# MODDP: A Multi-modal Open-domain Chinese Dataset for Dialogue Discourse Parsing

**Chen Gong***, **Dexin Kong***, **Suxian Zhao**, **Xingyu Li**, **Guohong Fu**†
Institute of Artificial Intelligence, School of Computer Science and Technology,
Soochow University, Suzhou, China
{gongchen18, ghfu}@suda.edu.cn;
{kongdx.nlp, zsx2000223, xingyuli7007}@gmail.com

## Abstract

Dialogue discourse parsing (DDP) aims to capture the relations between utterances in the dialogue. In everyday real-world scenarios, dialogues are typically multi-modal and cover open-domain topics. However, most existing widely used benchmark datasets for DDP contain only textual modality and are domain-specific. This makes it challenging to accurately and comprehensively understand the dialogue without multi-modal clues, and prevents them from capturing the discourse structures of the more prevalent daily conversations. This paper proposes MODDP, the first multi-modal Chinese discourse parsing dataset derived from open-domain daily dialogues, consisting 864 dialogues and 18,114 utterances, accompanied by 12.7 hours of video clips. We present a simple yet effective benchmark approach for multi-modal DDP. Through extensive experiments, we present several benchmark results based on MODDP. The significant improvement in performance from introducing multi-modalities into the original textual unimodal DDP model demonstrates the necessity of integrating multi-modalities into DDP.

## 1 Introduction

Dialogue Discourse Parsing (DDP) aims to identify the discourse relations between utterances in a dialogue using a dependency tree. The left part of Figure 1 shows an example, where the arcs represent the dependencies between utterances and the labels on arcs are discourse relations. As a fundamental task in natural language processing (NLP), DDP can contribute to a deeper understanding of the structure and semantics inherent in dialogues, and has been proven beneficial for various downstream tasks, including emotion recognition (Zhang et al., 2023), dialogue response generation(Jia et al., 2020), and meeting summarization (Feng et al., 2021).

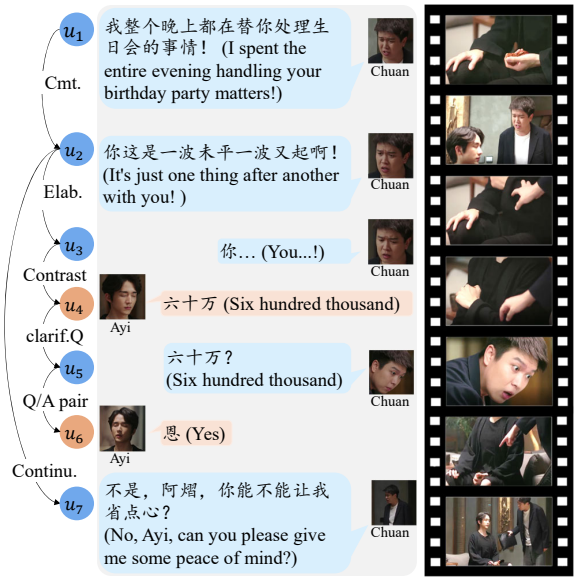


Figure 1: An example of discourse structures on multi-modal daily dialogues.

To support DDP research, high-quality labeled data is indispensable. To date, STAC (Nicholas Asher and Afantenos, 2016) and Molweni (Li et al., 2020b) are two existing publicly available benchmark datasets that widely used in prior research (Liu and Chen, 2021; Yu et al., 2022; Wang et al., 2023), promoting progress in the field of DDP. Both of the two datasets are collected from dialogues in English-language online forums, where STAC originates from the chat dialogues of an online game, and Molweni is sourced from an online forum about Ubuntu. Although the dialogue scenarios for the two sources actually contain multi-modalities beyond text, such as game graphics for the STAC source and screenshots of the Ubuntu interface for the Molweni source, these datasets are focused solely on textual modality, excluding the existing useful related multi-modal information. Consequently, based on these two datasets, most of

*Equal Contribution
† Corresponding author



*Findings of the Association for Computational Linguistics: ACL 2024*, pages 10561–10573

the DDP works only rely on the textual modality to parse discourse structures (Shi and Huang, 2019; Chi and Rudnicky, 2022; Li et al., 2023c).

However, despite significant progress in existing DDP research, the neglect of valuable multi-modal information for DDP still presents two challenges. First, conducting DDP within a single textual modality does not align with the real-world dialogue scenarios, since communication in real life often involves multiple modalities beyond text, such as visual and audio modalities. Second, solely relying on the textual modality for DDP may limit the ability to achieve a comprehensive and accurate understanding of the dialogue. Taking Figure 1 as an example, from the text only, it is difficult to understand the dialogue "你...！"，"六十万". When referring to the angry tone of the person named Chuan from audio modality, and his action of snatching the walnut from Ayi's hand according to visual modality, it becomes clear that Chuan is accusing Ayi, while Ayi warns Chuan about the walnut's value of six hundred thousand yuan. This illustrates that audio and visual modalities can provide essential supplementary cues for DDP that text alone cannot convey.

Zhao et al. (2022b) pioneered the introduction of multi-modalities to DDP, namely JDDC2.1, consisting of dialogues from a mainstream Chinese E-commerce platform. However, JDDC2.1 still mainly focuses on the textual modality, since each dialogue contains an average of only two images, and thus more helpful multi-modal information, such as tones in audio and actions in video, cannot be fully exploited for DDP.

Moreover, all the above existing datasets originate from task-specific domains, thereby limiting their abilities to reflect the dialogue discourse structures of the more natural and widespread open-domain real-world scenarios.

To address these limitations, this work proposes MODDP, the first **M**ulti-modal **O**pen-domain Chinese datasets for **D**ialogue **D**iscourse **P**arsing to the best of our knowledge, consisting of 864 two-party dialogues, and 18,114 utterances, with parallel video clips of 12.68 hours. Overall, MODDP has the following important features. First, it is sourced from open-domain dialogues across various TV series, containing textual, audio, and visual modalities simultaneously. This capability allows it to more effectively reflect discourse structures in multi-modal daily dialogues, thereby facilitating research in multi-modal DDP to better support practical applications in daily life. Second, we adopt the annotation scheme of SDRT (Segmented Discourse Representation Theory) (Lascarides and Asher, 2009) following STAC and Molweni, capturing the whole discourse structure of a dialogue as a dependency tree, with 16 labels to distinguish discourse relations. We also compile comprehensive annotation guidelines for annotators' reference. Third, to ensure data quality, we perform strict double annotation workflow. Each dialogue is assigned to two annotators to independently annotate discourse structures, with a third expert annotator to handle inconsistency annotations.

To provide benchmark results on our newly annotated MODDP, we propose a simple yet effective benchmark approach for multi-modal DDP. We conduct extensive experiments and provide several benchmark results based on MODDP, under both settings of using textual modality and using multi-modalities. Experimental results and further analysis show that the DDP performance increase by large margin after introducing multi-modalities, highlighting the substantial benefits of integrating multi-modalities for DDP.

We will release MODDP datasets, along with our compiled annotation guidelines and codes for research usage at https://github.com/Suda-iaiNLP/MODDP.

## 2 Related Work

**DDP Datasets.** To date, there are four representative datasets with annotations of dialogue discourse structures, i.e., English STAC (Nicholas Asher and Afantenos, 2016), Molweni (Li et al., 2020b), and GUM (Zeldes, 2017) with textual modality only, and Chinese JDDC 2.1 (Zhao et al., 2022b) with both text and visual modalities.

STAC (Nicholas Asher and Afantenos, 2016) is collected from chat dialogues of an online game and adopts the annotation scheme of SDRT (Lascarides and Asher, 2009), with 16 labels specifically designed to capture the discourse relations for dialogues. Following the annotation scheme of STAC, Li et al. (2020b) construct a larger dialogue discourse structure dataset called MolWeni, deriving from the multiparty dialogues dataset *The Ubuntu dialogue Corpus* (Lowe et al., 2015). GUM (Zeldes, 2017) also contains discourse annotations on dialogues. In their annotation schema, they adopt the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) to capture the dis-

course structure of dialogues. Though contribute to the progress for DDP field, above datasets contain only textual modality, overlooking the prevalent multi-modalities in real-world scenarios that could help understand dialogue better.

The first multi-modal DDP dataset is constructed by Zhao et al. (2022b), namely JDDC 2.1, which is derived from the online dialogues in a Chinese E-commerce platform and also adopt SDRT annotation scheme. Each dialogue in JDDC2.1 contains multiple text utterances and two image utterances in average. As pioneered work for multi-modal DDP, JDDC 2.1 can serve as a valuable benchmark dataset. However, each dialogue in JDDC2.1 still mainly present by texts with few images, without audio or video modalities. Besides, as a task-specific domain dataset, it is not able to reflect the dialogue discourse structure in the more natural and prevalent multi-modal open-domain real-world scenarios.

Therefore, in this paper, we propose the first high-quality multi-modal Chinese discourse structure dataset for open-domain daily dialogues, with parallel text, visual, and audio modalities simultaneously, aiming to facilitate comprehensive research in the field of open-domain daily dialogues.

**DDP Approaches.** Early studies utilize handcrafted features to train traditional models and improve DDP with diverse decoding algorithms (Afantenos et al., 2015; Perret et al., 2016), considering only local information of two concerned utterances when predicting discourse relations. Taking into account global information of the whole dialogue structure, Shi and Huang (2019) propose a deep sequential model, the first to employ deep learning methods for DDP. Recent works attempt to further enhance DDP by better representing dialogue structures (Wang et al., 2021; Chi and Rudnicky, 2022; Wang et al., 2023; Li et al., 2023c) and speaker interactions (Yu et al., 2022; Li et al., 2023b) with various neural architectures, or explore to utilize external knowledge from additional data (Liu and Chen, 2021), pre-trained language models (Li et al., 2023a) and auxiliary NLP tasks (Yang et al., 2021; Fan et al., 2023) for DDP.

The above prior studies have made great progress in DDP, however, they all focused solely on the text modality. Although real-world dialogues frequently occur in multi-modal forms, and integrating various modalities holds potential for improved DDP, to the best of our knowledge, no approach has been specifically tailored for multi-modal DDP. In this work, we present a benchmark approach specifically designed for multi-modal DDP, aiming to provide reliable benchmark results on our newly constructed MODDP.

## 3 Data Construction

In this section, we describe the annotation methodology and annotation procedure in detail.

**Data Selection.** In order to build a multi-modal open-domain DDP dataset that can illustrate the discourse structure in real-world daily conversations, we follow the work of Zhao et al. (2022a) on multi-modal emotion recognition in two-party conversations, and select their collected video dialogue clips from several different Chinese TV series with the categories of family, romance, soap opera, and modern opera. Finally, we obtain 864 dialogues, with a total of 18,114 utterances for annotation. Each utterance in the dialogue contains both text and its aligned video clip.

**Annotation Guideline.** After comparing three mainstream discourse formalisms, including Penn Discourse Treebank (PDTB) (Prasad et al., 2008), Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), and Segmented Discourse Representation Theory (SDRT) (Lascarides and Asher, 2009), we adopt the SDRT-style annotation guideline defined in STAC dataset (Nicholas Asher and Afantenos, 2016) based on the following considerations. Compared with PDTB that focus on shallow discourse relations between a pair of utterances without considering the whole discourse structure, and RST which can capture the whole discourse structure but does not allow non-adjacent relations, SDRT is able to represent the overall discourse structures of dialogues and handle the non-adjacent relations that can occur in dialogues. Moreover, the SDRT-style annotation guideline released by STAC is specifically designed for dialogues, with 16 labels to distinguish discourse relations as shown in Table1, and gives detailed illustrations and examples to ensure annotation quality.

**Quality Control.** We employ 7 postgraduate students as our part-time annotators, and select 2 capable annotators with linguistic background as expert annotators to deal with annotation inconsistency. We compile a comprehensive annotation guideline for annotators' reference, which we upload as supplementary materials. The annotators are asked to annotate all the utterances in the dialogue sequentially with reference to the correspond-

| Label | Meaning | Proportion | Absolute frequency | | |
|---|---|---|---|---|---|
| | | | Train | Dev | Test |
| Comment | Arg2 provides opinion or evaluation on Arg1. | 16.4% | 1,946 | 298 | 584 |
| Elaboration | Arg2 provides more information about Arg1. | 15.7% | 1,857 | 283 | 565 |
| QA pair | Arg2 is the answer to the question Arg1. | 9.4% | 1,147 | 147 | 323 |
| Continuation | Arg2 is the continuation of Arg1. | 9.3% | 1,160 | 154 | 293 |
| Result | The eventuality in Arg2 is caused by Arg1. | 6.8% | 802 | 112 | 259 |
| Contrast | Arg1/2 share semantic structure, differ in themes. | 6.6% | 771 | 120 | 243 |
| Q-Clarify | Arg2 clarifies Arg1. | 6.4% | 759 | 121 | 226 |
| Q-Elab | Arg2 elaborates the question Arg1. | 6.3% | 739 | 105 | 248 |
| Explanation | Arg2 explains Arg1. | 5.2% | 600 | 107 | 193 |
| Alteration | Arg1 and Arg2 are alternations. | 4.5% | 559 | 67 | 156 |
| Background | Arg2 is the background of Arg1. | 3.2% | 382 | 65 | 110 |
| Conditional | Arg2 is the condition of Arg1. | 3.2% | 398 | 56 | 97 |
| Acknowledgement | Arg2 acknowledges Arg1. | 2.7% | 330 | 49 | 85 |
| Parallel | Arg1/2 share both semantic structure and theme. | 2.3% | 273 | 42 | 85 |
| Narration | Arg2 narrates Arg1. | 1.8% | 214 | 42 | 62 |
| Correction | Arg2 corrects Arg1. | 0.2% | 17 | 2 | 6 |

Table 1: The 16 relation labels adopted in our MODDP and the label distribution.

ing video clip. The annotation process consists of three parts. First, given the text of an utterance, we ask the annotator to watch its corresponding video clip. Second, they need to recognize a previous utterance that has discourse relation with the current utterance after watching the video. Finally, they are asked to identify the relation type between the two discourse-related utterances from among the 16 labels in the annotation guideline. We build an annotation tool to support the above annotation process, as shown in Appendix C.

Before formal annotation, each annotator is trained for several hours to be familiar with annotation guidelines and our developed annotation tool. During annotation, we perform strict double annotation based on our annotation tool to guarantee the quality of the labeled data. Specifically, each dialog is randomly assigned to two different annotators to independently annotate the whole discourse structure. If the annotations submitted by two annotators are the same, we directly take the consistent annotation as the final answer, otherwise, a third expert annotator decides the final answer after analyzing the two inconsistent submissions.

## 4 Analysis on MODDP

In this section, we analyze our annotated data from different perspectives.

**Inter-annotator Consistency.** We use Cohen's kappa value (Cohen, 1960) to calculate the inter-annotator consistency in the above mentioned double annotation workflow. For the consistency on discourse dependency links, the kappa value is 0.95, which is a very high agreement because most of the dependency links occur between adjacent utterances and thus can be easily recognized. We will give more discussion on dependency distances below. For the consistency on both links and relations, the kappa value is 0.63, which is higher than that of in the DDP datasets with only textual modality, such as 0.56 in STAC and 0.58 in Molweni datasets. This demonstrates that the multi-modal information can help the annotators better determine discourse structures in dialogues, which we discuss in detail in AppendixB. Even with a higher kappa value attained, the relatively low consistency in both link and relation labels reflects the difficulty in distinguishing discourse relations. This demonstrates the importance of performing double annotation to ensure data quality.

**Label Distribution.** Table 1 shows the distribution of different discourse relation labels in MODDP. The most frequently occurred relation is "Comment", accounting for 16.4% of all the relations. This reflects the characteristic of daily dialogue interaction that people usually express their opinions by commenting on the words of others. "Elaboration" takes the second largest proportion

of 15.7%, since people usually prefer to emphasize something by providing additional details. The next frequent relations are "Answer/ Question answer pair" and "Continuation", which are 9.4% and 9.3%. Compared with the label distribution in specific-domain datasets STAC and Molweni, where the top four most frequent relations account for over "60%" and "80%" respectively, the proportion in our MODDP is much lower (about 50%). This means that the distribution of different labels in our MODDP is more balanced, indicating the discourse relation in daily conversations is more flexible than that in specific domains.

**Dependency Distances.** To gain more insights on daily two-party dialogue discourse structure, we analyze the dependency distances in MODDP. We divide all the dependencies in MODDP into five groups according to the absolute distance between the head utterance and the modifier utterance, i.e., distance = 1, 2, 3, 4, and > 4. First, We find that most of the dependencies have the distance of 1, accounting for 92.0%, meaning that the discourse relations usually occur between adjacent utterances. The reason is that different from previous DDP datasets such as STAC and Molweni which are sourced from online dialogues, MODDP dataset is sourced from daily offline two-party dialogues. In the offline scenario, speakers cannot directly see the dialogue history as they can online. This real-time, two-party daily dialogue scenario results in the speaker's utterance often directly related to the most recent utterance. Second, the longer the dependency distance, the fewer dependencies there are in the corresponding group, with the proportion of distance = 2, 3, 4, and >4 are 5.4%, 1.3%, 0.5% and 0.8%, respectively, indicating that in daily dialogue, speakers tend to respond to recent utterances rather than earlier ones. For the discourse relation occurs between non-adjacent utterances, we observe the annotated MODDP, and find that it happens in the situations when the speaker would like to disregard the current topic and continue with the historical discussion, or there is a need to review earlier utterances again.

## 5 Approach

In this section, we present a simple yet effective benchmark approach for multi-modal open-domain DDP, aiming to provide reliable benchmark results on our newly annotated dataset to facilitate researchers in their further exploration.

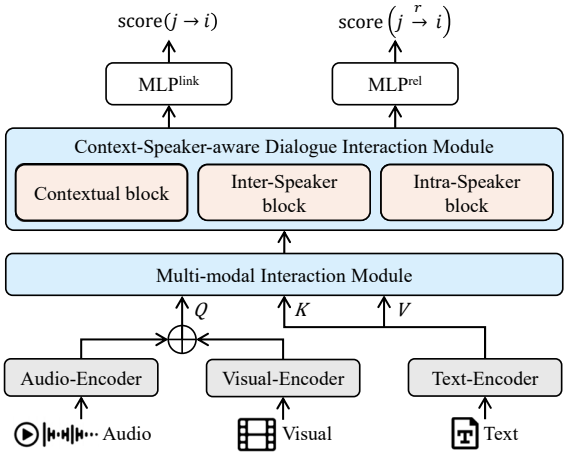


Figure 2: Architecture of our multi-modal dialogue discourse parser.

### 5.1 Problem Description

Formally, given a dialogue of $n$ utterances $\{u_0, u_1, ..., u_n\}$, where $u_0$ is a pseudo root utterance and $u_i$ represents the $i$-th utterance in the dialogue. Each utterance $u_i$ has the modalities of text, visual, and audio simultaneously, denoted as $u_i^t$, $u_i^v$, and $u_i^a$, respectively. The goal of discourse parsing on multi-modal dialogues is to predict a discourse dependency graph $\boldsymbol{d} = \{(u_j, u_i, r), 0 \leq j \leq n, 1 \leq i \leq n, r \in \mathcal{R}\}$ with consideration of all the three modalities, where $(u_j, u_i, r)$ is a dependency from the head utterance $u_j$ to the modifier utterance $u_i$ with the discourse relation $r$, and $\mathcal{R}$ is the relation label set.

### 5.2 Multi-modal Dialogue Discourse Parser

Figure 2 shows the architecture of the proposed multi-modal dialogue discourse parser.

**Utterance Representation.** For each utterance $u_i$, we employ three modality encoders to obtain the contextualized utterance representations for the three modalities respectively. Specifically, we adopt RoBERTa (Liu et al., 2019) to encode the textual modality $u_i^t$, Vision Transformer (ViT) (Dosovitskiy et al., 2021) to encode the visual modality $u_i^v$, and Wav2Vec2.0 (Baevski et al., 2020) to encode the audio modality $u_i^a$. We denote the corresponding utterance representations as $\mathbf{h}_i^t$, $\mathbf{h}_i^v$, and $\mathbf{h}_i^a$, where $\mathbf{h}_i^t$ and $\mathbf{h}_i^v$ are the representations of "[CLS]" position from RoBERTa and ViT respectively, and $\mathbf{h}_i^a$ is the last hidden state from Wav2Vec 2.0.

**Multi-modal Interaction.** After obtaining utterance representations of different modalities, we introduce a multi-modal interaction module to fuse

the information across modalities. Intuitively, for the task of multi-modal discourse parsing, the textual modality of utterances appear to take a predominant role over the visual and audio modalities, since it typically convey more explicit semantics for understanding the utterances in dialogue, while the audio and visual modalities can serve as supplements to the textual modality. With above consideration, we treat text as the core modality for multi-modal interaction by employing a cross-modality multi-head attention (CMA) mechanism, taking the non-textual modalities as query $\mathbf{q}_i$, textual modality as the key $\mathbf{k}_i$ and value $\mathbf{v}_i$. The multi-modal interaction representation $\mathbf{h}_i^m$ is:

$$\begin{aligned} \mathbf{h}_i^m &= \mathrm{CMA}(\mathbf{q}_i, \mathbf{k}_i, \mathbf{v}_i) \\ \mathbf{q}_i &= \mathbf{h}_i^v \oplus \mathbf{h}_i^a; \quad \mathbf{k}_i = \mathbf{v}_i = \mathbf{h}_i^t \end{aligned} \tag{1}$$

where $\oplus$ is a concatenate operation.

**Context-Speaker-aware Dialogue Interaction.** We then feed $\mathbf{h}_i^m$ into the context-speaker-aware dialogue interaction module to capture the interaction between utterances in the dialogue while considering both contextual and speaker information. This facilitates a comprehensive understanding of the dialogue. Following Li et al. (2020a), the utterance interaction module consists of three transformer blocks, i.e., contextual block, inter-speaker block, and intra-speaker block. They are realized by applying different masking strategies to the three transformer blocks: global mask to allow all the utterances in the dialogue to be accessed in the contextual block, intra-speaker mask to only allow the utterances that have the same speaker as $u_i$ to be accessed in the intra-speaker block, inter-speaker mask to only allow the utterances that have different speakers with $u_i$ to be accessed in the inter-speaker block. We denote the output of the three blocks as $\mathbf{h}_i^{\text{context}}$, $\mathbf{h}_i^{\text{intras}}$, and $\mathbf{h}_i^{\text{inters}}$. The final output of the utterance interaction module is:

$$\mathbf{h}_i^u = \alpha^0 \mathbf{h}_i^{\text{context}} + \alpha^1 \mathbf{h}_i^{\text{intras}} + \alpha^2 \mathbf{h}_i^{\text{inters}} \tag{2}$$

where $\alpha^0$, $\alpha^1$, $\alpha^2$ are learnable parameters, and their sum is 1.

**Link Prediction and Relation Classification.** Based on the representation obtained from the utterance interaction module, we perform link prediction and relation classification to predict the discourse structure. Link prediction aims to predict the dependency links between utterances, and relation classification aims to classify the discourse relation label of an existing dependency link. We compute the score of a dependency link $j \rightarrow i$, and the score of its relation as follows:

$$\begin{aligned} \mathbf{h}_{i,j} &= \mathrm{FC}(\mathbf{h}_i^u) \oplus \mathrm{FC}(\mathbf{h}_j^u) \\ \mathrm{score}(j \rightarrow i) &= \mathrm{MLP}^{\text{link}}(\mathbf{h}_{i,j}) \\ \mathrm{score}(j \xrightarrow{r} i) &= \mathrm{MLP}^{\text{rel}}(\mathbf{h}_{i,j}) \end{aligned} \tag{3}$$

where $\mathrm{MLP}^{\text{link}}$ and $\mathrm{MLP}^{\text{rel}}$ are two multi-layer perceptrons for computing link and relation scores respectively, and FC denotes a fully connected layer.

**Training Loss.** During training, we compute two independent cross-entropy losses for link prediction and relation classification, in order to maximize the probability of the correct dependency link and the correct relation on it.

| | Train | Dev | Test | Total |
|---|---|---|---|---|
| #Dialog | 604 | 87 | 173 | 864 |
| #Utt | 12,549 | 1,857 | 3,708 | 18,114 |
| #Turn | 5,559 | 788 | 1,640 | 7,987 |
| Avg. Utt/Dialog | 20.77 | 21.34 | 21.43 | 20.96 |
| Avg. Utt Length | 8.91 | 8.93 | 8.96 | 8.92 |
| Avg. Utt/Turn | 2.25 | 2.35 | 2.26 | 2.27 |
| Video Duration | 8.83h | 1.25h | 2.60h | 12.68h |

Table 2: Data statistics, including the number of dialogues (#Dialog), the number of utterances (#Utt), the number of turns (#Turn), the average length (in utterances) of dialogues (Avg. Utt/Dialog), the average length (in characters) of utterances (Avg. Utt Length), the average length (in utterances) of turns (Avg. Utt/Turn), and the video duration (in hours).

## 6 Experiments

**Data.** We randomly split our newly annotated MODDP datasets into train, dev, and test sets with the proportion of 7/1/2. Table 2 shows the data statistics.

**Evaluation Metrics.** Following previous works, we adopt the standard unlabeled attachment score (UAS) and labeled attachment score (LAS) as evaluation metrics. UAS focused solely on the correctness of the dependency link (also known as Link F1 score). ***LAS*** is typically considered as the ***primary evaluation metric*** for discourse parsing, since it considers the correctness of both the dependency link and the relation type (also known as Link&Rel F1 score).

**Implementation Detail.** We extract the fixed RoBERTa, ViT, and Wav2Vec2.0 representations as

| Model | Dev | | Test | |
|---|---|---|---|---|
| | UAS | LAS | UAS | LAS |
| with Textual Modality Only | | | | |
| DeepSeq | 92.23 | 41.91 | 92.55 | 43.39 |
| Hierarchical | 88.32 | 43.01 | 88.40 | 42.64 |
| Struct self-aware | 90.02 | 41.24 | 90.73 | 43.11 |
| SDDP | 90.02 | 41.19 | 90.67 | 42.77 |
| Speaker-aware | 91.89 | 42.56 | 92.05 | 42.65 |
| ours | 92.18 | 44.97 | 92.46 | 43.66 |
| with Multi-modalities | | | | |
| ours | 91.74 | 49.38 | 90.90 | 48.05 |
| Concat | 92.11 | 41.55 | 92.00 | 42.59 |
| Sum | 90.65 | 44.80 | 90.00 | 43.62 |
| Self-att | 91.93 | 45.57 | 92.02 | 45.51 |
| Cross-att (T2VA) | 91.91 | 45.45 | 91.60 | 45.19 |

Table 3: Dialogue discourse paring results on our MODDP dataset.

textual, visual, and audio utterance-level features, using chinese-roberta-wwm-ext[1], vit-base-patch16-224[2], and wav2vec2-large-xlsr-53-chinese-zh-cn[3], respectively. We use AdamW as the optimizer. The initial learning rate for encoders and other modules are 1e-5 and 1e-3, and decays at a rate of 1e-6. We train our model for 20 epochs and save the model with the best LAS performance on the dev data. We run each of our model for three times with different random seeds and report the average result. We provide more detailed parameter settings in the Table 5 of Appendix A.

### 6.1 Main Results

Table 3 shows the main results of dialogue discourse parsing on our MODDP dataset, comparing the performance under the settings of using textual modality only and using multi-modalities (i.e., textual, visual, and audio modalities).

**Results with Textual Modality Only.** In the first major row of Table 3, we show the results of several state-of-the-art DDP methods, i.e., the deep sequential model (DeepSeq) (Shi and Huang, 2019), the hierarchical model (Hierarchical) (Liu and Chen, 2021), the structure self-aware model (Wang et al., 2021), the structured dialogue discourse parsing (SDDP) model (Chi and Rudnicky, 2022), the speaker-aware model (Yu et al., 2022), and our proposed benchmark method (removing the visual/audio utterance representations and multi-modal interaction modal from Figure 2. For fair comparison, all the models use the same fixed RoBERTa representations as textual utterance-level features.

Looking into the results, we find that nearly all the models can achieve relatively high UAS of more than 90 since most of the dependency links in MODDP occur between adjacent utterances as discussed in Section 4. However, the LAS drops dramatically. This indicates that distinguishing discourse relations in open-domain daily dialogues is very challenging, especially without reference to multi-modal information.

Compared with the state-of-the-art models, our proposed benchmark method achieves better performance in the textual unimodal setting, demonstrating that our proposed method can be served as a strong benchmark model to provide reliable benchmark result. Therefore, we further investigate the effectiveness of integrating multi-modalities based on our proposed benchmark model.

**Results with Multi-modalities.** In the second major row of Table 3, we show the results of introducing multi-modal information for DDP, comparing the performance of replacing various multi-modal interaction modules in our proposed benchmark model. The "ours" row presents the result of our model described in Section 5. We further replace the original multi-modal interaction module, i.e., employing cross-attention with textual modality as query and non-textual modalities as key and value, with 1) concatenation of multiple modalities representations (Concat), 2) sum of multiple modalities representations (Sum), 3) passing the concatenated multiple modalities representations through self-attention (Self-att), 4) employing cross-attention with non-textual modalities (i.e., visual and audio) as query and textual modality as key and value. From the results, we have the following observations.

First, compared to our model that solely using textual modality with that using multi-modalities, we observe that incorporating multi-modalities significantly enhances the performance of DDP by 4.41/4.39 in LAS on dev/test sets. This demonstrates that multi-modal information greatly benefits DDP, as it offers a more comprehensive understanding of the dialogue context. Second, the performance is significantly affected by different multi-modal interaction methods. In particular,

[1] https://huggingface.co/hfl/chinese-roberta-wwm-ext
[2] https://huggingface.co/google/vit-base-patch16-224
[3] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-chinese-zh-cn

| Modalities | Dev | | Test | |
|---|---|---|---|---|
| | UAS | LAS | UAS | LAS |
| T | 92.18 | 44.97 | 92.46 | 43.66 |
| V | 91.96 | 25.67 | 91.89 | 24.41 |
| A | 91.93 | 34.54 | 91.74 | 33.00 |
| T+V | 91.36 | 48.70 | 90.36 | 47.59 |
| T+A | 91.39 | 48.18 | 90.53 | 47.21 |
| V+A | 91.63 | 35.92 | 90.96 | 34.28 |
| T+V+A | 91.74 | 49.38 | 90.90 | 48.05 |

Table 4: Results of using different modalities.

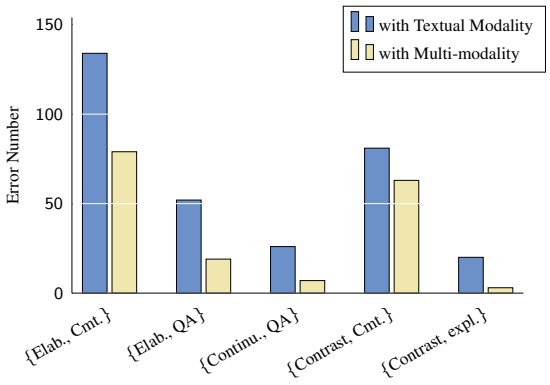


Figure 3: Statistics of different error patterns on test.

treating the three modalities equally such as the "Concat" and "Sum" methods can not help DDP or even can decrease the performance, meaning that visual and audio modalities sometimes may not provide semantically relevant information for discourse parsing, instead offering distracting information. Third, we observe that our multi-modal interaction module, which acquires textual information with reference to non-textual information (i.e., 'ours'), outperforms the module that acquires non-textual information with reference to textual information (i.e., Cross-att (T2VA)). This indicates that the textual modality holds greater importance than other modalities in discourse parsing, which is consistent with our intuition in Section 5.

Overall, we can conclude that introducing multi-modalities is of great benefit to DDP, with the textual modality playing a predominant role while the other modalities serve as auxiliary.

### 6.2 Analysis

**Effects of Different Modalities.** We analyze the effects of introducing different modalities for DDP, as shown in Table 4. We observe that, in the uni-modal setting, the DDP model using textual modality outperforms that using visual or audio modality by a very large margin, again demonstrates the predominant role of textual modality for DDP. We also find that introducing more modalities can improve the performance consistently, and the best result is achieved by fully exploiting all of the three modalities. Overall, we can conclude that the textual, visual, and audio modalities can provide complementary contributions for DDP.

**Analysis on Error Patterns.** To gain a deeper understanding of the improvements brought by multi-modalities, we analyze how the number of error patterns changes between the model using multi-modalities and that relying solely on textual modality. The error pattern "{X, Y}" means a dependency link with a gold relation label of "X" is incorrectly predicted to a label "Y" or vice versa. As shown in Figure 3, we select the error patterns that change most in number and present them in descending order of absolute change number.

We observe that the error pattern of "{Elaboration, Comment} " changes most in number. The reason is that compared to the "Comment" label, the "Elaboration" label usually occurs when people emphasize something by providing additional details, with a tone and expression that typically become more forceful. The audio and visual modalities can convey the tone and expression to help distinguish between these two labels. The reason of the decrease of the error pattern "{Elaboration, QA pair}" is similar. We also see that the model integrated with multi-modalities greatly reduces the error numbers of "{Contrast, Comment}" and "{Contrast, Explanation}", since when people express "Contrast", their tone and expression usually change suddenly, while they typically appear neutral and calm when comment or give explanations. This can be effectively conveyed by multi-modalities to reduce confusion on these labels.

## 7 Conclusions

This paper proposes MODDP, the first multi-modal open-domain dialogue discourse parsing dataset, consisting of 864 dialogues and 18,114 utterances with aligned video clips of daily conversations. We present a detailed description of the construction process for MODDP and give in-depth data analysis. We propose a simple yet effective benchmark approach for multi-modal DDP, and conduct ex-

tensive experiments, providing several benchmark results under both textual unimodal setting and multi-modal setting based on our newly constructed MODDP. Experimental results demonstrate the significant benefits of multi-modalities for DDP. We hope that MODDP will facilitate future research in the valuable yet under-explored field of multi-modal DDP.

## 8 Limitations

We think the limitations of our work are two-fold.

First, despite great efforts, the current size of our MODDP dataset remains relatively small. We will continue to collect more dialogues from real-world scenarios, and plan to construct more high-quality labeled data with discourse structures and multi-modalities to facilitate in-depth research in the field of multi-modal DDP.

Second, based on our newly annotated MODDP, we have introduced a straightforward benchmark approach to conduct preliminary experiments, reporting both textual unimodal and multi-modal performance. However, many other potentially beneficial approach to integrate multi-modalities into DDP can be further investigated. We believe that multi-modal open-domain DDP is a valuable research field and encourage future work to explore it using MODDP.

## 9 Ethical Considerations

For the copyright concerns related to our MODDP dataset, which includes video clips from TV series, we have consulted with professional legal advisors and reviewed the related copyright laws. The release of MODDP dataset will fully comply with the fair use principle in copyright laws. We will publicly release the texts of dialogues, the annotated discourse links, relation labels, speakers, timestamps of the start and end of utterances, and the names of the sourced TV series on https://github.com/Suda-iaiNLP/MODDP. Additionally, we will also publicly release the visual and audio representations extracted from ViT and Wav2Vec2.0. For access to the video source data, we will require the researchers to apply to us by committing that the data will be used solely for academic research and not for commercial or other non-academic purposes.

For the annotation payment, all the annotators are salaried for their work according to their annotation quality and quantity. The average salary is about 30 RMB per hour, which is a fair and reasonable wage in China.

## Acknowledgements

We would like to thank the anonymous reviewers for the helpful comments. This work was supported by National Natural Science Foundation of China (Grant No. 62306202 and 62076173), and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No.23KJB520034).

## References

Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 928–937, Lisbon, Portugal. Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ta-Chung Chi and Alexander Rudnicky. 2022. Structured dialogue discourse parsing. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 325–335, Edinburgh, UK. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yaxin Fan, Feng Jiang, Peifeng Li, Fang Kong, and Qiaoming Zhu. 2023. Improving dialogue discourse parsing via reply-to structures of addressee recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8484–8495, Singapore. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2021. Dialogue discourse-aware graph model and data augmentation for meeting summarization. In *Proceedings of the Thirtieth International*

*Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3808–3814.

Qi Jia, Yizhu Liu, Siyu Ren, Kenny Zhu, and Haifeng Tang. 2020. Multi-turn response selection using dialogue dependency relations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1911–1920, Online. Association for Computational Linguistics.

Alex Lascarides and Nicholas Asher. 2009. Agreement, disputes and commitments in dialogue. *Journal of semantics*, 26(2):109–158.

Chuyuan Li, Patrick Huber, Wen Xiao, Maxime Amblard, Chloe Braud, and Giuseppe Carenini. 2023a. Discourse structure extraction from pre-trained and fine-tuned language models in dialogues. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2562–2579, Dubrovnik, Croatia. Association for Computational Linguistics.

Jiangnan Li, Zheng Lin, Peng Fu, Qingyi Si, and Weiping Wang. 2020a. A hierarchical transformer with speaker modeling for emotion recognition in conversation. *arXiv preprint arXiv:2012.14781*.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020b. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jiaqi Li, Ming Liu, Yuxin Wang, Daxing Zhang, and Bing Qin. 2023b. A speaker-aware multiparty dialogue discourse parser with heterogeneous graph neural network. *Cognitive Systems Research*, 79:15–23.

Wei Li, Luyao Zhu, Wei Shao, Zonglin Yang, and Erik Cambria. 2023c. Task-aware self-supervised framework for dialogue discourse parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14162–14173, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Zhengyuan Liu and Nancy Chen. 2021. Improving multi-party dialogue discourse parsing via domain integration. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 122–127, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Mathieu Morey Benamara Farah Nicholas Asher, Julie Hunter and Stergos Afantenos. 2016. Fiscourse structure and dialogue acts in multiparty dialogue: the stac corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727.

Jérémy Perret, Stergos Afantenos, Nicholas Asher, and Mathieu Morey. 2016. Integer linear programming for discourse parsing. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 99–109, San Diego, California. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7007–7014.

Ante Wang, Linfeng Song, Hui Jiang, Shaopeng Lai, Junfeng Yao, Min Zhang, and Jinsong Su. 2021. A structure self-aware model for discourse parsing on multi-party dialogues. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 3943–3949.

Ante Wang, Linfeng Song, Lifeng Jin, Junfeng Yao, Haitao Mi, Chen Lin, Jinsong Su, and Dong Yu. 2023. $D^2$PSG: Multi-party dialogue discourse parsing as sequence generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:4004–4013.

Jingxuan Yang, Kerui Xu, Jun Xu, Si Li, Sheng Gao, Jun Guo, Nianwen Xue, and Ji-Rong Wen. 2021. A joint model for dropped pronoun recovery and conversational discourse parsing in Chinese conversational speech. 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 1752–1763, Online. Association for Computational Linguistics.

Nan Yu, Guohong Fu, and Min Zhang. 2022. Speaker-aware discourse parsing on multi-party dialogues. In *Proceedings of the 29th International Conference*

*on Computational Linguistics*, pages 5372–5382, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Duzhen Zhang, Feilong Chen, and Xiuyi Chen. 2023. DualGATs: Dual graph attention networks for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7395–7408, Toronto, Canada. Association for Computational Linguistics.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022a. 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)*, pages 5699–5710, Dublin, Ireland. Association for Computational Linguistics.

Nan Zhao, Haoran Li, Youzheng Wu, and Xiaodong He. 2022b. JDDC 2.1: A multimodal Chinese dialogue dataset with joint tasks of query rewriting, response generation, discourse parsing, and summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12037–12051, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

# Appendices

## A Parameter Settings

| Hyper-paramters | Value |
|---|---|
| Text Encoder | Roberta-base [4] |
| Visual Encoder | Vit-base [5] |
| Audio Encoder | Wav2Vec2.0-large [6] |
| Batch size | 1 Dialogue |
| Optimizer | AdamW<br>($\beta_1$, $\beta_2$) = (0.9, 0.9)<br>$\epsilon$ = 1e-12 |
| Weight decay | 1e-6 |
| Learning rate | 1e-3 |
| Learning rate-pm | 1e-5 |
| Learning rate scheduler | Linear |
| Dropout | 0.1 |
| Gradient clipping | 5 |
| Max train epochs | 20 |
| Devices | Nvidia V100 GPU |
| Total training time | About 8 hours |
| **Fusion Module** | |
| MultiHead-Att block num | 2 |
| Heads | 8 |
| Hidden size | 768 |
| **Interaction Module** | |
| Transformer block num | 6 |
| Heads | 4 |
| Hidden size | 300 |
| Learning rate | 1e-4 |

Table 5: Our hyper-parameter settings.

We list detailed experimental hyper-parameter settings in Table 5.

## B Analysis on Benefits of Multi-modalities for annotation

We interview all the annotators to investigate whether providing multi-modal information is helpful in determining discourse structures during their annotation process, and ask them for the reasons. All of the annotators hold the view that displaying multiple modalities instead of only text modality of the utterances can help them better figuring out the discourse structures of the dialogue. The reasons are as follows. First, the speakers' tone and the facial expressions in the audio modality and visual

modality can convey their emotions for better understanding the semantics of the utterances, and thus assisting annotators in determining the correct discourse structure. Second, when the mentions of the objects or speakers' actions in the utterance are not clearly expressed in the text modality, it is necessary to refer to visual modality to obtain the complete semantics. Another scenario where visual modality can be essential is that the current utterance is a response to the actions of the speaker. In summary, multi-modality is beneficial for annotators to comprehensively and accurately understand the semantics in dialogues for high-quality discourse structure annotation.

## C Annotation Tool

We show the annotation interface of our annotation tool in Figure 4. For each dialogue to be annotated, the annotation tool presents all its utterances, with both texts and video clips simultaneously. When clicking on the text of an utterance, its corresponding video clip will play for the annotators to reference. To annotate a dependency, annotators just need to first click an utterance, and then select its head from among previous utterances, and finally select the discourse relation label. The annotators must assign each utterance in the dialogue with a head and a relation label before submission.


Figure 4: Annotation interface of our annotation tool.