DialogCC: Large-scale Multi-Modal Dialogue Dataset

Young-Jun Lee[1] Byungsoo Ko [2] Han-Gyu Kim [3] Ho-Jin Choi [1]
[1] KAIST [2] NAVER Vision [3] NAVER Clova Speech
github.com/passing2961/DialogCC

Abstract

As sharing images in an instant message is a crucial factor, there has been active research on learning a image-text multi-modal dialogue model. However, training a well-generalized multi-modal dialogue model is challenging because existing multi-modal dialogue datasets contain a small number of data, limited topics, and a restricted variety of images per dialogue. In this paper, we present a multi-modal dialogue dataset creation pipeline that involves matching large-scale images to dialogues based on CLIP similarity. Using this automatic pipeline, we propose a large-scale multi-modal dialogue dataset, DialogCC, which covers diverse real-world topics and various images per dialogue. With extensive experiments, we demonstrate that training a multi-modal dialogue model with our dataset can improve generalization performance. Additionally, existing models trained with our dataset achieve state-of-the-art performance on image and text retrieval tasks. The source code and the dataset will be released after publication.

1. Introduction

People share various images with each other when communicating via instant messaging tools. Such behavior increases social bonding (rapport) as well as engagement. The ability to share images is also necessary for a dialogue model for better bonding conversations. In the visual dialogue domain, the majority of previous works have focused on image-grounded dialogues, where two persons talk about given images [2, 6, 18, 27, 29, 31, 38, 39, 46, 47, 51]. In practical situations, humans actively share images during conversations rather than merely talking about a given image, which is called *image sharing* behavior [49]. Recent studies for image sharing have proposed multi-modal dialogue datasets, which are constructed either manually by crowd-sourcing (PhotoChat [49]) or automatically by utilizing image-text similarity (MMDD [20][1]).

[1]Mutli-Modal Dialogue Dataset.

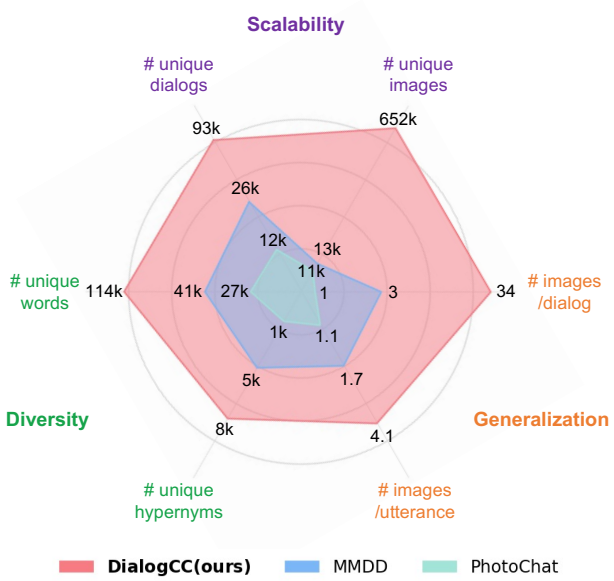


Figure 1. **Dataset statistics comparison.** We compare our proposed dataset DialogCC with existing multi-modal dialogue datasets: MMDD [20] and PhotoChat [49]. Compared to the existing datasets, DialogCC contains *large-scale* dialogues and images while covering more *diverse* words and hypernyms. Moreover, learning with various images for the same dialogue or utterance in DialogCC would benefit a model to obtain better *generalization*.

However, existing multi-modal dialogue datasets have three significant limitations; **(1) Scalability.** Recent years have witnessed the success of large-scale multi-modal pre-training models [1, 12, 17, 26, 32], which benefited from the power of large-scale image-caption datasets [4,7,35,37,42]. Nevertheless, as shown in Figure 1 (*# unique dialogs* and *# unique images*), existing image-dialogue datasets [20, 49] contain a small number of dialogues and images that are limited to training a large-scale multi-modal dialogue model. **(2) Diversity.** A large-scale multi-modal dialogue model should be able to cover open domain conversation. However, as illustrated in Figure 1 (*# unique words* and *# unique hypernyms*), existing datasets cover a limited number of words, topics, and domains. Such lack of diversity

can also be found in conversational skills (e.g., empathizing with situations [15, 24, 33] and understanding personal information [14, 50]), which is an important factor in human conversation. Existing datasets cover only few conversational skills (see Section 3.2). **(3) Generalization.** Given the same dialogue and context, people can share different types of images. For example, for an utterance of "*I love pets*," one can share an image of a dog, and the other can share an image of a cat. Nonetheless, as shown in Figure 1 (*# images / dialog* and *# images / utterances*), existing datasets consist of less than the average 3 images per dialogue and the average 1.7 images per utterance. A model trained with such datasets can be overfitted by memorizing those pairs of images and dialogues, resulting in a lack of generalization.

The objective of this work is to create a large-scale multi-modal dialogue dataset in order to train a well-generalized multi-modal dialogue model for open-domain conversation. To this end, we present an automatic pipeline for creating a multi-modal dialogue dataset and propose a large-scale multi-modal dialogue dataset, DialogCC, created by the automatic pipeline. The pipeline consists of two main filtering steps: source data filtering and multi-modal dialogue filtering. These filtering steps eliminate inappropriate images from large-scale images and dialogues based on CLIP similarity. As illustrated in Figure 1, DialogCC achieves better statistics compared to the existing datasets in terms of scalability, diversity, and generalization. In addition, extensive experiments demonstrate that a model trained with DialogCC achieves state-of-the-art performance in image and text retrieval tasks with enhanced generalization performance.

In summary, our main contributions are as follows: 1) We present an automatic pipeline for creating a multi-modal dialogue dataset that can create a large-scale dataset without human effort. 2) We propose a large-scale multi-modal dialogue dataset named DialogCC, which contains diverse images and dialogues consisting of various images per dialogue. 3) Extensive experiments demonstrate the effectiveness of our dataset, which achieves state-of-the-art performances.

2. Related Work

Image-Dialogue Dataset. In the visual dialogue domain, most previous studies are divided into two categories depending on whether the image is *grounded* or *sharing* in the dialogue. The image-grounded dialogue task aims to answer questions [2, 6, 18, 36] or generate natural conversations [27, 29, 38, 46, 51] about given images. These datasets require machines to perceive and understand the given images, but we sometimes share images relevant to dialogue contexts in daily conversations. Hence, it is difficult to train dialogue agents to retrieve an appropriate image based on dialogue contexts in image-grounded dialogue task.

Recently the image-sharing dialogue task has been proposed to overcome such limitation, which predicts images semantically relevant to given dialogue contexts. Since there were no existing datasets for image-sharing task, previous studies have focused on construction of the dataset. One of the existing datasets, named PhotoChat [49], is manually constructed through a crowd-sourcing platform with Open Image Dataset V4 [19] as source images. This dataset can provide a high-quality dialogue dataset, but the manual construction is time-consuming and expensive, which makes it hard to be expanded to a large-scale dataset. Another line of work [20] creates a 45k multi-modal dialogue dataset through an automatic pipeline composed of the Visual Semantic Reasoning Network [21] (VSRN). They replace utterances with semantically relevant images based on the similarity between the image and utterance obtained from VSRN. They also remove contextually incoherent images based on the threshold obtained through human annotation. However, both datasets are small and cover limited image objects and topics, as demonstrated in Figure 1. To this end, we construct a large-scale multi-modal dialogue dataset through the automatic pipeline.

Multi-Modal Dialogue Model. The multi-modal dialogue model is mainly categorized into retrieval and generative models. The retrieval model is to retrieve proper texts or images from the candidates given the dialogue contexts. The generative model is to generate responses given the dialogue contexts. For the retrieval model, most existing studies have adopted the dual encoder architecture consisting of a text encoder and image encoder [20, 38, 49]. For the generative model, many works are based on the encoder-decoder architecture [25, 39, 44, 47]. In this paper, for fair comparisons, we evaluate our dataset by adopting the text retrieval model [20] and image retrieval model [49] as our baselines.

3. Multi-Modal Dialogue Dataset Creation

In this section, we propose DialogCC, which is a large-scale multi-modal dialogue dataset. In order to construct DialogCC, we introduce an automatic pipeline, which consists of two steps: (1) source data filtering, (2) multi-modal dialogue filtering. Besides, we conduct a comprehensive analysis of our dataset with respect to scalability, diversity, and generalization by comparing two existing datasets, MMDD [20] and PhotoChat [49].

3.1. CLIP-based Automatic Pipeline

We present an automatic pipeline for constructing DialogCC. The key idea of our pipeline is considering two types of similarities: *utterance-image* similarity and *utterance-caption* similarity. The overall pipeline is illus-

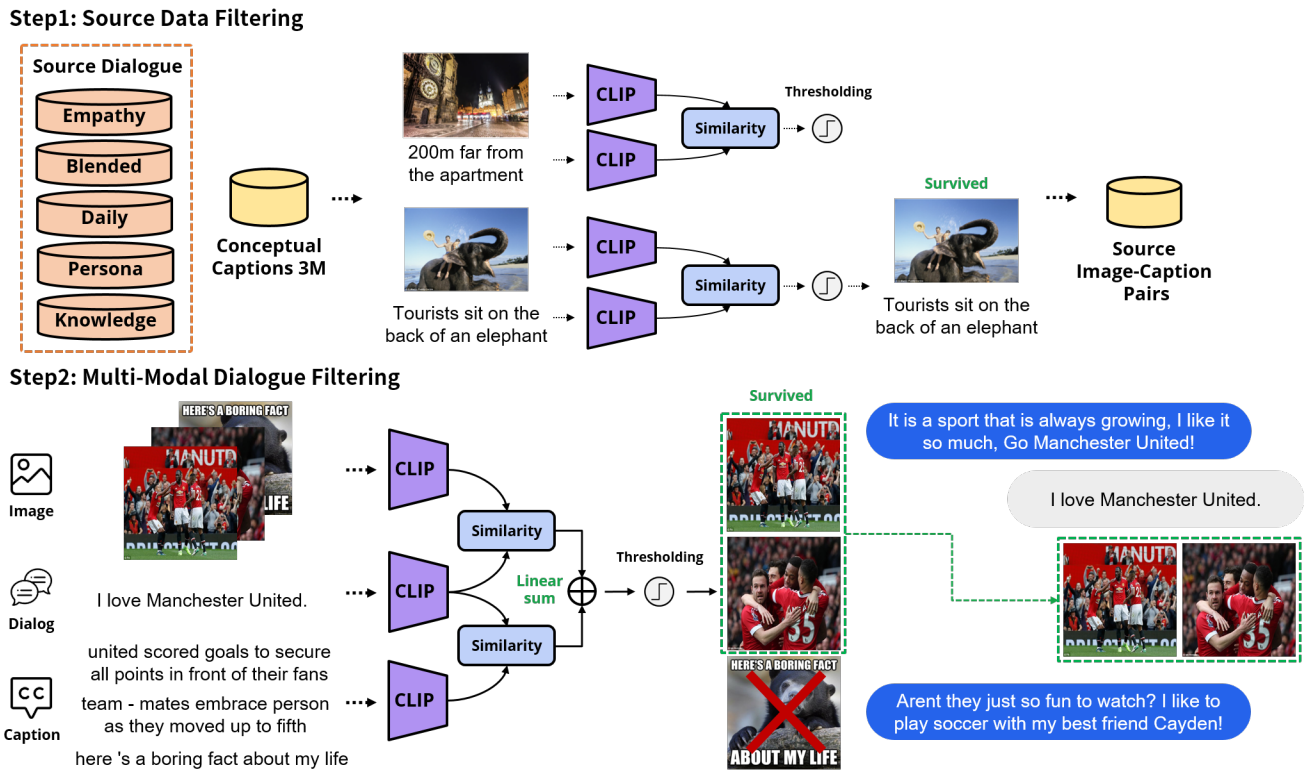


Figure 2. **An overview of the automatic pipeline for multi-modal dialogue dataset creation.** We collect source text-only dialogues and images, which survived by thresholding based on the similarity score between two modalities obtained from the CLIP model. Next, we combine two similarity types for each computed by utterance-image and utterance-caption. We then remove unrelated images from the matched results.

trated in Figure 2. In the following part of this section, we provide details about DialogCC construction pipeline.

3.1.1 Source Data Filtering

Source Dialogue. As a source data, we collect five multi-turn text-only dialogue datasets, which are publicly available online. Five dialogue datasets are Persona-Chat [50], EmpatheticDialogues [33], Wizard-of-Wikipedia [8], DailyDialog [22], and BlendedSkillTalk [40]. They are manually constructed via a crowd-sourcing platform, and each dataset is specialized in specific conversational skills. Persona-Chat dataset contains the ability to get to know each other based on given personal information. EmpatheticDialogues dataset contains the ability to understand and interpret interlocutors' emotional situations and express emotional reactions adequately. Wizard-of-Wikipedia contains the ability to generate specific responses using knowledge or topic. DailyDialog contains daily life conversations with aspects, such as emotion, topic, and dialog acts. Lastly, in the BlendedSkillTalk, multiple skills (i.e., persona, empathy and knowledge) are integrated into one conversation, as humans do. We incorporate five dialogue datasets into one large dialogue dataset by removing duplicated dialogues.

Source Image-Caption Pairs. We choose Conceptual Captions 3M [37] (CC3M), which is widely used in multi-modal modeling [26, 43] and creating multi-modality dataset [30]. We obtain 2,712,320 image-caption pairs for the training and validation set. The detailed collection process is described in Appendix. The duplicated images and images with image-caption similarity smaller than the threshold of 0.185 are filtered out. After the filtering, 2,440,485 image-caption pairs are obtained, which are divided into the training / validation / test set with a ratio of 5:1:1, resulting in 1.7M / 0.3M / 0.3M of unique images. Note that our pipeline can work with any image-caption datasets, such as Conceptual Captions 12M [4] and RedCaps [7].

3.1.2 Multi-Modal Dialogue Filtering

CLIP-based Similarity Calculation. In order to find images semantically relevant to given utterance, we should get meaningful textual and visual features through a multi-modal feature extractor $f(\cdot)$. The previous work [20] used a pre-trained Visual Semantic Reasoning Network [21] as $f(\cdot)$. In this work, we leverage CLIP [32] model as $f(\cdot)$,

Figure 3. **Comparison examples of relevant images in MMDD [20] vs. DialogCC.** In the first and second rows, both datasets contain semantically relevant images to the given utterances, while our dataset contains more and various images with different views or objects (e.g., dog breeds or backgrounds). In the last two rows, unlike the MMDD, our dataset is highly relevant for a given utterance because of matching from large-scale and diverse source data and considering image captions. Images in the green box are from MMDD, images in the red box are from DialogCC, and the blue box is utterances. More examples are in the Appendix.

which is widely used in previous studies [3, 5, 10, 11] because of a well-generalized open-domain model. We first extract utterance feature vector ($v_u = f(u)$), caption feature vector ($v_c = f(c)$), and image feature vector ($v_i = f(i)$). We then calculate the *utterance-image* similarity following [20] by computing the cosine similarity of v_u and v_i.

Besides, in order to enhance the quality of utterance-image matching by additionally adopting the information provided by image captions, we also calculate the *utterance-caption* similarity. The results from CLIP-based similarity calculation are shown in Figure 3. For example, in the last two rows of Figure 3, MMDD overlooks "hawaiian" or "play" in the utterances, which are important words, and only focuses on the words "dance" or "keyboard". This is because MMDD does not consider the image captions.

However, there is one problem that we have to consider about how to combine these two similarity types. As reported in [23, 41], there is a phenomenon called *modality gap* in multi-modal modeling, where two different modalities (e.g., image and text) are separately distributed in shared embedding space. Such phenomenon causes scale differences between utterance-image and utterance-caption similarities, so combining them directly would be biased to the larger scaled similarity. To alleviate this problem, the z-score normalization is conducted on both types of similarities where the mean and standard deviation values for each similarity type are calculated using training set. The normalized similarities are linearly combined as follows:

$$\mathcal{S} = \alpha f_Z\left(s_c(v_u, v_i)\right) + (1-\alpha) f_Z\left(s_c(v_u, v_c)\right), \quad (1)$$

where $s_c(x, y)$ denotes the cosine similarity and f_Z represents z-score normalization. In this paper, we set α as 0.5 to reflect two similarities equally. During the utterance-image matching process, the similarity matrix $\mathcal{S}$ of size of $N \times M$ is computed, where N and M are the number of utterances and images respectively.

Filtering. We have found out that there still exist unsuitable cases among the matched images found by CLIP-based similarity. To improve the quality of our dataset, we present two more simple filtering after similarity-based matching.

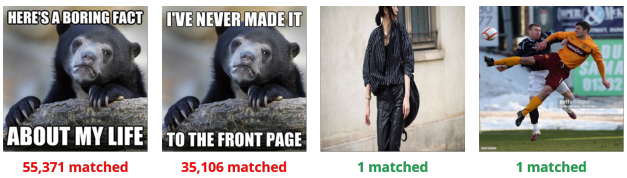


Figure 4. **Examples of frequently matched images.** We show representative examples of images matched with various utterances by CLIP-based similarity. The number under each image indicates the count of how many utterances are matched.

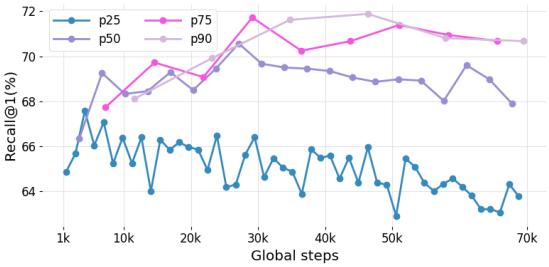


Figure 5. **Ablation study on the threshold τ_2.** We show the effect of the current turn prediction performance on DialogCC according to the threshold τ_2.

In the first filtering stage, images whose scores are lower than a threshold τ_1, where τ_1 is the median value of all scores. Median is used instead of heuristically determined threshold so that our pipeline can be applied to arbitrary datasets. Besides, we have observed that certain images are frequently matched with many utterances. As shown in Figure 4, the frequently matched images mostly contain general semantics (e.g., meme, questions), which goes along with general utterances, rather than object-centric or event-centric semantics (e.g., "playing a soccer game"). This phenomenon can result in the overfitting of the model to such frequently matched images, which is harmful to the generalization performance. Therefore, we propose to filter out such unsuitable images based on the frequency of being matched. In our method, a relatively determined threshold τ_2 is used to filter out a specific ratio of images. For example, using the threshold of $\tau_2 = p25$ denotes that only bottom 25% of images by frequency of being matched are included in the constructed dataset. We conduct an ablation study on the text retrieval performance (i.e., current turn prediction task) on our dataset by differentiating the relative determined threshold τ_2 as shown in Figure 5. The ablation study shows that the value of $\tau_2 = p75$ is the best for both the efficient training and performance. Unless otherwise noted, all experiments on DialogCC are conducted with $\tau_2 = p75$.

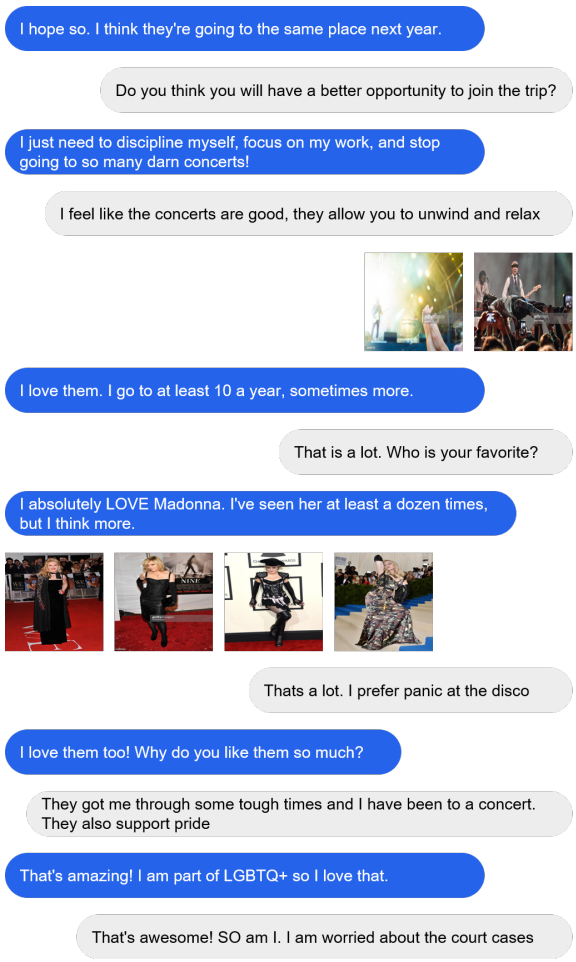


Figure 6. **An example of DialogCC.** We present an example of DialogCC with resized images. More examples are in Appendix.

3.2. Analysis of DialogCC

(1) Scalability: Table 1 shows the overall statistics of DialogCC compared to existing datasets MMDD [20] and PhotoChat [49]. In general, DialogCC comprises 92k unique dialogues in total, which is roughly 3.7× and 7.7× larger than MMDD and PhotoChat, respectively. While the average dialogue turn are shorter than that of other datasets, the average length of utterances is longer. Especially, DialogCC contains 651k unique images, which is approximately 54× larger than MMDD and PhotoChat. Our dataset is large-scale in both the number of dialogue and images.

(2) Diversity: In Table 2, we compare the diversity of datasets with the number of unique hypernyms from WordNet [28] and words in dialogues and image captions. As WordNet covers nouns, verbs, adjectives, and adverbs, we only count nouns with filtering out the hypernyms appearing less than ten times. In dialogue, compared to PhotoChat and MMDD, our dataset includes 3.7× and 1.4× more hy-

Datasets	Type	# Unique Dialog	Avg. # Turns	# Unique Image	Avg. # I./D.	Avg. # I./U.
MMDD [20]	train	21,411	13.0	12,272	3.3	1.9
	valid	2,400	13.6	334	1.1	1.0
	test	2,672	13.6	682	1.1	1.0
	total	26,434	13.1	13,288	3.4	1.7
PhotoChat [49]	train	10,286	12.7	8,899	1.0	1.0
	valid	1,000	12.7	1,000	1.0	1.0
	test	1,000	12.8	1,000	1.0	1.0
	total	12,286	12.7	10,889	1.2	1.1
DialogCC(ours)	train	77,354	8.3	467,944	32.2	3.9
	valid	7,940	8.2	92,723	43.0	5.2
	test	7,648	8.2	91,198	42.6	5.2
	total	**92,942**	**10.0**	**651,840**	**34.0**	**4.1**

Table 1. **Statistics of Datasets.** In total, DialogCC includes the largest number of unique dialogues and images than others. I./D. and I./U. denote images by an dialogue and images by an utterance, respectively. More detailed statistics are in the Appendix.

	Dialog		Image caption		Total	
	# hyp	# word	# hyp	# word	# hyp	# word
MMDD [20]	3,158	25,962	1,709	15,364	4,867	41,326
PhotoChat [49]	1,163	24,093	193	2,935	1,356	27,028
DialogCC(ours)	**4,337**	**82,912**	**4,097**	**30,911**	**8,434**	**113,826**

Table 2. **Diversity comparison.** We count the number of unique hypernyms from WordNet [28] and words in dialogues and image captions. We filter out a hypernym if it appears less than ten times in both dialogues and image captions. # hyp and # word denote the number of hypernyms and the number of unique words, respectively.

pernyms; 3.4× and 3.2× more words, which implies that our dataset covers more variety of open-domain topics. In image captions, compared to PhotoChat and MMDD, our dataset includes 2.4× and 21.2× more hypernyms; 2.0× and 10.5× more words. Such statistics of image captions shows the diversity of images.

Moreover, our dataset also shows better diversity in terms of conversational skills. Conversational skill is a general ability to lead a conversation, which includes empathetic listening, giving knowledgable responses, getting to know each other, talking about daily life topics, and blending all of these skills. As shown in Figure 7, MMDD covers three types of conversational skills mainly related to persona, while our dataset contains five types of conversational skills without overly focusing on one particular skill. Those skills enable a multi-modal dialogue model trained from our dataset to create engaging, vivid, and various dialogues.

(3) Generalization: In real-life scenarios, people can share images with different styles, views, or objects for the same dialogue and context. However, as shown in Table 1, the existing datasets include few images per dialogue and utterance. This does not reflect real-life scenarios and can

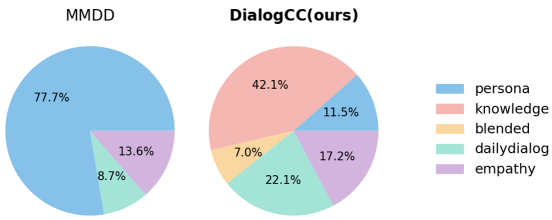


Figure 7. **Conversational Skills in DialogCC vs. MMDD [20].** We count the number of dialogues corresponding to the conversational skills. DialogCC covers two more various conversational skills (knowledge and blended skills) than MMDD.

cause an overfitting problem by forcing a model to memorize the pairs of images and dialogues. To handle this problem, our dataset has many and various images per dialogue and utterance, which is shown in Figure 3 and Figure 6. There are an average of 34 images per dialogue and 4.1 images per utterance. Training a model with our dataset would enhance the generalization performance, which is experimentally shown in Section 4.2.3.

4. Experiments

To explore how our dataset affects both text and image retrieval tasks, we evaluate two baseline models: text retrieval model used to evaluate MMDD [20] and image retrieval model used to evaluate PhotoChat [49].

4.1. Experimental Setting

4.1.1 Task Definition

We explain the formulation of two main tasks - text retrieval [20] and image retrieval [49]. In the text retrieval task, two different sub-tasks exist which are current turn prediction and next turn prediction. Let us assume that we have a multi-modal dialogue $\mathcal{D} = \{(u_j, i_j, c_j)\}_1^N$ where N denotes the number of dialogue turns, and $j = t$ is the turn that an image sharing behavior occurs. Then, each task is formulated as follows.

Current turn prediction is to predict the current response at turn t given the dialogue history ($\{u_j\}_1^{t-1}$) and image i_t.
Next turn prediction is to predict the next utterance at turn $t+1$ given the dialogue history ($\{u_j\}_1^{t}$) and image i_t.
Image retrieval is to retrieve relevant image at turn t given the dialogue history ($\{u_j\}_1^{t-1}$).

4.1.2 Baseline Model

As a baseline, we choose three models - BM25, text retrieval, and image retrieval. Followings are brief descriptions of each baseline model and more detailed information is provided in Appendix.

Models ↓	Task →	Current Turn Prediction		Next Turn Prediction	
	Eval →	MMDD	DialogCC	MMDD	DialogCC
	Train ↓	R@1	R@1	R@1	R@1
BM25	-	3.62	0.72	3.72	0.83
Text Retrieval	MMDD	5.28	2.65	2.93	1.56
	DialogCC	**6.64**	**10.17**	**4.31**	**11.07**

Table 3. **Text retrieval performance.** We compare the Recall@1 (%) performance of text retrieval model trained on DialogCC (when the maximum number of image is 10) and MMDD.

BM25 [34] retrieves response for the text retrieval task and image for the image retrieval task using captions.
Text retrieval [20] consists of a dialogue encoder, response encoder, image encoder, and multi-modal encoder. We use the sum module that adds two vectors element-wise for the multi-modal encoder.
Image retrieval [49] has a dual-encoder structure which consists of a dialogue encoder, photo description encoder, and image encoder.

4.1.3 Datasets

DialogCC (ours) is a large-scale multi-modal dialogue dataset created by the CLIP-based automatic pipeline described in Sec. 3.
MMDD [20] contains 45k multi-modal dialogues, where each utterance is replaced into a relevant image matched by their automatic pipeline.
PhotoChat [49] contains 10k multi-modal dialogues, where the dialogue is constructed via a crowd-sourcing platform.

4.1.4 Implementation Details

We implement baseline models based on PyTorch Lightning. All experiments are conducted on two A100 GPUs (40GB). To accelerate the training time, we apply distributed training to baselines. We follow the hyperparameter settings similar to the previous works [20, 49], which are described as follows:
Text retrieval. For the text encoder, we use the BERT-based architecture (12 layers, 12 attention heads, 768 dimensions, uncased version). For the image encoder, we use the ResNeXt-101 model (2048 dimensions). We use a negative log likelihood loss with in-batch negative samples, same as [20, 38]. In our experiment, we set the batch size to 32, the learning rate to 5e-5, and the gradient clipping value to 2.0. We use the Adam optimizer [16] without any learning rate scheduler. The maximum length of dialogue context and response is 150 and 30.
Image retrieval. We use the same BERT-based architecture and ResNet-152 model (2048 dimensions) as the image encoder. We use a hinge-based triplet ranking loss [9, 21, 49] with the hardest negatives, and we set the margin parameter to 0.2. We set the batch size to 128. We also use the Adam optimizer with initial learning rate to 5e-5 and decaying 0.1% at every 1,000 steps. We truncate the length of dialogue context and photo description longer than 128.
Training. Since our dataset contains several images per utterance, we randomly choose one image in each batch. For the text retrieval, we do not update the parameter of the image encoder as it helps achieve the best performance on the text retrieval task [20]). On the other hand, we update the parameters of all encoders in the image retrieval model. In the validation steps, for the memory efficiency and fast computation speed, we constitute the number of candidates as 100 for all retrieval tasks, which is the same setting in [20].
Inference. The settings of inference stage is almost the same as those of validation steps, except that the inference stage uses the entire test set as candidates rather than using only 100 of them.

Models ↓	Eval →	PhotoChat	DialogCC
	Train ↓	R@1	R@1
BM25	-	6.8	0.36
Image Retrieval	PhotoChat	7.15	1.35
	DialogCC	**7.58**	**14.85**

Table 4. **Image retrieval performance.** We report the Recall@1(%) performance on PhotoChat and DialogCC datasets.

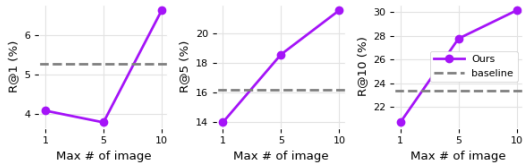


Figure 8. **Effect of maximum number of images.** We show the performance of the current turn prediction task by training the model with multiple images per utterance. In general, training with multiple images mostly improves performance.

4.2. Experimental Results

4.2.1 Text Retrieval Performance

We conduct an experiment by differentiating training and evaluation datasets to observe whether our dataset can boost performance in text retrieval tasks. As shown in Table 3, the model trained on MMDD performs poorly when evaluated on our dataset, implying that MMDD cannot help the model to understand various forms of images with similar semantic information. On the other hand, the model trained with our DialogCC achieves the best performance than the model trained with MMDD in all text retrieval tasks. This result indicates that our dataset improves the performance

	Current Turn Prediction		Next Turn Prediction	
Train → Aug ↓	MMDD R@1 (Δ)	DialogCC R@1 (Δ)	MMDD R@1 (Δ)	DialogCC R@1 (Δ)
Baseline	5.28	**6.64**	2.93	**4.31**
Shearing	3.75 (1.53)	**5.79 (0.85)**	1.01 (1.92)	**4.26 (0.05)**
Gaussian noise	3.40 (1.88)	**4.89 (1.75)**	1.05 (1.88)	**4.30 (0.01)**
Gaussian blur	3.91 (1.37)	**5.56 (1.08)**	1.05 (1.88)	**4.30 (0.01)**

Table 5. **Robustness comparisons on image augmentation.** We show the degree of decreased Recall@1(%) performance of the text retrieval model compared to the baseline score. We present detailed information on applied augmentation techniques in Appendix.

in open-domain conversation, which is benefited from large scalability, diversity, and images per dialogue.

4.2.2 Image Retrieval Performance

We also observe that training the image retrieval model with our dataset achieves the best performance on PhotoChat dataset, as shown in Table 4. However, the model trained on the PhotoChat dataset achieves lower performance when evaluated on DialogCC. This result indicates that our dataset also improves the performance in the image retrieval task, which is benefited from the largest number of images per dialogue and utterance.

4.2.3 Effect of Maximum Number of Images

We conduct an experiment to verify if learning with multiple images per utterance would be beneficial to model performance. Thus, we further evaluate the text retrieval performance by varying the maximum number of images to 1, 5, and 10. As shown in Figure 8, we confirm that the overall tendency of the model performance to increase the maximum number of images mostly increases across metrics (R@{1,5,10}). This demonstrates that showing various images per utterance enables the model to learn the semantic relationship between images and utterances rather than memorizing a pair of an image and an utterance, resulting in better generalization performance.

4.2.4 Robustness on Image Augmentation

To become a more generalized text retrieval model, the model should keep its performance as much as possible, even if the input image is distorted. Thus, we evaluate the performance of models trained on MMDD and DialogCC on the augmented MMDD dataset. We distort the input image of MMDD with several augmentation techniques, such as shearing and blurring provided by the imgaug [13]. In Table 5, the model trained with our dataset shows more robust performance even with input image variants, with a lower performance reduction compared to the model trained on MMDD dataset. This result indicates that our dataset makes the model robust to the input image distortion.

	Image Retrieval	
Train → Aug ↓	PhotoChat R@1 (Δ)	DialogCC R@1 (Δ)
Baseline	7.15	**7.58**
Synonym	5.02 (2.13)	**5.73 (1.85)**

Table 6. **Robustness comparisons on text augmentation.** We show the degree of decreased Recall@1(%) performance of the image retrieval model compared to the baseline score.

4.2.5 Robustness on Text Augmentation

To augment the dialogue, we replace randomly selected words (except stopwords) with synonyms [48]. Table 6 shows the performance of the model trained on our dataset are reduced less than before applying augmentation to input dialogue history. This results indicates that even if our dataset contains some noisy samples due to the automatic pipeline, the model trained on our dataset is more robust to the text distortion.

5. Conclusion

In this paper, we present the automatic pipeline for creating a multi-modal dialogue dataset that involves filtering with CLIP similarity. We also propose a large-scale multi-modal dialogue dataset, DialogCC, which is constructed by leveraging the automatic pipeline with five text-only dialogue datasets and an image-text pair CC3M dataset. In a comprehensive analysis, compared to existing datasets MMDD and PhotoChat, DialogCC contains a larger number of unique hypernyms, words, and conversational skills, which indicates better diversity. Moreover, our dataset consists of many and various images per dialogue that can be beneficial in model generalization performance. Extensive experiments demonstrate that a model trained with DialogCC achieves state-of-the-art performance in image and text retrieval tasks while increasing model robustness.

Societal Impact. As reported in [45], even if we give the gender-neutral query to CLIP [32] model, the CLIP model sometimes retrieves images causing gender-bias issues. We are concerned that this problematic issue may exist in our dataset because we use the CLIP model to match relevant images to given utterances. For example, most utterances related to the "hair designer" usually match images of women cutting hair. Therefore, the image retrieval model trained on our dataset may sometimes retrieve biased images. We should consider this problem important when building a multimodal search model.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 1

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1, 2

[3] Digbalay Bose, Rajat Hebbar, Krishna Somandepalli, Haoyang Zhang, Yin Cui, Kree Cole-McLaughlin, Huisheng Wang, and Shrikanth Narayanan. Movieclip: Visual scene recognition in movies. *arXiv preprint arXiv:2210.11065*, 2022. 4

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 1, 3

[5] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with clip reward. *arXiv preprint arXiv:2205.13115*, 2022. 4

[6] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017. 1, 2

[7] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021. 1, 3

[8] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*, 2018. 3

[9] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 7

[10] Kevin Frans, Lisa B Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *arXiv preprint arXiv:2106.14843*, 2021. 4

[11] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 4

[12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1

[13] Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. imgaug. https://github.com/aleju/imgaug, 2020. Online; accessed 01-Feb-2020. 8

[14] Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. Will i sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. *arXiv preprint arXiv:2004.05816*, 2020. 2

[15] Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. *arXiv preprint arXiv:2109.08828*, 2021. 2

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[17] Byungsoo Ko and Geonmo Gu. Large-scale bilingual language-image contrastive learning. *arXiv preprint arXiv:2203.14463*, 2022. 1

[18] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. *arXiv preprint arXiv:1903.03166*, 2019. 1, 2

[19] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 2

[20] Nyoungwoo Lee, Suwon Shin, Jaegul Choo, Ho-Jin Choi, and Sung-Hyun Myaeng. Constructing multi-modal dialogue dataset by replacing text with semantically relevant images. *arXiv preprint arXiv:2107.08685*, 2021. 1, 2, 3, 4, 5, 6, 7

[21] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4654–4662, 2019. 2, 3, 7

[22] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*, 2017. 3

[23] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053*, 2022. 4

[24] Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*, 2019. 2

[25] Hua Lu, Zhen Guo, Chanjuan Li, Yunyi Yang, Huang He, and Siqi Bao. Towards building an open-domain dialogue system incorporated with internet memes. *arXiv preprint arXiv:2203.03835*, 2022. 2

[26] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 1, 3

[27] Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. Openvidial: A large-scale, open-domain dialogue dataset with visual contexts. *arXiv preprint arXiv:2012.15015*, 2020. 1, 2

[28] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 5, 6

[29] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*, 2017. 1, 2

[30] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. *arXiv preprint arXiv:2204.00679*, 2022. 3

[31] Ramakanth Pasunuru and Mohit Bansal. Game-based video-context dialogue. *arXiv preprint arXiv:1809.04560*, 2018. 1

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 3, 8

[33] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*, 2018. 2, 3

[34] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009. 7

[35] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1

[36] Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. Visual reference resolution using attention memory for visual dialog. *Advances in neural information processing systems*, 30, 2017. 2

[37] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 1, 3

[38] Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. Image chat: Engaging grounded conversations. *arXiv preprint arXiv:1811.00945*, 2018. 1, 2, 7

[39] Kurt Shuster, Eric Michael Smith, Da Ju, and Jason Weston. Multi-modal open-domain dialogue. *arXiv preprint arXiv:2010.01082*, 2020. 1, 2

[40] Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. Can you put it all together: Evaluating conversational agents' ability to blend skills. *arXiv preprint arXiv:2004.08449*, 2020. 3

[41] Junhyuk So, Changdae Oh, Minchul Shin, and Kyungwoo Song. Multi-modal mixup for robust fine-tuning. *arXiv preprint arXiv:2203.03897*, 2022. 4

[42] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449, 2021. 1

[43] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 3

[44] Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. Multimodal dialogue response generation. *arXiv preprint arXiv:2110.08515*, 2021. 2

[45] Jialu Wang, Yang Liu, and Xin Eric Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv preprint arXiv:2109.05433*, 2021. 8

[46] Shuhe Wang, Yuxian Meng, Xiaoya Li, Xiaofei Sun, Rongbin Ouyang, and Jiwei Li. Openvidial 2.0: A larger-scale, open-domain dialogue generation dataset with visual contexts. *arXiv preprint arXiv:2109.12761*, 2021. 1, 2

[47] Shuhe Wang, Yuxian Meng, Xiaofei Sun, Fei Wu, Rongbin Ouyang, Rui Yan, Tianwei Zhang, and Jiwei Li. Modeling text-visual mutual dependency for multi-modal dialog generation. *arXiv preprint arXiv:2105.14445*, 2021. 1, 2

[48] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019. 8

[49] Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. *arXiv preprint arXiv:2108.01453*, 2021. 1, 2, 5, 6, 7

[50] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018. 2, 3

[51] Yinhe Zheng, Guanyi Chen, Xin Liu, and Ke Lin. Mmchat: Multi-modal chat dataset on social media. *arXiv preprint arXiv:2108.07154*, 2021. 1, 2

DialogCC: Large-Scale Multi-Modal Dialogue Dataset

Supplementary Material
github.com/passing2961/DialogCC

Contents

A. Details of Multi-Modal Dialogue Dataset Creation

A.1. Source Data Collection

This section describe how we collect the source dialogue and image-caption pairs used in creating DialogCC.

A.1.1 Source Dialogue Collection

We collect five text-only dialogue datasets (i.e., Wizard-of-Wikipedia [6], Persona-Chat [31], EmpatheticDialogues [20], DailyDialog [13], and BlendedSkillTalk [25]) through the ParlAI [16] framework, which provides many dialogue datasets online.

A.1.2 Source Image-Caption Collection

We download the Conceptual Captions 3M [24] (CC3M) dataset in here [1]. Since the CC3M dataset provides image URLs, we download images using img2dataset [2] library, which is a helpful library for quick downloading large-scale images based on URLs. We store downloaded images as a webdataset [3] format for efficiently extracting visual features by the CLIP [19] model. Note that because each image URL has the copyright, we only use opened URLs as source image-caption data when we create DialogCC.

A.2. Detailed Analysis of DialogCC

In this section, we provide a detailed analysis of DialogCC in terms of the *scalability*, *diversity*, and *generalization*.

A.2.1 Scalability

Table A shows the full statistics of DialogCC compared to existing datasets MMDD [12] and PhotoChat [30]. As we aforementioned, DialogCC comprises of 92k unique dialogues in total, which is roughly 3.7× and 7.7× larger than MMDD and PhotoChat, respectively. Although the number of dialogue turn on average is shorter than existing datasets, the utterance length is longer than others, which indicates that our dataset contains more specific utterances that can increase the interestingness and engaingness [23]. For example, given the utterance "how are you?", the response "I'm really good, because I'm going to paris today!" is more close to specific utterance than the response of "I'm good". Thus, the multi-modal dialogue generative model trained on our dataset can generate more specific responses, making conversation more attractive.

[1] https://ai.google.com/research/ConceptualCaptions/download

[2] https://github.com/rom1504/img2dataset

[3] https://github.com/webdataset/webdataset

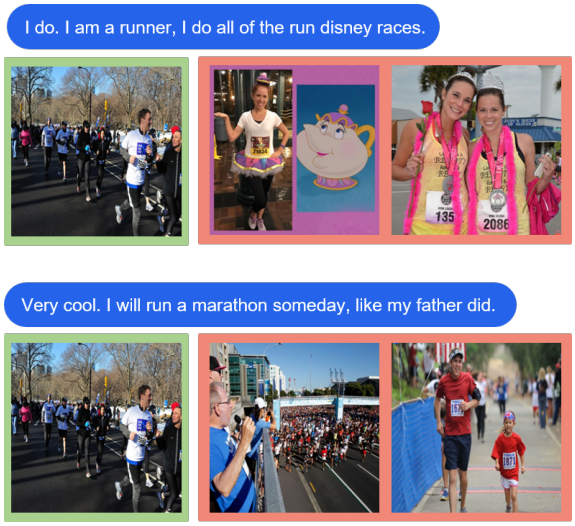


Figure A. **Problematic Examples in MMDD [12].** Even if the two given utterances have different semantic meanings, MMDD matches the same image. Hence, training a multi-modal dialogue model with MMDD will likely memorize this green image rather than a deep understanding of input utterances. Images in the green box are from MMDD, images in the red box are from DialogCC, and the blue box is utterances.

Furthermore, we provide the detailed statistics of DialogCC according to the source dialogue datasets, as shown in Table B. Overall, the KnowledgeCC dataset includes the largest number of images per utterance and dialogue than other source dialogue datasets and the EmpathyCC has the smallest number. This result indicates that the CLIP similarity between image and utterance, which contains more object information (e.g., dog, bus, soccer), is relatively higher than the similarity between image and utterance related to the emotional situation.

A.2.2 Diversity

As shown in Section 3.2, DialogCC contains the largest number of unique hypernyms and unique words in both dialogues and image captions. In addition, we show that our dataset covers various conversational skills with balanced distribution, as illustrated in Figure 7 in our main paper. We furthermore compare the diversity of datasets in terms of the part-of-speech (POS) by using the pos tagger provided by the spaCy [4]. In total, compared to PhotoChat and MMDD, our dataset includes 4.5× and 2.2× more noun words; 13.4× and 3.2× more verb words; 1.7× and 2.7× more adjective words, which suggests that our dataset covers more variety of words.

[4] https://spacy.io/

Datasets	Type	# Unique Dialog	# Utter	Avg. # Utter/Dialog	Avg. # Token/Utter	# Image	# Unique Image	Avg. # Image/Dialog	Avg. # Image/Utter	Avg. # Utter/Image
MMDD [12]	train	21,411	519,728	13.01	11.97	39,956	12,272	3.32	1.87	3.26
	valid	2,400	32,708	13.62	11.87	2,401	334	1.13	1.0	7.19
	test	2,672	36,315	13.59	11.85	2,673	682	1.13	1.0	3.92
	total	26,483	588,751	13.07	11.96	45,030	13,288	3.42	1.7	3.39
PhotoChat [30]	train	10,286	130,459	12.68	6.33	10,286	8,889	1.0	1.0	1.16
	valid	1,000	12,695	12.7	6.31	1,000	1,000	1.0	1.0	1.0
	test	1,000	12,841	12.84	6.29	1,000	1,000	1.0	1.0	1.0
	total	12,286	155,995	12.7	6.33	12,286	10,889	1.0	1.0	1.13
DialogCC (ours)	train	77,354	6,362,844	9.89	14.18	2,492,320	467,944	32.22	3.88	5.33
	valid	7,940	674,708	10.34	13.61	341,648	92,723	43.03	5.24	3.68
	test	7,648	644,968	10.3	13.77	325,627	91,198	42.58	5.2	3.57
	total	**92,942**	**7,682,520**	**9.97**	**14.09**	**3,159,595**	**651,840**	**34.0**	**4.1**	**4.85**

Table A. **Full Statistics of Datasets.** Compared to existing datasets MMDD and PhotoChat, DialogCC includes the largest number of unique dialogues and images.

Datasets	Type	# Unique Dialog	# Utter	Avg. # Utter/Dialog	Avg. # Token/Utter	# Image	# Unique Image	Avg. # Image/Dialog	Avg. # Image/Utter	Avg. # Utter/Image
KnowledgeCC	train	35,252	2,368,960	8.32	16.5	1,386,041	305,656	39.32	4.87	4.53
	valid	1,952	131,556	8.33	16.53	132,398	42,352	67.83	8.39	3.13
	test	1,917	127,684	8.29	16.53	129,317	41,580	67.46	8.39	3.11
	total	**39,121**	**2,628,200**	**8.32**	**16.5**	**1,647,756**	**389,574**	**42.12**	**5.22**	**4.23**
PersonaCC	train	8,763	1,924,764	14.92	11.67	268,399	145,069	30.63	2.08	1.85
	valid	1,000	244,500	15.67	11.92	54,333	31,169	54.33	3.48	1.74
	test	967	234,096	15.6	11.77	45,409	27,148	46.96	3.03	1.67
	total	**10,730**	**2,403,360**	**15.06**	**11.71**	**368,141**	**203,382**	**34.31**	**2.31**	**1.81**
EmpathyCC	train	11,830	209,280	4.26	13.73	130,441	84,072	11.03	2.65	1.55
	valid	2,151	38,280	4.27	14.62	27,739	19,597	12.9	3.1	1.42
	test	1,993	35,412	4.27	15.59	25,991	18,438	13.04	3.14	1.41
	total	**15,974**	**282,972**	**4.26**	**14.08**	**184,171**	**122,104**	**11.53**	**2.77**	**1.51**
DailyCC	train	16,934	1,240,428	9.69	14.11	550,047	161,406	32.48	4.3	3.41
	valid	1,845	128,612	9.4	13.67	68,686	24,760	37.23	5.02	2.77
	test	1,807	119,484	9.2	14.14	72,060	28,214	39.88	5.55	2.55
	total	**20,586**	**1,488,524**	**9.62**	**14.07**	**690,793**	**214,375**	**33.56**	**4.47**	**3.22**
BlendedCC	train	4,575	619,412	11.83	13.34	157,392	111,702	34.4	3.01	1.41
	valid	992	131,760	11.75	13.48	58,492	37,133	58.96	5.22	1.58
	test	964	128,292	11.77	13.85	52,850	34,590	54.82	4.85	1.53
	total	**6,531**	**879,464**	**11.81**	**13.44**	**268,734**	**183,418**	**41.15**	**3.61**	**1.47**

Table B. **Detailed Statistics of DialogCC.** We show the statistics on the results of matching each source dialogue dataset and CC3M. KnowledgeCC, PersonaCC, EmpathyCC, DailyCC, and BlendedCC denote the multi-modal dialogue dataset created from the Wizard-of-Wikipedia [6], Persona-Chat [31], EmpatheticDialogues [20], DailyDialog [13], and BlendedSkillTalk [25], respectively.

	Dialog			Image Caption			Total		
	# noun	# verb	# adj	# noun	# verb	# adj	# noun	# verb	# adj
MMDD	6,696	13,369	4,138	6,808	1,749	3,279	13,504	15,118	7,417
PhotoChat	5,240	3,270	10,918	1,529	412	377	6,769	3,682	11,295
DialogCC (Ours)	**15,976**	**35,522**	**11,652**	**14,822**	**13,921**	**8,567**	**30,798**	**49,443**	**20,219**

Table C. **Part-of-Speech (POS) Comparison Results.** We count the number of the unique noun, verb, and adjective words in dialogues and image captions. # noun, # verb, and # adj denote the number of the unique noun word, the number of the unique verb words, and the number of the unique adjective words, respectively.

A.2.3 Generalization

In real-life conversations, various images can be shared even with the same utterance, depending on who shares the image. As shown in Table A, DialogCC contains larger images per utterance and dialogue than existing datasets, which indicates that our dataset can successfully reflect this phenomenon. For example, as shown in Figure B, our dataset includes various images with similar semantic meanings to the given utterance, which can induce the model to be more robust than a model trained with MMDD. In addition, MMDD contains the same images on two different utterances, as illustrated in Figure A. For example, in the first row of Figure A, the image in our dataset is relevant to the keyword "disney races" in the utterance, while MMDD cannot. This result will induce the degradation of generalization performance because the model trained on MMDD may prefer to memorize specific images rather than

Figure B. **Comparison Examples of Relevant Images in MMDD [12] vs. DialogCC.** DialogCC contains various forms of images that are semantically similar to a given utterance. For example, in the top three rows, our dataset includes different kinds of "landscape", "a pair of shoes", and "two horses". Besides, in the fifth row, unlike the MMDD, our dataset contains highly relevant images to the given utterance, which benefitted from the utterance-caption similarity. Images in the green box are from MMDD, images in the red box are from DialogCC, and the blue box is utterances.

understanding the dependency between image and utterance. We show that our dataset improves the generalization performance in Section 4.2.4, Section 4.2.5, Section C.3, and Section C.4.

A.3. Case Studies of DialogCC

A.3.1 Comparison examples to MMDD

Unlike MMDD [12], we create DialogCC by using utterance-caption similarity as well as utterance-image similarity to improve the quality of DialogCC. To compare the different quality of multi-modal dialogue dataset, we present more comparison examples to MMDD in Figure B. In the last row of Figure B, our dataset includes more diverse images related to the "sneakers", which are in the given utterance.

A.3.2 Examples of DialogCC

We present more examples of DialogCC in Figure G and Figure D. Both Figure G and Figure D show that DialogCC covers a variety of images, including various objects or scenes which are relevant to the given utterance, regardless of whether the conversation is short or long. For example, there are diverse images of graduation ceremonies, christmas, traffic jams, and hiking with a dog.

However, we sometimes found poor-quality examples of DialogCC, as illustrated in Figure F. On the second turn in the red box with a dotted line, the images are related to the given utterance that includes the word or phrase "pretty animal" or "arctic". However, when we look at the previous utterance "Yes, I love huskys. They are a very pretty dog.", we can easily recognize that these matched images are inappropriate to the given utterance considering the whole dialogue context. This is because when we calculate two similarities between utterance-caption and utterance-image by using CLIP model, we do not extract dialogue feature vectors or calculate the similarity by considering the whole dialogue context. Such a problem also exists in MMDD, which can be regarded as a typical problem of automatic methods [2, 12, 17] that can be utilized in creating several datasets. Nevertheless, recent studies have shown large-scale multi-modal pretraining models [1, 9, 11, 15, 19] have achieved enormous performance on various downstream tasks, which benefitted from the noisy large-scale image-caption datasets [3, 4, 21, 22, 26]. Therefore, this paper aims to create a large-scale multi-modal dialogue dataset, even if there is some noise, rather than creating a cleaned dataset.

B. Details of Experimental Settings

B.1. Baseline Models

As illustrated in Figure C, we present the architecture of baseline models, which is the text retrieval model [12] and image retrieval model [30]. We provide a detailed description of baseline models below.

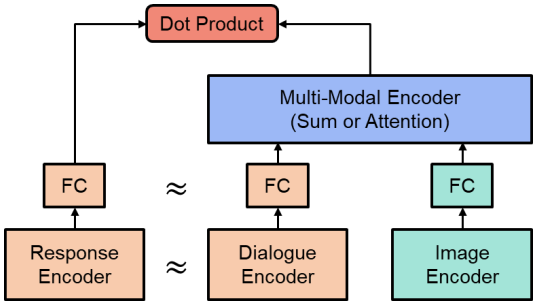
Text Retrieval Model [12]. The text retrieval model consists of the dialogue, response, image, and multi-modal encoder. The dialogue encoder encodes the whole dialogue history into a fixed-size representation using the BERT [5] model. The response encoder converts the response into a fixed-size representation using the BERT model, the same BERT used in the dialogue encoder. They use the pooled vector from the BERT model to get the representations for dialogue history and response. When they encode the dialogue history, they use up to three turns just before the current turn by concatenating each turn with `[SEP]` special token. The image encoder is to extract image feature vectors using ResNeXt-101 [29]. After extracting text and image features, these features are fed into the fully-connected layer with the ReLU [18] activation function. They use the sum and attention modules to make a fused representation between the dialogue history and image. The sum module adds two vectors in an element-wise manner. They first concatenate the dialogue history vector and image feature vector for the attention module. They then pass the concatenated vector into the Transformer [27] to obtain contextualized multi-modal representation. Lastly, they calculate the dot product between the response feature vector and the multi-modal feature vector to get the loss.

Image Retrieval Model [30]. The image retrieval model consists of a dialogue encoder, a caption encoder [5], and an image encoder. The dialogue encoder and caption encoder leverage two different BERT models, which means the parameters of the two BERT models are not shared. They use the contextualized representation vector corresponding to the position of `[CLS]` for dialogue history and a caption. For the image encoder, they use ResNet-152 [8]. They concatenate two feature vectors for image and caption to obtain a meaningful multi-modal representation. They then pass the concatenated vector into the fully-connected layers with the ReLU function. They encode the dialogue history into a fixed-size representation and pass it to the fully-connected layers with the ReLU function. They calculate the cosine similarity by computing the dot product of L2 normalized features for dialogue history and multi-modal.

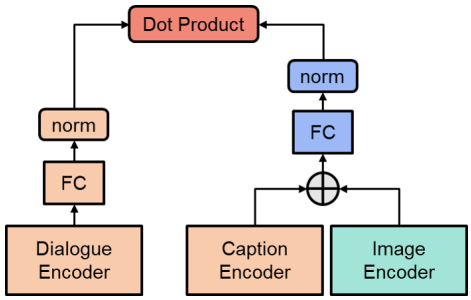
B.2. Implementation Details

As we described the implementation details in our main paper, we explain additional implementation details for the extended experiments. Since our dataset is much larger than

[5]The original paper calls it a photo description encoder because PhotoChat contains the photo description per photo (image). In this paper, we call it the caption encoder.



(a) An architecture of text retrieval model



(b) An architecture of image retrieval model

Figure C. **Architectures of Baseline Models.** We show the architecture of the text retrieval model [12] and image retrieval model [30], which are used in the experiment. We color the part corresponding to getting text representations as a light orange, the part corresponding to getting visual representation as a light green, and the part corresponding to getting a multi-modal representation as a light blue. The light red is the dot product.

the PhotoChat dataset, we use the cosine annealing schedule [14] with the same step size of 1,000. In all extended experiments, we train the models longer to achieve the best performance. We use the RandomResized cropping technique for image augmentation in both image retrieval and text retrieval experiments.

Image Augmentation Techniques For the experiment of robustness on image augmentation, we use the imgaug [10] library for image augmentation. We adopt eight image augmentation techniques, rotation, zoom-in, zoom-out, cutout, dropout, shearing, gaussian noise, and gaussian blur. We follow the same setting of each technique of the previous work [7].

Text Augmentation Techniques For the experiment of robustness on text augmentation, we use the synonym replacement techniques introduced in EDA [28].

C. Extended Experiments

In this section, we show the extended experiments of the text retrieval and image retrieval models.

C.1. Text Retrieval

Table D shows the text retrieval performance across all evaluation metrics, which is reported in our main paper. We train the model with more epochs to achieve improved performance. In addition, we conduct the ablation studies by applying different multi-modal encoders (i.e., sum and attention) and by changing model inputs, such as only providing the image or dialogue to the model, as shown in Table G. Even we further training, we observe a similar tendency that the model trained on our dataset improves the performance of MMDD. However, training the model with MMDD achieves considerably poor performance on our dataset. This result indicates that our dataset is more challenging than MMDD due to the most significant number of images per utterance and dialogue. Moreover, such lack of diversity in MMDD induces the model to memorize seen images in the training process. The following paragraph explains the effect of different multi-modal encoders and model inputs.

Effect of Multi-Modal Encoder. In the model trained on MMDD, the sum module achieves better performance on two text retrieval tasks than the attention module, similar to the results reported in [12]. However, in the model trained on DialogCC, the attention module performs better on current turn prediction task. This result indicates that the attention module benefitted from the scalability and diversity of our dataset. In addition, the next turn prediction task is more challenging than the current turn prediction task due to the overall lower performance.

Effect of Model Inputs. To understand which modality is important in the text retrieval task, we conduct another experiment by giving different model inputs, such as only the dialogue history, only the image, or both the dialogue history and the image. Overall, using both modalities as input achieves the best performance. We also observe that considering dialogue history is important in the multi-modal dialogue retrieval tasks due to the lower performance of "Image Only".

C.2. Image Retrieval

Table E shows the image retrieval performance across all evaluation metrics, which is reported in Table 4 in our main paper. Since our dataset is larger and more diverse than PhotoChat, we need to train the image retrieval model with more epochs and vary the learning rate to achieve the best performance. Thus, we adopt the cosine annealing scheduler with a step size of 1,000. As shown in Table F, we

Models↓	Task →	Current Turn Prediction						Next Turn Prediction					
	Eval →	MMDD			DialogCC			MMDD			DialogCC		
	Train ↓	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
BM25	-	3.62	7.91	10.46	0.72	4.48	8.05	3.72	8.34	12.58	0.83	4.61	8.53
Text Retrieval	MMDD	5.28	16.18	23.34	2.65	7.73	11.29	2.93	9.6	12.13	1.56	4.93	7.23
	DialogCC	**6.64**	**21.57**	**30.16**	**10.17**	**28.74**	**39.17**	**4.31**	**12.58**	**17.69**	**11.07**	**29.08**	**37.18**

Table D. **Text Retrieval Performance.** We compare the performance of text retrieval model trained on DialogCC (when the maximum number of image is 10) and MMDD across all metrics.

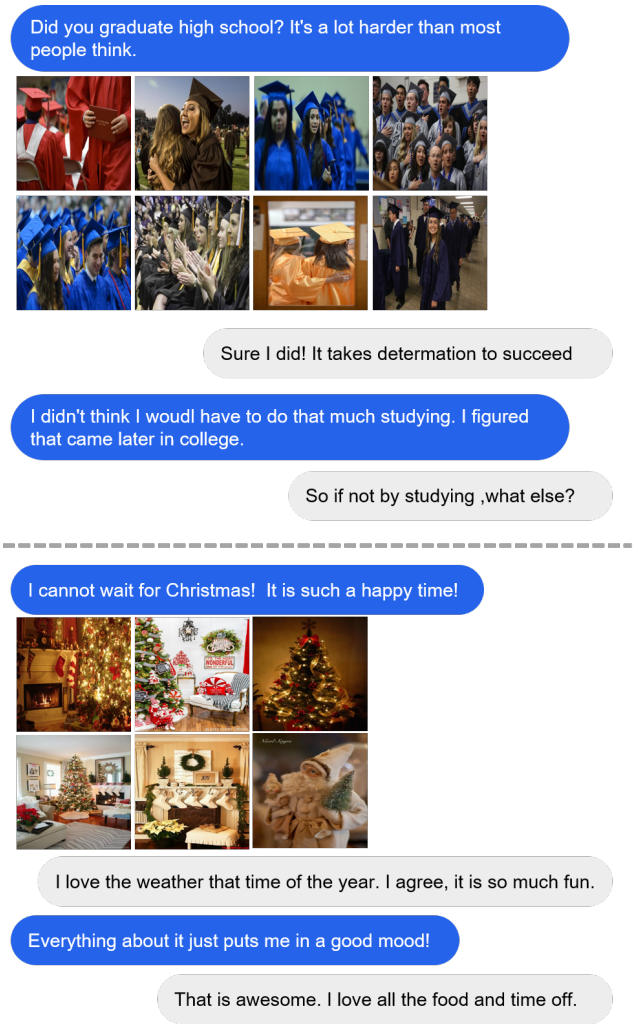


Figure D. **Case 1: Examples of DialogCC.** We present examples of DialogCC, and a gray dotted line separates each example. As illustrated in both examples, our dataset covers diverse images which are related to the utterance.

achieve the better performance than 7.58 (in Table E), which indicates that our dataset can make the model's training difficult due to the scalability and diversity. It will be helpful to improve the generalization performance.

Models ↓	Eval →	PhotoChat			DialogCC		
	Train ↓	R@1	R@5	R@10	R@1	R@5	R@10
BM25	-	6.8	15.9	22.5	0.36	1.21	1.77
Image Retrieval	PhotoChat	7.15	**25.02**	**37.20**	1.35	5.49	8.70
	DialogCC	**7.58**	17.52	25.00	**14.85**	**36.33**	**48.93**

Table E. **Image Retrieval Performance.** We report the image retrieval performance on PhotoChat and DialogCC datasets, which is reported in Table 4 in our main paper.

Eval →	PhotoChat			DialogCC		
Train ↓	R@1	R@5	R@10	R@1	R@5	R@10
PhotoChat	8.06	**23.76**	**38.02**	0.17	0.80	1.38
DialogCC (1)	7.03	18.70	26.86	2.45	9.30	15.19
DialogCC (5)	8.68	18.70	26.55	3.02	11.16	17.89
DialogCC (10)	7.44	18.08	26.34	2.95	11.17	18.04
DialogCC (15)	**8.99**	19.42	28.20	**3.37**	**11.61**	**18.12**
DialogCC (20)	8.68	18.70	26.14	2.92	11.15	18.01
DialogCC (30)	8.16	18.60	26.96	3.12	11.18	17.75

Table F. **Extended Image Retrieval Performance.** We report the image retrieval performance on PhotoChat and DialogCC dataset, where we train the image retrieval model with DialogCC by adopting the cosine annealing schedule. The number in parenthesis denotes the maximum number of images used in the training process.

However, the model trained on PhotoChat achieves better on R@5 and R@10. There are two possible reasons. The first reason is that the correlation between the training and test sets is larger in PhotoChat than in DialogCC. If the model trained on PhotoChat is well-generalized, it should have performed well in DialogCC, but it does not. On the other hand, training the model with our dataset shows high performance in both PhotoChat and DialogCC. The second reason is that the covered images in training, validation, and test set have similar domains in PhotoChat. PhotoChat only covers commonly shared objects, such as people, food, animal, and product. However, DialogCC is an open domain, which means that DialogCC covers many topics and conversational skills described in Section 3.2. This diversity would have decreased performance compared to the model trained on PhotoChat, which focuses on a specific domain. As shown in Figure E and Table F, the model trained on

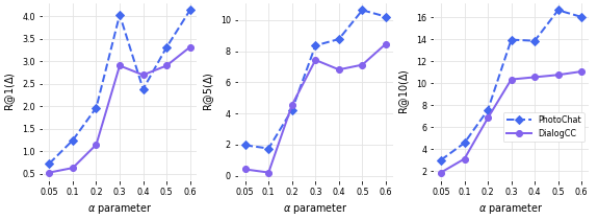


Figure E. **Performance Gap with Text Distortion.** We show the robustness performance of both models trained on our dataset or trained on PhotoChat by increasing the value of α. The y-axis denotes the performance gap between the baseline score (without augmentation) and the score (with augmentation). The higher the value of the y-axis, the less robust the trained model is. This graph shows that the model trained on our dataset mostly achieves better robustness across all α values.

DialogCC achieves better generalization performance than PhotoChat.

C.3. Robustness on Image Augmentation

To explore whether our dataset can make a well-generalized model, we conduct a robust experiment by distorting image input. As shown in Table H, the model trained on our dataset shows a more robust performance with a lower reduction than the model trained on the MMDD dataset. This result implies that our dataset induces the model to be more robust to image distortion, which benefitted from the diversity of our dataset.

C.4. Robustness on Text Augmentation

We evaluate the robustness of the model trained on our dataset by distorting input dialogue history. We replace randomly chosen words (except stopwords) with synonyms by adjusting the ratio of replacement α. As shown in Figure E, the model trained on our dataset shows a more robust performance when the value of α increases. This result indicates that while our dataset contains some noisy samples, our dataset makes the model more robust, which benefitted from the scalability and diversity of our dataset. Therefore, it is important to build a large-scale dataset to achieve improved generalization performance.

Figure F. **Case 3: Example of DialogCC that Unrelated to the Given Utterance.** In the red box with a dotted line, DialogCC contains diverse images that are relevant to the utterance. However, given the previous utterance, images should relate more to the "huskys" than many other animals, such as rabbits, polar bears, and penguins. Besides, it is more natural to share images related to "eskimo" or "arctic".

MME↓	Model Inputs↓	Task→	Current Turn Prediction						Next Turn Prediction					
		Eval→	MMDD			DialogCC			ACL			DialogCC		
		Train↓	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Sum	Image Only	MMDD	5.37	20.88	33.29	0.25	1.20	2.16	0.86	3.12	5.48	0.09	0.37	0.65
		DialogCC	3.90	15.67	24.90	1.01	4.27	6.94	1.67	5.82	9.07	0.23	0.92	1.61
	Dialogue Only	MMDD	8.55	19.53	26.53	2.64	9.47	14.91	7.02	15.79	21.22	3.52	10.45	14.96
		DialogCC	16.79	35.92	45.55	5.22	20.28	29.47	16.26	32.65	40.52	7.02	21.94	30.15
	Dialogue + Image	MMDD	13.72	35.68	49.48	0.95	3.95	6.70	6.80	15.28	20.63	2.41	8.15	12.25
		DialogCC	**15.16**	**37.51**	**49.64**	**6.31**	**21.63**	**30.90**	**13.35**	**30.81**	**39.62**	**6.64**	**21.22**	**29.51**
Attention	Image Only	MMDD	5.01	20.13	33.37	0.21	1.12	2.10	0.86	3.47	5.73	0.09	0.40	0.72
		DialogCC	4.97	16.99	26.93	1.08	2.94	6.74	1.37	6.29	9.76	0.25	1.13	2.04
	Dialogue Only	MMDD	8.83	19.81	26.93	1.90	7.57	12.01	6.76	15.32	20.92	3.28	10.45	15.55
		DialogCC	15.39	33.17	42.64	5.40	20.03	29.80	16.09	34.28	42.88	7.17	22.76	31.34
	Dialogue + Image	MMDD	12.97	33.13	45.55	0.92	3.57	6.23	4.66	12.11	17.37	2.24	7.23	10.54
		DialogCC	**19.49**	**44.03**	**55.69**	**6.87**	**22.26**	**31.58**	**13.61**	**30.51**	**38.55**	**5.04**	**15.76**	**22.71**

Table G. **Extended Text Retrieval Performance.** We report the text retrieval performance on MMDD and DialogCC. MME denotes the multi-modal encoder.

Train →	MMDD			DialogCC		
Aug ↓	R@1 (Δ)	R@5 (Δ)	R@10 (Δ)	R@1 (Δ)	R@5 (Δ)	R@10 (Δ)
Baseline	13.72	35.68	49.48	14.96	37.12	48.17
Rotation	10.46 (3.26)	30.27 (5.41)	41.89 (7.59)	**13.25 (1.71)**	32.62 (4.5)	43.16 (5.01)
Zoom-In	7.52 (6.2)	22.43 (13.25)	31.5 (17.98)	**9.83 (5.13)**	25.18 (11.94)	34.05 (14.12)
Zoom-Out	10.98 (2.74)	31.07 (4.61)	42.32 (7.16)	**13.6 (1.36)**	32.7 (4.42)	43.28 (4.89)
Cutout	11.14 (2.58)	31.11 (4.57)	43.56 (5.92)	**12.85 (2.11)**	32.54 (4.58)	43.44 (4.73)
Dropout	9.03 (4.69)	25.54 (10.14)	36.2 (13.28)	**10.5 (4.46)**	27.73 (9.39)	37.79 (10.38)
Shearing	9.67 (4.05)	27.69 (7.99)	38.27 (11.21)	**12.25 (2.71)**	31.19 (5.93)	41.25 (6.92)
Gaussian noise	10.46 (3.26)	28.4 (7.28)	39.62 (9.86)	**11.54 (3.42)**	30.11 (7.01)	40.37 (7.8)
Gaussian blur	10.78 (2.94)	29.44 (6.24)	40.69 (8.79)	**12.61 (2.35)**	32.34 (4.78)	42.8 (5.37)

Table H. **Robustness Performance on Image Distortion.** We report the robustness performance when input images are distorted.

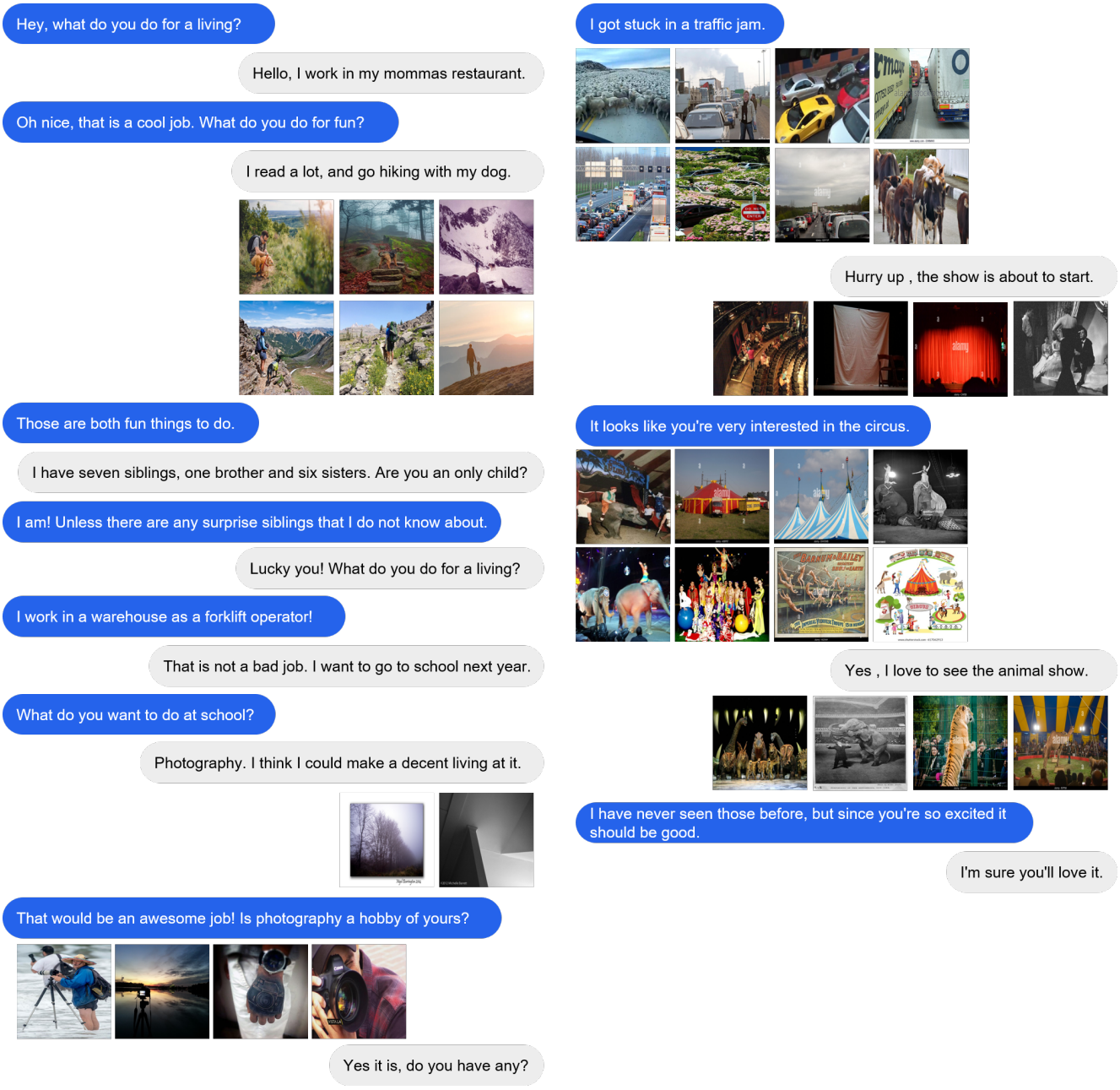


Figure G. **Case 2: Examples of DialogCC.** We present examples of DialogCC. Both examples also treat various images relevant to the utterance. For example, in the left figure, our dataset includes images related to an utterance in which the motion is revealed "go hiking with my dog".

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 5

[2] Digbalay Bose, Rajat Hebbar, Krishna Somandepalli, Haoyang Zhang, Yin Cui, Kree Cole-McLaughlin, Huisheng Wang, and Shrikanth Narayanan. Movieclip: Visual scene recognition in movies. *arXiv preprint arXiv:2210.11065*, 2022. 5

[3] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 5

[4] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021. 5

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[6] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*, 2018. 2, 3

[7] Geonmo Gu, Byungsoo Ko, and Han-Gyu Kim. Proxy synthesis: Learning with synthetic classes for deep metric learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1460–1468, 2021. 6

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[9] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 5

[10] Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. imgaug. https://github.com/aleju/imgaug, 2020. Online; accessed 01-Feb-2020. 6

[11] Byungsoo Ko and Geonmo Gu. Large-scale bilingual language-image contrastive learning. *arXiv preprint arXiv:2203.14463*, 2022. 5

[12] Nyoungwoo Lee, Suwon Shin, Jaegul Choo, Ho-Jin Choi, and Sung-Hyun Myaeng. Constructing multi-modal dialogue dataset by replacing text with semantically relevant images. *arXiv preprint arXiv:2107.08685*, 2021. 2, 3, 4, 5, 6

[13] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*, 2017. 2, 3

[14] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[15] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 5

[16] A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*, 2017. 2

[17] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. *arXiv preprint arXiv:2204.00679*, 2022. 5

[18] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010. 5

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 5

[20] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*, 2018. 2, 3

[21] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 5

[22] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5

[23] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*, 2019. 2

[24] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 2

[25] Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. Can you put it all together: Evaluating conversational agents' ability to blend skills. *arXiv preprint arXiv:2004.08449*, 2020. 2, 3

[26] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449, 2021. 5

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[28] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019. 6

[29] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 5

[30] Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. *arXiv preprint arXiv:2108.01453*, 2021. 2, 3, 5, 6

[31] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018. 2, 3