag{4}$$

where $M \in \mathbb{R}^{N \times H \times W}$ represents the generated semantic masks, N is the number of masks, and H and W denote the height and width of the image, respectively. Unlike traditional box-based or circle-based prompting methods, semantic masks can precisely capture target contours, reduce background interference, and preserve global contextual information.

Then, we calculate the similarity between the pseudo-label query Q_{label} and the masked region based on the semantic mask M, and perform matching to generate the visual prompt map I_p:

$$I_p = \text{FGVP}(M, Q_{label}) \tag{5}$$

Meanwhile, masks are applied to the synthesized images based on the sentiment polarity in the pseudo-labels, with the drawing strategy detailed in Figure 5.

Attention Prompt on Image

To further improve the quality and relevance of the generated images, we integrate Attention Prompting on Image (API) technology. This approach enhances the model's understanding of visual information by leveraging pseudo-label-guided attention heatmaps. Utilizing the visual prompt map I_p and the pseudo-label text query Q_{label}, we generate attention heatmaps I_a that are pertinent to aspects and sentiments:

$$I_a = \text{CLIP}(I_p, Q_{label}) \tag{6}$$

Then, the attention heatmap I_a is multiplied by the pixel values of the input image I_p to generate the attention prompt image I_{API}:

$$I_{API} = I_p \odot I_a \tag{7}$$

where $\odot$ denotes multiplication of pixel values.

4 Aspect-based Sentiment Analysis with Synthetic Image

After we obtain the generated synthetic images with visual refinement, we propose a multi-modal aspect-based sentiment analysis model to integrate both original text and the synthetic images.

4.1 Image and Text Encoding

We firstly utilize CLIP (Radford et al., 2021) as the generated image encoder to acquire the visual representation. For a given input image I, it is initially resized to a fixed resolution and subsequently partitioned into $P \times P$ non-overlapping image patches. Each of these image patches is then transformed into the feature space via a linear embedding process, yielding a sequence of image patch representations:

$$x_v[i] = W \cdot \text{Flatten}(P[i]) + b \tag{8}$$

where x_v denotes the encoded image sequence, $W \in \mathbb{R}^{D\times(P^2 \cdot C)}$ represents the weight matrix, and $b \in \mathbb{R}^D$ is the bias vector.

Subsequently, we employ MLLM (Dong et al., 2024) as our text encoder as well as the modality fusioner. We specifically craft fusion instructions in natural language, which are intended to direct the visual language model in the fusion of visual inputs:

$$x_t[i] = \text{TextTokenizer}(token[i]) \tag{9}$$

$$x = [x_v, x_t] \tag{10}$$

where x_t stands for the encoded text sequence.

	Twitter-2015			Twitter-2017		
	Train	Dev	Test	Train	Dev	Test
Positive	928	303	317	1,508	515	493
Neutral	1,883	670	607	1,638	517	573
Negative	368	149	113	416	144	168

Table 1: Statistics of the two benchmark datasets.

4.2 Decoding

The enhanced image I_{API} and text T are encoded by $E_I(\cdot)$ and $E_T(\cdot)$, respectively. Their features undergo cross-modal interaction ($\oplus$) and are projected into a joint representation via an MLP:

$$h_{\text{fusion}} = \text{MLP}\big(E_I(I_{\text{API}}) \oplus E_T(T)\big) \tag{11}$$

The decoder predicts the probability distribution of the current word conditioned on the fused feature h_{fusion} and the previously generated sequence $y_{<i}$:

$$P(y_i \mid y_{<i}, h_{\text{fusion}}) = \text{Decoder}(y_{<i}, h_{\text{fusion}}) \tag{12}$$

The final hidden state is linearly mapped to predict the vocabulary distribution, completing the multi-modal sentiment analysis.

4.3 Training

To optimize the model's performance, we adopt the cross-entropy loss function as the main objective for the sentiment classification task, defined as follows:

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{S}\log P(y_t|y_{<t}, X) \tag{13}$$

where y_t denotes the t-th word in the target sequence, P is the conditional probability, S is the length of the output sequence, and N is the batch size.

5 Experiments

In this section, we introduce the datasets used for evaluation and the baseline methods employed for comparison. We then report the experimental results conducted from different perspectives, and analyze the effectiveness of the proposed model with different factors.

5.1 Settings

We utilize two publicly available Twitter datasets, Twitter-2015 and Twitter-2017, originally introduced by Yu and Jiang (2019). As shown in Table 1, a substantial proportion of sentences in both datasets contain multiple aspects, making them well-suited for our aspect-based analysis.

Our proposed model is based on InternLM-XComposer2-VL (Dong et al., 2024) and fine-tuned using LoRA. We optimize the adapter parameters via grid search on the validation dataset, setting the LoRA alpha to 8 and the LoRA rank to 64. The model parameters are updated using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 5e-5. We use a batch size of 1, a maximum token length of 4096, and an image size of 490×490. All experiments are conducted on an Nvidia RTX 4090 GPU.

We evaluate the performance of our model on the JMABSA task using Micro-F1 score (F1), Precision (P), and Recall (R). A sample is considered correctly predicted only if both the aspect term and its corresponding polarity match the ground truth.

5.2 Main Results

To evaluate the proposed model, we implement several aspect-based sentiment analysis baseline models for comparison, including SPAN (Hu et al., 2019), D-GCN (Chen et al., 2020), GAS (Zhang et al., 2021c), Parapharse (Zhang et al., 2021a), OTG (Bao et al., 2023a), FaiMA (Yang et al., 2024c), BART (Lewis et al., 2020), Deepseek-llm (Bi et al., 2024), Qwen (Yang et al., 2024a), LLaMA (AI@Meta, 2024), and ChatGPT (OpenAI, 2024).

According to Table 2, traditional methods (such as SPAN and D-GCN) typically rely on graph neural networks or sequence models to model local text structures, but their generalization ability is limited by the scale of the data and the design of the tasks. Improvement methods based on pre-train models such as OTG, Paraphrase, and FaiMA demonstrate stronger generalization capabilities compared to traditional approaches by incorporating attention mechanisms and semantic reconstruction strategies. But these methods still exhibit limitations in modeling long-range emotional dependencies. New-generation large language models(LLMs) like LLaMA-3 and GPT-4o, trained on large-scale and diverse data, possess stronger text understanding and generation capabilities compared to past pure text models. However, they still primarily rely on textual information and struggle to capture complex emotions across modalities.

Our proposed model achieves the best results and significantly outperforms all other models ($p < 0.05$). This indicates that our approach, through supervised image generation and visual refinement, can generate high quality synthetic images for aspect-based sentiment analysis. Additional results on the aspect sentiment triplet extraction task are reported in Appendix A.

5.3 Impact of Supervised Synthetic Image Generation

To investigate the impact of different synthetic image generation methods on the experimental results, we designed the following experiment. As shown in Table 3, the “Direct (Raw)” approach involves generating images directly from the raw, unprocessed textual input. In contrast, the “Direct (Prompt)” method utilizes prompts generated by a large language model(LLM) (OpenAI, 2024) to synthesize the corresponding images. For the supervised generation experiments, we employed three distinct image augmentation strategies: “Object” (Object Detection), where training images are cropped based on detected object regions; “Random” (Random Cropping), which applies stochastic cropping for data augmentation; and “Ours” a composite approach that integrates all augmentation strategies to enhance model robustness and generalization.

The results indicate that incorporating visual features via generated images significantly improves performance. However, the gap between using raw text and prompts generated by LLMs remains relatively small.

Therefore, we aim to further enhance the quality of synthesized images through a supervised generation strategy. Although the basic supervised approach underperforms compared to direct generation, the gradual introduction of object detection and random cropping leads to consistent performance gains, ultimately surpassing direct generation methods. The complete method achieves the best overall results, demonstrating the effectiveness of each component.

5.4 Influence of Visual Refinement Strategies

To assess the contribution of visual refinement strategies, we conduct a set of comparative exper-

	Twitter-2015			Twitter-2017		
Method	P	R	F1	P	R	F1
SPAN*	53.7	53.9	53.8	59.6	61.7	60.6
D-GCN*	58.3	58.8	59.4	64.2	64.1	64.1
FaiMA	59.4	62.8	61.1	67.6	61.9	64.6
GAS	66.5	65.4	66.0	63.9	64.4	64.2
Parapharse	66.5	66.2	66.3	65.2	65.8	65.5
OTG	67.1	65.1	66.1	66.8	67.5	67.2
BART*	62.9	65.0	63.9	65.2	65.6	65.4
Deepseek-llm	63.6	65.7	64.6	65.5	67.7	66.6
Qwen-2.5	64.2	66.4	65.2	66.8	69.0	67.9
Llama-3	66.2	68.5	67.3	68.8	68.9	68.8
GPT-4o	65.1	65.6	65.4	69.3	70.3	69.8
Ours	**68.1**	**71.9**	**70.0**	**72.9**	**74.2**	**73.5**

Table 2: Comparison with baselines. *denotes the results from Zhou et al. (2023). The best results are bold-typed.

Method	Twitter15	Twitter17
Text-only	67.2	67.2
with Synthetic Images		
Direct (*Raw*)	68.3	71.0
Direct (*Prompt*)	67.9	71.6
Supervised	66.7	71.7
+Object	67.6	71.7
+Random	68.8	72.1
Ours	**70.0**	**73.5**

Table 3: Impact of different synthetic image generation methods.

Methods	Twitter15	Twitter17
Text-only	67.2	67.8
with Images		
Original	67.9	71.0
+Visual	69.0	72.1
+Attention	69.5	72.4
Synthetic	68.8	72.1
+Visual	69.9	71.4
+Attention	**70.0**	**73.5**

Table 4: Results of visual refinement strategies.

iments. As shown in Table 4, “Original” refers to directly using original images, while “Synthetic” denotes using supervised synthetic images. “Visual” and “Attention” correspond to the fine-grained visual prompting and image attention prompting introduced in Section 3.2, respectively.

The experimental results demonstrate that both fine-grained visual prompting and image attention prompting strategies effectively enhance the utility of both original and synthetic images for aspect-based sentiment analysis. These findings suggest that visual refinement strategies are indeed valuable for capturing sentiment-relevant information, thereby facilitating the reconstruction of images to better support sentiment analysis tasks.

5.5 Visual Augmentation Comparison

We compare various image augmentation approaches, which fall into two categories: original visual augmentation (augmented images are user-posted), including JML (Ju et al., 2021), VLP-MABSA (Ling et al., 2022), AoM (Zhou et al., 2023), and CORSA (Liu et al., 2025); and generated visual augmentation (augmented images are synthesized), represented by SIG4ABSA (Bao et al., 2025).

As shown in Table 6, generated visual augmentation methods consistently outperform original ones, suggesting that synthesized images provide more explicit, sentiment-related visual cues that effectively complement textual inputs.

In particular, our method achieves the best performance on both datasets, surpassing not only all original augmentation methods but also the state-of-the-art generated method, SIG4ABSA. This demonstrates the effectiveness of our proposed approach, which is able to precisely locate key image regions and highlight sentiment-related information, thereby leading to superior multimodal senti-

Original Text	Synthetic Image	Results
Former @ PracticeFusion CEO Ryan Howard turns painful lessons into new #startup		**Golden**: (Ryan Howard, Pos) **Baseline**: (Ryan Howard, Neu) **Ours**: (Ryan Howard, Pos)
Mad that this guy used to be the world's best footballer!! #R9 #ronaldo		**Golden**: (R9, Neg), (ronaldo, Neg) **Baseline**: (R9, Neu), (ronaldo, Pos) **Ours**: (R9, Neg), (ronaldo, Neg)

Table 5: Examples of case study.

Augmentation Type	Method	Twitter15	Twitter17
Original Visual	JML	64.1	66.0
	VLP-MABSA	66.6	68.0
	AoM	68.6	67.7
	CORSA	69.9	70.6
Generated Visual	SIG4ABSA	67.8	69.0
	Ours	**70.0**	**73.5**

Table 6: Results of different visual augmentations.

Model	Twitter15	Twitter17
LLaVA		
Text-only	66.8	69.7
+Original Image	68.3	70.3
+Synthetic Image	69.0	71.5
Qwen		
Text-only	65.2	67.9
+Original Image	68.8	70.5
+Synthetic Image	69.7	72.0
InternLM		
Text-only	67.2	67.8
+Original Image	67.9	70.9
+Synthetic Image	70.0	73.5

Table 7: Results with different multi-modals models.

ment analysis results.

5.6 Results with Different Multi-Modals Models

In this section, we compare several multi-modals models under different input configurations, including LLaVA-1.5 (Liu et al., 2024), Qwen2.5 (Yang et al., 2024a), and InternLM-XComposer2 (Dong et al., 2024).

As shown in Table 7, all models benefit from visual inputs, with our generated synthetic images consistently yielding the best performance. This demonstrates the robustness and effectiveness of our synthetic image generation strategy in providing sentiment-relevant visual cues. Compared to original images, our synthesized images offer more focused and sentimental content, leading to more accurate sentiment predictions. These findings highlight the value of task-specific image synthesis in multi-modal sentiment analysis.

5.7 Case Study

As shown in Table 5, we present a case study comparing our method with a baseline that performs sentiment analysis based solely on text.

In the first example, when only text features were used, the model predicted “Ryan Howard” as neutral. This misclassification indicates that text alone may not provide sufficient cues to clearly distinguish sentiment. The phrase “painful lessons” may introduce ambiguity, leading the model to interpret the statement as neutral rather than positive. After incorporating the enhanced image modality, the model successfully predicted the correct label, demonstrating that our method's improved image processing reduces noise and enhances emotional information, thereby decreasing ambiguity in sentiment interpretation.

In the second example, the phrase “Mad that this guy used to be the worlds best footballer ! !# R9 # ronaldo” conveys a mix of admiration and frustration, which may confuse the model. Without additional context, the sentiment was misinterpreted. After incorporating the generated visual enhancement images, the ambiguity in the text was reduced, making the negative sentiment features more prominent.

6 Conclusion

In this study, we introduce a novel approach for generating sentimental images that serve as ancillary visual semantics to enhance textual aspect-based sentiment analysis. Our method revolves around the development of a supervised image generation model. We further employ a visual refinement technique to substantially enhance the quality and pertinence of the generated images. Through extensive evaluations on multiple benchmark datasets, we have demonstrated that our proposed model significantly outperforms state-of-the-art methods in ABSA.

Limitations

While our method demonstrates strong performance in enhancing multi-modal sentiment analysis, there are still some open challenges worth exploring. The current approach may have relatively high computational complexity, which could affect its efficiency when scaling to larger datasets. In addition, the method has mainly been evaluated on specific benchmarks. Testing it on more diverse datasets, such as Chinese or domain-specific corpora, could further verify its generalizability. Future work could also explore more lightweight architectures and refined strategies for cross-modal fusion to improve robustness and adaptability.

Acknowledgments

We would like to thank Prof. Zhongqing Wang for his helpful advice and discussion during this work. Also, we would like to express our gratitude to the anonymous reviewers for their insightful comments on this paper. This research was supported by the National Natural Science Foundation of China (No. 62376178), and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

AI@Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/. Accessed: 2024-07-20.

Yinhao Bai, Yalan Xie, Xiaoyi Liu, Yuhua Zhao, Zhixin Han, Mengting Hu, Hang Gao, and Renhong Cheng. 2024. Bvsp: Broad-view soft prompting for few-shot aspect sentiment quad prediction. *arXiv preprint arXiv:2406.07365*.

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

Xiaoyi Bao, 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*, pages 7971–7984, Toronto, Canada. Association for Computational Linguistics.

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.

Xiaoyi Bao, Wang Zhongqing, 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-22*, pages 4044–4050. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Guimin Chen, Yuanhe Tian, and Yan Song. 2020. Joint aspect extraction and sentiment analysis with directional graph convolutional networks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 272–279, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Junjie Chen, Hao Fan, and Wencong Wang. 2024. Syntactic and semantic aware graph convolutional network for aspect-based sentiment analysis. *IEEE Access*, 12:22500–22509.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Rui Fan, Shu Li, Tingting He, and Yu Liu. 2025. Aspect-based sentiment analysis with syntax-opinion-sentiment reasoning chain. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3123–3137.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and

Weizhu Chen. 2021. Lora: Low-Rank Adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. *arXiv preprint arXiv:1906.03820*.

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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.

Sun-Kyung Lee and Jong-Hwan Kim. 2023. Sener: Sentiment element named entity recognition for aspect-based sentiment analysis. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. 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, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Fukun Ma, Xuming Hu, Aiwei Liu, Yawen Yang, Shu'ang Li, Philip S Yu, and Lijie Wen. 2023. Amr-based network for aspect-based sentiment analysis. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 322–337.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. 2023. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36:72983–73007.

OpenAI. 2024. Hello gpt-4o. `https://openai.com/blog/gpt-4o`. Accessed: 2024-07-20.

Joseph Peper and Lu Wang. 2022. Generative aspect-based sentiment analysis with contrastive learning and expressive structure. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6089–6095, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. `https://huggingface.co/CompVis/stable-diffusion-v1-4`. Accessed May 2025.

Xiangxiang Song, Guang Ling, Wenhui Tu, and Yu Chen. 2024. Knowledge-guided heterogeneous graph convolutional network for aspect-based sentiment analysis. *Electronics*, 13(3).

Luwei Xiao, Xingjiao Wu, Shuwen Yang, Junjie Xu, Jie Zhou, and Liang He. 2023. Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis. *Information Processing & Management*, 60(6):103508.

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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Hao Yang, Yanyan Zhao, and Bing Qin. 2022. Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3324–3335, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. ConFEDE: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7617–7630, Toronto, Canada. Association for Computational Linguistics.

Juan Yang, Mengya Xu, Yali Xiao, and Xu Du. 2024b. Amifn: Aspect-guided multi-view interactions and fusion network for multimodal aspect-based sentiment analysis. *Neurocomputing*, 573:127222.

Songhua Yang, Xinke Jiang, Hanjie Zhao, Wenxuan Zeng, Hongde Liu, and Yuxiang Jia. 2024c. FaiMA: Feature-aware in-context learning for multi-domain aspect-based sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7089–7100, Torino, Italia. ELRA and ICCL.

Yiqun Yao, Michalis Papakostas, Mihai Burzo, Mohamed Abouelenien, and Rada Mihalcea. 2021. MUSER: MUltimodal stress detection using emotion recognition as an auxiliary task. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2714–2725, Online. Association for Computational Linguistics.

Jianfei Yu and Jing Jiang. 2019. Adapting bert for target-oriented multimodal sentiment classification. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*.

Zhewen Yu, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2022. Dual-encoder transformers with cross-modal alignment for multimodal aspect-based sentiment analysis. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 414–423, Online only. Association for Computational Linguistics.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023. 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.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021c. Towards generative aspect-based sentiment analysis. Association for Computational Linguistics.

Fei Zhao, Chunhui Li, Zhen Wu, Yawen Ouyang, Jianbing Zhang, and Xinyu Dai. 2023. M2DF: Multi-grained multi-curriculum denoising framework for multimodal aspect-based sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9057–9070, Singapore. Association for Computational Linguistics.

Fei Zhao, Zhen Wu, Siyu Long, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2022. Learning from adjective-noun pairs: A knowledge-enhanced framework for target-oriented multimodal sentiment classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6784–6794, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ru Zhou, Wenya Guo, Xumeng Liu, Shenglong Yu, Ying Zhang, and Xiaojie Yuan. 2023. AoM: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8184–8196, Toronto, Canada. Association for Computational Linguistics.

A Results on Aspect Sentiment Triplet Extraction Task

To further demonstrate the effectiveness of our proposed method in improving ABSA, we evaluate its performance on the ASTE (Aspect Senti-

Dataset	Precision	Recall	F1 Score
Laptop14			
Text-only	66.07	61.55	63.73
Ours	68.04	64.14	**66.03**
Rest14			
Text-only	73.96	74.85	74.40
Ours	74.63	76.06	**75.34**
Rest15			
Text-only	65.35	68.04	66.67
Ours	68.60	72.99	**70.73**
Rest16			
Text-only	70.09	77.04	73.40
Ours	73.19	78.60	**75.80**

Table 8: Results on aspect sentiment triplet extraction task.

ment Triplet Extraction) task, which requires extracting sentiment triplets in the form of (aspect, opinion, polarity). We compare our method with a text-only baseline under the same experimental settings.

We conduct experiments on four widely-used benchmark datasets from the SemEval series: Laptop14, Rest14, Rest15, and Rest16, following the data splits and annotations provided in previous works (Zhang et al., 2021c). The experimental settings are consistent with those used in the main experiments: we use InternLM-XComposer2-VL (Dong et al., 2024) as the backbone, fine-tuned using LoRA. The model is optimized with Adam (Kingma and Ba, 2014) using a learning rate of 5e-5, a batch size of 1, and a maximum token length of 4096. The image size is set to 490 × 490.

As shown in Table 8, our method consistently improves the F1 score across all datasets, with particularly significant gains observed on the Rest15 and Laptop14 datasets. This indicates that integrating synthesized images enhances the model's understanding of aspect-opinion context, even in fine-grained sentiment triplet extraction scenarios.