DialogCC: An Automated Pipeline for Creating High-Quality Multi-Modal Dialogue Dataset

Young-Jun Lee [1] **Byungsoo Ko** [2] **Han-Gyu Kim** [3] **Jonghwan Hyeon** [1] **Ho-Jin Choi** [1]
[1] School of Computing, KAIST [2] NAVER Vision [3] NAVER Cloud Multimodal AI
{yj2961, jonghwanhyeon, hojinc}@kaist.ac.kr
kobiso62@gmail.com hangyu.kim@navercorp.com

Abstract

As sharing images in an instant message is a crucial factor, there has been active research on learning an image-text multi-modal dialogue models. However, training a well-generalized multi-modal dialogue model remains challenging due to the low quality and limited diversity of images per dialogue in existing multi-modal dialogue datasets. In this paper, we propose an automated pipeline to construct a multi-modal dialogue dataset, ensuring both dialogue quality and image diversity without requiring minimum human effort. In our pipeline, to guarantee the coherence between images and dialogue, we prompt GPT-4 to infer potential image-sharing moments - specifically, the utterance, speaker, rationale, and image description. Furthermore, we leverage CLIP similarity to maintain consistency between aligned multiple images to the utterance. Through this pipeline, we introduce DialogCC, a high-quality and diverse multi-modal dialogue dataset that surpasses existing datasets in terms of quality and diversity in human evaluation. Our comprehensive experiments highlight that when multi-modal dialogue models are trained using our dataset, their generalization performance on unseen dialogue datasets is significantly enhanced. We make our source code and dataset publicly available [1].

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 (Antol et al., 2015; Das et al., 2017; Mostafazadeh et al., 2017; Shuster et al., 2018; Pasunuru and Bansal, 2018; Kottur et al., 2019; Meng et al., 2020; Zheng et al., 2021; Shuster et al., 2020). In practical situations, humans actively share images during conversations rather than merely talking about a given image, which is called *image-sharing* behavior (Lobinger, 2016). Recent studies for the image-sharing have proposed multi-modal dialogue datasets, which are constructed through the crowd-sourcing (PhotoChat (Zang et al., 2021)), image-text similarity with human efforts (MMDD (Lee et al., 2021)), or social media platform (MMDialog (Feng et al., 2022)).

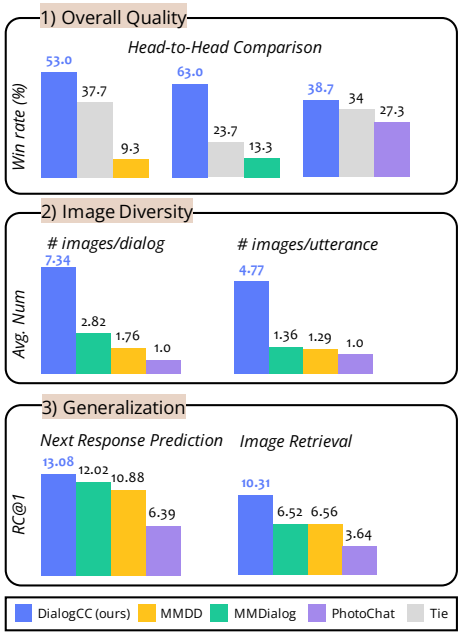


Figure 1: Comparing DialogCC (ours) to three existing multi-modal dialogue dataset in terms of a quality, diversity, and generalization. RC@1 denotes the averaged contributed R@1 performance.

However, existing multi-modal dialogue datasets have three significant limitations; **(1) Quality.** Recent studies have shown that a high-quality dataset enhances both the efficacy and the quality of the model training (Abbas et al., 2023; Zhou et al.,

[1] https://dialogcc.github.io/

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1938–1963
June 16-21, 2024

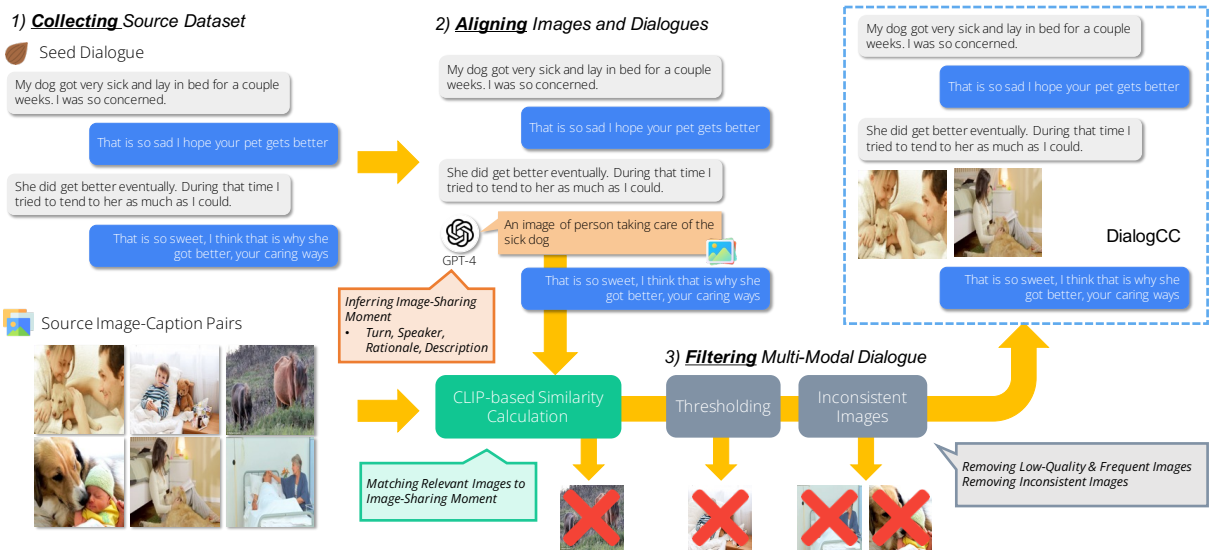


Figure 2: An overview of our proposed automatic pipeline for creating a high-quality and diverse multi-modal dialogue dataset.

2023). Nevertheless, as shown in Figure 1, existing datasets contain low-quality multi-modal dialogues (i.e., appearance of images in unnatural moments, inconsistency between the image and the context of the conversation) that hinder the training process of true multi-modal social dialogue agents. **(2) Diversity.** Given the same dialogue and context, people can share different types of images. For example, for an utterance of "*I love a dog*," one can share an image of a chihuahua, and the other can share an image of a poodle. Nonetheless, as shown in Figure 1 (*# images / dialog* and *# images / utterance*), existing datasets consist of less than the average 2.8 images per dialogue and the average 1.4 images per utterance. **(3) Generalization.** A model trained with conventional datasets can be overfitted by memorizing low-quality and limited pairs of images and dialogues, which can hinder its ability to handle unseen dialogue scenarios effectively by its lack of generalization. As shown in Figure 1, models trained on existing datasets show low performance on unseen dialogue datasets on both retrieval tasks. However, the model trained on our dataset achieves comparable performance, which benefited from the high quality and diversity.

This work aims to create a high-quality and diverse multi-modal dialogue dataset to train a well-generalized multi-modal dialogue model for open-domain conversation. To this end, we propose a fully automatic framework for creating a multi-modal dialogue dataset that involves three main steps: *collecting*, *aligning*, and *filtering*, as shown in Figure 2. After collecting source datasets, to ensure image-dialogue coherence, we ask GPT-4 (OpenAI, 2023) to infer all possible image-sharing moments via zero-shot prompting and leverage the CLIP (Radford et al., 2021) to increase the aligned image relevancy in the *aligning* step. In the *filtering* step, we eliminate inappropriate images based on CLIP similarity for image-image consistency. We propose a high-quality and diverse multi-modal dialogue dataset, DialogCC, constructed by our proposed pipeline without minimum human efforts, unlike the previous datasets. As illustrated in Figure 1, DialogCC achieves better statistics compared to the existing datasets in terms of quality, diversity, and generalization, indicating the effectiveness of our proposed pipeline. In addition, extensive experiments demonstrate that DialogCC can boost the generalization performance of trained models on unseen dialogue scenarios.

In summary, our main contributions are as follows: 1) We propose a fully automatic pipeline to create a multi-modal dialogue dataset that can achieve quality and diversity without human intervention. 2) We propose a high-quality and diverse multi-modal dialogue dataset named DialogCC, which contains various images per dialogue and utterance, respectively. 3) Extensive experiments demonstrate the effectiveness of our dataset, which enhances the generalization performance.

2 Related Work

Multi-Modal Dialogue Dataset. In the visual dialogue domain, most previous studies are divided into two categories depending on whether the im-

age is *grounded* or *sharing* in the dialogue. The image-grounded dialogue task aims to answer questions (Antol et al., 2015; Das et al., 2017; Seo et al., 2017; Kottur et al., 2019) or generate natural conversations (Mostafazadeh et al., 2017; Shuster et al., 2018; Meng et al., 2020; Wang et al., 2021b; Zheng et al., 2021) 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.

Image-Sharing Dialogue Dataset. 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 (Zang et al., 2021), is manually constructed through a crowd-sourcing platform with Open Image Dataset V4 (Kuznetsova et al., 2020) as source images. This dataset can provide a high-quality dialogue dataset, but the manual construction is time-consuming and expensive. Another line of work (Lee et al., 2021) creates a 45k multi-modal dialogue dataset by replacing an utterance with relevant images using image-text similarity, based on a threshold ensuring dialogue coherence as determined by human evaluation. Still, we need a human-in-the-loop process and the similarity of image and utterance result is not reliable in terms of the nature of dialogue context, such as coreference resolution. MMDialog dataset is a web-scale multi-modal dialogue dataset curated from a social media platform, but it lacks the natural conversational flow due to the nature of non-consecutive turn of social media interactions, resulting in highly low quality, which is also reported in the previous work (Han et al., 2023). All datasets cannot maintain both quality and diversity simultaneously, as demonstrated in Figure 1. Therefore, we construct a high-quality multi-modal dialogue dataset containing various images through the proposed 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 (Shuster et al., 2018; Lee et al., 2021; Zang et al., 2021). For the generative model, many works are based on the encoder-decoder architecture (Shuster et al., 2020; Wang et al., 2021c; Sun et al., 2021; Lu et al., 2022). Focusing on the *image-sharing* behavior, we train a cross-modal retrieval model on our dataset, highlighting potential future applications.

3 DialogCC

In this section, we propose DialogCC, a high-quality and diverse multi-modal social dialogue dataset. In order to construct DialogCC, we introduce an automatic pipeline, which consists of three steps: (1) *collecting*, (2) *aligning*, and (3) *filtering*. Besides, we conduct a comprehensive analysis of our dataset with respect to quality and diversity by comparing three existing datasets, MMDD (Lee et al., 2021), PhotoChat (Zang et al., 2021), and MMDialog (Feng et al., 2022). The overall pipeline is illustrated in Figure 2. In the following part of this section, we provide details about our proposed pipeline.

3.1 Collecting Source Dataset

Source Dialogue. As a source data, we collect five multi-turn text-only social dialogue datasets, which are publicly available online. Five dialogue datasets are Persona-Chat (Zhang et al., 2018), EmpatheticDialogues (Rashkin et al., 2018), Wizard-of-Wikipedia (Dinan et al., 2018), DailyDialog (Li et al., 2017), and BlendedSkillTalk (Smith et al., 2020). 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.

Source Image-Caption Pairs. We choose Conceptual Captions 3M (Sharma et al., 2018) (CC3M), which is widely used in multi-modal modeling (Lu et al., 2019; Su et al., 2019) and creating multi-modality dataset (Nagrani et al., 2022). We collect 2,796,458 image-caption pairs for the training and validation set. Then, we discard low-quality image-caption pairs based on our filtering criteria. First, we remove image-caption pairs with image-caption cosine similarity lower than the threshold of 0.2439 by leveraging CLIP ViT-L/14 model. Second, we remove watermark images using watermark detector [2]. Lastly, we remove image-caption pairs that contain copyright-related phrases (e.g., "royalty free") in captions. After the filtering, 692,292 image-caption pairs are obtained, which are divided into the training / validation / test set with a ratio of 5:1:1, resulting in 494K / 98K / 98K of unique images. Note that our pipeline can work with any image-caption datasets, such as Conceptual Captions 12M (Changpinyo et al., 2021) and RedCaps (Desai et al., 2021).

3.2 Aligning Images and Dialogues

After collecting a set of images and dialogues, we now describe how we create a high-quality and diverse multi-modal dialogue dataset starting from a seed text-only dialogue.

Inferring Image-Sharing Moments. While the image-sharing behavior naturally happens in existing human-authored datasets, we first should find potential image-sharing moments in the seed dialogue (Figure 2). However, it is challenging to determine the possible image-sharing moments in the given dialogue. Previously, the MMDD dataset is constructed by substituting utterances and images based on cosine similarities from an image-text matching model. This method can not guarantee the quality of the image-dialogue coherency (Figure 5), due to the nature of multi-turn conversation. Rather than directly measuring the similarity between utterances and images, we leverage GPT-4 (OpenAI, 2023) [3] in a zero-shot setting, inspired by its recent performance in the social dialogue domain (Kim et al., 2022; Lee et al., 2022a). Specifically, GPT-4 infers the appropriate turn and speaker to share the image. It also provides a contextual image description and explains why the image share is appropriate. The distribution of these rationales is in Table 15.

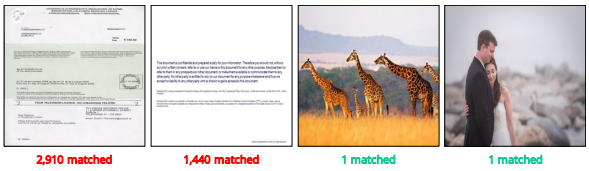


Figure 3: We show the examples of frequently matched images. The number under each image indicates the count of how many utterances are matched.

We use the carefully designed prompt template (detailed in the Appendix B.1). To make the dialogue inputs in the prompt more natural and soundness, we use Top-10K common names of US SSN applicants from 1990 to 2021 [4] for the speaker in a given dialogue, followed by a previous work (Kim et al., 2022). However, if the original dataset contains real speaker names, this could confuse the model. To avoid this, a named entity recognizer checks for person-related entities. If none is found, we select two names from the Top-10K list. If entities are detected, we ask GPT-4 to discern the actual speaker names. We then exclude non-human speakers, such as "hotel", "corporation". Finally, after constructing natural dialogue, we ask the model to infer potential image-sharing moments, specifying the image-sharing utterance, speaker, rationale, and image description, in a given dialogue, with a structured format of "`<utterance> | <speaker> | <rationale> | <image description>`". We parse each information in the structured format using the regex pattern (in Appendix B.1).

CLIP-based Similarity Calculation. In order to find images semantically relevant to a given dialogue context, we should get meaningful textual and visual features through a multi-modal feature extractor $f(\cdot)$. The previous work (Lee et al., 2021) used a pre-trained Visual Semantic Reasoning Network (Li et al., 2019) as $f(\cdot)$. In this work, we leverage CLIP (Radford et al., 2021) model as $f(\cdot)$, which is widely used in previous studies (Bose et al., 2022; Frans et al., 2021; Hessel et al., 2021; Cho et al., 2022; Zhu et al., 2023) because of a

[2] https://github.com/LAION-AI/LAION-5B-WatermarkDetection

[3] In this work, we use GPT-4 due to the high-quality, but our pipeline could work with any LLMs, such as LLaMa-2-Chat (Touvron et al., 2023) (See Appendix B.4). We use `gpt-4-0314` version, not using the recent version `gpt-4-0613` because of the lower performance reported in (Chen et al., 2023).

[4] https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-data

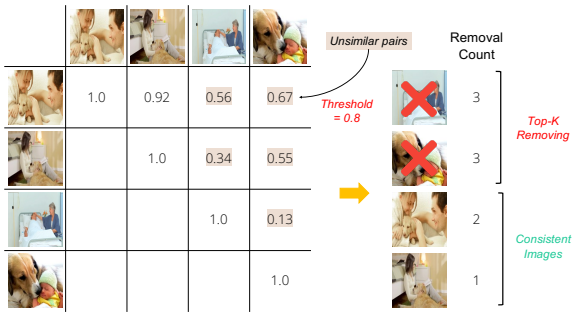


Figure 4: We illustrate the inconsistent images filtering process.

well-generalized open-domain model. We first extract LLM-generated description feature vector ($v_d = f(d)$), caption feature vector ($v_c = f(c)$), and image feature vector ($v_i = f(i)$). We then calculate the *description-image* similarity by computing the cosine similarity of v_d and v_i. Besides, to enhance the quality of utterance-image matching by additionally adopting the information provided by image captions, we also calculate the *description-caption* similarity.

However, there is one problem that we have to consider about how to combine these two similarity types. As reported in (Liang et al., 2022; So et al., 2022), there is a phenomenon called *modality gap* in multi-modal modeling, where two different modalities (i.e., image and text) are separately distributed in shared embedding space. Such phenomenon causes scale differences between description-image and description-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 a training set. The normalized similarities are linearly combined as follows:

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

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 the size of $N \times M$ is computed, where N and M are the number of utterances and images, respectively. We then select the top-100 samples based on the similarity scores.

3.3 Filtering Multi-Modal Dialogue

Thresholding-based 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 remove unsuitable images matched to utterances based on our criteria. Initially, we discarded images with cosine similarity scores below 2.702, retaining only 54.05% of the images. Moreover, we observe that certain images are frequently matched with many utterances. As shown in Figure 3, the frequently matched images mostly contain textual information (e.g., document) rather than object-centric or event-centric semantics (e.g., "giraffe" or "loving"). These frequent matches can lead to model overfitting, which is harmful to the generalization performance. To address this, we eliminate images that are matched more than 100 times.

Inconsistent Images Filtering. After we obtain multiple aligned images for each utterance, we should remove inconsistent images among multiple aligned images to ensure semantic similarity while maintaining diversity between multiple images. We illustrate the filtering process to help the understanding in Figure 4. First, we calculate a cosine similarity between multiple aligned images in a pairwise manner by leveraging the CLIP ViT-L/14 model. Next, we regard the image pair whose similarity score is lower than the threshold τ as unsimilar pair candidate. We set τ as 0.8. Then, we increase the removal candidate count of the image in this pair by 1. Finally, we sort by this count in descending order, and discard images in the Top-K% to have a high likelihood of being inconsistent with multiple images.

3.4 Analysis of DialogCC

High-Quality. To assess the quality of DialogCC, we conduct the human evaluations based on five criteria: (1) image-sharing turn relevance, (2) image-sharing speaker adequacy, (3) image-sharing rationale relevance, (4) aligned image relevance, and (5) image consistency. Each human rates 250 randomly chosen samples using a 4-point Likert scale for all criteria, except for (2) (i.e., "Yes" or "No"). Further details are in Appendix F.1. On average, we achieve higher scores across all evaluation criteria: 3.68 for (1), 95.1% ("Yes" ratio) for (2), 3.41 for (3), 3.30 for (4), and 3.57 for (5). In addition, we measure the inter-rater agreement using Krippendorff's α. On average, we get 0.39, which indicates

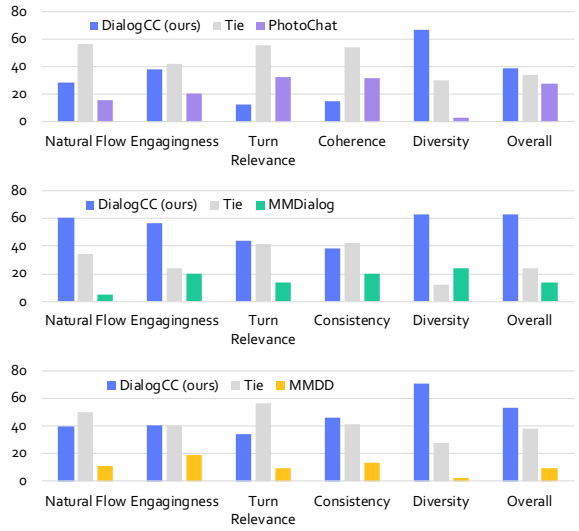


Figure 5: Results of head-to-head comparison between DialogCC (ours) and three existing datasets: PhotoChat, MMDialog, MMDD.

fair agreement. These results underscore the efficacy of our fully automatic pipeline, leveraging GPT-4 and CLIP. Breakdown analysis and details of human evaluation are shown in Appendix H.

To assess the quality gap between DialogCC and real-world scenarios, we conduct head-to-head human evaluations by comparing DialogCC with MMDD (Lee et al., 2021), PhotoChat (Zang et al., 2021), and MMDialog (Feng et al., 2022). We randomly sample 100 dialogues from each dataset and evaluate them based on six criteria: (1) natural flow, (2) engagingness, (3) turn relevance, (4) context consistency, (5) diversity, and (6) overall. Further details are in Appendix F.2. As shown in Figure 5, DialogCC achieves a higher score in overall quality, particularly surpassing MMDialog by a large margin. Furthermore, due to the nature of social media, MMDialog lacks natural conversational flow and engagingness compared to DialogCC by a large margin. This implies that while social media-sourced datasets may have significant advantages in terms of scale (in Table 1), their quality is not guaranteed for the social dialogue domain. Interestingly, compared to the PhotoChat, humans predominantly choose "Tie". This indicates that although DialogCC is built fully automatically, its quality closely matches human-authored datasets. Compared to the MMDD, DialogCC has more consistency between aligned images and dialogue context because we generate contextual image descriptions by prompting GPT-4.

Dataset	# Unique Dialog	# Unique Image	Avg. U./D.	Avg. I./D.	Avg. I./U.
PhotoChat	11,820	10,479	12.74	1.00	1.00
MMDD	17,679	13,288	11.56	1.76	1.29
MMDialog	1,079,117	1,556,868	4.56	2.82	1.36
DialogCC (ours)	**83,209**	**129,802**	**8.20**	**7.34**	**4.77**

Table 1: In total, DialogCC includes the largest number of Avg. I./D. and I./U. than others. I./D. and I./U. denote images by dialogue and images by an utterance, respectively. U./D. denotes utterances by a dialogue. More detailed statistics are in the Table 8.

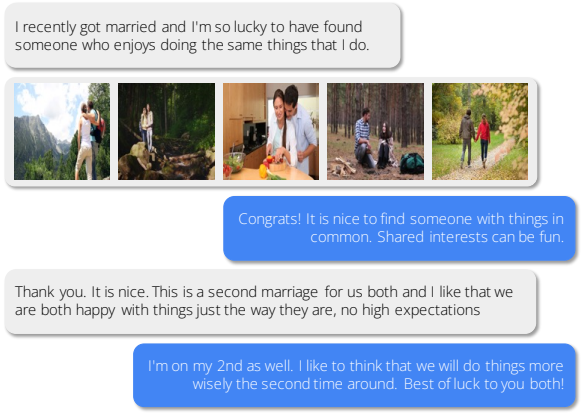


Figure 6: We present an example of DialogCC. More examples are in Appendix C.6. Note that during actual model training, one of these images is randomly sampled to enhance the model's generalization capability.

Image Diversity. 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 image-sharing turn. This does not reflect real-life scenarios and can 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 image-sharing turn, which is shown in Figure 6. In DialogCC, there are an average of 7.34 images per dialogue and 4.77 images per image-sharing turn, leading to enhanced generalization performance (in Section 4.4).

4 Experimentals

To explore how our dataset affects both text and image retrieval tasks, we implement two simple and standard baseline retrieval models for text-to-image and image-to-text settings.

4.1 Task Definition

Follwing (Lee et al., 2021; Zang et al., 2021), we explain the formulation of two main tasks - next response prediction and image retrieval. 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. **(1) Next response prediction** is to predict the next utterance at turn $t+1$ given the dialogue history ($\{u_j\}_1^t$) and image i_t. **(2) Image retrieval** is to retrieve relevant image at turn t given the dialogue history ($\{u_j\}_1^{t-1}$). Following (Shuster et al., 2018; Lee et al., 2021), we set the the number of retrieval candidates to 100 and use Recall@{1,5,10} and mean reciprocal rank (MRR) for the evaluation metrics.

4.2 Datasets

(1) DialogCC (ours) is a high-quality and diverse multi-modal dialogue dataset created by our proposed automatic pipeline powered by GPT-4 and CLIP models, which is described in Section 3. **(2) MMDD** (Lee et al., 2021) contains 45k multi-modal dialogues, where each utterance is replaced into a relevant image matched by their automatic pipeline. **(3) PhotoChat** (Zang et al., 2021) contains 10k multi-modal dialogues, where the dialogue is constructed via a crowd-sourcing platform. **(4) MMDialog** (Feng et al., 2022) contains 1M multi-modal dialogues, where the dialogue is obtained from the social media platform.

4.3 Baseline Models

The following are brief descriptions of two baseline retrieval models; more detailed information is provided in Appendix D.1. Two baseline models have a dual-encoder structure which consists of text encoder and image encoder. For the text encoder, we use the BERT-base (Devlin et al., 2018) architecture (12 layers, 12 attention heads, 768 dimensions, uncased version). For the image encoder, we use the CLIP-B/32 (Radford et al., 2021) model.

4.4 Main Results

DialogCC contributes to the model's robustness. To understand the contributed impact of DialogCC on other dialogue datasets, we evaluate the baseline models trained on DialogCC on unseen dialogue datasets, MMDD, PhotoChat, and MMDialog. In other words, we differentiate training and evaluation datasets to observe how much each dataset can

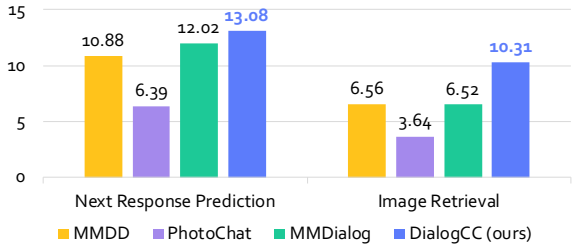


Figure 7: We report the average contributed performance on both tasks. Full results are in Appendix E.1.

	Image-Chat				MPChat			
Train Dataset	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR
MMDD	12.66	31.35	43.02	22.86	13.41	36.05	53.33	25.72
PhotoChat	6.89	21.41	32.54	15.76	7.52	29.23	43.57	19.15
MMDialog	14.37	32.45	43.39	24.29	21.94	51.01	66.90	36.08
DialogCC (ours)	**20.22**	**42.60**	**54.86**	**31.65**	**29.77**	**57.91**	**70.70**	**42.84**

Table 2: We report the next response prediction performance on Image-Chat (Shuster et al., 2018) and MPChat (Ahn et al., 2023) following the same evaluation setting.

boost the model's generalization performance on unseen dialogue scenarios in the next response prediction task and the image retrieval task. Figure 7 summarizes the average contributed performance of each datasets. Although the scale of DialogCC is significantly smaller than MMDialog (83K vs. 1M), DialogCC contributes to the model's understanding of the unseen dialogue dataset on both tasks. This suggests that increasing the quality of the dataset is more important than the scale, which is in line with the direction of recent studies (Zhou et al., 2023; Xu et al., 2023b) in data-centric AI.

In addition, PhotoChat, which is manually constructed, underperforms compared to other datasets as indicated by its limited diversity (see Table 1). This finding implies that, although our pipeline is automated compared to dialogue crowdsourcing, it not only ensures quality but is also more time and cost-efficient. This aligns with recent studies (Lee et al., 2022b; Kim et al., 2022) that generate dialogue datasets with the use of large language models, such as ChatGPT.

DialogCC improves the comprehension of the interaction between dialogue and images. We evaluate the baseline models on two unseen multi-modal dialogue datasets, Image-Chat (Shuster et al., 2018) and MPChat (Ahn et al., 2023), which belong to the image-grounded dialogue dataset. Table 2 summarizes the zero-shot results of next response prediction task of models trained on four

Model Inputs	R@1	R@5	R@10	MRR
Image Only	8.22	22.60	33.52	16.94
Dialogue Only	34.41	65.36	77.37	48.67
Dialogue + Image	**40.64**	**71.46**	**81.99**	**54.61**

Table 3: We show the effectiveness of image modality in DialogCC on the next response prediction task.

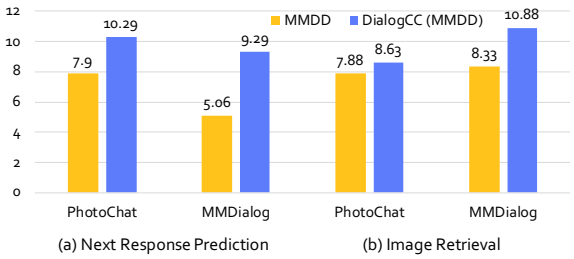


Figure 8: Results of the model trained on a sub-set of DialogCC using the same seed dialogue datasets as in MMDD.

different datasets: MMDD, PhotoChat, MMDialog, and DialogCC. The model trained on DialogCC outperforms those trained on other datasets. This indicates that DialogCC significantly improves the model's comprehension of the interaction between dialogue and images, even when the image-grounded dialogue datasets encompass various patterns in multi-modal dialogue scenarios. This improvement is attributed to DialogCC's high-quality and diverse images, as shown in Figure 5, underscoring the reliability of our pipeline.

Our pipeline effectively aligns dialogue with images. Since we align two distinct modalities – dialogue and images – using GPT-4 and CLIP automatically, we evaluate the model by varying the input modalities to investigate the correlation between dialogue and images. As shown in Table 3, providing only images to the model results in significantly lower performance, indicating the importance of dialogue context in multi-modal dialogue tasks. The model, when considering only dialogue, shows comparable performance, possibly because our dataset is based on original text-only dialogues. Notably, considering both dialogue and image leads to better performance. These results suggest that the image modality enhances the understanding of the dialogue without disrupting its flow, benefiting from our robust and reliable alignment process.

Our pipeline is better than the semi-automatic method. To validate our automatic pipeline, we evaluate a model trained on a subset of DialogCC,

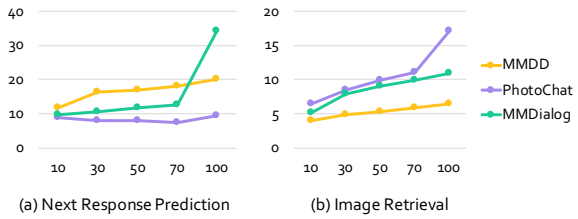


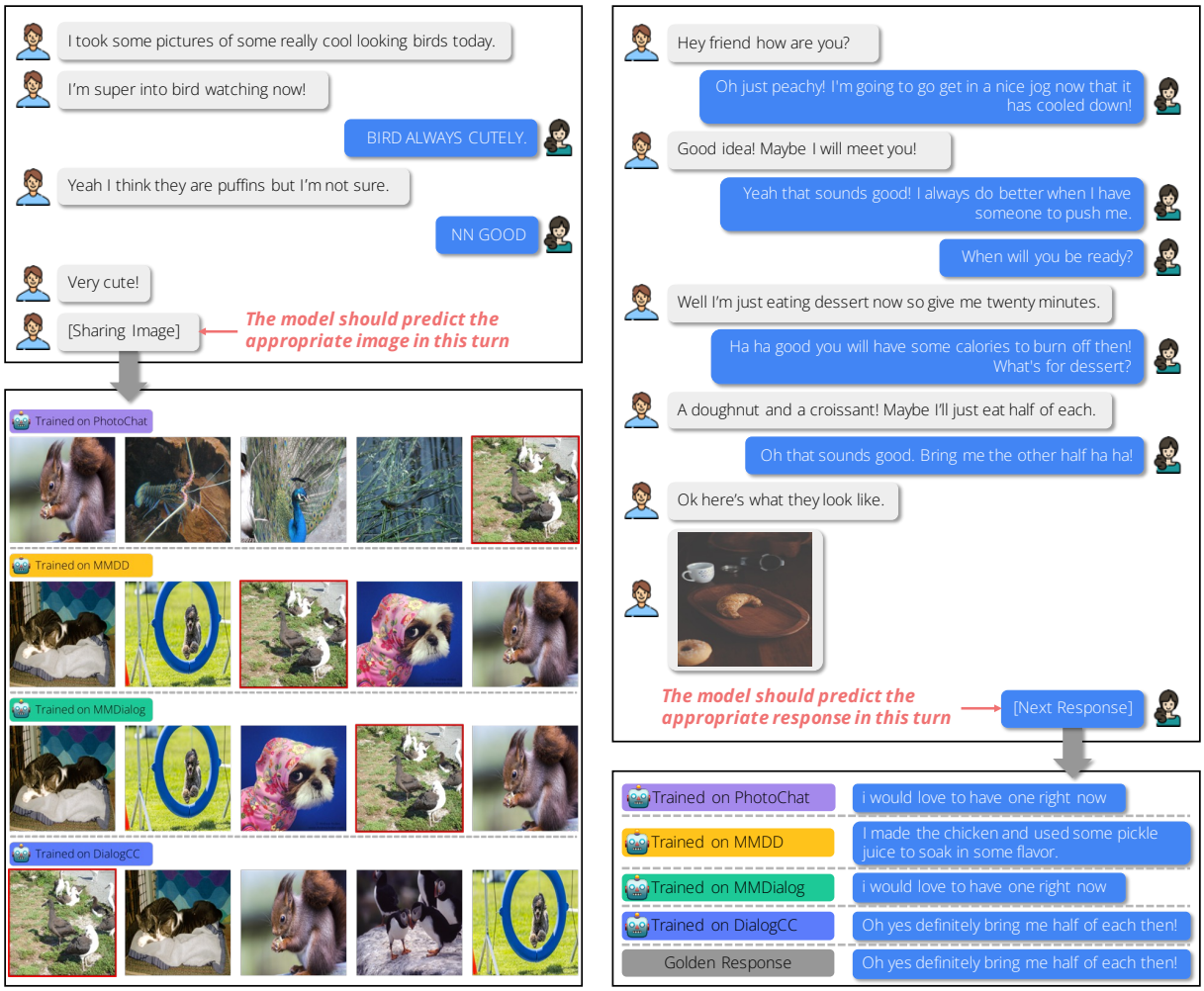
Figure 9: Scaling results of the model trained on DialogCC on both tasks.

using the same seed dialogue datasets as in MMDD (i.e., EmpatheticDialogues (Rashkin et al., 2018), Persona-Chat (Zhang et al., 2018), DailyDialog (Li et al., 2017)). Figure 8 demonstrates that, despite using identical dialogue datasets, our model significantly outperforms the one trained on MMDD. This result suggests that our pipeline more effectively discerns better image-sharing moments, taking advantage of GPT-4's capabilities. The performance gains are particularly notable in the MMDialog dataset, which includes a substantial number of images (see Table 1). This enhancement can be attributed to the use of a diverse image dataset (i.e., CC3M) as the seed dataset. Furthermore, our pipeline, requiring minimal human intervention, not only enhances the generalization performance of the trained model but also proves to be cost-effective, thereby ensuring both quality and performance.

Our pipeline benefits from scaling up the dataset size. To investigate whether our pipeline benefits from dataset scaling, we evaluate the model trained on DialogCC with varying dataset sizes. As shown in Figure 9, increasing the dataset size significantly enhances performance on three previously unseen dialogue datasets. These results indicate that our pipeline indeed benefits from scaling up the dataset size, thereby ensuring the creation of reliable and high-quality datasets. In future work, we plan to construct a million-scale, multi-modal dialogue dataset using the SODA (Kim et al., 2022) dataset in conjunction with our pipeline.

4.5 Case Study

As shown in Figure 10, we present two examples of results retrieved from models trained on four different datasets. In Figure 10-(a), three models trained on previous datasets (i.e., PhotoChat, MMDD, MMDialog) retrieve inappropriate images by focusing on the word "cute" in the last utterance. Conversely, the model trained with DialogCC accurately retrieves a suitable image of the cute ani-



(a) Image retrieval

(b) Next response prediction

Figure 10: Two examples of retrieved results (i.e., (a) image retrieval and (b) next response prediction) from models trained on four different datasets. Each provided dialogue is from the PhotoChat dataset. In (a), we display the top-5 ranked images from left to right, with the ground-truth image marked in red. In (b), only the top-1 ranked next response is shown. Note that neither the `[Sharing Image]` turn nor the `[Next Response]` turn is provided to the model's input during the inference stage. More examples are presented in Figure 22.

mal "puffins." This indicates that the model trained on DialogCC not only recognizes what a "puffin" looks like but also understands the contextual relevance of the word "cute" within the entire dialogue. This capability is attributed to the high-quality and diverse imagery of our dataset. In Figure 10-(b), DialogCC significantly enhances the model's ability to understand multi-modal dialogues, resulting in the accurate retrieval of the correct subsequent response. These results highlight the importance of both high-quality and image diversity in developing a more generalized and robust model.

5 Conclusion

In this paper, we propose the automatic pipeline for creating a multi-modal social dialogue dataset that involves aligning and filtering with GPT-4 and CLIP, respectively. We also propose a large-scale and high-quality 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, PhotoChat, MMDialog, using DialogCC helps achieve better quality in terms of various metrics. 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 increase model's robustness.

Limitations

Societal Impact. As reported in (Wang et al., 2021a), even if we give the gender-neutral query to CLIP (Radford et al., 2021) 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 the generated image description by GPT-4. A notable example of this bias is the association of women's images with the profession of "hair designer." Such biases are concerning as they could propagate stereotypes. Therefore, the image retrieval model trained on our dataset may sometimes retrieve biased images. We should consider this problem important when building a multi-modal search model. In the future work, we will mitigate this issue to be fairer and more generalized model.

Addressing Cross-Turn Image Inconsistency. In our effort to construct a natural and coherent multi-modal dialogue dataset, we utilize GPT-4 to identify appropriate moments for image sharing within text-only dialogues, ensuring conversational flow. We then generate image descriptions for these moments and align the corresponding images using CLIP. To maintain single-turn image consistency, we introduce a straightforward algorithm based on pairwise similarity comparisons through CLIP. Nevertheless, our approach currently overlooks cross-turn image inconsistencies within the same dialogue, and addressing this challenge is part of our future objectives.

Considering Personalization. Our dataset aims to enhance generalization performance by mapping multiple images to a single utterance. This dataset, while beneficial for model generalization, may occasionally result in the sharing of images unrelated to the speaker's specific subject, diminishing user interactability. For example, if a speaker refers to their Chihuahua, the model might incorrectly present an image of a Golden Retriever due to the broad mapping in our dataset. Recognizing these limitations, we emphasize the importance of not only improving generalization but also incorporating user preferences to bolster engagement. Our future work is thus dedicated to developing a personalized multi-modal dialogue dataset and system.

Improving Factuality in Alignment. Despite our meticulous efforts in developing DialogCC with carefully designed pipelines, the dataset may still include samples that are factually inaccurate. For example, an image meant to illustrate an utterance related to "race walking" might instead show a "marathon scene," or an utterance describing a "three-story hotel building" could be incorrectly matched with a photo of a "four-story hotel." In the future, we will consider real scene understanding (Lee et al., 2024a,b) to enhance the factual accuracy of the alignment process.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [No.2022-0-00641, XVoice: Multi-Modal Voice Meta Learning]

References

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*.

Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. Mpchat: Towards multimodal persona-grounded conversation. *arXiv preprint arXiv:2305.17388*.

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

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

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. 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.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is chatgpt's behavior changing over time? *arXiv preprint arXiv:2307.09009*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

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

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

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

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

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

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2022. Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. *arXiv preprint arXiv:2211.05719*.

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

Seungju Han, Jack Hessel, Nouha Dziri, Yejin Choi, and Youngjae Yu. 2023. Champagne: Learning real-world conversation from large-scale web videos. *arXiv preprint arXiv:2303.09713*.

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

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. 2020. spacy: Industrial-strength natural language processing in python.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, et al. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*.

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

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

Byung-Kwan Lee, Beomchan Park, Chae Won Kim, and Yong Man Ro. 2024a. Collavo: Crayon large language and vision model. *arXiv preprint arXiv:2402.11248*.

Byung-Kwan Lee, Beomchan Park, Chae Won Kim, and Yong Man Ro. 2024b. Moai: Mixture of all intelligence for large language and vision models. *arXiv preprint arXiv:2403.07508*.

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

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022a. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683.

Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022b. Personachatgen: Generating personalized dialogues using gpt-3. In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, pages 29–48.

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

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

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

Katharina Lobinger. 2016. Photographs as things–photographs of things. a texto-material perspective on photo-sharing practices. *Information, Communication & Society*, 19(4):475–488.

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

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

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

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

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

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

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

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Icml*.

OpenAI. 2023. Gpt-4 technical report. *arXiv*.

Ramakanth Pasunuru and Mohit Bansal. 2018. Game-based video-context dialogue. *arXiv preprint arXiv:1809.04560*.

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

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

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

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Image chat: Engaging grounded conversations. *arXiv preprint arXiv:1811.00945*.

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

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

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

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

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

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

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

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

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023a. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2023b. Demystifying clip data. *arXiv preprint arXiv:2309.16671*.

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

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? *arXiv preprint arXiv:1801.07243*.

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

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*.

A Details of Source Datasets

A.1 Source Dialogue Datasets

We collect the give text-only social dialogue datasets (i.e., Wizard-of-Wikipedia (Dinan et al., 2018), Persona-Chat (Zhang et al., 2018), EmpatheticDialogues (Rashkin et al., 2018), DailyDialog (Li et al., 2017), and BlendedSkillTalk (Smith et al., 2020)) through the ParlAI (Miller et al., 2017) framework, which provides many dialogue datasets online. The statistics of source dialogue datasets are shown in Table 4. The details of each dataset are described as follows:

Wizard-of-Wikipedia. This dataset aims to enable the dialogue agent to generate knowledgeable responses grounded in information retrieved from Wikipedia to enhance the engagement of the conversation. The dataset was constructed via a crowdsourcing platform, where two participants converse with each other on one of a total of 1,365 topics. One participant selects a conversational topic and assumes the role of a knowledgeable expert (referred to as the *wizard*), while the other acts as a curious learner (the *apprentice*). The dataset can be downloaded from the ParlAI framework by setting the task name as `wizard_of_wikipedia:basic_apprentice_dialog`.

Persona-Chat. This dataset is designed to enable the dialogue agent to generate responses based on personal information, whether their own or others. It was constructed using a crowdsourcing platform, where two participants engage in a conversation based on provided persona information. The persona is represented by a set of sentences that depict demographic and psychographic characteristics (Lee et al., 2022b). Examples of such sentences include "I am getting old." and "I love the color blue." Given that the original persona sentences exhibit simple linguistic structures, a revised version of these sentences is also provided to make the model training more challenging and thereby enhance performance. To download this dataset from ParlAI, set the task name to `personachat:both_original`.

EmpatheticDialogues. This dataset is designed to enable dialogue agents to generate empathetic responses by understanding and interpreting the interlocutor's emotional situation. It was constructed using a crowdsourcing platform where two turkers are assigned specific roles: *speaker* and *listener*. The speaker is provided with an emotional situation and one emotion label from a set of 32 labels, while the listener responds with empathy to the speaker's situation. The dataset can be downloaded from the ParlAI framework using the task name `empathetic_dialogues`.

Dataset	Type	# Dialog	# Utter	Avg. Utter. Len	Avg. Utter/Dialog
Blended Skill Talk	train	4,819	54,036	13.09	11.21
	valid	1,009	11,302	13.17	11.20
	test	980	10,964	13.60	11.19
	total	6,808	76,302	13.29	11.20
DailyDialog	train	21,753	152,104	11.44	6.99
	valid	1,960	14,138	11.36	7.21
	test	1,958	13,480	11.56	6.88
	total	25,671	179,722	11.45	7.03
EmpatheticDialogues	train	19,531	80,508	13.45	4.12
	valid	2,769	11,476	14.50	4.14
	test	2,547	10,518	15.33	4.13
	total	24,847	102,502	14.43	4.13
Persona-Chat	train	8,939	131,438	10.09	14.70
	valid	1,000	15,602	10.30	15.60
	test	968	15,024	10.19	15.52
	total	10,907	162,064	10.19	15.28
Wizard of Wikipedia	train	18,430	166,787	16.37	9.05
	valid	1,948	17,715	16.40	9.09
	test	1,933	17,497	16.26	9.05
	total	22,311	201,999	16.34	9.07

Table 4: We show the statistics of source dialogue datasets.

DailyDialog. This dataset was constructed by crawling daily-life conversations from various websites. It includes additional information crucial for understanding and proceeding with daily-life conversations between partners, such as emotion, topic, and dialog act. Specifically, there are seven emotion categories: anger, disgust, fear, happiness, sadness, surprise, and others. The dataset contains 10 daily topics: ordinary life, school life, culture & education, attitude & emotion, relationship, tourism, health, work, politics, and finance. Additionally, there are four dialog acts: inform, question, directive, and commission. The dataset can be downloaded from the ParlAI framework using the task name `dailydialog:no_start`.

Blended Skill Talk. This dataset is designed to help dialogue agents learn how to use multiple conversational skills interactively and naturally rather than relying on a single isolated skill. The dataset was constructed by integrating several skills

Prompt Template for Inferring Image-Sharing Moments:

The following is a dialogue between `[speaker1]` and `[speaker2]`. The dialogue is provided line-by-line. In the given dialogue, select all utterances that are appropriate for sharing the image in the next turn, and write the speaker who will share the image after the selected utterance. You should also provide a rationale for your decision and describe the relevant image concisely.

Dialogue:
`[dialogue]`

Restrictions:
(1) your answer should be in the format of "<UTTERANCE> | <SPEAKER> | <RATIONALE> | <IMAGE DESCRIPTION>".
(2) you MUST select the utterance in the given dialogue, NOT generate a new utterance.
(3) the rationale should be written starting with "To".

Answer:
1.

Prompt Template for Identifying Speaker Names in `[dialogue]`:

`[dialogue]`

Q: What are the names of Speaker A and Speaker B in the given dialogue? Your answer should be in the format of "<Speaker A> | <Speaker B>".
A:

Figure 11: A prompt template for inferring image-sharing moments (**top**). A prompt template for identifying speaker names in `[dialogue]` (**bottom**).

(i.e., empathetic, knowledgeable, and personalizing) into a single conversation via a crowdsourcing platform. Within this dataset, there are four skill annotations: (1) Knowledge, (2) Empathy, (3) Personal situations, and (4) Personal background. Each utterance in a conversation is annotated with a corresponding skill. The dataset can be downloaded from the ParlAI framework using the task name `blended_skill_talk`.

A.2 Source Image-Caption Pair Dataset

We download the Conceptual Captions 3M (Sharma et al., 2018) (CC3M) dataset in here [5]. Since the CC3M dataset provides image URLs, we download images using img2dataset [6] library, which is a helpful library for quick downloading large-scale images based on URLs. We downloaded images in March 2023 and we store downloaded images as a `jpg` format. We obtain 2,783,547 images from the train set and 12,911 from the valid set. Note that because each image URL has the copyright, we only use opened URLs as source image-caption data when we create DialogCC.

A.3 Licenses

We list the licenses of each source dataset that we utilized in the creation of DialogCC.

- Wizard-of-Wikipedia: CC-BY-4.0
- Persona-Chat: CC-BY-4.0
- EmpatheticDialogues: CC-BY-4.0
- DailyDialog: CC BY-NC-SA 4.0
- Blended Skill Talk: CC-BY-4.0

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

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

- Conceptual Caption 3M: Open License by Google

CC3M is under the Google open license, which allows for the free use of the dataset for any purpose. Since all the datasets except DailyDialog are permissible for commercial use, we will release our dataset DialogCC by following the "C BY-NC-SA 4.0" license. This means the dataset can only be used for academic or research purposes and is not permitted for commercial use.

B Details of Automated Pipeline

B.1 Prompt Templates

In order to infer image-sharing moments using GPT-4, we thoughtfully create the prompt template, as depicted in Figure 11. We provide GPT-4 with specific guidelines (i.e., *restrictions*) derived from insights gained in a preliminary study to ensure the generation of higher-quality results. Specifically, the model produces potential image-sharing utterances with speaker (*who*), rationale (*why*), and image description (*what*). Moreover, regarding the second sentence in the restrictions, if we omit this from the prompt, the model occasionally fails to infer the image-sharing utterance within the given dialogue. Instead, it creates a new utterance suggesting an event that might occur following the current dialogue context. For the `[dialogue]`, we provide the entire dialogue history into the model. The motivation behind this design decision is explained in Section B.2. Furthermore, as we mentioned in Section 3.2, to make the `[dialogue]` natural, we identify the actual speaker names within the given `[dialogue]` based on the designed prompt template as shown in Figure 11. To parse the *utterance*, *speaker*, *rationale*, and *image description* from the GPT-4 generation results, we implemented a simple parser using regex patterns, as depicted in Figure 12.

B.2 Motivation behind Providing Full Dialogue

The objective of this paper is to create a high-quality multi-modal dialogue dataset, building upon an existing text-only social dialogue dataset, as described in Section A.1. This implies that the source dialogue datasets already possess an inherent dialogue context, such as conversational flow, holistic meaning, and topic. Therefore, it's imperative to identify potential image-sharing moments without disturbing the established conversational flow, even after integrating relevant images into the inferred image-sharing utterances. As a result, we feed the complete dialogue history to the model.

	Similarity	Q1	Q2	Q3	Q4
Similarity		0.3066	0.3153	0.3573	0.2478
Q1			0.5566	0.4496	0.3313
Q2				0.7999	0.4461
Q3					0.5826
Q4					

Table 5: We show Spearman's correlation between four human evaluation items and utterance-image cosine similarity using CLIP ViT-L/14 model. Q1, Q2, Q3, and Q4 denote the turn relevance, rationale relevance, aligned image relevance, and image consistency, respectively.

Model	Recall (↑)
Tulu-13B (Wang et al., 2023)	2.27
WizardLM-13B (Xu et al., 2023a)	18.80
Vicuna-13B (Chiang et al., 2023)	29.96
LLaMA-2-Chat-13B (Touvron et al., 2023)	29.44
GPT-4 (OpenAI, 2023)	**35.23**

Table 6: We present a comparison of open-source LLMs, including GPT-4, based on the recall metric using the PhotoChat dataset.

B.3 Motivation behind Using GPT-4

The motivation behind using GPT-4 is to generate contextualized image descriptions rather than relying on the calculation of cosine similarity between a single utterance and an image, as done by image-text matching models (e.g., VSRN) in MMDD. Given that the image-text matching model is trained on image-caption pair datasets, it struggles to capture the holistic meaning from the dialogue context. For instance, it becomes challenging to identify relevant images for the sentence "I ate it yesterday. See this photo!" without access to the preceding dialogue context. Furthermore, as depicted in Table 5, the correlation between the CLIP similarity and the relevance of turns as rated by humans is low. This finding suggests that determining relevant images using only utterances is not effective. Therefore, inspired by the recent advancements of large language models in the social dialogue domain (Lee et al., 2022a,b; Kim et al., 2022), we choose to employ GPT-4 to generate contextualized image descriptions.

```python
import re
from typing import Dict

class Parser:

    PATTERN = r'^(?:\d+\.\s+)?\"?(?P<utterance>.*?)\"?\s+\|\s+(?P<speaker>.*?)(?:\s+\|\s+(?P<rationale>.*?))?(?:\s+\|\s+(?P<description>.*?))?$'

    def parse(pred: str) -> Dict:
        pred = pred.strip()

        matches = re.finditer(Parser.PATTERN, pred, re.MULTILINE)
        results = []
        for match in matches:
            utter = match.group('utterance')
            speaker = match.group('speaker')
            rationale = match.group('rationale')
            description = match.group('description')

            results.append({
                'utterance': utter,
                'speaker': speaker,
                'rationale': rationale,
                'description': description
            })

        return results
```

Figure 12: A Python code for parsing generated responses from GPT-4.

B.4 GPT-4 versus Open-Sourced LLM

GPT-4 can be replaced by open-sourced LLMs in our pipeline for cost reduction, leading to enhanced scalability of the dataset. To see their feasibility, we evaluate the LLM's ability to infer image-sharing moments in PhotoChat, using recall as the metric, measuring whether one of the generated imagesharing turns matches the ground-truth turn in PhotoChat. Table 6 shows that GPT-4 outperforms recent open-sourced LLMs. Thus, we used GPT-4 since our work focuses on the quality and diversity of the multi-modal dialogue dataset. However, it is possible to create a large-scale dataset using our automatic pipeline with GPT-4. With adequate budgets, we can increase the dataset size considerably, ensuring quality and diversity.

B.5 Details of Filtering Step

We determine the threshold scores used in the *filtering* step by manually evaluating randomly chosen 10,000 samples.

C Further Analyses on DialogCC

C.1 Comparing to Existing Multi-Modal Dialogue Datasets

Table 7 compares DialogCC with other multi-modal dialogue datasets. Unlike other image-grounded datasets, DialogCC falls under the category of image-sharing datasets in terms of multi-modal interaction type. Specifically, image-grounded datasets always begin with a given image. Both conversational partners perceive this image and discuss it, such as questioning. In other words, image-grounded datasets always start from the given image, then two conversational partners perceive the given image and then talk about the image. However, with image-sharing datasets, the two participants converse with each other before sharing an image. At some point, one of them shares a relevant image based on the preceding dialogue context. After this, the conversation continues, with both partners discussing the shared image. Thus, the image-sharing dialogue dataset is more challenging than image-grounded datasets, since the former encompasses the scope of the latter as well.

Among the existing image-sharing datasets, the alignment of two different modalities (i.e., image and dialogue) is typically performed by humans. However, we leverage the GPT-4 and CLIP models to align these modalities without human intervention. Although DialogCC is fully constructed by automatic pipeline, it achieves high-quality and diverse alignments compared to other image-sharing datasets, as depicted in Figure 5.

Dataset	Source Dialog	Source Image	Interaction Type	Aligning Two Modalities
VisualDialog	CS	COCO	grounding	Human
IGC	CS	VQG	grounding	Human
ImageChat	CS	YFCC100M	grounding	Human
OpenViDial	Movie & TV	Movie & TV	grounding	Human
MMChat	Social Media	Social Media	grounding	Human
MPChat	Reddit	Reddit	grounding	Human
PhotoChat	CS	Open Image Dataset V4	sharing	Human
MMDD	ED, PC, Daily	MS-COCO, Flicker 30k	sharing	VSRN + Human
MMDialog	Social Media		sharing	Human
DialogCC (ours)	ED, PC, Daily, BST, WoW	CC3M	sharing	GPT-4, CLIP

Table 7: Comparison of DialogCC with other multi-modal dialogue datasets: VisualDialog (Das et al., 2017), IGC (Mostafazadeh et al., 2017), ImageChat (Shuster et al., 2020), OpenViDial (Meng et al., 2020), MMChat (Zheng et al., 2021), MPChat (Ahn et al., 2023), PhotoChat (Zang et al., 2021), MMDD (Lee et al., 2021), and MMDialog (Feng et al., 2022). CS denotes crowdsourcing. ED, PC, Daily, BST, WoW denote EmpatheticDialogues, Persona-Chat, DailyDialog, BlendedSkillTalk, Wizard-of-Wikipedia. VSRN denotes the Visual Semantic Reasoning Network (Li et al., 2019).

C.2 Full Statistics of DialogCC

Table 8 presents a comprehensive comparison of the statistics for DialogCC against existing datasets, namely PhotoChat, MMDD, and MMDialog. DialogCC is constructed from five source dialogue datasets: Persona-Chat, EmpatheticDialogues, Blended Skill Talk, DailyDialog, and Wizard-of-Wikipedia. As a result, DialogCC consists of five sub-datasets: BlendedCC, DailyCC, EmpathyCC, PersonaCC, and KnowledgeCC. Taking PersonaCC as an example, this dataset is formulated by aligning images from the CC3M collection with the Persona-Chat dataset, achieved using our proposed automatic pipeline. The statistics of five sub-datasets are presented in Table 9.

C.3 Image-Sharing Moment Distribution

In Figure 13, we analyze the distribution of turns at which image-sharing occurs across various datasets. Unlike PhotoChat and MMDialog, DialoCC demonstrates that the moments that images are shared are evenly distributed throughout the conversation turns. This suggests that models trained on our dataset may better understand the optimal moments for image-sharing across diverse dialogue turns. Compared with MMDD, the turn distribution for image sharing in MMDD is also even. However, it's notable that in MMDD, images can be seen even in the initial dialogue turn. As highlighted in Section C.1, MMDD might not fully represent an image-sharing dataset, given it also encompasses image-grounded dialogues. This observation suggests that during the creation of the MMDD dataset, images were potentially matched with single utterances based on image-text similarity via the VSRN model. Such an approach might not truly reflect humans' cognitive processes when sharing images in real-life conversations. In contrast, DialoCC leveraged GPT-4 to determine appropriate moments to share an image in specific dialogues. This method results in a more naturally flowing dialogue with greater turn relevance, as shown in Figure 5. Consequently, in DialoCC, we can affirmatively state that no images are shared during the initial turn of our dialogues, unlike MMDD.

C.4 Diversity

In Table 10, we compare the diversity of datasets with the number of unique hypernyms from WordNet (Miller, 1995) and words in dialogues and image captions. As WordNet covers nouns, verbs, adjectives, and adverbs, we only count nouns by filtering out the hypernyms appearing less than ten times. Compared to PhotoChat and MMDD, DialogCC contains the largest number of unique hypernyms and unique words in both image captions and dialogues. Unfortunately, MMDialog does not include captions, so we cannot determine the number of unique hypernyms and unique words from that dataset. However, MMDialog has more hypernyms and unique words, likely attributed to its larger volume of dialogues. It's worth noting that despite MMDialog having the most extensive scale, its quality is subpar, as depicted in Figure 5.

C.5 Rationale Distribution

To gain a better understanding of the generated rationales, we conduct an analysis of their verb-noun patterns. Table 15 shows the rationale distribution obtained from GPT-4. We parse the rationales using spaCy (Honnibal et al., 2020) and extract the root verb along with its first direct noun object. Since we constrain a rationale to start with "To" in the prompt, we only consider rationales with a "To verb noun" structure during this analysis. Out of a total of 106,063 generated rationales, 102,554 rationales follow this structure, whereas 3,509 rationales contain more complex clauses (e.g., *To show how he spent his relaxing weekend.*).

Dataset	Type	# Unique Dialog	# Image	# Unique Image	# Utter	Avg. U/D	Avg. I/D	# Sharing Utter	Avg. S/D	Avg. I/S
PhotoChat	train	9,890	9,890	8,549	125,512	12.69	1.00	9,890	1.00	1.00
	valid	962	962	962	12,205	12.69	1.00	962	1.00	1.00
	test	968	968	968	12,421	12.83	1.00	968	1.00	1.00
	total	11,820	11,820	10,479	150,138	12.74	1.00	11,820	1.00	1.00
MMDD	train	13,141	39,956	12,272	131,392	10.00	3.04	21,525	1.64	1.86
	valid	2,148	2,401	334	26,576	12.37	1.12	2,401	1.12	1.00
	test	2,390	2,673	682	29,453	12.32	1.12	2,673	1.12	1.00
	total	17,679	45,030	13,288	187,421	11.56	1.76	26,599	1.29	1.29
MMDialog	train	1,059,117	2,981,568	1,509,284	4,825,053	4.56	2.82	2,193,816	2.07	1.36
	valid	10,000	27,944	23,812	45,382	4.54	2.79	20,546	2.05	1.36
	test	10,000	28,419	23,772	45,801	4.58	2.84	20,871	2.09	1.36
	total	1,079,117	3,037,931	1,556,868	4,916,236	4.56	2.82	2,235,233	2.07	1.36
DialogCC (ours)	train	68,269	699,505	101,877	552,991	8.10	10.25	106,063	1.55	6.60
	valid	7,635	44,093	13,842	63,074	8.26	5.78	11,662	1.53	3.78
	test	7,305	43,872	14,083	60,116	8.23	6.01	11,139	1.52	3.94
	total	**83,209**	**787,470**	**129,802**	**676,181**	**8.20**	**7.34**	**128,864**	**1.54**	**4.77**

Table 8: In total, DialogCC includes the largest number of Avg. I./D. and I./S. than others. I./D. and I./S. denote images by dialogue and images by an image-sharing utterance, respectively. U./D. denotes utterances by a dialogue.

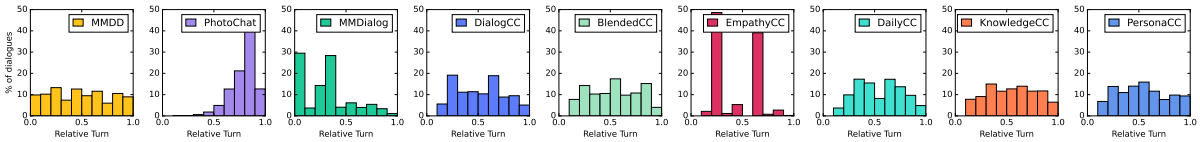


Figure 13: Comparison of DialogCC with other multi-modal dialogue datasets: PhotoChat (Zang et al., 2021), MMDD (Lee et al., 2021), and MMDialog (Feng et al., 2022), in terms of the distribution of image-sharing moments. The x-axis and y-axis represent the relative turn ratio and % of dialogues, respectively. We also show the distribution of a subset of DialogCC: BlendedCC, EmpathyCC, DailyCC, KnowledgeCC, and PersonaCC.

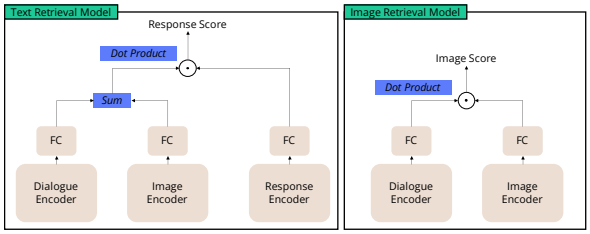


Figure 14: Architectures of two baseline models: Text retrieval and Image retrieval.

In this analysis, we observe that the verb "provide" is used most frequently. This indicates that image sharing is intended to provide additional information related to the context of the dialogue. The tendency to provide additional information through image sharing is also evident in the verbs "show" and "share".

C.6 More Examples of DialogCC

We present more examples of DialogCC in Figure 17, Figure 18, Figure 19, Figure 20, and Figure 21.

D Details of Experimental Settings

D.1 Baseline Models

As illustrated in Figure 14, we present the architecture of baseline models, which is the text retrieval model and image retrieval model. We provide a detailed description of baseline models below.

Text Retrieval Model. The text retrieval model comprises three main components: the dialogue encoder, the response encoder, and the image encoder. The dialogue encoder processes the entire dialogue history and transforms it into a fixed-size representation. To achieve this, we use the BERT model (Devlin et al., 2018). The dialogue history consists of up to three turns preceding the current turn. Each turn is concatenated using the `[SEP]` special token. The response encoder is responsible for converting the response into a fixed-size representation. While it also utilizes the BERT model, the specific BERT version used here is different from that employed in the dialogue encoder. For both the dialogue and response encoders, after pro-

Dataset	Type	# Unique Dialog	# Image	# Unique Image	# Utter	Avg. U/D	Avg. I/D	# Sharing Utter	Avg. S/D	Avg. I/S
BlendedCC	train	4,595	52,890	25,916	51,650	11.24	11.51	7,671	1.67	6.89
	valid	927	6,185	4,047	10,376	11.19	6.67	1,458	1.57	4.24
	test	872	5,962	3,856	9,790	11.23	6.84	1,394	1.60	4.28
	total	6,394	65,037	33,819	71,816	11.22	8.34	10,523	1.61	5.14
DailyCC	train	19,459	162,260	42,088	139,416	7.16	8.34	26,495	1.36	6.12
	valid	1,665	7,322	3,644	12,228	7.34	4.40	2,248	1.35	3.26
	test	1,641	7,610	4,138	11,562	7.05	4.64	2,183	1.33	3.49
	total	22,765	177,192	49,870	163,206	7.18	5.79	30,926	1.35	4.29
EmpathyCC	train	17,879	122,597	35,294	73,748	4.12	6.86	19,234	1.08	6.37
	valid	2,347	7,631	4,125	9,720	4.14	3.25	2,540	1.08	3.00
	test	2,165	7,924	4,402	8,932	4.13	3.66	2,344	1.08	3.38
	total	22,391	138,152	43,821	92,400	4.13	4.59	24,118	1.08	4.25
PersonaCC	train	8,798	150,818	41,579	129,404	14.71	17.14	20,648	2.35	7.30
	valid	956	10,289	4,406	14,916	15.60	10.76	2,278	2.38	4.52
	test	933	10,163	4,407	14,474	15.51	10.89	2,195	2.35	4.63
	total	10,687	171,270	50,392	158,794	15.27	12.93	25,121	2.36	5.48
KnowledgeCC	train	17,538	210,940	54,210	158,773	9.05	12.03	32,015	1.83	6.59
	valid	1,740	12,666	5,749	15,834	9.10	7.28	3,138	1.80	4.04
	test	1,694	12,213	5,915	15,358	9.07	7.21	3,023	1.78	4.04
	total	20,972	235,819	65,874	189,965	9.07	8.84	38,176	1.80	4.89

Table 9: Statistics of sub-dataset of DialogCC. I./D. and I./S. denote images by dialogue and images by an image-sharing utterance, respectively. U./D. denotes utterances by a dialogue.

cessing the text with BERT, we apply mean pooling to the text representations. The pooled representations are subsequently passed through a linear projection layer, which is then followed by the ReLU activation function (Nair and Hinton, 2010). The image encoder is to extract feature vectors from images, and for this purpose, we utilize the CLIP-base model (Radford et al., 2021). Once the feature vectors are extracted from the dialogue and images, we perform an element-wise addition of the image vectors and dialogue vectors. To compute the loss, we calculate the dot product between the response feature vector and the resulting summed vector.

Image Retrieval Model. The image retrieval model is composed of two main components: the dialogue encoder and the image encoder. The dialogue encoder utilizes the BERT-base model to transform the dialogue into a representation. After encoding, we apply mean pooling to the text representations derived from this dialogue encoder. For image representation, we employ the CLIP-base model. Following the encoding processes, both the image and dialogue vectors are passed through separate linear projection layers, each followed by a ReLU activation function. To determine the loss, we calculate the dot product between the image feature vector and the dialogue vector.

D.2 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 (Lee et al., 2021; Zang et al., 2021), which are described as follows:

Text retrieval. In our experiment, we set the batch size to 256, the learning rate to 5e-5, and the gradient clipping value to 2.0. We use the AdamW optimizer with a cosine learning rate scheduler. We set the warm-up ratio as 0.1% and weight decay as 0.2.

Image retrieval. We set the batch size to 256. We also use the AdamW optimizer with an initial learning rate of 2e-5 and decaying 0.1% at every 1,000 steps. We set the warm-up ratio as 0.1%.

Training. Since our dataset contains several images per utterance, we randomly choose one image in each batch. We do not update the parameter of the image encoder.

Dataset	Type	Image Caption			Dialogue		
		# hyp	# unigram	# bigram	# hyp	# unigram	# bigram
PhotoChat	train	293	4,203	10,772	1,203	18,252	179,904
	valid	72	1,001	2,059	348	4,994	32,883
	test	74	1,000	2,034	351	5,066	33,456
	total	439	6,204	14,865	1,902	28,312	246,243
MMDD	train	1,832	11,571	95,918	2,168	23,264	298,517
	valid	462	2,080	7,539	968	10,207	88,762
	test	463	2,337	8,867	1,033	11,055	96,891
	total	2,757	15,988	112,324	4,169	44,526	484,170
MMDialog	train	-	-	-	9,271	772,044	8,582,862
	valid	-	-	-	2,239	49,443	340,221
	test	-	-	-	2,247	49,310	339,883
	total	-	-	-	13,757	870,797	9,262,966
DialogCC	train	3,020	18,623	241,047	4,061	62,961	953,730
	valid	1,469	9,485	58,320	1,802	22,096	219,545
	test	1,493	9,725	59,529	1,819	21,873	216,436
	total	**5,982**	**37,833**	**358,896**	**7,682**	**106,930**	**1,389,711**

Table 10: We count the number of unique hypernyms from WordNet (Miller, 1995) 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, # unigram, and # bigram denote the number of hypernyms, the number of unique unigrams, and the number of unique bigrams. respectively

E Further Experiments

E.1 Full Results

Table 12 shows the full results of model trained on PhotoChat (Zang et al., 2021), MMDD (Lee et al., 2021), MMDialog (Feng et al., 2022), and DialogCC (ours). Then, we evaluate each trained model on four different datasets. We measure the average contributed performance only considering the out-of-domain datasets. For example, if the model is trained on the MMDD dataset, then we calculate the averaged contributed performance by evaluating this model on PhotoChat, MMDialog, and DialogCC.

E.2 Breakdown Results in DialogCC

We show the additional experiments in Table 13 and Table 14. We evaluate the trained retrieval model on the sub-dataset of DialogCC to other sub-dataset of DialogCC.

F Human Evaluation Questionnaire

This section presents the list of questions and multiple-choice options used for two human evaluations reported in Section 3.4: human ratings and head-to-head comparison.

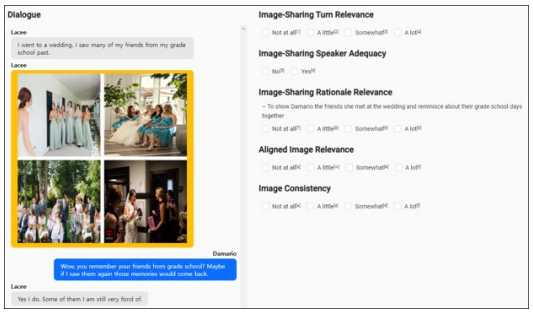


Figure 15: A screenshot of the human evaluation system for the human ratings.

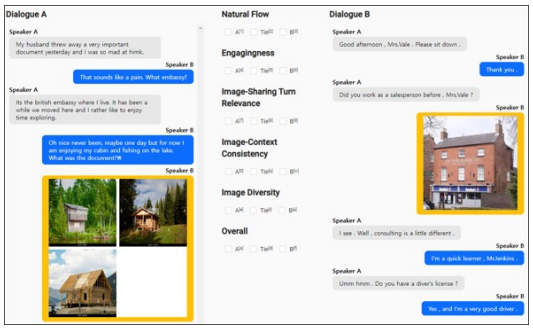


Figure 16: A screenshot of the human evaluation system for the head-to-head comparison.

F.1 Human Ratings

- **Image-Sharing Turn Relevance:** Do you think the image-sharing turn in the given dialogue is appropriate?

  Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

- **Image-Sharing Speaker Adequacy:** Do you think the speaker who shared the image in the given dialogue is appropriate?

  Options: No / Yes

- **Image-Sharing Rationale Relevance:** Do you think the reason for sharing the image in the given dialogue is valid?

  Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

- **Aligned Image Relevance:** How relevant do you think the aligned images are based on the dialogue context?

  Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

- **Image Consistency:** How consistent do you think there is between aligned images?

		Turn Relevance	Rationale Relevance	Aligned Image Relevance	Image Consistency	Speaker (% of Yes)
DialogCC	Avg.	3.68	3.41	3.30	3.57	95.07
	α	0.14	0.39	0.54	0.50	-
KnowledgeCC	Avg.	3.61	3.15	3.05	3.38	99.33
	α	0.14	0.38	0.64	0.59	-
PersonaCC	Avg.	3.84	3.71	3.69	3.80	92.67
	α	-0.03	0.24	0.27	0.59	-
EmpathyCC	Avg.	3.71	3.43	3.30	3.59	97.33
	α	0.12	0.31	0.55	0.56	-
BlendedCC	Avg.	3.67	3.49	3.37	3.61	88.67
	α	0.11	0.45	0.32	0.16	-
DailyCC	Avg.	3.54	3.27	3.11	3.47	97.33
	α	0.16	0.36	0.58	0.45	-

Table 11: Breakdown human evaluation results.

Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

F.2 Head-to-Head Comparison

- **Natural Flow:** Which dialogue has a more natural flow?

  Options: A / Tie / B

- **Engagingness:** Which dialogue has more interesting and engaging?

  Options: A / Tie / B

- **Image-Sharing Turn Relevance:** Which dialogue has a more appropriate image-sharing turn?

  Options: A / Tie / B

- **Image-Dialogue Consistency:** Which dialogue is more consistent between aligned images and dialogue context?

  Options: A / Tie / B

- **Image Diversity:** Which dialogue has more diverse images?

  Options: A / Tie / B

- **Overall:** Which dialogue has higher quality overall?

  Options: A / Tie / B

G Human Evaluation System

We show a screenshot of the human evaluation system in Figure 15 and Figure 16. We implement this system using Label Studio (Tkachenko et al., 2020-2022).

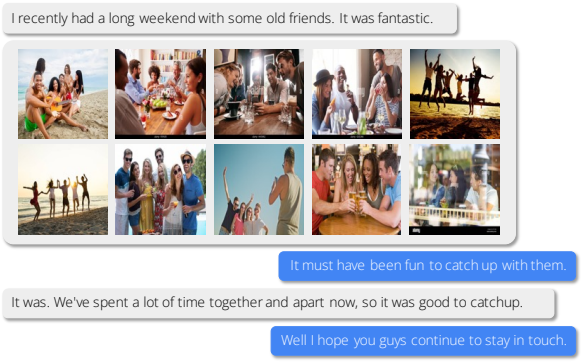


Figure 17: Case 1: An example of DialogCC.

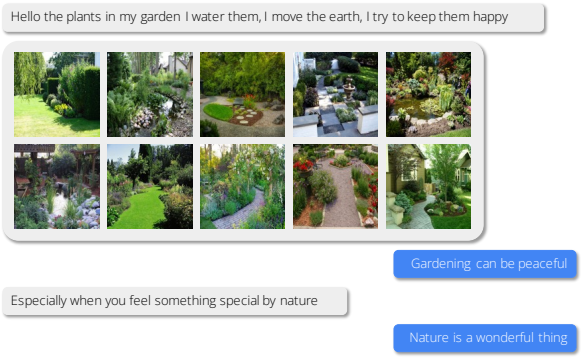


Figure 18: Case 2: An example of DialogCC.

H Details of Human Evaluation

We recruited three individuals, unknown to us, who are either graduate or undergraduate students. Prior to participating in the experiment, they were provided with comprehensive instruction on the task, an overview of the multi-modal dialogue dataset, and a detailed explanation of the evaluation criteria. This preparatory phase lasted approximately one hour. The detailed results of the human evaluation are presented in Table 11.

Eval	MMDD				PhotoChat				MMDialog				DialogCC			
Train Dataset	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR
Image Retrieval																
MMDD	4.20	14.26	22.22	10.91	7.90	24.01	35.34	17.16	5.06	17.47	27.45	12.91	6.72	23.38	36.07	16.35
PhotoChat	3.13	9.52	16.85	8.58	5.41	23.70	38.67	15.87	3.43	13.10	22.07	10.14	4.36	16.45	27.21	12.23
MMDialog	3.47	13.14	20.36	10.01	7.17	23.91	38.05	16.79	**19.79**	**51.13**	**66.56**	**34.66**	8.91	29.70	43.68	20.06
DialogCC	**6.45**	**17.33**	**26.27**	**13.32**	**13.51**	**37.32**	**51.14**	**25.63**	10.97	32.53	45.66	22.40	**17.09**	**46.53**	**62.29**	**31.36**
Next Response Prediction																
MMDD	**19.97**	**40.63**	**50.93**	**30.40**	7.88	21.25	29.45	15.91	8.33	24.08	36.14	17.62	16.44	41.74	55.16	29.16
PhotoChat	6.40	19.09	31.69	14.49	**9.39**	**25.03**	**39.05**	**19.08**	5.18	17.81	28.57	13.18	7.59	24.34	36.53	17.22
MMDialog	9.67	27.10	39.01	19.67	8.95	24.70	34.41	17.92	**34.21**	**61.22**	**72.88**	**46.98**	17.43	40.10	52.51	29.01
DialogCC	18.46	32.52	42.09	26.54	8.09	20.50	29.88	16.26	12.69	30.16	42.02	22.72	**40.64**	**71.46**	**81.99**	**54.61**

Table 12: We report the full results on the next response prediction and image retrieval tasks. The model with the best performance is indicated in **bold**.

Eval →	BlendedCC				DailyCC				EmpathyCC				PersonaCC				KnowledgeCC			
Train ↓	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR
BlendedCC	16.60	47.67	65.02	31.31	12.28	34.16	49.20	23.98	12.87	38.31	54.65	25.79	12.69	36.26	53.21	25.10	19.06	49.43	66.51	33.68
DailyCC	16.19	43.48	60.43	29.77	**21.22**	**52.76**	**68.86**	**36.25**	17.68	46.46	62.05	31.52	10.40	31.17	47.19	21.76	15.90	46.11	63.61	30.58
EmpathyCC	19.89	45.89	62.35	32.88	13.92	42.08	60.45	28.09	**19.80**	**51.22**	**67.64**	**34.82**	11.63	34.15	51.10	23.54	16.00	45.38	62.05	30.32
PersonaCC	17.56	49.11	66.46	32.66	12.19	32.30	47.33	22.99	13.62	38.90	54.06	26.40	14.18	39.86	**57.42**	27.29	17.72	47.80	65.30	32.21
KnowledgeCC	**22.91**	**54.46**	**69.75**	**37.38**	15.61	39.55	53.60	27.83	14.96	39.06	53.03	26.99	**14.75**	**41.22**	57.02	**28.03**	**26.83**	**65.14**	**79.45**	**43.63**

Table 13: We report the image retrieval performance on BlendedCC, DailyCC, EmpathyCC, PersonaCC, and KnowledgeCC. The model with the best performance is indicated in **bold**, while the second best is underlined.

Eval →	BlendedCC				DailyCC				EmpathyCC				PersonaCC				KnowledgeCC			
Train ↓	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR
BlendedCC	**16.51**	44.15	**58.94**	**30.19**	8.27	25.00	37.43	17.82	8.20	23.75	37.52	17.67	10.00	28.66	41.08	20.08	17.18	45.77	62.78	31.29
DailyCC	10.98	33.60	47.16	22.71	**23.70**	**55.71**	**68.69**	**38.57**	8.64	27.99	42.33	19.55	6.30	21.41	32.63	15.17	13.97	42.33	58.49	27.68
EmpathyCC	12.06	35.39	51.69	24.50	9.62	27.20	39.63	19.62	**18.78**	**47.82**	**62.68**	**32.79**	7.62	23.64	35.83	16.80	13.16	39.50	56.51	26.47
PersonaCC	14.50	41.13	54.06	27.28	7.16	20.99	31.41	15.68	6.14	19.95	30.78	14.45	**10.32**	**30.17**	**43.36**	**21.28**	12.07	37.66	53.92	24.78
KnowledgeCC	16.37	**44.44**	57.07	30.00	9.47	26.25	38.43	19.16	9.73	27.79	40.39	19.83	8.90	26.11	37.15	18.34	**25.19**	**62.68**	**75.97**	**41.79**

Table 14: We report the next response prediction performance on BlendedCC, DailyCC, EmpathyCC, PersonaCC, and KnowledgeCC. The model with the best performance is indicated in **bold**, while the second best is underlined.

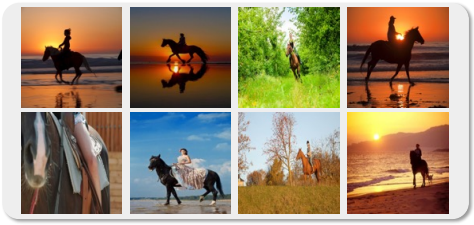
Verb	Object	Count	Example
provide	representation	13,181	To provide a visual representation of yoga practice and the use of a yoga mat
	evidence	7,108	To provide evidence of the value his company adds and support his argument
	example	3,387	To provide a visual example of the kids' behavior that led to her yelling.
	context	1,802	To provide context and show the positive change in the city's policy.
show	example	3,220	To show an example of a craft Delfina made using Dollar Tree items
	excitement	761	To show her excitement and happiness about having a little girl
	type	619	To show the type of tins used for making cupcakes in the past
	appreciation	607	To show appreciation for his friends and emphasize their importance in his life
share	image	1,350	To share a beautiful image of Hawaii that she remembers from her trip.
	experience	1,241	To share a personal experience and highlight the beauty of Savannah
	memory	868	To share a memory of their wedding day or a picture of his wife
	picture	751	To share a picture of the delicious pasta Denese's wife makes
give	idea	1,081	To give an idea of her living situation and the cost of her apartment
	representation	546	To give Tyana a visual representation of the history of horse domestication
	example	212	To give an example of the Beard of the Year award and its winners.
	visual	195	To give Brenner a visual of the Acura to compare with the Integra
showcase	skills	518	To showcase Tre's dancing skills or the type of dance they enjoy
	variety	235	To showcase the variety of species Courtlyn keeps in their aquariums
	work	182	To showcase his work and give Mary a better understanding of what he does.
	passion	170	To showcase her passion for dancing and her favorite Disney moment.
illustrate	concept	251	To illustrate the concept of hydraulic hybrids and how they store energy.
	difference	199	To illustrate the difference in the mountain scenery between January and April.
	process	107	To illustrate the batting process and the pitcher's role in the game
	connection	94	To illustrate the connection between the company's name and its inspiration.
emphasize	importance	211	To emphasize the importance of high-quality ingredients in Italian cooking
	love	50	To emphasize her love for 2pac and how it complements her black car
	popularity	42	To emphasize the popularity of My Little Pony toys in the 80s
	preference	38	To emphasize the preference for a kitten as a pet over a snake.
support	statement	127	To support their statement about liking pop music and finding it lovely.
	argument	30	To support the argument about the lack of educational programs and poorly done news shows.
	claim	27	To support his claim and provide evidence for his prediction.
	opinion	26	To support his opinion about Professor Wood and provide visual evidence
express	interest	45	To express his interest in trying mountain biking as another alternative sport
	love	32	To express her love for McDonald's breakfast and coffee
	gratitude	30	To express gratitude and acknowledge the teacher's role in their success.
	excitement	23	To express her excitement and share the news of winning the prize
demonstrate	process	43	To demonstrate the process of adding a web page to the favorites list.
	skills	38	To demonstrate her juggling skills and her work in the circus.
	technique	23	To demonstrate the technique of playing the guitar in rock music
	ability	21	To demonstrate the cat's ability to see in low light conditions
confirm	order	31	To confirm the order and show the specific items requested
	details	29	To confirm the booking details and provide a visual summary of the reservation.
	time	23	To confirm the appointment time and show that he will bring his husband.
	understanding	19	To confirm her understanding of desert classification and provide a visual aid
introduce	dog	24	To introduce her dog to Rance and show how it helps her
	pet	23	To introduce his pet and show how it looks.
	topic	19	To introduce the topic of baseball and initiate a conversation about it
	cat	18	To introduce her cat named after a Cars character
clarify	difference	31	To clarify the difference between divorce and annulment for Maxwell
	confusion	16	To clarify Ryley's confusion about Osiel's profession and provide a visual example
	concept	13	To clarify the concept of nearsightedness for Conrad.
	misconception	13	To clarify the misconception about black roses and show the actual dark red rose.
celebrate	achievement	47	To celebrate her achievement and share her excitement with Shanya
	promotion	9	To celebrate Britney's promotion and share the news with others.
	birthday	8	To celebrate Rupert's birthday and make the moment memorable
	accomplishment	7	To celebrate the accomplishment and share the excitement with Ayelet.
suggest	activity	15	To suggest an alternative activity for her kids instead of watching TV.
	place	14	To suggest a place to eat and provide a visual reference
	restaurant	12	To suggest a specific restaurant or location for their next hangout
	solution	9	To suggest a solution to make up for the lie and mend the relationship
explain	concept	24	To explain the concept of two hand touch and flag football visually
	process	10	To explain the process of setting the minimum wage and the parties involved.
	reason	8	To explain the reason for the stain and show their efforts to remove it

Table 15: The top 20 most common root verbs and their up to 4 direct noun objects in the generated rationale. Only pairs with a count of 5 or more are included.

Horse

They are so graceful and powerful. I would love to have horses.

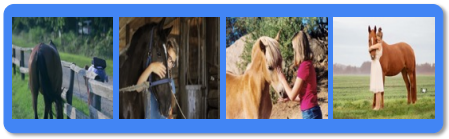
They are beautiful animals I have ridden horses my whole life.



Have you thought about becoming a veterinarian?

Yes I have I would like to go to school and become one.

You should totally be an equine veterinarian and work with horses every day



I would love to do that for a living.

Horses are so cool. My parents are doctors, it's a good life whether for people or for animals

Do you want to become a doctor?

Not sure if I could live up to the standard of my parents.

What do you want to do for work?

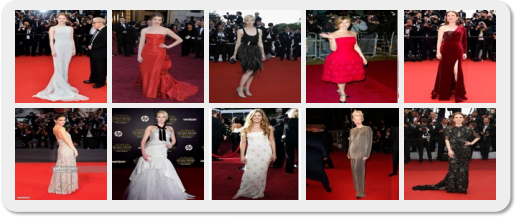
I'm more interested in the arts and theater.

Figure 19: Case 3: An example of DialogCC.

I would like to watch the oscars on tv tonight. How about you?

Yes, I'd love to. It's interesting to see who is considered the best in their field and which film are thought to be particularly good.

I like watching it for the fashion. I like to see what the ladies are wearing. Of course, the men nearly always just wear the traditional tuxedo.



Sometimes the men wear flamboyant colours. Which films do you think will win awards this year?

I'm really not sure. Usually just one or two films look set to sweep the awards ceremony, but this year there are several contenders.

You're right. This year should be much more exciting than usual. What's you favourite award category?

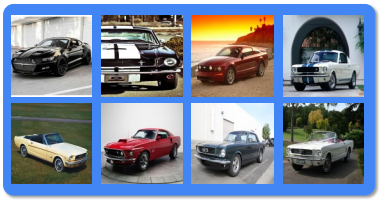
You might think this strange, but I like the category for best foreign language film.

It's nice to see foreign language films making a little impact on hollywood. I like the best actor and actress.

Figure 20: Case 4: An example of DialogCC.

Thinking about buying a Ford Mustang, I have wanted one for a long time

I just bought a Mustang yesterday! Awesome car! It's manufactured by Ford, and is American! I'm just about to take it for a spin to impress my girlfriend!



Nice! The Mustang created the "pony car" class of American muscle cars, affordable sporty coupes

They sure did! I went for the Shelby Mustang - high performance, built by Shelby American! It really takes those corners nicely! I think my girlfriend will love it!

The Mustang is also credited for inspiring the designs of coupés such as the Toyota Celica and Ford Capri, which were imported to the United States.

Really? Interesting stuff! As for my Shelby, it has a Cobra emblem and a "Cobra" valve cover - I love things like that! It looks so cool!

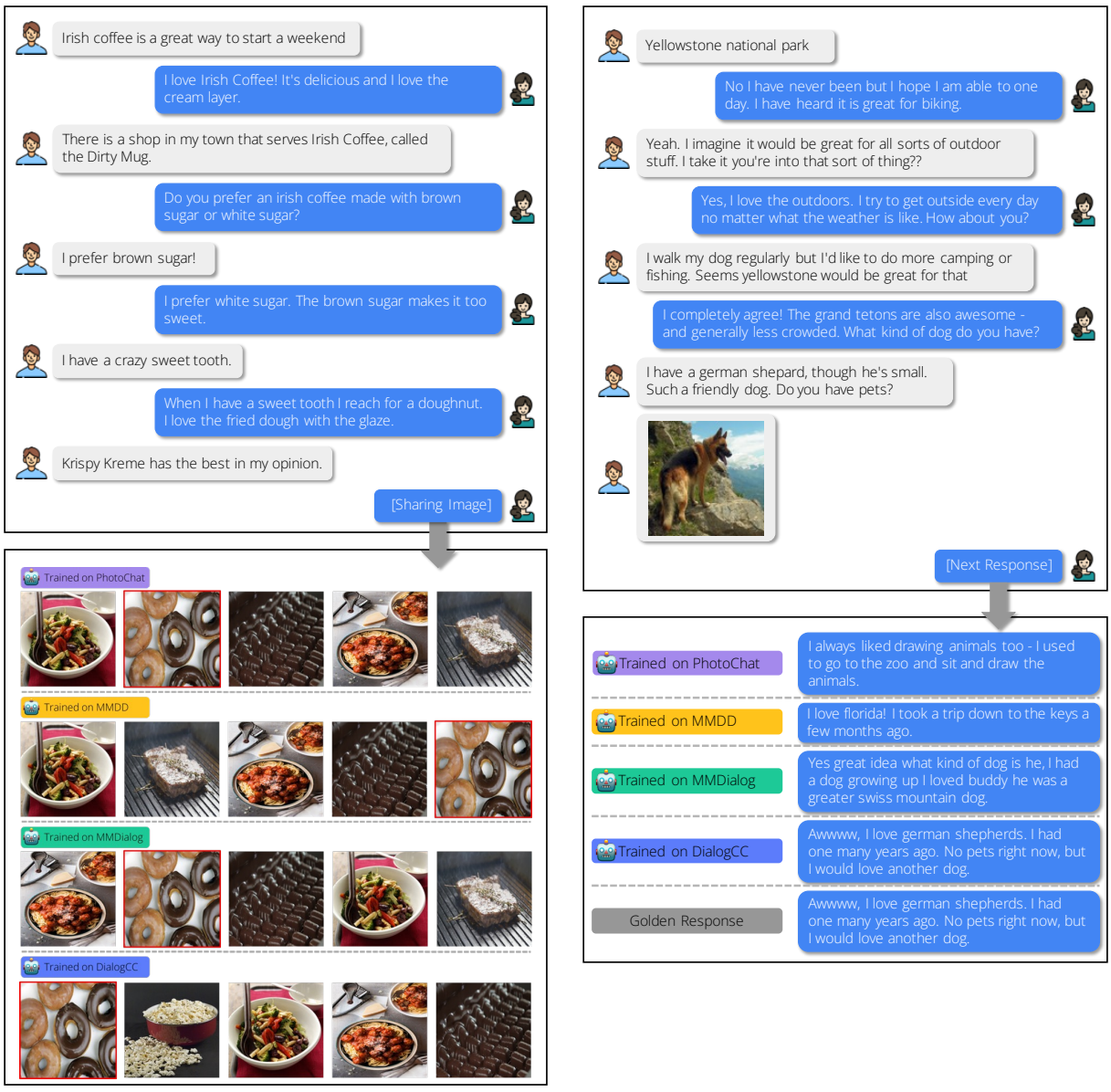
Mustangs are safe too, In February 2015, the Mustang earned a 5-star rating from the National Highway Traffic Safety Administration (NHTSA) for front, side, and rollover crash protection

They ARE safe! Something my mother insisted on! Mothers, eh! Interestingly, Mustangs have experienced several transformations to its current generation of cars!

Yes The 2018 model year Mustang was released in the third quarter of 2017 in North America and by 2018 globally

Yep! And the Mustang was actually based on the North American Ford Falcon - its second generation! Well, the girlfriend is getting impatient - time to fly!

Figure 21: Case 5: An example of DialogCC.



(a) Image retrieval

(b) Next response prediction

Figure 22: Two examples of retrieved results (i.e., (a) image retrieval and (b) next response prediction) from four different models. Each provided dialogue is from the DialogCC dataset. In (a), we display the top-5 ranked images from left to right, with the ground-truth image marked in red. In (b), only the top-1 ranked next response is shown. Note that neither the `[Sharing Image]` turn nor the `[Next Response]` turn is provided to the model's input during the inference stage.