

自我反馈多模态大模型:基于指令事实性判别与 音频辅理解的幻觉缓解方案

刘 帅¹, 陈 达¹, 潘以恒¹, 李 前², 蔺琛皓^{2*}, 沈 超²

(1. 西安交通大学软件学院, 陕西西安 710049; 2. 西安交通大学网络空间安全学院, 陕西西安 710049)

摘 要: 基于人类反馈的强化学习(Reinforcement Learning from Human Feedback, RLHF)能够有效推动模型输出与人类偏好对齐,已被广泛用于抑制多模态大语言模型(Multimodal Large Language Models, MLLMs)在实际应用中出现的幻觉问题. 其中,直接偏好优化(Direct Preference Optimization, DPO)方法通过避免显式奖励建模,以更稳定、高效的方式提升 MLLMs 的可信度与可用性,受到学术界与工业界的广泛关注. 然而, DPO 训练过程中仍存在若干挑战,如训练数据分布偏移、偏好数据构建过程中对指令事实性区分不足等,均可能加剧模型幻觉. 此外,现有方法对视频等多图数据中伴随的音频信息利用不足,而音频可作为视觉理解的有效补充信号,具备缓解幻觉的潜力. 针对上述问题,本文提出一种基于指令事实性判别与音频辅助的自对齐训练框架(Instruction Factuality and Audio Assistance, IFAA),通过四个核心模块生成高质量偏好数据,以抑制 MLLMs 的幻觉现象. 具体包括:(1)同风格响应采样,有效降低直接偏好优化训练中的数据分布偏移;(2)长响应分割策略,提升模型自我判别的准确性;(3)指令事实性判别模块,构建更具事实依据的偏好数据;(4)音频辅理解模块,通过融合音频信息提升偏好数据质量. 最后通过直接偏好优化训练增强模型的可靠性. 此外,创新性地引入基于 ROC(Receiver Operating Characteristic)曲线的置信平衡点选择机制,以有效缓解多模态大型语言模型的过度自信问题. 本文在五大主流 MLLM 评测基准上进行了实验,以验证所提框架的有效性 with 泛化能力. 以 LLaVA(Large Language and Vision Assistant)1.5 模型为例,经本框架优化后,其在 Object HalBench(Object Hallucination Benchmark)评测集上的句子级幻觉率降低 43.1%,实例级幻觉率下降 37.3%. 此外,在其他前沿模型上的迁移实验表明,基于 IFAA 构建的偏好数据具有良好的泛化性,能够显著降低不同模型的幻觉率. 该结果验证了本文框架在不同模型上的适用性,为 MLLMs 的幻觉抑制提供了新的有效途径.

关键词: 多模态大语言模型;幻觉缓解;偏好学习;自我对齐;指令事实性;音频辅助

基金项目: 科技部重点研发计划(No.2023YFB3107400);国家自然科学基金(No.T2341003, No.62376210, No.62161160337, No.62132011, No.U21B2018, No.U20A20177, No.622206217);陕西省重点研发计划重点产业创新链项目(No.2023ZDLGY38)

中图分类号: TP389.1 文献标识码: A 文章编号: 0372-2112(2025)12-4560-15

电子学报 URL: <http://www.ejournal.org.cn>

DOI:10.12263/DZXB.20250502

Self-Alignment Multimodal LLMs: Mitigating Hallucinations via Instruction Factuality and Audio Assistance

LIU Shuai¹, CHEN Da¹, PAN Yi-heng¹, LI Qian², LIN Chen-hao^{2*}, SHEN Chao²

(1. School of Software Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China;

2. School of Cyber Science and Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China)

Abstract: Reinforcement learning from human feedback (RLHF) can effectively align model outputs with human preferences and has been widely used to mitigate the hallucination problem of multimodal large language models (MLLMs) in practical applications. Among various RLHF approaches, direct preference optimization (DPO) avoids explicit reward modeling, enabling more stable and efficient improvement of MLLMs' reliability and usability. As a result, DPO has attracted extensive attention from both academia and industry. However, the DPO training process still faces several challenges: issues such as training data distribution shift and insufficient distinction of the factuality of instructions during preference data

construction may exacerbate model hallucinations. Additionally, existing methods underutilize the audio information accompanying multi-image data (e.g., videos). As an effective supplementary signal for visual understanding, audio has the potential to alleviate hallucinations. To address the aforementioned problems, this paper proposes an instruction factuality assessment and audio-aided self-alignment training framework (IFAA). This framework generates high-quality preference data through four core modules to suppress hallucinations in MLLMs. The specific modules are as follows: (1) Style-consistent response sampling, which effectively reduces data distribution shift in DPO training; (2) Long-response segmentation strategy, which improves the accuracy of the model's self-judgment; (3) Instruction factuality assessment module, which constructs preference data with stronger factual basis; (4) Audio-aided understanding module, which enhances the quality of preference data by fusing audio information. Finally, DPO training is conducted to further improve the model's reliability. In addition, this paper innovatively introduces a confidence balance point selection mechanism based on the receiver operating characteristic (ROC) curve to effectively mitigate the overconfidence issue of MLLMs. To verify the effectiveness and generalization ability of the proposed framework, experiments are conducted on five mainstream MLLM evaluation benchmarks. Taking the large language and vision assistant (LLaVA) 1.5 model as an example, after optimization with the IFAA framework, its sentence-level hallucination rate on the object hallucination benchmark (Object HalBench) dataset decreases by 43.1%, and the instance-level hallucination rate drops by 37.3%. Furthermore, transfer experiments on other cutting-edge models demonstrate that the preference data constructed based on IFAA exhibits excellent generalization, significantly reducing the hallucination rates of different models. These results confirm the applicability of the proposed framework across various models and provide a new effective approach for hallucination mitigation in MLLMs.

Key words: multimodal large language models; hallucination mitigation; preference learning; self-alignment; instruction factuality; audio assistance

Foundation Item(s): Ministry of Science and Technology Key Research and Development Program (No. 2023YFB3107400); National Natural Science Foundation of China (No. T2341003, No. 62376210, No. 62161160337, No. 62132011, No. U21B2018, No. U20A20177, No. 62206217); Shaanxi Provincial Key Research and Development Program Key Industrial Innovation Chain Project (No. 2023ZDLGY38)

1 引言

随着多模态研究的快速发展,融合图像与文本信息的多模态大语言模型(Multimodal Large Language Models, MLLMs)受到广泛关注^[1-4]. MLLMs通常将视觉特征编码器与大型语言模型(Large Language Models, LLMs)相结合,首先通过视觉编码器提取图像特征,再将其与文本提示共同输入语言模型进行分析与响应,展现出强大的图文理解能力^[5]. 该类模型突破了传统LLMs仅能处理单一模态信息的限制,通过在大规模多模态数据上进行训练,获得了处理跨模态信息的能力. MLLMs被视为实现通用人工智能的关键路径之一,显著扩展了LLMs的应用潜力,有望应用于自动驾驶^[6]、多模态机器人^[7]等领域.

然而,现有研究普遍指出,MLLMs存在明显的幻觉问题,即模型生成的文本内容可能与视觉输入或提示词不一致,具体表现为颜色错配、位置错误、虚构物体等. 此类问题严重制约了MLLMs的实际应用. 因此,抑制MLLMs的幻觉现象,使其输出更符合人类预期,已成为推动模型落地的重要研究方向.

近年来,已有诸多工作尝试揭示多模态幻觉的产生机制. 研究表明,幻觉可能源于模型对参数化知识的过度依赖(如统计偏差与语言先验)^[8,9],视觉编码器在

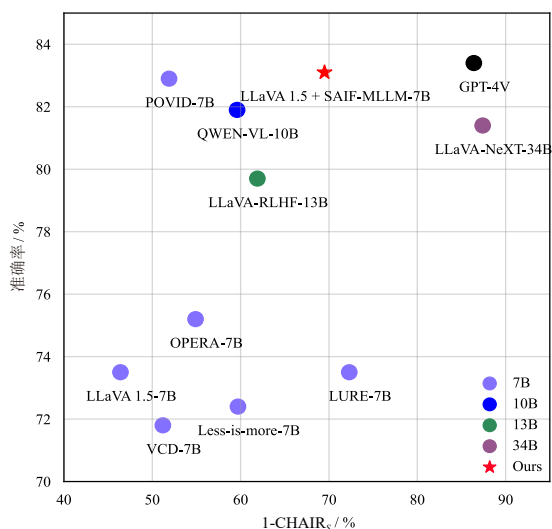
细粒度特征提取上的不足^[10],以及训练数据中存在错误标注^[11,12]等. 为缓解该问题,研究者提出多种策略,涵盖提升数据质量、优化模型结构、改进训练方法、调整解码机制等.

例如HalluciDoctor^[12]在数据预处理阶段引入一致性交叉检查机制,自动剔除训练数据中的幻觉样本,生成高质量数据集LLaVA+(Large Language and Vision Assistant),从而降低模型幻觉. Wu等人^[13]提出一种名为SEAL(Show, sEArch, and telL)的通用MLLM架构,通过语言模型引导的视觉搜索机制,在处理高分辨率图像时实现更精准的视觉定位,缓解因细节信息丢失引发的幻觉. 直接偏好优化(Direct Preference Optimization, DPO)^[14]采用隐式奖励建模方法,简化人类偏好对齐流程,提升训练稳定性与效率,使模型输出更贴合人类判断. VCD(Visual Contrastive Decoding)^[8]则引入视觉对比解码策略,通过对比原始视觉输入与失真视觉输入对应的输出分布差异,校正生成结果,以减轻统计偏差与语言先验的影响,增强响应与视觉内容的一致性.

针对当前MLLMs中出现的严重幻觉问题(如图1(b)上部所示),现有研究多依赖强大的商用模型提供修正信号^[15]. 该方法需投入大量资源,存在可持续性挑战. 而采用其他开源MLLMs进行反馈则易因模型间指令遵

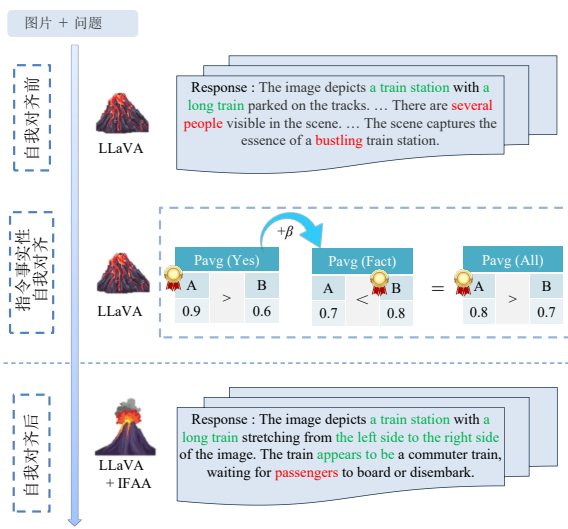
循能力差异、风格不一致等问题,导致训练效果不稳定,甚至出现幻觉较训练前恶化的情况^[2,5]。

我们认为,上述问题主要源于不同模型在训练数据风格、形式及内容上的显著差异。为此,本文探索利用模型自身生成反馈信号^[16],以保持数据风格的一致性,缓解数据与模型之间的分布偏移。此外,研究发现现有偏好数据构建方法过于强调数据的绝对真实性,未能根据指令类型动态调整评估标准,且多数 MLLMs 未能充分利用视频数据中的音频信息。



(a) IFAA 训练的模型与其他模型的可信度比较

基于此,本文提出基于指令事实性与音频辅助的视觉对齐策略(Instruction Factuality and Audio Assistance, IFAA)。对于事实性指令,该方法不仅考量响应的正确性,还综合评估其事实准确性,从而构建更高质量的偏好数据,提供更为精准有效的反馈指导。如图1(b)下部所示,通过对指令事实性加以区分,MLLMs 可实现自我对齐,显著提升模型的可信度。图中红色标注表示生成文本中存在幻觉的内容,绿色标注则为正确事实部分。



(b) LLaVA 模型经过 IFAA 缓解幻觉后的效果展示

图1 使用 IFAA 后对模型的性能提升和可视化效果展示

针对 MLLMs 的幻觉问题,本文提出一种自对齐模型框架。为解决模型自身反馈准确性不足的挑战,本文将长回复拆解为短因子(factor),以提升模型自评的精度。同时,采用同风格响应采样方法,确保反馈信息风格一致,使模型更专注于响应内容的正确性,有效避免了因使用异源 MLLMs 反馈而导致的训练波动,为生成高质量偏好数据奠定了基础。

其次,针对当前偏好数据构建方法单一的问题,本文提出一种新的偏好数据构造策略。具体而言,对于事实性相关指令,额外引入事实性评估机制,以衡量数据的质量,从而获取正确又符合事实的偏好数据。

此外,调研发现,现有大多数 MLLMs 或未利用视频中的音频信息,或将音频作为新增模态扩展至模型结构。本文创新性地提出以音频辅助视觉理解的新训练范式,通过融合音频信息构建偏好数据,进一步抑制模型幻觉。

以 LLaVA 1.5 7B 为例,如图1(a)所示,经 IFAA 算法优化后,模型幻觉显著减少,其性能从同规模模型中的最低水平提升至最优,并在多项指标上可媲美参数量为 10B 与 13B 的模型。本文选用 AMBER (An LLM-

free Multi-dimensional BEncH maRk) 的准确率(纵轴)与 Object HalBench (Object Hallucination Benchmark) 的句子级幻觉率 (Caption Hallucination Assessment with Image Relevance, CHAIR_s) (横轴) 综合评估模型效果。在五个主流评测集上的结果表明,IFAA 能有效提升基础模型的可信度与真实性。通过 IFAA 构建偏好对并进行 DPO 训练,模型在多个数据集上取得最优或次优性能。相较于基线模型,在 Object HalBench^[17] 上句子级幻觉率降低 43.1%,在 MMHal-Bench^[7] (Multimodal Models Hallucination Benchmark) 上相对幻觉减少 37.1%,体现 IFAA 在增强 MLLMs 可信性方面的显著优势。

本文的主要贡献可归纳为以下三点:

(1) 提出了 IFAA,一种基于指令事实性判别与音频辅助的偏好数据构建算法。经偏好对齐训练后,MLLMs 的幻觉现象显著降低,可信性得到提升。

(2) 创新性地提出同风格回复生成与长回复拆解策略,通过模型自生成偏好数据缓解 DPO 中的数据分布偏移问题;同时引入基于 ROC (Receiver Operating Characteristic) 的置信度平衡点选择方法,有效抑制模型过度自信。

(3)设计了对比实验与消融实验,验证所提框架在各维度上的有效性,并通过跨模型实验证明了所生成偏好数据具有良好的泛化能力。

2 相关工作

2.1 幻觉起源

MLLMs的幻觉是指模型生成的响应与用户输入、所提供的视觉信息不一致,与已有上下文冲突,或违背现实事实。通常,MLLMs产生幻觉的原因可归纳为以下四个方面:

(1)模型过度依赖参数化知识而非当前输入信息。MLLMs在推理过程中往往更倾向于调用训练时学到的参数化知识,而忽视当前提供的视觉内容与用户指令。Liu等人^[11]指出,模型在处理相关图像和人类指令时容易产生描述不一致的幻觉;Leng等人^[8]发现模型会生成看似合理但实际错误的输出,如颜色、位置等属性错乱;Zhai等人^[18]认为MLLMs在理解视觉细节方面仍存在较大不确定性;Guan等人^[19]在测试中观察到语言幻觉与视觉错觉并存的现象;McKenna等人^[20]指出模型对训练数据存在记忆依赖,导致性能受训练集分布影响显著;而Ren等人^[21]则发现模型过于依赖内部知识检索,忽略当前输入信息。

(2)指令微调数据质量不足。Liu等人^[11]与Yu等人^[12]分析指出,训练数据中因人工标注错误导致的标签噪声会引发模型幻觉;Yu等人^[7]进一步提出,微调数据规模不足且缺乏人类反馈,也会加剧幻觉现象。

(3)生成算法存在缺陷。Chuang等人^[22]指出,低层信息可能在生成过程中覆盖高层语义信息,从而引发幻觉;Lee等人^[23]发现,开放式文本生成中常用的采样算法可能损害模型的事实一致性。

(4)视觉编码器对细粒度信息捕捉能力有限。Tong等人^[10]指出,CLIP(Contrastive Language-Image Pre-training)等视觉编码器仅基于实例级对比学习进行预训练,可能导致“CLIP-blind”现象,即忽视图像间的显著差异而判定其相似;Jiang等人^[24]则认为现有方法多依赖CLIP或其变体作为视觉分支,且仅从深层提取特征,缺乏对视觉编码器的全面分析,导致视觉与文本模态间仍存在语义鸿沟。

2.2 幻觉消除

目前,消除MLLMs幻觉大致从以下五个方面着手:

(1)预处理。HalluciDoctor^[12]提出一种一致性交叉检查机制,可自动检测并剔除视觉指令数据中的多种幻觉错误,进而构建更高质量的数据集LLaVA+。Less-is-More^[25]采用基于句子结束符(End Of Sentence token, EOS)预测能力的数据过滤策略,通过评估训练数据对模型EOS预测的正负面影响,计算综合有害性评分

S_{final} ,据此过滤对模型终止生成能力有害的样本,提升模型的生成控制能力。

(2)模型结构改进。V*^[13]提出名为SEAL的多模态语言模型架构,通过集成语言模型引导的视觉搜索机制,在处理高分辨率图像时实现更精细的视觉定位。

(3)训练方法优化。Shikra^[26]引入与空间位置相关的数据,使模型能够理解图像中的空间关系,提升其在位置相关任务上的表现。DRESS^[27]采用条件强化学习框架,利用大型语言模型生成的自然语言反馈(Natural Language Feedback, NLF)训练视觉语言模型,增强其与人类偏好的对齐能力及多轮交互性能。Ziegler等人^[28]提出基于模型自身不确定性估计生成文本的真实性,并通过偏好学习算法直接优化模型以提升输出真实性。BPO^[29](Bootstrapped Preference Optimization)通过偏好学习抑制预训练阶段的偏差,强化模型对视觉输入的理解。

(4)解码策略优化。在文本生成模型LLMs中,ICD^[30](Induce-then-Contrast Decoding)在解码过程中通过对比“弱事实性”模型与原始模型的输出分布,增强原始模型的事实性信号,抑制不可靠生成。类似地,针对多模态MLLMs场景,VCD^[8]通过对比原始视觉输入与失真视觉输入对应的输出分布差异,校正生成结果,无需额外训练或外部工具,已被广泛应用于视觉-语言模型的解码过程中。OPERA(Over-trust PEnalty and Retrospection-Allocation)^[31]在解码阶段引入惩罚项与回溯重分配策略,有效降低MLLMs的幻觉生成。

(5)训练数据质量提升方法。

(a)基于人类反馈的强化学习(Reinforcement Learning from Human Feedback, RLHF)提出近端策略优化(Proximal Policy Optimization, PPO)框架,包含策略模型、奖励模型、评论模型与参考模型。DPO^[14]通过建立奖励函数与最优策略间的映射关系,可直接在偏好数据上以分类损失形式优化策略模型,实现人类偏好对齐。RLHF-V^[7]采用人工修正方法构建细粒度多模态偏好数据,并引导模型重点学习人工修改部分,充分利用标注信息实现更精准的对齐。

(b)基于自我纠正、反馈与学习的幻觉抑制方法。Yuan等人^[32]通过模型自我生成奖励信号进行训练,不依赖人类反馈,探索模型超越人类性能的潜力。基于人工智能反馈的强化学习(Reinforcement Learning from AI Feedback, RLAIFF)^[33]以开源LLM替代人类标注偏好,使用AI生成偏好训练强化学习奖励模型,达到与RLHF相近的性能。FGAIF(Fine-Grained AI Feedback)^[34]通过ChatGPT将响应拆分为子句进行评判,加权计算响应得分,进而收集LLaVA1.5 13B的反馈用于模型训练。Sun等人^[35]指出低层视觉感知与理解任务易

因模型缺乏自我意识产生幻觉,构建了首个面向低层视觉任务幻觉的HLPV(Hallucinations in Low level visual Perception and Understanding)指令库(含约20万问答对及4个子集),并提出融合图像特征、显著区域特征与质量特征模型,实验表明该方法可显著提升模型自我意识并降低幻觉。

(c)基于事实关联的对齐方法.在LLMs上,Zhang等人^[36]提出事实自对齐方法,利用LLM的自评估能力提供真实性训练信号,结合相关组件促使模型基于内部知识验证响应真实性.Liang等人^[37]开发自动幻觉标注工具DreamCatcher,融合知识探测与一致性检查对事实偏好数据排序,并提出基于知识反馈的强化学习(Reinforcement Learning with Knowledge Feedback, RLKF)框架以提升LLM真实性.在多模态领域,本文首次引入指令事实性区分机制,针对不同指令类型采取差异化偏好数据构建策略,通过IFAA框架生成高质量多模态偏好对数据,有效降低MLLMs幻觉并提升可靠性。

尽管现有方法多样,但仍然面临高质量反馈数据获取困难、训练效率低下及模型评估与人工判断一致性不足等问题,导致MLLMs幻觉现象依然严重.为缓解这些局限,本研究首次提出基于指令事实性区分机制,并结合音频辅助视觉理解,构建高质量多模态偏好数据.通过IFAA框架,有望显著降低MLLMs幻觉率,提升其可靠性。

3 基于指令事实性与音频辅助的自对齐式训练框架(IFAA)

在本节中,本文首先详细地展示了IFAA算法的工作流程,之后介绍了IFAA的四个关键组件,即相同风格响应生成算法、拆分长响应算法、区分指令事实性的偏好数据构建策略以及音频辅助视觉理解模块的详细内容,最后介绍如何使用IFAA生成的偏好对数据进行DPO训练以提升模型性能。

3.1 IFAA算法概览

IFAA算法包含四个核心步骤,整体流程如图2所示。

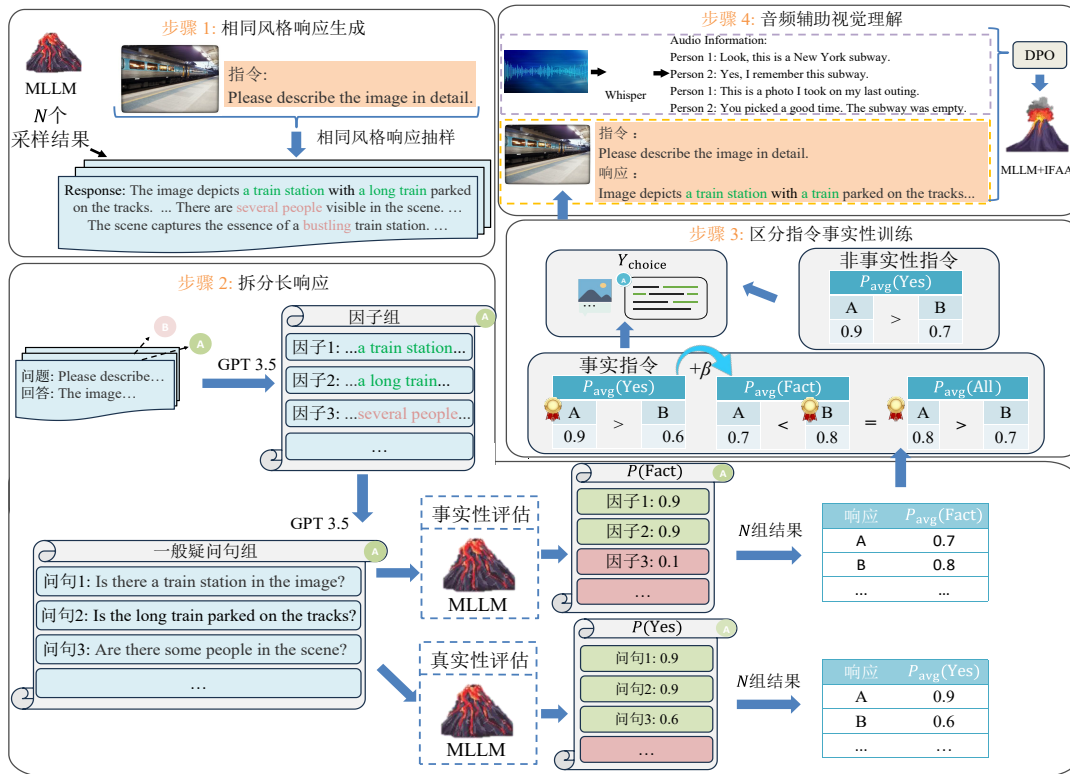


图2 本文提出的IFAA-MLLM框架概述(正确答案文本与幻觉内容分别以红色和绿色高亮标注)

步骤1:基于少样本同风格回复生成.针对输入的命令-图像对,首先使用MLLM生成若干初始响应作为少样本示例,随后基于这些示例引导MLLM生成多个风格一致的候选回复,确保生成内容在表达风格上具有一致性。

步骤2:长回复拆分与因子级评估.对步骤1所得候选回复,利用GPT-3.5进行事实信息提取,将其分解为多个独立的短因子(factor).随后,使用MLLM对每个因子进行事实性评估并打分,通过加权平均得到每个回复的事实性得分.进一步,将各因子转换为一般疑问

句形式,再次使用 MLLM 进行真实性评估(以“是/否”形式判断),加权计算每个回复的真实性得分。

步骤 3: 指令事实性区分的偏好对构建. 通过 GPT-3.5 判断指令是否为事实性指令,并据此采用差异化策略构建偏好对. (1)非事实性指令:选取候选回复中真实性得分最高的作为偏好响应,最低的作为非偏好响应;(2)事实性指令:将每个候选回复的真实性得分与事实性得分按预设权重加权,计算综合得分,选取最高分与最低分回复分别作为偏好与非偏好响应。

步骤 4: 音频辅助视觉理解与 DPO 训练. 针对视频输入,引入音频作为视觉语义的补充信号,通过五阶段处理实现多模态融合:(a)跨模态特征解耦;(b)时序对

齐的语义提取;(c)视觉引导的语义过滤;(d)视听联合文本生成;(e)幻觉抑制验证. 最终完成偏好数据的音频信息增强,并基于 DPO 对 MLLM 进行训练优化。

3.2 相同风格响应生成算法

DPO 方法所使用的偏好对数据来源多样,包括人工标注及其他 MLLMs 标注,可能引入主观偏差与文本风格差异,导致数据分布与模型分布之间的偏移. 为缓解此问题,本文采用初始 MLLM 自身生成偏好数据. 具体而言,对于给定的指令 $-x_{\text{inst}}$ 和图片 $-img$, 首先使用 MLLM 生成若干初始响应,作为少样本(few-shot)提示示例. 随后,依据图 3 上半部分所示的提示词格式,将示例输入初始 MLLM,生成一系列风格一致的候选回复,记为候选响应序列 $\{y_n\}_{n=1}^N$.

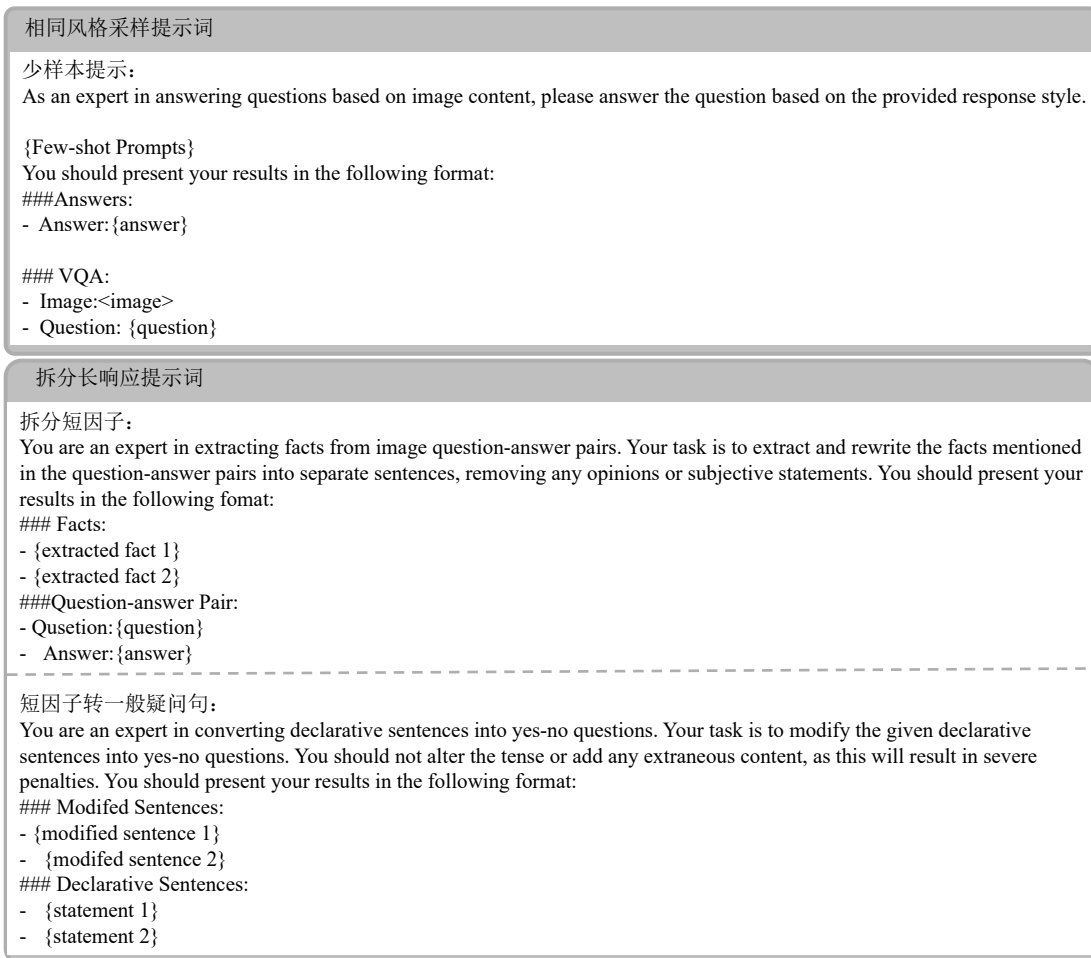


图 3 用于 IFAA 的模板

3.3 拆分长响应算法

实验表明,MLLMs 对长回复的幻觉判别准确性较低,可能因长文本结构复杂而难以精准评估. 为此,本文提出将长回复拆分为短因子(factor),以降低模型判别难度。

具体流程如下:针对第 3.2 节生成的候选序列 $\{y_n\}_{n=1}^N$. 使用可靠性较高的 GPT-3.5 对每个候选响应进行事实提取. 例如,对于包含 M 个事实的候选响应 $\{y_n\}_{n=1}^N$,经本步骤后得到其对应的事实因子集合 $\{y_x\}_{x=1}^M$,提示词示例见图 3 下半部分。

进一步,利用GPT-3.5将这些因子转换为一般疑问句形式,使MLLM能够以“是/否”形式进行判断,从而聚焦于内容真实性,避免受文本风格或句式干扰。

本文使用初始MLLM,基于事实性提示 q_f 为每一个factor f 对每个因子 f 进行事实性评估,得到分数 $P(\text{Fact}|q_f, f)$;基于正确性提示 $P(\text{Fact}|q_f, f)$ 对每个一般疑问句 gq 进行真实性评估,得到分数 $P(\text{Yes}|q_c, gq)$ 。以参数 θ 表示的MLLM评估公式如下:

$$\begin{cases} P(\text{Fact}|q_f, f) = f_\theta(f) \\ P(\text{Yes}|q_c, gq) = f_\theta(gq) \end{cases} \quad (1)$$

随后,对每个候选响应中的所有因子得分进行加权平均,得到该响应的事实性平均得分 $P_{\text{avg}}(\text{Fact})$ 和真实性平均得分 $P_{\text{avg}}(\text{Yes})$ 。

实验还发现,在计算 $P(\text{Yes}|q_c, gq)$ 时,基于ROC曲线选取置信度平衡点的方法优于固定阈值方案。该方法通过动态确定最优判别阈值,有效缓解了模型自我评估中因过度自信(over-confidence)导致的偏差问题。

3.4 区分指令事实性偏好训练算法

对于给定的 x_{inst} 和 x_{img} ,首先通过GPT-3.5判断 x_{inst} 是否为事实性指令。随后根据指令类型采取不同的偏好对构建策略。

(1)非事实性指令:通过选取 N 个候选响应中真实性得分最高的响应为chosen的偏好响应 y_c ,得分最低的响应为rejected的非偏好响应 y_r ,构造偏好对数据。

(2)事实性指令:将每个候选的真实性得分和事实性得分以 $1:\beta$ 的比例加权,选取候选响应中综合得分最高的响应为 y_c ,得分最低的响应为 y_r ,构造偏好对数据 $D = \{(x, y_c, y_r)\}$ 。接着本文将两种偏好对数据混合,并通过DPO训练初始MLLM,涉及公式如下:

$$\begin{cases} p_{\text{avg}}(\text{All}) = p_{\text{avg}}(\text{Yes}), & I_{\text{inst}} \notin I_{\text{fact}} \\ p_{\text{avg}}(\text{All}) = p_{\text{avg}}(\text{Yes}) + \beta p_{\text{avg}}(\text{Fact}), & I_{\text{inst}} \in I_{\text{fact}} \end{cases} \quad (2)$$

本文选取 β 为1,使模型在事实性指令中同等权衡真实性与事实性。实际应用中可调整 β 以灵活适配不同场景的需求。此外,本文中事实性指令与非事实性指令的数量比例为4:1,以保持和训练数据的原始分布对齐。我们保持训练数据中多项选择题、开放问答以及字幕任务的数据量不变。针对字幕任务,通过改动训练时该类型数据的采样比例(事实指令:非事实指令),在当前主流评测集如MMHal-Bench和AMBER上进行实验。此外,我们选取LLaVA-Instruct-150K数据集,针对性抽取并构建了一个创意描述改写评测集LLaVA-Creative Description,该数据集问题均由非事实性指令构建而成,来模拟评测数据全部为非事实性指令的极端情况。结果如表1所示。

表1 不同采样策略下的性能对比分析 单位:%

模型	数据比例	MMHal-Bench		AMBER		LLaVA-Creative Description
		Score	Hall.	Acc.	F1	Hall.
LLaVA 1.5	5:0	3.14	38.5	82.2	86.2	60.4
	4:1	3.20	32.1	83.1	87.4	59.2
	3:2	3.16	36.2	81.8	85.9	58.8
	2:3	3.09	39.7	80.1	84.7	57.4
	1:4	2.92	43.4	74.8	78.1	54.8
	0:5	2.81	48.3	72.4	74.3	52.9

可以看出,在主流评测集如MMHal-Bench和AMBER上,对该比例进行过采样或欠采样均会导致模型性能轻微下降。我们分析这可能是由于该数据比例更符合主流评测集的数据分布。此类评测集源于Open-Images和开源场景如医疗、遥感和日常生活,因此能取得较高指标。此外,我们提供在LLaVA-Creative Description这类全部为非事实性指令的特殊评测集上效果,可见随着数据中非事实性指令数据比例的提升,能有效降低关于非事实性指令数据的幻觉问题。

3.5 音频辅助视频理解

当前MLLMs在处理图像、多图及视频的文本生成任务时,对音频信息的处理主要存在两种方式:

(a)如Video-LLaMA、VideoMAE V2和InternVL等模型完全忽略音频信息,仅依赖视觉内容进行文本生成。

(b)如VITA-1.5和HumanOmni等异构模态统一模型,通过跨模态编码器实现音频理解,但其设计目标侧重于扩展模态覆盖范围,本质上属于模态容量扩充,而非利用音频提升生成文本的质量以抑制幻觉。

受蜜蜂、海豚等生物利用听觉增强视觉感知的机制启发,本文创新性地对音频定位为视觉语义的补充信号。通过利用与视觉内容高度相关且富含语义的音频信息(如环境音、语音旁白),提升生成字幕(CAP)文本的真实性、准确性与完整性。该方法包含以下五个处理阶段:

(1)跨模态特征解耦。使用FFmpeg解耦音视频流,对音频进行去噪与采样率归一化预处理。

(2)时间对齐语义提取。采用Whisper进行自动语音识别(Automatic Speech Recognition, ASR),生成带时间戳的句子级转录文本。

(3)视觉引导语义过滤。对视频帧进行均匀采样,利用跨模态对齐模型(如SigLIP)计算每帧与各音频文本片段的语义相似度。相似度得分计算如下:

$$\phi(s_t, v_t) = \text{Sigmoid}(E_T(S_t)^T E_V(V_t)) \quad (3)$$

其中, S_t 表示第 i 个音频句子的文本表示; v_t 表示时间 t 时视频帧的视觉特征; $E_T(\cdot)$ 和 $E_V(\cdot)$ 分别表示文本和图

像编码器;T表示矩阵转置;Sigmoid表示非线性激活函数; ϕ 表示视觉-文本特征匹配度,用于衡量第*i*个音频句子的文本表示 s_i 与时间*t*时视频帧的视觉特征 v_t 之间的相关性或对齐程度.

考虑到视频字幕需挖掘潜在语义关联而非严格匹配,经理论分析与实验验证,将语义补充信号的激活阈值 τ 设置为0.4.该阈值对应弱相关场景的中值区间 $\phi(s_i, v_t) \in [0.3, 0.5]$,可在捕捉跨模态潜在关联的同时抑制噪声干扰.此设置与SPECTRUM(Semantic Processing and Emotion-informed video-Captioning Through Retrieval and Understanding Modalities)在多模态语义情感分析中的结论一致,表明 $\tau=0.4$ 可实现跨模态关联的最优平衡.

(4)视听文本生成.基于预定义提示模板,利用GPT-4o将过滤后的音频转录按时间戳与内容连贯性整合至视频字幕中,确保描述与视频序列的时间一致性.

(5)示例演示.以厨房烹饪视频为例:

<视觉描述>:“一个人拿着一个白色粉末容器,将颗粒状物质倒入锅中.”

<音频描述>:“现在我们需要在煮熟的豌豆中加入大约2克盐.”

<视听合并描述>:“操作人员正在将2克白盐倒入煮熟豌豆的锅中.”

此外,我们考虑了两种极端情况下的模型表现:

(i)无音频数据.通过统计分析发现音频集中分布于解说、美食制作、vlog等视频类型.对于无音频的输入数据,Whisper提取内容为空.此时训练数据依赖前置因子拆分与指令事实性判断模块得分构建.为提升处理效率,本文额外引入前置筛选环节,使用FFmpeg预筛选输入视频,剔除无音轨或静音(分贝为负无穷)的数据,避免无效调用.

(ii)低资源环境适配.提供两种方案:(a)采用轻量模型(如OmniLMM-3B)进行偏好学习;(b)基于LLaVA 1.5-IFAA 7B,通过两阶段蒸馏结合LoRA(Low-Rank Adaptation of Large Language Models)得到轻量化LLaVA 1.5-IFAA 3B模型.实验结果表明,3B模型在MMHal Benchmark上Score为3.05(达7B模型的95.3%),Hall为34.9(达7B的91.3%),适用于资源受限场景.

3.6 基于DPO的MLLMs对齐

传统的RLHF包含两个核心步骤:奖励模型训练与强化学习微调.其基本思想是基于同一输入*x*,生成的两个不同响应构成偏好对,其中包含一个优选响应 y_c 和一个劣选响应 y_r .通过比较学习人类偏好.具体而言,奖励模型的损失函数定义为

$$L(\phi) = \log \sigma(r(x, y_c) - r(x, y_r)) \quad (4)$$

其中, σ 表示sigmoid函数; r 代表被参数化的 ϕ 的奖励模型的值; $r(x, y)$ 表示输入提示*x*和输出*y*通过奖励模型预测出的奖励值.

DPO^[14]作为一种高效的RLHF算法,通过数学推导将强化学习目标转化为等效的监督学习形式,省去了显式奖励模型与价值模型的复杂建模过程,已被广泛用于LLMs和MLLMs的对齐训练.

DPO推导表明,奖励函数 $r(x, y)$ 可表示为最优策略模型 $\pi_\theta(y|x)$ 与参考模型 $\pi_{\text{ref}}(y|x)$ 的对数几率比:

$$r(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x) \quad (5)$$

其中, β 是常数; $Z(x)$ 是配分函数.基于此,DPO的优化目标可简化为以下交叉熵损失:

$$L_D(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_c, y_r) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_c|x)}{\pi_{\text{ref}}(y_c|x)} - \beta \log \frac{\pi_\theta(y_r|x)}{\pi_{\text{ref}}(y_r|x)} \right) \right] \quad (6)$$

其中,最优策略模型 π_θ 由经过指令微调的基座模型初始化;参考模型 π_{ref} 在训练过程中保持固定; β 用于控制策略模型与参考模型的偏离程度.该目标函数直接基于偏好对数据优化策略模型,避免了复杂的强化学习流程.

4 实验

本节通过实验分析IFAA方法在缓解MLLMs幻觉问题上的效果.基于开源评测集,本文使用可信赖、准确率及 F_1 分数等指标评估了多种MLLMs的性能表现.此外,通过消融实验验证了所提框架中各组件的有效性,并测试了所构造偏好数据在不同MLLMs上的泛化能力.

4.1 实验设置

本节详细介绍了实验所使用的评测基准、对比模型以及偏好训练的具体设置.

评测基准和评估指标.对于图片描述中的常见幻觉,本文采用Object HalBench^[17]来进行评估,具体来说,本文参考less is more^[25]的方法,从MSCOCO验证集随机选择500张图片,以CHAIR_s(句子级别幻觉率)和CHAIR_r(实例级别幻觉率)指标来评价模型的幻觉程度.在响应级别幻觉与信息完整性评估方面,采用MMHal-Bench^[6].该基准通过GPT-4^[38]将模型输出与人工标注及目标标签进行对比,综合评价模型的可信度与实用性.此外,本文也采用了AMBER^[39]来对多模态大模型进行多维度的性能评测.该评测集包含15220个样本,包含生成任务、判别任务、存在维度等任务,本文选取判别任务部分,并提供了模型的准确度和 F_1 分数.

为考查 MLLMs 在复杂任务中的推理与表达能力, 本文进一步选用 MMStar^[40] 和 LLaVA Bench^[41] 两项评测基准, MMStar 整合了六个主流多模态评测集, 经人工筛选得到 1 500 个高质量样本, 从六项核心能力及其下属的 18 个细分类别对模型进行全面评估, 每项能力均包含 250 个平衡样本, 以确保评估的准确性与均衡性。

(1) 对比模型. 本文将所提模型与多种先进模型进行对比, 包括当前广泛使用的通用模型、基于解码策略优化的幻觉抑制模型、采用 RLHF 的模型以及闭源模型。

(a) 通用模型. 本文选择了包含 LLaVA 1.5^[42]、LLaVA-NeXT^[43] 以及 Qwen-VL^[44] 作为具有代表性的通用模型. 此类模型具有很高的关注度, 同时在精心挑选的数据上进行了指令微调, 具有强大的多模态能力。

(b) 幻觉优化模型. 选取了针对 MLLMs 幻觉问题进行专门优化的模型, 可分为两类. 一类方法通过后处理或修改解码方式轻量级别地缓解 MLLMs 的幻觉, 例如 LURE (LVLM (Large Vision-Language Model) hallucination REvisor)^[9] 通过简洁的后处理方法, 针对模型生成的文本描述, 采用训练的修正模型来更改模型中生成的幻觉内容. VCD^[8] 引入视觉对比解码方法, 作为校正机制, 通过对比一个倾向于产生幻觉的分布, 以减轻统计偏差和语言先验的影响, 确保内容生成和视觉输入更加一致. Less-is-more^[25] 修改最大似然估计 (Maximum Likelihood Estimate, MLE) 的训练目标, 使模型学习在适当的位置生成 EOS, 及时终止生成, 并采用额外的数据过滤方法以缓解幻觉问题. OPERA^[31] 在模型解码过程中引入惩罚项和回溯重新分配策略减轻幻觉. 另一类方法则利用 RLHF 技术来缓解幻觉问题, 例如 LLaVA-RLHF^[6] 通过人工标注了高达 10 000 条偏好数据, 并通过 RLHF 技术在偏好数据上训练奖励模型, 最后采用近端策略优化进行模型训练。

(c) 闭源模型. 选取当前公认性能较强的闭源模型 GPT-4V^[45] 作为对比, 以全面评估开源模型与闭源模型之间的性能差距。

(2) 实验方案. 以 LLaVA 1.5^[42] 作为基线模型, 首先评估其原始性能, 再采用 IFAA 框架对其进行偏好训练, 通过对比验证算法有效性。

(3) 实验细节. 对于对比模型^[46], VCD 在噪声步数为 700 的条件下进行测试, OPERA 采用标准推理超参数配置. 对于 IFAA, 通过所述算法构建 4 000 对偏好数据, 并基于 DPO 对 LLaVA 1.5 进行人类偏好对齐. 训练中学习率设为 5×10^{-7} , β 为 0.1, 批大小为 8. 数据生成耗时约 52 h, LLaVA 1.5-IFAA 训练耗时约 5 h. 实验使用 8 张 A800 80G GPU 进行。

4.2 主要结果

主要实验数据汇总展示于表 2, 其中 CHAIR_s 表示字幕任务中单个样本包含至少一个幻觉物体的百分比. CHAIR_t 表示以物体为统计维度, 整个字幕任务中出现幻觉的百分比. Score 为 GPT-4 作为评判者对模型响应质量 (信息量与幻觉程度) 的打分 (范围 0~6), Hall. 表示模型响应文本中包含幻觉的百分比. Acc. 为相应幻觉评测任务中所有样本的准确率, Overall 为各幻觉评测任务下所有子任务的加权平均得分。

通过分析实验结果, 可得出以下结论:

IFAA 有效地提升了 LLaVA 1.5 模型的可靠度, 在 7B 参数量级下, 该模型在几乎所有评测基准中均取得最优的可信度表现, 部分评测集上甚至可媲美更为复杂的 10B 模型. 具体来说, IFAA 算法显著降低了 LLaVA 1.5 在 Object HalBench 上的物体幻觉率, CHAIR_s 和 CHAIR_t 分别相对下降 43.1% 和 37.3%. 在 MMHal-Bench 中, 整体幻觉率相对降低 37.1%. 在 AMBER 评测中, 模型准确率位列次优, F₁ 分数达到最优, 性能接近闭源模型 GPT-4V.

在综合性更强的 LLaVA Bench 与 MMStar 评测集上, 经 IFAA 优化后的模型在复杂任务中表现出更优的可用性与可信性. 与原始 LLaVA 1.5 相比, IFAA 使模型在上述两个基准上的成绩均有提升, 说明该方法在增强 MLLMs 可信度的同时, 能够保持模型原有的复杂推理与表达能力。

4.3 分析

本文对 IFAA 算法框架进行了深入分析, 重点探讨以下几个问题:

(1) 将长回复分解为短因子 (factor) 是否必要; 依据指令的事实性类别区分构建偏好数据, 是否比不加以区分的构建方式更有效?

实验表明, 分解长回复有助于提升偏好数据的质量, 而区分事实性指令则能进一步提高模型在可信度与真实性方面的表现. 为验证各策略的有效性, 本文设计了消融实验. 在验证分解策略时, 对比了三种设置: 原始 LLaVA 1.5、基于长回复进行置信度判断的方法, 以及将长回复分解为短因子的方法. 实验在 MMHal-Bench 与 AMBER 基准上进行, 其余设置保持一致. 在验证指令区分策略时, 比较了“区分事实性指令”与“不区分指令类型”两种 DPO 策略配置. 结果展示如表 3 所示. 可以发现, 各个策略均有效地缓解了模型的幻觉率, 提升了准确度. 综合使用四种策略的完整 IFAA 方法取得了最低的幻觉度与最高的准确率, 表明多策略协同具有显著优势。

此外, 本文发现不同策略在不同评测集上的效果存在差异. 在 MMHal-Bench 中, 拆分长回复的策略效果

表2 主要实验结果

单位:%

模型	尺寸	Object HalBench		MMHal-Bench		AMBER		LLaVA ench	MMStar
		CHAIR _s	CHAIR _t	Score	Hall.	Acc.	F ₁	Overall	Overall
VCD	7B	48.8	24.3	2.12	54.2	71.8	74.9	65.8	33.8
Less-is-more	7B	40.3	17.8	2.33	50.0	72.4	75.8	60.9	32.9
OPERA	7B	45.1	22.3	2.15	54.2	75.2	78.3	60.3	32.9
CCA-LLaVA	7B	46.7	23.8	1.92	61.5	77.7	81.9	61.7	32.1
LURE	7B	27.7	17.3	1.64	60.4	73.5	77.7	36.9	33.3
POVID	7B	48.1	24.4	2.08	56.2	82.9	87.2	62.2	34.3
HA-DPO	7B	39.9	19.9	1.98	60.4	75.2	79.9	60.7	32.9
QWEN-VL	10B	40.4	20.7	2.76	38.5	81.9	86.4	71.9	34.5
Silkie	10B	27.1	16.4	3.19	32.3	82.2	87.6	61.8	33.6
LLaVA-RLHF	13B	38.1	18.9	2.02	62.5	79.7	83.9	61.6	34.2
AMP-MEG	13B	31.7	20.6	3.08	36.5	79.5	84.6	62.1	34.8
GPT-4V	—	13.6	7.3	3.49	28.1	83.4	87.4	93.1	50.4
OmnimMM	12B	19.4	10.9	3.06	36.5	86.5	89.5	65.7	39.7
+IFAA	12B	10.6	8.5	3.40	32.8	87.5	90.2	69.8	40.2
LLaVA 1.5	7B	53.6	25.2	2.36	51.0	73.5	77.7	59.7	33.3
+IFAA	7B	30.5	15.8	3.20	32.1	83.1	87.4	61.2	35.2

表3 不同策略下的性能对比分析

单位:%

模型	相同风格响应生成	拆分长响应	区分指令事实性	音频辅助视觉	MMHal-Bench		AMBER	
					Score	Hall.	Acc.	F ₁
LLaVA 1.5	—	—	—	—	2.36	51.0	73.5	77.7
	√	×	×	×	2.44	47.8	74.6	78.1
	√	√	×	×	2.95	37.3	76.8	79.2
	√	√	√	×	3.02	35.7	82.2	85.3
	√	√	√	√	3.20	32.1	83.1	87.4

尤为显著,这可能源于该基准包含较多长形式问答(Long-form QA)任务,将回复分解为短因子有助于模型更准确地评估每个局部的置信度,从而提升整体性能.而在 AMBER 评测基准中,区分指令事实性的策略对模型效果影响更大,原因可能是 AMBER 中包含大量事实性指令,因此针对此类指令额外引入真实性判断机制,能够更有效地提升准确率并抑制幻觉.

本文进一步对算法中各模块所使用的外部模型与工具进行了成本与效能分析:

(a)在长回复拆分与指令事实性判断环节中调用 GPT-3.5,其计算开销分别占整体数据生成与训练流程的 3.6% 与 1.4%. 偏好数据生成总耗时为 52 h,其中上述两个模块共计耗时约 6 h. 为降低对闭源商用模型的依赖,本文额外提供了通过监督微调(Supervised Fine-Tuning, SFT)得到替代模型 LLaVA-split 与 LLaVA-Fact,经人工抽检 1 000 条样本统计,其效果分别达到 GPT-3.5 的 98.4% 与 99.62%. 两个模块对性能的具体提升见表 3 拆分长响应和区分指令事实性.

(b)音频辅助视觉^[47]模块中调用 Whisper 进行音频

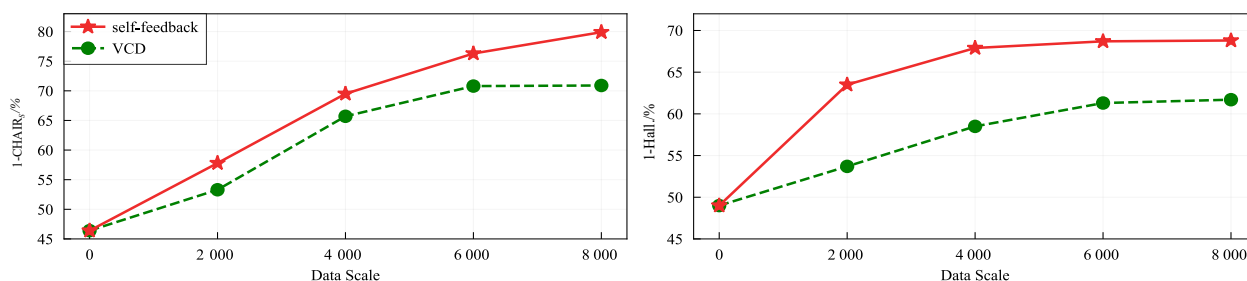
识别,其开销占总流程的 12.5%,耗时约 12 h. 由于 Whisper 为开源模型,可便捷部署于服务器中,具备较高的可靠性. 模块对效果的具体提升见表 3 音频辅助视觉.

(2)通过模型自我反馈生成的偏好数据相比于其他模型反馈生成的数据具有哪些优势?

自我反馈机制有助于缓解分布偏移问题. 为评估其有效性,本文在图 4 中对比了使用自我反馈构建的偏好对与基于 VCD 模型反馈生成的偏好对在训练过程中的表现. 实验结果表明,由 VCD 模型引导生成的偏好数据在训练早期即出现性能饱和,未能充分挖掘数据潜力;而基于自我反馈构建的偏好数据则能够持续优化模型性能,具体体现为在 Object HalBench 的句子级幻觉率 CHAIR_s 与 MMHal-Bench 的整体幻觉率(Hall.) 指标上均呈现稳定下降趋势.

(3)ROC 曲线确定置信度阈值的方法是否更为有效?

ROC 曲线动态确定置信度平衡点的方法表现出其优势. 本文选取了固定阈值下的错误百分比置信度



注:实验选取 Object HalBench 的 CHAIR_s 指标与 MMHal-Bench 的 Hall_s 指标进行分析。

图4 基于IFAA构建与VCD构建偏好数据的训练结果对比

(False Percentage Confidence, FPC)、真实百分比置信度 (True Percentage Confidence, TPC) 以及基于 ROC 曲线动态选取阈值的正确百分比置信度 (TPC-ROC). 从表4中可以看出,与固定阈值的置信度方法相比,该策略提高了训练效果. 本文认为这是因为该策略有效地减少了模型的过度自信问题,从而提高了训练效果.

表4 不同策略构建的偏好数据对模型性能的影响 单位:%

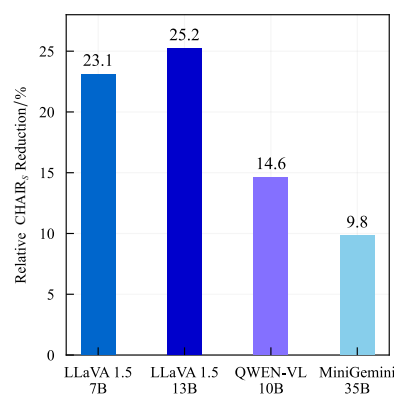
方法	Object HalBench		MMStar	AMBER	
	CHAIR _s	CHAIR _t	Overall	Acc.	F ₁
FPC	33.5	17.7	33.8	82.1	86.9
TPC	31.2	16.9	34.3	81.9	86.7
TPC-ROC	30.5	15.8	35.2	83.1	87.4

(4)IFAA 收集的偏好对数据是否可用于提升其他多模态大型语言模型的可信度?

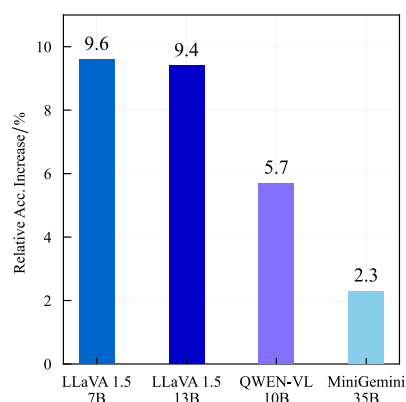
IFAA 算法能够产生高质量的通用偏好对数据. 为验证其泛化能力,本文基于 LLaVA 1.5 7B 模型,通过 IFAA 方法收集了 4 000 条偏好对数据,并利用 DPO 方法对另外三个不同规模的模型—LLaVA 1.5 13B、Qwen-VL 10B 及 MiniGemini-35B 进行训练. 图5展示了各模型在多个基准测试中可信度的提升情况. 实验结果表明,使用 IFAA 收集的偏好对数据能够有效降低不同 MLLMs 在多种评测任务中的幻觉现象,说明该方法所生成的数据具有良好的通用性,可用于增强各类 MLLMs 的可信度.

(5)GPT3.5 能否胜任判断指令事实性任务,能否使用开源方案或模型本身完成该任务?

GPT-3.5 与本文所使用的基座模型 LLaVA 1.5 的 LLM 模块均能有效完成指令事实性判断. 为评估 GPT-3.5 的准确性,本文随机抽取 1 000 条事实性与非事实性指令(如事实性指令:“图中男子手中是否拿着一个苹果?”,非事实性指令:“请描述一个小猪飞向月球的故事”)进行测试,其判断准确率高达 99.70%. 由于指令文本较短,该任务在整体流程中的计算开销仅占 1.4%. 考虑到闭源模型在稳定性与可持续性方面的限制,本文进一步使用 LLaVA 1.5 的 LLM 模块,通过对一批指令事



(a) 展示 Object HalBench 基准上的相对 CHAIR_s 下降率



(b) 呈现 AMBER 基准准确率的相对提升幅度

图5 基于IFAA构建的偏好对数据在其他MLLMs中的幻觉抑制效果

实性标注样本 SFT, 训练出开源模型 LLaVA-Fact, 以提供可替代的解决方案. 经相同抽样测试, LLaVA-Fact 的准确率达到 99.32%, 表明其能够可靠地完成指令事实性判断任务.

4.4 实例展示

为直观比较不同模型的性能差异,本文在图6中展示了两个典型场景的定性分析结果.

(1)在第一个案例中,三个模型均正确识别出图像为冬季雪景,并描述出“两匹马”和“一辆拖车”. 与



注:正确答案与幻觉内容分别以红色和绿色标注.

图6 LLaVA 1.5+IFAA-MLLM与LLaVA 1.5、GPT-4V的定性对比结果

LLaVA 1.5 和 GPT-4V 相比,本文提出的模型进一步准确描述出马匹的颜色. 尽管本文模型与 LLaVA 1.5 仍存在一定程度的事实性幻觉,但幻觉程度相较原始 LLaVA 1.5 有所降低.

(2)在第二个案例中,三个模型均能正确识别场景中的人物及足球运动. 本文的 LLaVA 1.5-IFAA 模型与 GPT-4V 均准确描述了人物的衣着颜色,且本文模型在可信度方面表现尤为突出,幻觉现象显著减少. 相比之下,LLaVA 1.5 与 GPT-4V 在物体识别上均出现错误,生成了图中不存在的实体或关系,产生明显幻觉.

5 结论

缓解幻觉是 MLLMs 迈向实际运用的关键问题. 本文提出了 IFAA,这是一种通过判别指令的事实性加以区分收集自我对齐偏好数据,并借助音频信息辅助视觉理解的新型 MLLMs 框架. 实验结果表明,本文的模型在多个当前主流数据集上的可信度达到了先进的性能. IFAA 通过区分指令的事实性类别,采取分治策略生成偏好数据,确保针对事实性指令的偏好数据同时兼顾可信度与真实性. 通过少样本学习引导模型生成

多样化回复,并将长响应拆解为短因子,有效缓解了 DPO 方法中存在的分布偏移问题.此外,采用 ROC 曲线确定置信度平衡点,显著减轻了模型自评估过程中的过度自信(over-confidence)现象.未来,我们将探索如何从 MLLMs 自身提取更多有效的反馈信号,以进一步提升模型在复杂场景下的适应与推理能力.

参考文献

- [1] AGRAWAL H, DESAI K R, WANG Y F, et al. Nocaps: Novel object captioning at scale[C]//2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2020: 8947-8956.
- [2] CHEN D P, CHEN R X, ZHANG S L, et al. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark[EB/OL]. (2024-06-11) [2025-10-30]. <https://arXiv.org/abs/2402.04788>.
- [3] HUANG Q Q, XIONG Y, RAO A Y, et al. MovieNet: A holistic dataset for movie understanding[C]//Computer Vision - ECCV 2020. Cham: Springer, 2020: 709-727.
- [4] KEMBHAVI A, SALVATO M, KOLVE E, et al. A diagram is worth a dozen images[C]//Computer Vision - ECCV 2016. Cham: Springer, 2016: 235-251.
- [5] LI J N, LI D X, SAVARESE S, et al. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//International Conference on Machine Learning. New York: ACM, 2023: 19730-19742.
- [6] SUN Z Q, SHEN S, CAO S C, et al. Aligning large multimodal models with factually augmented RLHF[C]//Findings of the Association for Computational Linguistics ACL 2024. Stroudsburg: ACL, 2024: 13088-13110.
- [7] YU T Y, YAO Y, ZHANG H Y, et al. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 13807-13816.
- [8] LENG S C, ZHANG H, CHEN G Z, et al. Mitigating object hallucinations in large vision-language models through visual contrastive decoding[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 13872-13882.
- [9] ZHOU Y Y, CUI C H, YOON J, et al. Analyzing and mitigating object hallucination in large vision-language models[EB/OL]. (2024-03-16) [2025-11-11]. <https://arXiv.org/abs/2310.00754>.
- [10] TONG S B, LIU Z, ZHAI Y X, et al. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 9568-9578.
- [11] LIU F X, LIN K, LI L J, et al. Mitigating hallucination in large multi-modal models via robust instruction tuning[EB/OL]. (2024-03-19) [2025-10-10]. <https://arXiv.org/abs/2306.14565>.
- [12] YU Q F, LI J C, WEI L H, et al. HalluciDoctor: Mitigating hallucinatory toxicity in visual instruction data[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 12944-12953.
- [13] WU P H, XIE S N. V*: Guided visual search as a core mechanism in multimodal LLMs[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 13084-13094.
- [14] RAFAILOV R, SHARMA A, MITCHELL E, et al. Direct preference optimization: Your language model is secretly a reward model[EB/OL]. (2024-07-29) [2025-10-10]. <https://arXiv.org/abs/2305.18290>.
- [15] LI L, XIE Z H, LI M K, et al. Silkie: Preference distillation for large visual language models[EB/OL]. (2023-12-17) [2025-10-10]. <https://arXiv.org/abs/2312.10665>.
- [16] CHEN Z X, DENG Y H, YUAN H Z, et al. Self-play fine-tuning converts weak language models to strong language models[EB/OL]. (2024-06-14) [2025-10-10]. <https://arXiv.org/abs/2401.01335>.
- [17] ROHRBACH A, HENDRICKS L A, BURNS K, et al. Object hallucination in image captioning[EB/OL]. (2019-03-29) [2025-10-10]. <https://arXiv.org/abs/1809.02156>.
- [18] ZHAI B H, YANG S J, ZHAO X C, et al. Halle-Switch: Rethinking and controlling object existence hallucinations in large vision-language models for detailed caption[C]//The 12th International Conference on Learning Representations. Appleton: ICLR, 2023: 3227.
- [19] GUAN T R, LIU F X, WU X Y, et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 14375-14385.
- [20] MCKENNA N, LI T Y, CHENG L, et al. Sources of hallucination by large language models on inference tasks[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg: ACL, 2023: 2758-2774.
- [21] REN R Y, WANG Y H, QU Y Q, et al. Investigating the factual knowledge boundary of large language models with retrieval augmentation[C]//Proceedings of the 31st

- International Conference on Computational Linguistics. Kerrville: Association for Computational Linguistics, 2025: 3697-3715.
- [22] CHUANG Y S, XIE Y J, LUO H Y, et al. DoLa: Decoding by contrasting layers improves factuality in large language models[EB/OL]. (2024-03-11)[2025-10-10]. <https://arXiv.org/abs/2309.03883>.
- [23] LEE N, PING W, XU P, et al. Factuality enhanced language models for open-ended text generation[EB/OL]. (2023-03-02)[2025-10-10]. <https://arXiv.org/abs/2206.04624>.
- [24] JIANG D S, LIU Y C, LIU S L, et al. From CLIP to DINO: Visual encoders shout in multi-modal large language models[EB/OL]. (2024-03-08)[2025-10-10]. <https://arXiv.org/abs/2310.08825>.
- [25] YUE Z H, ZHANG L, JIN Q. Less is more: Mitigating multimodal hallucination from an EOS decision perspective[EB/OL]. (2024-05-29)[2025-10-10]. <https://arXiv.org/abs/2402.14545>.
- [26] CHEN K Q, ZHANG Z, ZENG W L, et al. Shikra: Unleashing multimodal LLM's referential dialogue magic[EB/OL]. (2023-07-03)[2025-10-10]. <https://arXiv.org/abs/2306.15195>.
- [27] CHEN Y Y, SIKKA K, COGSWELL M, et al. DRESS: Instructing large vision-language models to align and interact with humans via natural language feedback[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 14239-14250.
- [28] ZIEGLER D M, STIENNON N, WU J, et al. Fine-tuning language models from human preferences[EB/OL]. (2020-01-08)[2025-10-10]. <https://arXiv.org/abs/1909.08593>.
- [29] PI R J, HAN T Y, XIONG W, et al. Strengthening multimodal large language model with bootstrapped preference optimization[C]//European Conference on Computer Vision - ECCV 2024. Cham: Springer, 2025: 382-398.
- [30] ZHANG Y, CUI L Y, BI W, et al. Alleviating hallucinations of large language models through induced hallucinations[C]//Findings of the Association for Computational Linguistics: NAACL 2025. Stroudsburg: ACL, 2025: 8218-8232.
- [31] HUANG Q D, DONG X Y, ZHANG P, et al. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 13418-13427.
- [32] YUAN W Z, PANG R Y, CHO K, et al. Self-rewarding language models[EB/OL]. (2025-03-28)[2025-10-10]. <https://arXiv.org/abs/2401.10020>.
- [33] YU T Y, ZHANG H Y, LI Q M, et al. RLAIIF-V: Open-source AI feedback leads to super GPT-4V trustworthiness[EB/OL]. (2025-10-29)[2025-10-10]. <https://arXiv.org/abs/2405.17220>.
- [34] JING L Q, DU X Y. FGAIF: Aligning large vision-language models with fine-grained AI feedback[EB/OL]. (2025-05-06)[2025-10-10]. <https://arXiv.org/abs/2404.05046>.
- [35] SUN Y N, MIN X K, ZHANG Z C, et al. Mitigating low-level visual hallucinations requires self-awareness: Database, model and training strategy[EB/OL]. (2025-03-27)[2025-10-10]. <https://arXiv.org/abs/2503.20673>.
- [36] ZHANG X Y, PENG B L, TIAN Y, et al. Self-alignment for factuality: Mitigating hallucinations in LLMs via self-evaluation[EB/OL]. (2024-06-11)[2025-10-10]. <https://arXiv.org/abs/2402.09267>.
- [37] LIANG Y X, SONG Z Y, WANG H, et al. Learning to trust your feelings: Leveraging self-awareness in LLMs for hallucination mitigation[EB/OL]. (2024-01-27)[2025-10-10]. <https://arXiv.org/abs/2401.15449>.
- [38] OPENAI, ACHIAM J, ADLER S, et al. GPT-4 technical report[EB/OL]. (2024-03-04)[2025-10-10]. <https://arXiv.org/abs/2303.08774>.
- [39] WANG J Y, WANG Y H, XU G H, et al. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation[EB/OL]. (2024-02-23)[2025-10-10]. <https://arXiv.org/abs/2311.07397>.
- [40] CHEN L, CHEN Z H, DONG X Y, et al. Are we on the right way for evaluating large vision-language models? [C]//Advances in Neural Information Processing Systems 37. Vancouver: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 27056-27087.
- [41] LIU H T, LI C Y, WU Q Y, et al. Visual instruction tuning[EB/OL]. (2023-12-11)[2025-10-10]. <https://arxiv.org/abs/2304.08485>.
- [42] LIU H T, LI C Y, LI Y H, et al. Improved baselines with visual instruction tuning[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 26286-26296.
- [43] LIU H T, LI C Y, LI Y, et al. Llava-next: Improved reasoning, ocr, and world knowledge[EB/OL]. (2024-01-30)[2025-10-10]. <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- [44] BAI J Z, BAI S, YANG S S, et al. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond[EB/OL]. (2023-10-13)[2025-10-

10]. <https://arXiv.org/abs/2308.12966>.

[45] OpenAI GPT. GPT-4V(ision) system card[EB/OL]. (2023-09-01) [2025-10-10]. <https://openai.com/index/gpt-4v-system-card/>.

[46] 陆庆阳, 袁广林, 朱虹, 等. 一种基于对比学习大模型的视觉定位方法[J]. 电子学报, 2024, 52(10): 3448-3458.
LU Q Y, YUAN G L, ZHU H, et al. A visual grounding

method with contrastive learning large model[J]. Acta Electronica Sinica, 2024, 52(10): 3448-3458. (in Chinese)

[47] 胡杰, 昌敏杰, 徐博远, 等. ConvFormer: 基于Transformer的视觉主干网络[J]. 电子学报, 2024, 52(1): 46-57.

HU J, CHANG M J, XU B Y, et al. ConvFormer: Vision backbone network based on transformer[J]. Acta Electronica Sinica, 2024, 52(1): 46-57. (in Chinese)

作者简介



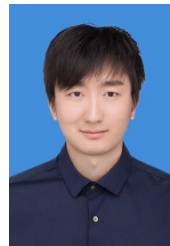
刘 帅 女, 1987年8月出生于陕西省西安市. 现为西安交通大学副教授. 主要研究方向为人工智能应用与安全、大型语言模型和软件设计. 中国电子学会会员编号: E190185359M.
E-mail: sh_liu@mail.xjtu.edu.cn



李 前 男, 1992年8月出生于陕西省宝鸡市. 现为西安交通大学助理教授. 主要研究方向为人工智能安全、对抗机器学习.
E-mail: qianlix@xjtu.edu.cn



陈 达 男, 2001年7月出生于安徽省六安市. 现为西安交通大学硕士研究生. 主要研究方向为多模态大模型.
E-mail: dachen@stu.xjtu.edu.cn



蔺琛皓 男, 1989年5月出生于陕西省西安市. 现为西安交通大学教授. 主要研究方向为人工智能安全、模式识别. 中国电子学会会员编号: E190029398M.
E-mail: linchenhao@xjtu.edu.cn



潘以恒 男, 2002年9月出生于辽宁省铁岭市. 现为西安交通大学硕士研究生. 主要研究方向为多模态大模型.
E-mail: panyiheng@stu.xjtu.edu.cn



沈 超 男, 1985年9月出生于重庆市. 现为西安交通大学特聘教授. 主要研究方向为人工智能安全、内部人员/入侵检测. 中国电子学会会员编号: E190181360M.
E-mail: chaoshen@xjtu.edu.cn