IntentionFrame: A Semi-Structured, Multi-Aspect Framework for Fine-Grained Conversational Intention Understanding

Jinggui Liang[1], **Dung Vo**[2*], **Lizi Liao**[1]

[1]Singapore Management University [2]Wayne State University

jg.liang.2023@phdcs.smu.edu.sg dung.vo@wayne.edu lzliao@smu.edu.sg

Abstract

Understanding user intentions in multi-turn dialogues is critical for conversational AI, yet existing approaches—relying on rigid slot-value structures or unstructured free-text—fail to fully capture conversational complexity. In this paper, we propose **IntentionFrame**, a semi-structured framework inspired by psychological and cognitive intention theories, which organizes conversational intents into four interrelated aspects: *situation, emotion, action,* and *knowledge*. This design not only retains interpretability but also provides LLMs with a rich context to accurately parse and respond to nuanced user inputs. To efficiently scale IntentionFrame annotations, we introduce a **We**akly-supervised **R**einforced **G**eneration (**WeRG**) method that leverages a small set of high-quality human annotations in conjunction with abundant coarsely labeled data. By applying reinforcement learning to balance these diverse signals, WeRG aims to effectively generate reliable IntentionFrame annotations, which serve as essential grounding for downstream tasks—leading to substantial improvements in response generation and task completion. Our experiments, supported by both automatic metrics and human evaluations, show that integrating IntentionFrame with WeRG significantly improves LLMs' conversational understanding and sets a new benchmark for intent analysis[1].

1 Introduction

Recent years have witnessed a surge of interest in developing conversational systems for social support and functional services, such as conversational recommendation (Kang et al., 2019) and emotional support (Liu et al., 2021a; Zheng et al., 2023). At the core of these systems lies Conversational Understanding (CU), which entails the accurate interpretation of user inputs across multiple dimensions (Chen et al., 2022b; Wang et al., 2023b).

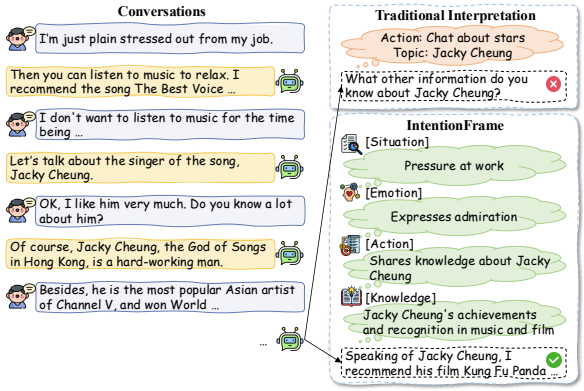

Figure 1: A comparison of existing structured interpretations and the proposed IntentionFrame framework.

Traditional CU approaches typically employ fixed ontologies with predefined intent classes and slot-value pairs (Casanueva et al., 2020; Tang et al., 2023). Yet, conversational systems may encounter rapidly evolving user needs and diverse expressions in real-world interactions. Thus, these static CU methods inevitably fall short of capturing such dynamics and dialogue nuances, leading to shallow or fragmented interpretations. (Zhang et al., 2021b; Liang et al., 2024b,a). An alternative line of work summarizes conversational content into free-form texts (Liu et al., 2019; Wu et al., 2021; Yang and Zhu, 2023), which offers greater flexibility to preserve nuanced details without rigid ontological constraints. Yet, unstructured summaries can easily become unfocused or inconsistent, making it difficult to model training and evaluation.

Meanwhile, the advent LLMs (OpenAI, 2023) has dramatically expanded the ability of conversational agents to handle complex contexts and subtle cues. Yet, prevailing CU paradigms—whether based on rigid schemas or free-text descriptions—have not kept pace with this progress and fail to fully leverage LLMs' potential. Structured

*This work was done during an internship at SMU.

[1]Dataset and code are available in https://github.com/liangjinggui/IntentionFame

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28097–28114

frames are too brittle to accommodate the fluid, open-ended nature of real dialogues, whereas unconstrained text representations lack clear focus and consistency. This discrepancy calls for a new approach to modeling conversational intentions that is both richly detailed and grounded in structure.

To fill this gap, we introduce **IntentionFrame**—a novel semi-structured framework for CU that offers a comprehensive, multi-aspect representation of user intents. Inspired by psychological and cognitive intention theories (Schröder et al., 2014), IntentionFrame decomposes user intentions into four key aspects: *situation* (conversational context), *emotion* (user's psychological state), *action* (intended behaviors), and *knowledge* (evolving dialogue information). As illustrated in Figure 1, a traditional interpretation might only identify an intent like "chat about a celebrity" with a topic slot (*e.g.*, "Jacky Cheung"), whereas IntentionFrame additionally encodes the user's circumstance (*e.g.*, feeling stressed at work), emotional attitude, and the evolving knowledge context. Unlike rigid CU interpretations, this structured yet adaptable format enables LLMs to capture a richer, more nuanced understanding of user queries, making it particularly well-suited for boosting the performance of downstream tasks like response generation.

To facilitate the large-scale adoption of IntentionFrame, we propose **We**akly-supervised **R**einforced **G**eneration (**WeRG**), a method that utilizes a small set of high-quality human annotations in combination with abundant coarsely labeled data, including a large proportion of existing ontology-based intents and LLM-annotated intentions. WeRG employs reinforcement learning to dynamically balance these diverse signals, assigning higher rewards to the high-quality annotations while still benefiting from the extensive coverage provided by the coarser labels. This approach enables the training of a conditional policy model that generates reliable and rich IntentionFrame annotations. Extensive evaluations show that the high-fidelity annotations produced by WeRG not only enhance LLMs' understanding of user intentions but also lead to significant improvements in downstream tasks such as response generation and task completion.

To sum up, our contributions are as follows:

- We formulate a semi-structured intention framework for effectively capturing the multi-faceted nature of human dialogues.
- We introduce the WeRG method for efficiently generating high-quality IntentionFrame annotations by integrating diverse supervision signals through reinforcement learning.
- Extensive experiments demonstrate significant improvements in CU and downstream conversation tasks, thereby highlighting the adaptability and robustness of IntentionFrame.

2 Related Works

Conversational Understanding. CU aims to accurately analyze user utterances in a conversation by delivering precise semantic interpretations (Chen et al., 2022b; Liu et al., 2023; Liang et al., 2024c). Prior CU studies primarily relied on static and structured conversational ontologies, delving into the individual tasks of intent detection and slot filling (Ravuri and Stolcke, 2015; Kurata et al., 2016; Xia et al., 2018; Lee and Jha, 2019; Casanueva et al., 2020; Zhang et al., 2023; Li et al., 2023; Mullick et al., 2024). Considering the close correlation between these tasks, recent efforts shifted toward joint intent-slot recognition, which employs a unified model to predict intents and slot sequences simultaneously (Zhang et al., 2019; Qin et al., 2021; Weld et al., 2023; Mirza et al., 2024; Yin et al., 2024; Pham and Nguyen, 2024). While these methods have shown progress, their reliance on static ontologies limits their applicability in real-world scenarios, where unforeseen user needs continually evolve.

Addressing this, recent CU studies also explore discovering new intents, slots, and values beyond the scope of static and structured ontologies. Innovations have developed techniques like unsupervised learning methods (Yang et al., 2017; Zhang et al., 2021a; Yu et al., 2022; De Raedt et al., 2023; Nguyen et al., 2023) and semi-supervised learning methods (Hsu et al., 2019; Zhang et al., 2021b, 2022; Zhou et al., 2023; Liang and Liao, 2023; Liang et al., 2024b,a; Wu et al., 2024). Extending beyond the inherently structured semantic interpretations, alternative methods (Liu et al., 2019; Wu et al., 2021; Chen et al., 2021, 2022a; Yang and Zhu, 2023) summarize conversation content into concise, free-form texts, offering greater flexibility for capturing conversational nuances without the constraints of rigid ontologies. Yet, the challenge remains in the lack of an effective framework that balances grasping in-depth conversational information with guiding the focus on producing accurate semantic interpretations—a gap this work

addresses by introducing the semi-structured IntentionFrame framework to CU.

Leveraging Diverse Annotations for Fine-tuning. In recent years, fine-tuning LLMs has been a key paradigm for improving their general performance and capabilities on unseen tasks (Wei et al., 2022; Ding et al., 2023). Generally, fine-tuning methods on LLMs can be broadly divided into two ways. The first focuses on Supervised Fine-Tuning (SFT) (Ding et al., 2023; Xu et al., 2024), which directly updates the LLM parameters using well-crafted SFT data with supervised learning objectives. Along this line, some studies (Chiang et al., 2023; Geng et al., 2023; Xu et al., 2024) have delved into designing high-quality data to facilitate the SFT process. Recent efforts also explore Parameter-Efficient Fine-Tuning (PEFT) methods (Lester et al., 2021; Hu et al., 2022; Zhang et al., 2024) and selecting high-quality data from varying quality supervision signals (Li et al., 2024) to balance the quality and efficiency of SFT.

The second fine-tuning method is Reinforcement Learning Fine-Tuning (RLFT), which employs a reward model trained on human preference data to fine-tune LLMs with RL objectives (Ouyang et al., 2022; Korbak et al., 2022; Rafailov et al., 2023; Wu et al., 2023; Wang et al., 2024). Recently, as LLMs evolve to be capable of supervising other models, RL from AI Feedback (RLAIF) has gained traction (Bai et al., 2022). RLAIF utilizes LLM-generated feedback to refine task instructions, optimizing LLMs to be harmless and detoxified (Shinn et al., 2023; Madaan et al., 2023; Hao et al., 2023). Inspired by this, some studies also explore using LLM-generated feedback combined with self-play mechanisms to enhance the abilities of the LLMs themselves (Chen et al., 2024b).

However, collecting high-quality supervision or reward signals to enhance LLM fine-tuning can be financially costly and prone to yielding substandard data, leading to compromised fine-tuning performance. This work addresses this with the WeRG method, which synergizes coarse-to-fine annotation data as weak supervision signals to facilitate the RLFT process.

3 Methodology

3.1 Preliminaries

Here, we study the CU problem formulated as follows: Let $\boldsymbol{c} = \{(x_1, y_1), \ldots, (x_T, y_T)\}$ represent a conversation, where x_t denotes the user's utterance at the t-th turn, y_t is the corresponding response, and T is the total number of dialogue turns. At each turn t, given the user utterance x_t and its associated dialogue history $h_t = \{(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})\}$, the primary objective is to learn a model $\mathcal{M}$ to generate the appropriate IntentionFrame data $\boldsymbol{o}_t$ using weak supervision signals from various data resources:

$$f_{\mathcal{M}} : (h_t, x_t) \rightarrow o_t, \tag{1}$$

where $o_t = \langle s_t, e_t, a_t, k_t \rangle$, comprising spans of defined situation, emotion, action, and knowledge.

3.2 IntentionFrame Framework

We present our framework for enhanced conversational understanding in Figure 2. Conventional CU approaches typically interpret user utterances through rigid, structured elements such as intents and slot-value pairs. However, these simplified representations often overlook the rich and nuanced information inherent in the conversational context—including aspects like conversational dynamics, emotional states, behavioral cues, and evolving contextual knowledge.

Motivated by these limitations, we introduce **IntentionFrame**, a semi-structured framework that provides a fine-grained, multidimensional representation of user intentions. IntentionFrame draws inspiration from intention theories in psychology and cognitive science (Eliasmith, 2013; Blouw et al., 2016) to encapsulate the nature of intentions and their decomposition in dialogue contexts. Specifically, Schröder et al. (2014) proposed a neural theory of intention as a brain process that functions as semantic pointers—binding together information about situations, emotional evaluations, actions, and sometimes also about self-knowledge.

Building upon this, we formalize IntentionFrame through four key aspects—*situation*, *emotion*, *action*, and *knowledge*—and elaborate on each as:

[Situation]: *This aspect describes physical or situational features of the current conversation.*
[Emotion]: *This aspect captures any emotional states or evaluations expressed by the user.*
[Action]: *This aspect refers to any actions the user mentions taking to achieve within their utterances.*
[Knowledge]: *This aspect identifies entities and relevant knowledge mentioned in the context.*

With this design, we can break down user inputs into four key aspects, gaining deeper insights into

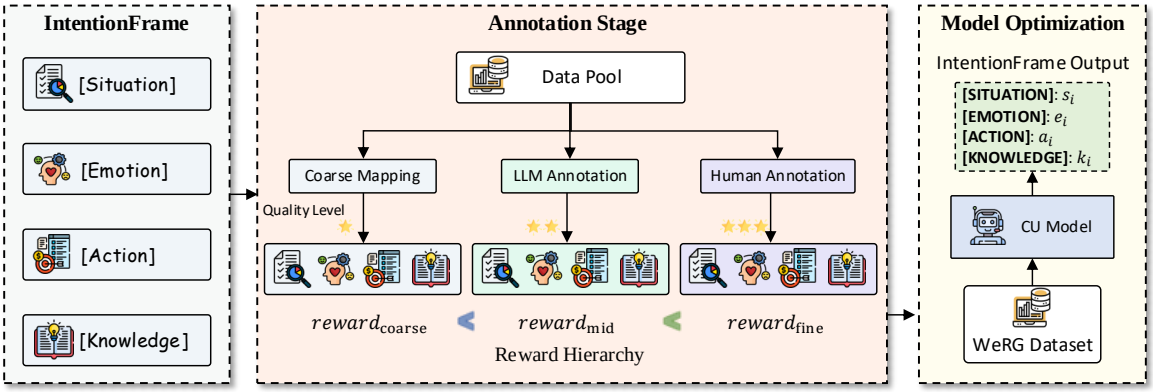

Figure 2: An overview of the proposed IntentionFrame framework and WeRG mechanism, which leverages a small set of high-quality human annotations together with abundant coarsely labeled data to train a better CU model.

the intentions behind their utterances. Additionally, each aspect of the IntentionFrame can be expressed in free-form natural language rather than being limited to a predefined and static conversational ontology, offering greater flexibility in accurately understanding users' evolving needs.

3.3 WeRG mechanism

After formulating the IntentionFrame framework that is capable of capturing enriched and in-depth information for understanding complex conversations, we need to acquire annotated IntentionFrame data for evaluation and further downstream applications. To accomplish this, a straightforward method involves directly employing human annotators to label high-quality IntentionFrame data, followed by supervised fine-tuning to optimize LLMs for generation. Despite its effectiveness, this approach is both labor-intensive and financially costly. Alternatives include leveraging cost-effective LLMs as annotators or directly transforming existing simplistic semantic interpretations—such as intents and slot-value pairs—into IntentionFrame labels for supervising LLMs. However, the resulting annotations are prone to noise and fail to cover the fine-grained aspects defined in the IntentionFrame schema, ultimately leading to degraded performance.

In light of the above considerations, we propose an effective weakly-supervised reinforced generation mechanism to facilitate the scalable adoption of IntentionFrame. Intuitively, WeRG is designed to synergistically integrate multiple sources of annotation—ranging from coarse-grained labels to fine-grained cues—as weak supervision signals. By leveraging this coarse-to-fine supervision hierarchy, WeRG enables efficient yet high-quality generation of IntentionFrame annotations.

Weak Supervision Construction. To implement WeRG mechanism, consider a conversation dataset $\mathcal{D} = \{(h_i, x_i, y_i)\}_{i=1}^{N}$, we first construct a fine-tuning dataset, $\mathcal{D}_{\text{WeRG}} = \mathcal{D}_{\text{coarse}} \cup \mathcal{D}_{\text{mid}} \cup \mathcal{D}_{\text{fine}}$, by employing a variety of annotation methods. Specifically, $\mathcal{D}_{\text{coarse}}$ uses hard mapping to transform existing structured interpretations into IntentionFrame labels, yielding coarse-level labels. In contrast, $\mathcal{D}_{\text{mid}}$ prompts cost-effective LLMs to annotate conversations within the IntentionFrame frame (details in Appendix A.2). Since LLMs can extract more nuanced information than existing structured interpretations, $\mathcal{D}_{\text{mid}}$ is thus endowed with mid-level labels. Unlike the above, $\mathcal{D}_{\text{fine}}$ employs human annotators to create IntentionFrame data, providing high-quality fine-level labels. Notably, due to the high cost of human annotation, the number of examples in $\mathcal{D}_{\text{fine}}$ is significantly less than those in $\mathcal{D}_{\text{mid}}$ and $\mathcal{D}_{\text{coarse}}$. Further details on these data segments are discussed in Appendix B.

To effectively utilize the coarse-to-fine level signals within $\mathcal{D}_{\text{WeRG}}$, following Wang et al. (2024), we further enhance $\mathcal{D}_{\text{WeRG}}$ by incorporating weak and tiered reward signals, which are meticulously calibrated to account for the variations across different annotation methods. Specifically, the reward is structured as a quadruple as follows:

$$r_c(h_i, x_i, o_i) = \langle r_s^{c_i}, r_e^{c_i}, r_a^{c_i}, r_k^{c_i} \rangle, \qquad (2)$$

where $\langle r_s^{c_i}, r_e^{c_i}, r_a^{c_i}, r_k^{c_i} \rangle$ are scalar rewards corresponding to the aspects $\langle s_i, e_i, a_i, k_i \rangle$ in o_i, with $c_i \in \{\text{coarse}, \text{mid}, \text{fine}\}$. Unlike previous studies, such as those described by Wang et al. (2024) that treat the entire ground-truth sequence equally, this

quadruple reward scheme allows for the allocation of distinct reward components to each aspect of the IntentionFrame. Notably, these reward signals are directly assigned according to the level of information provided by the annotations. By establishing the reward hierarchy—$r_{\text{coarse}} < r_{\text{mid}} < r_{\text{fine}}$—to reflect annotation quality, we can effectively guide the fine-tuning of LLMs to favor higher-quality IntentionFrame data without relying on strictly pairwise or ranking-based supervision for RL training.

Quality Assurance. To enhance the practical applicability of the constructed $\mathcal{D}_{\text{WeRG}}$, it is crucial to ensure the reliability and quality of the IntentionFrame annotations. We tackle this by conducting a human evaluation to assess their rationality across the four key aspects defined in Section 3.2. Following Cao et al. (2024), we adopt the Aspect Description Validity (ADV) criterion, which confirms that the annotated IntentionFrame aspects are contextually relevant, detailed, and accurate in capturing user intentions throughout the conversation. Evaluators are tasked with rating 200 randomly selected annotations, scoring each aspect for quality on a scale from 0 to 3. As shown in Table 1, the average scores range from 2.74 to 2.83, with $\mathcal{K}$ indicating a moderate to substantial level of inter-annotator agreement. This reflects the high quality of the IntentionFrame annotations collected in $\mathcal{D}_{\text{WeRG}}$.

Model Optimization. Given the constructed fine-tuning dataset $\mathcal{D}_{\text{WeRG}}$ and the reward information $r_c(h, x, o)$, we optimize a KL-regularized RL objective to fine-tune an LLM policy π_θ for efficiently generating high-quality IntentionFrame data. This widely used RL framework incorporates an additional KL penalty to constrain that the fine-tuned policy π_θ stays close to the base LLM, thereby avoiding distribution collapse. The objective can be formulated as follows:

$$J_{\text{WeRG}}(\theta) = \mathbb{E}_{\mathcal{O}\sim\pi_\theta}[r_c(h, x, o)] - \beta D_{KL}(\pi_\theta, \pi_w), \quad (3)$$

where π_w denotes the policy model augmented by the weak supervision signals in $\mathcal{D}_{\text{WeRG}}$. As demonstrated by previous works (Peters and Schaal, 2007; Korbak et al., 2022; Rafailov et al., 2023; Wang et al., 2024), the optimal solution π^* for the Equation (3) can be described as follows:

$$\begin{aligned}\pi^*(o|h, x, c) &= \arg\max_\theta J_{\text{WeRG}}(\theta) \\ &\propto \pi_w(o|h, x, c) \exp\left(\frac{1}{\beta} r_c(h, x, o)\right).\end{aligned} \quad (4)$$

	Situation	Emotion	Action	Knowledge	Ove.
		DuRecDial			
ADV	2.76	2.79	2.82	2.77	2.82
$\mathcal{K}$	0.48	0.58	0.64	0.56	0.62
		ESConv			
ADV	2.75	2.83	2.80	2.74	2.79
$\mathcal{K}$	0.50	0.63	0.56	0.55	0.56

Table 1: Human evaluation results. Scores (0 to 3) are averaged across all samples rated by evaluators. **Ove.** indicates overall performance across all four aspects, and $\mathcal{K}$ denotes Fleiss' Kappa (Fleiss, 1971) score.

Based on this optimal solution, the KL-regularized RL objective can be cast as minimizing the KL divergence between π_θ and π^* under the data distribution of $\mathcal{D}_{\text{WeRG}}$ (Nair et al., 2020; Korbak et al., 2022; Wang et al., 2024):

$$\begin{aligned}\pi_\theta = \arg\min_\theta \mathbb{E}_{(h,x,c)\sim\mathcal{D}_{\text{WeRG}}}\big[D_{KL}\big(\pi^*(\cdot|h, x, c) \,\|\, \\ \pi_\theta(\cdot|h, x, c)\big)\big].\end{aligned} \quad (5)$$

With this WeRG approach, we can effectively utilize weak supervision signals gathered from diverse data sources with coarse-to-fine labels, thereby enabling the LLM policy model to optimally generate IntentionFrame data.

4 Experiments

4.1 Datasets

We conduct experiments on two widely used conversational datasets: **DuRecDial** (Liu et al., 2021b) and **ESConv** (Liu et al., 2021a). Detailed dataset statistics are provided in Appendix A.1. We adhere to the same train, development, and test splits as prior studies (Dao et al., 2023; Deng et al., 2024; He et al., 2024). Additional experimental details can be found in Appendix A.2.

4.2 Evaluation Metrics

We employ both automatic and human evaluations to evaluate the effectiveness of the IntentionFrame framework and WeRG method. The automatic metrics include: (1) Content-based metrics (**F1** and **BLEU-1/2**), which measure the lexical overlap between generated outputs and ground-truth IntentionFrames; (2) Similarity-based metrics (**BERTScore** and **BARTScore**), which evaluate semantic alignment with the reference IntentionFrames; (3) LLM-as-a-Judge metric (**Judge**), where a strong LLM is employed to assess the alignment of the generated outputs with the ground-truth

Methods	DuRecDial				ESConv			
	F1 ↑	BLEU1/2 ↑	BERT/BARTScore ↑	Judge ↑	F1 ↑	BLEU1/2 ↑	BERT/BARTScore ↑	Judge ↑
DP	0.4851	0.3824/0.2015	0.5373/-3.5680	2.56	0.5279	0.4090/0.2386	0.5631/-3.2760	2.72
w/ Examples	0.5187	0.4021/0.2258	0.5554/-3.2842	2.82	0.5632	0.4376/0.2658	0.5903/-2.7149	2.91
CoTP	0.5135	0.4077/0.2331	0.5484/-3.2474	2.94	0.5695	0.4437/0.2718	0.5997/-2.6782	3.08
w/ Examples	0.5519	0.4354/0.2662	0.5897/-2.7762	3.11	0.6068	0.4912/0.3105	0.6431/-2.1365	3.25
SRT	0.5019	0.3923/0.2137	0.5264/-3.7261	2.73	0.5220	0.4035/0.2290	0.5525/-3.4100	2.86
SPIN	0.5423	0.4412/0.2589	0.5923/-2.8102	3.26	0.6112	0.4876/0.3145	0.6401/-2.1892	3.32
Ours	**0.5814**	**0.4715/0.2933**	**0.6232/-2.3652**	**3.65**	**0.6324**	**0.5127/0.3315**	**0.6721/-1.8863**	**3.74**

Table 2: Automatic evaluation of IntentionFrame generation performance. Results in bold indicate significant superiority over other methods. DP and CoTP represent baselines without examples.

IntentionFrames on a 0–4 scale; and (4) Dialogue-based metrics (**SR** and **AT**), which assess the success rate in guiding users to targets and the average turns of conversations during response generation. For human evaluation, we assess **Informativeness** (**Info.**), **Understanding** (**Und.**), and **Conciseness** (**Con.**). More details on these metrics are provided in Appendix A.3

4.3 Baselines

We compare the proposed method with the following baselines for generating IntentionFrame data: **Direct Prompting** (**DP**) ***w/o*** and ***w/*** Examples, **Chain-of-Thought Prompting** (**CoTP**) ***w/o*** and ***w/*** Examples, **SRT** (Li et al., 2024), and **SPIN** (Chen et al., 2024b). More details of these baselines are provided in Appendix A.4.

4.4 Main Results

4.4.1 Automatic Evaluation Results

To demonstrate the quality of the IntentionFrame data generated via the proposed WeRG mechanism, we compare our method against other baselines, with results reported in Table 2.

Firstly, regarding the content-based evaluation metrics, such as F1 and BLEU-1/2, our method consistently surpasses all baselines by a noticeable margin on both datasets. Among them, CoTP *w/o* Examples demonstrates superior performance compared to DP *w/o* Examples by enriching the LLM prompts with more detailed task descriptions and IntentionFrame explanations. CoTP *w/* Examples further amplifies this superiority by incorporating manually crafted IntentionFrame examples, showcasing the advantages of high-quality data in facilitating IntentionFrame data generation. Notably, our method synergistically integrates various sources of data annotated with coarse-to-fine labels to perform RLFT, allowing for a more effective and robust IntentionFrame model.

Methods	DuRecDial			ESConv		
	Info.	Und.	Con.	Info.	Und.	Con.
DP	2.88	3.74	2.55	2.52	3.17	2.75
w/ Examples	3.26	3.93	2.72	2.79	3.33	3.03
CoTP	3.31	4.05	2.83	2.76	3.40	2.97
w/ Examples	3.45	4.24	2.95	2.92	3.58	3.26
SRT	2.97	3.82	2.60	2.60	3.20	2.83
SPIN	3.38	4.12	2.88	2.85	3.47	3.15
Ours	**3.71**	**4.38**	**3.62**	**3.55**	**4.06**	**3.78**
$\mathcal{K}$	0.47	0.42	0.45	0.39	0.49	0.42

Table 3: Human evaluation results for IntentionFrame generation. Scores (0 to 5) are averaged across all samples rated by evaluators. $\mathcal{K}$ represents Fleiss' Kappa (Fleiss, 1971), indicating a moderate level of inter-annotator agreement ($0.2 < \mathcal{K} < 0.6$).

Secondly, in terms of similarity-based evaluation metrics, our method excels at generating more detailed and comprehensive content with broader inclusion of key information that semantically aligns with each aspect defined in the IntentionFrame schema. In contrast, the baseline methods—lacking explicit guidance to prioritize high-quality IntentionFrame data—struggle to yield outcomes that adequately capture the depth and richness required by the IntentionFrame framework. This suggests that the quadruple reward structure and tiered reward hierarchy implemented in the WeRG method enable LLMs to maximize the utility of high-quality data while compensating for the limitations of the substandard data during the fine-tuning process for IntentionFrame data generation.

4.4.2 Human Evaluation Results

To complement automatic evaluation, we further conduct human evaluations on the generated Inten-

Methods	DuRecDial				ESConv			
	F1 ↑	BLEU1 ↑	BLEU2 ↑	BERT/BARTScore ↑	F1 ↑	BLEU1 ↑	BLEU2 ↑	BERT/BARTScore ↑
Ours	**0.5814**	**0.4715**	**0.2933**	**0.6232/-2.3652**	**0.6324**	**0.5127**	**0.3315**	**0.6721/-1.8863**
- *w/o* $\mathcal{D}_{\text{coarse}}$	0.5744	0.4590	0.2811	0.6032/-2.6276	0.6231	0.5008	0.3197	0.6513/-2.1058
- *w/o* $\mathcal{D}_{\text{mid}}$	0.2355	0.1486	0.0832	0.2253/-4.5094	0.2547	0.1654	0.0968	0.2456/-4.3762
- *w/o* $\mathcal{D}_{\text{fine}}$	0.5488	0.4303	0.2622	0.5797/-2.8361	0.6015	0.4822	0.3035	0.6354/-2.2034
- *w/o* r_c	0.5347	0.4172	0.2430	0.5526/-3.1249	0.5792	0.4630	0.2851	0.6127/-2.5047

Table 4: Ablation study results for IntentionFrame generation. w/o denotes the model fine-tuned without the corresponding data source.

tionFrame examples with three student annotators. For both the DuRecDial and ESConv datasets, we randomly sampled 50 conversations from their respective test sets for validation. The annotators were asked to rate the performance of various methods. The evaluation results are reported in Table 3, which intuitively reveals the following findings: (1) It is evident that our proposed method consistently outperforms the baseline methods across all three human evaluation metrics, affirming the efficacy and practicality of our approach in generating high-quality IntentionFrame data. (2) We find that the WeRG mechanism, by applying quadruple rewards that separately emphasize different aspects as formulated in the IntentionFrame framework, effectively captures comprehensive information within conversations, including emotional cues. This nuanced approach leads to notable improvements, particularly in emotional support conversations, where our method demonstrates the most significant performance enhancements. Overall, the human evaluation results are consistent with those of the automatic evaluations, demonstrating that our method adeptly fine-tunes LLMs to generate IntentionFrame data of superior quality.

4.5 In-Depth Analysis

4.5.1 Ablation Studies

We conduct comprehensive ablation studies on the essential designs in our method—specifically, (1) the composition of weak supervision signals $\mathcal{D}_{\text{WeRG}}$ and (2) the reward module r_c—to analyze their individual contributions to overall generation performance using the DuRecDial dataset. The experimental results are detailed in Table 4. In the first setting, we selectively remove three types of supervision signals ($\mathcal{D}_{\text{coarse}}$, $\mathcal{D}_{\text{mid}}$, and $\mathcal{D}_{\text{fine}}$) from the fine-tuning dataset, where w/o denotes the configuration lacking the corresponding signals. As demonstrated in Table 4, excluding different sources of supervision from $\mathcal{D}_{\text{WeRG}}$ generally degrades the generation performance across both content-based and similarity-based evaluation metrics. In particular, the absence of supervision $\mathcal{D}_{\text{mid}}$, crucial for laying foundational insights into the IntentionFrame data, leaves the fine-tuning phase without essential guidance to extract the necessitated information that aligns with the defined IntentionFrame framework, leading to the most significant performance degradation. This suggests the effectiveness of these supervision signals with varying levels of annotated labels in supporting the model to generate higher-quality IntentionFrame data. In the second setting, we omit the quadruple reward r_c with its differential reward hierarchy during the model fine-tuning, which results in a notable decrease in performance. We hypothesize this can be attributed to the lack of explicit signals that enable the model to discern between coarse-to-fine annotated data without the differential rewards.

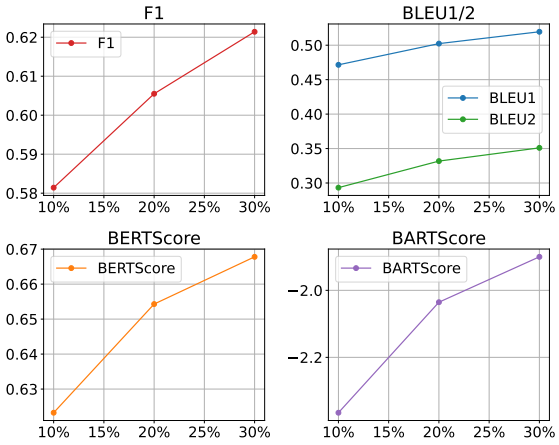

Figure 3: The impact of the proportion of fine-annotated data, ranging from 10% to 30%.

4.5.2 Impact of Annotated Data Proportion

We explore the effects of altering the proportion of human annotations $\mathcal{D}_{\text{fine}}$ on model performance.

Methods	F1 ↑	BLEU1/2 ↑	SR ↑	AT ↓
Direct Prompt	0.4297	0.3716/0.2147	0.7686	4.97
CoT Prompt	0.4427	0.3815/0.2243	0.7952	3.86
PPDPP	0.4362	0.3766/0.2195	0.79	3.43
T-EPL	0.4520	0.3920/0.2340	0.8120	3.39
CoT IntentionFrame	**0.4785**	**0.4107/0.3187**	**0.8537**	**3.37**

Table 5: Automatic evaluation results for the response generation task on the DuRecDial dataset, utilizing ChatGPT as the backbone generation model. ***CoT IntentionFrame*** denotes the CoT Prompt enhanced by the proposed IntentionFrame framework. PPDPP (Deng et al., 2024) and T-EPL (Dao et al., 2024) are two SOTA response generation methods.

Methods	F1 ↑	BLEU1/2 ↑	SR ↑	AT ↓
DuRecDial				
CoT IntentionFrame	**0.4785**	**0.4107/0.3187**	**0.8537**	**3.37**
- *w/o [SITUATION]*	0.4668	0.4012/0.3085	0.8404	3.48
- *w/o [EMOTION]*	0.4513	0.3908/0.2997	0.8309	3.60
- *w/o [ACTION]*	0.4321	0.3759/0.2904	0.8112	3.87
- *w/o [KNOWLEDGE]*	0.4365	0.3810/0.2952	0.8156	3.75
ESConv				
CoT IntentionFrame	**0.2979**	**0.2258/0.1370**	**0.8445**	**3.88**
- *w/o [SITUATION]*	0.2904	0.2158/0.1265	0.8292	4.10
- *w/o [EMOTION]*	0.2284	0.1758/0.0865	0.7692	5.34
- *w/o [ACTION]*	0.2746	0.2090/0.1205	0.8023	4.45
- *w/o [KNOWLEDGE]*	0.2679	0.1988/0.1141	0.7923	4.25

Table 6: Automatic evaluation results for the response generation task. *w/o* indicates the removal of the corresponding fine-grained aspect from the IntentionFrame during integration into generating responses.

In the standard experimental setting, we include human annotations that comprise 10% of the total dataset (*i.e.*, $|\mathcal{D}_{\text{fine}}|/N = 10\%$), primarily due to the costs associated with human annotators. Considering the pivotal role this high-quality data plays in steering the fine-tuning process towards generating more comprehensive IntentionFrame data, we experimentally increase this ratio to further examine its impact on model training using the DuRecDial dataset. Table 3 illustrates the performance trends across various ratios of fine-annotated IntentionFrame data. Notably, as the proportion of $\mathcal{D}_{\text{fine}}$ increases, the model performance improves with stable gains. This suggests that while the quantity of fine-annotated data $\mathcal{D}_{\text{fine}}$ is significantly less than $\mathcal{D}_{\text{coarse}}$ and $\mathcal{D}_{\text{mid}}$, it provides detailed insights into the human-preferred IntentionFrame data, continuously enhancing generation performance.

4.5.3 Effect on Downstream Applications

We further validate the effectiveness of applying the IntentionFrame data generated by the WeRG method to downstream conversational applications, specifically enhancing response generation in target-driven scenarios. We conduct experiments on the DuRecDial dataset by directly incorporating the IntentionFrame data into the inputs of the response generation model to enhance its output capabilities. Experimental results, detailing both dialogue-level and turn-level automatic evaluations, are presented in Table 5. By elucidating user utterances into fine-grained aspect information, our IntentionFrame framework markedly improves the ability of downstream response generation models, demonstrating the advantages of interpreting conversations in semi-structured natural language forms. Leveraging IntentionFrame, these models adeptly steer the flow of conversations by aligning subsequent turns with users' needs, thereby optimizing responses at each interaction to boost user engagement and successful target completion. Overall, the IntentionFrame framework lays a solid foundation for developing more sophisticated and effective conversational agents. More experimental results examining the robustness and adaptability of IntentionFrame across diverse dialogue scenarios are presented in Appendix C.

4.5.4 Effect of Different Fine-grained Aspects

The proposed IntentionFrame framework primarily establishes a multidimensional taxonomy, delving into aspects of situation, emotion, action, and knowledge to facilitate a comprehensive and multifaceted understanding of user utterances. To assess the individual contributions of these fine-grained aspects, we conduct experiments on the ESConv dataset by selectively omitting each of the four distinct aspects when applying the IntentionFrame framework to enhance downstream response generation. Results presented in Table 6 indicate a noticeable drop in performance whenever any detailed aspect is removed from the IntentionFrame framework. Notably, within the context of emotional support conversations, the removal of the ***[EMOTION]*** aspect—which is essential for revealing users' emotional cues throughout the conversation process—leads to the most substantial decrease in performance as the response generation model lacks specific guidance to tailor responses to users' emotional expectations. This underscores the potential of the IntentionFrame framework to support the customization of conversational agents for various real-world scenarios, aiding these agents in

accurately grasping users' diverse needs and delivering effective responses.

5 Conclusion

This work introduces the IntentionFrame, a novel fine-grained and aspect-aware formalism for understanding user intentions in intricate conversations. Building upon the semi-structured IntentionFrame framework, we propose WeRG, a mechanism that synergizes diverse sources of coarse-to-fine IntentionFrame annotations as weak supervision signals. By assigning varying quadruple rewards to each data source, WeRG facilitates the generation of high-quality IntentionFrame data. Overall, our method not only advances the capabilities of conversational agents in dialogue understanding but also offers insights into effectively leveraging coarse-to-fine supervision signals for generating large-scale, high-quality data. Extensive experiments validate the effectiveness of the IntentionFrame framework and demonstrate the superiority of the proposed WeRG approach. Consequently, this is a significant step towards building sophisticated conversational agents. Future research will focus on refining the IntentionFrame generation and validation process for even greater processing efficiency and accuracy.

Limitations

Despite the effectiveness of the IntentionFrame and the WeRG method, it is important to acknowledge several limitations. (1) We consider the introduced IntentionFrame to be a significant step toward enhancing conversational understanding for various downstream tasks that rely on robust dialogue comprehension. While we have explored its application in target-driven and emotional support dialogue contexts, IntentionFrame is anticipated to be applied to broader scenarios, such as harmful query detection and toxic behavior analysis in conversational safety. (2) The construction of weak supervision signals $\mathcal{D}_{\text{mid}}$ in $\mathcal{D}_{\text{WeRG}}$ relies on LLMs, making it susceptible to inherent issues such as biases in the training data and the potential for hallucinated or inaccurate outputs. While we have defined four well-structured dimensions in IntentionFrame to guide LLMs toward more focused, aspect-aware annotations—and combined $\mathcal{D}_{\text{mid}}$ with high-quality human annotations through the deliberately designed WeRG method—it remains a compelling direction for future work to systematically investigate the nature and impact of such issues in LLM-generated annotations. (3) The assumption that data quality varies by source might be overly simplistic, and the reward hierarchy could be refined to more precisely reflect the true quality of each annotation.

Acknowledgments

This research was supported by the Ministry of Education, Singapore, under its AcRF Tier 2 Funding (Proposal ID: T2EP20123-0052). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073.

Peter Blouw, Eugene Solodkin, Paul Thagard, and Chris Eliasmith. 2016. Concepts as semantic pointers: A framework and computational model. *Cognitive science*, 40 5:1128–62.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *NeurIPS*.

Yaru Cao, Zhuang Chen, Guanqun Bi, Yulin Feng, Min Chen, Fucheng Wan, Minlie Huang, and Hongzhi Yu. 2024. Enhancing emotional support conversation with cognitive chain-of-thought reasoning. In *NLPCC*, pages 175–187.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *NLP4ConvAI*.

Yue Chen, Chen Huang, Yang Deng, Wenqiang Lei, Dingnan Jin, Jia Liu, and Tat-Seng Chua. 2024a. STYLE: improving domain transferability of asking clarification questions in large language model powered conversational agents. In *Findings of ACL*, pages 10633–10649.

Yulong Chen, Naihao Deng, Yang Liu, and Yue Zhang. 2022a. Dialogsum challenge: Results of the dialogue summarization shared task. *CoRR*, abs/2208.03898.

Yulong Chen, Yang Liu, and Yue Zhang. 2021. Dialogsum challenge: Summarizing real-life scenario dialogues. In *INLG*, pages 308–313.

Zhi Chen, Lu Chen, Bei Chen, Libo Qin, Yuncong Liu, Su Zhu, Jian-Guang Lou, and Kai Yu. 2022b. Unidu: Towards A unified generative dialogue understanding framework. In *SIGDIAL*, pages 442–455.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024b. Self-play fine-tuning converts weak language models to strong language models. In *ICML*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, page 6.

Huy Dao, Yang Deng, Khanh-Huyen Bui, Dung D. Le, and Lizi Liao. 2024. Experience as source for anticipation and planning: Experiential policy learning for target-driven recommendation dialogues. In *Findings of EMNLP*, pages 14179–14198.

Huy Dao, Lizi Liao, Dung D. Le, and Yuxiang Nie. 2023. Reinforced target-driven conversational promotion. In *EMNLP*, pages 12583–12596.

Maarten De Raedt, Fréderic Godin, Thomas Demeester, and Chris Develder. 2023. IDAS: Intent discovery with abstractive summarization. In *NLP4ConvAI*, pages 71–88.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of EMNLP*, pages 10602–10621.

Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. 2024. Plug-and-play policy planner for large language model powered dialogue agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *EMNLP*, pages 3029–3051.

Chris Eliasmith. 2013. *How to build a brain: A neural architecture for biological cognition*. OUP USA.

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

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. Blog post.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *EMNLP*, pages 8154–8173.

Tao He, Lizi Liao, Yixin Cao, Yuanxing Liu, Ming Liu, Zerui Chen, and Bing Qin. 2024. Planning like human: A dual-process framework for dialogue planning. In *ACL*, pages 4768–4791.

Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. 2019. Multi-class classification without multi-class labels. In *ICLR*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *TICLR*.

Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul A. Crook, Y-Lan Boureau, and Jason Weston. 2019. Recommendation as a communication game: Self-supervised bot-play for goal-oriented dialogue. In *EMNLP-IJCNLP*, pages 1951–1961.

Tomasz Korbak, Ethan Perez, and Christopher L. Buckley. 2022. RL with KL penalties is better viewed as bayesian inference. In *Findings of EMNLP*, pages 1083–1091.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2022. Clam: Selective clarification for ambiguous questions with generative language models. *arXiv preprint arXiv:2212.07769*.

Gakuto Kurata, Bing Xiang, Bowen Zhou, and Mo Yu. 2016. Leveraging sentence-level information with encoder LSTM for semantic slot filling. In *EMNLP*, pages 2077–2083.

Sungjin Lee and Rahul Jha. 2019. Zero-shot adaptive transfer for conversational language understanding. In *AAAI*, pages 6642–6649.

Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020a. Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *WSDM*, pages 304–312.

Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020b. Interactive path reasoning on graph for conversational recommendation. In *KDD*, pages 2073–2083.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, pages 3045–3059.

Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Jiuxiang Gu, and Tianyi Zhou. 2024. Selective reflection-tuning: Student-selected data recycling for LLM instruction-tuning. In *Findings of ACL*, pages 16189–16211.

Xuefeng Li, Liwen Wang, Guanting Dong, Keqing He, Jinzheng Zhao, Hao Lei, Jiachi Liu, and Weiran Xu. 2023. Generative zero-shot prompt learning for cross-domain slot filling with inverse prompting. In *Findings of ACL*, pages 825–834.

Jinggui Liang and Lizi Liao. 2023. Clusterprompt: Cluster semantic enhanced prompt learning for new intent discovery. In *Findings of EMNLP*, pages 10468–10481.

Jinggui Liang, Lizi Liao, Hao Fei, and Jing Jiang. 2024a. Synergizing large language models and pre-trained smaller models for conversational intent discovery. In *Findings of ACL*, pages 14133–14147.

Jinggui Liang, Lizi Liao, Hao Fei, Bobo Li, and Jing Jiang. 2024b. Actively learn from llms with uncertainty propagation for generalized category discovery. In *NAACL-HLT*, pages 7845–7858.

Jinggui Liang, Yuxia Wu, Yuan Fang, Hao Fei, and Lizi Liao. 2024c. A survey of ontology expansion for conversational understanding. In *EMNLP*, pages 18111–18127.

Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019. Automatic dialogue summary generation for customer service. In *SIGKDD*, pages 1957–1965.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021a. Towards emotional support dialog systems. In *ACL/IJCNLP*, pages 3469–3483.

Xiao Liu, Jian Zhang, Heng Zhang, Fuzhao Xue, and Yang You. 2023. Hierarchical dialogue understanding with special tokens and turn-level attention. In *The First Tiny Papers Track at ICLR*. OpenReview.net.

Zeming Liu, Haifeng Wang, Zhengyu Niu, Hua Wu, and Wanxiang Che. 2021b. Durecdial 2.0: A bilingual parallel corpus for conversational recommendation. In *EMNLP*, pages 4335–4347.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *NeurIPS*.

Paramita Mirza, Viju Sudhi, Soumya Ranjan Sahoo, and Sinchana Ramakanth Bhat. 2024. ILLUMINER: instruction-tuned large language models as few-shot intent classifier and slot filler. In *LREC/COLING*, pages 8639–8651.

Ankan Mullick, Mukur Gupta, and Pawan Goyal. 2024. Intent detection and entity extraction from biomedical literature. In *LREC-COLING*.

Ashvin Nair, Murtaza Dalal, Abhishek Gupta, and Sergey Levine. 2020. Accelerating online reinforcement learning with offline datasets. *CoRR*, abs/2006.09359.

Hoang H. Nguyen, Chenwei Zhang, Ye Liu, and Philip S. Yu. 2023. Slot induction via pre-trained language model probing and multi-level contrastive learning. In *SIGDIAL*, pages 470–481.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Jan Peters and Stefan Schaal. 2007. Reinforcement learning by reward-weighted regression for operational space control. In *ICML*, volume 227, pages 745–750.

Thinh Pham and Dat Quoc Nguyen. 2024. JPIS: A joint model for profile-based intent detection and slot filling with slot-to-intent attention. In *ICASSP*, pages 10446–10450.

Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu. 2021. A survey on spoken language understanding: Recent advances and new frontiers. In *IJCAI*, pages 4577–4584.

Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and characterizing user intent in information-seeking conversations. In *SIGIR*, pages 989–992.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*.

Suman V. Ravuri and Andreas Stolcke. 2015. Recurrent neural network and LSTM models for lexical utterance classification. In *INTERSPEECH*, pages 135–139.

Tobias Schröder, Terrence C. Stewart, and Paul Thagard. 2014. Intention, emotion, and action: A neural theory based on semantic pointers. *Cogn. Sci.*, pages 851–880.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *NeurIPS*.

Yu-Chien Tang, Wei-Yao Wang, An-Zi Yen, and Wen-Chih Peng. 2023. RSVP: customer intent detection via agent response contrastive and generative pre-training. In *Findings of EMNLP*, pages 10400–10412.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2024. Openchat: Advancing open-source language models with mixed-quality data. In *ICLR*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *ACL*, pages 2609–2634.

Zhengjia Wang, Danding Wang, Qiang Sheng, Juan Cao, Silong Su, Yifan Sun, Beizhe Hu, and Siyuan Ma. 2023b. Understanding news creation intents: Frame, dataset, and method. *CoRR*, abs/2312.16490.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *ICLR*.

Henry Weld, Xiaoqi Huang, Siqu Long, Josiah Poon, and Soyeon Caren Han. 2023. A survey of joint intent detection and slot filling models in natural language understanding. *ACM Comput. Surv.*, pages 156:1–156:38.

Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. 2021. Controllable abstractive dialogue summarization with sketch supervision. In *Findings of ACL/IJCNLP*, pages 5108–5122.

Yuxia Wu, Tianhao Dai, Zhedong Zheng, and Lizi Liao. 2024. Active discovering new slots for task-oriented conversation. *IEEE ACM Trans. Audio Speech Lang. Process.*, pages 2062–2072.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. In *NeurIPS*.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S. Yu. 2018. Zero-shot user intent detection via capsule neural networks. In *EMNLP*, pages 3090–3099.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *ICLR*.

Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, and Mingyi Hong. 2017. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *ICML*, pages 3861–3870.

Diyi Yang and Chenguang Zhu. 2023. Summarization of dialogues and conversations at scale. In *EACL: Tutorial Abstracts*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *ICLR*.

Shangjian Yin, Peijie Huang, and Yuhong Xu. 2024. Uni-mis: United multiple intent spoken language understanding via multi-view intent-slot interaction. In *AAAI*, pages 19395–19403.

Dian Yu, Mingqiu Wang, Yuan Cao, Izhak Shafran, Laurent El Shafey, and Hagen Soltau. 2022. Unsupervised slot schema induction for task-oriented dialog. In *NAACL-HLT*, pages 1174–1193.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *NeurIPS*, pages 27263–27277.

Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and Philip S. Yu. 2019. Joint slot filling and intent detection via capsule neural networks. In *ACL*, pages 5259–5267.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen R. McKeown, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021a. Supporting clustering with contrastive learning. In *NAACL-HLT*, pages 5419–5430.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021b. Discovering new intents with deep aligned clustering. In *AAAI*, pages 14365–14373.

Haode Zhang, Haowen Liang, Li-Ming Zhan, Xiao-Ming Wu, and Albert Y. S. Lam. 2023. Revisit few-shot intent classification with plms: Direct fine-tuning vs. continual pre-training. In *Findings of ACL*, pages 11105–11121.

Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. 2024. Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *ICLR*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR*.

Yuwei Zhang, Haode Zhang, Li-Ming Zhan, Xiao-Ming Wu, and Albert Y. S. Lam. 2022. New intent discovery with pre-training and contrastive learning. In *ACL*, pages 256–269.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *Preprint*, arXiv:2308.11584.

Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024. Self-chats from large language models make small emotional support chatbot better. In *ACL*, pages 11325–11345.

Yunhua Zhou, Guofeng Quan, and Xipeng Qiu. 2023. A probabilistic framework for discovering new intents. In *ACL*, pages 3771–3784.

A Experimental Details

A.1 Dataset Statistics

In our experiments, we employ two commonly used conversational datasets—**DuRecDial** (Liu et al., 2021b) (recommendation dialogues) and **ESConv** (Liu et al., 2021a) (emotional support dialogues)—to evaluate the proposed IntentionFrame framework and WeRG mechanism. **DuRecDial** is a conversational recommendation dataset that contains 16.5K English-Chinese parallel dialogues and approximately 255K natural language utterances, spanning 14 goals and 646 topics. For our experiments, we utilize the English version of the dataset. **ESConv** is an emotional support conversation dataset consisting of 1,300 cases with 8 distinct support strategies. Each case includes a specified problem type, an emotion type, and a detailed situation description.

A.2 Implementation Details

For the construction of the dataset $\mathcal{D}_{\text{WeRG}}$, we employ *gpt-3.5-turbo* as the mid annotator to generate $\mathcal{D}_{\text{mid}}$ in our experiments. To ensure deterministic outputs during the acquisition of IntentionFrame annotations, the temperature parameter is fixed at 0, and the output is limited to a maximum of 1000 tokens. All other parameters are kept at their default settings. The prompts are designed to guide the LLMs, as detailed in Appendix D. For the dataset $\mathcal{D}_{\text{fine}}$, we randomly sample 10% of the conversations from the original dataset for fine-grained annotations.

For IntentionFrame policy model training, we use *llama-2-7b* as the backbone model and apply LoRA fine-tuning. The model is fine-tuned for 3 epochs on the constructed dataset $\mathcal{D}_{\text{WeRG}}$ using the AdamW optimizer, with a learning rate initialized at 6.7×10^{-5} and 100 warm-up steps. The fine-tuned parameters are saved every 1000 steps for subsequent evaluations. For the LoRA configuration, the rank is set to 8, the scaling factor to 16, and the dropout rate to 0.05. In the few-shot baseline setting, we utilize a one-shot demonstration randomly selected from the manually annotated dataset $\mathcal{D}_{\text{fine}}$.

For the reward setting, since the reward weight term in Equation (4) $\left(\exp\left(\frac{r_c}{\beta}\right)\right)$ remains constant within each class, we simplify the process by aligning the weights with the reward hierarchy $(r_{\text{fine}}) > r_{\text{mid}} > r_{\text{coarse}}$, assigning quadruple weights of $\langle 1.0, 1.0, 1.0, 1.0\rangle$ to $\mathcal{D}_{\text{fine}}$, $\langle 0.5, 0.5, 0.5, 0.5\rangle$ to $\mathcal{D}_{\text{mid}}$, and $\langle 0.1, 0.1, 0.1, 0.1\rangle$ to $\mathcal{D}_{\text{coarse}}$ for the conversation recommendation scenario. For emotional support conversations, we emphasize the emotion aspect, assigning fine-grained aspect weights of $\langle 0.9, 1.0, 0.9, 0.9\rangle$ to $\mathcal{D}_{\text{fine}}$, $\langle 0.4, 0.5, 0.4, 0.4\rangle$ to $\mathcal{D}_{\text{mid}}$, and $\langle 0.05, 0.1, 0.05, 0.05\rangle$ to $\mathcal{D}_{\text{coarse}}$.

A.3 Evaluation Matrics

In this work, the primary goal is to evaluate the quality of the IntentionFrame data generated via the WeRG approach, specifically its capability to capture the fine-grained aspect information as formulated by the IntentionFrame framework. To achieve this, we engage human annotators to label the IntentionFrame labels for the test set, thereby establishing the fundamental ground truth for the quality evaluation. After acquiring the IntentionFrame data, we also aim to validate its functionality in downstream applications. To this end, we further apply the generated IntentionFrame data to target-driven conversation scenarios, evaluating its effectiveness in enhancing the ability of conversational agents to respond to users and guide them toward the ultimate targets. In light of the above considerations, the evaluation protocols used in our experiments can be broadly categorized as follows:

Automatic Evaluation Protocols. The acquisition of IntentionFrame data via the WeRG method is fundamentally a generative process. In this sense, with the ground-truth labels previously established, most existing automatic generation metrics can be applied to assess the quality of the generated IntentionFrame data. Specifically, we utilize word-level F1 (**F1**) and **BLEU-N** (N=1, 2) metrics (Papineni et al., 2002) to compute the lexical overlap between the generated IntentionFrame data and the ground-truth labels, offering a quantitative measure of the precision and syntactic accuracy of the WeRG method. Additionally, we adopt **BERTScore** (Zhang et al., 2020) and **BARTScore** (Yuan et al., 2021) to measure the semantic similarity, further evaluating how well the generated data contextually aligns with the ground truth. To provide more reliable results for the automatic evaluation at scale, we also utilize the LLM-as-a-Judge (**Judge**) approach to assess the alignment of the generated outputs with the ground-truth IntentionFrames on a 0–4 scale. For validating the effectiveness of the IntentionFrame data in downstream tasks, we measure the dialogue-level Success Rate (**SR**) and the Averaged number of conversation

Data Segment	# Dialogues	Avg. Turns	Avg. Sit. Len	Avg. Emo. Len	Avg. Act. Len	Avg. Know. Len
$\mathcal{D}_{\text{coarse}}$	4154	8.0	6.87	1.0	2.1	2.72
$\mathcal{D}_{\text{mid}}$	4154	8.0	4.7	1.7	5.6	8.0
$\mathcal{D}_{\text{fine}}$	410	8.3	5.1	2.5	6.3	5.6

Table 7: Statistics of the annotated IntentionFrame data for the DuRecDial dataset.

Turns (**AT**) necessitated to successfully guide users to targets (Lei et al., 2020a,b).

Human-centered Evaluation Protocols. Generally, the most effective method for evaluating such texts is still human evaluation, wherein human annotators assess the quality of the generated IntentionFrame data. This evaluation can be approached from various perspectives, and we suggest several commonly used methodologies (Zheng et al., 2024): (1) **Informativeness (Info.)**: can the IntentionFrame data capture the key information throughout the conversation process? (2) **Understanding (Und.)**: whether the IntentionFrame data is clear and easy to understand in accurately describing users' real intentions? (3) **Conciseness (Con.)**: does the IntentionFrame data effectively communicate the necessary details without superfluous content? For these evaluations, we engaged three students as annotators, each tasked with assessing the IntentionFrame labels generated by various methods in 50 randomly selected conversations to ensure a comprehensive comparison.

A.4 Baselines

Direct Prompting (Brown et al., 2020): Directly provide LLMs with the necessary instructions as prompts to generate IntentionFrame data that grasps user intentions throughout the conversation process, including zero-shot and few-shot settings. In particular, the few-shot demonstrations are randomly selected from a set of manually constructed IntentionFrame examples.

Chain-of-Thought (CoT) Prompting: Building upon manually created examples provided, equip LLMs with detailed task descriptions and explanations of the IntentionFrame framework, specifying the criteria for generating IntentionFrame data by referring to the CoT method (Yao et al., 2023; Wang et al., 2023a), also including zero-shot and few-shot settings similar to the Direct Prompt baseline.

SRT (Li et al., 2024): A novel method synergizes the reflection and introspection of a teacher LLM with the data selection capabilities of a student LLM to automatically refine existing instruction-tuning data and improve data quality.

Methods	Recall@5 ↑	SR@3 ↑	SR@5 ↑
CLAM (Qu et al., 2018)	0.4950	0.5700	0.5933
ProCoT (Deng et al., 2023)	0.4950	0.6067	0.6233
STYLE (Chen et al., 2024a)	0.4956	0.6144	0.6511
STYLE w/ IntentionFrame	**0.5014**	**0.6237**	**0.6667**

Table 8: Evaluation of IntentionFrame's effectiveness on information-seeking dialogues.

SPIN (Chen et al., 2024b): A new fine-tuning method begins with a supervised fine-tuned model, which leverages a self-play mechanism that allows the LLM to refine its capabilities by playing against instances of itself.

B Details of Data Collection

Table 7 presents the statistical overview of the annotated IntentionFrame for the DuRecDial dataset. As shown in the table, the proportion of different data segments is: $\mathcal{D}_{\text{coarse}} : \mathcal{D}_{\text{mid}} : \mathcal{D}_{\text{fine}} = 10 : 10 : 1$. This ratio is deliberately set to balance the cost-effectiveness of annotating D_{fine} with the efficiency of guiding model training.

Generally speaking, the average length of each IntentionFrame-defined aspect indicates an increasing trend across $\mathcal{D}_{\text{coarse}} : \mathcal{D}_{\text{mid}} : \mathcal{D}_{\text{fine}}$, showcasing that higher-quality aspect annotations capture more comprehensive and intention-relevant information. Notably, the average lengths of situations in D_{coarse} and knowledge in D_{mid} deviate from this trend, which can be attributed to potential redundancy in these coarser annotations.

C Impact on Additional Downstream Applications

Notably, IntentionFrame is designed to be domain-agnostic. Its four aspects—Situation, Emotion, Action, and Knowledge—are grounded in established psychological and cognitive theories of intention, enabling application across diverse dialogue contexts. To examine its generalization and adaptabil-

ity, we additionally conduct an experiment on the complex information-seeking dataset MSDialog (Qu et al., 2018), with the corresponding results reported in Table 8. It can be observed that incorporating IntentionFrame consistently improves performance over the top-performing baselines (Kuhn et al., 2022; Deng et al., 2023; Chen et al., 2024a), highlighting the framework's broad applicability across domains.

D Prompt Details

In this section, we present the prompting details in our experiments.

D.1 DP Prompt

The prompts used for implementing the Direct Prompting baseline are presented in Table 9, including Direct Prompt *w/o* Example and Direct Prompt *w/* Example.

D.2 CoTP Prompt

The prompts used for implementing the Chain-of-Thought Prompting baseline are presented in Table 10, including CoTP *w/o* Example and CoTP *w/* Example.

D.3 Prompt to Response Generation

The prompts used for implementing the downstream response generation model are presented in Table 11.

E Case Study

Table 12 presents the cases of existing intention interpretations and the IntentionFrame examples.

DP Prompt

Please extract the conversational intentions based on the target-driven conversation provided below, where the {**target_goal**} guides the conversation. The intentions should concisely capture the user's focus conveyed in the [USER]-marked utterances. For each user utterance, identify the four aspects of user intentions—[SITUATION], [EMOTION], [ACTION], and [KNOWLEDGE]—and label them accordingly.
Please mark the input conversation according to the requirements and examples, ensuring each aspect is clearly addressed and provided. The marked intention numbers must strictly correspond one-to-one with conversation turns, with no merging or omissions allowed.
Here are some examples:
Conversation: ${Conversation}
IntentionFrame: ${IntentionFrame}
Input Conversation: ${Conversation}
IntentionFrame: [Provide the final output here]

Table 9: The prompt for implementing DP baseline.

CoTP Prompt

Description: I want you to apply your expertise in philosophy, psychology, and cognitive science to analyze and extract user intentions from a target-driven conversation, where the AI aims to make a {**target_goal**} to the user. The conversation is target-driven, meaning it strategically shifts towards the AI's goal.
Requirements: User intentions should succinctly reflect the user's focus conveyed within [USER]-marked utterances during conversations. Below are the detailed definitions and marking requirements for four aspects of user intentions:
[SITUAION]: Describe any physical or situational context mentioned by the user. If not applicable, mark as [SITUATION]: None.
[EMOTION]: Capture any emotional states or evaluations expressed by the user. If no emotions are expressed, mark as [EMOTION]: None.
[ACTION]: List any actions the user mentions taking to achieve the goal. If no actions are taken, mark as [ACTION]: None.
[KNOWLEDGE]: Identify entities and relevant knowledge mentioned in the conversation. If no specific knowledge is referenced, mark as [KNOWLEDGE]: None.
Please mark the input conversation according to the requirements and examples, ensuring each aspect is clearly addressed and provided. The marked intention numbers must strictly correspond one-to-one with conversation turns, with no merging or omissions allowed.
Here are some examples:
Conversation: ${Conversation}
IntentionFrame: ${IntentionFrame}
Input Conversation: ${Conversation}
IntentionFrame: [Provide the final output here]

Table 10: The prompt for implementing CoTP baseline.

Response Generation Prompt

Your task is to generate the next-turn response based on the provided IntentionFrame and Context. The Context is part of a **scenario** conversation between a user and an AI agent. The AI's goal is to guide the conversation toward achieving the {**target_goal**}. Generate a dialogue with frequent topic changes to extend the conversation over multiple rounds. Throughout the chat, smoothly switch topics back to the target goal, making the interaction feel like a natural, real-life conversation. Be specific in your responses. In the example below, [User] and [SYS] indicate the speaker of each turn.
Here are some examples:
Target: $Target
Conversation: ${Conversation}
IntentionFrame: ${IntentionFrame}
Next-turn response: ${Response}
Input:
Target: $Target
Conversation: ${Conversation}
IntentionFrame: ${IntentionFrame}
Next-turn response: [Provide the final output here]

Table 11: The prompt for implementing response generation model.

Example conversation

Target Music: Departing One

Conversations:
[SYS] Hello, Mr.Si! How are you doing?
[USR] 1. Hello! I'm good. It's just that the work is very boring.
[SYS] You can do something to enrich your life when you are bored.
[USR] 2. OK, I will. Thank you for reminding me!
[SYS] You can go to see a movie when you are bored. The movie 1:99 Shorts is wonderful. Give it a try!
[USR] 3. Not really for movie at the moment
[SYS] It's starred by Jacky Cheung who has won the Golden Horse Award for Best Actor.
[USR] 4. I like him very much.
[SYS] You've got good taste. He is a real singer.
[USR] 5. Thanks! The point is that he's very excellent. That's why I like him so much.
[SYS] Do you want to listen to Love You More Each Day by him? It's a golden melody in Hong Kong. It's very catchy.
[USR] 6. I've listened to it. Are there any other songs?
[SYS] Yes. How about Departing One? There are too many details. This brewer is gaudy but not pretentious.
[USR] 7. The name sounds so sad!
[SYS] You'll know after listening. Everyone feels differently.
[USR] 8. OK, I'll give it a try.

Existing Interpretations:

1. **[Action]** Greetings **[Topic]** None
2. **[Action]** Greetings **[Topic]** None
3. **[Action]** Movie recommendation **[Topic]** 1:99 Shorts
4. **[Action]** Chat about stars **[Topic]** Jacky Cheung
5. **[Action]** Chat about stars **[Topic]** Jacky Cheung
6. **[Action]** Music recommendation **[Topic]** Love You More Each Day
7. **[Action]** Music recommendation **[Topic]** Departing One
8. **[Action]** Music recommendation **[Topic]** Departing One

IntentionFrame:

Turn 1:
[Situation] Boring work. **[Emotion]** Neutral, but implies dissatisfaction with work. **[Action]** Describes current emotional state regarding work. **[Knowledge]** None.
Turn 2:
[Situation] None. **[Emotion]** Agreeable. **[Action]** Acknowledges the assistant's suggestion. **[Knowledge]** None.
Turn 3:
[Situation] None. **[Emotion]** Disinterest in movies at the moment. **[Action]** Declines the movie suggestion. **[Knowledge]** The movie 1:99 Shorts.
Turn 4:
[Situation] None. **[Emotion]** Positive, expresses admiration. **[Action]** Expresses liking for Jacky Cheung. **[Knowledge]** Jacky Cheung, Golden Horse Award for Best Actor.
Turn 5:
[Situation] None. **[Emotion]** Positive, expresses strong admiration. **[Action]** Explains the reason for liking Jacky Cheung. **[Knowledge]** Jacky Cheung's excellence.
Turn 6:
[Situation] None. **[Emotion]** None . **[Action]** Requests other song suggestions after acknowledging previous experience. **[Knowledge]** The song Love You More Each Day.
Turn 7:
[Situation] None. **[Emotion]** Emotional anticipation, slight negativity due to the sad connotation of the song's name. **[Action]** Reacts to the name of the song. **[Knowledge]** The song Departing One.
Turn 8:
[Situation] None. **[Emotion]** Willingness to try. **[Action]** Agrees to try listening to the suggested song. **[Knowledge]** None.

Table 12: Example conversation with the existing CU interpretations and our IntentionFrame.