

结合少样本逻辑推理的多模态机器人故障解释一致性方法

翁 华^{1,2}, 韩宇辰¹, 张旭龙², 刘发贵¹, 倪晓俊³, 田晖⁴, 吴建汉², 王健宗²

1. 华南理工大学计算机科学与工程学院, 广州 广东 510006;

2. 平安科技(深圳)有限公司, 深圳 广东 518063;

3. 中国科学院计算技术研究所, 海淀 北京 100190;

4. 华侨大学, 厦门 福建 361021

摘要

针对多模态机器人故障解释中模态间逻辑不一致的问题, 本文提出一种结合少样本逻辑推理 (few-shot logical reasoning) 的解释一致性方法。该方法由两个核心智能体组成: 逻辑判断智能体与故障解释再生智能体。前者通过少样本学习结合预训练大模型, 对不同模态间的逻辑关系 (支持、矛盾、无关) 进行两两推理与权重融合, 实现高效的逻辑一致性判断; 后者基于逻辑判断结果与推理链条, 生成逻辑一致且语义更连贯的多模态故障解释。实验结果表明, 本文方法在 RoboC 数据集上的逻辑判断准确率较现有最优方法提升 2.77%, 故障解释再生准确率提升 19.88%, 显著提高了解释的一致性与可解释性。该研究有效降低了模型的标注依赖与训练成本, 为提升机器人系统的透明性与用户信任度提供了新思路。

关键词

具身智能; 多模态故障解释; 逻辑判断; 少样本推理

中图分类号: TP39

文献标志码: A

doi:10.11959/j.issn.2096-0271.2026060

A Consistency Method for Multimodal Robot Fault Explanation via Few-shot Logical Reasoning

WENG Qihua^{1,2}, HAN Yuchen¹, ZHANG Xulong², LIU Fagui¹, NI Xiaojun³, TIAN Hui⁴, WU Jianhan², WANG Jianzong²

1. South China University of Technology, School of Computer Science and Engineering, Guangzhou 510006, China;

2. Ping An Technology (Shenzhen) Co., Ltd., Shenzhen 518063, China;

3. Institute of Computing Technology, Chinese Academy of Sciences (ICT), Haidian, Beijing, China 100190;

4. Huaqiao University, Xiamen, Fujian, China 361021

Abstract

To address the logical inconsistency among modalities in multimodal robot fault explanations, this paper proposes a consistency method based on few-shot logical reasoning. The proposed framework consists of two core agents: a logical judgment agent and a fault explanation regeneration agent. The former employs few-shot learning with pre-trained models to infer pairwise logical relations (support, contradiction, or irrelevance) between modalities and integrates

them through scenario-dependent weighting. The latter utilizes both the logical reasoning process and the inferred relations to regenerate logically consistent and semantically coherent multimodal fault explanations. Experimental results on the RoboC dataset demonstrate that the proposed method improves logical judgment accuracy by 2.77% and regeneration accuracy by 19.88% over state-of-the-art approaches. This method enhances the consistency, interpretability, and efficiency of robot fault explanations, providing a new direction for trustworthy human-robot interaction.

Key words

embodied intelligence, multimodal fault explanation, logical reasoning, few-shot learning

0 引言

随着智能机器人的不断发展，并逐步渗透至工业制造、家庭服务和医疗康复等关键领域，其可靠性与可信度已成为人机交互研究的核心议题^{[1][2]}。随着任务复杂度不断提高，机器人发生故障几乎不可避免。对这些故障进行解释，是用户与机器人交互中的重要环节。有效的故障解释不仅能帮助用户理解机器人的能力局限，还能增强机器人决策的可解释性，并显著提升用户对其的信任度^[3]。研究表明^{[4][5]}，用户在面对机器人故障时常常产生认知偏差，导致任务执行中断，甚至在某些情况下放弃继续使用系统。这凸显了故障解释与修复机制在人机协作中的重要价值。

在现有研究中^{[6][7]}，机器人故障解释主要依赖文本模态来阐述故障原因。尽管文本型解释能够提供基本的分析信息，但其存在多模态信息割裂和技术参数模糊化等局限^{[8][9]}。已有研究表明，多模态解释在处理复杂信息时更具直观性和效率，这进一步强调了开展机器人故障多模态解释研究的必要性^{[10][11]}。然而，在现有多模态故障解释研究中^{[7][12]}，多模态信息之间仍存在逻辑不一致的问题。图1展示了两种典型的多模态信息不一致案例。

在图1(a)中，故障原因在于世界模型中咖啡机只能容纳一个杯子，但机器人却判断可以放置两个杯子，因而在放置操作时失败。然而，系统却将故障原因判定为缺少放置动作，这与机器人参数信息相互矛盾。在图1(b)中，系统认为故障原因是放置动作错误，导致杯子放置失败，但图像信息和参数信息均显示机器人已成功完成放置操作，说明多模态信息之间存在不一致。通过我们的方法，可以识别出多模态信息间的不一致，并结合历史元组信息发现，真正的故障原因是机器人在先前的清洗操作中未正确打开水龙头，从而生成逻辑一致的多模态故障解释。

然而，尽管多模态故障解释中模态间逻辑不一致的情况较为常见，但针对这一问题的系统性研究仍然十分匮乏^[12]。

在最新的研究中^[12]，多模态故障解释中的模态间逻辑关系被划分为三类：互相支持、无直接关系和相互矛盾。研究者通过构建自定义数据集并对分类模型进行微调，实现了对多模态信息逻辑关系的判断。然而，该研究仍存在一定局限性：首先，训练数据集的构建往往需要耗费大量时间和资金成本，同时模型的泛化能力有限；其次，该研究仅将多模态信息的逻辑关系简化为语义逻辑分类问题，将其划分为三类并进行判断。但在获得判断结果后，并未提出进一步的解决方案来生成逻辑一致的故障解释。这不仅增加了用户理解故障

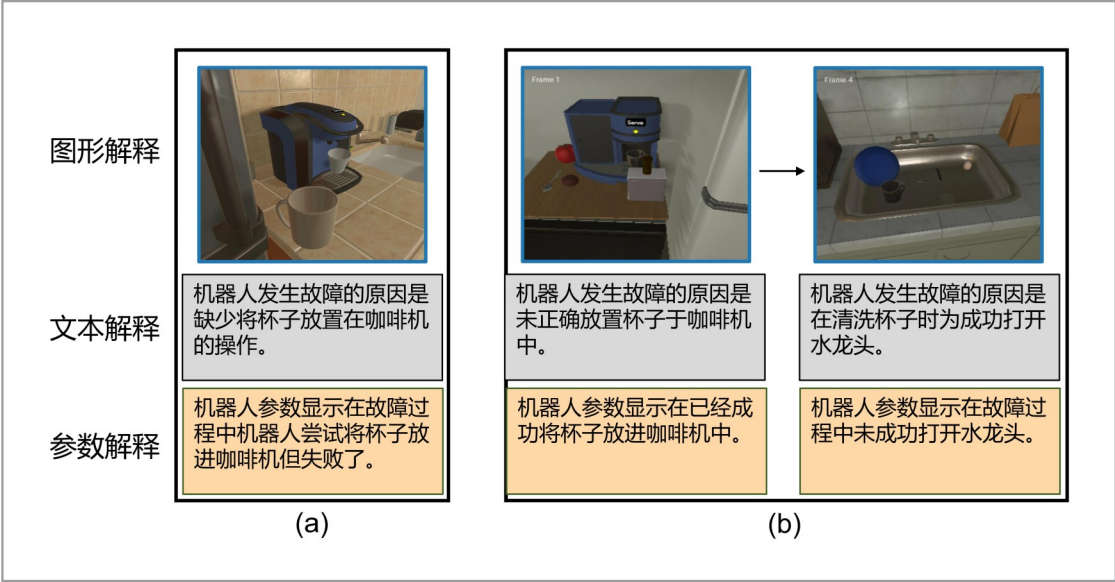


图1 逻辑不一致的多模态故障解释案例

信息的负担，甚至可能造成误导。

为提升多模态机器人故障解释的泛化能力并降低构建成本，同时生成逻辑一致的 解释结果，本文在现有研究基础上^{[7][12][13]}提出了一种结合少样本逻辑推理的多模态机器人故障解释一致性方法。

该方法由两个核心智能体组成：逻辑判断智能体与故障解释再生成智能体。首先，逻辑判断智能体对机器人系统生成的故障解释以及采集到的多模态信息进行分析，输出模态间的逻辑关系（如支持、无关或矛盾），并提供详细的推理过程。和基线^[12]对比，我们的方法具有明显的少样本优势：得益于预训练大模型的强泛化能力，以及我们在提示词中嵌入的逻辑推理示例和结构化的机器人参数先验，逻辑判断过程不再依赖大规模标注数据，而是能够在不到 1% 的数据支持下实现有效学习和迁移。随后，故障解释再生成智能体综合多模态信息、逻辑关系及其推理过程，生成逻辑一致且更具准确性的故障解释。在这一过程中，逻辑判断智能体不仅提供了逻辑

辑关系的判定结果，还增强了推理链条的透明性与可解释性，从而提升用户对系统的信任，并有效支撑故障解释再生成智能体的优化输出。

我们将本文的主要贡献总结如下：

- 1、提出少样本逻辑判断方法：结合机器人参数信息，实现对机器人故障解释中多模态信息间逻辑关系的高效判断，显著降低了对大规模标注数据的依赖；
- 2、引入推理过程指导的解释生成机制：利用逻辑判断的推理链条指导故障解释再生成，从而获得逻辑一致且准确率更高的解释结果；
- 3、验证方法有效性与先进性：通过系统化实验评估，证明了所提方法的有效性，并在多项指标上取得了具有竞争力的最优性能。

1 相关工作

1.1 多模态机器人故障解释

机器人故障解释是 HRI (Human-Robot Interaction, 人机交互) 领域长期关注的重要研究任务, 其核心目标在于增强人类对机器人系统的信任^{[14][15]}, 并帮助非专业用户在故障场景下更有效地协助机器人^{[6][16]}。已有研究表明^{[17][18]}, 当机器人能够清晰地解释其任务失败原因时, 可有效降低用户的认知负担并改善人机协作体验。近期的研究^{[19][20][21][22]}进一步指出, 相较于单一文本模态, 多模态解释(如结合语言、视觉与内部参数信息)能够更直观地呈现故障机理, 从而增强解释的透明性与可理解性。这些研究为多模态解释的研究奠定了基础, 但对于不同模态间逻辑一致性问题的系统性分析仍较为不足。

1.2 多模态信息的逻辑一致性

多模态信息处理必须确保不同数据源之间的逻辑一致性, 才能实现准确的综合判断。近年来, 大量研究围绕多模态融合中的一致性问题展开。早期研究^{[23][24][25][26]}主要聚焦于跨模态语义匹配与对齐问题, 如图文连贯性建模、联合嵌入表示与跨模互信息优化等。事实核查^[27]则和多模态姿态检测^[28]一样, 将一致性评估建模为一个语义识别问题。然而, 除了研究^[12]以外, 现有的分类与预测方法大多面向图像字幕生成或多模态信息检索, 难以直接应用于机器人故障多模态解释的相干性建模。

研究^[12]尝试将机器人故障解释中的文本解释与图形解释归类为三类语义关系问题, 并通过自建数据集对模型进行微调, 从而实现逻辑一致性判断。然而, 该方法仍存在两方面局限。首先, 模型仅输出模态关系的分类结果, 缺乏推理链条与中间因果信息, 难以支持后续的解释优化; 同时, 模型训练依赖大量标注样本, 泛化能

力有限, 难以适应多场景任务。

针对上述问题, 本文提出的少样本逻辑一致性方法在以下方面进行了改进:

本文利用预训练大模型的强泛化能力, 仅依赖少量示例即可完成模态逻辑判断, 显著降低数据标注与训练成本。同时, 本文不仅给出模态关系结果, 还提供完整的逻辑推理链条, 为后续再生成提供因果依据。此外, 本文在文本、视觉与参数三类模态间引入权重分配机制, 实现不同场景下的自适应逻辑一致性判断。

1.3 多模态信息的逻辑判断

逻辑判断在机器人任务执行与故障分析中发挥着至关重要的作用。现有研究虽已探讨了将大型语言模型(LLMs)融入规划与控制系统以提升推理能力^{[29][30]}, 但在多模态信息间逻辑关系建模与可解释性推理方面仍显不足。在整合多源信息以进行故障解释时, 引入逻辑判断机制是确保解释结果合理性与可验证性的关键^{[25][26]}。然而, 在机器人与LLMs结合的场景下, 相关研究依旧十分匮乏。据我们所知, 研究^[12]率先提出了对故障解释中的文本解释和场景图信息进行连贯性分析, 将其处理为分类问题。但是将其处理为分类问题导致缺少判断过程的细节信息, 难以进行下一步的故障解释优化。Reflect^[7]作为当前性能最优的故障解释生成方法, 提出了基于任务经验总结的文本生成机制, 但其生成过程未显式建模多模态解释间的逻辑一致性, 易导致解释语义割裂或推理冲突。

为克服上述不足, 本文提出了一种基于双智能体协同机制的多模态故障解释方法。具体而言: 逻辑判断智能体通过少样本推理实现不同模态间的逻辑关系分析, 并输出推理依据与结果; 故障解释再生成

智能体则基于逻辑判断输出的推理链条，对不一致或矛盾信息进行优化重构，从而生成逻辑一致的多模态故障解释。这种“判断-再生成”的闭环架构不仅能够识别模态间冲突，还能实现逻辑一致性的主动修复，使得模型在逻辑推理深度、一致性生成机制与可解释性等方面均超越 Coherent^[12]与 Reflect^[7]。

2 研究方法

针对机器人故障解释中多模态信息存在逻辑不一致的问题，本文提出了一种新方法，用于生成逻辑一致且准确率更高的多模态故障解释，从而帮助用户更好地理解机器人的能力局限，并协助其应对故障情境。该方法由两个核心智能体协同完成：逻辑判断智能体与解释再生成智能体。整体实现流程如图2所示。

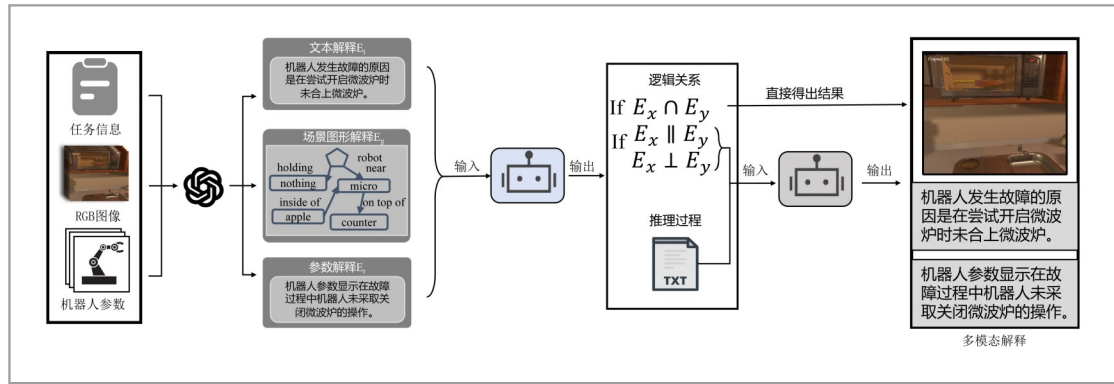


图2 方法流程图

2.1 逻辑判断

在机器人执行高级任务中的离散时间步长 $T = \{T_1, T_2, \dots, T_i\}$ 上，若在步长上检测到故障，我们的目的是生成一个关于该故障的多模态解释 E 。其中的 E 由系统生成的文本解释 E_t 、场景图形解释 E_g 和机器人内部参数解释 E_r 叠加组成。

在本论文中，我们参考 Reflect^[7]生成故障解释的思路来生成 E_t 。具体来说，我们将机器人高级任务的描述、机器人执行动作后的视觉观察描述以及机器人执行的子任务列表作为一个元组，将其作为子任务执行摘要嵌入到提示词模板中，用于提

示一个大型语言模型并生成故障的文本解释。我们使用了与 Reflect 相同的提示词模板。对于场景图形解释 E_g ，我们将故障发生时刻中记录的场景图像表示为一个场景图。我们构建一个场景图 $G_t = \{N + \{n_{robot}\}, E\}$ ，它描述了对对象节点 (N) 及其空间关系 (E)。每个节点定义为 $n_i = (c_i, s_i)$ ，其中 c_i 是对象类， s_i 是对象状态（如果存在的话）。每条边都包含两个对象之间的空间关系。为了使生成的解释更加简洁，并减少计算内存占用，我们对场景图中的对象进行过滤，只保留与任务相关的对象。对于机器人内部参数解释 E_r ，我们将故障时刻正在执行的子任务期间记录的每个步长

的机器人内部参数作为一个元组，组合为一个元组列表，嵌入到提示词模板中，用于分析子任务期间的机器人状态变化。然后将其输入到大型语言模型中，得到机器人状态变化体现的故障解释。

由于不同模态的故障解释之间存在着彼此之间逻辑关系不一致的情况，因此在得到不同模态的故障解释后，我们需要对不同模态的故障解释 $\{E_t, E_g, E_r\}$ 进行逻辑关系的判断。

为了量化故障解释的逻辑关系，与 Coherent 相似，我们提出将关系分为三个类别：互相支持 $\cap$ 、互相矛盾 $\perp$ 、无关联 $\parallel$ 。首先，我们需要对其进行两两的语义逻辑判断得到逻辑关系 C_{xy} ：

$$C_{xy} = \{E_x \cap E_y, E_x \perp E_y, E_x \parallel E_y\}, E_x, E_y \in \{E_t, E_g, E_r\} \quad (1)$$

其中 $E_x \cap E_y$ 代表着在逻辑关系上互相支持， $E_x \perp E_y$ 代表着在逻辑关系上没有直接关联关系， $E_x \parallel E_y$ 代表着在逻辑关系上互相矛盾。

然后在进行综合逻辑判断时，我们给各个逻辑关系分配不同的权重，从而得到综合的逻辑关系。

$$C = xC_{gt} + yC_{gr} + zC_{tr}, x + y + z = 1 \quad (2)$$

具体实现过程如图 2 所示。在对不同模态的故障解释进行逻辑关系判断之前，我们首先为大型语言模型编制了三个不同类型的逻辑关系判断示例。上下文示例是 LLMs 识别任务特定信息的重要来源。不同的示例可以为当前任务提供不同的信息。接着，我们将三个不同模态的故障解释嵌入到逻辑关系判断的提示词当中，要求大型语言模型按照示例中的逻辑关系判断过程进行思考，并输出详细的判断过程以及最终的判断结果。

在进行综合逻辑关系判断时，我们认为 C_{gt}, C_{gr}, C_{tr} 的权重不应该是平均的，而是应该根据不同的任务执行环境不同的机器人能力来调节权重占比。例如对于视觉能力突出的机器人，应该增加 C_{gr} 的权重占比；对于在黑暗环境下执行任务的机器人，应当增加 C_{tr} 的权重占比。

进一步来说，我们对多模态信息的逻辑一致判断方法可以延伸到更加多维的信息之间的判断，同样地先通过两两之间的判断得到初步的关系，然后在进行综合的权重分配得到综合的信息逻辑关系。我们将公式进一步改写如下：

$$C_{xy} = \{E_x \cap E_y, E_x \perp E_y, E_x \parallel E_y\}, E_x, E_y \in \{E_1, E_2, \dots, E_I\} \quad (3)$$

$$C = \sum_{i=1}^I w_i C_{xy}, \sum_{i=1}^I w_i = 1 \quad (4)$$

经过多次实验验证，实验使用的仿真平台中在系数之间的差值不超过 0.3 的区间范围内，权重的设置对实验的影响不显著，所以为了简化实验，我们将权重设置如下所示：

$$C = 0.5C_{gt} + 0.3C_{gr} + 0.2C_{tr} \quad (5)$$

2.2 故障解释的再生成

在获得不同模态之间的逻辑关系后，本文进一步针对逻辑关系为无直接联系或相互矛盾的故障解释进行再生成，因为此类解释往往无法全面、准确地揭示故障发生的真实原因。

为确保生成的故障解释在逻辑上的一致性，我们将逻辑判断过程显式嵌入提示词中，使 LLMs 能够直接利用各模态之间的逻辑关系信息。这一设计不仅有助于 LLMs 在历史元组中检索与其他模态保持一致的内容，还能引导其生成更加连贯、

互相支持的解释。基于此，我们提出了以下再生成策略：

1、部分一致场景：当三个模态中有两个模态信息保持一致时，以一致模态为参照，LLMs 从不一致模态的历史元组中自起点遍历，寻找能够支持前两者的证据，并据此生成逻辑一致的故障解释；

2、完全不一致场景：当三个模态之间均存在冲突时，以机器人参数这一结构化模态的故障解释为基准，优先在其余两个模态中检索出与其逻辑一致的元组信息，再据此生成解释。

最终，我们将经过逻辑优化的多模态解释结果进行整合与叠加，形成一个新的全局多模态故障解释，为用户提供更直观、更一致且可验证的说明。

与现有方法相比，该环节的创新性在于：不仅将逻辑判断结果直接引入生成过程中，显式增强了故障解释的逻辑一致性，还通过历史元组的动态检索与推理链约束，实现了多模态解释的自适应优化。这一机制有效降低了模态冲突带来的歧义风险提升了解释的准确性、透明性和研究参考价值。

3 实验

3.1 实验数据集

考虑到 RoboFail 数据集^[7]并不是一个为了研究多模态故障解释一致性的数据集，仅包含了约 11% 的逻辑不一致案例。为了评估我们的方法，我们在 RoboFail 数据集的基础上进行扩展，从而构建新的数据集。

RoboFail 数据集包含十个任务类别，每类包含十个任务，共计 100 个任务执行数据。基于 AI2THOR 仿真平台^[31]，这是一个用于具身智能研究的交互式三维仿真

平台，支持超过 100 种可交互物体的操作（如开关、拾取、放置、移动等），并可以在其中获取多模态感知信息，包括 RGB 图像、深度图、语义分割图和物体检测信息，从而便于开展任务规划、视觉导航、物体操作以及多模态感知等研究。我们对数据集中的任务进行了仿真，并记录了任务执行过程中机器人的视觉图像、动作指令序列以及机器人参数信息。在此基础上，我们通过修改场景图中物体的状态、调整故障发生的时间步以及更新视觉观察与机器人参数，生成新的故障样本数据集。参考研究^[12]中数据集的构建方法，我们在十个任务类别中随机注入故障，并对一组或两组视觉观察或计划元组进行修改，以生成反事实任务执行样本，最终我们收集到 570 个样本。这些样本专门用于多模态逻辑判断实验，从而评估方法在逻辑一致性分析和故障解释生成上的有效性。

3.2 实验评估指标和实验设置

本研究旨在对多模态故障解释中存在的逻辑不一致进行判定，并生成逻辑一致的多模态故障解释。基于此，实验主要采用两项指标进行评估：第一项指标是逻辑判断准确率，它衡量模型对于生成的不同模态解释之间的逻辑关系判断准确率，而不管它们是否正确地指出故障原因。第二项指标是故障解释再生成准确率，它衡量模型所重新生成的多模态解释是否正确地指出了故障原因，并且保持模态之间的逻辑一致。对于这一项指标的评估，我们邀请了三位人类专家来进行评估，如果两位及以上专家认同重新生成的多模态故障解释，则判定为正确。

在实验中，我们将所提方法与以下基线 and 消融设置进行对比：Coherent^[12]、

Reflect^[7]和 TEM-MLLM^[32]（原始的多模态大型语言模型）。在我们的方法中，我们制作了4个由人类专家进行标注的样本用于模型的学习。Coherent（研究^[12]中所提出的方法，在此处用Coherent指代）是第一个提出研究多模态故障解释的逻辑一致性的方法，它采用与我们的数据集类似的构建方法来基于RoboFail数据集构建自己的数据集，并用于微调使用多个NLI (natural language inference)数据集进行预训练的模型，从而对不同模态之间的故障解释进行逻辑分类。为了评估所提出的方法在重新生成故障解释的能力，我们还将我们的方法与目前性能最好的故障解释生成方法Reflect作对比。此外，我们还设置了TEM-MLLM，使用与我们所提出的方法同样的示例样本进行学习，完成逻辑关系的判断和重新生成故障解释。所有实验中所用的模型均为GPT-4o，以保证方法评估的公平性与一致性。

3.3 实验结果

从实验表1可以看出，在逻辑一致性

表1 逻辑判断准确率实验结果(%)			
模型	Coherent	TEM-MLLM	Ours
Boil water	57.85	47.82	60.86
Toast bread	61.07	43.93	62.86
Fry egg	60.95	47.30	65.1
Heat potato	58.34	49.27	60.10
Serve coffee	58.84	50.71	60.34
Store egg	61.75	51.57	65.75
Make salad	55.86	42.93	55.97
Water plant	56.67	43.67	58.67
Switch devices	58.17	49.82	62.84
Serve warm water	57.24	48.06	61.97
平均	58.67	47.51	61.44

表2 重新生成故障解释准确率实验结果(%)			
模型	Coherent	TEM-MLLM	Ours
Boil water	52.17	39.13	73.06
Toast bread	53.37	38.92	72.91
Fry egg	55.11	38.83	74.53
Heat potato	51.72	41.03	74.49
Serve coffee	51.86	42.87	73.66
Store egg	56.52	43.47	74.49
Make salad	49.24	37.18	68.59
Water plant	50.37	39.93	71.59
Switch devices	55.02	42.66	74.53
Serve warm water	56.62	38.21	72.91
平均	53.20	40.22	73.08

判断方面，相比于依赖大规模数据集微调的Coherent方法，我们的方法采用3个样例的少样本学习训练，并引入了机器人参数的结构化模态信息作为逻辑判断参考。实验结果显示，我们的方法在平均逻辑判断准确率上比Coherent提高了2.77%。这表明，相较于需要大量数据进行微调的训练方法，我们的方法不仅节省了数据收集的时间和成本，同时在判断准确性上也获得了提升，验证了方法的有效性。此外，相较于直接使用LLMs进行逻辑判断，我们的方法通过对不同模态的故障解释进行两两判断并加权计算，得到最终的模态一致性判断结果。在RoboC数据集中，我们的方法相较TEM-LLM的47.51%准确率提升了13.93%，进一步证明了方法不仅仅依赖于LLMs的模型能力，同时增强了逻辑判断的可靠性。

由于Coherent并未提供完整的逻辑一致多模态故障解释生成策略，本研究选取在故障解释生成上表现最好的Reflect方法作为对比基线。在实验中，我们将Coherent判定为“模态间相互矛盾”或“无直接联系”的任务案例，多模态信息作

为 Reflect 方法的输入，以生成新的故障解释。实验表 2 显示，在重新生成的故障解释中，相较于 Reflect 方法，我们的方法取得了 19.88% 的性能提升，验证了方法在处理逻辑不一致多模态故障解释上的有效性。

同时，我们还对被判断为“模态间相互矛盾”或“无直接联系”的任务案例，直接使用 TEM-LLM 进行故障解释再生成，而不利用逻辑判断智能体输出的推理过程。实验结果表明，我们的方法相较于 TEM-LLM 实现了 32.88% 的性能提升。这证明，将逻辑判断智能体的推理过程嵌入提示词的策略是有效的，不仅显著提升了故障解释再生成的准确率，也增强了方法决策的透明性和可解释性。

4 结束语

4.1 总结

本文针对多模态机器人故障解释中常见的模态间逻辑不一致问题，提出了一种结合少样本逻辑推理的解释一致性方法。通过引入逻辑判断智能体与再生成智能体的协同机制，我们能够对不同模态解释进行逻辑关系分析，并基于推理链条引导生成逻辑一致的故障解释。实验结果表明，该方法在逻辑判断准确率和故障解释生成准确率上均显著优于现有方法，验证了其有效性与先进性。与此同时，方法在提升解释透明性与用户信任度方面展现出良好的潜力。

4.2 局限性

本文方法主要基于受控实验环境下的数据进行验证，而在真实机器人应用中，

传感器噪声、环境动态变化以及任务干扰因素可能导致多模态输入间存在较大偏差，从而影响逻辑判断的稳定性与再生成解释的可靠性。未来可引入环境感知增强机制，结合时序一致性建模或跨时间模态追踪，以提升模型在动态环境中的适应能力。此外，当前模型在多模态数据均可用的情况下表现良好，但当部分模态（如视觉或参数数据）缺失时，逻辑判断会受到影响导致准确率下降。后续可考虑采用自适应模态补全机制，例如通过生成模型或跨模态推断在缺失模态上生成补充特征，从而缓解数据缺失对逻辑一致性判定的影响。

综上，本文方法在逻辑推理深度、少样本能力及一致性生成方面已取得较好的研究进展，但在真实复杂环境中仍需进一步提升鲁棒性与自适应能力。未来研究可重点探索基于因果推理的逻辑一致性建模与自监督语义约束生成框架，以实现更具通用性与解释透明度的多模态机器人故障解释系统。

参考文献：

[1] Massaguer Gómez G. Should we Trust Social Robots? Trust without Trustworthiness in Human-Robot Interaction[J]. Philosophy & Technology, 2025, 38 (1): 24.

[2] Furuya H, Battistel L, Datta Choudhary Z, et al. Difficulties in Perceiving and Understanding Robot Reliability Changes in a Sequential Binary Task [C]//Proceedings of the 2024 ACM Symposium on Spatial User Interaction. New York, NY, USA: Association for Computing Machinery, 2024: 1-11.

[3] Nettet B, Robb D A, Lopes J, et al. Transparency in HRI: Trust and Deci-

- sion Making in the Face of Robot Errors [C]//Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. Boulder CO USA: ACM, 2021: 313–317.
- [4] Liu D, Li C, Zhang J, et al. Robot service failure and recovery: Literature review and future directions[J]. International Journal of Advanced Robotic Systems, 2023, 20(4): 17298806231191606.
- [5] Liu D, Zhang J, Qi J. Exploring customers' reuse intention to robots under different service failures: A mind perception perspective[J]. Acta Psychologica, 2025, 256: 105030.
- [6] Das D, Banerjee S, Chernova S. Explainable AI for robot failures: Generating explanations that improve user assistance in fault recovery[C]//Bethel C L, Paiva A, Broadbent E, et al. ACM/IEEE international conference on human-robot interaction, HRI 2021, boulder, CO, USA, march 8–11, 2021. ACM, 2021: 351–360.
- [7] Liu Z, Bahety A, Song S. REFLECT: Summarizing Robot Experiences for Failure Explanation and Correction[C]//7th Annual Conference on Robot Learning. 2023.
- [8] Park D H, Hendricks L A, Akata Z, et al. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. [C]//2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018. 2018: 8779–8788.
- [9] Chen S, Zhao Q. REX: Reasoning-aware and Grounded Explanation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 15565–15574.
- [10] Anjomshoe S, Najjar A, Calvaresi D, et al. Explainable Agents and Robots: Results from a Systematic Literature Review.[C]//Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13–17, 2019. 2019: 1078–1088.
- [11] Zhu H, Yu C, Cangelosi A. Affective Human-Robot Interaction with Multimodal Explanations. [C]//Social Robotics – 14th International Conference, ICSR 2022, Florence, Italy, December 13–16, 2022, Proceedings, Part I. 2022: 241–252.
- [12] Pramanick P, Rossi S. Multimodal coherent explanation generation of robot failures[C]//IEEE/RSJ international conference on intelligent robots and systems, IROS 2024, abu dhabi, united arab emirates, october 14–18, 2024. IEEE, 2024: 2487–2493.
- [13] Wang Z, Liang B, Dhat V, et al. I can tell what I am doing: Toward real-world natural language grounding of robot experiences[C]//Agrawal P, Kroemer O, Burgard W. Conference on robot learning, 6–9 november 2024, munich, germany: 卷 270. PMLR, 2024: 1863–1890.
- [14] Khanna P, Yadollahi E, Björkman M, et al. User Study Exploring the Role of Explanation of Failures by Robots in Human Robot Collaboration Tasks. [J]. CoRR, 2023, abs/2303.16010.
- [15] Ye S, Neville G, Schrum M, et al. Human Trust After Robot Mistakes: Study of the Effects of Different Forms of Robot Communication. [C]//28th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2019, New Delhi, India, October 14–18, 2019. 2019: 1–7.
- [16] Das D, Chernova S. Semantic-Based Explainable AI: Leveraging Semantic

- Scene Graphs and Pairwise Ranking to Explain Robot Failures. [C]//IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2021, Prague, Czech Republic, September 27 – Oct. 1, 2021. 2021: 3034–3041.
- [17] Apraiz, Ainhua, et al. The user experience in industrial human-robot interaction: A comparative analysis of Unimodal and Multimodal interfaces for disassembly tasks. *Robotics and Computer-Integrated Manufacturing* 95 (2025): 103045.
- [18] Sobrí n-Hidalgo D, Santamarta M G, Guerrero-Higueras M, et al. Explaining Autonomy: Enhancing Human-Robot Interaction through Explanation Generation with Large Language Models. [J]. *CoRR*, 2024, abs/2402.04206.
- [19] Alipour K, Schulze J P, Yao Y, et al. A Study on Multimodal and Interactive Explanations for Visual Question Answering. [C]//Proceedings of the Workshop on Artificial Intelligence Safety, co-located with 34th AAAI Conference on Artificial Intelligence, SafeAI@AAAI 2020, New York City, NY, USA, February 7, 2020. 2020: 54–62.
- [20] Sch tt S Y, Amin R M, Butz A. A Literature Survey of How to Convey Transparency in Co-Located Human-Robot Interaction. [J]. *Multimodal Technol. Interact.*, 2023, 7(3): 25.
- [21] Schreiter T, Morillo-Méndez L, Chaldavada R T, et al. Advantages of Multimodal versus Verbal-Only Robot-to-Human Communication with an Anthropomorphic Robotic Mock Driver. [C]//32nd IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2023, Busan, Republic of Korea, August 28–31, 2023. 2023: 293–300.
- [22] Bensch S, Eriksson A. Mining multimodal communication patterns in interaction with explainable and non-explainable robots[C]//IEEE RO-MAN 2023, 32nd IEEE International conference on Robot and Human Interactive Communication; Workshop Human-Robot Interaction for Explainability in Robotics, Busan, Korea, August 28–31, 2023. 2023.
- [23] Alikhani M, Sharma P, Li S, et al. Cross-modal Coherence Modeling for Caption Generation. [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020. 2020: 6525–6535.
- [24] Otto C, Springstein M, Anand A, et al. Understanding, Categorizing and Predicting Semantic Image-Text Relations. [C]//Proceedings of the 2019 International Conference on Multimedia Retrieval, ICMR 2019, Ottawa, ON, Canada, June 10–13, 2019. 2019: 168–176.
- [25] Alikhani M, Sharma P, Li S, et al. Cross-modal Coherence Modeling for Caption Generation. [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020. 2020: 6525–6535.
- [26] Alikhani M, Khalid B, Stone M. Image-text coherence and its implications for multimodal AI. [J]. *Frontiers Artif. Intell.*, 2023, 6.
- [27] Yao B M, Shah A, Sun L, et al. End-to-End Multimodal Fact-Checking and Explanation Generation: A Challenging Dataset and Models[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. Taipei Taiwan:

ACM, 2023: 2733–2743.

[28] Yuan X, Guo J, Qiu W, et al. Support or Refute: Analyzing the Stance of Evidence to Detect Out-of-Context Mis- and Disinformation. [C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6–10, 2023. 2023: 4268–4280.

[29] Cai Y, Chen X, Mao Y, et al. Hbtp: Heuristic behavior tree planning with large language model reasoning[C]// 2025 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2025: 13706–13713.

[30] Luan Z, Lai Y, Huang R, et al. Enhancing Robot Task Planning and Execution through Multi-Layer Large Language Models. Sensors. 2024; 24(5):1687.

[31] Kolve E, Mottaghi R, Han W, et al. AI2-THOR: An Interactive 3D Environment for Visual AI[M]. arXiv, 2022.

[32] Wu Y, Hu X, Fu Z, et al. GPT-4o: Visual perception performance of multi-modal large language models in piglet activity understanding[M]. arXiv, 2024.

翁啟華（2003–），男，华南理工大学计算机科学与工程学院研究生，主要研究方向为大模型、具身智能。



韩宇辰（2002–），男，华南理工大学计算机科学与工程学院硕士研究生，主要研究方向为人工智能大模型、具身智能等。

收稿日期: XXXX-XX-XX
通信作者: jerric_wu@163.com
基金项目: 深港联合资助项目(A类)(No.SGDX20240115103359001);国家重点研发计划资助项目(青年科学家项目)(No.2024YFB4504300)
Foundation Items: The Shenzhen-Hong Kong Joint Funding Project (Category A) (No. SGDX20240115103359001), The National Key Research and Development Program of China (Youth Scientist Project)(No.2024YFB4504300)