

# LLM-Guided Semantic Relational Reasoning for Multimodal Intent Recognition

Qianrui Zhou<sup>1</sup>, Hua Xu<sup>1\*</sup>, Yifan Wang<sup>2,1</sup>, Xinzhi Dong<sup>1</sup>,  
Hanlei Zhang<sup>1</sup>

<sup>1</sup>Department of Computer Science and Technology, Tsinghua University

<sup>2</sup>School of Information Science and Engineering, Hebei University of Science and Technology  
zgr22@mails.tsinghua.edu.cn, xuhua@tsinghua.edu.cn

## Abstract

Understanding human intents from multimodal signals is critical for analyzing human behaviors and enhancing human-machine interactions in real-world scenarios. However, existing methods exhibit limitations in their modality-level reliance, constraining relational reasoning over fine-grained semantics for complex intent understanding. This paper proposes a novel LLM-Guided Semantic Relational Reasoning (LGSRR) method, which harnesses the expansive knowledge of large language models (LLMs) to establish semantic foundations that boost smaller models' relational reasoning performance. Specifically, an LLM-based strategy is proposed to extract fine-grained semantics as guidance for subsequent reasoning, driven by a shallow-to-deep Chain-of-Thought (CoT) that autonomously uncovers, describes, and ranks semantic cues by their importance without relying on manually defined priors. Besides, we formally model three fundamental types of semantic relations grounded in logical principles and analyze their nuanced interplay to enable more effective relational reasoning. Extensive experiments on multimodal intent and dialogue act recognition tasks demonstrate LGSRR's superiority over state-of-the-art methods, with consistent performance gains across diverse semantic understanding scenarios. The complete data and code are available at <https://github.com/thuiar/LGSRR>.

## 1 Introduction

Multimodal intent recognition aims to utilize information from both natural language and other non-verbal modalities (e.g. video and audio) to enable machines to discern intents within real-world scenarios. It holds significant research importance and has broad applications in domains such as human-computer interaction (Xu, 2019), chatbot (Fan et al., 2022), intelligent transportation system (Kaffash

et al., 2021), medical diagnosis (Tiwari et al., 2022; Moon et al., 2022) and other human-robot interaction systems (Paul et al., 2022; Mi et al., 2019).

Prior works (Zhang et al., 2022, 2024a) have pioneered this area by introducing large-scale multimodal intent datasets, attracting increasing research attention. These studies also adapt fusion strategies from multimodal sentiment analysis to construct initial baselines (Hazarika et al., 2020; Tsai et al., 2019; Rahman et al., 2020), laying the foundation for subsequent advancements. Recently, numerous studies focus on extracting and aligning modality-level semantics to improve multimodal fusion. For instance, TCL-MAP (Zhou et al., 2024) uses token-level contrastive learning and modality-aware prompts to enhance fusion between text and non-verbal modalities, while SDIF-DA (Huang et al., 2024) applies a shallow-to-deep framework that aligns modalities before fusing them through a Transformer (Vaswani et al., 2017) layer. Beyond fusion mechanisms, several approaches investigate diverse perspectives to enhance intent understanding, such as leveraging global video context (Sun et al., 2024), reducing noise and redundancy in non-verbal streams (Zhu et al., 2024) and employing multi-task optimization (Zhang et al., 2024b).

Despite these significant advancements, two critical issues remain in multimodal intent recognition. First, existing approaches primarily emphasize coarse-grained and modality-level semantics, which introduces substantial redundancy and noise (Zhu et al., 2024), resulting in a considerable gap between the extracted features and the true intent. Second, these methods generally model relationships between semantic concepts with basic fusion mechanisms, capturing only a limited subset of the complex reasoning relationships essential for accurate intent recognition. Consequently, there is a need for relational reasoning methods concentrating on fine-grained semantics for multimodal intent recognition, presenting two main challenges: (1)

\* Hua Xu is the corresponding author.

extracting fine-grained and intent-related semantics from diverse modalities, and (2) modeling complex reasoning relationships between these semantics.

Given the aforementioned limitations, the demonstrated strengths of LLMs in semantic understanding tasks (Lai et al., 2024; Xu et al., 2024a) offer a promising solution for capturing fine-grained semantics. While Xu et al. (2024b) first leverages LLMs to extract commonsense knowledge, the approach remains constrained to modality-level cues and requires manually specified information. In this work, we aim to further unlock the potential of LLMs by enabling them to independently discover high-level multimodal semantic concepts and provide reasoning guidance. Moreover, due to the inherent complexity of semantic relations, we draw inspiration from classical logic, mapping basic operators (“or,” “and,” “not”) to their semantic counterparts (relative importance, complementarity, and inconsistency) to model intricate semantic interactions through their structured composition.

Consequently, we propose the LLM-Guided Semantic Relational Reasoning (LGSRR) framework, as illustrated in Figure 1. To address the first challenge, we design an LLM-Guided Semantic Extraction module that employs a shallow-to-deep CoT for high-quality semantic discovery. Specifically, GPT-3.5 (Zhou et al., 2023) is first prompted to identify fine-grained semantic cues relevant to multimodal intents, from which the top- $K$  most frequent cues are selected to conduct semantic extraction. These initial cues are then enriched by VideoLLaMA2 (Cheng et al., 2024), which extracts detailed descriptive features from both text and video. Finally, GPT-3.5’s abductive reasoning capabilities are leveraged to generate a ranked semantic list, which serves as supervised guidance for subsequent relational reasoning. To tackle the second challenge, we introduce the Semantic Relational Reasoning module, which models the complex interplay among semantic cues by focusing on three fundamental logic-inspired relations. The relative importance of different semantics is learned through a unified weighting network optimized with NeuralNDCG (Pobrotyn and Białobrzęski, 2021) ranking loss based on the semantic rankings, while complementarity and inconsistency are naturally captured through cosine similarity and mean squared error between semantic representations respectively. To construct cohesive and discriminative intent representation, we integrate importance and complementarity as weighted factors to enrich

the semantic space, while using inconsistency as a regularization term to ensure balanced reasoning.

Our contributions are summarized as follows: (1) We introduce an LLM-Guided framework which utilizes LLMs to autonomously acquire fine-grained semantics and deliver effective supervision for reasoning. To the best of our knowledge, this is the first attempt to employ LLMs to guide the learning of reasoning networks for multimodal intent recognition. (2) We propose a Semantic Relational Reasoning module that formally establishes three logic-driven relations and captures dynamic interactions among nuanced semantics, enabling structured and interpretable enhancement of multimodal reasoning. (3) We conduct extensive experiments on two challenging datasets for multimodal intent and dialogue act recognition respectively, achieving state-of-the-art performance.

## 2 Related Works

### 2.1 Multimodal Intent Recognition

Multimodal intent recognition is crucial for understanding human behavior by integrating verbal and nonverbal cues to capture real-world intents. While early datasets focus on simple semantic tasks (Kruk et al., 2019; Saha et al., 2020), MIntRec (Zhang et al., 2022) advances the field through its diverse, fine-grained collection of multimodal samples, setting the first benchmark for multimodal intent recognition. Expanding on this, MIntRec2.0 (Zhang et al., 2024a) increases the data scale and label diversity, while establishing multimodal fusion techniques (Zadeh et al., 2017; Liu et al., 2018; Zadeh et al., 2018; Hazarika et al., 2020; Tsai et al., 2019; Rahman et al., 2020) as baselines, thereby offering a more robust foundation for advancing intent recognition in complex contexts.

Recently, specialized intent recognition models have emerged to tackle unique challenges in this area. TCL-MAP (Zhou et al., 2024) focuses on fusion strategies, using token-level contrastive learning and modality-aware prompts to enhance semantic depth. SDIF-DA (Huang et al., 2024) employs a shallow-to-deep interaction module to align modalities across various levels, yielding high-quality fusion features. Additionally, CAGC (Sun et al., 2024) uses video context to address perception biases and reduce intent uncertainty from multimodal inconsistencies, while InMu-Net (Zhu et al., 2024) applies an information bottleneck and multi-objective optimization to filter noise and redun-

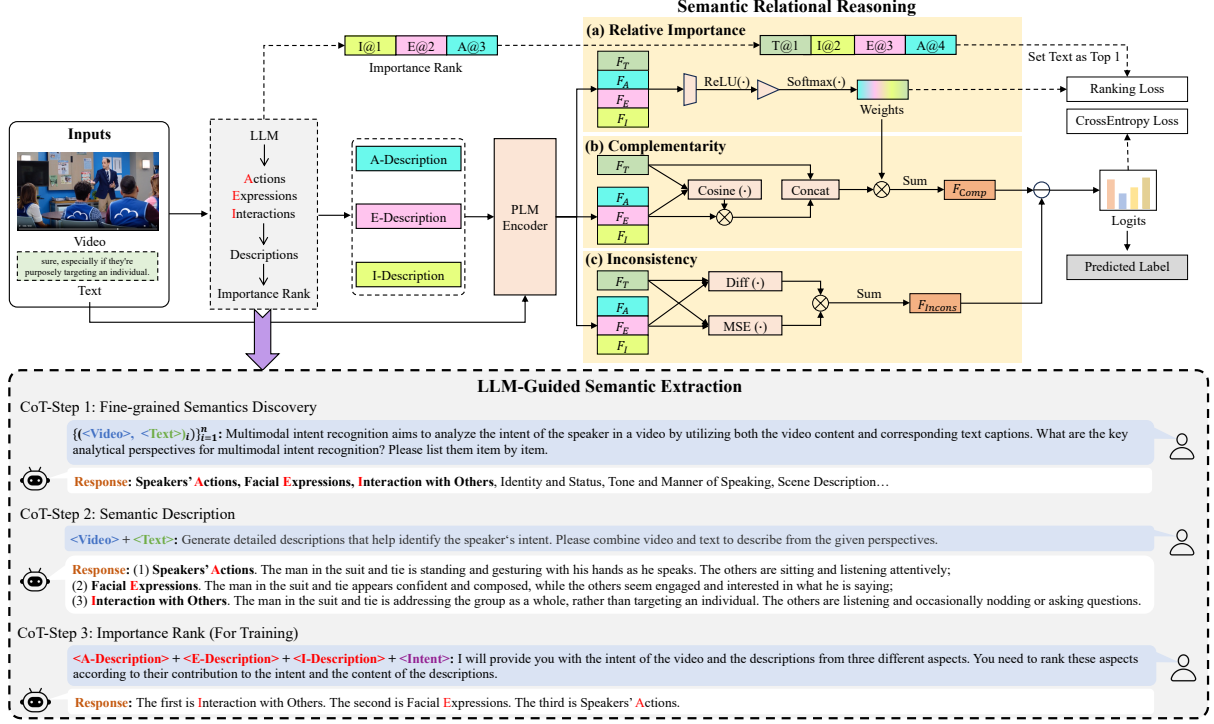


Figure 1: Overall architecture of our proposed LLM-guided semantic relational reasoning (LGSRR) method.

dancy in nonverbal data. MIntOOD (Zhang et al., 2024b) introduces multi-granularity optimized objectives and employs dynamic weight fusion to enhance the robustness of multimodal representation.

## 2.2 Multimodal Large Language Models

Multimodal large language models (MLLMs) build on the success of LLMs in natural language processing by extending their capabilities to multimodal understanding. Early MLLMs primarily focus on aligning nonverbal modalities with LLMs' input space. For instance, Flamingo (Alayrac et al., 2022) introduces gated cross-attention to handle interleaved multimodal data, while BLIP-2 (Li et al., 2023) uses a Q-Former to map visual representations for integration with LLMs. Advanced video-capable MLLMs (Li et al., 2024; Zhang et al., 2023; Ataallah et al., 2024) address the challenge of extracting essential information from extensive visual content. VideoLLaMA2 (Cheng et al., 2024) employs spatial-temporal convolutions to capture dynamic visual details, while Qwen2-VL (Wang et al., 2024a) and LLaVA-Next-Video (Zhang et al., 2024c) implement techniques like dynamic resolution and linear scaling for improved handling of varied frame sizes and longer videos. Recent models further leverage reasoning strategies to enhance understanding. Techniques such as CoT prompting (Wei et al., 2022) improve complex reasoning,

and models like HuggingGPT (Shen et al., 2023) and VideoAgent (Wang et al., 2024b) introduce agent-based planning, allowing the LLM to select or retrieve relevant frames and expert models, enhancing multimodal comprehension.

## 3 Methodology

### 3.1 LLM-Guided Semantic Extraction

Given raw video  $V$  and text  $T$ , our goal is to extract fine-grained semantic features that closely align with intents for nuanced multimodal reasoning. To this end, we design a shallow-to-deep CoT consisting of the following three progressive steps. Details of CoT design are shown in Appendix C.

#### CoT-Step 1: Fine-grained Semantics Discovery.

To determine essential fine-grained semantics, we begin by randomly selecting a subset of samples  $B = \{(T_i, V_i) \mid i \in \{1, 2, \dots, n\}\}$  as background knowledge, in which each sample pair consists of corresponding text  $T_i$  and video  $V_i$ . We then design Template<sub>1</sub> to prompt GPT-3.5 (Zhou et al., 2023) to analyze these samples within the task context, and autonomously uncover salient semantic aspects, yielding an initial set  $S_{init}$  of fine-grained concepts such as Speakers' Actions (A), Facial Expressions (E), Interactions with Others (I), Identity and Status (IS), Tone and Manner of Speaking

(TMS), Scene Description (SD), among others.

$$S_{\text{init}} = \text{GPT-3.5}(B; \text{Template}_1). \quad (1)$$

From the set, we select top- $K$  frequent cues to form the refined set  $S$  for further analysis, where the choice of  $K = 3$  is supported by the detailed comparison in Appendix B.

**CoT-Step 2: Semantic Description.** Building on the fine-grained semantic set  $S = \{A, E, I\}$ , we employ VideoLLaMA2 (Cheng et al., 2024) to extract semantic descriptions  $D_M$  for each concept  $M \in S$ . To ensure both descriptive quality and conceptual relevance, we craft structured prompts  $\text{Template}_2$  specifically tailored to each semantic.

$$D = \text{VideoLLaMA2}(V, T; S; \text{Template}_2). \quad (2)$$

Consequently, we obtain high-quality descriptions  $D = \{D_A, D_E, D_I\}$  for intent-related semantics, serving as a rich semantic foundation

**CoT-Step 3: Importance Rank.** To further incorporate LLM guidance into reasoning, we analyze the connections among the fine-grained semantics  $S$  derived from the previous steps. Specifically, to bridge the comprehension gap between LLM and lightweight reasoning model, GPT-3.5 is utilized to generate a generalizable ranking of semantic contributions based on the ground-truth label  $y$ , using structured instructions  $\text{Template}_3$ :

$$S_{\text{rank}} = \text{GPT-3.5}(D; y; \text{Template}_3), \quad (3)$$

where the label  $y$  is used only during training, allowing GPT-3.5 to perform abductive reasoning and provide supervisory signals that guide the reasoning module in evaluating semantic significance. The ranking statistics are presented in Appendix D.

### 3.2 Semantic Relational Reasoning

To move beyond simple fusion, we extend three core logical operations (“or,” “and,” and “not”) to semantic level, forming relations of relative importance, complementarity, and inconsistency. These relations underpin our Semantic Relational Reasoning module, capturing complex interactions for nuanced multimodal intent understanding. We argue that these core relational structures effectively represent the intricate dynamics necessary for robust intent recognition, drawing from logical reasoning principles in which complex relational patterns are built on basic operations.

For feature extraction from the input text and fine-grained semantic descriptions, we use BERT

(Devlin et al., 2019), a pre-trained language model commonly applied in intent recognition (Zhang et al., 2022). To unify the feature space, we encode the text and semantic descriptions separately with BERT, accounting for stylistic differences. Specifically, given the text  $T$  and the concatenated semantic descriptions  $[D_A, D_E, D_I]$ , we obtain their corresponding token representations as follows:

$$Z_{\text{text}} = \text{BERT}(T), \quad Z_{\text{desc}} = \text{BERT}(D), \quad (4)$$

where  $Z_{\text{text}}$  and  $Z_{\text{desc}}$  are token embeddings with dimensions  $\mathbb{R}^{(l_T+1) \times d_T}$  and  $\mathbb{R}^{(3 \times l_D+1) \times d_T}$ , respectively. After encoding, we separate embeddings for actions  $Z_A$ , expressions  $Z_E$ , and interactions  $Z_I$ . To acquire reasoning features, we apply mean pooling over the token embeddings:

$$F_T = \text{Mean-Pooling}(Z_{\text{text}}), \quad (5)$$

$$F_M = \text{Mean-Pooling}(Z_M), \quad M \in \{A, E, I\}. \quad (6)$$

**Relative Importance.** To account for the “or” operation among semantic components and highlight varying contributions, we apply weighted importance scores, a common approach in multimodal intent recognition (Rahman et al., 2020). Due to the limited supervision for these scores, we employ LLM-derived rankings with a ranking loss, NeuralNDCG (Pobrotyn and Białobrzewski, 2021), to capture each semantic element’s relative significance. Specifically, we use a unified weight network with two linear layers, ReLU, and Softmax activations to produce normalized importance scores  $\alpha_{T,A,E,I}$  for each semantic feature:

$$h_{\{T,A,E,I\}} = \text{ReLU}(W_1 F_{\{T,A,E,I\}} + b_1), \quad (7)$$

$$\alpha_{\{T,A,E,I\}} = \text{Softmax}(W_2 h_{\{T,A,E,I\}} + b_2), \quad (8)$$

where  $W_1$ ,  $W_2$ ,  $b_1$ , and  $b_2$  are parameters. Given the central role of text, we prioritize the text feature  $F_T$ , assigning it the top rank in  $R = \{R_T, R_{M_1}, R_{M_2}, R_{M_3}\}$ , where  $M_i$  denotes other semantic features. We then apply NeuralNDCG loss to align learned importance scores with LLM-derived rankings, formalized as:

$$\mathcal{L}_{\text{rank}} = \frac{1}{N_R} \sum_{j \in R} \text{scale}(\hat{P})_j g(\alpha_j) d(j), \quad (9)$$

where  $N$  is the size of  $R$ ,  $g(\alpha_j)$  is the gain function,  $d(j)$  is the rank discount, and  $\text{scale}(\hat{P})_j$  is the softmax-based similarity matrix, with a detailed explanation available in Appendix E.



**Complementarity.** In the intricate landscape of multimodal interactions, semantic components inherently possess complementary features that actively reinforce and validate one another, playing a pivotal role in decoding composite meanings. To capture this, we extend the logical “and” operation to emphasize complementarity between text and other semantic features. Specifically, we calculate the cosine similarity between the text feature  $F_T$  and each fine-grained feature  $F_M$ , capturing cross-modal complementarity for richer representations. The complementarity score  $\beta_{T,M}$  is given by:

$$\beta_{T,M} = \frac{F_T \cdot F_M}{\|F_T\| \cdot \|F_M\|}. \quad (10)$$

To integrate complementarity relationship within the semantic space, we weight fine-grained semantic feature  $F_M$  by its complementarity score  $\beta_{T,M}$ , resulting in an enhanced feature representation:

$$C_M = \beta_{T,M} \cdot F_M. \quad (11)$$

We then combine relative importance and complementarity by concatenating each complementarity-enhanced feature  $C_M$  with  $F_T$  and applying weighted averaging using importance scores for an integrated representation:

$$F_{\text{Comp}} = \sum_M \alpha_M \cdot \text{Concat}(F_T, C_M). \quad (12)$$

**Inconsistency.** Modeling inconsistency among semantics is vital to understanding complex multimodal intents. For example, text might imply the *Inform* intent, while gestures or expressions suggest *Joke*. To capture such nuances, we extend the logical “not” operation to identify inconsistencies, enabling the model to interpret conflicting signals. Aligned with complementarity, we examine  $F_T$  and other semantic features  $F_M$ , creating an inconsistency penalty by calculating their differences:

$$I_M = F_T - F_M. \quad (13)$$

On the relational side, we use mean squared error to quantify this divergence, yielding inconsistency score  $\gamma_{T,M}$  to reflect the degree of misalignment:

$$\gamma_{T,M} = \frac{1}{d} \sum_{i=1}^d (F_T^{(i)} - F_M^{(i)})^2, \quad (14)$$

where  $d$  is the dimension of  $F_M$ . Finally, we obtain the combination by weighting each  $I_M$  with

its corresponding score  $\gamma_{T,M}$ , yielding the overall penalty feature  $F_{\text{Incons}}$ :

$$F_{\text{Incons}} = \sum_M \gamma_{T,M} \cdot I_M. \quad (15)$$

This composite penalty feature  $F_{\text{Incons}}$  captures complex semantic contradictions, enabling the model to identify subtle inconsistencies within multimodal intents, thereby refining its interpretative accuracy in complex contexts.

### 3.3 Training Objective

For classification, we calculate predicted output  $\hat{y}$  with the obtained features and apply cross-entropy loss to optimize the model under the supervision of multi-class labels:

$$\hat{y} = W(F_{\text{Comp}} - F_{\text{Incons}}) + b, \quad (16)$$

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N} \sum_{i=1}^N \sum_{c \in \mathcal{Y}} y_i^c \log(\text{Softmax}(\hat{y}_i)^c), \quad (17)$$

where  $N$  is the batch size and  $\mathcal{Y} = \{0, 1, \dots, K-1\}$  denotes the label set. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{rank}}, \quad (18)$$

where  $\lambda$  denotes the weight parameter.

## 4 Experiments

### 4.1 Experimental Setting

**Datasets.** MIntRec2.0 is a large-scale dataset with 30 fine-grained intent labels spanning text, video, and audio modalities. We adopt the original split with 6,165 training samples, 1,106 for validation, and 2,033 for test. For dialogue act classification, IEMOCAP-DA offers 12 annotated labels across the same three modalities, with 6,590 training samples, 942 for validation, and 1,884 for test.

**Baselines.** We compare LGSRR with state-of-the-art methods for multimodal intent recognition and dialogue act classification: (1) MulT (Tsai et al., 2019) introduces directional cross-modal attention to model interactions between modalities without requiring strict alignment; (2) MISA (Hazarika et al., 2020) projects each modality into modality-specific and modality-invariant subspaces, followed by self-attention for effective fusion; (3) MAG-BERT (Rahman et al., 2020) incorporates a multimodal adaptive gating mechanism that adjusts the text representation in semantic space using offsets computed from nonverbal

Methods	MIntRec2.0						IEMOCAP-DA					
	ACC (↑)	F1 (↑)	P (↑)	R (↑)	WF1 (↑)	WP (↑)	ACC (↑)	F1 (↑)	P (↑)	R (↑)	WF1 (↑)	WP (↑)
MISA	55.16	49.51	51.80	49.92	55.05	57.06	73.76	72.26	73.03	72.51	73.59	73.87
MAG-BERT	60.38	54.74	57.51	54.54	59.61	60.00	74.25	72.07	73.18	72.33	74.03	74.28
MuT	<b>60.66</b>	54.12	58.02	53.77	59.55	60.12	73.74	72.28	73.40	72.21	73.66	74.15
TCL-MAP	58.24	52.25	54.28	52.41	57.24	57.55	74.37	72.63	74.02	72.39	74.21	74.76
SDIF-DA	58.06	51.95	53.17	52.16	57.47	57.85	74.19	72.34	73.76	72.39	74.04	74.77
MIntOOD	58.25	51.73	56.79	50.99	57.11	58.65	74.56	71.31	72.70	70.89	74.40	74.65
LGSRR	<u>60.46</u>	<b>55.35</b>	<b>59.33</b>	<b>55.09</b>	<b>59.72</b>	<b>60.85</b>	<b>74.95</b>	<b>72.99</b>	<b>74.27</b>	<b>72.74</b>	<b>74.88</b>	<b>75.47</b>

Table 1: Main Results comparing LGSRR with baselines on the MIntRec2.0 and IEMOCAP-DA datasets.

modalities; (4) TCL-MAP (Zhou et al., 2024) leverages token-level contrastive learning to enhance the textual modality by integrating visual and acoustic information, thereby promoting semantic acquisition and multimodal integration; (5) SDIF-DA (Huang et al., 2024) utilizes a shallow-to-deep interaction module that aligns and fuses modalities at both shallow and deep levels, capturing relations on all granularities; (6) MIntOOD (Zhang et al., 2024b) employs a weighted feature fusion network to effectively model multimodal representations across multiple optimization granularities.

**Evaluation Metrics.** We evaluate model performance using following metrics: accuracy (ACC), F1-score (F1), precision (P), recall (R), weighted F1-score (WF1), and weighted precision (WP), where higher values indicate better performance.

**Implementation details.** For the LLM-Guided Semantic Extraction module, We utilize GPT-3.5-turbo for the first and third step and VideoLLaMA2-7B-16F for the second step. For feature extraction, the sequence lengths  $l_T$  and  $l_D$  are set as follows: (30, 50) for MIntRec2.0 and (70, 50) for IEMOCAP-DA, with a feature dimension  $d_T$  of 1024. The training process includes 100 epochs, with a batch size of 32. We employ the PyTorch library via HuggingFace (Wolf et al., 2020) for the pre-trained BERT model, optimized using AdamW (Loshchilov and Hutter, 2019) with learning rates of (6e-6, 1e-5) for the respective datasets. For consistency, all reported results are the averages of five runs, using random seeds from 0 to 4, conducted on NVIDIA Tesla V100-SXM2s. The training cost is discussed in Appendix F.

## 4.2 Results

**Results on the MIntRec2.0 Dataset.** As illustrated in Table 1, our LGSRR consistently surpasses existing SOTA methods, achieving the highest performance across five key metrics. In particular, LGSRR improves precision by 1.31% over the top baseline, attributed to its ability to capture fine-

grained semantics and perform effective relational reasoning. Moreover, LGSRR demonstrates substantial gains across the F1, R, and WP metrics with respective enhancements of 0.61%, 0.55%, and 0.73%, further validating the effectiveness and robustness of our approach. Although MuT achieves a comparable ACC score, LGSRR maintains a clear edge across all other metrics, underscoring its comprehensive superiority in semantic understanding. The experimental results strongly demonstrate that our method excels in recognizing complex intent within multimodal scenarios by effectively leveraging fine-grained semantic relations.

**Results on the IEMOCAP-DA Dataset.** To comprehensively assess LGSRR’s effectiveness in various multimodal semantic tasks, we conduct experiments on the IEMOCAP-DA dataset. As shown in Table 1, LGSRR exceeds state-of-the-art methods across all six metrics, demonstrating its robust capacity for nuanced semantic understanding. Notably, LGSRR achieves substantial gains with improvements of 0.58%, 0.67%, and 0.70% in ACC, WF1 and WP, respectively, which underscores its ability to capture ambiguous concepts like dialogue acts. These results affirm LGSRR’s strong generalizability and position it as a promising framework for advancing multimodal semantic understanding.

## 4.3 Ablation Study

We evaluate the effects of key components, including the LLM-Guided Semantic Extraction module (LGSE), the ranking loss ( $\mathcal{L}_{\text{rank}}$ ) and the Semantic Relational Reasoning module (SRR). Additionally, we also compare the performance of employing different MLLMs (VideoLLaMA (Zhang et al., 2023) and Qwen2-VL (Wang et al., 2024a)) in LGSE. The results are shown in Table 2, confirming the individual contributions of the modules to the overall performance. The ablation study on the semantic relations within SRR is provided in Appendix G.

First, by removing the LGSE module, we utilize coarse-grained modality-level features from text,

Ablation	MIntRec2.0						IEMOCAP-DA					
	ACC (↑)	F1 (↑)	P (↑)	R (↑)	WF1 (↑)	WP (↑)	ACC (↑)	F1 (↑)	P (↑)	R (↑)	WF1 (↑)	WP (↑)
w / o LGSE	58.57	52.83	54.37	52.72	58.22	58.69	74.55	70.40	71.32	70.64	74.40	74.73
LGSE (VideoLLaMA)	<u>60.27</u>	<u>54.91</u>	<u>57.07</u>	<u>54.78</u>	<u>59.56</u>	<u>60.00</u>	74.26	71.62	72.51	71.46	74.14	74.55
LGSE (Qwen2-VL)	59.59	54.62	56.63	54.58	58.89	59.20	74.54	<u>72.19</u>	<u>73.05</u>	<u>72.43</u>	74.42	<u>75.05</u>
w / o $\mathcal{L}_{\text{rank}}$	58.55	53.30	55.43	53.23	58.07	58.98	73.69	71.50	72.38	71.63	73.54	74.02
w / o SRR	60.04	54.31	55.98	54.47	59.34	59.82	<u>74.57</u>	71.80	71.82	<u>72.43</u>	<u>74.44</u>	74.63
Full	<b>60.46</b>	<b>55.35</b>	<b>59.33</b>	<b>55.09</b>	<b>59.72</b>	<b>60.85</b>	<b>74.95</b>	<b>72.99</b>	<b>74.27</b>	<b>72.74</b>	<b>74.88</b>	<b>75.47</b>

Table 2: Ablation studies on the MIntRec2.0 and IEMOCAP-DA datasets.

MLLMs	MIntRec2.0						IEMOCAP-DA					
	ACC (↑)	F1 (↑)	P (↑)	R (↑)	WF1 (↑)	WP (↑)	ACC (↑)	F1 (↑)	P (↑)	R (↑)	WF1 (↑)	WP (↑)
Qwen2-VL	<u>28.63</u>	7.18	8.68	7.74	<u>28.94</u>	<u>37.53</u>	<u>23.83</u>	<u>7.39</u>	10.14	9.43	<u>20.10</u>	<u>37.09</u>
MiniCPM-o 2.6	17.46	<u>14.75</u>	<u>27.32</u>	<u>14.97</u>	17.98	37.11	14.01	6.93	<u>10.65</u>	<u>11.62</u>	13.56	36.29
VideoLLaMA2	17.17	4.95	12.32	4.65	13.10	25.55	15.02	2.97	7.05	4.11	7.59	29.77
LGSRR	<b>60.46</b>	<b>55.35</b>	<b>59.33</b>	<b>55.09</b>	<b>59.72</b>	<b>60.85</b>	<b>74.95</b>	<b>72.99</b>	<b>74.27</b>	<b>72.74</b>	<b>74.88</b>	<b>75.47</b>

Table 3: Comparison of LGSRR with MLLMs across both datasets, with MLLMs evaluated via direct inference.

video, and audio extracted via BERT, Swin Transformer (Liu et al., 2021), and WavLM (Chen et al., 2022), respectively, to perform relational reasoning. This leads to substantial declines in all metrics across both datasets, with F1, P, and R scores dropping by over 2%, demonstrating the improved alignment of fine-grained semantics with intent by effectively narrowing the semantic gap. When employing different MLLMs in LGSE, we observe no significant performance drop across most metrics, demonstrating the generalizability of our approach. Besides, removing the ranking loss ( $\mathcal{L}_{\text{rank}}$ ) causes noticeable declines across metrics, underscoring the effectiveness of LLM guidance in capturing semantic feature importance and enhancing relational reasoning. Finally, replacing the SRR module with a simple summation-based fusion retains competitive ACC scores, suggesting the robustness of other components. However, declines in other metrics emphasize the SRR module’s role in refining semantic interactions and enhancing feature synergy, affirming each component’s unique contribution to nuanced multimodal understanding.

#### 4.4 Comparison to Frozen MLLMs

To substantiate the rationale and effectiveness of the proposed LGSRR framework, we benchmark it against leading MLLMs on the MIntRec2.0 and IEMOCAP-DA datasets. To maintain consistency with the MLLM in LGSRR, we evaluate Qwen2-VL, MiniCPM-O 2.6 (Yao et al., 2024), and VideoLLaMA2 in their 7B configurations using zero-shot reasoning without fine-tuning.

As Table 3 shows, the experimental results on both datasets underscore the remarkable superior-

MLLMs	IEMOCAP-DA					
	ACC (↑)	F1 (↑)	P (↑)	R (↑)	WF1 (↑)	WP (↑)
Qwen2-VL	68.10	59.85	56.91	66.75	67.09	68.72
VideoLLaMA2	71.07	68.57	71.90	68.91	71.32	73.62
LGSRR	<b>74.95</b>	<b>72.99</b>	<b>74.27</b>	<b>72.74</b>	<b>74.88</b>	<b>75.47</b>

Table 4: Comparison of LGSRR with MLLMs on IEMOCAP-DA, with MLLMs evaluated via SFT.

ity of our proposed method over the state-of-the-art MLLMs. On the MIntRec2.0 dataset, LGSRR achieves an impressive 30% improvement over the best-performing MLLM on all metrics other than WP, demonstrating its exceptional ability to model complex multimodal semantics. Likewise, on the IEMOCAP-DA dataset, LGSRR consistently outperforms MLLMs by over 35% across all evaluation metrics, further affirming its robustness and adaptability. These results highlight LGSRR as a groundbreaking advancement in multimodal semantic understanding, offering a more efficient, scalable, and generalizable solution compared to MLLM-based approaches.

#### 4.5 Comparison to Fine-tuned MLLMs

To further evaluate the practicality of our proposed method, we compared LGSRR with two cutting-edge MLLMs under the supervised fine-tuning settings on IEMOCAP-DA, as shown in Table 4.

The results indicate that our method achieves SOTA performance across all metrics, showcasing its strong comprehensive capabilities. Specifically, compared to VideoLLaMA2 and Qwen2-VL, LGSRR yielded performance improvements of 3.88% to 6.85% and 4.42% to 13.14% on ACC and F1, highlighting our method’s prominent capability for modeling intricate semantic relations.

Methods	MIntRec2.0					
	ACC ( $\uparrow$ )	F1 ( $\uparrow$ )	P ( $\uparrow$ )	R ( $\uparrow$ )	WF1 ( $\uparrow$ )	WP ( $\uparrow$ )
Or	59.46	53.51	54.69	53.44	58.90	59.12
And	<u>59.76</u>	54.08	56.07	54.01	59.02	59.36
Not	59.53	<u>54.58</u>	55.98	54.46	59.04	59.31
Combination	59.63	54.13	<u>56.77</u>	<u>54.95</u>	<u>59.06</u>	<u>59.41</u>
LGSRR	<b>60.46</b>	<b>55.35</b>	<b>59.33</b>	<b>55.09</b>	<b>59.72</b>	<b>60.85</b>

Table 5: Comparison of LGSRR with classic relations on the MIntRec2.0 dataset.

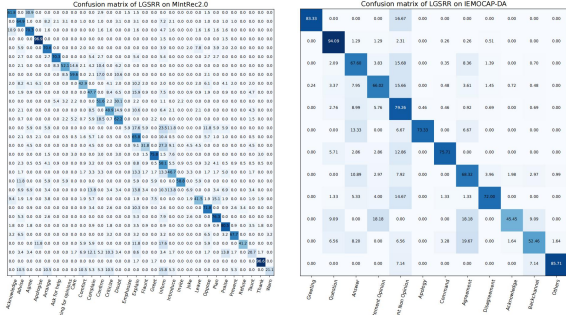


Figure 2: Confusion matrices on both datasets.

Moreover, the performance gains of LGSRR on P and R metrics, ranging from 2.37% to 17.36% and 3.83% to 5.99%, underscore the superiority of LGSRR in capturing deep semantic associations. Meanwhile, the stable performance on WF1 and WP metrics confirms the robustness of LGSRR when faced with imbalanced datasets. Overall, our method not only delivers more competitive performance but also avoids the computational burden of supervised fine-tuning, providing strong support and valuable insights for the precise optimization of smaller models for vertical tasks.

#### 4.6 Comparison with Classic Relations

To compare with classic relations, we define the corresponding feature-level operations and evaluate them on MIntRec2.0, as shown in 5. Specifically, *Or* is implemented by summing all semantic features, where the presence of any single salient cue can contribute effectively to intent recognition. *And* is realized via the Hadamard product of semantic features, requiring all contributing cues to be jointly aligned with the target intent. *Not* captures semantic inconsistency by computing pairwise feature differences, concatenating them, and projecting through a nonlinear layer. The *Combination* setting integrates all three classic relation features by concatenation, followed by a nonlinear transformation to obtain a unified relational representation.

The experimental results demonstrate the com-

prising superiority over classic relations across all metrics. Specifically, LGSRR outperforms the top classic relation method, achieving gains of 0.70% in ACC and 0.77% in F1, respectively. This reveals the enhanced feature representation capability of our semantic relationship reasoning module in complex multimodal scenarios. The significant 2.56% performance increase on the P metric indicates that LGSRR is capable of identifying genuinely relevant semantic cues, thereby avoiding the misclassification of extraneous information as target intention features. Meanwhile, the model achieves performance improvements of 0.66% and 1.44% on WF1 and WP, respectively, which highlights that LGSRR can effectively alleviate the class imbalance problem. Overall, our semantic relational reasoning module outperforms classic relations methods, validating the efficacy of LGSRR in modeling fine-grained and intricate semantic interactions.

#### 4.7 Analysis of Classification Performance

To investigate the fine-grained performance of classification, we plot the confusion matrices across all classes of our method on the MIntRec2.0 dataset and IEMOCAP-DA dataset, as shown in Figure 2. Each score on the main diagonal represents the accuracy for the corresponding class.

On the MIntRec2.0 dataset, our method achieves over 70% accuracy in 9 out of 30 classes. On the IEMOCAP-DA dataset, this ratio increases to 7 out of 12, highlighting LGSRR’s superior performance. Specifically, LGSRR not only maintains high accuracy for simple intents like *Apologize* and *Thank*, but also excels in identifying complex intents like *Arrange*, *Ask for help* and *Plan*, which require integrating multiple semantic cues. This demonstrates LGSRR’s capability to handle various levels of intent complexity. Limited performance on classes such as *Joke* and *Emphasize* is primarily due to less data in these categories. For dialogue act classification, LGSRR performs well across most categories, with only slight underperformance in more ambiguous classes such as *Acknowledge* and *Backchannel*. These results validate that LGSRR excels in distinguishing both fine-grained intents and broad communicative semantics through fine-grained semantic extraction and advanced relational reasoning.

#### 4.8 Case Study

To offer deeper insight into semantic descriptions and rankings, we select three representative samples from the MIntRec2.0 dataset, covering diverse






Sample	Description	Rank
 <p>Good job!</p>	<p><b>Speakers' Actions:</b> The woman is <b>holding hands with the man</b>, while the other woman is standing nearby. They are all engaged in a conversation, with the woman in the middle speaking and gesturing with her hands.</p> <p><b>Facial Expressions:</b> The woman in the middle is <b>smiling</b> and appears to be <b>happy and relaxed</b>. The other two individuals are also smiling and seem to be enjoying the conversation.</p> <p><b>Interaction with Others:</b> The three individuals are <b>standing close to each other</b>, indicating a <b>friendly and intimate</b> atmosphere. They are all actively participating in the conversation, with the woman in the middle <b>taking the lead in speaking</b>.</p>	<p>1<sup>st</sup>: Interaction with Others</p> <p>2<sup>nd</sup>: Facial Expressions</p> <p>3<sup>rd</sup>: Speakers' Actions</p> <p>Label: Praise</p>
 <p>An employee's inventing fake customers.</p>	<p><b>Speakers' Actions:</b> The man in the white shirt with a blue tie is seen <b>pointing his finger at the other man</b>, who is wearing a blue shirt. He then proceeds to talk to the camera, gesturing with his hands. The man in the blue shirt listens attentively and occasionally nods his head.</p> <p><b>Facial Expressions:</b> The man in the white shirt with a blue tie appears to be <b>angry and frustrated</b>, as evidenced by his facial expressions and gestures. The man in the blue shirt maintains a calm and composed demeanor throughout the video.</p> <p><b>Interaction with Others:</b> The two men seem to have a <b>professional relationship</b>, as they engage in a conversation <b>in a business setting</b>. The man in the white shirt with a blue tie appears to be <b>in a position of authority</b>, as he is addressing the camera and the other man.</p>	<p>1<sup>st</sup>: Interaction with Others</p> <p>2<sup>nd</sup>: Speakers' Actions</p> <p>3<sup>rd</sup>: Facial Expressions</p> <p>Label: Criticize</p>
 <p>Oh, look! It's a two-for-one sale!</p>	<p><b>Speakers' Actions:</b> The woman is <b>taking a picture of herself using her phone</b>. She is <b>holding the phone in front of her face</b> and pointing it towards herself.</p> <p><b>Facial Expressions:</b> The woman appears to be smiling and enjoying herself while taking the picture. Her facial expression is <b>relaxed and happy</b>.</p> <p><b>Interaction with Others:</b> There is no interaction with others in this video. The woman is alone and <b>focused on taking her picture</b>.</p>	<p>1<sup>st</sup>: Speakers' Actions</p> <p>2<sup>nd</sup>: Facial Expressions</p> <p>3<sup>rd</sup>: Interaction with Others</p> <p>Label: Introduce</p>

Figure 3: Examples from the MIntRec2.0 dataset, showcasing descriptions and rankings of fine-grained semantics.

scenarios, multiple characters, varied emotions and distinct intents, as shown in Figure 3. Each sample is presented with the raw data, detailed descriptions from three semantic perspectives and contribution rankings. Additional case studies from both datasets are provided in Appendix H.

In terms of fine-grained semantic descriptions, our method accurately identifies significant details, including expressions like “smiling”, actions like “taking a picture of herself using her phone” and interactions like “standing close to each other”. Beyond basic cues, it also handles complex scenarios, such as ambiguous hand gestures and varying numbers of individuals, enabled by the effective CoT mechanism which significantly boosts the MLLM’s generative capabilities for nuanced semantics. For semantic ranking, our method leverages the LLM’s abductive reasoning ability to weigh the contribution of each fine-grained semantic relative to the true intent. In the first two examples, interpersonal interactions are prioritized consistent with the interaction concepts defined in the intent label (Zhang et al., 2024a), with actions and expressions ranked based on their direct relevance. In the third example, where a shopkeeper films a product introduction alone, the ranking correctly emphasizes the action, further demonstrating the strong capability of our semantic ranking approach.

## 5 Conclusion

In this paper, we present the LLM-Guided Semantic Relational Reasoning (LGSRR) framework

to tackle the challenges of fine-grained semantic extraction and relational reasoning in multimodal intent recognition. Utilizing a shallow-to-deep CoT strategy, LGSRR harnesses the LLMs to autonomously uncover detailed semantics across modalities. By capturing logic-inspired relational patterns from logical, LGSRR effectively models intricate semantic relations to achieve superior representations. Our work not only demonstrates notable improvements across challenging multimodal classification tasks, but also carries significant implications for advancing LLM-guided frameworks in complex semantic understanding.

## 6 Limitations

This work has two primary limitations that warrant careful consideration. First, despite promising experimental results, the performance on these datasets still indicates substantial room for improvement, given the inherent complexity and variability of multimodal data. Second, although the study formally models basic semantic relations from a logical perspective, it does not fully account for the nuanced and context-dependent interactions in real-world scenarios. Future research could address these gaps by exploring more expressive relational structures or integrating adaptive reasoning mechanisms to better capture semantic complexity.

## 7 Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No.62173195).

## References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, and 8 others. 2022. [Flamingo: a visual language model for few-shot learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.
- Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. 2024. [Minigt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens](#). *Preprint*, arXiv:2404.03413.
- Remi Cadene, Hedi Ben-younes, Matthieu Cord, and Nicolas Thome. 2019. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. [Wavlm: Large-scale self-supervised pre-training for full stack speech processing](#). *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.
- Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. 2019. Graph-based global reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024. [Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms](#). *Preprint*, arXiv:2406.07476.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yan Fan, Chengyu Wang, Peng He, and Yunhua Hu. 2022. [Building multi-turn query interpreters for e-commercial chatbots with sparse-to-dense attentive modeling](#). In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, WSDM '22*, page 1577–1580, New York, NY, USA. Association for Computing Machinery.
- Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. 2024. Question aware vision transformer for multimodal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13861–13871.
- Aditya Grover, Eric Wang, Aaron Zweig, and Stefano Ermon. 2019. [Stochastic optimization of sorting networks via continuous relaxations](#). In *International Conference on Learning Representations*.
- Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. [Misa: Modality-invariant and -specific representations for multimodal sentiment analysis](#). In *Proceedings of the 28th ACM International Conference on Multimedia, MM '20*, page 1122–1131, New York, NY, USA. Association for Computing Machinery.
- Shijue Huang, Libo Qin, Bingbing Wang, Geng Tu, and Ruifeng Xu. 2024. [Sdif-da: A shallow-to-deep interaction framework with data augmentation for multimodal intent detection](#). In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10206–10210.
- Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.
- Ding Jiang and Mang Ye. 2023. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2787–2797.
- Sepideh Kaffash, An Truong Nguyen, and Joe Zhu. 2021. [Big data algorithms and applications in intelligent transportation system: A review and bibliometric analysis](#). *International Journal of Production Economics*, 231:107868.
- Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. [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 (EMNLP-IJCNLP)*, pages 4622–4632, Hong Kong, China. Association for Computational Linguistics.
- Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. [Lisa: Reasoning segmentation via large language model](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9579–9589.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. [BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202

- of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- KunChang Li, Yanan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2024. [Videochat: Chat-centric video understanding](#). *Preprint*, arXiv:2305.06355.
- Yabo Liu, Jinghua Wang, Chao Huang, Yaowei Wang, and Yong Xu. 2023. Cigar: Cross-modality graph reasoning for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23776–23786.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. [Swin transformer: Hierarchical vision transformer using shifted windows](#). In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002.
- Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Jinpeng Mi, Song Tang, Zhen Deng, Michael Görner, and Jianwei Zhang. 2019. [Object affordance based multimodal fusion for natural human-robot interaction](#). *Cogn. Syst. Res.*, 54:128–137.
- Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. 2022. [Multimodal understanding and generation for medical images and text via vision-language pre-training](#). *IEEE Journal of Biomedical and Health Informatics*, 26(12):6070–6080.
- Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Sheuli Paul, Michael Sintek, Veton Këpuska, Marius Silaghi, and Liam Robertson. 2022. [Intent based multimodal speech and gesture fusion for human-robot communication in assembly situation](#). In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 760–763.
- Przemysław Pobrotyn and Radosław Białobrzęski. 2021. [NeuralIndcg: Direct optimisation of a ranking metric via differentiable relaxation of sorting](#). *Preprint*, arXiv:2102.07831.
- Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. [Integrating multimodal information in large pretrained transformers](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online. Association for Computational Linguistics.
- Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. [Towards emotion-aided multimodal dialogue act classification](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372, Online. Association for Computational Linguistics.
- Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. [Hugging-gpt: Solving ai tasks with chatgpt and its friends in hugging face](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 38154–38180. Curran Associates, Inc.
- Kaili Sun, Zhiwen Xie, Mang Ye, and Huyin Zhang. 2024. Contextual augmented global contrast for multimodal intent recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26963–26973.
- Pallavi Tiwari, Bhaskar Pant, Mahmoud M. Elarabawy, Mohammed Abd-Elnaby, Noor Mohd, Gaurav Dhimman, and Subhash Sharma. 2022. [Cnn based multiclass brain tumor detection using medical imaging](#). *Computational Intelligence and Neuroscience*, 2022(1):1830010.
- Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. [Multimodal transformer for unaligned multimodal language sequences](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024b. Videoagent: Long-form video understanding with large language model as agent. *European Conference on Computer Vision (ECCV)*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.



- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Wei Xu. 2019. [Toward human-centered ai: a perspective from human-computer interaction](#). *Interactions*, 26(4):42–46.
- Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. 2024a. [Exploring chain-of-thought for multimodal metaphor detection](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 91–101, Bangkok, Thailand. Association for Computational Linguistics.
- Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. 2024b. [Exploring chain-of-thought for multimodal metaphor detection](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 91–101, Bangkok, Thailand. Association for Computational Linguistics.
- Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. [Mm-react: Prompting chatgpt for multimodal reasoning and action](#). *Preprint*, arXiv:2303.11381.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. [Minicpm-v: A gpt-4v level mllm on your phone](#). *arXiv preprint arXiv:2408.01800*.
- Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. [Tensor fusion network for multimodal sentiment analysis](#). *arXiv preprint arXiv:1707.07250*.
- Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. [Memory fusion network for multi-view sequential learning](#). In *Proc. AAAI Conf. Artif. Intell.*
- Hang Zhang, Xin Li, and Lidong Bing. 2023. [Video-LLaMA: An instruction-tuned audio-visual language model for video understanding](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore. Association for Computational Linguistics.
- Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, jinyue Zhao, Wenrui Li, and Yanting Chen. 2024a. [MIntrec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations](#). In *The Twelfth International Conference on Learning Representations*.
- Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. 2022. [Mintrec: A new dataset for multimodal intent recognition](#). In *Proceedings of the 30th ACM International Conference on Multimedia*, MM ’22, page 1688–1697. ACM.
- Hanlei Zhang, Qianrui Zhou, Hua Xu, Jianhua Su, Roberto Evans, and Kai Gao. 2024b. [Multimodal classification and out-of-distribution detection for multimodal intent understanding](#). *arXiv preprint arXiv:2412.12453*.
- Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024c. [Llava-next: A strong zero-shot video understanding model](#).
- Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, and Lichao Sun. 2023. [A comprehensive survey on pretrained foundation models: A history from bert to chatgpt](#). *Preprint*, arXiv:2302.09419.
- Qianrui Zhou, Hua Xu, Hao Li, Hanlei Zhang, Xiaohan Zhang, Yifan Wang, and Kai Gao. 2024. [Token-level contrastive learning with modality-aware prompting for multimodal intent recognition](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15):17114–17122.
- Zhihong Zhu, Xuxin Cheng, Zhaorun Chen, Yuyan Chen, Yunyan Zhang, Xian Wu, Yefeng Zheng, and Bowen Xing. 2024. [Inmu-net: Advancing multimodal intent detection via information bottleneck and multi-sensory processing](#). In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM ’24, page 515–524, New York, NY, USA. Association for Computing Machinery.

## A Related Work for Multimodal Reasoning

Multimodal reasoning serves as a cornerstone for advancing intelligent systems, empowering them to integrate and interpret information from diverse sources to tackle complex, real-world challenges. Early approaches primarily focus on attention-based mechanisms to extract essential features from multimodal data. For instance, DANs (Nam et al., 2017) employs joint visual and textual attention to capture interactions between modalities. To further improve reasoning abilities, alignment between textual and nonverbal modalities becomes a focal point. MUREL (Cadene et al., 2019) models intra-modal interactions through rich vectorial representations, and IRRA (Jiang and Ye, 2023)



leverages implicit fine-grained relation learning to align multimodal data effectively. Graph-based methods also gain traction for characterizing semantic information and relations across modalities. GloRe (Chen et al., 2019) proposes a global reasoning framework to learn relationships between distant regions, and CIGAR (Liu et al., 2023) combines linguistic and visual knowledge to construct multimodal graph representations. Recently, the reasoning capabilities of large language models (LLMs) are explored to tackle multimodal tasks. MM-REACT (Yang et al., 2023) utilizes ChatGPT to identify visual experts for task-specific solutions, while QA-ViT (Ganz et al., 2024) integrates query-related information into vision transformers to enhance reasoning capabilities. These advancements highlight the continuous evolution of multimodal reasoning techniques and their potential for addressing increasingly complex scenarios.

## B Selection of Semantic Aspects

To validate the effectiveness and rationale of our selected semantics, we conduct experiments by progressively incorporating additional semantic aspects, following their occurrence frequency, into the three most frequent categories: Speaker’s Actions (A), Facial Expressions (E), and Interactions with Others (I). The newly introduced semantic aspects include Tone and Manner of Speaking (TMS), Identity and Status (IS), and Scene Description (SD). The results with additional semantic aspects, compared against those from the main experiment, are presented in Table 6.

The experimental results clearly demonstrate that incorporating additional semantic aspects leads to a significant performance decline, reaffirming the validity of our selection strategy. Across both datasets, adding one, two, or all three new aspects results in a drop of over 1% across nearly all evaluation metrics. This suggests that human intent is primarily conveyed through facial expressions, actions, and social interactions, whereas incorporating excessive semantic aspects introduces redundancy or less relevant information, thereby diminishing the model’s ability to focus on essential multimodal cues. Furthermore, results from the IEMOCAP-DA dataset provide deeper insights into the relative impact of different semantic aspects. The negative effects of TMS and IS are more pronounced than those of SD, likely because TMS and IS capture intrinsic speaker character-

istics, which are challenging to infer from brief, single-turn conversations. In contrast, SD primarily characterizes the dialogue environment, posing minimal ambiguity but contributing less to intent recognition. These findings strongly confirm the adequacy of our selected semantic aspects, emphasizing the importance of carefully choosing those most relevant to multimodal intent recognition.

## C Details of CoT

This section details the progressive design of the CoT in LGSRR, providing a clearer illustration of the underlying reasoning mechanism. For CoT-Step 1, the prompt includes an introduction to the task background along with raw data descriptions, enabling the LLM to develop a more comprehensive understanding of the target semantics, as shown in Template<sub>1</sub>. It is then asked to list key analytical perspectives, prompting it to autonomously identify fine-grained semantic aspects that are crucial for multimodal understanding.

**Template<sub>1</sub>:** Multimodal intent recognition aims to analyze the intent of the speaker in a video by utilizing both the video content and corresponding text captions. What are the key analytical perspectives for multimodal intent recognition? Please list them.

For CoT-Step 2, we guide MLLMs to generate structured and detailed semantic descriptions across key semantic cues from the previous step, such as speakers’ actions, facial expressions, and interaction with others. This step ensures that the extracted semantics are not only modality-aware but also intent-relevant, laying a solid foundation for reasoning over semantic relationships.

**Template<sub>2</sub>:** Generate detailed descriptions that help identify the speaker’s intent. Please combine video and text to describe from the following perspectives: (1) Speakers’ Actions: <Action Instruction>; (2) Facial Expressions: <Expression Instruction>; (3) Interaction with Others: <Interaction Instruction>. Focus on these aspects to create a comprehensive description that would aid in recognizing the intentions behind the speakers’ actions and words.

To further incorporate LLM guidance in establishing reasoning relationships, we take fine-grained semantic descriptions as input and prompt the LLM to rank them according to the contribution

Settings	MIntRec2.0						IEMOCAP-DA					
	ACC (↑)	F1 (↑)	P (↑)	R (↑)	WF1 (↑)	WP (↑)	ACC (↑)	F1 (↑)	P (↑)	R (↑)	WF1 (↑)	WP (↑)
A+E+I+TMS	58.99	53.61	55.62	<u>53.70</u>	<u>58.27</u>	58.79	73.41	70.41	72.37	69.66	73.27	73.69
A+E+I+TMS+IS	<u>59.03</u>	53.26	56.24	53.22	<u>58.27</u>	<u>59.29</u>	73.24	69.69	70.69	69.41	73.09	73.28
A+E+I+TMS+IS+SD	59.00	<u>53.63</u>	<u>56.29</u>	53.25	58.25	58.94	<u>73.96</u>	<u>70.46</u>	<u>72.57</u>	<u>69.71</u>	<u>73.79</u>	<u>74.33</u>
LGSRR (A+E+I)	<b>60.46</b>	<b>55.35</b>	<b>59.33</b>	<b>55.09</b>	<b>59.72</b>	<b>60.85</b>	<b>74.95</b>	<b>72.99</b>	<b>74.27</b>	<b>72.74</b>	<b>74.88</b>	<b>75.47</b>

Table 6: Comparison of using different semantic aspects, where A, E, I, TMS, IS and SD represent Speaker’s Actions, Facial Expressions, Interaction with Others, Tone and Manner of Speaking, Identity and Status, and Scene Description, respectively.

Semantics	MIntRec2.0			IEMOCAP-DA		
	Rank@1	Rank@2	Rank@3	Rank@1	Rank@2	Rank@3
Speakers’ Actions	1,476	3,814	875	1,675	3,982	933
Facial Expressions	523	1,450	4,192	604	1,542	4,444
Interactions with Others	4,166	901	1,098	4,311	1,066	1,213

Table 7: The semantic ranking results of Speakers’ Actions, Facial Expressions, and Interactions with Others on the MIntRec2.0 and IEMOCAP-DA Dataset.

of each semantic description for CoT-Step 3. This ranking provides explicit supervision signals for the reasoning module to learn semantic importance in a more interpretable and intent-aware manner.

**Template<sub>3</sub>:** I will provide you with the intent of the video and descriptions of the video from three different semantics. You need to rank these semantics according to their contribution to the intent and the content of the descriptions. The three aspects are Speakers’ Actions, Facial Expressions and Interaction with Others. The intent of the video is <Intent>. The descriptions are <Action Description>, <Expression Description> and <Interaction Description>.

## D Statistics on Ranking Results

To thoroughly evaluate the significance of different fine-grained semantics, we analyze the semantic ranking results on the MIntRec2.0 and IEMOCAP-DA datasets, as illustrated in Table 7. The table summarizes the ranking distributions for Speakers’ Actions, Facial Expressions, and Interactions with Others across both datasets.

From the statistics, we observe a consistent trend across both datasets: Interactions with Others consistently has the highest number of Rank@1 samples, with counts of 4,166 and 4,311, followed by Speakers’ Actions with 1,476 and 1,675, while Facial Expressions ranks the lowest with 523 and 604. This ranking distribution reflects the interaction-centric nature of intent labels such as *criticize* and *question*, which are deeply rooted in social dynamics and align with real-world intent distributions.

Given the complexity of intent semantics, facial expressions generally serve as a coarse indicator of intents, whereas actions provide more decisive clues, as seen in intents like *Leave* or *Criticize*. When examining each fine-grained semantic, LGSRR effectively differentiates the importance of actions and expressions, as reflected in their distinct ranking distributions. For instance, in MIntRec2.0, the number of Rank@2 and Rank@3 samples for actions is 3,814 and 875, respectively, compared to 1,450 and 4,192 for expressions. These results underscore LGSRR’s nuanced understanding of semantic contributions and potential in handling complex multimodal semantic tasks.

## E Explanation of NeuralNDCG Loss

The NeuralNDCG loss (Pobrotyn and Białobrzewski, 2021) is a differentiable reformulation of the Normalized Discounted Cumulative Gain (NDCG) (Järvelin and Kekäläinen, 2002) metric, tailored for learning-to-rank tasks by aligning model optimization directly with ranking performance. It approximates the sorting operation using a normalized soft permutation matrix  $\hat{P}$  and integrates gain and discount functions to compute ranking quality. The loss is defined as:

$$\mathcal{L}_{\text{NeuralNDCG}} = \frac{1}{N_R} \sum_{j \in R} \text{scale}(\hat{P})_j g(s_j) d(j), \quad (19)$$

where  $R$  represents the set of ranks,  $N_R$  is its size,  $j \in R$  corresponds to individual ranks,  $s_j$  is the predicted score (interpreted as the importance score  $\alpha_j$  in our work),  $g(s_j) = 2^{s_j} - 1$  is the gain function that emphasizes the relevance of high-importance

Ablations	MIntRec2.0						IEMOCAP-DA					
	ACC (↑)	F1 (↑)	P (↑)	R (↑)	WF1 (↑)	WP (↑)	ACC (↑)	F1 (↑)	P (↑)	R (↑)	WF1 (↑)	WP (↑)
w / o Relative Importance	58.76	52.99	54.11	52.84	58.21	58.31	74.65	<u>72.29</u>	<u>73.41</u>	72.02	74.56	74.91
w / o Complementarity	58.80	<u>53.93</u>	55.90	53.75	58.28	58.96	73.67	70.85	72.40	70.43	73.49	73.86
w / o Inconsistency	<u>59.53</u>	53.88	<u>56.39</u>	<u>54.01</u>	<u>58.72</u>	<u>59.30</u>	<u>74.81</u>	71.78	72.26	<u>72.31</u>	<u>74.65</u>	<u>75.03</u>
Full	<b>60.46</b>	<b>55.35</b>	<b>59.33</b>	<b>55.09</b>	<b>59.72</b>	<b>60.85</b>	<b>74.95</b>	<b>72.99</b>	<b>74.27</b>	<b>72.74</b>	<b>74.88</b>	<b>75.47</b>

Table 8: Ablation studies for the reasoning relations on the MIntRec2.0 and IEMOCAP-DA datasets, with each configuration presenting results for the exclusion of relative importance, complementarity, or inconsistency. Bold text denotes the best performance, while underlined text indicates the second-best.


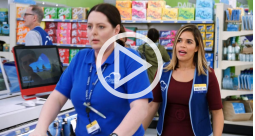



Sample	Description	Rank
 <p>Look, i'm sorry, i did not know that she was gonna say those things.</p>	<p><b>Speakers' Actions:</b> The two women are sitting at the table, with one of them talking to the other. The woman who is <b>speaking is gesturing with her hands as she talks</b>.</p> <p><b>Facial Expressions:</b> The woman who is speaking seems to be <b>angry and frustrated</b>, as indicated by her facial expressions and tone of voice. The other woman appears to be listening attentively.</p> <p><b>Interaction with Others:</b> The two women seem to be engaged in a <b>serious conversation</b>, but there are no visible signs of tension or hostility between them.</p>	<p>1<sup>st</sup>: Speakers' Actions</p> <p>2<sup>nd</sup>: Facial Expressions</p> <p>3<sup>rd</sup>: Interaction with Others</p> <p>Label: Apologise</p>
 <p>As a friend, you wanna be a friend?</p>	<p><b>Speakers' Actions:</b> The two women are engaged in a conversation, with one woman looking at the other while speaking. The other woman is <b>holding a shopping cart</b> and appears to be listening attentively.</p> <p><b>Facial Expressions:</b> Both women have a <b>calm and neutral expression</b> on their faces, suggesting a friendly and <b>non-confrontational</b> conversation.</p> <p><b>Interaction with Others:</b> The two women seem to be the only people in the scene, and there is <b>no interaction with others</b> visible in the video.</p>	<p>1<sup>st</sup>: Facial Expressions</p> <p>2<sup>nd</sup>: Speakers' Actions</p> <p>3<sup>rd</sup>: Interaction with Others</p> <p>Label: Confirm</p>
 <p>Uh, he's weird, and when you mentioned that he worked in insurance, you could've mentioned it was horse insurance.</p>	<p><b>Speakers' Actions:</b> The woman is standing and <b>gesturing with her hands</b> while talking to the man. The man is sitting at the table and listening to her.</p> <p><b>Facial Expressions:</b> The woman appears to be <b>angry and frustrated</b>, as indicated by her facial expressions and gestures. The man seems to be calmly listening to her. The woman and the man are the only ones interacting in the scene.</p> <p><b>Interaction with Others:</b> They seem to be engaged in a <b>serious conversation</b>, but there is <b>no visible tension or hostility</b> between them.</p>	<p>1<sup>st</sup>: Interaction with Others</p> <p>2<sup>nd</sup>: Speakers' Actions</p> <p>3<sup>rd</sup>: Facial Expressions</p> <p>Label: Complain</p>
 <p>It'll be fine, you guys.</p>	<p><b>Speakers' Actions:</b> The man in the blue shirt is seen sitting at the table with the two women. He is engaged in conversation with them, occasionally <b>gesturing with his hands to emphasize his points</b>. The women are attentive and responsive to his comments.</p> <p><b>Facial Expressions:</b> The man appears to be <b>relaxed and friendly</b>, with a <b>smile on his face</b> and a <b>calm demeanor</b>. The women seem to be enjoying the conversation and are engaged in the discussion.</p> <p><b>Interaction with Others:</b> The man, woman in red, and woman in blue <b>are all familiar with each other</b>, as evidenced by their <b>relaxed posture and comfortable interaction</b>. The conversation is friendly and informal, with occasional moments of laughter and camaraderie.</p>	<p>1<sup>st</sup>: Facial Expressions</p> <p>2<sup>nd</sup>: Interaction with Others</p> <p>3<sup>rd</sup>: Speakers' Actions</p> <p>Label: Comfort</p>
 <p>Sorry, everybody. Looks like we're gonna starve down here because any thought we'd only be trapped for 15 minutes.</p>	<p><b>Speakers' Actions:</b> The woman is standing in front of the cookie shelf, <b>gesturing with her hands</b> as she talks to the other woman. The other woman is listening and occasionally <b>nodding</b> her head.</p> <p><b>Facial Expressions:</b> The woman talking seems to be expressing <b>frustration or annoyance</b>, as indicated by her gestures and facial expressions. The other woman appears to be calmly listening to her.</p> <p><b>Interaction with Others:</b> The two women appear to be <b>acquaintances or friends</b>, as they are engaged in a <b>casual conversation</b>. There is no visible tension or hostility between them.</p>	<p>1<sup>st</sup>: Interaction with Others</p> <p>2<sup>nd</sup>: Facial Expressions</p> <p>3<sup>rd</sup>: Speakers' Actions</p> <p>Label: Inform</p>

Figure 4: Samples from the MIntRec2.0, showcasing descriptions and ranking results of fine-grained semantics.

items,  $d(j) = \frac{1}{\log_2(j+1)}$  is the discount function that reduces the weight of lower-ranked elements, and  $\text{scale}(\hat{P})_j$  represents the row-stochastic approximation of the sorting operator. The matrix  $\hat{P}$  is derived by approximating the hard permutation matrix  $P_{\text{sort}(s)}$ , induced by sorting the predicted scores  $s = f(x)$ , with the formula (Grover et al., 2019) as follows:

$$P = \frac{(n+1-2u)s - A_s \mathbf{1}}{\tau}, \quad (20)$$

$$\hat{P}_{\text{sort}(s)}[u, :](\tau) = \text{softmax}(P), \quad (21)$$

where  $A_s[u, v] = |s_u - s_v|$  represents pairwise differences between scores,  $\mathbf{1}$  is a column vector of ones, and  $\tau > 0$  is the temperature parameter that controls the trade-off between the approximation accuracy and gradient stability. As  $\tau \rightarrow 0$ ,  $\hat{P}$  converges to the true permutation matrix  $P_{\text{sort}(s)}$ , closely approximating the hard sorting process. To ensure both row- and column-stochasticity, Sinkhorn normalization is applied to  $\hat{P}$ , further sta-


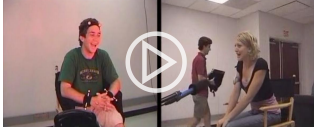
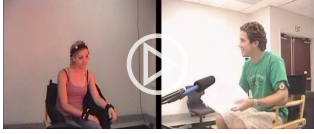


Sample	Description	Rank
 <p>I got this idea watching them go down. everything was being destroyed, see, but it seemed that there was one new thing being made. a sort of responsibility. man for man.</p>	<p><b>Speakers' Actions:</b> The woman is sitting in one of the chairs, <b>leaning forward with her hands resting on her lap</b>. The man is sitting in the other chair, holding a piece of paper in his hand. Both speakers appear to be engaged in a conversation, occasionally <b>gesturing with their hands</b>.</p> <p><b>Facial Expressions:</b> The woman's facial expression is <b>neutral</b>, while the man seems to be smiling at certain points in the conversation. Their tone of voice is <b>friendly</b> and conversational.</p> <p><b>Interaction with Others:</b> The speakers appear to be having a <b>one-on-one conversation</b>, with no other individuals present in the scene. Their body language suggests a <b>comfortable and familiar relationship</b>.</p>	<p>1<sup>st</sup>: Interaction with Others</p> <p>2<sup>nd</sup>: Speakers' Actions</p> <p>3<sup>rd</sup>: Facial Expressions</p> <p>Label: Statement opinion</p>
 <p>What? What? You look really happy what's going on?</p>	<p><b>Speakers' Actions:</b> he woman is sitting on a chair with her hands resting on her lap. She occasionally <b>leans forward and gestures with her hands while speaking</b>. The man is sitting on the other chair, wearing a green shirt and a headband. He listens attentively and occasionally <b>nods his head</b>.</p> <p><b>Facial Expressions:</b> Both the woman and the man appear to be in a <b>good mood, smiling and laughing</b> at certain points in the conversation. The woman's facial expressions are more <b>animated</b>, while the man maintains a more composed demeanor.</p> <p><b>Interaction with Others:</b> The woman and the man seem to be engaged in a <b>friendly conversation</b>, exchanging ideas and opinions. There is a sense of <b>mutual respect and understanding between them</b>.</p>	<p>1<sup>st</sup>: Facial Expressions</p> <p>2<sup>nd</sup>: Interaction with Others</p> <p>3<sup>rd</sup>: Speakers' Actions</p> <p>Label: Question</p>
 <p>Honey, this is a natural phenomenon. i mean this only happens once a year, you know.</p>	<p><b>Speakers' Actions:</b> The woman is sitting in one of the chairs, wearing a pink tank top and a black leather wristband. She is looking directly at the camera and occasionally <b>gestures with her hands</b>. The man is sitting in the other chair, wearing a green t-shirt and <b>holding a piece of paper</b>. He is <b>looking down at the paper</b> and occasionally glances up at the camera.</p> <p><b>Facial Expressions:</b> Both speakers appear to be <b>calm and composed</b> throughout the video. Their facial expressions are neutral, and their tone of voice is measured and professional.</p> <p><b>Interaction with Others:</b> They appear to <b>be focused solely on the conversation</b> with the camera.</p>	<p>1<sup>st</sup>: Speakers' Actions</p> <p>2<sup>nd</sup>: Facial Expressions</p> <p>3<sup>rd</sup>: Interaction with Others</p> <p>Label: Statement non opinion</p>
 <p>Yeah. I'm so sorry.</p>	<p><b>Speakers' Actions:</b> The speakers are engaged in a conversation, with one person speaking and the other listening attentively. The speaker uses hand gestures to emphasize their points, while the listener maintains a <b>composed posture</b>. Both speakers appear to be <b>calm</b> and composed, with no visible signs of strong emotions.</p> <p><b>Facial Expressions:</b> Their facial expressions are <b>neutral</b>, and their tone of voice is measured.</p> <p><b>Interaction with Others:</b> The speakers are the only individuals in the scene, and their interaction is <b>focused and one-on-one</b>. There is no visible tension or familiarity between them.</p>	<p>1<sup>st</sup>: Facial Expressions</p> <p>2<sup>nd</sup>: Speakers' Actions</p> <p>3<sup>rd</sup>: Interaction with Others</p> <p>Label: Apology</p>
 <p>Calm yourself.</p>	<p><b>Speakers' Actions:</b> The woman is seen sitting in one of the chairs, leaning forward with her hands clasped. She occasionally <b>nods her head and gestures with her hands</b>. The man is seated in the other chair, holding a microphone and speaking directly to the camera. He occasionally <b>looks towards the woman and gestures with his hands</b>.</p> <p><b>Facial Expressions:</b> Both the woman and the man appear to be engaged in a <b>serious conversation</b>. They maintain a neutral facial expression throughout the video, with occasional slight smiles.</p> <p><b>Interaction with Others:</b> The woman and the man seem to be having a <b>one-on-one conversation</b>, with no other individuals present in the scene. They appear to be <b>actively listening and responding to each other</b>.</p>	<p>1<sup>st</sup>: Interaction with Others</p> <p>2<sup>nd</sup>: Speakers' Actions</p> <p>3<sup>rd</sup>: Facial Expressions</p> <p>Label: Command</p>

Figure 5: Samples from the IEMOCAP-DA, showcasing descriptions and ranking results of fine-grained semantics.

bilizing its use in optimization by resolving inconsistencies in quasi-sorted outputs. NeuralNDCG integrates these components to enable gradient-based learning directly aligned with ranking performance, providing a powerful mechanism for optimizing ranking tasks in diverse applications.

## F Training Cost

The additional computational cost primarily arises from the LLM-Guided Semantic Extraction process, which requires approximately two hours per dataset under the experimental conditions outlined. This duration is comparable to the training time of the model itself, which also takes around two hours. Despite this added computational expense, the trade-off is justified by the substantial benefits it brings in enhancing semantic understanding. Moreover, the training cost introduced by this approach remains considerably lower than that of fine-tuning, which typically demands 8–10 hours of computation. Thus, the proposed method achieves a more efficient balance between computational efficiency

and semantic extraction performance.

## G Ablations for Semantic Relations

To further validate the effectiveness of the three proposed reasoning relations, we conduct ablation studies on relative importance, complementarity, and inconsistency across the MIntRec2.0 (Zhang et al., 2024a) and IEMOCAP-DA (Saha et al., 2020) datasets, with results summarized in Table 8. For relative importance and complementarity, we exclude their respective weight generation and weighting processes, while for inconsistency, the entire penalty feature is removed.

From the experimental results, the absence of relative importance leads to a notable drop in performance across all metrics on both datasets, highlighting its role as a cornerstone of our reasoning framework. On the MIntRec2.0 dataset, performance decreases range from 1.51% to 5.22%, which are significantly higher than the declines observed on IEMOCAP-DA, ranging from 0.30% to 0.86%. This emphasizes the critical importance



of differentiating the relative significance of fine-grained semantics in understanding complex intents. Similarly, the removal of complementarity leads to metric reductions exceeding 1% across the board, illustrating LGSRR’s success in capturing the inherent synergy between semantic elements. The absence of inconsistency results in the most significant declines in F1 and P scores, with reductions of 1.47% and 2.94% on MIntRec2.0 and 1.21% and 2.01% on IEMOCAP-DA. This decline is especially pronounced in complex intent categories such as *Joke* and *Flaunt*, which depend heavily on identifying contradictory semantic cues and are underrepresented in the dataset. These findings underscore LGSRR’s robustness in addressing nuanced inconsistencies, even in categories with limited data. These ablation studies confirm the essential role of the proposed reasoning relations and their integration, forming the basis of LGSRR’s capability to handle diverse intent semantics with precision and adaptability.

ability of LGSRR but also underscore its strength in navigating the intricacies of multimodal reasoning across datasets with distinct characteristics.

## H Additional Case Studies

Figure 4 and Figure 5 present a diverse set of samples from the MIntRec2.0 and IEMOCAP-DA datasets, offering a detailed glimpse into the fine-grained semantic descriptions and importance rankings produced by LGSRR. The selected cases cover a range of intent labels and scenarios, each presenting semantic challenges and opportunities. On MIntRec2.0, LGSRR excels at identifying critical cues, such as “gesturing with hands” or “expressing frustration,” which are particularly relevant in contexts like *Complain* and *Criticize*. These results demonstrate the model’s ability to capture subtle yet impactful cues that are essential for intent understanding. In contrast, IEMOCAP-DA poses a greater challenge, featuring two-person dialogue scenes from varied perspectives that demand more nuanced reasoning. Fine-grained semantics in this dataset involve conversational dynamics rather than explicit physical actions, making it difficult to disentangle key semantic elements. For example, interactions are frequently characterized by subtle cues such as mutual respect, attentiveness, or slight gestures, which require precise modeling to capture effectively. Despite these complexities, LGSRR achieves impressive performance, consistently identifying the most relevant interactions and prioritizing critical semantic details. These case studies not only highlight the versatility and adapt-