# 鲁棒对话智能：复杂场景下人机语音交互感知与表达




内蒙古大学计算机学院、人工智能学院
蒙古文智能信息处理技术国家地方联合工程研究中心
内蒙古自治区多语种人工智能技术重点实验室

**刘瑞**

**2026年11月06日**

# 报告人介绍



1. 内蒙古大学人工智能学院副院长、教授、博士生导师；入选中国科协青年人才托举工程
2. 主持**国自然面上/青年、自治区杰青、中国科协青年人才托举工程**等 10 余项项目
3. 研究方向：多语种信息处理、共情人机语音交互、多模态情感计算
4. TASLP/TAFFC/ACL/ACMMM/AAAI等发表成果，国际会议最佳论文奖2项
5. Information Fusion、IEEE-TAFFC、ACM-TALLIP 期刊编委

# 从训练数据到真实复杂对话场景

对话是人类交流信息、表达情感和传递意图的重要方式，广泛存在于日常交流、会议讨论和人机交互等场景中。通过语言、语音及非语言行为交换信息、理解意图、情绪并作出回应的交互过程。

**训练阶段的理想场景：**

- ***轮流发言、信息完整、语音单模态交互***

**真实世界的复杂场景：**

- ***多人交互、信息缺失、跨模态协同表达***



# 真实世界的复杂对话场景主要面临以下三方面挑战：

## 1.多说话人场景下的语音感知



多人交替发言，导致语音重叠，使语音内容提取与说话人归属更加困难。

# 真实复杂的对话场景主要面临以下三方面挑战：

## 2.模态缺失下的情感-意图理解



**软硬件设备现实问题，导致语音、文本或视觉信息的缺失，导致情感-意图理解偏差。**

# 真实复杂的对话场景主要面临以下三方面挑战：

## 3.多模态交互中的语音与视觉协同表达

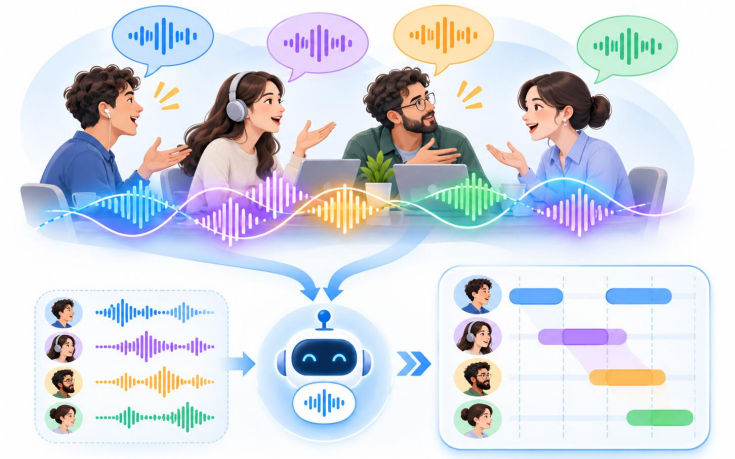


**单模态语音对话虽然实现自然交互，但只有实现语音、人脸、面部表情甚至肢体动作的跨模态协调表达才能实现更加友好的人机交互**

# 本教程希望回答三个问题

## 感知

如何在多人交替、重叠发言的情况下，准确识别语音内容并确定说话人归属?



## 理解

如何在模态信息不完整的条件下，准确理解用户的情感与交互意图?



## 表达

如何实现语音、面部表情与动作在时间、语义和情感上的协调一致?



# 教程结构与讲述安排

1. 多人对话场景下的语义感知
2. 模态缺失下的情感意图理解
3. 对话交互中多模态协同表达
4. 总结与展望

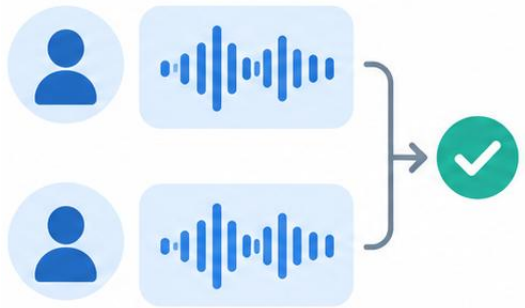
# 第一部分

# 多人对话场景下的语义感知

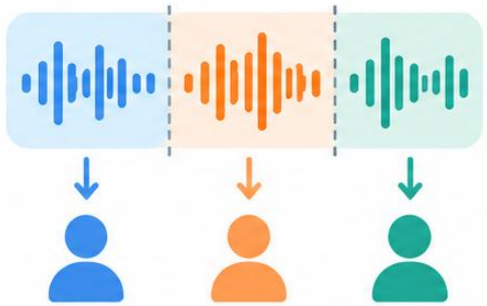
# 说话人验证、说话人分离与多说话人语音识别

## 三个任务的目标、输入和输出并不相同

- 说话人验证：判断两段声音是否属于同一人
- 说话人分离：标出每个时间区间由谁说话（When+Who）
- 多说话人语音识别：输出含转写内容、说话人归属和时序信息（What+Who+When）



说话人验证



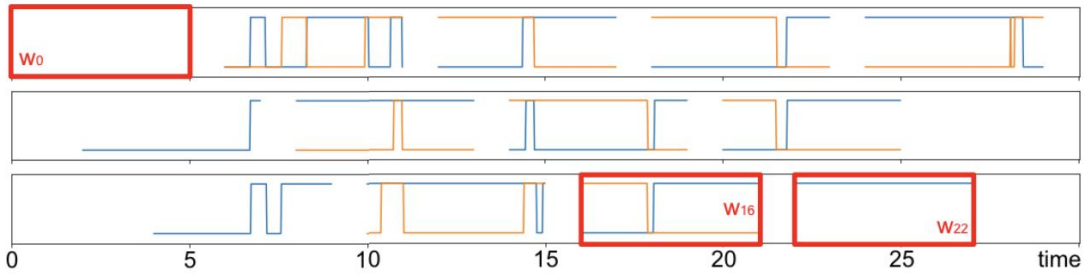
说话人分离



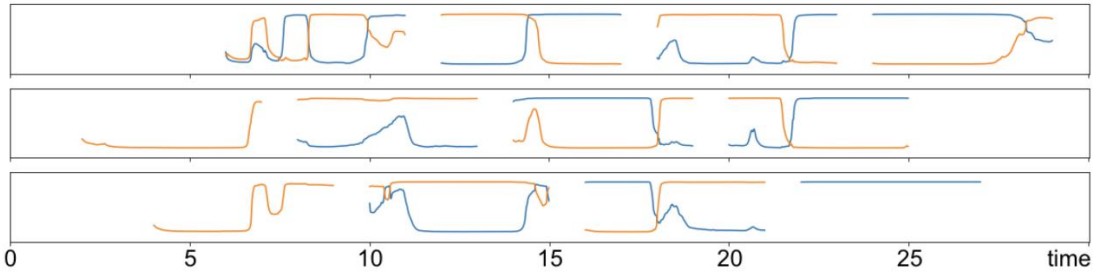
多说话人语音识别

# pyannote.audio：局部分割与全局聚类的说话人标注

**目标：在多人轮换与重叠发言的情况下，准确标注各说话人的发言时间，并保持跨片段身份一致。**




Binary local speaker segmentation



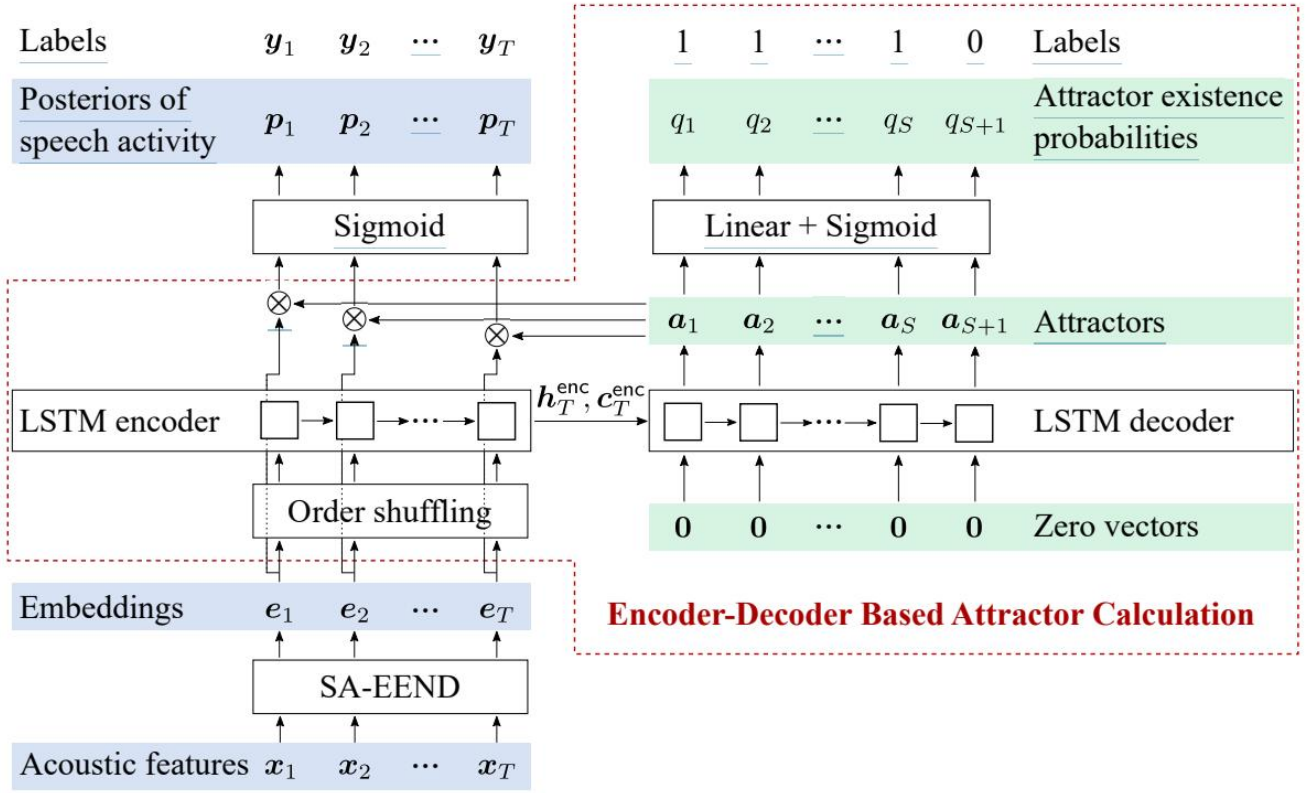

Local speaker segmentation after global clustering

**通过滑动窗口预测局部说话人活动，利用重叠感知特征提取减少语音混叠干扰，并结合全局聚类与时序聚合形成整段录音的说话人标注。**

Bredin H. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe[C]//**24th Interspeech Conference (INTERSPEECH 2023)**. ISCA, 2023: 1983-1987.

# EEND-EDA：多说话人活动的端到端联合预测

**目标：在说话人数未知且存在重叠发言的情况下，准确判断各说话人在不同时间的活动状态。**




**通过时序编码提取帧级语音表示，利用编码器-解码器生成说话人表示，并估计说话人数，联合预测允许重叠的多说话人活动概率。**

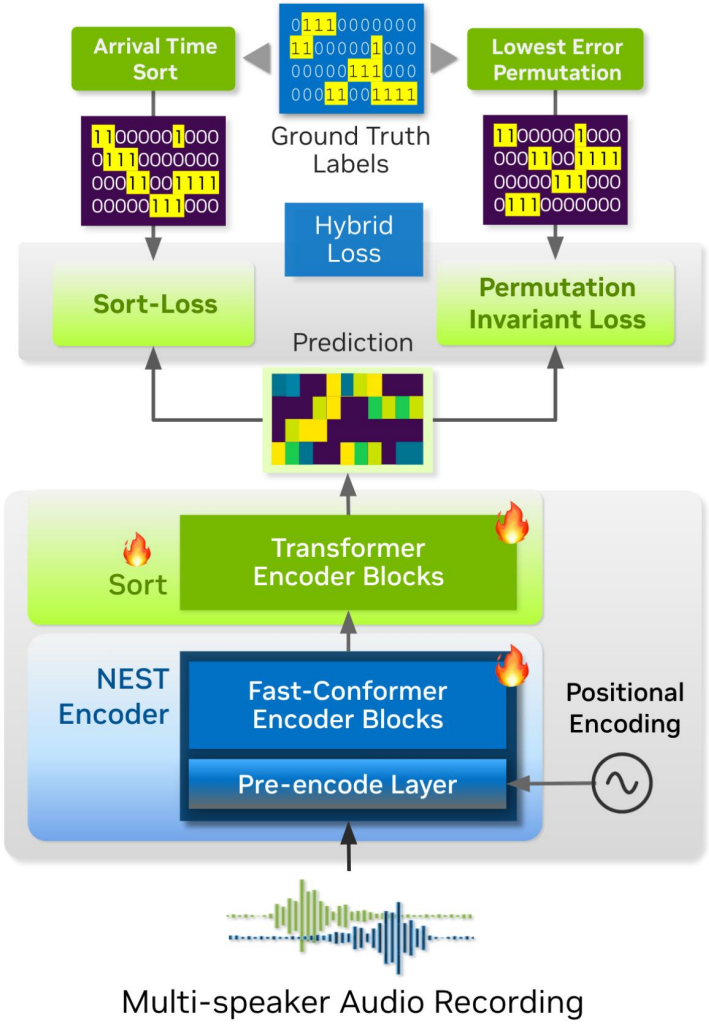
Horiguchi S, Fujita Y, Watanabe S, et al. Encoder-decoder based attractors for end-to-end neural diarization[J]. **IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022**, 30: 1493-1507.

# Sortformer：首次发言顺序约束的说话人标签建模

ICML 2025

**目标：在多说话人活动预测中明确标签输出顺序，缓解标签置换带来的监督歧义，并辅助转写文本的说话人归属。**

**核心技术：依据首次发言时间排序说话人标签，结合 排序损失 与 排列不变损失 进行混合优化，并将说话人活动信息注入语音识别模型。**




Park T, Medennikov I, Dhawan K, et al. Sortformer: A novel approach for permutation-resolved speaker supervision in speech-to-text systems[C]//**Proceedings of the 42nd International Conference on Machine Learning. PMLR**, 2025, 267: 48153-48169.

# 从Diarization到多说话人语音识别

## 说话人日志

| | |
|---|---|
| 00-03s | 说话人 A |
| 02-05s | 说话人 B |
| 05-07s | 说话人 A |

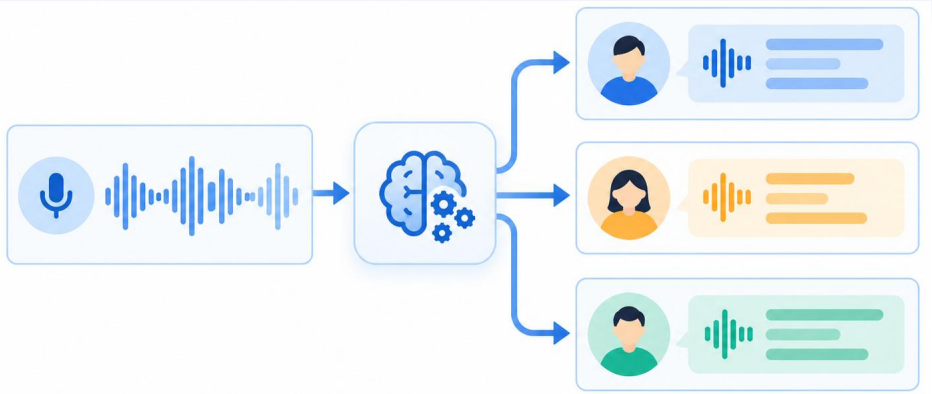
**仅包含了时间与身份**



## 多说话人识别

| | |
|---|---|
| 00-03s | 说话人 A：“大家好，今天...” |
| 02-05s | 说话人 B：... |
| 05-07s | 说话人 A：... |

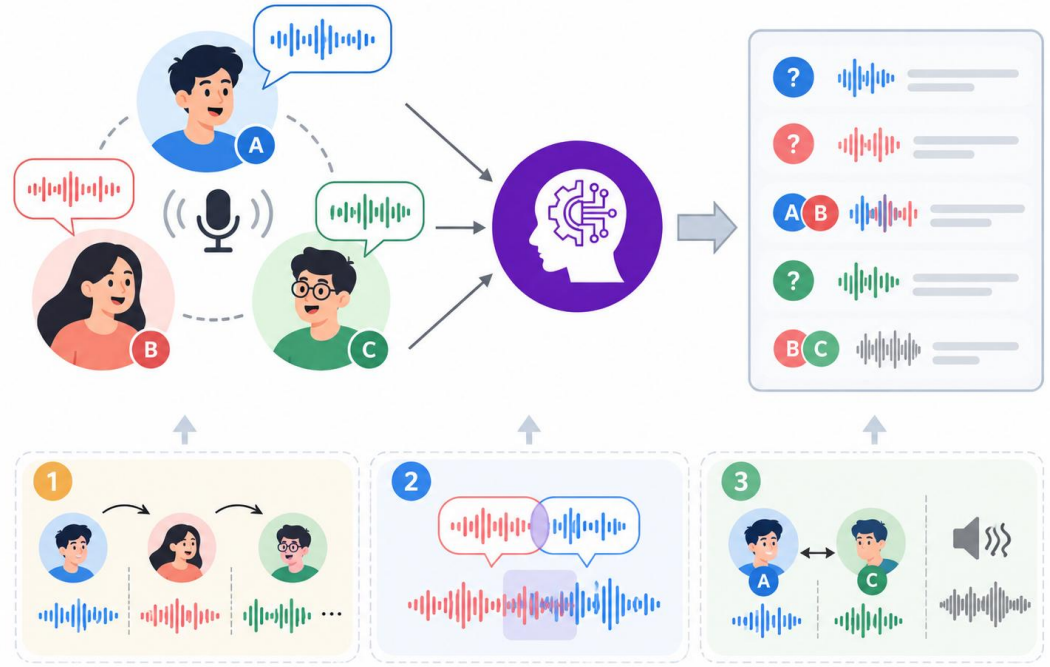
**身份 + 时间 + 内容**



# 多说话人语音识别为何比单人ASR、或对话日志更难?

➢ **正确转写词语，不代表正确理解多人对话**

- 说话人快速轮换导致边界感知错误
- 重叠发言使同一时刻存在多个语音源
- 音色相近、距离变化、混响造成身份混淆



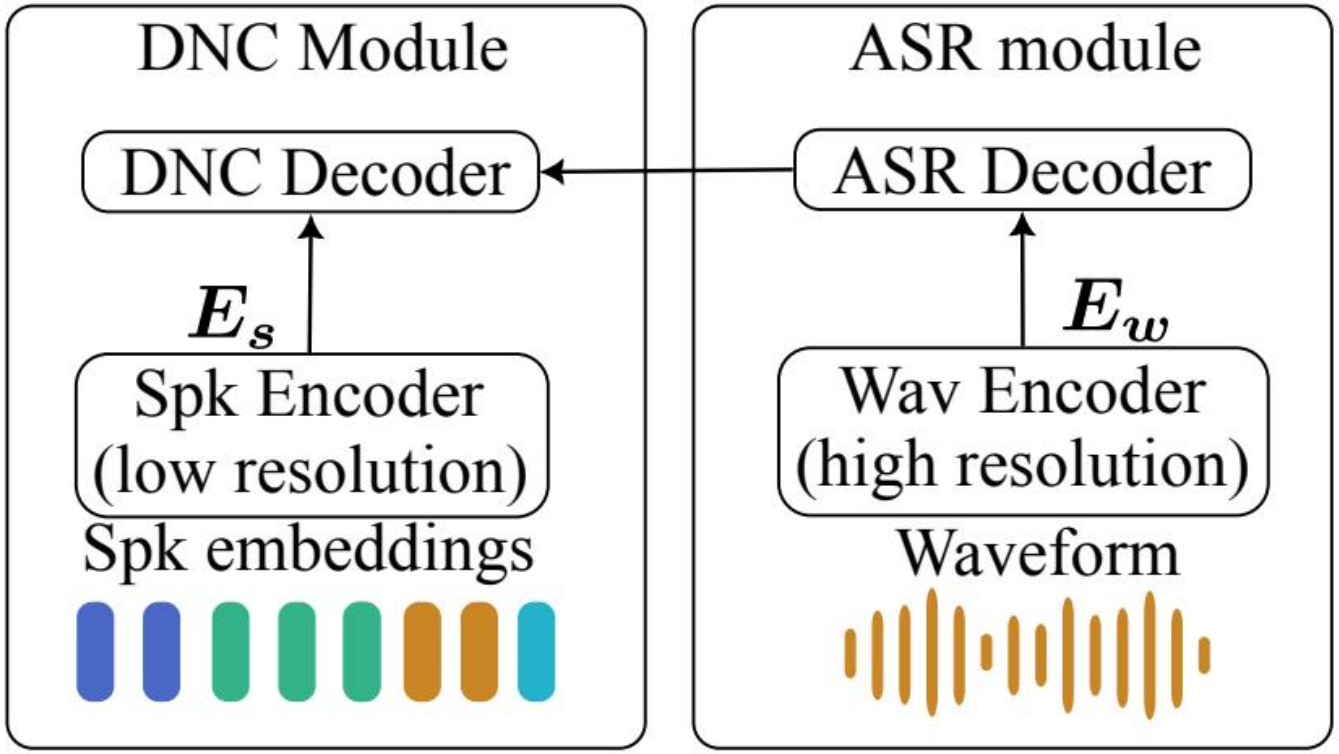
# Who，What，When 形式化表示

1. 对一段多说话人波形 X，预测序列 {(说话人身份,开始时间,结束时间,文本内容)}
2. 说话人标记可以是匿名的 Speaker 1/2，不必对应真实身份
3. 重叠区间可以出现两个同时活动的标签



# DNCASR：说话人聚类与语音转写的联合建模

**目标：在长时多人会议中，准确识别发言内容，并保持转写文本与说话人归属的一致性。**

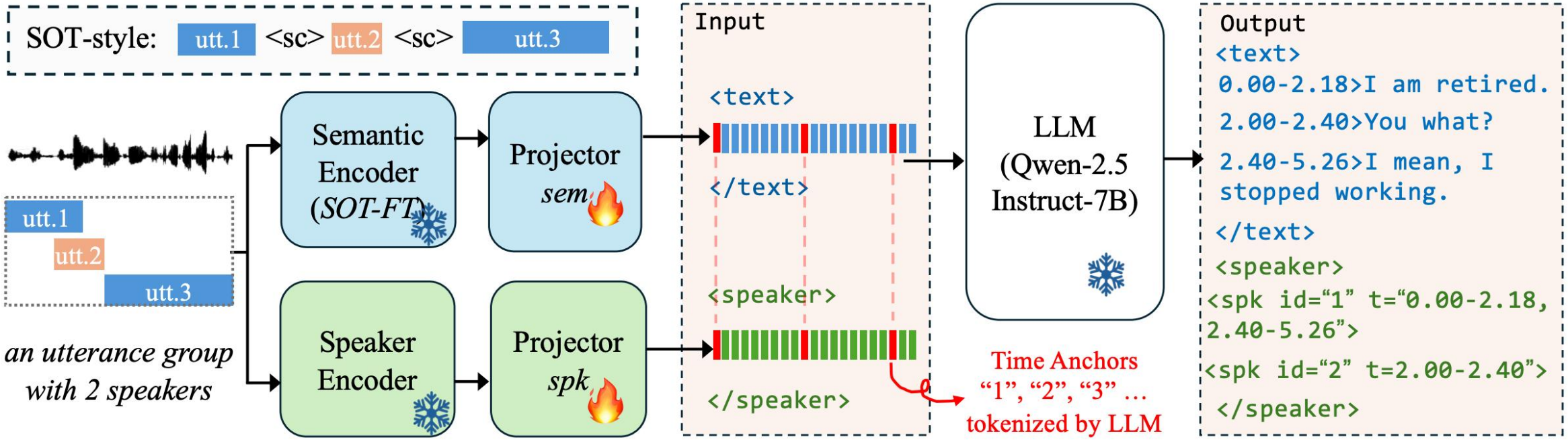



**核心技术：通过双路编码提取全局说话人特征与局部语音内容特征，利用跨解码器注意力关联身份与转写信息，并采用两阶段训练实现联合优化。**

Zheng X, Zhang C, Woodland P. DNCASR: End-to-end training for speaker-attributed ASR[C]//**Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025**: 18369-18383.

# TagSpeech：时间锚点驱动的语音内容与说话人对齐

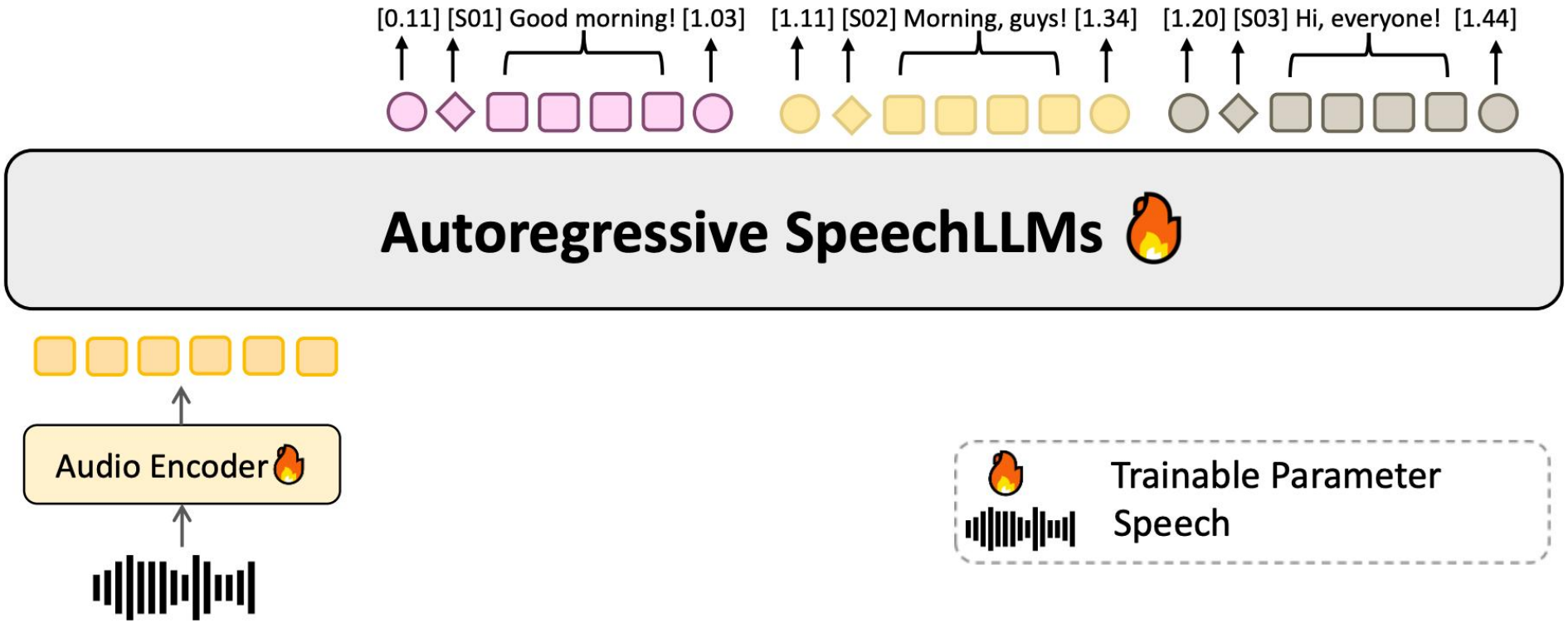
**目标：在多说话人语音中，实现转写内容、说话人身份与时间边界的精确对齐。**




**核心技术：通过语义与说话人双流编码提取内容和身份特征，利用交错时间锚点建立细粒度时序关联，给模型提供"时间刻度"。**

Huo M, Shao Y, Zhang Y. Tagspeech: End-to-end multi-speaker asr and diarization with fine-grained temporal grounding[C]//Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2026: 41847-41862.

# MOSS-Transcribe-Diarize：长上下文多说话人语音识别

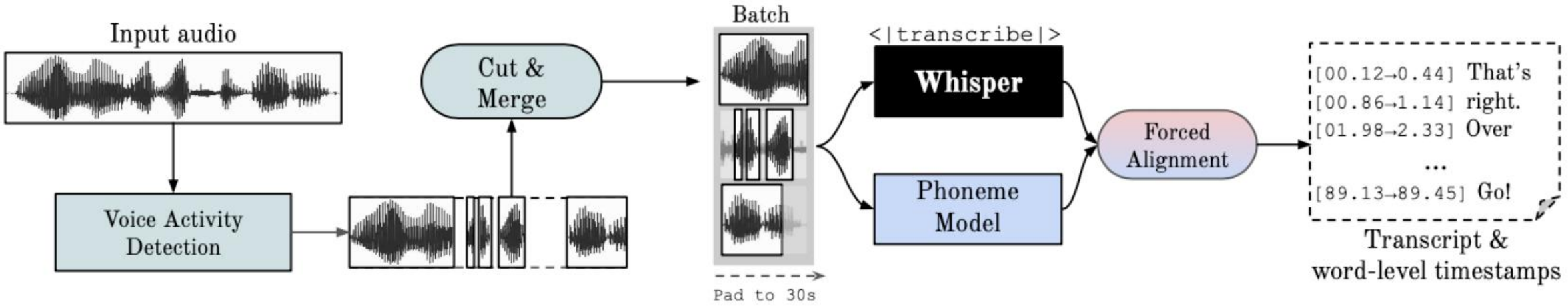



**核心技术：通过长上下文语音大模型联合建模转写内容、说话人身份与时间戳，实现最长90分钟多人语音的端到端结构化识别。**

https://github.com/OpenMOSS/MOSS-Transcribe-Diarize

# WhisperX：高效语音转写与词级时间对齐

**WhisperX 并不是将所有任务完全统一为一个端到端模型，而是利用模块化工具建立高效处理流程。**

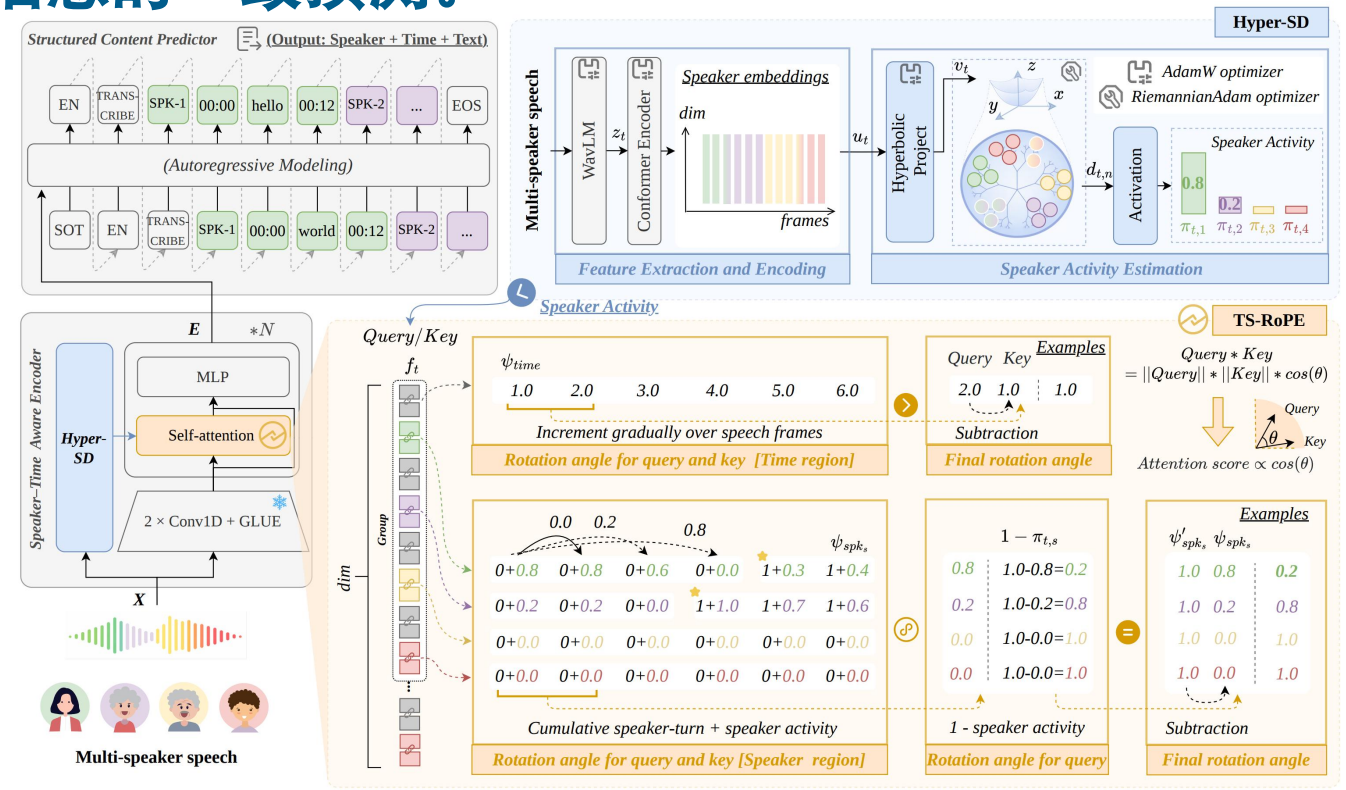



**核心技术：通过语音活动检测与分段合并实现批量转写，利用强制对齐获取词级时间戳，并可结合说话人分离标注确定词语归属。**

https://github.com/m-bain/whisperX

# TellWhisper：让语音编码器同时感知时间与身份

**目标：在多人语音中，加强时间结构与说话人身份的关联，实现语音内容、说话人归属与时间信息的一致预测。**



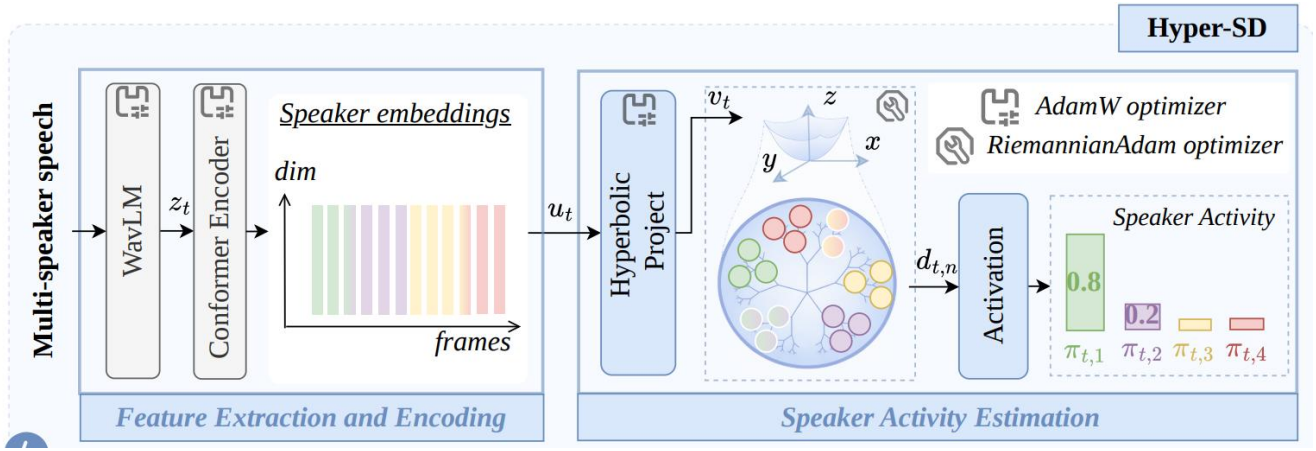

团队工作

ACL 2026

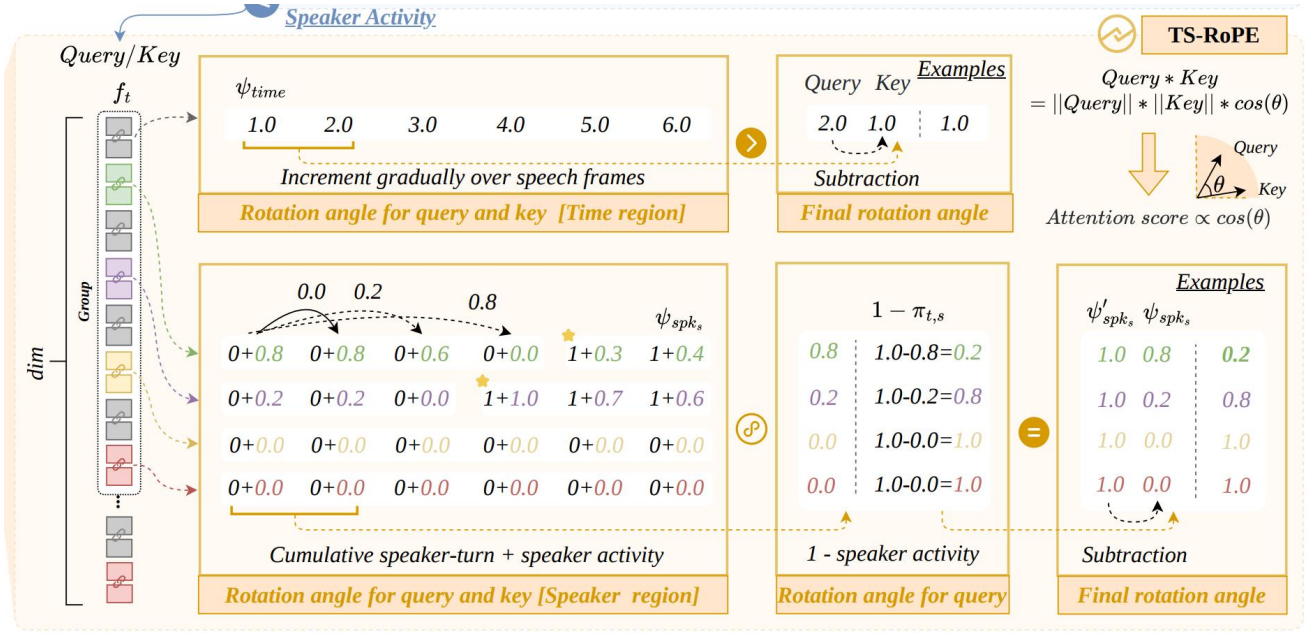
**核心技术：通过Hyper-SD在双曲空间估计帧级说话人活动，利用TS-RoPE将时间与身份信息融入语音编码，并通过结构化解码联合输出文本、说话人标签与时间戳。**


**Hu Y**, Yang P, Wang Z, et al. TellWhisper: Tell Whisper Who Speaks When[C]//**Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2026**: 18884-18898.

# TS-RoPE与Hyper-SD：从何时到何人




Hyper-SD：利用**双曲空间**表示与说话人原型，估计帧级说话人活动。




TS-RoPE：将**时间位置与说话人活动融入旋转位置编码**，联合建模时间与身份关联。

# 研究总结｜多说话人场景下的语音感知

01

**分割聚类**

- 识别说话人活动区间，保持跨时间身份一致性
- pyannote.audio（InterSpeech 23）

02

**身份建模**

- 通过动态身份表示与顺序约束，优化说话人身份预测
- EEND-EDA（TASLP 22）、Sortformer（ICML 25）

03

**联合识别**

- 关联语音内容、说话人身份与时间信息，实现协同预测
- DNCASR （ACL 25）、WhisperX、TagSpeech（ACL 26）

04

**最新策略**

- 利用长上下文与联合编码，增强身份与时间建模
- MOSS、TellWhisper （ACL 2026）

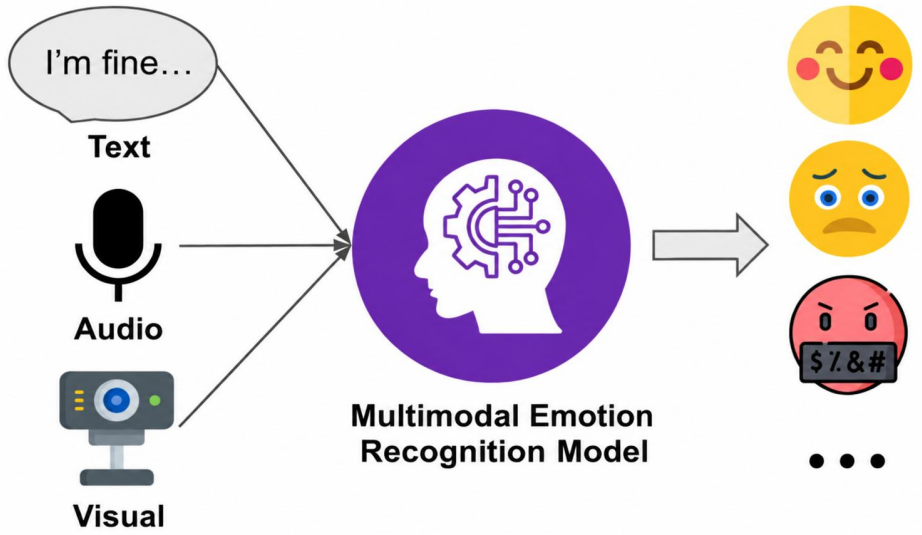
# 第二部分

# 模态缺失下的情感意图理解

# 多模态情感识别

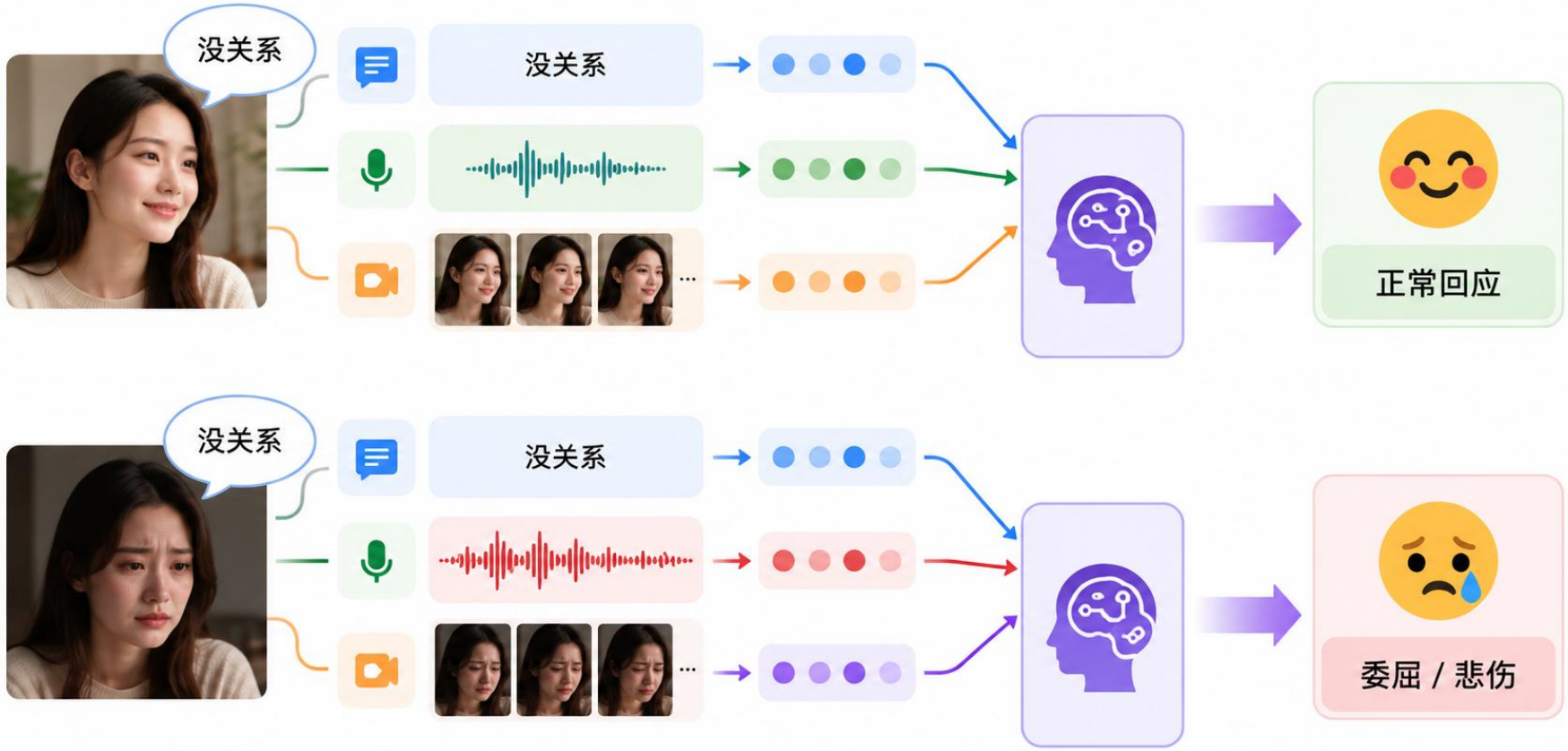
- 语言、声学和视觉共同表达人的情绪状态

- 文本主要承载词义和事件语义等
- 语音通过音高、节奏、能量传达非语言信息
- 视觉提供表情、头姿、动作等线索
- 感知情感类别或连续情感维度



# 情感理解不仅取决于字面语义

- 相同的文本，语音与表情线索可能揭示不同的真实情感



# 模态缺失下的情感理解挑战

- 多模态情感理解依赖跨模态互补证据，而模态缺失会导致关键情感信息不完整

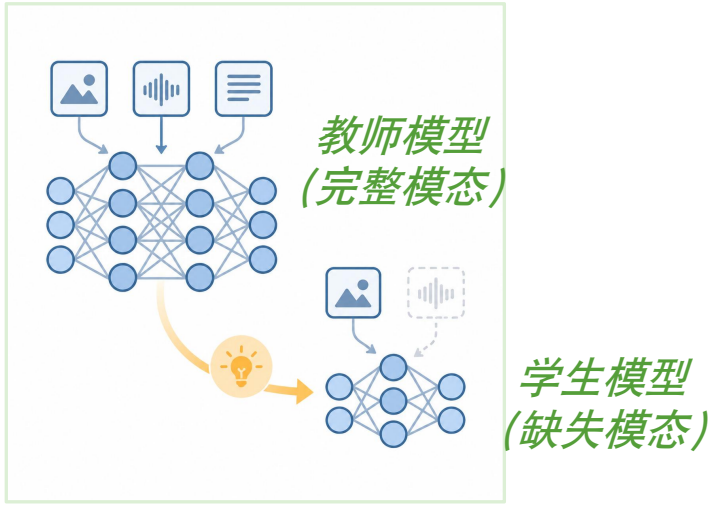


# 技术全景：缺失模态学习的四条路线

- **重构补全：** 显式恢复缺失特征



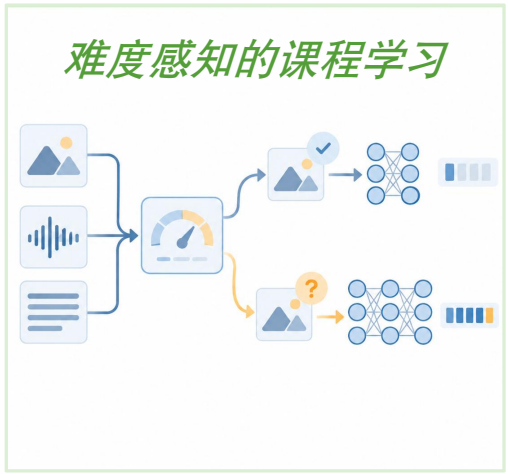

- **知识蒸馏：** 将完整模态教师的知识传给学生




- **直接增强：** 直接增强可用模态信息




- **训练策略：** 根据样本难度配置学习资源



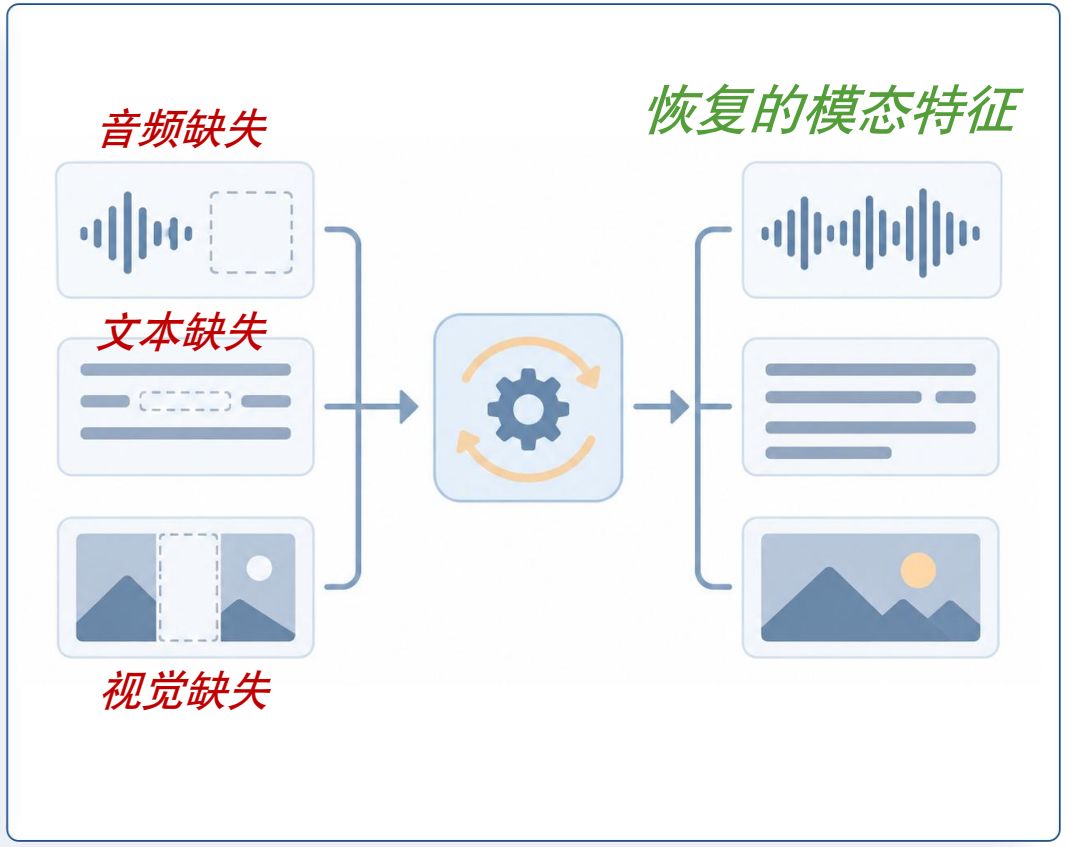
# 路线一：缺失信息的补全与重构

**重构不一定能恢复原始数据，但需尽量保留情感相关信息**

➢ 重构形式可为模态特征或条件生成分布

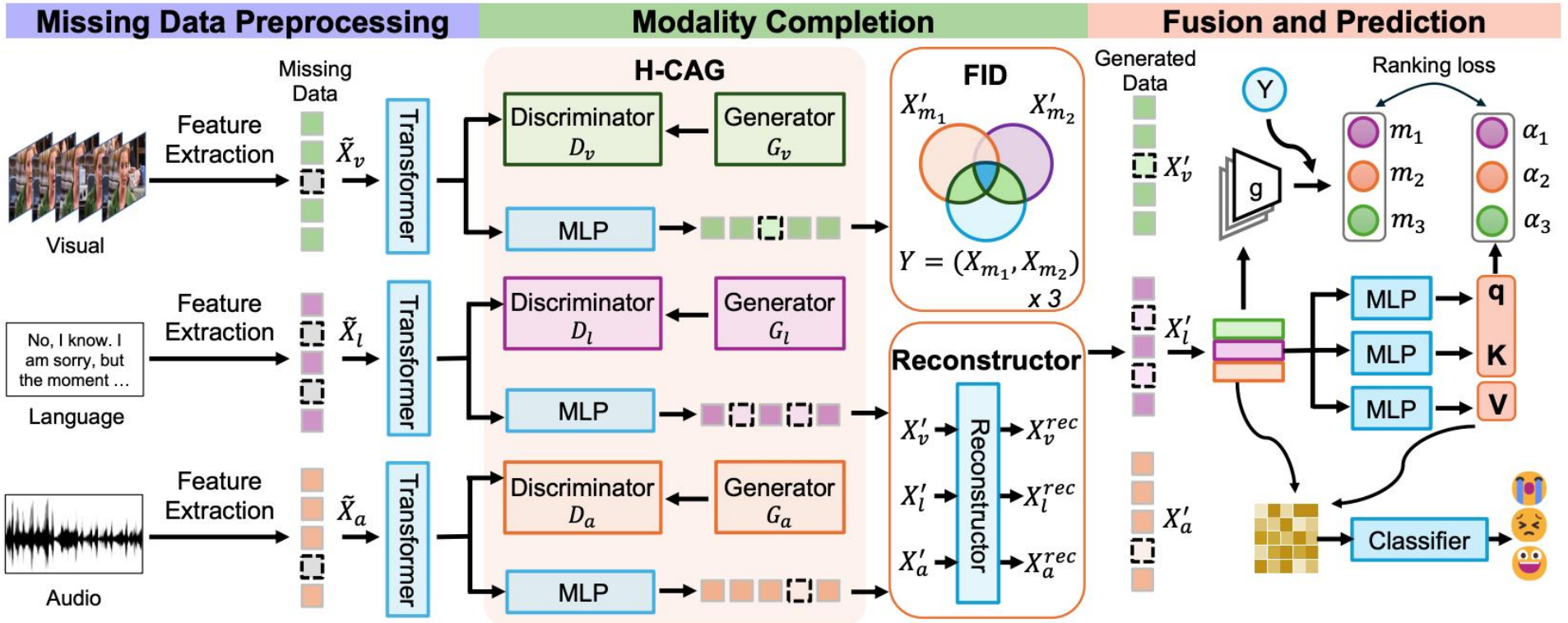
**优点：** 显式补全便于复用原有融合模型

**局限：** 欠约束估计可能造成情感语义偏差，（尤其在强缺失时）



# RECAP：从特征重建到情感模式恢复

**目标：在模态不完整的情况下，恢复具有时间连续性与跨模态一致性的情感信息。**

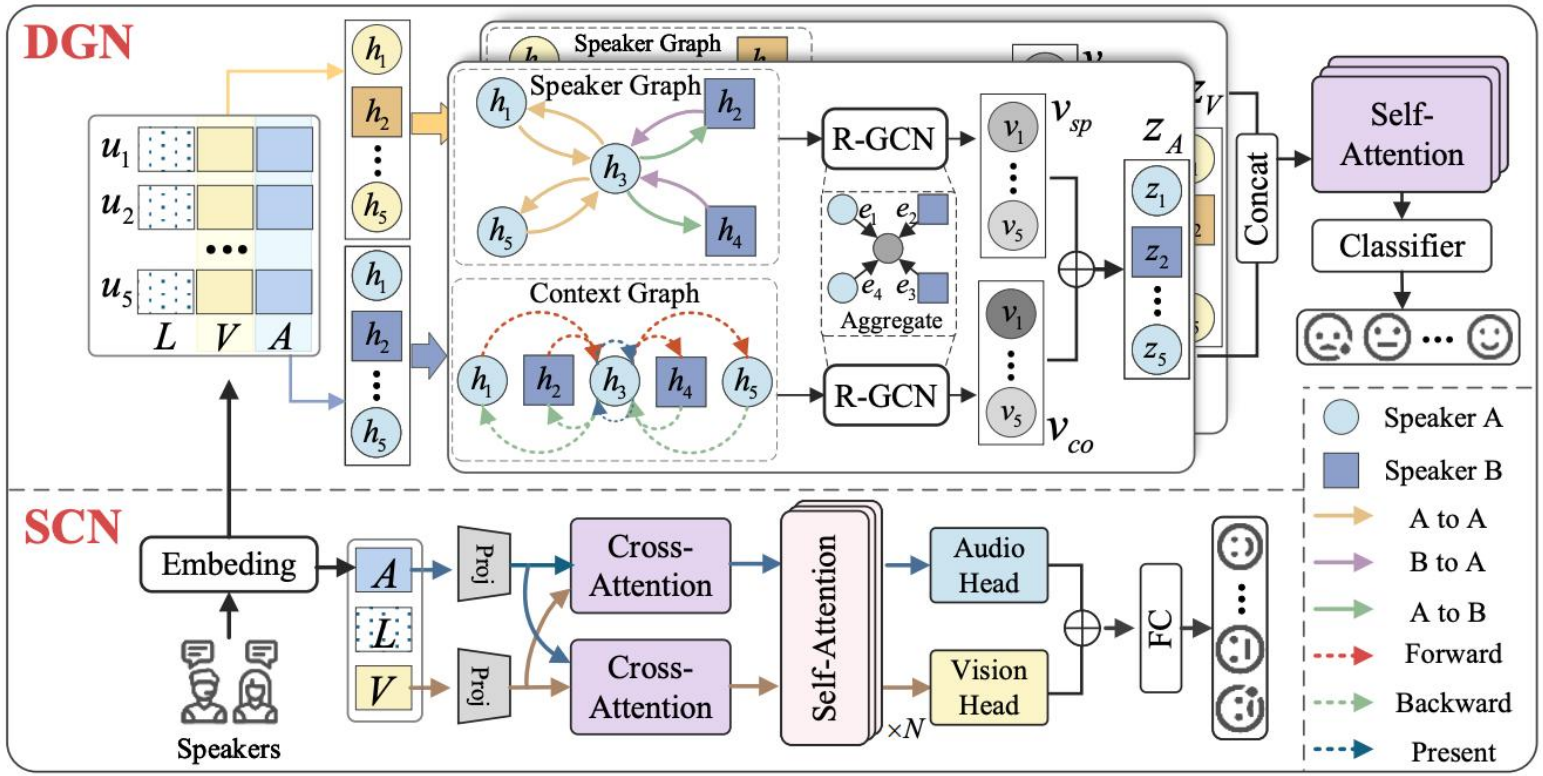



**核心技术：通过多粒度对抗生成恢复情感连续性，利用互信息分解保留共享与特有信息，并根据信息贡献排序实现自适应融合。**

Huang H, Gong T, He K, et al. Recovering coherent affective patterns: Addressing modality missing in multimodal sentiment analysis[C]//**Proceedings of the AAAI Conference on Artificial Intelligence. 2026**, 40(26): 21957-21965.

# FedDISC：利用对话结构约束缺失恢复

**目标：在不同客户端模态不完整的情况下，恢复符合对话语境的缺失信息，提升情感识别的鲁棒性。**

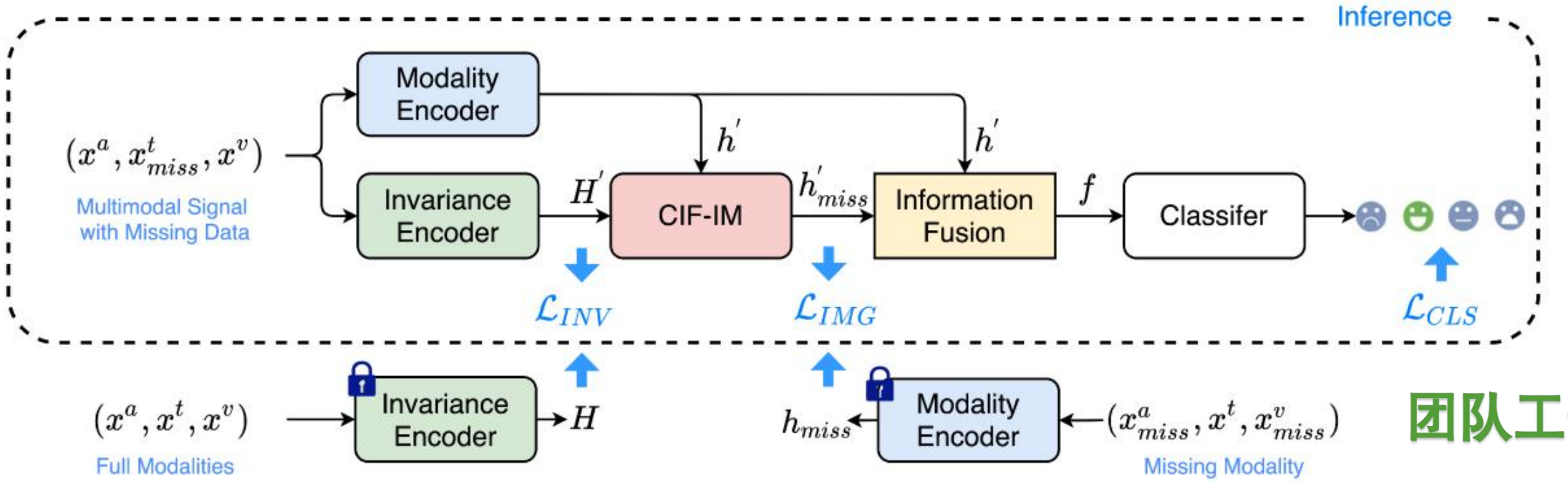



**核心技术：通过联邦扩散共享模态恢复的生成知识，利用对话关系与跨模态语义图结构约束引导特征恢复，并采用交替冻结策略协调模态恢复与情感分类。**

Qiu X, Cheng J, Fang Y, et al. Federated dialogue-semantic diffusion for emotion recognition under incomplete modalities[J]. **Advances in Neural Information Processing Systems, 2025**, 38: 95833-95862.

# CIF-MMIN：跨模态共同情感语义引导的缺失重构

**目标：在模态缺失的情况下，缓解不同模态间的表示差异，恢复具有一致情感语义的缺失信息。**




团队工作

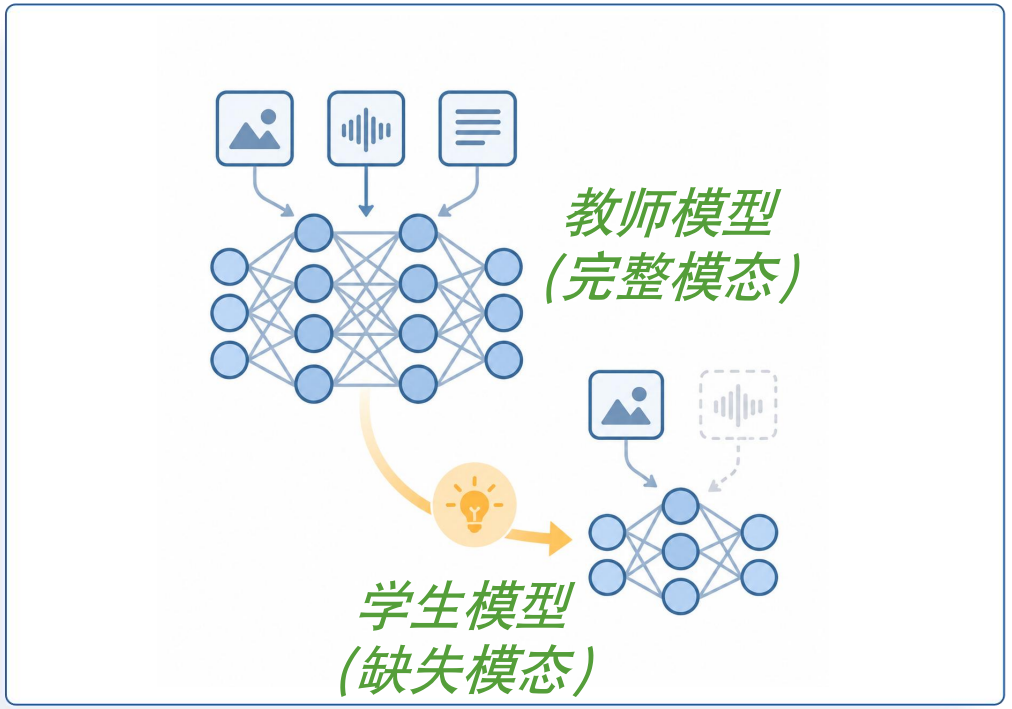
**核心技术：通过对比学习提取跨模态不变特征，引导缺失模态信息重构，并融合补全信息与可用模态特征，实现鲁棒情感识别。**

**Liu R**, Zuo H, Lian Z, et al. Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities[J]. **IEEE Transactions on Affective Computing, 2024**, 15(4): 1856-1873.

# 路线二：完整教师到缺失学生的知识蒸馏

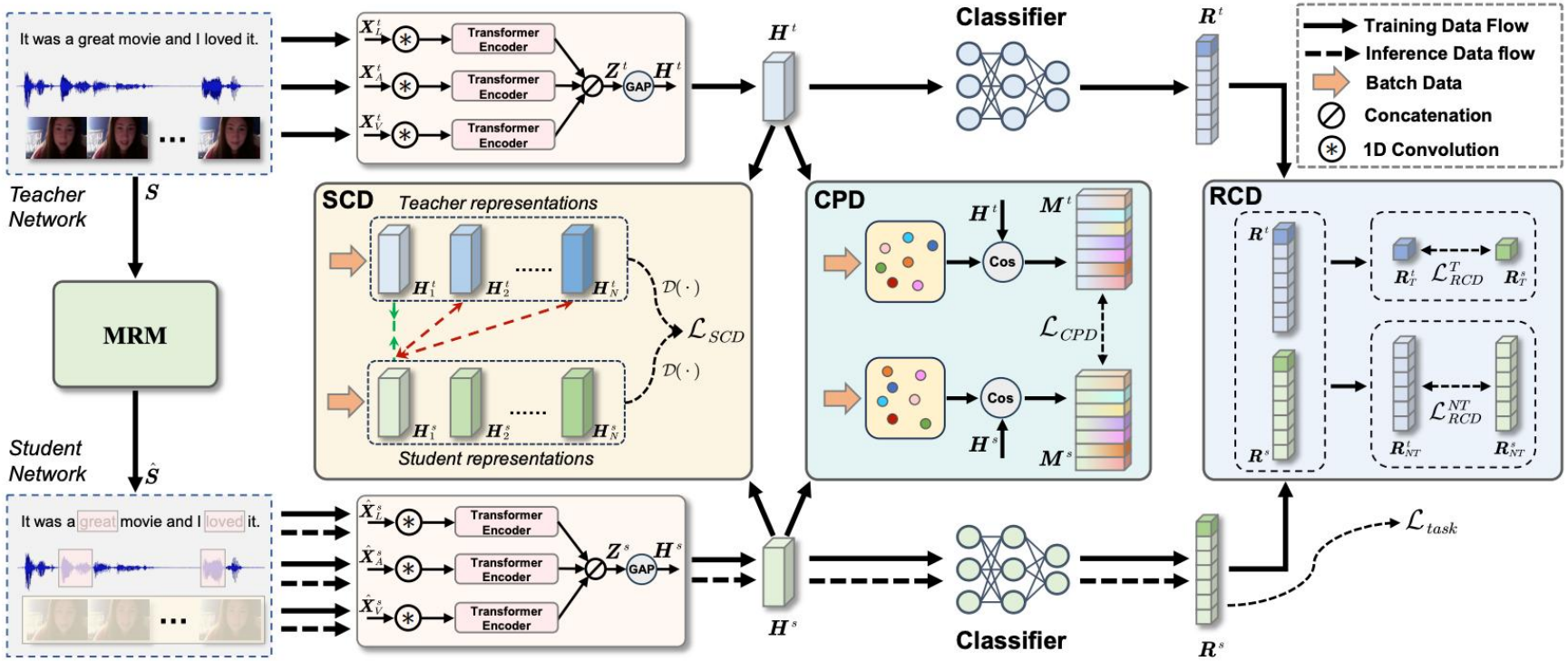
**不直接恢复输入，也可以迁移“如何做出可靠判断”**

- 完整输入Teacher提供高质量表示或软标签
- 残缺输入Student学习一致的情感判别结构
- 蒸馏应区分特征空间对齐、样本关系和决策边界



# CorrKD：样本、类别与预测响应的多层次知识蒸馏

**目标：在模态缺失的情况下，充分迁移完整模态教师的情感判别知识，提升学生模型的识别能力。**

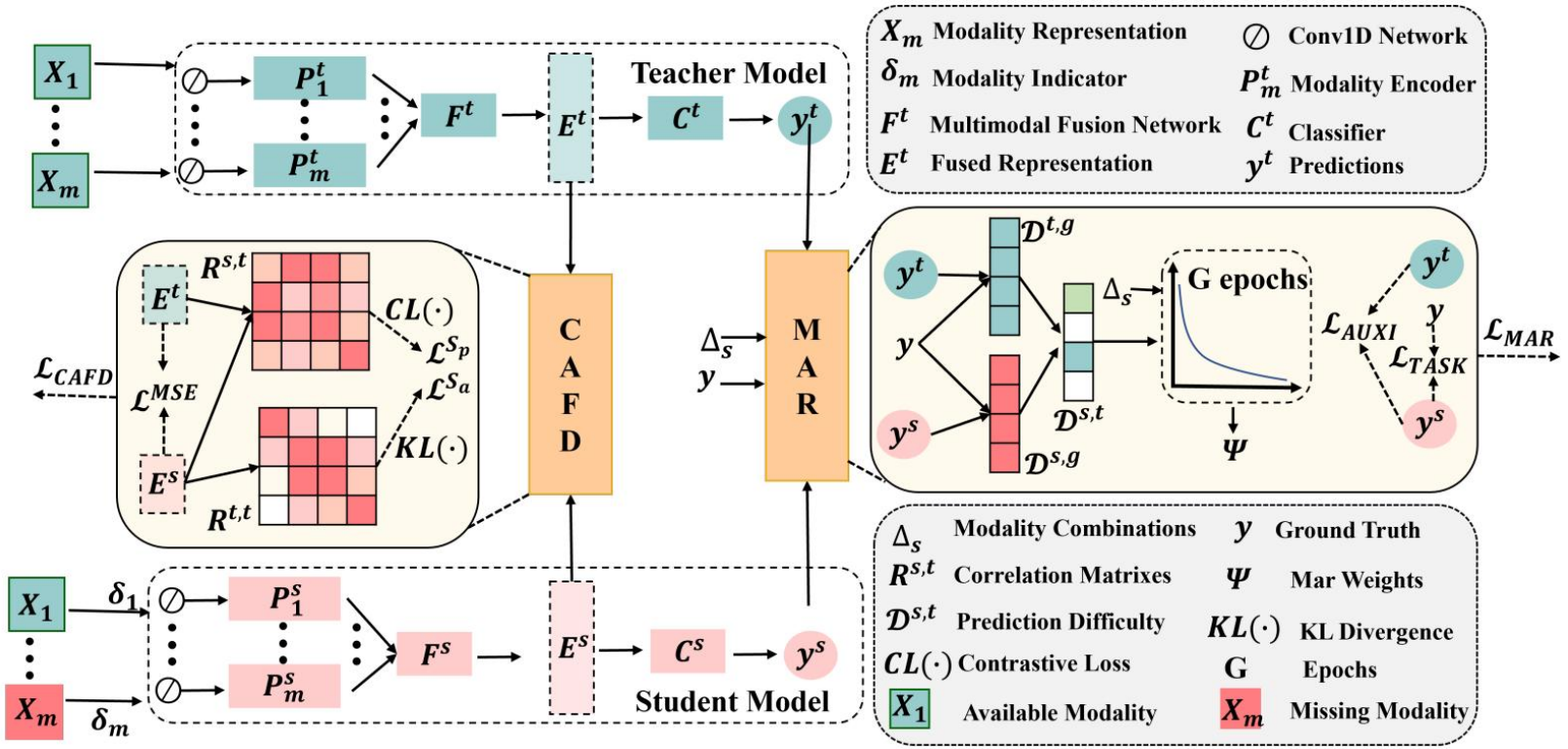



**核心技术：通过样本级对比蒸馏保持样本关系，利用类别原型蒸馏迁移类别结构，并通过预测响应解耦蒸馏传递决策知识。**

Li M, Yang D, Zhao X, et al. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities[C]//**2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. IEEE, 2024: 12458-12468.

# CMAD：关联知识蒸馏与模态难度感知学习

**目标：适应不同缺失模态组合的学习难度差异，提升模型在多种模态缺失条件下的情感识别能力。**




**核心技术：通过师生特征对齐迁移表示知识，利用跨样本关联蒸馏保持样本关系，并通过模态难度感知课程学习动态调整训练过程。**

Zhuang Y, Liu M, Bai W, et al. Cmad: Correlation-aware and modalities-aware distillation for multimodal sentiment analysis with missing modalities[C]//**2025 IEEE/CVF International Conference on Computer Vision (ICCV)**. IEEE, 2025: 4626-4636.

# 路线三：只增强可用模态，不显式重构

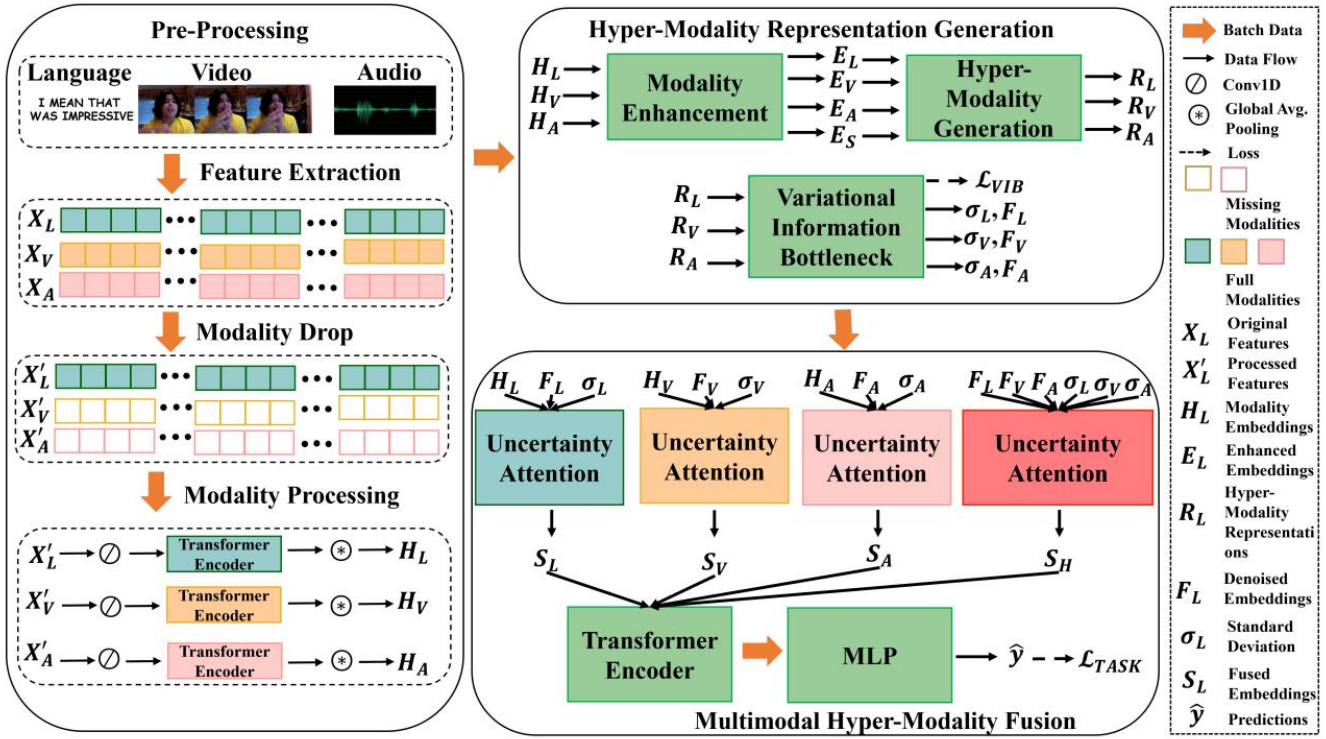
**缺失情形下，检索与提示可能比生成假模态更直接**

- 通过跨样本语义检索丰富可用模态
- 减少重构偏差，但需要防止错误检索与提示噪声



# HME：跨样本语义增强与超模态融合

**目标：在不显式重构缺失模态的情况下，增强可用模态的情感信息，并降低不可靠辅助信息的干扰。**

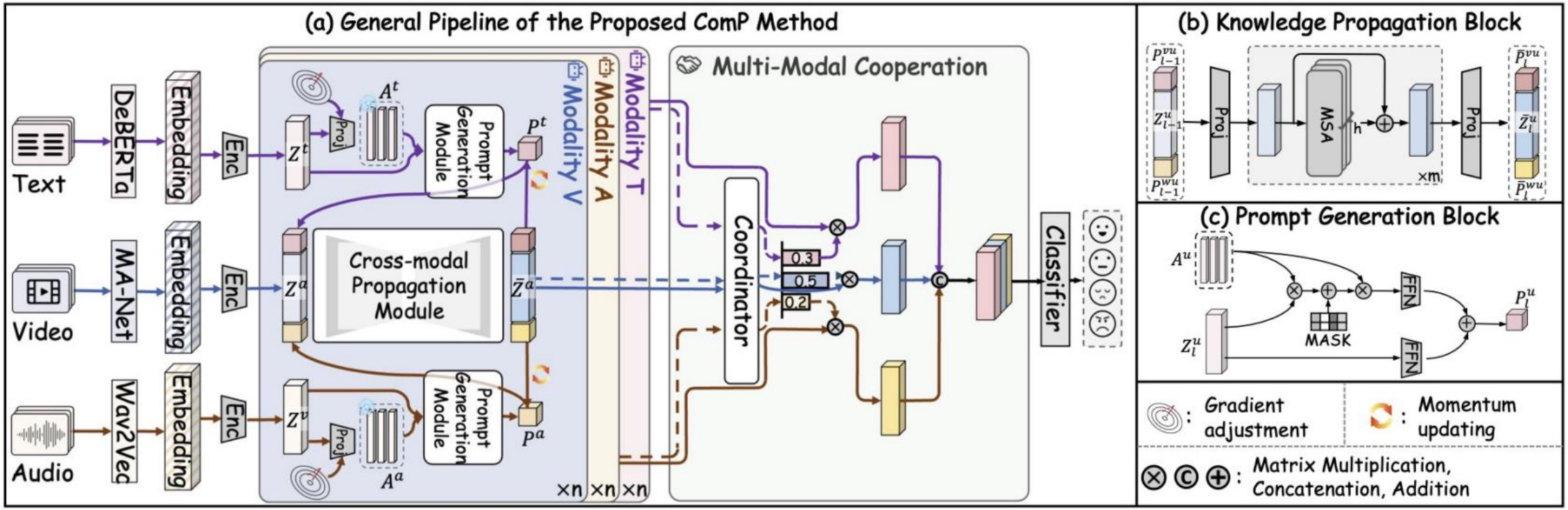



**核心技术：通过跨样本检索与可学习提示增强语义，聚合互补信息形成超模态表示，利用信息瓶颈抑制冗余，并通过不确定性感知融合动态协调信息贡献。**

Zhuang Y, Liu M, Bai W, et al. Hyper-modality enhancement for multimodal sentiment analysis with missing modalities[J]. **Advances in Neural Information Processing Systems, 2025**, 38: 142100-142129.

# ComP：跨模态提示驱动的均衡学习

**目标：在模态缺失的情况下，缓解模态间的学习不均衡，降低模型对优势模态的依赖。**



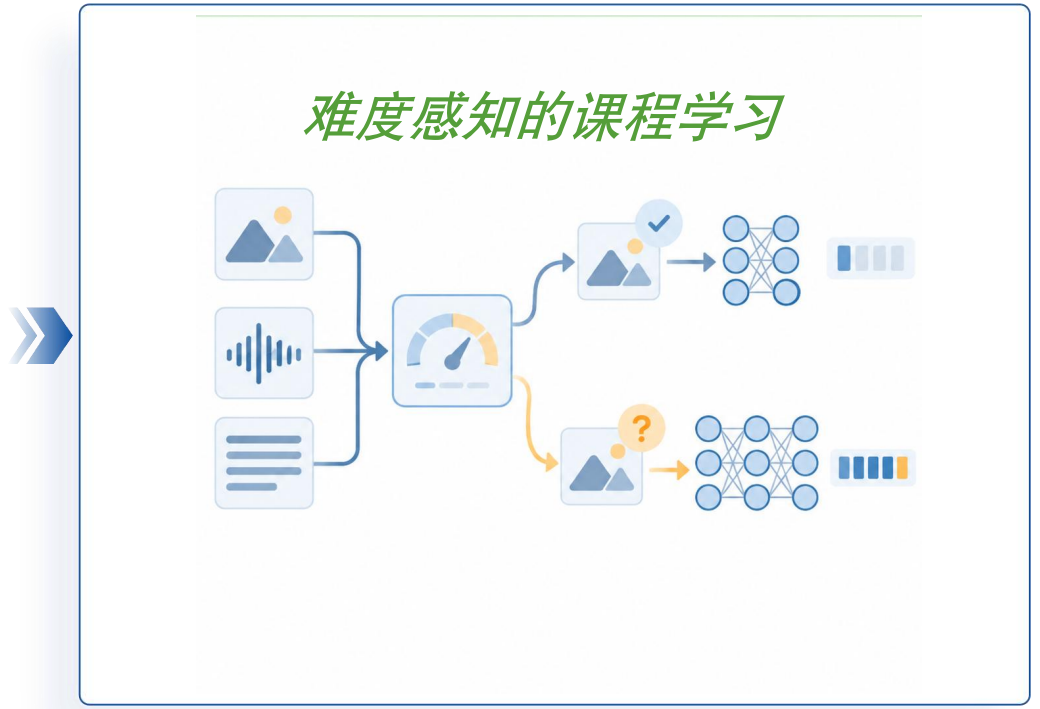

**核心技术：通过渐进式提示生成提取具有情感代表性的模态信息，利用跨模态知识传播增强各模态表示，并通过动态协同融合协调不同模态的预测贡献。**

He W J, Zhu X, Zhang Z. Cross-modal prompting for balanced incomplete multi-modal emotion recognition[C]//**Proceedings of the AAAI Conference on Artificial Intelligence. 2026**, 40(21): 17463-17471.

# 路线四：样本难度感知的课程学习

**相同缺失组合的样本，难度仍然可能不同**

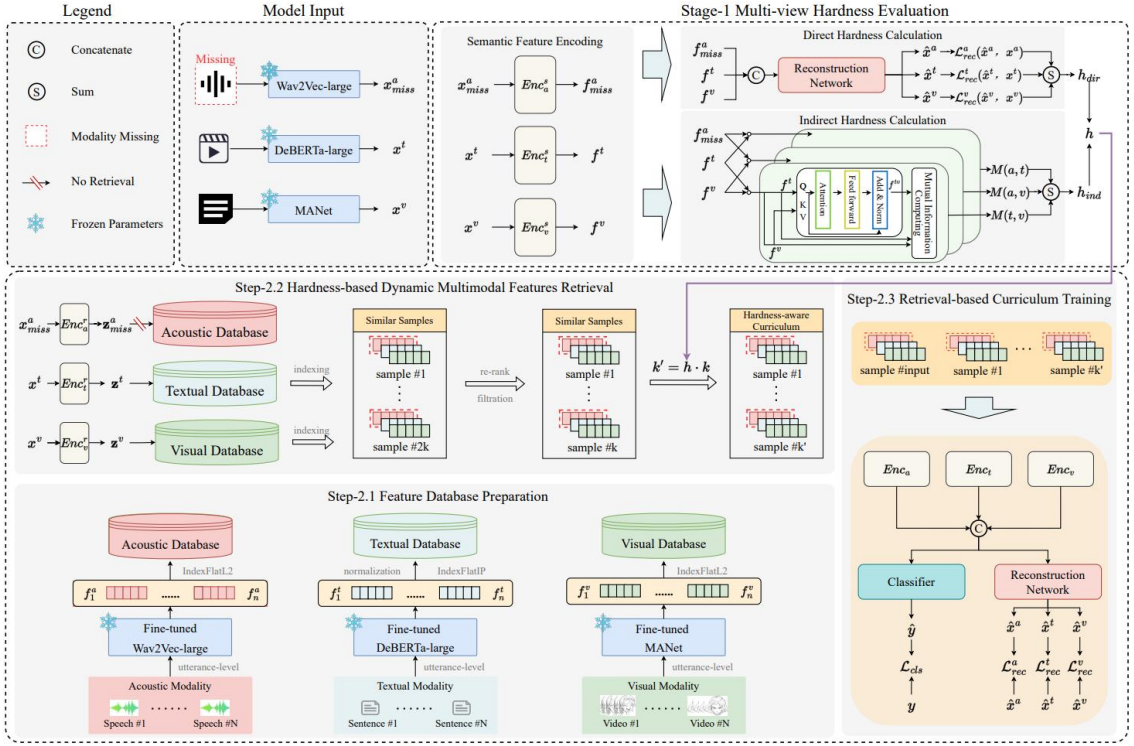
- 困难可能来自跨模态差异、情绪强度和历史线索贫乏
- 检索语义相关辅助样本并动态调整学习内容



# HARDY-MER：样本难度感知的检索增强课程学习

**目标：在模态缺失的情况下，根据样本难度分配辅助信息与训练资源，提升困难样本的特征恢复与情感识别能力。**

团队工作



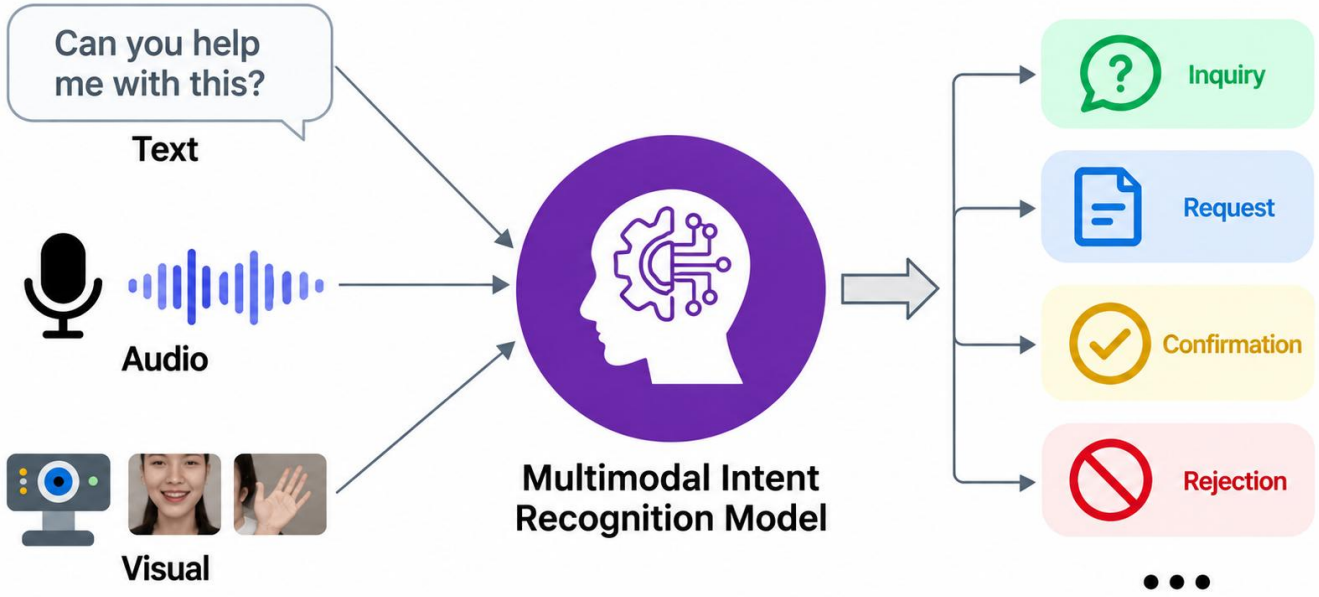

ACM MM 2025

**核心技术：结合重构误差与跨模态信息关系评估样本难度，按难度动态调整辅助样本检索规模，并通过检索增强课程学习协同优化特征恢复与情感分类。**


**Liu R**, Zuo H, Lian Z, et al. Hardness-aware dynamic curriculum learning for robust multimodal emotion recognition with missing modalities[C]//**Proceedings of the 33rd ACM International Conference on Multimedia. 2025**: 5755-5764.

# 多模态意图识别的任务定义

**综合语言内容、声学表达与视觉行为，识别说话人在当前交互中的实际目的。**

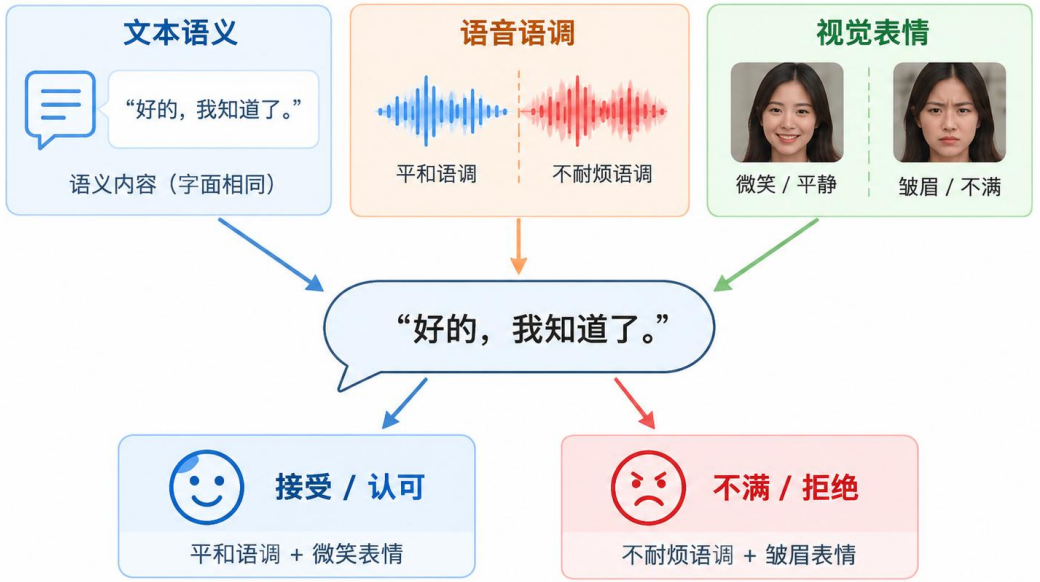



- **文本模态：** 提供话语语义，反映基本交互目的。
- **声学模态：** 利用语调、重音与节奏，辅助意图判断。
- **视觉模态：** 通过表情、视线与手势，补充非语言线索。
- **输出结果：** 识别交互意图类别，支持后续交互决策。

# 相同的话语，为什么会表达不同的交互意图?

**相同的语言内容在不同表达方式下，可能对应不同的交互意图。**

- 字面内容难以完整表达说话人的真实意图
- 语调、情绪变化和面部表情提供重要的补充证据
- 不同模态之间存在复杂的语义互补与关联关系



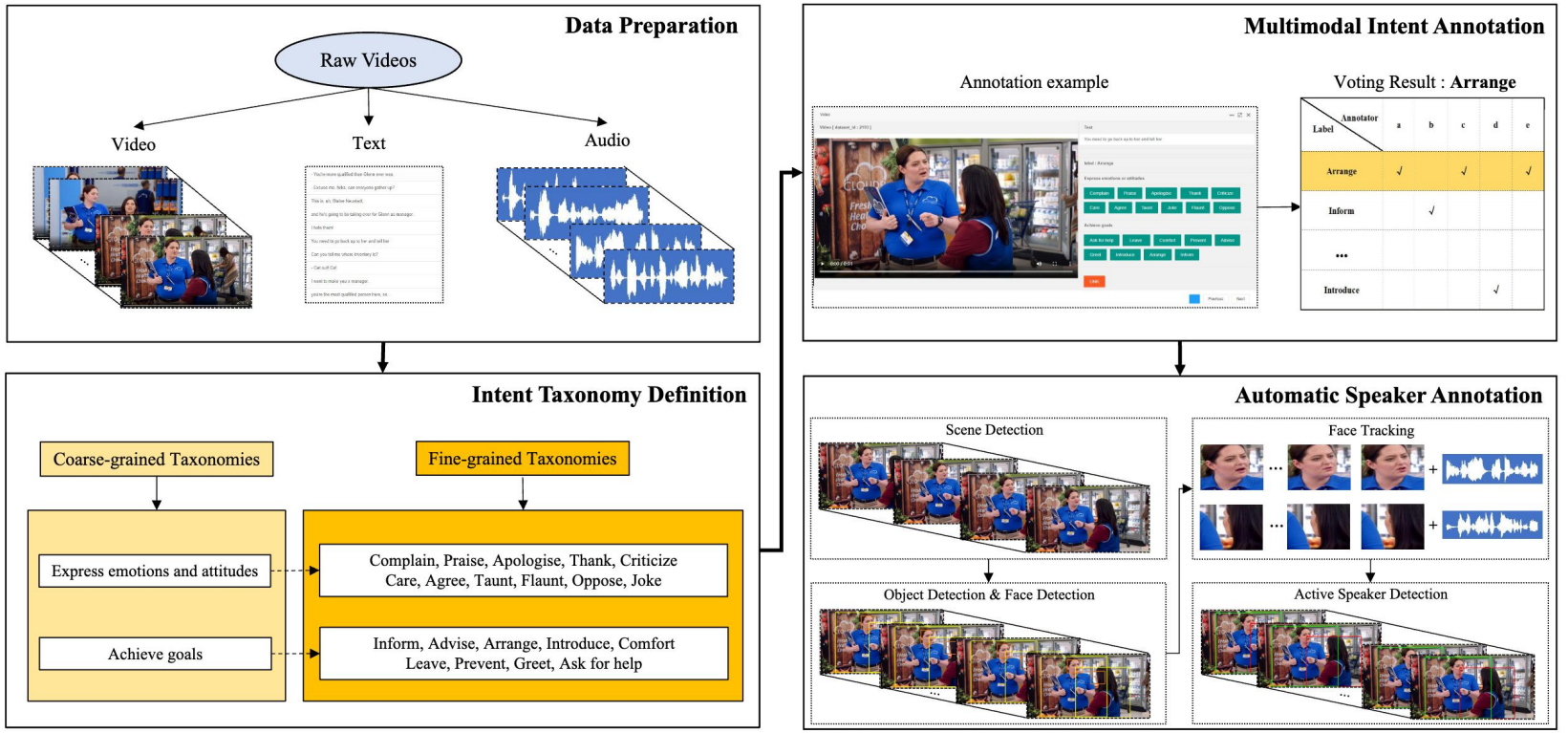
# 多模态意图识别的关键挑战

**多模态信息提供了更丰富的交互线索，但也带来了互补信息提取、跨模态关联及意图泛化等挑战。**

- 如何挖掘文本之外的声学与视觉线索?
- 如何融合不同模态的一致信息与互补信息?
- 如何结合交互情境区分相似或隐含意图?

# MIntRec：面向真实交互的多模态意图识别基准

**目标：构建融合文本、语音与视觉的意图识别基准，评估非语言线索对交互意图理解的作用。**

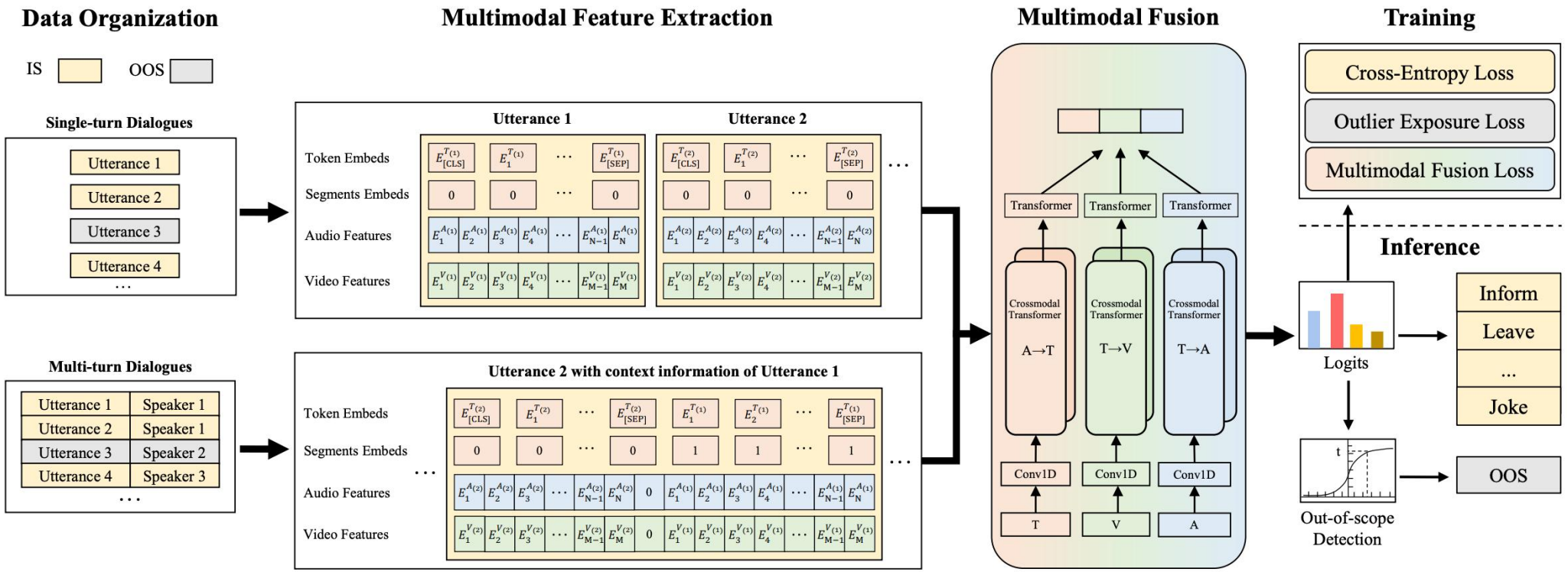



**核心技术：通过层次化分类体系定义20种意图，定位当前说话人的视觉信息，并利用多种多模态融合模型开展基准评估。**

Zhang H, Xu H, Wang X, et al. Mintrec: A new dataset for multimodal intent recognition[C]//**Proceedings of the 30th ACM international conference on multimedia. 2022**: 1688-1697.

# MIntRec2.0：面向多方对话意图检测的多模态基准

**目标：在多轮、多说话人交互中，结合对话历史识别当前意图，并检测预定义类别之外的交互意图。**




**核心技术：构建包含30类意图及范围外样本的多模态对话数据，利用说话人感知上下文建模辅助意图判断，并联合评估意图分类与范围外检测能力。**

Zhang H, Wang X, Xu H, et al. Mintrec2. 0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations[C]//**International Conference on Learning Representations. 2024**.

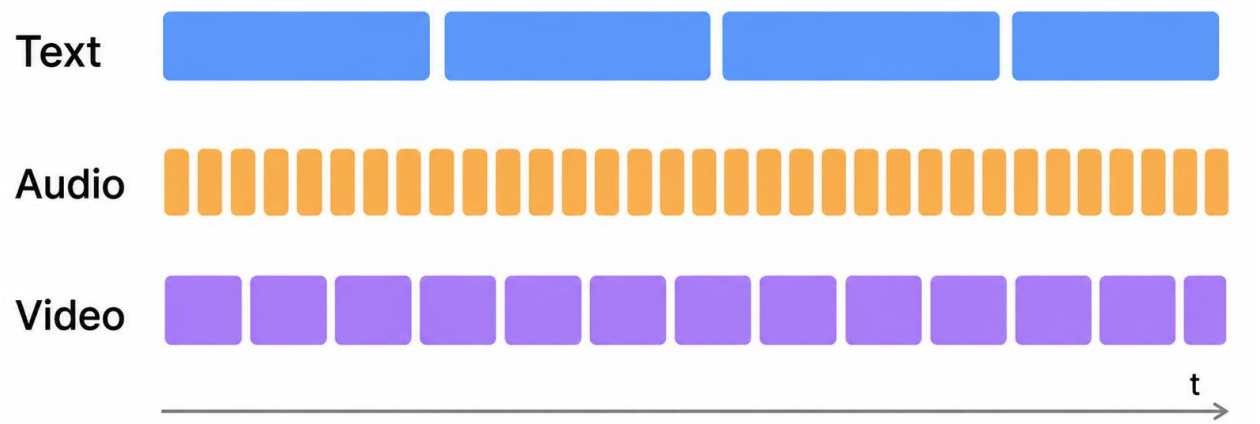
# 多模态信息如何有效融合?

**不同融合策略决定了模态间的信息交互方式，需要兼顾模态互补性与表征一致性。**

- 早期融合：　特征拼接与联合建模，实现直接信息整合。
- 后期融合：　独立建模各模态，融合预测结果。
- 中间融合：　利用跨模态注意力等机制，学习深层交互关系。

# 跨模态信息如何建立时间与语义对应?

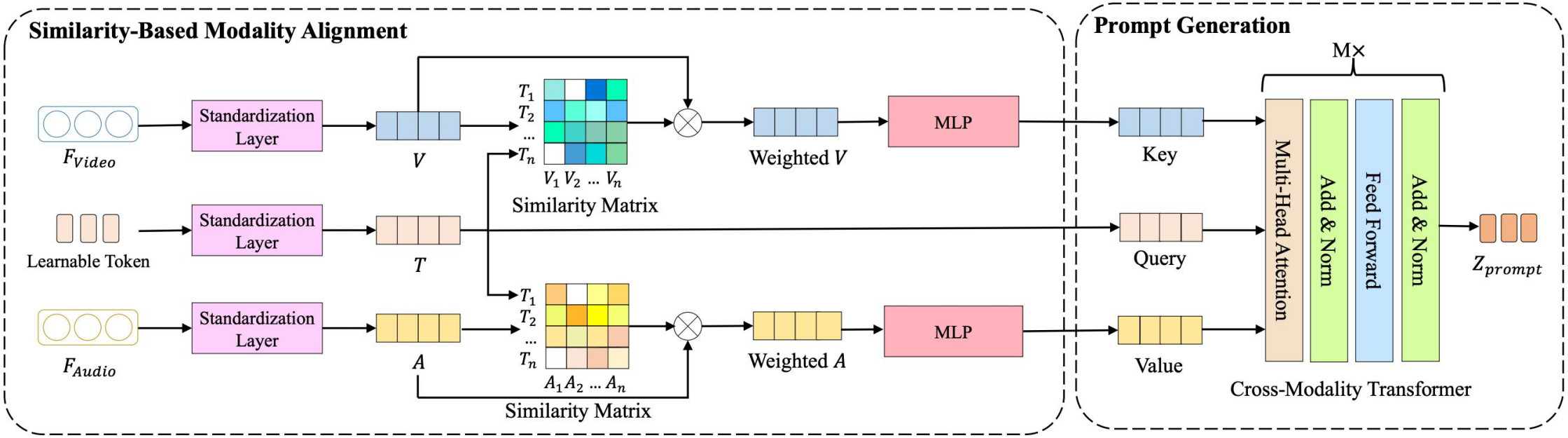
**文本、语音与视觉具有不同的时间粒度和表达方式，跨模态建模需要建立有效的时间与语义关联。**




- 时间粒度差异：　词元、声学帧与视频帧具有不同的采样尺度。
- 语义对应差异：　语调、表情和动作不一定与词元一一对应。
- 显式对齐：　利用时间戳、分段或插值建立时间对应。
- 隐式对齐：　通过跨模态注意力学习动态关联。

# TCL-MAP：提示引导的跨模态语义对齐

**目标：在多模态意图识别中，有效利用语音与视觉的补充线索，增强意图语义表示的判别能力。**

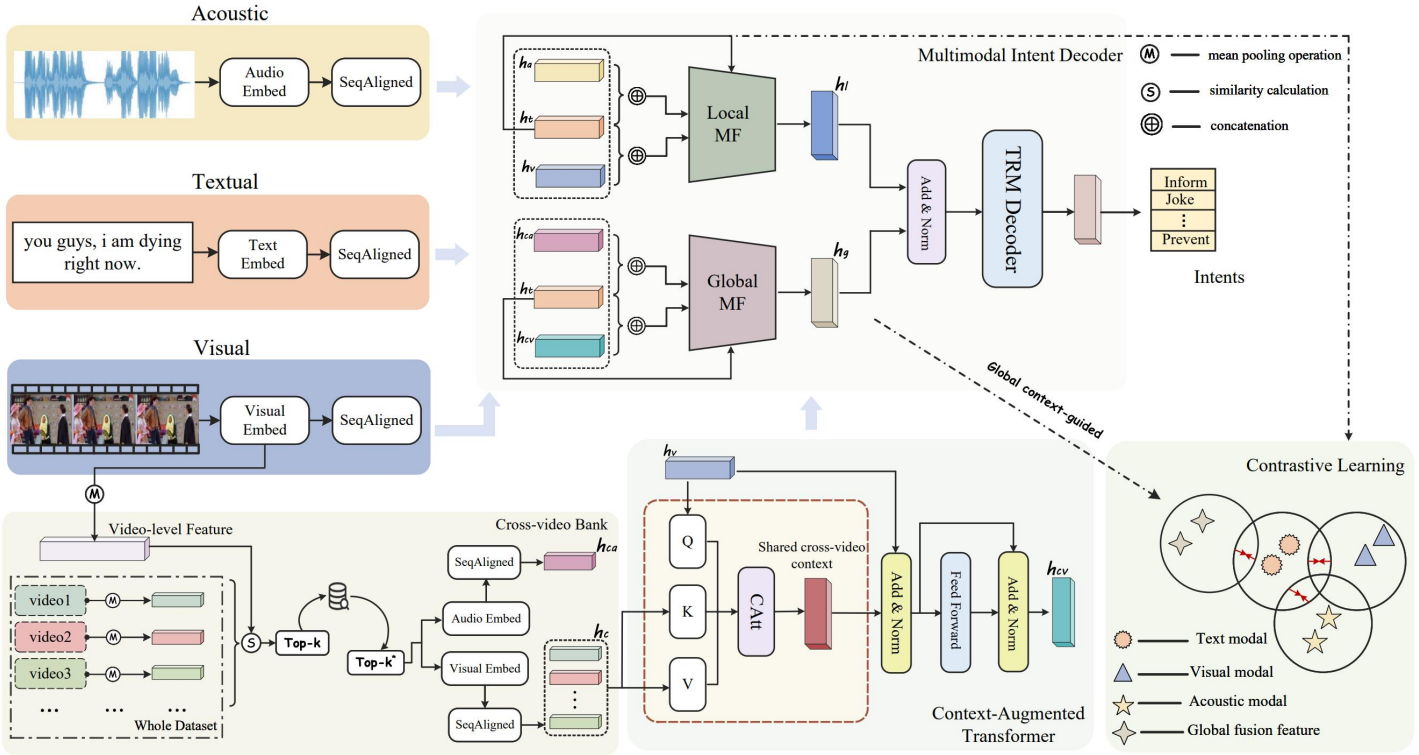



**核心技术：通过相似性引导的模态对齐缓解表示差异，利用跨模态注意力生成模态感知提示，并通过标签引导的Token级对比学习区分不同意图。**

Zhou Q, Xu H, Li H, et al. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition[C]//**Proceedings of the AAAI conference on artificial intelligence. 2024**, 38(15): 17114-17122.

# CAGC：跨视频上下文增强与局部-全局对比学习

**目标：在当前片段意图线索不足的情况下，利用相似视频的语义信息辅助理解，缓解模态差异与意图歧义。**

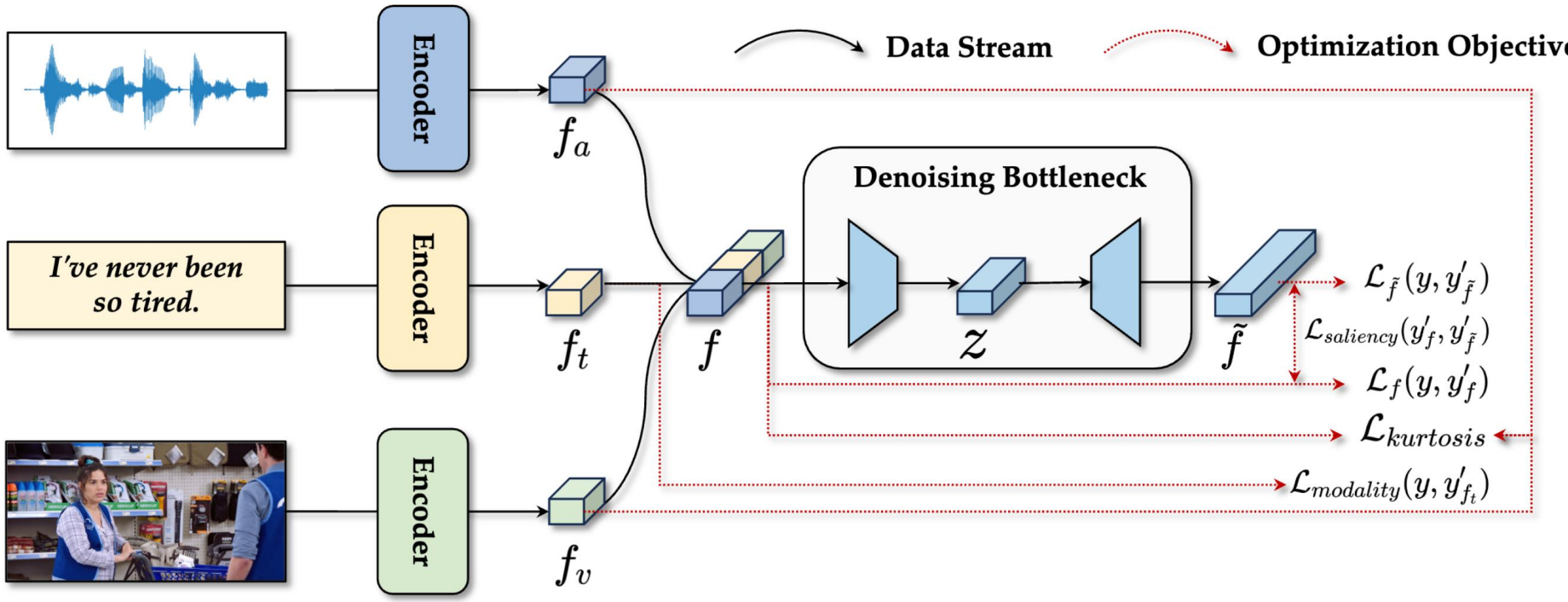



**核心技术：通过意图感知的跨视频检索获取相关信息，构建上下文增强的全局表示，并利用局部—全局对比学习强化跨模态语义关联。**

Sun K, Xie Z, Ye M, et al. Contextual augmented global contrast for multimodal intent recognition[C]//**2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. IEEE, 2024: 26953-26963.

# InMu-Net：信息瓶颈去噪与长尾意图识别

**目标：在多模态意图识别中，减少无关信息干扰，保留关键意图线索，并提升长尾类别的识别能力。**




**核心技术：通过信息瓶颈压缩冗余特征，利用意图相关信息保留机制避免有效线索丢失，并通过峰度正则化改善类别表示分布。**

Zhu Z, Cheng X, Chen Z, et al. InMu-Net: advancing multi-modal intent detection via information bottleneck and multi-sensory processing[C]//**Proceedings of the 32nd ACM International Conference on Multimedia. 2024**: 515-524.

# 自适应融合：不同模态的可信度并不恒定

**不同模态在不同交互场景中的有效性存在差异，需要根据模态质量与意图相关性动态调整融合贡献。**




- 模态可靠性评估： 感知不同模态的信息质量与意图相关性。
- 动态权重分配： 通过注意力或门控机制，自适应调整模态贡献。
- 鲁棒互补融合： 抑制低质量模态干扰，保留有效的互补信息。

# 自适应融合：全局与局部信息如何协同

## 全局语义与局部信息缺乏有效的协同建模

- 全局上下文：话题、交互对象、前文关系与整体情绪
- 局部证据：短暂语音片段、关键词和面部关键区域

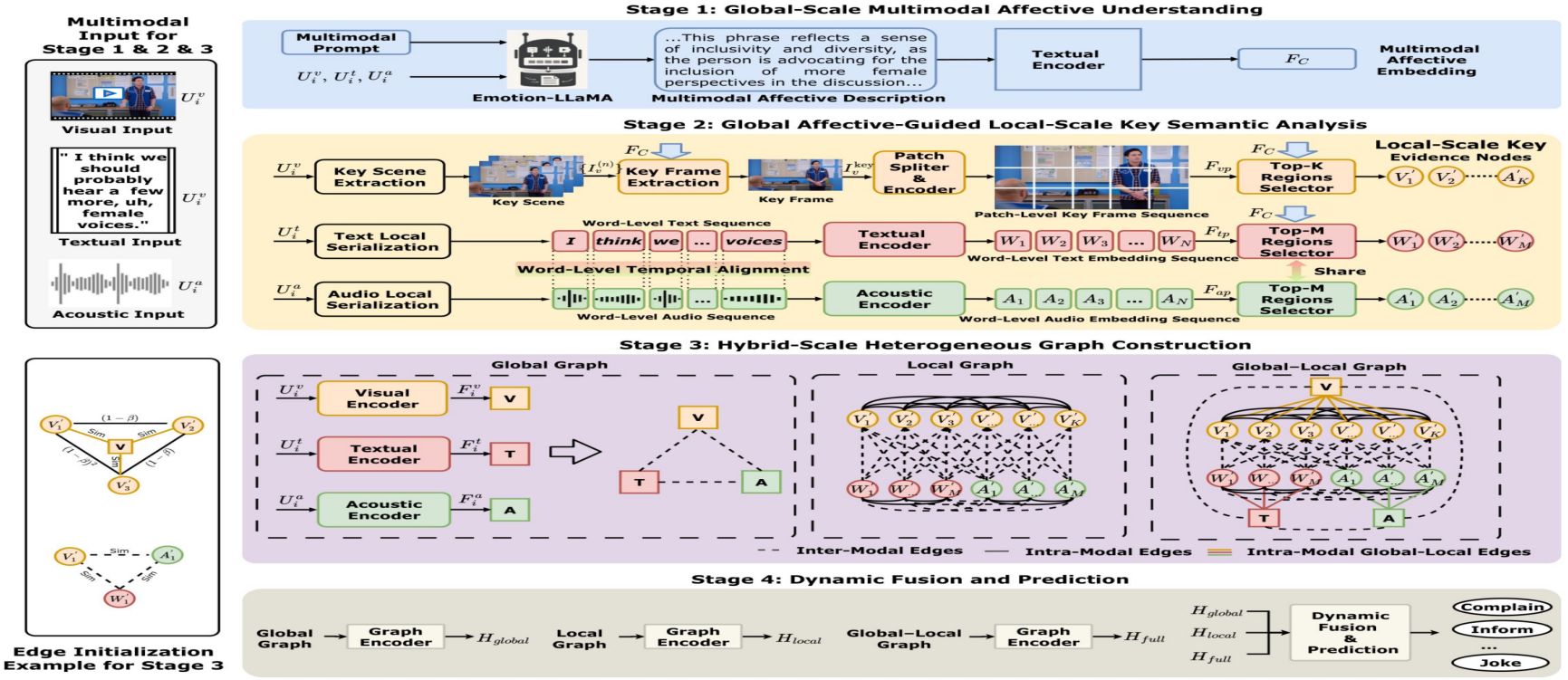



- 现存问题： 全局建模容易忽略细粒度线索，局部建模又缺乏全局语义指导与跨尺度关联

# H²GF-Net：全局语义指导的混合尺度图融合

**目标：在多模态意图识别中，建立全局语义与局部关键线索的有效关联，充分利用不同模态与粒度的互补信息。**

**团队工作**



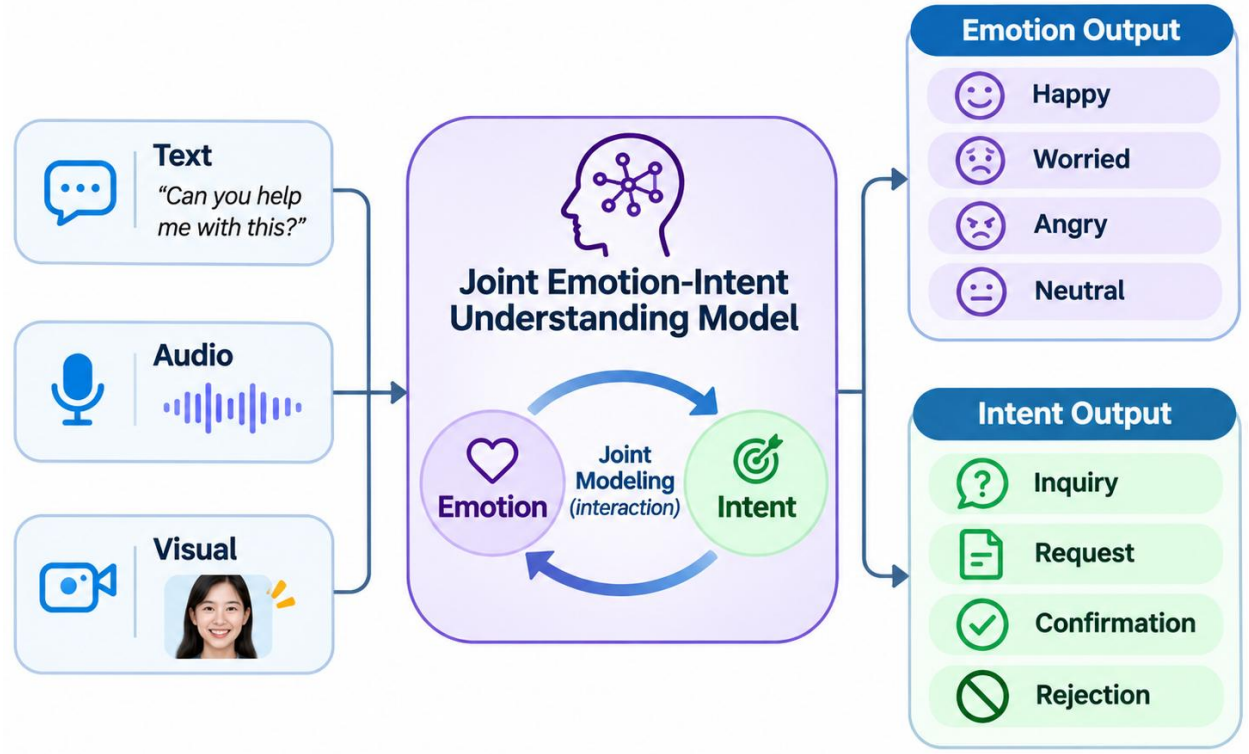

**核心技术：通过全局情感语义引导局部线索选择，构建涵盖模态内、跨模态及局部—全局关系的异构图，并动态融合多尺度图表示完成意图识别。**

Liu R, Ma Y, Li H. Hybrid-Scale Heterogeneous Graph Fusion for Multimodal Intent Recognition[J]. **IEEE Transactions on Affective Computing, 2026**.

# 从单独识别到情感-意图联合识别

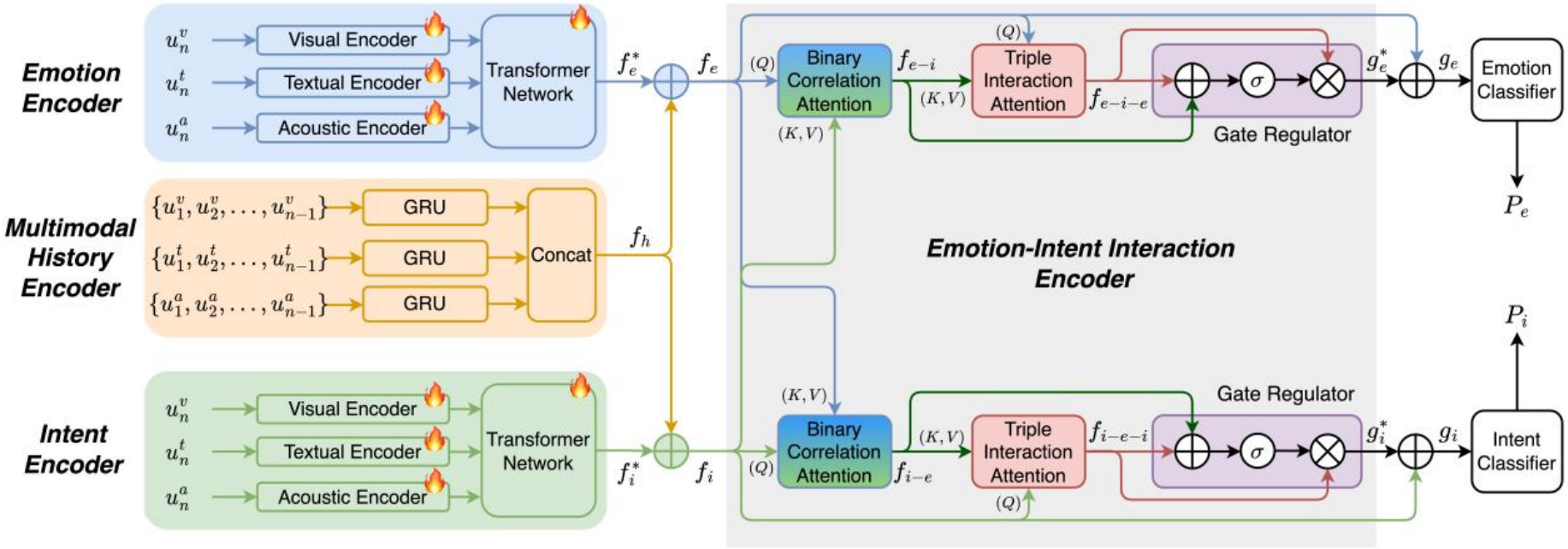
**情感反映说话人的情绪状态，意图体现其交互目的，二者相互关联、相互影响。**

- **情感-意图关联**：二者相互影响、相互补充。
- **联合理解**：建模对话上下文与情感-意图交互关系，实现联合预测。



# MC-EIU：多模态对话情感-意图联合理解

**目标：构建中英双语多模态情感-意图联合理解基准，结合对话历史与多模态信息，实现当前话语情感和意图的联合识别。** Data Intelligence 2026




团队工作

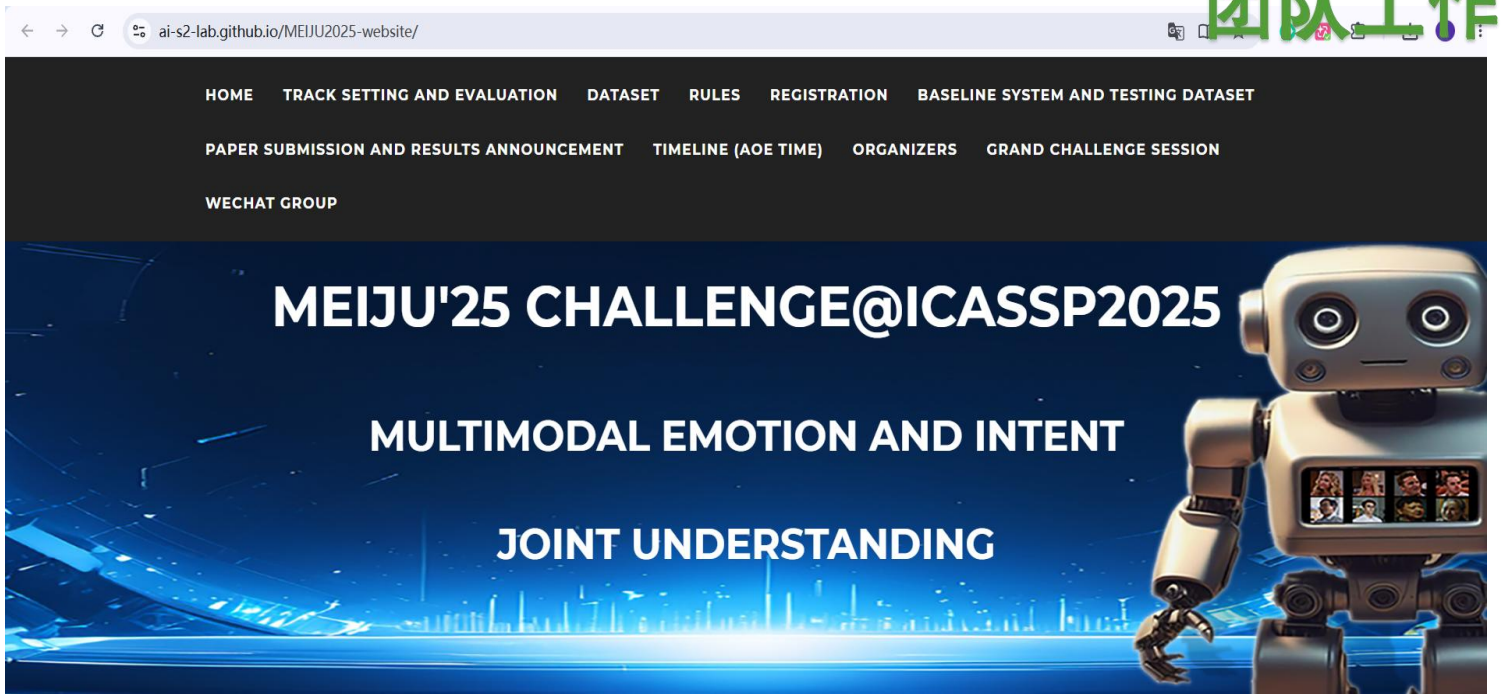
**核心技术： 构建包含7类情感、9类意图的中英双语多模态数据集；提出EI²网络，通过多模态历史编码、关联注意力、交互注意力及门控调节，建模情感与意图的深层交互关系，实现联合预测。**

Liu R, Zuo H, Lian Z, et al. Emotion and intent joint understanding in multimodal conversation: A benchmarking dataset[J]. **Data Intelligence 2026**.

# MEIJU’25：多模态情感-意图联合理解挑战赛

**目标：　面向多模态情感-意图联合识别中的标注数据不足与类别分布不均衡问题，推动复杂条件下的联合理解研究。**

团队工作




- **低资源联合识别（Track 1）：**
  利用少量标注数据与大量未标注数据，提升情感和意图的联合识别能力。

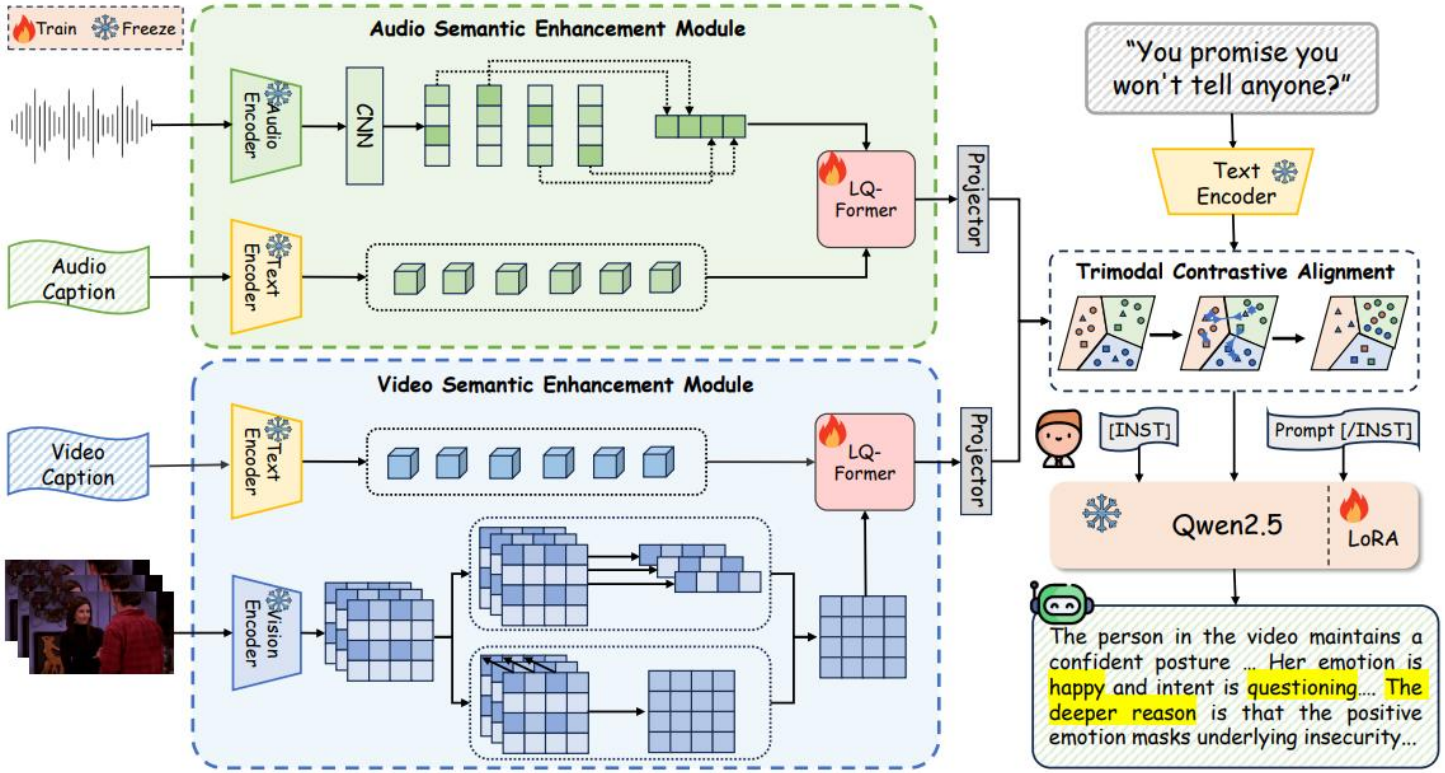
- **类别不均衡联合识别（Track 2）：**
  针对情感与意图类别分布不均的问题，提升少数类别的识别性能。

**评测方式：**　两个赛道均设置中文和英文子赛道，共4个子赛道，采用联合识别平衡指标（JRBM）综合评价情感与意图识别性能。

https://ai-s2-lab.github.io/MEIJU2025-website/

# JX4MEI：从情感-意图联合分类到联合解释

**目标：在情感与意图联合识别的基础上，进一步解释二者的形成原因及内在关联，提升多模态对话理解的可解释性。**




**核心技术： 构建包含情感、意图及其关联解释的XMEI-dataset，提出XMEI-Qwen模型，利用LQ-Former引入模态语义描述，引导音视频特征对齐，实现情感-意图联合分类与解释生成。**


Huang Y, Yang X, Feng S, et al. JX4MEI: Multimodal Semantically-Enhanced LLM for Joint Multimodal Emotion-Intent Explanation and Classification[C]//**Findings of the Association for Computational Linguistics: ACL 2026**. 2026: 20242-20261.

# 研究总结｜模态缺失下的情感-意图理解

01

**缺失模态鲁棒识别**

- 通过模态重构、知识蒸馏与特征增强，缓解模态缺失造成的情感信息损失。
- CIF-MMIN（TAFFC 2024）、HARDY-MER（ACM MM 2025）...

02

**多模态意图识别**

- 结合跨模态语义关联与上下文建模，提升意图识别能力。
- TCL-MAP（AAAI 2024）、H²GF-Net（TAFFC 2026）...

03

**情感-意图联合建模**

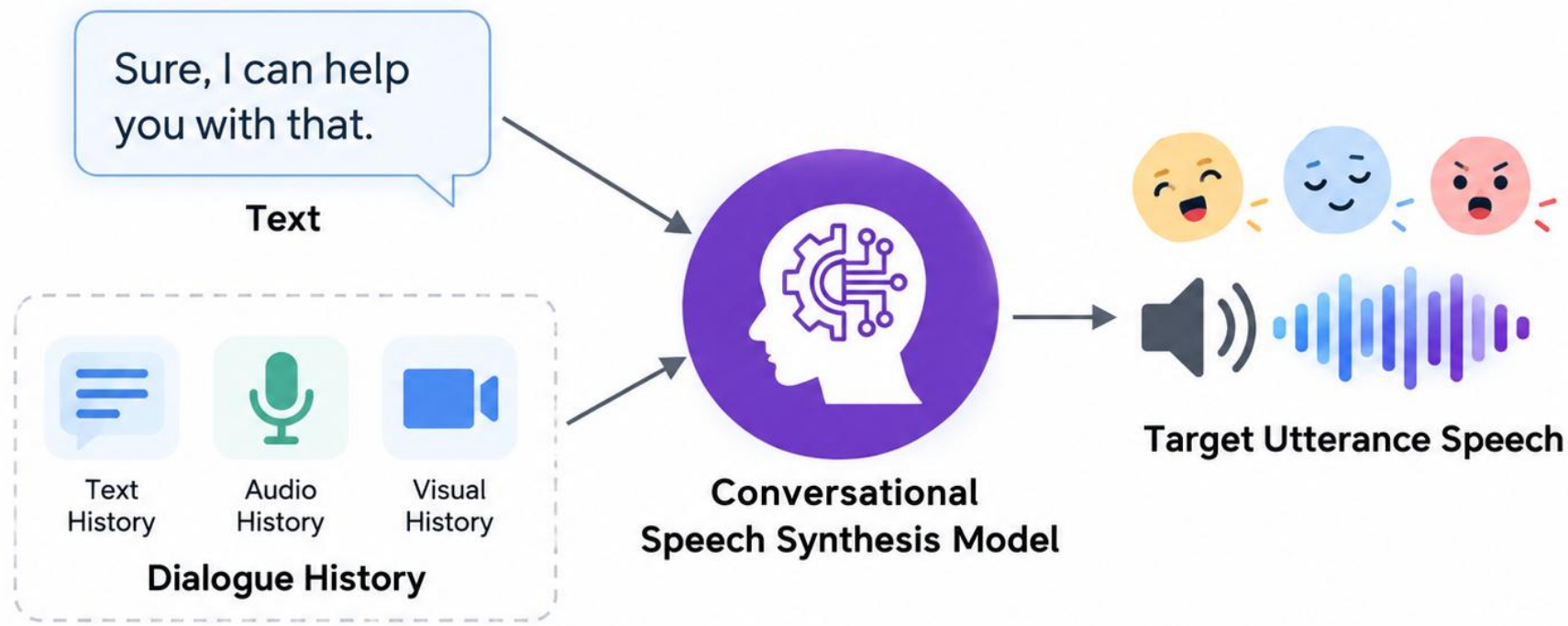
- 建模多模态对话上下文及情感与意图之间的关联，实现双任务联合预测。
- MEIJU'25（ICASSP 2025）、JX4MEI（ACL 2026）...

# 第三部分

# 对话交互中多模态协同表达

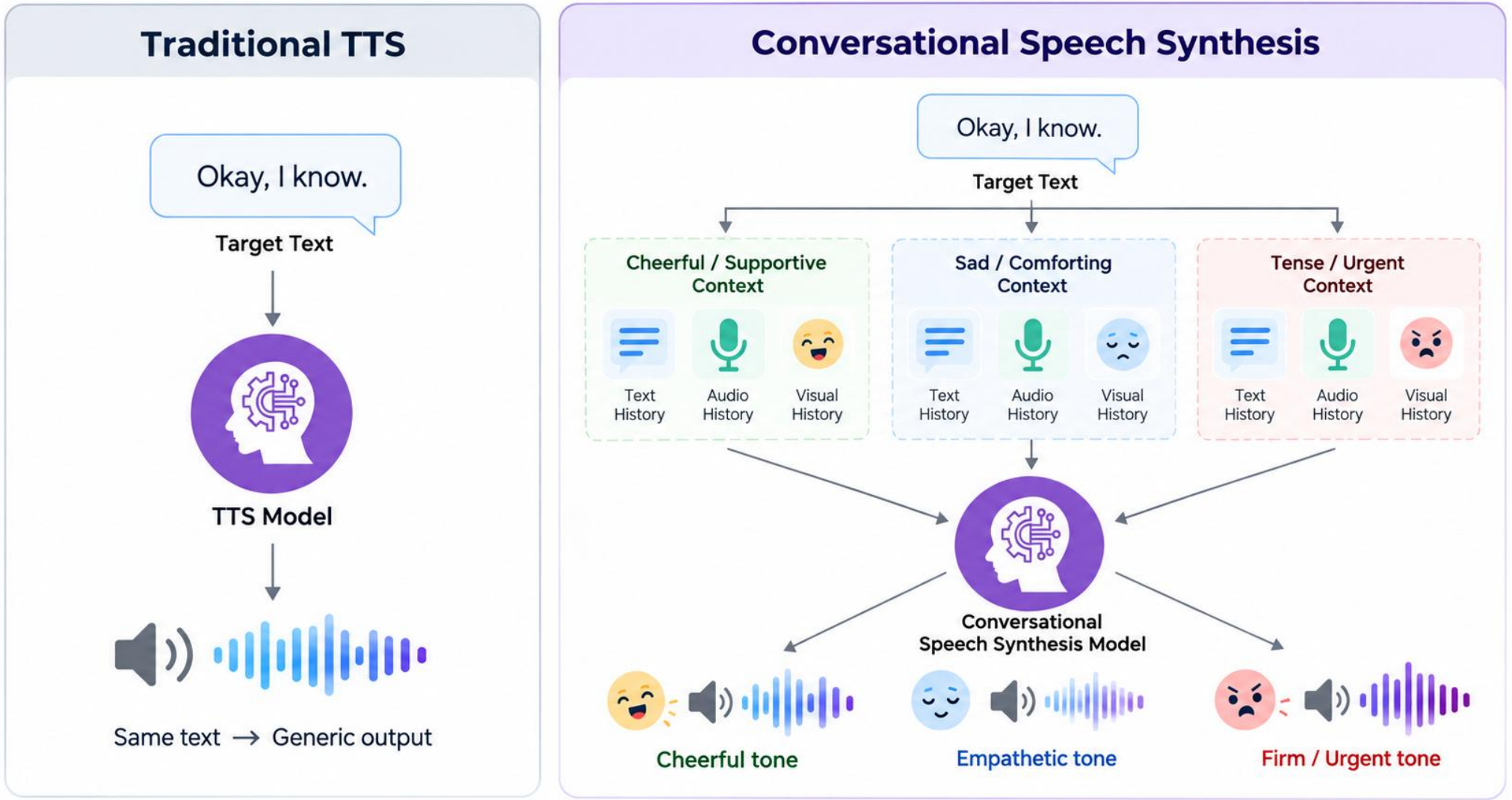
# 对话语音合成的任务定义

- 输入：目标回复文本、多模态对话历史
- 输出：目标句语音



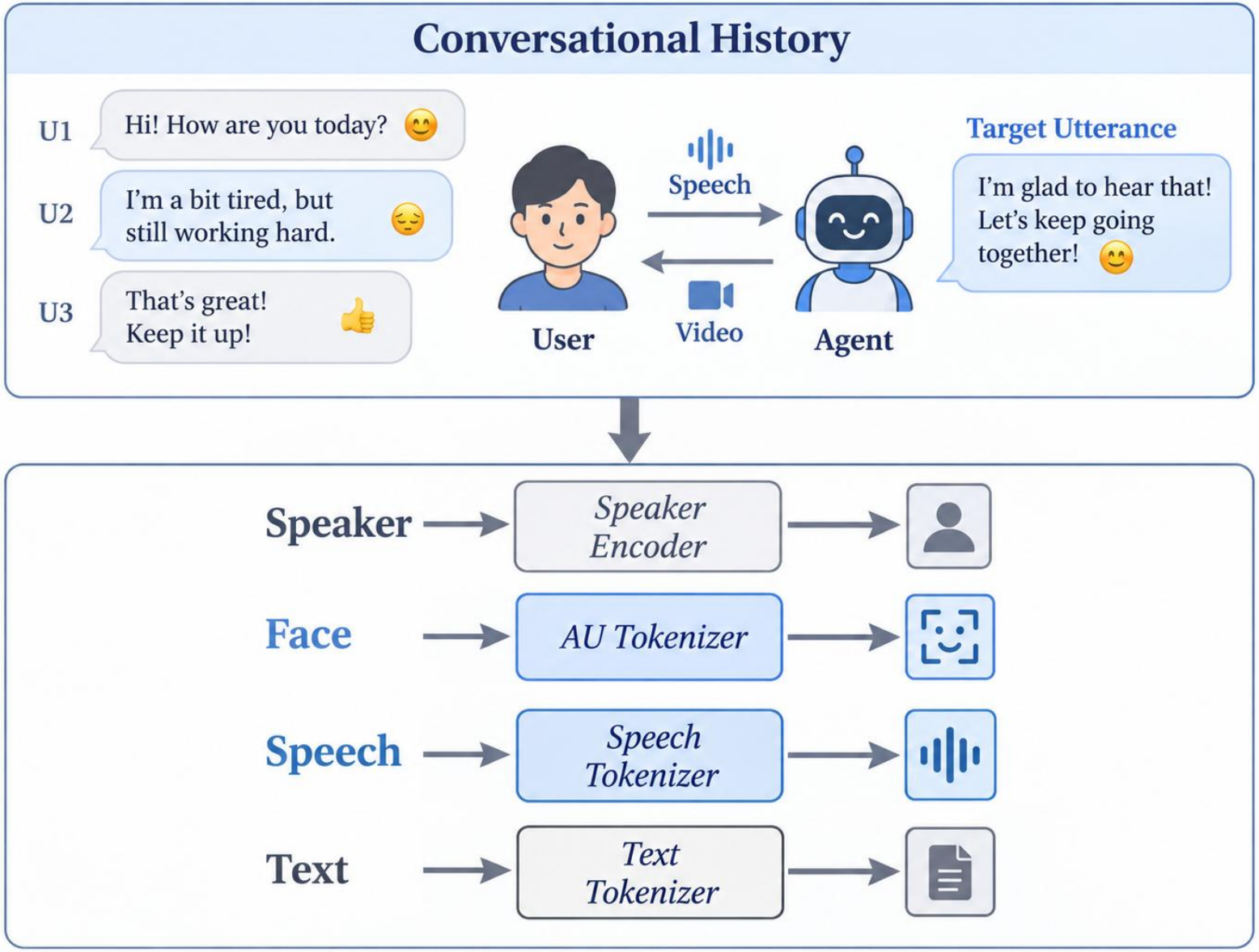
# 独立TTS为什么不能直接替代CSS?

## 相同目标话语在不同对话语境中的表达可能完全不同



# 对话历史中需要提取哪些信息？

- 文本语义
- 语音表达
- 说话人信息
- 视觉线索（可选）



# 对话语音数据集

- DailyTalk[1]： 具有自然话轮结构的双人对话语音数据集。
- MultiDialog[2]： 面向面对面交互的音视频多模态对话数据集。
- NCSSD[3]（Ours）： 融合自然录制与影视对话的中英双语情感对话数据集。
- CSS-EmCap[4]（Ours）： 带有语义-共情描述的对话数据集。
- VSDD-1K[5]（Ours）： 面向真实多轮交互的大规模视听对话数据集

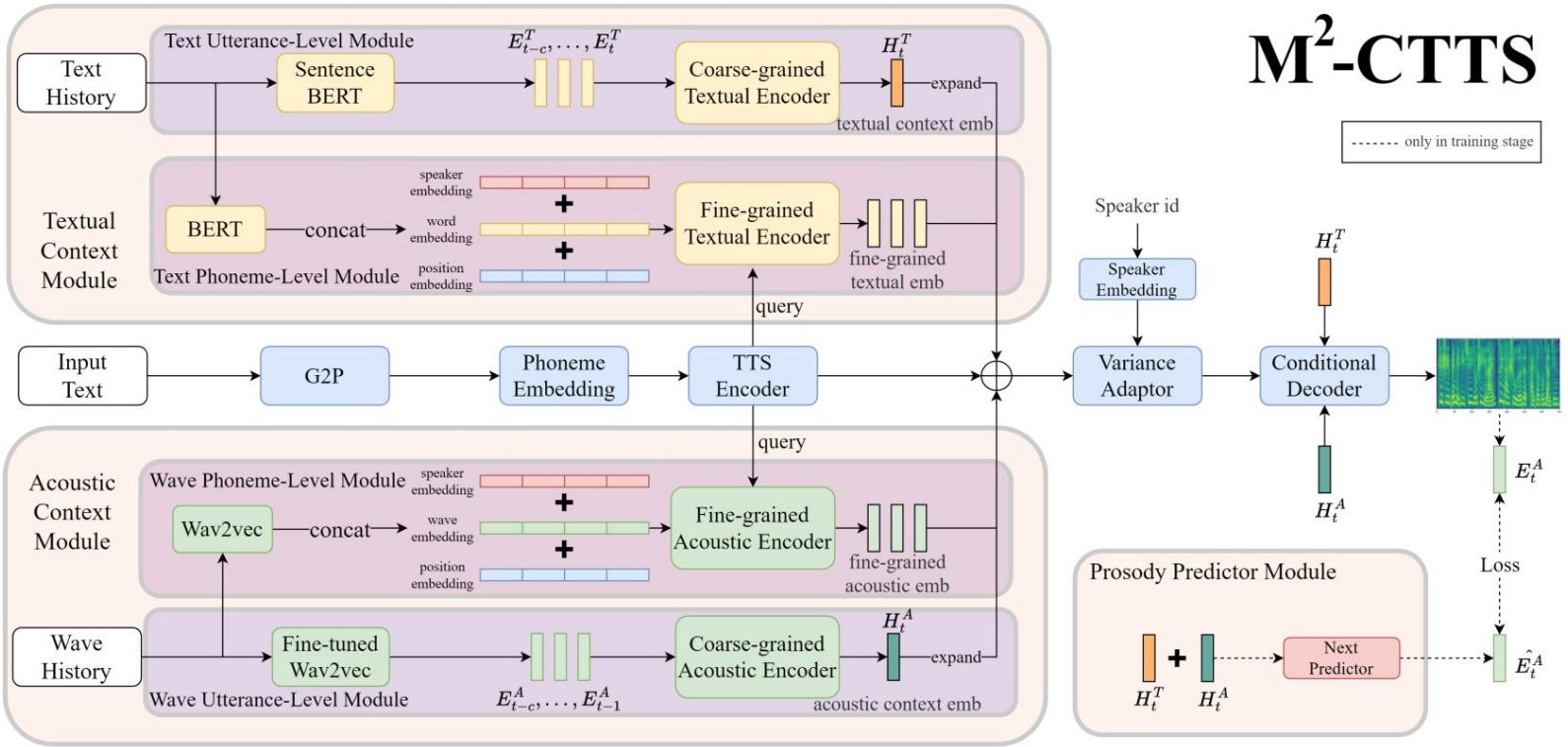

[1] Lee K, Park K, Kim D. Dailytalk: Spoken dialogue dataset for conversational text-to-speech[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.
[2] Park S, Kim C, Rha H, et al. Let's go real talk: Spoken dialogue model for face-to-face conversation[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 16334-16348.
[3] Liu R, Hu Y, Ren Y, et al. Generative expressive conversational speech synthesis[C]//Proceedings of the 32nd ACM International Conference on Multimedia. 2024: 4187-4196.
[4] Hu Y, Liu R, Ren Y, et al. Chain-talker: Chain understanding and rendering for empathetic conversational speech synthesis[C]//Findings of the Association for Computational Linguistics: ACL 2025. 2025: 1988-2003.
[5] Hu Y, He S, Liu R, et al. Let Me Look at You: Advanced Facial Expression Modeling for Conversational Speech Synthesis[J]. ACMMM2026

# 对话语音合成的发展框架

- **阶段一：** 时序历史编码和注意力提取
- **阶段二：** 关系图与多尺度上下文建模
- **阶段三：** 检索增强与自回归上下文建模
- **阶段四：** 真实情感理解及音视联合生成

# M²-CTTS：多尺度多模态对话语音合成

**目标：在对话语音合成中，兼顾整体语境与局部表达变化，生成韵律自然、符合上下文的目标语音。**

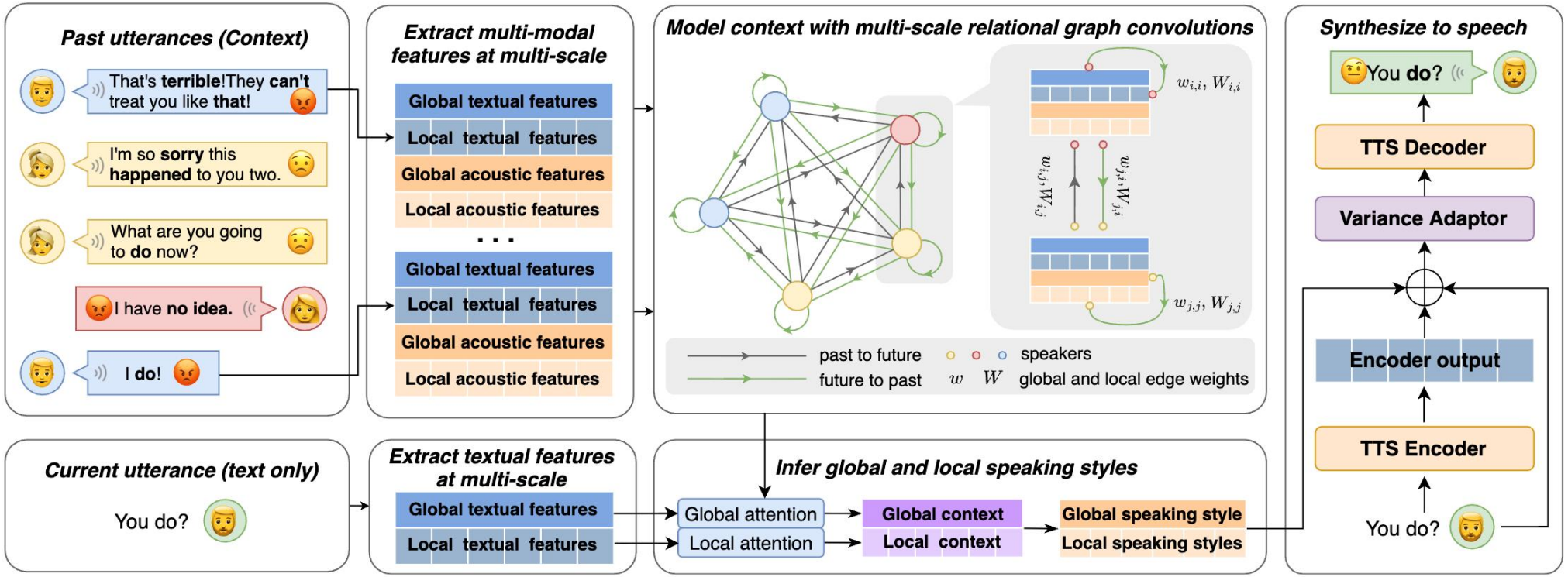



**核心技术：通过文本与声学双模态建模提取句子级和音素级上下文，利用跨注意力关联目标话语与历史信息，并结合韵律预测指导语音生成。**

Xue J, Deng Y, Wang F, et al. M 2-ctts: End-to-end multi-scale multi-modal conversational text-to-speech synthesis[C]//**ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. IEEE, 2023: 1-5.

# MSRGCN：多尺度关系图驱动的说话风格推断

**目标：在多轮对话中，结合不同说话人的历史表达与话语依赖，推断符合语境的整体风格和局部表达。**



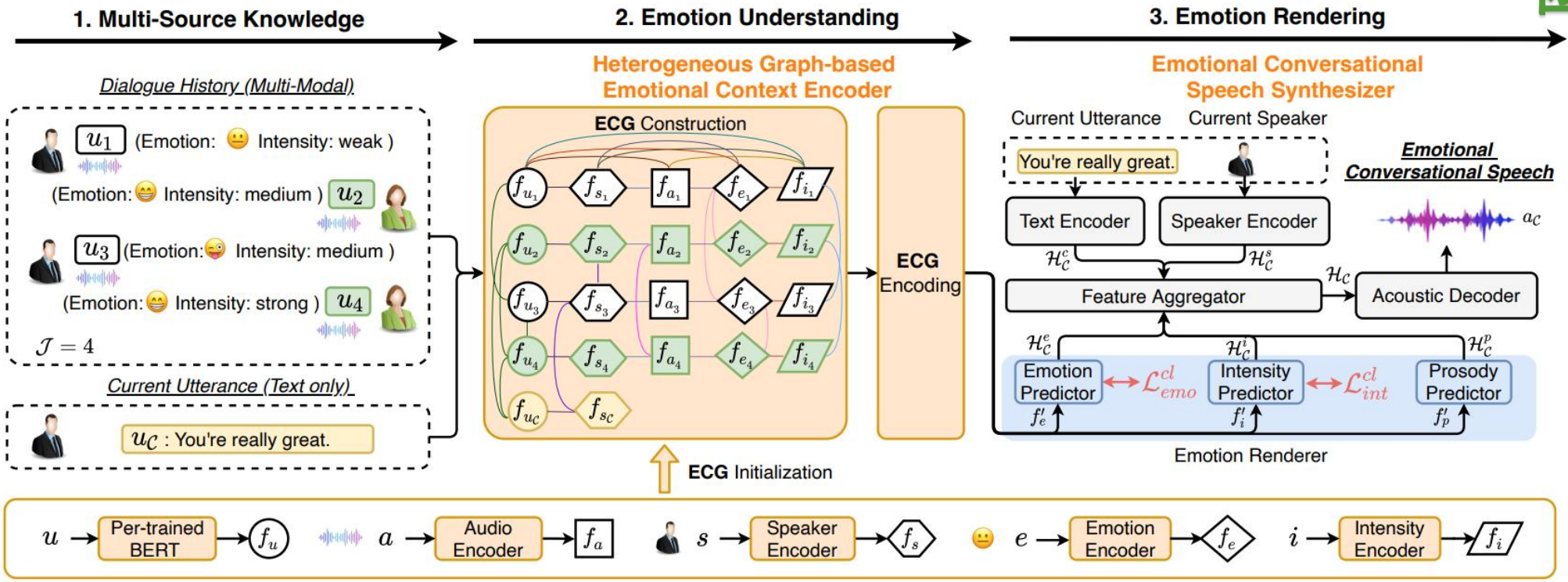

**核心技术：提取文本与声学的多尺度上下文特征，通过关系图卷积建模说话人及跨话语词级依赖，并利用全局与局部注意力推断说话风格，指导语音合成。**

Li J, Meng Y, Wu X, et al. Inferring speaking styles from multi-modal conversational context by multi-scale relational graph convolutional networks[C]//**Proceedings of the 30th ACM International Conference on Multimedia. 2022**: 5811-5820.

# ECSS：异构图驱动的情感上下文理解与渲染

**目标：在对话语音合成中，理解历史对话的情感状态，生成情感与韵律符合当前语境的目标语音。**

团队工作



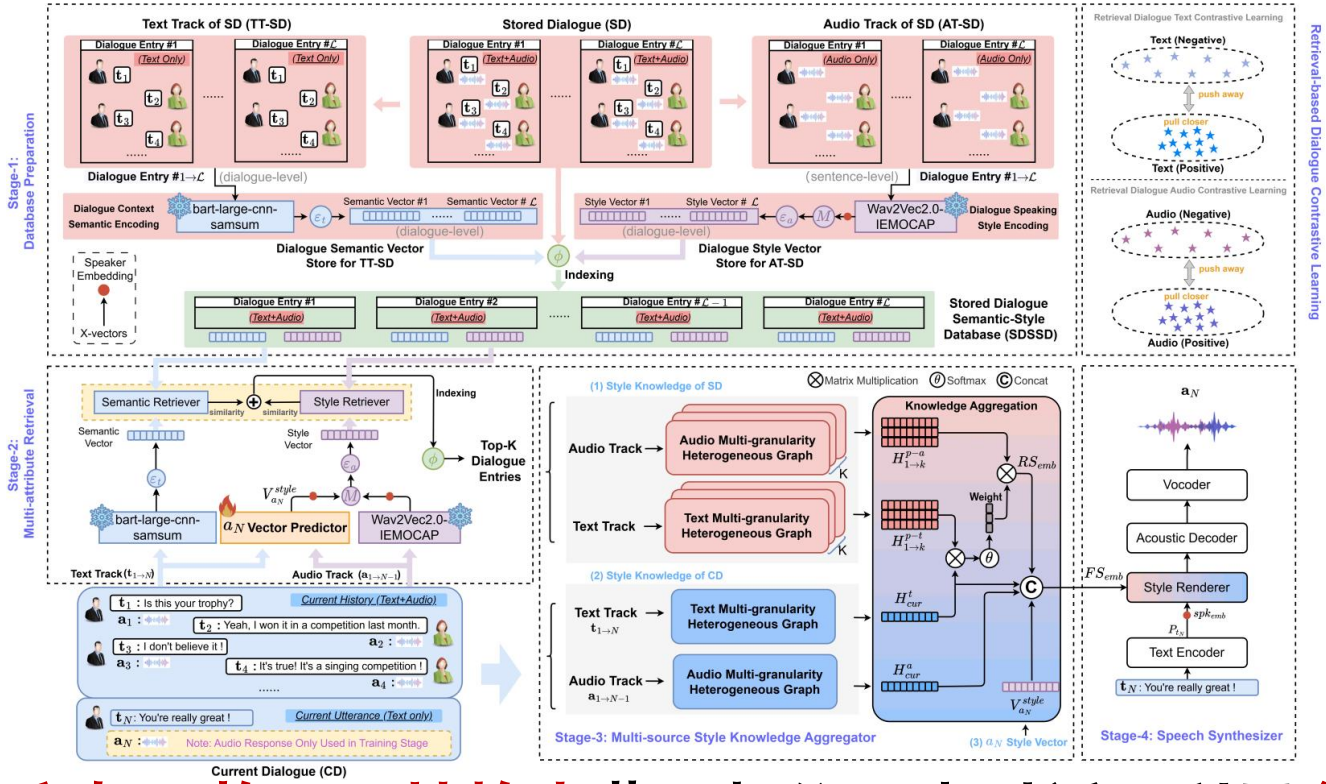

**核心技术：构建融合文本、语音、说话人及情感类别与强度的异构图，建模跨话轮情感依赖，并通过监督对比学习驱动的情感渲染器生成情感与韵律条件。**

Liu R, Hu Y, Ren Y, et al. Emotion rendering for conversational speech synthesis with heterogeneous graph-based context modeling[C]//**Proceedings of the AAAI Conference on Artificial Intelligence. 2024**, 38(17): 18698-18706.

# RADKA-CSS：检索增强的对话知识聚合

**目标：在当前对话的风格线索有限时，利用相似历史交互中的表达经验，生成符合语境的表现力语音。**

团队工作



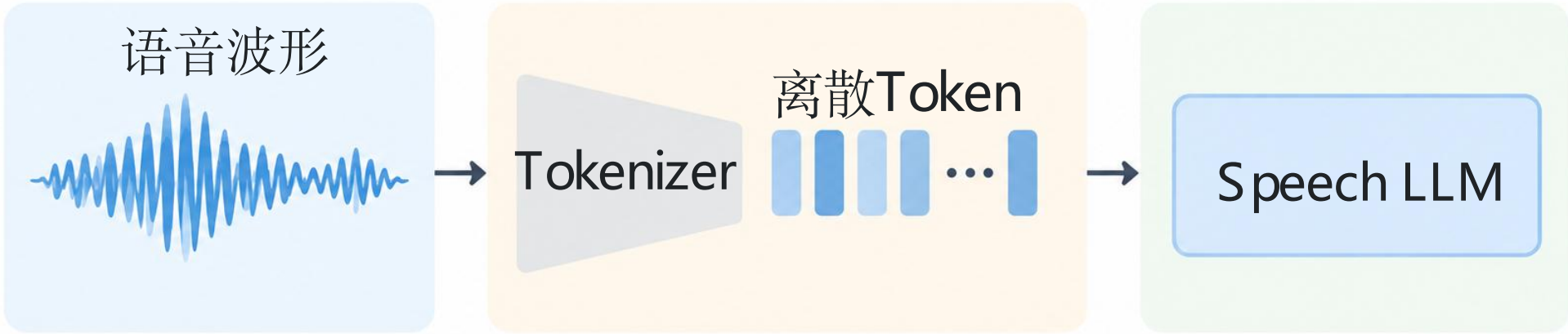

**核心技术：通过语义与风格双属性检索获取相似历史对话，利用多粒度异构图建模当前与检索对话的文本和声学关系，并聚合多源风格知识指导语音生成。**

Liu R, Jia Z, Bao F, et al. Retrieval-augmented dialogue knowledge aggregation for expressive conversational speech synthesis[J]. **Information Fusion, 2025**, 118: 102948.

# 离散语音Token与生成式建模

## 语音既可用连续声学特征，也可用离散单位建模

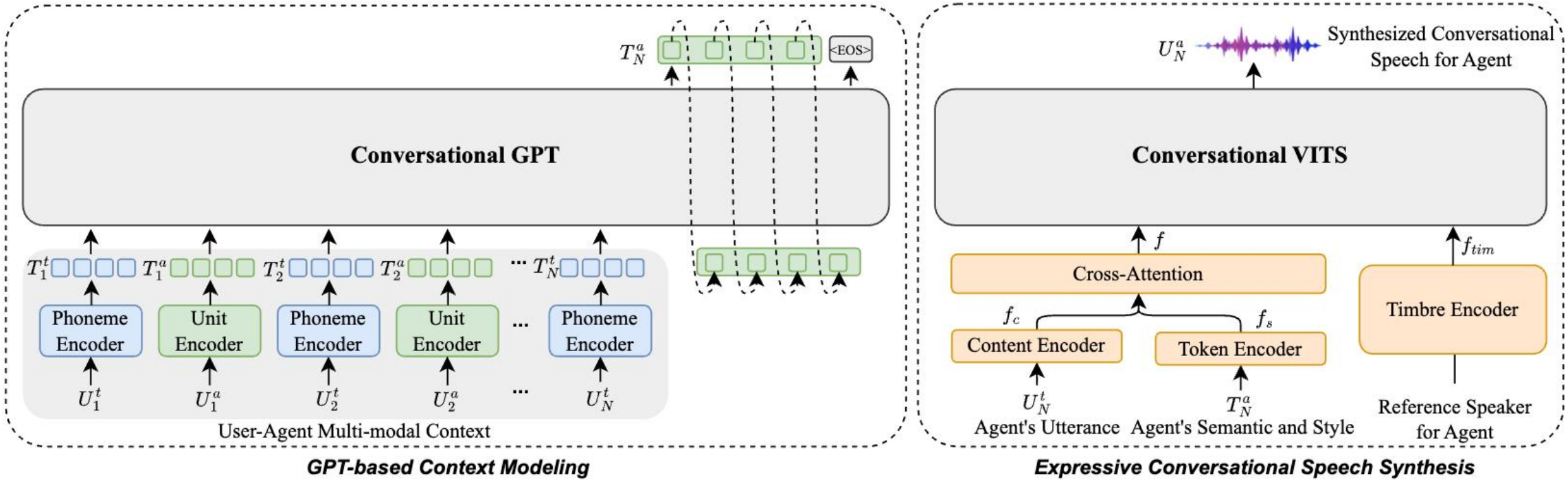
- Speech Tokenizer将语音压缩为离散单位，供序列模型预测
- 离散Token建模利于与LLM融合，但需权衡码率、延迟和表现力



# GPT-Talker：自回归序列建模的上下文感知语音合成

**目标：在多轮对话中，统一学习语义与表达依赖，生成内容准确、风格符合语境且音色一致的目标语音。**

**团队工作**

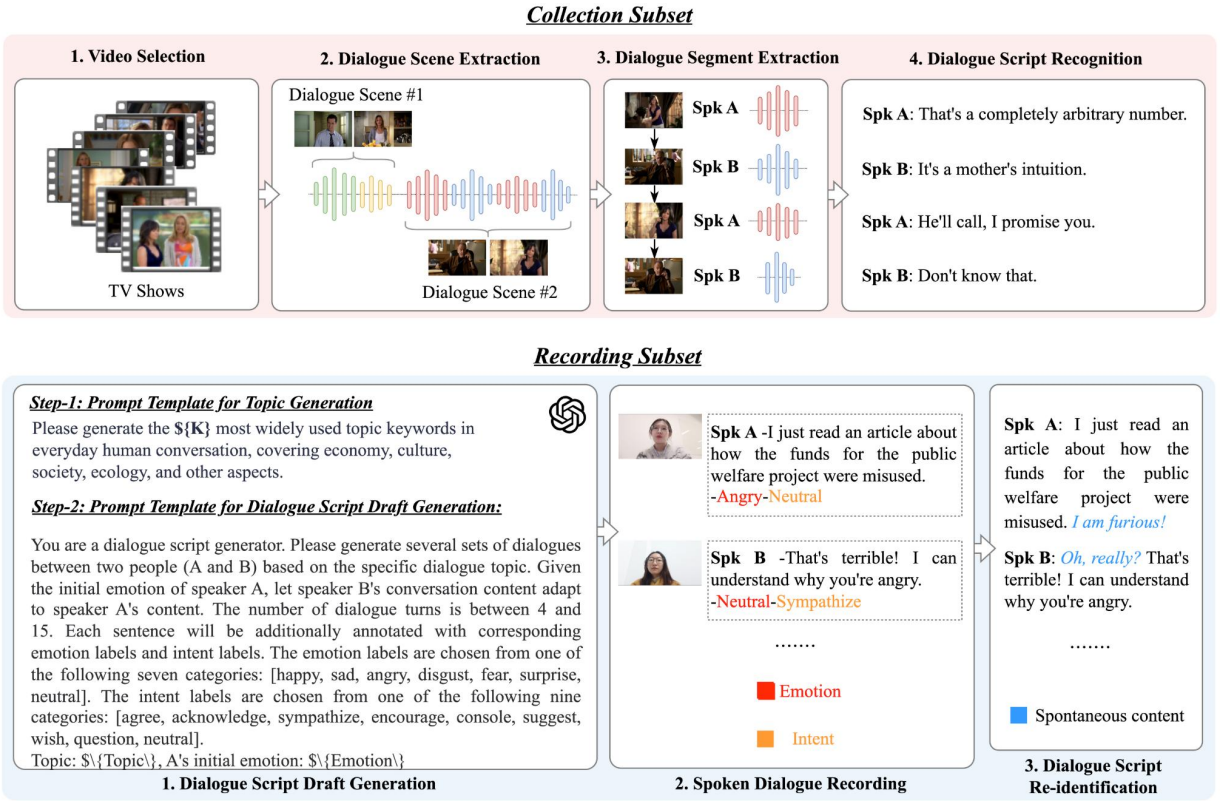



**核心技术：将历史文本与语音组织为多模态Token序列，通过ConGPT自回归预测目标语音的语义与风格表示，再结合目标文本与参考音色，由ConVITS生成对话语音。**

Liu R, Hu Y, Ren Y, et al. Generative expressive conversational speech synthesis[C]//**Proceedings of the 32nd ACM International Conference on Multimedia. 2024**: 4187-4196.

# GPT-Talker：自回归序列建模的上下文感知语音合成

## NCSSD：大规模中英双语对话语音数据集



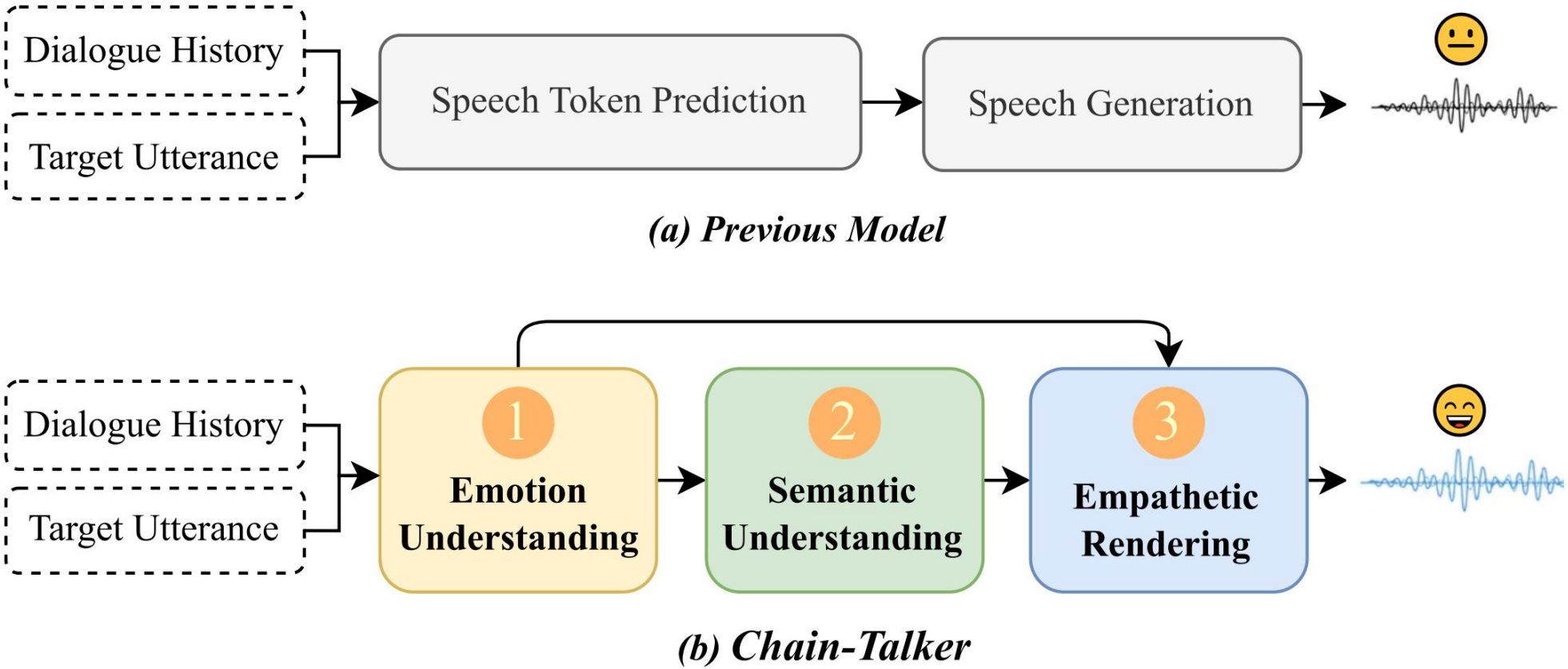

| Item | Collection | | Recording | |
|---|---|---|---|---|
| Language | *EN* | *ZH* | *EN* | *ZH* |
| Dialogues | 7,033 | 8,776 | 1,196 | 2,451 |
| Utterances | 62,603 | 99,126 | 10,033 | 21,688 |
| Words | 856,011 | 1,688,778 | 157,967 | 507,008 |
| Min words per utterance | 1 | 1 | 2 | 2 |
| Max words per utterance | 299 | 322 | 53 | 92 |
| Duration(h) | 72.94 | 115.22 | 19.10 | 29.57 |
| Max dialogue duration(s) | 177.51 | 308.21 | 108.52 | 75.82 |
| Min dialogue duration(s) | 4.74 | 5.54 | 20.06 | 15.73 |
| Mean utterance duration(s) | 4.19 | 4.18 | 6.85 | 4.90 |
| Max dialogue turns | 39 | 69 | 15 | 14 |
| Mean dialogue turns | 8.90 | 11.29 | 8.38 | 8.84 |
| Min dialogue turns | 4 | 4 | 5 | 6 |
| Speakers | > 339 | > 410 | 11 | 16 |

Liu R, Hu Y, Ren Y, et al. Generative expressive conversational speech synthesis[C]//**Proceedings of the 32nd ACM International Conference on Multimedia. 2024**: 4187-4196.

# 从生成式语音建模到链式同理心表达

## 仅依靠目标语音预测，难以刻画对话中的情感与语义理解过程

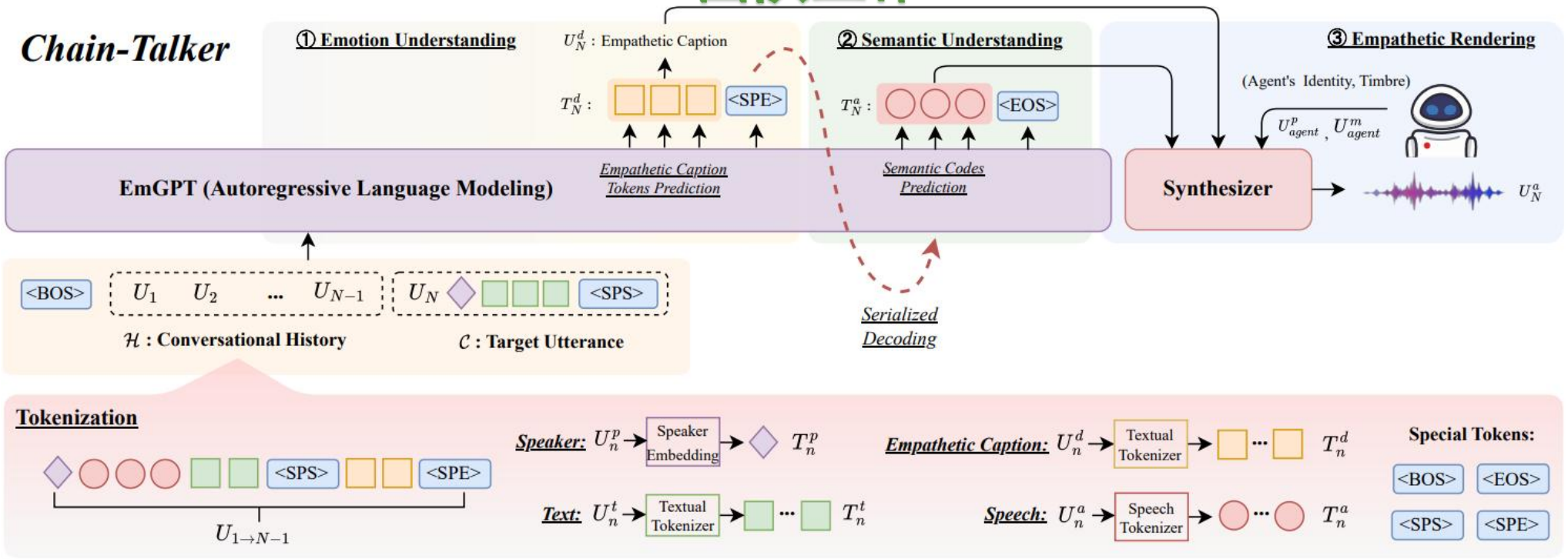
- 面向同理心交互，需要进一步建立情感理解、语义理解与语音表达之间的关联。



# Chain-Talker：情感理解、语义建模与共情渲染的链式生成

**目标：在对话语音合成中，明确目标回复的情感与语气，使生成语音的表达符合对话语境。**

团队工作

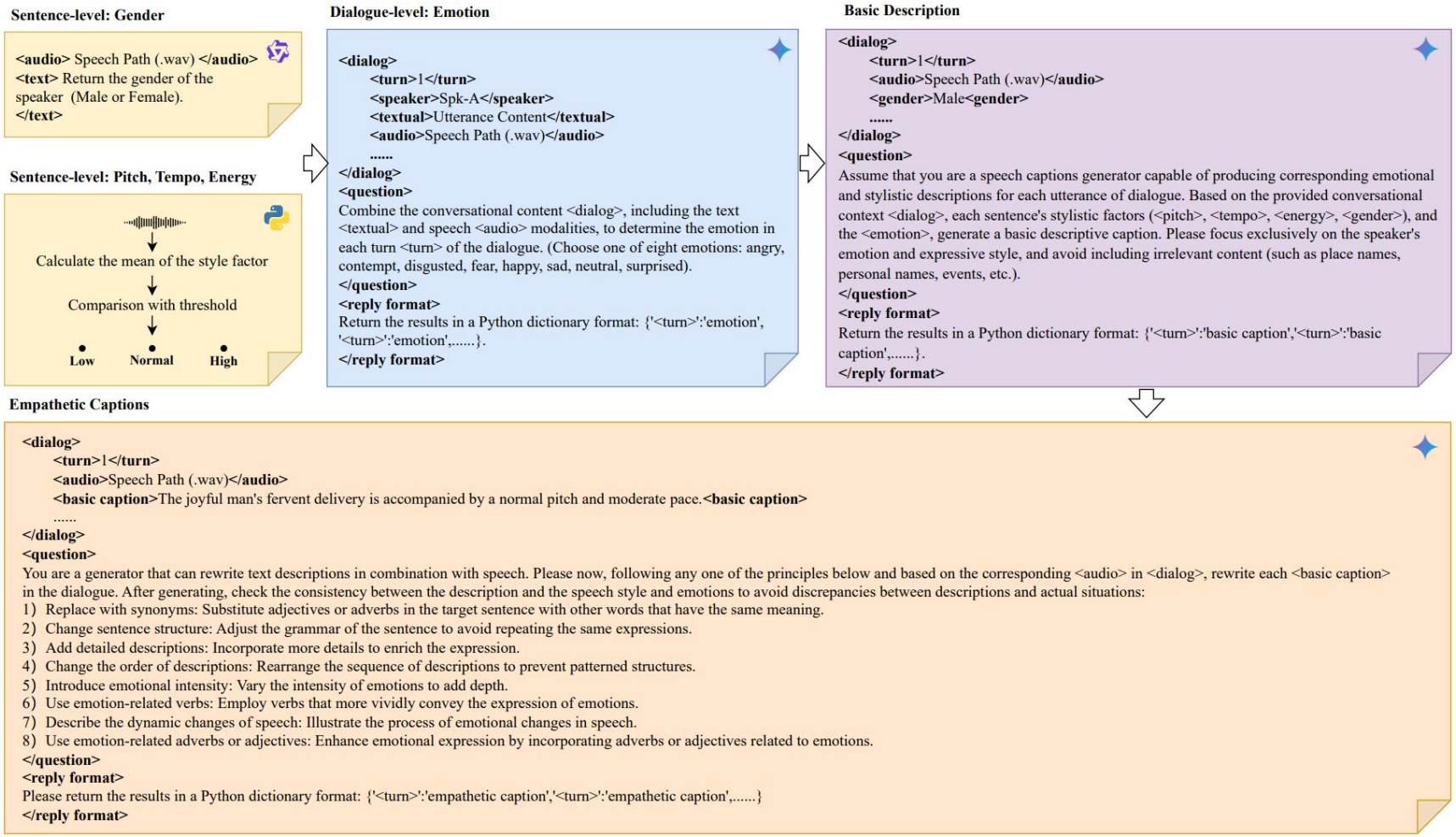



**核心技术：通过EmGPT根据对话历史与目标文本生成情感描述，在其引导下预测语义Token，再结合说话人条件，通过条件流匹配实现共情语音渲染。**

Hu Y, Liu R, Ren Y, et al. Chain-talker: Chain understanding and rendering for empathetic conversational speech synthesis[C]//**Findings of the Association for Computational Linguistics: ACL 2025**. 2025: 1988-2003.

# CSS-EmCap：自然语言情感描述

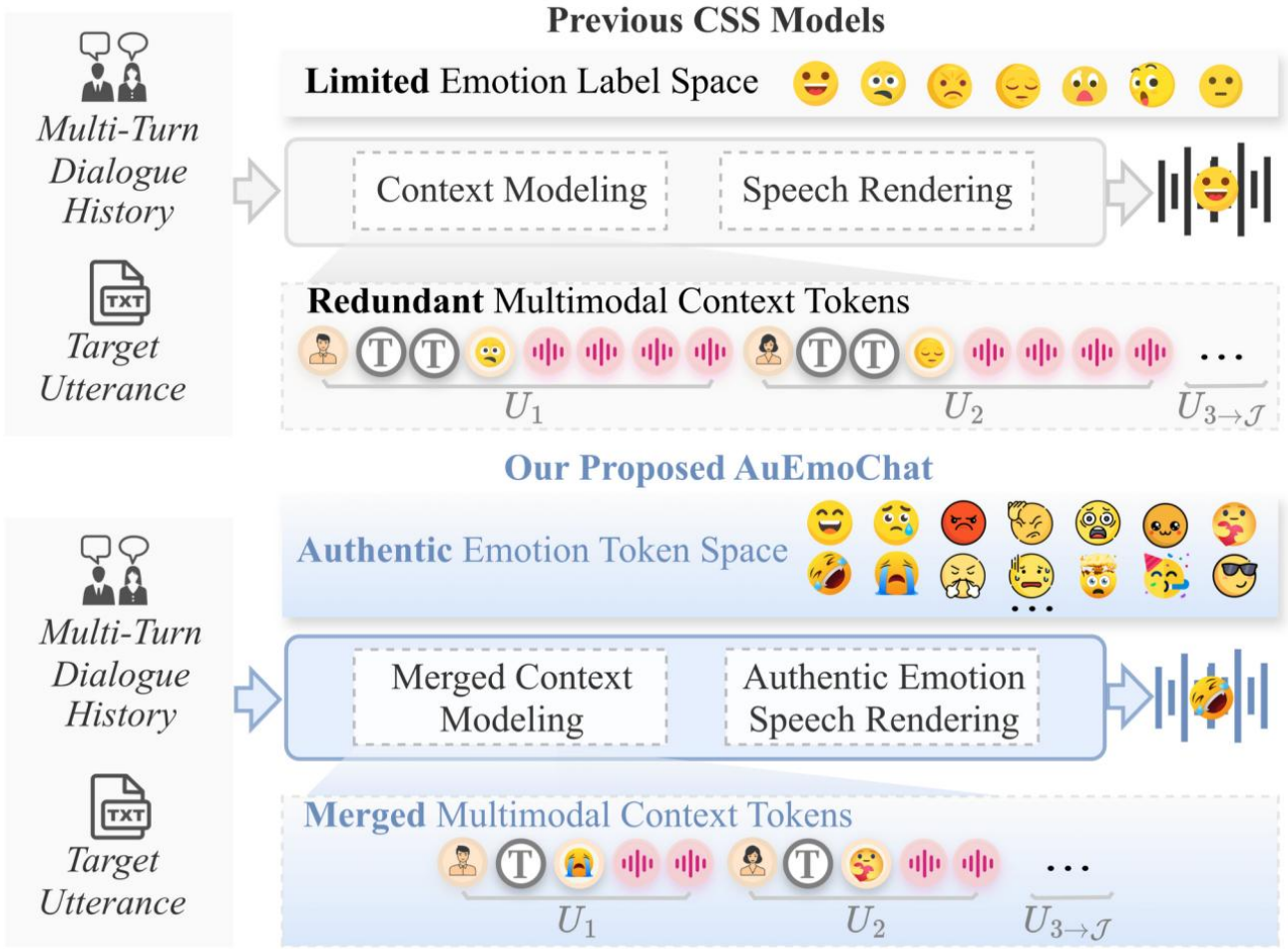
- 单一情感类别难以标注丰富的语气与共情信息




Hu Y, Liu R, Ren Y, et al. Chain-talker: Chain understanding and rendering for empathetic conversational speech synthesis[C]//Findings of the Association for Computational Linguistics: ACL 2025. 2025: 1988-2003.

# 从有限情感标签走向真实细粒度情感表达

- 预定义情感类别难以覆盖复杂、细腻的真实情感
- 冗余的多模态历史 Token 则容易干扰上下文理解

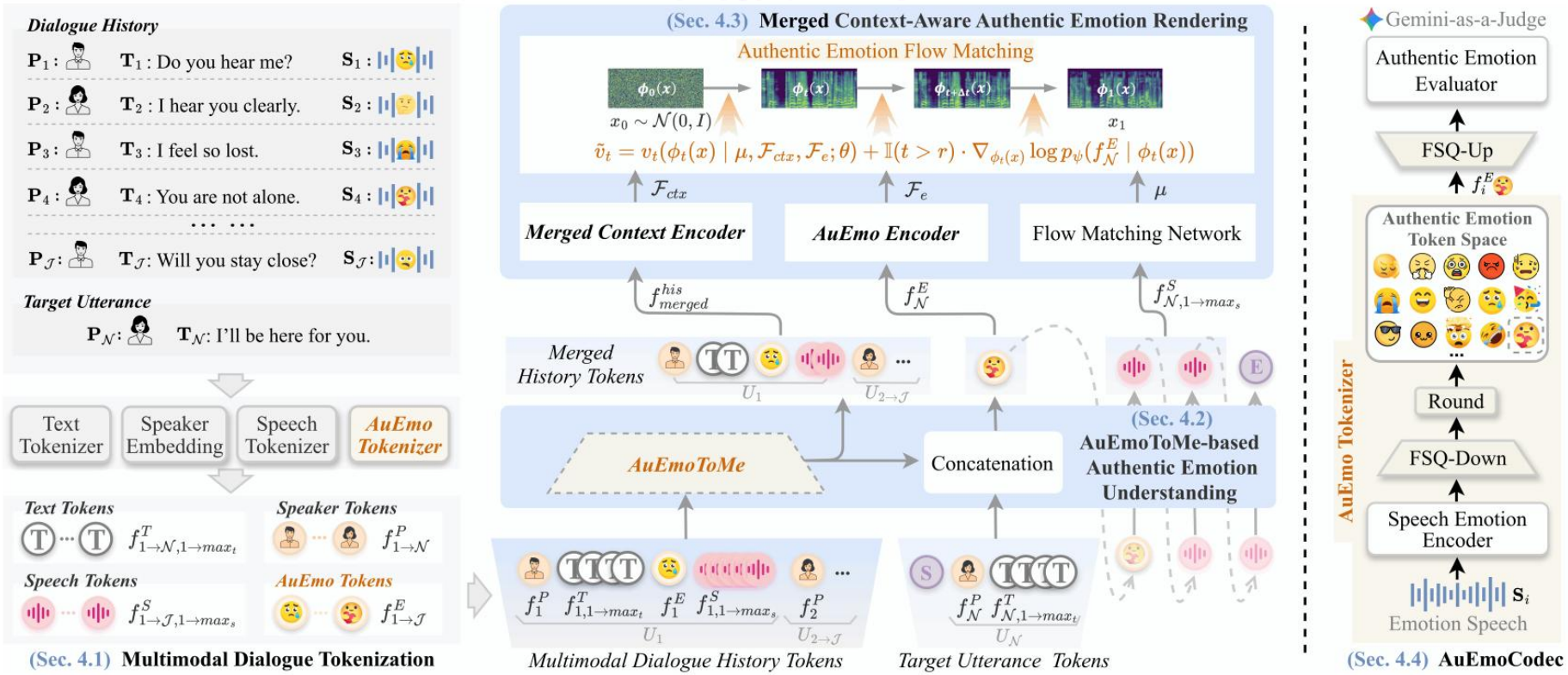



Jia Z, Zhao Y, Liu R, et al. AuEmoChat: Authentic Emotion Understanding and Rendering for Conversational Speech Synthesis[J]. arXiv preprint arXiv:2607.15755, 2026.

# AuEmoChat：细粒度真实情感理解与高效上下文渲染

**目标：在多轮对话中，捕捉超出预定义类别的细腻情感变化，兼顾长上下文处理效率与语音情感表现力。**

**团队工作**

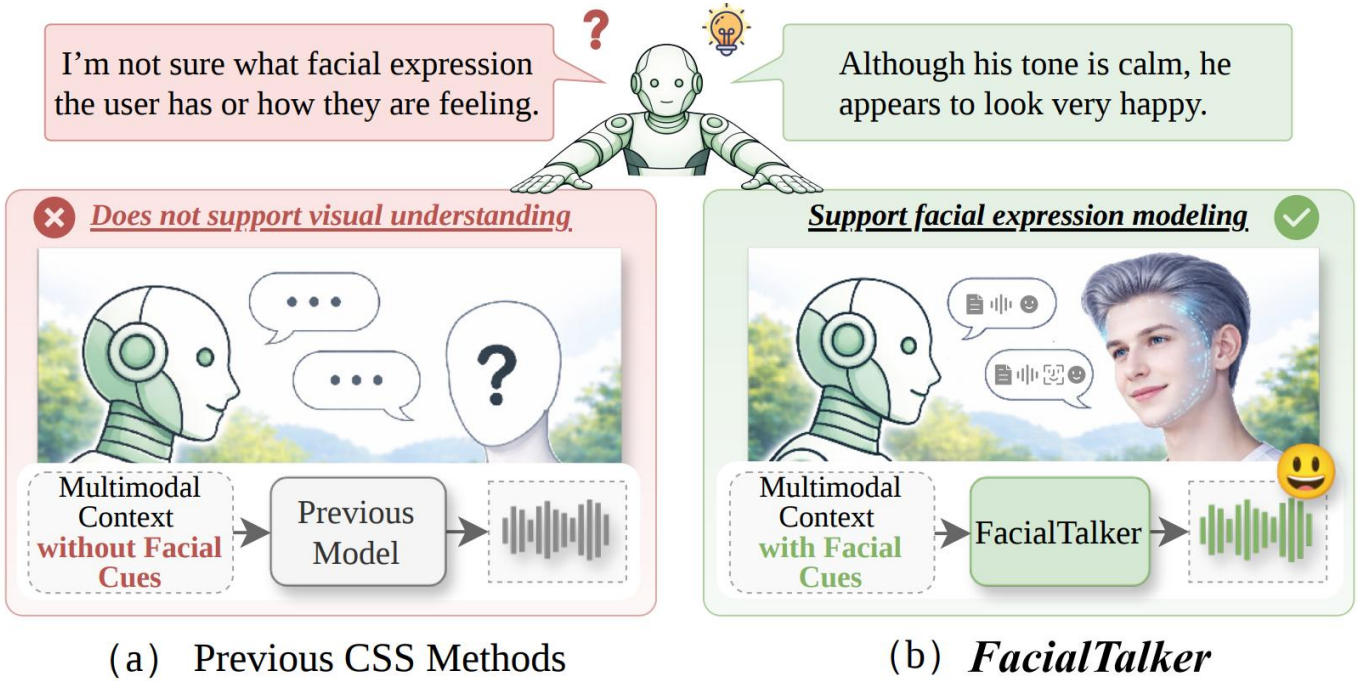



**核心技术：通过AuEmoCodec学习离散真实情感Token，利用AuEmoToMe合并冗余上下文并保留情感线索，自回归预测目标情感与语音Token，再通过真实情感条件流匹配完成语音**

Jia Z, Zhao Y, Liu R, et al. AuEmoChat: Authentic Emotion Understanding and Rendering for Conversational Speech Synthesis[J]. **Proceedings of the 34nd ACM International Conference on Multimedia. 2026**

# 为什么音频之外还需要人脸表情?

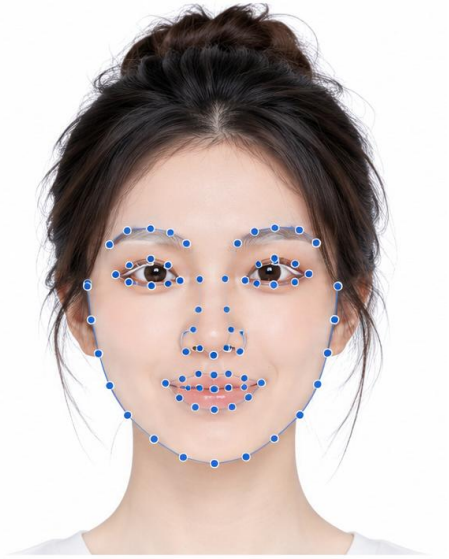
- 文本与语音线索难以覆盖全部情感信息
- 面部表情能够补充非语言情感线索
- 视听信息融合有助于消除情感歧义




(a) Previous CSS Methods (b) *FacialTalker*

# 面部表情的三种常见表示

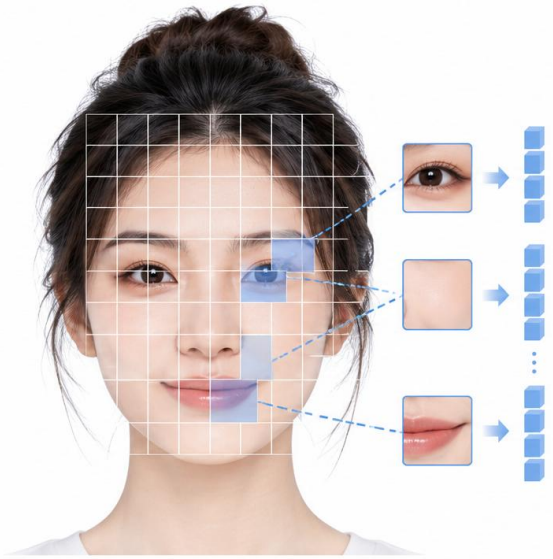
- Landmarks：几何关键点，适于动态同步与轨迹
- Action Units：面部肌肉动作组合，更接近表情机制
- Visual Tokens：压缩视觉动态，便于LLM序列建模



Landmarks



Action Units

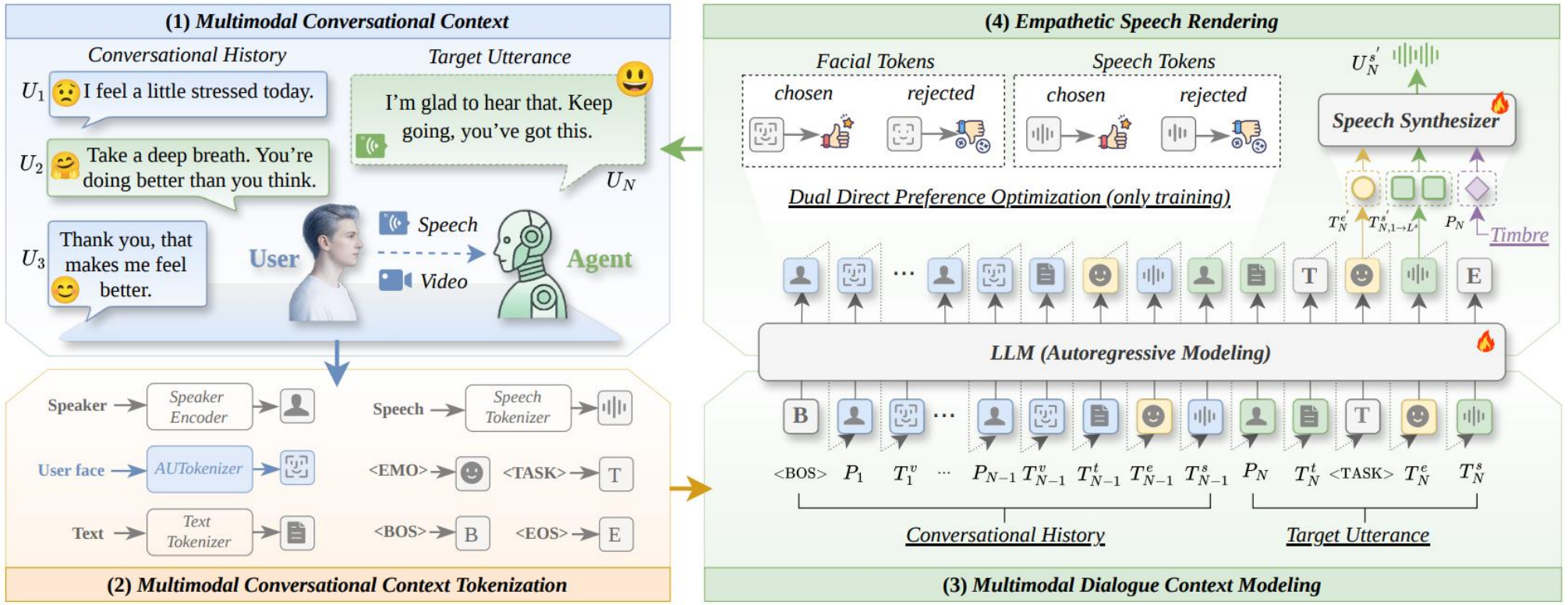


Visual Tokens

# FacialTalker：细粒度面部表情感知的对话语音合成

**目标：在多模态对话中，充分利用面部动作蕴含的情感线索，生成情感自然、符合交互语境的目标语音。**

**团队工作**

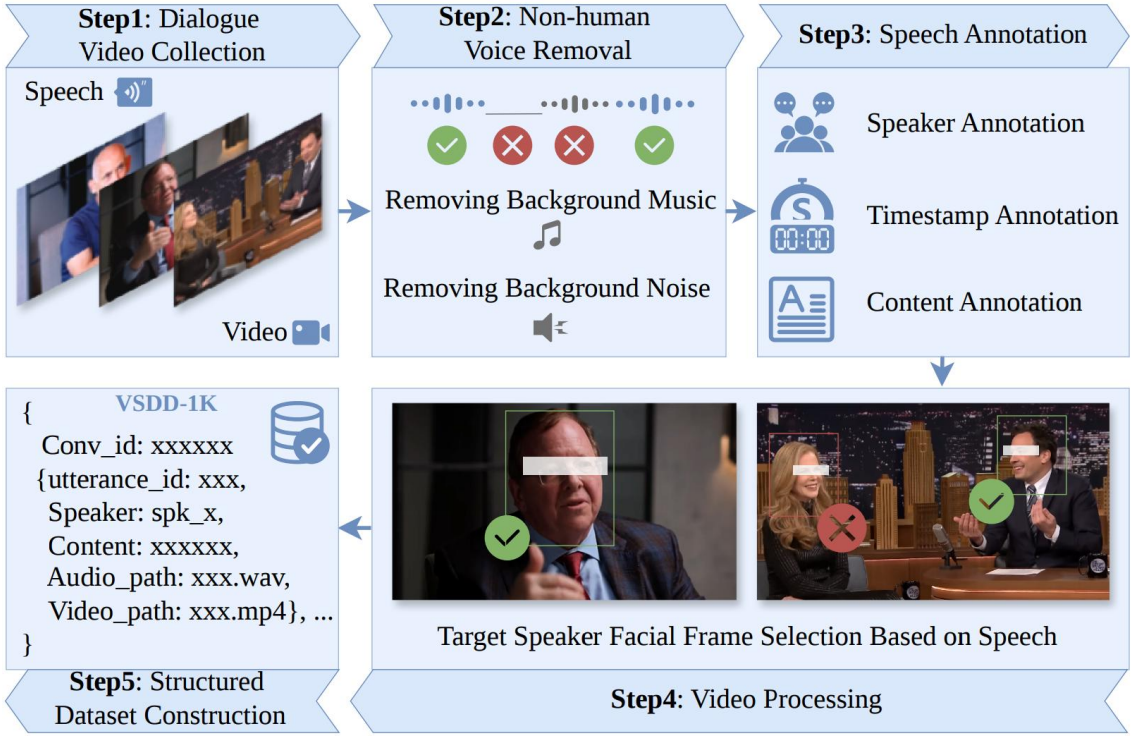



**核心技术：通过AU监督的AUTokenizer提取离散表情Token，结合多模态上下文预测目标情感与语音Token，利用DualDPO优化视觉与语音建模，并通过情感引导的条件流匹配完成语音渲染。**

Hu Y, He S, Liu R, et al. Let Me Look at You: Advanced Facial Expression Modeling for Conversational Speech Synthesis[J]. **Proceedings of the 34nd ACM International Conference on Multimedia. 2026.**

# VSDD-1K：大规模真实视听对话数据集

- 通过自动化数据处理与视听对齐，构建约 1,033 小时的真实面对面视听对话数据集。



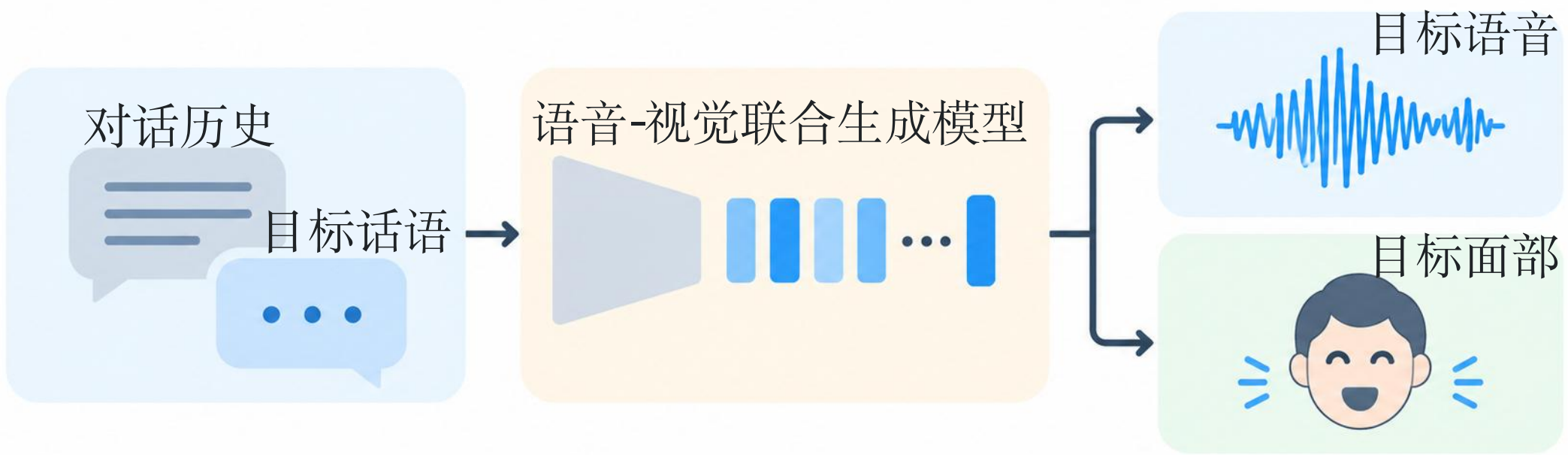

| Dataset | Language | Modal | Duration (h) | Speaker |
|---|---|---|---|---|
| NCSSD [34] | CN, EN | ( ,a,t) | 236 | >776 |
| RAMC [65] | CN | ( ,*a*,*t*) | 180 | 663 |
| DailyTalk [22] | EN | ( ,*a*,*t*) | 20 | 2 |
| STUDIES [52] | JP | ( ,*a*,*t*) | 8.2 | 3 |
| RyanSpeech [66] | EN | ( ,*a*,*t*) | 10 | 1 |
| CALLS [51] | JP | ( ,*a*,*t*) | 6.5 | 1 |
| ECC [24] | EN | (*v*,*a*,*t*) | 24 | > 26 |
| MultiDialog [44] | EN | (*v*,*a*,*t*) | 340 | 12 |
| AvaMERG [69] | EN | (*v*,*a*,*t*) | 239 | 65 |
| **VSDD-1K** | **EN** | **(*v*,*a*,*t*)** | **1033** | **>1498** |

Hu Y, He S, Liu R, et al. Let Me Look at You: Advanced Facial Expression Modeling for Conversational Speech Synthesis[J]. **Proceedings of the 34nd ACM International Conference on Multimedia. 2026.**

# 语音与视觉联合生成

## 不仅需要决定语音“怎么说”，还需要考虑面部“怎么动”

- 能否联合生成语音与面部动作，实现情感表达一致?



# 语音与视觉联合生成的挑战

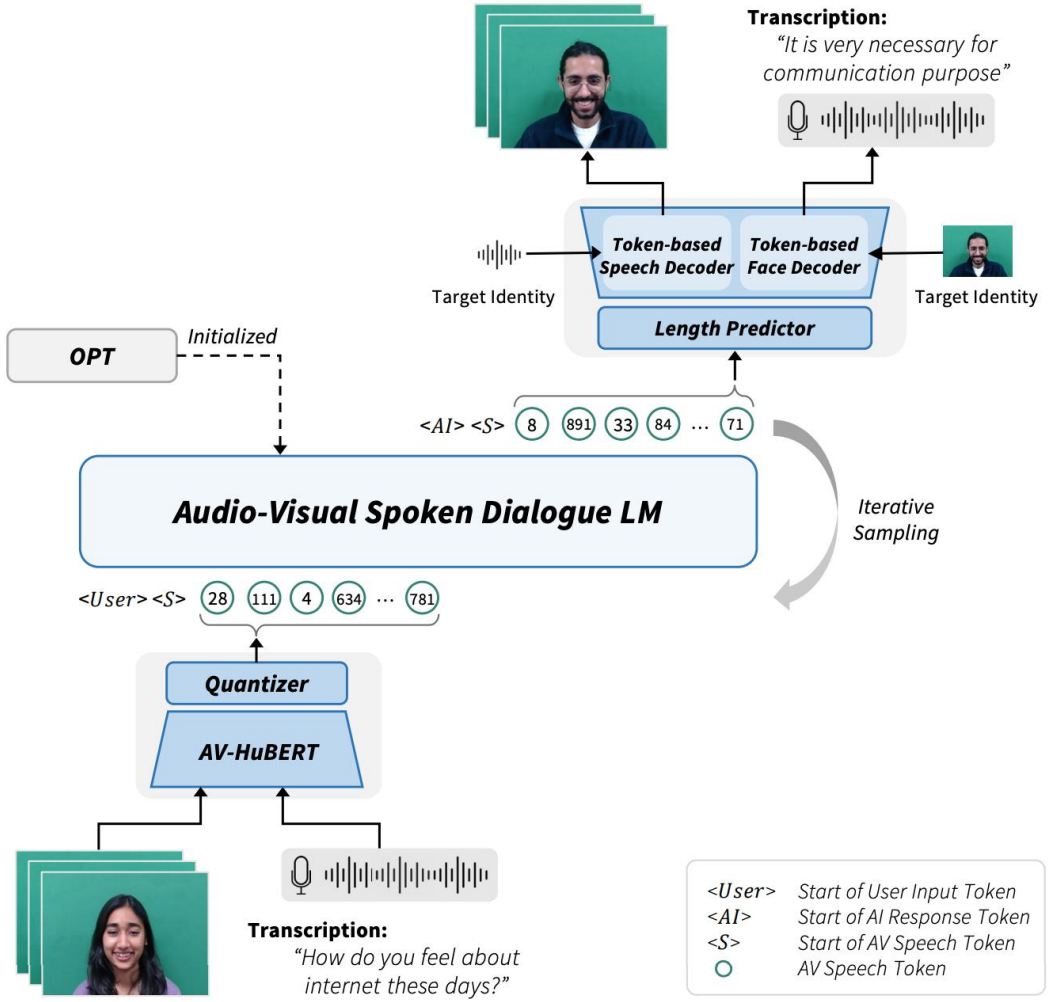
- 时序同步：　语音发音与口型动作需要精确对齐。
- 情感协调：　面部表情与语音情感需要保持一致。
- 稳定生成：　兼顾低延迟生成与面部身份一致性。



# Let's Go Real Talk：面对面音视对话模型

**目标：根据用户语音与面部视频理解交互内容，生成内容连贯、声画同步的语音与人脸视频回复。**

**核心技术：通过AV-HuBERT与量化构建离散视听语音Token，结合文本与视听联合预训练进行自回归对话建模，再通过时长恢复及语音、人脸双解码器生成视听回复。**

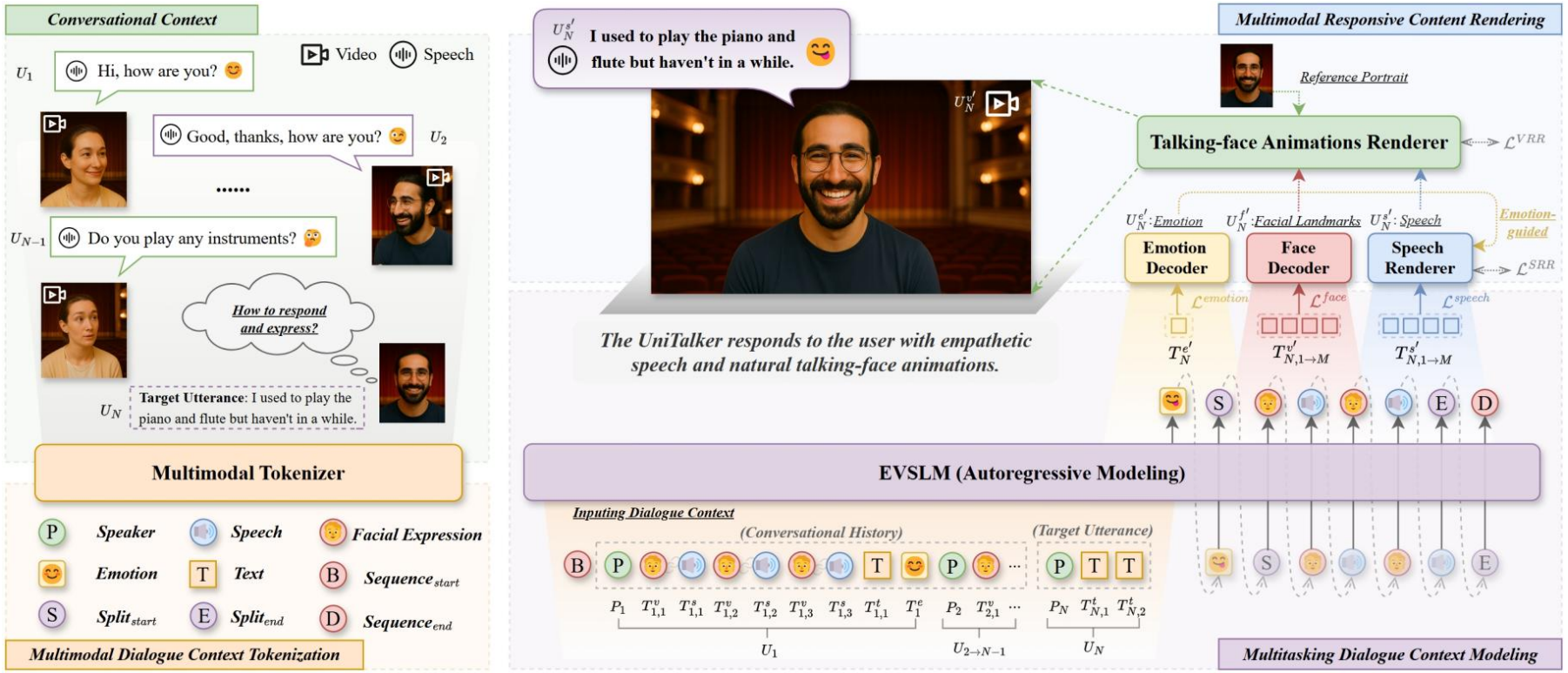



Park S, Kim C, Rha H, et al. Let's go real talk: Spoken dialogue model for face-to-face conversation[C]//**Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024**: 16334-16348.

# UniTalker：统一对话语音-视觉生成

**目标：在对话交互中，生成符合语境的语音与面部动作，保持两者在时间与情感上一致。**

**团队工作**




**核心技术：通过LmkCodec构建与语音同采样率的面部关键点Token，利用EVSLM先预测情感，再交替生成硬对齐的面部与语音Token，结合情感引导的条件流匹配与人脸渲染生成视听表达。**

Hu Y, Liu R, Ren Y, et al. UniTalker: Conversational speech-visual synthesis[C]//**Proceedings of the 33rd ACM International Conference on Multimedia. 2025**: 10248-10257.

# 研究总结｜多模态交互中的语音与视觉协同表达

01

**对话数据集**

- 构建多轮、多模态对话语音数据，并逐步丰富情感描述与面部视觉信息。
- NCSSD（ACM MM 2024）、CSS-EmCap（ACL 2025）...

02

**上下文建模**

- 从多尺度语义与韵律建模，扩展到情感关系建模和外部风格知识增强。
- ECSS（AAAI 2024）、RADKA-CSS（INFFUS 2025）...

03

**生成式表达**

- 通过自回归上下文建模，实现显式情感理解、细粒度情感渲染。
- Chain-Talker（ACL 2025）、AuEmoChat（ACM MM 2026）...

04

**视听协同**

- 从面部表情辅助情感感知，发展到语音与面部动作的联合生成及时序对齐。
- UniTalker（ACM MM 2025）、FacialTalker（ACM MM 2026）...

# 第四部分

# 总结与展望

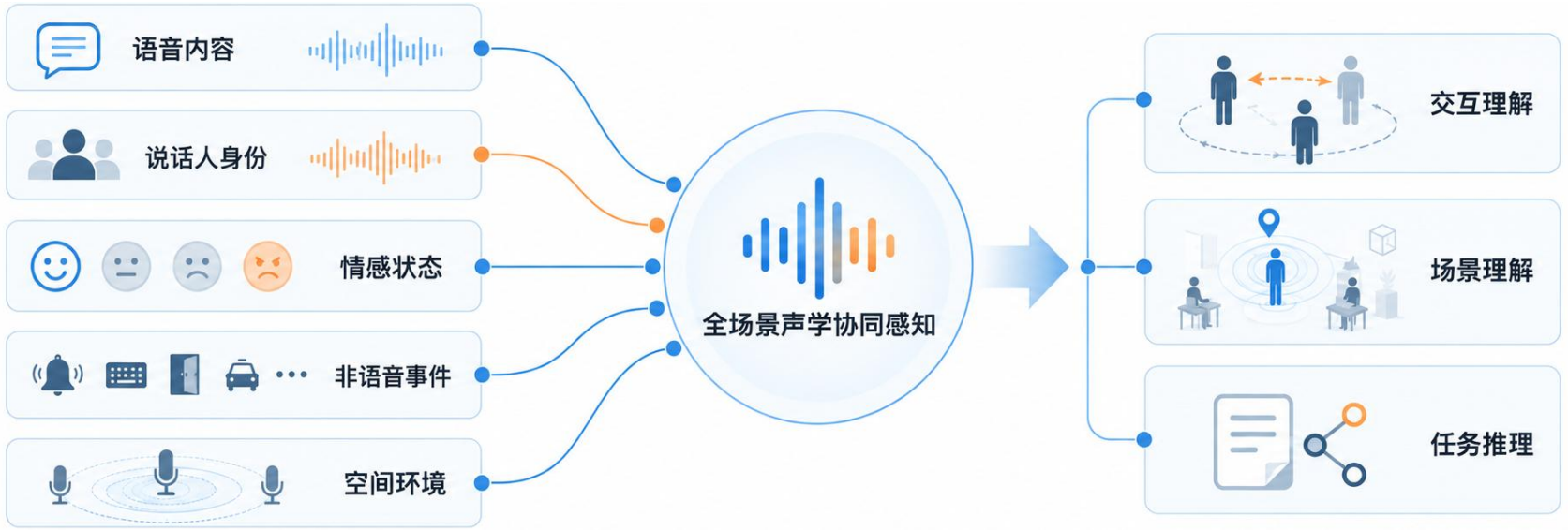
# 总结与展望

## 面向复杂人机交互，构建“鲁棒感知—精准理解—共情表达”的对话智能技术体系

- **多说话人场景下的语音感知：** 联合识别语音内容、说话人身份与时序关系，提升复杂场景下的语音感知能力。
- **模态缺失下的情感-意图理解：** 缓解模态缺失造成的信息损失，融合多模态线索与对话上下文，实现情感与意图的准确理解。
- **多模态对话交互中的语音与视觉协同表达：** 协调语音、面部表情与动作在时间、语义和情感上的一致性，实现自然、富有共情能力的表达。

# 趋势一：从多说话人识别走向全场景声学协同感知

## 复杂交互中的声学感知不应局限于语音内容与说话人身份，还需要联合理解情感状态、声学事件及空间环境




- 多维声学信息协同： 联合感知语音内容、说话人身份、情感与非语音事件。
- 声学场景关系建模： 关联声源位置、时序活动与交互状态，增强复杂环境理解。
- 任务驱动的动态感知： 根据交互目标选择关键声学线索，实现多任务协同推理。

# 趋势二：模态鲁棒性从补全走向可信推断

## 复杂场景下不仅要恢复缺失信息，更需要判断多模态证据是否可信

- 模态缺失与噪声干扰可能同时存在，单纯重构难以保证信息可靠性。
- 显式评估模态质量与预测不确定性，降低错误证据的干扰。
- 动态选择补全、蒸馏、检索与融合策略，实现可信情感推断。



# 趋势三：从短期上下文走向长期记忆与个性化表达

## 拟人化对话表达不仅取决于当前情绪，还受到长期互动关系与个人偏好的影响

- 短期上下文反映当前语义、意图和情感变化。
- 长期记忆保留用户偏好、交互习惯与关系演变。
- 通过相关记忆检索和动态更新，实现个性化、上下文适切的表达。



# 实验室介绍

实验室主页
(https://ttslr.github.io)

实验室Github地址
(https://github.com/AI-S2-Lab)

实验室公众号
(IMU语音理解与生成实验室)

# 致谢



**左昊麟**
22级博士研究生
内蒙古大学教师
在 ACM MM、ICASSP、TAFFC发表多篇论文



**胡一帆**
23级博士研究生
在 ACL、AAAI、ACM MM、TASLP发表多篇论文



**麻宇轩**
23级博士研究生
在 TAFFC、Data Intelligence发表多篇论文



**贾真琦**
25级博士研究生
在ACM MM、EMNLP、ICASSP、Information Fusion发表多篇论文

# 致谢

## 感谢聆听

## 敬请批评指正！

