

文章编号: 1003-0077(2026)05-0001-25

从感知到推理: 深度思考赋能多模态大语言模型

祝文鑫¹, 陈安东¹, 宋宇宸¹, 陈科海², 朱聪慧¹, 陈自岩³, 赵铁军¹

(1. 哈尔滨工业大学 计算学部, 黑龙江 哈尔滨 150001;
2. 哈尔滨工业大学(深圳) 计算机科学与技术学院, 广东 深圳 518055;
3. 中译语通科技股份有限公司, 北京 100043)

摘要: 随着多模态大语言模型在感知任务中取得显著成果, 如何提升其复杂推理能力成为研究重点。当前模型存在推理路径不透明、泛化能力不足等问题, 而思维链推理有助于增强模型的推理透明性和输出可解释性, 且已在多模态大语言模型中取得显著效果, 将该范式扩展至多模态领域, 有望提升模型的推理能力。该文围绕“多模态思维链”展开系统梳理。首先, 从技术演进与任务需求角度分析其提出背景与理论动因; 其次, 从思维链范式、后训练阶段与推理阶段三个方面介绍了当前多模态思维链的主流方法, 并分析了其内在机理。此外, 该文还总结了现有评测基准与指标, 并探讨了多模态思维链的应用场景。最后, 分析了当前多模态思维链面临的挑战并展望其未来发展方向。

关键词: 多模态大语言模型; 思维链推理; 多模态思维链

中图分类号: TP391

文献标识码: A

From Perception to Reasoning: Deep Thinking Empowers Multimodal Large Language Models

ZHU Wenxin¹, CHEN Andong¹, SONG Yuchen¹, CHEN Kehai², ZHU Conghui¹, CHEN Ziyang³, ZHAO Tiejun¹

(1. Faculty of Computing, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China;
2. School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Guangdong 518055, China;
3. Global Tone Communication Technology, Beijing 100043, China)

Abstract: With the remarkable progress of multimodal large language models (MLLMs) in perception tasks, enhancing their reasoning capability has become a key research focus. Chain-of-Thought (CoT) reasoning improves reasoning transparency and output interpretability, and extending this paradigm to the multimodal domain is expected to enhance the reasoning ability of MLLMs. This paper presents a systematic review of multimodal CoT. It first analyzes the background and theoretical motivations from the perspectives of technological evolution and task demands. Then, it introduces the major approaches to enabling CoT reasoning in MLLMs, including the CoT paradigm, post-training phase, and inference phase. Furthermore, it summarizes existing benchmarks and evaluation metrics, and explores potential application scenarios of multimodal CoT. Finally, the paper discusses the current challenges in this field and outlines future research directions.

Keywords: multimodal large language models; chain of thought reasoning; multimodal chain of thought

0 引言

依托海量数据和强大计算资源, 近年来多模态大语言模型(Multimodal Large Language Models, MLLMs)在跨模态内容理解与生成任务中取得了显著进展^[1-5], 广泛应用于图像字幕生成(Image Captioning)、视觉问

答(Visual Question Answering, VQA)^[6]以及视频摘要(Video Captioning)^[7]等任务。尽管多模态大语言模型在感知与生成方面表现优异, 其在应对涉及复杂推理的任务时仍存在明显不足^[8-11]。具体来说, 现有多模态大语言模型主要依赖隐式推理, 即基于训练数据的统计模式进行预测^[12-13], 缺乏明确、可解释的中间推理步骤, 因而在多步逻辑推导、因果关系分析和组合

收稿日期: 2025-07-13 定稿日期: 2025-12-13

基金项目: 国家重点研发计划(2024YFC3809100); 国家自然科学基金(62376075, 62276077, U23B2055); 广东省基础与应用基础研究基金(2024A151011205); 深圳市科技计划项目(GXWD20220811170358002, GXWD20220817123150002)

泛化等方面能力有限^[14-15]。

对于文本大语言模型 (Large Language Models, LLMs) 中的类似问题, 研究人员通过引入思维链 (Chain-of-Thought, CoT)^[16-18] 推理机制以提升其逻辑推理能力。思维链的核心思想在于将复杂问题拆解为一系列明确的中间推理步骤, 模拟人类逐步构建逻辑链条的思维过程。该方法不仅在复杂推理任务中展现出显著优势, 还增强了模型的可解释性与决策透明度^[19-21]。在大语言模型领域, OpenAI o1 和 DeepSeek-R1^[22] 等代表模型通过集成思维链策略, 在各推理任务中取得了显著突破, 推动了大语言模型推理能力的发展。

受这一成功范式启发, 研究人员开始探索将思维链推理机制拓展至多模态大语言模型中, 即发展基于思维链的多模态推理模型 (CoT-MLLMs)^[23-24]。这种新范式旨在将结构化的推理步骤嵌入多模态建模框架中, 使模型能够在处理视觉、语言等多种模态时捕捉不同模态信息之间的深层语义关联, 提升整体推理能力与可解释性^[25-27]。相较于文本大语言模型, 激发多模态大语言模型的推理能力受到诸多限制。其需要同时处理多种模态信息, 因此如何在不同模态间建立有效的思维链是一个重要挑战。这不仅涉及模态间的信息对齐问题, 还包括如何构建多模态间的层级

推理结构, 以实现跨模态的深度推理。

为整合领域发展现状, 学术界已出现一些总结性工作^[13,15,25], 为研究人员提供该领域的全面概览。在前人工作的基础上, 本文旨在提供一个差异化且更侧重于理论剖析的视角。现有工作侧重于归纳技术方法, 而本综述的核心贡献在于进一步分析讨论其内在机理。此外, 在多模态思维链的评测以及挑战和未来方向等章节, 本文进行了更加体系化的分类和总结。总结来说, 本文系统地梳理了 CoT-MLLMs 的最新研究进展, 涵盖其核心方法、评估基准与指标、典型应用领域以及面临的挑战与未来方向, 期望所提供的总结和见解能够为多模态思维链推理这一新兴方向提供系统性的参考与理论支持, 推动该领域的持续发展。

如图 1 所示, 本文首先在第 1 节介绍了与 CoT-MLLMs 相关的基本概念和背景知识。随后, 在第 2 节总结了现有 CoT-MLLMs 中的思维链范式, 以及在训练和推理阶段的主流方法, 并分析了其内在机理。第 3 节提供了对目前已有评价基准和指标的总结。在第 4 节中, 对多模态思维链的几个具体应用领域进行总结。最后, 在第 5 节详细讨论了该领域面临的挑战和未来的发展方向。

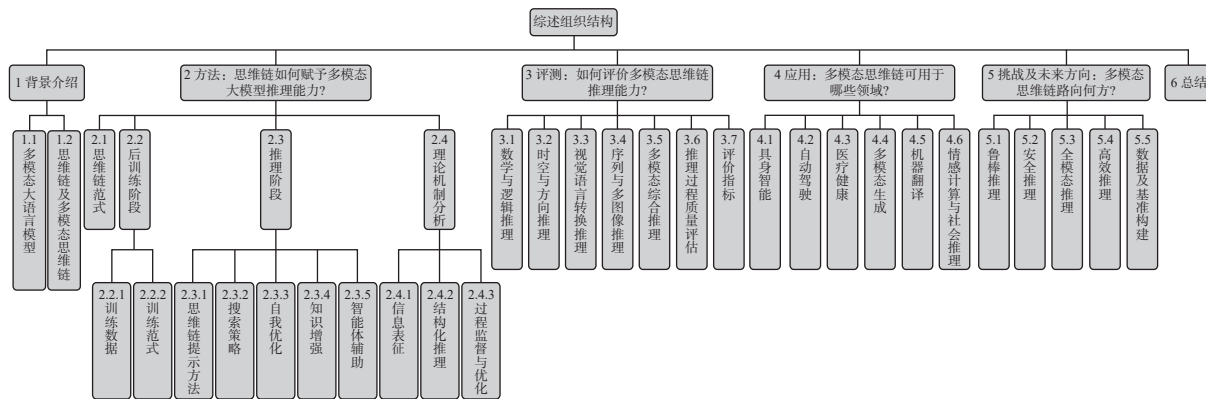


图 1 综述主体部分的组织结构

1 背景介绍

在多模态大语言模型不断突破传统人工智能瓶颈的背景下, 研究者逐步认识到其在复杂推理任务中的局限性。现有模型的推理能力受到端到端学习范式的制约, 难以完成逻辑推导、因果推理或多步推理任务。因此, 引入思维链推理成为一种潜在解决方案。构建 CoT-MLLMs, 即具备思维链推理能力的多

模态大语言模型, 不仅涉及如何在多模态大语言模型中构建显式的推理链, 还需要应对跨模态信息融合、知识对齐以及训练优化等挑战。随着研究的深入, CoT-MLLMs 正逐步展现其在多模态复杂推理任务中的潜力, 为具身智能、医疗健康和机器翻译等领域的发展注入新动力。

1.1 多模态大语言模型

近年来, 多模态大语言模型在跨模态理解与生成

任务中取得了显著进展,并涌现出一系列代表模型。例如,Flamingo^[28]通过视觉-语言桥接模块在零样本和少样本任务中表现卓越;BLIP-2^[1]采用双阶段训练策略,实现了高效的图文对齐能力;LLaVA^[4]通过对视觉编码器与语言模型的高效适配,使得多模态推理能力得到进一步增强。此外,GPT-4V作为GPT-4的多模态版本,在图像理解任务中展现出强大的综合推理能力。多模态大语言模型的基本架构通常由模态编码器、模态融合模块和大语言模型三个核心部分组成。模态融合模块是多模态大模型的关键组成部分,其主要任务是将不同模态的数据进行语义对齐和信息融合,从而实现跨模态理解与生成。其常见方法有BLIP-2^[1]、Video-LLaMA^[3]、InstructBLIP^[29]等模型采用的Q-Former方法,以及MiniGPT-4^[30]、LLaVA^[4]和PMC-VQA^[31]等模型采用的简单线性层或多层感知机投影。通过三个核心模块的协同工作,多模态大语言模型能够在输入模态之间进行信息交互,实现跨模态推理与内容生成。

目前多数多模态大模型在一些基础模态感知任务中表现较好,但其在复杂推理任务上仍然存在明显不足。例如,Yang等^[32]发现在进行区域涂色任务时,Qwen2.5-VL^[2]虽然具备一定的多模态处理能力,但在涉及空间关系和逻辑判断时易出现推理错误。Toh等^[33]在实验中发现OpenAI o1模型在视觉推理基准PUZZLEVQA上的表现低于人类水平,且GPT-4o在感知和归纳推理方面存在明显短板。究其原因,首要问题在于当前多模态大模型大多依赖统计模式进行预测,缺乏明确的推理机制。这使得他们在逻辑推理任务中容易出现跳步推理、错误归因等问题,难以复现人类那种逐步演绎的推理过程。

另一方面,当前多模态大模型在视觉感知能力上也并不完备,尤其是在面向推理任务时,感知和理解之间的信息传导常常存在断层^[34-37]。多模态推理要求模型能够从图像、文本等多源输入中高效提取、融合并组织信息,而现有模型在这一过程中仍存在明显瓶颈。Ghosal等^[38]在实验中发现,当为模型提供更准确、结构化的感知信息后,其在视觉推理任务中的表现显著提升,这进一步印证了当前多模态大模型在感知与推理协同机制上的不足。

1.2 思维链及多模态思维链

针对当前多模态大模型在复杂推理任务中存在的两大关键挑战,即缺乏明确、结构化的推理机制,以及在感知与推理之间存在断层,研究者引入思维链推

理机制作为解决方案。思维链通过引入显式的中间推理步骤,使模型能模拟人类逐步推理的认知过程,并被证明能显著提升大模型在复杂认知任务中的表现^[16]。形式化地,给定文本输入 x ,传统大模型在回答问题时直接建模 $P(y|x)$,即直接进行基于概率的自回归形式的回答生成。而引入思维链后的推理过程可重构为如式(1)所示。

$$P(y|x) = \sum_{r_1, \dots, r_n} P(y|r_n) \prod_{t=1}^n P(r_t|r_{t-1}, x) \quad (1)$$

其中, r_t 表示第 t 步的文本推理过程。该方法通过逐步构建推理路径,有效缓解了模型跳步、模糊推理的问题,并提高了推理链条的可追踪性。在此基础上,研究者进一步发展出多种扩展形式。例如,Long等^[39]及Yao等^[40]提出树状思维链(Tree-of-Thought, ToT),通过构建分支推理路径来增强探索与评估能力。

为进一步解决“感知与推理断层”问题,研究者将思维链范式扩展至多模态领域,提出多模态思维链(Multimodal Chain-of-Thought, MCoT),实现了异构模态信息的协同推理^[23]。图2展示了常规多模态大模型的黑盒式推理与朴素多模态思维链推理的一个样例对比。可以看到,多模态思维链的主要优势为显式的推理过程,一些方法还会将其他模态信息融入推理过程,以更好地实现多模态协同推理。

多模态思维链的推理过程可分为三个阶段:模态内语义演化阶段,其不仅依赖自身的历史状态与原始输入,还受到上一步模态融合状态的影响;跨模态融合表示阶段,当前融合状态基于所有模态当前语义状态与前一融合状态,并进一步实现语义对齐;输出阶段,在融合思维链末端状态基础上生成最终输出,确保推理路径的连贯性与回答的可解释性。上述过程可形式化为如式(2)~式(4)所示。

$$P_{\text{modal}} = \prod_{t=1}^T \prod_{m \in \mathcal{M}} P(r_t^m | r_{t-1}^m, r_{t-1}, x^m) \quad (2)$$

$$P_{\text{fusion}} = \prod_{t=1}^T P(r_t | r_{t-1}, \{r_t^m\}_{m \in \mathcal{M}}) \quad (3)$$

$$P(\mathcal{Y}|X) = \sum_{r_1, \dots, r_T} P(\mathcal{Y}|r_T) \cdot P_{\text{fusion}} \cdot P_{\text{modal}} \quad (4)$$

其中, $\mathcal{M}$ 表示所有模态的集合,如视觉模态 v 、文本模态 l 等; $X = \{x^m\}_{m \in \mathcal{M}}$ 为多模态输入, x^m 为模态 m 的原始输入; r_t^m 是模态 m 在时间步 t 的中间推理过程,而 r_t 表示各模态融合后的联合推理过程; P_{modal} 建模模态内的状态演化, P_{fusion} 描述模态间的语义交互与推理整合, $P(\mathcal{Y}|X)$ 是最终输出,其可以是多种模态。

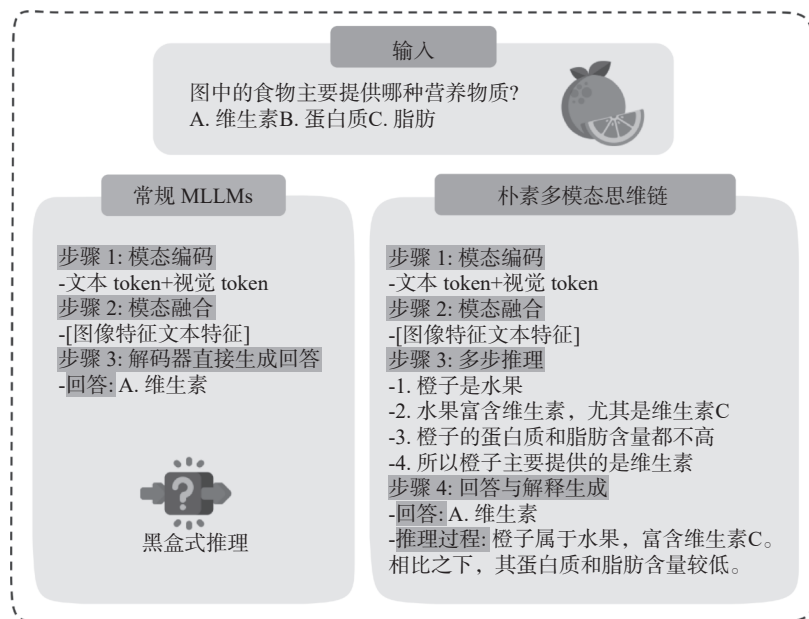


图 2 常规多模态大模型与朴素多模态思维链推理对比

综上所述, 多模态思维链通过结构化推理路径增强了传统多模态大模型对复杂语义关系的建模能力, 有效缓解了黑盒式预测不可解释的问题。作为一种新兴推理范式, 多模态思维链架起了模态感知与逻辑演绎之间的桥梁, 为提升多模态任务的性能与可解释性提供了有力支持。

2 方法: 思维链如何赋予多模态大模型推理能力?

本节围绕“思维链如何赋予多模态大模型推理能

力”展开, 首先介绍不同思维链范式的核心特点, 随后探讨其在后训练阶段与推理阶段的关键实现策略与内在机理, 并在表 1 中总结了主流的多模态思维链方法与代表模型。

2.1 思维链范式

当前多数多模态大模型采用“模态编码器编码+跨模态融合+解码器生成”的端到端黑盒式推理。其缺乏显式推理路径, 容易陷入表层语义对齐, 更适用于感知主导、任务复杂度较低的场景。

为激活模型更深层次的推理能力, 研究者提出思

表 1 多模态思维链相关方法和模型总结

模型/方法	时间	研究机构	支持模态	后训练方法	推理时缩放方法	任务/领域
HoT ^[41]	2023-8	中国科学院大学	图像, 文本	仅微调	—	科学问答
DDCoT ^[42]	2023-10	上海科技大学	图像, 文本	—	思维链提示	视觉问答
Chain-of-look ^[43]	2023-10	纽约州立大学布法罗分校	视频, 文本	—	思维链提示	医学影像分析
GCoT ^[44]	2023-12	威斯康星大学麦迪逊分校	视频, 文本	—	—	自动驾驶
CoCoT ^[45]	2024-1	罗切斯特大学	图像, 文本	—	思维链提示	多图像理解
KAM-CoT ^[46]	2024-1	三星	图像, 文本	—	知识增强	科学问答
BBA ^[47]	2024-2	腾讯	图像, 文本	—	—	多领域
CoS ^[48]	2024-3	清华大学	图像, 文本	—	—	图像问答
CCoT ^[49]	2024-4	加州大学伯克利分校	图像, 文本	—	思维链提示	多领域
MI-CoT ^[50]	2024-4	IIIT-D	图像, 文本	—	思维链提示	物理推理
VoT ^[51]	2024-5	新加坡国立大学	视频, 文本	—	自我优化	视觉理解
Multimodal-CoT ^[23]	2024-5	上海交通大学	图像, 文本	—	—	科学问答

续表						
模型/方法	时间	研究机构	支持模态	后训练方法	推理时缩放方法	任务/领域
IoT ^[52]	2024-5	西湖大学	图像, 文本	—	思维链提示	视觉理解
CoT-ST ^[53]	2024-9	哈尔滨工业大学	语音, 文本	仅微调	—	语音翻译
ARES ^[54]	2024-10	俄亥俄州立大学	图像, 文本	微调+强化学习	—	视觉问答
SKETCHPAD ^[55]	2024-11	华盛顿大学	图像, 文本	—	—	数学推理
LLaVA-o1 ^[56]	2024-11	清华大学	图像, 文本	仅微调	集束搜索	多模态推理
VIC ^[57]	2024-11	宾夕法尼亚大学	图像, 文本	—	—	多模态推理
Insight-V ^[58]	2024-11	南洋理工大学	图像, 文本	微调+强化学习	—	多模态推理
VisVm ^[59]	2024-12	微软	图像, 文本	—	搜索方法	多模态推理
AtomThink ^[60]	2024-12	中山大学	图像, 文本	仅微调	搜索方法	数学推理
EMMA-X ^[61]	2024-12	新加坡科技设计大学	图像, 文本	—	—	具身智能
AR-MCTS ^[62]	2024-12	中国人民大学	图像, 文本	—	树搜索	多领域
Mulberry ^[63]	2024-12	南洋理工大学	图像, 文本	仅微调	树搜索	多任务
LlamaV-o1 ^[64]	2025-1	MBZUAI	图像, 文本	仅微调	集束搜索	多任务
MVoT ^[65]	2025-1	微软	图像, 文本	仅微调	—	空间推理
CoR ^[66]	2025-1	清华大学	图像, 文本	—	—	数学推理
DeepSeek-R1 ^[22]	2025-1	深度求索	图像, 文本	微调+强化学习	—	多模态推理
SpatialCoT ^[67]	2025-1	华为	图像, 文本	仅微调	—	空间推理
AStar ^[68]	2025-2	清华大学	图像, 文本	—	树搜索	多模态推理
LLaVA-CoT ^[56]	2025-2	清华大学	图像, 文本	仅微调	集束搜索	多模态推理
MedVLM-R1 ^[69]	2025-2	伦敦帝国理工学院	图像, 文本	强化学习	—	医学影像分析
Visual-RFT ^[70]	2025-3	上海交通大学	图像, 文本	强化学习	—	视觉感知
Audio-Reasoner ^[71]	2025-3	南洋理工大学	音频, 文本	监督微调	—	音频推理
R1-Omni ^[72]	2025-3	阿里巴巴	视频, 音频, 文本	强化学习	—	情感识别
Vision-R1 ^[73]	2025-3	华东师范大学	图像, 文本	微调+强化学习	—	数学推理
DriveLMM-o1 ^[74]	2025-3	MBZUAI	图像, 文本	仅微调	—	自动驾驶
EmbodiedVSR ^[75]	2025-3	北京人形机器人创新中心	图像, 文本	—	—	具身智能
ICoT ^[76]	2025-3	苏州大学	图像, 文本	—	思维链提示	视觉推理
R1-Onevision ^[32]	2025-3	浙江大学	图像, 文本	微调+强化学习	—	多模态推理
Embodied-Reasoner ^[77]	2025-3	浙江大学	图像, 文本	—	—	具身智能

维链机制,并进一步拓展初始链式结构,衍生出多种推理路径的拓扑变体^[20],典型的有链结构^[16]、树结构^[39-40]和图结构^[78-79]。这些结构将复杂的推理任务分解为一系列更简单可控的中间步骤,降低错误累积风险。此外,以不同方式组织推理,也提升了模型的多路径思考能力和灵活性。

如图 3 所示,本文对比了常规多模态大模型以及文本思维链和多模态思维链的推理范式。思维链是

最基本的推理结构,它按照线性顺序逐步生成中间步骤,中间推理结果直接影响下一步骤。Wang 等^[18]在此基础上提出 CoT-SC,引入多条独立推理链,并选择评分最高的答案。

在多模态场景中,多数工作仍沿用初始的单条思维链^[32,48,76],其将隐性的跨模态对齐过程显性化,通过中间语言描述为视觉感知和语言逻辑搭建桥梁,从而缓解模态间的表示鸿沟。一些工作在此基础上融入多

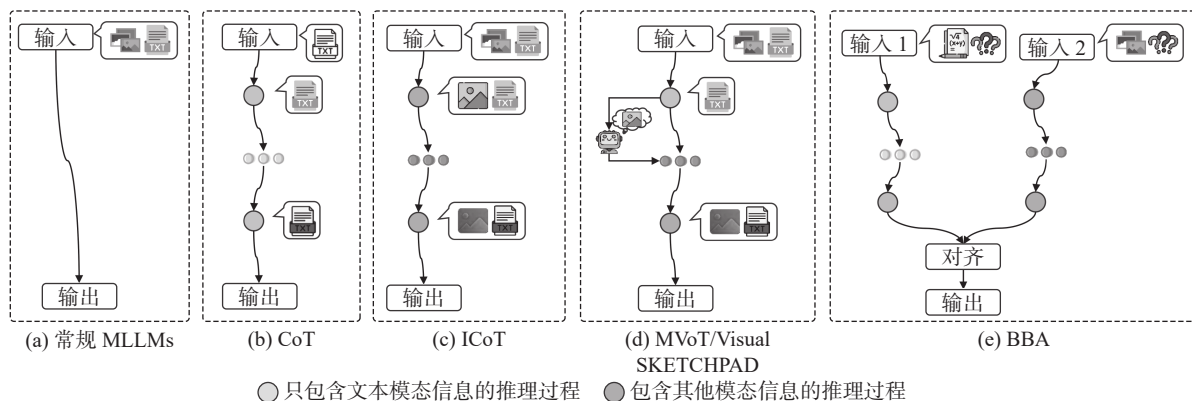


图3 常规多模态大模型与思维链和多模态思维链的推理范式对比

(a) 常规多模态大模型的端到端“黑盒式”推理, (b) 文本思维链推理, (c) ICoT 方法将图像作为推理内容的一部分, (d) MVoT 和 Visual SKETCHPAD 通过生成图像辅助模型进行推理, (e) BBA 方法为文本模态和视觉模态构建不同的推理链并进行对齐整合

模态元素以进一步增强这一桥梁。例如, CoS 方法^[48]强调图像的关键区域感知以增强视觉推理, ICoT^[76]则将图像加入中间推理步骤。此外, MVoT^[65]和 Visual SKETCHPAD^[55]则通过生成推理过程的可视化图像辅助模型推理。

其他方法, 如 CoCoT^[45]通过让模型比较多个图像输入之间的相似性和差异性来指导模型回答多图像问题。BBA^[47]采用与 CoT-SC 类似的策略, 为视觉模态和领域特定语言表示分别构建独立推理链, 而后对其进行对齐整合。VisualCoder^[80]则将思维链推理与可视化控制流图相结合, 用于代码推理任务。为缓解过度思考并使输出结构化, Xiang 等提出 SCoT^[81], 其能在保证推理效率的同时生成各种原子步骤。对于视频和语音等其他模态, VoT^[51]利用思维链技术实现人类级别的视频推理能力, CoT-ST^[53]则通过引入多模态思维链提升了模型在语音翻译任务中的性能。

然而, 链式思维的拓扑结构比较单调, 模型只能遵循线性推理过程, 一旦某个步骤出错, 错误便会沿着推理链传播、放大。由于无法回溯以进行错误检查, 线性推理范式对复合错误高度敏感, 这限制了模型探索的自由度和深度。

为解决这一问题, 思维树 (Tree of Thoughts, ToT)^[39-40]通过模仿人类解决问题时的树状思维过程赋予模型回溯的能力。在思维树中, 树节点表示局部解, 对于某个节点, 模型为其生成多个子节点 (也即多个可选解) 并对其分别评分, 最后利用广度优先和深度优先等搜索算法决定树的拓展路径。思维树的本质是将单一脆弱的推理路径拓展为一个稳健的可探索的解空间, 并引入主动剪枝和错误修正能力, 从而克服线性链条的复合错误问题。此外, BoT 框架^[82]通过迭代

探索和评估思维树以获得试错经验, 从而得到精确的解决方案。思维树可广泛探索问题的解决空间, 取得局部甚至全局的最优解。

思维图 (Graph of Thoughts, GoT)^[78-79]的拓扑结构更为复杂, 其将问题的概念建模为图节点, 并用边来表示关系。与思维树不同, 思维图允许子节点有多个父节点, 即引入了循环和 N -to-1 连接, 这使得模型不仅能够探索不同的推理路径, 还能将多个并行的推理结果进行聚合与协同, 形成全面的解决方案。思维图中的每个子图代表一个子问题的求解过程, 所有子图的结合即为最终解决方案。思维树在深度与广度上探索, 思维图则增加合并的能力, 构建更为灵活和强大的非线性推理结构。在多模态场景中, BDoG^[83]将辩论过程限制在图中, 并将观点和证据等存储在图分支中, 减少了出现频繁但不相关的概念引起的干扰。此外, 思维超图 (Hypergraph-of-Thought, HoT)^[41]通过引入超边对高阶关系进行建模, 进一步增强了对复杂跨模态语义关系的对齐与整合能力。

2.2 后训练阶段

多模态大模型在预训练阶段主要学习世界知识并对齐不同模态, 实现基础的跨模态理解。然而多数 CoT-MLLMs 并不从头预训练, 而是直接在已有模型的基础上引入显式推理机制。因此, 后训练阶段对多模态思维链建模尤为关键, 其核心目标是引导模型学习链式思维的推理模式与输出格式, 并进一步优化多步推理过程。这一阶段的常用方法包括监督微调 (Supervised Fine-tuning, SFT) 和强化学习 (Reinforcement Learning, RL)。

2.2.1 训练数据

微调 CoT-MLLMs 需要专门的数据集, 其为思考过程提供直接的监督信号, 而不仅仅为模型输出提供

监督。为填补这类数据的空缺, 一些研究建立了特定领域内带有推理过程的微调数据集, 本节在表 2 中对相关数据集进行了整理总结。

表 2 用于多模态思维链推理模型训练的数据集

数据集	年份	任务	领域	包含模态	形式	规模
COCO-MMR ^[84]	2023	视觉问答	多领域	文本, 图像	开放式问答	62 351
MM-PhyQA ^[50]	2024	物理问答	物理	文本, 图像	开放式问答	4 500
HaloQuest ^[85]	2024	视觉问答	模型幻觉	文本, 图像	开放式问答	7.7 K
MathV360K ^[86]	2024	数学问答	数学	文本, 图像	开放式问答	360 K
CMM-Math ^[87]	2024	数学问答	数学	文本, 图像	多项选择、填空及开放式问答	28 K
Visual CoT ^[88]	2024	视觉问答	多领域	文本, 图像	开放式问答	438 K
GPSM4K ^[89]	2024	数学问答	几何	文本, 图像	开放式问答	4 440
AMATH-SFT ^[60]	2024	数学问答	数学	文本, 图像	多项选择及开放式问答	157 K
LLaVA-CoT-100k ^[56]	2024	视觉问答	常识, 科学	文本, 图像	多项选择及开放式问答	834 K
MLRQA ^[90]	2025	视觉问答	多领域	文本, 图像	多项选择	4 356
MM-Verify ^[91]	2025	数学问答	数学	文本, 图像	多项选择及开放式问答	59 772
CoTA ^[71]	2025	音频问答	多领域	文本, 音频	开放式问答	1.2 M
Vision-R1-cold ^[73]	2025	视觉问答	多领域	文本, 图像	多项选择及开放式问答	200 K
R1-Onevision ^[32]	2025	多领域	多领域	文本, 图像	多项选择及开放式问答	155 K
LongPerceptual Thoughts ^[92]	2025	视觉问答	视觉感知	文本, 图像	多项选择	30 K
DriveLMM-o1 ^[74]	2025	视觉问答	自动驾驶	文本, 图像, 点云	多项选择, 开放式问答	18 K
Emma-X ^[61]	2024	机器人操作	具身智能	文本, 视频	机器人动作	60 K
Embodied-Reasoner ^[77]	2025	机器人操作	具身智能	文本, 图像	多项选择, 开放式问答	9 390

COCO-MMR 数据集^[84]从 COCO 数据集^[93]中导出大量问题、推理过程和答案, 并使用开放式问题训练和评估模型的思维链推理能力。MM-PhyQA^[50]包含结构良好的高中物理题, 促进模型在多步物理推理任务中的性能提升。HaloQuest^[85]则以超过 7.7K 的数据范例助力缓解多模态推理模型的幻觉问题。MathV360K^[86]、AMATH-SFT^[60]为多模态数学推理提供了高质量数据。此外, CMM-Math^[87]提供了一个中文数学推理数据集, GPSM4K^[89]以一致的格式为几何推理提供详细的分步解答。Visual CoT^[88]、LLaVA-CoT-100K^[56]、Vision-R1-cold^[73]以及 R1-Onevision^[32]等为多个领域的视觉推理任务提供问答数据, 且包含详细的推理步骤。MLRQA^[90]包含由领域专家编写的逻辑推理数据, Sun 等^[91]提出两种有效的多模态思维链数据合成方法, 为生成大规模的多模态思维链数据提供参考。对于音频数据, CoTA^[71]提供了包含结构化思维链过程的多任务音频数据。

由于当前数据集中包含数学和代码等硬推理场景, 一些研究开始探索其他应用场景。这一拓展至关重要, 因为它验证了思维链并不仅仅是一种适用于数理逻辑的特殊技巧, 更是一种具有广泛适用性的通用框架。LongPerceptual Thoughts^[92]致力于填补视觉感知任务中推理数据的空白, Emma-X^[61]和 Embodied-Reasoner^[77]为具身智能模型提供了训练数据, DriveLMM-o1^[74]则为自动驾驶的逐步视觉推理提供视觉问答训练集实例, 涵盖感知、预测和规划等不同场景。

2.2.2 训练范式

如图 4 所示, 多模态大模型在后训练阶段可通过仅监督微调、仅强化学习以及监督微调+强化学习三种训练范式获得思维链推理能力。以下对其进行逐一介绍, 并结合常规多模态大模型的训练流程和目的进行对比说明。

(1) 仅监督微调

如图 4 (a) 所示, 该范式通过大规模思维链标注数

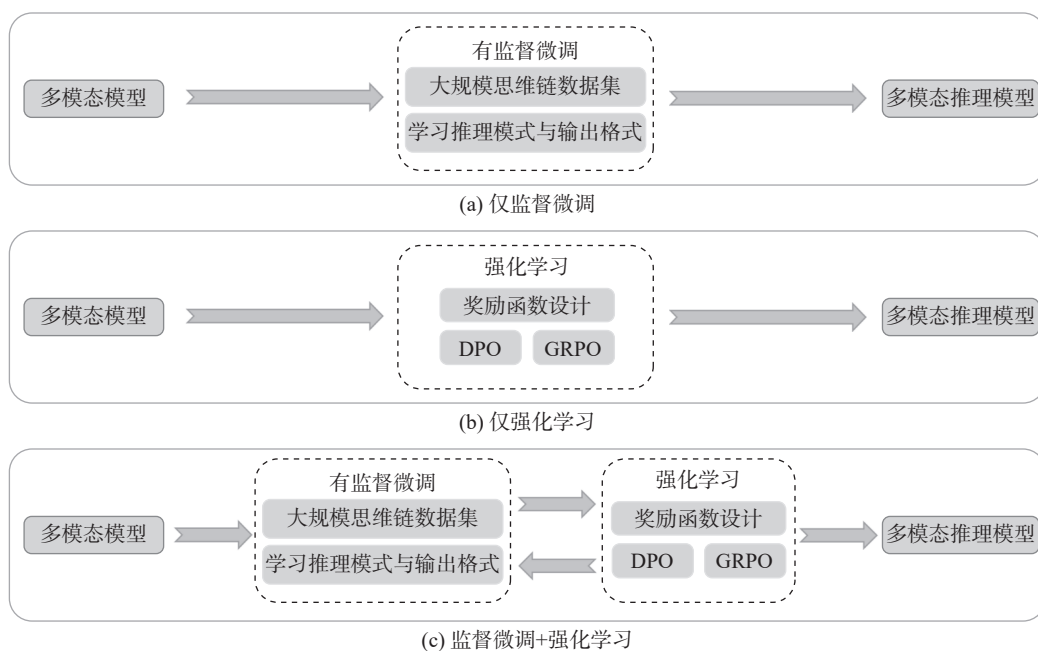


图4 多模态思维链推理模型的三种训练范式

据对模型进行微调,模型在训练过程中学习从处理输入数据到逐步推理再到最终输出的推理模式。与常规多模态大模型通过监督微调方法提升模型泛化能力不同,多模态思维链中的监督微调更关注模型能否生成符合逻辑结构的推理过程,因此其更加依赖数据中思维链结构的质量与多样性。这种方法训练稳定、高效,但在数据不足或问题类型复杂时泛化能力有限。

为解决这一局限并提升学习效率,研究者参考人类学习过程,引导模型逐步获得推理能力。例如, Llamav-o1^[64]采用多步骤对课程学习方式进行训练,模型先学习子任务,然后再训练多步推理能力。类似地, CoR^[66]、LLaVA-CoT^[56]采用渐进式训练方法,通过分阶段引入不同类型的推理数据,使模型能逐渐拓展其推理能力。这种由易到难的训练策略能帮助模型更稳健地构建复杂推理能力,避免因任务难度过高而导致的训练失败。

(2) 仅强化学习

如图4(b)所示,该范式跳过监督微调阶段,直接对预训练模型进行强化微调。其从模仿学习转向目标驱动的探索学习,模型不再被动复制数据中的固定路径,而是主动探索各种可能的推理路径,并通过奖励信号来判断路径优劣,从而发现数据中不曾出现过的更优解。

与常规多模态大模型使用基于人类反馈的强化学习(Reinforcement Learning from Human Feedback, RLHF)以对齐人类偏好、减少模型幻觉不同,多模态思维链

的强化学习更注重提升推理链条的逻辑性、完整性与问题求解效果。在这种训练范式中,使用直接偏好优化(Direct Preference Optimization, DPO)^[94]算法或群体相对策略优化(Group Relative Policy Optimization, GRPO)^[95]的强化微调(Reinforcement Fine-Tuning, RFT)方法对模型进行强化学习。

例如, Seg-Zero^[96]采用五种强化学习奖励函数,在不依赖任何监督推理数据的情况下使模型获得相当水平的推理能力。Vision-R1^[73]仿照 DeepSeek-R1^[22]中的冷启动策略,并提出渐进式思维抑制训练策略缓解冷启动后过度思考的问题,其在强化学习的早期阶段抑制推理长度,并逐渐放宽限制,使模型能够自主学习并利用更长的思维链来解决复杂问题。此外, Visual-RFT^[70]证明了强化微调方法可用于数学和代码领域之外的视觉感知任务,并将基于可验证奖励的强化学习(Reinforcement Learning with Verifiable Rewards, RLVR)用于图像多模态模型。而 R1-Omni^[72]则首次将 RLVR 应用于视频多模态模型。

(3) 监督微调+强化学习

如图4(c)所示,该范式将监督微调与强化学习相结合,分阶段训练模型。这种混合范式在理论上最为完备,它首先通过监督微调方法使模型掌握基础的推理模式,再借助强化学习过程优化推理路径,使模型在合理性和任务适配性等方面获得进一步提升。

ARES^[54]提出一种两阶段算法,交替使用强化学习和监督微调,首先让教师模型进行反馈打分并进行强化学习,而后使用教师模型纠正错误的推理过程,

并进行监督微调以稳定强化学习后的模型表现。此外, Insight-V^[58]在强化学习阶段进行多轮 DPO 训练和采样,使模型更好地模拟在线 DPO 过程,解决了离线生成的偏好数据逐渐偏离模型分布的问题。为增强模型性能的稳定性, Wang 等^[97]提出混合偏好优化 (Mixed Prefer Optimization, MPO) 方法,其损失函数包含偏好损失、质量损失和生成损失三部分,引导模型在优化某一特定目标的同时兼顾其他重要属性,从而培养出更全面、更鲁棒的推理能力。

2.3 推理阶段

在模型训练阶段,监督微调和强化学习等方法通过调整模型参数使其具备思维链推理能力。而在模型推理阶段,存在无须更新模型参数的方法与策略,其意图不在于教会模型新知识,而是通过巧妙引导激活其潜在推理能力。这类方法包括思维链提示、搜索策略、自我优化、知识增强以及智能体辅助等推理时缩放(Inference-Time Scaling)方法,我们在图 5 中对其进行了总结。

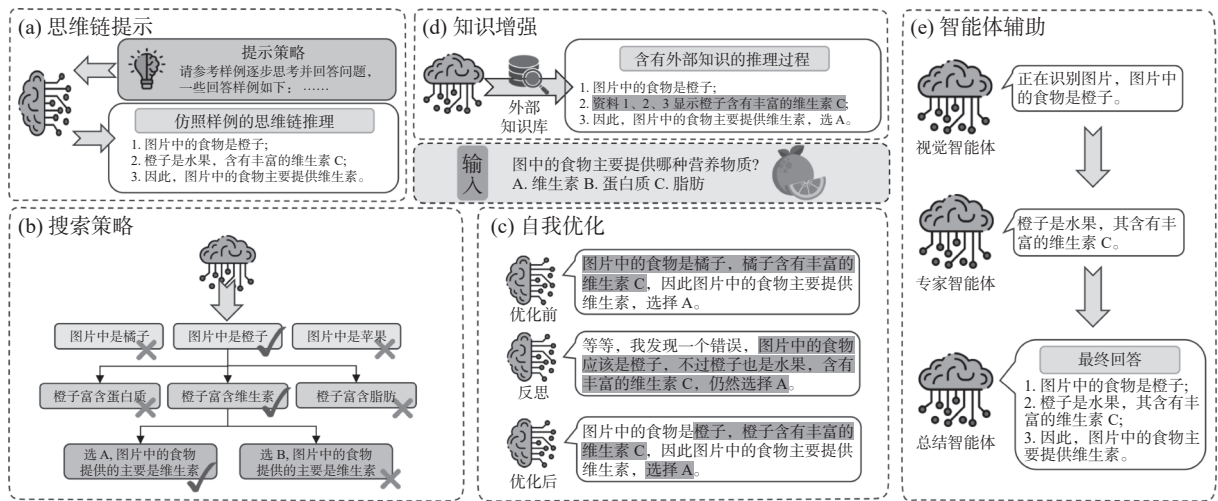


图 5 多模态思维链推理模型的推理时缩放方法总结

2.3.1 思维链提示

思维链提示 (Chain-of-Thought Prompting) 是一种低成本且泛化能力强的方法,如图 5 (a) 所示,其通过少样本或零样本提示策略显式地提供推理样例,并利用语境学习 (In-Context Learning, ICL) 设定临时的模型推理范式,引导模型将复杂的单步预测任务分解为一系列简单自治的生成步骤。

Wei 等^[16]提出在 LLM 中引入思维链提示方法,提高了模型在算数、常识和逻辑推理任务上的性能。Kojima 等^[17]进一步提出零样本思维链 (Zero-shot CoT) 提示,仅需在提示词中添加“让我们一步步思考”的指令,就能引导模型进行思维链推理。

在多模态场景中,思维链提示的核心挑战在于提供引导框架,协同处理和转化不同模态信息以跨越表示鸿沟。BBA 方法^[47]为视觉表示和领域特定语言表示创建独立推理链,而后对其进行对齐整合。CCoT^[49]首先利用模型生成场景图,再将其嵌入到提示词中生成回答。Zhou 等^[52]的 IoT 提示方法引导模型根据图像和问题自动设计视觉信息提取操作,为每一步推理获

取视觉辅助信息,使模型从被动的信息接收者转变为主动获取信息的智能体。此外,DDCoT^[42]、MI-CoT^[50]和 ICoT^[76]方法通过在提示词中加入图像指示,强制模型在每一步推理时回溯检验视觉证据,实现语言符号与视觉信息的持续接地,有效抑制了模型幻觉。

2.3.2 搜索策略

大模型作为一种自回归生成模型,在推理时遵循贪心策略,每一步都选择局部最优的下一词元。这种决策机制容易导致复合错误,即早期的微小偏差在后续步骤中被不断放大,最终偏离正确推理轨道。如图 5 (b) 所示,搜索策略以增加计算为代价,将单一脆弱的推理路径拓展为可探索的解空间,从而在全局寻找更优的推理路径。常见搜索策略包括集束搜索 (Beam Search)、蒙特卡洛树搜索 (Monte Carlo Tree Search, MCTS) 和最优 N (Best-of-N, BoN) 搜索^[98-99]等。

集束搜索 作为一种启发式搜索算法,其在每一步生成过程中保留固定数量 (称为“束宽”) 的概率最高的候选序列,并在后续步骤中进行扩展。它是介于贪心搜索和穷举搜索之间的实用折中,通过有限的并

行探索缓解了单一路径的脆弱性。然而,其可能因为启发式假设的短视而错失全局最优解。在多模态场景中,为平衡搜索的粒度和复杂度,Xu 等^[56]提出阶段级束搜索方法,利用 LLaVA-CoT 的阶段输出特点执行搜索。

蒙特卡洛树搜索 MCTS 是一种基于随机抽样和树状搜索的决策算法,其主要包括选择、拓展、模拟和回溯更新四个步骤。相较于集束搜索,MCTS 的核心优势在于其非对称的树拓展策略,它能够智能分配计算预算,更深入地探索有潜力的推理分支,充分体现了探索-利用的权衡。

针对多模态思维链推理路径扩展的可靠性问题,AR-MCTS^[62]集成 MCTS 和主动检索机制,实现了自动生成分步标注,确保搜索过程逻辑连贯且有事实依据。类似地,Yao 等^[63]提出 CoMCTS,将集体学习引入树搜索,利用多个模型的集体知识来协同猜想、搜索和识别有效推理路径,从而平滑单一模型的认知偏差,提升搜索覆盖率和找到最优解的概率。

为平衡 AR-MCTS 和 CoMCTS 的性能和效率问题,Wu 等^[68]提出 AStar,将搜索对象从低维的词元序列提升至高维、预定义的认知动作层面,以更高效率推导出高级认知推理模式。

最优 N 搜索 BoN 的思路是模型先生成 N 个可能输出,再根据预设评估标准从中选出最优解。该方法通过增加候选解的数量提高获得高质量输出的可能性。Speculative Rejection 方法^[100]通过停止生成评分较低的候选回答,降低了模型对计算资源的需求。TreeBoN 方法^[101]融合推测式树搜索策略与 BoN 方法,通过迭代拓展分支并修剪低质量回答,实现了高输出质量与高效推理的平衡。

其他搜索方法 一些工作综合采用多种搜索策略来提升搜索的稳定性和准确性,或提出新的过程奖励模型(Process Reward Model, PRM)指导树搜索。例如,AtomThink 框架^[60]结合四种搜索策略和奖励模型,并将搜索分为路径搜索和逐步搜索。为生成具有更好视觉理解的模型回答,Wang 等^[59]提出 VisVM 作为信号来引导模型进行推理时搜索,通过提供长期价值避免局部最优。

2.3.3 自我优化

一些工作^[102-104]表明,大模型具备自我修正能力,即模型可以通过反思提升回答质量。如图 5(c) 所示,自我优化将模型的生成过程从线性的“思考-输出”模式,转变为“思考-输出-反思-修正”的迭代循环,在不更新模型参数的情况下,通过引入内部反馈机制优化

推理路径。

在多模态推理中, R^3V ^[105]引入自我修正和自我选择,使模型能够修正有缺陷的推理过程。EVLm^[106]则引入包含初始推理、中间输出生成和反思优化的多步反思推理框架,支持模型通过迭代推理内化和泛化变换原理。InfiGUIAgent^[107]利用期望-反射推理的循环过程增强 GUI 智能体的行为一致性和自我修正能力。

2.3.4 知识增强

如图 5(d) 所示,近期一些工作通过引入外部知识增强 LLM 的推理能力。这类方法旨在克服模型知识更新不及时与模型幻觉问题,使模型能利用更准确、更具时效性的外部事实来印证其推理过程。RAGAR^[108]和 RMR^[109]利用检索增强生成技术在推理过程中融入特定领域知识以增强模型推理能力。MR-MKG 方法^[110]通过引入多模态知识图谱,跨模态地学习丰富的语义知识,显著增强了多模态大模型的推理能力。

在思维链推理场景中,KAM-CoT 框架^[46]在推理过程中整合了思维链、知识图谱和多模态信息,实现了更深入的上下文理解,进而减少模型幻觉并提升回答质量。

2.3.5 智能体辅助

多智能体协作将复杂的认知任务分解为多个专业化的子角色,如图 5(e) 所示,每个大模型(智能体)被分配特定角色,并通过与其他智能体的结构化交互来迭代地改进其输出。

在多模态推理中,Elhenawy 等^[111]提出 Multi-Agent 1 和 Multi-Agent 2 解决旅行商问题和多旅行商问题,两种多智能体框架均显著提高了模型的问题求解质量。Insight-V^[58]将问题求解过程分解为推理和总结两个步骤,降低了单个智能体的认知负荷,从而提升了整体效率和产出质量。此外,MultiMedRes^[112]、InfiGUI-Agent^[107]和 Zhai 等^[113]的工作利用单一智能体来增强模型的视觉推理能力。

2.4 理论机制分析

在前面几节中,本文系统梳理了多模态思维链的实现路径,即“如何做”,并对应解释了其内在机理。为了更系统全面地回答多模态思维链“为什么有效”这一核心问题,本节将深入分析其运作的三个关键环节,包括信息表征、结构化推理、过程监督与优化。

2.4.1 信息表征

多模态大模型在融合异构模态信息时,面临模态表示鸿沟的根本挑战,即如何对齐视觉信号与语言表达,这直接影响模型后续的推理结果。

对于这一问题,思维链通过强制模型生成中间文本,将视觉感知过程显式化。这一过程创造了一个共享语义空间,更为推理步骤提供了回溯检验视觉证据的锚点,实现了语言表达与视觉事实的持续接地,从而减少表示鸿沟。具体而言,诸如 ICoT^[76]等方法将图像本身作为推理步骤的一部分,而 CoS^[48]则引导模型聚焦于图像中的关键区域, IoT^[52]方法更是让模型能够主动规划视觉信息的提取动作。这些方法均验证了显式的中间步骤能够增强跨模态理解与对齐,提升模型回答的事实一致性,进而抑制模型幻觉问题。

2.4.2 结构化推理

思维链重塑了模型推理范式,其能够降低模型在单一决策点上的认知负荷,并抑制其在自回归生成中的复合错误。这种范式促使模型从快速、直觉式的“系统1”思维模式,切换至更缓慢、逻辑严谨的“系统2”思维模式。正如 Kojima 等^[17]所证实的,简单的指令就能激活模型的潜在推理能力。更进一步,这种显式、模块化的推理结构,为复杂的计算策略提供了必要基础。它为思维树^[39-40]、思维图^[78-79]等高级拓扑结构以及集束搜索^[56]、蒙特卡洛树搜索^[62-63]等搜索策略开辟了可探索的解空间,并为 Insight-V^[58]等智能体辅助框架提供协同推理基础。

2.4.3 过程监督与优化

在训练和推理中,思维链范式改变了模型监督优化的方式。传统多模态模型采用结果监督方式,即只对最终答案进行评估,而思维链的显式推理步骤则为更具稳健性的过程监督提供了形式基础,其监督信号贯穿训练与推理的全过程。

在后训练阶段,监督的重点从结果正确转移到过程合理, Visual CoT^[88]和 LLaVA-CoT-100K^[56]等数据集为模型提供了详细的推理路径,使其能通过监督微调进行高效的行为克隆,或通过强化学习基于奖励信号进行目标导向型优化。

在推理阶段,监督机制则更动态、更即时。思维链提示^[16-17]作为一种即时监督方法,能高效地引导模型进行结构化推理。而 VisVM^[59]提出的过程奖励模型,以及 Elhenawy 等^[111]采用的批判者智能体,则构成了外部动态监督信号。与此同时,以 R³V^[105]和 EVLM^[106]为代表的自我优化机制,实现了更深层的内部动态监督。无论监督信号来自外部还是内部,都对模型的推理行为进行了更细致的实时校准,进而提升模型的推理性能。

3 评测:如何评价多模态思维链推理能力?

多模态思维链推理的形式比较特殊,因此需要专门的评价基准来对模型进行性能评估。本节对现有的基准进行了全面调查和总结,表3是对这些基准的总结。此外,我们在图6中对这六类评价基准进行了直观的对比总结,并对每种类型的评价基准各举了一些例子。

3.1 数学与逻辑推理

如图6(a)所示,此类基准评估模型具有对数学及逻辑关系的理解与推理能力。它们要求模型从视觉输入中提取信息并进行形式化推理,不仅衡量结果的准确性,更关注模型推理过程的合理性。

表3 多模态思维链的评价基准总结

基准	年份	任务	包含模态	指标	形式	规模
数学与逻辑推理						
MathVista ^[114]	2024	数学问答	文本, 图像	准确率	多项选择, 开放式问答	6 141
MATH-Vision ^[115]	2024	数学问答	文本, 图像	准确率	多项选择, 开放式问答	3 040
We-Math ^[116]	2024	数学问答	文本, 图像	自定义指标	多项选择	6.5 K
LogicVista ^[117]	2024	逻辑推理	文本, 图像	准确率	多项选择, 开放式问答	448
MathVerse ^[118]	2024	数学问答	文本, 图像	准确率+自定义指标	多项选择, 开放式问答	15 672
ErrorRadar ^[119]	2024	错误检测	文本, 图像	准确率	开放式问答	2 500
CMM-Math ^[87]	2024	数学问答	文本, 图像	准确率	多项选择, 开放式问答	28 069
MuCR ^[120]	2025	因果推理	文本, 图像	自定义指标	开放式问答	—
EnigmaEval ^[121]	2025	谜题解决	文本, 图像	准确率	开放式问答	1 184
MME-Reasoning ^[122]	2025	逻辑推理	文本, 图像	自定义指标	多项选择, 开放式问答, 基于规则的问答	1 188

续表

基准	年份	任务	包含模态	指标	形式	规模
时空与方向推理						
CDR ^[123]	2024	方向推理	文本, 图像	准确率	多项选择	100 K
GSR-Bench ^[124]	2024	空间推理	文本, 图像	准确率	多项选择	4 958
TOMATO ^[125]	2024	时序推理	文本, 视频	自定义指标	多项选择	1 484
PulseCheck457 ^[126]	2025	空间推理	文本, 图像	自定义指标	开放式问答	—
DriveLMM-o1 ^[74]	2025	自动驾驶	文本, 图像, 点云	自定义指标	多项选择, 开放式问答	—
视觉语言转换推理						
Plot2Code ^[127]	2024	视觉编码	文本, 图像	自定义指标	开放式问答	132
P ² G ^[128]	2024	视觉推理	文本, 图像	准确率	多项选择	2 130
HumanEval-V ^[129]	2025	视觉编码	文本, 图像	pass@k	开放式问答	100 K
Plot2XML ^[130]	2025	视觉编码	文本, 图像	自定义标准	开放式问答	247
序列与多图像推理						
MIMU ^[131]	2025	多任务	文本, 图像	准确率	开放式问答	—
多模态综合推理						
NPHardEval4V ^[132]	2024	视觉问答	文本, 图像	自定义指标	开放式问答	—
M ³ CoT ^[24]	2024	视觉问答	文本, 图像	准确率	多项选择	11 459
EMMA ^[133]	2025	视觉问答	文本, 图像	pass@k	多项选择, 开放式问答	2 788
VRC-Bench ^[64]	2025	多任务	文本, 图像	自定义指标	多项选择, 开放式问答	—
MME-CoT ^[134]	2025	视觉问答	文本, 图像	自定义指标	多项选择, 开放式问答	1 130
R1-Onevision-Bench ^[32]	2025	视觉问答	文本, 图像	准确率	多项选择, 开放式问答	942
VisualPuzzles ^[135]	2025	视觉问答	文本, 图像	准确率	多项选择	1 168
推理过程质量评估						
MiCEval ^[136]	2025	视觉问答	文本, 图像	自定义指标	开放式问答	2 130
MMIR ^[137]	2025	一致性推理	文本, 图像	准确率	多项选择, 开放式问答	534

MathVista^[114]、MATH-Vision^[115]、We-Math^[116]和CMM-Math^[87]等基准将视觉信息融入数学推理任务中。而 MathVerse^[118]致力于评估带图表的数学问题, 并使用 GPT-4(V) 对输出答案进行细粒度评估。ErrorRadar^[119]则引入多模态错误步骤识别和错误分类任务, 为评估多模态大模型的复杂数学推理能力提供全面框架。

此外, LogicVista^[117]致力于评估模型在视觉环境中的综合推理能力, MME-Reasoning^[122]全面评估模型的逻辑推理能力, 包括归纳、演绎和溯因三种类型。EnigmaEval^[121]通过给出谜题来探究模型执行隐性知识综合和多步骤演绎推理的能力。MuCR^[120]利用合成的孪生图像和文本对来评估模型识别多模态信息间因果联系的能力。

3.2 时空与方向推理

此类基准旨在评估模型对空间和时序概念的理解与推理能力。如图 6 (b) 所示, 这些任务要求模型从视觉场景中推断物体间的空间拓扑关系、理解时间维度的事件连续性并解析方向性指令。

PulseCheck457^[126]提供了全面评估 6D 空间推理的框架, 包含从基本单物体识别到复杂 6D 空间推理任务。GSR-Bench^[124]扩展了 What'sUp 数据集, 并提出一种综合评估方法来理解空间关系。CDR^[123]旨在评估模型的方向推理能力, 包含测试空间关系和罗盘方向的图像。在时序推理方面, TOMATO^[125]提供了评估模型在视频理解中时间推理能力的基准。在自动驾驶场景中, DriveLMM-o1^[74]提供了包含超 4 000 个视觉

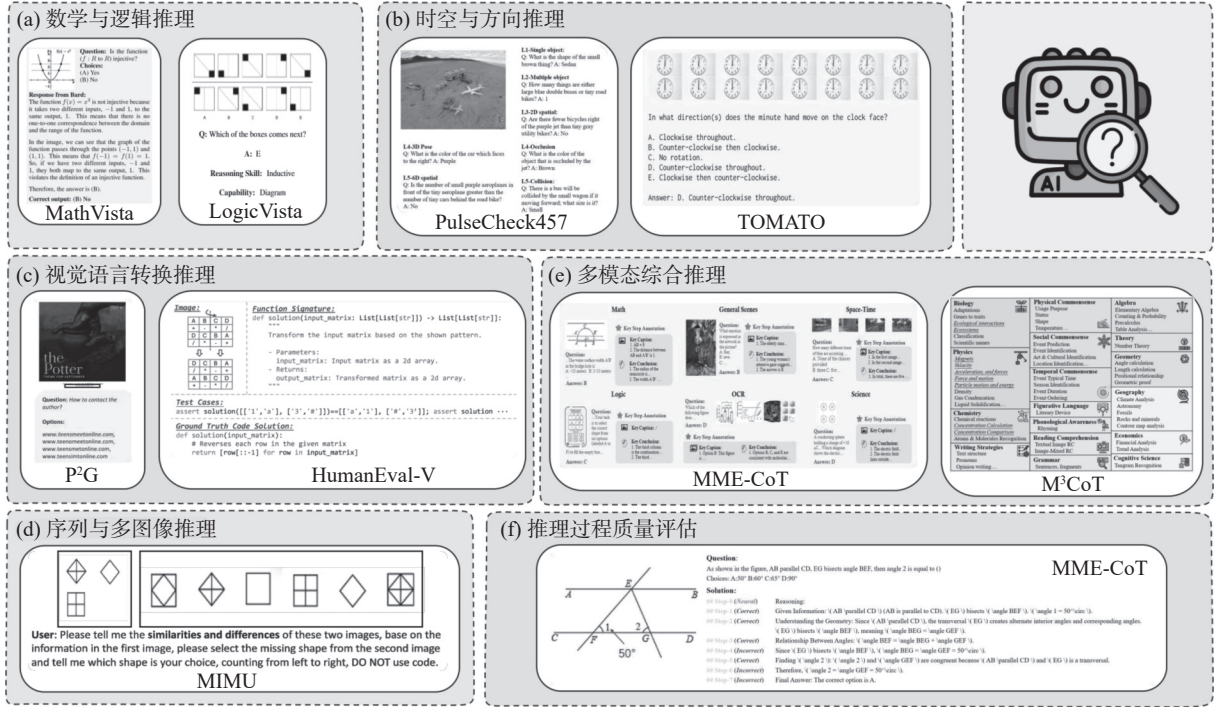


图6 多模态思维链的六类评价基准

问答实例的基准, 涵盖感知、预测和规划等任务。

3.3 视觉语言转换推理

如图6(c)所示, 此类基准旨在评估模型将视觉内容转换为结构化语言(如代码、文本描述)的能力。这些任务检验模型跨模态理解、信息提取与表征转换的能力, 要求模型建立视觉元素与语言符号之间的语义映射关系。

HumanEval-V^[129]采用代码生成任务来评估模型的图表理解能力, 其中每个子任务都包含图表以及对应的函数签名和测试用例。Plot2Code^[127]包含 matplotlib 图库中的图片, 并对应给出源代码以及 GPT-4 总结的描述指令。类似的, Plot2XML^[130]提供了将科学图表重构为 XML 代码的评价基准。P2G^[128]则提供了对模型在富文本或高分辨率图像场景下的视觉推理能力的定量评估。

3.4 序列与多图像推理

此类基准旨在评估模型处理时序数据和多图像输入的能力。如图6(d)所示, 这些任务要求模型在多个视觉输入间建立关联, 理解图像序列中的事件演变并识别相关模式。MIMU^[131]用于评估多图像输入场景中的模型性能, 注重模型对细粒度视觉细节的感知能力, 并集中在图像到图像的匹配和多图像到文本的匹配这两类场景中。

3.5 多模态综合推理

如图6(e)所示, 此类基准旨在全面评估模型在跨领域多模态任务上的思维链推理能力。这些基准要求模型同时处理多模态输入、整合异构信息, 并在复杂情境中应用推理能力。

MME-CoT^[134]是评估多模态思维链推理能力的首个综合基准, 涵盖数学、科学、光学字符识别、逻辑、时空和一般场景六个领域。M3CoT^[24]则在思维链的多领域、多步骤和多模态场景中迈出了第一步。EMMA^[133]和 VRC-Bench^[64]同样为评估多模态思维链提供了跨模态、跨领域的多步骤基准。此外, R1-Onevision-Bench^[32]横跨数学、物理、化学、生物和逻辑推理等领域, 涵盖初中及以上水平的考试内容。而 NPHardEval4V^[132]旨在提供一个场景, 在其中分离出图像识别和指令跟随等各种影响模型性能的因素, 使其能专注于评估模型的推理能力。类似的, VisualPuzzles^[135]通过减少所需的领域特定知识并增加推理复杂性, 使其能专注于评估模型的多模态推理能力。

3.6 推理过程质量评估

此类基准着重于评估模型推理过程的质量, 如图6(f)所示, 这些任务要求模型提供推理链、解释决策过程并保持逻辑一致性。

MPBench^[138]使用步骤正确识别、答案聚合和推理

过程搜索三种评价范式来系统评估过程奖励模型的有效性。MiCEval^[136]通过评估模型对图像的描述准确性和推理步骤的质量来评估推理链的正确性。此外, MMIR^[137]和 Jia 等^[139]的工作着眼于评估模型的推理不一致问题, 为多模态大模型的未来改进方向提供了有价值的意见。

3.7 评价指标

目前多数工作^[60,64,120,125,127]通过调用 LLM 和 MLLM 来评估模型回答的准确性, 并以此作为模型思维链推理能力的指标, 或提出几个适用于特定任务的指标, 但这些指标存在评价不全面、不准确的问题。基于这些问题, MME-CoT^[134]率先提出用于评估多模态大模

型思维链推理能力的具体方案, 其中包括思维链质量、思维链鲁棒性和思维链效率三个方面的评价。此外, MMR^[140]引入一个模块化的推理步骤评估流程, 从与问题的相关性、与答案的相关性和推理步骤的一致性三个方面对推理步骤进行综合质量评估。

4 应用: 多模态思维链可用于哪些领域?

多模态思维链通过引入逐步推理机制, 有效提升了模型在复杂任务中的推理能力与可解释性。基于这一优势, 其在具身智能、自动驾驶、医疗健康、多模态生成、机器翻译和情感计算等领域展现出广阔应用前景, 本文在图 7 中进行了总结。



图 7 多模态思维链的应用领域

4.1 具身智能

具身智能 (Embodied Intelligence) 作为实现通用人工智能的重要路径, 旨在将智能系统置于物理实体 (如机器人) 中, 以赋予其环境感知、自主学习与动态交互的能力。

如图 7 (a) 所示, 多模态思维链显著提升了具身智能在复杂环境中的任务规划、空间推理、导航与操作等核心能力。

在任务分解与操作执行方面, 多模态思维链帮助模型将操作指令或任务目标划分为具体细致的执行步骤。Hao 等^[141]通过引入基础模型增强了人形机器人在非结构化复杂场景中执行局部操作任务的能

力。Embodied Reasoner^[77]则面向交互式搜索任务, 展现了智能体自发进行目标识别、推理判断与行动决策的能力, 并强化其适应性。

在任务规划与高低层协同方面, EmbodiedGPT^[142]提出一种融合高层策略规划与底层控制执行的统一架构, 能够分步骤规划实现复杂任务的执行流程。MCot-Memory 框架^[143]引入基于记忆检索的场景图驱动的思维链机制, 通过保留高置信度经验支持智能体在动态环境中进行长期任务规划。ECot^[144]进一步在视觉-语言-动作模型中引入思维链机制, 使模型能进行多步推理和子任务分解, 并展现出良好泛化能力。

在空间推理任务中, EMMA-X^[61]具备前瞻动作

生成能力与基于事实的思维链推理能力,能有效预测环境变化并对应制定行动策略。SpatialCoT^[67]采用双向空间坐标对齐和思维链空间对齐的两阶段训练范式,实现了空间信息与思维链间的深度协同。EmbodiedVSR^[75]则通过动态场景图显式建模结构化知识,赋予模型在零样本场景中的空间推理能力。此外,NavCoT^[145]通过参数高效地实现领域内训练,低成本地实现自引导导航决策。

此外,COUT^[146]将用户个性、偏好等个体因素纳入自主智能体的学习过程,拓展了具身智能在人机个性化互动方向的应用前景。

综上所述,将多模态思维链与视觉语言模型或视觉-语言-动作模型结合,不仅能提升具身智能体的多步推理能力,也能增强其对不同环境和任务的泛化与适应能力。同时,在系统架构、任务执行效率与个性化方面取得一系列突破,推动具身智能向高效、低成本与个性化的方向发展。

4.2 自动驾驶

近年来多模态大语言模型在自动驾驶(Autonomous Driving, AD)领域展现出巨大潜力,为构建新一代端到端自动驾驶系统提供了新范式。如图7(b)所示,将思维链机制引入自动驾驶系统,不仅能提升系统在复杂交通环境中的认知精度与决策稳健性,还降低了模型的开发与部署成本。

在系统构建方面,Dolphins^[44]作为对话式驾驶助手,基于场景上下文引入多模态思维链,增强其对驾驶指令的理解与执行能力。Agent-Driver^[147]采用大语言模型作为核心认知代理,通过整合通用工具库改造传统自动驾驶流程,实现了从感知、理解到决策的端到端建模。

在数据资源方面,多个研究工作构建了支持多模态推理与端到端驾驶任务的开放数据集。DriveCoT^[148]构建了数据集并使用其训练模型 DriveCoT-Agent,首次将多模态思维链集成进端到端自动驾驶系统。Reason2Drive^[149]提供超 60 万对包含显式推理过程的视频、文本样本,促进可解释自动驾驶推理的研究。DriveLMM-ol^[74]则专注于逐步视觉推理任务,涵盖感知、预测与规划三个环节,为系统评估多模态推理模型提供数据集。

在模型推理框架方面,PKRD-CoT^[150]为自动驾驶中的多模态大模型提供统一的提示词设计框架,引导模型进行实时驾驶决策,增强了系统的适应性与实时性。OpenEMMA^[151]融合思维链推理与外部视觉专家

模型,在汽车轨迹规划任务中表现优异。与此同时,Islam 等^[152]将自动驾驶的决策过程分解为场景理解、行为预测与决策生成三个阶段,并引入思维链连接不同阶段的推理路径,进而提升了整体决策的合理性与连贯性。

综上所述,多模态思维链与自动驾驶系统的深度融合,显著增强了系统对复杂多变交通环境的建模能力与理解能力,并提高了决策的连贯性、可解释性与鲁棒性。这些进展推动自动驾驶技术向实现类人驾驶迈出了关键的一步。

4.3 医疗健康

如图7(c)所示,在医疗健康领域,多模态思维链为融合医学影像和临床文本等多源信息提供了统一的推理框架。它能模拟医生的诊疗思维,为疾病诊断、治疗规划以及个性化健康管理提供更透明、可解释的智能辅助,从而提升人工智能在医疗场景中的实用性与可信度。

针对医学视觉问答任务,MedThink^[153]同时生成决策结果与推理理由来展示推理路径,提高决策的可解释性。MedCoT^[154]引入层级式专家验证推理链,融合多个专家模型进行推理与校正,以获得更可靠且具临床一致性的解答。对于医学影像分析,Chain-of-Look^[43]将内窥镜视频中的三联体识别任务建模为视觉提示生成问题,引入视觉语言模型,并显式地将任务分解为一系列具有时序结构的视频推理过程。同时,MedVLM-R1^[69]通过引入强化学习机制提升视觉语言模型在医学影像分析任务中的推理能力,并通过显式自然语言推理增强推理路径的透明性与可追溯性。

此外,Dai 等^[155]将思维链推理引入压力检测系统,使模型在评估主观性强的心理状态时具备更强的解释力。而 TI-PREGO^[156]结合语境学习与自动思维链,利用动作识别与预测任务对自我中心视角下的医疗视频进行程序性错误检测。

综上所述,在高度依赖决策可解释性与可控性的医疗健康领域,多模态思维链技术展现出巨大潜力。它通过模拟医疗专家的决策逻辑,实现从感知到认知的统一建模,有望为构建可信赖、高性能的人机协同医疗系统奠定基础。

4.4 多模态生成

如图7(d)所示,在多模态生成任务中,多模态思维链引导模型逐步完成跨模态推理与内容生成,显著提升了生成结果的连贯性、逻辑一致性及模态对齐

性。该方法通过模拟人类认知路径,有效缓解了传统生成模型在处理异构模态信息时易出现的模态偏置与信息融合不充分的问题。

MINT 模型^[157]采用“思考—推理—反思”的多阶段思维链范式,显著提高生成图像的语义一致性与视觉质量。在物理一致性建模方面,Phys-AR 框架^[158]结合符号推理与强化学习,构建了具备因果理解能力的多模态生成系统。在生成内容的真实性与可追溯性方面,CoT-Finetuned 模型^[159]将思维链推理用于 AI 生成内容检测,不仅能判断文本是否由 LLM 生成,还能进一步识别具体模型,有效提升了检测系统的判别能力。

综上所述,多模态思维链在生成任务中发挥着多重作用,它优化了跨模态信息融合的路径,提高了生成内容的逻辑与视觉一致性,同时也推动了生成内容的可信性验证与责任归属研究。

4.5 机器翻译

相较于基于神经网络以及基于大语言模型的翻译范式,多模态思维链为机器翻译任务开辟了新路径。如图 7(e) 所示,该范式引入视、语音和环境上下文等外部模态信息作为推理的辅助输入,以更系统地理解源语言语境和文化背景,从而生成更自然、准确且具有文化适应性的目标语言表达。

CoT-ST^[53]利用多模态思维链将语音翻译任务分解为语音识别和翻译的连续步骤,增强了子任务间的协同和互补,显著提升了语言语音模型的翻译性能。IMAGE^[160]结合多模态大模型与 Stable Diffusion 模型,为待翻译句子显式生成图像,并利用图像信息提升模型的机器翻译性能。此外,针对低资源语言的多模态翻译,Rajpoot 等^[161]提出一种结合图像描述与思维链推理的策略,通过多阶段提取图像语义描述信息并嵌入到翻译推理路径中,在英语到低资源印度语种的翻译任务中取得了良好效果。

综上所述,多模态思维链的引入使机器翻译更具人类认知特征,通过显式推理连接语言与多模态信息间的深层语义关系,全面提升了翻译系统的准确性、文化适应性与上下文一致性。

4.6 情感计算与社会推理

如图 7(f) 所示,在情感计算与社会推理等涉及人类认知与行为建模的领域,多模态思维链能有效整合多模态信息,提升模型在情绪识别、情感理解及复杂社会行为建模中的综合表现。其推理机制不仅提供了结构化的决策依据,也增强了模型对人类语境与社

会知识的泛化能力。

在情感计算方面,MM-PEAR-CoT^[162]首次将思维链引入多模态情感识别任务,通过融合文本、视频和音频信息,缓解了单一模态偏误导致的推理偏差,从而取得更稳定、可解释的情绪标签分类表现。此外,Empatheia 系统^[163]融合多模态大模型与共情推理链,在共情响应生成任务中展现出更自然且具有个体适应性的语言生成能力。

在社会推理方面,Social genome^[164]提供了首个评估多模态模型社会推理能力的基准框架,包含细粒度的推理步骤,并引用外部知识辅助评估。Yu 和 Luo^[165]探索了融合思维链提示的多模态大模型在人口统计推断中的应用,显著减少了预测偏差。此外,CPFEND^[166]和 RAGAR^[108]致力于提升模型在虚假信息识别与政治事实核查中的推理与判断能力。

综上所述,上述研究充分展示了多模态思维链推理机制在提升模型结构化认知能力、实现情境感知与社会行为建模方面的巨大潜力。

5 挑战及未来方向:多模态思维链路向何方?

尽管当前多模态思维链已取得一些进展,并表现出良好的发展潜力,但仍面临一些关键问题和挑战。本节总结和介绍了一些挑战和未来发展方向,希望对相关研究者有所启发。

5.1 鲁棒推理

随着多模态推理模型日益复杂,鲁棒性成为衡量其实际应用能力的关键指标。推理鲁棒性体现为模型对输入扰动和任务复杂性等的容忍度,以及持续生成稳定、可信结论的能力。当前模型主要面临幻觉与推理链长度控制两大鲁棒性问题。

LLM 中的幻觉问题往往源于缺乏明确的推理过程,易产生上下文矛盾或事实不准确的回答。虽然思维链推理在一定程度上缓解了幻觉问题,但也存在过度思考等问题。现有幻觉缓解策略多作用于整体回答,存在错误累积问题。未来研究应探索步骤级优化策略,如 HaluSearch^[167]使用 MCTS 将回答生成建模为逐步思考过程,并借助步骤级别的奖励机制探索最优推理路径。

此外,多模态环境中不同模态信息间的不充分对齐也会诱发幻觉。Zheng 等^[57]指出“边看边想”范式(即在接收视觉输入的同时生成推理链)易受误导性视觉信息影响,导致推理结果偏差。因此,多模态思

维链的跨模态对齐机制是关键研究方向。比如, VIC 框架^[57]在引入视觉信息前基于文本上下文构建推理链, 减少了视觉干扰带来的跨模态偏差, 从而增强推理准确性。

除了幻觉问题, 推理链长度也直接影响推理鲁棒性。推理链过短可能导致思考不充分, 影响准确性, 而链条过长则易引发信息遗忘与过度思考。因此, 设计合理的推理长度控制机制是一个具有前景的研究方向。目前已有部分工作对此提出解决方案, Sun 等^[168]针对模型推理中对视觉信息关注逐渐减弱的问题, 提出 TVC 策略, 将图像信息延续至关键推理阶段, 并结合动态剪枝以保持模型对图像的关注。同时, ThinkEdit^[169]提出一种权重编辑方法, 缓解推理链过短导致的数学问题求解表现不佳问题。Yang 等^[170]针对过长思维链削弱模型数学推理能力问题, 提出 TOPS 策略, 使模型通过自我优化学会用较短推理链获得正确答案。此外, Yang 等^[171]提出一种“即插即用”机制, 通过检测模型在潜在推理转折点(如“wait”标记)处的行为表现, 在高置信度时动态终止后续推理链生成, 实现高效的推理路径规划。

5.2 安全推理

推理模型的安全与伦理问题是一项重要议题, 多模态推理模型因其跨模态对齐和逐步推理的特性, 展现出独特的脆弱性, 更易遭受恶意攻击并生成有害信息。Zhou 等^[172]评估了主流推理模型面对越狱攻击和提示词注入等对抗性攻击时的鲁棒性, 发现 DeepSeek-R1 等开源模型和经过蒸馏的模型存在安全劣势。Fang 等^[173]在多模态场景中揭示了推理税(reasoning tax)现象, 即模型在获得复杂推理能力的同时, 会削弱其安全对齐特性。

目前针对多模态推理模型的特定攻击手段与防御机制的研究尚处于初级阶段, 而针对文本推理模型或通用多模态模型的相关研究则相对成熟。Wang 等^[174]将推理模型的攻击方法归纳为四类, 包括推理长度攻击、答案正确性攻击、提示词注入攻击和越狱攻击。推理长度攻击通过提示词诱导模型进行过度思考或欠思考^[175]。答案正确性攻击通过后门攻击^[176-178]或错误注入^[179]篡改中间推理步骤, 破坏结果完整性。提示词注入攻击将恶意指令伪装成正常用户输入, 绕过模型预设的安全协议^[180-181]。越狱攻击则利用提示词来规避模型安全限制。针对视觉语言模型, Ye 等^[182]根据攻击者对模型架构的了解程度, 将攻击手段分为白盒、灰盒和黑盒三类, 并分别采用对抗样本攻击以及

设计提示词等方法进行攻击。

在防御机制方面, 当前研究主要聚焦于训练时防御、推理时防御以及外部防护模型三个层面^[174]。在模型对齐训练阶段, 通过构建思维链安全数据集, 并使用监督微调^[183-184]或强化学习^[185-186]进行安全推理对齐。为避免安全训练可能牺牲模型性能的安全对齐税问题, 研究者探索在模型推理阶段应用安全提示词^[187-188]、检测并去除对抗性噪声^[189]和安全解码等技术提升模型的安全性。防护模型方法^[190-191]利用独立的审核模型对主模型的输入与输出进行实时监控和过滤, 从而在不用修改原模型的情况下实现安全防护。

总结来说, 多模态推理的特异性为模型安全带来了新挑战, 未来研究需要构建集数据、模型和推理于一体的多层次联动防御体系, 以期在不牺牲模型推理性能的前提下全面提升其安全性。

5.3 全模态推理

尽管 GPT-4o(mini) 模型展示了大模型感知和处理多模态数据的巨大潜力, 但其不具备思维链推理能力。目前多模态思维链主要局限于文本-图像^[70, 192]、文本-视频^[51, 193-194]或文本-音频^[53, 71]等双模态组合, 难以满足更复杂和通用的推理场景。

全模态思维链推理旨在拓展模型对多种模态信息的统一处理能力, 使其能够在单一框架下实现全面、多源的信息融合与推理支持。该方向对模态间表示对齐、信息互补机制以及推理路径的统一建模提出了更高要求。R1-Omini^[72]在情感识别任务中实现了思维链推理的全模态扩展, 并首次将可验证奖励的强化学习引入面向情感识别的全模态大语言模型中, 其中文本、视频与音频等模态信息均发挥了重要作用。

5.4 高效推理

受人类智能启发, 研究者将 LLM 中的直接推理与深度推理分别类比为快速、直观的系统 1 推理和缓慢、深思熟虑的系统 2 推理。系统 1 推理在计算效率上具有优势, 系统 2 推理能提升模型的推理性能。因此, 平衡模型性能与计算资源需求是实现思维链高效推理的关键。

在推理效率评估方面, MME-CoT^[134]引入关联度与反思质量指标, 用以评估多模态思维链的推理效能。针对冗长文本导致的低效问题, Shen 等^[195]提出 Heima 框架实现隐藏思维的高效推理, 将每条思维链编码为单个 token 来微调 Heima 编码器, 并在解码器中引入解释性提示词来解码思维 token 中的隐藏表

示。Lu 等^[196]提出基于确定性的推理框架 CAR, 根据模型的困惑度动态切换系统 1 与系统 2 推理, 当系统 1 生成的答案困惑度较高时即转入系统 2 进行深度推理。此外, Wang 等^[197]以及 Feng 等^[198]系统调研了 LLM 的推理效率, 提出一系列高效推理策略, 其中部分方法已成功应用于多模态大模型中。

总的来说, 当前多模态思维链推理的经济性研究仍处于初步探索阶段, 相关评估基准及优化策略有待进一步深入研究。

5.5 数据及基准建设

系统性的数据集与科学的评价基准是推动多模态思维链研究的关键支撑。由于多模态思维链数据具有跨模态和多步骤推理的特点, 其标注难度大、构建成本高, 在模态覆盖范围、推理复杂度及数据规模等方面仍显不足。同时, 相关研究也缺乏统一的评估框架以及具备挑战性的评价基准。

未来应加强高质量多模态推理数据的构建, 并建立多维度、多层次的完善评价体系, 以全面衡量模型的思维链推理能力。此外, 已有的数据集和基准多集中于逻辑结构强的数学、科学与代码领域, 后续研究应拓展更多实际应用场景, 构建覆盖更广泛领域的的数据资源, 推动思维链推理能力的泛化与实用化。

6 总结

本文对多模态思维链进行了系统而全面的综述。首先介绍了多模态思维链的基本概念及其研究必要性, 阐述了当前多模态大模型在发展中面临的局限, 以及思维链推理所具备的独特优势。随后, 从思维链范式的构建、后训练阶段的设计以及推理阶段的方法三个维度, 探讨了如何赋予多模态大模型以思维链推理能力, 并深入分析其内在机理。在评估层面, 本文整理并总结了现有的多模态思维链评测基准与指标体系。与此同时, 本文也回顾了多模态思维链在多个领域的应用。最后, 本文深入分析了当前多模态思维链推理面临的关键挑战, 并展望了未来的研究方向, 希望为构建更加鲁棒、安全和高效的思维链推理模型提供清晰的发展路径与理论基础。

参考文献

[1] LI J, LI D, SAVARESE S, et al. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//Proceedings of the 40th Interna-

tional Conference on Machine Learning, PMLR. Hawaii, USA, 2023: 19730-19742.

[2] BAI S, CHEN K, LIU X, et al. Qwen2. 5-VL technical report[J]. arXiv preprint arXiv: 2502.13923, 2025.

[3] ZHANG H, LI X, BING L. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Singapore, 2023: 543-553.

[4] LIU H, LI C, WU Q, et al. Visual instruction tuning [C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA, 2023: 34892-34916.

[5] YIN S, FU C, ZHAO S, et al. A survey on multimodal large language models[J]. arXiv preprint arXiv: 2306.13549, 2023.

[6] ANTOL S, AGRAWAL A, LU J, et al. VQA: Visual question answering[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 2425-2433.

[7] VENUGOPALAN S, ROHRBACH M, DONAHUE J, et al. Sequence to sequence-video to text[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 4534-4542.

[8] GHAFARI S, KRISHNASWAMY N. Exploring failure cases in multimodal reasoning about physical dynamics [C]//Proceedings of the AAAI Spring Symposium Series, Stanford, USA, 2024: 105-114.

[9] MALKIŃSKI M, PAWLONKA S, MAŃDZIUK J. Reasoning limitations of multimodal large language models: A case study of bongard problems[C/OL]//Proceedings of the 42nd International Conference on Machine Learning, 2025. <https://openreview.net/pdf?id=BTk1hNuIPq>.

[10] SHIRI F, GUO X Y, FAR M, et al. An empirical analysis on spatial reasoning capabilities of large multimodal models[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Miami, USA, 2024: 21440-21455.

[11] IIMAM M F, LYU C, AJI A F. Can multimodal LLMs do visual temporal understanding and reasoning?: the answer is no![J]. arXiv preprint arXiv: 2501.10674, 2025.

[12] BAI T, LIANG H, WAN B, et al. A survey of multimodal large language model from a data-centric perspective [J]. arXiv preprint arXiv: 2405.16640, 2024.

[13] WANG Y, WU S, ZHANG Y, et al. Multimodal chain-of-thought reasoning: A comprehensive survey[J]. arXiv preprint arXiv: 2503.12605, 2025.

[14] LU P, MISHRA S, XIA T, et al. Learn to explain: Multimodal reasoning via thought chains for science question answering[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 2507-2521.

[15] LI Y, LIU Z, LI Z, et al. Perception, reason, think,

- and plan: A survey on large multimodal reasoning models[J]. arXiv preprint arXiv: 2505.04921, 2025.
- [16] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models [C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 24824-24837.
- [17] KOJIMA T, GU S S, REID M, et al. Large language models are zero-shot reasoners[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 22199-22213.
- [18] WANG X, WEI J, SCHUURMANS D, et al. Self-consistency improves chain of thought reasoning in language models[C/OL]//Proceedings of ICLR, 2023. <https://openreview.net/pdf?id=1PL1NIMMrw>.
- [19] HUANG J, KEVIN C. Towards reasoning in large language models: A survey[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada, 2023: 1049-1065.
- [20] CHU Z, CHEN J, CHEN Q, et al. Navigate through enigmatic labyrinth: A survey of chain of thought reasoning: Advances, frontiers and future[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 2024: 1173-1203.
- [21] XIA Y, WANG R, LIU X, et al. Beyond chain-of-thought: A survey of chain-of-x paradigms for LLMs[C]//Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, UAE, 2025: 10795-10809.
- [22] GUO D, YANG D, ZHANG H, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning[J]. Nature, 2025, 645: 633-638. <https://doi.org/10.1038/s41586-025-09422-z>.
- [23] ZHANG Z, ZHANG A, LI M, et al. Multimodal chain-of-thought reasoning in language models[J]. Transactions on Machine Learning Research, 2024, 5. <https://openreview.net/pdf?id=y1pPWFVfR>.
- [24] CHEN Q, QIN L, ZHANG J, et al. M³CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand, 2024: 8199-8221.
- [25] BI J, LIANG S, ZHOU X, et al. Why reasoning matters? A survey of advancements in multimodal reasoning (v1) [J]. arXiv preprint arXiv: 2504.03151, 2025.
- [26] LIN Z, GAO Y, ZHAO X, et al. Mind with eyes: From language reasoning to multimodal reasoning[J]. arXiv preprint arXiv: 2503.18071, 2025.
- [27] CHEN Q, QIN L, LIU J, et al. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models[J]. arXiv preprint arXiv: 2503.09567, 2025.
- [28] ALAYRAC J B, DONAHUE J, LUC P, et al. Flamingo: A visual language model for few-shot learning[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 23716-23736.
- [29] DAI W, LI J, LI D, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA, 2023: 49250-49267.
- [30] ZHU D, CHEN J, SHEN X, et al. MiniGpt-4: Enhancing vision-language understanding with advanced large language models[J]. arXiv preprint arXiv: 2304.10592, 2023.
- [31] ZHANG X, WU C, ZHAO Z, et al. Pmc-vqa: Visual instruction tuning for medical visual question answering [J]. arXiv preprint arXiv: 2305.10415, 2023.
- [32] YANG Y, HE X, PAN H, et al. R1-Onevision: Advancing generalized multimodal reasoning through cross-modal formalization[J]. arXiv preprint arXiv: 2503.10615, 2025.
- [33] TOH V Y H, CHIA Y K, GHOSAL D, et al. The jumping reasoning curve? tracking the evolution of reasoning performance in gpt-[n] and o-[n] models on multimodal puzzles[J]. arXiv preprint arXiv:2502.01081, 2025.
- [34] WANG G, WEI X, LIU J, et al. MR-mLLM: Mutual reinforcement of multimodal comprehension and vision perception[J]. arXiv preprint arXiv: 2406.15768, 2024.
- [35] JIANG Q, LUO G, YANG Y, et al. Chatrex: Taming multimodal LLM for joint perception and understanding[J]. arXiv preprint arXiv: 2411.18363, 2024.
- [36] CHEN S, ZHANG J, ZHU T, et al. Bring reason to vision: Understanding perception and reasoning through model merging[J]. arXiv preprint arXiv: 2505.05464, 2025.
- [37] YAN H, ZHENG H, WANG H, et al. Visuriddles: Fine-grained perception is a primary bottleneck for multimodal large language models in abstract visual reasoning[J]. arXiv preprint arXiv: 2506.02537, 2025.
- [38] GHOSAL D, HAN V T Y, KEN C Y, et al. Are language models puzzle prodigies?: Algorithmic puzzles unveil serious challenges in multimodal reasoning[J]. arXiv preprint arXiv: 2403.03864, 2024.
- [39] LONG J. Large language model guided tree-of-thought[J]. arXiv preprint arXiv: 2305.08291, 2023.
- [40] YAO S, YU D, ZHAO J, et al. Tree of thoughts: Deliberate problem solving with large language models[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA, 2023: 11809-11822.
- [41] YAO F, TIAN C, LIU J, et al. Thinking like an expert: Multimodal hypergraph-of-thought (HOT) reasoning to boost foundation modals[J]. arXiv preprint arXiv: 2308.06207, 2023.
- [42] ZHENG G, YANG B, TANG J, et al. DDCoT: Duty-distinct chain-of-thought prompting for multimodal reason-

- ing in language models[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, USA, 2023: 5168-5191.
- [43] XI N, MENG J, YUAN J. Chain-of-Look prompting for verb-centric surgical triplet recognition in endoscopic videos [C]//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada, 2023: 5007-5016.
- [44] MA Y, CAO Y, SUN J, et al. Dolphins: Multimodal language model for driving[C]//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy, 2024: 403-420.
- [45] ZHANG D, YANG J, LYU H, et al. CoCoT: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs[J]. arXiv preprint arXiv: 2401.02582, 2024.
- [46] MONDAL D, MODI S, PANDA S, et al. KAM-CoT: Knowledge augmented multimodal chain-of-thoughts reasoning[C]//Proceedings of the 38th Annual AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 18798-18806.
- [47] ZHAO X, HUANG X, FU T, et al. BBA: Bi-modal behavioral alignment for reasoning with large vision-language models[C]//Proceedings of the Association for Computational Linguistics: ACL. Bangkok, Thailand, 2024: 7255-7279.
- [48] LIU Z, DONG Y, RAO Y, et al. Chain-of-spot: Interactive reasoning improves large vision-language models[J]. arXiv preprint arXiv: 2403.12966, 2024.
- [49] MITRA C, HUANG B, DARRELL T, et al. Compositional chain-of-thought prompting for large multimodal models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024: 14420-14431.
- [50] ANAND A, KAPURIYA J, SINGH A, et al. MM-PhyQa: Multimodal physics question-answering with multi-image cot prompting[C]//Proceedings of the 28th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Taipei, China, 2024: 53-64.
- [51] FEI H, WU S, JI W, et al. Video-of-thought: Step-by-step video reasoning from perception to cognition[C]//Proceedings of the 41st International Conference on Machine Learning, PMLR. Vienna, Austria, 2024: 13109-13125.
- [52] ZHOU Q, ZHOU R, HU Z, et al. Image-of-thought prompting for visual reasoning refinement in multimodal large language models[J]. arXiv preprint arXiv: 2405.13872, 2024.
- [53] DU Y, MA Z, YANG Y, et al. CoT-ST: Enhancing LLM-based speech translation with multimodal chain-of-thought[J]. arXiv preprint arXiv: 2409.19510, 2024.
- [54] BYUN J S, CHUN J, KIL J, et al. ARES: Alternating reinforcement learning and supervised fine-tuning for enhanced multi-modal chain-of-thought reasoning through diverse AI feedback[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Miami, USA, 2024: 4410-4430.
- [55] HU Y, SHI W, FU X, et al. VisualSKETCHPAD: sketching as a visual chain of thought for multimodal language models[J]. arXiv preprint arXiv: 2406.09403, 2024.
- [56] XU G, JIN P, HAO L, et al. LLaVA-o1: Let vision language models reason step-by-step[J]. arXiv preprint arXiv: 2411.10440, 2024.
- [57] ZHENG H, XU T, SUN H, et al. Thinking before looking: improving multimodal LLM reasoning via mitigating visual hallucination[J]. arXiv preprint arXiv: 2411.12591, 2024.
- [58] DONG Y, LIU Z, SUN H L, et al. Insight-V: Exploring long-chain visual reasoning with multimodal large language models[J]. arXiv preprint arXiv: 2411.14432, 2024.
- [59] WANG X, YANG Z, LI L, et al. Scaling inference-time search with vision value model for improved visual comprehension[J]. arXiv preprint arXiv: 2412.03704, 2024.
- [60] XIANG K, LIU Z, JIANG Z, et al. AtomThink: A slow thinking framework for multimodal mathematical reasoning[J]. arXiv preprint arXiv: 2411.11930, 2024.
- [61] SUN Q, HONG P, PALA T D, et al. Emma-X: An embodied multimodal action model with grounded chain of thought and look-ahead spatial reasoning[J]. arXiv preprint arXiv: 2412.11974, 2024.
- [62] DONG G, ZHANG C, DENG M, et al. Progressive multimodal reasoning via active retrieval[J]. arXiv preprint arXiv: 2412.14835, 2024.
- [63] YAO H, HUANG J, WU W, et al. Mulberry: Empowering MLLM with o1-like reasoning and reflection via collective Monte Carlo tree search[J]. arXiv preprint arXiv: 2412.18319, 2024.
- [64] THAWAKAR O, DISSANAYAKE D, MORE K, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms[J]. arXiv preprint arXiv: 2501.06186, 2025.
- [65] LI C, WU W, ZHANG H, et al. Imagine while reasoning in space: Multimodal visualization-of-thought[J]. arXiv preprint arXiv: 2501.07542, 2025.
- [66] YU Y, ZHANG Y, ZHANG D, et al. Chain-of-reasoning: Towards unified mathematical reasoning in large language models via a multi-paradigm perspective[J]. arXiv preprint arXiv: 2501.11110, 2025.
- [67] LIU Y, CHI D, WU S, et al. SpatialCoT: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning[J]. arXiv preprint arXiv: 2501.10074, 2025.
- [68] WU J, FENG M, ZHANG S, et al. Boosting multimodal reasoning with mcts-automated structured thinking[J]. arXiv preprint arXiv: 2502.02339, 2025.
- [69] PAN J, LIU C, WU J, et al. MedVLM-R1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning[J]. arXiv preprint arXiv:

- 2502.19634, 2025.
- [70] LIU Z, SUN Z, ZANG Y, et al. Visual-RFT: Visual reinforcement fine-tuning[J]. arXiv preprint arXiv: 2503.01785, 2025.
- [71] XIE Z, LIN M, LIU Z, et al. Audio-Reasoner: Improving reasoning capability in large audio language models[J]. arXiv preprint arXiv: 2503.02318, 2025.
- [72] ZHAO J, WEI X, BO L. R1-Omni: Explainable omni-multimodal emotion recognition with reinforcement learning[J]. arXiv preprint arXiv: 2503.05379, 2025.
- [73] HUANG W, JIA B, ZHAI Z, et al. Vision-R1: Incentivizing reasoning capability in multimodal large language models[J]. arXiv preprint arXiv: 2503.06749, 2025.
- [74] ISHAQ A, LAHOUD J, MORE K, et al. DriveLMM-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding[J]. arXiv preprint arXiv: 2503.10621, 2025.
- [75] ZHANG Y, ZHANG Q, JU X, et al. EmbodiedVSR: Dynamic scene graph-guided chain-of-thought reasoning for visual spatial tasks[J]. arXiv preprint arXiv: 2503.11089, 2025.
- [76] GAO J, LI Y, CAO Z, et al. Interleaved-modal chain-of-thought[J]. arXiv preprint arXiv: 2411.19488, 2024.
- [77] ZHANG W, WANG M, LIU G, et al. Embodied-Reasoner: Synergizing visual search, reasoning, and action for embodied interactive tasks[J]. arXiv preprint arXiv: 2503.21696, 2025.
- [78] BESTA M, BLACH N, KUBICEK A, et al. Graph of thoughts: Solving elaborate problems with large language models[C]//Proceedings of the 38th Annual AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 17682-17690.
- [79] LEI B, LIAO C, DING C. Boosting logical reasoning in large language models through a new framework: The graph of thought[J]. arXiv preprint arXiv: 2308.08614, 2023.
- [80] LE C C, TRUONG-VINH H C, PHAN H N, et al. VisualCoder: Guiding large language models in code execution with fine-grained multimodal chain-of-thought reasoning[J]. arXiv preprint arXiv: 2410.23402, 2024.
- [81] XIANG K, LIU Z, JIANG Z, et al. Can atomic step decomposition enhance the self-structured reasoning of multimodal large models?[J]. arXiv preprint arXiv: 2503.06252, 2025.
- [82] CHEN S, LI B, NIU D. Boosting of thoughts: Trial-and-error problem solving with large language models[J]. arXiv preprint arXiv: 2402.11140, 2024.
- [83] ZHENG C, LIANG D, ZHANG W, et al. A picture is worth a graph: A blueprint debate paradigm for multimodal reasoning[C]//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia, 2024: 419-428.
- [84] WEI J, TAN C, GAO Z, et al. Enhancing human-like multimodal reasoning: A new challenging dataset and comprehensive framework[J]. *Neural Computing and Applications*, 2024, 36(33): 20849-20861.
- [85] WANG Z, BINGHAM G, YU A W, et al. HaloQuest: A visual hallucination dataset for advancing multimodal reasoning[C]//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy, 2024: 288-304.
- [86] SHI W, HU Z, BIN Y, et al. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models[C]//Proceedings of the Association for Computational Linguistics: EMNLP. Miami, USA, 2024: 4663-4680.
- [87] LIU W, PAN Q, ZHANG Y, et al. CMM-Math: A Chinese multimodal math dataset to evaluate and enhance the mathematics reasoning of large multimodal models[J]. arXiv preprint arXiv: 2409.02834, 2024.
- [88] SHAO H, QIAN S, XIAO H, et al. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning[C]//Proceedings of the 38th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2024: 8612-8642.
- [89] ANAND A, JAISWAL R, DHARMADHIKARI A, et al. Improving multimodal LLMs ability in geometry problem solving, reasoning, and multistep scoring[J]. arXiv preprint arXiv: 2412.00846, 2024.
- [90] XIAO J, LIN G, LI P. MLRQA: A dataset with multimodal logical reasoning challenges[C]//Proceedings of the Pacific Rim International Conference on Artificial Intelligence. Kyoto, Japan, 2024: 3-14.
- [91] SUN L, LIANG H, WEI J, et al. MM-Verify: Enhancing multimodal reasoning with chain-of-thought verification[J]. arXiv preprint arXiv: 2502.13383, 2025.
- [92] LIAO Y H, ELFLEIN S, HE L, et al. Long Perceptual Thoughts: Distilling system-2 reasoning for system-1 perception[J]. arXiv preprint arXiv: 2504.15362, 2025.
- [93] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland, 2014: 740-755.
- [94] RAFAILOV R, SHARMA A, MITCHELL E, et al. Direct preference optimization: Your language model is secretly a reward model[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA, 2023: 53728-53741.
- [95] SHAO Z, WANG P, ZHU Q, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models[J]. arXiv preprint arXiv: 2402.03300, 2024.
- [96] LIU Y, PENG B, ZHONG Z, et al. Seg-Zero: Reasoning-chain guided segmentation via cognitive reinforcement[J]. arXiv preprint arXiv: 2503.06520, 2025.
- [97] WANG W, CHEN Z, WANG W, et al. Enhancing the

- reasoning ability of multimodal large language models via mixed preference optimization[J]. arXiv preprint arXiv: 2411.10442, 2024.
- [98] AMINI A, VIEIRA T, ASH E, et al. Variational best-of-n alignment[J]. arXiv preprint arXiv: 2407.06057, 2024.
- [99] WANG X, LI J, ZHANG Y. Improved value alignment in large language models using variational best-of-n techniques [J]. arXiv preprint arXiv: 2407, 2024.
- [100] SUN H, HAIDER M, ZHANG R, et al. Fast best-of-n decoding via speculative rejection[J]. arXiv preprint arXiv: 2410.20290, 2024.
- [101] QIU J, LU Y, ZENG Y, et al. TreeBon: Enhancing inference-time alignment with speculative tree-search and best-of-n sampling[J]. arXiv preprint arXiv: 2410.16033, 2024.
- [102] GOU Z, SHAO Z, GONG Y, et al. Critic: Large language models can self-correct with tool-interactive critiquing [J]. arXiv preprint arXiv: 2305.11738, 2023.
- [103] PAUL D, ISMAYILZADA M, PEYRARD M, et al. Refiner: Reasoning feedback on intermediate representations[C]//Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, St. Julians, Malta, 2024: 1100-1126.
- [104] LI Y, LIN Z, ZHANG S, et al. Making language models better reasoners with step-aware verifier[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada, 2023: 5315-5333.
- [105] CHENG K, LI Y, XU F, et al. Vision-language models can self-improve reasoning via reflection[J]. arXiv preprint arXiv: 2411.00855, 2024.
- [106] KHALID U, IQBAL H, FAROOQ A, et al. EVLM: Self-reflective multimodal reasoning for cross-dimensional visual editing[J]. arXiv preprint arXiv: 2412.10566, 2024.
- [107] LIU Y, LI P, WEI Z, et al. InfiGUIAgent: A multimodal generalist GUI agent with native reasoning and reflection[J]. arXiv preprint arXiv: 2501.04575, 2025.
- [108] KHALIQ M, CHANG P, MA M, et al. RAGAR, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models [C]//Proceedings of the Seventh Fact Extraction and VERification Workshop, Miami, USA, 2024: 280-296.
- [109] TAN C, WEI J, SUN L, et al. Retrieval meets reasoning: even high-school textbook knowledge benefits multimodal reasoning[J]. arXiv preprint arXiv: 2405.20834, 2024.
- [110] LEE J, WANG Y, LI J, et al. Multimodal reasoning with multimodal knowledge graph[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 2024: 10767-10782.
- [111] ELHENAWY M, ABUTAHOUN A, ALHADIDI T I, et al. Visual reasoning and multi-agent approach in multimodal large language models (mlms): Solving tsp and mtsp combinatorial challenges[J]. *Machine Learning and Knowledge Extraction*, 2024, 6(3): 1894-1920.
- [112] GU Z, LIU F, YIN C, et al. Inquire, interact, and integrate: A proactive agent collaborative framework for zero-shot multimodal medical reasoning[J]. arXiv preprint arXiv: 2405.11640, 2024.
- [113] ZHAI S, BAI H, LIN Z, et al. Fine-tuning large vision-language models as decision-making agents via reinforcement learning[C]//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada, 2024: 110935-110971.
- [114] LU P, BANSAL H, XIA T, et al. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts[J]. arXiv preprint arXiv: 2310.02255, 2023.
- [115] WANG K, PAN J, SHI W, et al. Measuring multimodal mathematical reasoning with math-vision dataset[C]//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada, 2024: 95095-95169.
- [116] QIAO R, TAN Q, DONG G, et al. We-Math: Does your large multimodal model achieve human-like mathematical reasoning? [J]. arXiv preprint arXiv: 2407.01284, 2024.
- [117] XIAO Y, SUN E, LIU T, et al. LogicVista: Multimodal LLM logical reasoning benchmark in visual contexts[J]. arXiv preprint arXiv: 2407.04973, 2024.
- [118] ZHANG R, JIANG D, ZHANG Y, et al. MathVerse: Does your multi-modal llm truly see the diagrams in visual math problems?[C]//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy, 2024: 169-186.
- [119] YAN Y, WANG S, HUO J, et al. ErrorRadar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection[J]. arXiv preprint arXiv: 2410.04509, 2024.
- [120] LI Z, WANG H, LIU D, et al. Multimodal causal reasoning benchmark: Challenging vision large language models to discern causal links across modalities[J]. arXiv preprint arXiv: 2408.08105, 2024.
- [121] WANG C J, LEE D, MENGHINI C, et al. EnigmaEval: A benchmark of long multimodal reasoning challenges[J]. arXiv preprint arXiv: 2502.08859, 2025.
- [122] YUAN J, PENG T, JIANG Y, et al. MME-Reasoning: A comprehensive benchmark for logical reasoning in MLLMs[J]. arXiv preprint arXiv: 2505.21327, 2025.
- [123] YIN H, LIN Z, LIU X, et al. Do multimodal language models really understand direction?: A benchmark for compass direction reasoning[C]//Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Hyderabad, India, 2025: 1-5.
- [124] RAJABI N, KOSECKA J. Gsr-Bench: A benchmark for grounded spatial reasoning evaluation via multimodal llms[J]. arXiv preprint arXiv: 2406.13246, 2024.

- [125] SHANGGUAN Z, LI C, DING Y, et al. TOMATO: Assessing visual temporal reasoning capabilities in multimodal foundation models[J]. arXiv preprint arXiv: 2410.23266, 2024.
- [126] WANG X, MA W, ZHANG T, et al. PulseCheck457: A diagnostic benchmark for 6d spatial reasoning of large multimodal models[J]. arXiv e-prints, 2025: arXiv: 2502.08636.
- [127] WU C, GE Y, GUO Q, et al. Plot2Code: A comprehensive benchmark for evaluating multi-modal large language models in code generation from scientific plots[J]. arXiv preprint arXiv: 2405.07990, 2024.
- [128] CHEN J, LIU Y, LI D, et al. Plug-and-play grounding of reasoning in multimodal large language models[J]. arXiv preprint arXiv: 2403.19322, 2024.
- [129] ZHANG F, WU L, BAI H, et al. HumanEval-V: Evaluating visual understanding and reasoning abilities of large multimodal models through coding tasks[J]. arXiv preprint arXiv: 2410.12381, 2024.
- [130] CUI Z, YUAN J, WANG H, et al. Draw with thought: unleashing multimodal reasoning for scientific diagram generation[J]. arXiv preprint arXiv: 2504.09479, 2025.
- [131] ZHANG D, YANG J, LYU H, et al. A benchmark and chain-of-thought prompting strategy for large multimodal models with multiple image inputs[C]//Proceedings of the 27th International Conference on Pattern Recognition. Kolkata, India, 2024: 226-241.
- [132] FAN L, HUA W, LI X, et al. NPhardEval4V: A dynamic reasoning benchmark of multimodal large language models[J]. arXiv preprint arXiv: 2403.01777, 2024.
- [133] HAO Y, GU J, WANG H W, et al. Can MLLMs reason in multimodality? emma: An enhanced multimodal reasoning benchmark[J]. arXiv preprint arXiv: 2501.05444, 2025.
- [134] JIANG D, ZHANG R, GUO Z, et al. MME-CoT: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency[J]. arXiv preprint arXiv: 2502.09621, 2025.
- [135] SONG Y, OU T, KONG Y, et al. VisualPuzzles: Decoupling multimodal reasoning evaluation from domain knowledge[J]. arXiv preprint arXiv: 2504.10342, 2025.
- [136] ZHOU X, HE J, CHEN L, et al. MiCEval: Unveiling multimodal chain of thought's quality via image description and reasoning steps[J]. arXiv preprint arXiv: 2410.14668, 2024.
- [137] YAN Q, FAN Y, LI H, et al. Multimodal inconsistency reasoning (mmir): A new benchmark for multimodal reasoning models[J]. arXiv preprint arXiv: 2502.16033, 2025.
- [138] XU Z, ZHOU P, AI J, et al. Mpbench: A comprehensive multimodal reasoning benchmark for process errors identification[J]. arXiv preprint arXiv: 2503.12505, 2025.
- [139] JIA B, ZHANG J, ZHANG H, et al. Exploring and evaluating multimodal knowledge reasoning consistency of multimodal large language models[J]. arXiv preprint arXiv: 2503.04801, 2025.
- [140] TIE G, ZHOU X, GU T, et al. MMR: Benchmarking massive multi-modal reasoning tasks[J]. arXiv preprint arXiv: 2505.16459, 2025.
- [141] HAO Y, BETHALA G C R, PUDASAINI N, et al. Embodied chain of action reasoning with multi-modal foundation model for humanoid loco-manipulation[J]. arXiv preprint arXiv: 2504.09532, 2025.
- [142] MU Y, ZHANG Q, HU M, et al. EmbodiedGPT: Vision-language pre-training via embodied chain of thought [C]//Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, USA, 2023: 25081-25094.
- [143] LIANG X, LIN M, RUAN W, et al. Memory-driven multimodal chain of thought for embodied long-horizon task planning[J]. arXiv preprint arXiv: 2409, 2024.
- [144] ZAWALSKI M, CHEN W, PERTSCH K, et al. Robotic control via embodied chain-of-thought reasoning[J]. arXiv preprint arXiv: 2407.08693, 2024.
- [145] LIN B, NIE Y, WEI Z, et al. Navcot: Boosting LLM-based vision-and-language navigation via learning disentangled reasoning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025: 1-13.
- [146] ZHANG J, GAO C, ZHANG L, et al. Smartagent: Chain-of-user-thought for embodied personalized agent in cyber world[J]. arXiv preprint arXiv: 2412.07472, 2024.
- [147] MAO J, YE J, QIAN Y, et al. A language agent for autonomous driving[J]. arXiv preprint arXiv: 2311.10813, 2023.
- [148] WANG T, XIE E, CHU R, et al. DriveCoT: Integrating chain-of-thought reasoning with end-to-end driving[J]. arXiv preprint arXiv: 2403.16996, 2024.
- [149] NIE M, PENG R, WANG C, et al. Reason2Drive: Towards interpretable and chain-based reasoning for autonomous driving[C]//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy, 2024: 292-308.
- [150] LUO X, DING F, SONG Y, et al. PKRD-CoT: A unified chain-of-thought prompting for multi-modal large language models in autonomous driving[J]. arXiv preprint arXiv: 2412.02025, 2024.
- [151] XING S, QIAN C, WANG Y, et al. OpenEMMA: Open-source multimodal model for end-to-end autonomous driving[C]//Proceedings of the Winter Conference on Applications of Computer Vision. Tucson, USA, 2025: 1001-1009.
- [152] ISLAM M R. Application of multimodal large language models in autonomous driving[J]. arXiv preprint arXiv: 2412.16410, 2024.

- [153] GAI X, ZHOU C, LIU J, et al. MedThink: Explaining medical visual question answering via multimodal decision-making rationale[J]. arXiv preprint arXiv: 2404.12372, 2024.
- [154] LIU J, WANG Y, DU J, et al. MedCoT: Edical chain of thought via hierarchical expert[J]. arXiv preprint arXiv: 2412.13736, 2024.
- [155] DAI Y. Interpretable video based stress detection with self-refine chain-of-thought reasoning[J]. arXiv preprint arXiv: 2410.09449, 2024.
- [156] PLINI L, SCOFANO L, DE MATTEIS E, et al. TI-PREGO: Chain of thought and in-context learning for online mistake detection in procedural egocentric videos[J]. arXiv preprint arXiv: 2411.02570, 2024.
- [157] WANG Y, LIU M, HE W, et al. Mint: Multi-modal chain of thought in unified generative models for enhanced image generation[J]. arXiv preprint arXiv: 2503.01298, 2025.
- [158] LIN W, JIA L, HU W, et al. Reasoning physical video generation with diffusion timestep tokens via reinforcement learning[J]. arXiv preprint arXiv: 2504.15932, 2025.
- [159] AGRAHARI S, SINGH S R. Tracing thought: Using chain-of-thought reasoning to identify the llm behind AI-generated text[J]. arXiv preprint arXiv: 2504.16913, 2025.
- [160] CHEN A, SONG Y, CHEN K, et al. Make imagination clearer!: stable diffusion-based visual imagination for multimodal machine translation[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Vienna, Austria, 2025: 26567-26583.
- [161] RAJPOOT P, BHAT N, SHRIVASTAVA A. Multimodal machine translation for low-resource indic languages: A chain-of-thought approach using large language models [C]//Proceedings of the 9th Conference on Machine Translation. Miami, USA, 2024: 833-838.
- [162] LI Y, LAN X, CHEN H, et al. Multimodal pear chain-of-thought reasoning for multimodal sentiment analysis[J]. [ACM Transactions on Multimedia Computing, communications and Applications](#), 2025, 20(9): 1-23.
- [163] ZHANG H, MENG Z, LUO M, et al. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark[C]//Proceedings of the ACM on Web Conference, Sydney, Australia, 2025: 2872-2881.
- [164] MATHUR L, QIAN M, LIANG P P, et al. Social genome: Grounded social reasoning abilities of multimodal models[J]. arXiv preprint arXiv: 2502.15109, 2025.
- [165] YU Y, LUO J. Chain-of-thought prompting for demographic inference with large multimodal models[C]//Proceedings of the IEEE International Conference on Multimedia and Expo. Niagra Falls, Canada, 2024: 1-7.
- [166] XU Y, GE J, LYU G, et al. Multimodal fake news detection based on chain-of-thought prompting large language models[C]//Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics. Sarawak, Malaysia, 2024: 559-566.
- [167] CHENG X, LI J, ZHAO W X, et al. Think more, hallucinate less: Mitigating hallucinations via dual process of fast and slow thinking[J]. arXiv preprint arXiv: 2501.01306, 2025.
- [168] SUN H L, SUN Z, PENG H, et al. Mitigating visual forgetting via take-along visual conditioning for multi-modal long cot reasoning[J]. arXiv preprint arXiv: 2503.13360, 2025.
- [169] SUN C E, YAN G, WENG T W. ThinkEdit: Interpretable weight editing to mitigate overly short thinking in reasoning models[J]. arXiv preprint arXiv: 2503.22048, 2025.
- [170] YANG W, MA S, LIN Y, et al. Towards thinking-optimal scaling of test-time compute for LLM reasoning[J]. arXiv preprint arXiv: 2502.18080, 2025.
- [171] YANG C, SI Q, DUAN Y, et al. Dynamic early exit in reasoning models[J]. arXiv preprint arXiv: 2504.15895, 2025.
- [172] ZHOU K, LIU C, ZHAO X, et al. The hidden risks of large reasoning models: A safety assessment of R1[J]. arXiv preprint arXiv: 2502.12659, 2025.
- [173] FANG J, WANG Y, WANG R, et al. Safemllm: Demystifying safety in multi-modal large reasoning models[J]. arXiv preprint arXiv: 2504.08813, 2025.
- [174] WANG C, LIU Y, LI B, et al. Safety in large reasoning models: A survey[J]. arXiv preprint arXiv: 2504.17704, 2025.
- [175] ZAREMBA W, NITISHINSKAYA E, BARAK B, et al. Trading inference-time compute for adversarial robustness [J]. arXiv preprint arXiv: 2501.18841, 2025.
- [176] XIANG Z, JIANG F, XIONG Z, et al. Badchain: Backdoor chain-of-thought prompting for large language models[J]. arXiv preprint arXiv: 2401.12242, 2024.
- [177] GUO Z, TOURANI R. Darkmind: Latent chain-of-thought backdoor in customized LLMs[J]. arXiv preprint arXiv: 2501.18617, 2025.
- [178] ZHAO G, WU H, ZHANG X, et al. Shadowcot: Cognitive hijacking for stealthy reasoning backdoors in LLMs[J]. arXiv preprint arXiv: 2504.05605, 2025.
- [179] CUI Y, HOOI B, CAI Y, et al. Process or result?: Manipulated ending tokens can mislead reasoning llms to ignore the correct reasoning steps[J]. arXiv preprint arXiv: 2503.19326, 2025.
- [180] KUMAR S S, CUMMINGS M L, STIMPSON A. Strengthening LLM trust boundaries: A survey of prompt injection attacks[C]//Proceedings of the 4th IEEE International Conference on Human-Machine Systems. Toronto, Canada, 2024: 1-6.
- [181] LIU Y, DENG G, LI Y, et al. Prompt injection attack

- against LLM-integrated applications[J]. arXiv preprint arXiv: 2306.05499, 2023.
- [182] YE M, RONG X, HUANG W, et al. A survey of safety on large vision-language models: Attacks, defenses and evaluations[J]. arXiv preprint arXiv: 2502.14881, 2025.
- [183] LOU X, LI Y, XU J, et al. Think in safety: Unveiling and mitigating safety alignment collapse in multimodal large reasoning model[J]. arXiv preprint arXiv: 2505.06538, 2025.
- [184] XIA Y, JIANG Y, TAN Y, et al. MSR-align: Policy-grounded multimodal alignment for safety-aware reasoning in vision-language models[J]. arXiv preprint arXiv: 2506.19257, 2025.
- [185] GUAN M Y, JOGLEKAR M, WALLACE E, et al. Deliberative alignment: Reasoning enables safer language models[J]. arXiv preprint arXiv: 2412.16339, 2024.
- [186] ZHANG Y, ZHANG S, HUANG Y, et al. Stair: Improving safety alignment with introspective reasoning[J]. arXiv preprint arXiv: 2502.02384, 2025.
- [187] WANG Y, LIU X, LI Y, et al. Adashield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting[C]//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy, 2024: 77-94.
- [188] ZHAO Y, ZHENG X, LUO L, et al. Bluesuffix: Reinforced blue teaming for vision-language models against jail-break attacks[J]. arXiv preprint arXiv: 2410.20971, 2024.
- [189] SUN J, WANG C, WANG J, et al. Safeguarding vision-language models against patched visual prompt injectors[J]. arXiv preprint arXiv: 2405.10529, 2024.
- [190] LIU Y, GAO H, ZHAI S, et al. Guardreasoner: Towards reasoning-based llm safeguards[J]. arXiv preprint arXiv: 2501.18492, 2025.
- [191] WEN X, ZHOU W, MO W J, et al. Thinkguard: Deliberative slow thinking leads to cautious guardrails[J]. arXiv preprint arXiv: 2502.13458, 2025.
- [192] LI Z, XU J, WANG S, et al. Streetviewllm: Extracting geographic information using a chain-of-thought multimodal large language model[J]. arXiv preprint arXiv: 2411.14476, 2024.
- [193] HIMAKUNTHALA V, OUYANG A, ROSE D, et al. Let's think frame by frame with vip: A video infilling and prediction dataset for evaluating video chain-of-thought[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Singapore, Singapore, 2023: 204-219.
- [194] ZHANG Y, ZHANG K, LI B, et al. WorldQA: Multimodal world knowledge in videos through long-chain reasoning[J]. arXiv preprint arXiv: 2405.03272, 2024.
- [195] SHEN X, WANG Y, SHI X, et al. Efficient reasoning with hidden thinking[J]. arXiv preprint arXiv: 2501.19201, 2025.
- [196] LU J, YU H, XU S, et al. Prolonged reasoning is not all you need: Certainty-based adaptive routing for efficient llm/mlm reasoning[J]. arXiv preprint arXiv: 2505.15154, 2025.
- [197] WANG R, WANG H, XUE B, et al. Harnessing the reasoning economy: A survey of efficient reasoning for large language models[J]. arXiv preprint arXiv: 2503.24377, 2025.
- [198] FENG S, FANG G, MA X, et al. Efficient reasoning models: A survey[J]. arXiv preprint arXiv: 2504.10903, 2025.



祝文鑫(2003—), 硕士, 主要研究领域为多模态大语言模型、机器翻译等。

E-mail: wenzinzh@stu.hit.edu.cn



宋宇宸(2003—), 硕士, 主要研究领域为多模态大语言模型、机器翻译等。

E-mail: songyuchn@126.com



陈安东(1995—), 博士, 主要研究领域为大语言模型、自然语言处理、机器翻译等。

E-mail: ands691119@gmail.com