DraDDP: A Multimodal Multi-Party Dialogue Discourse Parsing Dataset

Shannan Liu, Peifeng Li*, Yaxin Fan, Qiaoming Zhu
School of Computer Science and Technology, Soochow University, Suzhou, China
20234027002@stu.suda.edu.cn
{pfli, qmzhu}@suda.edu.cn, yxfansuda@stu.suda.edu.cn

Abstract

Multi-party dialogue discourse parsing aims to identify dependency structures and relation types between utterances in conversations. Previous studies are mostly limited to textual modality or two-party dialogue, failing to meet the multimodal and multi-party settings. In this paper, we construct the first publicly available English multimodal dataset DraDDP for multi-party dialogue discourse parsing, based on American TV dramas. DraDDP contains 495 dialogue segments with 6,374 utterances and 9.1 hours of parallel video content, covering rich multi-party interaction scenarios. Moreover, we establish comprehensive benchmarks by evaluating this task on DraDDP and conducting in-depth analysis on the impact of different modalities. Experimental results demonstrate the value of multimodal information in capturing dialogue structures and relation types. We will publicly release the dataset, annotation guidelines, and code to promote future research in multimodal dialogue understanding.[1]

1 Introduction

Multi-party dialogue discourse parsing aims to identify dependency structures and semantic relation types (e.g., *Comment*, *Background*, and *Alternation*) between utterances in multi-party conversations. As shown in Figure 1, this example contains 5 utterances, where the arcs represent dependency structures between utterances, and the labels on the arcs indicate the types of discourse relations. This task is of significant value for downstream applications such as meeting summarization (Feng et al., 2021; Gao et al., 2023; Rennard et al., 2024), dialogue generation (Fan et al., 2024b; Li et al., 2024b), and emotion recognition (Zhang et al., 2023; Hao et al., 2024).

Previous research on multi-party dialogue discourse parsing focused on two public datasets: STAC (Asher et al., 2016) and Molweni (Li et al., 2020). However, these datasets only consider the textual modality, ignoring the complexity and richness of multimodal interactions in real-world scenarios (Zhang et al., 2022; Ju et al., 2024). As shown in Figure 1, when relying solely on textual modality, it is difficult to understand why *Ross* would suddenly mention *"green"* (a seemingly unrelated response) after *Rachel* expresses personal emotions. When audio-visual modalities are introduced, we observe that *Rachel* is in a private phone call scenario, while *Ross* is engaged in gaming interaction with friends nearby, with both being in parallel and independent dialogue contexts. This indicates that multimodal information not only supplements scene details not covered by text, but also plays an irreplaceable role in identifying dependency structures in multi-party dialogues and ensuring contextual semantic coherence.

Currently, there are only two available datasets for multimodal dialogue discourse parsing: JDDC 2.1 (Zhao et al., 2022) and MODDP (Gong et al., 2024). Both datasets focus on two-party dialogues and only support Chinese, failing to meet the needs of multi-party dialogue and cross-lingual research. Compared to two-party dialogues, multi-party dialogues involve multiple participants and have more complex structures. Therefore, understanding the discourse structure of multi-party dialogues is more valuable and challenging.

In this paper, we construct the first English multimodal dataset **DraDDP** (**Dra**ma-based **D**ialogue **D**iscourse **P**arsing) for the task of multi-party dialogue discourse parsing. DraDDP is annotated based on the classic American TV dramas (e.g., *Friends*), covering rich multi-party interaction scenarios and emotional expression patterns. To the best of our knowledge, DraDDP is the first publicly available English multimodal multi-party dia-

* Corresponding author
[1] https://github.com/DraDDP

Findings of the Association for Computational Linguistics: ACL 2026, pages 27359–27373

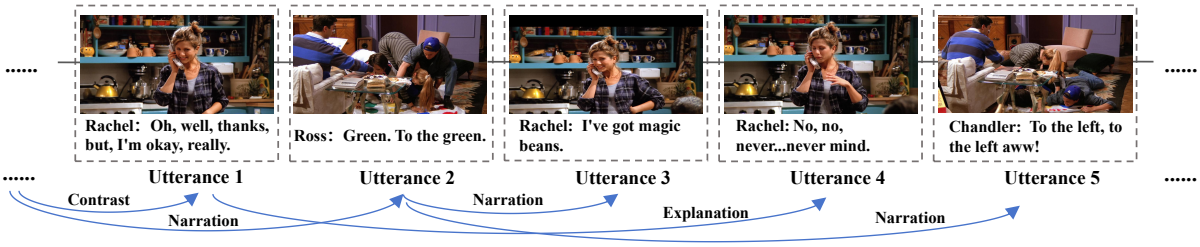


Figure 1: An example of multimodal dialogue discourse parsing.

Modalities	Dataset	#Dialogues	#Utterances	Data Source	Language	Participants
T	STAC	1.1K	10.7K	Online game	English	Multi-party
T	Molweni	10K	88.3K	Online forums	English	Multi-party
T	DialogueDSA	705	24.5K	TV drama	Chinese	Multi-party
T	MSDC	541	22.6K	Online game	English	Two-party
T+I	JDDC 2.1	246K	3.46M	E-commerce platform	Chinese	Two-party
T+V+A	MODDP	864	18K	TV drama	Chinese	Two-party
T+V+A	DraDDP	495	6.4K	TV drama	English	Multi-party

Table 1: Comparison with existing dialogue discourse parsing datasets (T/V/A/I: Text/Video/Audio/Image).

logue discourse parsing dataset, providing a novel research benchmark for this field. The main contributions of this paper include:

1) We construct the first English multimodal dataset DraDDP for multi-party dialogue discourse parsing, containing 495 dialogue segments, 6,374 utterances, and 9.1 hours of parallel video content, providing rich resources for multimodal dialogue understanding research.

2) We establish comprehensive benchmarks by evaluating multiple dialogue discourse parsing models on DraDDP, and conducting systematic analysis to reveal the impact of multimodal information on parsing performance.

2 Related Work

Datasets Table 1 presents the core attributes of available datasets on dialogue discourse parsing. The textual datasets include STAC (Asher et al., 2016), Molweni (Li et al., 2020), DialogueDSA (Jiang et al., 2023), and MSDC (Thompson et al., 2024b). Notably, while MSDC incorporates textually described non-verbal actions (e.g., "picks up blue (-1,1,1)") as discrete symbolic elementary discourse unit (EDU) nodes, real-world non-verbal signals (e.g., facial expressions and intonation) are continuous and temporally synchronized, exhibiting higher semantic complexity.

There are only two public datasets on multimodal dialogue discourse parsing: JDDC 2.1 (Zhao et al., 2022) and MODDP (Gong et al., 2024). Although JDDC 2.1 introduces image modality, the image content is relatively scarce and limited to specific domains. MODDP, sourced from TV drama dialogue scenarios, achieves significant improvements in modality completeness and scenario authenticity. However, both multimodal datasets only cover two-party dialogues and exclusively support Chinese, making it difficult to meet the urgent needs for multi-party dialogue and cross-lingual research.

Textual Methods Current mainstream research on dialogue discourse parsing primarily leverages pre-trained language models, which enhance parsing performance through strategies such as modeling key dialogue elements (Wang et al., 2024; Li et al., 2024a), injecting external information (Li et al., 2023; Ma et al., 2023), or joint learning (Xu et al., 2024; Fan et al., 2025a). With the rapid development of Large Language Models (LLMs), Chan et al. (2023) and Fan et al. (2024a) found that ChatGPT performed poorly on dialogue discourse parsing. Thompson et al. (2024a) proposed LLaMIPa (LLaMA Incremental Parser), which achieved incremental prediction based on historical discourse structures through fine-tuning LLaMA3. Besides, Liu et al. (2025) and Fan et al. (2025b) improved LLMs through explanatory prompts and dialogue clarification. However, these advances remain limited to text-only scenarios.

Multimodal Methods There are only a few studies on multimodal dialogue discourse parsing. Only MODDP (Gong et al., 2024) provides a basic multimodal benchmark, employing cross-modal attention to fuse multimodal features. However, it fo-

cuses solely on two-party Chinese dialogues and does not explore multimodal large language models (MLLMs). We address this gap by constructing the first English multimodal dataset for the task of multi-party dialogue discourse parsing and benchmarking both traditional and MLLM approaches.

3 Data Construction

3.1 Data Preparation

DraDDP uses the first season of the American TV series *Friends* (1994) as its data source, covering all 24 episodes. This choice offers three main advantages: 1) the dialogue participants include 6 core protagonists and over 20 supporting characters, generating rich multimodal interaction information such as body language, facial expressions, and intonation; 2) the dialogue scenarios are diverse, covering homes, cafes, and workplaces, with topics spanning multiple dimensions including emotions, daily life, humor, and negotiation, providing representative multi-party interaction patterns for discourse parsing research; 3) the subtitle data from Season 1 has been widely used in NLP research (e.g., the emotion recognition dataset MELD, the dialogue generation and question answering dataset FriendsQA), providing a well-established foundation within the research community.

For EDU segmentation, we adopted each official subtitle line as a basic discourse unit based on three considerations: 1) subtitle lines are professionally produced with moderate length; 2) segmentation follows speaker turns and semantic boundaries without crossing scene transitions; 3) precise timestamps enable accurate alignment of text, video frames, and audio segments.

3.2 Annotation Guidelines

The dialogue datasets shown in Table 1 are all constructed based on Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003). SDRT employs directed graph structures to represent dependency relations between discourse units, which can effectively capture complex interaction patterns and dynamic contextual changes in dialogues. Furthermore, we utilized the 16 relation labels from the STAC (Asher et al., 2016) system (detailed in Appendix A) to distinguish different types of discourse relations. This labeling system provides comprehensive definitions and rich annotation examples, offering guarantees for annotation quality.

3.3 Annotation Quality Control

To ensure the dataset annotation quality meets academic standards, we designed a rigorous four-stage quality control system. The entire annotation work was completed through collaboration among 2 doctoral students and 4 master's students, whose research fields are dialogue or discourse analysis. All 6 annotators possessed advanced English reading and listening proficiency suitable for research settings. Before the formal annotation process began, they received approximately 20 hours of systematic training, including instruction in SDRT, detailed explanations of the 16 STAC relation types, and collaborative practice on real-world data.

1) We introduced a pre-annotation mechanism to improve data annotation efficiency and provide reliable initial references for human annotation. Specifically, we employed LLaMA3[2] fine-tuned on the STAC dataset to perform preliminary discourse structure prediction on all data based on the text modality (section 3.4 for details).

2) The annotators collectively watched corresponding video segments based on model predictions and collaboratively corrected discourse structures. The main goal was to establish unified annotation standards, during which 1/6 of the dataset was annotated collaboratively. We systematically compiled problems encountered during annotation and developed comprehensive annotation guidelines, which will be publicly released as supplementary materials alongside the dataset.

3) Each episode's data was randomly assigned to two different annotators to independently complete full discourse structure annotation. When the results from both annotators were completely consistent, the annotation was directly adopted as the final result; when disagreements occurred, consensus was reached through collective discussion, and related issues were incorporated into the annotation guidelines for improvement. This stage completed annotation of 1/3 of the dataset.

4) The remaining data was randomly assigned to two annotators for preliminary annotation, with disagreed portions adjudicated by a third annotator (a doctoral student).

We used Fleiss' Kappa coefficient (Fleiss, 1971) to evaluate inter-annotator agreement. The Kappa value for discourse dependency structures was 0.91, showing high consistency, mainly attributed to

[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B

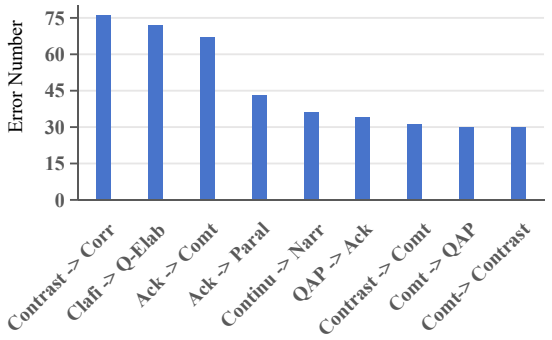


Figure 2: Error statistics from pre-annotation where *{X → Y}* indicates misclassifying relation X as Y and the relations are defined in Appendix A.

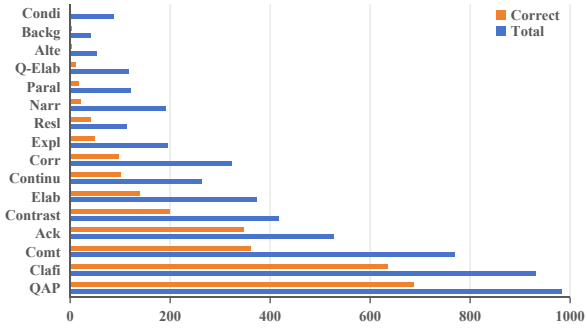


Figure 3: Analysis of pre-annotation results, including total number and correct number of each relation type.

most dependency structures occurring between adjacent utterances, making identification relatively straightforward. The Kappa value for relation types was 0.60, which, although exceeding the STAC corpus's 0.58, reflects the inherent complexity and challenges of distinguishing discourse relations through its relatively low consistency. We present several challenging annotation cases in Appendix B to further illustrate the difficulties and complexities encountered during the annotation process.

3.4 Pre-annotation Analysis on LLaMA3

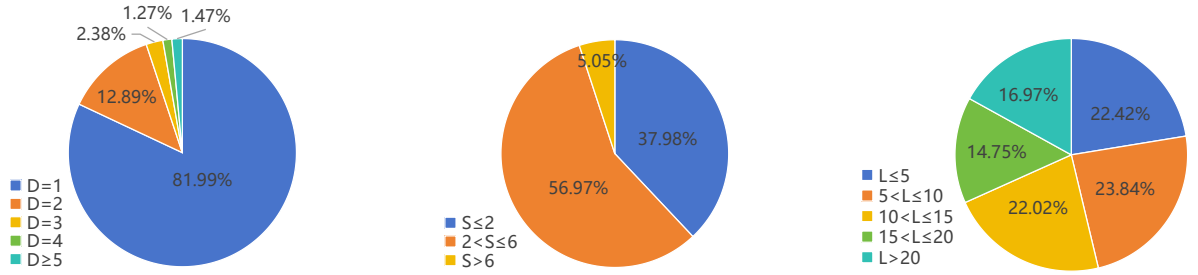
To improve annotation efficiency and provide initial references for human annotators, we employed a pre-annotation mechanism. Specifically, we used LLaMIPa†, a variant of LLaMIPa without predicted discourse context (Thompson et al., 2024a), which was trained on the STAC dataset to generate preliminary discourse structure predictions for DraDDP based on textual modality. Given the text sequence from the beginning of a dialogue to the current utterance, $\{u_0, u_1, \ldots, u_i\}$, the model predicts the dependency parent and relation type for the current utterance u_i.

Compared with final human annotations, the model achieved an F1-score of 72.69% on dependency structures and 41.31% on relation types. It is important to emphasize that the pre-annotation results served only as a reference for annotators and were not decisive. Annotators primarily made judgments based on watching video clips and text content independently. During the pre-annotation correction phase, annotators made an average of 3.8 modifications to the dependency structure of each dialogue segment and an average of 7.4 adjustments to the relation types.

To further verify that pre-annotation did not introduce systematic bias, we conducted a controlled experiment: selecting 50 dialogue samples, two groups of annotators independently completed the same annotation tasks, with one group correcting pre-annotations and the other annotating from scratch. The inter-group Kappa value reached 0.90 for dependency structure and 0.58 for relation type, almost consistent with the overall annotation reported in section 3.3. Given that nearly 82% of dependency relations occur between adjacent utterances, the pre-annotation model more easily captures short-distance dependencies, thereby significantly reducing annotators' workload on short-distance dependency structures.

As shown in Figure 2, the model most frequently misclassifies *Contrast* as *Correction* (76 instances). Both involve responsive expressions to preceding content. However, *Contrast* emphasizes different viewpoints or situations while *Correction* explicitly indicates modifications of previous information. The model struggles to accurately capture these subtle semantic distinctions. Another significant error is misclassifying *Clarification_Question* as *Q-Elab* (72 instances), further demonstrating the inadequacy of text-based models in handling semantically similar relation types. At the same time, we note that there are non-negligible label distribution differences between STAC and DraDDP, which may also affect the performance of cross-domain pre-annotation. For example, in STAC, *Q-Elab* is more frequent than *Clarification_Question*, whereas the opposite pattern holds in DraDDP; such a distribution mismatch may also contribute to certain confusion patterns.

Nevertheless, as shown in Figure 3, the model performs well on structurally regular discourse relations (e.g., *Question-Answer Pair (QAP)* and *Clarification_Question*). This indicates that LLMs can



(a) Dependency Relation Distance (b) Speaker Statistics (c) Dialogue Length Statistics

Figure 4: Data analysis of the DraDDP dataset.

identify regular discourse patterns but still show deficiencies in distinguishing semantically similar categories, highlighting the importance of introducing multimodal information for improving dialogue discourse parsing performance.

In summary, pre-annotation can effectively reduce annotators' workload on short-distance dependency structures and regular discourse relations without compromising annotation quality.

4 Dataset Analysis

To gain deeper insights into the feature distribution and complexity of the DraDDP dataset, we conducted statistical analysis across four dimensions: dependency distance, dialogue participant scale, dialogue length, and label distribution.

4.1 Dependency Distance

Dependency distance represents the positional span between the "current utterance" and its "parent utterance", serving as a key indicator of dialogue semantic coherence and structural complexity. As shown in Figure 4(a), 81.99% of dependencies have a distance of 1, meaning most dependencies occur between adjacent utterances. Dyadic dialogues in DraDDP exhibit a higher proportion of short-distance dependency relations than multi-party dialogues (see Appendix C). This suggests that as the number of participants increases, the discourse structure becomes less localized and more prone to topic branching and shifts.

4.2 Speaker Statistics

We analyze the speakers in each dialogue segment. As shown in Figure 4(b), single/two-party dialogues account for 37.98%, 3-6 person conversations account for 56.97% (the highest proportion), and conversations with more than 6 participants account for 5.05%. This aligns with the setting of *Friends*, which centers on the daily interactions of 6 core protagonists, making 3-6 person conversations the predominant dialogue type. The increase in participants leads to models needing to traverse more dialogue history turns to capture authentic discourse dependency structures (see Appendix C), significantly increasing parsing difficulty.

4.3 Dialogue Length Statistics

Dialogue length determines the span of semantic context and serves as an important metric for evaluating models' long-text modeling capabilities. As shown in Figure 4(c), the distribution across different dialogue length intervals is relatively uniform, with an average dialogue length of 12.88. This uniform distribution is beneficial for effective model training in different dialogue length scenarios.

4.4 Label Distribution

Appendix A presents the distribution of discourse relation types across DraDDP and existing datasets. The *Question-Answer Pair* relation dominates in multi-party datasets (DraDDP and STAC) and accounts for 9.4% in dyadic MODDP, indicating that question-answer exchanges serve as core communication patterns for introducing topics and exchanging information. Compared to STAC where the top two relations account for 41.8%, DraDDP (34.6%) and MODDP (32.1%) show more balanced distributions. This difference stems from data sources: STAC's game scenario concentrates relations on question-answer and comment patterns, while DraDDP and MODDP from TV dramas encompass richer social interactions and emotional expressions. Additionally, multimodal cues (tone, facial expressions) help identify non-core relations that are difficult to understand through text alone (details in section 6.3 and section 6.5), further promoting label balance.

Category	Model	DraDDP		MODDP	
		Link	Link&Rel	Link	Link&Rel
Unimodal	RLTST	76.94	40.10	90.76	41.41
	BERTLine	78.65	45.82	90.73	40.96
	MODDP_{T}	83.31	45.06	**92.46**	43.66
	LLaMIPa	84.71	53.39	91.32	53.05
	LLaMIPa†	**85.03**	54.58	91.55	53.68
	LLaMIPa‡ (Qwen2.5)	84.14	53.55	91.26	52.82
Multimodal	$\text{MODDP}_{\text{Multi}}$	83.65	46.06	90.90	48.05
	LLaMIPa‡ (Qwen2.5-VL)	84.15	53.15	91.33	51.11
	LLaMIPa‡ (Qwen2-Audio)	84.90	**55.09**	92.43	**54.88**
	LLaMIPa‡ (Qwen2.5-Omni)	84.55	53.34	91.46	52.27

Table 2: Experimental results (F1-score) on DraDDP and MODDP test sets. †: Without concatenating dialogue structure history previously predicted by the model. ‡: Built upon the LLaMIPa† framework with the backbone model replaced by Qwen-series (7B). Qwen2.5-VL: Text+Video; Qwen2-Audio: Text+Audio; Qwen2.5-Omni: Text+Video+Audio.

Category	Train	Dev	Test	Total
Dialogues	345	75	75	495
Utterances	4,447	962	965	6,374
Avg. Utt/Dialog	12.89	12.83	12.87	12.88
Avg. Utt Length	9.54	10.19	9.59	9.64

Table 3: Data statistics of the DraDDP dataset.

5 Approach

5.1 Problem Definition

Given a dialogue sequence $D = \{U_0, U_1, \ldots, U_n\}$, where each utterance U_i contains the textual modality U_i^t, the audio modality U_i^a, and the video modality U_i^v, our goal is to predict the utterance dependency graph $G = \{(U_i, U_j : R_{ij}) \mid i < j, R_{ij} \in \mathcal{R}\}$, where $(U_i, U_j : R_{ij})$ represents a dependency arc from utterance U_i to U_j with the relation type R_{ij}, and $\mathcal{R}$ is the predefined set of relation types in Appendix A.

5.2 Benchmarks

We employed four advanced dialogue discourse parsing systems to validate DraDDP's effectiveness: **RLTST** (Fan et al., 2023), integrates multi-task learning with reinforcement learning to address data sparsity; **BERTLine** (Bennis et al., 2023), fine-tunes BERT to encode EDU pairs and handle multi-parent structures; **MODDP** (Gong et al., 2024) fuses text, visual, and audio features via cross-modal attention with context-aware modules; and **LLaMIPa** (Thompson et al., 2024a) performs incremental parsing by incorporating historical discourse structures during fine-tuning LLaMA3 (details in Appendix D).

Moreover, we introduced two variants of LLaMIPa: **LLaMIPa**† removes the historical structure concatenation mechanism from the original LLaMIPa. **LLaMIPa**‡ adopts the LLaMIPa† framework while replacing the backbone LLaMA3 (8B version) with Qwen-series (7B version).

6 Experimentation

6.1 Experimental Setup

Data As shown in Table 3, we randomly split DraDDP proportionally into training, development, and test sets, and evaluate the above benchmarks on both DraDDP and MODDP (using the split in Gong et al. (2024)).

Evaluation Metrics Following previous research (Fan et al., 2023; Thompson et al., 2024a; Gong et al., 2024), we use micro F1-score as evaluation metric: **Link-F1** evaluates the accuracy of dependency edge identification; **Link&Rel-F1** requires both edges and relation types to be correct.

Implementation Details All experiments were conducted on a server equipped with two NVIDIA RTX 4090D GPUs. All LLMs were fine-tuned with LoRA using the LLaMA-Factory framework[3], setting the rank to 8 and the scaling parameter to 16. We employed the AdamW optimizer with a learning rate of 1×10^{-4}. The batch size per GPU

[3] https://github.com/hiyouga/LLaMA-Factory

Model	LLaMIPa‡ (Qwen2.5)		LLaMIPa‡ (Qwen2.5-VL)		LLaMIPa‡ (Qwen2-Audio)		LLaMIPa‡ (Qwen2.5-Omni)	
	Link	Link&Rel	Link	Link&Rel	Link	Link&Rel	Link	Link&Rel
$s \leq 2$	88.87	53.20	88.28	55.28	88.8	54.09	89.17	55.05
$2 < s \leq 6$	80.80	53.30	81.60	51.30	83.20	54.90	80.60	48.90
$s > 6$	79.85	55.42	81.77	51.58	87.54	61.19	87.54	59.27

Table 4: Performance on DraDDP with different models and number of speakers.

was set to 1 with gradient accumulation steps of 8. All models were trained for 3 epochs using mixed precision training. For video processing, frames were sampled at 1 fps (up to 16 frames maximum). Audio signals were encoded after converting 16 kHz sampled audio into 80-channel Mel spectrograms. During training, checkpoints were saved every 500 steps, and we retained the model with the best Link&Rel-F1 performance on the development set.

6.2 Experimental Results

Table 2 presents the results on the DraDDP and MODDP datasets. Compared to dyadic dialogues, multi-party dialogues exhibit higher complexity. On the DraDDP dataset, LLaMIPa‡ (Qwen2.5) achieves a Link-F1 of only 84.14%, which is 7.12 lower than on the MODDP dataset, reflecting the inherent challenges posed by the more complex discourse interaction structures in multi-party dialogue scenarios.

Comparing incremental LLaMIPa with non-incremental LLaMIPa†, we observe that removing history concatenation improves Link&Rel-F1 performance by 1.19, indicating that early prediction errors can mislead subsequent parsing decisions. Therefore, we adopt the LLaMIPa† framework for subsequent experiments.

Introducing the audio modality can significantly improve performance. On the DraDDP dataset, LLaMIPa‡ (Qwen2-Audio) achieves a 1.54 improvement in Link&Rel-F1 over LLaMIPa‡ (Qwen2.5), and a 2.06 improvement on the MODDP dataset. This demonstrates that audio cues (e.g., intonation, emotional tendencies, and voice intensity) can provide fine-grained relation discrimination information beyond text.

In contrast, the visual modality shows limited effectiveness. The reasons are: 1) the current video encoding approach (1 fps sampling) struggles to capture temporal variations in fine-grained facial expressions and body movements; 2) visual information in TV drama scenes contains substantial background noise unrelated to discourse relations. This reveals that in multimodal dialogue discourse parsing, different modalities contribute significantly differently, requiring targeted design of modality fusion strategies. Additionally, we evaluate the performance of the SOTA LLMs including GPT-4o and Claude in Appendix E.

Notably, we further analyze the performance of different modalities across varying speaker counts. As shown in Table 4, we divide the DraDDP test set into three groups based on participant numbers ($s \leq 2$, $2 < s \leq 6$, $s > 6$).

In the $s \leq 2$ scenario, the video modality contributes significantly to relation type identification. LLaMIPa‡ (Qwen2.5-VL) outperforms the pure text model LLaMIPa‡ (Qwen2.5) by 2.08 of Link&Rel-F1. This is because in two-party dialogues, visual cues such as facial expressions and eye contact are more focused and prominent, effectively assisting in relation type judgment.

In the $2 < s \leq 6$ scenario, the audio modality performs better. LLaMIPa‡ (Qwen2-Audio) achieves improvements of 2.40 and 1.60 in Link-F1 and Link&Rel-F1, respectively, compared to the pure text model. This indicates that as dialogue complexity increases, audio features such as intonation variations and emotional coloring begin to provide valuable discriminative information. However, the multimodal fusion model (LLaMIPa‡ (Qwen2.5-Omni)) shows a performance decline to 48.90% in Link&Rel-F1, suggesting that current multimodal LLMs still have significant deficiencies in effectively fusing audio and visual information. The modality fusion introduces noise interference, weakening the model's ability to identify both dependency structures and relation types.

In the complex multi-party dialogue scenario with $s > 6$, the advantage of audio modality becomes more pronounced, achieving improvements of 7.69 and 5.77 over the pure text model, respectively. This indicates that as the number of dialogue participants increases, the audio modality can more effectively help the model identify discourse dependency structure and semantic relation types in complex multi-party interactions by capturing features

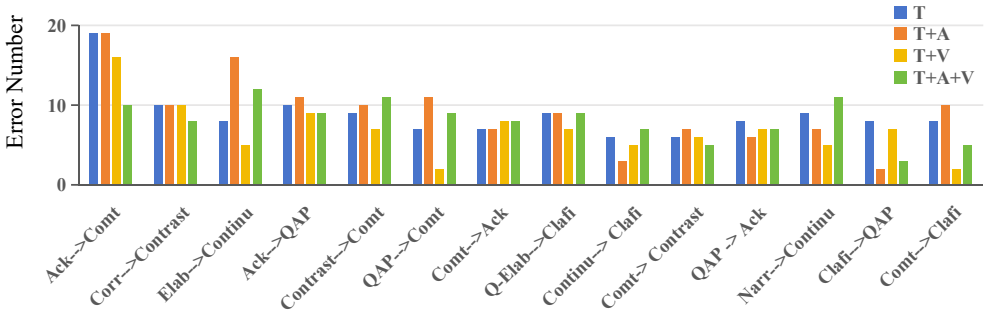


Figure 5: Error pattern statistics of the DraDDP test set under different modalities, where *{X → Y}* indicates misclassifying relation X as Y.

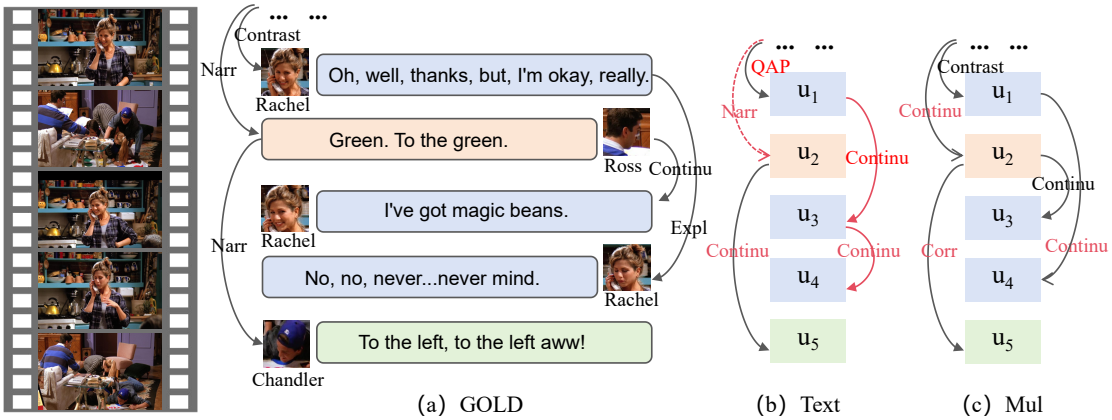


Figure 6: A case study on unimodal (T) and multimodal (T+V+A) dialogue parsing using the DraDDP dataset.

such as speakers' tone variations and emotional intensity.

6.3 Analysis on Relation and Modality

To explore the impact of different modalities on discourse relation types, we analyzed typical error patterns under different modality combinations using the LLaMIPa[‡] (Qwen2.5-Omni) model, as shown in Figure 5.

Audio demonstrates significant advantages in distinguishing emotion-oriented and question-oriented relations. For example, *{Comt → Clafi}* errors decrease by 71.4%, as *Clafi* typically carries confused questioning intonation while *Comt* exhibits clear emotional tendencies (e.g., sarcasm, praise, criticism). Similarly, *{QAP → Comt}* errors decrease by 75%, since answers to questions usually maintain stable intonation whereas comments carry richer emotional coloring.

Visual cues play an important role in identifying certain interactive discourse relations. For example, *{Continu → Clafi}* errors decrease by 50% and *{Clafi → QAP}* by 75%. The visual modality provides crucial cues through facial expressions, gestures, and body language. For example, sustained eye contact and unchanged gaze help identify *Continu*, while confused expressions and questioning gestures help distinguish *Clafi* from *QAP*. However, it also increases certain errors such as *{Elab → Continu}* and *{QAP → Comt}*, as visual cues (e.g., sustained eye contact) may sometimes be misinterpreted as topic continuation signals, and expressions during answers may be mistaken for emotionally evaluative behaviors.

When simultaneously introducing audio and visual information, *{Ack → Comt}* decrease by 47.4%, *{Clafi → QAP}* by 62.5%, and *{Comt → Clafi}* by 35%. These improvements demonstrate the complementary nature of audio-visual fusion. However, certain error types increase, such as *{Elab → Continu}* by 50% and *{Contrast → Comt}* by 22.2%, suggesting that multimodal fusion may cause information interference for specific relation types. For instance, synchronized body movements and speech rhythms may lead to misidentifying elaborative content as simple continuation, while contrastive tones combined with rich facial expressions may be confused with evaluative comments.

Overall, multimodal information provides valuable semantic cues for emotion-related and interaction-related discourse relations, though not universally beneficial. Future research should focus

on refined fusion strategies to maximize modality advantages while minimizing information conflicts.

Modalities	Link	Link&Rel
T	84.67	53.69
V	43.38	22.21
A	47.39	38.83
T+V	83.61	52.97
T+A	84.83	54.76
V+A	50.12	40.39
T+A+V	84.55	53.34

Table 5: Results (F1-score) of different modalities.

6.4 Ablation Study

To explore the contributions of different modalities in the DraDDP dataset, we conducted ablation experiments based on the LLaMIPa‡ (Qwen2.5-Omni) model. As shown in Table 5, under the unimodal setting, the text modality performs best (Link-F1 of 84.67 and Link&Rel-F1 of 53.69), far surpassing the visual and audio modalities. This indicates that text carries the most direct and complete semantic information and serves as the foundation for identifying discourse relations.

Among bimodal combinations, text+audio achieves optimal performance on the test set, with Link&Rel-F1 reaching 54.76, an improvement of 1.07 over pure text, significantly outperforming the text+visual combination. This is mainly because the audio modality more directly conveys speakers' intonation changes, emotional tendencies, and voice intensity, which have strong correlations with semantic relations between utterances (e.g., *Acknowledgement*, *Comment*, *Clarification_Question*, etc.) (see section 6.3 and section 6.5). In contrast, although the visual modality provides cues such as facial expressions and body movements, its correlation with semantic relations is relatively weak, and complex scene backgrounds introduce substantial visual noise, which to some extent interferes with the model's semantic relation judgment.

When fusing all three modalities, model performance is slightly lower than the pure text modality, indicating that under the current model architecture and video processing approach, the introduction of visual information partially offsets the performance gains brought by audio information. As analyzed in section 6.2, the effectiveness of multimodal information exhibits significant differences across various dialogue scenarios. In two-party dialogues ($s \leq 2$), visual cues such as facial expressions and eye contact are more focused and prominent, enabling the video modality to effectively enhance relation type identification. As the number of participants increases ($s > 6$), visual information becomes more scattered and noisier, while audio signals (e.g., speakers' intonation variations and emotional intensity) demonstrate stronger advantages in capturing discourse dependencies and semantic relations in complex multi-party interactions. Therefore, how to dynamically weight different modalities according to dialogue contextual features and more effectively extract spatiotemporal video features will be an important direction for future multimodal dialogue discourse parsing research.

6.5 Case Study

To demonstrate the role of multimodal information in dialogue discourse parsing, we analyzed a typical case from DraDDP. As shown in Figure 6, we compare the LLaMIPa‡ (Qwen2.5-Omni) performance under unimodal and multimodal settings.

In this scenario, the unimodal model fails to identify the sudden topic shift and cannot understand why u_2 would suddenly mention content related to "green" after u_1 expresses personal emotions. With the help of multimodal information (u_3 shows Rachel looking at Ross while speaking), the model identifies the dependency structures of u_2-u_3 and u_1-u_4, where the mentioned *green* and *magic beans* indicate a gaming interaction scenario, while u_1 and u_4 are in a private phone call scenario. This result demonstrates that in multi-party dialogue discourse parsing, effectively integrating multimodal information such as audio and video that contains "interactive intentions" and "scene associations" can enable more precise parsing of complex dialogue semantic structures.

7 Conclusion

This paper fills the research gap in multimodal multi-party English dialogue discourse parsing by constructing the DraDDP dataset. We establish comprehensive benchmarks by evaluating multiple state-of-the-art models, including traditional discourse parsing systems and LLMs. Experiments and analysis validate the importance of multimodal information in understanding dialogue semantics and relation types, providing data support and technical reference for subsequent research in this field.

Limitations

We identify two main limitations of our work. First, despite substantial annotation efforts, the DraDDP dataset remains relatively small due to the inherent complexity of multimodal multi-party dialogue discourse parsing and intensive annotation requirements. The annotation process requires simultaneous consideration of textual content, audio-visual cues, speaker interactions, and discourse structures, which is significantly more time-consuming than traditional text-only annotation tasks. Second, the dataset is sourced from a specific TV series, which may carry domain bias from sitcom-specific dialogue patterns, humor styles, and scripted dialogue dynamics. Future work should focus on expanding dataset scale and diversifying data sources to cover more spontaneous, cross-cultural, and contemporary dialogue scenarios.

Ethical Considerations

Regarding copyright, DraDDP was constructed based on content from Season 1 of *Friends*. The annotation targets dialogues between fictional characters in the show and does not involve the collection of private data from real individuals. We release only the annotated utterances and the annotation guidelines, exclusively for academic research purposes, in accordance with the fair use principle of copyright law. Regarding compensation, all annotators were members of our research group and received research stipends in accordance with institutional standards.

Acknowledgments

The authors would like to thank the three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (Nos. 62276177 and 62376181), and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the stac corpus. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2721–2727.

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Zineb Bennis, Julie Hunter, and Nicholas Asher. 2023. A simple but effective model for attachment in discourse parsing with multi-task learning for relation labeling. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3412–3417.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *arXiv preprint arXiv:2304.14827*.

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.

Yaxin Fan, Feng Jiang, Peifeng Li, and Haizhou Li. 2024a. Uncovering the potential of chatgpt for discourse analysis in dialogue: An empirical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics*, pages 16998–17010.

Yaxin Fan, Peifeng Li, Fang Kong, and Qiaoming Zhu. 2025a. Enhancing multiparty dialog discourse parsing with dynamic task-adaptive graph transformer and difficulty-aware task scheduling. *IEEE Transactions on Neural Networks and Learning Systems*, 36(9):16492–16506.

Yaxin Fan, Peifeng Li, and Qiaoming Zhu. 2024b. Improving multi-party dialogue generation via topic and rhetorical coherence. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3240–3253.

Yaxin Fan, Peifeng Li, and Qiaoming Zhu. 2025b. Improving dialogue discourse parsing through discourse-aware utterance clarification. *arXiv preprint arXiv:2506.15081*.

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*, pages 3808–3814.

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

Shen Gao, Xin Cheng, Mingzhe Li, Xiuying Chen, Jinpeng Li, Dongyan Zhao, and Rui Yan. 2023. Dialogue summarization with static-dynamic structure fusion graph. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 13858–13873.

Chen Gong, DeXin Kong, Suxian Zhao, Xingyu Li, and Guohong Fu. 2024. Moddp: A multi-modal open-domain chinese dataset for dialogue discourse

parsing. In *Findings of the Association for Computational Linguistics*, pages 10561–10573.

Xiulan Hao, Shaohua Wei, Qian Cao, and Xiongtao Zhang. 2024. Emotion recognition in conversations based on discourse parsing and graph attention network. *Telecommunications Science*, 40(5):100–111.

Yuru Jiang, Yu Li, Weikai He, Jie Chen, Yanchao Yu, and Yangsen Zhang. 2023. A new dataset and parsing model for chinese multiparty dialogue discourse structure. In *Proceedings of the 2023 International Conference on Asian Language Processing*, pages 221–227.

Xincheng Ju, Dong Zhang, Suyang Zhu, Junhui Li, Shoushan Li, and Guodong Zhou. 2024. Ecfcon: Emotion consequence forecasting in conversations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2233–2241.

Chuyuan Li, Chloé Braud, Maxime Amblard, and Giuseppe Carenini. 2024a. Discourse relation prediction and discourse parsing in dialogues with minimal supervision. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse*, pages 161–176.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. 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.

Jingyang Li, Shengli Song, Yixin Li, Hanxiao Zhang, and Guangneng Hu. 2024b. Chatmdg: A discourse parsing graph fusion based approach for multi-party dialogue generation. *Information Fusion*, 110:102469.

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

Shannan Liu, Peifeng Li, Yaxin Fan, and Qiaoming Zhu. 2025. Enhancing multi-party dialogue discourse parsing with explanation generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1531–1544.

Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. 2023. Enhanced speaker-aware multi-party multi-turn dialogue comprehension. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2410–2423.

Virgile Rennard, Guokan Shang, Michalis Vazirgiannis, and Julie Hunter. 2024. Leveraging discourse structure for extractive meeting summarization. *arXiv preprint arXiv:2405.11055*.

Kate Thompson, Akshay Chaturvedi, Julie Hunter, and Nicholas Asher. 2024a. Llamipa: an incremental discourse parser. In *Findings of the Association for Computational Linguistics*, pages 6418–6430.

Kate Thompson, Julie Hunter, and Nicholas Asher. 2024b. Discourse structure for the minecraft corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 4957–4967.

Chengrui Wang, Shaoming Ji, and Fang Kong. 2024. Local or global optimization for dialogue discourse parsing. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 149–161. Springer.

Jiahui Xu, Feng Jiang, Anningzhe Gao, Luis Fernando D'Haro, and Haizhou Li. 2024. Unsupervised mutual learning of discourse parsing and topic segmentation in dialogue. *arXiv preprint arXiv:2405.19799*.

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*, pages 7395–7408.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. 2022. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1688–1697.

Nan Zhao, Haoran Li, Youzheng Wu, and Xiaodong He. 2022. 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.

A Discourse Relations

Table 6 presents the 16 discourse relations used in our annotation. For each relation, we provide its definition and distribution across the DraDDP, STAC, and MODDP datasets. The percentages in Table 6 reveal significant variations in how frequently different relations appear, reflecting the unique conversational focuses of each dataset. Notably, *Question-Answer Pair* relation dominates in DraDDP and STAC (17.8% and 24.2%, respectively), but is far less common in MODDP (9.4%), suggesting the latter involves fewer direct Q&A exchanges. Conversely, MODDP shows a higher prevalence of the *Elaboration* (15.7%) and *Continuation* (9.3%) relations, indicating a stronger tendency toward extended, descriptive turns. *Acknowledgement (Ack)* is prominent in both DraDDP and STAC (9.6% each) but minimal in MODDP (2.7%), potentially pointing to differences in interactive feedback or social rapport. Furthermore, the high frequency of *Clarification_Question* in DraDDP (16.8%) compared to the other two datasets highlights its distinctive focus on resolving misunderstandings.

B Analysis on Controversial Cases

To illustrate the complexity of discourse relation annotation and the challenges faced by annotators, we selected several controversial cases from the annotation process for analysis. These cases demonstrate the subtle boundaries between different discourse relation types and highlight the difficulty of achieving perfect agreement among annotators.

As shown in Figure 7, although annotators reached consensus on the dependency structures between utterances, they disagreed on the specific classification of relation types. Regarding the relation between utterances u_1 and u_0, some annotators classified it as *Comment*, believing that u_1 expressed Joey's evaluative opinion about Monica's situation. Other annotators tended to mark it as *Contrast*, considering the contrasting relation between romantic partners and colleagues. After thorough discussion, we ultimately classified this relation as *Correction*, because u_1 actually corrects Monica's description of her relationship status, correcting the nature of her relationship with colleagues to a romantic one. For the relation between utterances u_2 and u_1, some annotators considered this to be *Elaboration*, interpreting Chandler's question as providing specific details about the "certain mistakes" mentioned in u_1. Other annotators argued this was a *Clarification_Question*, believing that u_2 seeks to clarify the specific meaning of the "certain mistakes" Joey mentioned. We ultimately chose the *Clarification_Question* label, because the primary purpose of u_2 is to clarify the nature of the problem Joey implied, similar to u_3.

The above examples show that although each type of discourse relation has clear definitional distinctions, in real daily conversations, they often lack clear boundaries and have semantic overlaps and ambiguous areas. This phenomenon illustrates the difficulty of the annotation task and the complexity of multimodal dialogue discourse structure parsing. In Table 9, we have identified five groups of discourse relations that are easily confused during the annotation process and clarified their core criteria for discrimination, aiming to provide a reference for subsequent work.

C Impact of Participant Numbers on Dependency Distance

To comprehensively validate the relation between participant numbers and dialogue complexity, we analyzed the correlation between dependency distance and speaker count. As shown in Table 7, we divided the dialogues in the DraDDP dataset into three groups based on the number of speakers ($s \leq 2$, $2 < s \leq 6$, $s > 6$) and calculated the distribution of different dependency distances (d) within each group.

With the increase in speaker count, the proportion of local dependencies gradually decreases, while the proportion of long-distance dependency increases significantly. This trend fully validates the characteristic that topic branching and switching occur more frequently in multi-party dialogues. This means that in scenarios with more participants, models need to traverse more dialogue history turns to capture authentic discourse dependency structures, thereby significantly increasing the difficulty of discourse relation identification and parsing.

D Baselines

We employed four advanced dialogue discourse parsing systems (RLTST (Fan et al., 2023), BERT-Line (Bennis et al., 2023), MODDP (Gong et al., 2024), and LLaMIPa (Thompson et al., 2024a)) to validate the effectiveness and applicability of the DraDDP dataset.

Discourse Relation	Definition	DraDDP (%)	STAC (%)	MODDP (%)
Question-Answer Pair (QAP)	Arg2 is the answer to the question raised by Arg1.	17.8	24.2	9.4
Clarification_Question (Clafi)	Arg2 provides clarification for Arg1.	16.8	2.5	6.4
Comment (Comt)	Arg2 expresses a viewpoint or evaluation of Arg1.	14	17.6	16.4
Acknowledgement (Ack)	Arg2 shows approval or acknowledgement for Arg1.	9.6	9.6	2.7
Contrast (Contrast)	Arg1 and Arg2 differ on a shared theme.	7.6	4.7	6.6
Elaboration (Elab)	Arg2 elaborates on Arg1 in detail.	6.8	8.3	15.7
Correction (Corr)	Arg2 corrects Arg1.	5.9	2	0.2
Continuation (Continu)	Arg2 is a continuation of the content of Arg1.	4.8	9.4	9.3
Explanation (Expl)	Arg2 provides an explanation for Arg1.	3.5	4.2	5.2
Narration (Narr)	Arg2 narrates or describes Arg1.	3.5	1.2	1.8
Parallel (Paral)	Arg1/2 have alike semantic structures and themes.	2.2	2	2.3
Q-Elab (Q-Elab)	Arg1 is a question, and Arg2 elaborates on it in detail.	2.1	5.7	6.3
Result (Resl)	Arg2 is the result of the situation described in Arg1.	2.1	5.5	6.8
Conditional (Condi)	Arg2 is the condition for Arg1.	1.6	1.2	3.2
Alternation (Alte)	Arg1 and Arg2 represent interchangeable situations.	1	1.4	4.5
Background (Backg)	Arg2 provides background information for Arg1.	0.7	0.6	3.2

Table 6: Definition and proportion of discourse relations in DraDDP, MODDP, and STAC datasets.

Speaker Count	D=1	D=2	D=3	D=4	D≥5
$s \leq 2$	88.6%	9.0%	1.2%	0.8%	0.3%
$2 < s \leq 6$	80.2%	13.5%	3.0%	1.5%	1.8%
$s > 6$	78.1%	12.6%	3.3%	2.1%	3.9%

Table 7: Distribution of dependency relation distances across different speaker counts.

RLTST (Fan et al., 2023): A multi-task learning-based dialogue discourse parser that integrates complementary information from dialogue discourse parsing and addressee recognition, alleviating data sparsity issues without requiring additional manual annotations. Its core mechanism employs reinforcement learning to filter samples with significant gains in addressee recognition, and utilizes a task-aware structure transformer to distinguish between task-shared and task-private structures, avoiding mutual interference.

BERTLine (Bennis et al., 2023): It fine-tunes BERT to directly encode EDU pairs and employs simple linear layers to perform dependency structure predictions. This model is capable of handling multi-parent structures and can effectively capture complex dependencies found in multi-turn dialogues.

MODDP (Gong et al., 2024): It is a discourse parser for multimodal open-domain Chinese dialogues, which employs RoBERTa, ViT, and Wav2Vec2.0 to encode textual, visual, and audio modality features respectively, and achieved multimodal information interaction through cross-modal multi-head attention mechanisms. This model introduces context-aware and speaker-aware dialogue interaction modules to enhance overall understanding of dialogue structure.

LLaMIPa (Thompson et al., 2024a): It is an incremental parser based on fine-tuned LLaMA3. When processing each new utterance unit, the model not only considers the current textual content but also incorporates previously predicted discourse structure information to synchronously perform dependency structure and relation prediction, representing a strong baseline model in the dialogue discourse parsing field.

E Evaluation on Different LLMs

To further evaluate the performance of the DraDDP dataset on current mainstream LLMs, we selected Claude-Sonnet-4[4], GPT-4o[5], and Qwen3-235B[6] as representative unimodal text large models, while selecting Gemini 2.5 Pro[7] and Qwen3-VL-235B[8] as representative multimodal large models. For model input, we adopted a prompt template consistent with the "Vanilla" type in Fan et al. (2024a), which only provides textual dialogue content. For multimodal models, we additionally provided audio-visual content and specified in the template: "*Analyze the following dialogue by integrating textual content with multimodal cues from the video, including but not limited to speakers' facial expressions, body language, tone variations, and emo-*

[4] https://www.anthropic.com/claude/sonnet
[5] https://openai.com/
[6] https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507
[7] https://gemini.google.com/
[8] https://huggingface.co/Qwen/Qwen3-VL-235B-A22B-Instruct

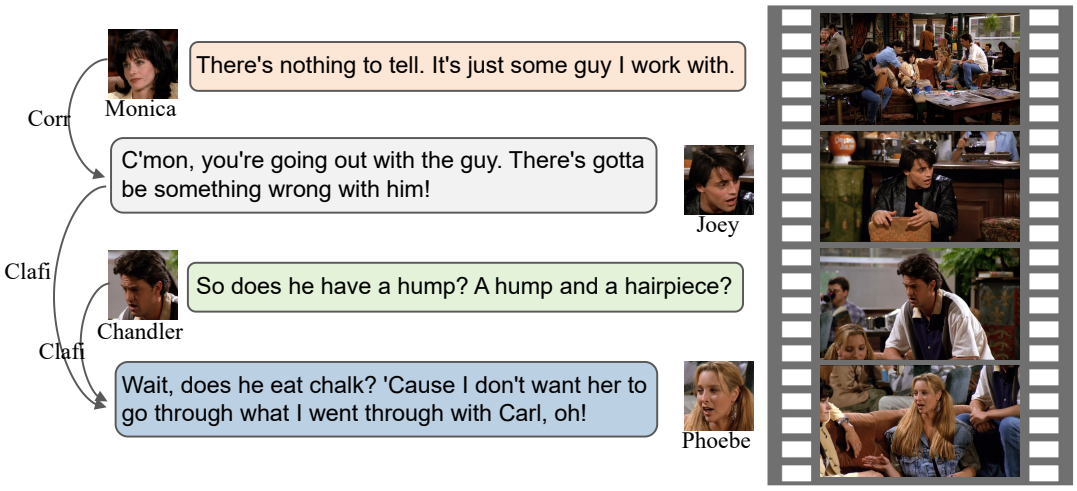


Figure 7: Controversial annotation cases in the DraDDP dataset.

Category	Model	Link	Link&Rel
Unimodal	Claude-Sonnet-4	71.46	23.18
	GPT-4o	78.11	35.87
	Qwen3-235B	71.48	42.18
Multimodal	Gemini 2.5 Pro	68.17	21.69
	Qwen3-VL-235B	71.64	37.71

Table 8: Baseline performance of LLMs on the DraDDP dataset (F1-score).

tional signals. Identify discourse dependency structures and relation types, then complete the annotation according to the specified format.".

The experimental results are shown in Table 8. We observe that even on the most advanced large models currently available, the parsing performance on DraDDP remains significantly lower than specialized models, particularly showing poor performance in Link&Rel-F1. This finding echoes the research conclusions of Chan et al. (2023) and Fan et al. (2024a), further validating the extremely high complexity of multi-party dialogue discourse parsing.

In the pure text modality, GPT-4o performs best on linking (Link at 78.11%), while Qwen3-235B performs relatively better on relation (Link&Rel-F1 at 42.18%), reflecting the differences in models' capabilities to capture dialogue structure and semantic relations.

Notably, the multimodal models Gemini 2.5 Pro and Qwen3-VL-235B did not demonstrate significant advantages over the pure text models on this task. This phenomenon may stem from the following reasons: 1) general-purpose multimodal large models lack targeted training on the specific task of dialogue discourse parsing; 2) the models may introduce semantic interference when fusing text-video cross-modal information; 3) current large models' capabilities in deep fusion of multimodal information and understanding of complex dialogue structures remain immature.

Type	Distinction	Example
Clafi	*Clarification_Question* is to ask about the "current situation, details, or ambiguities of an existing statement" to confirm facts or eliminate information bias.	u_0: There's nothing to tell. It's just some guy I work with. u_1: C'mon, you're going out with the guy. There's gotta be something wrong with him!
Q-Elab	*Q-Elab* is to further inquire or expand on an "already raised question" to obtain more comprehensive information, focusing on the "question itself".	u_0: Who's Paul? u_1: Paul, the wine guy? Paul?
Continu	*Continuation* is generally a "parallel extension" to the current topic, with no strict order, which is just horizontal expansion of the topic.	u_0: I say push her down the stairs. u_1: Push her down the stairs! Push her down the stairs! Push her down the stairs!
Narr	*Narration* is a "vertical expansion" of the current topic, with a clear implicit sequence, process, or logical connection, presenting complete events or information.	u_0: Yeah, right! See, he gave up something, but then he got those magic beans. u_1: And then he woke up, and there was a big plant outside his window, full of possibilities and stuff...
Expl	*Explanation* provides a clear reason, basis, or reasoning process for a "certain viewpoint, phenomenon, or result", focusing on "answering why".	u_0: Oh my god, oh, you guys are great. u_1: We all chipped in.
Backg	*Background* has a weaker causal relationship, providing relevant background information (such as time, scene, premise, etc.) for the "current topic".	u_0: I'm smoking. I'm smoking, I'm smoking. u_1: Oh, I can't believe you! You've been so good, for three years!
Corr	*Correction* directly points out and gives the correct content for "errors in the other party's statement", focusing on "correcting errors", can be replaced with "no no no".	u_0: ...That's it. I'm getting cigarettes. u_1: No no no!
Contrast	*Contrast* highlights the difference or opposition, focusing on "presenting differences", not involving "error correction".	u_0: How does she do that? u_1: I cannot sleep in a public place, libraries, airplanes, movie theaters....
Ack	The two types have a "containment and independence" relationship. *Acknowledgement* expresses clear agreement with the other's statement/viewpoint, and is very brief.	u_0: Would you look at her? She is so peaceful. u_1: yeah.
QAP	*Question-Answer Pair* may include a brief acknowledgement and then add specific information. In such cases, "yes" serves merely as a prefix to the answer, and the whole is still classified as *Question-Answer Pair*.	u_0: Alright. Phoebe? u_1: Okay, okay. If I were omnipotent for a day, I would want, um, world peace, no more hunger, good things for the rain forest. And bigger boobs.

Table 9: Distinction between easily confusable discourse relations.