MMDialog: A Large-scale Multi-turn Dialogue Dataset Towards Multi-modal Open-domain Conversation

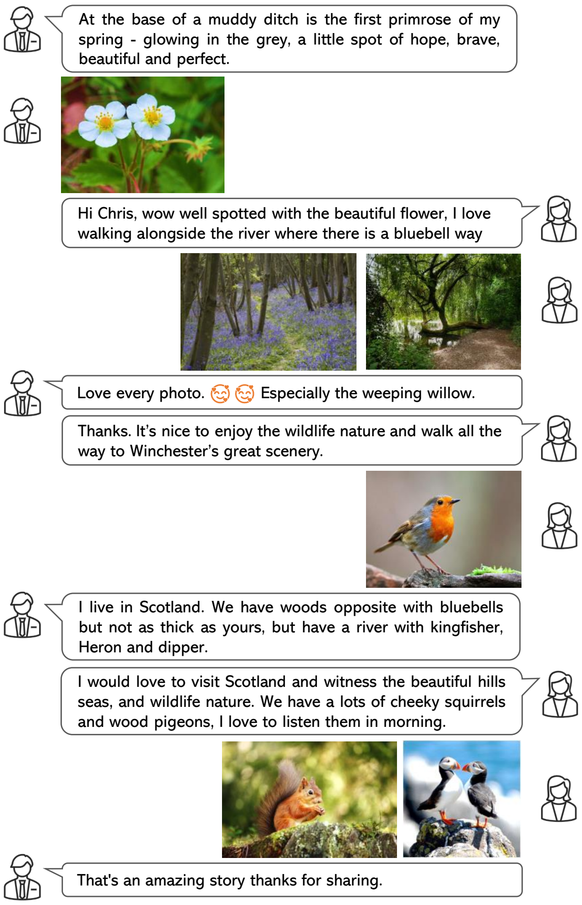
Jiazhan Feng[1]* **Qingfeng Sun**[2] **Can Xu**[2] **Pu Zhao**[2]
Yaming Yang[2] **Chongyang Tao**[2] **Dongyan Zhao**[1,3] **Qingwei Lin**[2]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]Microsoft Corporation
[3]National Key Laboratory of General Artificial Intelligence, BIGAI
{fengjiazhan,zhaody}@pku.edu.cn
{qins,caxu,puzhao,yayaming,chotao,qlin}@microsoft.com

Abstract

Responding with multi-modal content has been recognized as an essential capability for an intelligent conversational agent. In this paper, we introduce the MMDialog dataset to facilitate multi-modal conversation better. MMDialog is composed of a curated set of 1.08 million real-world dialogues with 1.53 million unique images across 4,184 topics. MMDialog has two main and unique advantages. First, it is the largest multi-modal conversation dataset by the number of dialogues by 88x. Second, it contains massive topics to generalize the open domain. To build an engaging dialogue system with this dataset, we propose and normalize two response prediction tasks based on retrieval and generative scenarios. In addition, we build two baselines for the above tasks with state-of-the-art techniques and report their experimental performance. We also propose a novel evaluation metric MM-Relevance to measure the multi-modal responses. Our dataset is available in https://github.com/victorsungo/MMDialog.

1 Introduction

Empowering machines to converse like humans is a long-cherished goal of AI community, and there is growing interest in developing open-domain conversational agents (Li et al., 2017a; Gao et al., 2018; Ghazvininejad et al., 2018; Zhou et al., 2018a; OpenAI, 2022; Thoppilan et al., 2022). To usher machines into the real visual physical world of human, it is a desirable trait of conversational agents to understand, perceive, and respond appropriately to multi-modality contexts beyond pure text (Das et al., 2017; Mostafazadeh et al., 2017; Shuster et al., 2020; Sun et al., 2022), which is similar to communicating through messenger tools (e.g., Facebook, WhatsApp, and WeChat) in reality.



At the base of a muddy ditch is the first primrose of my spring - glowing in the grey, a little spot of hope, brave, beautiful and perfect.

Hi Chris, wow well spotted with the beautiful flower, I love walking alongside the river where there is a bluebell way

Love every photo. 😊😊 Especially the weeping willow.

Thanks. It's nice to enjoy the wildlife nature and walk all the way to Winchester's great scenery.

I live in Scotland. We have woods opposite with bluebells but not as thick as yours, but have a river with kingfisher, Heron and dipper.

I would love to visit Scotland and witness the beautiful hills seas, and wildlife nature. We have a lots of cheeky squirrels and wood pigeons, I love to listen them in morning.

That's an amazing story thanks for sharing.

Figure 1: An example of human conversation in our MMDialog dataset. They are talking about scenery and wildlife with both text and various images.

Yet existing approaches to building multi-modal dialogue systems are primarily data-driven, requiring the collection of a large-scale dataset first. To facilitate this line of research, the community emerges a few dialogue datasets incorporating visual information (Meng et al., 2020; Wang et al.,

[*] Work done during the internship at Microsoft Research Asia. Corresponding author: caxu@microsoft.com and puzhao@microsoft.com

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7348–7363
July 9-14, 2023 ©2023 Association for Computational Linguistics

Dataset	Genre	Open-domain	Dialog Source	Response Modalities	#(Dialog Session)
Visual Dialog v1.0 (Das et al., 2017)	VQA	✔	Crowdsourcing	Plain Text	133.35K
IGC (Mostafazadeh et al., 2017)	Image-grounded	✔	Crowdsourcing	Plain Text	4.22K
MMD (Saha et al., 2018)	Goal-oriented	✘	Domain Experts	Image+Text	150K
MDMMD (Firdaus et al., 2020)	Goal-oriented	✘	Domain Experts	Plain Text	132K
Image-Chat (Shuster et al., 2020)	Image-grounded	✔	Crowdsourcing	Plain Text	202K
OpenViDial (Meng et al., 2020)	Video Frames & Subtitles	✔	Movies&TVs	Plain Text	≈78.57K
OpenViDial 2.0 (Wang et al., 2021)	Video Frames & Subtitles	✔	Movies&TVs	Plain Text	≈116.67K
MMChat (Zheng et al., 2022)	Image-grounded	✔	Social Media	Plain Text	120.84K
PhotoChat (Zang et al., 2021)	Casual Visual+Text	✘	Crowdsourcing	Image+Text	12.29K
MMDialog (Ours)	Casual Visual+Text	✔	Social Media	Image+Text	1.08M (Million-scale)

Table 1: A summary of main multi-modal dialogue datasets. We can conclude that MMDialog is the only multi-modal dialogue dataset that satisfies the following criteria simultaneously: 1) Million-scale multi-turn dialogue sessions; 2) The modality of the response can be image or text or a flexible combination; 3) Casual, open-domain human-human visual+text dialogues taken from daily life.

2021; Zang et al., 2021; Zheng et al., 2022). For example, Visual Dialog (Das et al., 2017) is set up for visual question answering involving image inputs. IGC (Mostafazadeh et al., 2017) and Image-Chat (Shuster et al., 2020) are constructed in a crowd-sourcing method in which annotators are employed to chat about given images only. MM-Chat (Zheng et al., 2022) initiates the conversation with images but does not constrain the topic of subsequent responses. OpenViDial (Meng et al., 2020; Wang et al., 2021) cuts out the image frames and captions in the video as dialogue. PhotoChat (Zang et al., 2021) is also built via crowd-sourcing and contains photo sharing in conversations.

Despite the diversity of multi-modal dialogue corpora, these datasets still have limitations: i) Except for PhotoChat, the responses in the above datasets are all plain text. ii) Visual Dialog, IGC and Image-Chat are all defined as limited scenarios, focusing only on the QA or conversation based on the topics in given images. iii) OpenViDial is mined from movies and TV series. The paired frame-caption is far from real-world dialogue. iv) PhotoChat is the closest to real-world multi-modal dialogue. However, it is still limited by small scale, only having single-modal responses (text or image), and lack of domain diversity, impeding further explorations on multi-modal dialogue modeling.

To address the aforementioned issues, we present MMDialog, a large-scale multi-turn dialogue dataset containing multi-modal open-domain conversations derived from real-world human-human interactions in an online social media platform. Comprising 1.08M dialogue sessions and 1.53M associated images, MMDialog boasts an average of 2.59 images per dialogue session, which can appear anywhere at any conversation turn. Figure 1 depicts an example of human conversations in our MMDialog. To the best of our knowledge, this is the first million-scale open-domain multi-modal dialogue corpus. We anticipate that the large amount of dialogues and images can shed light on this line of research. We summarize the main characteristics of multi-modal dialogue datasets in Table 1.

Furthermore, we define the multi-modal response generation and retrieval tasks based on MMDialog that are essential for building a more engaging multi-modal dialogue agent. We also build baseline models and conduct several analyses of their performance. For the generative task, we follow Sun et al. (2022) and implement the models for multi-modal response generation. For the retrieval task, we also propose a CLIP-based retrieval model inspired by Zang et al. (2021). Evaluation results on MMDialog demonstrate that our designed baselines can achieve considerable performance on generation and retrieval tasks of both modalities.

Since the modality orders and amounts of predicted responses may not be aligned with the ground-truth responses, it is difficult for us to evaluate the response quality comprehensively and reasonably using single-modal metrics (e.g., BLEU, PPL, FID, Recall@K, etc.). To tackle the above challenges, we propose a novel evaluation metric called MM-Relevance, which performs visual-language matching based on the large-scale pretrained multi-modal CLIP model (Radford et al., 2021). Experiments on both retrieval and generative multi-modal dialog systems indicate that MM-Relevance outperforms existing single-modal metrics in terms of correlation with human judgments.

Contributions of this work are four-fold:

1. We construct a novel multi-turn dialogue

dataset **MMDialog** that contains 1.08M multi-modal open-domain conversations and 1.53M associated images derived from social media. To the best of our knowledge, this is the first million-scale multi-turn open-domain multi-modal dialogue corpus.

2. We propose two benchmark tasks including generative and retrieval scenarios on MMDialog that are essential for building more engaging multi-modal dialogue systems.
3. We design two baselines for corresponding tasks to promote future research on this dataset and achieve considerable performance.
4. We propose a novel evaluation metric **MM-Relevance** measuring the relevance between generated multi-modal response and ground truth. It can specifically mitigate the modal misalignment issues.

2 Related Works

2.1 Multi-Modal Dialogue Datasets

Recently, there emerge several multi-modal dialogue datasets. Das et al. (2017) introduce the task of Visual Dialog, which requires an AI agent to hold a meaningful dialogue with humans in natural, conversational language about visual content. Mostafazadeh et al. (2017) propose IGC, which contains 4K dialogues where each includes an image with a textual description, along with the questions and responses around the image. However, IGC is usually used for evaluation due to its small scale. Shuster et al. (2020) release Image-Chat that is larger than IGC and consists of 202K image-grounded dialogues. However, the above three datasets were created by asking the crowd workers to talk about a shared image to generate the conversation. Therefore, the utterances are often triggered and grounded by these images. In contrast, human daily communication utterances are not always image-related (Zheng et al., 2022), which retain gaps with open-domain multi-modal conversation scenarios. Then, other groups of works proposed to derive the images from the multi-turn conversations: Meng et al. (2020); Wang et al. (2021) construct OpenViDial 1.0/2.0 by directly extracting dialogues and their visual contexts from movies and TV series. Lee et al. (2021) also build a multi-modal dialogue dataset by replacing the selected utterances with retrieved relevant images. However, although these corpora were constructed from open-domain conversations with images, they did not originate from a real multi-modal conversation scenario. Therefore, recently some researchers begin to introduce real human-human conversations. Zang et al. (2021) create the first human-human dialogue dataset with photo-sharing acts via crowd-sourcing. Zheng et al. (2022) collect multi-modal dialogues from real conversations on social media. Nevertheless, they were still limited by their small scale or lack of domain diversity, which may hinder further explorations on multi-modal dialogue modeling. To address the above issue, we make the first attempt to construct a million-scale multi-turn dialogue dataset, namely **MMDialog**, derived from social media and conduct data filtering and post-processing elaborately.

2.2 Open-Domain Conversation

Open-domain dialogue systems can converse on a broad range of topics. Users often do not have any specific goal when participating in open-domain interactions. Early on, several open-domain dialog datasets are constructed from social media such as Twitter, Weibo and Reddit. Advanced datasets including (Ritter et al., 2010; Wang et al., 2013; Sordoni et al., 2015; Shang et al., 2015; Wu et al., 2017; Li et al., 2017b; Zhang et al., 2018b; Adiwardana et al., 2020). Later, researchers begin to incorporate knowledge into conversations (Zhang et al., 2018a; Dinan et al., 2018; Zhou et al., 2018b; Gopalakrishnan et al., 2019; Qin et al., 2019; Moon et al., 2019; Rashkin et al., 2019; Tuan et al., 2019; Song et al., 2020). Meanwhile, there emerge several advanced models based on these corpora including generative ones (Zhang et al., 2020; Shuster et al., 2022; OpenAI, 2022) and retrieval-based ones (Wu et al., 2017; Whang et al., 2021; Han et al., 2021; Feng et al., 2022).

2.3 Multi-Modal Dialogue Modeling

Based on the multi-modal dialogue datasets, many advanced works have been proposed. Several works (Niu et al., 2019; Gan et al., 2019; Qi et al., 2020) investigate how to escalate the performance of conversational agents in image-grounded dialogue. Afterward, researchers (Yang et al., 2021; Liang et al., 2021) explore enriching textual expression of generated dialogue responses through associative vision scenes. Zang et al. (2021) propose two tasks, including photo-sharing intent prediction to predict whether model should intend to share the photo in the next dialogue turn and a dialogue-based image retrieval task to retrieve the most proper photo given the dialogue context. They also propose a dual-encoder model that uses object labels to encode image features. However, the authors do not conduct textual response retrieval tasks. Zheng et al. (2022) propose a Seq2Seq based multi-modal dialogue generation model. However, it only generates plain textual responses, which is not in line with the open domain multi-modal response generation scenario. Recently, Sun et al. (2022) make the first attempt to build a multi-modal dialogue response generation model named Divter that can effectively understand multi-modal dialogue context and generate informative text and high-resolution images. In this paper, we adapt Divter to our multi-modal response generation settings and extend the dual-encoder (Zang et al., 2021) to the retrieval-based scenarios as baselines.

3 Dialogue Creation

MMDialog is a large-scale multi-turn dialogue dataset towards multi-modal open-domain conversations. It derives from a worldwide social media platform where users can converse with each other and share daily life experiences freely through various modalities, including plain text, photos, or even videos. We design the data collection process into three phases: In the first phase, we extensively manually collect the hashtags commonly used by users and cover as many domains as possible. The second phase starts from the seed hashtags collected in phase one and focuses on collecting all turns that contain at least one image, which we subsequently refer to as *anchors*. For each anchor, we retrieve all the turns that replied to it and the turn it replied to. We could obtain a complete dialogue session by performing the above process iteratively. In the final phase, we also elaborately design a series of data filtering and post-processing steps to eliminate invalid cases and improve the quality of multi-modal dialogues in MMDialog. To protect the privacy and security of the data, users, and platform, MMDialog is only available to academic researchers under strict terms.

3.1 Hashtag Collection

To collect MMDialog, we crawl one of the most influential online social platform utilizing its academic access API. To improve the data quality, we consider extracting dialogues with their hashtags (e.g., '#travel', '#friends', '#golf'), as hashtags tend to show the primary topic of the textual utterances and accompanying visual media. Specifically, we manually screen out 4,184 popular hashtags, each with a minimum of 1,000 dialogues. In this way, our dataset can not only satisfy the properties of open-domain, but also ensure a large scale. We depict the most popular hashtags in Appendix A.1.

3.2 Multi-modal Conversations Construction

Then, we leverage these hashtags as initial seeds for constructing multi-turn dialogues. Initially, for each hashtag, we crawl the turns containing the corresponding hashtag and only retain those incorporating at least one image object (i.e., *anchors*). Obviously, dialogues containing the anchors are the multi-modal multi-turn dialogues we pursue. Next, within the same conversation, for each anchor, we could seek out all the other turns i) that replied to the anchor until they reach the leaf node (i.e., the end of conversation), and ii) that anchor replied to up to the root node (i.e., the start of conversation). Thus, we could recursively follow the chain of replies to recover the entire conversation.

3.3 Data Filtering and Post-processing

Since the style of messages posted on social media platforms is widely varied, the initial version of MMDialog contains a significant number of invalid, noisy, and even harmful conversations, which may impede research utilizing this dataset. To tackle the above issue, we design a series of sophisticated data filtering processes to filter out high-quality multi-modal conversations: a) We remove dialogues containing toxic statements with explicit offensive words; b) We ignore and discard dialogues with GIFs and other modalities (such as videos) which cannot be downloaded immediately. We leave this part of research as future work; c) We remove irregular characters from the dialogue content. For example, we do not consider any urls and '@' items (i.e., expression items for mentioning somebody); d) In particular, we convert emojis and hashtags into corresponding natural language forms to guarantee the coherence of the dialogues; e) We remove all self-talking cases (such as replying to themselves for two or more consecutive dialogue turns) to enhance the integrity of the conversations; f) We discard dialogues with incomplete or missing images; g) We only keep the conversations of no less than three dialogue turns. We believe that by adopting the above data-filtering and post-processing procedure, the final large-scale multi-turn dialogues can be better leveraged to develop

Statistics	PhotoChat	MMDialog
Language	English	English
Open-domain	✘	✔
#Dialogues	12.29K	1.08M
#Images	10.92K	1.53M
#Turns	156.10K	4.92M
#Topics/Objects	89	4,184
Avg. #Turns per Dialogue	12.71	4.56
Avg. #Images per Dialogue	0.89	2.59
Avg. #Tokens per Turn	6.33	15.90

Table 2: Statistics of MMDialog and previous multi-modal dialogue dataset PhotoChat.

multi-modal open-domain conversation models.

4 Corpus Statistics

MMDialog consists of 1,079,117 unique dialogues and 1,531,341 images. The statistics of several multi-modal open-domain dialogue corpora are detailed in Table 2. On average, each dialogue session in MMDialog has 2.59 images and 4.56 turns, with the images being located arbitrarily within any conversation turns. This reflects the freedom with which individuals can choose any conversational modality at any conversation stage in daily life. Compared to the recently released multi-modal open domain dialogue dataset PhotoChat (Zang et al., 2021), MMDialog enjoys a significantly larger scale of dialogue data and more visual objects, especially that the volume of dialogue sessions has reached million-level. The dialogues in MMDialog, which originate from a wide range of hashtags representing broad domains, are open-domain and cover diverse topics, shedding light on research of multi-modal dialogue modeling. Besides, on average each dialogue turn in MMDialog contains more text tokens than PhotoChat, demonstrating that our proposed data may convey more semantic information in textual utterances.

5 Task Definition

Suppose that we have a multi-modal dialogue dataset $\mathcal{D} = \{(U_i, R_i)\}_{i=1}^{n}$, where $\forall i \in \{1, ..., n\}$, U_i is the dialogue context, R_i is the response regarding to U_i. U_i and R_i could contain multi-modal components: textual elements (e.g., utterances) and visual elements (e.g., images). For any U and R, we denote $U_i = \{u_k^m\}_{k=1}^{K}$ and $R_i = \{r_l^m\}_{l=1}^{L}$ as sequence of multi-modal elements including textual utterances and visual images. K and L are the number of elements in context and response respectively. $m \in \{t, v\}$ indicates the modal type of elements where t represents textual utterances while v signifies visual images.

Since advanced works on pure-text open-domain dialogue systems mainly include retrieval and generative methods. We adapt them to multi-modal scenarios and define the following two tasks that are essential for building a multi-modal open-domain dialogue system:

Task-1: Multi-modal Response Generation To generate a multi-modal response R, one should learn a multi-modal generation model $g(R|U)$ from D. Thus, given a new dialogue context U, following $g(R|U)$, one can directly synthesize a multi-modal response $\tilde{R}$ consisting of textual utterances or visual images, or both of them.

Task-2: Multi-modal Response Retrieval The goal of this task is to learn a multi-modal response matching model $s(\cdot, \cdot)$ from D. For any context-response elements (U, R) pair, $s(U, R)$ gives a score that reflects the matching degree between U and R, and thus allows one to rank a set of response candidates C according to the scores for response matching, C consists of both textual utterances and visual images. Thus we can obtain a multi-modal response $\tilde{R}$ consisting of textual utterances or visual images, or both of them.

6 Evaluation of Multi-Modal Responses

6.1 General Metrics

In **Task-1**, we follow Sun et al. (2022) and evaluate the quality of the generated responses by using metrics from both text and visual modalities. For the text modality, we compare the generated text to the ground-truth responses using metrics like **BLEU** (Papineni et al., 2002) and **ROUGE** (Lin, 2004). For the visual modality, we measure the quality of the generated images using metrics like Inception Score (Salimans et al., 2016) (**IS**) and CLIP Score (Radford et al., 2021) (**CLIP-Sim**). Suppose the model does not generate a corresponding textual or visual element, we assign it a score of zero in that case, indicating no relevance between the "empty" element and the ground truth. However, we could not directly adopt the same strategy for Perplexity (**PPL**, for text) and Frechet Inception Distance (Heusel et al., 2017) (**FID**, for image). As a lower PPL (or FID) implies a closer distance between generated text (or image) distribution and

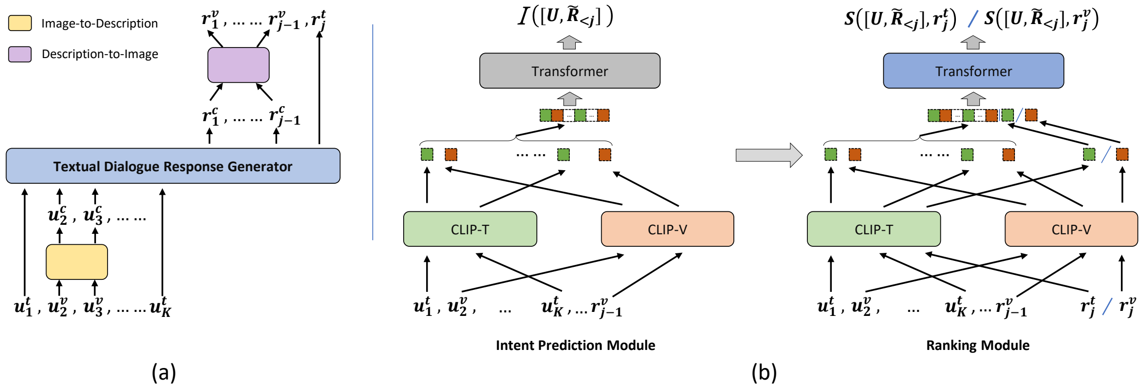


Image-to-Description
Description-to-Image
r_1^v , …… r_{j-1}^v , r_j^t
r_1^c , …… r_{j-1}^c
Textual Dialogue Response Generator
u_2^c , u_3^c , ……
u_1^t , u_2^v , u_3^v , …… u_K^t
$I([U, \tilde{R}_{<j}])$
Transformer
CLIP-T
CLIP-V
u_1^t , u_2^v , … u_K^t , … r_{j-1}^v
Intent Prediction Module
$S([U, \tilde{R}_{<j}], r_j^t)$ / $S([U, \tilde{R}_{<j}], r_j^v)$
Transformer
CLIP-T
CLIP-V
u_1^t , u_2^v , … u_K^t , … r_{j-1}^v r_j^t / r_j^v
Ranking Module
(a)
(b)

Figure 2: The overview of multi-modal response generation (a) and retrieval (b) baselines.

real-world text (or image) distribution. Unfortunately, we could not assign an infinity value (similar with zero value for BLEU/IS metrics) standing for "empty" element for both metrics.

In **Task-2**, we could also compute the **Recall@N** scores for text and image in similar way. Specifically, for each retrieval step, if the predicted modality intent of next element mismatches with the ground-truth, we could also assign the recall score of this step to zero values. The final recall scores of an example is the average scores of all elements group by the same modality.

Inspired by Zang et al. (2021) which use F1 score to evaluate intent prediction, we extend it from single to multi elements: our **F1** score could measure whether the models predict right order of elements modality in a response. Specifically, we first get the modality sequence of generated/retrieved and ground-truth responses as $\tilde{M} = \{\tilde{m}_j\}_{j=1}^{J}$ and $M = \{m_l\}_{l=1}^{L}$ respectively. Then, the F1 score is formulated as:

$$\begin{aligned} \text{F1}_{\text{Intent}} &= \frac{2\text{P}_{\text{intent}}\text{R}_{\text{intent}}}{\text{P}_{\text{intent}} + \text{R}_{\text{intent}}} \\ \text{P}_{\text{intent}} &= \frac{\text{Match}(M, \tilde{M})}{J} \\ \text{R}_{\text{intent}} &= \frac{\text{Match}(M, \tilde{M})}{L} \qquad (1) \\ \text{Match}(M, \tilde{M}) &= \sum_{i=1}^{\min\{L,J\}} \mathbb{1}(m_i = \tilde{m}_i) \end{aligned}$$

where $\mathbb{1}$ is an indicator function that has value 1 when $m_i = \tilde{m}_i$, otherwise 0. J and L are the number of elements in responses. More details are described in Appendix A.2.

6.2 MM-Relevance

However, there exist two shortcomings of above metrics: i) They can only reveal the model performance of a single modality, and can not comprehensively measure the overall quality of multi-modal responses. ii) They address the modal-misalignment issue between prediction and ground truth via assigning zero value, which overlooks the relevance between image and text elements.

Thus we propose a novel evaluation metric, dubbed MM-Relevance, which performs visual-language matching based on the large-scale pre-trained multi-modal CLIP model (Radford et al., 2021) for multi-modal dialogue response generation and retrieval tasks. In specific, suppose we obtain a multi-modal response $\tilde{R} = \{\tilde{r}_j^m\}_{j=1}^{J}$, and the corresponding ground-truth $R = \{r_l^m\}_{l=1}^{L}$. We first align the two sequences from the left. Then, the representation vector of each element is obtained by encoding it through CLIP-Text-Encoder or CLIP-Image-Encoder respectively. We denote the encoded vectors of two responses as: $\tilde{E} = \{\tilde{e}_j^m\}_{j=1}^{J}$ and $E = \{e_l^m\}_{l=1}^{L}$. Then, we compute the CLIP relevance of the two elements from left to right until they cannot be aligned:

$$\text{MM}_{\text{Rel}}(R, \tilde{R}) = \sum_{i=1}^{\min\{L,J\}} (e_i^m)^{\mathrm{T}} \cdot \tilde{e}_i^m \qquad (2)$$

In order to penalize the $\tilde{R}$ that is too long or short, we further improve MM-Relevance as F1_{MM}:

$$\begin{aligned} \text{F1}_{\text{MM}} &= \frac{2\text{P}_{\text{MM}}\text{R}_{\text{MM}}}{\text{P}_{\text{MM}} + \text{R}_{\text{MM}}} \\ \text{P}_{\text{MM}} &= \frac{\text{MM}_{\text{Rel}}(R, \tilde{R})}{J} \qquad (3) \\ \text{R}_{\text{MM}} &= \frac{\text{MM}_{\text{Rel}}(R, \tilde{R})}{L} \end{aligned}$$

The relevance can be computed thoroughly between two modal-misaligned responses R and $\tilde{R}$.

6.3 Correlation With Human Annotation

To verify the effectiveness of MM-Relevance in measuring the relevance between a response and

Statistics	Training	Validation	Test
#Dialogues	1,059,117	10,000	10,000
#Images	1,509,284	23,812	23,772
#Turns	4,825,053	45,382	45,801
Avg. #Turns per Dialogue	4.56	4.54	4.58
Avg. #Images per Dialogue	2.59	2.58	2.62
Avg. #Tokens per Turn	15.90	15.98	15.84
Avg. #(Neg. Images) per Dialogue	-	999	999
Avg. #(Neg. Utterances) per Dialogue	-	999	999

Table 3: Statistics of training, validation, and test sets.

the ground truth, we consider evaluating the correlation between metrics in Task-1/2, including MM-Relevance and human annotation on MMDialog. Following (Tao et al., 2018; Gao et al., 2021; Ghazarian et al., 2022), we randomly selected 200 contexts from the test set and solicited ratings from three English-speaking volunteers on a 1-5 Likert scale, reflecting their satisfaction with the response (retrieved or generated) to the given context. We ensure that each response receives three valid ratings, where the average value is used as the final human judgment. We excluded reference-free metrics such as IS and Recall@K, as they do not consider the ground truth. Besides, as Intent-F1 only measures the order accuracy of elements' modalities in response while ignoring the information of each modal element, we also get rid of this metric.

7 Baselines

7.1 Multi-modal Response Generation Model

We consider to implement the state-of-the-art multi-modal response generation model Divter (Sun et al., 2022) (Figure 2a), which consists of two components: a textual dialogue response generator $\mathcal{G}$ and a description-to-image translator $\mathcal{F}$.

Specifically, $\mathcal{G}$ takes the dialogue context U as input, then generates a textual sequence which may contains a textual response r^t or a textual image description r^c or both of them. Noting that in our settings on MMDialog, there may also be several images u^v in multi-turn dialogue context, we thereby replace these images by their descriptions u^c with the help of an image-to-description translation model. In this way, we could concatenate the textual utterances u^t and descriptions u^c into a sequence as the input of $\mathcal{G}$. Then, for a generated description r^c , $\mathcal{F}$ would take them as condition input, and generate a realistic and consistent high resolution image r^v as the real response.

7.2 Multi-modal Response Retrieval Model

Inspired by Parekh et al. (2021) and Zang et al. (2021), we also build a retrieval model $\mathcal{R}$ named DE++ which consists of a modality intent prediction module $\mathcal{I}$ and a ranking module $\mathcal{S}$. As shown in Figure 2b, before each ranking action, $\mathcal{I}$ firstly takes the dialogue context U and previous retrieved $j-1$ response elements $\tilde{R}_{<j}$ as input, then predicts the intent $\mathcal{I}([U, \tilde{R}_{<j}])$ as: i) the response is completed and model should stop retrieving new elements. or ii) the modality of next elements. If ii), $\mathcal{S}$ will calculate the relevance score $\mathcal{S}([U, \tilde{R}_{<j}], r_z^m)$. In which $C = \{r_z^m\}_{z=1}^{Z}$ is the candidates set. Eventually, $\mathcal{S}$ will select the one with highest relevance score as j-th element r_j^m of $\tilde{R}$.

Specifically, $\mathcal{I}$ and $\mathcal{S}$ have similar architecture, we adopt CLIP-Text-Encoder and CLIP-Image-Encoder to represent text and image respectively. In $\mathcal{I}$, we concatenate all the context embeddings with a `[CLS]` embedding prepending at the first and feed them into a transformer module to predict the intent. In $\mathcal{S}$, we prepend the `[CLS]` embedding to the context and candidate embeddings, then feed them into a transformer module separately. After that we can obtain the representation vectors of context and candidate, and compute relevance scores by conducting dot-product of them.

8 Experiments

Experiments are conducted on MMDialog to assess two baselines. We perform response/intent predictions for **all turns except the first turn** of each dialogue and consider all previous turns as context.

8.1 Experimental Setup

We first sample 10K and 10K dialogue sessions for validation and testing respectively. The statistics are shown in Table 3. For detailed experimental and computational setup, please refer to Appendix A.3.

8.2 Results of Multi-modal Baselines

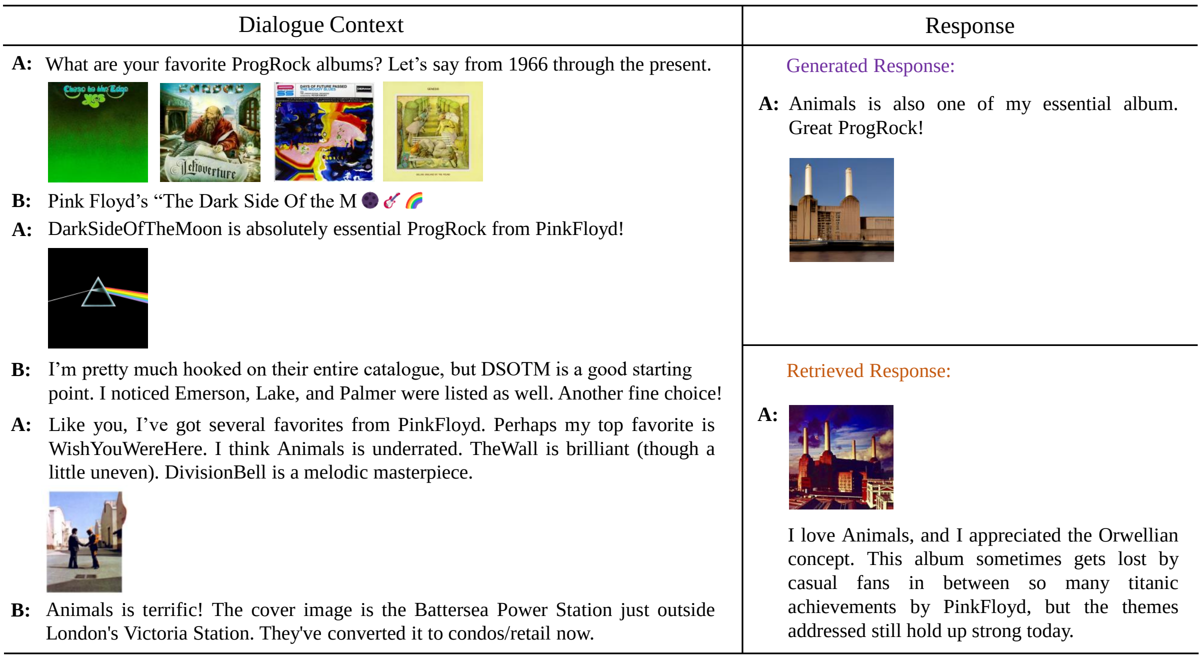
Table 4 reports the evaluation results of the generative baseline. Firstly, Divter achieves relatively low textual response generation performance (9.44 on BLEU-1 and 11.19 on ROUGE-L) on MMDialog. This emphasizes the complexity of the multi-modal response generation task and the need for constructing a large-scale multi-modal dialogue dataset to build data-driven models. Secondly, the model shows better performance in image generation, which may be attributed to the use of pre-

Models	Intent	Image Generation		Text Generation			Multi-Modal Generation
	F1	IS↑	CLIP-Sim↑	BLEU-1	BLEU-2	ROUGE-L	MM-Relevance↑
Divter (Sun et al., 2022)	71.77	20.53 ± 0.50	26.07	9.44	7.45	11.19	61.85

Table 4: Automatic evaluation results of the generative baseline on the test set of MMDialog.

Models	Intent	Image Retrieval			Text Retrieval			Multi-Modal Retrieval
	F1	R@1	R@5	R@10	R@1	R@5	R@10	MM-Relevance↑
DE++ (Zang et al., 2021)	82.69	18.23	26.99	31.73	23.07	39.21	47.05	68.91

Table 5: Automatic evaluation results of the retrieval baseline on the test set of MMDialog.



Dialogue Context

A: What are your favorite ProgRock albums? Let's say from 1966 through the present.

B: Pink Floyd's "The Dark Side Of the M 🌑🎸🌈

A: DarkSideOfTheMoon is absolutely essential ProgRock from PinkFloyd!

B: I'm pretty much hooked on their entire catalogue, but DSOTM is a good starting point. I noticed Emerson, Lake, and Palmer were listed as well. Another fine choice!

A: Like you, I've got several favorites from PinkFloyd. Perhaps my top favorite is WishYouWereHere. I think Animals is underrated. TheWall is brilliant (though a little uneven). DivisionBell is a melodic masterpiece.

B: Animals is terrific! The cover image is the Battersea Power Station just outside London's Victoria Station. They've converted it to condos/retail now.

Response

Generated Response:

A: Animals is also one of my essential album. Great ProgRock!

Retrieved Response:

A: I love Animals, and I appreciated the Orwellian concept. This album sometimes gets lost by casual fans in between so many titanic achievements by PinkFloyd, but the themes addressed still hold up strong today.

Figure 3: An example of MMDialog test set. **Left**: the multi-modal dialogue context between "A" and "B". **Right**: the multi-modal responses generated or retrieved by our designed baselines.

Metrics	Generation			Retrieval		
	Pearson (p)	Spearman (p)	Kendall (p)	Pearson (p)	Spearman (p)	Kendall (p)
BLEU-1	0.3609 (< 0.01)	0.3419 (< 0.01)	0.2471 (< 0.01)	0.2138 (< 0.01)	0.2648 (< 0.01)	0.2047 (< 0.01)
BLEU-2	0.3323 (< 0.01)	0.3711 (< 0.01)	0.2706 (< 0.01)	0.1660 (0.019)	0.2611 (< 0.01)	0.2026 (< 0.01)
ROUGE-L	0.3533 (< 0.01)	0.2579 (< 0.01)	0.2031 (< 0.01)	0.2540 (< 0.01)	0.2044 (< 0.01)	0.1745 (< 0.01)
CLIP-Sim	0.2723 (< 0.01)	0.1847 (< 0.01)	0.1469 (< 0.01)	0.2664 (< 0.01)	0.1743 (0.014)	0.1538 (0.011)
MM-Relevance	**0.4043** (< 0.01)	**0.3814** (< 0.01)	**0.2818** (< 0.01)	**0.2863** (< 0.01)	**0.3408** (< 0.01)	**0.2635** (< 0.01)

Table 6: Correlation between the metrics and human annotation. p denotes p-value.

trained DALL-E. Thirdly, Divter achieves a 71.77 F1 score in modality intent prediction, indicating its ability to determine whether to generate text or an image during the conversation. Finally, we leverage the proposed MM-Relevance metric to evaluate the overall relevance degree between generated multi-modal responses and ground-truth responses, and Divter achieves a score of 61.85.

Table 5 shows the results of the retrieval baseline. DE++ achieves 18.23% R@1 and 23.07% R@1 on image and text retrieval respectively, demonstrating the capacity of multi-modal retrieval model. Furthermore, we can also find that DE++ obtains a better intent F1 and MM-Relevance than Divter, which may be attributed to a limited retrieval space of test set (1K), as we observe that above metrics would decrease when the retrieval space increases. And we found out that the alignment of the modality would considerably improve the CLIP scores, which benefits the MM-Relevance.

8.3 Case Study

To further investigate the quality of multi-modal responses predicted by our proposed baselines, we display an example on the MMDialog test set in Figure 3. As demonstrated, Divter produces a textual response coherent with the dialogue context, and also can generate a realistic high-resolution image about the "Power Station" in the last turn of context, exhibiting the multi-modal generative capability of our proposed generative baseline. For retrieval, our baseline also retrieved a textual response about "PinkFloyd" and an image on "Power Station" semantically related to the context, validating the effectiveness of retrieval baseline.

8.4 Metrics Correlation With Human

Table 6 illustrates the Pearson, Spearman, and Kendall Correlation between various metrics and human judgments. The Fleiss' kappa (Fleiss, 1971) of the generative and retrieval labeling is 0.52 and 0.47 respectively, indicating a relatively high agreement among the three labelers. MM-Relevance outperforms other baselines across both generative and retrieval scenarios, indicating a stronger correlation with human annotation compared to other metrics. Besides, the metrics in generative task exhibit a higher correlation than those in retrieval task. This discrepancy may be attributed to the lower relevance between retrieved responses and ground-truth responses, resulting in a relatively low agreement and degrading the correlation.

9 Conclusion

We presented MMDialog, a large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. By extracting turns associated with images and their surrounding contexts from more than 4K topics, MMDialog provides a diverse and open-domain dataset. To facilitate research on building a more engaging multi-modal dialogue system, we define multi-modal response generation and retrieval tasks, and the MM-Relevance metric based on MMDialog. We also build baseline models and conduct comprehensive performance analyses. We believe MMDialog can serve as a rich resource to propel research in the multi-modal conversation, for years to help the community propose better methods suited to more scenarios.

Limitations

Besides its merits, this work still has limitations that could be further explored. On the one hand, we collect source data of MMDialog with only English language. Thus the applicability to other languages would be restricted. On the other hand, we get rid of GIFs and video-modality elements in MMDialog. In the future, we hope to extend MMDialog to multilingual scenarios and include more modalities such as audio and video.

Ethics Statement

This paper presents a large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation MMDialog. Data in MMDialog are collected from a public worldwide social media platform on which users can converse with each other and share their daily lives messages freely in multiple modalities, including plain text, photos, or even videos. The data-collecting API is only available for academic purposes. And to protect the privacy and security of data, users and platform, MMDialog is just released under strict terms for academic people only. This action fully aligns with the data using and sharing regulations of source data providers. Thus, there will not be any ethical problems or negative social consequences from the research. The proposed method does not introduce ethical/social bias in the data.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments and suggestions. This work was supported by the National Key R&D Program of China under Grant No. 2020AAA0106600.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

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.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Jiazhan Feng, Chongyang Tao, Zhen Li, Chang Liu, Tao Shen, and Dongyan Zhao. 2022. Reciprocal learning of knowledge retriever and response ranker for knowledge-grounded conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 389–399, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Mauajama Firdaus, Nidhi Thakur, and Asif Ekbal. 2020. MultiDM-GCN: Aspect-guided response generation in multi-domain multi-modal dialogue system using graph convolutional network. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2318–2328, Online. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Zhe Gan, Yu Cheng, Ahmed El Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. *arXiv preprint arXiv:1902.00579*.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1371–1374.

Jun Gao, Wei Bi, Ruifeng Xu, and Shuming Shi. 2021. REAM♯: An enhancement approach to reference-based evaluation metrics for open-domain dialog generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2487–2500, Online. Association for Computational Linguistics.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022. DEAM: Dialogue coherence evaluation using AMR-based semantic manipulations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 771–785, Dublin, Ireland. Association for Computational Linguistics.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895.

Janghoon Han, Taesuk Hong, Byoungjae Kim, Youngjoong Ko, and Jungyun Seo. 2021. Fine-grained post-training for improving retrieval-based dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1549–1558, Online. Association for Computational Linguistics.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Nyoungwoo Lee, Suwon Shin, Jaegul Choo, Ho-Jin Choi, and Sung-Hyon Myaeng. 2021. Constructing multi-modal dialogue dataset by replacing text with semantically relevant images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 897–906, Online. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017a. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169, Copenhagen, Denmark. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017b. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zujie Liang, Huang Hu, Can Xu, Chongyang Tao, Xiubo Geng, Yining Chen, Fan Liang, and Daxin Jiang. 2021. Maria: A visual experience powered conversational agent. *arXiv preprint arXiv:2105.13073*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. 2019. Recursive visual attention in visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6679–6688.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *https://openai.com/blog/chatgpt/*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. 2021. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2855–2870, Online. Association for Computational Linguistics.

Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. 2020. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10860–10869.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5427–5436, Florence, Italy. Association for Computational Linguistics.

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.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of Twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, Los Angeles, California. Association for Computational Linguistics.

Amrita Saha, Mitesh Khapra, and Karthik Sankaranarayanan. 2018. Towards building large scale multimodal domain-aware conversation systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. Image-chat: Engaging grounded conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429, Online. Association for Computational Linguistics.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Haoyu Song, Yan Wang, Wei-Nan Zhang, Zhengyu Zhao, Ting Liu, and Xiaojiang Liu. 2020. Profile consistency identification for open-domain dialogue agents. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6651–6662, Online. Association for Computational Linguistics.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. 2022. Multimodal dialogue response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2854–2866, Dublin, Ireland. Association for Computational Linguistics.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, and et al. Yu Du. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Yi-Lin Tuan, Yun-Nung Chen, and Hung-yi Lee. 2019. DyKgChat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1855–1865, Hong Kong, China. Association for Computational Linguistics.

Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A dataset for research on short-text conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 935–945, Seattle, Washington, USA. Association for Computational Linguistics.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

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

Taesun Whang, Dongyub Lee, Dongsuk Oh, Chanhee Lee, Kijong Han, Dong-hun Lee, and Saebyeok Lee. 2021. Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14041–14049.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Ze Yang, Wei Wu, Huang Hu, Can Xu, Wei Wang, and Zhoujun Li. 2021. Open domain dialogue generation with latent images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14239–14247.

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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6142–6152, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018b. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yinhe Zheng, Guanyi Chen, Xin Liu, and Jian Sun. 2022. MMChat: Multi-modal chat dataset on social media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5778–5786, Marseille, France. European Language Resources Association.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018a. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018b. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

A Appendix

A.1 The Most Popular Hashtags

We also depict the most popular hashtags in Figure 4.

A.2 Response Modal Intent Prediction

In MMDialog, the textual utterances and visual images can be freely located anywhere in the response. Therefore, the generation or retrieval order of the modality of response elements is also of great importance for the multi-modal conversation. The intent prediction task aims to predict the order of different modalities in response $\tilde{R}$ given the dialogue context U. Therefore, the intent prediction can be formulated as a classification task:

$$\forall j \in [1, J], \mathcal{I}(U, \tilde{R}_{<j}) \in \{0, 1, 2\} \qquad (4)$$

where $\mathcal{I}(\cdot, \cdot)$ is the intent prediction model which takes the dialogue context U and previously generated/retrieved response elements $\tilde{R}_{<j}$ before j-th step as inputs and provides the modality of next element. Specifically, the model should predict 0 when r_j is a textual utterance, 1 when r_j is a visual image, and 2 which indicates that the response $\tilde{R}$ is completed and the model should stop generating/retrieving new elements.

A.3 Detailed Experimental Setup

For retrieval task, we randomly sample 999 negative textual utterances and 999 negative visual images from the same split set for each dialogue, maintaining the total number of candidate elements at 1K. While in training phase, the negative ones are in-batch sampled similar to Radford et al. (2021). To be consistent with the Divter of Sun et al. (2022), we fine-tune DialoGPT-medium (345M) (Zhang et al., 2020) as the textual dialogue response generator with Hugging Face *transformers*[1]. For the description-to-image translator, we implement DALL-E (Ramesh et al., 2021) using the code of "mega" version in https://github.com/borisdayma/dalle-mini. We fine-tune DALL-E for one epoch with initial learning rate 1e-8 and batch size of 64. We process all images into 256 × 256 RGB format for DALL-E. To obtain the description of images in MMDialog, we adopt OFA-huge (Wang et al., 2022) using the code https://github.com/OFA-Sys/OFA/tree/feature/add_transformers for image captioning. The version of CLIP model is "openai/clip-vit-base-patch32" [2] that has 151M model parameters. As for the retrieval baseline, the representation vectors for both modality obtained by CLIP are fixed during training. The transformers of DE++ consist of 4 Transformer layers with a hidden size of 512 and 8 heads. We train the retrieval model with an initial learning rate of 5e-7 and batch size of 512. For all baselines, early stopping on the validation set is adopted as a regularization strategy and the best model is selected based on the validation performance. The training of both tasks is conducted on 8 Nvidia Tesla A100 80G GPU cards for about 20 hours. The BLEU and ROUGE scores are computed by codes in https://github.com/Maluuba/nlg-eval, while the IS is obtained by https://github.com/toshas/torch-fidelity.

[1] https://github.com/huggingface/transformers

[2] https://huggingface.co/openai/clip-vit-base-patch32

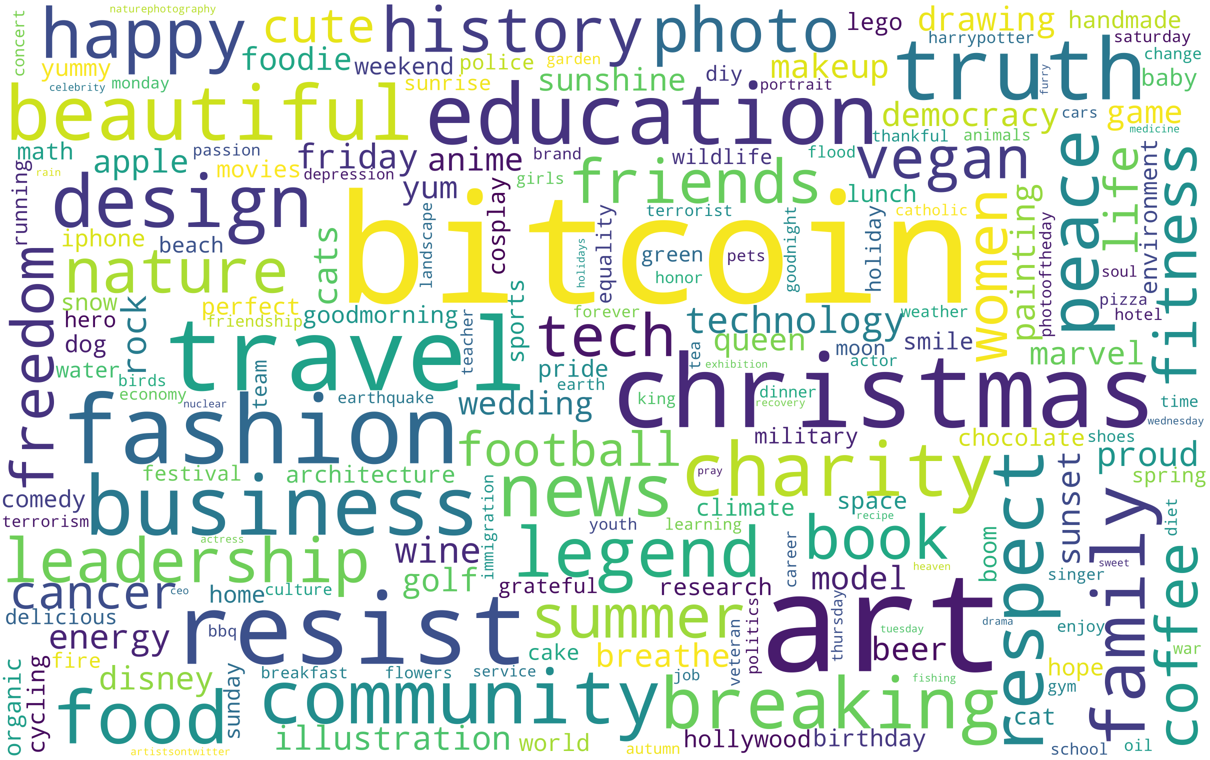


Figure 4: 200 most popular hashtags in MMDialog weighted by their frequencies.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section "Limitations"

☑ A2. Did you discuss any potential risks of your work?
Section "Ethics Statement"

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Section 8

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 8.1 and Appendix A.3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 6, Section 8.1 and Appendix A.3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4 and Section 8

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 8.1 and Appendix A.3

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 6.3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 6.3

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 6.3 and they are volunteers.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 3, Section 4, Section 6.3 and Ethics Statement

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section 3, Section 4, Section 6.3 and Ethics Statement

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 3, Section 4, Section 6.3 and Ethics Statement