Recent Advances on Multi-modal Dialogue Systems: A Survey

Fenghua Cheng[1][0009−0004−8075−7528], Xue Li[1][0000−0002−4515−6792], Haoyang Wu[2][0009−0007−7634−8758], Jiangcheng Sang[3][0009−0007−4092−1161], and Wenqi Zhao[4][0009−0004−4919−7075]

[1] The University of Queensland, Brisbane QLD 4101, AUS
[2] Xidian University, Xi'an Shaanxi 710071, CHN
[3] Sichuan University, Chengdu Sichuan 610065, CHN
[4] Chongqing University, Chongqing 400044, CHN
fenghua.cheng@uq.edu.au

Abstract. Empowering conversational agents to see the world and interact with humans using all their senses is one of the long-term goals of Artificial Intelligence (AI). Multi-modal interactions are crucial in real-world conversations, and compensation between different modalities helps improve the quality of the conversation. To this end, a growing research interest has been devoted to developing a multi-modal conversational agent with visual ability. Different from traditional unimodal dialogue systems, a multi-modal dialogue system can read context from multiple modalities and respond based on the understanding of them. In this work, we provide a comprehensive review of recent advances achieved in multi-modal dialogue generation. First of all, we categorize the multi-modal dialogue systems according to the tasks that they aim to address. Then, we review benchmark datasets as well as evaluation metrics. Finally, we discuss some existing challenges and promising directions for future work.

Keywords: Multi-modal Dialogue · Large Language Model · Vision Language.

1 Introduction

Recent years have witnessed the extraordinary success achieved in Large Language Models (LLMs) [1, 64, 75]. Benefiting from the development of LLMs, increasingly versatile and powerful conversational agents [2, 62] have been endowed with the remarkable capability of communicating with humans in pure language. However, most of them are still limited by incapacity of reading visual information due to unimodal large-scale training corpus. Text-only dialogue systems cannot interact with humans in a multi-modal way, which is more naturally and intuitively in human-human conversations. On the other hand, Large Vision Model (LVMs) [49, 71] can see clearly but cannot talk. Therefore, significant research efforts have been committed to designing a multi-modal dialogue system which can see and talk at the same time by incorporating visual modality.

An ideal multi-modal dialogue system should understand information from multiple modalities, capture relationships and generate textual or visual responses by bridging effective connections and alignment between different modalities.

Leading by GPT-4 [1], astonishing advances have been witnessed in multi-modal dialogue systems. Under such rapid development in this field, it is necessary to summarize recent progress [38, 65]. To this end, we present a comprehensive overview of recent contributions made in multi-modal dialogue systems and provide some possible future research directions. The following parts of the survey will be structured as follows: (1) Categories of existing multi-modal dialogues and corresponding tasks they aim to address; (2) Existing benchmark datasets used in multi-modal dialogue generation; (3) Popular evaluation methods and (4) A summary and potential future directions.

2 Categories of Multi-Modal Dialogue Systems

Starting from VQA [6] and image captioning [26], recent years have witnessed a surge in new fine-grained tasks in multi-modal dialogue research. Depending on different tasks that need to be addressed, multi-modal dialogue systems can be roughly categorized into the following: 1) image-grounded question-answering dialogue system, 2) image-grounded chit-chat dialogue system, 3) visual-evidence-embedded dialogue system, 4) image-response dialogue system, 5) video-grounded question-answering dialogue system, 6) video-grounded chit-chat dialogue system and 7) video-grounded real-time comment dialogue. We present a comparison between these tasks in Table 1.

Table 1. Comparison between Different Tasks in Multi-Modal Dialogue Systems

Task	Input Modalities	Output Modalities
Image-Grounded Question-Answering	Text & Image	Text
Image-Grounded Chit-Chat	Text & Image	Text
Visual-Evidence-Embedded	Text	Text
Image-Response	Text & Image	Text & Image
Video-Grounded Question-Answering	Video & Audio & Text	Text
Video-Grounded Chit-Chat	Video&Audio&Text	Text
Video-Grounded Commenting	Video & Audio & Text	Text

Although these different multi-modal dialogue systems address distinct tasks, they share the same ultimate goal: to generate responses given contextual information from multiple modalities, including visual information, audio information and basic text information. In this section, we elaborate on these specialized tasks and provide an introduction to pioneering works in these categories.

2.1 Image-Grounded Question-Answering Dialogue System

Image-grounded question-answering dialogue system aims to generate answers to given questions based on a related image. The agent needs to ground questions in the given visual content, take conversation history into consideration, and finally generate accurate answer responses. Difference from VQA [6] that only has single-turn question-answering, image-grounded question-answering agents are normally designed to conduct multi-turn QA conversations. This raises a challenge of co-reference resolution because subjects in follow-up questions are often specified by using pronouns.

[14] first presented this task and proposed the VisDial dataset specialized for it. In their work, they combined three different types of encoders: Late Fusion (LF) Encoder, Hierarchical Recurrent Encoder (HRE) and Memory Network (MN) Encoder, along with two kinds of decoders, Generative (LSTM) decoder and Discriminative (softmax) decoder to generate answers. The results show that the model with MN encoder, which takes question, image and dialogue history into account, outperforms other baselines.

[47] introduced a pre-trained LLM into the image-grounded question-answering dialogue system. They proposed a method that transfers a large language model, ViLBERT [39], to this task. Instead of focusing on a specific vision-language task, ViLBERT is proposed to learn basic visual knowledge, which can be pre-trained and transferred to any vision-language task. ViLBERT has two transformer-based [66] encoders: one for linguistic modality and one for visual modality. There are co-attention layers to jointly model inputs from different modalities. To transfer ViLBERT to visual dialogue, the author employed four phases of training. Firstly, the language stream was pre-trained on English Wikipedia and the BooksCorpus datasets. Then the entire ViLBERT was trained on Conceptual captions and VQA datasets. To tackle the problem that captions and question-answering pairs differ a lot in input format, the model is fine-tuned on VisDial. By being pre-trained on large-scale datasets and transferred to visual dialogue task, ViLBERT can provide reasonable and accurate answers to given question and visual content.

[48] presented a new neural architecture, the Light-weight Transformer for Many Inputs (LTMI), to handle the heterogeneity gap between different modalities in visual dialogue. LTMI computes the attention between different tokens, as the original transformer does. Unlike projecting input embeddings to lower-dimensional spaces in transformers, LTMI splits the input feature space into low-dimensional spaces according to indexes, which saves computation. In addition to less computation, LTMI can take inputs from multiple modalities to compute the attention between them. These two features make LTMI suitable for understanding both visual information and textual questions and generating responses.

2.2 Image-Grounded Chit-Chat Dialogue System

Different from image-grounded question-answering in which images and conversation history are used as input to generate answers to specified questions, the

image-grounded chit-chat dialogue system is designed to generate entertaining answers that are reasonable and empathetic. The conversational agent needs to maintain chit-chat conversations with humans based on visual contents.

[60] presented Image-Chat, a dataset specialized for image chit-chat, and explored the performance of both generative and retrieval-based models on this task. Both generative and retrieval-based models utilize the same encoders, including an image encoder that is ResNet [24] with 152 layers, a dialogue encoder based on transformer architecture, and a style encoder. For the generative model, there is another decoder transformer to generate responses, and for the retrieval-based model, the matching score between multi-modal inputs and candidate responses are calculated. Results demonstrate that the retrieval model outperforms the generative one, and human evaluation shows it is more popular compared to human conversationalists.

[61] explored the combination of two different image encoders ResNeXt [70] and Fast R-CNN [22], which are pre-trained on object detection tasks with a text encoder transformer that is pre-trained on massive pure-language dialogue datasets to generate responses. Two ways of fusion, late fusion and early fusion, were also explored. For the late fusion method, the encoding of image is projected to the same dimensional space as the text encoding from transformer and concatenated after text encoding. However, the early fusion method concatenates the image encoding with token embeddings and jointly encodes the text and image using a transformer, which means the transformer can calculate self-attention between textual and visual inputs.

2.3 Visual-Evidence-Embedded Dialogue System

Visual evidence is frequently involved in human-human conversations to create associations, but it is hard to document by pure-text. To this end, non-paired images, acting as visual evidence, are introduced in this kind of dialogue system to generate more related responses. Unlike other types of multi-modal dialogue systems, which either take input from multiple modalities or output responses in multiple modalities, a visual-evidence-embedded dialogue system takes text-only context as input and generates text-only responses as well. The multi-modality of visual-evidence-embedded dialogue system reflects in the fusion of implicit visual evidence generated or retrieved based on given context.

[59] proposed a two-stage method to integrate visual evidence into dialogue generation. In the first step, words that may induce visual evidence in the given context are extracted, and then the most related images are selected by a word-image mapping model as the visual evidence. At the second stage, the authors proposed VisAD, an encoder-decoder framework in which 1) text context and retrieved visual information are encoded by a co-attention encoder; 2) visual words in response are generated and response visual evidence generated based on visual words by two sub-decoders; and 3) final response is generated based on integration of context and response visual evidence.

Maria [35] used pipeline workflow to realize the integration of visual experience. Maria contains three main components. The first is a text-to-image

retriever, which maps dialogue context with a candidate image. The retriever calculates a related score between images and text encoding obtained by a pre-trained BERT text encoder and an image encoder, respectively. The training of the retriever is on image captioning dataset. Then the retrieved image will be fed into the second component, Visual Concept Detector, which is a pre-trained object detection model [4]. The detector can extract visual knowledge from the given image, including salient object features and visual concepts. Finally, the dialogue context and visual knowledge will be fed into the last component, visual-knowledge-grounded response generator, a multi-layer transformer, to generate responses.

2.4 Image-Response Dialogue System

Image-Response dialogue system is required to generate not only textual but also visual responses to the given dialogue context. In real-world scenarios, human-to-human conversations can highly involve visual responses like sharing photos or memes as well. Generating visual response when conversing with humans is advantageous to maintain a long conversation and attract human's interest in the conversation. To this end, some research raised image-response dialogue system. The biggest challenge in image-response generation is that visual concepts may not be explicitly mentioned in the dialogue context. On the contrary, the generated image should be mainly based on the completed understanding of the historical context.

To resolve this problem, [72] presented a human-to-human conversation dataset PhotoChat in which photos are sent to each other and utilized a retrieval-based method specialized for this task. The task is divided into two separate sub-tasks in their work: photo-sharing intent prediction to predict whether a photo should be shared in the next turn and image retrieval that retrieves images from a candidate image set. For photo-sharing intent prediction model, the authors explored three pre-trained large-scale models, BERT [15], ALBERT [29], and T5 [54] and for image retrieval, they developed dual encoder which encodes candidate image and context respectively and gives a final matching score between them. Finally, visual response is generated under the cooperation of these two components.

On the contrary, [63] resolved this problem by proposing an entire generation model called Divter. Divter has two transformer-based components: a text-to-image translator and a response generator. The response generator outputs a text response or a textual image description, and then the text-to-image translator can take the image description as input and generate a reasonable and corresponding visual response. Text-to-image translator and response generator are separate, which means they can be pre-trained independently on image captioning datasets and text-only dialogue datasets.

2.5 Video-Grounded Question-Answering Dialogue System

Similar to image-grounded question-answering, video-grounded question-answering dialogue system requires a conversational agent to generate an answer to a given

question and context in multiple modalities, including video and audio. The conversational agent needs to understand visual and audio information from the whole video, and textual information from QA context and question. Challenge arises from understanding of video content because of more complex features in video and heavy computation load in video understanding.

[31] utilized transformer architecture to overcome this challenge. The sequential visual information from multiple video frames is captured by a multi-head attention mechanism in which visual, audio, and textual information is co-attended repeatedly. Additionally, an auto-encoder is applied to extract features from non-text input. A decoder can integrate encodings from input, context history, video caption and question to generate a final response.

[3] presented a video-dialog answerer model and compared several baseline video-grounded question-answering dialogue systems. The answerer model is a retrieval-based model. Video input, audio input, dialogue history and question are encoded by 3D CNN [9], CNN, LSTM respectively. The final input representation is the concatenation of these encodings. Meanwhile, all candidate answers are encoded by LSTM. Finally, the mapping score is calculated by dot product. Results demonstrate that the full utilization of information from multiple modalities is necessary for generating a more complete, related and natural response.

Neural Modules Network [5] was first introduced to video-grounded dialogues by VGNMN [30]. By decomposing questions into a set of components that are resolved by NMN and extracting video features, VGNMN can focus on actions in the video which are highly related to decomposed components. Text encoder is shared for history context, question and video captions. Encoding of them is fed to question parser to decompose language components and then entity/action-level video context is obtained by NMN based on decomposed entities together with textual and video encodings. Finally, a response is generated by a vanilla transformer decoder.

2.6 Video-Grounded Chit-Chat

Similar to image-grounded chit-chat dialogue system, video-grounded chit-chat dialogue system is designed to produce chit-chat responses. However, it is required to understand the given video and respond based on the video contents.

CHAMPAGNE [23] is a generative model that produces real-world conversations based on a given video. The model architecture is based on Unified-IO [40] and in an encoder-decoder manner. Video positional embeddings are utilized to help specify multiple frames in a video. Experiments demonstrate strong ability to understand the video content and converse based on it.

Tiktalk [37] proposed a large-scale video-dialogue chit-chat dataset along with two methods to realize video-grounded chit-chat, one based on Maria and the other based on BLIP-2. Audio information, visual information, and external knowledge are all inputs. The dataset was extracted from TikTok, containing 38K videos and 367K corresponding dialogues.

SportsVD [12] presented a novel idea of creating an event-content-oriented video-dialogue dataset. Rather than simple entity movement in previous work, SportsVD focuses on complicated sports-domain event understanding. The authors used BERT-based method and unified input representation to generate response.

2.7 Video-Grounded Real-Time Comments System

With the popularity of danmaku on major video websites, such as YouTube and Bilibili, video-grounded real-time commenting task has arisen at the right moment. It aims to generate reasonable live comments as opinions of a specific video clip near a timestamp or response to chit-chat of other live comments. Given a video, a timestamp, its corresponding frame, and prior comments before this timestamp, a video-grounded real-time comments system should generate a comment that is relevant to the clip of the video near the timestamp and other comment context. The biggest challenge in video-grounded live commenting is the complicated relationship between video clip and other comments. The generated comments may be about the video itself or about comments prior to it.

LiveBot [42] introduces two models for this task. The first model is fusional RNN model, which is a typical encoder-decoder framework. It consists of a video encoder, a text encoder, and a comment decoder, all based on LSTM. Video is encoded first by the video encoder and then attended to the text encoding. With the help of attention, the text representation contains information about video. The decoder generates a response based on the fusional encoding of video and text. The second model is a unified transformer model which is similar to fusional RNN model with video encoder, text encoder and comment decoder. However, different from fusional RNN method, unified transformer model uses transformers for encoders and decoder.

MML-CG [69], in short for multi-modal multitask learning based comments generation framework, resolves the challenge by introducing multitask learning. Similar to LiveBot, MML-CG uses a text encoder to encode other context and uses a video encoder to obtain video representations, where video encoder is under transformer architecture and text encoder is based on LSTM. There is another transformer-based multi-modal encoder to integrate video encoding and text encoding. To improve the fusion between text and video clips, MML-CG introduces a new learning task, temporal relation prediction. The newly introduced task can help model the dependencies between comments and video clips. Results show the introduction of temporal relation prediction task results in an improvement on comments generation.

3 Datasets

In this section, we summarize popular choices of datasets used in multi-modal dialogue system. Due to the training of multi-modal, the conversational agent

undergoes two phases: pre-training and fine-tuning for specialized tasks . We introduce datasets used in two phases separately.

3.1 Pre-Training Data

Since the main purpose of the pre-training phase is to align between different modalities, most pre-training datasets are in image-text pairs, e.g., image captioning datasets. We list common pre-training datasets used in pre-training phase below.

- **Conceptual Caption:** CC3M [58] and its following work CC12M [10] is a large-scale image caption datasets consisting of 3.3M, 12.4M image-caption pairs respectively. All data was from web and a well-designed pipeline was used to clean data.
- **LAION:** LAION family is a series of large-scale image-text datasets from web. A complicated cleaning process was conducted as well. LAION family includes LAION-400M [57], LAION-5B [56] and LAION-COCO [13].
- **COYO-700M:** COYO-700M [8] is a large-scale dataset consists of 747M image-text pairs. It also provides meta-attributes to make it easier to provide help for various tasks.
- **SBU Captioned Photo:** SBU Captioned Photo Dataset [50] is an image-caption dataset in which images were from the web and associated captions were written by people. It contains 1M image-caption pairs.
- **WebLI:** WebLI [11] consists of 10B images and tens of billions of image-text pairs. The dataset is multilingual, covering over 100 languages.
- **DataComp:** DataComp [20] is an experimental testbed with 12.8B image-text pairs from Common Crawl, currently the largest public image-text dataset.

3.2 Fine-Tuning Data

Unlike the use of image-text pairs in pre-training datasets, fine-tuning datasets emphasize on dialogues that can be single-turn or multi-turn and two-party or multi-party. There has been a surge in pure-text dialogue datasets [25, 21] and attempts to extend them to multi-modal dialogue datasets. Dialogues in these multi-modal fine-tuning datasets are tightly connected to the visual information. We present fine-tuning datasets specially designed for different categories of multi-modal dialogue systems in Table 2.

4 Evaluation

Evaluation in multi-modal dialogue system research is fundamental to provide a systematic method to test how well a model performs and to make comparison. However, compared with traditional NLP generative tasks like machine translation or question-answering, the evaluation of multi-modal dialogue systems faces new challenges.

Table 2. Benchmark Fine-Tuning Datasets Used in Multi-Modal Dialogue Systems. We use number 1 to 7 to denote corresponding sub-tasks mentioned in section 2.

Name	Task	#Dialogues	#Images	#Videos	Language
VisDial [14]	1	1.2M	120K	-	English
IGC [46]	1	250K	250K	-	English
PlotQA [45]	1	28.9M	224.3K	-	English
ChartQA [43]	1	21.9K	32.7K	-	English
SciGraphQA [34]	1	657K	295K	-	English
MMChat [76]	2	120K	204K	-	Chinese
MMDialog [16]	2	1.0M	1.5M	-	English
Image-Chat [60]	2	202K	202K	-	English
OpenViDial [44]	2	1.1M	1.1M	-	English
OpenViDial2.0 [68]	2	5.6M	5.6M	-	English
SIMMC 2.0 [28]	2	11K	1.5K	-	English
Maria [35]	3	1.4M	50K	-	English
PhotoChat [72]	4	12K	12K	-	English
MMD [55]	4	45K	45K	-	English
MDMMD [17]	4	131K	1.5M	-	English
MDMMD++ [18]	4	203K	2.0M	-	English
DialogCC [32]	2&4	83K	787K	-	English
AVSD [3]	5	18K	-	18K	English
M^3ED [74]	5	2.4K	-	2.4K	Chinese
VideoIC [69]	6	5.3M (comments)	-	4.9K	Chinese
LiveBot [42]	6	896K (comments)	-	2.3K	Chinese
YTD-18M [23]	7	18M	-	18M	English
MELD [53]	7	1.4K	-	1.4K	English
TikTalk [37]	7	367K	-	38K	Chinese

The first challenge lies on the trait of open-domain dialogue: the response maybe diverse, flexible yet reasonable given the context. The second challenge is measuring how the generated response relates to input from non-linguistic modalities. We can roughly categorize the evaluation metrics into: automatic metric and human metric. In this section, we list several popular evaluation methods employed in multi-modal dialogue systems.

4.1 Automatic Metric

We summarize some dominant automatic metrics along with their advantages and drawbacks when applied into multi-modal dialogue systems in Table 3.

4.2 Human Metric

Since determining whether the response content is highly relevant to non-linguistic modalities could be tricky and tough, many works in multi-modal dialogue

Table 3. Automatic Metrics Used in Multi-Modal Dialogue System

Metric	Advantages	Disadvantages
BLEU [51]	Simple; Explainable; Language-independent	×Meaning of word; ×Importance of word; ×Order of words
Perplexity [27]	Simple; Robust	×Accurate expression; Dataset dependency
Meteor [7]	Semantic similarity; Order matters	Complex computation
Rouge-L [36]	Simple; Explainable	×Meaning of word
CIDEr [67]	Multi-modality; Confidence considering	×Meaning of word
Dist-n [33]	Diversity evaluation	×Meaning of word; ×Importance of word; ×Order of words
MRR	Accurate; Explainable	Retrieval-based (inflexible)
Recall-K	Accurate; Explainable	Retrieval-based (inflexible)
Accuracy	Accurate; Explainable	Limited to specific task
BERTScore [73]	More accurate to text similarity; Comprehensiveness; Reference-free	Bias to BERT; ×Grammar
GPTScore [19]	More accurate to text similarity; Comprehensiveness	Weak generalization; ×Grammar

systems introduced human metrics in evaluation methods. The manual evaluation could be categorized into different aspects for various sub-tasks including logic, fluency, relevance to context, relevance to non-linguistic modalities, and relevance to external knowledge. Although human assessment could make intuitive and directive judgements on generated responses, there are still two challenges. Human evaluation could be labor-intensive, limiting the testing size, which means bias may exist. On the other hand, bias may also exist between different annotators. Hence, bias is the biggest problem in human evaluation.

5 Conclusion and Future Direction

In this work, we present a comprehensive review of existing multi-modal dialogue system designs. We conclude categories of multi-modal dialogue systems based on different tasks and recent advances in multi-modal dialogue generation methods within the corresponding categories. Moreover, we summarize some benchmark datasets and popular evaluation methods used in multi-modal dialogue systems. Despite all the academic attention and efforts devoted to multi-modal dialogue systems, there are still open challenges to be addressed. We discuss several possible future research directions for this field.

– **Modality Gaps:** Although recent works explore methods to bridge gap between different modalities, reasoning between different modalities is still

weak, as [41] reported. Research on how to reduce gaps and reason more effectively can result in a significant improvement in the multi-modal dialogue field.

- **Miss of Comprehensive Evaluation Metric:** As we mentioned before, few automatic metrics can truly understand the relationship between dialogue and other modalities and take it into consideration. On the contrary, human evaluation can relate between multiple modalities, but the bias is an inevitable problem. Hence, a more comprehensive evaluation method is an urgent need.
- **Long Non-Linguistic Context:** Existing multi-modal dialogue system can hardly handle long contexts in non-linguistic modalities, e.g., a long video. How to read long contexts is still an open task.
- **Harmful Contents:** Due to the large-scale pre-training corpus from the web, a multi-modal conversational agent could generate inappropriate and harmful responses [52]. Further efforts are still necessary to prevent harmful content.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Adiwardana, D., Luong, M.T., So, D.R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., et al.: Towards a human-like open-domain chatbot. arXiv preprint arXiv:2001.09977 (2020)
3. Alamri, H., Cartillier, V., Das, A., Wang, J., Cherian, A., Essa, I., Batra, D., Marks, T.K., Hori, C., Anderson, P., et al.: Audio visual scene-aware dialog. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7558–7567 (2019)
4. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
5. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Neural module networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 39–48 (2016)
6. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
7. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
8. Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset (2022)
9. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)

10. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: 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. pp. 3558–3568 (2021)
11. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al.: Pali: A jointly-scaled multilingual language-image model. In: The Eleventh International Conference on Learning Representations (2022)
12. Cheng, F., Li, X., Huang, Z., Wang, J., Wang, S.: Event-content-oriented dialogue generation in short video. In: 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (2024)
13. Christoph Schuhmann, Andreas Köpf, R.V.T.C.R.B.: Laion coco: 600m synthetic captions from laion2b-en (2022)
14. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D.: Visual dialog. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 326–335 (2017)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
16. Feng, J., Sun, Q., Xu, C., Zhao, P., Yang, Y., Tao, C., Zhao, D., Lin, Q.: Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. arXiv preprint arXiv:2211.05719 (2022)
17. Firdaus, M., Thakur, N., Ekbal, A.: 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. pp. 2318–2328 (2020)
18. Firdaus, M., Thakur, N., Ekbal, A.: Aspect-aware response generation for multimodal dialogue system. ACM Transactions on Intelligent Systems and Technology (TIST) **12**(2), 1–33 (2021)
19. Fu, J., Ng, S.K., Jiang, Z., Liu, P.: Gptscore: Evaluate as you desire. arXiv preprint arXiv:2302.04166 (2023)
20. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next generation of multimodal datasets. Advances in Neural Information Processing Systems **36** (2024)
21. Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al.: The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027 (2020)
22. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
23. Han, S., Hessel, J., Dziri, N., Choi, Y., Yu, Y.: Champagne: Learning real-world conversation from large-scale web videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15498–15509 (2023)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
25. Henderson, M., Budzianowski, P., Casanueva, I., Coope, S., Gerz, D., Kumar, G., Mrkšić, N., Spithourakis, G., Su, P.H., Vulić, I., et al.: A repository of conversational datasets. arXiv preprint arXiv:1904.06472 (2019)

26. Hossain, M.Z., Sohel, F., Shiratuddin, M.F., Laga, H.: A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR) **51**(6), 1–36 (2019)
27. Jelinek, F., Mercer, R.L., Bahl, L.R., Baker, J.K.: Perplexity—a measure of the difficulty of speech recognition tasks. The Journal of the Acoustical Society of America **62**(S1), S63–S63 (1977)
28. Kottur, S., Moon, S., Geramifard, A., Damavandi, B.: Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations. arXiv preprint arXiv:2104.08667 (2021)
29. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. In: International Conference on Learning Representations (2019)
30. Le, H., Chen, N.F., Hoi, S.C.: Vgnmn: Video-grounded neural module network to video-grounded language tasks. arXiv preprint arXiv:2104.07921 (2021)
31. Le, H., Sahoo, D., Chen, N., Hoi, S.: Multimodal transformer networks for end-to-end video-grounded dialogue systems. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5612–5623 (2019)
32. Lee, Y.J., Ko, B., Kim, H.G., Choi, H.J.: Dialogcc: Large-scale multi-modal dialogue dataset. arXiv preprint arXiv:2212.04119 (2022)
33. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, W.B.: A diversity-promoting objective function for neural conversation models. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 110–119 (2016)
34. Li, S., Tajbakhsh, N.: Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. arXiv preprint arXiv:2308.03349 (2023)
35. Liang, Z., Hu, H., Xu, C., Tao, C., Geng, X., Chen, Y., Liang, F., Jiang, D.: Maria: A visual experience powered conversational agent. arXiv preprint arXiv:2105.13073 (2021)
36. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
37. Lin, H., Ruan, L., Xia, W., Liu, P., Wen, J., Xu, Y., Hu, D., Song, R., Zhao, W.X., Jin, Q., et al.: Tiktalk: A video-based dialogue dataset for multi-modal chitchat in real world. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 1303–1313 (2023)
38. Liu, G., Wang, S., Yu, J., Yin, J.: A survey on multimodal dialogue systems: recent advances and new frontiers. In: 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE). pp. 845–853. IEEE (2022)
39. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019)
40. Lu, J., Clark, C., Zellers, R., Mottaghi, R., Kembhavi, A.: Unified-io: A unified model for vision, language, and multi-modal tasks. In: The Eleventh International Conference on Learning Representations (2022)
41. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems **35**, 2507–2521 (2022)
42. Ma, S., Cui, L., Dai, D., Wei, F., Sun, X.: Livebot: Generating live video comments based on visual and textual contexts. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 6810–6817 (2019)

43. Masry, A., Do, X.L., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 2263–2279 (2022)
44. Meng, Y., Wang, S., Han, Q., Sun, X., Wu, F., Yan, R., Li, J.: Openvidial: A large-scale, open-domain dialogue dataset with visual contexts. arXiv preprint arXiv:2012.15015 (2020)
45. Methani, N., Ganguly, P., Khapra, M.M., Kumar, P.: Plotqa: Reasoning over scientific plots. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1527–1536 (2020)
46. Mostafazadeh, N., Brockett, C., Dolan, B., Galley, M., Gao, J., Spithourakis, G.P., Vanderwende, L.: Image-grounded conversations: Multimodal context for natural question and response generation. arXiv preprint arXiv:1701.08251 (2017)
47. Murahari, V., Batra, D., Parikh, D., Das, A.: Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In: European Conference on Computer Vision. pp. 336–352. Springer (2020)
48. Nguyen, V.Q., Suganuma, M., Okatani, T.: Efficient attention mechanism for visual dialog that can handle all the interactions between multiple inputs. arXiv preprint arXiv:1911.11390 (2019)
49. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2023)
50. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. Advances in neural information processing systems **24** (2011)
51. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
52. Pi, R., Han, T., Xie, Y., Pan, R., Lian, Q., Dong, H., Zhang, J., Zhang, T.: Mllm-protector: Ensuring mllm's safety without hurting performance. arXiv preprint arXiv:2401.02906 (2024)
53. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508 (2018)
54. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)
55. Saha, A., Khapra, M., Sankaranarayanan, K.: Towards building large scale multimodal domain-aware conversation systems. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
56. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
57. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
58. Sharma, P., Ding, N., Goodman, S., Soricut, R.: 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). pp. 2556–2565 (2018)

59. Shen, L., Zhan, H., Shen, X., Song, Y., Zhao, X.: Text is not enough: Integrating visual impressions into open-domain dialogue generation. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 4287–4296 (2021)
60. Shuster, K., Humeau, S., Bordes, A., Weston, J.: Image-chat: Engaging grounded conversations. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 2414–2429 (2020)
61. Shuster, K., Smith, E.M., Ju, D., Weston, J.: Multi-modal open-domain dialogue. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 4863–4883 (2021)
62. Shuster, K., Xu, J., Komeili, M., Ju, D., Smith, E.M., Roller, S., Ung, M., Chen, M., Arora, K., Lane, J., et al.: Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. arXiv preprint arXiv:2208.03188 (2022)
63. Sun, Q., Wang, Y., Xu, C., Zheng, K., Yang, Y., Hu, H., Xu, F., Zhang, J., Geng, X., Jiang, D.: Multimodal dialogue response generation. arXiv preprint arXiv:2110.08515 (2021)
64. Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., Wang, H.: Ernie 2.0: A continual pre-training framework for language understanding. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 8968–8975 (2020)
65. Sundar, A., Heck, L.: Multimodal conversational ai: A survey of datasets and approaches. In: Proceedings of the 4th Workshop on NLP for Conversational AI. pp. 131–147 (2022)
66. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
67. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
68. Wang, S., Meng, Y., Li, X., Sun, X., Ouyang, R., Li, J.: Openvidial 2.0: A larger-scale, open-domain dialogue generation dataset with visual contexts. arXiv preprint arXiv:2109.12761 (2021)
69. Wang, W., Chen, J., Jin, Q.: Videoic: A video interactive comments dataset and multimodal multitask learning for comments generation. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2599–2607 (2020)
70. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
71. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021)
72. Zang, X., Liu, L., Wang, M., Song, Y., Zhang, H., Chen, J.: Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. arXiv preprint arXiv:2108.01453 (2021)
73. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2019)
74. Zhao, J., Zhang, T., Hu, J., Liu, Y., Jin, Q., Wang, X., Li, H.: M3ed: Multi-modal multi-scene multi-label emotional dialogue database. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5699–5710 (2022)

75. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 (2023)
76. Zheng, Y., Chen, G., Liu, X., Sun, J.: Mmchat: Multi-modal chat dataset on social media. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 5778–5786 (2022)