arXiv:2603.03827v1 [cs.MM] 4 Mar 2026

基于层次语义表示的进化式多模态推理，用于意图识别

Qianrui Zhou[1], Hua Xu[1*], Yunjin Gu[1,3], Yifan Wang [1,2], Songze Li [1,2], Hanlei Zhang [1]清华大学 [2]河北科技大学[3]香港中文大学

zgr22@mails.tsinghua.edu.cn, xuhua@tsinghua.edu.cn

摘要

多模态意图识别旨在通过联合建模多种模态来推断人类意图，在现实世界的对话系统中发挥着关键作用。然而，当前方法难以建模复杂意图所蕴含的层次语义，也缺乏基于多模态表示进行自我进化推理的能力。为解决这些问题，我们提出了一种新方法 *HIER*，将层次语义表示与进化式 推理相结合，并基于多模态大语言模型（*MLLM*）。受人类认知启发，*HIER* 引入了一种结构化推理范式，将多模态语义组织为三个逐步抽象的层级。首先，利用模态专属令牌捕捉局部语义线索；随后，通过标签引导策略对其聚类，形成中层语义概念。为捕捉更高阶结构，*HIER* 使用 *JS* 散度分数筛选概念间关系，以突出概念之间显著的依赖关系。然后，通过思维链驱动的提示将这些层次表示注入 *MLLM*，从而实现逐步推理。此外，*HIER* 还采用自演化机制，通过多模态大语言模型反馈来优化语义表示，使其能够在推理过程中动态适应。在三个具有挑战性的基准数据集上进行的实验表明，*HIER* 始终优于最先进的方法和多模态大语言模型，在所有指标上均提升了 *1–3%*。代码及更多结果见 https://github.com/thuiar/HIER。

1. 引言

多模态意图识别是一项关键任务，是人机交互的基础。通过利用多种模态，模型能够获取更广泛的语义线索，从而全面、细致地解读复杂意图。随着其重要性日益凸显，多模态意图识别已广泛应用于下游应用，包括虚拟人 [38], 智能交通系统 [19], 医学诊断 [25, 36], 以及其他交互系统 [24, 26]。

尽管近年来多模态意图理解日益受到关注，但该领域的研究仍处于早期阶段。其中一个核心挑战在于构建能够捕捉复杂多模态语义多层级结构的层次表示。先驱工作 [47, 48] 通过发布涵盖多种社会意图的大规模基准数据集，并采用多模态情感分析中的融合策略 [12, 28, 37] 作为基线方法，为后续研究奠定了基础。近期方法更加关注对细粒度多模态线索进行建模，以增强意图理解，具体包括强化不同模态之间的交互 [4, 15, 18, 51, 55], 、利用视频内及跨视频上下文 [42], , 以及引入多粒度优化以提升意图理解能力 [49]。然而，这些方法忽视了语义信息的层次化特性，限制了其进行连贯且可靠推理的能力。另一项挑战在于依赖次优且静态的推理过程，这阻碍了对多模态语义进行自适应和层次化解释。尽管近期多模态大语言模型（MLLMs）展现出强大的推理能力 [1, 5, 20, 21, 39, 53], , 但由于缺乏细粒度层次推理路径，它们在处理复杂多模态语义时仍然面临困难。此外，自演化优化对于动态且可靠的推理至关重要，而现有工作缺乏跨复杂多模态场景进行泛化的能力，也无法在推理过程中进行自适应调整 [14, 31–33]。

为应对上述挑战，我们提出了一种名为 HIER 的新方法，如图1所示，该方法将层次化语义建模与自进化推理相结合，用于多模态意图识别。其动机源于人类认知固有的层次性：人们首先建立情境意识，随后识别并关联显著的语义线索，最后综合

* Hua Xu 是通讯作者。

这些线索，并通过迭代自我精炼强化的关系推理加以整合。为建模层次化语义表示，HIER 首先使用 Qwen2-VL [39] 编码器对文本和视觉输入进行编码，形成基础的上下文表示。为捕捉中层语义，我们提出一种多模态概念聚类策略，通过球面 K-Means++ [10] 将语义相似的词元归并为统一概念；其中，每个质心都通过余弦相似度加权以及与标签向量的凸组合，动态地与意图语义对齐。为建模高层关系，我们将成对概念输入信息瓶颈网络，以提炼紧凑的关系；通过原始概念与其组合关系的意图分类逻辑值之间的詹森-香农（JS）散度，量化该关系的重要性。保留散度较高的关系，以捕捉新颖且具有判别性的语义，反映超越单个概念的有意义交互。为实现自进化推理，我们设计了一种结构化思维链（CoT），引导 Qwen2-VL 依次经历上下文理解、概念分析和关系推理三个渐进阶段，并使其与层次化语义表示相对应。在概念和关系阶段，我们使用详细提示评估语义的合理性，并以此作为自我进化的基础。接下来，我们引入一种由反馈驱动的机制，通过内部推理过程精炼语义表示。具体而言，概念和关系的隐藏状态均通过共享生成头映射为词汇表逻辑值。在先前提示的条件下，这些逻辑值编码评估反馈，据此得出肯定和否定响应的归一化置信分数，从而动态精炼语义特征。

我们的贡献总结如下：（1）我们提出了 HIER，这是一种新颖的框架，通过建模层次化语义表征来捕捉人类认知的分层特性。据我们所知，这是首项为多模态意图识别建立多层渐进式推理范式的工作。（2）我们介绍了一种自演化推理机制，将结构化思维链指导与多模态大语言模型反馈相结合，以增强推理深度并自适应地优化语义表征。（3）在三个具有挑战性的基准数据集上，HIER 相较于最先进方法和领先的多模态大语言模型始终取得 1–3% 的提升，并能有效泛化至不同骨干网络，增强其推理能力。

2. 相关工作

2.1. 多模态意图识别

多模态意图识别旨在通过整合现实场景中的言语与非言语线索来理解用户意图。受多模态情感与情绪基准数据集 [2, 27, 44–46], MIntRec [47] 以一个捕捉多样交互意图的细粒度数据集开创了该任务，并采用情感融合技术 [12, 28, 37] 作为强基线方法，而 MIntRec2.0 [48] 则通过扩大数据规模、丰富意图分类体系并建立更全面的基准数据集进一步推进了该任务。随后，专门的意图识别方法涌现，以应对该领域的独特挑战。例如，TCL-MAP [51] 与MVCL-DAF [15] 通过词元级对比学习改进多模态融合，而InMu-Net[55] 利用信息瓶颈应对噪声非言语线索并缓解长尾偏差。此外，CAGC[42]与MuProCL [8] 分别聚焦视频与音频模态，捕捉时间依赖与语义上下文，从而增强对非文本模态的理解。为加强多模态整合，SDIF-DA [18] 采用由浅入深交互框架，而MIntOOD [49]则通过加权融合与多粒度优化来精炼意图表示。此外，DuoDN [4] 利用带有反事实推理的双网络解耦共享特征与模态特定特征。最近，LGSRR [52]探索利用大语言模型的推理能力来增强意图理解，并取得了令人鼓舞的结果，凸显了多模态语义推理的关键作用。然而，其推理过程仍然相对浅层，且依赖于特定的语义概念。

2.2. 多模态大语言模型的自进化

自我进化机制旨在使模型能够通过自我生成的经验自主获取并完善知识 [13, 22, 35]。这些策略源于迭代式基于人类反馈的强化学习 [7], ，该方法利用模型自身生成的数据，使模型与人类偏好对齐。在此基础上，自我进化经历了多个发展阶段，包括自我指令 [41], 、自我博弈 [3], 、自我改进 [17], 和自我训练 [9]。近年来，多模态大语言模型（MLLM）的推理能力不断增强，推动了多模态场景中自我进化方法的出现。例如，RLAIF-V [43] 利用能力更强的模型评估获得的回答，从而构建用于对齐的偏好对。类似地，SIMA [40] 直接利用标准答案比较回答，并生成偏好监督数据。其他方法则通过常规方式生成正向回答，并通过有意扰动（例如使用受损图像 [54] 或误导性提示 [6]）合成否定响应，以此构建偏好数据。此外，为摆脱对标注的依赖，SENA [34] 通过图像驱动的自我提问机制，自主生成高质量的问答对。然而，这些方法主要依赖基于偏好的对齐，缺乏

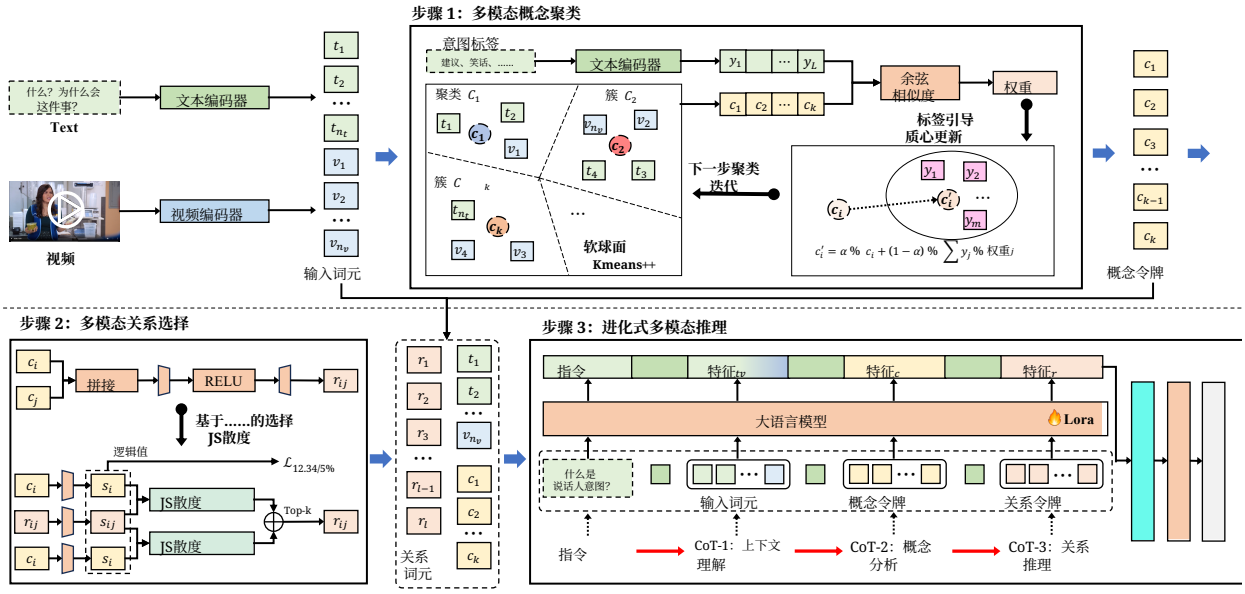


图1. 所提出的 HIER 方法概览。该模型包含三个关键步骤：(1) 多模态概念聚类，通过软球面 K-Means++ 并结合意图标签，将语义相关的词元归为中层概念表示；(2) 多模态关系选择，利用信息瓶颈网络和 JS 散度捕获有信息量的概念间依赖；(3) 进化式多模态推理，通过结构化思维链和自我进化机制开展层次化推理，从而提升推理深度和鲁棒性。

自反思推理过程，这限制了它们动态完善多模态理解的能力。

3. 方法

3.1. 总体架构

本节介绍 HIER 方法的架构，如图 1 所示。该方法包含三个主要步骤：多模态概念聚类、多模态关系选择和进化式多模态推理。前两个步骤从原始数据中逐步提取多模态输入词元、语义概念和概念间关系，从而构建结构化的层次化表示。在此基础上，最后一步通过自我进化进行多模态推理，以实现深度感知推理，并针对复杂的意图理解进行自适应细化。

3.2. 多模态表示

给定视频数据和文本输入，我们首先使用Qwen2-VL编码器提取对齐的模态特定表示。文本编码器将原始文本映射为语言词元 $T = \{t_1, t_2, \ldots, t_{n_t}\} \in \mathbb{R}^{n_t \times d}$，而视频编码器生成视觉词元 $V = \{v_1, v_2, \ldots, v_{n_v}\} \in \mathbb{R}^{n_v \times d}$，其中 n_t 和 n_v 分别为文本和视频的词元序列长度，d为统一特征维度。得益于Qwen2-VL的预训练，所提取的视觉词元和文本词元被嵌入共享语义空间，从而确保跨模态的语义兼容性。

这些词元随后被拼接为统一序列 $Z = \{z_1, z_2, \ldots, z_n\} \in \mathbb{R}^{n \times d}$，其中 n 是 n_t 与 n_v之和。所得词元序列表示局部语义单元，构成层次化语义建模的基础层级，并作为多模态概念聚类过程的基本元素。

3.3. 多模态概念聚类

人类意图理解依赖于能够连接原始多模态信号与复杂意图的中间线索，从而缩小语义鸿沟，实现精准解读。受这一认知过程启发，我们提出一种标签引导的聚类策略，将语义对齐的词元归为中层概念，以实现结构化且可靠的层次建模。具体而言，我们采用球面 K 均值++算法 [10]，并使用余弦相似度，从而更好地捕捉语义邻近性并减轻幅值偏差。给定输入词元 Z，我们的目标是将其划分为k个簇 $\{C_1, \ldots, C_k\}$，每个簇由一个质心$c_i \in \mathbb{R}^d$ 表示，作为中层语义概念。为了保持聚类的可微性，我们采用软概率形式进行词元分配 [29]。在聚类的第 u次迭代中，我们计算概率 $p_{i\,m}^{(u)}$，该概率用于量化词元 z_i 对更新后质心 $c_m^{(u)}$ 的贡献，如下所示：

$$p_{i,m}^{(u)} = \frac{\exp(\cos(z_i, c_m^{(u-1)}))}{\sum_{j=1}^{k} \exp(\cos(z_i, c_j^{(u-1)}))}, \tag{1}$$

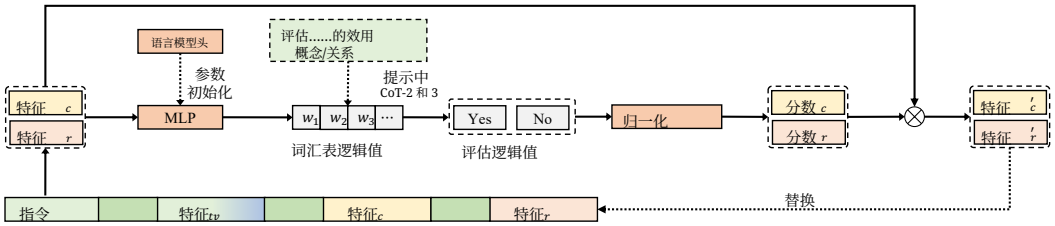


图 2。自进化模块的详细信息。我们首先复制 Qwen2-VL 的生成头，将概念和关系特征投影到词汇表逻辑值中。在有用性评估提示的条件下，我们随后提取并归一化肯定响应和否定响应的逻辑值，以得到置信度分数，进而引导特征细化，实现自适应且鲁棒的推理。

该概率用于将质心更新为 $c_m^{(u)} = \sum_{i=1}^{n} p_i^{(u)} \cdot {}_m z_i$。

为了使聚类与意图语义保持一致，我们将意图标签嵌入作为语义锚点，并在每次迭代后细化质心。具体而言，我们从文本编码器中获得意图向量 $\{y_1, y_2, \ldots, y_L\}$，其中 L 表示标签数量，并计算其与质心之间的余弦相似度，以得到对齐权重，形式化表示为：

$$\text{Weight}_{i,m}^{(u)} = \frac{\exp(\cos(c_m^{(u)}, y_i))}{\sum_{j=1}^{k} \exp(\cos(c_j^{(u)}, y_i))}, \tag{2}$$

从而强化从局部化词元到语义概念，再到意图层级理解的连贯语义层次结构。我们利用这些权重构建标签引导的语义向量，并通过凸组合将其与当前质心 $c_m^{(u)}$融合：

$$\tilde{c}_m^{(u)} = \alpha \cdot c_m^{(u)} + (1-\alpha) \cdot \sum_{i=1}^{L} \text{Weight}_{i,m}^{(u)} y_i, \tag{3}$$

其中 $\alpha \in [0,1]$ 平衡了两部分的贡献。通过标签引导的聚类，我们获得了高质量的语义概念 $\{c_1, c_2, \ldots, c_k\}$，这些概念捕捉了对意图理解至关重要的显著中层模式。

3.4. 多模态关系选择

为了提升中层语义概念的高层推理能力，我们通过关系选择策略对其交互进行建模，以捕捉概念间富有信息的依赖关系。具体而言，我们采用信息瓶颈网络 [11] 来编码紧凑的关系表示，并使用詹森-香农（JS）散度量化每种关系的语义新颖性。与库尔贝克-莱布勒（KL）散度相比，JS 散度具有更好的平衡性和对称性。至关重要的是，高散度关系能够捕捉超越单个概念的互补语义和涌现语义，这与人类的推理过程相契合：语义信息不仅源自单个概念，也源自它们之间的交互，从而提升意图理解的深度和连贯性。

给定概念表示 $\{c_1, c_2, \ldots, c_k\}$，我们考虑所有两两组合（$c_i$， c_j），并使用由带有 ReLU 激活的非线性层构成的信息瓶颈网络对它们的交互 r_{ij} 进行编码：

$$r_{ij} = \text{MLP}(\text{ReLU}([c_i; c_j])), \tag{4}$$

其中 $[c_i;\ c_j]$ 表示概念向量的拼接。为确保与意图语义保持一致，概念（c_i， c_j）及其关系 r_{ij} 通过意图分类目标进行联合优化：

$$s_i = \text{MLP}_c(c_i),\ s_j = \text{MLP}_c(c_j), \tag{5}$$

$$s_{ij} = \text{MLP}_r(r_{ij}), \tag{6}$$

其中，$\text{MLP}_c(\cdot)$ 和 $\text{MLP}_r(\cdot)$ 分别表示应用于概念表示和关系表示的分类头。此外， $\mathcal{L}_{\text{relation}}$ 定义为：

$$\mathcal{L}_{\text{relation}} = \text{CE}(s_i, g) + \text{CE}(s_j, g) + \text{CE}(s_{ij}, g), \tag{7}$$

$$\text{CE}(s, g) = -\frac{1}{B} \sum_{i=1}^{B} \sum_{j=0}^{|\mathcal{Y}|} \mathbb{I}_{[g_i=j]} \log \left(\text{Softmax}(s_i)^j\right), \tag{8}$$

其中， g 表示真实意图， B 为批次大小，$\mathcal{Y} = \{0, 1, \cdots, L-1\}$ 表示标签集。接下来，我们通过计算关系 logits s_{ij}与单个概念 logit之间的 JS散度，评估结合 c_i 和 c_j 所带来的语义增益。以 JS_i 为例：

$$m_i = \frac{1}{2}(\text{Softmax}(s_i) + \text{Softmax}(s_{ij})), \tag{9}$$

$$\text{JS}_i = \frac{1}{2}(\text{KL}(\text{Softmax}(s_i) \| m_i) + \text{KL}(m_i \| \text{Softmax}(s_{ij}))), \tag{10}$$

其中，$\text{KL}(\cdot|\cdot)$表示 KL 散度的计算。我们将最终分数计算为 $\text{JS}_{ij} = \text{JS}_i + \text{JS}_j$，并选择分数最高的 top-$k$ 个关系。鉴于不同实例的散度尺度可能有所不同，我们采用固定的保留比例，在每个样本中自适应地保留排名靠前的关系，从而有效筛选出信息丰富的关系表示 $\{r_1, r_2, \ldots, r_l\}$。

3.5. 进化式多模态推理

基于层次语义，我们通过结构化的思维链（CoT）提示引导 Qwen2- VL 进行渐进式多模态推理，该提示模拟人类认知过程，依次经历上下文感知、概念抽象和关系分析。思维链驱动的多模态推理分为三个渐进阶段，分别对应不同的语义层级。在 CoT-1（上下文理解）中，首先处理特定模态的输入词元 $Z = \{z_1, z_2, \ldots, z_n\}$，以建立对多模态场景的整体理解。随后，CoT-2（概念分析）聚焦于中层语义概念，利用词元 $\{c_1, c_2, \ldots, c_k\}$ 从上下文中提炼关键的意图相关线索。最后，CoT-3（关系推理）引入关系令牌 $\{r_1, r_2, \ldots, r_l\}$，对概念之间的交互进行建模，从而捕捉对于细致意图推断至关重要的高阶语义依赖关系。值得注意的是，在最后两个阶段，我们明确提示模型 <.style id='4'>判断其有用性，鼓励模型反思概念与关系，为自我进化奠定基础。此外，层次化语义特征最初由特殊词元表示（例如，<概念>和 <关系>），随后无缝整合到编码后的提示中，共同引导 Qwen2-VL 进行连贯而深入的推理。附录 2.1 和 4.1 详细介绍了提示设计和敏感性分析，附录 4.2 则提供了与思维链增强型多模态大语言模型的比较。

为增强复杂场景中的适应性，我们引入一种自演化机制，以实现对语义特征的内部评估与精炼，如图2所示。在 Qwen2-VL 处理提示表示后，我们分离概念与关系特征（特征$_c$与特征$_r$），并将其投影为词汇表逻辑值（逻辑值$_c$ 与逻辑值$_r$），所用线性层以 Qwen2-VL 的生成头初始化以实现对齐：

$$\text{Logits}_c = \text{MLP}_{lm}(\text{Feature}_c), \tag{11}$$

$$\text{Logits}_r = \text{MLP}_{lm}(\text{Feature}_r), \tag{12}$$

其中多层感知机$_{lm}$(·)指共享线性层。借助 CoT-2 与 CoT-3 中的反思提示作为上下文线索，我们直接提取肯定与否定回答（例如“是”和“否”）的逻辑值，并将其归一化为自我进化过程所用的精炼分数：

$$\text{Score}_c = \frac{\exp(\text{Logits}_c[Idx^+])}{\exp(\text{Logits}_c[Idx^+]) + \exp(\text{Logits}_c[Idx^-])}, \tag{13}$$

$$\text{Score}_r = \frac{\exp(\text{Logits}_r[Idx^+])}{\exp(\text{Logits}_r[Idx^+]) + \exp(\text{Logits}_r[Idx^-])}, \tag{14}$$

其中 Idx^+ 和 Idx^- 分别表示与肯定响应和否定响应对应的逻辑值。精炼分数用于调制每个语义特征

表1. MIntRec、MIntRec2.0 与 MELD-DA 数据集的统计信息。#C 表示类别数量。

数据集	#C	#训练集	#验证集	#测试集
MIntRec	20	1,334	445	445
MIntRec2.0（范围内）	30	6,165	1,106	2,003
MELD-DA	12	6,991	999	1,998

由其对应的置信度分数进行调制，从而增强推理的稳健性与适应性。最后，精炼特征计算为特征′ = 分数 · 特征，并替换推理感知特征中的原始表示。

将自演化特征输入生成头，以计算自回归语言建模损失 [39]$\mathcal{L}_{\text{task}}$。整体训练目标将此任务特定损失与关系引导辅助损失相结合：

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \beta \mathcal{L}_{\text{relation}}, \tag{15}$$

其中，β 用于平衡两种损失的贡献。

4. 实验

4.1. 数据集与评估指标

我们使用 MIntRec [47] 和 MIntRec2 的范围内子集 0 [48] 进行多模态意图识别，并使用 MELD-DA [30] 进行多模态对话分类。详细统计数据见表1。遵循 [47, 48]，我们采用分类指标进行评估，包括准确率（ACC）、F1 分数（F1）、精确率（P）、召回率（R）、加权F1分数（WF1）和加权精确率（WP）。

4.2. 基线方法

我们将 HIER 与最先进方法及领先的多模态大语言模型进行比较，包括：(1) **MISA** [12], **MAG-BERT**[28], 和 **MulT** [37]均改编自多模态情感分析；(2) **TCL-MAP** [51] 通过模态感知提示、基于标签的数据增强和词元级对比对齐来增强文本线索；(3) **SDIF-DA** [18]通过由浅入深的交互模块改进非语言对齐；(4)**MIntOOD** [49] 结合加权融合和多粒度建模，实现稳健的表征学习；(5)**MVCL-DAF** [15] 采用结合动态注意力分配融合的多视图对比学习；(6) **Qwen2-VL** [39], **VideoLLaMA2** [5], 和**LLaVA-NEXT** [21]是通用多模态推理领域近期进展的有力代表。

4.3. 实验设置

我们基于 Qwen2-VL 构建 HIER，并使用低秩适应 [16] 进行微调，秩设为 8、批次大小设为 2，通过 LLaMA-Factory [50] 实现。我们使用 AdamW [23] 优化器进行优化，并将损失平衡系数 β 设为 0.01。输入序列长度

表 2. 对比　　结果与最先进方法和领先的多模态大语言模型的对比，最佳分数以粗体标出，次佳分数加下划线　　.

数据集	MIntRec						MIntRec2.0						MELD-DA					
方法	准确率	F1值	P	R	WF1	加权精确率	准确率	F1值	P	R	WF1	加权精确率	准确率	F1值	P	R	WF1	加权精确率
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**

根据 Qwen2-VL 的配置动态调整，隐藏维度设为 3,584。聚类过程运行 30 次迭代，标签引导的插值系数 α 在训练期间学习。对于 MIntRec，我们使用 50 个概念词元和 25 个关系令牌；对于 MIntRec2.0 和 MELD-DA，则分别设为 40 和 20。遵循先前工作，我们报告 5 个随机种子 0–4 的平均结果。对于我们的方法和多模态大语言模型，使用 NVIDIA A100-PCIE GPU 进行 5 个周期的微调。详细的训练成本分析和超参数见附录 3.1和 3.2。

4.4. 主要结果

表2 展示了三个具有挑战性的数据集上的主要结果，其中最佳分数以粗体标出，次佳分数以下划线标出。与多模态意图识别基线相比，HIER 在 MIntRec 和 MIntRec2.0 的所有指标上均提升了 3–10%。尽管 TCL-MAP 和 SDIF-DA 在 MELD-DA 上仍具竞争力，HIER 在一半指标上仍实现了超过 2% 的提升。在具有细粒度意图的数据集上，这些改进尤为显著，凸显了 HIER 在捕捉复杂层次语义和支持渐进式意图推理方面的优势。此外，HIER 在三个数据集上均取得最高的 F1 分数，领先幅度分别为 7.67%、5.57% 和 3.34%，体现了其在不同意图类别上的稳健性，以及借助演化推理范式理解复杂语义的能力。与领先的多模态大语言模型相比，HIER 在三个数据集上均持续取得更优表现。在 MIntRec 上，其表现超过强劲的 Qwen2-VL 基线在所有指标上均领先 2% 以上；而在更具挑战性的 MIntRec2.0 上，其 F1值、精确率和召回率均展现出绝对优势，分别提升 10.36%、6.67% 和 10.46%。尽管 VideoLLaMA2 在 MELD-DA 上表现颇具竞争力，尤其是在 WF1 和精确率指标上，HIER 仍保持明显的性能领先，提升幅度超过 0.5%，进一步验证了其在多样化场景中的优越性。尽管多模态大语言模型具备出色的语义推理能力，HIER 通过利用显式的结构化语义层次和自演化推理机制，仍能超越这些模型；其机制与人类认知过程高度契合，适用于多模态意图理解。更多性能变化情况和显著性检验结果见附录 3.3，详细案例研究见附录 4.5。

4.5. 消融研究

为评估HIER中各模块的贡献，我们在四种设置下进行消融实验：去除概念、去除关系、去除思维链和去除自我进化，如表3所示。**（1）去除概念：**为评估语义概念的贡献，我们移除多模态概念聚类步骤，并直接对模态特定词元之间的关系进行建模。这导致MIntRec2.0和MELD-DA上的所有指标均出现超过1%的显著下降，凸显了结构化概念表示的重要性。尽管MIntRec上的性能仍基本相当，这可能是由于其规模较小限制了中层语义的学习，但所有指标仍观察到超过0.3%的一致下降。**（2）去除关系：**禁用关系选择模块会导致显著的性能下

表 3. 不同模块在 MIntRec、MIntRec2.0 和 MELD-DA 数据集上的消融结果。

	消融实验	ACC	F1	P	R	WF1 加权精确率
MIntRec	无概念	79.55	76.11	78.06	76.31	79.25 80.30
	无关系	77.08	70.76	71.55	71.52	76.92 77.83
	无思维链	75.51	71.57	72.03	72.36	74.95 75.37
	无自我进化	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	无概念	63.07	50.13	50.39	50.44	62.80 62.99
	不含关系	62.67	55.73	57.18	56.19	62.95 64.09
	不含思维链	63.35	56.81	58.59	56.19	62.87 63.27
	不含自我进化	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	不含概念	59.90	51.85	55.24	51.13	59.07 59.25
	不含关系	60.01	51.85	55.84	50.14	58.76 58.49
	不含思维链	59.91	52.08	56.73	50.13	58.62 58.46
	无自我进化	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**

降，在所有数据集的大多数指标上下降1%–7%，MIntRec2.0上的加权精确率除外。这一结果与先前研究 [18, 51] 相一致，这些研究强调了建模跨模态依赖的关键作用，并证实了我们的选择策略在识别具有语义信息的概念间关系方面的有效性。**（3）去除思维链：**为证明渐进式推理在利用层次语义方面的益处，我们将思维链提示替换为直接指令。在所有数据集上均观察到显著的性能下降，尤其是F1值、精确率和召回率指标的降幅超过2%，这表明由浅入深的推理轨迹能够通过层次化语义表示增强复杂意图理解。**（4）去除自我进化：**移除自我评估模块导致最大的性能下降，尤其是在MELD-DA上，精确率指标下降5.69%，证实了其在精炼语义表示并提升推理鲁棒性方面的作用。该模块的缺失还导致MIntRec上持续下降2%–6%，MIntRec2.0上下降1%–2%，凸显了自演化机制对整体性能的重要性。

4.6. 概念与关系数量的影响

为考察概念与关系数量的影响，我们在 MIntRec2.0 上使用准确率评估 HIER，并与表现最强的基线方法（包括 MulT 和 VideoLLaMA2）进行比较，如图 3所示。首先，我们将概念令牌数量从 20 调整到 60，步长为 10，同时保持概念与关系的数量比固定为 2:1，以确保关系对的数量处于合理范围内。接着，在概念令牌数量固定为 40 的情况下，我们调整

表4. HIER在MIntRec数据集上使用不同多模态大语言模型骨干网络的性能。

方法	ACC	F1	P	R	WF1	WP
LLaVA-NeXT	72.65	64.94	66.21	64.65	72.63	73.59
+HIER	**78.20**	**75.33**	**75.32**	**76.30**	**77.97**	**78.59**
VideoLLaMA2	74.61	71.64	71.82	73.35	74.37	75.43
+HIER	**77.18**	**74.56**	**75.33**	**75.06**	**76.35**	**77.27**

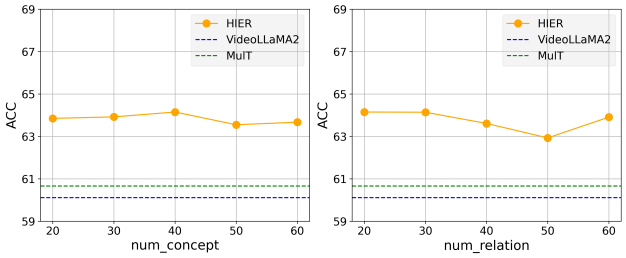


图 3. HIER 中概念与关系数量对性能的影响，在 MIntRec2.0 数据集上进行评估。

关系令牌数量从 20 调整到 60，以评估其独立影响。这些设置使我们能够系统地分析层次语义结构的深度和粒度如何影响整体模型准确率。随着概念令牌数量增加，HIER 始终保持出色的性能，并在 40 个概念时达到 64.15% 的最高准确率，显著优于基线方法。这证实了标签引导的聚类能够稳健地发现意图相关语义概念，无论目标簇数量如何变化。对于关系令牌，HIER 表现出显著的稳定性，仅在令牌数量为 50 时准确率略微降至 62.93%，仍比两种基线方法高出 1% 以上。这种稳定性凸显了多模态关系选择策略在提炼语义丰富且具有意图区分性的关系方面的优势。通过有效过滤冗余或噪声相关性，模型仅保留真正有助于理解复杂意图的关系线索，进一步彰显了多模态语境下结构化推理的价值。

4.7. 不同多模态大语言模型骨干网络的性能

为进一步评估 HIER 在不同骨干模型上的泛化能力，我们将其与另外两个具有代表性的多模态大语言模型 LLaVA-NeXT 和 VideoLLaMA2 集成，并采用与主要实验相同的超参数和实验设置。如表4所示，将 HIER 与 LLaVA-NeXT 集成后，各项指标均取得了稳定且显著的提升，增幅均超过 5%。其中，F1值和召回率的提升尤为突出，超过 10%，这表明 HIER 不仅通过分层语义建模加深并构建推理过程，使其更深入、逻辑更连贯，还利用自演化机制持续优化中间表示，从而显著增强骨干模型的推理能力和推理鲁棒性。分层

表5。 HIER与优于基线方法在MIntRec数据集上各类别的F1分数比较。

方法	抱怨	赞扬	道歉	感谢	批评	Care	同意	嘲弄	炫耀	反对	Joke
MISA	63.91	86.63	97.78	**98.03**	53.44	87.14	92.05	22.15	46.44	36.15	38.74
TCL-MAP	68.70	87.20	97.70	97.00	51.30	86.80	93.10	17.20	50.80	35.90	29.00
VideoLLaMA2	75.00	**93.02**	98.18	92.00	63.16	82.93	**96.00**	**35.29**	**63.64**	50.00	47.62
Qwen2-VL	74.14	88.10	94.55	95.83	56.00	89.47	77.42	31.58	63.16	**80.00**	63.16
HIER	**79.67**	88.37	**100.00**	97.96	**69.57**	**94.74**	85.71	23.53	60.87	66.67	**66.67**
人类	80.08	93.44	96.15	96.90	72.21	96.09	87.21	65.55	78.10	69.04	72.22

方法	告知	建议	安排	介绍	安慰	离开	预防	问候	寻求帮助
MISA	70.18	69.56	67.32	67.22	78.78	77.23	83.30	82.71	67.57
TCL-MAP	72.80	68.90	65.40	68.40	79.80	**83.40**	**83.60**	**90.10**	66.40
VideoLLaMA2	62.86	**81.48**	51.06	71.79	88.89	64.86	75.86	84.62	54.55
Qwen2-VL	70.27	79.17	66.67	**85.00**	84.85	68.57	81.25	80.00	62.50
HIER	**75.47**	78.26	**72.73**	76.19	**91.89**	70.59	81.48	80.00	**77.78**
人类	79.69	87.14	81.40	84.09	95.95	97.06	86.43	94.15	88.54

语义建模能够加深并构建推理过程，使其更深入、逻辑更连贯；自演化机制则能持续优化中间表示，从而提升推理鲁棒性。将 HIER 应用于 VideoLLaMA2 后，除 F1值外，其余各项指标也均显著提升至 75% 以上，而原始模型仅在加权精确率指标上超过这一阈值。这些结果表明，HIER 在不同骨干模型上具有很强的泛化能力，能够有效增强现有多模态大语言模型对复杂多模态语义的推理能力。采用最新骨干模型的结果以及与闭源模型的比较见附录4.3和4.4。

4.8. 细粒度意图类别上的性能

为深入洞察我们方法的有效性，我们报告了 MIntRec 上每个细粒度意图标签的 F1分数，如表5所示。比较对象包括排名前两位的多模态意图识别方法（TCL-MAP 和 MISA）以及两种领先的多模态大语言模型（Qwen2-VL 和 VideoLLaMA2），这些方法和模型均根据总体 F1 性能选定。此外，人工评估结果 [47] 被用于识别潜在的改进方向，为发现模型局限性提供了有价值的参考。从总体来看，HIER 在细粒度意图识别方面展现出显著优势，在 20 个类别中的 9 个类别上取得最高 F1 分数，几乎是排名第二的 VideoLLaMA2（5 个类别）的两倍。这一明显差距表明，HIER 在处理大多数细微意图类别时具备更强的能力和鲁棒性。具体而言，在这 9 个意图类别中的 5 个类别上，我们的方法相比第二名取得了超过 4% 的显著提升，凸显了其捕捉复杂且具有区分性的意图语义的有效性。此外，HIER 在道歉 意图类别上取得了令人瞩目的 100% F1 分数，有力地证明了我们所提出的层次表示和演化推理方法的有效性。与人工评估相比，HIER 取得了极具竞争力的结果，在道歉和感谢 类别上超越了人类表现，并在另外三个类别上将差距控制在 2% 以内。这一强劲表现彰显了我们推理范式的重要性，该范式模拟了人类意图理解的逐步性和层次性。然而，在嘲弄、炫耀和离开 类别上，HIER 的表现明显低于人类，这可能是由于语义不一致所致，例如面部表情与口头内容不匹配。这些挑战指出了未来探索的关键方向，尤其是增强模型在复杂场景中解决冲突的能力。

5. 结论

本文介绍了 HIER，这是一种新颖的方法，将层次语义表示与自进化推理相结合，以应对多模态意图识别中捕捉渐进语义和开展自进化推理的挑战。HIER 构建了一个分层语义体系，从模态专属令牌，到中层的意图感知概念，最终延伸至概念间关系。该体系通过标签引导聚类策略形成语义对齐的概念表示，并借助信息论机制有效筛选判别性概念间关系。在此层次结构的基础上，HIER 通过结构化思维链策略开展受认知启发的渐进式推理，并进一步利用自进化模块通过内部反馈优化语义特征。在三个具有挑战性的数据集上进行的实验表明，HIER 的表现优于最先进基线和领先的多模态大语言模型，充分彰显了其在建立新颖的进化式多模态推理范式方面的开创性作用。

致谢

本文由中国国家自然科学基金资助（项目编号：62173195）。

参考文献

[1] 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 Menick、Sebastian Borgeaud、Andrew Brock、Aida Nematzadeh、Sahand Sharifzadeh、Mikolaj Binkowski、Ricardo Barreira、Oriol Vinyals、Andrew Zisserman 和 Karen Simonyan。Flamingo：一种用于少样本学习的视觉语言模型。载于《神经信息处理系统进展》，2022。1[2]Carlos Busso、Murtaza Bulut、Chi-Chun Lee、Abe Kazemzadeh、Emily Mower、Samuel Kim、Jeannette N Chang、Sungbok Lee 和 Shrikanth S Narayanan。Iemocap：交互式情感双人动作捕捉数据库。《语言资源与评测》，42(4):335–359，2008年。2[3] Zixiang Chen、Yihe Deng、Huizhuo Yuan、Kaixuan Ji 和 Quanquan Gu。自博弈微调可将弱语言模型转化为强语言模型。载于《第41届国际机器学习会议论文集》。JMLR.org，2024。2[4] Zhanpeng Chen、Zhihong Zhu、Xianwei Zhuang、Zhiqi Huang 和 Yuexian Zou。采用反事实干预的双向解耦网络，用于多模态意图检测。载于《2024年自然语言处理经验方法会议论文集》，第17554–17567页，迈阿密，佛罗里达州，美国，2024。计算语言学协会。1, 2[5] Zesen Cheng、Sicong Leng、Hang Zhang、Yifei Xin、Xin Li、Guanzheng Chen、Yongxin Zhu、Wenqi Zhang、Ziyang Luo、Deli Zhao 等。VideoLLaMA 2：提升视频大语言模型中的时空建模与音频理解。arXiv预印本 arXiv:2406.07476，2024。1, 5[6] Yihe Deng、Pan Lu、Fan Yin、Ziniu Hu、Sheng Shen、Quanquan Gu、James Zou、Kai-Wei Chang 和 Wei Wang。通过在图像理解任务上进行自我训练，增强大型视觉语言模型。载于《第三十八届神经信息处理系统年会》，2024。2[7] Hanze Dong、Wei Xiong、Bo Pang、Haoxiang Wang、Han Zhao、Yingbo Zhou、Nan Jiang、Doyen Sahoo、Caiming Xiong 和 Tong Zhang。RLHF工作流程：从奖励建模到在线RLHF。《机器学习研究汇刊》，2024。2[8] Qian Dong、Yuezhou Dong、Ke Qin、Guiduo Duan 和 Tao He。无偏多模态音频到意图识别。载于《ICASSP 2025-2025 IEEE国际声学、语音与信号处理会议（ICASSP）》，第1–5页。IEEE，2025。2[9] Zi-Yi Dou、Cheng-Fu Yang、Xueqing Wu、Kai-Wei Chang 和 Nanyun Peng。Re-ReST：面向语言智能体的反思强化自我训练。载于《2024年自然语言处理实证方法会议，页码 15394–15411，美国佛罗里达州迈阿密，2024。计算语言学协会。2[10] Yasunori Endo 和 Sadaaki Miyamoto。球面 k-均值++聚类。载于《人工智能决策建模：第12届国际会议 MDAI 2015》，瑞典舍夫德，2015年9月21日至23日，论文集 12，页码 103–114。施普林格，2015。2, 3[11] Marco Federici、Anjan Dutta、Patrick Forr´e、Nate Kushman 和 Zeynep Akata。通过多视图信息瓶颈学习鲁棒表征。载于《国际学习表征会议》，2020。4[12] Devamanyu Hazarika、Roger Zimmermann 和 Soujanya Poria。MISA：用于多模态情感分析的模态不变及特定表征。载于《第28届 ACM 国际多媒体会议论文集》，页码 1122–1131，2020。1, 2, 5[13] Tao He、Hao Li、Jingchang Chen、Runxuan Liu、Yixin Cao、Lizi Liao、Zihao Zheng、Zheng Chu、Jiafeng Liang、Ming Liu 和 Bing Qin。打破推理障碍：从自我进化的视角综述大语言模型复杂推理。载于《计算语言学协会论文集：ACL 2025》，页码 7377–7417，奥地利维也纳，2025。计算语言学协会。2[14] Arian Hosseini、Xingdi Yuan、Nikolay Malkin、Aaron Courville、Alessandro Sordoni 和 Rishabh Agarwal。V-star：训练自我训练推理器的验证器。arXiv 预印本 arXiv:2402.06457，2024。1[15] Bo Hu、Kai Zhang、Yanghai Zhang 和 Yuyang Ye。自适应多模态融合：面向意图识别的动态注意力分配。载于《AAAI人工智能会议论文集》，页码 17267–17275，2025。1, 2, 5[16] Edward J Hu、yelong shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。低秩适应：大语言模型的低秩适应。载于《国际学习表征会议》，2022。5[17] Jiaxin Huang、Shixiang Gu、Le Hou、Yuexin Wu、Xuezhi Wang、Hongkun Yu 和 Jiawei Han。大语言模型能够自我改进。载于《2023年自然语言处理实证方法会议论文集》，页码 1051–1068，新加坡，2023。计算语言学协会。2[18] Shijue Huang、Libo Qin、Bingbing Wang、Geng Tu 和 Ruifeng Xu。SDIF-DA：一种采用数据增强的由浅入深交互框架，用于多模态意图检测。载于《ICASSP 2024—2024年IEEE国际声学、语音与信号处理会议》，页码 10206–10210。IEEE，2024。1, 2, 5, 7[19] Sepideh Kaffash、An Truong Nguyen 和 Joe Zhu。智能交通系统中的大数据算法及其应用：综述与文献计量分析。《国际生产经济学杂志》，231:107868，2021。1[20] Junnan Li、Dongxu Li、Silvio Savarese 和 Steven Hoi。BLIP-2：利用冻结图像编码器和大语言模型进行自举式语言-图像预训练。载于《国际机器学习会议》，页码 19730–19742。PMLR，2023。1

[21] Haotian Liu、Chunyuan Li、Yuheng Li、Bo Li、Yuanhan Zhang、Sheng Shen 和 Yong Jae Lee。LLaVA-NeXT：改进推理、光学字符识别和世界知识，2024。1, 5[22] Zihan Liu、Shun Zheng、Xumeng Wen、Yang Wang、Jiang Bian 和 Mao Yang。深度自我演化推理，2025 2 。[23] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。arXiv 预印本 arXiv:1711.05101，2017。5[24] Jinpeng Mi、Song Tang、Zhen Deng、Michael Goerner 和 Jianwei Zhang。基于物体可供性的多模态融合，实现自然的人机交互。《认知系统研究》，54:128–137，2019。1[25] Jong Hak Moon、Hyungyung Lee、Woncheol Shin、Young-Hak Kim 和 Edward Choi。通过视觉语言预训练实现医学图像与文本的多模态理解和生成。《IEEE生物医学与健康信息学杂志》，26(12):6070–6080，2022。1[26] Sheuli Paul、Michael Sintek、Veton K¨epuska、Marius Silaghi 和 Liam Robertson。面向装配场景人机通信的基于意图的多模态语音与手势融合。载于《2022年第21届IEEE机器学习与应用国际会议（ICMLA）》，页码 760–763。IEEE，2022。1[27] Soujanya Poria、Devamanyu Hazarika、Navonil Majumder、Gautam Naik、Erik Cambria 和 Rada Mihalcea。MELD：用于对话情绪识别的多模态多方数据集。载于《第57届计算语言学协会年会论文集》，页码 527–536，2019。2[28] Wasifur Rahman、Md Kamrul Hasan、Sangwu Lee、Amir Zadeh、Chengfeng Mao、Louis-Philippe Morency 和 Ehsan Hoque。在大型预训练 Transformer 中整合多模态信息。载于会议论文集。计算语言学协会年会，页码 2359，2020 1, 2, 5。[29] Bishwajit Saha、Dmitry Krotov、Mohammed J Zaki 和 Parikshit Ram。利用联想记忆实现端到端可微聚类。载于《国际机器学习会议》，页码 29649–29670。PMLR，2023。3[30] Tulika Saha、Aditya Patra、Sriparna Saha 和 Pushpak Bhattacharyya。迈向情绪辅助的多模态对话行为分类。载于《第58届计算语言学协会年会论文集》，页码 4361–4372，2020。5[31] Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Xiao Bi、Haowei Zhang、Mingchuan Zhang、YK Li、Y Wu 等。DeepSeekMath：推动开放语言模型数学推理能力的极限。arXiv 预印本 arXiv:2402.03300，2024。1[32] Gaurav Srivastava、Zhenyu Bi、Meng Lu 和 Xuan Wang。辩论、训练、演化：语言模型推理的自我进化。载于《2025年自然语言处理实证方法会议论文集》，页码 32752–32798，中国苏州，2025。计算语言学协会。[33] Zhiqing Sun、Longhui Yu、Yikang Shen、Weiyang Liu、Yiming Yang、Sean Welleck 和 Chuang Gan。由易到难泛化：超越人工监督的可扩展对齐。arXiv 预印本 arXiv:2403.09472，2024。1[34] Wentao Tan、Qiong Cao、Yibing Zhan、Chao Xue 和 Changxing Ding。超越人类数据：通过迭代自我进化对齐多模态大语言模型。arXiv预印本 arXiv:2412.15650，2024。2[35] Zhengwei Tao、Ting-En Lin、Xiancai Chen、Hangyu Li、Yuchuan Wu、Yongbin Li、Zhi Jin、Fei Huang、Dacheng Tao 和 Jingren Zhou。大语言模型自我进化综述。arXiv预印本 arXiv:2404.14387，2024。2[36] Abhisek Tiwari、Manisimha Manthena、Sriparna Saha、Pushpak Bhattacharyya、Minakshi Dhar 和 Sarbajeet Tiwari。Dr. can see：迈向多模态疾病诊断虚拟助手。发表于第31届ACM国际信息与知识管理会议论文集，第1935–1944页，2022。1[37] Yao-Hung Hubert Tsai、Shaojie Bai、Paul Pu Liang、J Zico Kolter、Louis-Philippe Morency 和 Ruslan Salakhutdinov。用于未对齐多模态语言序列的多模态 Transformer。发表于计算语言学协会会议论文集。计算语言学协会年会，第6558页，2019。1、2、5[38] Kai Wang、Shiguo Lian、Haiyan Sang、Wen Liu、Zhaoxiang Liu、Fuyuan Shi、Hui Deng、Zeming Sun 和 Zezhou Chen。面向虚拟人自然交互的多模态活动检测。发表于2023年IEEE虚拟现实与3D用户界面会议摘要与研讨会（VRW），第671–672页。IEEE，2023。1[39] Peng Wang、Shuai Bai、Sinan Tan、Shijie Wang、Zhihao Fan、Jinze Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge 等。Qwen2-VL：增强视觉语言模型在任意分辨率下对世界的感知。arXiv预印本 arXiv:2409.12191，2024。1、2、5[40] Xiyao Wang、Ji uhai Chen、Zhaoyang Wang、Yuhang Zhou、Yiyang Zhou、Huaxiu Yao、Tianyi Zhou、Tom Goldstein、Parminder Bhatia、Furong Huang 等。通过自我改进增强大型视觉语言模型中的视觉-语言模态对齐。arXiv预印本 arXiv:2405 15973，2024 2 . .[41] Yizhong Wang、Yeganeh Kordi、Swaroop Mishra、Alisa Liu、Noah A. Smith、Daniel Khashabi 和 Hannaneh Hajishirzi。Self-Instruct：通过自生成指令对齐语言模型。发表于计算语言学协会第61届年会论文集（第1卷：长篇论文），第13484–13508页，加拿大多伦多，2023。计算语言学协会。2[42] Suzhen Wu、Chunfeng Du、Haijun Li、Hong Jiang、Zhirong Shen 和 Bo Mao。CAGC：一种面向超低延迟闪存固态硬盘的内容感知垃圾回收方案。发表于2021年IEEE国际并行与分布式处理研讨会（IPDPS），第162–171页。IEEE，2021。1、2[43] Tianyu Yu、Haoye Zhang、Yuan Yao、Yunkai Dang、Da Chen、Xiaoman Lu、Ganqu Cui、Taiwen He、Zhiyuan Liu、Tat-Seng Chua 等。RLAIF-V：通过开源AI反馈对齐多模态大语言模型，以提升GPT-4V的可信度。arXiv预印本 arXiv:2405.17220，2024。2[44] Wenmeng Yu、Hua Xu、Fanyang Meng、Yilin Zhu、Yixiao Ma、Jiele Wu、Jiyun Zou 和 Kaicheng Yang。CH-SIMS：一个带有模态细粒度标注的中文多模态情感分析数据集。发表于第58届

计算语言学协会年会，页码3718–3727，2020。2[45] Amir Zadeh、Rowan Zellers、Eli Pincus 和 Louis-Philippe Morency。MOSI：在线观点视频中情感强度与主观性分析的多模态语料库。arXiv预印本 arXiv:1606.06259，2016。[46] AmirAli Bagher Zadeh、Paul Pu Liang、Soujanya Poria、Erik Cambria 和 Louis-Philippe Morency。真实场景中的多模态语言分析：CMU-MOSEI数据集与可解释动态融合图。发表于计算语言学协会第56届年会论文集（第1卷：长篇论文），第2236–2246页，2018。2[47] Hanlei Zhang、Hua Xu、Xin Wang、Qianrui Zhou、Shaojie Zhao 和 Jiayan Teng。MIntRec：一个用于多模态意图识别的新数据集。发表于第30届ACM国际多媒体会议论文集，第1688–1697页。ACM，2022。1、2、5、8[48] Hanlei Zhang、Xin Wang、Hua Xu、Qianrui Zhou、Kai Gao、Jianhua Su、jinyue Zhao、Wenrui Li 和 Yanting Chen。MIntRec2.0：用于对话中多模态意图识别和范围外检测的大规模基准数据集。发表于第十二届国际学习表征会议，2024。1、2、5[49] Hanlei Zhang、Qianrui Zhou、Hua Xu、Jianhua Su、Roberto Evans 和 Kai Gao。面向多模态意图理解的多模态分类与分布外检测。IEEE多媒体汇刊，2025。1、2、5[50] Yaowei Zheng、Richong Zhang、Junhao Zhang、Yanhan Ye、Zheyan Luo、Zhangchi Feng 和 Yongqiang Ma。Llamafactory：统一且高效的 100+ 语言模型微调。arXiv预印本 arXiv:2403.13372，2024。5[51] Qianrui Zhou、Hua Xu、Hao Li、Hanlei Zhang、Xiaohan Zhang、Yifan Wang 和 Kai Gao。通过模态感知提示进行词元级对比学习，以实现多模态意图识别。发表于AAAI人工智能会议论文集，第17114–17122页，2024。1、2、5、7[52] Qianrui Zhou、Hua Xu、Yifan Wang、Xinzhi Dong 和 Han- lei Zhang。大语言模型引导的语义关系推理，用于多模态意图识别。发表于2025年自然语言处理实证方法会议论文集，第22221–22237页，中国苏州，2025。计算语言学协会。2[53] Deyao Zhu、Jun Chen、Xiaoqian Shen、Xiang Li 和 Mohamed Elhoseiny。MiniGPT-4：利用先进的大语言模型增强视觉语言理解。arXiv预印本 arXiv:2304.10592，2023。1[54] Ke Zhu、Liang Zhao、Zheng Ge 和 Xiangyu Zhang。自监督视觉偏好对齐。发表于第32届ACM国际多媒体会议论文集，第291–300页，2024。2[55] Zhihong Zhu、Xuxin Cheng、Zhaorun Chen、Yuyan Chen、Yunyan Zhang、Xian Wu、Yefeng Zheng 和 Bowen Xing。InMu-Net：通过信息瓶颈和多感官处理推进多模态意图检测。发表于第32届ACM国际多媒体会议论文集，第515–524页，2024。1、2