A Multimodal Dataset and Benchmark for Tourism Review Generation

Hiromasa Yamanishi, Ling Xiao, Toshihiko Yamasaki

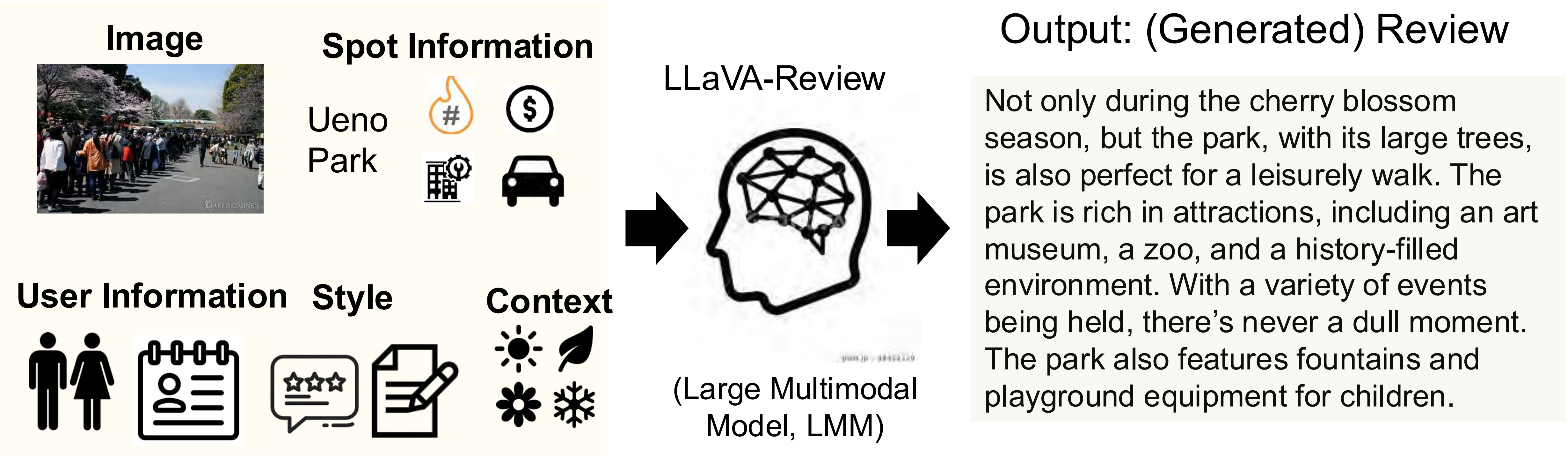
The University of Tokyo

UTokyo

RecTour2024

Overview

Input: Multimodal Information



Image

Spot Information

Ueno Park

User Information

Style

Context

LLaVA-Review

(Large Multimodal Model, LMM)

Output: (Generated) Review

Not only during the cherry blossom season, but the park, with its large trees, is also perfect for a leisurely walk. The park is rich in attractions, including an art museum, a zoo, and a history-filled environment. With a variety of events being held, there’s never a dull moment. The park also features fountains and playground equipment for children.

TourMix1M dataset

Github: HiromasaYamanishi/TourMix1M

Background: Review Generation

- Task
  - Generate informative and structural review based on user/item Information, rating

- Large Application Field

User Assistance

- Only few users write reviews
- **Hard to write informative and structured reviews.**

Marketing

- User demand is getting **diverse**
- Can be used as **recommendation text**

Simulation

- Predict the reaction of the unpopular segment
- Predict the reaction to **events, facility**

Motivations

- Creation of tourism multimodal dataset
  - Previous works is based on **user information, item information, rating**
  - Previous dataset domain is **book, movie, restaurant, beer** etc
  - **Lacks tourism-specific multimodal context** such as image, user attribute, visit time

→ We create **tourism specific large scale (1M) dataset**

- Generate the baseline model with open-source LMM
  - LMM have high instruction-following and multimodal understanding
  - ChatGPT is costly for generating review

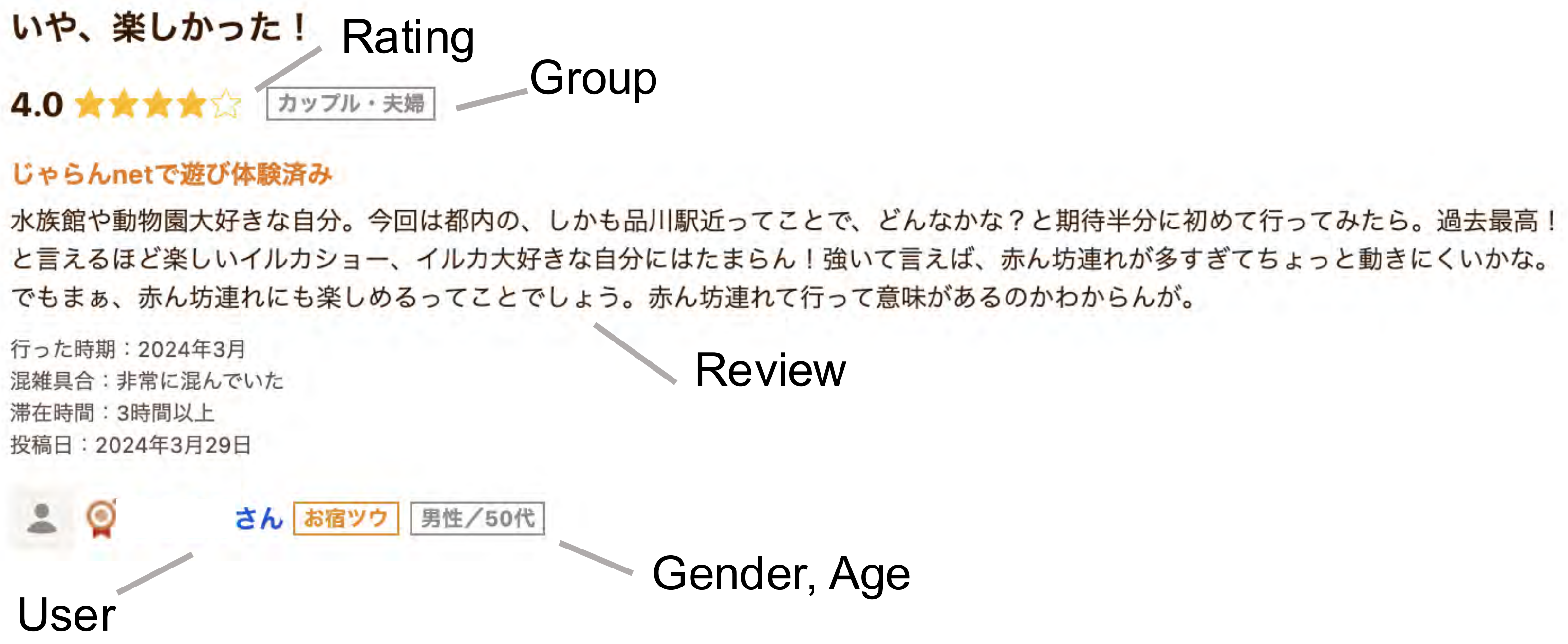
→ We explore **open-source LMM + retrieval augmented fine tuning**

ChatGPT

LLaVA
[NeurIPS23]

Data Collection

Review Sample



いや、楽しかった！

4.0 ★★★★☆ カップル・夫婦

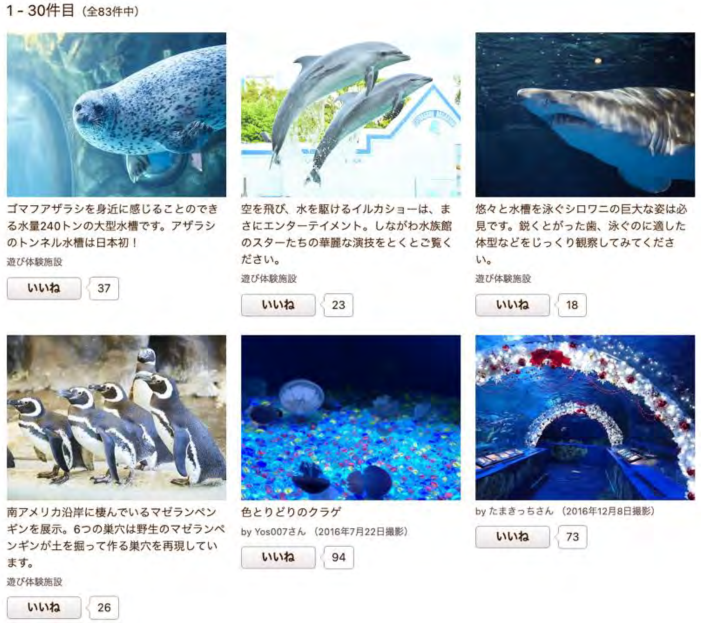
じゃらんnetで遊び体験済み

水族館や動物園大好きな自分。今回は都内の、しかも品川駅近ってことで、どんなかな？と期待半分に初めて行ってみたら。過去最高！と言えるほど楽しいイルカショー、イルカ大好きな自分にはたまらん！強いて言えば、赤ん坊連れが多すぎてちょっと動きにくいかな。でもまぁ、赤ん坊連れにも楽しめるってことでしょう。赤ん坊連れて行って意味があるのかわからんが。

行った時期：2024年3月
混雑具合：非常に混んでいた
滞在時間：3時間以上
投稿日：2024年3月29日

さん お宿ツウ 男性／50代

Image sample



1 - 30件目（全83件中）

ゴマフアザラシを身近に感じることのできる水量240トンの大型水槽です。アザラシのトンネル水槽は日本初！
遊び体験施設
37

空を飛び、水を駆けるイルカショーは、まさにエンターテイメント。しながわ水族館のスターたちの華麗な演技をとくとご覧ください。
遊び体験施設
23

悠々と水槽を泳ぐシロワニの巨大な姿は必見です。鋭くとがった歯、泳ぐのに適した体型などをじっくり観察してみてください。
遊び体験施設
18

南アメリカ沿岸に棲んでいるマゼランペンギンを展示。6つの巣穴は野生のマゼランペンギンが土を掘って作る巣穴を再現しています。
遊び体験施設
26

色とりどりのクラゲ
by Yos007さん（2016年7月22日撮影）
94

by たまきっちさん（2016年12月8日撮影）
73

- The data was collected from **Japanese tourism website** Jalan.net, consisting of **470k images, 540k reviews, and 50,000 tourist spots**
- **10 types of context**: gender, age, group, season, month, rating, two types of user profiles (short and long), review length, and keywords
- Image-review pairs are created with CLIP [ICLR21] in same tourism spot

TourMix1M Dataset: Tasks and Dataset Statistics

☐ Prompt Examples

Short Review Generation

You are a tourist who visits sensoji. Based on the given image, please write a **concise** review.

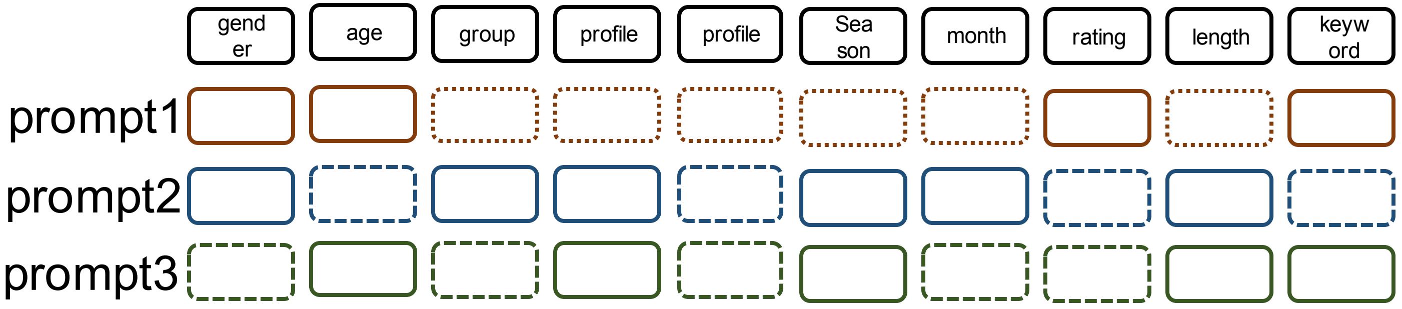
General Review Generation

You are a tourist who visits sensoji. Based on the given image, please write a review.

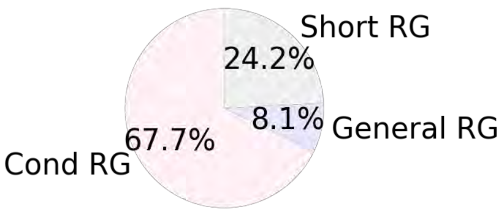
Conditional Review Generation

You are a tourist visiting Jufuku-ji with friends. You are interested in historical buildings and world heritage sites, particularly enjoying visiting places like Nikko and temples or shrines. You also like capturing Japan's traditional culture and landscapes in photos. Please create a 3-star review based on the input image.

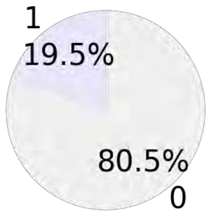
☐ Attribute sampling



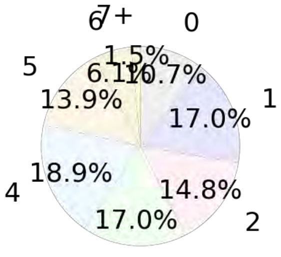
☐ Dataset Statistics



Task Distribution (Total: 1,318,813)



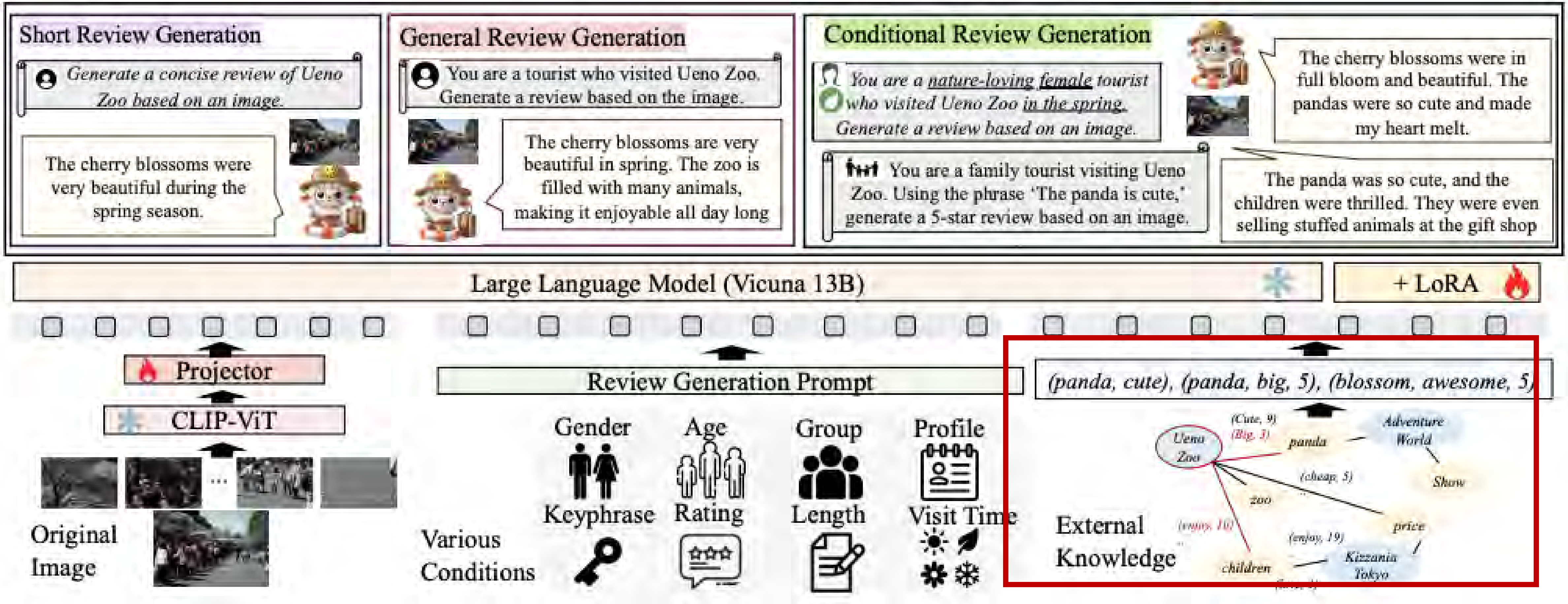
Short RG Condition Number Distribution (Total: 318,813)



Long RG Condition Number Distribution (Total: 1,000,000)

- **Three tasks**, Short/General/Conditional review generation, **1.3M prompts** in total.
- For Conditional RG, attributes are **sampled to increase diversity**
  - ➢ The ratio remains near to original proportion after sampling

Proposed LLaVA-Review model



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

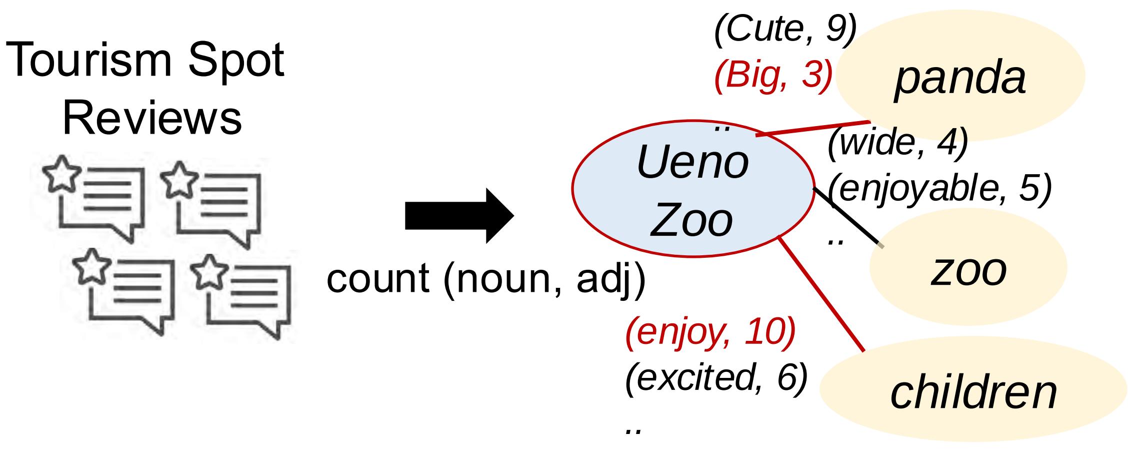
(panda, cute), (panda, big, 5), (blossom, awesome, 5)

External Knowledge

- Created a model based on **LLaVA, an open-source LMM**. Vicuna 13B for LLM
- **Retrieval-Augmented Fine-Tuning** by adding aggregated information

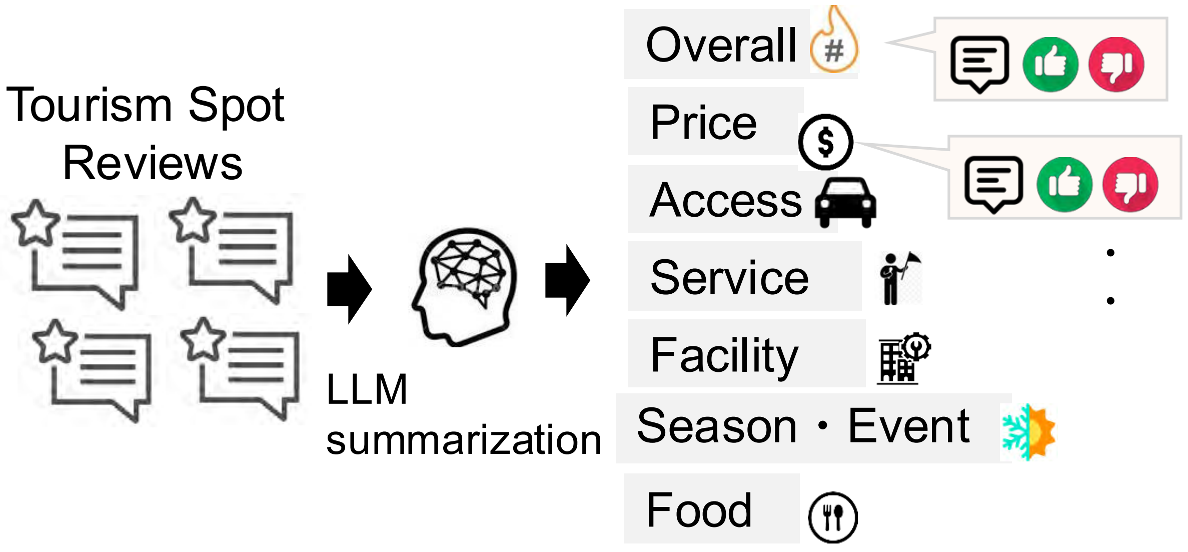
Retrieval Augmentation Knowledge

☐ Subgraph of sentiment-aware KG



- Sampled **noun and adjective** according to frequency

☐ Aspect-based summary

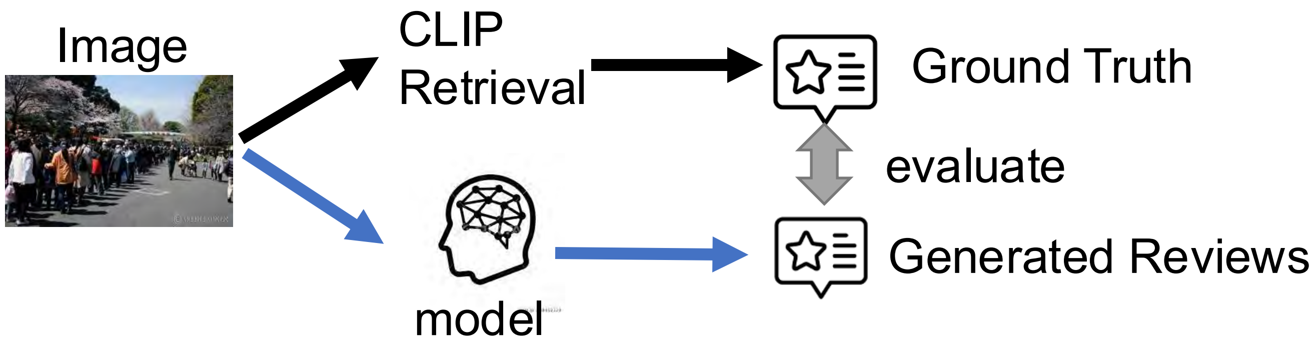


- For each **aspect**, summarize the **overview, positive, negative**
- Gemma2-9B for summarization

Experimental Settings

- Evaluation Setup
  - Evaluation Data: 1,000 pairs of image-review combinations not included in training
  - Evaluation Metrics: BLEU, ROUGE1, ROUGEL, CIDEr (quality assessment), DIV, PROPN (diversity), TFIDF-F1 (domain knowledge), Senti-F1 (user opinion)
- Training Setup
  - GPUs Used: Ada RTX A6000 x 8 , Training Time: 37 hours (1 epoch), 45-60 hours (RAFT, 1 epoch)
- Experiments

Experiment	Compared Model	Prompt
General Review Generation	PETER [Li+ACL21], PEPLER [Li+TOIS23], MRG [Truong+WWW19], LLaVA1.5 [Liu+CVPR24], ChatGPT-4v,4o	Spot name + Image
Conditional Review Generation (gender, age…)	LLaVA1.5 [Liu+CVPR24]	Spot Name + Image + each conditioning



Image

CLIP Retrieval

Ground Truth

evaluate

model

Generated Reviews

Quality

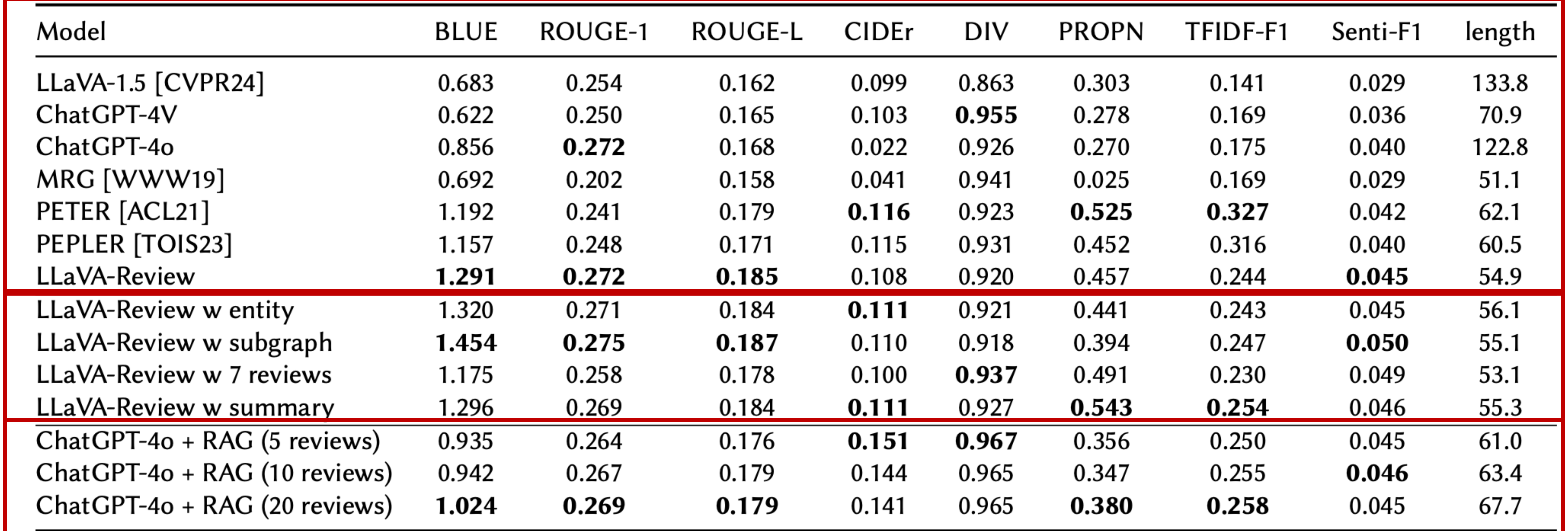
Diversity

Domain Knowledge

User Opinion

Experiments: General Review Generation

- Quantitative Results



Model	BLUE	ROUGE-1	ROUGE-L	CIDEr	DIV	PROPN	TFIDF-F1	Senti-F1	length
LLaVA-1.5 [CVPR24]	0.683	0.254	0.162	0.099	0.863	0.303	0.141	0.029	133.8
ChatGPT-4V	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 [WWW19]	0.692	0.202	0.158	0.041	0.941	0.025	0.169	0.029	51.1
PETER [ACL21]	1.192	0.241	0.179	**0.116**	0.923	**0.525**	**0.327**	0.042	62.1
PEPLER [TOIS23]	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

- Standard instruction-tuning achieves better performance than other baselines
- Retrieval-Augmented Fine-Tuning using subgraphs and summary improves performance
- While ChatGPT + RAG has high accuracy, LLaVA-Review outperforms in BLEU and ROUGE scores

Experiments: General Review Generation

- ☐ Qualitative Results

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

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

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

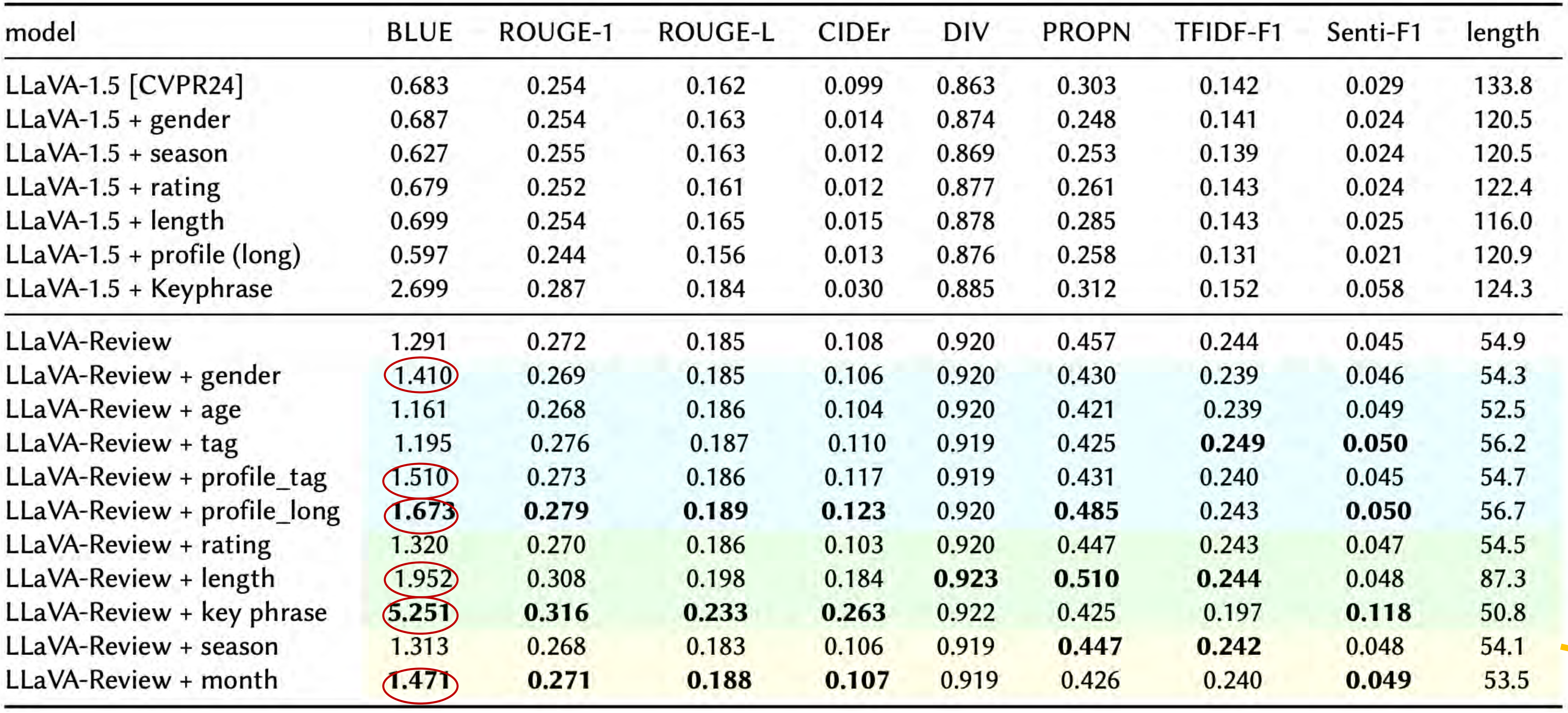
The Isawa Onsen area is a hot spring destination located in the southern part of Yamanashi Prefecture. **The hotel where I stayed also had a natural hot spring on the premises. It is said that this hot spring has the characteristic of making your skin smooth.**

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

- ■ LMMs lacks detailed information. Fine-tune models contain detailed information
- ■ LLaVA-Review captures image information and the characteristics of the hot springs

Experiments: Conditional Review Generation



model	BLUE	ROUGE-1	ROUGE-L	CIDEr	DIV	PROPN	TFIDF-F1	Senti-F1	length
LLaVA-1.5 [CVPR24]	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

User-Base Conditioning

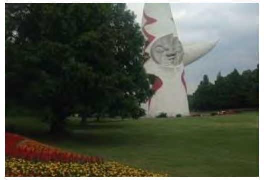
Style-Base Conditioning

Time-Base Conditioning

- Conditioning is effective in LLaVA-Review, while not in LLaVA-1.5
- Specifying gender, visit month, profile, key phrases, and review length is effective
- The error in review length remaining within approximately ±4.

Experiments: Conditional Review Generation

- ☐ Result for attribute-base generation



You are a tourist from ○○ 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.

Experiments: Conditional Review Generation

- Result for user-profile base generation

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

Prof: Nature lover.

Profile (short)

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.

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.

Profile (Long)

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.

Experiments: Conditional Review Generation

☐ Result for length & rating conditioned generation

Generate a review for Nagoya Tokyu Hotel with a condition of 〇〇

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

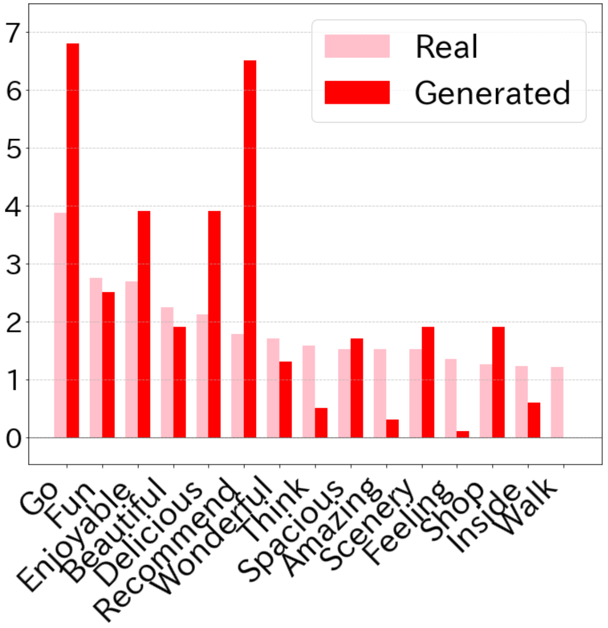
Rating

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

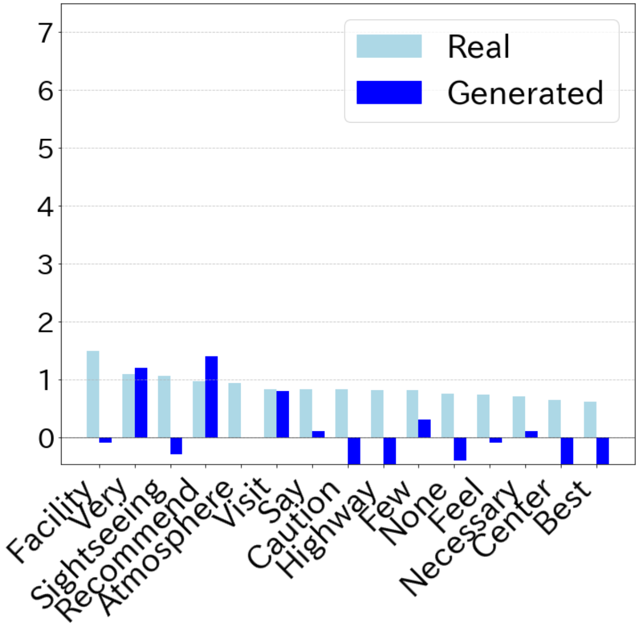
Aggregated evaluation of conditioned generation

☐ Conditioning based on gender

Words more appear in female

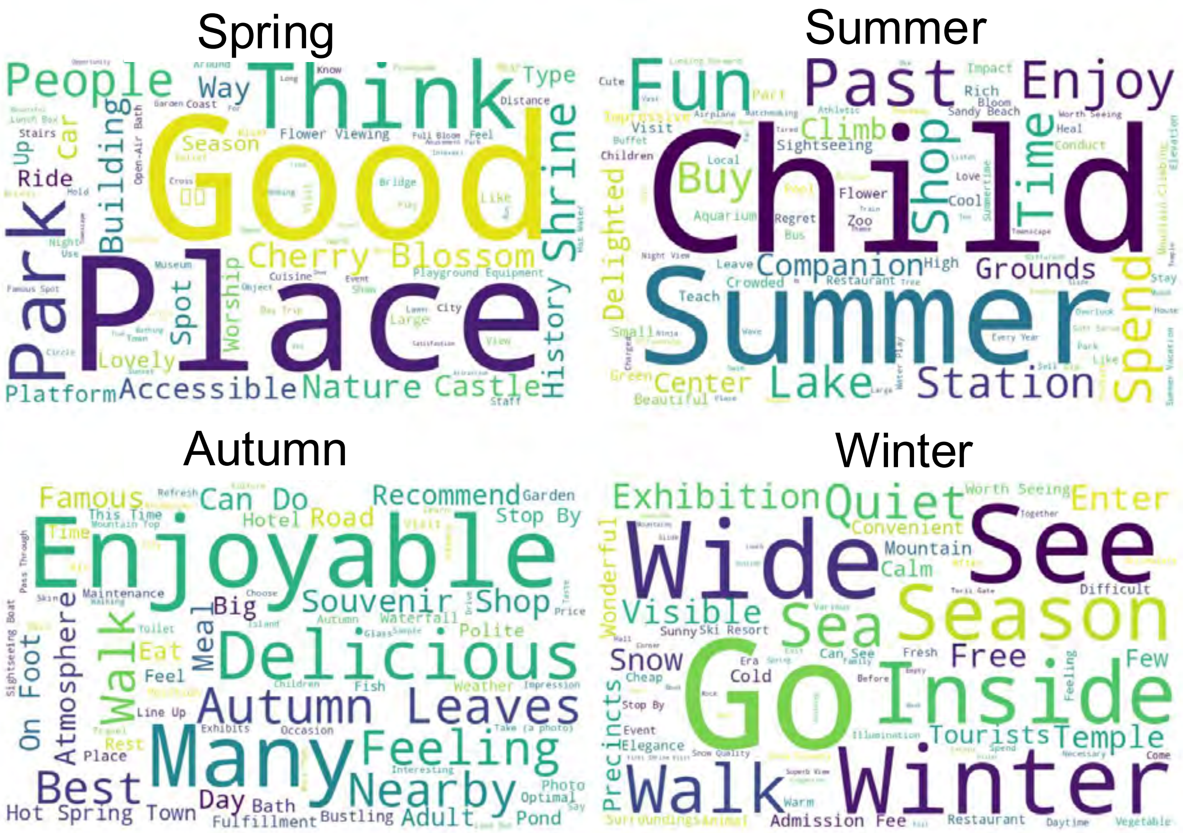


Words more appear in male



- Words like **"Spacious," "Scenery," and "Delicious" appear frequently among women**, both in generated and actual data

☐ Conditioning based on seasons



- Cherry blossoms in spring, children in summer, autumn leaves in fall, and snow in winter can be observed

Conclusion

■Summary

- Creation of the **TourMix1M dataset**, a diverse multimodal tourism dataset
- Confirmed the **effectiveness of LLaVA + RAFT (LLaVA-Review)** as a baseline model
- Verified the **effectiveness of conditioning** and identified **important attributes**

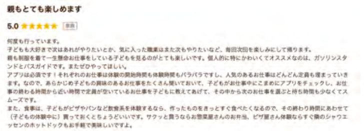
■Future Work

- ***Dataset:*** include more countries and domains, multimodal information like user behaviors
- ***Architecture:*** longer and structural review generation
- ***Application:*** proposals for tourism site improvements using a multi-agent system

global

domain

long, informative review



親もとても楽しめます

5.0 ★★★★★

何度も行っています。
子どもも大好きで次はあれがやりたいとか、気に入った職業はまた次もやりたいなど、毎回次回を楽しみにして帰ります。
親も制服を着て一生懸命お仕事をしている子どもを見るのがとても楽しいです。個人的に特にかわいくてオススメなのは、ガソリンスタンドとバスガイドです。またぜひやってほしい。
アプリは必須です！それぞれのお仕事は体験の開始時間も体験時間もバラバラですし、人気のあるお仕事はどんどん定員も埋まっていきます。なので、あらかじめ子どもの興味のあるお仕事をたくさん聞いておいて、子どもがお仕事中にこまめにアプリをチェックし、お仕事の終わる時間から近い時間で定員が空いているお仕事を子どもに教えてあげて、その中から次のお仕事を選ぶと待ち時間も少なくてスムーズです。
また、食事は、子どもがピザやパンなど飲食系を体験するなら、作ったものをきっとすぐ食べたくなるので、その終わり時間にあわせて（子どもの体験中に）買っておくとちょうどいいです。サクッと買うならお惣菜屋さんのお弁当、ピザ屋さん体験ならすぐ隣のシャウエッセンのホットドックもお手軽で美味しいですよ。

description

simulation

Appendix: Attribute ratio

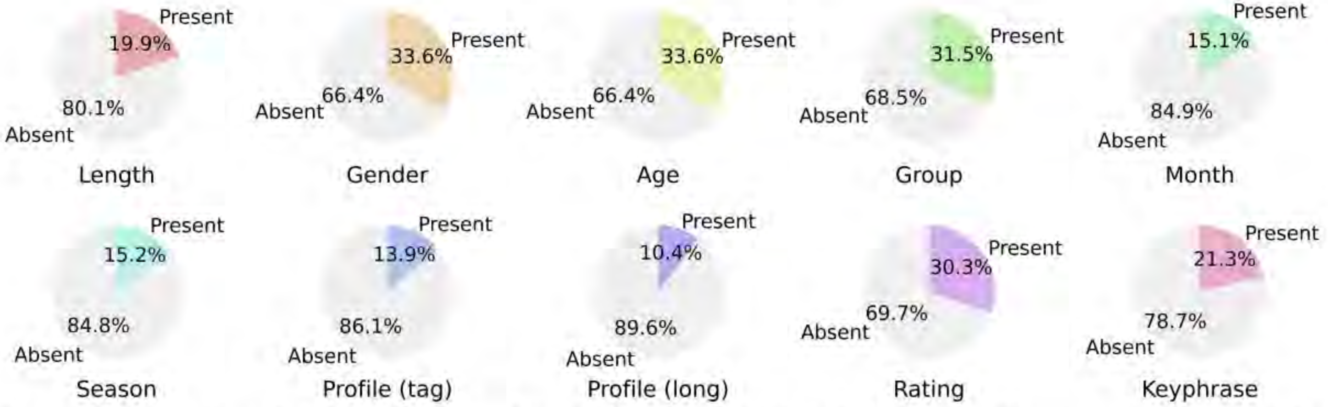


Figure 2: Proportion of each of the 10 conditioning types within the whole dataset. We constructed prompts by sampling attributes to increase the diversity of conditioning.

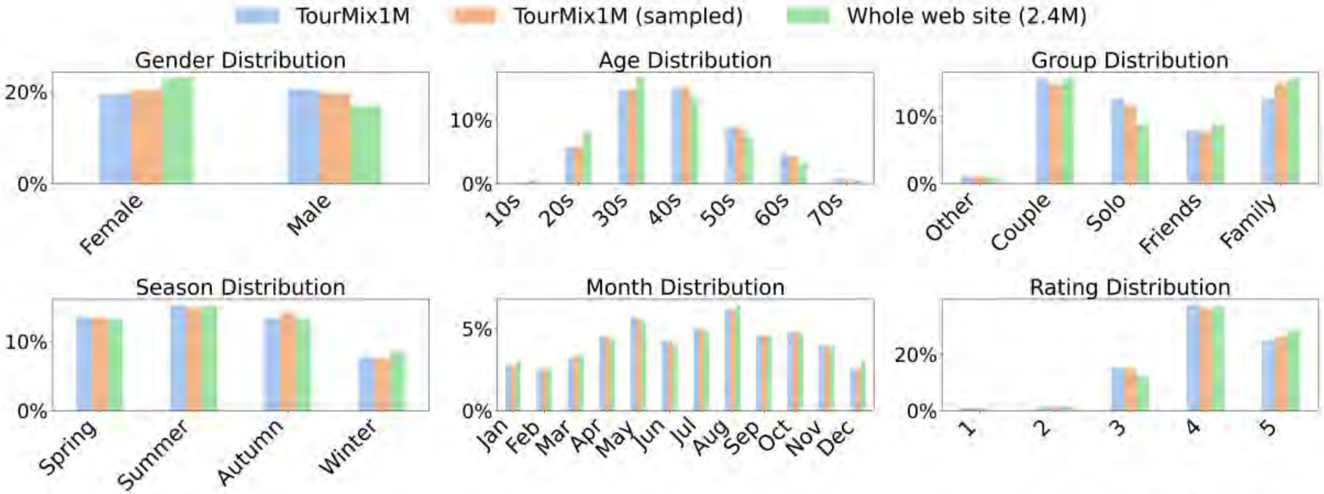


Figure 3: Comparison of categorical attribute distribution in: 1) the TourMix1M dataset, and 2) the sampled TourMix1M dataset 3) the original web data, .

Appendix: Image-Review pair example

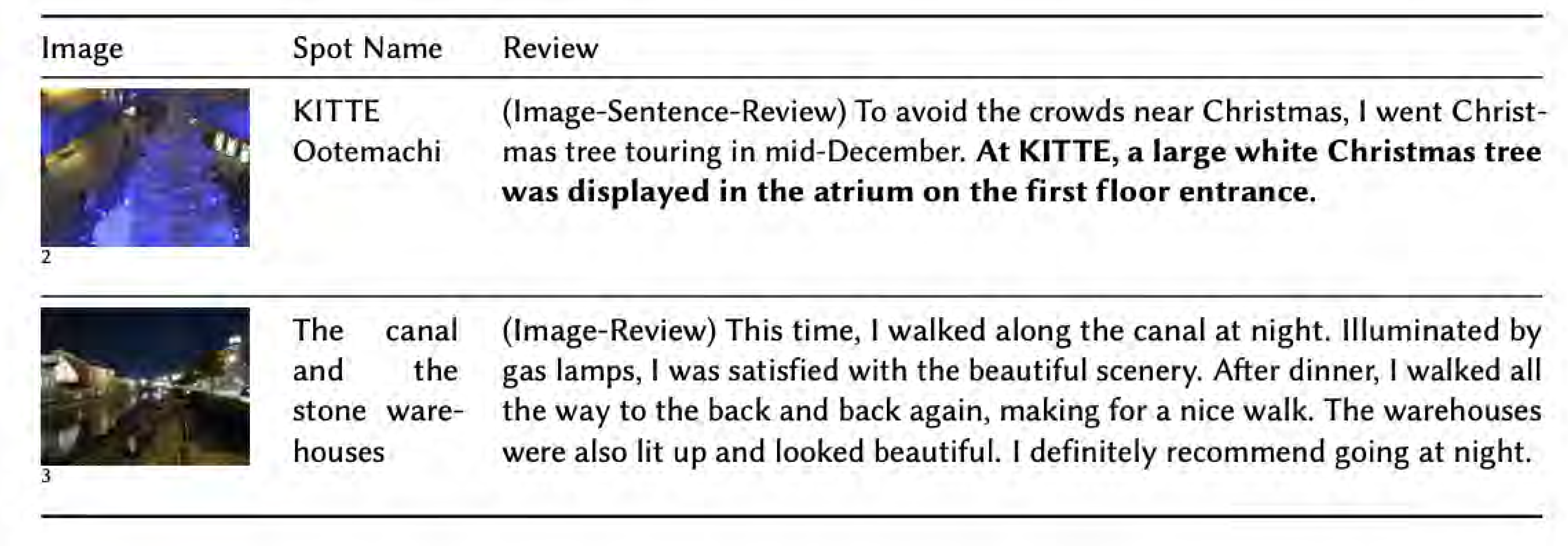
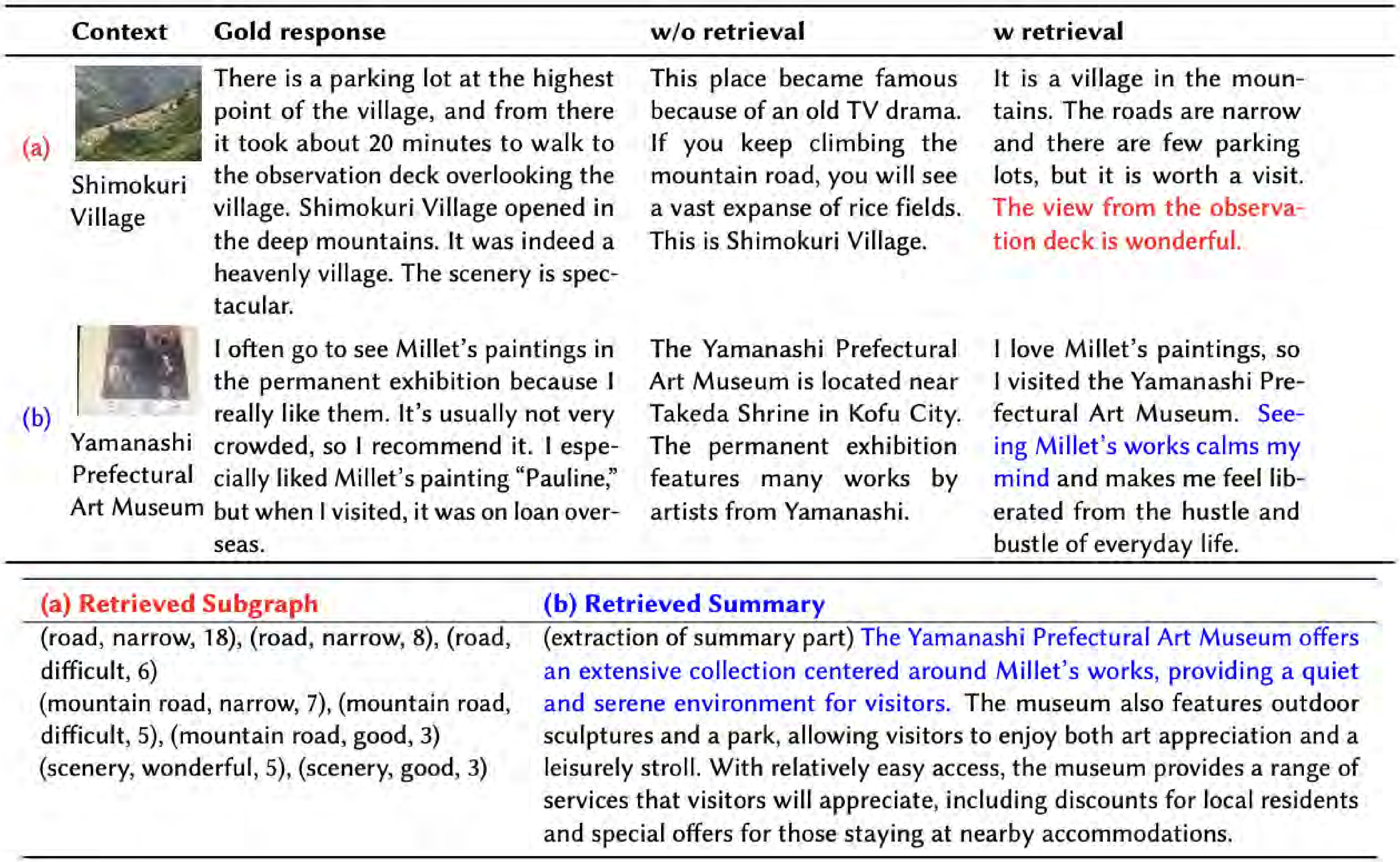


Image	Spot Name	Review
2	KITTE Ootemachi	(Image-Sentence-Review) To avoid the crowds near Christmas, I went Christmas tree touring in mid-December. **At KITTE, a large white Christmas tree was displayed in the atrium on the first floor entrance.**
3	The canal and the stone warehouses	(Image-Review) This 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.

Appendix: The effectiveness of RAFT



	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.

Appendix: Visualization of Attention