A Multimodal Dataset and Benchmark for Tourism Review Generation

Hiromasa Yamanishi[1], Ling Xiao[1,*] and Toshihiko Yamasaki[1]

[1] *The University of Tokyo, Hongō, Bunkyo, Tokyo 113-8656, Japan*

Abstract

This paper addresses the challenge of generating accurate and contextually relevant tourism reviews, which are essential to assist travelers in creating reviews and allowing businesses to predict the reactions of different users to tourist spots. To address this problem, we introduce the first multimodal dataset for Japanese tourism review generation, TourMix1M, which contains one million review instances under various conditions, including images, user attributes, user profiles, review ratings, review length, key phrases, and visit seasons, collected from Japanese tourism websites. Based on this dataset, we develop a baseline model for multimodal review generation, LLaVA-Review, by performing instruction tuning of the LLaVA model. Furthermore, to enhance domain knowledge, we apply retrieval-augmented fine-tuning with aggregated tourism review data, exploring two types of knowledge representations: one incorporating noun and adjective information from Sentiment Aware Knowledge Graph, and another using aspect-based summaries from reviews. Experimental results show that LLaVA-Review outperforms existing models in review generation and adapts well to various conditioning factors, with improved accuracy by conditioning the gender, visiting month, review length, key phrases, and user profiles information into the prompt. Furthermore, retrieval-augmented fine-tuning using tourism information effectively improved accuracy across both types of knowledge representations.

Keywords

Tourism, Review Generation, Conditional Review Generation, Large Multimodal Model

1. Introduction

Tourism is one of the most crucial sectors in the global economy, contributing more than 15 trillion US dollars [1], and is enjoyed for various purposes, such as relaxation, exploration, and education. Online platforms such as TripAdvisor and Google Maps play essential roles [2] for tourism. In tourism services, user-generated content (UGC), such as reviews and photos of tourist spots, plays a vital role for both the tourism business and individual tourists. By collecting diverse aspects and opinions about these locations, the tourism business can further improve the services, enhance their credibility, and make more profits. At the same time, UGC helps shape tourists' perceptions of tourist destinations and significantly influences travel planning [3, 4]. As tourism demand diversifies, understanding the preferences, needs, and objectives of different segments is crucial for identifying market opportunities. The literature showed that UGC reflects differences in trends based on user attributes and preferences [5, 6, 7], and analyzing these trends has provided valuable insights for improving tourist destinations and for effective marketing targeting [8, 9, 10].

This paper focuses specifically on review generation. There are two primary applications for review generation. First, presenting automatically generated reviews to users can significantly minimize their effort [11] and encourage more review posting. Additionally, presenting generated reviews to users is beneficial for recommendation systems [12]. Second, predicting user reactions, such as "what a man in his 50s who prefers leisurely tourism would say" can be valuable for businesses in marketing and improving tourist destinations. Recent advancements in deep learning and natural language processing (NLP) have enabled the generation of high-quality reviews [11, 13, 14, 15]. In [11], they developed a

---

Workshop on Recommenders in Tourism (RecTour 2024), October 18th, 2024, co-located with the 18th ACM Conference on Recommender Systems, Bari, Italy

[*] Corresponding author. Email: ling@cvm.t.u-tokyo.ac.jp

yamanishi@cvm.t.u-tokyo.ac.jp (H. Yamanishi); ling@cvm.t.u-tokyo.ac.jp (L. Xiao); yamasaki@cvm.t.u-tokyo.ac.jp (T. Yamasaki)

0009-0006-8289-757X (H. Yamanishi); 0000-0002-4650-8841 (L. Xiao); 0000-0002-1784-2314 (T. Yamasaki)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3886/paper6.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

recurrent neural networks based on long short-term memory (LSTM) for user conditioned product review generation. Truong et al. [13] improved review generation by leveraging image information. Li et al. [14] enhanced product-related factual information by leveraging knowledge graphs. Xie et al. [15] utilized existing reviews as additional context and generated reviews using GPT-2. However, although various datasets exist, relatively few studies focus on tourism-related data in review generation research literature. In particular, there is a significant gap in research that employs multimodal datasets combining diverse contexts, such as images, rich user attributes, and review content, specifically tailored for tourism. To address this gap, this research creates a multimodal review generation dataset, TourMix1M, specifically tailored for tourism. The TourMix1M is created from data collected from Japanese tourism websites, comprising 470,000 images, 510,000 reviews, and a total of one million training instances. We will release a list of URLs for the collected data at https://github.com/HiromasaYamanishi/TourMix1M. Our dataset supports the generation of reviews based on various contexts such as images, age, gender, season, user profiles, review length, and ratings. This is the first dataset specialized in generating tourism reviews that incorporates diverse information such as images and user data, promoting the understanding and advancement of review generation in the tourism sector.

We also propose a baseline model for multimodal review generation, LLaVA-Review. LLaVA-Review was developed by instruction-tuning the open-source LLaVA model [16]. Compared to previous multimodal review generation models [13], LLaVA-Review offers greater flexibility in handling different conditions by changing natural language instructions. LLaVA-Review outperformed state-of-the-art Large Multimodal Model (LMM) ChatGPT-4o and review generation models on various metrics in experiments conducted on our dataset, particularly in terms of BLEU, ROUGE, and consistency with user opinions. By incorporating tourism knowledge and user opinions to fine-tune LLaVA, LLaVA-Review also demonstrated stronger adaptability to diverse tourism related conditioning. Furthermore, inspired by recent research in enhancing domain specificity [17, 18, 19] with retrieval-augmented fine-tuning, we incorporate aggregated information from tourism spot reviews during fine-tuning and inference stages. We explored two approaches: subgraph method and aspect-based summary. The subgraph method means adding noun and adjective information extracted from a Sentiment Aware Knowledge Graph [20]. The aspect-based summary denotes using aspect-based summaries generated from reviews. Experimental results showed that the subgraph method increased BLEU by 12% and user opinion accuracy by 11%, while using aspect-based summary improved proper noun variety by 18% and domain knowledge by 4%.

One related submission to this paper is [21], which develops a multimodal model for landmark recognition and review generation. However, this work differs from [21] in two key aspects: 1) Dataset: The dataset used in [21] addresses limited conditions such as gender, age, group, and key phrases, whereas the TourMix1M includes a wider range of factors such as visit season, user profiles, ratings, and review length, providing a more comprehensive foundation for tourism review generation. 2) Model design: [21] addresses review generation as part of a multitask framework, which is trained by standard instruction-tuning. Our research explores the usage of information aggregated from reviews to improve quality. Overall, our contributions are threefold:

- We created a multimodal dataset, TourMix1M, for Japanese tourism review generation. The TourMix1M has one million instances and ten types of diverse conditions. The URLs used in the dataset construction will be made publicly available.
- We proposed LLaVA-Review, which outperformed state-of-the-art large multimodal models and review generation models in general review generation. LLaVA-Review showed strong adaptability to different conditions, and we analyzed variations in its outputs across these conditions.
- We propose two retrieval-augmented fine-tuning methods: aggregating sub-graphs from a Sentiment Aware Knowledge Graph and aspect-based summaries from reviews, both of them enhanced accuracy.

2. Related Works

2.1. User-Generated Content in Tourism

Tourism is a key sector in the global economy, accounting for over 10% of global economic output [1]. As tourism is often a high-cost, once in a lifetime experience, UGC on social platforms plays a crucial role in helping tourists decide what to visit. UGC, typically in the form of reviews and images, offers valuable insights. Reviews express user opinions on aspects such as content, price, and service, while images enhance understanding of intangible experiences such as tourist destinations [2]. Furthermore, images complement textual information and function as visual evidence to substantiate the content, thus enhancing the usefulness of reviews when combined with images [22]. UGC significantly influences travel planning and behavior by shaping both cognitive and emotional perceptions of tourist destinations [4, 23, 3].

Analyzing UGC is vital for marketing and improving tourist destinations. Text mining has shown that reviews reflect user satisfaction and preferences based on factors such as gender [5], age, companions [6], and season [7]. For example, [5] finds that men are more interested in historical sites, while women prefer landscapes and rural areas. Similarly, [6] shows that couples tend to leave reviews with higher satisfaction levels. Analyzing tourist reviews with deep learning techniques, such as topic modeling and sentiment classification, is effective for marketing and destination improvement [8, 9]. Recently, large language models such as ChatGPT have shown potential for applications such as automated replies to customer questions and requests in tourism through prompt engineering [24, 25]. Our research focuses on developing large-scale models using tourism-specific corpora and images.

2.2. Review Generation

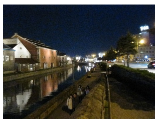
Review generation helps users write reviews [11] and improve the trustworthiness of recommendations [26, 12]. Unlike explanation generation [26, 27] or tip generation [28], review generation aims to create longer texts that cover multiple aspects. Various models have been developed for review generation using recurrent neural networks (RNNs) [11, 14, 29, 30], generative adversarial networks (GANs) [31, 32], and large-scale models [33]. Typically, these models generate reviews based on inputs such as user data, product information, ratings, or sentiment polarity. Furthermore, some research explores generation conditioned on images [13] or aspects [34] such as Enviroment, Service, and Price. Additionally, guiding the generation process with information extracted from reviews is effective in improving quality and factuality. For instance, [29] identified relevant aspects between users and items and guided generation using words from those aspects. Similarly, [35] proposed a method that first generates an aspect sequence and then performs review generation from coarse to fine. In addition, [14] expanded a Freebase-based knowledge graph using user and item reviews, capturing user preferences for each aspect via Caps GNN. Furthermore, [15] generated reviews by inputting past reviews and key terms into GPT-2 and asking questions like "What was great?" to guide the generation.

Research on review generation in the tourism domain remains underexplored. One of the biggest challenges is the lack of a dataset. One related work under submission is [21], where review generation is addressed as part of a multitask framework, with conditions limited to user attributes and keywords. This paper focuses on creating a dataset specifically for review generation with a broader range of ten conditions such as visit timing, user profiles, ratings, and review length. We also apply a large-scale multimodal model to multimodal review generation for the first time.

2.3. Retrieval-augmented Large-Scale Model Fine-tuning

Large-scale models (LMs) based on the Transformer architecture [36] have shown high performance across various tasks, with notable examples including large language models (LLMs) such as LLaMA [37] and GPT [38, 39], large multimodal models (LMMs) such as LLaVA [40, 16], ChatGPT-4 [41], and QwenVL [42]. These LMMs align language with images and can generate text across diverse inputs and conditions.

Table 1
Examples of text-image pair construction using CLIP. The bold text represents sentences retrieved through image-sentence retrieval.

Image	Spot Name	Review
[2]	KITTE Ootemachi	(Image-Sentence-Review) To avoid the crowds near Christmas, I went Christmas tree touring in mid-December. **At KITTE, a large white Christmas tree was displayed in the atrium on the first floor entrance.**
 [3]	The canal and the stone warehouses	(Image-Review) This time, I walked along the canal at night. Illuminated by gas lamps, I was satisfied with the beautiful scenery. After dinner, I walked all the way to the back and back again, making for a nice walk. The warehouses were also lit up and looked beautiful. I definitely recommend going at night.

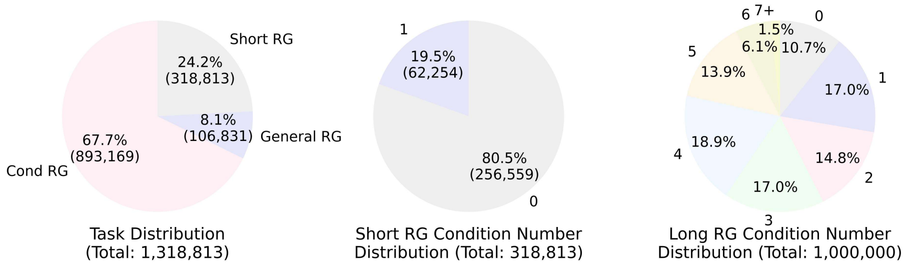
However, these models often lack domain-specific knowledge. To address this, methods such as instruction-tuning [43] and retrieval-augmented generation (RAG) [44, 45, 46, 47, 48, 49, 17, 50, 18, 51] have been developed to enhance performance in specific domains. RAG, which incorporates external knowledge, including structured knowledge such as Knowledge Graphs [44, 45, 46, 47] and unstructured sources [48, 49, 17, 50, 18, 51], has proven effective for domain-specific tasks. Knowledge integration has proven to be beneficial either during pre-training [48, 49, 44], fine-tuning [17, 18, 45], or inference [51, 46]. Our research aligns closely with studies such as [17, 18, 45], which focus on retrieval-augmentation during both fine-tuning and inference stages. Most research in this area is applied to question answering, where a single or a few lengthy documents are retrieved. However, this approach may not be ideal for review generation. A small number of reviews may not capture all aspects and opinions of a certain tourism spot, while too many reviews can introduce noise, potentially degrading quality. Although frameworks such as [19] train large language models using concept retrieval, the knowledge may still be insufficient. So, effective knowledge retrieval methods for toursim review generation requires careful design.

3. Proposed Method

Considering the fact that most tourists take photos and their experiences vary depending on context information such as user attributes and visit timing, we have constructed a multimodal tourism review generation dataset. The created dataset accounts for ten conditions, including review length, user gender, user age, group, visiting month, visiting season, two types of user profiles, and rating. Moreover, we propose a high-performance LLaVA-Review model for review generation. We also propose retrieval-augmented fine tuning where aggregated review information is incorporated into the prompt during both training and inference. Each of these aspects will be detailed in the following sections. We propose two formats for aggregating review information: one based on subgraphs consisting of noun and adjective information extracted from a Sentiment Aware Knowledge Graph [20], and another based on aspect based review summaries. Each of these aspects will be detailed in the following sections. Figures 4, 6, 7, 8, and 10 as well as Tables 1 and 3 present the original reviews alongside the generated examples, all of which were originally in Japanese and have been translated into English for demonstration purposes in this paper.

---

[2] https://cdn.jalan.jp/jalan/img/6/kuchikomi/3776/KL/6b4e1_0003776908_1.webp
[3] https://cdn.jalan.jp/jalan/img/0/kuchikomi/4140/KL/3a837_0004140092_1.webp



Component	Count
Dialogues	1,000,000
Prompts	1,310,000
Reviews	545,891
Images	476,167
Tourism Spots	51,011

Figure 1: Dataset statistics. Left: proportions of the three tasks—Short Review Generation, General Review Generation, and Conditional Review Generation, with the latter two collectively referred to as long review generation. "RG" refers to review generation. Middle and right: distribution of conditions in short and long review generation. The table on the right summarizes the dataset components.

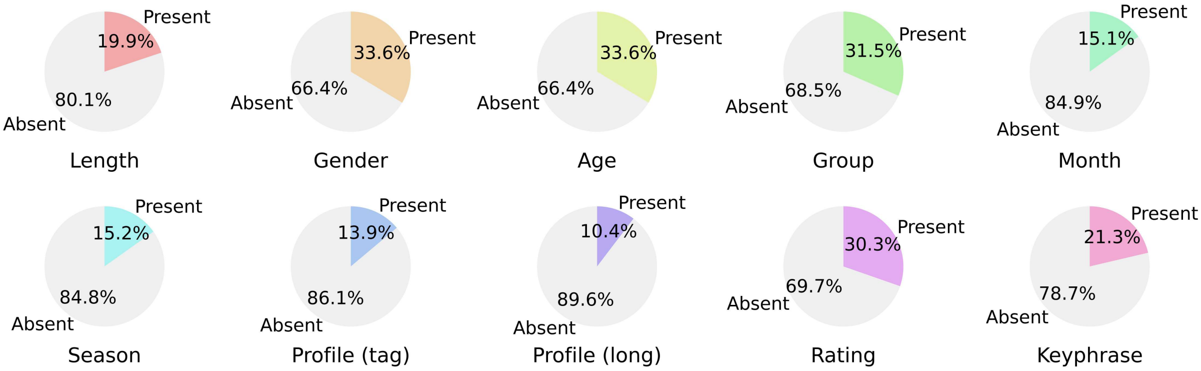


Figure 2: Proportion of each of the 10 conditioning types within the whole dataset. We constructed prompts by sampling attributes to increase the diversity of conditioning.

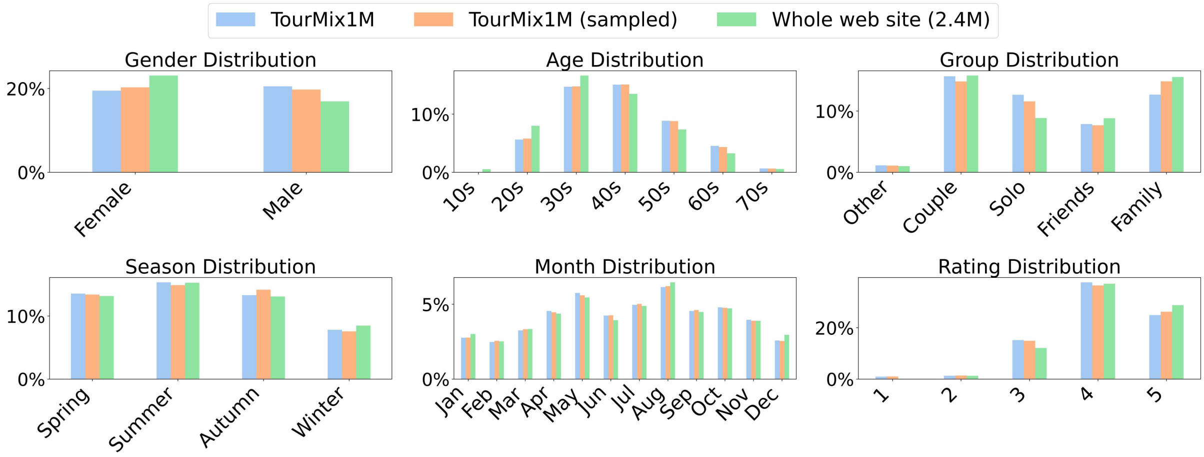


Figure 3: Comparison of categorical attribute distribution in: 1) the TourMix1M dataset, and 2) the sampled TourMix1M dataset 3) the original web data, .

3.1. TourMix1M Dataset

We created a multimodal tourism dataset for review generation by utilizing data collected from the Japanese tourism website jalan.net[4], with permission to use its data for non-commercial purposes. We originally collected 470k images and 2.4 million reviews from the website for 51k tourist spots across Japan that had a sufficient number of images and reviews. This represents a large portion of the content available on the website. Initially, the collected images and texts were not always paired. To create image-text pairs, we employed the Contrastive Language-Image Pretraining (CLIP) [52] method,

---

[4] https://www.jalan.net/

specifically focusing on the model tailored for the Japanese language[5]. The data were generated by pairing images and reviews based on the nearest neighbors in the embedding space. To create training pairs, we used three retrieval methods for diverse image-text matching: image-to-review retrieval, review-to-image retrieval, and image-sentence-review retrieval. Table 1 shows examples of obtained image-text pairs. In the image-sentence-review retrieval process, sentences were first extracted from the images, after which the full corresponding reviews were identified by locating the reviews that contained those sentences.

We constructed three review generation tasks based on the generated image-text pairs. The tasks consist of Short Review Generation, General Review Generation, and Conditional Review Generation. General and Conditional Review Generation are collectively referred to as Long Review Generation. For Short Review Generation, image-sentence pairs obtained from image-sentence retrieval were used to generate concise reviews. For Long Review Generation, image-review pairs obtained from three retrieval methods were used to generate reviews. The left side of Figure 1 shows the distribution of each task. As input for each task, in General Review Generation, instructions were given to generate reviews based solely on images and place names. In Short Review Generation, only one condition, rating, was applied. For Conditional Review Generation, instructions were given to generate reviews based on combinations of images, place names, and various conditions. The conditions considered include ten categories: review length, gender, age, groups, visit month, season, two types of user profiles (tag and long), rating, and key phrase in the review. These variables encompass a wide range of conditions specific to tourism. While this research does not perform conditioning based on user ID, making the setup less personalized, it provides a more general context and attribute based framework, which is applicable even to cold-start users. For example, a new user, such as a man in his 50s who enjoys leisurely activities, could request a review for a spring visit to a tranquil garden or scenic nature trail, allowing the system to generate a review tailored to his preferences without prior interactions. Moreover, our dataset enables analysis of how different conditioning factors, such as age, gender, and season, influence the generated reviews, providing insights into the diverse user experiences.

Specifically, for categorical variables, gender is either male or female; age is in ten-year increments from the 10s to the 90s; groups include family, couple, friends, single, or other; rating is an integer between 1 and 5; visit month is an integer between 1 and 12; and the season is either spring, summer, autumn, or winter. For user profiles, we use two types: “tag” profiles, which are simple, keyword-based summaries such as “history enthusiast” and “long” profiles, which provide more detailed descriptions such as “a curious traveler with a deep interest in local history and culture, enjoying museum visits and exploring traditional cuisine.” These profiles are generated by prompting a large language model[6] based on past reviews. For key phrases, we use an LLM to extract important parts from sentences, ranging from single words to short sentences, that users found positive or negative. The attributes for each review are derived from the accompanying metadata of reviews and the data of the user who wrote them. During training, instead of using all conditions for each instance, we perform condition sampling based on a pre-defined probability to increase the diversity of condition combinations. The probabilities were chosen so that each conditioning factor would appear in approximately 10% to 30% of the entire dataset. Figure 2 illustrates how frequently each attribute appears throughout the entire dataset. Figure 3 shows the distribution of categorical attributes such as Gender, Age, Group, Season, Month, and Rating across the original 2.4 million reviews for both the TourMix1M dataset and the sampled TourMix1M dataset. The results indicate that the TourMix1M dataset closely replicates the original distribution. We release both the full set of conditions and the specific conditions used in the experiments. The middle and right sections of Figure 1 show the distribution of the number of conditions in short and long review generation, respectively. For prompt construction, in the General Review Generation task, prompts are structured as “Generate a review for Sensoji based on the image.” For Short Review Generation, prompts are phrased as “Generate a concise review for Sensoji based on the image,” and for Conditional Review Generation, such as “Generate a 4-star review for Sensoji

---

[5] https://huggingface.co/rinna/japanese-clip-vit-b-16

[6] https://huggingface.co/UCLA-AGI/Gemma-2-9B-It-SPPO-Iter3

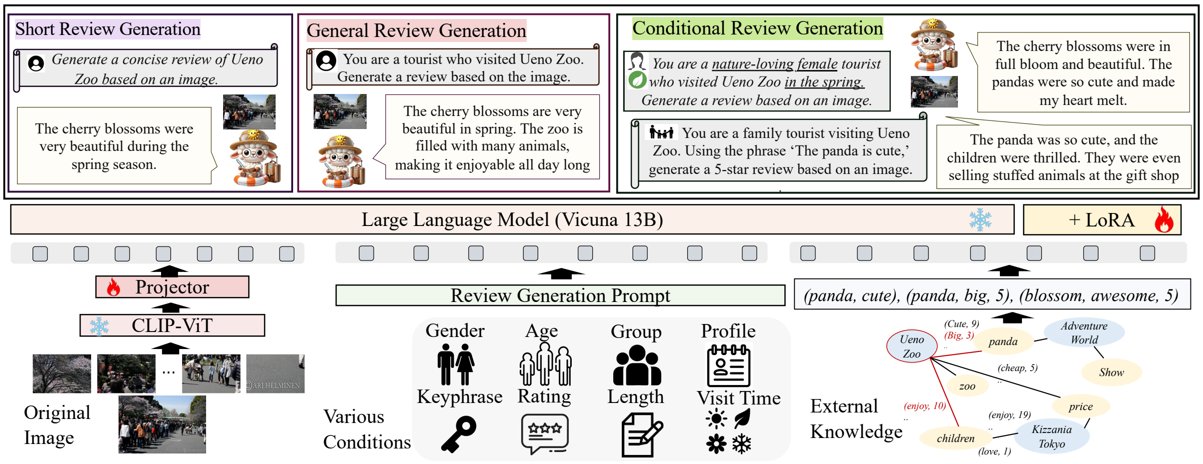

Short Review Generation
Generate a concise review of Ueno Zoo based on an image.
The cherry blossoms were very beautiful during the spring season.
General Review Generation
You are a tourist who visited Ueno Zoo. Generate a review based on the image.
The cherry blossoms are very beautiful in spring. The zoo is filled with many animals, making it enjoyable all day long
Conditional Review Generation
You are a nature-loving female tourist who visited Ueno Zoo in the spring. Generate a review based on an image.
The cherry blossoms were in full bloom and beautiful. The pandas were so cute and made my heart melt.
You are a family tourist visiting Ueno Zoo. Using the phrase 'The panda is cute,' generate a 5-star review based on an image.
The panda was so cute, and the children were thrilled. They were even selling stuffed animals at the gift shop
Large Language Model (Vicuna 13B)
+ LoRA
Projector
CLIP-ViT
Original Image
Review Generation Prompt
Various Conditions
Gender
Age
Group
Profile
Keyphrase
Rating
Length
Visit Time
(panda, cute), (panda, big, 5), (blossom, awesome, 5)
External Knowledge
Ueno Zoo
(Cute, 9)
(Big, 3)
panda
Adventure World
(cheap, 5)
Show
zoo
(enjoy, 10)
price
(enjoy, 19)
children
Kizzania Tokyo
(love, 1)

Figure 4: LLaVA-Review architecture: Images are split into patches and projected with an MLP layer. Image tokens are combined with review prompts and external information. Instruction-tuning is done using instruction-response pairs, as shown at the top, for three review generation tasks. [7]

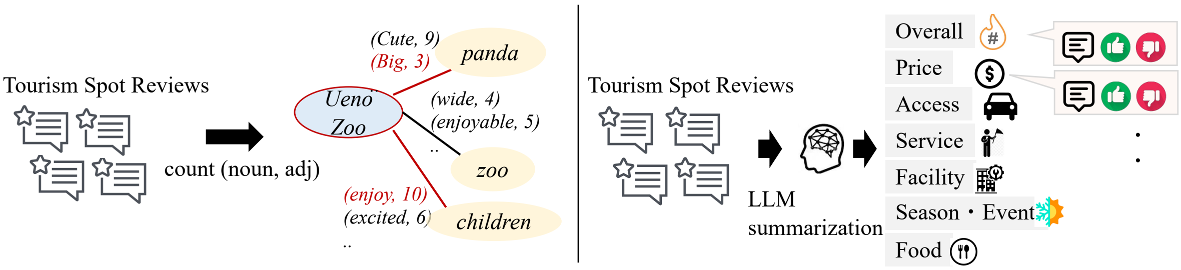


Figure 5: External Knowledge for retrieval-augmented fine-tuning. Left: subgraph from sentiment aware knowledge graph. Right: aspect and sentiment based review summary

using the keyword 'blossom' as written by a male in his 20s based on the image." The resulting training dataset comprises 1 million dialogues, 1.31 million prompts, 545,891 reviews, 476,167 images, and 51,011 tourist spots.

3.2. LLaVA-Review

3.2.1. Model Architecture

The architecture of our proposed model is illustrated in Figure 4. The icons in the figure will represent LLaVA-Review in subsequent sections. The baseline model we developed, LLaVA-Review, is based on LLaVA [40, 16], a prominent open-source large-scale multimodal model. In LLaVA, the image is initially divided into patches, which are then converted into image tokens by the image encoder. These tokens are transformed into the language space via a projector consisting of multilayer perceptrons. By simultaneously feeding instruction tokens and image tokens into the large language model, the model generates responses that incorporate image information. The training process consists of two stages: pretraining, where only the projector is trained, and fine-tuning, where both the large language model (LLM) and the projector are optimized. In the proposed LLaVA-Review, we performed only fine-tuning based on the model pretrained in [16]. Instruction-tuning [43] was employed as the fine-tuning method, which learns to generate responses using instruction-response pairs. Optimization is conducted by minimizing the negative log-likelihood of the response generation using the following loss function, where x represents tokens, k denotes the length of the instruction part, and T is the total length of the text. We used Vicuna-13B [53], which has strong capabilities in Japanese, as the language model,

---

[7] https://farm6.static.flickr.com/5018/5580357389_a472ea2466_z.jpg

and CLIP as the image encoder. Additionally, to ensure efficient training and to prevent degradation in language capabilities, we employed a Low-Rank Adaptation (LoRA) [54]-based strategy for model training.

$$\mathcal{L} = -\sum_{t=k+1}^{T} \log P\left(x_t \mid x_1, \ldots, x_{t-1}\right). \tag{1}$$

3.2.2. Instruction-Tuning with External Knowledge

We propose utilizing knowledge extracted from existing reviews as references to enhance the quality of review generation. This external knowledge is incorporated during both the training and inference phases, with the extracted sentences added to the end of the review generation prompts. Tourist destination reviews typically reflect diverse perspectives and opinions, resulting in a large volume of content. Relying on only a few reviews may fail to fully capture this diversity, while using the entire review text as input risks introducing noise. To address these challenges, we empirically evaluated various external knowledge aggregation methods that effectively capture domain knowledge and user opinions. Figure 5 represents two proposed methods. We utilized external knowledge constructed from 2.4 million reviews that do not overlap with test reviews.

Subgraph-based method. This method involves sampling from a Sentiment Aware Knowledge Graph (SAKG). SAKG is a knowledge graph that incorporates user opinions and sentiment information. The SAKG in this research is represented as $G = \{(e_h, r, e_t) | e_h, e_t \in E, r \in R\}$, where E represents entities, R represents relationships, and a triplet (e_h, r, e_t), which denotes a relationship r from a head entity e_h (e.g., Ueno Park) to a tail entity e_t (e.g., Panda). Unlike previous work [20], this graph uses edges to represent adjectives and their frequency of use, such as (cute, 9) and (big, 4). The graph is constructed by extracting noun-adjective pairs from reviews using syntactic parsing and aggregating them for each tourist spot. During training, first, between 1 to 5 entities related to the target destination are sampled based on edge frequency. Subsequently, relationships associated with the sampled entities are selected similarly. During inference, the top-k entities and relationships by total edge frequency are sampled and incorporated into the prompt as natural language.

Summary-based method. This method involves adding aspect-based summaries. Previous research on review summarization has shown that aspect-based summaries facilitate a broader understanding of items [55, 10]. Furthermore, incorporating aspect information is crucial in review generation [55]. In this research, we input reviews and prompt a large-scale language model[5] to generate a summary in a single stage. Specifically, for each tourist destination, up to 50 reviews are selected as input, with a total length less than 5,000 characters. The prompt instructs to create summaries of around 1000 characters for reviews from the perspectives of overviews, positive opinions, and negative opinions for the elements of key content, including price, service, food and drink, facilities, transportation access, and seasonal events.

4. Experimental Results

4.1. Experimental Setup

The TourMix1M dataset was employed to train a model for generating reviews based on inputs such as images, tourist destination names, and various contextual conditions. Specifically, to create the test data, images not used in the training set were first selected, and then the nearest corresponding reviews were retrieved with CLIP embeddings. Only image-review pairs where both the image and review were absent from the training data were included in the test set.

The training of LLaVA-Review was conducted using eight 48GB Ada 6000 RTX GPUs with a batch size of 80 and a learning rate of 2×10^{-4}, taking approximately 37 hours per epoch. When incorporating external knowledge, the training time increased to 41 hours for the subgraph-based method and 60 hours for the summary-based method. Evaluation was performed using 1,000 image-review pairs that did not overlap with the training data.

In this research, the key characteristics for effective review generation are 1) the integration of image information, 2) maintaining high text quality, 3) incorporating detailed information about tourist destinations, and 4) accounting for user opinions, particularly the collective sentiment regarding tourist experiences.

The evaluation metrics included BLEU [56], ROUGE-1, ROUGE-L [57], CIDEr [58], diversity (DIV), the number of unique proper nouns (PROPN), TFIDF-F1 score, and Senti-F1 score. BLEU, ROUGE, and CIDEr were used for overall text quality. BLEU and ROUGE were calculated with the sumeval library [8] and CIDEr was calculated with pycocoevalcap library [9]. Diversity was assessed based on adjectives, nouns, proper nouns, and verbs. It was calculated by measuring the overlap of these features between generated sentences, using part-of-speech information from GiNZA[10], following the approach of [59]. The number of unique proper nouns was calculated by dividing the total number of unique proper nouns in all generated reviews by the number of reviews. To evaluate domain knowledge, the TFIDF-F1 metric was used by identifying the top 15 TFIDF words for each tourist spot and calculating the F1-score between these top words and those in the generated reviews. The Senti-F1 metric, developed using Aspect-Based Sentiment Analysis (ABSA) [60], measured user opinion consideration by extracting (feature, opinion, sentiment) triplets from the text via a large language model[5]. F1 scores for the alignment of (feature, sentiment) and (feature, opinion) were averaged to produce the Senti-F1 score. For all metrics, except length, higher values indicate better performance.

The comparison methods include MRG [13], PETER [26], PEPLER [27], LLaVA-1.5 [16], ChatGPT-4V, and ChatGPT-4o. MRG is a multimodal review generation model based on LSTM [61]. We use VGG16 [62] for vision backbone. PETER is an explanation generation model based on transformer structure, while PEPLER is based on GPT-2 [38]. We utilized a GPT-2 model [11] trained on Japanese data for PEPLER. In PETER and PEPLER, image features were extracted using ResNet [63], then reduced via PCA [64], and clustered with KMeans [65] to generate photo_id, which was used in place of user_id. LLaVA-1.5 is an open-source large multimodal model, while ChatGPT-4V [41] and ChatGPT-4o are closed-source large multimodal models known for their state-of-the-art knowledge and language capabilities. For large-scale multimodal models, prompts such as "You are a tourist who visited location. Generate a review based on the image" are used. Since models other than the proposed method tend to generate verbose text, we prompt the length of generated review to be approximately 100 characters. For retrieval-augmented fine-tuning, we employed two methods: one that extracts entities for comparison with subgraphs, and another that retrieves up to seven reviews based on CLIP image similarity for comparison with summaries. During inference, we used four entities and three relations in the subgraph method.

4.2. General Review Generation

Table 2 presents the quantitative evaluation results of review generation. General LMMs such as LLaVA-1.5 and GPT-4v achieved high ROUGE scores due to their strong linguistic capabilities. However, their knowledge related to tourism domain and user opinions is limited. ChatGPT-4o demonstrates high quality in terms of domain specificity and understanding user opinions. However, it sometimes generates factually incorrect outputs, such as describing a crowded place as quiet. Additionally, some explanations lean towards generalizations and lack detailed knowledge of tourism spots. For fine-tuned models such as MRG, PETER, PEPLER and LLaVA-Review, accuracy was generally improved in terms of BLEU and ROUGE-L quality metrics, as well as domain specificity and user opinion metrics. For the

---

[8] https://github.com/chakki-works/sumeval

[9] https://github.com/sks3i/pycocoevalcap

[10] https://github.com/megagonlabs/ginza

[11] https://huggingface.co/rinna/japanese-gpt2-medium

[12] https://cdn.jalan.jp/jalan/img/2/kuchikomi/3622/KL/ed832_0003622962_1.webp

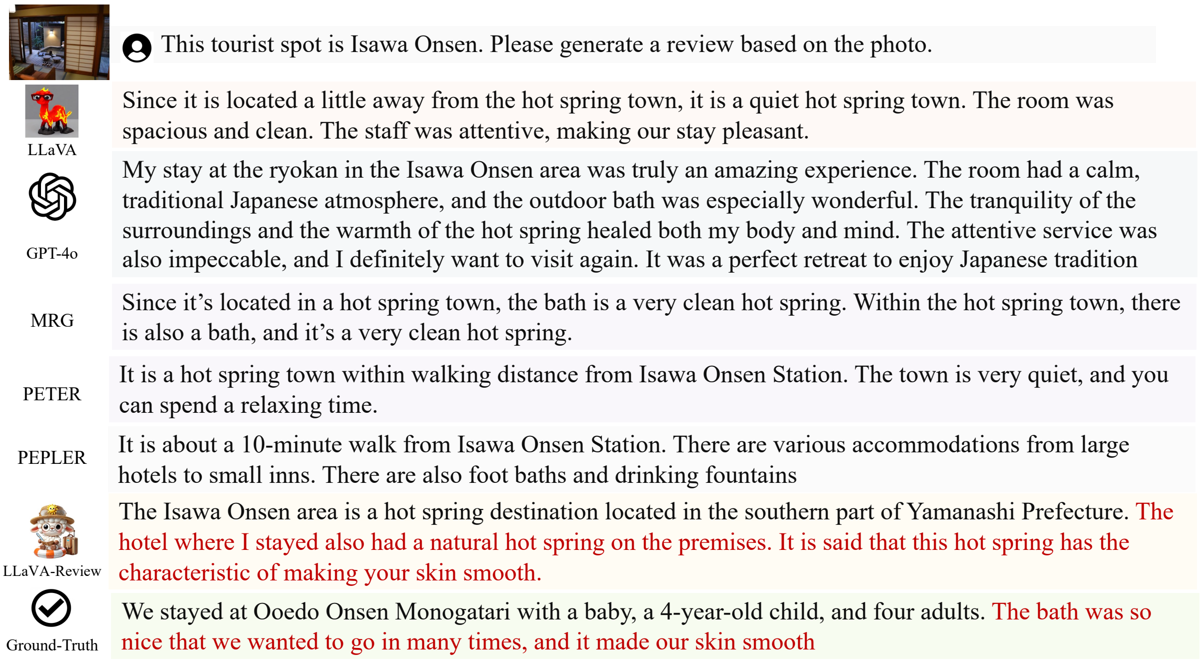
[13] https://cdn.jalan.jp/jalan/img/5/kuchikomi/3905/KL/52041_0003905143_1.webp

[14] https://cdn.jalan.jp/jalan/img/4/kuchikomi/0894/KL/de82f_0000894193_1.webp

Table 2

Results of General Review Generation: The first group compares models. The second group compares retrieval-augmented fine-tuning with different knowledge sources. The last group shows results of RAG during ChatGPT-4o inference. Bold text indicates the best performance in each groups.

Model	BLUE	ROUGE-1	ROUGE-L	CIDEr	DIV	PROPN	TFIDF-F1	Senti-F1	length
LLaVA-1.5 [16]	0.683	0.254	0.162	0.099	0.863	0.303	0.141	0.029	133.8
ChatGPT-4V [41]	0.622	0.250	0.165	0.103	**0.955**	0.278	0.169	0.036	70.9
ChatGPT-4o	0.856	**0.272**	0.168	0.022	0.926	0.270	0.175	0.040	122.8
MRG [13]	0.692	0.202	0.158	0.041	0.941	0.025	0.169	0.029	51.1
PETER [26]	1.192	0.241	0.179	**0.116**	0.923	**0.525**	**0.327**	0.042	62.1
PEPLER [27]	1.157	0.248	0.171	0.115	0.931	0.452	0.316	0.040	60.5
LLaVA-Review	**1.291**	**0.272**	**0.185**	0.108	0.920	0.457	0.244	**0.045**	54.9
LLaVA-Review w entity	1.320	0.271	0.184	**0.111**	0.921	0.441	0.243	0.045	56.1
LLaVA-Review w subgraph	**1.454**	**0.275**	**0.187**	0.110	0.918	0.394	0.247	**0.050**	55.1
LLaVA-Review w 7 reviews	1.175	0.258	0.178	0.100	**0.937**	0.491	0.230	0.049	53.1
LLaVA-Review w summary	1.296	0.269	0.184	**0.111**	0.927	**0.543**	**0.254**	0.046	55.3
ChatGPT-4o + RAG (5 reviews)	0.935	0.264	0.176	**0.151**	**0.967**	0.356	0.250	0.045	61.0
ChatGPT-4o + RAG (10 reviews)	0.942	0.267	0.179	0.144	0.965	0.347	0.255	**0.046**	63.4
ChatGPT-4o + RAG (20 reviews)	**1.024**	**0.269**	**0.179**	0.141	0.965	**0.380**	**0.258**	0.045	67.7



This tourist spot is Isawa Onsen. Please generate a review based on the photo.

LLaVA

Since it is located a little away from the hot spring town, it is a quiet hot spring town. The room was spacious and clean. The staff was attentive, making our stay pleasant.

GPT-4o

My stay at the ryokan in the Isawa Onsen area was truly an amazing experience. The room had a calm, traditional Japanese atmosphere, and the outdoor bath was especially wonderful. The tranquility of the surroundings and the warmth of the hot spring healed both my body and mind. The attentive service was also impeccable, and I definitely want to visit again. It was a perfect retreat to enjoy Japanese tradition

MRG

Since it's located in a hot spring town, the bath is a very clean hot spring. Within the hot spring town, there is also a bath, and it's a very clean hot spring.

PETER

It is a hot spring town within walking distance from Isawa Onsen Station. The town is very quiet, and you can spend a relaxing time.

PEPLER

It is about a 10-minute walk from Isawa Onsen Station. There are various accommodations from large hotels to small inns. There are also foot baths and drinking fountains

LLaVA-Review

The Isawa Onsen area is a hot spring destination located in the southern part of Yamanashi Prefecture. The hotel where I stayed also had a natural hot spring on the premises. It is said that this hot spring has the characteristic of making your skin smooth.

Ground-Truth

We stayed at Ooedo Onsen Monogatari with a baby, a 4-year-old child, and four adults. The bath was so nice that we wanted to go in many times, and it made our skin smooth

Figure 6: Comparison of General Review Generation using different models. Each model is given an image and tourism spot name as input. The proposed baseline accurately captures the image information, user opinions, and domain knowledge. [12]

MRG, due to the limited expressiveness of the model, the generated output tended to have a relatively small vocabulary and frequent repetitions, leading to lower overall accuracy. However, the PETER achieved the highest CIDEr score and performed well in domain specificity and the number of proper nouns used, as the expressiveness was improved by using the Transformer. The PETER also exhibits limitations of repetitive outputs at the sentence or phrase level, likely due to a limited number of Transformer layers. In contrast, the PEPLER consistently delivered superior performance in both review quality and domain relevance, producing natural and coherent outputs.

The proposed method, LLaVA-Review, leverages a large corpus of UGC and demonstrates excellent review performance across metrics, particularly in BLEU, ROUGE-L, and sentiment expression. This improvement is attributed to the model's effective capture of image features and its capability to incorporate relevant tourist information and user opinions, facilitated by its large parameterization. While it tends to generate shorter reviews, the length can be adjusted by specifying the desired review

Table 3

Examples showing the generated reviews with different retrieval strategies. (a) shows subgraph retrieval while (b) shows summary retrieval. Subgraph enhances user opinion while summary enhances domain knowledge.[13] [14]

	Context	Gold response	w/o retrieval	w retrieval
(a)	 Shimokuri Village	There is a parking lot at the highest point of the village, and from there it took about 20 minutes to walk to the observation deck overlooking the village. Shimokuri Village opened in the deep mountains. It was indeed a heavenly village. The scenery is spectacular.	This place became famous because of an old TV drama. If you keep climbing the mountain road, you will see a vast expanse of rice fields. This is Shimokuri Village.	It is a village in the mountains. The roads are narrow and there are few parking lots, but it is worth a visit. The view from the observation deck is wonderful.
(b)	 Yamanashi Prefectural Art Museum	I often go to see Millet's paintings in the permanent exhibition because I really like them. It's usually not very crowded, so I recommend it. I especially liked Millet's painting "Pauline," but when I visited, it was on loan overseas.	The Yamanashi Prefectural Art Museum is located near Takeda Shrine in Kofu City. The permanent exhibition features many works by artists from Yamanashi.	I love Millet's paintings, so I visited the Yamanashi Prefectural Art Museum. Seeing Millet's works calms my mind and makes me feel liberated from the hustle and bustle of everyday life.

(a) Retrieved Subgraph	**(b) Retrieved Summary**
(road, narrow, 18), (road, narrow, 8), (road, difficult, 6) (mountain road, narrow, 7), (mountain road, difficult, 5), (mountain road, good, 3) (scenery, wonderful, 5), (scenery, good, 3)	(extraction of summary part) The Yamanashi Prefectural Art Museum offers an extensive collection centered around Millet's works, providing a quiet and serene environment for visitors. The museum also features outdoor sculptures and a park, allowing visitors to enjoy both art appreciation and a leisurely stroll. With relatively easy access, the museum provides a range of services that visitors will appreciate, including discounts for local residents and special offers for those staying at nearby accommodations.

length in the prompt as discussed later. Figure 6 illustrates the generated review results for Isawa Onsen. ChatGPT-4o produces relatively descriptive and typical outputs, while PETER and PEPLER effectively capture the domain-specific information of the tourist spot. Notably, the proposed LLaVA-Review accurately identifies the presence of indoor hot springs and their characteristics from the images, resulting in high-quality generation.

Table 2 also presents the effects of incorporating external knowledge. While knowledge acquisition improved accuracy overall, the extent of improvement varied. Sampling from the Knowledge Graph contributed to enhancements in BLEU by 12% and Senti-F1 by 11%, leading to better consideration of user opinions. In contrast, simply retrieving nouns showed limited improvements, highlighting the importance of including adjective information. Aspect-based summaries notably enhanced domain specificity and informativeness, increasing TFIDF-F1 by 4% and the number of proper nouns by 12%, without negatively affecting quality metrics. However, directly retrieving reviews introduced noise, which lowered overall quality despite improved user opinion consideration. Table 3 shows specific examples.

Finally, for ChatGPT-4o, we retrieved five, ten, and twenty reviews similar to the image using CLIP, and added them to the original prompt during inference. This approach significantly boosted the original ChatGPT-4o's performance, achieving the highest CIDEr score and high TFIDF-F1 and Senti-F1 score. However, the variants of LLaVA-Review show comparable accuracy across almost all metrics, confirming its high review generation performance.

4.3. Conditional Review Generation

Table 4 presents the quantitative results of conditioned review generation. LLaVA-1.5 shows minimal to no performance improvement when considering attributes, as reflected in its low CIDEr scores for different conditions, such as gender (CIDEr: 0.014) and rating (CIDEr: 0.012). In contrast, LLaVA-Review

Table 4
Conditioned Review Generation Results. The first group shows results of conditioning on LLaVA. The second group shows results of conditioning on LLaVA-Review. Light blue indicates user attribute conditions, green represents style-related information, and yellow indicates time-related conditions.

model	BLUE	ROUGE-1	ROUGE-L	CIDEr	DIV	PROPN	TFIDF-F1	Senti-F1	length
LLaVA-1.5 [16]	0.683	0.254	0.162	0.099	0.863	0.303	0.142	0.029	133.8
LLaVA-1.5 + gender	0.687	0.254	0.163	0.014	0.874	0.248	0.141	0.024	120.5
LLaVA-1.5 + season	0.627	0.255	0.163	0.012	0.869	0.253	0.139	0.024	120.5
LLaVA-1.5 + rating	0.679	0.252	0.161	0.012	0.877	0.261	0.143	0.024	122.4
LLaVA-1.5 + length	0.699	0.254	0.165	0.015	0.878	0.285	0.143	0.025	116.0
LLaVA-1.5 + profile (long)	0.597	0.244	0.156	0.013	0.876	0.258	0.131	0.021	120.9
LLaVA-1.5 + Keyphrase	2.699	0.287	0.184	0.030	0.885	0.312	0.152	0.058	124.3
LLaVA-Review	1.291	0.272	0.185	0.108	0.920	0.457	0.244	0.045	54.9
LLaVA-Review + gender	1.410	0.269	0.185	0.106	0.920	0.430	0.239	0.046	54.3
LLaVA-Review + age	1.161	0.268	0.186	0.104	0.920	0.421	0.239	0.049	52.5
LLaVA-Review + tag	1.195	0.276	0.187	0.110	0.919	0.425	**0.249**	**0.050**	56.2
LLaVA-Review + profile_tag	1.510	0.273	0.186	0.117	0.919	0.431	0.240	0.045	54.7
LLaVA-Review + profile_long	**1.673**	**0.279**	**0.189**	**0.123**	0.920	**0.485**	0.243	**0.050**	56.7
LLaVA-Review + rating	1.320	0.270	0.186	0.103	0.920	0.447	0.243	0.047	54.5
LLaVA-Review + length	1.952	0.308	0.198	0.184	**0.923**	**0.510**	**0.244**	0.048	87.3
LLaVA-Review + key phrase	**5.251**	**0.316**	**0.233**	**0.263**	0.922	0.425	0.197	**0.118**	50.8
LLaVA-Review + season	1.313	0.268	0.183	0.106	0.919	**0.447**	**0.242**	0.048	54.1
LLaVA-Review + month	**1.471**	**0.271**	**0.188**	**0.107**	0.919	0.426	0.240	**0.049**	53.5

exhibits significant accuracy gains for certain attributes. LLaVA-1.5 fails to capture the characteristics of reviews influenced by user attributes and context, often directly outputting the conditions in the reviews. Although conditioning with key phrases improves accuracy in LLaVA-1.5, it is less effective than in LLaVA-Review.

Due to differing attribute frequencies as shown in Figure 2, directly comparing the effects of conditioning is difficult, but certain trends in accuracy improvements are observed across different attributes. Among user attributes, gender significantly improves generation quality, while age and group information have little impact, suggesting that gender has a stronger influence on shaping individual user reviews compared to age or group composition. For profiles, the short tag profile did not significantly improve accuracy, but long-form and detailed profiles reflecting styles and preferences, such as "relaxed" or "preference for natural landscapes," enhanced review generation accuracy. Incorporating more detailed profiles could be a direction for future work. For style conditions such as review length and rating, review length proved especially effective, significantly improving accuracy. The average deviation between generated and target lengths was just 4.1 characters, reflecting precise conditioning. Low ratings had some effect, but imbalanced training data led to occasional under-representation of negative opinions. Keyword conditioning showed the highest accuracy gains. For temporal context, month-based conditioning outperformed seasonal, indicating greater effectiveness at a finer granularity.

Figure 7 illustrates examples of different conditioning: user demographic, age, user profile, rating and review lengths, and key phrases. For user demographic conditioning, characteristics of men, women, and families are reflected, such as mentioning "playground equipment" for families and "beautiful flowers" for women. For age conditioning, in the twenties conditioning, a lively writing style using exclamations was observed, while in the sixties conditioning, more factual reviews related to historical knowledge were produced. For review length, it was observed that even with the requirement of 290 characters, reviews were generated without content redundancy or structural inconsistencies. For rating, for the negative conditioning of one star, reviews expressing disappointment over the lack of menu options were generated. Generation based on keywords also demonstrated high fidelity, producing results that were very close to the ground truth.

---

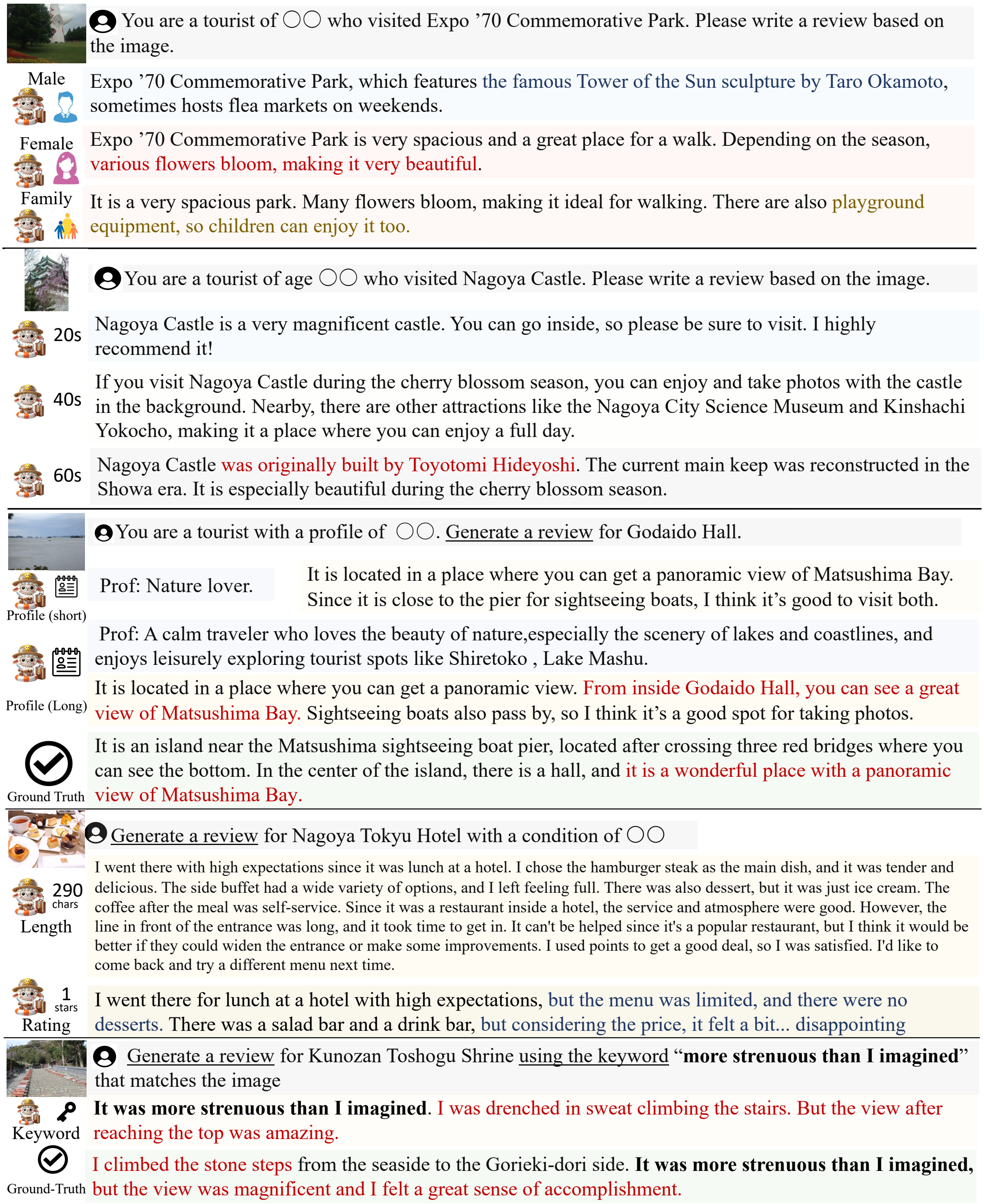
[15] https://cdn.jalan.jp/jalan/img/7/kuchikomi/2837/KL/d72d0_0002837701_1.webp

[16] https://cdn.jalan.jp/jalan/img/8/kuchikomi/1388/KL/e88ab_0001388176_2.webp

[17] https://cdn.jalan.jp/jalan/img/5/kuchikomi/4115/KL/09030_0004115266_1.webp

[18] https://cdn.jalan.jp/jalan/img/3/kuchikomi/5203/KL/02e5b_0005203413_1.webp

[19] https://cdn.jalan.jp/jalan/img/5/kuchikomi/4745/KL/be8e4_0004745614_4.webp



You are a tourist of ○○ who visited Expo ’70 Commemorative Park. Please write a review based on the image.

Male

Expo ’70 Commemorative Park, which features the famous Tower of the Sun sculpture by Taro Okamoto, sometimes hosts flea markets on weekends.

Female

Expo ’70 Commemorative Park is very spacious and a great place for a walk. Depending on the season, various flowers bloom, making it very beautiful.

Family

It is a very spacious park. Many flowers bloom, making it ideal for walking. There are also playground equipment, so children can enjoy it too.

You are a tourist of age ○○ who visited Nagoya Castle. Please write a review based on the image.

20s

Nagoya Castle is a very magnificent castle. You can go inside, so please be sure to visit. I highly recommend it!

40s

If you visit Nagoya Castle during the cherry blossom season, you can enjoy and take photos with the castle in the background. Nearby, there are other attractions like the Nagoya City Science Museum and Kinshachi Yokocho, making it a place where you can enjoy a full day.

60s

Nagoya Castle was originally built by Toyotomi Hideyoshi. The current main keep was reconstructed in the Showa era. It is especially beautiful during the cherry blossom season.

You are a tourist with a profile of ○○. Generate a review for Godaido Hall.

Profile (short)

Prof: Nature lover.

It is located in a place where you can get a panoramic view of Matsushima Bay. Since it is close to the pier for sightseeing boats, I think it’s good to visit both.

Profile (Long)

Prof: A calm traveler who loves the beauty of nature,especially the scenery of lakes and coastlines, and enjoys leisurely exploring tourist spots like Shiretoko , Lake Mashu.

It is located in a place where you can get a panoramic view. From inside Godaido Hall, you can see a great view of Matsushima Bay. Sightseeing boats also pass by, so I think it’s a good spot for taking photos.

Ground Truth

It is an island near the Matsushima sightseeing boat pier, located after crossing three red bridges where you can see the bottom. In the center of the island, there is a hall, and it is a wonderful place with a panoramic view of Matsushima Bay.

Generate a review for Nagoya Tokyu Hotel with a condition of ○○

290 chars
Length

I went there with high expectations since it was lunch at a hotel. I chose the hamburger steak as the main dish, and it was tender and delicious. The side buffet had a wide variety of options, and I left feeling full. There was also dessert, but it was just ice cream. The coffee after the meal was self-service. Since it was a restaurant inside a hotel, the service and atmosphere were good. However, the line in front of the entrance was long, and it took time to get in. It can't be helped since it's a popular restaurant, but I think it would be better if they could widen the entrance or make some improvements. I used points to get a good deal, so I was satisfied. I'd like to come back and try a different menu next time.

1 stars
Rating

I went there for lunch at a hotel with high expectations, but the menu was limited, and there were no desserts. There was a salad bar and a drink bar, but considering the price, it felt a bit... disappointing

Generate a review for Kunozan Toshogu Shrine using the keyword “**more strenuous than I imagined**” that matches the image

Keyword

It was more strenuous than I imagined. I was drenched in sweat climbing the stairs. But the view after reaching the top was amazing.

Ground-Truth

I climbed the stone steps from the seaside to the Gorieki-dori side. **It was more strenuous than I imagined,** but the view was magnificent and I felt a great sense of accomplishment.

Figure 7: Conditional review generation examples. The first example shows generation based on user demographics. The second example shows generation based on age. The third example illustrates generation based on the user’s profile. The fourth example shows generation based on rating and review length. The fifth example illustrates generation based on key phrases.[15] [16] [17] [18] [19]

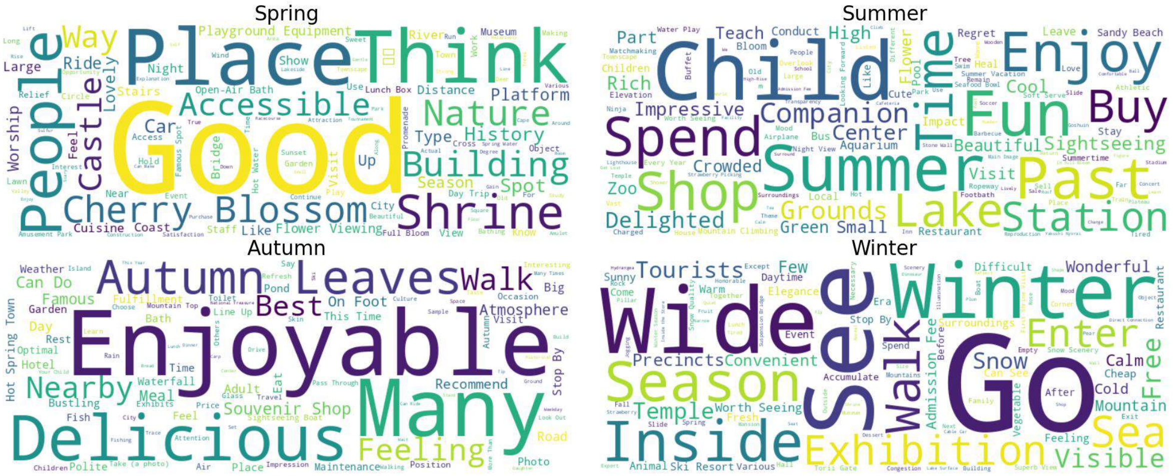


Figure 8: Word Cloud of Season Conditioned Review Generation, showing reflections such as cherry blossoms in spring, children in summer, and fall leaves in autumn.

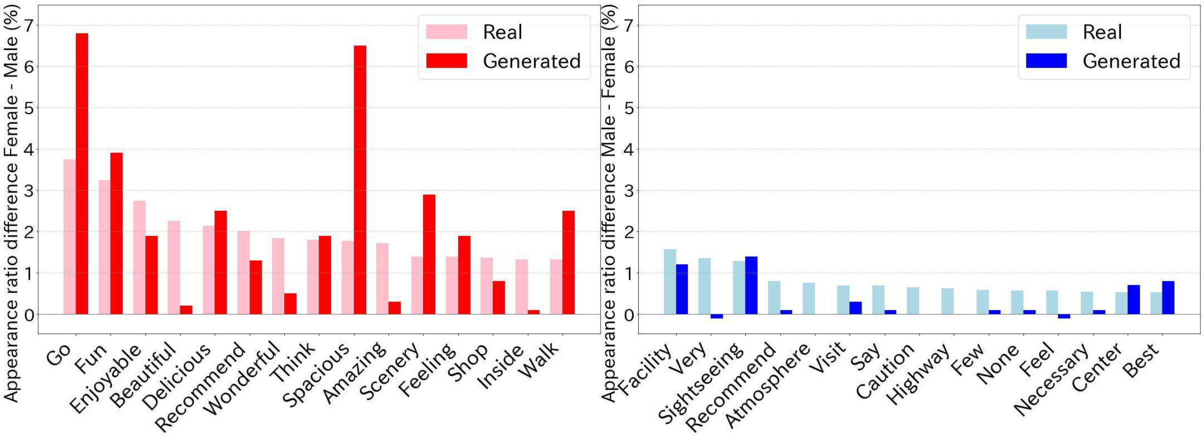


Figure 9: Frequency differences in generated reviews for the top 15 words with notable gender-based frequency differences in the original reviews. Left: Words with higher frequency for females, Right: Words with higher frequency for males.

As an aggregate analysis, we generated reviews for the same image by changing only the seasonal instructions across the four seasons, then compared word frequency in the output. As shown in Figure 8, distinct features appear: cherry blossoms in spring, children in summer, autumn leaves in fall, and hot springs in winter, highlighting the impact of seasonal conditioning. Figure 9 provides a quantitative view of the conditioning effects under gender conditioning. In generated reviews, we compared word frequencies between male and female conditioned reviews using 1,000 test image-text pairs. For real reviews, we sampled ten reviews from men and three from women for the same tourist destination. The table highlights the top 15 words with the largest (female - male) and (male - female) frequency differences in real reviews. In reviews originally written by women, certain words appear more frequently, and this pattern is reflected in generated reviews. All 15 words with higher frequency in women's original reviews also show higher frequency under female conditioning, with notable differences for “Go,” “Beautiful,” and “Spacious.” In contrast, men's reviews show fewer specific words, and the male-to-female ratio remains small in generated reviews, maintaining the overall trend of higher male-conditioned word frequency.

Table 5
Results of Short Review Generation.

Model	BLUE	ROUGE-1	ROUGE-L	CIDEr	length
LLaVA-Review	**1.172**	**0.215**	**0.190**	0.146	19.1
ChatGPT-4o	0.596	0.167	0.144	**0.170**	16.3

Table 6
Long Review Generation Results with and without Short Reivew Generation.

Model	BLUE	ROUGE-1	ROUGE-L	CIDEr	DIV	PROPN	TFIDF-F1	Senti-F1	length
LLaVA-Review w/o Short Review	1.291	0.272	0.185	0.108	0.920	0.457	0.244	0.045	54.9
LLaVA-Review w Short Review	1.310	0.269	0.183	0.104	0.921	0.447	0.248	0.046	55.6

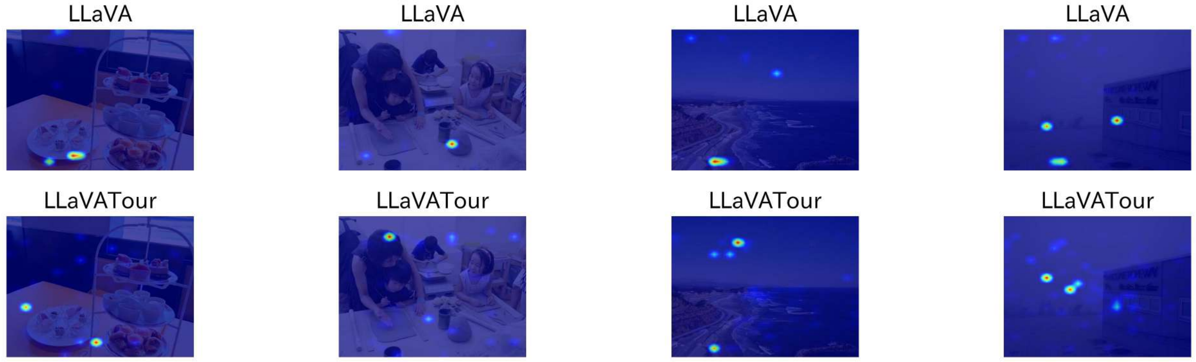


Figure 10: Attention visualization for image patches. The first row shows visualization results of LLaVA-1.5, and the second row refers to outputs of the LLaVA-Review. [20] [21] [22] [23]

4.4. Short Review Generation

Figure 5 presents the results of Short Review Generation. The ground truth, similar to General Review Generation, is a sentence retrieved by CLIP for each image, ensuring no overlap with the training data. A comparison between ChatGPT-4o and LLaVA-Review reveals that both models generate concise sentences of around 20 characters, with LLaVA-Review achieving higher ROUGE and BLEU scores. Additionally, an evaluation of the effect of fine-tuning with and without Short Review Generation on General Review Generation performance, as shown in Table 6, indicated that all metrics changed by no more than 4%, suggesting minimal impact.

4.5. Visualization of Attention

The visualization results of the attention weights in the final layer of the language model during the generation of the first token in the review generation process are shown in Figure 10. Specifically, the visualization highlights the 576 tokens that correspond to image positions within the cs-token attention, resized to 24×24 and presented as a heatmap. In the standard LLaVA-1.5 model, high attention values are concentrated in narrow regions of the image. In contrast, LLaVA-Review exhibits high attention values over a broader area of the image and across more extensive objects.

---

[20] https://cdn.jalan.jp/jalan/img/1/kuchikomi/4251/KL/2c80c_0004251538_1.webp
[21] https://cdn.jalan.jp/jalan/img/8/kuchikomi/4228/KL/2c27e_0004228816_3.webp
[22] https://cdn.jalan.jp/jalan/img/1/kuchikomi/0011/KL/c5437_0000011649.webp
[23] https://cdn.jalan.jp/jalan/img/1/kuchikomi/0821/KL/d60dd_0000821087_2.webp

5. Limitations and Future Works

In terms of dataset, we are considering updates to improve the fidelity of image-text pair creation. Leveraging more detailed user information, such as utilizing all past reviews written by the user, or considering their actual behavior at tourist destinations when possible, is also a promising direction. Additionally, while this research developed a dataset specific to Japanese data, it is known that tourist destinations and perceptions of tourism vary by country. Building a more comprehensive dataset that encompasses diverse languages and cultures remains a challenge for future research. Moreover, expanding the dataset to support more tasks such as personalized product description generations and recommendations are also future works. In terms of the model design, we will leverage more powerful language models as the backbone, incorporate more robust image information and domain knowledge, and develop more adaptive external documents. We will also consider the application to real-world scenarios, such as simulations and marketing. Our model is capable of generating virtual user experiences, which could be used to improve tourist destinations and simulate travel experiences. However, when utilizing such pseudo-reviews, it is essential to address potential issues related to privacy, bias, and reputation harms that these reviews might cause.

6. Conclusion

In this research, we developed TourMix1M, the first multimodal dataset for tourism review generation. We also introduced LLaVA-Review, a large-scale multimodal model for review generation. Furthermore, we researched two knowledge retrieval methods for tourism review generation. Experiments with the proposed dataset showed LLaVA-Review's superior performance in domain specificity and user sentiment expression. The proposed two retrieval-augmented fine-tuning strategies further improved accuracy. Additionally, additional attention to factors such as gender, user profiles, month, review length, and key phrases significantly enhanced review generation. This work is expected to advance research in tourism and broader review generation fields.

7. Acknowledgments

This work is partially financially supported by Center for Real Estate Innovation, The University of Tokyo.

References

[1] WTTC, Economic Impact Report: Global Infographic, Technical Report, World Travel Tourism Council (WTTC), 2023. URL: https://wttc.org/Research/Economic-Impact.

[2] Anonymous, Tourist experiences at overcrowded attractions: A text analytics approach, in: Information and Communication Technologies in Tourism, 2022, pp. 231–243.

[3] S.-E. Kim, K. Y. Lee, S. I. Shin, S.-B. Yang, Effects of tourism information quality in social media on destination image formation: The case of sina weibo, Information & Management 54 (2017) 687–702.

[4] M. del Carmen Hidalgo Alcázar, M. S. Piñero, S. R. de Maya, The effect of user-generated content on tourist behavior: The mediating role of destination image, Tourism & Management Studies 10 (2014) 158–164.

[5] Y. L. . H. S. Dogan Gursoy, Gender difference on destination image and travel options: An exploratory text-mining study., in: PloS one, volume 30, 2018, pp. 1–5.

[6] Y. L. . H. S. Dogan Gursoy, Does traveler satisfaction differ in various travel group compositions?, in: International Journal of Contemporary Hospitality Management, volume 30, 2018, pp. 1663–1685.

[7] J. J. Padilla1, H. Kavak, C. J. Lynch, R. J. Gore1, S. Y. Diallo, Temporal and spatiotemporal investigation of tourist attraction visit sentiment on twitter., PloS one 13 (2018).

[8] M. Rossetti1, F. Stella1, M. Zanker, Analyzing user reviews in tourism with topic models., Information Technology Tourism 316 (2016) 5–21.

[9] S. A. C. Estela Marine-Roig, Tourism analytics with massive user-generated content: A case study of barcelona., Journal of Destination Marketing Management 4 (2015) 162–172.

[10] C.-F. Tsai, K. Chen, Y.-H. Hu, W.-K. Chen, Improving text summarization of online hotel reviews with review helpfulness and sentiment, Tourism Management 80 (2020).

[11] L. Dong, S. Huang, F. Wei, M. Lapata, M. Zhou, K. Xu, Learning to generate product reviews from attributes, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017, pp. 623–632.

[12] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[13] Q.-T. Truong, H. Lauw, Multimodal review generation for recommender systems, in: The World Wide Web Conference, 2019, pp. 1864–1874.

[14] J. Li, S. Li, W. X. Zhao, G. He, Z. Wei, N. J. Yuan, J.-R. Wen., Knowledge-enhanced personalized review generation with capsule graph neural network, in: Proceedings of the 29th ACM International Conference on Information Knowledge Management, 2020, pp. 735–744.

[15] Z. Xie, S. Singh, J. McAuley, Bodhisattwa, P. Majumder, Factual and informative review generation for explainable recommendation., in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 13816–13824.

[16] H. Liu, C. Li, Y. Li, Y. J. Lee, Improved baselines with visual instruction tuning, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 26296–26306.

[17] X. V. Lin, X. Chen, M. Chen, W. Shi, M. Lomeli, R. James, P. Rodriguez, J. Kahn, G. Szilvasy, M. Lewis, L. Zettlemoyer, S. Yih, Ra-dit: Retrieval-augmented dual instruction tuning, in: The Twelfth International Conference on Learning Representations., 2024, pp. 2206–2240.

[18] T. Zhang, N. J. Shishir G. Patil, S. Shen, M. Zaharia, I. Stoica, J. E. Gonzalez, Raft: Adapting language model to domain specific rag, in: arXiv preprint, 2024.

[19] M. Yang, M. Zhu1, Y. Wang, L. Chen, Y. Zhao, X. Wang, B. Han, X. Zheng, J. Yin, Fine-tuning large language model based explainable recommendation with explainable quality reward., in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 9250–9259.

[20] S.-J. Park, D.-K. Chae, H.-K. Bae, S. Park, S.-W. Kim, Reinforcement learning over sentiment-augmented knowledge graphs towards accurate and explainable recommendation, in: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 2022, pp. 784–793.

[21] Anonymous, Llava-tour: Creation of a large-scale multimodal model specializing in japanese tourism data., in: Proceedings of IEEE Visual Communications and Image Processing (under review), 2024.

[22] M. L. Cheung, W. K. S. Leung, J.-H. Cheah, H. Ting, Effects of user-provided photos on hotel review helpfulness: An analytical approach with deep leaning, International Journal of Hospitality Management 71 (2018) 120–131.

[23] M. L. Cheung, W. K. S. Leung, J.-H. Cheah, H. Ting, Exploring the effectiveness of emotional and rational user-generated contents in digital tourism platforms, Vacation Marketing 28 (2022) 152–170.

[24] Y. L. . H. S. Dogan Gursoy, Chatgpt and the hospitality and tourism industry: an overview of current trends and future research directions, in: Journal of Hospitality Marketing & Management, volume 32, 2023, pp. 579–592.

[25] W. C. Yogesh K Dwivedi, Neeraj Pandey, A. Micu, Leveraging chatgpt and other generative artificial intelligence (ai)-based applications in the hospitality and tourism industry: practices, challenges and research agenda, in: International Journal of Contemporary Hospitality Management, volume 36, 2024, pp. 1–12.

[26] C. Zong, F. Xia, W. Li, R. Navigli, Personalized transformer for explainable recommendation, in: Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021, pp. 4947–4957.

[27] L. Li, Y. Zhang, L. Chen, Personalized prompt learning for explainable recommendation, in: ACM

Transactions on Information Systems, volume 41, 2023, pp. 1–26.

[28] P. Li, Z. Wang, Z. Ren, L. Bing, W. Lam, Neural rating regression with abstractive tips generation for recommendation, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 345–354.

[29] J. Ni, J. McAuley, Personalized review generation by expanding phrases and attending on aspect-aware representations, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 706–711.

[30] Z. C. Lipton, S. V. andand Julian McAuley, Generative concatenative nets jointly learn to write and classify reviews, in: arxiv preprint, 2015.

[31] P. Li, A. Tuzhilin, Towards controllable and personalized review generation, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, volume 1, 2019, pp. 3237–3245.

[32] J. Guo, S. Lu, H. Cai, W. Zhang, Y. Yu, J. Wang, Long text generation via adversarial training with leaked information., in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2019.

[33] D. V. Hada, V. M., S. K. Shevade, Rexplug: Explainable recommendation using plug and play language model, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, volume 1, 2021, pp. 81–91.

[34] H. Chen, Y. Lin, F. Qi, J. Hu, P. Li, J. Zhou, M. Sun, Aspect-level sentiment-controllable review generation with mutual learning framework, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 12639–12647.

[35] J. Li, W. X. Zhao, J.-R. Wen, , Y. Song, Generating long and informative reviews with aspect-aware coarse-to-fine decoding, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1969–1979.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, volume 30, 2017.

[37] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, et al., Llama: Open and efficient foundation language models, arXiv preprint (2023).

[38] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models are Unsupervised Multitask Learners, Technical Report, OpenAI, 2019. URL: https://api.semanticscholar.org/CorpusID:160025533.

[39] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. N. P., et al., Language models are few-shot learners, in: Proc. Advances in Neural Information Processing Systems, volume 33, 2020, pp. 1877–1901.

[40] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, in: Proc. Advances in Neural Information Processing Systems, volume 36, 2023.

[41] OpenAI, Gpt-4 technical report, arXiv preprint (2023).

[42] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, J. Zhou, Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, arxiv preprint (2023).

[43] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, et al., Finetuned language models are zero-shot learners, in: Proc. International Conference on Learning Representations, 2022.

[44] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, Q. Liu, Ernie: Enhanced language representation with informative entities, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1441–1451.

[45] M. Kang, J. M. Kwak, J. Baek, S. J. Hwang, Knowledge graph-augmented language models for knowledge-grounded dialogue generation, in: arxiv preprint, 2023.

[46] X. He, Y. Tian, Y. Sun, N. V. Chawla, T. Laurent, Y. LeCun, X. Bresson, B. Hooi, G-retriever: Retrieval-augmented generation for textual graph understanding and question answering, in: arxiv preprint, 2024.

[47] L. Luo, Y.-F. Li, G. Haffari, S. Pan, Reasoning on graphs: Faithful and interpretable large language

model reasoning, in: International Conference on Learning Representations, 2024.

[48] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. van den Driessche, J.-B. Lespiau, B. Damoc, , A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. W. Rae, E. Elsen, L. Sifre, Improving language models by retrieving from trillions of tokens., in: International Conference on Machine Learning, 2022, pp. 2206–2240.

[49] K. Guu, K. Lee, Z. Tung, P. Pasupat, M.-W. Chang, Realm: Retrieval-augmented language model pre-training, in: International Conference on Machine Learning, 2020, pp. 3929–3938.

[50] W. Shi, S. Min, M. Yasunaga, M. Seo, R. James, M. Lewis, L. Zettlemoyer, W. tau Yih, Replug: Retrieval-augmented black-box language models, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2024, pp. 8371–8384.

[51] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in: Advances in Neural Information Processing Systems, 2024, pp. 9459–9474.

[52] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, 2021, pp. 8748–8763.

[53] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing, Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. URL: https://lmsys.org/blog/2023-03-30-vicuna/.

[54] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, in: Proc. International Conference on Learning Representations, 2022.

[55] R. K. Amplayo, S. Angelidis, M. Lapata, Aspect-controllable opinion summarization, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, p. 6578–6593.

[56] K. Papineni, S. Roukos, T. Ward, W. Zhu, Bleu: A method for automatic evaluation of machine translation, in: Proc. 40th Annual Meeting on Association for Computational Linguistics, 2002, p. 311–318.

[57] C.-Y. LIN, Rouge: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81.

[58] R. Vedantam, C. L. Zitnick, D. Parikh, Cider: Consensus-based image description evaluation., in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2015, pp. 4566–4575.

[59] B. Smyth, P. McClave., Similarity vs. diversity., in: Proceedings of the 4th International Conference on Case-Based Reasoning, 2001, pp. 347–361.

[60] H. Peng, L. Xu, L. Bing, F. Huang, W. Lu, L. Si, Knowing what, how and why: A near complete solution for aspect-based sentiment analysis,, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 8600–8607.

[61] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

[62] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.

[63] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[64] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometrics and intelligent laboratory systems (1987) 37–52.

[65] S. P. Lloyf, Least squares quantization in pcm, IEEE transactions on information theory, 28 (1982) 129–137.