Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

多模态意图识别

硕士研究生 杨桢弘

2026年07月12日

问题回溯

- 总结反思
  - 评价指标描述不清晰
  - 算法原理讲解不深入
- 相关内容
  - 2025.11.30 王旭《缓解多模态大语言模型的幻觉问题》
  - 2025.12.28 杨桢弘《缺失模态的情绪变化识别》

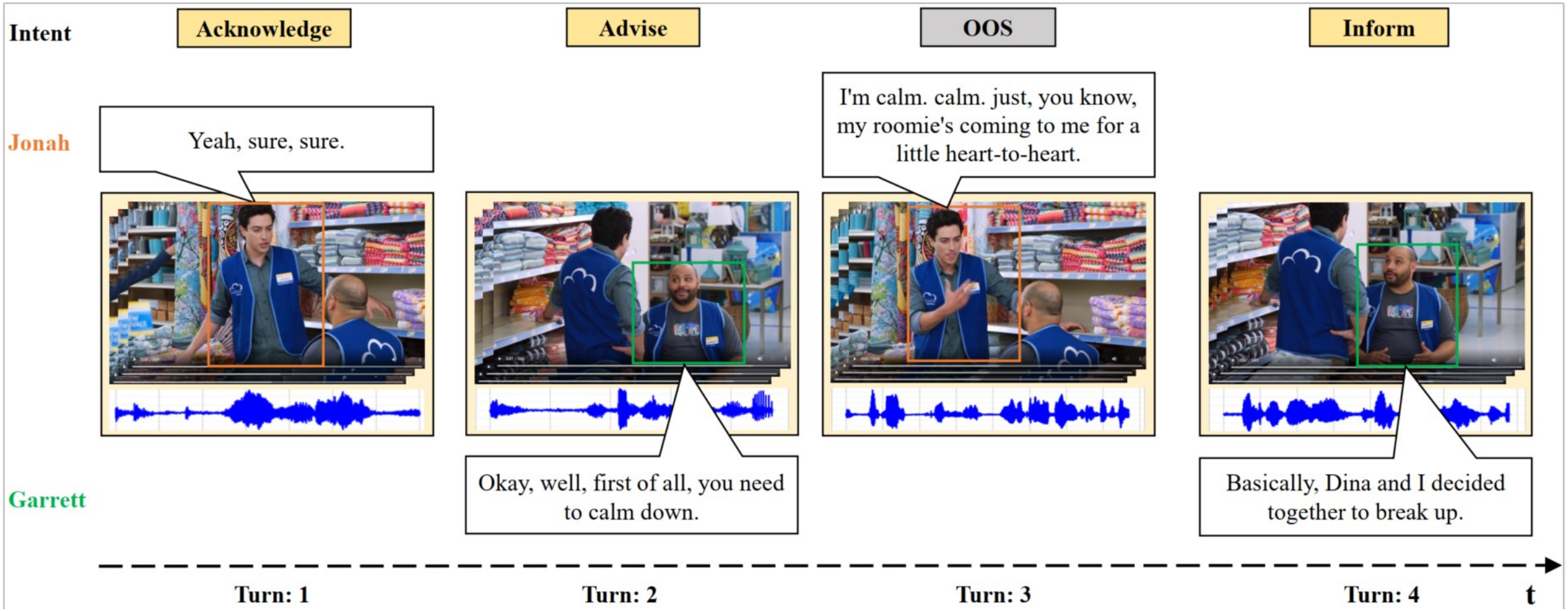
内容提要

- 预期收获
- 内容引入
- 内涵解析与研究目标
- 研究背景与意义
- 研究历史与现状
- 知识基础
- 算法原理
  - LGSRR
  - HIER
- 特点总结与未来展望
- 参考文献

预期收获

- 预期收获
  - 掌握多模态意图识别的研究现状与基本概念
  - 理解多模态意图识别的基本方法及其原理
  - 了解多模态意图识别未来发展方向

内容引入


Intent
Acknowledge
Advise
OOS
Inform
Jonah
Yeah, sure, sure.
I'm calm. calm. just, you know, my roomie's coming to me for a little heart-to-heart.
Garrett
Okay, well, first of all, you need to calm down.
Basically, Dina and I decided together to break up.
Turn: 1
Turn: 2
Turn: 3
Turn: 4
t

意图识别场景

内涵解析与研究目标

- 内涵解析
  - 意图识别
    - 根据用户的语言、行为或交互信息，判断其背后的目标或交际目的
    - 请求、感谢、建议、抱怨等
  - 多模态意图识别
    - 语音+文本+视频→意图类别
    - 综合语言内容、面部表情、动作姿态、人物互动、语音语调等信息，推断用户真实意图

- 研究目标
  - 结合多模态大模型、语义推理等技术
  - 在模态信息不完整、不对齐或质量受损的现实条件下，构建能够从文本、视觉等多模态信息中识别用户真实意图的模型

研究背景与意义

- 研究背景
  - 在实际交流中，同一句话可能因为表情、动作、语气和上下文不同而表达不同意图，因此单一文本信息往往难以准确反映用户真实需求
  - 不同模态之间存在异质性、噪声干扰和语义冲突

- 研究意义
  - 多模态意图识别是人机交互的重要基础技术，可应用于智能客服、虚拟助手和对话系统等场景
  - 能够利用更丰富的交互线索，提升模型对复杂语义、隐含意图和细粒度类别的识别能力

研究历史与现状

Rahman等人提出MAG-BERT，通过多模态适配门控机制将视觉和音频信息注入BERT的文本表示中，将预训练语言模型与非语言模态融合，使文本主干模型能够吸收视觉和声学信息

Zhang等人提出MIntOOD，通过加权特征融合和多粒度表示学习同时处理已知意图分类与未知意图检测，将多模态意图识别从封闭类别分类扩展到开放场景鲁棒识别

Zhou等人提出HIER，将多模态语义组织为输入token语义概念和概念关系三个层次，并通过MLLM反馈进行自进化推理

2019 2020 2024 2026

2019

Tsai等人提出MulT，通过方向性成对跨模态注意力建模不同模态序列之间的交互关系，在不显式对齐数据的情况下处理多模态序列的异步问题

2024

Zhou等人提出TCL-MAP，通过token级对比学习和模态感知提示增强文本、视觉、音频之间的语义融合

2025

Zhou等人提出LGSRR，通过LLM自动发现动作、表情、交互等细粒度语义，并建模语义之间的重要性、互补性和不一致性，把多模态意图识别从“模态级融合”推进到“细粒度语义关系推理”

跨模态融合阶段

细粒度对齐与鲁棒融合阶段

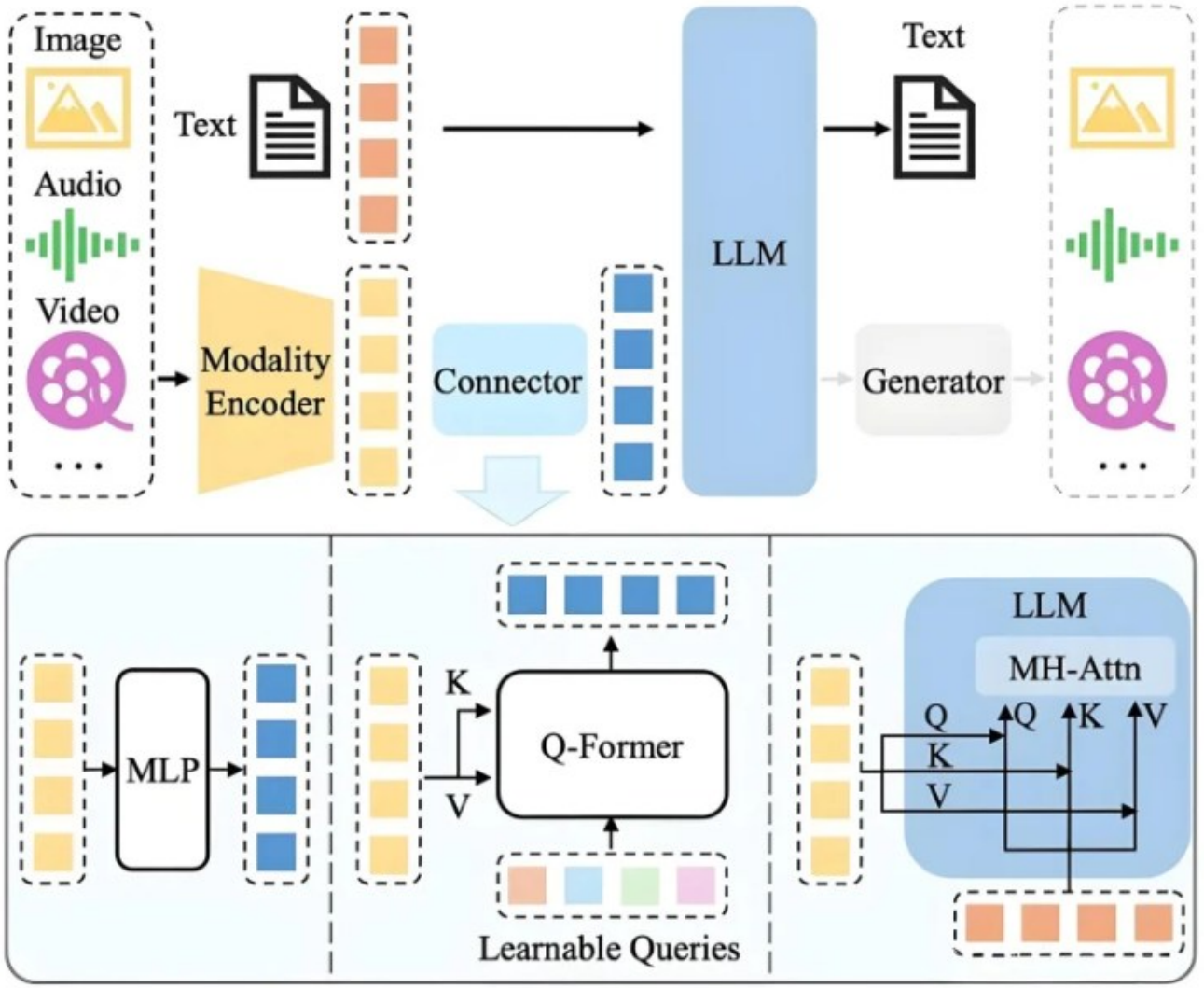
大模型辅助语义推理阶段

研究历史与现状

- 现有存在问题
  - 依赖模态级粗粒度融合，难以提取细粒度意图线索
  - 语义关系建模不足，复杂意图推理能力有限
  - 大模型直接推理效果不稳定，层次化推理能力不足
- 现有的常见方法
  - 基于多模态融合与对齐的方法
    - 利用跨模态注意力、对比学习等方法，对齐异质模态语义
    - 通过特征级融合、决策级融合或混合融合，整合不同模态互补信息
    - 能够缓解单模态语义歧义，但多数仍停留在模态级表示融合
  - 基于大模型与语义推理的方法
    - 利用LLM/MLLM的语义理解能力，提取意图相关知识、描述或推理线索
    - 通过CoT将意图判断拆分为场景理解、语义概念分析和关系推理
    - 进一步建模细粒度语义关系和层次化表示，提升复杂意图理解能力

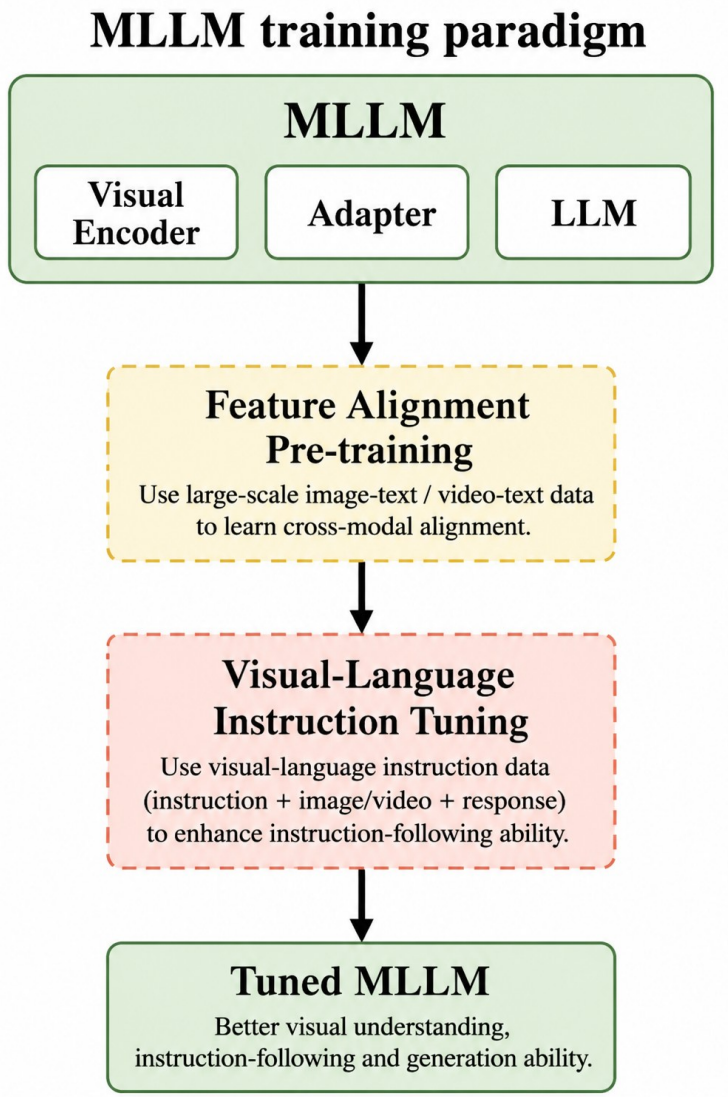
知识基础

- 多模态大语言模型MLLM
  - 输入、模态编码器、模态接口、大语言模型、输出
- 重要组成部分和特点：
  - 特征投影：视觉特征直接映射到语言模型的嵌入空间
  - 查询式融合：引入查询向量，通过交叉注意力机制，从视觉特征中提取语义相关的关键信息，生成固定数量的视觉上下文token
  - 感知重采样注入：利用感知重采样模块，将高维视觉token压缩为少量语义密集的latent token



知识基础

- 预训练
  - 学习不同模态之间的基础语义对应关系
  - 使模型具备基本的图像理解、视频理解和文本生成能力
- 指令微调
  - 使用“多模态输入 + 任务指令 + 标准回答”进行训练
  - 让模型能够按照自然语言指令完成具体任务
- 对齐微调
  - 根据人类反馈或偏好数据进一步优化模型输出
  - 减少多模态幻觉，提高回答准确性和可靠性
  - 使模型输出更符合人类需求和真实视觉内容



- 思维链
  - 基本思想
    - 将复杂任务拆分为多个简单步骤
    - 让模型按照“问题理解 → 线索分析 → 逻辑推理 → 最终判断”的路径进行推断
    - 提升模型处理复杂语义、因果关系和多步推理任务的能力
  - 主要特点
    - 显式中间过程：模型生成可读的推理步骤
    - 分阶段分析：将复杂问题转化为多个子问题
    - 可解释性更强：便于观察模型依据哪些线索做出判断
    - 适合复杂任务：常用于问答、数学推理、常识推理和多模态理解

思维链技术	介绍
CoT	引入解释性的中间步骤，在模型内部完成推理计算
ToT	引进额外结构，构建树状思维过程，允许回溯
SoT	首先生成思维“骨架”，在此基础上填充
GoT	使用图形结构，顶点和边表示信息单元和依赖关系
PoT	产生程序语句，使用程序解释器得到推理结果

- 多模态意图识别的评价指标
  - 精确率（P）：

$$P_i = \frac{TP_i}{TP_i + FP_i}$$

  - 召回率（R）：

$$R_i = \frac{TP_i}{TP_i + FN_i}$$

  - F1值：

$$F1_i = \frac{2 \cdot P_i \cdot R_i}{P_i + R_i}$$

  - 加权F1值：

$$WF1 = \sum_{i=1}^{C} w_i F1_i$$

  - 加权精确率：

$$WP = \sum_{i=1}^{C} w_i P_i$$

LLM-Guided Semantic Relational Reasoning for Multimodal Intent Recognition

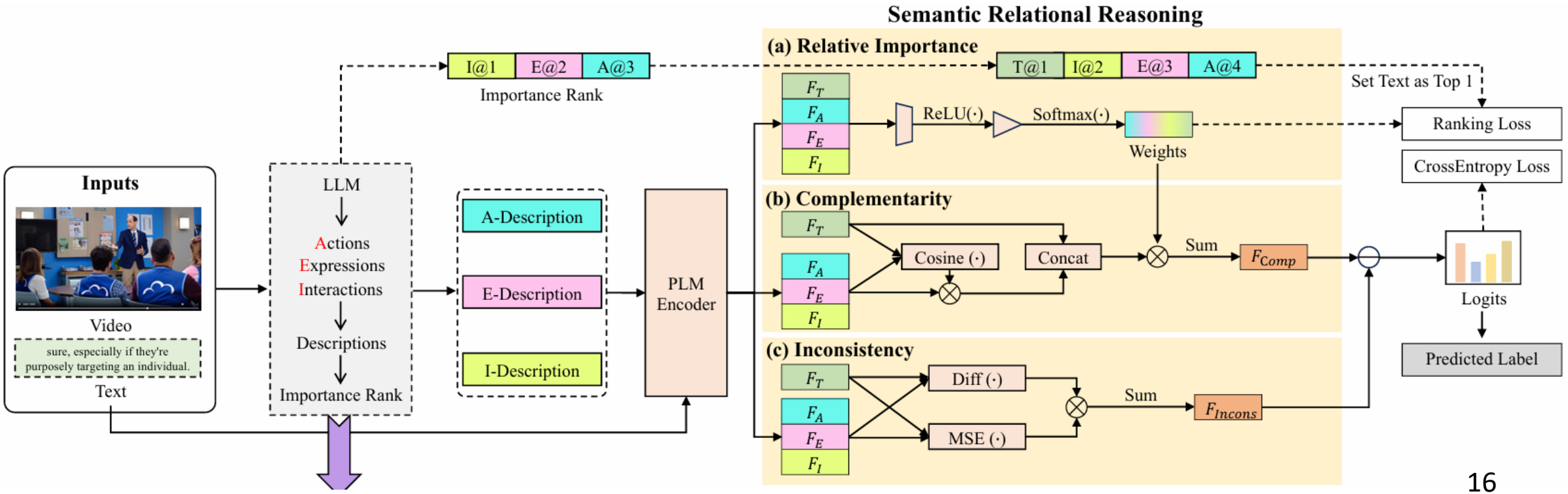
LGSRR TIPO

T	目标	提升多模态意图识别中复杂意图和细粒度意图的识别能力，避免模型只依赖粗粒度模态融合
I	输入	2个数据集（IEMOCAP-DA、MIntRec2.0数据集）
P	处理	1.大模型引导细粒度语义提取 2.语义关系推理 3.意图分类
O	输出	意图类别（12/30类）

P	问题	1.现有方法过度依赖模态级粗粒度语义 2.现有方法缺少细粒度语义关系推理能力
C	条件	需要预训练LLM/MLLM支持
D	难点	1.如何自动提取与意图相关的细粒度语义 2.如何建模语义线索之间的复杂关系
L	水平	2025 CCF B类

LGSRR 算法原理

- 算法原理图
  - 大模型引导细粒度语义提取
  - 语义关系推理
  - 意图分类



LGSRR 创新点分析

- 现有方法存在问题
  - 现有方法大多在文本、视频、音频的整体特征上进行融合，容易停留在模态级粗粒度表示，对于讽刺、批评、玩笑等细粒度意图，仅依靠整体模态特征容易忽略关键语义线索

- LGSRR的解决方法
  - 利用大语言模型自动提取与意图相关的细粒度语义，不直接融合原始模态特征，而是先将视频和文本转换为更清晰的语义描述，后续语义关系推理提供更明确的语义基础

LGSRR 创新点分析

- 解决方法
  - 大模型引导细粒度语义提取
    - 细粒度语义发现：$S_{init} = GPT\_3.5(B; Template_1)$
    - 语义描述生成：$D = VideoLLaMA2(V, T; S; Template_2)$
    - 语义重要性排序：$S_{rank} = GPT\_3.5(D; y; Template_3)$

LLM-Guided Semantic Extraction

CoT-Step 1: Fine-grained Semantics Discovery

$\{(\langle Video\rangle, \langle Text\rangle_i)\}_{i=1}^{n}$: 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 item by item.

Response: Speakers' Actions, Facial Expressions, Interaction with Others, Identity and Status, Tone and Manner of Speaking, Scene Description…

CoT-Step 2: Semantic Description

<Video> + <Text>: Generate detailed descriptions that help identify the speaker's intent. Please combine video and text to describe from the given perspectives.

Response: (1) **Speakers' Actions**. The man in the suit and tie is standing and gesturing with his hands as he speaks. The others are sitting and listening attentively;
(2) **Facial Expressions**. The man in the suit and tie appears confident and composed, while the others seem engaged and interested in what he is saying;
(3) **Interaction with Others**. The man in the suit and tie is addressing the group as a whole, rather than targeting an individual. The others are listening and occasionally nodding or asking questions.

CoT-Step 3: Importance Rank (For Training)

<A-Description> + <E-Description> + <I-Description> + <Intent>: I will provide you with the intent of the video and the descriptions from three different aspects. You need to rank these aspects according to their contribution to the intent and the content of the descriptions.

Response: The first is Interaction with Others. The second is Facial Expressions. The third is Speakers' Actions.

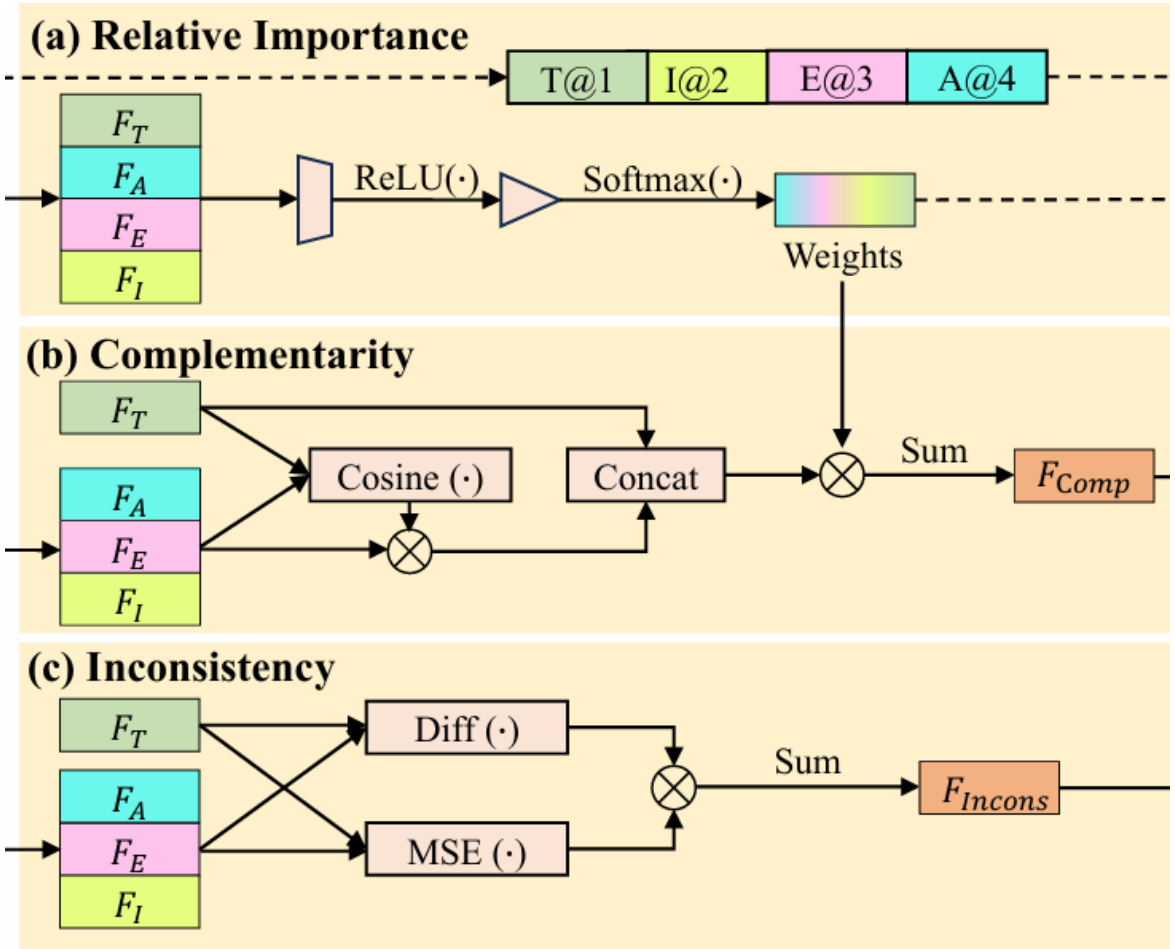
LGSRR 创新点分析

- 现有方法存在问题
  - 现有方法通常使用拼接、加权或注意力等方式进行融合，但这种方式主要关注“哪些模态重要”，没有进一步分析不同语义线索之间的关系，不同语义线索之间可能相互支持，也可能相互冲突，简单融合难以判断这些线索到底是补充信息还是干扰信息

- LGSRR的解决方法
  - 提出语义关系推理，从逻辑关系角度建模三类语义关系：相对重要性、互补性、不一致性，通过关系推理，进一步判断不同语义线索如何共同影响最终意图识别

LGSRR 创新点分析

- 解决方法
  - 语义关系推理
    - 语义特征编码：使用BERT对原始文本和语义描述进行编码
    - 相对重要性建模：利用GPT-3.5生成的语义重要性排序作为监督信号，引导模型学习重要性
    - 互补性建模：比较文本特征和细粒度语义特征之间的相似程度，判断它们是否具有互补关系
    - 不一致性建模：比较文本特征和细粒度语义特征之间的差异，判断它们是否存在冲突



实验设计　数据资源

- 数据集
  - IEMOCAP-DA（12类对话行为标签；训练集：6590；验证集：942；测试集：1884；2020年改自IEMOCAP）
  - MIntRec2.0（30类细粒度意图；训练集：6165；验证集：1106；测试集：2033；2024年发布，多轮多方对话）
- 对比方法

  MulT(2019)、MISA(2020)、MAG-BERT(2020)、TCL-MAP(2024)、SDIF-DA(2024)、MIntOOD(2024)
- 评价指标
  - ACC:准确率　　F1：F1值
  - P：精确率　　R：召回率
  - WF1：加权F1　　WP：加权精确率

实验设计　对比实验

- 评估LGSRR在数据集上的表现
  - LGSRR 在 MIntRec2.0 上虽然 ACC 与 MulT 接近，但其余指标均更优；在 IEMOCAP-DA 上则超过所有对比方法
  - 细粒度语义提取和语义关系推理能够提升复杂意图理解能力

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
MulT	**60.66**	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	60.46	**55.35**	**59.33**	**55.09**	**59.72**	**60.85**	**74.95**	**72.99**	**74.27**	**72.74**	**74.88**	**75.47**

实验设计　消融实验

- 评估LGSRR的不同模块在数据集上的表现
  - w/o LGSE：去掉 LLM 引导细粒度语义提取，改用粗粒度模态级特征
  - w/o Lrank：去掉语义重要性排序损失
  - w/o SRR：去掉语义关系推理模块，改用简单融合
  - LGSE(VideoLLaMA)、LGSE(Qwen2-VL)：替换语义描述生成模型

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)	60.27	54.91	57.07	54.78	59.56	60.00	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	72.19	73.05	72.43	74.42	75.05
w / o $\mathcal{L}_{rank}$	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	74.57	71.80	71.82	72.43	74.44	74.63
Full	**60.46**	**55.35**	**59.33**	**55.09**	**59.72**	**60.85**	**74.95**	**72.99**	**74.27**	**72.74**	**74.88**	**75.47**

算法总结 LGSRR

- 算法贡献
  - 大模型引导细粒度语义提取：自动发现动作、表情、互动等意图相关语义，缓解粗粒度模态融合中语义不明确、噪声冗余较多的问题
  - 语义关系推理：从重要性、互补性和不一致性三个角度建模复杂语义关系，提升了模型对相似意图和隐含意图的区分能力
- 算法不足
  - 依赖GPT-3.5和VideoLLaMA2生成语义描述，预处理成本较高
  - 语义维度主要固定为动作、表情、互动，无法覆盖所有复杂场景
  - 关系建模仍基于三类基本关系，对更复杂的上下文语义关系刻画有限

Evolutionary Multimodal Reasoning via Hierarchical Semantic Representation for Intent Recognition

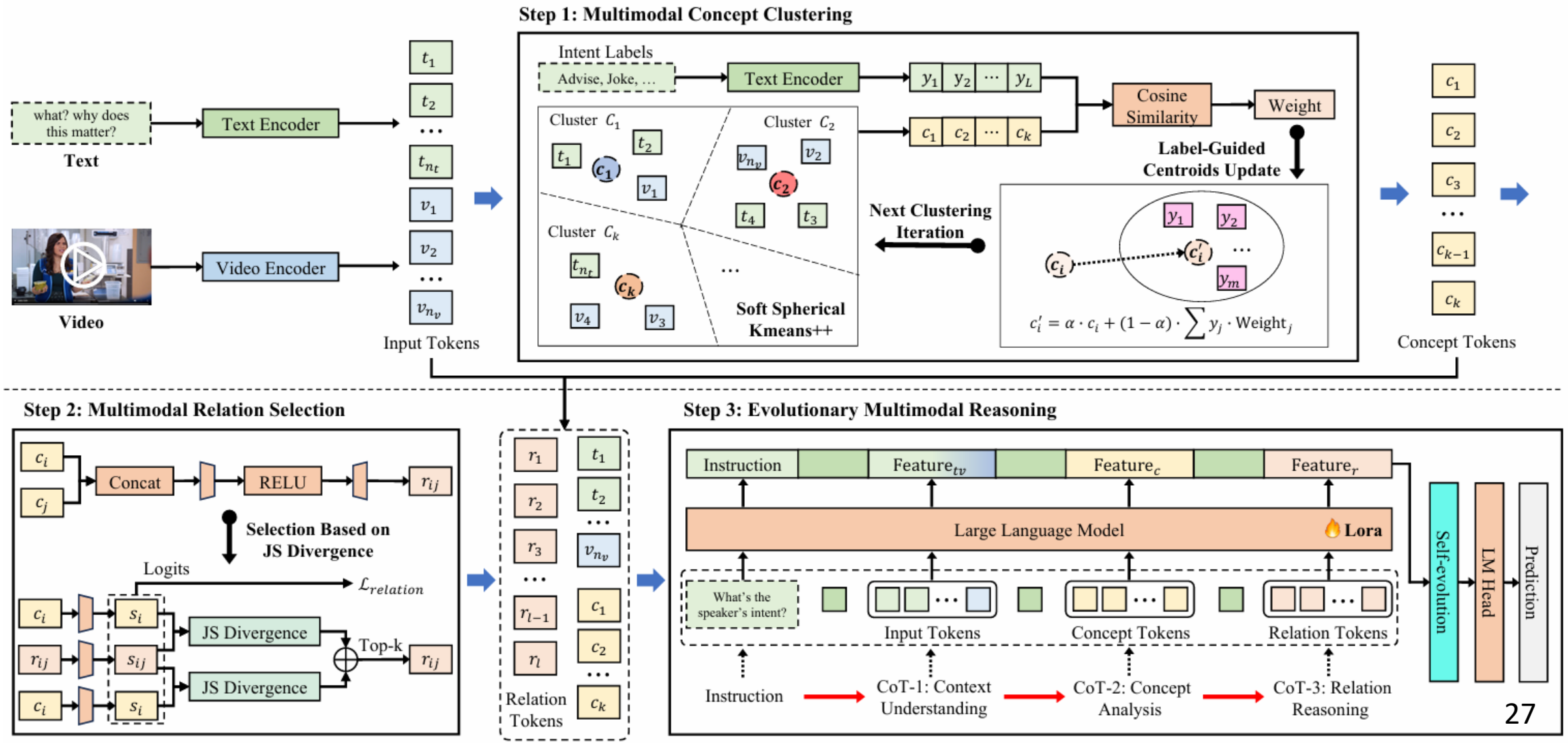
HIER TIPO

T	目标	提升复杂多模态意图识别中的层次化语义理解能力和推理鲁棒性
I	输入	3个数据集（MIntRec、MIntRec2.0、MELD-DA数据集）
P	处理	1.多模态概念聚类 2.多模态关系选择 3.自进化多模态推理 4.预测意图
O	输出	意图类别（20/30/12类）

P	问题	1.现有方法缺少对多模态语义层次结构的建模 2.现有大模型推理过程较静态，难以根据当前样本动态调整语义表示
C	条件	需要多模态大模型作为基础模型
D	难点	1.如何抽象出意图相关概念并筛选真正有用的概念关系 2.如何让模型在推理过程中自我评估和修正语义表示
L	水平	2026 CCF A类

HIER 算法原理

- 算法原理图



HIER 创新点分析 多模态概念聚类

- 现有方法存在问题
  - 现有方法大多细粒度线索提取层面，没有建模语义层次结构，容易出现推理不连贯、语义理解不稳定的问题
- 解决方法
  - 输入：多模态大模型提取文本token和视频token
  - 利用多模态概念聚类构建中层语义概念
    - Token 表示：保留文本和视频中的局部语义线索
    - Concept 聚类：将语义相近的token聚合为概念表示
    - Label-guided：引入意图标签语义，引导概念更贴近分类目标
  - 得到更稳定、更具有意图判别性的概念表示

低层token零散 → 中层概念聚合；标签语义引导 → 概念更贴近意图判断

- 现有方法存在问题
  - 单个语义概念难以表达复杂意图中的高阶语义依赖，如果保留所有概念组合，会引入大量无关关系和噪声
- 解决方法
  - 输入：上一阶段得到的多模态概念表示
  - 构建概念之间的候选关系，并筛选有价值的关系
    - Pairwise combination：概念两两组合，形成候选概念关系
    - Relation representation：利用信息瓶颈网络得到紧凑关系表示
    - JS divergence：衡量概念关系是否带来新的判别信息
    - Top-k selection：保留高价值关系，过滤冗余和无关关系
  - 得到能够辅助复杂意图判断的高阶关系表示

单个概念有限 → 概念关系补充；全部关系有噪声 → 筛选高价值关系

- 现有方法存在问题
  - 多模态大模型容易直接根据输入生成答案，推理过程不够明确，传统推理过程较静态，难以判断哪些概念和关系真正有用
- 解决方法
  - 输入：原始上下文、概念表示和关系表示
  - 构建结构化思维链，并利用模型反馈动态修正语义表示
    - Context Understanding：理解视频和文本的整体语境
    - Concept Analysis：分析哪些概念与当前意图相关
    - Relation Reasoning：推理概念关系对意图判断的作用
    - Self-evolution：根据模型反馈增强有用语义，削弱无关语义
  - 使模型按照“语境 → 概念 → 关系”的路径逐步完成意图判断

直接分类 → 语境、概念、关系分步推理

实验设计　数据资源

- 数据集
  - MELD-DA（12类对话行为标签；训练集：6991；验证集：999；测试集：1998；2020年改自MELD）
  - MIntRec（20类细粒度意图；训练集：1334；验证集：445；测试集：445；2022年发布，来源于情景喜剧 Superstore）
  - MIntRec2.0（30类细粒度意图；训练集：6165；验证集：1106；测试集：2033；2024年发布，支持 OOS 范围外意图检测）
- 对比方法

  MISA(2020)、MAG-BERT(2020)、MulT(2019)、TCL-MAP(2024)、SDIF-DA(2024)、MIntOOD(2024)、MVCL-DAF(2025)
- 评价指标
  - ACC、F1、P、R、WF1、WP

实验设计　对比实验

- 评估HIER在数据集上的表现
  - HIER在三个数据集上整体取得最优或接近最优表现

Datasets	MIntRec						MIntRec2.0						MELD-DA					
Methods	ACC	F1	P	R	WF1	WP	ACC	F1	P	R	WF1	WP	ACC	F1	P	R	WF1	WP
MISA	72.29	69.24	72.38	73.48	69.32	70.85	55.16	49.51	51.80	49.92	55.05	57.06	60.86	49.45	52.70	49.14	58.80	59.55
MAG-BERT	72.40	68.29	68.87	69.22	72.06	72.94	60.38	54.74	57.51	54.54	59.61	60.00	61.08	50.02	52.29	49.85	59.59	59.60
MulT	72.31	68.97	69.73	68.83	72.07	72.24	60.66	54.12	58.02	53.77	59.55	60.12	59.99	50.69	54.83	50.59	58.67	59.39
TCL-MAP	73.17	68.92	68.90	69.99	72.66	72.97	58.24	52.25	54.28	52.41	57.24	57.55	61.63	50.25	53.32	49.64	59.74	60.16
SDIF-DA	71.64	68.19	69.08	68.30	71.34	71.74	58.06	51.95	53.17	52.16	57.47	57.85	60.91	51.46	57.57	50.80	59.58	60.17
MIntOOD	73.48	70.51	71.19	70.45	72.39	73.24	58.73	52.40	55.48	51.20	58.03	58.34	61.58	50.57	54.97	50.63	59.46	60.37
MVCL-DAF	73.63	70.41	71.07	70.11	73.57	74.31	59.64	53.41	54.90	53.24	58.67	58.57	60.78	48.88	54.05	47.73	59.16	59.83
Qwen2-VL	76.56	74.59	75.85	74.76	76.44	77.19	59.82	47.73	55.23	47.25	58.44	62.51	59.71	52.44	55.19	51.93	59.11	59.15
LLaVA-NeXT	72.65	64.94	66.21	64.65	72.63	73.59	50.61	45.33	50.79	44.14	50.70	54.21	51.30	39.87	43.68	38.36	49.77	50.52
VideoLLaMA2	74.61	71.64	71.82	73.35	74.37	75.43	60.11	49.95	51.58	49.13	59.71	59.90	60.81	51.24	56.20	48.79	59.53	59.79
HIER	**80.00**	**76.91**	**78.98**	**77.11**	**79.59**	**80.67**	**64.15**	**60.31**	**61.90**	**59.59**	**63.79**	**64.17**	**61.95**	**54.80**	**59.41**	**52.94**	**60.38**	**60.44**

实验设计　消融实验

- 评估HIER的不同模块在数据集上的表现
  - -w/o Concept：不构建中层语义概念，直接使用原始token
  - -w/o Relation：不筛选概念关系，缺少高层关系推理
  - -w/o COT：不使用场景理解、概念分析、关系推理的分步推理过程
  - -w/o Self-evolution：不对概念和关系进行有用性评估与动态修正

概念、关系、思维链和自进化模块都对最终性能有贡献

	Ablations	ACC	F1	P	R	WF1	WP
MIntRec	w/o Concept	79.55	76.11	78.06	76.31	79.25	80.30
	w/o Relation	77.08	70.76	71.55	71.52	76.92	77.83
	w/o CoT	75.51	71.57	72.03	72.36	74.95	75.37
	w/o Self-evolution	77.75	71.50	72.05	72.31	77.53	78.20
	HIER	**80.00**	**76.91**	**78.98**	**77.11**	**79.59**	**80.67**
MIntRec2.0	w/o Concept	63.07	50.13	50.39	50.44	62.80	62.99
	w/o Relation	62.67	55.73	57.18	56.19	62.95	64.09
	w/o CoT	63.35	56.81	58.59	56.19	62.87	63.27
	w/o Self-evolution	63.06	58.38	60.61	57.11	62.39	62.88
	HIER	**64.15**	**60.31**	**61.90**	**59.59**	**63.79**	**64.17**
MELD-DA	w/o Concept	59.90	51.85	55.24	51.13	59.07	59.25
	w/o Relation	60.01	51.85	55.84	50.14	58.76	58.49
	w/o CoT	59.91	52.08	56.73	50.13	58.62	58.46
	w/o Self-evolution	60.21	50.50	53.72	48.90	58.74	58.53
	HIER	**61.95**	**54.80**	**59.41**	**52.94**	**60.38**	**60.44**

算法总结 HIER

- 算法贡献
  - 多模态概念聚类：将文本和视频token聚合为中层语义概念，并利用意图标签语义引导概念中心更新，缓解了原始多模态特征零散、语义层次不足的问题
  - 多模态关系选择：将语义概念两两组合，建模概念之间的高阶有价值关系，缓解了单个概念难以表达复杂意图依赖的问题
  - 自进化多模态推理：通过语境理解、概念分析和关系推理完成分步判断，缓解了大模型推理过程难以自适应调整的问题

- 算法不足
  - 依赖多模态大模型，整体训练和推理成本较高
  - 对高度冲突的语义场景仍存在不足

特点总结与未来展望

特点总结与未来展望

- 特点总结
  - LGSRR
    - 利用大模型引导的细粒度语义提取并建模语义关系
      - 缓解了传统方法只做粗粒度模态融合、语义线索不明确的问题
  - HIER
    - 通过概念聚类、关系选择和自进化推理完成意图判断
      - 缓解了原始多模态特征零散、缺少层次化推理的问题

- 未来发展
  - 如何减少对外部大模型生成语义描述的依赖
  - 如何处理文本、表情、动作之间存在冲突的多模态样本

参考文献

[1] Zhou, Qianrui, et al. "LLM-Guided Semantic Relational Reasoning for Multimodal Intent Recognition." Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025.

[2] Zhou, Qianrui, et al. "Evolutionary Multimodal Reasoning via Hierarchical Semantic Representation for Intent Recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2026.

道德经

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！