MIntRec2.0: A Large-scale Benchmark Dataset for Multimodal Intent Recognition and Out-of-scope Detection in Conversations

Anonymous authors
Paper under double-blind review

Abstract

Multimodal intent recognition poses significant challenges, requiring the incorporation of non-verbal modalities from real-world contexts to enhance the comprehension of human intentions. However, most existing multimodal intent benchmark datasets are limited in scale and suffer from difficulties in handling out-of-scope samples that arise in multi-turn conversational interactions. In this paper, we introduce MIntRec2.0, a large-scale benchmark dataset for multimodal intent recognition in multi-party conversations. It contains 1,245 high-quality dialogues with 15,040 samples, each annotated within a new intent taxonomy of 30 fine-grained classes, across text, video, and audio modalities. In addition to more than 9,300 in-scope samples, it also includes over 5,700 out-of-scope samples appearing in multi-turn contexts, which naturally occur in real-world open scenarios, enhancing its practical applicability. Furthermore, we provide comprehensive information on the speakers in each utterance, enriching its utility for multi-party conversational research. We establish a general framework supporting the organization of single-turn and multi-turn dialogue data, modality feature extraction, multimodal fusion, as well as in-scope classification and out-of-scope detection. Evaluation benchmarks are built using classic multimodal fusion methods, ChatGPT, and human evaluators. While existing methods incorporating nonverbal information yield improvements, effectively leveraging context information and detecting out-of-scope samples remains a substantial challenge. Notably, powerful large language models exhibit a significant performance gap compared to humans, highlighting the limitations of machine learning methods in the advanced cognitive intent understanding task. We believe that MIntRec2.0 will serve as a valuable resource, providing a pioneering foundation for research in human-machine conversational interactions, and significantly facilitating related applications.

1 Introduction

Understanding human intentions in multimodal scenarios holds significant research importance and has broad applications, such as human-computer interaction (Xu, 2019), intelligent transportation system (Kaffash et al., 2021), and medical diagnosis (Tiwari et al., 2022; Moon et al., 2022). For instance, perceiving user tones, expressions, and body language enables better capture of user needs in intelligent customer systems. This also leads to more personalized, efficient, and natural interactions (Luo et al., 2022). While there emerge numerous multimodal language datasets in recent years, particularly in multimodal sentiment analysis and emotion recognition (Li et al., 2019; Chudasama et al., 2022; Hu et al., 2022b), few datasets provide high-quality multimodal intent resources, which significantly hampers related research. Zhang et al. (2022) pioneered this area by formulating intent taxonomies in multimodal conversational scenarios and providing 2,224 annotated utterances with text, video, and audio information. However, it has three major limitations: First, its scale is relatively small compared to other multimodal datasets (Zadeh et al., 2018b; Poria et al., 2019), leading to potential overfitting and impacting the generalization ability. Second, it only includes utterances from single-turn dialogues, neglecting context and multi-party information. Third, it fails to consider out-of-scope utterances, which commonly occur in dialogue systems (Larson et al., 2019) and are crucial for improving system robustness.

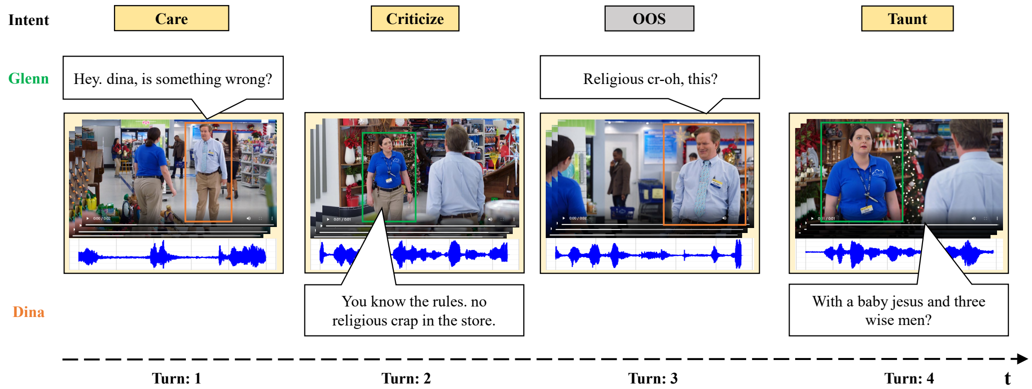


Figure 1: An example from the MIntRec2.0 dataset. More examples are provided in the Appendix A.

To address these issues, we propose MIntRec2.0, a large-scale multimodal multi-party benchmark dataset that comprises 1,245 high-quality dialogues, totaling 12.3 hours. A representative sample is depicted in Figure 1. The construction of this dataset involves four main steps. Initially, raw videos from three TV series are collected and segmented into utterance-level portions based on timestamps. These segments are then manually grouped into dialogues in alignment with the conversational scenes and events. Subsequently, each utterance is annotated with speaker identity information to leverage specific contextual information. Following this, we propose a new intent taxonomy incorporating 30 fine-grained intent classes. An *OOS* tag is also added to identify utterances that do not belong to any known classes, a phenomenon commonly occurred in real-world, open-ended scenarios. Lastly, six experienced workers annotate each piece of data using text, video, and audio information. The final dataset contains 9,304 in-scope and 5,736 out-of-scope samples.

We develop a general framework for multimodal intent recognition and out-of-scope detection within single-turn and multi-turn conversations. First, data inputs are organized at both utterance and dialogue levels, where the latter retrieves all the context information corresponding to the speaker in the current dialogue turn. Secondly, we extract text, video, and audio features for each utterance. For multi-turn dialogues, context information is concatenated to the utterance in the current turn using a special token as a separator. Third, we perform multimodal fusion on the extracted features. Specifically, we employ two strong multimodal fusion methods (Tsai et al., 2019; Rahman et al., 2020) to leverage nonverbal information by capturing cross-modal interactions. In the training stage, in addition to the multimodal fusion loss, cross-entropy loss is applied under the supervision of hard and soft targets for learning in-scope and out-of-scope data, respectively. During inference, a threshold-based method (Shu et al., 2017) is adopted to both identify high-confidence in-scope and detect low-confidence out-of-scope samples. Experimental results demonstrate that using multimodal information can effectively improve in-scope intent recognition accuracy and enhance out-of-scope detection robustness. Furthermore, we evaluate ChatGPT and human performance under a challenging setting with few-shot samples as prior knowledge. The results reveal a significant performance gap of over 30% absolute scores between large language models (LLMs) and humans. Humans achieve the state-of-the-art benchmark performance of 71% accuracy with merely 7% of the training data, indicating this dataset is extremely challenging for existing machine learning methods.

Contributions. (1) This paper presents MIntRec2.0, the first large-scale multimodal multi-party conversational intent dataset. This dataset provides detailed annotations for both intent and speaker identity for each utterance within multimodal contexts and enables out-of-scope detection in open-world scenarios. (2) We establish a universal framework for in-scope classification and out-of-scope detection, applicable to both single-turn and multi-turn conversations, and introduce strong benchmark baselines. (3) Extensive experiments demonstrate the effectiveness of leveraging multimodal information in intent recognition. However, considerable opportunities for enhancement persist in existing methods when compared with human performance, highlighting the challenges inherent in high-level cognitive intent recognition tasks and underscoring the value of this dataset in advancing related research. This dataset will be released under the CC BY-NC-SA 4.0 license, and codes will be publicly available as open source. A portion of the data are accessible in supplementary materials.

Table 1: Comparison of the MIntRec2.0 dataset with previous intent datasets. #I and #U represent the number of intent classes and utterances. Conv. denotes the conversational nature of the dataset. OOS and Multi-Party indicate the inclusion of out-of-scope examples and multiple speakers per dialogue, respectively. T, V, and A represent text, video, and audio information.

Datasets	#I	#U	Conv. Scenes	Conv. Type	OOS	Multi-Party	T	V	A
ATIS (Tür et al., 2010)	17	6,371	✓	Single-turn	✗	✗	✓	✗	✗
Snips (Coucke et al., 2018)	7	14,484	✓	Single-turn	✗	✗	✓	✗	✗
CLINC150 (Larson et al., 2019)	150	23,700	✓	Single-turn	✓	✗	✓	✗	✗
MDID (Kruk et al., 2019)	7	1,299	✗	-	✗	✗	✓	✓	✗
Intentonomy (Jia et al., 2021)	28	14,455	✗	-	✗	✗	✗	✓	✗
MIntRec (Zhang et al., 2022)	20	2,224	✓	Single-turn	✗	✗	✓	✓	✓
MIntRec2.0	30	15,040	✓	Multi-turn	✓	✓	✓	✓	✓

2 RELATED WORK

This section provides a brief overview of the existing literature in benchmark datasets, multimodal fusion methods, and multimodal multi-turn conversations. Further related works focusing on video understanding and out-of-scope intent detection are detailed in Appendix B.

Benchmark Datasets. Intent recognition is a substantial task in NLP, supported by numerous benchmark datasets. These datasets can be broadly categorized into two branches. The first branch, originating from task-oriented dialogues, includes datasets like ATIS (Tür et al., 2010), SNIPS (Coucke et al., 2018), CLINC150 (Larson et al., 2019), BANKING77 (Casanueva et al., 2020). Notably, CLINC150 incorporates out-of-scope data to test system robustness. SIMMC 2.0 (Kottur et al., 2021) is a multimodal dataset focusing on the shopping domain but lacks intent annotations. The second branch derives from open-ended dialogues, represented by multi-turn dialogue datasets such as DailyDialog (Li et al., 2017) and SWBD (Godfrey et al., 1992). However, these datasets primarily offer dialogue acts and are less suited for specific intent classes. Recent trends show a growing interest in multimodal language datasets for both single-turn (Zadeh et al., 2016; 2018b; Yu et al., 2020) and multi-turn dialogues (Busso et al., 2008; Poria et al., 2019; Saha et al., 2020). A major difference between these two branches is that the former usually stems from human-computer interactions, while the latter originates from human-human interactions (detailed in Appendix C). Some studies have also explored visual or multimodal intents using image modality (Jia et al., 2021; Kruk et al., 2019). MIntRec (Zhang et al., 2022) stands as the first multimodal intent recognition dataset for open-ended dialogues. MIntRec2.0 significantly expands in scale from 2,224 to 15,040 utterances and is designed to handle both out-of-scope utterances and multi-turn dialogues. A comparison between MIntRec2.0 and other benchmark intent datasets is presented in Table 1.

Multi-modal Fusion Methods. Multimodal fusion presents prosperous development in multimodal language understanding. Early methods aim to learn cross-modal relations and single-modal properties (Fukui et al., 2016; Zadeh et al., 2017; 2018a; Hazarika et al., 2020) or efficient multimodal representations (Liu et al., 2018). MulT (Tsai et al., 2019) designs an effective crossmodal attention module to learn adaptations across different modalities. MAG-BERT (Rahman et al., 2020) integrates nonverbal information into pre-trained language models using a multimodal adaptation gate. Deep-HOSeq (Verma et al., 2020) combines LSTMs and CNNs to capture intra-modality and inter-modality dynamics, incorporating temporal-granularity information. MBT (Nagrani et al., 2021) restricts cross-modal information flow through tight fusion bottlenecks, facilitating the connection of relevant inputs in each modality. We also explore state-of-the-art methods in multimodal sentiment analysis (MSA), such as Self-MM (Yu et al., 2021) and MMIM (Han et al., 2021). However, these methods rely on specific sentiment properties (e.g., polarity) that are not applicable to our task.

Multimodal Multi-turn Conversations. Leveraging multimodal information is a hot topic in multi-turn conversations (Ghosal et al., 2019; Majumder et al., 2019; Ghosal et al., 2020a). For instance, DialogueRNN (Majumder et al., 2019) uses GRU networks to track important temporal information, including the history of speaker states and global states. MM-DFN (Hu et al., 2022a) proposes a graph-based dynamic fusion module to reduce historical redundancy while tracking the history of speaker states. Another approach is to construct multimodal fusion networks to integrate con-

Table 2: Expanded intent classes in the MIntRec2.0 dataset with brief interpretations.

Intent Categories		Interpretations
Express emotions or attitudes	Doubt	Convey a sense of mistrust or uncertainty regarding someone or something (e.g., questioning with an expression of disbelief).
	Acknowledge	Indicate comprehension or agreement (e.g., using affirming words such as alright, well).
	Refuse	Show unwillingness or rejection (e.g., using negative words to decline an offer or request).
	Warn	Alert to potential dangers or risks (e.g., signaling alarm with a serious expression and tone).
	Emphasize	Highlight the importance or significance of something (e.g., speaking with stress and a determined attitude).
Achieve goals	Ask for opinions	Request others' views or thoughts on a particular matter (e.g., asking for others' perspectives).
	Confirm	Validate or ascertain the truth or accuracy of something (e.g., affirming certainty without raising doubts).
	Explain	Provide an elaborate account or clarification (e.g., using explanatory and causal words such as because).
	Invite	Offer someone to participate in an activity or event (e.g., asking someone to join in activities like going out).
	Plan	Organize or schedule an event or action (e.g., deliberating on schedules and making commitments for the future).

textual information between different modalities, such as M2FNet (Chudasama et al., 2022) and MMGCN (Hu et al., 2021). However, modeling temporal contextual information with multimodal fusion representations does not yield good results (see Appendix D). Therefore, we propose a simple baseline that concatenates the context information of inputs before multimodal fusion.

3 The MIntRec2.0 Dataset

Data Sources & Dialogue Division. First, we collect raw videos from three different TV series: Superstore, The Big Bang Theory, and Friends on YouTube and obtain subtitles from OpenSubtitles. These TV series cover a range of various scenes and topics relevant to daily life (Appendix E). We ensure that the selected videos do not infringe on user privacy and are free from malicious content (Appendix F). The videos are segmented into continuous clips according to timestamps in the transcripts, and we take care to exclude audience signals, such as laugh tracks, to maintain the quality of the dataset. We then organize these clips into a series of dialogues for multi-turn dialogue intent analysis. Specifically, we manually annotate the start and end points of video segments for each dialogue and distinguish different dialogues based on whether they occur in the same scene and episode, as suggested in (Poria et al., 2019). Additionally, we establish a baseline to estimate the utterance boundary in each segmented dialogue (Appendix G).

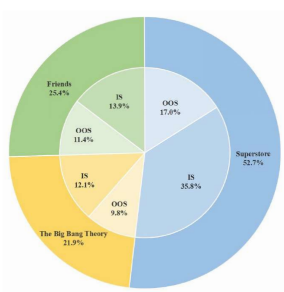


Figure 2: In-scope and out-of-scope data distribution.

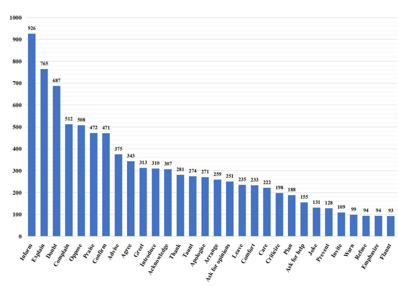


Figure 3: Distribution of in-scope intents in the MIntRec2.0 dataset.

Table 3: Data statistics. # denotes the total number.

# data sources	3
# intents classes	30
# dialogues	1,245
# utterances	15,040
# in-scope utterances	9,304
# out-of-scope utterances	5,736
# words in utterances	118,477
# unique words in utterances	9,524
Average length of utterances	7.9
Maximum length of utterances	46
Average video clip duration	3.0 (s)
Maximum video clip duration	19.9 (s)
Video hours	12.3 (h)

Speaker Information. In multi-turn conversations, we can leverage context information to help analyze the intent conveyed by the speaker in each dialogue turn. However, context information may involve multiple speakers (e.g., there are a total of 51.5% dialogues with more than two speakers). As using context information of speakers is helpful for intent analysis (Ghosal et al., 2020b), we aim to differentiate different speakers in each dialogue and annotate the identities of the speakers. Specifically, we perform annotation of 21, 7, and 6 main characters in Superstore, The Big Bang Theory, and Friends, respectively, which account for 90.4% of the data. The remaining data include other characters with fewer appearances (Refer to Appendix H for statistics of different characters).

Expanded Intent Classes. In this work, we utilize the established intent taxonomy from the MIntRec dataset (Zhang et al., 2022). However, as the dataset primarily focuses on discrete single-turn conversations, and the existing 20 intent classes are insufficient for capturing the diverse range of intents in continuous multi-turn conversations. To address this issue, we conduct a comprehensive analysis of the divided dialogues and collect 10 additional high-frequency intent tags for the two coarse-grained intent classes (i.e., *express emotions or attitudes* and *achieve goals*). Specifically, we add *doubt, acknowledge, refuse, warn, emphasize* to the former category, and *ask for opinions, confirm, explain, invite, plan* to the latter. Interpretations of both the expanded and existing intent categories can be found in Table 2 and Appendix I, respectively. Notably, these newly introduced classes account for 37.3% of the utterances in our dataset, highlighting their significance in intent understanding. The intent taxonomies are highly applicable across various domains, offering considerable promise for real-world applications (Further discussions can be found in Appendix J).

Out-of-scope Utterances. Given that intents usually reside within particular contextual events (Schröder et al., 2014), there inevitably exist some utterances that fall outside the predefined intent categories in continuous conversational interactions, as suggested in (Larson et al., 2019). There are two common types of such utterances. First, there are statements that primarily convey personal views or factual information, which correspond to the *statement-opinion* and *statement-non-opinion* categories, as defined in the 42 dialogue acts (Godfrey et al., 1992). While this type of dialogue act covers a significant proportion of utterances in multi-turn conversations, it provides limited contribution to understanding specific and applicable intents. Second, given the diversity and uncertainty of human intentions, the predefined intent classes cannot encompass all possible intentions in an open-world environment (Zhang et al., 2023), and there may exist utterances falling under open intent classes (e.g., *help, drive person away, wish*). Given the ambiguous boundary in determining specific out-of-scope utterances, we adopt a similar manner as in (Larson et al., 2019) and define them as those that do not belong to any of the existing intent classes. Incorporating these utterances in multi-turn conversations brings us closer to real-world scenarios and presents many practical applications.

Annotation Process. We employ six college students proficient in English for multimodal label annotation. They receive a detailed guidebook explaining interpretations of intents and relevant scenarios and begin annotation only after demonstrating high accuracy with seed examples. The annotators are split into two groups, with each group responsible for a distinct half of the data. A user-friendly annotation platform with a unified database is used to assist their work (details on the platform are in Appendix K). Each annotator assesses the speaker's intention in a video segment by considering text, video, audio, and context information. The relative importance of these modalities during annotation is discussed in Appendix L. Annotators select from 30 predefined intent tags or an *OOS* tag for each utterance. The final label is assigned by majority vote, requiring a consensus of at least two out of three annotators. We assume each utterance expresses a single intent. The reasoning behind not using multi-intent labeling is detailed in Appendix M. Utterances receiving three differing votes are excluded from the dataset to ensure labeling consistency.

Annotation Results. We have successfully collected 1,245 high-quality dialogues to create the MIntRec2.0 dataset. This dataset consists of 9,304 in-scope and 5,736 out-of-scope utterances with multimodal labels. The statistics of the dataset are presented in Table 3. To assess annotation reliability, we calculate the Fleiss's kappa statistics for each of our six annotators to measure interrater reliability. The Fleiss's kappa scores range from 0.66 to 0.70, averaging 0.69. This indicates a level of *substantial* agreement, as defined in (McHugh, 2012). The distribution of the dataset across three different data sources is illustrated in Figure 2. Superstore, The Big Bang Theory, and Friends contribute 53%, 22%, and 25% of the dataset, respectively. Each data source contains between 54.5% and 67.9% of in-scope utterances. The intent distribution of in-scope utterances is depicted in Figure 3, demonstrating a common long-tailed distribution similar to real-world scenarios. As expected, some intents such as *inform, explain, doubt*, and *complain* are more prevalent in daily life, while others like *warn, refuse, emphasize*, and *flaunt* tend to occur less in specific occasions and scenes. To ensure adequate training, each intent class contains more than 90 samples.

4 BENCHMARK FRAMEWORK

This section presents a general benchmark framework, illustrated in Figure 4. It includes data organization, multimodal feature extraction, multimodal fusion, training, and evaluation.

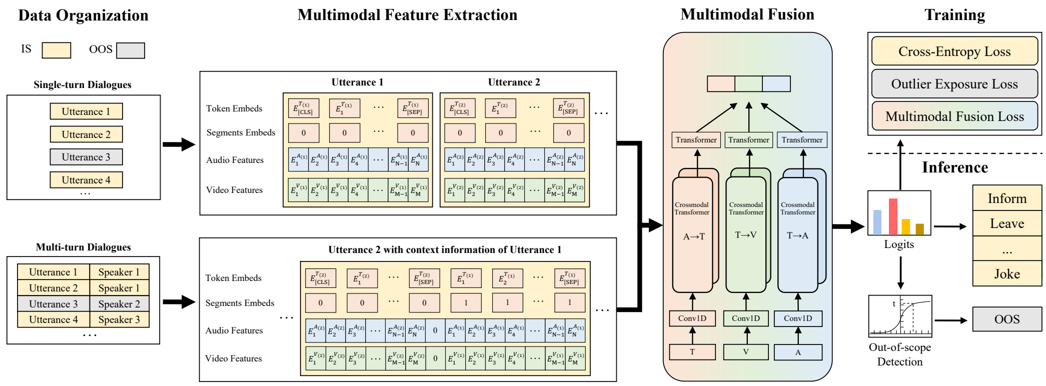


Figure 4: Overview of the benchmark framework for the MIntRec2.0 dataset.

Data Organization. In the case of single-turn dialogues, we utilize the pre-segmented utterance-level samples. Each individual utterance represents a complete turn of dialogue and includes corresponding text, video, and audio information of one speaker. For multi-turn dialogues, we employ well-divided dialogues as described in Section 3. In particular, the utterances within each dialogue are arranged chronologically based on the order in which the speakers take their turn. To further leverage the context of the respective speaker, we attribute the corresponding speaker identity information to each utterance, as suggested in (Poria et al., 2019).

Text Feature Extraction. We select the pre-trained BERT (Devlin et al., 2018) language model as a powerful backbone for processing the text modality, which has demonstrated strong performance when fine-tuned on our dataset. For each text utterance s, we first tokenize it in the required format, i.e., [CLS], $s_1, \cdots, s_n$, [SEP], and then obtain the token embeddings $\mathbf{E}^T \in \mathbb{R}^{L^T \times D^T}$, where L^T is the sequence length, and D^T is the feature dimension.

Video Feature Extraction. Video features are extracted at the frame-level, as suggested in (Yu et al., 2020; Zadeh et al., 2018b). Since video frames often contain multiple individuals, we begin by identifying regions of interest (RoIs) for the speakers, using a sequence of automated procedures. This involves scene detection, object detection (Ren et al., 2015), face detection (Zhang et al., 2017), face tracking, and audio-visual active speaker detection (Tao et al., 2021), as described in (Zhang et al., 2022). This process can generate more than 1,000K high-quality keyframes with speaker bounding boxes in approximately 5 days. Next, we use these annotated RoIs and employ the instance segmentation method, Mask R-CNN (He et al., 2017), pre-trained on the COCO (Lin et al., 2014) dataset to extract visual features. We utilize the well-initialized Swin Transformer (Liu et al., 2021), pre-trained on the ImageNet-1K (Deng et al., 2009) dataset, as the backbone due to its superior vision task performance. We use it to extract feature maps of each keyframe and apply RoIAlign (He et al., 2017) to convert them into fixed sizes using annotated RoIs. Finally, applying average pooling to these feature maps yields the overall RoI feature embeddings $\mathbf{E}^V \in \mathbb{R}^{L^V \times D^V}$.

Audio Feature Extraction. To process the audio modality, we first use the librosa toolkit (McFee et al., 2015) to load the audio waveform data with a sampling rate of 16,000 Hz. Then, we employ WavLM (Chen et al., 2022), a speech pre-trained model to extract audio feature representations. Due to its masked speech prediction and denoising pre-training strategy, it has shown remarkable performance in a wide range of speech tasks, outperforming other powerful speech pre-trained models such as wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021). Notably, it excels in speaker verification and speech separation tasks, which is suitable for conversational scenarios involving multiple speakers. By utilizing WavLM, we acquire audio embeddings $\mathbf{E}^A \in \mathbb{R}^{L^A \times D^A}$.

Incorporating Context Information. In single-turn dialogues, we can directly extract embeddings for text, video, and audio modalities, as mentioned previously. However, in multi-turn dialogues, it is substantial to consider the context information of different modalities to gain a better understanding of the conversation. To address this, we utilize the context information based on different speakers, as suggested in (Majumder et al., 2019; Ghosal et al., 2019). Specifically, for the utterance in the current turn, we first obtain the speaker identity information and then retrieval all the content from the previous dialogue turns corresponding to this speaker, which serves as the context information. Next, we employ a simple and effective method to leverage the context information by concatenating it with the utterance in the current turn. Taking the context information from one turn of utterance as an example, for the text modality, the first sequence comprises all the token embeddings in the current turn: $\mathbf{E}^{T_{(1)}}_{[\mathrm{CLS}]}, \mathbf{E}^{T_{(1)}}_1, \cdots, \mathbf{E}^{T_{(1)}}_2, \mathbf{E}^{T_{(1)}}_{[\mathrm{SEP}]}$. The second sequence comprises the context information. We remove the first token [CLS] and concatenate the remaining embeddings with the first sequence: $\mathbf{E}^{T_{(1)}}_{[\mathrm{CLS}]}, \cdots, \mathbf{E}^{T_{(1)}}_{[\mathrm{SEP}]}, \mathbf{E}^{T_{(2)}}_1, \cdots, \mathbf{E}^{T_{(2)}}_{[\mathrm{SEP}]}$. Besides, we include segment embeddings to aid in understanding the relationships between current and contextual utterances. The segment embeddings for the first and second sequences are encoded as zero and one vectors, respectively, with the same length as the token embeddings. For nonverbal modalities, we insert a one-dimensional zero vector between the feature embeddings of the two sequences to distinguish them. If additional context information is available, such as more contextual utterances, we append each of them to the end of the latest context utterance using the same operation as the second sequence.

Multimodal Fusion. After extracting multimodal features, our goal is to utilize multimodal fusion techniques to capture cross-modal interactions and exploit complementary information from different modalities to further enhance intent recognition capability. Specifically, we use $\mathbf{E}^T$, $\mathbf{E}^V$, and $\mathbf{E}^A$ as inputs and feed them into a multimodal fusion network $\mathcal{F}$ to obtain multimodal representations $\mathbf{z} = \mathcal{F}(\mathbf{E}^T, \mathbf{E}^V, \mathbf{E}^A)$. In this work, we adopt two strong multimodal fusion methods, namely MAG-BERT (Rahman et al., 2020) and MulT (Tsai et al., 2019) as baselines.

Training. Following multimodal fusion, we employ the multimodal representations $\mathbf{z}$ for training. For in-scope samples $\mathbf{z}^{\mathrm{in}} = \{\mathbf{z}_i | y_i \in \mathcal{Y}\}_{i=1}^N$, we perform classification on $\mathbf{z}^{\mathrm{in}}$ using the cross entropy loss $\mathcal{L}_{\mathrm{CE}}$, where N is the number of training samples, and $\mathcal{Y}$ is the set of K known intent labels. For out-of-scope samples $\mathbf{z}^{\mathrm{out}} = \{\mathbf{z}_i | y_i \notin \mathcal{Y}\}_{i=1}^N$, we apply the outlier exposure (OE) (Hendrycks et al., 2018) loss, denoted as $\mathcal{L}_{\mathrm{OE}}$, to distinguish them from the in-scope samples and enhance the model's robustness and its generalization ability for out-of-scope samples. Specifically, we use a uniform distribution over the K known classes as soft targets. The definitions for losses are as follows:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\phi(\mathbf{z}_i^{\mathrm{in}})^{y_i})}{\sum_{j=1}^{K}\exp(\phi(\mathbf{z}_i^{\mathrm{in}})^{j})}, \mathcal{L}_{\mathrm{OE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\frac{1}{K}\log\frac{\exp(\phi(\mathbf{z}_i^{\mathrm{out}})^{j})}{\sum_{m=1}^{K}\exp(\phi(\mathbf{z}_i^{\mathrm{out}})^{m})},$$

where $\phi(\cdot)$ is the classifier with a linear layer. The training loss $\mathcal{L}_{\mathrm{Train}} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{OE}} + \mathcal{L}_{\mathrm{Fusion}}$, where $\mathcal{L}_{\mathrm{Fusion}}$ is the loss specified in different multimodal fusion methods. Besides, we also conduct experiments by training a (K+1)-way classifier with out-of-scope samples grouped as the $(K{+}1)^{\mathrm{th}}$ class, resulting in significant decrease in the performance of in-scope classification (Appendix N).

Inference. During inference, our goal is to both identify in-scope classes and detect out-of-scope samples. To accomplish this, we employ a threshold-based open world classification method in NLP called DOC (Shu et al., 2017), which performs well in our experiments. This method rejects low-confidence samples, assigning statistical thresholds to each known class. For each sample $\mathbf{z}_i$, the predicted probability of the k^{th} class is given by $p(k|\mathbf{z}_i) = \mathrm{Sigmoid}(\phi(\mathbf{z}_i)^k)$. We use the output probabilities from each class of the training samples to calculate the corresponding class threshold δ_k. Specifically, we fit them to one half of the Gaussian distribution with μ = 1 and calculate the standard deviations σ_k using two symmetric halves of the probabilities. The class threshold is then given by $\delta_k = \max(0.5, 1 - \alpha\sigma_k)$, where $\alpha = 1$ usually works well. A test sample is detected as out-of-scope if $p(k|\mathbf{z}_i) < \delta_k, \forall k \in \mathcal{Y}$. Otherwise, it is considered as an in-scope sample and is assigned the predicted class with the maximum probability, denoted as $y_p = \mathrm{argmax}_{k\in\mathcal{Y}}\, p(k|\mathbf{z}_i)$.

5 Experiments

Implementation Details. We partition our dataset into training, validation, and testing sets, maintaining an approximate ratio of 7:1:2 for both dialogues and utterances. (Further details are provided in Appendix O). For the text modality, we utilize $\mathrm{BERT}_{\mathrm{LARGE}}$ as a powerful backbone consisting

Table 4: Benchmark baseline results on the MIntRec2.0 dataset.

		In-scope Classification						In-scope + Out-of-scope Classification			
Train	Methods	F1	P	R	ACC	WF1	WP	F1-IS	ACC	F1-OOS	F1
w / o OOS	TEXT	51.60	55.47	51.31	59.30	58.01	58.85	43.37	43.24	30.40	42.96
	MAG-BERT	55.17	57.78	55.10	60.58	59.68	59.98	46.48	44.80	34.03	46.08
	Δ(MAG-BERT)	3.57↑	2.31↑	3.79↑	1.28↑	1.67↑	1.13↑	3.11↑	1.56↑	3.63↑	3.12↑
	MulT	54.12	58.02	53.77	60.66	59.55	60.12	45.65	46.14	38.57	45.42
	Δ(MulT)	2.52↑	2.55↑	2.46↑	1.36↑	1.54↑	1.27↑	2.28↑	2.90↑	8.17↑	2.46↑
w OOS	TEXT	52.08	54.57	52.11	59.99	58.62	58.65	45.83	55.61	61.54	46.34
	MAG-BERT	53.64	54.84	53.79	60.12	59.11	58.83	47.52	56.20	62.47	48.00
	Δ(MAG-BERT)	1.56↑	0.27↑	1.68↑	0.13↑	0.49↑	0.18↑	1.69↑	0.59↑	0.93↑	1.66↑
	MulT	52.72	56.45	52.56	60.18	58.82	59.38	46.88	56.00	61.66	47.35
	Δ(MulT)	0.64↑	1.88↑	0.45↑	0.19↑	0.20↑	0.73↑	1.05↑	0.39↑	0.12↑	1.01↑
w OOS	Context TEXT	53.61	54.46	54.10	59.04	58.69	59.27	46.42	56.12	63.56	46.98
	Context MAG-BERT	53.89	55.72	54.21	59.84	59.41	60.22	46.74	56.20	62.52	47.25
	Δ(Context MAG-BERT)	0.28↑	1.26↑	0.11↑	0.80↑	0.72↑	0.95↑	0.32↑	0.08↑	1.04↓	0.27↑
	Context MulT	53.96	54.91	54.15	59.48	59.33	60.04	46.45	56.07	62.93	46.98
	Δ(Context MulT)	0.35↑	0.45↑	0.05↑	0.44↑	0.64↑	0.77↑	0.03↑	0.05↓	0.63↓	0.00

of 24 transformer layers implemented in the Huggingface transformers library (Wolf et al., 2020), to extract features with the dimension D^T of 1024. For the video modality, we employ well-trained checkpoints of Mask R-CNN from the MMDetection toolbox (Chen et al., 2019a) to extract features with the dimension D^V of 256. For the audio modality, we use the pre-trained model WavLM, implemented in (Wolf et al., 2020) to extract features with the dimension D^A of 768. In single-turn dialogues, we apply zero-padding with a maximum sequence length of 50, 180, and 400 for text, video, and audio features, respectively. The number of training epochs is set to 40, and the training batch size is set to 16 for all baselines. We employ AdamW (Loshchilov & Hutter, 2019) for optimization, implement our approach using PyTorch 1.13.1, and conduct experiments on Tesla V100-SXM2-32GB GPUs. For all experiments, we report the results averaged over five runs, using random seeds ranging from 0 to 4 (Additional hyper-parameters details are available in Appendix P).

Benchmark Baselines. As text is the predominant modality in conversational multimodal intent recognition (Zhang et al., 2022), we establish a robust baseline by fine-tuning $\text{BERT}_{\text{LARGE}}$, comparing its performance with two multimodal fusion methods: MAG-BERT and MulT. We evaluate these methods in both single-turn and multi-turn conversations, focusing on in-scope classification and out-of-scope detection. For single-turn conversations, we use only in-scope utterances for training. The out-of-scope utterances are included in the testing set and treated as a separate class, following (Lin & Xu, 2019; Zhang et al., 2023). For multi-turn conversations, we consider both in-scope and out-of-scope samples at the dialogue-level during training, and all the baselines utilize the context information as described in section 4. We conduct additional baselines related to dialogue intent classification in NLP and out-of-distribution detection across different sources in Appendices Q and R, respectively. Besides, we test the performance of ChatGPT on our dataset using both zero-shot and few-shot settings. In the zero-shot setting, ChatGPT is provided with the prompts of the label sets (e.g., 30 intent labels and one *OOS*) and an introduction to the task. In the few-shot setting, we use 10 dialogues with 227 utterances that cover all intent classes for learning (Details of the utilized prompts can be found in Appendix S). Finally, we invite ten evaluators to assess human performance. Each worker is assigned an equal portion of the testing set, ensuring they have not seen the data before. They receive the same background information of 10 dialogues as that provided to ChatGPT to ensure a fair comparison. Besides, we provide them with more prior knowledge, consisting of 100 dialogues and 997 utterances, to explore human potential in addressing this complex problem. We average the predictions from all evaluators to obtain the final score.

Evaluation Metrics. To evaluate the in-scope classification performance, we adopt six commonly used metrics: F1-score (F1), Precision (P), Recall (R), Accuracy (ACC), Weighted F1 (WF1), and Weighted Precision (WP). To evaluate out-of-scope detection performance, we utilize four metrics commonly employed in open intent classification (Shu et al., 2017; Zhang et al., 2023): Accuracy, Macro F1-score over all classes, In-scope classes (F1-IS), and the Out-of-scope class (F1-OOS).

Results. Table 4 shows the MIntRec2.0 dataset performance, with Δ indicating improvements of multimodal fusion methods over the text baseline. In single-turn dialogues, we test two scenarios:

without out-of-scope (OOS) samples (w / o OOS) and with OOS samples (w OOS). Multimodal fusion methods outperform the text baseline significantly in the in-scope only setting, with 1-4% score increases across all metrics. When including OOS samples, these methods show even larger improvements, suggesting enhanced in-scope identification and out-of-scope detection robustness due to cross-modal interactions. Training with OOS data leads to slight decreases in some in-scope metrics but a notable over 30% increase in F1-OOS scores for all baselines, highlighting the challenges in leveraging multimodal information for OOS data. Additionally, we conduct a case study on a selected dialogue, as detailed in Appendix T, to further explore the effect of incorporating multimodal information.

In multi-turn dialogues, multimodal methods improve all in-scope metrics compared to the text baseline but have minimal or negative effects when testing with mixed in-scope and OOS data. This suggests further potential for multimodal information in conversational contexts. For detailed analysis on fine-grained intent performance in single-turn and multi-turn conversations, see Appendix U.1.

ChatGPT v.s. Humans. Finally, we present the performance of ChatGPT and humans in Table 5. Humans typically excel at learning from few-shot samples and quickly grasping new concepts (Lake et al., 2015), leading us to apply a challenging setting of only 10 dialogues comprising 227 utterances. Multimodal fusion baselines, such as MAG-BERT-10, struggle significantly in this setting by easily overfitting and resorting to trivial solutions, like predicting the most frequent in-scope or out-of-scope class, due to the challenges posed by imbalanced and few-shot training samples. In contrast, ChatGPT demonstrates better performance even without prior knowledge of labeled data (ChatGPT-0), exhibiting strong language understanding and reasoning capabilities, comprehending complex textual semantics and understanding human intentions (Bang et al., 2023). However, ChatGPT shows slight improvements or, in some cases, a decrease in most metrics with merely 10 dialogues for learning (ChatGPT-10). This indicates a struggle in learning from limited prior knowledge to enhance intent recognition capability. Notably, it achieves a significant 6% improvement in F1-OOS, underscoring its improved robustness in out-of-scope detection. When provided with the same prior knowledge as ChatGPT-10, humans (Humans-10) achieve an increase of over 30% in scores across almost all metrics compared to ChatGPT. This highlights the significant limitations of existing AI methods in this challenging task, as humans effectively leverage limited multimodal information to understand high-level intentions and discern between known and unknown boundaries. To further explore human potential, we observe the performance of Humans-100 with additional knowledge of 100 dialogues comprising 997 utterances. Compared with Humans-10, Humans-100 achieve 7-13% improvements in almost all metrics and attain state-of-the-art benchmark performance. This underscores the advantages of humans in mastering this complex task by leveraging multimodal information. Additionally, we conduct experiments on fine-grained intent performance with ChatGPT and humans, which are detailed in Appendix U.2.

Table 5: Performance of ChatGPT and humans on the MIntRec2.0 dataset.

	In-scope			In-scope + Out-of-scope		
Methods	ACC	WF1	WP	ACC	F1-OOS	F1
MAG-BERT-10	9.82	11.58	13.34	34.28	50.57	3.75
ChatGPT-0	35.27	37.10	48.22	27.68	21.21	28.34
ChatGPT-10	34.53	36.39	49.27	29.72	27.85	28.41
Humans-10	64.34	67.82	72.80	60.43	62.83	57.83
Humans-100	**71.03**	**75.63**	**81.83**	**71.86**	**75.41**	**69.49**

6 Conclusions

This paper presents MIntRec2.0, a pioneering dataset for multimodal intent recognition, encompassing 1,245 dialogues and 15,040 multimodal utterances. This marks MIntRec2.0 as the first large-scale dataset in this domain. The dataset includes annotations for speaker identity and introduces a comprehensive taxonomy of 30 intent classes, spanning 9,304 in-scope utterances. To evaluate model robustness, 5,736 out-of-scope utterances are also annotated. We propose a general framework for organizing data, extracting multimodal features, and performing multimodal fusion for in-scope classification and out-of-scope detection in both single-turn and multi-turn conversations. Extensive experiments reveal the substantial potential of using multimodal information and uncover significant opportunities for improvement in effectively utilizing out-of-scope data and context information. Moreover, even with a strong LLM such as ChatGPT, using text-only modality remains challenging in scenarios with limited prior knowledge, highlighting the importance and challenge of using multimodal information compared to human performance. The limitations and potential negative societal impacts of this work are discussed in Appendix V.

REFERENCES

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460, 2020.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning*, volume 2, pp. 4, 2021.

Hervé Bredin and Antoine Laurent. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proceedings of the 22nd Annual Conference of the International Speech Communication Association*, pp. 3111–3115, 2021.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. Pyannote.audio: Neural building blocks for speaker diarization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7124–7128, 2020.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Inigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. Efficient intent detection with dual sentence encoders. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics WorkShop*, 2020.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019a.

Qian Chen, Zhu Zhuo, and Wen Wang. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*, 2019b.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Zifeng Cheng, Zhiwei Jiang, Yafeng Yin, Cong Wang, and Qing Gu. Learning to classify open intent via soft labeling and manifold mixup. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:635–645, 2022.

Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. M2fnet: multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4652–4661, 2022.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*, pp. 12–16, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009.

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

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2625–2634, 2015.

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1933–1941, 2016.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6202–6211, 2019.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 154–164, 2019.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2470–2481, 2020a.

Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. Utterance-level dialogue understanding: An empirical study. *arXiv preprint arXiv:2009.13902*, 2020b.

John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 517–520, 1992.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9180–9192, 2021.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6546–6555, 2018.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131, 2020.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *Proceedings of the International Conference on Learning Representations*, 2018.

Yutai Hou, Yongkui Lai, Yushan Wu, Wanxiang Che, and Ting Liu. Few-shot learning for multi-label intent detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 13036–13044, 2021.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7037–7041. IEEE, 2022a.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 7837–7851, 2022b.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 5666–5675, 2021.

Menglin Jia, Zuxuan Wu, Austin Reiter, Claire Cardie, Serge Belongie, and Ser-Nam Lim. Intentonomy: a dataset and study towards human intent understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12986–12996, 2021.

Sepideh Kaffash, An Truong Nguyen, and Joe Zhu. Big data algorithms and applications in intelligent transportation system: A review and bibliometric analysis. *International Journal of Production Economics*, 231:107868, 2021.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4903–4912, 2021.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. Integrating text and image: Determining multimodal document intent in Instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 4622–4632, 2019.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 1311–1316, 2019.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pp. 986–995, 2017.

Yuanchao Li, Carlos Toshinori Ishi, Koji Inoue, Shizuka Nakamura, and Tatsuya Kawahara. Expressing reactive emotion based on multimodal emotion recognition for natural conversation in human–robot interaction. *Advanced Robotics*, 33(20):1030–1041, 2019.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proceedings of the International Conference on Learning Representations*, 2018.

Ting-En Lin and Hua Xu. Deep unknown intent detection with margin loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5491–5496, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European conference on computer vision*, pp. 740–755. Springer, 2014.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3202–3211, 2022.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2247–2256, 2018.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*, 2019.

Bei Luo, Raymond YK Lau, Chunping Li, and Yain-Whar Si. A critical review of state-of-the-art chatbot designs and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(1):e1434, 2022.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 6818–6825, 2019.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pp. 18–25, 2015.

Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE J. Biomed. Health Informatics*, 26(12):6070–6080, 2022.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Proceedings of the Advances in Neural Information Processing Systems*, 2021.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *Proceedings of the European Conference on Computer Vision*, pp. 1–18. Springer, 2022.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, 2019.

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. A stack-propagation framework with token-level intent detection for spoken language understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 2078–2087, 2019.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2359–2369, 2020.

Shaoqing Ren, Kaiming He, Ross B Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in neural information processing systems*, 2015.

Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. Towards emotion-aided multi-modal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4361–4372, 2020.

Tobias Schröder, Terrence C Stewart, and Paul Thagard. Intention, emotion, and action: A neural theory based on semantic pointers. *Cognitive science*, 38(5):851–880, 2014.

Lei Shu, Hu Xu, and Bing Liu. Doc: Deep open classification of text documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2911–2916, 2017.

Janienke Sturm, Olga Houben-van Herwijnen, Anke Eyck, and Jacques M. B. Terken. Influencing social dynamics in meetings through a peripheral display. In *Proceedings of the 9th International Conference on Multimodal Interfaces*, pp. 263–270, 2007.

Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3927–3935, 2021.

Abhisek Tiwari, Manisimha Manthena, Sriparna Saha, Pushpak Bhattacharyya, Minakshi Dhar, and Sarbajeet Tiwari. Dr. can see: Towards a multi-modal disease diagnosis virtual assistant. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pp. 1935–1944, 2022.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, 2019.

Gökhan Tür, Dilek Z. Hakkani-Tür, and Larry P. Heck. What is left to be understood in atis? *2010 IEEE Spoken Language Technology Workshop*, pp. 19–24, 2010.

Sunny Verma, Jiwei Wang, Zhefeng Ge, Rujia Shen, Fan Jin, Yang Wang, Fang Chen, and Wei Liu. Deep-hoseq: Deep higher order sequence fusion for multimodal sentiment analysis. In *Proceedings of the 2020 IEEE International Conference on Data Mining*, pp. 561–570. IEEE, 2020.

Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4305–4314, 2015.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European conference on computer vision*, pp. 20–36. Springer, 2016.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, 2018.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision*, pp. 305–321, 2018.

Puyang Xu and Ruhi Sarikaya. Exploiting shared information for multi-intent natural language sentence classification. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association*, pp. 3785–3789, 2013.

Wei Xu. Toward human-centered ai: a perspective from human-computer interaction. *interactions*, 26(4):42–46, 2019.

Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu, and Albert Y. S. Lam. Unknown intent detection using gaussian mixture model with an application to zero-shot intent classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1050–1060. Association for Computational Linguistics, 2020.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727, 2020.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 10790–10797, 2021.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, 2017.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018a.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2236–2246, 2018b.

Hanlei Zhang, Xiaoteng Li, Hua Xu, Panpan Zhang, Kang Zhao, and Kai Gao. TEXTOIR: an integrated and visualized platform for text open intent recognition. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 167–174, 2021a.

Hanlei Zhang, Hua Xu, and Ting-En Lin. Deep open intent classification with adaptive decision boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14374–14382, 2021b.

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

Hanlei Zhang, Hua Xu, Shaojie Zhao, and Qianrui Zhou. Learning discriminative representations and decision boundaries for open intent detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1611–1623, 2023.

Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE International Conference on Computer Vision*, Oct 2017.

Yunhua Zhou, Peiju Liu, and Xipeng Qiu. Knn-contrastive learning for out-of-domain intent classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 5129–5141, 2022.

A Sample Selection within the MIntRec2.0 Dataset

Figure 5 illustrates a diverse selection of samples from our MIntRec2.0 dataset to showcase representative examples. The selected samples cover all 30 intent categories and the *OOS* label.

B Additional Related Work

Video Understanding. As a significant research field within computer vision, video understanding involves the extraction of valuable information from video content. Numerous methods have been developed to handle spatial and temporal data in videos, including the Two-Stream method, which comprises TDD (Wang et al., 2015), LRCN (Donahue et al., 2015), Fusion (Feichtenhofer et al., 2016), and TSN (Wang et al., 2016). This methodology integrates a secondary path to learn a video's temporal information by training a convolutional neural network on the optical flow stream. However, these methods require extensive computation and storage capacity due to the pre-computation of optical flow.

To address this, researchers introduce 3D convolutional neural networks (3D CNNs) such as I3D (Carreira & Zisserman, 2017), R3D (Hara et al., 2018), S3D (Xie et al., 2018), Non-local (Wang et al., 2018), and SlowFast (Feichtenhofer et al., 2019). More recently, self-attentive mechanisms like TimeSformer (Bertasius et al., 2021) and Video Swin Transformer (Liu et al., 2022) are demonstrating exceptional performance in image and video tasks. TimeSformer encodes video frames into a sequence of two-dimensional images, employing temporal self-attention to understand temporal relationships, while Video Swin Transformer partitions the input video into two-dimensional spatial and one-dimensional temporal patches, applying self-attention and cross-attention to manage long-distance temporal dependencies. X-CLIP (Ni et al., 2022), a CLIP-based method, has achieved state-of-the-art performance in video understanding by processing video content through matching video frames with text data.

While these techniques show proficiency in action recognition, they encounter difficulties when attempting to understand fine-grained intentions with high-level semantics and require considerable computational resources. For instance, X-CLIP demonstrates subpar performance on our task and demands a substantial amount of GPU memory, underscoring the need to incorporate other modalities such as language and acoustics in multimodal intent recognition tasks. Consequently, we have established baselines using multimodal fusion methods in this work.

Out-of-scope Intent Detection. As a significant task in natural language processing (NLP), out-of-scope intent detection has attracted considerable attention in recent years. Lin & Xu (2019) pioneers this task by employing margin loss to detect unknown intent. Zhang et al. (2021b) learns adaptive decision boundaries for each known class, thereby further reducing the open space risk. Yan et al. (2020) uses Gaussian mixture models to tackle this problem and extends the task to zero-shot intent detection. Cheng et al. (2022) constructs out-of-scope samples using manifold mixup technologies and employed soft labels for representation learning. Zhou et al. (2022) enhances intent representations to balance both empirical and open space risks with the aid of contrastive learning in the K-nearest neighbors space.

C Human-Computer v.s. Human-Human Interaction Datasets

The difference between human-computer and human-human interaction datasets is substantial, primarily due to the inherent differences in the nature of communication and interaction in each setting. Here are the primary three distinctions:

- **Interaction Dynamics**. In human-computer interactions, the dynamics are typically unidirectional or asymmetrical. Users often lead the conversational directions and give commands to generate dialogue utterances (e.g., in goal-oriented dialogue systems (Coucke et al., 2018; Larson et al., 2019)). In contrast, human-human interactions are more dynamic and bidirectional (e.g., in open-ended dialogue systems Busso et al. (2008); Poria et al. (2019)), with both parties actively contributing, responding, and adapting to the conversation flow.
- **Complexity of Communication**. Human-computer interactions are generally more structured and predictable, with a limited range of intents that follow specific orders or needs and relatively simple responses. Human-human interactions are far more complex, involving a wider range of intents, subtleties, emotions, and unpredictability.
- **Non-Verbal Cues**. Non-verbal cues are often limited or absent in human-computer interactions, as seen in many task-oriented datasets in NLP Coucke et al. (2018); Larson et al.

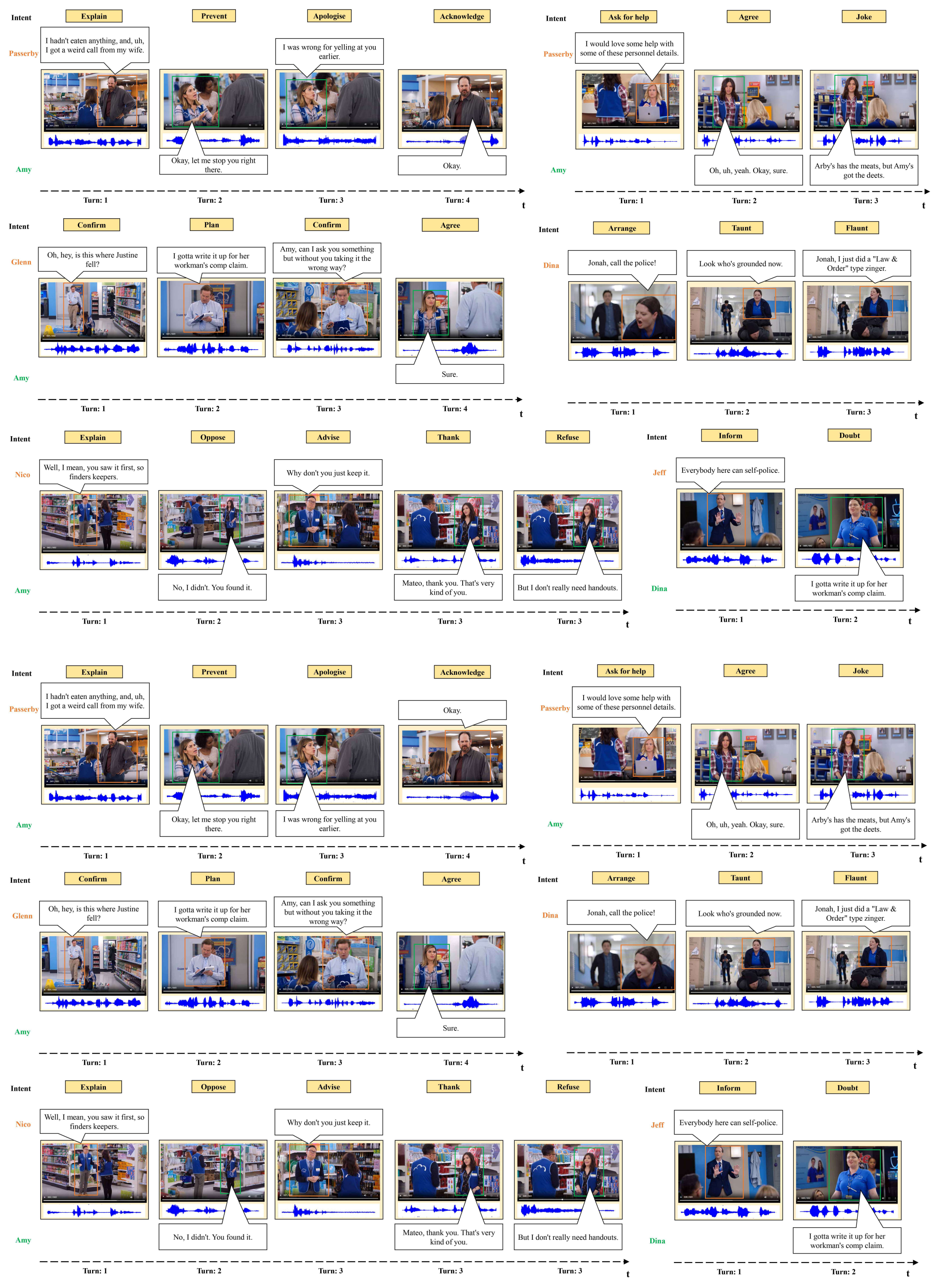


Figure 5: Samples of the MIntRec2.0 dataset.

(2019); Casanueva et al. (2020). Even with advanced multimodal fusion algorithms, interpreting non-verbal cues from humans remains challenging for computers. In human-human interactions, non-verbal cues play a crucial role in understanding intent, with nuances in body language, facial expressions, and tone carrying significant information.

Therefore, constructing a human-human interaction dataset for multimodal intent recognition is more appropriate, as it satisfies interaction dynamics, supports complex communication, and incorporates non-verbal cues. MIntRec2.0 makes a pioneering contribution in this area and aims to facilitate related research and application.

D Performance of DialogueRNN

Table 6: Results of DialogueRNN on the MIntRec2.0 dataset.

	In-scope Classification						In-scope + Out-of-scope Classification			
Setting	F1	P	R	ACC	WF1	WP	F1-IS	ACC	F1-OOS	F1
K+1	0.67	0.58	3.34	10.7	2.15	1.77	0.36	16.65	34.82	1.47
Outlier Exposure	2.75	4.19	3.74	3.89	3.23	5.29	2.21	11.10	23.67	2.91

To leverage context information, existing methods typically use multimodal fusion representations to directly model the temporal information of contexts. However, we find this approach to be ineffective for our task. Specifically, we select DialogueRNN (Majumder et al., 2019), a method specifically designed for multimodal emotion detection in conversations, for evaluation. We conduct experiments under two settings: K+1 and Outlier Exposure. The former treats the out-of-scope class as the $(K+1)^{\text{th}}$ class and trains using both K intent classes and one out-of-scope class, while the latter employs the outlier exposure loss on out-of-scope data during training.

As illustrated in Table 6, DialogueRNN demonstrates significantly low performance across all metrics. Furthermore, we observe that it tends to fall into trivial solutions, predominantly predicting most utterances as the out-of-scope class. This observation suggests that leveraging temporal information with fused multimodal representations remains a considerable challenge. Consequently, we adopt a simple method to leverage context information by concatenating the context information from the inputs of each modality.

E Conversational Scenes and Topics

The MIntRec2.0 dataset contains three popular TV series, Superstore, Friends and The Big Bang Theory. With 34 main characters and more than 10 primary types of conversational scenes and 15 distinct topics, covering a wide range of common intents encountered in daily life. Specifically, these conversational scenes and topics are diverse and include:

E.1 Scenes and Settings

Superstore provides a unique retail environment with scenes in the store, cash registers, break room, parking lot, managerial offices, and warehouse, reflecting workplace dynamics and customer interactions. Friends showcases diverse social settings like the Central Perk, apartments, travel locations, and various city spots, emphasizing personal and relational interactions. The Big Bang Theory offers academic and living spaces, including apartments and the university, highlighting intellectual and social engagements.

Each of these series brings a unique set of environments and interaction dynamics, ranging from personal and intimate to professional and public. The diversity in character backgrounds, professions, and social settings across these shows ensures a wide-ranging exploration of human interactions and conversational intents. Moreover, these series are culturally iconic and have significantly influenced societal communication patterns, making them highly relevant for studying contemporary conversational trends. Their popularity also ensures that the dataset is relatable and accessible for a broad range of researchers and applications.

E.2 TOPICS AND THEMES

Superstore touches on workplace relations, management challenges, customer service scenarios, labor issues, and social issues like immigration and corporate dynamics. Friends explores friendship, romantic relationships, career challenges, and urban living, offering insights into a variety of emotional and relational topics. The Big Bang Theory delves into scientific discourse, geek culture, technological advancements, scientific research, social awkwardness, and the balance between intellectual pursuits and everyday life.

The combination of these series presents an extensive range of human experiences and topics, from the mundane to the complex. This diversity enriches our dataset, making it an invaluable tool for studying and understanding the nuances of multimodal intent recognition in varied conversational contexts. Furthermore, these series, with their wide cultural impact, provide a relatable and realistic reflection of contemporary social dynamics, essential for developing robust and applicable AI models in the field of human-computer interaction.

In summary, the chosen TV series offer a balanced mix of scenes and topics, providing a comprehensive resource that captures the complexity of human interactions and conversational intents. We are confident that our dataset's scope and diversity significantly contribute to the advancement of multimodal intent recognition research.

F DATA PRIVACY AND CONTENT CONSIDERATIONS

Our dataset is meticulously curated and consists exclusively of character names and dialogues sourced from television shows, ensuring no infringement on the privacy or disclosure of personal information pertaining to real individuals. We have rigorously reviewed the content to maintain a high standard of decorum, assiduously avoiding any material that could be construed as offensive. Our focus remains strictly confined to the dialogues and interactions, all contextualized within the narrative framework of the respective shows, allowing for a comprehensive understanding of character dynamics without compromising ethical standards.

G UTTERANCE BOUNDARY ESTIMATION

To further validate the accuracy of these boundaries, we conduct additional experiments using a metric known as Speaker Boundary Error Rate (SBER), commonly employed in speech diarization tasks (Sturm et al., 2007). This metric quantifies the difference between predicted and reference speaker boundaries, with a lower SBER indicating better performance and serving as a proxy for sentence boundary accuracy. We utilize an end-to-end method implemented with pyannote (Bredin et al., 2020; Bredin & Laurent, 2021), a pre-trained speaker change detection model, to predict speaker IDs, starting times, and durations for each utterance within a dialogue segment. These predictions are then compared to the ground truth.

The results show an average SBER of 0.59 across all dialogues, suggesting considerable room for improvement in automatic sentence boundary segmentation. We believe this approach offers a reasonable method for evaluating utterance boundary performance.

H STATISTICS OF CHARACTERS

To further analyze the character distribution in each of the three data sources (i.e., Superstore, Friends, The Big Bang Theory) within our dataset, we present the proportions of characters from these sources in Figure 6, Figure 7, and Figure 8.

In Superstore, seven main characters and 21 recurring characters are observed. It can be noted that the seven main characters represent a significant proportion of nearly 80%, distributed uniformly. Friends have six main characters who constitute about 85% of the data, also distributed uniformly. The Big Bang Theory has seven main characters, while their distribution is imbalanced, a property we preserve due to the distinctive nature of each speaker. It is worth noting that there are other characters involved in the conversations, contributing 9.3%, 14.4%, and 5.9% respectively in each of the three TV series. These characters are also differentiated within each dialogue in our experiments.

I Intent Taxonomies Defined in the MIntRec Dataset

The MIntRec dataset (Zhang et al., 2022) introduces a hierarchical intent taxonomy, including two coarse-grained and 20 fine-grained intent categories. The two coarse-grained classes include *Express Emotions or Attitudes* and *Achieve Goals*. Based on these, it further includes 11 and 9 fine-grained classes for them, respectively. In particular, *Express Emotions or Attitudes* contains *complain, praise, apologize, thank, criticize, care, agree, oppose, taunt, flaunt*, and *joke*. *Achieve Goals* contains *inform, advise, arrange, introduce, comfort, leave, prevent, greet*, and *ask for help*. The interpretations of these categories are shown in Table 7, referring to (Zhang et al., 2022).

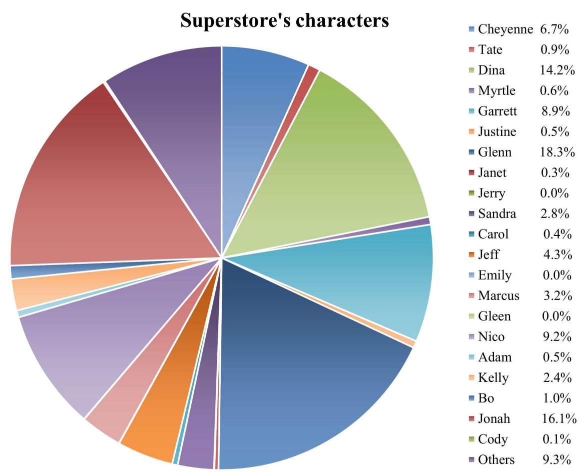


Figure 6: Proportions of characters from the TV series of Superstore.

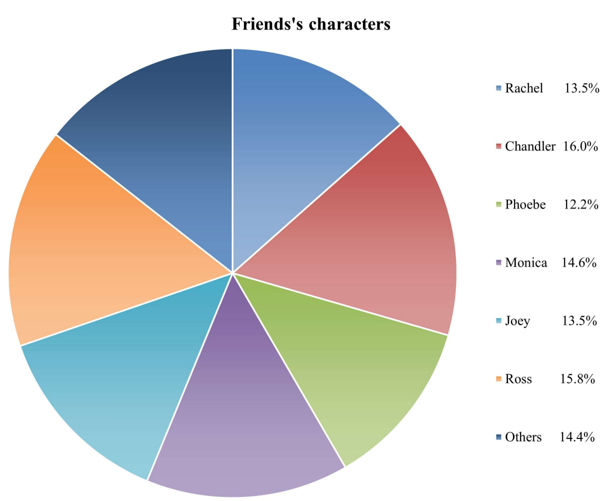


Figure 7: Proportions of characters from the TV series of Friends.

J Application of Intent Labels

Our intent labels can be generalized to many domains, including intelligent customer service, healthcare, mental health therapy, hazard detection, virtual assistants, and personalized recommendation systems. For instance:

- *complain, criticize, comfort*: These labels are instrumental in identifying potential mental health concerns in patients and can be pivotal in healthcare settings.
- *warn, prevent, OOS*: These labels can be employed effectively in systems designed for hazard detection.

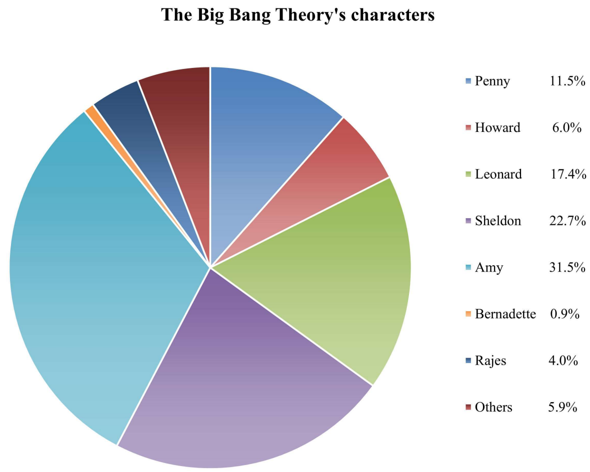


Figure 8: Proportions of characters from the TV series of The Big Bang Theory.

Table 7: Intent taxonomies of the MIntRec dataset with brief interpretations.

Intent Categories		Interpretations
Express emotions or attitudes	Complain	Express dissatisfaction with someone or something (e.g., saying unfair encounters with a sad expression and helpless motion).
	Praise	Express admiration for someone or something (e.g., saying with an appreciative expression).
	Apologize	Express regret for doing something wrong (e.g., saying words of apology such as sorry).
	Thank	Express gratitude in word or deed for the convenience or kindness given or offered by others (e.g., saying words of appreciation such as thank you).
	Criticize	Point out and emphasize someone's mistakes (e.g., yelling out someone's problems).
	Care	Concern about someone or be curious about something (e.g., worrying about someone's health).
	Agree	Have the same attitude about something (e.g., saying affirmative words such as yeah and yes).
	Oppose	Have an inconsistent attitude about something (e.g., saying negative words to express disagreement)
	Taunt	Use metaphors and exaggerations to accuse and ridicule (e.g., complimenting someone with a negative expression).
	Flaunt	Boast about oneself to gain admiration, envy, or praise (e.g., saying something complimentary about oneself arrogantly).
	Joke	Say something to provoke laughter (e.g., saying something funny and exaggerated with a cheerful expression).
Achieve goals	Inform	Tell someone to make them aware of something (e.g., broadcasting something with a microphone).
	Advise	Offer suggestions for consideration (e.g., saying words that make suggestions).
	Arrange	Plan or organize something (e.g., requesting someone what they should do formally).
	Introduce	Communicate to make someone acquaintance with another or recommend something (e.g., describing the identify of a person or the properties of an object).
	Comfort	Alleviate pain with encouragement or compassion (e.g., describing something is hopeful).
	Leave	Get away from somewhere (e.g., saying where to go while turning around or getting up).
	Prevent	Make someone unable to do something (e.g., stop someone from doing something with a hand).
	Greet	Express mutual kindness or recognition during the encounter (e.g., waving to someone and saying hello).
	Ask for help	Request someone to help (e.g., asking someone to deal with the trouble).

- *ask for help, inform*: These labels are particularly suited for customer service platforms.
- *praise, complain, agree*: These labels can be harnessed in personalized recommendation engines.
- *the majority of these intent labels*: These labels are ideal for virtual robots designed to interact naturally with users.

K Multimodal Intent Annotation Platform

We have developed an efficient platform featuring a unified database for multimodal label annotation, aiming to facilitate seamless interaction between annotators and the diverse set of multimodal data. The interface of this platform is depicted in Figure 9. This user-friendly interface allows annotators to access transcripts and associated videos from the dialogues and data sources easily, thereby ensuring accurate and consistent annotations. Annotators simply need to select one label from the 30 intent classes and an out-of-scope (*OOS*) tag by clicking a button. This intuitive design minimizes the learning curve for annotators and accelerates the annotation process. Once annotation is complete, the selected labels are automatically recorded in the database for statistical analysis.

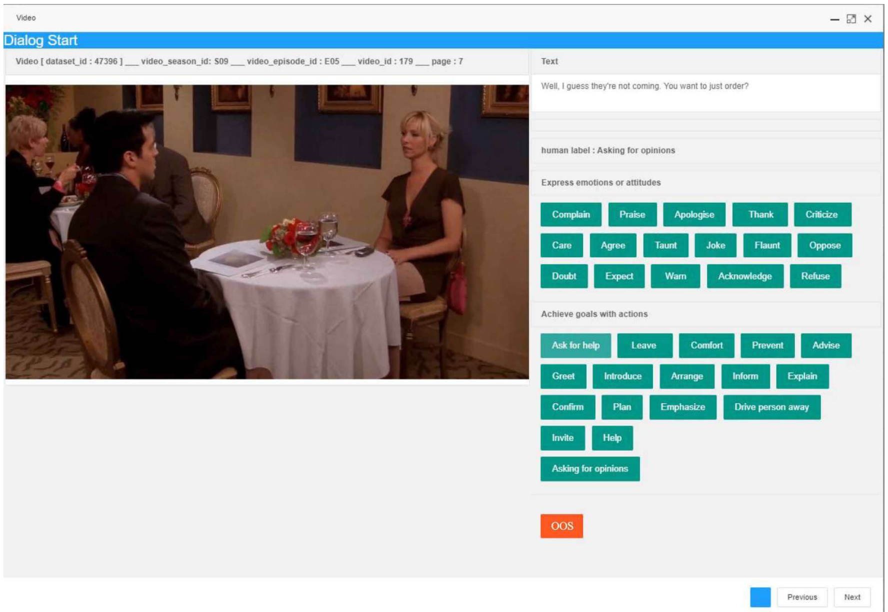


Figure 9: The interface of the annotation platform.

Table 8: Annotator rankings on the significance of modalities and background information in multimodal intent annotation.

rank	spoken language	facial expressions	tone	context information	body language	conversational scenes
1	**6**	0	2	1	0	0
2	2	**4**	2	1	0	0
3	1	3	**3**	1	1	0
4	0	2	1	**4**	2	0
5	0	0	0	2	**6**	1
6	0	0	0	0	1	**8**

This systematic approach ensures the reliability and consistency of the annotated data, which is crucial for training robust and high-performing models. The platform not only aids in the efficient collection of annotated data but also serves as a valuable tool for exploring and understanding the intricate relationships between different modalities and intents.

L The Significance of Modalities and Background Information

Intent recognition originates from natural language processing, with significant foundational research and advancements in this area (Chen et al., 2019b; Qin et al., 2019; Zhang et al., 2023), including high performance in some goal-oriented dialogue systems. As a result, the text modality often plays a central role in deciphering complex human intentions, as evidenced in (Zhang et al., 2022). Nevertheless, in real-world settings, integrating non-verbal modalities, such as video and audio, along with background information, such as conversational context and scene information, is crucial to accurately infer intentions. Recognizing the limitations of existing datasets, which are often small-scale, single-turn, and limited to closed-world classification, our MIntRec2.0 dataset aims to facilitate research into the effectiveness of non-verbal modalities in multi-turn conversations and in contexts that more closely resemble real-world situations, including out-of-scope utterances. To investigate the relative importance of different modalities in understanding conversational intent, we gather insights on key aspects including text modality (*spoken language*), video modality (*facial expressions* and *body language*), audio modality (*tone*), and background information (*context information* and *conversational scenes*). We ask 9 annotators involved in both annotation and human evaluation to rank these six aspects based on their importance in understanding human intentions, drawing from their experience.

As shown in Table 8, the results suggest that *spoken language* is the most critical factor, followed by *facial expressions, tone, context information, body language*, and *conversational scenes*. While *spoken language* is predominant, non-verbal cues like *facial expressions* and *tone* are valuable in perceiving emotions or attitudes, especially in the *express emotions or attitudes* coarse-grained category. *context information* provides essential background knowledge about the speakers, aiding in a more profound understanding of intentions. *body language*, though complex and implicit, is also insightful, particularly in the *achieve goals* coarse-grained category. Conversational scenes, while less critical in open-ended dialogues, still contribute to understanding intent in specific contexts.

M Single-intent Assumption

In real-world scenarios, it is possible for multiple intents coexist among the 30 pre-defined classes in a single utterance. In this work, we obey the single-intent assumption due to the following two reasons:

- **Single vs. Multi-Intent Datasets**: Most existing single-turn intent datasets in NLP, such as SNIPS, CLINC, and BANKING, focus on single-intention labeling. This is also true for multi-turn dialogue datasets like SWBD (Godfrey et al., 1992) and DailyDialog (Li et al., 2017), which generally assume a single dialogue act label at the utterance level. Therefore, while multiple intentions could theoretically exist in an utterance, the prevailing practice is to identify a primary intent for the sake of clarity and brevity.
- **Applicability to Real-World Scenarios**: We have examined multi-intent datasets like Standford_LU (Hou et al., 2021) and (Xu & Sarikaya, 2013). These datasets often include action and slot labels (e.g., find music or movie, request address or route), which are more suited for task-oriented dialogue systems. Such labeling is generally not applicable in real-world multimodal scenarios, as suggested in (Zhang et al., 2022).

To verify our assumption, we conduct an additional multi-intent annotation on the testing set. Six annotators are asked to identify up to three probable intents for each utterance. The results are shown in Table 9.

Table 9: Statistics of multiple intents in one utterance.

Express emotions or attitudes	Classes	complain, praise, apologize, thank, criticize, care, agree, warn
	Number	9, 5, 2, 1, 8, 1, 6, 1,
	Classes	oppose, taunt, flaunt, joke, doubt, acknowledge, refuse, emphasize
	Number	7, 4, 1, 2, 14, 3, 1, 8
Achieve goals	Classes	inform, advise, arrange, introduce, comfort, leave, prevent
	Number	5, 1, 1, 1, 2, 4, 0
	Classes	greet, ask for help, ask for opinions, confirm, explain, invite, plan
	Number	1, 1, 4, 5, 35, 1, 2

The results show that only 136 out of 3,230 utterances (4.2%) have a second most probable intent, and none have a third. This suggests that multi-intent scenarios are relatively rare, reinforcing the adequacy of our single-intent taxonomy. In summary, our findings align with those of most existing benchmark intent datasets, indicating that our intent taxonomy is both general and distinguishable enough for real-world applications.

N (K+1)-way Classification Performance

We also investigate another prevalent method, the (K+1)-way classification, to utilize the out-of-scope samples during training. In other words, we train on both the K known classes and one out-of-scope class. The results of this approach are displayed in Table 10. A noticeable decrease of approximately 10% in in-scope classification performance across numerous metrics (e.g., F1-score, recall, accuracy, weighted F1) is observed, compared to the results obtained with outlier exposure (OE) as depicted in Table 4 in the paper. Although there are slight improvements in F1-OOS (2% score increase) for out-of-scope detection in most methods, these methods still underperform when recognizing known classes and in overall performance. Therefore, we opt for outlier exposure as a more effective technique to deal with out-of-scope samples and adopt this approach in our work.

Table 10: K+1 classification results on the MIntRec2.0 dataset.

	In-scope Classification						In-scope + Out-of-scope Classification			
Methods	F1	P	R	ACC	WF1	WP	F1-IS	ACC	F1-OOS	F1
TEXT	42.23	55.34	37.42	43.84	49.60	64.28	40.52	55.69	64.28	41.29
MAG-BERT	40.68	53.34	36.57	43.75	48.95	63.14	38.87	55.76	64.41	39.70
MulT	39.48	54.96	34.90	42.47	48.04	64.17	38.26	56.33	65.48	39.14
Context TEXT	40.33	50.45	36.97	43.72	47.80	59.18	38.21	54.65	63.79	39.04
Context MAG-BERT	43.14	53.20	39.34	47.09	51.70	62.53	40.87	55.65	64.04	41.62
Context MulT	42.46	54.72	38.28	31.54	35.80	65.88	40.38	42.59	50.02	40.69

O DATA SPLITS

We partition our dataset into training, validation, and testing sets at an approximate ratio of 7:1:1 for both utterances and dialogues. Detailed statistics for each set, encompassing both in-scope and out-of-scope data, are presented in Table 11.

Table 11: Data splits of the MIntRec2.0 dataset. # denotes the number.

Item	# Dialogues	# Utterances	# In-scope Utterances	# Out-of-scope Utterances
Total	1,245	15,040	9,304	5,736
Training	871	9,989	6,165	3,824
Validation	125	1,821	1,106	715
Testing	249	3,230	2,033	1,197

P HYPER-PARAMETER CONFIGURATIONS

The comprehensive configurations of hyper-parameters used in our experiments are presented in Table 12, Table 13, Table 14, Table 15, Table 16, and Table 17.

Table 12: The hyperparameters of the TEXT baseline in single-turn conversations.

Setting	hyperparameters	value
w / o OOS	*eval_monitor*:	*accuracy*
	train_batch_size:	16
	eval_batch_size:	8
	test_batch_size:	8
	wait_patience:	8
	num_train_epochs:	40
	warmup_proportion:	0.1
	lr:	2e-5
	weight_decay:	0.1

Setting	hyperparameters	value
w OOS	*eval_monitor*:	*accuracy*
	train_batch_size:	16
	eval_batch_size:	8
	test_batch_size:	8
	wait_patience:	8
	num_train_epochs:	40
	warmup_proportion:	0.1
	lr:	1e-5
	weight_decay:	0.1

Q DIALOGUE INTENT CLASSIFICATION IN NLP

We have conducted experiments to benchmark our dataset with two state-of-the-art algorithms in open intent detection for NLP: DA-ADB (Zhang et al., 2023) and KNNCL (Zhou et al., 2022) with the open-source TEXTOIR platform (Zhang et al., 2021a). Consistent with the original settings of these algorithms, they are trained on in-scope samples and tested on both in-scope and out-of-scope samples. The results are shown in Table 18.

Table 13: The hyperparameters of the MAG-BERT baseline in single-turn conversations.

Setting	hyperparameters	value
w / o OOS	*need_aligned*:	*True*
	eval_monitor:	*accuracy*
	train_batch_size:	16
	eval_batch_size:	8
	test_batch_size:	8
	wait_patience:	8
	num_train_epochs:	40
	beta_shift:	0.005
	dropout_prob:	0.5
	warmup_proportion:	0.1
	lr:	5e-6
	aligned_method:	*ctc*
	weight_decay:	0.03

Setting	hyperparameters	value
w OOS	*need_aligned*:	*True*
	eval_monitor:	*accuracy*
	train_batch_size:	16
	eval_batch_size:	8
	test_batch_size:	8
	wait_patience:	8
	num_train_epochs:	40
	beta_shift:	0.005
	dropout_prob:	0.5
	warmup_proportion:	0.1
	lr:	5e-6
	aligned_method:	*ctc*
	weight_decay:	0.1

Table 14: The hyperparameters of the MulT baseline in single-turn conversations.

Setting	hyperparameters	value
w / o OOS	*padding_mode*:	*zero*
	padding_loc:	*end*
	need_aligned:	*False*
	eval_monitor:	*accuracy*
	train_batch_size:	16
	eval_batch_size:	8
	test_batch_size:	8
	wait_patience:	8
	num_train_epochs:	40
	dst_feature_dims :	80
	nheads:	4
	n_levels:	8
	attn_dropout:	0.0
	attn_dropout_v:	0.1
	attn_dropout_a:	0.1
	relu_dropout:	0.3
	embed_dropout:	0.0
	res_dropout:	0.0
	output_dropout:	0.2
	text_dropout:	0.1
	grad_clip:	0.5
	attn_mask:	*True*
	conv1d_kernel_size_l:	5
	conv1d_kernel_size_v:	1
	conv1d_kernel_size_a:	1
	lr:	5e-6

Setting	hyperparameters	value
w OOS	*padding_mode*:	*zero*
	padding_loc:	*end*
	need_aligned:	*False*
	eval_monitor:	*accuracy*
	train_batch_size:	16
	eval_batch_size:	8
	test_batch_size:	8
	wait_patience:	8
	num_train_epochs:	40
	dst_feature_dims :	80
	nheads:	4
	n_levels:	8
	attn_dropout:	0.0
	attn_dropout_v:	0.1
	attn_dropout_a:	0.1
	relu_dropout:	0.3
	embed_dropout:	0.0
	res_dropout:	0.0
	output_dropout:	0.0
	text_dropout:	0.0
	grad_clip:	0.5
	attn_mask:	*True*
	conv1d_kernel_size_l:	5
	conv1d_kernel_size_v:	1
	conv1d_kernel_size_a:	1
	lr:	5e-6

The results show that even state-of-the-art methods for open intent detection generally underperform compared to the $\text{BERT}_{\text{LARGE}}$ text classifier across most metrics. However, they do excel in identifying out-of-scope utterances, typically achieving higher F1-OOS scores. Notably, KNNCL also scores higher in accuracy.

R Out-of-distribution Detection Across Different Sources

We also explore the model performance in an out-of-distribution (OOD) setting across different sources. To address this, we have conducted experiments where we use data from one source as the in-distribution dataset for training, validation, and testing. We then use data from the other two sources exclusively for OOD testing, in accordance with (Hendrycks & Gimpel, 2017; Liang et al., 2018). For evaluation, we utilize a comprehensive set of metrics: AUROC (Area Under the

Table 15: The hyperparameters of the TEXT baseline in multi-turn conversations.

hyperparameters	**value**
eval_monitor:	*accuracy*
train_batch_size:	2
eval_batch_size:	2
test_batch_size:	2
wait_patience:	3
num_train_epochs:	40
warmup_proportion:	0.1
lr:	1e-5
weight_decay:	0.1
train_batch_size:	16

Table 16: The hyperparameters of the MAG-BERT baseline in multi-turn conversations.

hyperparameters	**value**
need_aligned:	*True*
eval_monitor:	*accuracy*
train_batch_size:	2
select_batch_size:	16
eval_batch_size:	2
test_batch_size:	2
wait_patience:	3
num_train_epochs:	40
context_len:	0.5
beta_shift:	0.05
dropout_prob:	0.05
warmup_proportion:	0.01
lr:	4e-6
aligned_method:	*conv1d*
weight_decay:	0.1

Receiver Operating Characteristic Curve), AUPR-In (Area Under the Precision-Recall Curve for in-distribution detection), AUPR-Out (Area Under the Precision-Recall Curve for OOD detection), FPR-95 (False Positive Rate at 95% True Positive Rate), and EER (Equal Error Rate). Higher scores are preferable for the first three metrics, while lower scores are desirable for the last two.

As shown in Table 19, the results indicate that MAG-BERT shows lower performance on OOD detection compared with the text baseline on most metrics. Both text and multimodal fusion methods achieve very low performance on OOD detection metrics, highlighting the substantial challenges presented by this setting. This opens up an intriguing avenue for future research in OOD detection under these conditions.

S CHATGPT PROMPTS

We provide prompts for both zero-shot (ChatGPT-0) and few-shot (ChatGPT-10) settings of ChatGPT. The detailed prompts are as follows:

ChatGPT-0 Prompts: Here is a set of given intent labels: [*Acknowledge*, *Advise*, *Agree*, *Apologise*, *Arrange*, *Ask for help*, *Asking for opinions*, *Care*, *Comfort*, *Complain*, *Confirm*, *Criticize*, *Doubt*, *Emphasize*, *Explain*, *Flaunt*, *Greet*, *Inform*, *Introduce*, *Invite*, *Joke*, *Leave*, *Oppose*, *Plan*, *Praise*, *Prevent*, *Refuse*, *Taunt*, *Thank*, *Warn*, *OOS*]. Additionally, *OOS* represents an unknown intent that does not belong to the known set of intents. Next, I will provide you with a collection of dialogs: *utterances*. The collection contains multiple utterances presented in sequential order, and they can be considered as contextualized conversations. When considering each sample and taking into account its contextual information, please select an appropriate label from the intent label set (emphasis: you can only choose intent labels from the given set of intent labels). If there are no suitable labels in the

Table 17: The hyperparameters of the MulT baseline in multi-turn conversations.

hyperparameters	**value**
padding_mode:	*zero*
padding_loc:	*end*
need_aligned:	*False*
eval_monitor:	*accuracy*
train_batch_size:	2
select_batch_size:	16
eval_batch_size:	2
test_batch_size:	2
wait_patience:	3
context_length:	1
num_train_epochs:	40
dst_feature_dims:	80
nheads:	4
n_levels:	8
attn_dropout:	0.0
attn_dropout_v:	0
attn_dropout_a:	0.1
relu_dropout:	0.2
embed_dropout:	0.1
res_dropout:	0
output_dropout:	0
text_dropout:	0.4
grad_clip:	0.5
attn_mask:	*True*
conv1d_kernel_size_l:	5
conv1d_kernel_size_v:	1
conv1d_kernel_size_a:	1
lr:	5e-6

Table 18: Performance of open intent detection on the MIntRec2.0 dataset.

	In-scope Classification						Out-of-scope Classification			
Methods	F1	P	R	ACC	WF1	WP	F1-IS	ACC	F1-OOS	F1
TEXT	51.60	55.47	51.31	59.30	58.01	58.85	43.37	43.24	30.40	42.96
DA-ADB	46.16	51.28	46.08	57.44	54.96	55.66	39.60	39.18	36.17	39.49
KNNCL	50.64	51.19	50.71	56.54	56.27	56.39	35.58	48.58	55.77	36.23

set, assign the label of the sample as *OOS*. Please provide the output in the following format: *Serial number and original text of the sample: Intent label*. Apart from that, do not output anything else.

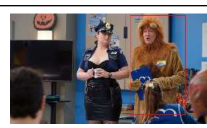
ChatGPT-10 Prompts: Here is a list of multiple multi-turn conversations. Each dictionary in the list represents a conversation paragraph, where each key-value pair represents an intent example as

Table 19: OOD detection performance across different sources.

ID source	OOD source(s)	Methods	AUROC	AUPR-In	AUPR-Out	FPR95	EER
Superstore	Bigbang & Friends	TEXT	51.33	21.75	80.25	93.47	49.43
		MAG-BERT	50.96	21.28	80.14	93.74	49.21
Bigbang	Superstore & Friends	TEXT	51.33	21.75	80.25	93.47	49.43
		MAG-BERT	50.96	21.28	80.14	93.74	49.21
Friends	Bigbang & Superstore	TEXT	55.97	26.17	80.56	91.22	45.40
		MAG-BERT	51.01	25.62	79.81	92.57	45.90

a key and its corresponding label as a value. Next time I will enter my request, please only reply "received".

Table 20: Case study on the impact of multimodal information.

Index	Speaker	Utterance	Video	TEXT: Predicted Label / Confidence	MAG-BERT: Predicted Label / Confidence	Ground Truth
0	Glenn	Salvatore kazlauskas.		Greet / 0.0735	OOS / 0.8289	OOS
1	Nico	Wait, you mean creepy sal?		Confirm / 0.9563	Confirm / 0.6321	Confirm
2	Glenn	The man is dead.		Inform / 0.4575	Inform / 0.3950	Inform
3	Dina	Police said he's been dead for at least a year.		Inform / 0.9842	Inform / 0.6941	Inform
4	Amy	Are you crying?		Care / 0.7637	Doubt / 0.3580	Confirm
5	Kelly	poor guy.		Taunt / 0.1573	Taunt / 0.1045	OOS
6	Amy	But you didn't know him.		Comfort / 0.0932	Explain / 0.1131	OOS
7	Kelly	But he was a human being.		Emphasize / 0.0693	Introduce / 0.0603	OOS
8	Cheyenne	When he looked at you, it felt like he was grabbing you.		Complain / 0.3858	Complain / 0.1209	OOS
9	Glenn	Apparently he was doing some work behind the drywall outside the women's washroom and then his foot got caught in a beam, and he starved to death.		Inform / 0.3145	Inform / 0.5194	Inform
10	Glenn	We're not sure.		Inform / 0.1238	Inform / 0.1233	OOS

11	Glenn	He drilled a hole into the women's washroom so...		Inform / 0.2779	Inform / 0.1861	Inform
12	Glenn	Why?		Doubt / 0.9722	Doubt / 0.8096	Doubt
13	Dina	I know we all assumed that was amy.		Emphasize / 0.0682	Explain / 0.0996	Explain
14	Amy	Why–why me?		Doubt / 0.7256	Doubt / 0.6292	Doubt
15	Dina	Cause, you know...		Explain / 0.9794	Explain / 0.4273	Explain
16	Dina	I'm sure there's a lot of churn going on in there.		Comfort / 0.1254	Explain / 0.0841	Taunt
17	Cheyenne	Wait, so it's just going to sit in the store?		Doubt / 0.9335	Doubt / 0.8337	Doubt
18	Nico	Uh... i'm not working next to a dead body.		Oppose / 0.9767	Oppose / 0.3158	Taunt
19	Dina	technically we've all been working next sal's dead body for the past year.		Explain / 0.1855	Explain / 0.22875768	Inform
20	Dina	Nobody complained until now.		Emphasize / 0.0534	Inform / 0.0770	OOS
21	Jonah	That must've been sal's foot we found.		Inform / 0.078	Introduce / 0.1448	OOS
22	Dina	Actually he still had both of his feet.		Doubt / 0.0581	Inform / 0.0656	Oppose

This is a list of given intent labels: [*Acknowledge*, *Advise*, *Agree*, *Apologise*, *Arrange*, *Ask for help*, *Asking for opinions*, *Care*, *Comfort*, *Complain*, *Confirm*, *Criticize*, *Doubt*, *Emphasize*, *Explain*, *Flaunt*, *Greet*, *Inform*, *Introduce*, *Invite*, *Joke*, *Leave*, *Oppose*, *Plan*, *Praise*, *Prevent*, *Refuse*, *Taunt*, *Thank*, *Warn*, *OOS*], where *OOS* represents an unknown intent that is not intended otherwise. Now, you need to learn from the conversations that you were given in the last Q&A, and then I'll provide

Table 21: Results of the text and MAG-BERT baselines for each fine-grained intent category in single-turn conversations.

Methods	Acknowledge	Advise	Agree	Apologise	Arrange	Ask for help	Ask for opinions	Care	Comfort	Complain	Confirm
TEXT	49.51	55.03	50.45	88.96	48.34	57.97	47.89	45.98	43.55	31.02	46.38
MAG-BERT	54.86	53.86	54.54	93.02	52.04	61.14	48.89	50.05	41.69	32.26	45.68
$\triangle$(MAG-BERT)	5.35↑	1.17↓	4.09↑	4.06↑	3.70↑	3.17↑	1.00↑	4.07↑	1.86↓	1.24↑	0.70↓

Methods	Criticize	Doubt	Emphasize	Explain	Flaunt	Greet	Inform	Introduce	Invite	Joke	Leave
TEXT	33.84	52.38	1.10	46.43	13.41	74.59	40.83	30.26	27.85	3.14	47.26
MAG-BERT	37.24	50.68	1.18	47.80	23.70	77.09	41.25	31.80	47.36	8.85	47.61
$\triangle$(MAG-BERT)	3.40↑	1.70↓	0.08↑	1.37↑	10.29↑	2.50↑	0.42↑	1.54↑	19.51↑	5.71↑	0.35↑

Methods	Oppose	Plan	Praise	Prevent	Refuse	Taunt	Thank	Warn	OOS
TEXT	56.88	51.43	61.03	52.41	20.64	12.61	91.12	18.97	30.40
MAG-BERT	56.60	54.73	63.30	51.32	29.37	13.76	91.85	30.89	34.03
$\triangle$(MAG-BERT)	0.28↓	3.30↑	2.27↑	1.09↓	8.73↑	1.15↑	0.73↑	11.92↑	3.63↑

Table 22: Results of the text and MAG-BERT baselines for each fine-grained intent category in multi-turn conversations.

Methods	Acknowledge	Advise	Agree	Apologise	Arrange	Ask for help	Ask for opinions	Care	Comfort	Complain	Confirm
TEXT	55.04	55.28	51.43	91.65	48.00	56.94	45.71	55.76	48.98	40.30	47.66
MAG-BERT	61.67	55.45	53.22	91.98	52.78	60.65	51.88	56.72	50.50	40.82	46.85
$\triangle$(MAG-BERT)	6.63↑	0.17↑	1.79↑	0.33↑	4.78↑	3.71↑	6.17↑	0.96↑	1.52↑	0.52↑	0.81↓

Methods	Criticize	Doubt	Emphasize	Explain	Flaunt	Greet	Inform	Introduce	Invite	Joke	Leave
TEXT	36.32	49.89	2.86	47.28	15.46	78.15	43.40	32.85	39.58	4.98	53.43
MAG-BERT	37.22	51.46	2.11	48.60	8.34	79.43	45.64	32.35	36.46	4.66	52.01
$\triangle$(MAG-BERT)	0.90↑	1.57↑	0.75↓	1.32↑	7.12↓	1.28↑	2.24↑	0.50↓	3.12↓	0.32↓	1.42↓

Methods	Oppose	Plan	Praise	Prevent	Refuse	Taunt	Thank	Warn	OOS
TEXT	58.00	51.91	66.23	56.90	31.60	15.33	84.72	27.01	63.56
MAG-BERT	55.87	48.44	67.65	46.67	21.36	13.27	85.23	28.19	62.52
$\triangle$(MAG-BERT)	2.13↓	3.47↓	1.42↑	10.23↓	10.24↓	2.06↓	0.51↑	1.18↑	1.04↓

you with a dialog that contains utterances in it, and these utterances are given in order and can be considered as contextual. Now, for each utterance that requires you to use the knowledge you gained from the given conversations, select a label as output from the given list of labels: for the following given dialog, in this format: *Original sample: Intent labels* output.

T Case Study

To investigate the impact of multimodal information in intent recognition, we select a specific dialogue from our dataset for a detailed case study. We compare the predicted intent and ground truth for each utterance using the text and MAG-BERT models, calculate the accuracy scores, and use the predicted probabilities as confidence scores for both models.

As Table 20 shows, MAG-BERT generally achieves high accuracy in most in-scope classes, except for some with complex semantics like *Taunt*, *Oppose*, and *Confirm*. Furthermore, in many correctly predicted in-scope classes, MAG-BERT demonstrates a high confidence level, often exceeding a 0.5 probability. In contrast, the text model exhibits higher confidence and more errors than MAG-BERT in certain utterances, such as the dialogues with indices 4, 5, 8, and 18. This comparison highlights the effectiveness of incorporating non-verbal modalities over relying solely on textual information. However, it is important to note that MAG-BERT struggles with out-of-scope (OOS) utterances, often making errors in these categories. This observation suggests that while existing multimodal fusion methods have capabilities in recognizing known intents, their performance in detecting out-of-scope utterances is limited, pointing to a significant area for future research and development.

Table 23: Comparison of ChatGPT-10, Humans-10, and Humans-100 across fine-grained intent categories.

Methods	Acknowledge	Advise	Agree	Apologise	Arrange	Ask for help	Ask for opinions	Care	Comfort	Complain	Confirm
ChatGPT-10	24.44	30.84	35.29	62.94	17.39	28.39	22.95	0.00	40.00	34.16	25.00
Humans-10	46.27	67.76	61.90	93.02	55.46	69.33	50.00	50.57	66.67	49.54	60.19
Δ (Humans-10)	21.83$\uparrow$	36.92$\uparrow$	26.61$\uparrow$	30.08$\uparrow$	38.07$\uparrow$	40.94$\uparrow$	27.05$\uparrow$	50.57$\uparrow$	26.67$\uparrow$	15.38$\uparrow$	35.19$\uparrow$
Humans-100	69.23	70.73	65.67	93.02	64.22	79.45	61.11	62.34	77.23	64.11	62.43
Δ (Humans-100)	44.79$\uparrow$	39.89$\uparrow$	30.38$\uparrow$	30.08$\uparrow$	46.83$\uparrow$	51.06$\uparrow$	38.16$\uparrow$	62.34$\uparrow$	37.23$\uparrow$	29.95$\uparrow$	37.43$\uparrow$

Methods	Criticize	Doubt	Emphasize	Explain	Flaunt	Greet	Inform	Introduce	Invite	Joke	Leave
ChatGPT-10	14.12	15.05	6.45	33.55	18.46	64.52	31.76	8.22	47.06	16.00	33.33
Humans-10	59.05	60.67	31.03	55.67	40.00	86.13	50.41	47.76	38.71	28.17	71.30
Δ (Humans-10)	44.93$\uparrow$	45.62$\uparrow$	24.58$\uparrow$	22.12$\uparrow$	21.54$\uparrow$	21.61$\uparrow$	18.65$\uparrow$	39.54$\uparrow$	8.35$\downarrow$	12.17$\uparrow$	37.97$\uparrow$
Humans-100	67.96	68.35	42.86	65.79	67.92	90.91	67.09	66.67	70.97	50.00	79.28
Δ (Humans-100)	53.84$\uparrow$	53.30$\uparrow$	36.41$\uparrow$	32.24$\uparrow$	49.46$\uparrow$	26.39$\uparrow$	35.33$\uparrow$	58.45$\uparrow$	23.91$\uparrow$	34.00$\uparrow$	45.95$\uparrow$

Methods	Oppose	Plan	Praise	Prevent	Refuse	Taunt	Thank	Warn	OOS
ChatGPT-10	25.00	29.27	50.00	12.12	16.87	13.48	59.09	37.04	27.85
Humans-10	67.51	57.78	73.06	70.77	47.83	44.62	93.91	34.78	62.83
Δ (Humans-10)	42.51$\uparrow$	28.51$\uparrow$	23.06$\uparrow$	58.65$\uparrow$	30.96$\uparrow$	31.14$\uparrow$	34.82$\uparrow$	2.26$\downarrow$	34.98$\uparrow$
Humans-100	67.56	70.89	79.65	77.97	53.66	63.93	94.83	62.86	75.41
Δ (Humans-100)	42.56$\uparrow$	41.62$\uparrow$	29.65$\uparrow$	65.85$\uparrow$	36.79$\uparrow$	50.45$\uparrow$	35.74$\uparrow$	25.82$\uparrow$	47.56$\uparrow$

U Fine-grained Intent Performance

U.1 Text v.s. Multimodal Fusion Methods

To further demonstrate the effectiveness of nonverbal modalities, we conduct experiments across 30 specific intent classes and one out-of-scope category, using the F1-score as the evaluation metric, similar to (Zhang et al., 2022). The average performance over 5 experimental runs is presented in Table 21. The results show significant enhancements in understanding 26 intent classes when integrating nonverbal modalities. Notably, 14 classes show improvements of over 3 points, including *Acknowledge, Agree, Apologise, Arrange, Ask for help, Care, Criticize, Flaunt, Invite, Joke, Plan, Refuse, Warn*, and *OOS*. These classes represent a mix of common and challenging scenarios, as well as out-of-scope instances, all requiring high-level cognitive inference and semantic understanding. Remarkably, we observe substantial improvements of over 10 points in challenging classes like *Flaunt*, *Invite*, and *Warn*. This underscores the importance of nonverbal modalities in recognizing human intentions. While some classes show lesser improvements or rely more on text modality, the performance with non-verbal modalities is competitive, demonstrating their substantial benefit in intent recognition in multimodal scenarios.

Furthermore, we extend our research to multi-turn conversations, maintaining the same experimental settings as in single-turn conversations and show the results in Table 22. In these settings, MAG-BERT outperforms the text-only modality in 18 classes. Specifically, it achieves improvements of over 3 points in four classes, including *Acknowledge*, *Arrange*, *Ask for help*, and *Ask for opinions*, and 1-2% improvements in 8 classes, such as *Agree*, *Comfort*, *Doubt*, *Explain*, *Greet*, *Inform*, *Praise*, and *Warn*. These classes encompass a significant portion of common interaction intents. However, the gains from nonverbal modalities in multi-turn conversations are not as pronounced as in single-turn conversations, indicating existing methodological limitations in handling out-of-scope utterances and fully utilizing context information. Addressing these challenges is crucial for future research and highlights both the importance and complexity of the MIntRec2.0 dataset.

U.2 ChatGPT v.s. Humans

To delve deeper into the performance on specific intent classes, we conduct experiments for ChatGPT-10, Humans-10, and Humans-100. We calculate the F1-score for each class and present these results in Table 23. The results illustrate that humans significantly outperform ChatGPT. With the same foundational knowledge of 10 dialogues encompassing 227 utterances, Humans-10 exhibits superior performance across nearly all intent classes and the out-of-scope category, outperforming ChatGPT-10 by over 10 points in most cases. Notably, 15 intent classes and one out-of-scope category show improvements of over 30 points. Classes like *Care* and *Prevent* achieve improvements of over 50 points, while *Ask for help*, *Criticize*, *Doubt*, and *Oppose* see over 40 points improvement. These findings highlight a significant gap between ChatGPT and human capabilities, underscoring humans' adeptness at using limited prior knowledge from multimodal contexts, such as body language and facial expressions, to infer and synthesize complex intents at a cognitive level, a skill where current machine learning methods, including large language models, fall short.

Additionally, with more extensive prior knowledge of 100 dialogues comprising 997 utterances, Humans-100 performs even better compared to Humans-10 and achieves state-of-the-art performance across all classes. This includes markedly improved performance of over 10 points in 16 intent classes and the out-of-scope category. This demonstrates the remarkable potential of humans to leverage multimodal knowledge and their ability to learn effectively with only a marginally larger dataset (7% of all training data). This proficiency even surpasses current fully supervised multimodal fusion methods, as shown in Table 4. The detailed intent performance comparison between ChatGPT and humans further validates the challenges presented by the MIntRec2.0 dataset, indicating that there is still considerable progress to be made in AI for the complex task of multimodal intent recognition.

V Limitations and Potential Negative Societal Impacts

Limitations: This study presents several limitations that warrant acknowledgment. First, deploying this system in real-world settings necessitates collecting personal data, including facial expressions, voice, and text, thereby raising critical privacy concerns requiring meticulous attention. Second, the issue of liability remains ambiguous, especially in sensitive applications such as medical diagnosis, should the technology produce erroneous results. Third, our training dataset may lack comprehensive representation across diverse cultural backgrounds, potentially resulting in misunderstandings or the perpetuation of stereotypes. Lastly, substantial opportunities exist for enhancing the system's performance, particularly in effectively utilizing context information and out-of-scope sample data and incorporating non-verbal modalities.

Potential Negative Societal Impacts: While our work contributes valuable advancements in the field of multimodal intent recognition, it also has the potential to introduce negative societal impacts.

Firstly, there is the potential for misuse of our dataset if it becomes publicly available under an open-source license. Such misuse could include unauthorized commercial applications or other nefarious purposes that could result in harm. To mitigate this, we strongly urge users to adhere strictly to the licensing terms associated with this dataset.

Secondly, as AI systems like ours become increasingly sophisticated and prevalent, there is the risk of over-reliance on these technologies. This could lead to a decline in certain human skills, especially those related to understanding and interpreting conversational cues. As researchers and developers, we must continue to balance the advancement of AI with the preservation and enhancement of human capabilities.

Thirdly, the baseline system might be used with malicious intent. While any technology can be used for both beneficial and harmful purposes, our system is designed to detect out-of-scope (OOS) categories, which could be exploited to identify harmful or malicious intents. By integrating robust OOS detection, our system can flag conversations or utterances that deviate from predefined, acceptable intents. This feature could act as a first line of defense against technology misuse, as it can be tailored to detect and flag potentially harmful conversation intents.

Furthermore, establishing a benchmark in this field can have numerous positive societal impacts, such as enhancing human-computer interactions, aiding mental health assessments, and improving customer service automation. We believe the ethical deployment of this technology largely hinges on implementation safeguards and the specific contexts in which it is used.