基于思维链推理的可控多模态决策方法

胡宇航[1)] 王诗涵[1)] 刘利龙[1)] 杨振宇[2)] 钱胜胜[2)]

[1)](郑州大学河南先进技术研究院 郑州 450003)

[2)](中国科学院自动化研究所多模态人工智能系统全国重点实验室 北京 100190)

摘 要 近年来,多模态大语言模型(Multimodal Large Language Models, MLLMs)在人工智能领域取得了显著的进展,特别是逻辑推理方面。思维链推理的出现显著提升了大语言模型的能力,尤其是在复杂推理任务中。尽管取得了这些进展,多模态大语言模型在可控推理方面仍面临挑战,特别是在视觉问答环境中。传统方法通常导致无结构的推理过程或受限于僵化的框架,限制了其适应性和有效性。为了解决这些问题,本研究提出了基于思维链推理的可控多模态决策方法(Controlled Multimodal Decision-making Method based on Chain-of-Thought reasoning,CMDM-CoT),一种旨在增强MLLMs推理能力的新型可控框架。CMDM-CoT引入了自适应问题解决决策集,使模型能够根据任务复杂性自主选择适当的推理路径,从而克服固定框架的局限性。此外,CMDM-CoT还包含状态评估机制,通过对每个推理状态进行评分,确保逻辑一致性和高质量学习。这种方法不仅促进了简单任务的最小化推理,还支持复杂问题的详细推理。值得注意的是,CMDM-CoT在应用于Llama、Qwen2-VL、InternVL2三种主流模型时表现出色,与基线模型相比,这些模型平均提高了7.3%。此外,这些模型甚至超过了参数量更大的闭源模型GPT-4V,显示出本研究的开源模型在多个基准测试中具有竞争力。

关键词 思维链推理;多模态大语言模型;视觉问答;自适应决策机制;复杂推理

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2026.00841

Controllable Multimodal Decision-Making Method Based on Chain of Thought Reasoning

HU Yu-Hang[1)] WANG Shi-Han[1)] LIU Li-Long[1)] YANG Zhen-Yu[2)] QIAN Sheng-Sheng[2)]

[1)](Henan Institute of Advanced Technology, Zhengzhou University, Zhengzhou 450003)

[2)](National Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190)

Abstract In recent years, Multimodal Large Language Models (MLLMs) have made significant and remarkable progress in the field of artificial intelligence, particularly demonstrating unprecedented potential in logical reasoning. With the continuous evolution of deep learning techniques and enhanced capabilities for cross-modal data fusion, MLLMs are gradually narrowing the gap with human cognitive abilities. The emergence of chain-of-thought reasoning, as a method mimicking human step-by-step problem analysis, has significantly enhanced the capabilities of large language models, especially in complex reasoning tasks involving multiple steps and diverse information, where its importance is increasingly prominent. Despite these encouraging advancements, MLLMs still face substantial challenges in controllable reasoning, particularly in visual question answering (VQA) environments that require simultaneous

收稿日期:2025-06-05;在线发布日期:2025-11-03。本课题得到国家重点研发计划(2023YFC3310700)、北京市自然科学基金(JQ23018)、国家自然科学基金(62276257)资助。**胡宇航**,硕士研究生,主要研究领域为多模态理解与应用。E-mail:hyh422904725@gs.zzu.edu.cn。**王诗涵**,硕士研究生,主要研究领域为多模态理解与应用。**刘利龙**,硕士研究生,主要研究领域为人工智能与自然语言处理。**杨振宇**,博士研究生,主要研究领域为多模态理解与应用。**钱胜胜**(通信作者),博士,副研究员,中国计算机学会(CCF)高级会员,主要研究领域为多媒体内容分析、数据挖掘、跨模态检索与个性化推荐。E-mail:shengsheng.qian@nlpr.ia.ac.cn。

processing and understanding of both visual and textual information. Traditional methods often lead to unstructured reasoning processes that are difficult to trace and verify, or are constrained by predefined, rigid frameworks. This not only limits the model's adaptability and generalization across different tasks but also compromises the effectiveness and interpretability of final decisions. A core research difficulty lies in balancing the flexibility and structure of the reasoning process to meet both the efficiency needs of simple tasks and the in-depth analysis requirements of complex scenarios.

To address these key issues, this study innovatively proposes the Controlled Multimodal Decision-making Method based on Chain-of-Thought reasoning (CMDM-CoT), a novel controllable framework designed to systematically enhance the reasoning capabilities of MLLMs. A core innovation of CMDM-CoT is the introduction of an adaptive problem-solving decision set. This set functions like a dynamic "reasoning toolkit," enabling the model to autonomously and intelligently select the most appropriate reasoning paths and strategies based on the intrinsic complexity and characteristics of the specific task, thereby effectively overcoming the limitations and rigidity associated with traditional fixed frameworks. Furthermore, to ensure the quality and robustness of the entire reasoning process, CMDM-CoT meticulously designs and integrates a state evaluation mechanism. This mechanism provides real-time, quantitative scores for each intermediate state within the reasoning chain, continuously monitoring logical consistency, rationality, and informational validity, thus promoting high-quality learning and decision-making and effectively preventing error accumulation and logical deviation. This comprehensive methodological design allows the framework not only to facilitate efficient, minimal reasoning for simple tasks, avoiding unnecessary computational overhead, but also to support detailed, step-by-step reasoning for complex and ambiguous problems, demonstrating powerful task adaptability.

Notably, to validate the effectiveness and universality of the CMDM-CoT framework, we applied it to three representative mainstream open-source models—Llama, Qwen2-VL, and InternVL2—for extensive experimental evaluation. The results demonstrated that CMDM-CoT performs excellently. Compared to the baseline models without this framework enhancement, these models achieved an average performance improvement of 7.3% across multiple visual reasoning and question-answering tasks, fully attesting to the method's effectiveness and generalizability. More remarkably, these open-source models enhanced by CMDM-CoT even surpassed the larger-parameter, computationally intensive closed-source commercial model GPT-4V in certain evaluation scenarios. This finding strongly indicates that through the introduction of advanced reasoning control frameworks like CMDM-CoT, the open-source models involved in this study exhibit strong competitiveness across multiple authoritative benchmarks. This provides a new direction and robust technical support for the future development and application of the open-source community, and also suggests that controllable and interpretable multimodal reasoning is moving towards a more mature and practical stage.

Keywords chain-of-thought reasoning; multimodal large language models; visual question answering; adaptive decision making mechanism; complex reasoning

1 引 言

近年来，思维链推理（Chain of Thought，CoT）概念[1-3]引起广泛关注，成为人工智能领域的焦点。从少样本CoT到零样本CoT，大型语言模型（Large Language Models，LLMs）在推理能力方面取得了显著进展，尤其是在推理阶段的扩展方面，如OpenAI

的o1模型[4]所示。然而，当前的多模态大语言模型（Multimodal Large Language Models，MLLMs）在进行系统和结构化推理时常面临挑战，特别是在复杂的视觉问答（Visual Question Answering，VQA）任务中[5-6]。

最初，多模态大语言模型旨在提供直接答案，如图1(a)所示。然而，当面对复杂的视觉问题时，这些模型在没有经过深思熟虑的过程中难以提供准确的答案。随后，许多研究采用了零样本CoT方法，如图1(b)所示，利用经典提示“Let's think step by step”来“唤醒”MLLMs的推理能力，使其能够自主生成思维链[7-10]。

在VQA任务中，思维链的结构和长度是两个关键方面。最近的研究主要集中在扩展这些思维链的结构和长度。Zhang等[11]提出一种两阶段多模态CoT方法，首先利用文本和视觉特征推导出一个合理性，然后根据合理性以及文本和视觉特征生成最终答案。LLaVA-CoT[12]包含四个不同阶段，使模型能够独立执行多阶段逐步推理，从而实现长思维链。

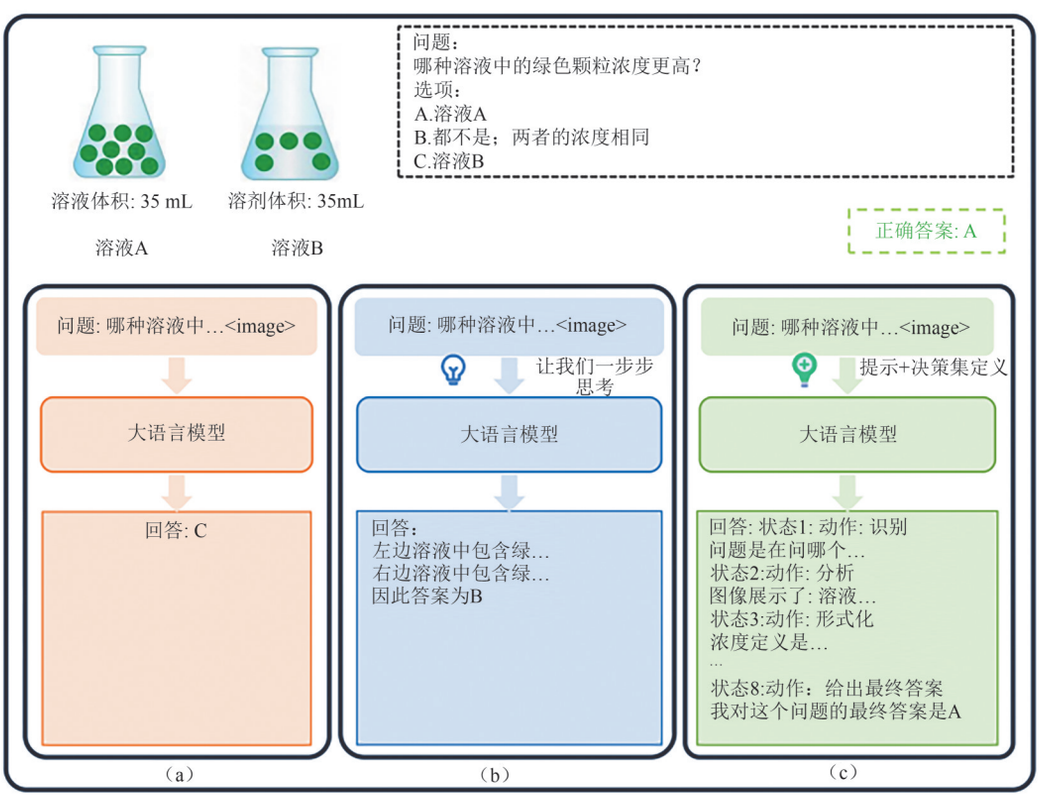

图1　多模态大语言模型中三种响应策略的比较

然而，VQA任务的多样性、复杂性和格式差异显著。零样本CoT可能导致无结构的思维链，其中无目的推理常常无法产生正确结果。相反，固定结构可能限制灵活性，多步骤思维链推理可能阻碍实际应用。受此启发，本研究旨在解决挑战1：如何使MLLMs能够自适应地生成结构化和有效的思维链，以应对多样化的VQA任务，而不受固定框架的限制？

此外，视觉思维链要求模型在整个推理过程中保持逻辑一致性。现有的多模态大语言模型主要关注最终任务的准确性，往往忽略中间推理步骤的质量。质量差的数据可能导致模型学习错误信息，从而在处理类似推理路径时产生错误结果。根据上述讨论，本研究旨在解决挑战2：如何确保MLLMs在推理过程中保持逻辑一致性，并仅学习高质量信息？

为了解决这些挑战，本研究提出了基于思维链推理的可控多模态决策方法（Controlled Multimodal Decision-making Method based on Chain-of-Thought reasoning，CMDM-CoT），一种旨在增强多模态大语言模型推理能力的新方法。针对挑战1，CMDM-CoT引入了自适应问题解决决策集，使模型能够在问题解决过程中自主选择必要决策，摆脱固定框架的限制。这使得在简单问题中进行最小化推理，在复杂问题中进行详细推理，然后提供答案。针对挑

战2，CMDM-CoT包含一个状态评估机制，对每个状态进行评分，确保模型仅学习对问题解决有积极贡献的信息，从而避免低质量数据的干扰。

综上所述，本研究的贡献可以总结如下：

（1）本研究设计了CMDM-CoT，一种旨在通过允许自适应问题解决决策来增强多模态大语言模型推理能力的新方法，从而克服现有模型中固定框架的限制。

（2）本研究开发了一个自适应问题解决决策集，使模型能够在问题解决过程中自主选择必要决策，从而在简单问题中进行最小化推理，在复杂问题中进行详细推理。

（3）本研究在CMDM-CoT中引入了一个状态评估机制，对每个状态进行评分，确保模型仅学习对问题解决有积极贡献的信息，从而避免低质量数据的干扰。

（4）广泛的评估表明，CMDM-CoT在应用于三种主流模型时，在多个基准测试中表现优异，超越了基线模型，甚至超过了参数量更大的闭源模型。

2 相关工作

2.1 思维链推理

思维链[1,13-16]已成为人工智能研究的一个关键领域，特别是在增强大型语言模型推理能力方面。CoT范式促进了将复杂推理任务分解为可管理的步骤，从而提高了模型输出的可解释性和准确性。该领域的早期工作集中在少样本CoT上[17]，模型通过提示几个示例来生成连贯的思维链[16]。后来，这种方法扩展到零样本CoT[18]，利用诸如“Let's think step by step”的提示来激发推理，而无需明确的示例。值得注意的进展包括Zhang等[11]的工作，他们引入了一个两阶段的多模态CoT方法，将文本和视觉特征整合起来以推导理由，然后生成最终答案。同样，Xu等[12]提出了LLaVA-CoT，结合多个推理阶段以实现扩展的思维链。Zheng等[19]提出了DDCoT，通过负空间认知记忆提示增强模型推理能力。Zhu等[20]提出了DYVAL，基于有向无环图动态生成推理数据。Jiang等[21]提出了Corvid，利用大模型对多源推理数据进行标准化重写和质量筛选。尽管取得了这些进展，但在生成结构化和有效的思维链方面仍然存在挑战，特别是在多样化和复杂任务中，本研究提出的CMDM-CoT旨在解决这些问题。

2.2 多模态语言模型

多模态语言模型通过整合文本和视觉数据显著推进了视觉问答（Vision Question Answer，VQA）领域[5-6,22-23]发展，提供了更全面的答案。初始模型旨在直接回答视觉查询，但它们常常在需要更深入推理的复杂问题上表现不佳[24-25]。这一限制促使在MLLMs中探索CoT推理，如零样本CoT方法的采用[18]。然而，现有模型在整个推理过程中经常遇到保持逻辑一致性和质量的困难[26]。VQA任务的多样性和复杂性需要能够动态调整以适应不同任务需

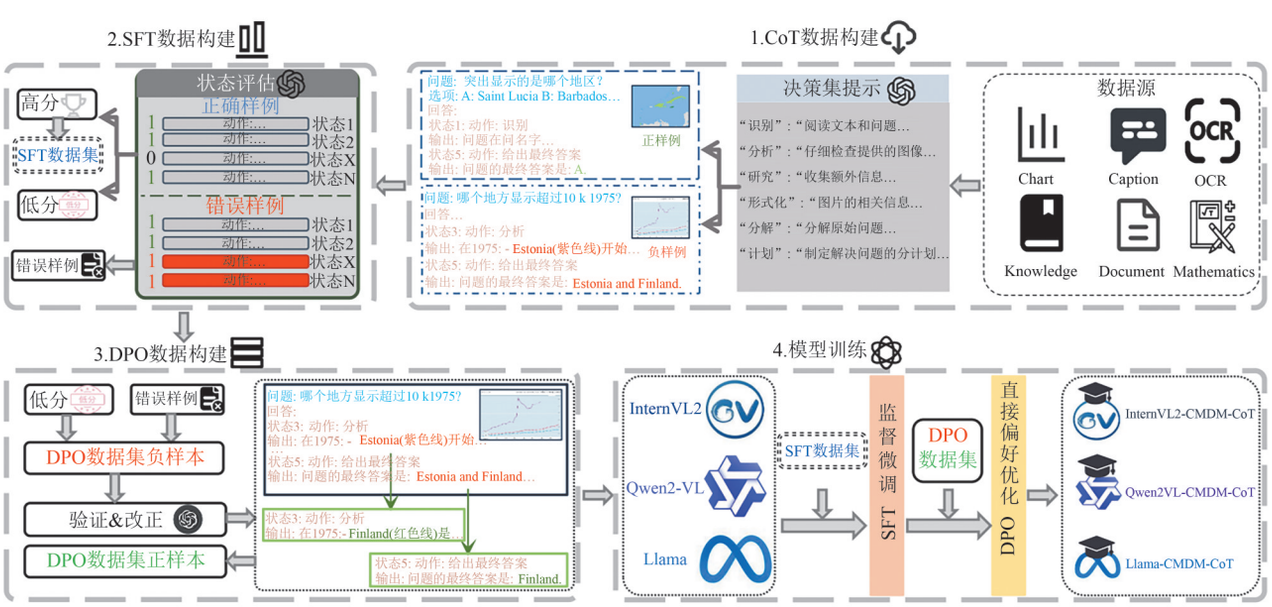

图2 基于思维链推理的可控多模态决策方法四个阶段

求的自适应推理策略。本研究的CMDM-CoT通过引入自适应问题解决决策集和状态评估机制来解决这些挑战，确保推理过程中的逻辑一致性和高质量学习。这种方法不仅增强了MLLMs的推理能力，还提高了它们在多样化VQA基准上的表现。

3　方法论

在本节中，本研究详细介绍了生成思维链数据的流程，这对于增强多模态大语言模型的推理能力至关重要。如图2所示，思维链数据的生成分为两个主要阶段：准备阶段和CoT生成流程。每个阶段都经过精心设计，以确保生成高质量的数据，用于训练先进的多模态大语言模型。

3.1　初步准备

3.1.1　数据收集

本研究的数据来源于各种开源数据集，分为六种类型的多模态问答任务，总计约57,000个多模态问答对，并对4,000个多项选择题的选项进行打乱以创建新数据，使模型能够从不同角度进行推理以得出正确答案。这些数据集包括ChartQA[27]、DocVQA[28]、TextVQA[29]、InfoVQA[30]和ScienceQA[31]，它们专注于简单的视觉任务，总计18,200个问答对。相比之下，Geometry3K[32]、GeoQA[33]和Super-CLEVR[34]则针对复杂推理任务。MathV360K[35]是一个多源数据集，包含简单视觉任务和复杂推理任务。这种多样化的数据集确保本研究的CoT生成流程具有鲁棒性，能够处理广泛的多模态问答任务。

3.1.2　决策集定义

决策集A代表一个精心设计的框架，类似于人类在解决问题时固有的认知过程。如表1所示，该框架包含十二个不同的决策，系统地分为四个主要组成部分：理解、问题解决、验证和回答。每个组成部分反映了人类认知处理的关键阶段，从而增强模型处理和响应复杂查询的能力，提高精确度和深度。这种结构化的方法确保问题解决的每个阶段都得到彻底解决，促进对复杂问题的更细致和准确的响应。具体定义如附录A所示。

理解：理解阶段是整个问题解决过程的基础。它包括识别、分析、研究和形式化等决策。这些决策至关重要，因为它们专注于文本和图像的初步理解和识别。该阶段的灵感来自人类感知和解释信息的能力，为更复杂的推理任务奠定基础。

问题解决：问题解决阶段包括分解、计划、解决子问题和解决父问题等决策。该阶段受到人类在面对复杂问题时常用的分而治之策略的启发。

表1　决策集定义

阶段	决策	操作
理解	识别	彻底阅读文本和问题...
	分析	仔细检查提供的图像...
	研究	收集额外的信息或...
	形式化	将图像中的信息联系起来...
解决问题	分解	将原始问题分解为...
	计划	制定解决问题的逐步计划...
	解决子问题	解决未解决的子问题之一...
	解决父问题	使用子问题的解决方案...
确认	重新审视	重新检查图像或文本以确保...
	验证	检查准确性和有效性...
	回顾	回顾整个解决过程...
回答	给出最终答案	根据... 提供最终答案

验证：验证阶段包括重新审视、验证和审查等决策。该阶段允许模型自我检查、核对和纠正其推理过程，类似于人类检查自己工作的做法。

回答：回答阶段由一个决策组成：给出最终答案。该阶段是模型综合整个推理过程，根据给定的范式输出答案的地方。

3.2　CoT生成流程

3.2.1　CoT数据生成

本研究将定义的决策集整合到提示词中，以指导多模态大语言模型G在每个状态下自主选择适当的决策，从而生成既连贯又具有上下文相关性的状态转换数据。为此，本研究将A融入提示P(A)，使其能够利用前面步骤中收集的多样化数据集。具体而言，对于每个输入图像I和文本T，本研究使用P(A)和G(如GPT-4o[36])自适应地生成本研究的状态转换数据R。相关提示如附录B所示。该过程具体表述如下：

$$R = G\left(P(A); \{I_i, T_i\}_{i=1}^{N_1}\right) \tag{1}$$

3.2.2　高质量SFT数据的构建

过程评估：以往的工作常常忽略了CoT过程的评估。随着OpenAI的o1模型的发布，对CoT质量分析的重视日益增加。本研究的方法引入了一种简单而有效的方法，如图6所示，通过六个维度来评估CoT质量：(1)准确性：评估每一步是否正确处理输入信息并产生准确的中间结果，可以与标准答案进行比较。(2)相关性：评估每一步生成的中间结果是否与最终答案相关。高度相关的步骤通常对最终结果有积极贡献。(3)连贯性：检查思维链中步骤之间

的逻辑连贯性。每一步的输出应自然地引导到下一步的输入。(4)信息增益:评估每一步是否增加了对问题的理解或提供了新信息,帮助更好地回答问题。(5)效率:评估每一步的计算资源和时间消耗。高效的步骤在较短时间内提供有用信息。(6)可解释性:评估每一步的输出是否易于理解和解释。可解释的步骤有助于分析整个思维链的有效性。

数据构建过程:本研究引入另一个多模态判断模型J,如Qwen2-VL-7B,将R与真实值GT进行比较以获得正确数据R_t和错误数据R_f。然后,本研究使用多模态评估模型ε,如GPT-4o,评估正确数据R_t的每个状态,基于其对问题解决的正面或负面影响从−1到1进行评分,并提供理由。最后,人工验证将平均得分不超过0.8的数据标记为低质量数据R_t^l,平均得分为0.8或更高的数据标记为高质量数据R_t^h,形成本研究的高质量监督微调(Supervised Fine-Tuning,SFT)数据D_{SFT},数据集的相关信息如表2所示。该过程具体表述如下:

$$R_t, R_f = J\left(P_J; \{R_i, GT_i\}_{i=1}^{N_1}\right) \tag{2}$$

$$R_t^h, R_t^l = \varepsilon(P_\varepsilon; R_t) \tag{3}$$

$$D_{SFT} = R_t^h \tag{4}$$

3.2.3 高质量DPO数据的构建

本研究的高质量直接偏好优化(Direct Preference Optimization,DPO)数据D_{DPO}来自三个部分。第一部分是状态转换数据生成过程中产生的错误数据R_f。本研究允许多模态大语言模型V(如GPT-4o)识别错误步骤并继续生成正确答案R_{f-t}。第二部分来自低质量数据R_t^l。本研究将R_t^l和状态评估结果添加到提示P_I中,允许V基于评估生成高质量答案R_t^{l-h}。第三部分来自后SFT模型M_{SFT}。本研究使用M_{SFT}对训练数据D_{SFT}进行推理,ε评判训练数据D_{SFT}中的错误数据R_{SFT_f}及其对应的高质量答案,形成本研究的高质量DPO数据D_{DPO}。该过程具体表述如下:

$$R_{f-t} = V(P_V; R_f) \tag{5}$$

$$R_t^{l-h} = \varepsilon(P_I; R_t^l) \tag{6}$$

$$R_{SFTf} = \varepsilon(P_\varepsilon; M_{SFT}(D_{SFT})) \tag{7}$$

$$D_{DPO} = (R_{f-t}, R_f) \cup (R_t^{l-h}, R_t^l) \cup (D_{SFT}, R_{SFT_f}) \tag{8}$$

4 模型训练

在Llama上使用基于思维链推理的可控多模态决策方法(Controlled Multimodal Decision-making Method based on Chain-of-Thought reasoning,CMDM-CoT)也就是Llama-CMDM-CoT,Llama-CMDM-CoT的训练过程类似于之前的工作。主要分为两个阶段:监督微调和直接偏好优化。在第一阶段,本研究从预训练的多模态大语言模型开始,使用精心构建的高质量SFT数据对模型进行监督和微调,训练模型学习自适应生成状态转换数据的能力。在第二阶段,本研究使用构建的高质量DPO数据直接优化模型的偏好。这种优化进一步提高了模型输出的思维链数据质量。

表2 数据分布

任务	数据集	大小	平均状态数
图表	ChartQA	6.6 k	4.6
文档	DocVQA	4.2 k	4.0
通用问答	MathV360K	6.9 k	5.5
知识	ScienceQA	9.9 k	6.6
数学	Geometry3K	1.1 k	7.8
数学	GeoQA	2.9 k	7.9
数学	Super-CLEVR	2.0 k	5.3
OCR	TextVQA	2.9 k	4.0
OCR	InfoVQA	4.5 k	4.7

4.1 监督微调

然而,鉴于复杂推理任务的数据集稀缺且复杂,本研究提出将监督微调过程分为两个不同阶段。这种方法旨在解决这些数据集的有限可用性和高难度所带来的挑战,确保Llama-CMDM-CoT的训练更加稳健和全面。在监督微调过程的第一阶段,本研究利用简单视觉任务和多源数据集来训练多模态大语言模型。该阶段重点在于赋予模型执行自适应状态转换推理的基础能力。通过利用简单的视觉任务,模型可以在一个可控且不太复杂的环境中逐步建立其推理能力。多源数据集提供多样化的数据输入,使模型能够从各种场景和上下文中学习。这种基础训练至关重要,因为它使模型具备处理后续阶段更复杂推理任务的必要技能。监督微调过程的第二阶段针对使用专门设计的复杂推理任务数据集对模型进行监督和微调。该阶段对于细化模型处理复杂和具有挑战性的推理场景的能力至关重要。通过专注于复杂推理数据集,模型可以进一步磨炼其技能,深入理解这些任务中的细微差别,确保模型能够很好地应对现实世界中的推理挑战并提高其整体性能和准确性。

4.2 直接偏好优化

在对模型监督微调之后,直接偏好优化阶段仍

然是训练过程的重要组成部分。在这一阶段，本研究利用高质量DPO数据直接优化模型的偏好。这一优化过程对于微调模型的输出质量，特别是在CoT推理方面至关重要。通过直接优化模型的偏好，本研究可以确保输出符合所需的推理模式和质量标准。该阶段补充了增强的SFT过程，进一步细化模型的能力，确保在视觉推理任务中表现出色。

5 实 验

在本节中，本研究进行了一系列实验以验证CMDM-CoT的有效性。本研究的实验重点关注以下几个方面：(1) 实验设置，(2) 整体性能，(3) 决策集的合理性，(4) 评分模型的有效性，(5) 消融实验，(6)定性分析。

5.1 实验设置

本研究使用官方推荐的框架训练模型。对于Llama-3.2-11B-Vision-Instruct[37]的训练，由于官方训练代码和推荐框架没有开源，本研究使用Swift框架进行训练。对于InternVL2[38]的训练，本研究使用官方提供的训练代码。对于Qwen2-VL[39]的训练，本研究采用Qwen2-VL团队推荐的LLaMA-Factory框架。所有训练参数均设置为各自模型的推荐值。为了确保本研究方法的有效性，本研究在多个视觉语言基准上进行了广泛的实验，包括ChartQA、MathVerse、MathVision、MathVista、MME、ScienceQA。这些基准在评估大型视觉语言模型(Large Vision Language Models，LVLMs)的性能方面被广泛认可。为了公平和一致的评估，本研究使用了VLMEvalKit框架，这是一个专为评估LVLMs设计的开源工具包。

5.2 整体性能

基准数据集：本研究选择了六个知名且具有代表性的基准，涵盖综合多模态评估、知识评估、图表理解和数学问题解决。ChartQA[26]包括9,600个人工编写的问题和23,100个从人工编写的图表摘要生成的问题，提供了对图表理解的全面评估。MathVerse[40]包含2,612个高质量、多学科的数学问题及图表，精心收集自公开可用的来源。MathVision[41]是一个精心策划的集合，包含3,040个具有视觉背景的高质量数学问题，来源于真实的数学竞赛。MathVista[42]包括从31个不同数据集中收集的6,141个例子，提供了多样化的数学挑战。MME[43]测量了14个子任务中的感知和认知能力。ScienceQA[31]涵盖26个主题、127个类别和379项技能，涉及广泛的领域。值得注意的是，MathVerse和MathVista分别使用MathVerse mini和MathVista mini数据集进行特定评估。

主要结果：实验结果如表3所示，清楚地展示了本研究提出的CMDM-CoT方法在增强多模态大语言模型推理能力方面的有效性。本研究的模型Llama-CMDM-CoT、Qwen2VL-CMDM-CoT和InternVL2-CMDM-CoT始终优于基线模型，突显了CMDM-CoT的有效性。Llama-CMDM-CoT、Qwen2VL-CMDM-CoT和InternVL2-CMDM-CoT分别获得了61.2、62.7和62.8的平均分。这些分数显著超过了基线模型，如Llama-3.2-11B-Vision-Instruct、InternVL2-8B和Qwen2-VL-7B-Instruct，分别得分为53.3、61.5和59.7。此外，本研究的模型即使与闭源模型相比也表现出竞争力。例如，本研究的模型优于GPT-4V，其平均得分为57.8，强调了本研究的开源模型能够与参数量更大的闭源模型竞争并超越的潜力。

除了整体竞争性能外，本研究的模型在多个基准上表现出色。InternVL2-CMDM-CoT在MathVision基准上获得了最高分20.9，而Qwen2VL-CMDM-CoT在MathVerse中表现出色，得分为38.2。这些结果强调了CMDM-CoT增强模型在多样化和具有挑战性的任务中的稳健性和适应性，进一步验证了本研究方法在推进多模态推理领域的先进性方面的优越性。因时间和资源有限，实验结果中“-”表示未评测且官方数据缺失。值得注意的是，本研究的模型在特定基准上也表现出色。Llama-CMDM-CoT在ChartQA上获得了86.6的最高分，超过了闭源的GPT-4o-20240513和开源的Llama-3.2-90B-Vision-Instruct。这些结果突显了CMDM-CoT在增强多模态大语言模型推理能力方面的有效性，使其能够自适应地生成结构化和有效的思维链，以应对多样化的VQA任务。相关比较如附录C所示。

5.3 决策集的合理性

为评估本研究决策集的合理性，本研究进行涉及CMDM-CoT和其他基线模型的比较分析，特别是Llama-3.2V-11B-zero-shot-CoT模型和Llama-3.2V-11B-CoT[6]，Llama-3.2V-11B-CoT是对Llama-3.2-11B-Vision-Instruct模型采用思维链推理方法开发，而Llama-3.2V-11B-zero-shot-CoT模型，是在Llama-3.2-11B-Vision-Instruct视觉语言

表 3 多模态大语言模型在不同基准上的性能比较

	模型	ChartQA	MathVerse	MathVision	MathVista	MME	ScienceQA	平均
闭源模型	GPT-4o-20240513	85.7	50.2	30.4	63.8	2328.7	90.1	67.2
	Claude3.5-Sonnet	90.8	-	38.0	67.7	1920.0	88.9	-
	Gemini-1.5-Pro	87.2	51.1	19.2	63.9	2110.6	85.7	63.7
	GPT-4o-mini-20240718	-	36.7	28.8	52.4	2003.4	85.4	-
	GPT-4V	78.5	32.8	24.0	58.1	1926.6	84.8	57.8
	Cambrian-34B	75.6	-	-	50.3	2049.9	85.6	-
	GLM-4V-9B	71.1	35.7	15.3	51.1	2018.8	96.7	57.0
	Pixtral-12B	81.8	-	-	56.9	1921.7	87.2	-
	LLaVA-NeXT-Yi-34B	67.6	-	-	40.4	2006.5	82.0	-
开源模型	Llama-3.2-90B-Vision-Instruct	85.5	-	-	58.3	1741.0	87.1	-
	Phi-3.5-vision-instruct	81.8	28.0	15.5	43.9	1838.1	91.3	54.4
	Llama-3.2-11B-Vision-Instruct	83.4	22.9	13.1	51.5	1820.5	83.9	53.3
	Qwen2-VL-7B-Instruct	83.0	33.6	16.3	58.2	**2326.8**	85.8	59.7
	InternVL2-8B	82.0	33.3	18.4	58.3	2227.7	**97.3**	61.5
	Llama-CMDM-CoT	**86.6**	32.3	20.5	54.5	2198.3	94.6	61.2
所提方法	Qwen2VL-CMDM-CoT	85.0	**38.2**	19.8	57.9	2311.3	92.8	62.7
	InternVL2-CMDM-CoT	83.2	36.3	**20.9**	**59.5**	2311.4	94.7	**62.8**

表 4 使用 CMDM-CoT 效果对比

模型	ChartQA	MathVerse	MathVision	MathVista	MME	ScienceQA	平均
Qwen2-VL-7B-Instruct	83.0	33.6	16.3	**58.2**	2326.8	85.8	59.7
Qwen2VL-CMDM-CoT w/o DPO	84.4	**39.0**	**20.1**	56.9	**2347.6**	91.2	**62.8**
Qwen2VL-CMDM-CoT	**85.0**	38.2	19.8	57.9	2311.3	**92.8**	62.7
InternVL2-8B	82.0	33.3	18.4	58.3	2227.7	**97.3**	61.5
InternVL2-CMDM-CoT w/o DPO	**84.3**	34.9	**21.0**	56.5	2305.2	95.0	62.3
InterVL2-CMDM-CoT	83.2	**36.3**	20.9	**59.5**	**2311.4**	94.7	**62.8**
Llama-3.2-11B-Vision-Instruct	83.4	22.9	13.1	51.5	1820.5	83.9	53.3
Llama-3.2V-11B-zero-shot-CoT	81.0	25.2	13.4	50.4	1900.3	90.8	54.8
Llama-3.2V-11B-CoT	75.6	32.7	15.8	51.9	2166.4	94.4	57.9
Llama-CMDM-CoT w/o DPO	85.2	29.6	18.2	52.1	**2203.2**	**95.9**	59.9
Llama-CMDM-CoT w/o SE	84.7	**33.0**	17.9	52.5	2154.0	94.8	60.0
Llama-CMDM-CoT	**86.6**	32.3	**20.5**	**54.5**	2198.3	94.6	**61.2**

模型基础上的创新改进，模型通过引入零样本思维链（Zero-shot Chain-of-Thought，Zero-shot CoT）提示机制，在基准测试中采用“let's think step by step”的提示策略，有效激活了预训练模型的潜在推理能力。其中使用GPT-4o生成CoT数据以微调Llama-3.2-11B-Vision-Instruct模型。相比之下，Llama-3.2V-11B-CoT为多模态大语言模型提供了一个固定的四步框架，但缺乏明确的决策指导。

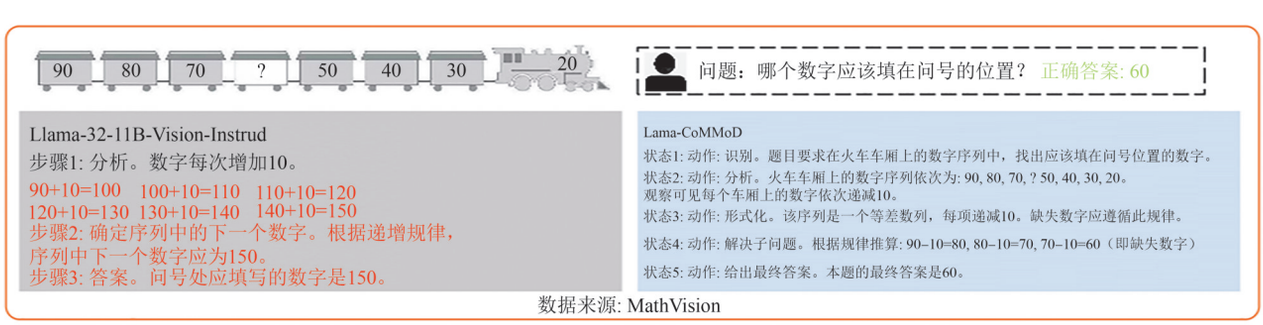

图 3 定性分析

5.3.1　性能比较分析

通过CMDM-CoT与其他基线模型（如Llama-3.2V-11B-zero-shot-CoT模型和Llama-3.2V-11B-CoT）的比较分析，突显了决策集的合理性。CMDM-CoT模型在各种基准上表现出色，如表4所示，其更高的平均分数表明CMDM-CoT采用的决策集有效地增强了模型处理和解释复杂多模态数据的能力。相关的具体资源损耗如附录D所示。

5.3.2　决策过程的重要性

此外，尽管Llama-3.2V-11B-zero-shot-CoT模型在使用零样本思维链推理方法方面具有创新性，但其性能未达到CMDM-CoT的水平。这表明决策集中包含明确的决策过程对于优化模型性能至关重要。Llama-3.2V-11B-CoT模型由于其固定的四步框架，在比较中也显得不足，强调了决策框架中的灵活性和适应性的重要性。

5.4　评分模型的有效性

在本研究的实验分析中，本研究进行了比较研究，以评估状态评估（State Evaluation，SE）组件的有效性，通过对比Llama-CMDM-CoT在有无SE情况下在各种基准上的表现。结果如表4所示，显示出在引入SE后模型性能显著提升。配备SE的Llama-CMDM-CoT模型在大多数基准上始终优于其无SE的对应模型，即Llama-CMDM-CoT w/o SE模型。具体而言，在ChartQA基准上，Llama-CMDM-CoT获得了86.6的分数，超过了无SE的Llama-CMDM-CoT的84.7分。这一趋势在MathVision和MathVista基准中同样得到体现，其中引入SE后分别从17.9提高到20.5和从52.5提高到54.5。这些增量突显了SE对模型理解和处理复杂数学和视觉数据能力的积极影响。此外，所有基准的平均表现进一步证实了SE的有效性。配备SE的Llama-CMDM-CoT获得了61.2的平均分，而无SE的模型为60.0。尽管这一提升较为温和，但它强调了SE在增强模型整体能力方面的累积效益。

表5　消融实验分析

模型	平均
Qwen2VL-CMDM-CoT w/o DPO	**62.8**
Qwen2VL-CMDM-CoT	62.7
InternVL2-CMDM-CoT w/o DPO	62.3
InternVL2-CMDM-CoT	**62.8**
Llama-CMDM-CoT w/o DPO	59.9
Llama-CMDM-CoT	**61.2**

5.5　消融实验

在消融实验中，DPO的影响在不同模型中表现出显著的差异。在表5中显示了每个模型在应用SFT和DPO两种方法后的平均结果。CMDM-CoT w/o DPO表示模型仅经过SFT过程，而CMDM-CoT则是在SFT过程之后还进行了DPO。Qwen2VL-CMDM-CoT w/o DPO的平均分是62.8，而Qwen2VL-CMDM-CoT的平均分是62.7，虽然在这个例子中提升不明显，但在其他模型中，如InternVL2-CMDM-CoT w/o DPO的平均分是62.3，而InternVL2-CMDM-CoT的平均分是62.8，Llama-CMDM-CoT w/o DPO的平均分是59.9，而Llama-CMDM-CoT的平均分是61.2，这些例子都显示了DPO的应用对模型性能的提升作用。总体来看，DPO在大多数情况下都能提高模型的平均效果，说明在SFT过程之后进行DPO优化是有益的。

5.6　定性分析

本研究的模型的定性分析如图3所示，突显了Llama-CMDM-CoT相较于基线模型Llama-3.2-11B-Vision-Instruct的卓越推理能力。在此任务中，目标是识别火车车厢上显示的序列中缺失的数字，其中数字从一个车厢到下一个车厢递减10。Llama-CMDM-CoT展示了一个结构化和逻辑的方法来解决问题。它首先准确识别任务要求，然后详细分析序列模式。模型正确识别出算术级数，其中每项递减10，并形式化此模式以解决子问题。通过继续这一模式，Llama-CMDM-CoT成功确定缺失的数字为60，展示了其保持逻辑一致性和应用系统推理的能力。相比之下，基线模型Llama-3.2-11B-Vision-Instruct误解了模式为递增10，导致错误结论。此错误突显了基线模型推理过程中的一个关键限制，即未能准确分析和应用正确的模式来解决问题。

6　结　论

在本文中，本研究介绍了CMDM-CoT，这是一种旨在增强多模态大语言模型推理能力的新框架。本研究的方法解决了视觉问答任务中的两个主要挑战：对自适应和结构化思维链的需求，以及在整个推理过程中保持逻辑一致性和高质量学习的要求。通过结合自适应问题解决决策集和状态评估机制，CMDM-CoT使模型能够自主选择适当的推理

路径，并确保中间推理步骤的完整性。本研究广泛的实验评估表明，基于思维链推理的可控多模态决策方法显著优于基线模型，验证了本研究方法的有效性。

致 谢 在此我们向对本文的工作给予支持和宝贵建议的评审老师表示衷心的感谢！

参 考 文 献

[1] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 24824-24837

[2] Li Y, Liu Z, Li Z, et al. Perception, reason, think, and plan: A survey on large multimodal reasoning models. arXiv preprint arXiv:2505.04921, 2025

[3] Wang Y, Wu S, Zhang Y, et al. Multimodal chain-of-thought reasoning: A comprehensive survey. arXiv preprint arXiv: 2503.12605, 2025

[4] Jaech A, Kalai A, Lerer A, et al. Openai o1 system card. arXiv preprint arXiv:2412.16720, 2024

[5] Antol S, Agrawal A, Lu J, et al. Vqa: Visual question answering//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 2425-2433

[6] Wu Q, Teney D, Wang P, et al. Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding, 2017, 163: 21-40

[7] Bi J, Liang S, Zhou X, et al. Why reasoning matters? A survey of advancements in multimodal reasoning (v1). arXiv preprint arXiv:2504.03151, 2025

[8] Sun J, Zheng C, Xie E, et al. A survey of reasoning with foundation models. arXiv preprint arXiv:2312.11562, 2023

[9] Yan Y, Su J, He J, et al. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. arXiv preprint arXiv:2412.11936, 2024

[10] Chen Q, Qin L, Liu J, et al. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. arXiv preprint arXiv:2503.09567, 2025

[11] Zhang Z, Zhang A, Li M, et al. Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923, 2023

[12] Xu G, Jin P, Hao L, et al. Llava-o1: Let vision language models reason step-by-step. arXiv preprint arXiv: 2411.10440, 2024

[13] Qiao S, Ou Y, Zhang N, et al. Reasoning with language model prompting: A survey//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada, 2023: 5368-5393

[14] Sun S, An W, Tian F, et al. A review of multimodal explainable artificial intelligence: Past, present and future. arXiv preprint arXiv:2412.14056, 2024

[15] Xia Y, Wang R, Liu X, et al. Beyond chain-of-thought: A survey of chain-of-x paradigms for llms// Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, UAE, 2025: 10795 - 10809

[16] Song Y, Wang T, Cai P, et al. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. ACM Computing Surveys, 2022, 55(13s): 1-40

[17] Zhang Z, Zhang A, Li M, et al. Automatic chain of thought prompting in large language models. arXiv preprint arXiv: 2210.03493, 2022

[18] Kojima T, Gu S S, Reid M, et al. Large language models are zero-shot reasoners//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 22199-22213

[19] Zheng G, Yang B, Tang J, et al. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA, 2023: 5168-5191

[20] Zhu K, Chen J, Wang J, et al. Dyval: Dynamic evaluation of large language models for reasoning tasks. arXiv preprint arXiv: 2309.17167, 2023

[21] Jiang J, Ma C, Song X, et al. Corvid: Improving multimodal large language models towards chain-of-thought reasoning. arXiv preprint arXiv:2507.07424, 2025

[22] Kafle K, Kanan C. Visual question answering: Datasets, algorithms, and future challenges. Computer Vision and Image Understanding, 2017, 163: 3-20

[23] Hartsock I, Rasool G. Vision-language models for medical report generation and visual question answering: A review. Frontiers in Artificial Intelligence, 2024, 7: 1430984

[24] Wang Z, Wan W, Lao Q, et al. Towards top-down reasoning: An explainable multi-agent approach for visual question answering. arXiv preprint arXiv:2311.17331, 2023

[25] Yu Z, He L, Wu Z, et al. Towards better chain-of-thought prompting strategies: A survey. arXiv preprint arXiv: 2310.04959, 2023

[26] Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171, 2022

[27] Masry A, Long D X, Tan J Q, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning//Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland, 2022: 2263 - 2279

[28] Mathew M, Karatzas D, Jawahar C V. Docvqa: A dataset for vqa on document images//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2021: 2199-2208

[29] Singh A, Natarajan V, Shah M, et al. Towards vqa models that can read//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 8309-8318

[30] Mathew M, Bagal V, Tito R, et al. Infographicvqa//

Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2022: 2582-2591

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

[32] Lu P, Gong R, Jiang S, et al. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Online, 2021: 6774-6786

[33] Chen J, Tang J, Qin J, et al. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning// Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online, 2021: 513-523

[34] Li Z, Wang X, Stengel-Eskin E, et al. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada, 2023: 14963-14973

[35] Shi W, Hu Z, Bin Y, et al. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models// Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024. Miami, USA, 2024: 4663-4680

[36] Hurst A, Lerer A, Goucher A P, et al. Gpt-4o system card. arXiv preprint arXiv:2410.21276, 2024

[37] Grattafiori A, Dubey A, Jauhri A, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024

[38] Chen Z, Wang W, Tian H, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. Science China Information Sciences, 2024, 67(12): 220101

[39] Wang P, Bai S, Tan S, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024

[40] Zhang R, Jiang D, Zhang Y, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? // Proceedings of the 18th European Conference on Computer Vision. Milan, Italy, 2024: 169-186

[41] Wang K, Pan J, Shi W, et al. Measuring multimodal mathematical reasoning with math-vision dataset//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada, 2025: 95095-95169

[42] Lu P, Bansal H, Xia T, et al. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. CoRR, 2023

[43] Fu C, Chen P, Shen Y, Qin Y, Zhang M, Lin X, Yang J, Zheng X, Li K, Sun X, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394, 2023

附录A. 决策集

在本节中，我们提供了决策集的全面定义，该决策集概述了系统化的问题解决方法。决策集的结构如下：

• 识别：仔细阅读文本和问题，以掌握上下文和关键要求。识别解决问题所需的关键字和句子。

• 分析：仔细检查提供的图像，记录所有相关细节。从图像中提取解决问题所需的具体信息。

• 研究：如果需要，收集额外的信息或资源来解决问题。这可能包括查找定义、公式或相关概念以增强理解。

• 形式化：将图像中的信息与文本信息和问题联系起来。识别连接，得出新的结论，并构建所有可用信息。重新表达问题并识别需要解决的具体内容。

• 分解：如果原始问题涉及多个步骤或条件，则将其分解为更简单的子问题。列出在解决原始问题之前需要解决的所有必要问题、知识和定理。

• 计划：制定逐步解决问题的计划，概述达到解决方案所需的行动顺序。

• 重新审视：重新检查图像或文本，以确保对问题有更准确和全面的理解，从而提供更好和更精确的答案。

• 验证：检查获得的信息和解决方案的准确性和有效性。确保所有步骤和计算都是正确的。

• 解决子问题：解决在分解操作中识别出的未解决的子问题之一。

• 解决父问题：使用子问题的解决方案构建对原始问题的全面答案。

• 审查：审查整个解决过程以确保完整性和正确性。进行任何必要的调整或更正。

• 给出最终答案：根据分析和子问题的解决方案提供最终答案。输出声明为："我对问题的最终答案是：…"

附录B. 提　示

本节详细介绍了用于生成思维链推理（Chain of Thought，CoT）数据的提示，如图4所示，这是训练我们的模型熟练解决图文问题的关键要素。提示经过精心设计，引导模型通过预定义的决策集，促进系统的、逐步的推理过程，并能够构建全面的解决方案。

这种结构化的方法使模型能够通过将复杂的图文挑战分解为更简单、更易管理的任务，从而提高解决方案的准确性和可靠性。提示作为我们方法的基

```
你是解答图像相关问题的专家。你需要解答一些图文结合的问题，每个问题由一个图像和一个问题组成。你需要使用预定义操作集中的一系列操作逐步解决问题。
你需要输出一系列状态，其中初始状态是原始输入，每个后续状态都通过特定操作从前一个状态过渡而来。每个状态都应包含其操作名称和输出。
以下是解决图文结合问题的具体步骤：
1.识别: 仔细阅读文本和问题，理解上下文和关键要求。识别出对解决问题至关重要的关键词和句子。
2.分析: 仔细检查提供的图像，记录所有相关细节。从图像中提取解决问题所需的具体信息。详细列举图表中与问题相关的信息，但不要过度推理。
3.研究: 收集解决问题所需的其他信息或资源。这可能包括查找定义、公式或相关概念。
4.形式化: 将图像信息与文本信息和问题联系起来。找出其中的联系，得出新的结论，并构建所有可用信息。重新表述问题，并确定需要解决的具体内容。
5.分解: 如果原始问题涉及多个步骤或条件，则将其分解为更简单的子问题。列出解决原始问题之前需要解决的所有必要问题、知识和定理。
6.计划: 制定解决问题的分步计划，概述所需的操作顺序。
7.重新审视: 重新审视图像或文本，以确保对问题有更准确、更全面的理解，并提供更好、更精确的答案。
8.验证: 检查所获得信息和解决方案的准确性和有效性。确保所有步骤和计算均正确无误。
9.解决子问题: 解决在分解过程中发现的其中一个未解决的子问题。
10.解决父问题: 利用子问题的解构建原问题的综合答案。
11.复习: 回顾整个解答过程，确保其完整性和正确性。并进行必要的调整或修正。
12给出最终答案：根据对子问题的分析和解，给出最终答案。输出语句: “我对这个问题的最终答案是: …”
请注意，您可以根据需要，按任意顺序、任意次数执行这些操作。
以下两个示例可以帮助您理解如何按照前面的规则解决数学问题:
……
现在解答这个问题：
问题: {your_problem_here}
```

图4 生成CoT数据提示

础组件，支持生成高质量的CoT数据，这是有效模型训练所必需的。

附录C. 比 较

在本节中，我们详细比较了我们提出的模型Llama-CMDM-CoT与基线模型Llama-3.2V-11B-Instruct和Llama-3.2V-11B-CoT的性能。图5的雷达图提供了模型在六个不同基准测试中的性能的可视化表示：MathVision、MathVerse、ChartQA、ScienceQA、MME和MathVista。雷达图的每个轴对应于这些基准测试之一，分数表示模型在处理每个领域特定任务方面的熟练程度。Llama-CMDM-CoT在ChartQA和ScienceQA方面表现出显著的进步，强调了其在处理和理解复杂数据方面的增强能力。

附录D. 资 源

为了测试CMDM-CoT方法的资源消耗，我们在A800 GPU（8卡，单卡内存为81 920 MB）环境下，运用CMDM-CoT方法的各模型在内存占用和推理时间上表现出不同的特点。如表6所示，Qwen2VL-CMDM-CoT在SFT阶段的单卡内存占用为21 886 MB，DPO阶段增加至57 070 MB，而推理时单卡内存占用降至19 948 MB，其推理速度为8.15秒/迭代。InternVL2-CMDM-CoT的SFT阶段单卡内存占用达到50 767 MB，DPO阶段略降至49 097 MB，推理时单卡内存占用为14 503 MB，推理时间为9.08秒/迭代。Llama-CMDM-CoT在SFT阶段的单卡内存占用为20 861 MB，DPO阶段增至25 952 MB，推理时单卡内存占用为188 35 MB，推理时间为12.20秒/迭代。所有模型的内存占用均未超过单卡内存上限，展现了CMDM-CoT方法在不同模型上的适应性。

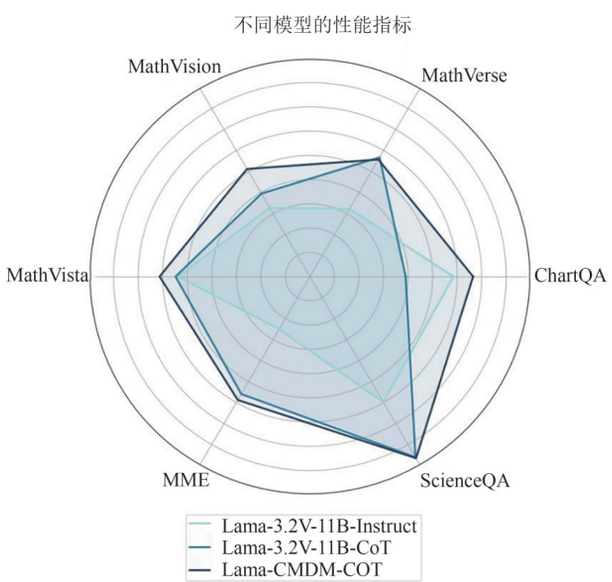

图5 各模型在基准测试上的结果雷达图

```
你是一位专注于推理链分析的数据评估专家。你的任务是通过六个关键维度来评估思维链(CoT)数据的质量和有效性。评估将考察CoT推理过程在多大程度上展示了准确的逻辑进展、保持与原始问题的相关性、提出连贯的论点、提供有意义的信息价值、实现高效的问题解决以及提供清晰的可解释性。每个维度都有助于理解推理过程的整体质量。
评估维度：
1.准确性: 衡量推理步骤和最终结论是否正确地解决了问题，并与真实答案一致。逻辑流程应事实正确且无错误。
2.相关性: 评估每个推理步骤与最终答案的关联程度。所有步骤都应有意义地贡献于解决原始问题，避免离题。
3.连贯性: 评估推理过程的逻辑流和结构完整性。论证应组织良好，连续步骤之间有清晰的联系。
4.信息增益: 判断推理是否提供了实质性的新见解，还是仅仅重新表述了现有信息。高质量的CoT应通过有价值的中间结论加深理解。
5.效率: 评估每一步所需的计算资源和时间消耗。该过程应在保持全面性的同时避免不必要的复杂性。
6.可解释性: 评估人类理解和跟随推理过程的难易程度。解释应透明，并对每一步给出清晰的理由。
输入格式：
问题: {question}
CoT数据: {cot_data}
真实答案: {Groundtruth}
输出格式:
总体评分: {一个1-5的数值评分，其中5为优秀}
评分理由: {涵盖所有六个评估维度的详细解释，突出优点和缺点}
示例评估:
问题: 如果一列火车在5小时内行驶了300英里，它的平均速度是多少?
CoT数据：首先，我们知道平均速度是指在给定时间内覆盖了多少距离。题目给出了总距离为300英里，总时间为5小时。为了求速度，我们可以看看每小时行驶多少英里。如果300英里是在5小时内完成的，那么在一小时内，火车将覆盖该距离的五分之一。计算得出，300英里平均分配到5小时，意味着火车每小时行驶60英里。我对这个问题的最终答案是60英里/小时。
真实答案: 60英里/小时。
总体评分5
评分理由: 该CoT通过正确应用速度公式并得出正确答案，展现了完美的准确性。所有步骤都与速度计算问题高度相关。推理过程完全连贯，步骤之间有清晰的联系。它通过明确陈述公式并展示其应用，提供了良好的信息增益。解决方案高效，没有冗余步骤。该过程具有高度可解释性，因为每个数学运算都得到了清晰的论证。这体现了一个最优的思维链过程。
```

图6 状态评估标准

表 6 各模型内存占用和推理时间

模型	SFT内存占用(MB)	DPO内存占用(MB)	推理内存占用(MB)	推理时间(s/it)
Qwen2VL-CMDM-CoT	21 886	57 070	19 948	8.15
InternVL2-CMDM-CoT	50 767	49 097	14 503	9.08
Llama-CMDM-CoT	20 861	25 952	18 835	12.20

HU Yu-Hang, M. S. candidate. His research interest is multimodal understanding and application.

WANG Shi-Han, M. S. candidate. His research interest is multimodal understanding and application.

LIU Li-Long, M. S. candidate. His research interests include artificial intelligence and natural language processing.

YANG Zhen-Yu, Ph. D. candidate. His research interests include multimodal understanding and application.

QIAN Sheng-Sheng, Ph. D. associate professor. His research interests include multimedia content analysis, data mining, cross-modal retrieval, and personalized recommendation.

Background

The research presented in this paper falls within the domain of artificial intelligence, specifically focusing on enhancing the reasoning capabilities of multimodal large language models (MLLMs) in the context of visual question answering (VQA). VQA is a challenging task that requires models to interpret and reason about visual and textual data to provide accurate answers to questions related to images. Despite significant advancements in the field, current MLLMs often struggle with structured and effective reasoning, particularly when faced with complex and diverse VQA tasks. Internationally, the concept of Chain of Thought (CoT) reasoning has garnered considerable attention as a means to improve the reasoning abilities of large language

models. CoT reasoning involves breaking down complex reasoning tasks into manageable steps, thereby enhancing the interpretability and accuracy of model outputs. Early work in this area focused on few-shot CoT, where models were prompted with a few examples to generate coherent reasoning chains. This approach was later extended to zero-shot CoT, which uses prompts like "let's think step by step" to elicit reasoning without explicit examples.

Despite few advancements, generating structured and effective reasoning chains remains a challenge, especially for diverse and complex tasks. Existing models often face difficulties in maintaining logical consistency and quality throughout the reasoning process. The diversity and complexity of VQA tasks necessitate adaptive reasoning strategies that can dynamically adjust to different task requirements. This is where the proposed Controlled Multimodal Decision-making Method based on Chain-of-Thought reasoning (CMDM-CoT) aims to make a significant contribution.

CMDM-CoT introduces an adaptive problem-solving decision set, allowing models to autonomously select appropriate reasoning paths based on task complexity. This approach overcomes the limitations of fixed frameworks, enabling minimal reasoning for simple tasks and detailed reasoning for complex problems. Additionally, CMDM-CoT incorporates a state evaluation mechanism that scores each reasoning state to ensure logical consistency and high-quality learning. This mechanism prevents the model from learning incorrect information from poor-quality data, thereby enhancing the overall reasoning process. The experiments demonstrates that CMDM-CoT significantly improves the performance of MLLMs in VQA tasks. When applied to three mainstream models—Llama, Qwen2VL, and InternVL2—CMDM-CoT outperformed baseline models by an average of 7.3%. Moreover, these models even surpassed the larger proprietary model GPT-4V in several benchmark tests, showcasing the competitiveness of the open-source models developed in this study.

This paper addresses the challenges of structured and effective reasoning in MLLMs for VQA tasks by introducing the CMDM-CoT framework. The framework's adaptive decision set and state evaluation mechanism ensure logical consistency and high-quality learning, significantly enhancing the reasoning capabilities of MLLMs and improving their performance on diverse VQA benchmarks.

This project was supported by the National Key Research and Development Program of China (2023YFC3310700), the Beijing Natural Science Foundation (JQ23018), and the National Natural Science Foundation of China (62276257).