Constructing Multi-Modal Dialogue Dataset by Replacing Text with Semantically Relevant Images

Nyoungwoo Lee*, **Suwon Shin***, **Jaegul Choo**, **Ho-Jin Choi**, and **Sung-Hyon Myaeng**

KAIST, Daejeon, South Korea

{*leenw2, ssw0093, jchoo, hojinc, myaeng*}*@kaist.ac.kr*

Abstract

In multi-modal dialogue systems, it is important to allow the use of images as part of a multi-turn conversation. Training such dialogue systems generally requires a large-scale dataset consisting of multi-turn dialogues that involve images, but such datasets rarely exist. In response, this paper proposes a 45k multi-modal dialogue dataset created with minimal human intervention. Our method to create such a dataset consists of (1) preparing and pre-processing text dialogue datasets, (2) creating image-mixed dialogues by using a text-to-image replacement technique, and (3) employing a contextual-similarity-based filtering step to ensure the contextual coherence of the dataset. To evaluate the validity of our dataset, we devise a simple retrieval model for dialogue sentence prediction tasks. Automatic metrics and human evaluation results on such tasks show that our dataset can be effectively used as training data for multi-modal dialogue systems which require an understanding of images and text in a context-aware manner. Our dataset and generation code is available at https://github.com/shh1574/multi-modal-dialogue-dataset.

1 Introduction

Humans often use images in instant messaging services to express their meaning and intent in the dialogue context. For a dialogue system such as a chatbot to respond to human users adequately in this kind of multi-modal situations, it is necessary to understand both images and texts in their context and incorporate them in the dialogue generation process.

Training such a multi-modal dialogue system generally requires a large amount of training data involving images and texts in various contexts. However, numerous existing approaches relying on image captioning (Lin et al., 2014; Young et al., 2014) or visual question answering (Mostafazadeh et al., 2016; Das et al., 2017) techniques had to be trained with the datasets mostly irrelevant to the dialogue context. In other words, images were interpreted independently of the dialogue context, due to the lack of sufficient multi-modal dialogue datasets.

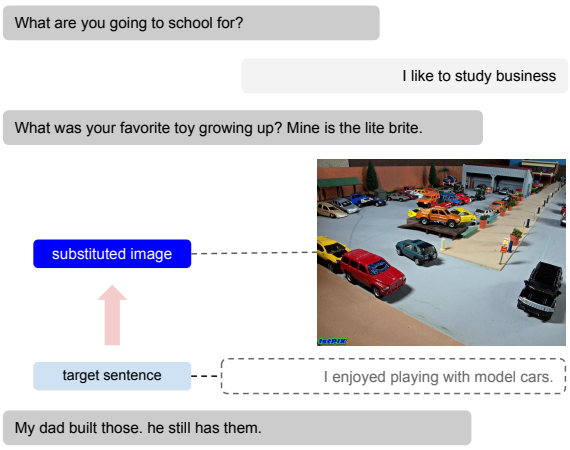

Figure 1: Example of multi-modal dialogue dataset.

Those datasets containing image-grounded conversations (Mostafazadeh et al., 2017; Shuster et al., 2020a) do not even cover the situations related to dialogue context before the image, because all conversations in the dataset always start from the given image. Although the relationship between images and texts can be learned using image-grounded conversations (Lu et al., 2019; Chen et al., 2020; Tan and Bansal, 2019; Su et al., 2020; Li et al., 2019b), it cannot still learn the dependency between the dialogue context before and after the image.

In this paper, we propose a 45k multi-modal dialogue dataset in the form of Fig. 1. Each multi-modal dialogue instance consists of a textual response and a dialogue context with multiple text

* Equal contribution.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 897–906
August 1–6, 2021.

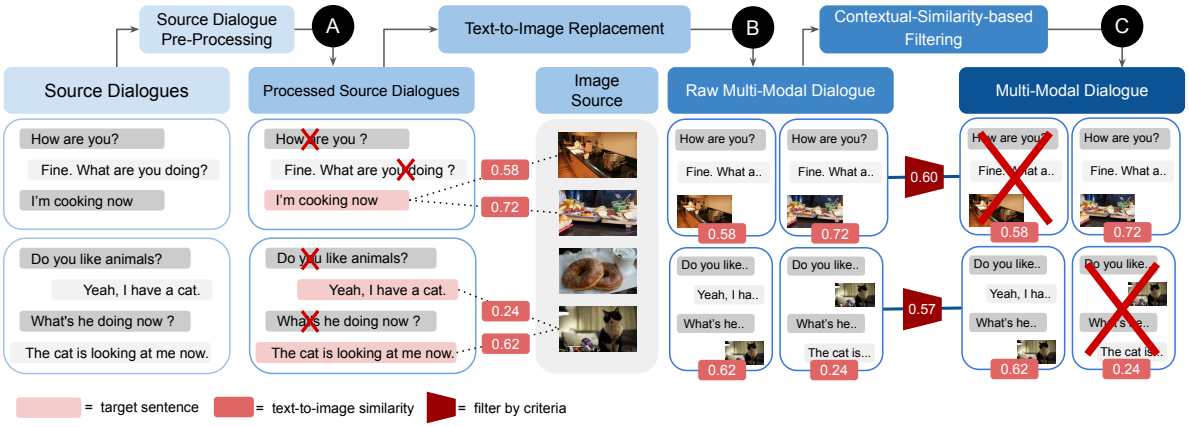

Figure 2: Overall pipeline for multi-modal dialogue dataset creation.

utterances and an image. To create this dataset, we start with existing text-only dialogue datasets as source dialogues, and then replace part of sentences in source dialogues with their semantically relevant images. The detailed steps include (1) source dialogue pre-processing, such as deleting a stop word, to improve the quality of similarity calculations, (2) creating dialogues containing an image by replacing a sentence with a similarity-based text-to-image replacement technique, and (3) pruning low-quality dialogues by employing a contextual-similarity-based filtering method. The overall process ensures that the created dataset consists of natural dialogue examples containing diverse images.

In order to validate our dataset creation process and examine the quality of our multi-modal dialogue dataset, we devise the task of predicting current and next dialogue sentences while considering the dialogue context and images. We also develop simple retrieval models to learn the relationship between images and texts for the tasks. Human evaluation results for predicting dialogue tasks show that the sentences are predicted as intended, i.e., in a context-aware manner, using the images. The results also show that our dataset can serve as practical training resources for multi-modal dialogue tasks that involve both image and dialogue context.

2 Multi-Modal Dialogue Generation

Our multi-modal dialogue dataset is constructed based on three source dialogue datasets and two image captioning datasets: DailyDialog (Li et al., 2017), EmpatheticDialogues (Rashkin et al., 2019), and Persona-Chat (Zhang et al., 2018) for the former and the MS-COCO (Lin et al., 2014) and Flicker 30k (Young et al., 2014) for the latter. The statistics of each dataset are summarized in Appendix A. After obtaining the source datasets, we replace sentences in the source dialogues with proper images by searching the image dataset to create image-mixed dialogues that maintain semantic coherence. To this end, we apply the three-stage method as shown in Fig. 2: (1) source dialogue pre-processing, (2) text-to-image replacement, and (3) contextual-similarity-based filtering.

Source Dialogue Pre-Processing We pre-process source dialogue datasets for the subsequent text-to-image replacement (A in Fig. 2). To select candidate dialogue sentences to be replaced by images, we first exclude the question sentences from the candidate dialogues because it is not realistic to infer back a question out of an image to put in the place of the question. This step filters out 25.08% of the total sentences in the source dialogue datasets. Second, we remove stop words from the source dialogue datasets, because they do not contain meaningful information. All the remaining sentences in dialogue contexts after the pre-processing step are considered as potential target sentences to replace.

Text-to-Image Replacement In this step, we create multi-modal dialogues containing images by replacing target sentences from the candidate dialogue sentences with appropriate images in the image dataset based on text-to-image similarity (B in Fig. 2). We calculate the similarity by the pre-trained Visual Semantic Reasoning Network (VSRN) (Li et al., 2019a), a state-of-the-art image-

text matching model based on text-to-image similarity. We first identify target sentences and then select candidate images for replacement using the threshold ensuring context coherence, as will be discussed in the subsequent contextual-similarity-based filtering step. Because we aim to maintain the comprehensive flow of the dialogue, we replace only one sentence with an image per dialogue. If multiple image candidates exist for a single sentence, we separate them into distinct image-mixed dialogue instances. In detail, such separated instances are all made up of the same dialogue context and text response except for substituted images.

Contextual-Similarity-based Filtering We employ a contextual-similarity-based filtering step to enhance the context coherence of the created image-mixed dialogues (C in Fig. 2). We filter out the dialogues where text-to-image similarity does not exceed the threshold determined by human annotators. For human annotators on the matching quality of an image, a total of 300 test dialogues are selected for each combination. Since we used three source dialogue datasets and two image datasets, we create six combinations of each dialogue dataset and each image dataset. Automatically created image-mixed dialogue instances are divided into ten segments based on the similarity values, and 30 are selected randomly from each. We hired a total of 18 annotators to evaluate 1,800 instances sampled from these six combinations. The evaluation system is described in Appendix C.

The human evaluation was conducted based on three questions for each instance:

- Q1: How well does the substituted image contain **key objects** in the target sentence?
- Q2: How well does the substituted image represent the **meaning** in the target sentence?
- Q3: When the image is substituted for the target sentence, how **consistent** is it with the **context** of the conversation?

Q1 and Q2 ask whether the substituted image contains the core meaning of the target sentence (on a 3-point scale). Q3 evaluates the context coherence of the created dialogue containing the image (on a 5-point scale). We assume that dialogues above the median of the evaluation score (2 for Q1, Q2, and 3 for Q3) are suitable for use as training instances. Based on this assumption, we determine the threshold for each combination by interpolating the median in the correlation graph of the evaluation results and the similarity (Appendix B). We then analyze the correlation between the score for each question and text-to-image similarity using Spearman's correlation analysis as shown in Table 1. Overall, the similarity values are positively correlated with the scores obtained for the questions. Since Q2 and Q3 are reasonably correlated with semantic similarity, the substituted images tend to reflect the meaning of the target and context sentences. Thus, the evaluation results indicate that the automatically created image-text pairs with high similarity can be used as multi-modal dialogues. We filter the generated multi-modal dialogues based on the determined similarities, and then set the filtered dialogues as our final dataset. The statistics of the final dataset are summarized in Table 2.

	Similarity	Q1	Q2	Q3
Similarity		**0.5893**	0.4422	0.4334
Q1			0.7103	0.6646
Q2				**0.7570**
Q3				

Table 1: Spearman's correlation ρ between three questions and text-to-image similarity.

	train	valid	test
# total dataset	**39956**	**2401**	**2673**
Avg length of dialogue turns	13.01	13.62	13.59
Avg length of sentences	51.47	50.76	50.70
# total unique images	12272	334	682
# total unique dialogues	13141	2148	2390
# total unique target sentences	21495	2400	2671
Avg # of substituted images in a dialogue	1.86	1.00	1.00
Avg # of targets in a dialogue	1.64	1.12	1.12

Table 2: Multi-modal dialogue dataset statistics for splits of training, validation, and test set.

Data Quality We evaluate the quality of our dataset to validate the proposed dataset creation method. To this end, we randomly sample 300 image-mixed dialogues from our final dataset. The evaluation proceeds in the same manner as before, but we add a new question Q4, which asks to choose the intent of the image used in the dialogue as one among (1) answering the question, (2) expressing emotional reactions, (3) proposing a new topic, and (4) giving additional explanations for the previous context. For Q1, Q2, and Q3, the average scores evaluated by three annotators are shown to be 2.56, 2.17, and 3.13, respectively, indicating that

Model	Task	R@1	R@5	Mean Rank
IR Baseline	Current	21.62	49.49	30.04
IR Baseline	Next	8.13	21.07	29.41
Retrieval Model	Current	**50.35**	**86.64**	**3.11**
Retrieval Model	Next	14.38	36.10	20.58

Table 3: Automatic evaluation results about retrieval models and an information retrieval baseline on the current and next dialogue prediction task.

the context of the conversation containing the substituted image is consistent in our dataset. For Q4, the responses from the annotators are distributed with 27.3%, 20.0%, 32.7%, and 14.7%, for the four intent types as mentioned above, indicating our dataset contains balanced intent types.

3 Experiments

3.1 Experimental Setup

We consider two dialogue sentence prediction tasks given an image and a dialogue: current dialogue prediction and next dialogue prediction for a given image. We use a simple retrieval model composed of three modules (Shuster et al., 2020a,b): Resnext-101 (Xie et al., 2017) for an image encoder, BERT (Devlin et al., 2019) for a text encoder, and the fusion module. As input for training the model, we use images and up to three dialogue sentences immediately before the images as dialogue context.

3.2 Automatic Evaluation

We perform quantitative comparisons that follow recent work (Shuster et al., 2020a) to find the optimal setting for our retrieval model (Appendix D). To evaluate the retrieval accuracy, we use the recall at 1 and 5 out of 100 candidates consisting of 99 candidates randomly chosen from the test set and 1 ground-truth sentence, called R@1/100 and R@5/100, respectively. We also use the mean reciprocal rank. We compare our model with a simple information retrieval baseline. The candidates of the baseline model are ranked according to their weighted word overlap between the target sentence and an image caption followed by dialogue context.

As shown in Table 3, the R@1 performance of the retrieval model obtained 50.35 and 14.38 on the current and next sentence prediction task, outperforming the baseline on both tasks. This result indicates that our dataset properly works as the training data to learn the relationship between images and dialogue context in dialogue sentence prediction tasks where images and dialogue context have to be considered together.

Model inputs	R@1	R@5	Mean Rank
Image Only	37.30	80.66	3.91
Dialogue Context Only	28.06	56.83	12.57
Image + Dialogue Context	**51.21**	**86.34**	**3.08**

Table 4: Ablation studies of our retrieval models on the current dialogue prediction task.

Model inputs	R@1	R@5	Mean Rank
Image Only	7.29	21.92	31.78
Dialogue Context Only	11.90	29.89	23.95
Image + Dialogue Context	**14.38**	**36.10**	**20.58**

Table 5: Ablation studies about our retrieval models on the next dialogue prediction task.

3.3 Ablation Study

We then conduct ablation studies by removing modalities (image and dialogue context) in turn to check whether unwanted correlations exist in our dataset. Since we created our training and test datasets by a semi-automatic data creation method, unwanted correlations can exist in datasets that can infer the correct answer without using the image and context simultaneously. Such correlations would prevent the model from properly learning the relationship between images and context.

As shown in Tables 4 and 5, the results first show that the recall measure for ground-truth answers in the model that considers both context and image is higher than the model considering only images. It indicates that the models in each task properly consider both images and dialogue context to predict sentences. To elaborate, the model that only considers images are likely to choose responses that do not match the dialogue context before the image. For example in a given dog photo shown during a sad mood conversation, the model that only considers images can generate an out-of-context response, such as "It is so cute.". On the other hand, in the same context, the model that considers both the context and the image could generate appropriate responses, such as "what is wrong with your dog?" or "I miss your dog.".

The overall tendency also shows that the model performance degrades when we delete each modality one by one. Such results suggest that our data creation process did not generate correlations that interfere with forming the relationship between images and dialogue context.

3.4 Human Evaluation

We create a new test set to confirm that the model can predict sentences well even on test dialogues that are not constructed in the same manner. To this end, two researchers manually created 100 multi-modal dialogues by adding images to source dialogues that were not used in our dataset generation process for human evaluation. We proceed with the evaluation with three annotators per each prediction task, using a question (on a 5-point scale) asking how much the sentences predicted by the model are relevant to the image and dialogue context. The average scores of three annotators for each task were shown to be 3.36 for the current turn prediction and 3.06 for the next turn prediction. The results indicate that the models can predict sentences in a context-aware manner even with dialogues organized by humans.

4 Conclusions

We present the multi-modal dialogue dataset consisting of 45k multi-turn dialogues containing semantically coherent images as well as the dataset creation method. Human evaluation results of our multi-modal dialogues reveal that context coherence is well maintained even if the sentence is replaced by an image, showing the validity of our dataset and data creation approach. We then evaluate our dataset using two multi-modal dialogue prediction tasks, demonstrating its effectiveness when training a dialogue system to learn the relationship between images and dialogue contexts. Our proposed data creation method can be applied when efficiently preparing large-scale multi-modal dialogue datasets that cover diverse multi-modal situations.

Acknowledgments

This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2013-2-00131, Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services, No. 2019-0-00075, Artificial Intelligence Graduate School Program(KAIST), and No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques)

References

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: learning universal image-text representations. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 104–120.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 326–335.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019a. Visual semantic reasoning for image-text matching. In *Proc. of the IEEE international conference on computer vision (ICCV)*, pages 4654–4662.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019b. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *International Joint Conference on Natural Language Processing (IJCNLP)*, pages 986–995.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 740–755.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, pages 13–23.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *International Joint Conference on Natural Language Processing (IJCNLP)*, pages 462–472.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proc. the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1802–1813.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proc. the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5370–5381.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020a. Image-chat: Engaging grounded conversations. In *Proc. the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2414–2429.

Kurt Shuster, Eric Michael Smith, Da Ju, and Jason Weston. 2020b. Multi-modal open-domain dialogue. *arXiv preprint arXiv:2010.01082*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *Proc. the International Conference on Learning Representations (ICLR)*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5100–5111.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1492–1500.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, pages 67–78.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proc. the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2204–2213.

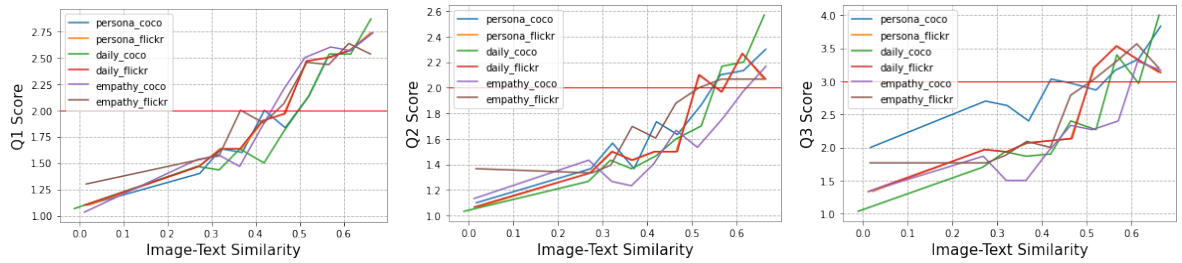

Figure 3: Correlation between text-to-image similarity and question scores (Q1, Q2, and Q3) for six combinations.

A Source Datasets Statistics

	type	training	validation	test
DailyDialogue	dialog	11118	1000	1000
Persona-Chat	dialog	8938	999	967
EmpatheticDialogues	dialog	17792	2758	2539
MS-COCO	image	113287	5000	5000
Flickr 30k	image	28000	1000	1000

Table 6: Source dialogue and image captioning dataset statistics for splits of training, validation, and test set.

B Detailed Description of Contextual-Similarity-based Filtering

	threshold	train	valid	test
Persona-COCO	0.546	11606	411	1136
Persona-Flickr	0.509	19148	1654	1014
Daily-COCO	0.555	3418	47	319
Daily-Flickr	0.619	141	6	5
Empathetic-COCO	0.623	245	2	11
Empathetic-Flickr	0.516	5398	281	188
Total		**39956**	**2401**	**2673**

Table 7: Number of data instances filtered by the thresholds for each combination

In this section, we analyze the human evaluation results for contextual-similarity-based filtering and determine thresholds for each dataset combination. The correlations between the similarity and evaluation results for each question are shown in Fig. 3. We assume that dialogue instances above the median of the evaluation score (2 for Q1, Q2, and 3 for Q3) are suitable for use in training. Based on the assumption, we determine the threshold for each combination by interpolating the median in the correlation graph of the evaluation results and the similarity. We select the largest one of three interpolated values of each question (Q1, Q2, and Q3). The data statistics for each combination filtered by the threshold are shown in Table 7.

Since the thresholds for each combination are determined differently, there are differences in the number of dialogue instances by combination. Such results suggest that the quality of multi-modal dialogue generation may vary depending on combining the text and image datasets. For example, the DailyDialog goes well with the MS-COCO but not with Flicker 30k. On the contrary, the EmpatheticDialogues goes well with the Flicker 30k but not with MS-COCO. Thus, we must consider finding the right combination among text and image datasets in the multi-modal dialogues generation process.

C Human Evaluation System

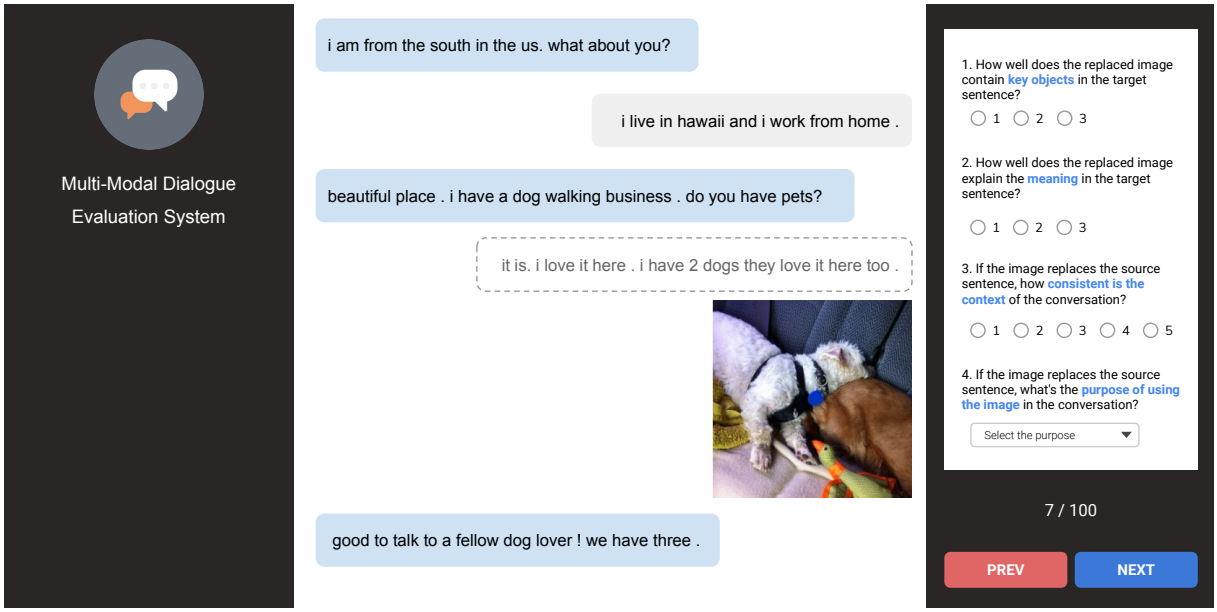

Figure 4: Human evaluation system for testing our multi-modal dialogue dataset.

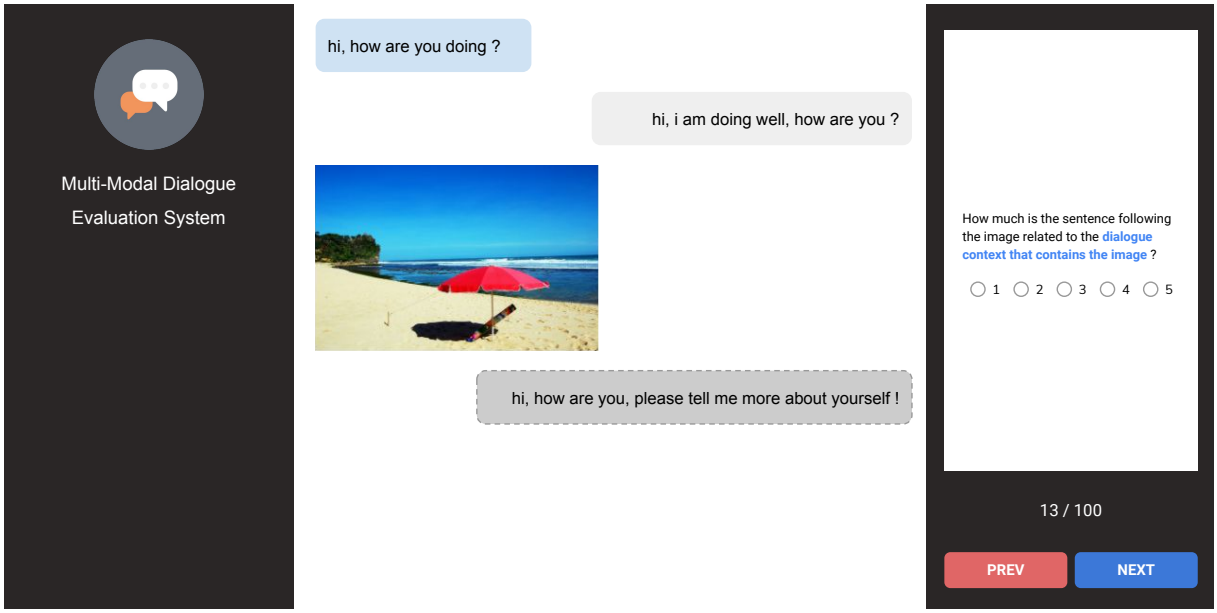

Figure 5: Human evaluation system for testing two dialogue sentence prediction tasks using our retrieval models.

In this section, we introduce the human evaluation system. We develop the system using a JavaScript library called ReactJS. Fig. 4 shows the implemented system for evaluating our multi-modal dialogue dataset. In this system, we ask users to evaluate a total of 100 dialog instances and answer three or four questions per instance. In addition to three questions described in Section 2, Q4[1] is added depending on the purpose of use. Fig. 5 shows the system for evaluating the performance of a retrieval model that performs dialog sentence prediction tasks. Similarly, we also ask users to evaluate a total of 100 dialog instances and answer one question per instance.

[1]Q4: If the image replaces the source sentence, what is the purpose of using the image in the conversation?

D Best Model Search

Model	Fusion Module	Image Encoder	Text Encoder	R@1	R@5	Mean Rank
$IRBaseline$	n/a	n/a	n/a	21.62	49.49	30.04
$RetrievalModel_{Att}$	Attention	Unfreeze	Freeze	11.74	39.13	15.73
$RetrievalModel_{Sum}$	Sum	Unfreeze	Freeze	9.95	35.13	15.73
$RetrievalModel_{Att}$	Attention	Unfreeze	Unfreeze	43.51	80.55	4.13
$RetrievalModel_{Sum}$	Sum	Unfreeze	Unfreeze	48.19	84.21	3.66
$RetrievalModel_{Att}$	Attention	Freeze	Unfreeze	48.41	85.97	3.40
$RetrievalModel_{Sum}$	Sum	Freeze	Unfreeze	**50.35**	**86.64**	**3.11**

Table 8: Comparison tests of the current dialogue prediction task on the multi-modal dialogue dataset. We compare different module variations and training strategies for our retrieval models.

Model	Fusion Module	Image Encoder	Text Encoder	R@1	R@5	Mean Rank
$IRBaseline$	n/a	n/a	n/a	8.13	21.07	29.41
$RetrievalModel_{Att}$	Attention	Unfreeze	Freeze	2.04	9.50	40.99
$RetrievalModel_{Sum}$	Sum	Unfreeze	Freeze	3.08	12.46	36.36
$RetrievalModel_{Att}$	Attention	Unfreeze	Unfreeze	4.09	15.95	32.07
$RetrievalModel_{Sum}$	Sum	Unfreeze	Unfreeze	13.38	33.93	21.10
$RetrievalModel_{Att}$	Attention	Freeze	Unfreeze	10.02	28.49	23.71
$RetrievalModel_{Sum}$	Sum	Freeze	Unfreeze	**14.38**	**36.10**	**20.58**

Table 9: Comparison tests of the next dialogue prediction task on the multi-modal dialogue dataset. We compare different module variations and training strategies for our retrieval models.

We compare different module options of our model. Each encoder has two options: whether to freeze or not during training, and the fusion module has two options: summation, and the attention-based transformer encoder. For final image-context fused representation, context and image representations are added in the summation fusion method, while two representations are concatenated, and then fed into the attention-based two-layer transformer encoder in the attention-based method. By this comparison, we decide to freeze only the image encoder and use the summation fusion method for both current and next dialogue prediction tasks.

We additionally show the results of an information retrieval baseline, which retrieves target dialogue using the tf-idf method between candidate dialogues and the caption of an image followed by dialogue context. As shown in Tables 8 and 9, our retrieval model significantly outperforms the information retrieval baseline, indicating that comprehensive understanding of context and images is helpful in multi-modal dialogues.

Our implementation uses an NVIDA TITAN RTX GPU for training, and training each epoch takes about 15 minutes. Our retrieval model using the summation fusion method has 204M parameters, while that using the attention-based fusion method has 254M parameters.

E Multi-Modal Dialogue Dataset Example

Figure 6: Our multi-modal dialogue dataset examples

For easy understanding of our dataset, we provide two additional examples of the multi-modal dialogue dataset in Fig. 6.

F Selected Example of Current Dialogue Prediction Task

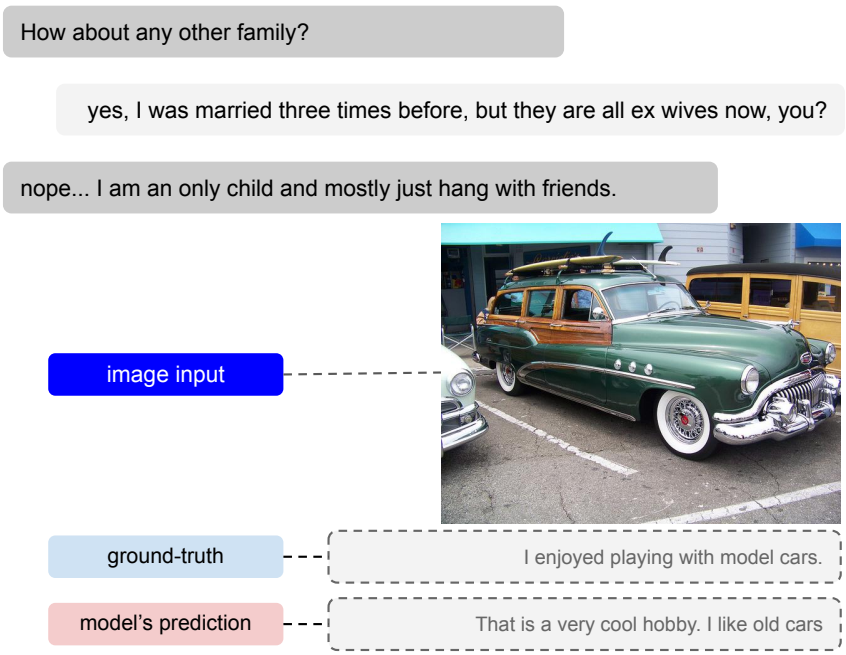

Figure 7: Ground-truth and dialogue sentence prediction example by our retrieval model used in the current turn prediction task.

Fig. 7 shows a reasonable example of a retrieved dialogue sentence by the retrieval model used in the current turn prediction task. Even if the model does not predict the ground-truth sentence, it can predict a plausible dialogue sentence.