

Multimodal Customized Review Generation

Anonymous ACL submission

Abstract

In this study, we introduce a new task called *customized review generation*. This task aims to generate a personalized review that a specific user would give to a product that they have not yet reviewed. This can help users write high-quality reviews for products they have not previously reviewed, providing them with valuable insights. Additionally, customized reviews can offer a tailored summary of all reviews for a product, catering to the individual preferences of the reader. To achieve this goal, we explore the use of multimodal information for customized review generation. Specifically, we utilize a *multimodal pre-trained language model* that takes a picture of a product and a set of words as input and generates a customized review using both visual and textual information. Our experimental results demonstrate the effectiveness of the proposed model in generating customized reviews that are often of high quality.

1 Introduction

Review websites have gained immense popularity as they provide a platform for customers to voice their opinions and rate products or services based on their personal experiences. These websites have become a valuable source of information for potential customers who are looking for unbiased and honest feedback before making a purchase decision. By providing both an overall rating score and detailed user reviews, these websites offer a comprehensive view of a product or service, allowing customers to make informed decisions.

Sentiment analysis and recommendation systems are two of the most important research areas for analyzing reviews and rating scores on these review websites. Sentiment analysis aims to extract aspect and opinion terms from review text, assign a unique predefined category for each aspect, and give a semantic orientation (e.g., positive, negative, or neutral) toward the aspect (Qiu et al., 2011;

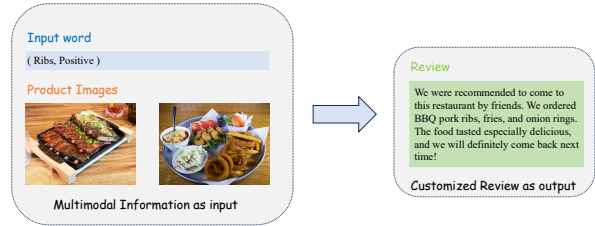


Figure 1: Example of multimodal review generation.

Chen and Qian, 2020; Bao et al., 2022). The recommendation system generates a ranking score for a specific product or service based on the user’s purchase history and other customers who have purchased the target product (Sarwar et al., 2001; He et al., 2020; Tian et al., 2022). While these studies have achieved significant success, they primarily analyze existing reviews or numerical purchase history. However, they cannot generate a *customized review* that a user would have written for a specific product if they had not reviewed it before. Such a customized review can provide a concise summary of all reviews tailored to the individual reader’s preferences.

In this study, we introduce a new task called *customized review generation*. The goal of this task is to generate a customized review for an unreviewed product tailored to the user’s preferences. This task can assist users in writing high-quality reviews by providing them with a starting point that is tailored to their preferences. Additionally, by analyzing the customized review, users can gain insights into their interests and preferences for a particular product.

A straightforward way to generate customized reviews is by using a picture of the product. However, pictures alone cannot fully describe the product details or effectively convey opinions about the product. Therefore, we integrate the picture with a set of words as input to generate a more comprehensive and customized review. These words include

aspect and opinion terms related to the product, which help to describe the product in detail and reflect the opinions towards it. This visual and textual information integration allows for a more complete and nuanced review that accurately captures the product’s features and the user’s opinions.

Therefore, we utilize a *multimodal pre-trained language model* that takes both a picture of a product and a set of words as input by leveraging visual and textual information. Furthermore, we generate a caption for the picture to bridge the gap between text and image, and we employ a text-guided fusion module to effectively fuse the information from multiple modalities. Finally, we generate the customized review based on the fused representation output by the modality fusion module.

Our experimental results demonstrate the importance of this new task and show that our proposed model outperforms existing competitive models, achieving state-of-the-art results. Overall, our findings suggest that multimodal pre-trained language models can effectively generate customized reviews by combining visual and textual information, paving the way for future research in this area.

2 Related Work

In this section, we introduce three related topics of this study: sentiment analysis, recommendation systems and multimodal fusion.

2.1 Sentiment Analysis

Early research on sentiment analysis primarily focused on document-level sentiment classification (Pang et al., 2002; Yu and Hatzivassiloglou, 2003; Yang et al., 2016; Nguyen and Le Nguyen, 2018).

Recently, Aspect-based sentiment analysis (ABSA) has obtained much more attention. The progression of ABSA research typically begins with tackling individual sub-tasks such as Aspect Term Extraction (Tulkens and van Cranenburgh, 2020), Aspect Category Detection (Shi et al., 2021) and Aspect Sentiment Classification (Wu and Ong, 2021). Then, some work start to consider more complex combinations, such as extracting both aspect and opinion terms (Gao et al., 2021; Li et al., 2022b), as well as detecting the specific aspect category and its corresponding sentiment polarity simultaneously (Cai et al., 2020; Bu et al., 2021).

More recently, end-to-end models have also been employed to extract sentiment elements in triplet

or quadruple formats (Zhao et al., 2022; Gou et al., 2023) and achieved impressive performance in multiple sentiment element extraction tasks.

2.2 Recommendation System

Recommendation system is a widely applied task aiming to provide customized suggestions to users based on their preferences and historical behavior. Early research primarily focused on collaborative filtering methods (Sarwar et al., 2001; Wu et al., 2016; Choi et al., 2023).

Due to limitations imposed by data sparsity on recommendation performance, some work consider leveraging historical reviews to alleviate the aforementioned problem. These studies can be categorized into two approaches: historical reviews method (Sun et al., 2021; Shuai et al., 2022), which utilize reviews to better learn embeddings of users and items and target reviews method (Ni and McAuley, 2018; Li and Tuzhilin, 2019; Sun et al., 2020; Xi et al., 2021), which uses reviews to model interactions between users and items more effectively. The classical idea behind this method is learning user-item interactions during the training stage and utilizing a transformer layer during the inference stage to approximate target reviews.

2.3 Multimodal Fusion

Multimodal fusion aims to leverage information from different modalities to enhance the performance of the model (Atrey et al., 2010; Bramon et al., 2011). In multimodal sentiment analysis scenarios, Zadeh et al. (2017) proposed a novel fusion model to model intra-modality and inter-modality dynamics. With the popularity of Transformers, Tsai et al. (2019) and Huang et al. (2020) introduced the multimodal Transformer to alleviate the problem of data misalignment and long-range dependencies. Recently, Yang et al. (2023) was not satisfied with simply concatenating modal features, but treated them differently depending on the modal contribution to fully exploit the modal interaction.

Our proposed task differs significantly from several similar tasks in terms of input and output. For example, in *sentiment analysis*, the input is usually existing reviews and the output is aspect terms or sentiment words. In *recommender systems*, the inputs are usually user or item IDs and user preferences and item attributes derived from historical reviews. The output is the degree of recommendation and the corresponding recommendation reason.

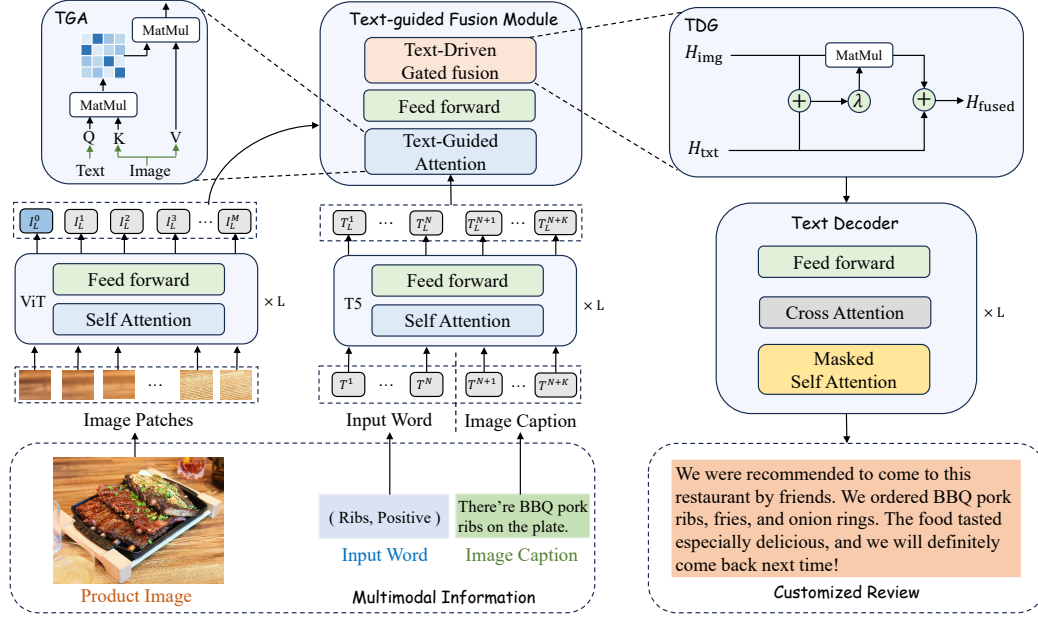


Figure 2: Overview of proposed model.

In our task, the inputs are keywords and product images, and the output is a customized review based on the above information.

Therefore, the key novelty of your task lies in the focus on generating reviews for unreviewed products, using only keywords and product images as input. In addition, this approach differs from *personalized review generation*, which typically relies on historical data or user profiles to craft personalized reviews.

3 Multimodal Review Generation

In this study, we introduce a new task called *Multimodal Customized Review Generation*. This task involves generating a customized review for a product based on relevant images and input words. Formally, the input to our task is a tuple $\{I, T\}$, where I represents the product images and T represents the input words describing the product or user’s opinion. The output generated by our model is a customized review C that provides a detailed and customized description of the product. Our proposed task is challenging as the review must be customized to the specific product and user preferences, making it a complex task that requires a deep understanding of language and visual information.

As shown in Figure 2, we propose a novel framework based on a *Multimodal Pre-trained Language Model* to tackle the above challenges. The proposed framework commences by utilizing a *Text Encoder* to transform the input text into a rich textual feature

representation. Concurrently, an *Image Encoder* is employed to encode the corresponding picture into a distinct visual feature representation. To establish a seamless connection between the text and image modalities, we introduce a *Image Caption Generation* component that produces a descriptive and contextually relevant caption for the picture. Subsequently, a *Text-guided Fusion* module is leveraged to integrate the textual and visual feature representations. Finally, the fused representation is utilized to generate the customized review.

3.1 Text Encoder

The input to the text encoder is a piece of text T , which consists of aspect terms and opinion terms provided by the user. These aspect terms describe specific attributes or features of the product, while the sentiment polarities indicate the user’s opinion or feeling towards each aspect.

To process the input text, we first tokenize the words into individual tokens and create an input sequence X of these tokens. To incorporate positional information into the input sequence, we add positional encodings to the input sequence pointwisely. Then, we feed the input feature into the encoder. The encoder consists of stacking L identical layers, each composed of a Multi-head Self-Attention (MSA) sub-layer and a feed-forward network (FFN) sub-layer,

$$T_\ell = \text{FFN}(\text{MSA}(T_{\ell-1})), T_\ell \in \mathbb{R}^{N \times D} \quad (1)$$

where T_ℓ is the hidden state of the ℓ -th encoder

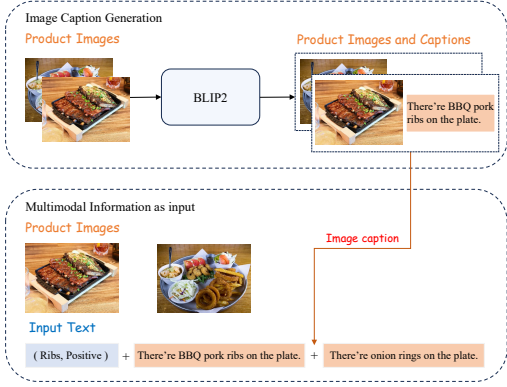


Figure 3: The process of image caption generation.

layer. In order to simplify, we omit the layer normalization operation in the formula.

3.2 Image Encoder

We employ a pre-trained Vision Transformer (ViT) model to learn the image representation from the product picture. ViT is a state-of-the-art model for image representation learning that uses a transformer architecture to process images (Dosovitskiy et al., 2020).

We first split the image into a sequence of m flattened 2D patches. These patches are small, fixed-size regions of the image that are used as input to the transformer model. Next, we add a special token $[CLS]$ at the beginning of the sequence to form an input sequence X . Finally, we feed the input feature into the pre-trained ViT model and obtain the visual representation H_{img} ,

$$I_\ell = \text{MLP}(\text{MSA}(I_{\ell-1})), I_\ell \in \mathbb{R}^{(M+1) \times D} \quad (2)$$

$$H_{img} = I_L^0 \quad (3)$$

where I_ℓ is the hidden state of the ℓ -th encoder layer and I_L^0 denotes the output of $[CLS]$ token from the L -th encoder layer.

3.3 Image Caption Generation

The existing multimodal work has shown that the addition of visual modality may not necessarily have a significant improvement or even a negative impact compared to typical unimodal tasks. (Zhu et al., 2018) argues that there is noise present in the images, and the input text already contains sufficient information to generate the target text. We conjecture that the current methods do not fully harness the potential of the visual modality to provide meaningful information.

In this study, we propose using *image captions* as a bridge between images and texts to enhance multimodal fusion. By generating the natural language caption of images, we can effectively utilize visual information and facilitate cross-modal interaction between the visual and textual domains.

As shown in figure 3, we employ the BLIP2 model (Li et al., 2023) to generate image captions. We fine-tune it on the COCO dataset (Lin et al., 2014) for image captioning tasks. Specifically, given multimodal data $\{I, T\}$, we generate a corresponding caption for image I using the BLIP2 model. We then concatenate the image caption with the original text T , resulting in a final input text format of “ \langle aspect terms \rangle ; \langle sentiment polarities \rangle ; \langle image captions \rangle ”.

3.4 Text-guided Fusion

The Text-guided Fusion module is designed to perform multimodal fusion by effectively combining textual representation and visual representation. As shown in Figure 2, this module consists of several sub-layers that work together to achieve the desired fusion.

The *Text-Guided Attention* (TGA) sub-layer is responsible for learning different attention scores for each object in the image based on the aspect terms and sentiment polarities in the query. By computing attention scores specific to the given aspect terms, the model can focus on relevant parts of the image and effectively integrate visual information related to the aspect terms.

$$H_{\text{txt}} = \text{FFN}(\text{TGA}(H_{\text{txt}}, H_{\text{img}}, H_{\text{img}})) \quad (4)$$

where $H_{\text{txt}} (= T_L)$ denotes textual representation.

The *Text-Driven Gated* (TDG) sub-layer controls how much visual information is preserved through the gate λ . This gate is learned during training and allows the model to filter out redundancy and noise from the visual information, ensuring that only relevant visual features are incorporated into the final fused representation.

$$\lambda = \text{Sigmoid}(W^T H_{\text{txt}} + W^I H_{\text{img}}) \quad (5)$$

$$H_{\text{fused}} = H_{\text{txt}} + \lambda \cdot H_{\text{img}} \quad (6)$$

where W^T and W^I are trainable parameters.

By combining the TGA and TDG sub-layers, the Text-guided Fusion Module effectively performs multimodal fusion, enabling the model to generate informative and expressive reviews that effectively integrate both visual and textual features.

3.5 Customied Review Generation

After performing multimodal fusion on the encoded features of the image and text, the decoder predicts the output sequence token-by-token with the multimodal fused representation.

The generated output sequence ends with the end token " $\langle /s \rangle$ ". The conditional probability of the whole output sequence $p(y|I, T)$ is progressively combined by the probability of each step $p(y_t|y_{<t}, I, T; \theta)$:

$$p(y|I, T) = \prod_{t=1}^{|y|} p(y_t|y_{<t}, I, T; \theta) \quad (7)$$

$$p(y_t|y_{<t}, I, T; \theta) = \sigma(W^o O_{L,t} + b^o) \quad (8)$$

where $O_{L,t}$ is the hidden state of the L -th decoder layer at the t -th decoding step, $\{W^o, b^o\}$ are trainable parameters, $\sigma(\cdot)$ is a softmax function, $y_{<t} = y_1 \dots y_{t-1}$ and $p(y_t|y_{<t}, I, T; \theta)$ are the probabilities over target vocabulary V normalized by softmax.

4 Experiment

In this section, we conduct extensive experiments to evaluate the performance of our proposed model. Additionally, we provide various analyses and discussions to demonstrate the effectiveness of our model.

4.1 Data and Setting

In this study, we construct a novel *multimodal dataset* for customized review generation, derived from the GEST dataset (Yan et al., 2023). We choose GEST-s2 from GEST, in which the reviews and images are aligned and therefore of higher quality compared to GEST-s1. Each review text in our dataset is associated with at least one corresponding image. We further adopt a sentiment analysis model (Bao et al., 2022) to extract sentiment elements and utilize these as input words from the review text. The detailed statistics of the dataset are presented in Table 1. For our experimental setup, we randomly selected 4,000 samples as training data, 500 samples as development data, and the remainder as test data.

We use T5-Base¹ (Raffel et al., 2020) and ViT² (Dosovitskiy et al., 2020) to initialize the model for our two tasks. We use adam (Kingma and Ba, 2014)

¹<https://huggingface.co/t5-base>

²<https://huggingface.co/google/vit-base-patch16-224>

Category	Account
Samples	5000
Avg. product images per sample	2.29
Avg. input words per sample	4.99
Avg. customized review length	47.62

Table 1: Statistics of the dataset.

as our optimizer to finetune hyper-parameters with a momentum of $\beta = 0.1$. We set the model learning rate as $1e-4$ and the batch size as 4. To generate higher quality reviews, we use beam search with a beam size of 5 and refrain from repeating n-grams of size 3 (Paulus et al., 2017). Our experiments are carried out with an NVIDIA Tesla V100 16G GPU.

We employ ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and METEOR (Denkowski and Lavie, 2014) as evaluation metrics to analyze the quality of reviews generated by the model and the effectiveness of our proposed model.

4.2 Main Results

In this study, we compare the proposed model with three kinds of models, including *text-only* models, *image-only* models, and *multimodal* models.

In text-only models, **Opword** directly employs the given input word as the foundation for producing personalized reviews. **BART** (Lewis et al., 2020) and **T5**, both pre-trained models, are designed to excel in text generation. **LLaMA** (Touvron et al., 2023), a large language model, is finetuned with Alpaca-LoRA, allowing us to achieve comparable performance to full-parameter training while using only a fraction of the parameters. Lastly, **ChatGPT** (Ouyang et al., 2022) is a conversational model developed by OpenAI.

Furthermore, **ViT-GPT2** features an image encoder and a text decoder, enabling the generation of natural language based on given images. **GIT** (Wang et al., 2022) is a generative Image-to-text transformer that unifies vision-language tasks such as image captioning and question answering. **Pix2Struct** (Lee et al., 2023) is a pre-trained image-to-text model designed for purely visual language understanding, suitable for finetuning on tasks containing visually-situated language.

In multimodal models, **BLIP2** (Li et al., 2023) is a large multimodal model capable of leveraging pre-trained frozen image encoders and large-scale language models to guide visual language pre-training. **Selective Attention** (Li et al., 2022a)

Method	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-4	METEOR
Text-Only						
OpWord	7.81	3.97	7.81	0.03	0.03	2.06
T5-ImgCap	28.32	4.04	18.17	23.45	10.47	15.12
T5	29.34	6.85	18.78	24.76	12.16	16.95
BART	29.18	6.59	19.09	24.83	12.23	16.74
LLaMA	26.97	6.48	19.02	21.81	11.28	15.02
ChatGPT*	27.57	6.02	18.23	22.70	11.40	15.63
Image-Only						
ViT-GPT2	23.84	3.20	15.41	16.86	9.57	15.98
GIT	24.33	3.15	16.60	21.05	10.15	14.11
Pix2Struct	25.31	2.38	16.14	21.78	10.07	14.01
Multimodal						
BLIP2	30.25	7.58	20.61	21.02	9.50	15.25
Selective Attn	30.23	6.79	19.00	25.23	12.14	17.44
VLP-MABSA	29.81	8.76	22.73	23.25	11.94	17.40
AoM	30.40	9.14	23.71	21.70	10.67	16.79
Ours	31.76	7.65	19.99	26.9	12.73	18.57

Table 2: Comparison with baselines. The model with "-ImgCap" uses image captions as input to incorporate visual information. * denotes we finetune the gpt-3.5-turbo-1106 on our dataset, through the OpenAI Finetune API, with the prompt "Please write a customized review for the user based on the prompted words entered!", which is determined empirically to perform well, and the temperature is set to 0.7 in the stage of generation.

is a method that utilizes a mechanism of selective attention to enhance the contribution of textual and visual features to the model’s performance. Both **VLP-MABSA** (Ling et al., 2022) and **AoM** (Zhou et al., 2023) are unified frameworks based on the BART model for realizing MABSA. We modify their models to complete generative tasks.

As shown in Table 2, input words cannot directly serve as customized reviews because they are too simplistic and contain only limited information. The performance of the pre-trained language models is generally higher than that of the image-only approach, which suggests that text can usually provide more detailed and direct information than images, which usually contain noise and invalid information.

Furthermore, we have observed that the performance of ChatGPT is relatively lower than that of T5 and BART. One plausible explanation for this could be that ChatGPT is less controllable even after fine-tuning. Additionally, we have noticed that the multimodal methods consistently outperform all unimodal methods. This observation implies that information derived from multiple modalities can effectively complement each other, thereby enriching the overall representation and leading to improved performance.

Method	FL	Info	HL	SR
T5	3.78	3.90	3.94	3.45
ChatGPT	4.28	3.67	4.17	4.03
LLaMA	4.02	3.54	3.92	3.89
Pix2Struct	3.02	2.17	3.21	2.50
BLIP2	4.18	4.26	4.02	4.14
Ours	4.68	4.76	4.49	4.55

Table 3: Results of human evaluation. **FL** denotes Fluency; **Info** denotes Informativeness; **HL** denotes Humanlike; **SR** denotes Sentiment Relevance.

Besides, our proposed model significantly outperforms all baseline models ($p < 0.05$), which underscores the efficacy of our proposed multimodal review generation framework. This framework integrates image caption generation and text-guided fusion, thereby demonstrating its effectiveness in leveraging multimodal information for enhanced performance.

4.3 Human Evaluation

We conduct the human evaluation for our proposed model and baseline models from four perspectives: *Fluency (FL)* is used to assess grammatical accuracy, expression fluency, and language readability; *Informativeness (Info)* is used to evaluate the over-

Method	R-1	B-1	METEOR
Text-Only	29.34	24.76	16.95
Image-Only	26.59	22.24	14.24
Add	29.81	24.94	16.96
Concat	29.76	25.23	17.14
Gate	30.33	25.32	17.28
Attention	30.74	26.18	17.83
Ours	31.76	26.90	18.57

Table 4: Impact of multimodal fusion strategies.

lap with key information in the Reference; *Human-like (HL)* is used to assess the degree of having a human-like language style; *Sentiment Relevance (SR)* is used to evaluate the relevance of sentiment expressed in the reference. We randomly select 300 examples from the test set and ask human annotators to evaluate the generated customized reviews, with evaluation scores ranging from 1 to 5 for each aspect.

As presented in Table 3, it is apparent that the customized reviews generated by the proposed multimodal model consistently outperform the unimodal model in all evaluated aspects. Furthermore, the *sentiment relevance metrics* specifically indicate that our proposed model excels in generating sentiment accuracy within the customized reviews. In conclusion, our proposed model surpasses the baseline model in multiple aspects and demonstrates its capability to generate reviews that are not only richer in content but also more precise in sentiment expression, thereby enhancing the overall quality of the customized reviews.

5 Analysis and Discussion

In this section, we give some analysis and discussion to show the effectiveness of the proposed multimodal model for customized review generation.

5.1 Impact of Multimodal Fusion Strategies

We first analyze the effect of different multimodal fusion strategies for generating the customized review.

As shown in Table 4, our analysis reveals that the performance of naive approaches, such as simply *adding* or *concatenating* the representation vectors of text and images, is inferior to more sophisticated methods. This observation implies that basic fusion techniques may not adequately capture the complex interactions and relationships between textual and visual information.

Input	R-1	B-1	METEOR
Image	26.45	22.86	15.16
+ Aspect	30.73	26.15	17.79
+ Polarity	27.35	22.95	14.83
+ Caption	28.39	23.86	15.44
Ours	31.76	26.90	18.57

Table 5: Effect of text input prompts. R-1 and B-1 are the abbreviations of ROUGE-1 and BLEU-1, respectively.

Moreover, our findings indicate that more advanced fusion strategies, such as *Gate* and *Attention*, exhibit superior effectiveness compared to simpler methods. These strategies demonstrate their ability to better model and leverage the interdependencies between the different modalities.

Most importantly, our proposed multimodal fusion strategy achieves the highest performance among all the compared methods. This achievement underscores the importance of employing an effective combination of gate and attention mechanisms to comprehensively harness and integrate the information and relationships present within text and image data.

5.2 Effect of Text Input

In Table 5, we delve into the impact of various input prompt strategies on the performance of review generation. To conduct the experiments, we categorize them into two main groups: those relying on images alone and those utilizing a combination of images with three distinct types of input texts (i.e., aspect term, opinion term, and image caption).

The results of our analysis reveal that solely relying on images to generate customized reviews is inadequate. This is because images alone lack the necessary semantic richness and contextual information required to generate meaningful and targeted reviews. Furthermore, our experimental results demonstrate that models that leverage input prompts in conjunction with images consistently yield higher performance compared to those that only use images. This underscores the crucial role of input strategies in guiding the review generation process and enhancing the overall quality and relevance of the output customized review. A more detailed discussion can be found in Appendix A.

5.3 Influence of Number of Images

We examine the influence of the number of images on the performance of review generation in


Input word	(Burger , Positive)	Product image	
Reference	<i>Partly disappointed, the burger was really good, I ordered #8 which is the bbq bacon with onion rings, this onion rings were not even rings, they were fried onions and you could not use the sauces with these onions. And the burger as I said was really good but not that big for the cost.</i>		
T5	<i>The burger was good, but a little dry. The fries were a bit salty. I ordered a burger with a side of fries and it was ok. If you are looking for a good burger, this is the place for you! I highly recommend this burger.</i>		
BLIP2	<i>I had the burger with bacon and onion rings. The brioche bun was a little too thick for my taste, but the fried onion rings were delicious. The bacon was melted perfectly.</i>		
Ours	<i>I got the burger and it was delicious. The onion rings were a bit dry but the BBQ burger was ok. I'm not a big fan of burgers, but if you're looking for a good burger, this is a great place to get a burger.</i>		

Table 6: Example of the case study. The content related to key information in the reference is highlighted in color (e.g., **Burger**), while the remaining content is displayed in italics..

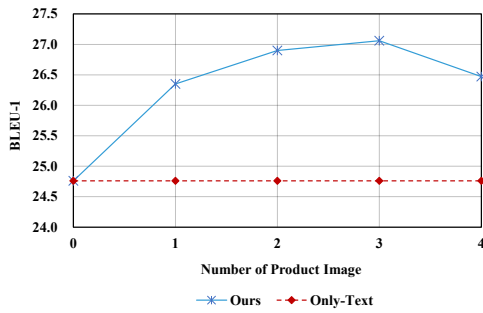


Figure 4: Influence of the number of images.

Figure 4.

As shown in the figure, it is evident that the method relying only on text as input yields the lowest performance ($number = 0$), highlighting the insufficiency of text-only models in generating satisfactory reviews.

Moreover, as the number of images increases, there is a noticeable upward trend in the performance. This observation underscores the beneficial role of images in enhancing the quality and relevance of customized reviews. The improved performance can be attributed to the richer contextual information provided by images, which aids in generating more specific and tailored reviews.

In conclusion, the number of images plays a pivotal role in determining the quality of the generated customized reviews. Notably, an optimal image size can significantly enhance the overall quality and effectiveness of the generated reviews.

5.4 Case Study

We give an example from the test data to compare the quality of customized reviews generated by our

proposed model and other baseline models.

As demonstrated in Table 6, it is evident that the reviews generated by the unimodal pre-trained model exhibit limitations in the amount of information provided. In contrast, both BLIP2 and our proposed model effectively convey the key information in the References comprehensively. Notably, our proposed model distinguishes itself by surpassing the others in terms of sentiment expression and human-like qualities of the reviews. An exemplar from our model mentions that the onion rings are slightly dry, which closely aligns with the viewpoints expressed in the reference. Conversely, BLIP2 describes the onion rings as delicious, which diverges from the sentiment expressed in the reference. In summary, the customized reviews generated by our proposed model demonstrate exceptional informativeness, sentiment expression, and human-like characteristics.

6 Conclusion

In this study, we propose a new task called Multimodal Customized Review Generation, aimed at generating a customized review based on product images and input words. To this end, we introduce a Multimodal Pre-trained Language Model for generating customized reviews. In particular, we incorporate image captions to establish a bridge between the images and texts, efficiently leveraging information from images. We further design a Text-Driven Fusion module to integrate representations between different modalities. Experimental results show that, our proposed model is capable of generating higher-quality customized reviews.

574
575
576
577
578
579
580
581

582

583
584
585
586

587
588
589
590
591
592

593
594
595
596
597

598
599
600
601
602

603
604
605
606
607
608

609
610
611
612
613
614

615
616
617
618
619
620

621
622
623
624
625

Limitations

Although our proposed model has achieved the best performance in the task of generating customized reviews, it is still necessary for us to explore how to make the generated reviews more diverse and rich in content. In addition, we also need to explore how to capture more multimodal information for generating customized reviews.

References

Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16:345–379.

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.

Roger Bramon, Imma Boada, Anton Bardera, Joaquim Rodriguez, Miquel Feixas, Josep Puig, and Mateu Sbert. 2011. Multimodal data fusion based on mutual information. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1574–1587.

Jiahao Bu, Lei Ren, Shuang Zheng, Yang Yang, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. Asap: A chinese review dataset towards aspect category sentiment analysis and rating prediction. *arXiv preprint arXiv:2103.06605*.

Hongjie Cai, Yaofeng Tu, Xiangsheng Zhou, Jianfei Yu, and Rui Xia. 2020. Aspect-category based sentiment analysis with hierarchical graph convolutional network. In *Proceedings of the 28th international conference on computational linguistics*, pages 833–843.

Zhuang Chen and Tiejun Qian. 2020. [Enhancing aspect term extraction with soft prototypes](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2107–2117, Online. Association for Computational Linguistics.

Jeongwhan Choi, Seoyoung Hong, Noseong Park, and Sung-Bae Cho. 2023. Blurring-sharpening process models for collaborative filtering. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1096–1106.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. 626
627
628
629
630
631
632

Lei Gao, Yulong Wang, Tongcun Liu, Jingyu Wang, Lei Zhang, and Jianxin Liao. 2021. Question-driven span labeling model for aspect–opinion pair extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12875–12883. 633
634
635
636
637

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. 638
639
640
641
642
643
644

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648. 645
646
647
648
649
650

Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue Niu. 2020. Multimodal transformer fusion for continuous emotion recognition. In *ICASSP 2020-IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3507–3511. IEEE. 651
652
653
654
655
656

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv: Learning*. 657
658

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR. 659
660
661
662
663
664
665

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 *ACL 2020*, pages 7871–7880. 666
667
668
669
670
671

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022a. [On vision features in multimodal machine translation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6327–6337, Dublin, Ireland. Association for Computational Linguistics. 672
673
674
675
676
677
678

Junjie Li, Jianfei Yu, and Rui Xia. 2022b. [Generative cross-domain data augmentation for aspect and opinion co-extraction](#). In *Proceedings of the 2022 Conference of the North American Chapter of the* 679
680
681
682

683		Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion Word Expansion and Target Extraction through Double Propagation . <i>Computational Linguistics</i> , 37(1):9–27.	739
684			740
685			741
686			742
687	Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models . <i>arXiv preprint arXiv:2301.12597</i> .	Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer . <i>Journal of Machine Learning Research</i> , 21(140):1–67.	743
688			744
689			745
690			746
691	Pan Li and Alexander Tuzhilin. 2019. Towards controllable and personalized review generation. <i>arXiv preprint arXiv:1910.03506</i> .		747
692			748
693		Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In <i>Proceedings of the 10th international conference on World Wide Web</i> , pages 285–295.	749
694	Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. <i>meeting of the association for computational linguistics</i> .		750
695			751
696			752
697	Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In <i>Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13</i> , pages 740–755. Springer.	Tian Shi, Liuqing Li, Ping Wang, and Chandan K Reddy. 2021. A simple and effective self-supervised contrastive learning framework for aspect detection. In <i>Proceedings of the AAAI conference on artificial intelligence</i> , volume 35, pages 13815–13824.	754
698			755
699			756
700			757
701			758
702		Jie Shuai, Kun Zhang, Le Wu, Peijie Sun, Richang Hong, Meng Wang, and Yong Li. 2022. A review-aware graph contrastive learning framework for recommendation. In <i>Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval</i> , pages 1283–1293.	759
703			760
704	Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. <i>arXiv preprint arXiv:2204.07955</i> .		761
705			762
706			763
707			764
708	Huy Thanh Nguyen and Minh Le Nguyen. 2018. Effective attention networks for aspect-level sentiment classification. In <i>2018 10th International Conference on Knowledge and Systems Engineering (KSE)</i> , pages 25–30. IEEE.		765
709			766
710			767
711			768
712			769
713	Jianmo Ni and Julian McAuley. 2018. Personalized review generation by expanding phrases and attending on aspect-aware representations. In <i>Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)</i> , pages 706–711.	Peijie Sun, Le Wu, Kun Zhang, Yanjie Fu, Richang Hong, and Meng Wang. 2020. Dual learning for explainable recommendation: Towards unifying user preference prediction and review generation. In <i>Proceedings of The Web Conference 2020</i> , pages 837–847.	770
714			771
715			772
716			773
717			774
718			775
719	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 . <i>CoRR</i> , abs/2203.02155.	Peijie Sun, Le Wu, Kun Zhang, Yu Su, and Meng Wang. 2021. An unsupervised aspect-aware recommendation model with explanation text generation. <i>ACM Transactions on Information Systems (TOIS)</i> , 40(3):1–29.	776
720			777
721			778
722			779
723			780
724			781
725		Changxin Tian, Yuexiang Xie, Yaliang Li, Nan Yang, and Wayne Xin Zhao. 2022. Learning to denoise unreliable interactions for graph collaborative filtering. In <i>Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval</i> , pages 122–132.	782
726			783
727	Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. <i>arXiv preprint cs/0205070</i> .	Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models .	784
728			785
729			786
730			787
731	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In <i>Proceedings of the 40th annual meeting of the Association for Computational Linguistics</i> , pages 311–318.		788
732			789
733			790
734			791
735			792
736	Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. <i>Learning</i> .	Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In <i>Proceedings of the conference. Association for Computational Linguistics. Meeting</i> , volume 2019, page 6558. NIH Public Access.	793
737			794
738			795

796 Stéphan Tulkens and Andreas van Cranenburgh. 2020.
797 Embarrassingly simple unsupervised aspect extrac-
798 tion. *arXiv preprint arXiv:2004.13580*.

799 Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie
800 Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and
801 Lijuan Wang. 2022. Git: A generative image-to-text
802 transformer for vision and language. *arXiv preprint*
803 *arXiv:2205.14100*.

804 Yao Wu, Christopher DuBois, Alice X Zheng, and
805 Martin Ester. 2016. Collaborative denoising auto-
806 encoders for top-n recommender systems. In *Pro-*
807 *ceedings of the ninth ACM international conference*
808 *on web search and data mining*, pages 153–162.

809 Zhengxuan Wu and Desmond C Ong. 2021. Context-
810 guided bert for targeted aspect-based sentiment anal-
811 ysis. In *Proceedings of the AAAI conference on arti-*
812 *ficial intelligence*, volume 35, pages 14094–14102.

813 Wu-Dong Xi, Ling Huang, Chang-Dong Wang, Yin-
814 Yu Zheng, and Jian-Huang Lai. 2021. Deep rating
815 and review neural network for item recommendation.
816 *IEEE Transactions on Neural Networks and Learning*
817 *Systems*, 33(11):6726–6736.

818 An Yan, Zhankui He, Jiacheng Li, Tianyang Zhang,
819 and Julian McAuley. 2023. Personalized showcases:
820 Generating multi-modal explanations for recommen-
821 dations. In *Proceedings of the 46th International*
822 *ACM SIGIR Conference on Research and Develop-*
823 *ment in Information Retrieval*, pages 2251–2255.

824 Haozhe Yang, Xianqiang Gao, Jianlong Wu, Tian Gan,
825 Ning Ding, Feijun Jiang, and Liqiang Nie. 2023. Self-
826 adaptive context and modal-interaction modeling for
827 multimodal emotion recognition. In *Findings of*
828 *the Association for Computational Linguistics: ACL*
829 *2023*, pages 6267–6281.

830 Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He,
831 Alex Smola, and Eduard Hovy. 2016. Hierarchical at-
832 tention networks for document classification. In *Pro-*
833 *ceedings of the 2016 conference of the North Ameri-*
834 *can chapter of the association for computational lin-*
835 *guistics: human language technologies*, pages 1480–
836 1489.

837 Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards
838 answering opinion questions: Separating facts from
839 opinions and identifying the polarity of opinion sen-
840 tences. In *Proceedings of the 2003 conference on*
841 *Empirical methods in natural language processing*,
842 pages 129–136.

843 Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cam-
844 bria, and Louis-Philippe Morency. 2017. Tensor
845 fusion network for multimodal sentiment analysis.
846 *arXiv preprint arXiv:1707.07250*.

847 Shiman Zhao, Wei Chen, and Tengjiao Wang. 2022.
848 [Learning cooperative interactions for multi-overlap](#)
849 [aspect sentiment triplet extraction](#). In *Findings of the*
850 *Association for Computational Linguistics: EMNLP*
851 *2022*, pages 3337–3347, Abu Dhabi, United Arab
852 Emirates. Association for Computational Linguistics.

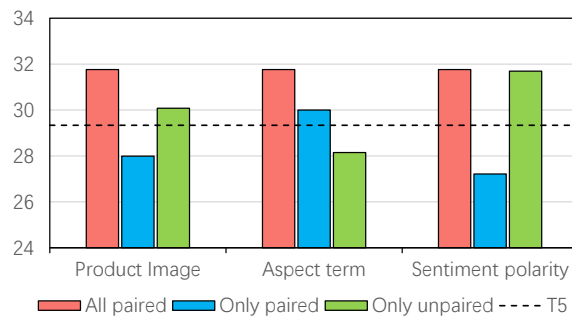


Figure 5: Performance of unpaired and paired multi-modal data

853 Ru Zhou, Wenya Guo, Xumeng Liu, Shenglong Yu,
854 Ying Zhang, and Xiaojie Yuan. 2023. Aom: De-
855 tecting aspect-oriented information for multimodal
856 aspect-based sentiment analysis. *arXiv preprint*
857 *arXiv:2306.01004*.

858 Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Ji-
859 ajun Zhang, and Chengqing Zong. 2018. Msmo:
860 Multimodal summarization with multimodal output.
861 In *Proceedings of the 2018 conference on empiri-*
862 *cal methods in natural language processing*, pages
863 4154–4164.

A Visual and Textual Ablation Study 864

865 In Figure 5, we conduct the ablation study on image
866 and text inputs to ascertain their contributions to
867 the performance. We divided them into three input
868 scenarios. **All paired** denotes that both image and
869 text inputs are correct, serving as a benchmark for
870 comparison. **Only paired** denotes that only the cur-
871 rent element is paired, while the rest are unpaired.
872 **Only unpaired** denotes that only the current ele-
873 ment is unpaired, while the rest are paired.

874 From the experimental results, we can intuitively
875 see that the contribution of sentiment polarity is
876 the smallest, which might be because it does not
877 provide substantially effective information. In the
878 case where only the image is unpaired, the model
879 performance drops significantly, confirming our
880 model’s reliance on visual information. In the sce-
881 nario where only the aspect term is paired, the per-
882 formance still surpasses the baseline, suggesting
883 that visual information not only provides multi-
884 modal information but also acts as a regularization
885 term. Moreover, we find that when the aspect term
886 is unpaired, the model performance drops below
887 the baseline level, implying its critical role in the
888 model.