多模态思维链推理：一项综合调查

Yaoting Wang[1], Shengqiong Wu[1], Yuechen Zhang[2],
Shuicheng Yan[1], Ziwei Liu[3], Jiebo Luo[4], Hao Fei[1*]
[1]NUS, [2]CUHK, [3]NTU, [4]UR

Survey Project: https://github.com/yaotingwangofficial/Awesome-MCoT

Abstract

通过将人类类似逐步推理的链-of-thought（CoT）优势扩展到多模态环境中，多模态 CoT（MCoT）推理最近引起了显著的研究关注，尤其是在与多模态大型语言模型（MLLMs）的整合方面。现有的 MCoT 研究设计了各种方法和创新的推理范式，以应对图像、视频、语音、音频、3D 和结构化数据等不同模态的独特挑战，在机器人、医疗保健、自动驾驶和多模态生成等领域取得了广泛的成功。然而，MCoT 仍然呈现出独特的挑战和机遇，需要进一步关注，以确保该领域的持续繁荣发展，遗憾的是，目前缺乏对该领域的最新综述。为填补这一空白，我们呈现了首个系统性的 MCoT 推理综述，阐明了相关基础概念和定义。我们从多个应用情景的不同视角提供了全面的分类和深入分析当前的方法。此外，我们还探讨了现有挑战和未来的研究方向，旨在推动向多模态通用人工智能（AGI）的创新。

Keywords— Multimodal Reasoning, Chain-of-Thought, Multimodal Large Language Models

*对应作者.（haofei37@nus.edu.sg）

目录

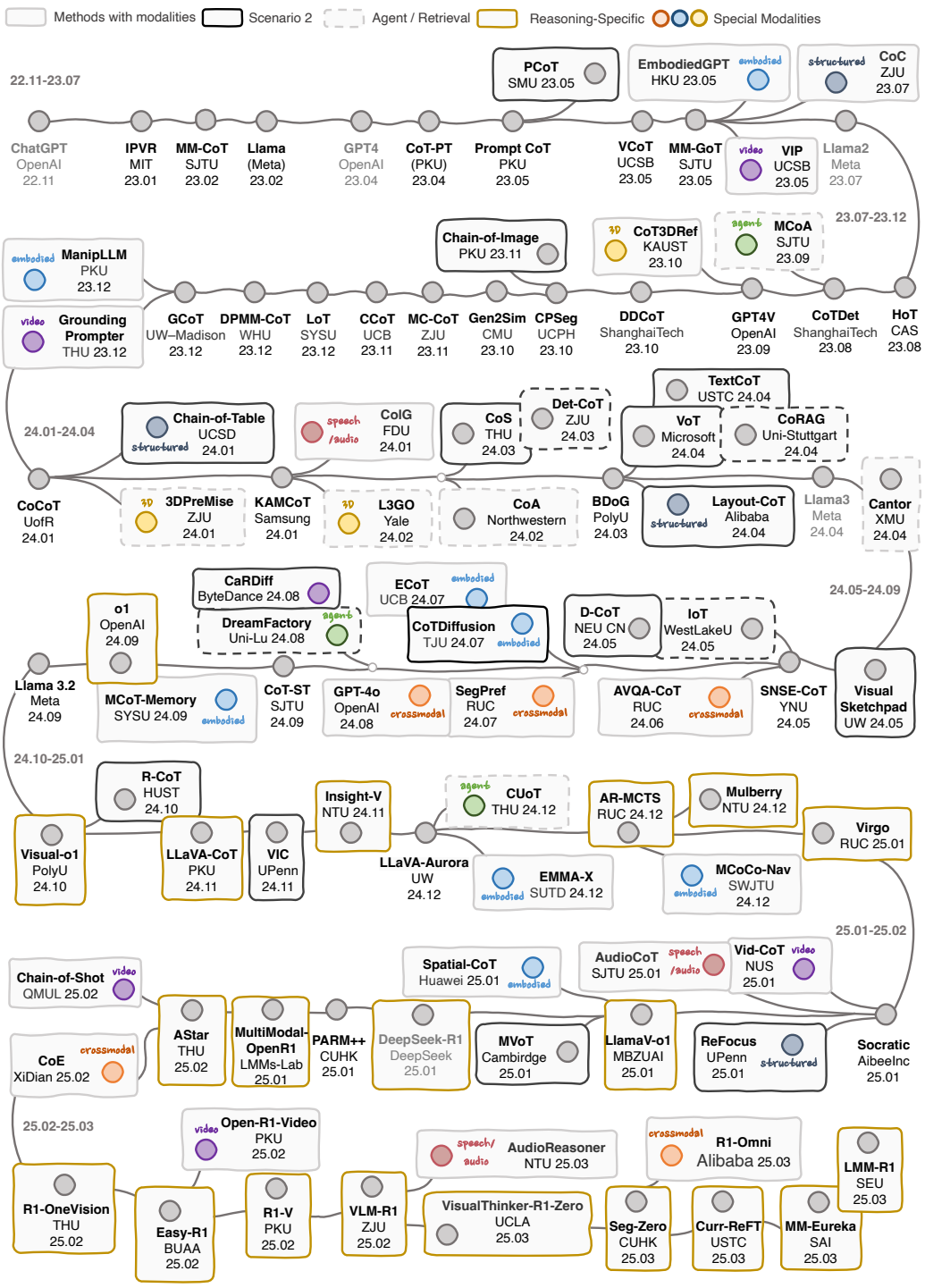


图 1: 开发多模态思维链（MCoT）推理的时间线。名称在gray中的模型是仅文本的 LLMs。为清晰起见，图中的模型默认包括图像模态，除非用彩色圆圈指示特殊模态。

1 引言

大型语言模型（LLMs）[1–7] 的出现开启了人工智能（AI）领域前所未有的时代。长期以来，人们认识到与现实世界环境中固有的多模态特性保持一致的必要性，相应地，AI 领域从 LLMs 演化到多模态 LLMs（MLLMs）[8–18]，将多种模态整合到语言智能中。实现人类水平的智

能需要超越基本的感知能力，达到复杂的认知推理——这是人类认知的一个标志，它通过上下文理解和自我纠正实现迭代推理。受此启发，在上下文学习（ICL）技术已赋予 LLMs 展示逐步推理的能力——通常称为思维链（CoT）推理机制 [19–24]。该技术使模型能够将问题分解为一系列中间步骤，提高决策透明度并在复杂推理任务中的表现。CoT 推理在各种下游复杂任务中的卓越成功推动了其在学术界和工业界的广泛应用。特别是最近在像 OpenAI 的 o1/o3 [25] 和 DeepSeek R1 [26] 这样的尖端系统中隐式集成这一能力，引起了广泛关注。

将思维链推理整合到多模态情境中，随后催化了人工智能领域的变革性进展，催生了多模态思维链（MCoT）推理 [27, 28]。由于思维链特性和跨模态数据交互的异质性，MCoT 主题产生了广泛创新成果。一方面，原始的思维链框架已演变为先进的推理架构，从线性序列 [19] 到基于图的表示 [23]，包含分层的思维结构。另一方面，与单模态文本设置不同，视觉、听觉和时空数据等多样化模态需要专门的处理策略——视觉推理要求精确感知和分析静态场景、对象关系，而视频理解则需要强大的时序动态建模。这些需求推动了多种复杂的 MCoT 方法的发展，以适应模态特定特性，如 Multimodal-CoT[29]、MVoT[30]、Video-of-Thought[31]、Audio-CoT[32]、Cot3DRef[33] 和 PARM++[34]。MCoT 的显著有效性也促使其在自动驾驶 [35–38]、具身 AI[39–41]、机器人技术 [42–45] 和医疗保健 [46–50] 等关键领域成功应用，确立其作为实现多模态通用人工智能的基础技术地位。

近年来，关于 MCoT 的研究引起了越来越多的重视。图1展示了这一新兴领域的关键里程碑的时间线。尽管它在增强多模态推理方面展现出巨大潜力，但 MCoT 也带来了重大挑战，并留下了一些关键问题尚未解决——例如，确定利用多样化多模态上下文的最有效策略，设计真正提升 MLLMs 推理能力的 CoT 过程，以及在这些模型中实现隐式推理。值得注意的是，缺乏全面的综述阻碍了该新兴领域中的知识整合。为填补这一关键空白，本文提供了对 MCoT 推理的第一个系统性概述，提供了一个技术发展、方法论、实际应用和未来方向的结构化分析。我们希望这篇综述能成为权威参考，推动该快速发展的领域进一步创新和进步。

1.1 贡献

- **第一项调查：** 本文代表了首次专门针对 MCoT 推理进行全面回顾的调查。
- **综合分类法：** 我们提出了一种细致的分类法（参见 Figure 2），该分类法对 MCoT 研究中的各种方法进行了分类。
- **前沿与未来方向：** 我们讨论了 emerging 挑战并概述了未来研究的有前景的方向。
- **资源共享：** 我们整理并公开所有相关资源，以支持和加速研究社区内的进展。

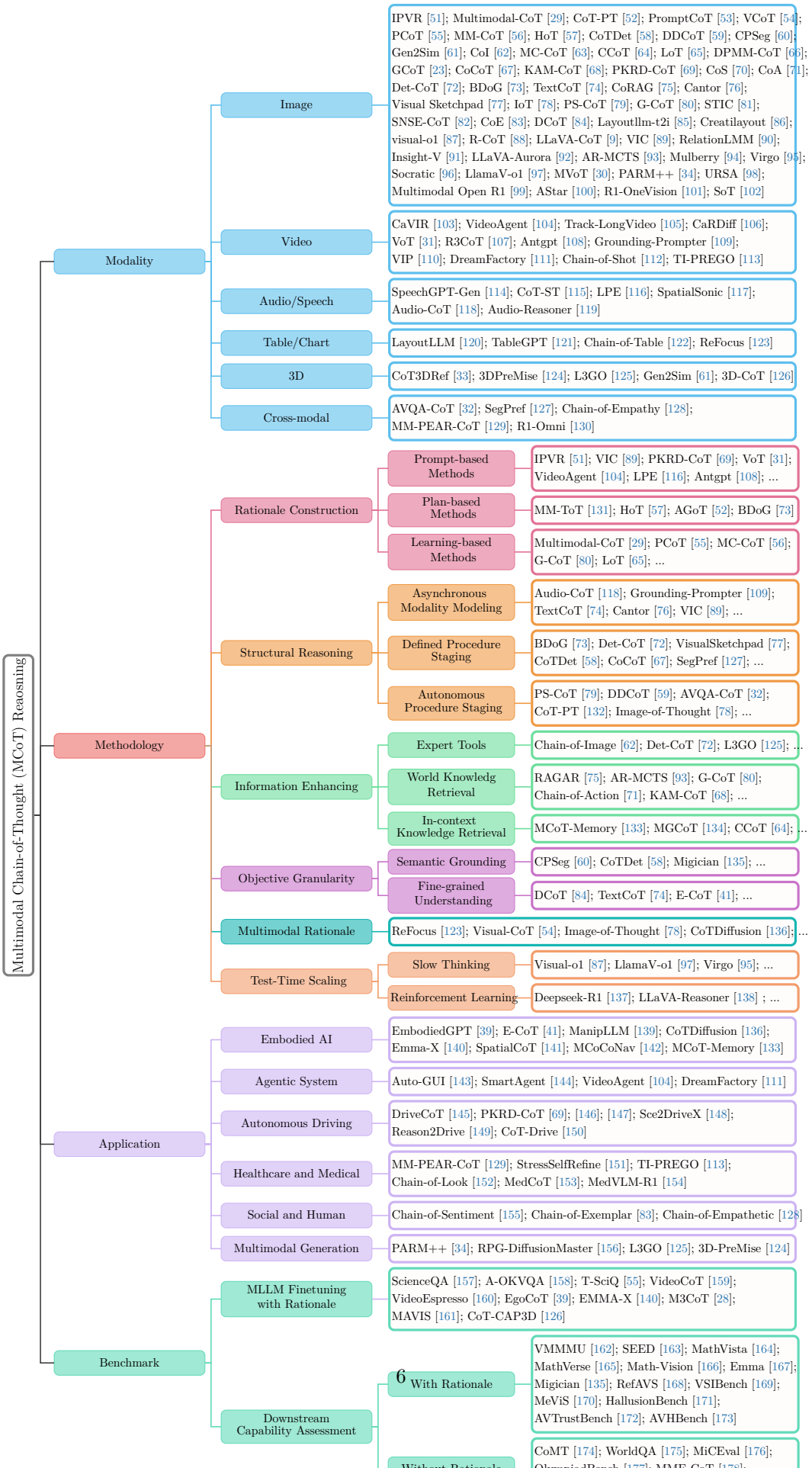
1.2 调查组织

本文的其余部分组织如下。我们首先介绍与 MCoT 相关的基础知识和背景知识（§ 2）。然后，我们回顾不同模态下 MCoT 的最新研究（§ 3）。接下来，我们从多个视角提供一个分类，并整合 MCoT 中的主流方法（§ 4）。随后，我们总结 MCoT 的广泛下游应用（§ 5）。接着，我们从多个角度概述数据集和基准（§ 6）。最后，我们讨论该领域的挑战和未来方向（§ 7）。

2 背景与初步研究

近年来，模型预训练规模的最新进展推动了语言模型应用范式的重大转变，从传统的“先预训练再微调”方法转移到更具适应性的“先预训练再提示”框架 [180–184]。在这一不断发展的背景下，研究者们探索了创新技术以增强 LLM 在复杂任务（尤其是 ICL[180, 185, 186] 和 CoT）中的推理能力。[1] 推理 [19]。ICL 的本质在于在提示中提供与任务相关的示例或演

[1]为了一致性和清晰性，我们用“CoT”表示多步推理技术，用“拓扑”描述不同的思维结构（例如，链式或图拓扑）。特定的方法被赋予唯一的标识符，如 vanilla CoT。


Multimodal Chain-of-Thought (MCoT) Reasoning
Modality
Image
IPVR [51]; Multimodal-CoT [29]; CoT-PT [52]; PromptCoT [53]; VCoT [54]; PCoT [55]; MM-CoT [56]; HoT [57]; CoTDet [58]; DDCoT [59]; CPSeg [60]; Gen2Sim [61]; CoI [62]; MC-CoT [63]; CCoT [64]; LoT [65]; DPMM-CoT [66]; GCoT [23]; CoCoT [67]; KAM-CoT [68]; PKRD-CoT [69]; CoS [70]; CoA [71]; Det-CoT [72]; BDoG [73]; TextCoT [74]; CoRAG [75]; Cantor [76]; Visual Sketchpad [77]; IoT [78]; PS-CoT [79]; G-CoT [80]; STIC [81]; SNSE-CoT [82]; CoE [83]; DCoT [84]; Layoutllm-t2i [85]; Creatilayout [86]; visual-o1 [87]; R-CoT [88]; LLaVA-CoT [9]; VIC [89]; RelationLMM [90]; Insight-V [91]; LLaVA-Aurora [92]; AR-MCTS [93]; Mulberry [94]; Virgo [95]; Socratic [96]; LlamaV-o1 [97]; MVoT [30]; PARM++ [34]; URSA [98]; Multimodal Open R1 [99]; AStar [100]; R1-OneVision [101]; SoT [102]
Video
CaVIR [103]; VideoAgent [104]; Track-LongVideo [105]; CaRDiff [106]; VoT [31]; R3CoT [107]; Antgpt [108]; Grounding-Prompter [109]; VIP [110]; DreamFactory [111]; Chain-of-Shot [112]; TI-PREGO [113]
Audio/Speech
SpeechGPT-Gen [114]; CoT-ST [115]; LPE [116]; SpatialSonic [117]; Audio-CoT [118]; Audio-Reasoner [119]
Table/Chart
LayoutLLM [120]; TableGPT [121]; Chain-of-Table [122]; ReFocus [123]
3D
CoT3DRef [33]; 3DPreMise [124]; L3GO [125]; Gen2Sim [61]; 3D-CoT [126]
Cross-modal
AVQA-CoT [32]; SegPref [127]; Chain-of-Empathy [128]; MM-PEAR-CoT [129]; R1-Omni [130]
Methodology
Rationale Construction
Prompt-based Methods
IPVR [51]; VIC [89]; PKRD-CoT [69]; VoT [31]; VideoAgent [104]; LPE [116]; Antgpt [108]; ...
Plan-based Methods
MM-ToT [131]; HoT [57]; AGoT [52]; BDoG [73]
Learning-based Methods
Multimodal-CoT [29]; PCoT [55]; MC-CoT [56]; G-CoT [80]; LoT [65]; ...
Structural Reasoning
Asynchronous Modality Modeling
Audio-CoT [118]; Grounding-Prompter [109]; TextCoT [74]; Cantor [76]; VIC [89]; ...
Defined Procedure Staging
BDoG [73]; Det-CoT [72]; VisualSketchpad [77]; CoTDet [58]; CoCoT [67]; SegPref [127]; ...
Autonomous Procedure Staging
PS-CoT [79]; DDCoT [59]; AVQA-CoT [32]; CoT-PT [132]; Image-of-Thought [78]; ...
Information Enhancing
Expert Tools
Chain-of-Image [62]; Det-CoT [72]; L3GO [125]; ...
World Knowledg Retrieval
RAGAR [75]; AR-MCTS [93]; G-CoT [80]; Chain-of-Action [71]; KAM-CoT [68]; ...
In-context Knowledge Retrieval
MCoT-Memory [133]; MGCoT [134]; CCoT [64]; ...
Objective Granularity
Semantic Grounding
CPSeg [60]; CoTDet [58]; Migician [135]; ...
Fine-grained Understanding
DCoT [84]; TextCoT [74]; E-CoT [41]; ...
Multimodal Rationale
ReFocus [123]; Visual-CoT [54]; Image-of-Thought [78]; CoTDiffusion [136]; ...
Test-Time Scaling
Slow Thinking
Visual-o1 [87]; LlamaV-o1 [97]; Virgo [95]; ...
Reinforcement Learning
Deepseek-R1 [137]; LLaVA-Reasoner [138] ; ...
Application
Embodied AI
EmbodiedGPT [39]; E-CoT [41]; ManipLLM [139]; CoTDiffusion [136]; Emma-X [140]; SpatialCoT [141]; MCoCoNav [142]; MCoT-Memory [133]
Agentic System
Auto-GUI [143]; SmartAgent [144]; VideoAgent [104]; DreamFactory [111]
Autonomous Driving
DriveCoT [145]; PKRD-CoT [69]; [146]; [147]; Sce2DriveX [148]; Reason2Drive [149]; CoT-Drive [150]
Healthcare and Medical
MM-PEAR-CoT [129]; StressSelfRefine [151]; TI-PREGO [113]; Chain-of-Look [152]; MedCoT [153]; MedVLM-R1 [154]
Social and Human
Chain-of-Sentiment [155]; Chain-of-Exemplar [83]; Chain-of-Empathetic [128]
Multimodal Generation
PARM++ [34]; RPG-DiffusionMaster [156]; L3GO [125]; 3D-PreMise [124]
Benchmark
MLLM Finetuning with Rationale
ScienceQA [157]; A-OKVQA [158]; T-SciQ [55]; VideoCoT [159]; VideoEspresso [160]; EgoCoT [39]; EMMA-X [140]; M3CoT [28]; MAVIS [161]; CoT-CAP3D [126]
Downstream Capability Assessment
With Rationale
VMMMU [162]; SEED [163]; MathVista [164]; MathVerse [165]; Math-Vision [166]; Emma [167]; Migician [135]; RefAVS [168]; VSIBench [169]; MeViS [170]; HallusionBench [171]; AVTrustBench [172]; AVHBench [173]
Without Rationale
CoMT [174]; WorldQA [175]; MiCEval [176];

Terms	Abbrev.	Description
In-context Learning	ICL	Prompting LLMs with task-specific examples without additional explicit training.
Chain-of-Thought	CoT	Prompting LLMs to reason step-by-step or breaks complex problems into logical steps.
Multimodal CoT	MCoT	Extends CoT to reason with multimodalities, e.g., audio, image.
Cross-modal CoT		Reasoning with two or more multimodalities, e.g., audio-visual.
Thought		A single reasoning step in CoT.
Rationale		Built upon multiple thoughts to support the final answer.

表 1: MCoT 相关术语的解释。

示，使大语言模型能够更好地理解用户意图并生成符合预期的输出。该方法利用上下文引导将模型导向任务适当的响应。相比之下，CoT 推理通过将复杂任务分解为一系列可管理的子任务来模拟人类解决问题的过程，并系统地构建解决方案。中间的推理步骤或轨迹，称为理由，阐明了模型结论背后的逻辑进展。在此基础上，MCoT 推理通过整合多种数据模态（如图像、视频和音频）扩展了 CoT 范式。这种增强扩大了多步推理的范围，提高了其在日益复杂的场景中的适用性。

2.1 从 CoT 到 MCoT

我们提供了解释 Table 1 中与 MCoT 相关的术语。为了形式化 MCoT 框架，我们首先定义 $\mathcal{P}$、$\mathcal{S}$、$\mathcal{Q}$、$\mathcal{A}$ 和 $\mathcal{R}$ 分别表示提示、指令、查询、答案和理由。这些元素中的每一个都表示为语言 Token 的序列，长度用 $|\cdot|$ 表示。我们还使用小写字母表示单个 Token，例如，a_i 表示答案 $\mathcal{A}$ 的第 i 个 Token。接下来，我们定义一个标准的 ICL 过程，该过程集成了 few-shot 示例对，可以表示如下：

$$\mathcal{P}_{ICL} = \{\mathcal{S}, (x_1, y_1), \ldots, (x_n, y_n)\}, \tag{1}$$

其中 $\mathcal{P}_{ICL}$ 表示用于 ICL 的提示，由指令 $\mathcal{S}$ 以及 n 组成的问题 x 和其对应答案 y 的演示对组成。然后，给定提示 $\mathcal{P}_{ICL}$ 和查询 $\mathcal{Q}$ 时生成答案序列 $\mathcal{A}$ 的概率在数学上定义为：

$$p(\mathcal{A} \mid \mathcal{P}_{ICL}, \mathcal{Q}) = \prod_{i=1}^{|\mathcal{A}|} \mathcal{F}(a_i \mid \mathcal{P}_{ICL}, \mathcal{Q}, a_{<i}), \tag{2}$$

其中 $\mathcal{F}$ 表示概率语言模型。注意，当 $n = 0$ 时，该过程简化为标准的 zero-shot 提示场景。

然后，我们可以将 vanilla CoT 定义为：

$$\mathcal{P}_{CoT} = \{\mathcal{S}, (x_1, e_1, y_1), \ldots, (x_n, e_n, y_n)\}, \tag{3}$$

其中 $\mathcal{P}_{CoT}$ 表示用于 CoT 推理的提示，e_i 表示示例推理。接下来，我们定义在给定输入提示 $\mathcal{P}_{CoT}$ 和查询 $\mathcal{Q}$ 的情况下生成答案 $\mathcal{A}$ 和理由 $\mathcal{R}$ 的联合概率：

$$p(\mathcal{A}, \mathcal{R} \mid \mathcal{P}_{CoT}, \mathcal{Q}) = p(\mathcal{A} \mid \mathcal{P}_{CoT}, \mathcal{Q}, \mathcal{R}) \cdot p(\mathcal{R} \mid \mathcal{P}_{CoT}, \mathcal{Q}), \tag{4}$$

其中右侧表示生成答案 $\mathcal{A}$ 和理由 $\mathcal{R}$ 的两个条件概率，可以定义为：

$$p(\mathcal{R} \mid \mathcal{P}_{CoT}, \mathcal{Q}) = \prod_{i=1}^{|\mathcal{R}|} \mathcal{F}(r_i \mid \mathcal{P}_{CoT}, \mathcal{Q}, r_{<i}), \tag{5}$$

$$p(\mathcal{A} \mid \mathcal{P}_{CoT}, \mathcal{Q}, \mathcal{R}) = \prod_{i=1}^{|\mathcal{A}|} \mathcal{F}(a_i \mid \mathcal{P}_{CoT}, \mathcal{Q}, a_{<i}). \tag{6}$$

与 ICL 方法相反，如公式(2)所示，CoT 框架要求在得出答案 $\mathcal{A}$ 之前生成一个推理 $\mathcal{R}$，这体现在公式(5)和(6)中。

当考虑 MCoT 时，至关重要的是指出，与 CoT 不同，MCoT 将多模态信息引入组件 $\mathcal{P}$、$\mathcal{Q}$、$\mathcal{A}$ 和 $\mathcal{R}$ 中。然而，并非所有这些组件都必须同时包含多模态信息。也就是说，给定基于语言的输入 $\mathcal{T}$ 和排除语言的多模态上下文 $\mathcal{M}$，我们有 $\exists \vartheta \in \{\mathcal{P}, \mathcal{Q}, \mathcal{A}, \mathcal{R}\} : \mathcal{M}(\vartheta)$。因此，我们根据推理的组成将 MCoT 分为两种不同的场景：一种完全依赖于基于语言的信息，另一种结合了超出语言内容的多模态信号。

► Scenario-1: MCoT with text-only thought to tackle multimodal input and output:

$$\mathcal{R} \in \mathcal{L}\,.$$

► Scenario-2: MCoT with multimodal thought to tackle unimodal or multimodal scenes:

$$\mathcal{R} \in \{\mathcal{M}, \mathcal{M} \oplus \mathcal{L}\}\,.$$

场景 1 旨在解决涉及输入或输出中多模态信息的任务，同时利用仅由语言组成的推理依据。相比之下，场景 2 强调在推理依据中整合给定的、检索到的或生成的多模态信息。

2.2 思维范式

自从引入 vanilla CoT [19] 后，各种范式出现以增强多模态和多步推理。基于推理过程中思想生成的构建，社区将推理结构 [187] 或拓扑 [188] 分类为链式、树状和图三种类型，如 Figure 3 所示。在这些拓扑中，思想被表示为结点，边表示它们之间的依赖关系 [188]。链式拓扑 [19, 189, 190] 促进线性和顺序的思想生成，逐步收敛到最终答案。然而，链式拓扑在推理过程中缺乏对个体思想进行深入探索的能力。

相比之下，树形拓扑结构 [191, 192] 在推理过程中启用了探索和回溯。在树形拓扑的每个结点（即，想法）上，一个想法生成器会生成多个子结点，如 Figure 3.C 左侧所示。然后这些子结点由状态评估器进行评估，并为它们分配得分。这些得分可以来自 LLM 本身或基于特定规则。搜索算法（例如广度优先搜索 (BFS) 或深度优先搜索 (DFS)）则指导树的扩展。

图拓扑结构 [23] 还允许从单一父结点生成多个子结点。然而，它们引入了循环和 N-to-1 连接，这意味着一个结点可以有多个父结点。这促进了多个结点之间的聚合，如 Figure 3中的蓝色箭头所示。超图拓扑结构 [57] 通过使用超边扩展了图拓扑结构，超边连接了两个以上的想法。这种结构通过整合来自不同模态的信息，天然支持联合推理。此外，自一致性 [193] 可以无缝集成到各种推理方法中。例如，以链式拓扑作为基准 (Figure 3.A)，可以并行执行多个基于链的推理过程，最终答案由绝对多数投票决定，以确保多个理由之间的一致性。总体而言，推理拓扑的演化反映了从线性依赖到分支探索、聚合与精炼以及高阶关联的进展。

2.3 多模态大型语言模型

如 GPT-4V [181]、Gemini 2.0 [7] 和 Claude3 [194] 等模型的发布，在多模态理解方面展示了非凡的能力，引发了研究社区对 MLLMs 的极大兴趣。对 MLLMs 的初步探索集中在

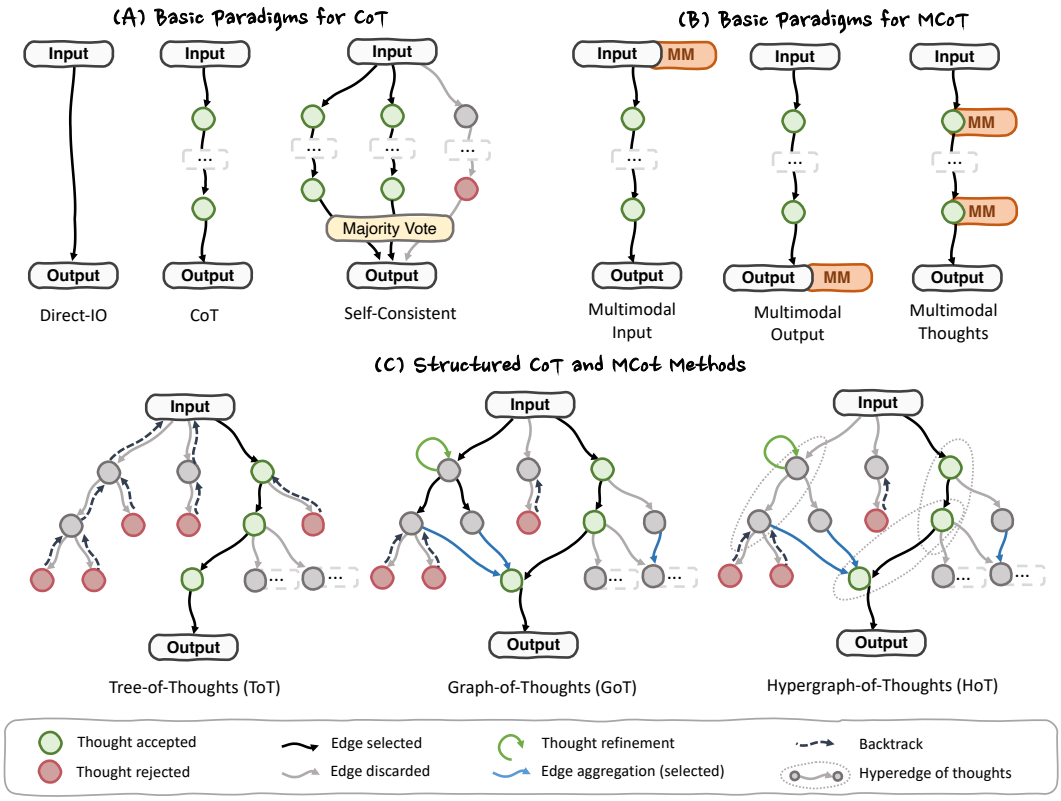


图 3: CoT 和 MCoT 的不同思维范式。

开发能够解释多模态内容并生成文本响应的强大语言模型上。在图像-文本理解领域，通过 BLIP2 [195]、OpenFlamingo [196]、MiniGPT-4 [197] 和 LLaVA [13] 等视觉大型语言模型（VLLMs）取得了显著进展。同时，视频-文本理解领域的进步也显现出来，VideoChat [198] 和 Video-ChatGPT [17] 做出了重要贡献。音频和语音理解也引起了关注，例如 Qwen-Audio [199, 16] 和 LLaSM [200] 等模型。一个值得注意的发展是 VideoLLaMA [18]，它利用 Qformer [195] 实现对音频和视频的理解。简而言之，主流的 MLLMs 通常遵循一致的模型架构，通过将多模态嵌入或词元处理到解码器结构中，并以自回归的方式生成上下文相关的输出，如 Figure 4左侧所示。

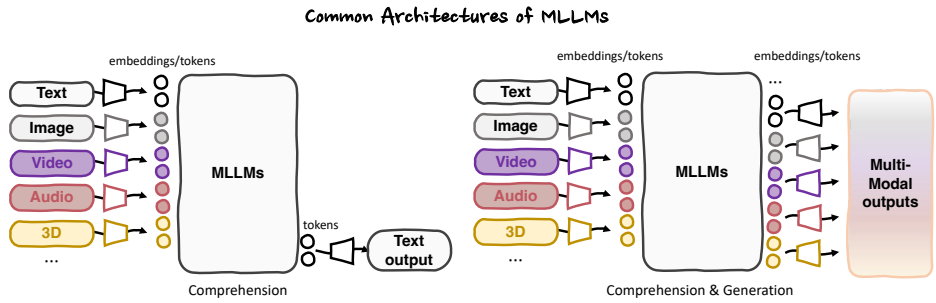


图 4: 仅理解型和理解生成型 MLLMs 的常见架构。

与这些关于多模态理解的工作平行，研究者还探索了多模态内容生成。在图像生成方面，Kosmos-2、GILL、Emu 和 MiniGPT-5 等模型取得了突破。音频生成领域通过 SpeechGPT

和 AudioPaLM 取得了进展，而视频生成研究，包括 CogVideo、VideoPoet、Video-Lavit 和 StreamingT2V，则为多模态内容创作奠定了基础。最近引入的 GPT-4o，能够理解和生成图像和音频，将注意力转向了“任意到任意”范式的模型。先前的工作，如 NExT-GPT 首次通过整合多模态适配器与各种扩散模型来实现这一目标。AnyGPT 利用多模态离散词元促进多样化多模态内容的生成。随后，Mini-Omni2 引入了一种基于命令的中断机制，增强了用户交互，并进一步与 GPT-4o 的能力保持一致。与仅支持理解的多模态大型语言模型相比，如 Figure 4 所示，集成理解与生成的多模态大型语言模型要么采用自回归方法生成多模态词元 [213]，要么连接不同模态的解码器以解码多模态嵌入 [14]。

最近，专注于推理的 OpenAI o1 [216] 模型的发布引起了通过有意识的、长时间的处理和测试时缩放来增强推理能力的兴趣。像 Mulberry [94]、AStar [100] 和 LlamaV-o1 [97] 这样的模型，通过采用长 MCoT 推理策略，在多模态推理方面表现出稳健的性能，进一步推动了多模态理解领域的发展。

3 多模态推理中的 MCoT（思维链）reasoning

MCoT 通过采用链式思维推理，将 LLMs/MLLMs 的推理能力扩展到处理跨多样模态的复杂任务，e.g.，图像、视频、音频、3D、表格/图表等。如 Figure 5所示，将 MCoT 与这些模态相结合，促进了众多基础且重要的应用的实现。本节系统回顾了这些模态下的 MCoT 研究，重点介绍关键进展及其对多模态推理发展的贡献。

3.1 图像的思维链推理

图像数据及其相关任务的普及推动了 MCoT 在视觉问答（VQA）中的广泛应用。早期实现，如 IPVR[51] 和多模态-CoT[29]，通过在最终预测之前生成中间理由来建立基础的 MCoT 框架。后续进展进一步完善了这一范式：MC-CoT[56] 将自我一致性 [193] 与 MCoT 相结合，在训练中采用词级绝对多数投票以提高生成理由的质量。SoT[102] 利用路由模型动态选择推理范例（即概念链、分块符号主义和专家词典），这些推理范例受到人类认知策略的启发，以提高推理效率。CoCoT[67] 通过输入之间的相似性和差异性分析改进了 MLLMs 中的多图像理解，而 RelationLMM[90] 则通过任务分解显式地解决对象关系建模问题。HoT[57] 通过引入超边连接多个推理节点来扩展思想图框架，从而增强多模态推理能力。

结构化的推理机制被提出以增强可控性和可解释性。DDCoT [59] 和 Socratic Questioning [96] 使用分阶段的推理过程来系统地优化多模态结果。文本和视觉模态之间的交互方法也严重影响了理由生成。Chain-of-Spot [70]、TextCoT [74] 和 DCoT [84] 优先进行感兴趣区域的分析以提高上下文理解。RAGAR [75] 和 Cantor [76] 将自动化过程与低级图像属性相结合以加强推理，而 KAM-CoT [68] 和 PKRD-CoT [69] 结合了外部知识库，并通过在 [134] 和 [73] 中描述的基于图的技术进一步增强。MCoT 对标注的推理数据的依赖推动了对自动化数据增强的研究。G-CoT [80]、STIC [81]、PS-CoT [79]、SNSE-CoT [82]、Chain-of-Exemplar [83] 和 R-CoT [88] 通过创新方法来自动化和增强训练数据生成，从而解决了这一限制。此外，静态图像特征提取在处理复杂的推理需求时常常导致不一致。为了解决这个问题，DPMM-CoT [66] 和 LLavA-AURORA [92] 从潜在空间重新生成图像特征。除了基于文本的理由外，最近的方法利用多模态理由进行全面推理，例如，Visual-CoT [54]、Chain-of-Image [62]、VisualSketchpad [77]、MVoT [30] 和 Visualization-of-Thought [217] 有效地处理多模态场景和多模态思想。

MCoT 的适用性也扩展到了 VQA 之外的专门领域。对于细粒度实例级任务，CoTDet [58]、Det-Cot [72] 和 CPSeg [60] 展示了显著的进步。在图像生成方面，PromptCoT [53] 专注于

图 5: 多模态和任务中 MCoT 应用的示例。

优化输入提示，PARM++ [34] 优化了奖励机制，而 LayoutLLM-T2I [85] 和 CreatiLayout [86] 在合成之前使用基于文本的布局构建先验，极大地提高了输出质量。

3.2 视频推理

视频理解同样依赖于基本的推理能力，因为在处理静态视觉内容和空间关系之外，视频还带来了时间动态性的挑战，特别是在长视频的情况下。作为一种基本应用，CaVIR [103] 通过实现零样本 MCoT 方法来增强需要上下文和常识理解的意图问答。类似地，VideoAgent [104] 和 HM-Prompt [105] 使用零样本 MCoT 来改善长视频推理并减少幻觉。AntGPT [108] 将少量样本 MCoT 扩展到以自我为中心的视频中的动作分类。在生成任务中，DreamFactory [111] 使用少量样本 MCoT 为长视频合成生成一致的关键帧。

对于复杂的视频理解，Video-of-Thought [31] 提出了一种全面的五阶段框架：任务和目标识别、对象跟踪、动作分析、排名问题回答和答案验证。这种结构化方法确保了对视频内容的透彻解释。同样地，CaRDiff [106] 将复杂的视频任务分解为子组件——标题生成、显著性推理和边界框生成——以引导扩散过程进行显著物体掩码创建。R3CoT [107] 引入了一个三阶段模型（即细化、检索、推理），专门用于视频谣言检测，而 Grounding-Prompter [109] 集成了全局和局部感知进行时间句子定位，基于语言查询对视频时刻进行定位。长视频分析效率由诸如 VIP [110] 等框架解决，这些框架优先选择关键帧并提取关键特征（例如焦点、动作、

情感、物体、背景）以通过中间和未来帧的属性预测来评估推理。Chain-of-Shot [112] 通过在训练期间使用二进制视频摘要进一步优化帧采样，并评估帧-任务相关性以实现高效推理。

MCoT 的效用同样张成于专门的领域，例如医学视频分析 [113, 151] 和情感计算 [218, 219, 130]。总体而言，这些进展强调了 MCoT 在分解复杂视频任务、提高推理准确率以及在各种应用中提升计算效率中的作用，标志着长视频理解领域的一个重要进步。

3.3 三维 MCoT 推理

由于需要整合包含形状、空间关系和物理属性等复杂高维数据，三维场景中的推理面临重大挑战。传统方法依赖于手动标注和僵化规则，促使采用 MCoT 将复杂的任务分解为可管理且结构化的流程。

几个框架展示了 MCoT 在三维生成方面的有效性。3D-PreMise [124] 使用 MCoT 引导 LLMs 生成三维形状和编程参数，简化对象合成。同样，L3GO [125] 引入了三维思维链，通过在仿真环境中进行迭代试错和工具调用实现三维图像生成，提高了适应性和准确性。Gen2Sim [61] 利用 MCoT 通过生成三维资产作为 MCoT 的输入来推进机器人技能学习，随后提示 LLMs 生成任务描述和奖励函数。这种方法减少了人为干预，同时促进了仿真中可扩展且多样化的任务获取。

当遇到语言指令时，CoT3DRef [33] 将复杂的 3D 定位分解为可解释的步骤。同时，3D-CoT [126] 通过将 MCoT 与包含结构推理标注的数据集结合，提高了 3D 视觉-语言对齐性能，涵盖形状识别、功能推理和因果推理等方面。这些进展共同突显了 MCoT 在高效解决复杂且组合式的 3D 任务中的关键作用。

3.4 MCoT 音频与语音推理

MCoT 已被有效扩展到逐步且可管理的语音和音频处理中，缩小了波形信号与语言语义之间的差距。CoT-ST [115] 将语音翻译分解为离散的语音识别和随后的翻译阶段。Xie et al. [116] 在共情对话生成之前整合了自动语音识别和情感检测的先验知识。Audio-CoT [118] 将原始的 CoT 融入音频理解和推理任务中。此外，Audio-Reasoner [119] 通过集成四步结构化推理框架（例如，规划、描述、推理、总结）实现了首个长 MCoT 推理。对于生成任务，SpatialSonic [117] 使用原始 MCoT 推导相关属性和描述，支持空间音频生成的创建。SpeechGPT-Gen [114] 进一步引入了信息生成链的方法，该方法在顺序步骤中系统地建模语义和感知信息，以促进自然语音生成。这些发展突显了 MCoT 在增强语音和音频处理方面的适应性和有效性，促进了更自然和上下文响应的结果。

3.5 表格和图表中的推理

LLMs 在文档理解方面表现出色，但由于表格和图表等结构化数据的复杂布局和隐式模式，面临挑战。LayoutLLM [120] 通过在文档、区域和段落级别上集成布局感知的预训练，并使用 vanilla MCoT 来增强处理能力，从而提升文档理解。同样，Dai et al. [220]结合场景图来解释图表，利用 vanilla MCoT 减少 LLMs 响应中的幻觉。尽管这些努力提供了粗粒度分析，但近期方法通过将任务分解为顺序可执行的操作来解决这些限制。TableGPT [121] 引入了命令链方法，利用命令集（例如，SelectCondition 和 GroupBy）系统地处理表格问题。Wang et al. [122]提出链表方法，使 LLMs 能够动态生成必要的操作和参数，重构表格以保留相关信息。相比之下，ReFocus [123] 通过编辑操作（如在表格中添加高亮或屏蔽区域）模拟人类注意力，产生视觉思维，从而提高理解能力。这些进展共同展示了 MCoT 在处理结构化数据复杂性方面的有效性。

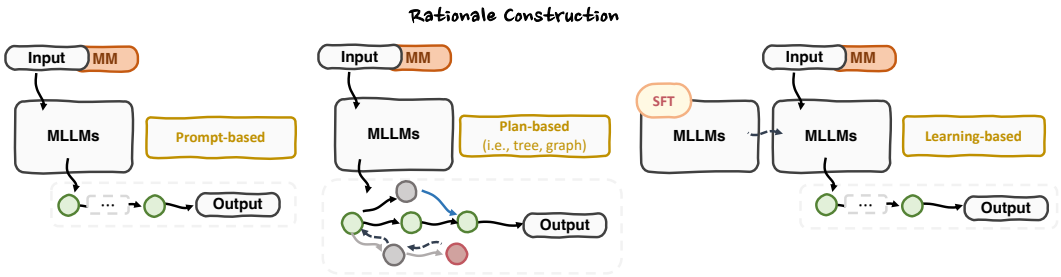


图 6: 在不同**推理构造**视角下的 MCoT 推理方法。

3.6 跨模态 CoT 推理

CoT 类似的推理在整合多种模态（而不仅仅是文本和单一附加模态）方面也表现出色，即跨模态 CoT 推理。AVQA-CoT [32] 将复杂的查询分解为更简单的子问题，通过大语言模型（LLMs）和预训练模型依次解决音频-视觉问答（AVQA）。类似地，SegPref [127] 使用视觉语言大模型（VLLMs）检测视觉场景中的潜在发声对象，随后结合文本理由与掩码解码器进行音频-视觉分割（AVS），从而减少对视觉特征的过度依赖。同时，Chain-of-Empathy [128] 借助纯 MCoT 与心理治疗原则相结合，提升大语言模型（LLMs）对人类情感的推理能力，促进富有同理心的响应。同样，MM-PEAR-CoT [129] 将纯 MCoT 应用于分析语言情感，并将其与音频和视频输入集成以改进多模态情感识别并减轻幻觉现象。R1-Omni [130] 在情感识别背景下首次将具有可验证奖励的强化学习（RLVR）应用于全模态大语言模型（Omni-multimodal LLM）。尽管跨模态 CoT 推理主要依赖于基于文本的理由，但这些 MCoT 的进展在各种下游任务中展示了卓越的性能。

4 MCoT 推理的方法论

为了全面研究 MCoT 在多模态环境中的稳健推理能力，研究社区围绕 MCoT 开发了多种方法和策略。为进行系统且全面的分析，我们从多个角度对这些方法进行了分类：理由构建、结构推理、信息增强、目标粒度、多模态理由以及测试时缩放。

4.1 从理性构建的角度来看

这部分总结了构建 MCoT 推理依据所采用的方法论。与传统直接输入-输出方法不同，后者优先考虑最终答案，而 CoT 和 MCoT 强调通过推理过程得出正确答案。因此，MCoT 推理方法主要关注依据的构建，可分为三种不同的类型：基于提示的方法、基于计划的方法和基于学习的方法，如 Figure 6所示。

基于提示的方法。 基于提示的 MCoT 推理采用精心设计的提示，包括指令或上下文演示，以引导模型在推理过程中生成理由，通常在 zero-shot 或 few-shot 设置中使用。例如，最简单的指令是“逐步思考以理解给定的文本和图像输入”，作为 zero-shot 提示 [110]，以引出解决多模态问题的理由。然而，大多数 MCoT 方法会指定明确的步骤，以确保推理遵循特定指导 [31, 51, 62, 69, 89, 104, 116]。此外，通常会整合专家工具以深入洞察详细信息 [72, 76]，或将多模态数据融入到语言推理中 [62, 77, 217]，特别是在图像和视频理解中。在 few-shot 场景中，提示可能包含明确的推理示例，以进一步引导推理过程 [108, 109, 113]。这种方法展示了显著的灵活性，在计算资源受限或需要快速响应的情景下具有优势。

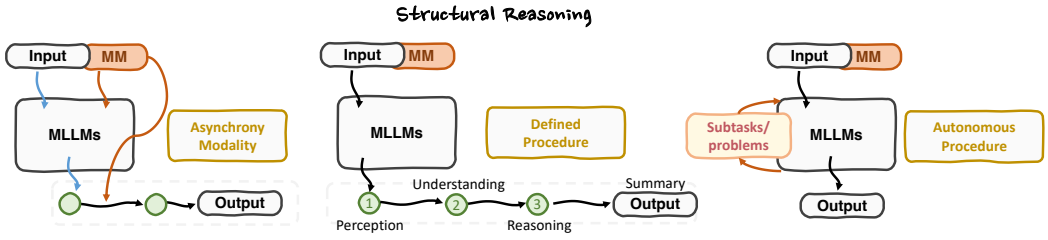


图 7: 在不同**结构推理**视角下的 MCoT 方法。

基于计划的方法。 基于计划的 MCoT 推理使模型能够在推理过程中动态地探索和优化思路。MM-ToT [131] 利用 GPT-4 [181] 和 Stable Diffusion [221] 生成多模态输出，通过 DFS 和 BFS 根据 0.0–1.0 的度量尺度选择最优输出。HoT [57] 从多模态输入中产生相互连接的思路，并封装在一个单一的超边内。相比之下，聚合思维图（AGoT）[52] 构建了一个推理聚合图，在每个推理步骤中整合多个推理方面，然后结合视觉数据。蓝图辩论图（BDoG）[73] 采用了一种独特的方法，放弃了搜索算法，转而使用三个智能体——肯定辩论者、否定辩论者和主持人。通过迭代辩论，这些智能体解决多模态问题，主持人综合出最终答案，隐式形成一个探索和聚合多样化思路的思维图。PARM++ [34] 使用链式验证步骤训练图像生成模型，例如潜在评估，以在图像生成过程中过滤掉不良输出。总之，与基于提示的方法不同，后者具有线性、示例驱动的推理，基于计划的 MCoT 变体使模型能够遍历多个推理路径，增强适应性和解决问题的深度。

基于学习的方法。 基于学习的 MCoT 推理将理由构建嵌入到训练或微调过程中，要求模型在多模态输入的同时显式地学习推理技能。Multimodal-CoT [29] 通过用包含理由的推理数据微调模型，开创了这种方法，培养了内在的推理能力。PCoT [55] 优化了这一范式用于理由生成，而 MC-CoT [56] 在训练期间结合多模态一致性与多数投票以增强小型模型的推理能力。G-CoT [80] 使用 ChatGPT 生成推理数据，通过微调激活可转移至自动驾驶的推理潜力。LoT [65] 通过使用跳跃思维数据进行微调来提升创造力，而 PromptCoT [53] 通过针对性微调增强图像合成的提示生成。总之，基于学习的方法侧重于在训练过程中嵌入推理模式。然而，在 2024 年末 OpenAI o1 [216] 发布后，人们对利用规模化的测试时计算 [91, 137, 222, 223] 来增强长 CoT 推理的兴趣激增，我们将在 Section 4.6 中进一步讨论这一点。

4.2 从结构推理的角度来看

除了基于下一个词元预测的简单化理由生成之外，最近的研究提出了结构化推理框架，以增强理由生成过程的可控性和可解释性。Figure 7展示了分为三种类型的结构化格式：异步模态建模、定义过程阶段化和自主过程阶段化。

异步模态建模。 MCoT 早期的研究探索直接从多模态上下文中生成论据，例如 Multimodal-CoT [29]、Audio-CoT [118] 和 Grounding-Prompter [109]。然而，Wu et al. [224]的神经科学研究表明，识别和推理在不同的认知模块中运作，遵循“先描述后决策”的策略。这一见解激发了异步模态处理方法。例如，IPVR [51] 引入了一个三阶段的“看、想、确认”框架用于 VQA，将感知与推理分离。Visualization-of-Thought [217] 通过生成基于 2D 网格的文本表示来模拟心智图像，以指导搜索和导航任务。同样，TextCoT [74] 采用两阶段过程：首先总结图像上下文，然后生成基于视觉输入的响应。Cantor [76] 将

感知和决策阶段分离，其中感知阶段从图像或文本描述中提取低级属性（例如对象、颜色、形状），而决策阶段整合这些特征以准确解决问题。相比之下，VIC [89] 在整合视觉输入之前将任务分解为基于文本的子步骤以得出最终论据。这些方法通过隔离感知编码与高级推理，从而增强了可解释性和与人类认知过程的对齐。

定义的步骤划分。 多项研究明确定义了结构化推理阶段以增强过程的可解释性。BDoG [73] 使用带有专门智能体的固定辩论-总结流水线，而 Det-CoT [72] 将视觉问答（VQA）推理形式化为模板指令解析、子任务分解、执行和验证。VisualSketchpad [77] 将理由结构化为“思考、动作、观察”阶段，而 CoTDet [58] 通过对象列表、功能分析和视觉特征总结实现目标检测。苏格拉底提问法 [96] 将 VQA 分解为自我引导的子问题生成、详细描述和总结。Grounding-Prompter [109] 在最终决策之前进行全局理解、噪声评估和分区理解。对于多图像理解，CoCoT [67] 系统地比较输入之间的相似性和差异性。LLaVA-CoT [225] 通过摘要、描述、分析和结论阶段实现长时多步骤推理（long-MCoT），Audio-Reasoner [119] 也以相同方式实现。

结构化分阶段方法也促进了数据集构建和下游应用开发。URSA [98] 通过推理蒸馏和轨迹重写生成数学推理数据集。与教育和情感计算领域的 Chain-of-Sentiment [155]、Chain-of-Exemplar [83] 和 Chain-of-Empathetic [128] 并行。SmartAgent [144] 通过图形用户界面导航、推理和推荐阶段构建个人助手。CoT-ST [115] 结合了语音识别和机器翻译用于语音翻译。SegPref [127] 利用全局理解、发声目标过滤和噪声信息移除，鲁棒地定位视觉空间中的发声对象。在生成任务中，PARM [34] 生成具有明确判断、潜在评估和最佳-N 选择的图像，而 SpeechGPT-Gen [114] 从感知角度到语义角度合成语音。

自主程序分阶段。 近期研究探索了自主程序分阶段方法，使 LLM 能够自行确定推理步骤的顺序。PS-CoT [79] 允许 LLM 在生成理由之前自主生成解决问题的计划，而 DDCoT [59] 和 AVQA-CoT [32] 将问题分解为子问题以进行迭代解决。CoT-PT [132] 采用从抽象到具体概念的分层推理（例如，物体 → 动物 → 狗）。Image-of-Thought [78] 自动将 VQA 任务分割为具有相应图像操作动作的子任务。Insight-V [91] 动态确定每个推理步骤的重点，并自主决定是继续还是总结中间结果。Chain-of-Table [122] 生成逐步查询以修改表结构（例如，添加“国家”标题），合成操作参数，并优化数据存储以高效推导答案。在具身智能任务中，E-CoT [41] 和 Emma-X [140] 使 LLM 能够推断可执行的子任务序列。

4.3 从信息增强的角度来看

增强多模态输入通过集成专家工具和内部或外部知识，促进了全面推理。

使用专家工具。 近期研究利用专业工具通过结构化的视觉或几何操作来增强多模态推理能力。对于数学和几何任务，方法如 Chain-of-Image [62] 和 VisualSketchpad [77] 通过专家工具或代码生成辅助可视化。同样，Det-CoT [72]、Cantor [76] 和 Image-of-Thought [78] 使用图像操作工具（例如，放大、标尺标记）以改进细粒度的视觉分析。同时，L3GO [125] 和 3D-Premise [124] 集成三维生成工具以支持空间推理工作流。这些方法强调了在多模态推理任务中整合领域特定工具包的重要性，以提升可解释性和精度。

使用世界知识检索。 近期研究通过整合外部知识源来增强推理过程。像 RAGAR [75]、AR-MCTS [93] 和 Chain-of-Action [71] 等方法利用检索增强生成（RAG）在推理过程中引入领域特定或常识知识。G-CoT [80] 从 ChatGPT 中提炼与任务相关的常识信息，而 CoTDet [58] 检索对象功能以提供检测任务的上下文。KAM-CoT [68] 联合推理图像、文本数据和结

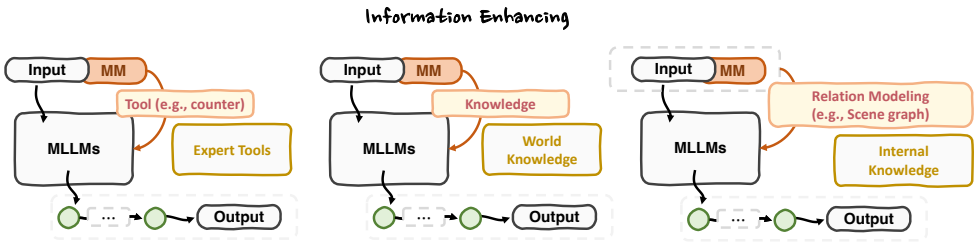


图 8: MCoT 推理在具有 **信息增强**的视角下。

构化知识图谱，以提升多模态理解能力。这些方法展示了知识感知架构在连接感知输入与概念理解中的关键作用。

利用上下文知识检索。 除了外部知识增强外，多项研究通过从输入内容或 LLMs/MLLMs 自身生成的论证中检索和组织信息来改进推理。DCoT [84] 专注于在推理过程中优先处理感兴趣的图像区域。相比之下，MCoT-Memory [133]、MGCoT [134]、Video-of-Thought [31]、CCoT [64] 和 BDoG [73] 通过场景图表示建模对象或概念之间的关系，隐式地检索上下文知识。类似地，CoT3DRef [33] 在定位参考句子时生成目标锚点，有效地充当简化的场景图。这些方法共同展示了结构化上下文知识提取在提高推理逼真度方面的有效性。

4.4 从目标粒度的角度来看

研究方法通常与目标的粒度保持一致，如 Figure 9所示。虽然大多数问答任务强调概览信息，即粗略理解，但一些细粒度任务（例如语义接地）更重视个体实例，即语义接地和细粒度理解。

粗理解层次。 作为探索最广泛的信息处理层次，粗理解在 VQA 和 AQA 等任务中被广泛使用，例如 Multimodal-CoT [29] 和 Audio-CoT [118] 等方法就体现了这一点。这些方法旨在对给定的多模态信息有一个大致了解，而不关注细节。

语义接地级别。 通过专门的推理范式解决语义接地任务。CPSeg [60]、CoTDet [58] 和 Migician [135] 利用大型语言模型（LLMs）细化接地参考，以增强文本提示与目标视觉实例之间的对齐，从而提高下游掩码解码器或边界框提议器的精度。类似地，SegPref [127] 使用视觉大语言模型（VLLMs）从全局场景信息中推断潜在的声音对象，然后结合音频信息定位视觉空间中的声音对象。

细粒度理解层次。 细粒度理解还需要在多模态背景下捕获详细信息。DCoT [84]、TextCoT [74] 和 Spot 链 [70] 首先关注图像中的感兴趣区域，然后从这些检索到的区域中识别出细粒度信息。E-CoT [41] 检索目标对象的边界框并确定机器人任务中的抓取位置，促进具身交互。

4.5 从多模态理性视角

如引言部分 §2所述，推理过程可以采用纯文本或多种模态的论证。主要的关注点集中在以文本为中心的方法上，例如 Multimodal-CoT [29]、PCoT [55]、MC-CoT [63]、LLaVA-CoT [225] 和 Grounding-Prompter [109]。这些方法主要使用文本表示来对多模态信息进行编码，从而与 LLMs 或 MLLMs 的推理机制无缝集成。

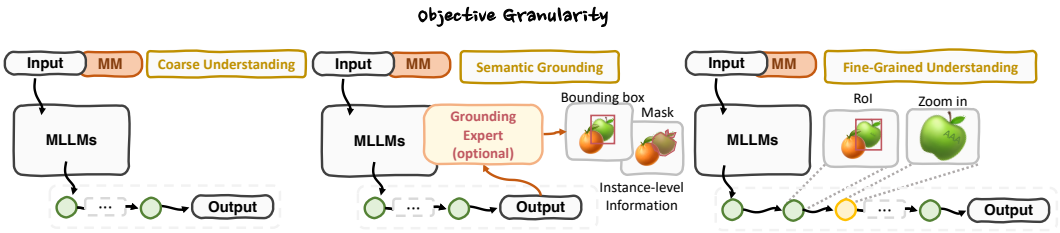


图 9: 在各种**目标粒度**的视角下进行 MCoT 推理。

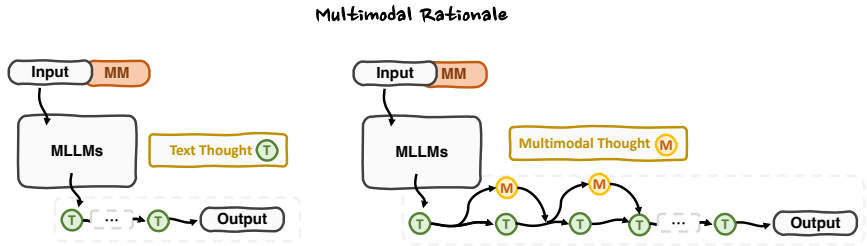


图 10: MCoT 推理与 **多模态推理**。

新兴方法却探索了受人类认知过程启发的多模态推理构造。例如，ReFocus [123] 通过在表格数据上叠加高亮区域来模拟视觉注意力，而 Visual-CoT [54] 通过生成中间虚拟状态来解决顺序图像推理中的逻辑缺口。Chain-of-Image [62] 通过为数学或几何问题求解生成辅助图表来模拟人类工具辅助推理，而 Image-of-Thought [78] 结合文本和视觉检索以动态定位对象，这些对象应被引用以回答细粒度问题。Visualization-of-Thought [217] 构建二维网格以表示人类解决空间推理任务时的心理图像，而 MVoT [30] 进一步可视化每个推理步骤。同样，CoTDiffusion [136] 利用扩散模型将机器人操作任务分解为连贯的视觉子目标计划，从而弥合抽象推理与物理执行之间的差距。这种从文本为中心到多模态推理的进展反映了对模仿类人认知机制的日益重视。

4.6 **从测试时缩放** perspective

LLMs 的推理响应可以分为直接响应和 CoT 响应，类似于人类认知中存在的两种不同的推理系统 [226]。系统 1 以快速、启发式驱动的决策为特征，与系统 2 形成对比，系统 2 采用深思熟虑的逻辑推理，产生更准确且较少有偏的结果 [227]。Snell et al. [222]进一步证实了“慢思考” 在 LLMs 中的有效性，基于系统 2，在推理过程中最优地扩展测试时间计算可能比扩展模型参数更高效。OpenAI 的 o1[216] 的发布进一步激发了对大规模推理模型的兴趣，这些模型结合了内部和外部的慢思考机制 [228–231]，为解决复杂的挑战提供了潜在解决方案，特别是在数学和编码等领域，而 Deepseek-R1[137] 展示了单独使用强化学习（RL）可以唤醒长期 CoT 推理能力。

基于慢思考的模型。 内部慢思考通过训练或微调增强推理的深度和质量。相反，外部慢思考在推理过程中通过迭代采样和优化解决方案来改善推理。作为开创性工作，Qwen-QwQ [232] 经过带有 7,000 个长 CoT 样本的监督微调（SFT），解锁了长 CoT 推理能力。Macro-o1 [233] 也通过 CoT 微调集成了内部慢思考，但进一步采用了启发式搜索算法蒙特卡洛树搜索（MCTS）来激活外部慢思考能力。

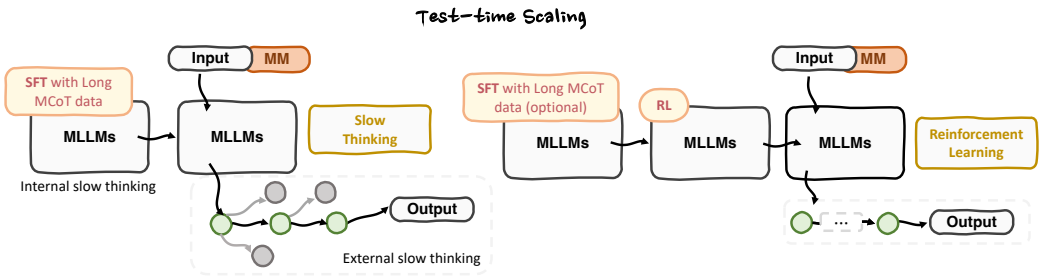


图 11: MCoT 推理结合**测试时缩放**策略。RL 可以在没有标注的长链推理训练数据的情况下，帮助提升推理质量或主动长链推理能力。SFT 是可选的。

在此基础上，研究已扩展到大型多模态推理模型。Visual-o1 使用多模态和多轮 CoT 框架解决模糊指令问题，而 LLaVA-CoT [225] 和 Audio-Reasoner [119] 通过 SFT 和结构化推理实现长-MCoT 推理。LlamaV-o1 [97] 通过课程学习方法激活长-MCoT 推理能力。Virgo [95] 通过微调一个带有紧凑长文本数据集的 MLLM 构建一个多模态慢思考系统。AR-MCTS [93] 在 MCTS 扩展过程中动态检索多模态洞察，以丰富采样的多样性和可靠性。AStar [100] 通过 MCTS 从仅 500 个数据样本中提炼推理模式来引导推理，而 Mulberry [94] 利用多个 MLLM 的集体学习增强树搜索，利用负路径生成反思数据以提高自我反思能力。RedStar [234] 表明几千个样本足以激活长-CoT 能力，并且其有效性随着模型大小的增加而扩展，可以提升 MLLM 的泛化能力。这些发展突显了慢思考范式在推进多模态推理能力方面的变革潜力。

基于强化学习的模型。 强化学习在提升大型语言模型（LLMs）的推理能力方面展示了显著的效果。Deepseek-R1 通过仅使用强化学习激活长链思维（long-CoT）推理，展示了这一点，并且通过监督微调（SFT）冷启动和迭代自我改进进一步超越了 OpenAI 的 o1，在某些方面表现出色，激发了对 Open-R1 和 TinyZero 等强化学习驱动模型的兴趣。在多模态背景下，早期的工作如 LLaVA-Reasoner 和 Insight-V 通过使用长链多步骤思维（long-MCoT）数据微调以及由人类反馈引导的直接偏好优化（DPO），改进了推理能力。在此基础上，Multimodal-Open-R1 集成了 GPRO 框架来开发类似于 R1 的多模态大型语言模型（MLLM），而 R1-V 验证了强化学习可以增强视觉推理任务中的泛化能力。RL 在视觉推理中的有效性也得到了其他同期工作的验证，例如 R1-OneVision、VLM-R1、LMM-R1 和 Easy-R1。此外，与上述结果奖励模型（ORM）相比，进步奖励模型（PRM）如 MSTaR 和 VisualPRM 在每个推理步骤中评估并提供反馈，进一步增强了 MLLMs 的自我一致性与自我进化能力。

基于 RL 的推理范式进一步扩展到视频推理任务中（Open-R1-Video [243]），检测任务中（Curr-ReFT [244]），分割任务中（Seg-Zero [245]）以及通过 R1-Omni 将音频纳入多模态情感识别任务中（[130]）。此外，RL 促进了“啊哈时刻”的出现，这使得在推理过程中能够进行反思和回溯，这一现象首先由 Deepseek-R1 在仅文本场景中确定。MM-Eureka [246] 和 VisualThinker-R1-Zero [247] 在视觉推理中成功重现了这一现象。Table 2 总结了 MLLMs 使用 RL 进行更好的长链推理的技术。总之，RL 解锁了复杂的推理和“啊哈时刻”，而无需 SFT，展示了其通过迭代自我改进和基于规则的方法增强模型能力的潜力，最终为更先进、更自主的多模态推理系统铺平了道路。

5 具有 MCoT 推理的应用程序

MCoT 将复杂任务分解为可管理子任务的强大能力促使其在不同领域中的应用，包括具身系统、智能体、自动驾驶、医疗创新以及多模态生成框架等。每个领域都展示了 MCoT 推

Model	Foundational LLMs	Modality	Learning	Cold Start	Algorithm	Aha-moment
Deepseek-R1-Zero [137]	Deepseek-V3	T	RL	✗	GRPO	✔
Deepseek-R1 [137]	Deepseek-V3	T	SFT+RL	✔	GRPO	-
LLaVA-Reasoner [138]	LLaMA3-LLaVA-NEXT-8B	T,I	SFT+RL	✔	DPO	-
Insight-V [91]	LLaMA3-LLaVA-NEXT-8B	T,I	SFT+RL	✔	DPO	-
Multimodal-Open-R1 [99]	Qwen2-VL-7B-Instruct	T,I	RL	✗	GRPO	✗
R1-OneVision [101]	Qwen2.5-VL-7B-Instruct	T,I	SFT	-	-	-
R1-V [237]	Qwen2.5-VL	T,I	RL	✗	GPRO	✗
VLM-R1 [238]	Qwen2.5-VL	T,I	RL	✗	GPRO	✗
LMM-R1 [239]	Qwen2.5-VL-Instruct-3B	T,I	RL	✗	PPO	✗
Curr-ReFT [244]	Qwen2.5-VL-3B	T,I	RL+SFT	✗	GPRO	-
Seg-Zero [245]	Qwen2.5-VL-3B + SAM2	T,I	RL	✗	GPRO	✗
MM-Eureka [246]	InternVL2.5-Instruct-8B	T,I	SFT+RL	✔	RLOO	-
MM-Eureka-Zero [246]	InternVL2.5-Pretrained-38B	T,I	RL	✗	RLOO	✔
VisualThinker-R1-Zero [247]	Qwen2-VL-2B	T,I	RL	✗	GPRO	✔
Easy-R1 [240]	Qwen2.5-VL	T,I	RL	✗	GRPO	-
Open-R1-Video [243]	Qwen2-VL-7B	T,I,V	RL	✗	GRPO	✗
R1-Omni [130]	HumanOmni-0.5B	T,I,V,A	SFT+RL	✔	GRPO	-
VisRL [248]	Qwen2.5-VL-7B	T,I	SFT+RL	✔	DPO	-
R1-VL [249]	Qwen2-VL-7B	T,I	RL	✗	StepGRPO	-

表 2: 使用强化学习的多模态推理模型。Deepseek-R1 用作仅文本的基础大型语言模型(LLM)进行对比。

理如何增强任务分解、决策制定和泛化能力，从而为其在解决现实世界 AI 挑战中的变革潜力提供了重要见解。

5.1 具身人工智能

近期具身 AI 的进步显著提升了机器人在规划、操作和导航方面的能力。EmbodiedGPT [39] 和 E-CoT [41] 利用 MCoT 推理将任务分割为可执行的子目标。值得注意的是，EmbodiedGPT 引入了 EgoCoT 数据集用于视觉-语言预训练，而 E-CoT 则专注于文本命令的顺序执行。ManipLLM [139] 通过针对对象为中心的任务微调 MLLMs 来增强操作能力，而 CoTDiffusion [136] 使用扩散生成的视觉子目标以实现长时程活动中的精确性。在空间推理方面，Emma-X [140] 集成了基于地面的规划和预测运动，而 SpatialCoT [141] 利用坐标对齐进行复杂的空间推理。在导航方面，MCoCoNav [142] 通过全局语义地图和基于得分的合作优化多机器人协调，而 MCoT-Memory [133] 通过结合记忆检索和场景图更新来改善长时程规划，保留高置信度的经验以支持稳健决策。总体而言，这些研究强调了一种趋势，即整合多模态数据和链式推理，以构建适应性强且具有通用性的具身系统。

5.2 自主系统

AI 驱动的智能体系统的发展扩展了自主交互和内容生成能力。Auto-GUI [143] 使用多模态动作链(MCoA)直接操作图形界面，提高了效率而不依赖外部工具或 API。同样，SmartAgent [144] 将 GUI 导航与用户思维链 (CoUT) 推理相结合，为具身智能体提供个性化推荐。在视频理解方面，VideoAgent [104] 利用 LLMs 结合反射性的三步决策过程，对长篇内容进行准

确解释。补充这些研究，DreamFactory [111] 通过多智能体框架开创了长视频生成技术，通过关键帧迭代和 MCoT 推理确保场景一致性。这些研究共同展示了链条机制和智能体协作在解决复杂现实世界 AI 挑战中的关键作用。

此外，最近人工智能智能体系统中出现了一个重要的范式转变，将 “感知-推理” 与 “规划-执行” 相结合。Manus [250] 的出现体现了这一转变，激发了对像 OpenManus [251] 这样的工具使用型智能体的兴趣。利用大语言模型进行自然语言理解和生成，Manus 通过目标导向的自我反思迭代优化解决方案。作为一个工具使用型智能体，它集成了多种功能，如网络搜索、数据查询和代码执行，采用工具链方法来解决现实世界中的复杂多模态任务。未来工具使用型智能体的发展预计将建立在基础模型之上，增强长程 MCoT 推理能力，并整合各种多模态接口和工具。这种发展路径表明了向着具有越来越接近人类能力的 AI 智能体发展的趋势。

5.3 自动驾驶

近期自动驾驶领域的进展越来越多地利用了 MLLMs 和 MCoT 推理来提高决策能力和适应性。DriveCoT [145] 将 MCoT 集成到端到端驾驶系统中，并通过定制的数据集提供支持，而 PKRD-CoT [69] 利用 zero-shot MCoT 提示处理动态环境中的感知、知识、推理和决策。Ma et al. [147] 和 Cui et al. [146] 强调人类交互，他们结合 LLMs 有效解释反馈和口头指令。Sce2DriveX [148] 通过多模态场景理解增强端到端控制，并展示出强大的泛化能力。此外，Reason2Drive [149] 提供超过 600K 的视频-文本对，以探索可解释的推理，通过对象级感知增强 LLMs 以强化规划能力。总体而言，这些努力表明自动驾驶系统正朝着类人推理、增强交互性和改进泛化的方向转变。

5.4 医学和医疗保健

创新的 AI 在医疗领域的应用通过链式推理增强了各种医学任务。StressSelfRefine [151] 通过受心理学启发的 “描述、评估、突出” 过程检测视频中的压力，并通过 DPO 优化以提高准确率。TI-PREGO [113] 将 ICL 与自动思维链（ACoT）相结合，识别自视角视频中的程序错误，利用动作序列和逻辑推理。Chain-of-Look [152] 通过将任务分解为视频推理阶段并使用视觉语言提示来解决内窥镜视频中的手术三元组识别问题。同时，MedCoT [153] 通过分层专家系统改进了医学视觉问答，最终实现专家混合诊断。此外，MedVLM-R1 [154] 使用仅 600 个医学 VQA 样本的 RL 方法，旨在增强视觉语言模型的医学推理能力。总体而言，这些努力展示了 MCoT 推理在提高多种医学 AI 应用的可解释性和精度方面的有效性。

5.5 社会与人文

MCoT 已被有效扩展到人文和社会科学领域，利用其任务分解能力。例如，Chain-of-Empathetic [128] 使用 MCoT 来促进共情对话生成。MM-PEAR-CoT 框架 [129] 通过结构化的预备问题答案推理方法增强了多模态情感分析，在后期多模态融合之前生成文本解释。在情感计算领域，Chain-of-Sentiment [155] 在会话背景下优化了情感分析，同时并发研究做出了补充贡献 [218, 219]。此外，Chain-of-Exemplar [83] 将 MCoT 扩展到教育领域，X-Reflect [252] 将 MCoT 应用于多模态推荐系统，而 Yu and Luo [253] 使用零样本 MCoT 进行人口统计推断。这些进展突显了 MCoT 在以人类为中心和社科领域的复杂挑战中所具有的潜力。

5.6 多模态生成

近期，AI 驱动的图像和 3D 生成领域取得了显著进展，展示了多种多样的多模态合成策略。PARM 和 PARM++ [34] 使用逐步迭代推理方法，并辅以清晰的潜在评估和反思机制，生成高质量图像。GoT [254] 通过引入清晰的语言推理过程构建生成链式思维，该过程在生成和编辑图像之前分析语义关系和空间布局。RPG-DiffusionMaster [156] 利用 MLLMs 进行文本到图像的扩散生成，将提示分解为详细的子区域以实现连贯输出，而 L3GO [125] 借助语言智能体采用 3D 思维链方法创建非传统的 3D 对象，在分布外描述上优于传统扩散模型。此外，3D-PreMise [124] 将 LLMs 与程序合成结合，生成参数化 3D 形状，当由明确的推理示例指导时，可产生有前景的工业成果。这些研究共同强调了增强推理的 AI 克服数据驱动生成局限性的潜力，实现了精确且创新的多模态输出。

6 思维导图 (MCoT) 数据集和基准

MCoT 推理需要专门的数据集和基准来支持模型微调和性能评估。结合 Table 3，本节综述了与 MCoT 相关的资源全景，分为两个关键领域：设计用于带有推理理由微调 MLLM 的数据集，以及开发用于评估下游能力的基准，带或不带相应的理由。这些资源共同满足了在多个领域、模态和推理复杂度下训练和评估 MLLM 的多样化需求。

6.1 带有推理的 MLLMs 微调数据集

几个研究探索了通过特定数据集激活 MLLMs 的 MCoT 推理能力。ScienceQA [157] 提供了带有标注答案、课程和解释的多模态科学问题，展示了语言模型如何利用 MCoT 增强多跳推理。A-OKVQA [158] 通过提供丰富的常识和世界知识为 VQA 提供了重要数据。基于 ScienceQA，T-SciQ [55] 通过高级 LLMs 丰富了推理理由。VideoCoT [159] 提供了逐步视频问答（VideoQA）的推理数据，尽管其推理仅限于文本解释。相比之下，VideoEspresso [160] 提供了保持空间和时间连贯性以及多模态推理标注的 VideoQA 对。此外，MAVIS [161] 数据集通过自动生成数学视觉数据推动了 MLLMs 在数学领域的训练和应用，从而提供了丰富的对齐的视觉-语言对和推理理由。EgoCoT [39] 和 EMMA-X [140] 提出了一个以自我为中心的数据集，用于训练模型执行具身任务和子目标任务。M3CoT [28] 进一步推进了 VLLMs 在多领域和多跳推理样本上的推理能力。同时，MAmmoTH-VL-Instruct [255] 在 118 个数据集和 10 个类别中构建了 1200 万长 MCoT 推理数据，提升了 MLLMs 的长 MCoT 推理能力。此外，像 LLaVA-CoT-100k [225]、Mulberry-260k [94]、MM-Verify [256] 和 VisualPRM400K [242] 等数据集由相应的推理模型提出，以激活长 MCoT 推理能力。这些数据集通常与基于学习的推理构造方法相关联，如我们在 Section 4.1 中提到的。

6.2 下游能力评估基准

各种各样的基准已被开发出来以评估下游能力，特别是在常识和科学推理的领域。如 Table 4所示，我们展示了来自不同机构的 MLLMs 在四个基准上的性能比较：MMMU[162]、MathVista[164]、Math-Vision[166] 和 EMMA[167]。虽然 MMMU 和 EMMA 侧重于多学科场景，但 MathVista 和 Math-Vision 主要评估数学推理。

没有原理的数据集。 已经广泛采用多个多模态评估基准来评估 MLLMs 的表现。尽管这些基准没有提供原理，但它们的多样性和挑战表明，在 MCoT 的帮助下，MLLMs 可以在这些基准上进一步提升性能。MMMU[162] 涉及六个核心学科的视觉-语言问题，旨在衡量 LMMs 的三种基本能力：感知、知识和推理。SEED[163] 进一步引入视频模态，以评估 MLLMs 的理

解和生成能力。MathVista[164]、MathVerse[165] 和 Math-Vision[166] 特别关注数学领域的视觉感知和推理。Emma[167] 引入了无法通过各模态独立推理解决的推理挑战，为 MLLMs 的多模态推理能力提供了全面评估。

除了通用和数学领域外，MCoT 由于其逐步推理能力，在各种下游任务中展示了其有效性。Migician [135] 提供了对多图像定位的评估，并展示了 MCoT 在此类任务中的有效性。RefAVS [168] 将定位引入到视听上下文中，结合时间信息和空间信息，这些问题可以通过 MCoT 进一步解决。VSIBench [169] 提供了对 MLLMs 空间推理能力的评估。MeViS [170] 提供了对带有运动表达的视频分割的评估。特别地，通过 MCoT 的逐步推理，预计可以进一步解决 MLLMs 中出现的幻觉现象。HallusionBench [171] 评估了 VLLMs 中的幻觉现象，而 AVTrustBench [172] 和 AVHBench [173] 评估了视听环境中的幻觉现象，这些问题可以通过 MCoT 推理进一步缓解。OSWorld [257] 和 AgentClinic [258] 提供了用于评估智能体在多模态场景中能力的基准，这些能力可以通过推理增强。

带有理由的数据集。 随着 OpenAI o1 [216] 和 Deepseek-r1 [137] 的出现，对扩展测试时计算和慢思考的兴趣稳步增长，从而推动了评估基准的发展，旨在评估 MLLMs 在推理过程中生成的理由的质量。CoMT [174] 被引入以解决传统多模态基准的局限性，这些基准仅通过语言进行推理，通过要求多模态输入和输出，旨在更好地模拟人类式的推理并探索复杂的视觉操作。WorldQA [175] 挑战 MLLMs 使用语言、视觉和音频回答问题，同时结合长链推理和世界知识。MiCEval [176] 精心设计用于评估推理链的准确性，通过仔细评估描述部分和每个单独推理步骤的质量。OlympiadBench [177] 包含 8000 多个双语奥林匹克水平的数学和物理问题以及注释的理由，可以有效地评估 MLLMs 的高级能力。MME-CoT [178] 提供了系统性的 MCoT 推理评估，揭示了关键见解，包括反思机制如何提高推理质量以及 MCoT 提示如何可能对感知密集型任务产生负面影响，这可能是由于过度思考造成的。OmniBench [179] 是首个涉及文本、视觉和音频的综合性评估基准。

7 局限性、挑战与未来方向

尽管越来越多的注意力和研究努力集中在 MCoT 上，但仍有一些关键方面尚未得到解决和深入探索，这些问题可能是实现人类水平多模态通用人工智能的关键瓶颈。以下，我们总结了一些挑战，以阐明未来的研究方向。

计算可持续性与慢思考悖论。 尽管取得了显著进展，但过度依赖测试时的缩放和慢思考来支持长链多步推理仍带来了重大挑战。维持深度推理过程所需的计算资源和训练数据呈指数级增长，仍然是一个关键瓶颈，需要在算法效率（如强化学习）和硬件加速方面进行创新。

通识场景中缺乏推理。 现有的长链思维（long-MCoT）框架主要关注数学和科学领域可验证的数据，但在通识场景中的稳健推理能力不足。数学和科学任务因其严格的逻辑结构和独特的解决方案，在测试时使用 ORM[99, 246] 和 PRM[241, 242] 进行了大量关于缩放的研究。然而，在通识场景中，答案很少是固定的，通常包含多个可能的解释和推理。这种变异性使得基于数学和科学的推理框架效果不佳。推理能力的不足表现为当模型处理涉及复杂情况、分歧和多因素影响的一般上下文任务时，无法充分评估推理过程。未来的研究应探索开放式的奖励模型，以在多模态通识场景中实现稳健的长链推理。

扩展推理链中的误差传播。 当前 MCoT 系统中长链推理的主要担忧之一是误差雪球效应[230]，即早期步骤中的小误差可能会通过后续阶段放大，并导致灾难性结果。传统的置信度校准技术无法解决多模态误差传播的独特挑战，其中不同模态可能相互矛盾，同时保持较高

的自一致性得分。开发定量指标以诊断、量化和缓解这些累积误差是一个尚未解决的问题，需要进行严格的研究。

符号-神经集成差距。 虽然神经模型在模式识别方面表现出色，但其无法进行严格的符号操作限制了复杂推理的能力。混合神经符号架构 [277–279] 可以帮助改进数学证明，但仍可能受到知识接地问题的影响。根本的挑战在于开发分布式表示与离散符号系统之间的无缝接口，特别是在跨模态符号操作方面（例如，将几何图转换为形式化证明）。

动态环境适应与自适应链长。 大多数 MCoT 系统假设静态输入条件，严重限制了其在现实世界中的适用性。当处理流式多模态输入时，“冻结推理”悖论出现——大多数现有架构无法在不重新开始整个链条的情况下根据新证据修订早期结论。此外，开发能够根据计算约束动态调整链长或推理步骤数量的资源高效推理框架至关重要。基于实时评估和反馈机制的自适应策略将是平衡推理准确性与资源效率的关键。

幻觉预防。 现有 MCoT 推理中的一个关键挑战是减轻幻觉现象，其中模型生成看似合理但事实错误或不一致的输出。[89] 这一问题削弱了推理过程的可靠性，并在多模态情景中尤为突出，在这种情景中，整合多样化数据源可能导致上下文错配和虚假信息。未来的工作应着重于每一步推理中稳健的跨模态对齐方法和验证机制。诸如不确定性量化、对抗训练以及利用外部知识库等技术在减少幻觉方面显示出潜力。此外，人类错误检测的见解可能启发增强多模态推理系统准确性和可解释性的新策略。

数据选择、标注和增强策略。 近期研究表明，精心策划的数据集可以激活模型的长时-MCoT 推理能力 [225, 94, 280]。然而，自动选择和标注适合扩展推理的数据仍是一个开放的挑战。结合半监督或自监督学习策略以及强化学习方法，可以减少对大量人工标注的依赖。

模态不平衡与高维模态集成。 当前研究表明，不同模态的发展存在不均衡现象，例如文本和图像等模态相较于其他模态进展更快。未来的研究应着重解决更高维度模态（如 3D 数据、传感器信息等）的集成问题，以构建一个更加平衡且全面的多模态推理框架，充分利用每个模态的独特特性。

与认知科学的跨学科集成。 MCoT 推理的固有复杂性要求采用一种跨学科方法，该方法整合来自认知科学、心理学和神经科学的见解。借鉴人类决策和认知理论 [224, 151, 217] 可以启发新的推理架构，这些架构更接近人类思维过程，从而提高性能和可解释性。

具身推理的局限性。 大多数 MCoT 系统在与物理实体分离的抽象符号空间中运行。弥合这一仿真到现实的差距需要紧密集成本体感觉反馈、触觉理解以及动态世界建模——这些挑战是当前架构设计几乎未涉及的问题。

可解释推理与理论支持。 随着模型变得越来越复杂，决策过程的可解释性对于建立信任并实现实际部署至关重要。尽管 CoT 或 MCoT 推理提供了中间步骤，但其潜在的理论机制仍然很大程度上不透明，使模型成为一个“黑盒子”，仅仅产生预测和输出。开发提供透明、可追溯推理路径的方法不仅会增强模型的可解释性，还将促进 MCoT 推理系统的调试和进一步优化。因此，加强理论支持对于实现真正可解释的推理至关重要。

伦理、鲁棒性和安全性推理。 随着 MCoT 系统变得越来越强大，确保 AI 在对抗扰动下的安全性和鲁棒性是至关重要的。集成更强大的推理技术和方法可以提高系统的透明度，并为潜在故障提供安全保障。此外，随着这些系统接近实际部署，开发能够解析并跨模态应用伦

理约束的多模态宪法级 AI 框架变得至关重要。需要进一步研究来量化和减轻与对抗攻击和其他多模态推理环境中安全问题相关的风险。

8 结论

本文综述了多模态思维链（MCoT）推理的首次系统性回顾。我们首先提供定义并阐明基础概念，为方法论奠定基础。全面的分类法从不同角度对各种方法和推理范式进行分类，以应对图像、视频、语音、音频、3D 和结构化数据等模态特有的挑战。我们整合了全面的 MCoT 相关数据集和基准，以准确概述当前资源状况。此外，我们的综述探讨了相关应用，突出了在机器人技术、医疗保健、自动驾驶、社会科学和多模态生成等重要领域的成功案例。最后，我们概述了有前景的未来研究方向，旨在克服这些限制，并推动多模态通用人工智能（AGI）的发展。我们将所有与 MCoT 相关的资源和信息公开共享，以促进这一快速发展的领域的后续研究。

参考文献

[1] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. CoRR, abs/2407.10671, 2024.

[2] Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from GLM-130B to GLM-4 all tools. CoRR, abs/2406.12793, 2024.

[3] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor

Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. CoRR, abs/2307.09288, 2023.

[4] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai. CoRR, abs/2403.04652, 2024.

[5] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. Palm 2 technical report. CoRR, abs/2305.10403, 2023.

[6] Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone. CoRR, abs/2404.14219, 2024.

[7] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican,

et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.

[8] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: An LMM perceiving any aspect ratio and high-resolution images. In ECCV, pages 390–406, 2024.

[9] Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. CoRR, abs/2411.10440, 2024.

[10] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In NeurIPS, 2023.

[11] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. CoRR, abs/2409.12191, 2024.

[12] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In CVPR, pages 26753–26763, 2024.

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36:34892–34916, 2023.

[14] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-any multimodal llm. In International Conference on Machine Learning, pages 53366–53397, 2024.

[15] Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision LLM for understanding, generating, segmenting, editing. In Advances in neural information processing systems, 2024.

[16] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. arXiv preprint arXiv:2407.10759, 2024.

[17] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424, 2023.

[18] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858, 2023.

[19] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837, 2022.

[20] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. arXiv preprint arXiv:2210.03493, 2022.

[21] Xuezhi Wang and Denny Zhou. Chain-of-thought reasoning without prompting. In NeurIPS, 2024.

[22] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In NeurIPS, 2023.

[23] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In AAAI, pages 17682–17690, 2024.

[24] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In NeurIPS, 2023.

[25] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, and Ilge Akkaya. Openai o1 system card. CoRR, abs/2412.16720, 2024.

[26] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin

Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. CoRR, abs/2501.12948, 2025.

[27] Xiongtao Zhou, Jie He, Lanyu Chen, Jingyu Li, Haojing Chen, Víctor Gutiérrez-Basulto, Jeff Z. Pan, and Hanjie Chen. Miceval: Unveiling multimodal chain of thought's quality via image description and reasoning steps. CoRR, abs/2410.14668, 2024.

[28] Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M^3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In ACL, pages 8199–8221. Association for Computational Linguistics, 2024.

[29] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923, 2023.

[30] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. arXiv preprint arXiv:2501.07542, 2025.

[31] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In Forty-first International Conference on Machine Learning, 2024.

[32] Guangyao Li, Henghui Du, and Di Hu. Avqa-cot: When cot meets question answering in audio-visual scenarios. In CVPR Workshops, 2024.

[33] Eslam Abdelrahman, Mohamed Ayman, Mahmoud Ahmed, Habib Slim, and Mohamed Elhoseiny. Cot3dref: Chain-of-thoughts data-efficient 3d visual grounding. arXiv preprint arXiv:2310.06214, 2023.

[34] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. arXiv preprint arXiv:2501.13926, 2025.

[35] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. CoRR, abs/2402.12289, 2024.

[36] Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao Ma, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. Dilu: A knowledge-driven approach to autonomous driving with large language models. In ICLR, 2024.

[37] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and José M. Álvarez. Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. CoRR, abs/2405.01533, 2024.

[38] Yifan Bai, Dongming Wu, Yingfei Liu, Fan Jia, Weixin Mao, Ziheng Zhang, Yucheng Zhao, Jianbing Shen, Xing Wei, Tiancai Wang, and Xiangyu Zhang. Is a 3d-tokenized LLM the key to reliable autonomous driving? CoRR, abs/2405.18361, 2024.

[39] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. Advances in Neural Information Processing Systems, 36:25081–25094, 2023.

[40] Ming-Yi Lin, Ou-Wen Lee, and Chih-Ying Lu. Embodied AI with large language models: A survey and new HRI framework. In ICARM, pages 978–983, 2024.

[41] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. arXiv preprint arXiv:2407.08693, 2024.

[42] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In ICRA, pages 11523–11530, 2023.

[43] Xuan Xiao, Jiahang Liu, Zhipeng Wang, Yanmin Zhou, Yong Qi, Qian Cheng, Bin He, and Shuo Jiang. Robot learning in the era of foundation models: A survey. CoRR, abs/2311.14379, 2023.

[44] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian D. Reid, and Niko Sünderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. In CoRL, pages 23–72, 2023.

[45] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong T. Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: vision-language-action models transfer web knowledge to robotic control. In CoRL, pages 2165–2183, 2023.

[46] Yuting He, Fuxiang Huang, Xinrui Jiang, Yuxiang Nie, Minghao Wang, Jiguang Wang, and Hao Chen. Foundation model for advancing healthcare: Challenges, opportunities, and future directions. CoRR, abs/2404.03264, 2024.

[47] Sagar Goyal, Eti Rastogi, Sree Prasanna Rajagopal, Dong Yuan, Fen Zhao, Jai Chintagunta, Gautam Naik, and Jeff Ward. Healai: A healthcare LLM for effective medical documentation. In WSDM, pages 1167–1168, 2024.

[48] Zhiyao Ren, Yibing Zhan, Baosheng Yu, Liang Ding, and Dacheng Tao. Healthcare copilot: Eliciting the power of general llms for medical consultation. CoRR, abs/2402.13408, 2024.

[49] Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In ACM WWW, pages 2627–2638, 2024.

[50] Ziyu Wang, Hao Li, Di Huang, and Amir M. Rahmani. Healthq: Unveiling questioning capabilities of LLM chains in healthcare conversations. CoRR, abs/2409.19487, 2024.

[51] Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Hao Zhang, and Chuang Gan. See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning. arXiv preprint arXiv:2301.05226, 2023.

[52] Juncheng Yang, Zuchao Li, Shuai Xie, Wei Yu, Shijun Li, and Bo Du. Soft-prompting with graph-of-thought for multi-modal representation learning. arXiv preprint arXiv:2404.04538, 2024.

[53] Junyi Yao, Yijiang Liu, Zhen Dong, Mingfei Guo, Helan Hu, Kurt Keutzer, Li Du, Daquan Zhou, and Shanghang Zhang. Promptcot: Align prompt distribution via adapted chain-of-thought. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7027–7037, 2024.

[54] Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. Visual chain of thought: bridging logical gaps with multimodal infillings. arXiv preprint arXiv:2305.02317, 2023.

[55] Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 19162–19170, 2024.

[56] Cheng Tan, Jingxuan Wei, Zhangyang Gao, Linzhuang Sun, Siyuan Li, Ruifeng Guo, Bihui Yu, and Stan Z Li. Boosting the power of small multimodal reasoning models to match larger models with self-consistency training. In European Conference on Computer Vision, pages 305–322. Springer, 2024.

[57] Fanglong Yao, Changyuan Tian, Jintao Liu, Zequn Zhang, Qing Liu, Li Jin, Shuchao Li, Xiaoyu Li, and Xian Sun. Thinking like an expert: Multimodal hypergraph-of-thought (hot) reasoning to boost foundation modals. arXiv preprint arXiv:2308.06207, 2023.

[58] Jiajin Tang, Ge Zheng, Jingyi Yu, and Sibei Yang. Cotdet: Affordance knowledge prompting for task driven object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3068–3078, 2023.

[59] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. Advances in Neural Information Processing Systems, 36:5168–5191, 2023.

[60] Lei Li. Cpseg: Finer-grained image semantic segmentation via chain-of-thought language prompting. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 513–522, 2024.

[61] Pushkal Katara, Zhou Xian, and Katerina Fragkiadaki. Gen2sim: Scaling up robot learning in simulation with generative models. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 6672–6679. IEEE, 2024.

[62] Fanxu Meng, Haotong Yang, Yiding Wang, and Muhan Zhang. Chain of images for intuitively reasoning. arXiv preprint arXiv:2311.09241, 2023.

[63] Lai Wei, Wenkai Wang, Xiaoyu Shen, Yu Xie, Zhihao Fan, Xiaojin Zhang, Zhongyu Wei, and Wei Chen. Mc-cot: A modular collaborative cot framework for zero-shot medical-vqa with LLM and MLLM integration. CoRR, abs/2410.04521, 2024.

[64] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14420–14431, 2024.

[65] Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13246–13257, 2024.

[66] Liqi He, Zuchao Li, Xiantao Cai, and Ping Wang. Multi-modal latent space learning for chain-of-thought reasoning in language models. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 18180–18187, 2024.

[67] Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. arXiv preprint arXiv:2401.02582, 2024.

[68] Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 18798–18806, 2024.

[69] Xuewen Luo, Fan Ding, Yinsheng Song, Xiaofeng Zhang, and Junnyong Loo. Pkrd-cot: A unified chain-of-thought prompting for multi-modal large language models in autonomous driving. arXiv preprint arXiv:2412.02025, 2024.

[70] Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. Chain-of-spot: Interactive reasoning improves large vision-language models. arXiv preprint arXiv:2403.12966, 2024.

[71] Zhenyu Pan, Haozheng Luo, Manling Li, and Han Liu. Chain-of-action: Faithful and multimodal question answering through large language models. arXiv preprint arXiv:2403.17359, 2024.

[72] Yixuan Wu, Yizhou Wang, Shixiang Tang, Wenhao Wu, Tong He, Wanli Ouyang, Philip Torr, and Jian Wu. Dettoolchain: A new prompting paradigm to unleash detection ability of mllm. In European Conference on Computer Vision, pages 164–182. Springer, 2024.

[73] Changmeng Zheng, Dayong Liang, Wengyu Zhang, Xiao-Yong Wei, Tat-Seng Chua, and Qing Li. A picture is worth a graph: A blueprint debate paradigm for multimodal reasoning. In Proceedings of the 32nd ACM International Conference on Multimedia, pages 419–428, 2024.

[74] Bozhi Luan, Hao Feng, Hong Chen, Yonghui Wang, Wengang Zhou, and Houqiang Li. Textcot: Zoom in for enhanced multimodal text-rich image understanding. arXiv preprint arXiv:2404.09797, 2024.

[75] M Abdul Khaliq, P Chang, M Ma, Bernhard Pflugfelder, and F Miletić. Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models. arXiv preprint arXiv:2404.12065, 2024.

[76] Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, et al. Cantor: Inspiring multimodal chain-of-thought of mllm. In Proceedings of the 32nd ACM International Conference on Multimedia, pages 9096–9105, 2024.

[77] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. arXiv preprint arXiv:2406.09403, 2024.

[78] Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. arXiv preprint arXiv:2405.13872, 2024.

[79] Qun Li, Haixin Sun, Fu Xiao, Yiming Wang, Xinping Gao, and Bir Bhanu. Ps-cot-adapter: adapting plan-and-solve chain-of-thought for scienceqa. Science China Information Sciences, 68(1):119101, 2025.

[80] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. In European Conference on Computer Vision, pages 403–420. Springer, 2024.

[81] Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Zou, Kai-Wei Chang, and Wei Wang. Enhancing large vision language models with self-training on image comprehension. arXiv preprint arXiv:2405.19716, 2024.

[82] Guangmin Zheng, Jin Wang, Xiaobing Zhou, and Xuejie Zhang. Enhancing semantics in multimodal chain of thought via soft negative sampling. arXiv preprint arXiv:2405.09848, 2024.

[83] Haohao Luo, Yang Deng, Ying Shen, See-Kiong Ng, and Tat-Seng Chua. Chain-of-exemplar: enhancing distractor generation for multimodal educational question generation. In ACL, 2024.

[84] Zixi Jia, Jiqiang Liu, Hexiao Li, Qinghua Liu, and Hongbin Gao. Dcot: Dual chain-of-thought prompting for large multimodal models. In The 16th Asian Conference on Machine Learning (Conference Track), 2024.

[85] Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In Proceedings of the 31st ACM International Conference on Multimedia, pages 643–654, 2023.

[86] Hui Zhang, Dexiang Hong, Tingwei Gao, Yitong Wang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. arXiv preprint arXiv:2412.03859, 2024.

[87] Minheng Ni, Yutao Fan, Lei Zhang, and Wangmeng Zuo. Visual-o1: Understanding ambiguous instructions via multi-modal multi-turn chain-of-thoughts reasoning. arXiv preprint arXiv:2410.03321, 2024.

[88] Linger Deng, Yuliang Liu, Bohan Li, Dongliang Luo, Liang Wu, Chengquan Zhang, Pengyuan Lyu, Ziyang Zhang, Gang Zhang, Errui Ding, et al. R-cot: Reverse chain-of-thought problem generation for geometric reasoning in large multimodal models. arXiv preprint arXiv:2410.17885, 2024.

[89] Haojie Zheng, Tianyang Xu, Hanchi Sun, Shu Pu, Ruoxi Chen, and Lichao Sun. Thinking before looking: Improving multimodal llm reasoning via mitigating visual hallucination. arXiv preprint arXiv:2411.12591, 2024.

[90] Chi Xie, Shuang Liang, Jie Li, Zhao Zhang, Feng Zhu, Rui Zhao, and Yichen Wei. Relationlmm: Large multimodal model as open and versatile visual relationship generalist. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.

[91] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. arXiv preprint arXiv:2411.14432, 2024.

[92] Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. arXiv preprint arXiv:2412.03548, 2024.

[93] Guanting Dong, Chenghao Zhang, Mengjie Deng, Yutao Zhu, Zhicheng Dou, and Ji-Rong Wen. Progressive multimodal reasoning via active retrieval. arXiv preprint arXiv:2412.14835, 2024.

[94] Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. arXiv preprint arXiv:2412.18319, 2024.

[95] Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Virgo: A preliminary exploration on reproducing o1-like mllm. arXiv preprint arXiv:2501.01904, 2025.

[96] Wanpeng Hu, Haodi Liu, Lin Chen, Feng Zhou, Changming Xiao, Qi Yang, and Changshui Zhang. Socratic questioning: Learn to self-guide multimodal reasoning in the wild. arXiv preprint arXiv:2501.02964, 2025.

[97] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. arXiv preprint arXiv:2501.06186, 2025.

[98] Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. arXiv preprint arXiv:2501.04686, 2025.

[99] EvolvingLMMs Lab. Multimodal open r1, 2025. URL https://github.com/EvolvingLMMs-Lab/open-r1-multimodal. Accessed: 2025-02-28.

[100] Jinyang Wu, Mingkuan Feng, Shuai Zhang, Ruihan Jin, Feihu Che, Zengqi Wen, and Jianhua Tao. Boosting multimodal reasoning with mcts-automated structured thinking. arXiv preprint arXiv:2502.02339, 2025.

[101] R1-onevision: open-source multimodal large language model with reasoning ability, 2025. URL https://yangyi-vai.notion.site/r1-onevision#198b1e4047f780c78306fb451be7160d.

[102] Simon A. Aytes, Jinheon Baek, and Sung Ju Hwang. Sketch-of-thought: Efficient llm reasoning with adaptive cognitive-inspired sketching, 2025.

[103] Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. Intentqa: Context-aware video intent reasoning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 11963–11974, 2023.

[104] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In European Conference on Computer Vision, pages 58–76. Springer, 2024.

[105] Yiwei Sun, Zhihang Liu, Chuanbin Liu, Bowei Pu, Zhihan Zhang, and Hongtao Xie. Hallucination mitigation prompts long-term video understanding. arXiv preprint arXiv:2406.11333, 2024.

[106] Yunlong Tang, Gen Zhan, Li Yang, Yiting Liao, and Chenliang Xu. Cardiff: Video salient object ranking chain of thought reasoning for saliency prediction with diffusion. arXiv preprint arXiv:2408.12009, 2024.

[107] Rongpei Hong, Jian Lang, Jin Xu, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Following clues, approaching the truth: Explainable micro-video rumor detection via chain-of-thought reasoning. In THE WEB CONFERENCE 2025, 2025.

[108] Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Antgpt: Can large language models help long-term action anticipation from videos? arXiv preprint arXiv:2307.16368, 2023.

[109] Houlun Chen, Xin Wang, Hong Chen, Zihan Song, Jia Jia, and Wenwu Zhu. Grounding-prompter: Prompting llm with multimodal information for temporal sentence grounding in long videos. arXiv preprint arXiv:2312.17117, 2023.

[110] Vaishnavi Himakunthala, Andy Ouyang, Daniel Rose, Ryan He, Alex Mei, Yujie Lu, Chinmay Sonar, Michael Saxon, and William Yang Wang. Let's think frame by frame with vip: A video infilling and prediction dataset for evaluating video chain-of-thought. arXiv preprint arXiv:2305.13903, 2023.

[111] Zhifei Xie, Daniel Tang, Dingwei Tan, Jacques Klein, Tegawend F Bissyand, and Saad Ezzini. Dreamfactory: Pioneering multi-scene long video generation with a multi-agent framework. arXiv preprint arXiv:2408.11788, 2024.

[112] Jian Hu, Zixu Cheng, Chenyang Si, Wei Li, and Shaogang Gong. Cos: Chain-of-shot prompting for long video understanding. arXiv preprint arXiv:2502.06428, 2025.

[113] Leonardo Plini, Luca Scofano, Edoardo De Matteis, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Andrea Sanchietti, Giovanni Maria Farinella, Fabio Galasso, and Antonino Furnari. Ti-prego: Chain of thought and in-context learning for online mistake detection in procedural egocentric videos. arXiv preprint arXiv:2411.02570, 2024.

[114] Dong Zhang, Xin Zhang, Jun Zhan, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechgpt-gen: Scaling chain-of-information speech generation. arXiv preprint arXiv:2401.13527, 2024.

[115] Yexing Du, Ziyang Ma, Yifan Yang, Keqi Deng, Xie Chen, Bo Yang, Yang Xiang, Ming Liu, and Bing Qin. Cot-st: Enhancing llm-based speech translation with multimodal chain-of-thought. arXiv preprint arXiv:2409.19510, 2024.

[116] Jingran Xie, Shun Lei, Yue Yu, Yang Xiang, Hui Wang, Xixin Wu, and Zhiyong Wu. Leveraging chain of thought towards empathetic spoken dialogue without corresponding question-answering data. arXiv preprint arXiv:2501.10937, 2025.

[117] Peiwen Sun, Sitong Cheng, Xiangtai Li, Zhen Ye, Huadai Liu, Honggang Zhang, Wei Xue, and Yike Guo. Both ears wide open: Towards language-driven spatial audio generation. arXiv preprint arXiv:2410.10676, 2024.

[118] Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. arXiv preprint arXiv:2501.07246, 2025.

[119] Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models. arXiv preprint arXiv:2503.02318, 2025.

[120] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15630–15640, 2024.

[121] Liangyu Zha, Junlin Zhou, Liyao Li, Rui Wang, Qingyi Huang, Saisai Yang, Jing Yuan, Changbao Su, Xiang Li, Aofeng Su, et al. Tablegpt: Towards unifying tables, nature language and commands into one gpt. arXiv preprint arXiv:2307.08674, 2023.

[122] Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, et al. Chain-of-table: Evolving tables in the reasoning chain for table understanding. arXiv preprint arXiv:2401.04398, 2024.

[123] Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. arXiv preprint arXiv:2501.05452, 2025.

[124] Zeqing Yuan, Haoxuan Lan, Qiang Zou, and Junbo Zhao. 3d-premise: Can large language models generate 3d shapes with sharp features and parametric control? arXiv preprint arXiv:2401.06437, 2024.

[125] Yutaro Yamada, Khyathi Chandu, Yuchen Lin, Jack Hessel, Ilker Yildirim, and Yejin Choi. L3go: Language agents with chain-of-3d-thoughts for generating unconventional objects. arXiv preprint arXiv:2402.09052, 2024.

[126] Yanjun Chen, Yirong Sun, Xinghao Chen, Jian Wang, Xiaoyu Shen, Wenjie Li, and Wei Zhang. Integrating chain-of-thought for multimodal alignment: A study on 3d vision-language learning, 2025.

[127] Yaoting Wang, Peiwen Sun, Yuanchao Li, Honggang Zhang, and Di Hu. Can textual semantics mitigate sounding object segmentation preference? In European Conference on Computer Vision, pages 340–356. Springer, 2024.

[128] Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark. arXiv preprint arXiv:2502.04976, 2025.

[129] Yan Li, Xiangyuan Lan, Haifeng Chen, Ke Lu, and Dongmei Jiang. Multimodal pear chain-of-thought reasoning for multimodal sentiment analysis. ACM Transactions on Multimedia Computing, Communications and Applications, 2024.

[130] Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning, 2025.

[131] Kye Gomez. Multimodal-tot. https://github.com/kyegomez/MultiModal-ToT, 2023.

[132] Jiaxin Ge, Hongyin Luo, Siyuan Qian, Yulu Gan, Jie Fu, and Shanghang Zhang. Chain of thought prompt tuning in vision language models. arXiv preprint arXiv:2304.07919, 2023.

[133] Xiwen Liang, Min Lin, Weiqi Ruan, Yuecheng Liu, Yuzheng Zhuang, and Xiaodan Liang. Memory-driven multimodal chain of thought for embodied long-horizon task planning. Openreview, 2025.

[134] Yao Yao, Zuchao Li, and Hai Zhao. Beyond chain-of-thought, effective graph-of-thought reasoning in language models. arXiv preprint arXiv:2305.16582, 2023.

[135] You Li, Heyu Huang, Chi Chen, Kaiyu Huang, Chao Huang, Zonghao Guo, Zhiyuan Liu, Jinan Xu, Yuhua Li, Ruixuan Li, et al. Migician: Revealing the magic of free-form multi-image grounding in multimodal large language models. arXiv preprint arXiv:2501.05767, 2025.

[136] Fei Ni, Jianye Hao, Shiguang Wu, Longxin Kou, Jiashun Liu, Yan Zheng, Bin Wang, and Yuzheng Zhuang. Generate subgoal images before act: Unlocking the chain-of-thought reasoning in diffusion model for robot manipulation with multimodal prompts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13991–14000, 2024.

[137] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.

[138] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. arXiv preprint arXiv:2410.16198, 2024.

[139] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18061–18070, 2024.

[140] Qi Sun, Pengfei Hong, Tej Deep Pala, Vernon Toh, U Tan, Deepanway Ghosal, Soujanya Poria, et al. Emma-x: An embodied multimodal action model with grounded chain of thought and look-ahead spatial reasoning. arXiv preprint arXiv:2412.11974, 2024.

[141] Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang, Shuang Wu, Tongtong Cao, Guowei Huang, et al. Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning. arXiv preprint arXiv:2501.10074, 2025.

[142] Zhixuan Shen, Haonan Luo, Kexun Chen, Fengmao Lv, and Tianrui Li. Enhancing multi-robot semantic navigation through multimodal chain-of-thought score collaboration. arXiv preprint arXiv:2412.18292, 2024.

[143] Zhuosheng Zhang and Aston Zhang. You only look at screens: Multimodal chain-of-action agents. arXiv preprint arXiv:2309.11436, 2023.

[144] Jiaqi Zhang, Chen Gao, Liyuan Zhang, Yong Li, and Hongzhi Yin. Smartagent: Chain-of-user-thought for embodied personalized agent in cyber world. arXiv preprint arXiv:2412.07472, 2024.

[145] Tianqi Wang, Enze Xie, Ruihang Chu, Zhenguo Li, and Ping Luo. Drivecot: Integrating chain-of-thought reasoning with end-to-end driving. arXiv preprint arXiv:2403.16996, 2024.

[146] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Receive, reason, and react: Drive as you say, with large language models in autonomous vehicles. IEEE Intelligent Transportation Systems Magazine, 2024.

[147] Yunsheng Ma, Xu Cao, Wenqian Ye, Can Cui, Kai Mei, and Ziran Wang. Learning autonomous driving tasks via human feedbacks with large language models. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 4985–4995, 2024.

[148] Rui Zhao, Qirui Yuan, Jinyu Li, Haofeng Hu, Yun Li, Chengyuan Zheng, and Fei Gao. Sce2drivex: A generalized mllm framework for scene-to-drive learning. arXiv preprint arXiv:2502.14917, 2025.

[149] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In European Conference on Computer Vision, pages 292–308. Springer, 2024.

[150] Haicheng Liao, Hanlin Kong, Bonan Wang, Chengyue Wang, Wang Ye, Zhengbing He, Chengzhong Xu, and Zhenning Li. Cot-drive: Efficient motion forecasting for autonomous driving with llms and chain-of-thought prompting, 2025.

[151] Yi Dai. Interpretable video based stress detection with self-refine chain-of-thought reasoning. arXiv preprint arXiv:2410.09449, 2024.

[152] Nan Xi, Jingjing Meng, and Junsong Yuan. Chain-of-look prompting for verb-centric surgical triplet recognition in endoscopic videos. In Proceedings of the 31st ACM International Conference on Multimedia, pages 5007–5016, 2023.

[153] Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. Medcot: Medical chain of thought via hieratrchical expert. arXiv preprint arXiv:2412.13736, 2024.

[154] Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. arXiv preprint arXiv:2502.19634, 2025.

[155] Meng Luo, Hao Fei, Bobo Li, Shengqiong Wu, Qian Liu, Soujanya Poria, Erik Cambria, Mong-Li Lee, and Wynne Hsu. Panosent: A panoptic sextuple extraction benchmark for multimodal conversational aspect-based sentiment analysis. In Proceedings of the 32nd ACM International Conference on Multimedia, pages 7667–7676, 2024.

[156] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and CUI Bin. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In Forty-first International Conference on Machine Learning, 2024.

[157] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems, 35:2507–2521, 2022.

[158] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In European conference on computer vision, pages 146–162. Springer, 2022.

[159] Yan Wang, Yawen Zeng, Jingsheng Zheng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. Videocot: A video chain-of-thought dataset with active annotation tool. arXiv preprint arXiv:2407.05355, 2024.

[160] Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. arXiv preprint arXiv:2411.14794, 2024.

[161] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, et al. Mavis: Mathematical visual instruction tuning with an automatic data engine. arXiv preprint arXiv:2407.08739, 2024.

[162] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9556–9567, 2024.

[163] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13299–13308, 2024.

[164] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255, 2023.

[165] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multimodal llm truly see the diagrams in visual math problems? In European Conference on Computer Vision, pages 169–186. Springer, 2024.

[166] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. Advances in Neural Information Processing Systems, 37:95095–95169, 2025.

[167] Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. arXiv preprint arXiv:2501.05444, 2025.

[168] Yaoting Wang, Peiwen Sun, Dongzhan Zhou, Guangyao Li, Honggang Zhang, and Di Hu. Ref-avs: Refer and segment objects in audio-visual scenes. In European Conference on Computer Vision, pages 196–213. Springer, 2024.

[169] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. arXiv preprint arXiv:2412.14171, 2024.

[170] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In Proceedings of the IEEE/CVF international conference on computer vision, pages 2694–2703, 2023.

[171] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14375–14385, 2024.

[172] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Yaoting Wang, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Avtrustbench: Assessing and enhancing reliability and robustness in audio-visual llms. arXiv preprint arXiv:2501.02135, 2025.

[173] Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. arXiv preprint arXiv:2410.18325, 2024.

[174] Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. arXiv preprint arXiv:2412.12932, 2024.

[175] Yuanhan Zhang, Kaichen Zhang, Bo Li, Fanyi Pu, Christopher Arif Setiadharma, Jingkang Yang, and Ziwei Liu. Worldqa: Multimodal world knowledge in videos through long-chain reasoning. arXiv preprint arXiv:2405.03272, 2024.

[176] Xiongtao Zhou, Jie He, Lanyu Chen, Jingyu Li, Haojing Chen, Víctor Gutiérrez-Basulto, Jeff Z Pan, and Hanjie Chen. Miceval: Unveiling multimodal chain of thought's quality via image description and reasoning steps. arXiv preprint arXiv:2410.14668, 2024.

[177] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. arXiv preprint arXiv:2402.14008, 2024.

[178] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. arXiv preprint arXiv:2502.09621, 2025.

[179] Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, et al. Omnibench: Towards the future of universal omni-language models. arXiv preprint arXiv:2409.15272, 2024.

[180] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[181] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[182] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[183] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

[184] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. arXiv preprint arXiv:2303.18223, 2023.

[185] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682, 2022.

[186] Libo Qin, Qiguang Chen, Hao Fei, Zhi Chen, Min Li, and Wanxiang Che. What factors affect multi-modal in-context learning? an in-depth exploration. arXiv preprint arXiv:2410.20482, 2024.

[187] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. A survey of chain of thought reasoning: Advances, frontiers and future. arXiv preprint arXiv:2309.15402, 2023.

[188] Maciej Besta, Florim Memedi, Zhenyu Zhang, Robert Gerstenberger, Nils Blach, Piotr Nyczyk, Marcin Copik, Grzegorz Kwaśniewski, Jürgen Müller, Lukas Gianinazzi, et al. Topologies of reasoning: Demystifying chains, trees, and graphs of thoughts. arXiv preprint arXiv:2401.14295, 2024.

[189] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. Advances in neural information processing systems, 35:22199–22213, 2022.

[190] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. arXiv preprint arXiv:2211.12588, 2022.

[191] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. Advances in neural information processing systems, 36:11809–11822, 2023.

[192] Jieyi Long. Large language model guided tree-of-thought. arXiv preprint arXiv:2305.08291, 2023.

[193] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171, 2022.

[194] Anthropic. The claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf, 2024. Preprint.

[195] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning, pages 19730–19742. PMLR, 2023.

[196] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390, 2023.

[197] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592, 2023.

[198] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355, 2023.

[199] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. arXiv preprint arXiv:2311.07919, 2023.

[200] Yu Shu, Siwei Dong, Guangyao Chen, Wenhao Huang, Ruihua Zhang, Daochen Shi, Qiqi Xiang, and Yemin Shi. Llasm: Large language and speech model. arXiv preprint arXiv:2308.15930, 2023.

[201] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824, 2023.

[202] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. Advances in Neural Information Processing Systems, 36: 21487–21506, 2023.

[203] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. arXiv preprint arXiv:2307.05222, 2023.

[204] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. arXiv preprint arXiv:2310.02239, 2023.

[205] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. arXiv preprint arXiv:2305.11000, 2023.

[206] Dong Zhang, Xin Zhang, Jun Zhan, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechgpt-gen: Scaling chain-of-information speech generation. arXiv preprint arXiv:2401.13527, 2024.

[207] Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. Audiopalm: A large language model that can speak and listen. arXiv preprint arXiv:2306.12925, 2023.

[208] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868, 2022.

[209] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. arXiv preprint arXiv:2312.14125, 2023.

[210] Yang Jin, Zhicheng Sun, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, et al. Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. arXiv preprint arXiv:2402.03161, 2024.

[211] Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. arXiv preprint arXiv:2403.14773, 2024.

[212] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. arXiv preprint arXiv:2410.21276, 2024.

[213] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. arXiv preprint arXiv:2402.12226, 2024.

[214] Zhifei Xie and Changqiao Wu. Mini-omni: Language models can hear, talk while thinking in streaming. arXiv preprint arXiv:2408.16725, 2024.

[215] Zhifei Xie and Changqiao Wu. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. arXiv preprint arXiv:2410.11190, 2024.

[216] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. arXiv preprint arXiv:2412.16720, 2024.

[217] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's eye of llms: Visualization-of-thought elicits spatial reasoning in large language models. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.

[218] Jaewook Lee, Yeajin Jang, Hongjin Kim, Woojin Lee, and Harksoo Kim. Analyzing key factors influencing emotion prediction performance of vllms in conversational contexts. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 5801–5816, 2024.

[219] Yuxuan Lei, Dingkang Yang, Zhaoyu Chen, Jiawei Chen, Peng Zhai, and Lihua Zhang. Large vision-language models as emotion recognizers in context awareness. arXiv preprint arXiv:2407.11300, 2024.

[220] Yue Dai, Soyeon Caren Han, and Wei Liu. Multimodal graph constrastive learning and prompt for chartqa. arXiv preprint arXiv:2501.04303, 2025.

[221] Rombach Robin, Blattmann Andreas, Lorenz Dominik, Esser Patrick, and Ommer Björn. High-resolution image synthesis with latent diffusion models, 2021.

[222] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. arXiv preprint arXiv:2408.03314, 2024.

[223] Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms, 2025.

[224] Yifan Wu, Pengchuan Zhang, Wenhan Xiong, Barlas Oguz, James C Gee, and Yixin Nie. The role of chain-of-thought in complex vision-language reasoning task. arXiv preprint arXiv:2311.09193, 2023.

[225] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. arXiv preprint arXiv:2411.10440, 2024.

[226] Jonathan St BT Evans. In two minds: dual-process accounts of reasoning. Trends in cognitive sciences, 7(10):454–459, 2003.

[227] Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. From system 1 to system 2: A survey of reasoning large language models. arXiv preprint arXiv:2502.17419, 2025.

[228] Jinhao Jiang, Zhipeng Chen, Yingqian Min, Jie Chen, Xiaoxue Cheng, Jiapeng Wang, Yiru Tang, Haoxiang Sun, Jia Deng, Wayne Xin Zhao, et al. Technical report: Enhancing llm reasoning with reward-guided tree search. arXiv preprint arXiv:2411.11694, 2024.

[229] Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, et al. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. arXiv preprint arXiv:2412.09413, 2024.

[230] Zeyu Gan, Yun Liao, and Yong Liu. Rethinking external slow-thinking: From snowball errors to probability of correct reasoning. arXiv preprint arXiv:2501.15602, 2025.

[231] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wangxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. arXiv preprint arXiv:2503.09567, 2025.

[232] Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown, November 2024. URL https://qwenlm.github.io/blog/qwq-32b-preview/.

[233] Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. Marco-o1: Towards open reasoning models for open-ended solutions. arXiv preprint arXiv:2411.14405, 2024.

[234] Haotian Xu, Xing Wu, Weinong Wang, Zhongzhi Li, Da Zheng, Boyuan Chen, Yi Hu, Shijia Kang, Jiaming Ji, Yingying Zhang, et al. Redstar: Does scaling long-cot data unlock better slow-reasoning systems? arXiv preprint arXiv:2501.11284, 2025.

[235] Hugging Face. open-r1. https://github.com/huggingface/open-r1, 2025. GitHub Repository.

[236] Jiayi Pan. Tinyzero. https://github.com/Jiayi-Pan/TinyZero, 2025. GitHub Repository.

[237] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. https://github.com/Deep-Agent/R1-V, 2025. Accessed: 2025-02-02.

[238] Haozhan Shen, Zilun Zhang, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model. https://github.com/om-ai-lab/VLM-R1, 2025. Accessed: 2025-02-15.

[239] Peng Yingzhe, Zhang Gongrui, Zhang Miaosen, You Zhiyuan, Liu Jie, Zhu Qipeng, Yang Kai, Xu Xingzhong, Geng Xin, and Yang Xu. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl, 2025.

[240] Zheng Yaowei, Lu Junting, Wang Shenzhi, Feng Zhangchi, Kuang Dongdong, and Xiong Yuwen. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1, 2025.

[241] Wei Liu, Junlong Li, Xiwen Zhang, Fan Zhou, Yu Cheng, and Junxian He. Diving into self-evolving training for multimodal reasoning. arXiv preprint arXiv:2412.17451, 2024.

[242] Wang Weiyun, Gao Zhangwei, Chen Lianjie, Chen Zhe, Zhu Jinguo, Zhao Xiangyu, Liu Yangzhou, Cao Yue, Ye Shenglong, Zhu Xizhou, Lu Lewei, Duan Haodong, Qiao Yu, Dai Jifeng, and Wang Wenhai. Visualprm: An effective process reward model for multimodal reasoning. 2025.

[243] Wang Xiaodong and Peng Peixi. Open-r1-video. https://github.com/Wang-Xiaodong1899/Open-R1-Video, 2025.

[244] Deng Huilin, Zou Ding, Ma Rui, Luo Hongchen, Cao Yang, and Kang Yu. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning, 2025. URL https://arxiv.org/abs/2503.07065.

[245] Liu Yuqi, Peng Bohao, Zhong Zhisheng, Yue Zihao, Lu Fanbin, Yu Bei, and Jia Jiaya. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement, 2025. URL https://arxiv.org/abs/2503.06520.

[246] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning, 2025. URL https://github.com/ModalMinds/MM-EUREKA.

[247] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's "aha moment" in visual reasoning on a 2b non-sft model, 2025. URL https://arxiv.org/abs/2503.05132.

[248] Zhangquan Chen, Xufang Luo, and Dongsheng Li. Visrl: Intention-driven visual perception via reinforced reasoning. arXiv preprint arXiv:2503.07523, 2025.

[249] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. arXiv preprint arXiv:2503.12937, 2025.

[250] Monica AI. Manus, 2025. URL https://manus.im/.

[251] Liang Xinbin, Xiang Jinyu, Yu Zhaoyang, Zhang Jiayi, and Hong Sirui. Openmanus: An open-source framework for building general ai agents. https://github.com/mannaandpoem/OpenManus, 2025.

[252] Hanjia Lyu, Ryan Rossi, Xiang Chen, Md Mehrab Tanjim, Stefano Petrangeli, Somdeb Sarkhel, and Jiebo Luo. X-reflect: Cross-reflection prompting for multimodal recommendation. arXiv preprint arXiv:2408.15172, 2024.

[253] Yongsheng Yu and Jiebo Luo. Chain-of-thought prompting for demographic inference with large multimodal models. In 2024 IEEE International Conference on Multimedia and Expo (ICME), pages 1–7. IEEE, 2024.

[254] Fang Rongyao, Duan Chengqi, Wang Kun, Huang Linjiang, Li Hao, Yan Shilin, Tian Hao, Zeng Xingyu, Zhao Rui, Dai Jifeng, Liu Xihui, and Li Hongsheng. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing, 2025. URL https://arxiv.org/abs/2503.10639.

[255] Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. arXiv preprint arXiv:2412.05237, 2024.

[256] Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification. arXiv preprint arXiv:2502.13383, 2025.

[257] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Jing Hua Toh, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. Advances in Neural Information Processing Systems, 37:52040–52094, 2024.

[258] Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. arXiv preprint arXiv:2405.07960, 2024.

[259] Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, et al. Av-odyssey bench: Can your multimodal llms really understand audio-visual information? arXiv preprint arXiv:2412.02611, 2024.

[260] S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In The Thirteenth International Conference on Learning Representations, 2024.

[261] Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, et al. Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning. arXiv preprint arXiv:2503.07459, 2025.

[262] OpenAI. Openai gpt-4.5 system card, 2025. URL https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf.

[263] OpenAI. Openai gpt-4v system card, 2024. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf.

[264] Google. Introducing gemini 2.0: our new ai model for the agentic era, 2024. URL https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#ceo-message.

[265] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530, 2024.

[266] xAI. Grok-3 beta release, 2025. URL https://x.ai/news/grok-3.

[267] xAI. Grok-2 beta release, 2024. URL https://x.ai/news/grok-2.

[268] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. arXiv preprint arXiv:2501.12599, 2025.

[269] Qwen Team. Qvq: To see the world with wisdom, December 2024. URL https://qwenlm.github.io/blog/qvq-72b-preview/.

[270] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.

[271] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271, 2024.

[272] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 24185–24198, 2024.

[273] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

[274] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326, 2024.

[275] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895, 2024.

[276] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 26296–26306, 2024.

[277] Saeed Amizadeh, Hamid Palangi, Alex Polozov, Yichen Huang, and Kazuhito Koishida. Neuro-symbolic visual reasoning: Disentangling. In International Conference on Machine Learning, pages 279–290, 2020.

[278] Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13326–13365, 2024.

[279] Lauren Nicole DeLong, Ramon Fernández Mir, and Jacques D Fleuriot. Neurosymbolic ai for reasoning over knowledge graphs: A survey. IEEE Transactions on Neural Networks and Learning Systems, 2024.

[280] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. arXiv preprint arXiv:2502.03387, 2025.

Datasets	Year	Task	Domain	Modality	Format	Samples
Training with rationale						
ScienceQA [157]	2022	VQA	Science	T, I	MC	21K
A-OKVQA [158]	2022	VQA	Common	T, I	MC	25K
EgoCoT [39]	2023	VideoQA	Common	T, V	Open	200M
VideoCoT [159]	2024	VideoQA	Human Action	T, V	Open	22K
VideoEspresso [160]	2024	VideoQA	Common	T, V	Open	202,164
EMMA-X [140]	2024	Robot Manipulation	Indoor	T, V	Robot Actions	60K
M3CoT [28]	2024	VQA	Science, Math, Common	T, I	MC	11.4K
MAVIS [161]	2024	ScienceQA	Math	T, I	MC and Open	834K
LLaVA-CoT-100k [225]	2024	VQA	Common, Science	T, I	MC and Open	834K
MAmmoTH-VL [255]	2024	Diverse	Diverse	T, I	MC and Open	12M
Mulberry-260k [94]	2024	Diverse	Diverse	T, I	MC and Open	260K
MM-Verify [256]	2025	MathQA	Math	T, I	MC and Open	59,772
VisualPRM400K [242]	2025	ScienceQA	Math, Science	T, I	MC and Open	400K
R1-OneVision [101]	2025	Diverse	Diverse	T, I	MC and Open	155K
Evaluation without rationale						
MMMU [162]	2023	VQA	Arts, Science	T, I	MC and Open	11.5K
SEED [163]	2023	VQA	Common	T, I	MC	19K
MathVista [164]	2023	ScienceQA	Math	T, I	MC and Open	6,141
MathVerse [165]	2024	ScienceQA	Math	T, I	MC and Open	15K
Math-Vision [166]	2024	ScienceQA	Math	T, I	MC and Open	3040
OSWorld [257]	2024	Agent	Real Comp. Env.	T,I	Agent Actions	369
AgentClinic [258]	2024	MedicalQA	Medical	T,I	Open	335
MeViS [170]	2023	Referring VOS	Common	T, V	Dense Mask	2K
VSIBench [169]	2024	VideoQA	Indoor	T, V	MC and Open	5K
HallusionBench [171]	2024	VQA	Common	T, I	Yes-No	1,129
AV-Odyssey [259]	2024	AVQA	Common	T, V, A	MC	4,555
AVHBench [173]	2024	AVQA	Common	T, V, A	Open	5,816
RefAVS-Bench [168]	2024	Referring AVS	Common	T, V, A	Dense Mask	4,770
MMAU [260]	2024	AQA	Common	T, A	MC	10K
AVTrustBench [172]	2025	AVQA	Common	T, V, A	MC and Open	600K
MIG-Bench [135]	2025	Multi-image Grounding	Common	T, I	BBox	5.89K
MedAgentsBench [261]	2025	MedicalQA	Medical	T, I	MC and Open	862
Evaluation with rationale						
CoMT [174]	2024	VQA	Common	T, I	MC	3,853
OmniBench [179]	2024	VideoQA	Common	T, I, A	MC	1,142
WorldQA [175]	2024	VideoQA	Common	T, V, A	Open	1,007
MiCEval [176]	2024	VQA	Common	T, I	Open	643
OlympiadBench [177]	2024	ScienceQA	Maths, Physics	T, I	Open	8,476
MME-CoT [178]	2025	VQA	Science, Math, Common	T, I	MC and Open	1,130
EMMA [167]	2025	VQA	Science	T, I	MC and Open	2,788
VisualProcessBench [242]	2025	ScienceQA	Math, Science	T, I	MC and Open	2,866

表 3: 数据集和用于 MCoT 训练和评估的基准。“MC” 和 “Open” 分别指代多项选择题和开放式回答格式，而 “T”、“I”、“V” 和 “A” 分别表示文本、图像、视频和音频。

Model	Params (B)	MMMU (Val)	MathVista (mini)	Math-Vision	EMMA (mini)
Human	-	88.6	60.3	68.82	77.75
Random Choice	-	22.1	17.9	7.17	22.75
OpenAI					
o1 [216]	-	78.2	73.9	-	45.75
GPT-4.5 [262]	-	74.4	-	-	-
GPT-4o [212]	-	69.1	63.8	30.39	36.00
GPT-4o mini [212]	-	59.4	56.7	-	-
GPT-4V [263]	-	56.8	49.9	23.98	-
Google & DeepMind					
Gemini 2.0 Pro [264]	-	72.7	-	-	-
Gemini 2.0 Flash [264]	-	71.7	-	41.3	48.00
Gemini 1.5 Pro [265]	-	65.8	63.9	19.24	-
Anthropic					
Claude 3.7 Sonnet [194]	-	75	-	-	56.50
Claude 3.5 Sonnet [194]	-	70.4	67.7	37.99	37.00
Claude 3 Opus [194]	-	59.4	50.5	27.13	-
Claude 3 Sonnet [194]	-	53.1	47.9	-	-
xAI					
Grok-3 [266]	-	78.0	-	-	-
Grok-2 [267]	-	66.1	69.0	-	-
Grok-2 mini [267]	-	63.2	68.1	-	-
Moonshot					
Kimi-k1.5 [268]	-	70	74.9 (test)	38.6	33.75
Alibaba					
QVQ-72B-Preview [269]	72	70.3	71.4	35.9	32.00
Qwen2.5-VL-72B [270]	72	70.2	74.8	38.1	-
Qwen2-VL-72B [11]	72	64.5	70.5	25.9	37.25
Qwen2.5-VL-7B [270]	7	58.6	68.2	25.1	-
Qwen2-VL-7B [11]	7	-	-	16.3	-
OpenGVLab					
InternVL2.5 [271]	78	70.1	-	-	35.25
InternVL2 [272]	76	58.2	65.5	-	-
LLaMA					
Llama-3.2-90B [273]	90	60.3	57.3	-	-
Llama-3.2-11B [273]	11	-	48.6	-	-
LLaVA					
LLaVA-OneVision [274]	72	56.8	67.5	-	27.25
LlaVA-NEXT-72B [275]	72	49.9	46.6	-	-
LLaVA-NEXT-34B [275]	34	48.1	46.5	-	-
LLaVA-NEXT-8B [275]	8	41.7	37.5	-	-
LLaVA-Reasoner [138]	8	40.0	50.6	-	-
LLaVA-1.5 [276]	13	36.4	27.6	11.12	-
Community					
Mulberry [94]	7	55.0	63.1	-	-
MAmmoTH-VL [255]	8	50.8	67.6	24.4	-
MM-Eureka [246]	8	-	67.1	22.2	-
MM-Eureka-Zero [246]	38	-	64.2	26.6	-
Curr-ReFT [244]	7	-	64.5	-	-
Curr-ReFT [244]	3	-	58.6	-	-
LMM-R1 [239]	3	-	63.2	26.35	-
LlamaV-o1 [97]	11	-	54.4	-	-
R1-Onevision [101]	7	-	-	26.16	-
Virgo [95]	7	46.7	-	24.0	-
Insight-V [91]	8	42.0	49.8	-	-
R1-VL [249]	7	63.5	24.7	-	-

表 4: 不同机构的 MLLM 模型在四个基准上的性能比较：MMMU（Val）、MathVista（Mini）、Math-Vision 和 EMMA（Mini）。