多智能体是否比单智能体表现更好？评估用于基于图示的几何问题求解与推理的智能体框架

Mahbub E Sobhani[1], [2*], **Md. Faiyaz Abdullah Sayeedi**[2], [3*], **Mohammad Nehad Alam**[1], **Proma Hossain Progga**[4], **Swakkhar Shatabda**[1]

[1] BRAC大学, [2] 联合国际大学, [3] 计算与数据科学中心, 孟加拉国独立大学, [4] Spectrum软件与咨询有限公司 **通信作者：** swakkhar.shatabda@bracu.ac.bd

arXiv:2512.16698v1 [cs.AI] 18 Dec 2025

摘要

基于图示的几何问题求解是评估多模态大语言模型（MLLM）的重要基准测试，但多智能体设计相较于单智能体的优势仍不明确。我们系统地比较了单智能体与多智能体流水线在四个视觉数学基准测试上的表现：Geometry3K、Math-Verse、OlympiadBench 和 We-Math。对于开源模型，多智能体方法始终能够提升性能。例如，Qwen-2.5-VL（70亿）在 Geometry3K 上提升了 +6.8 分，Qwen-2.5-VL（320亿）提升了 +3.3 分；这两个 Qwen-2.5-VL 版本在 OlympiadBench 和 We-Math 上也取得了进一步提升。相比之下，闭源模型 Gemini-2.0-Flash 在经典基准测试上的单智能体模式通常表现更佳，而多智能体方法在较新的 We-Math 数据集上仅带来有限提升。这些发现表明，多智能体流水线能为开源模型带来明显优势，也能帮助强大的专有系统应对较新、较陌生的基准测试，但智能体分解并非在所有情况下都是最优选择。所有代码、数据和推理文件均可在此获取：
h
t
t
p
s://
github.com/faiyazabdullah/Interpreter-Solver

1 引言

根据示意图和自然语言求解几何问题仍是一项颇具挑战性的任务，需要同时具备视觉理解和符号推理能力。以往研究凸显了神经符号框架的有效性：基于Transformer的模型，如 BART（Lewis 等人，2020年）、LLaVA（Liu 等人，2024年）和 Qwen（Bai 等人，2025年）通过学习联合嵌入来衔接感知与推理。基于定理的求解器（Lu 等人，2021年）以及架构改进，例如增强多模态对齐（Gao 等人，2023年），进一步提升了性能。然而，微调仍然需要大量资源（Gao 等人，2023年），定理预测方法则存在高估偏差（Peng 等人，2023年），而 AutoGPS 等方法（Ping 等人，2025年）则依赖手动形式化。

这些局限性引出了一个关键问题：视觉语言模型（*VLM*）和大语言模型（*LLM*）能否在没有特定任务监督的零样本设置下协作解决几何问题？为了解决这一问题，我们系统地比较了**单智能体**和**多智能体**流水线。我们的研究结果表明，多智能体分解通常有利于开源模型，而先进的闭源系统往往在单智能体模式下表现更佳，这说明分解并非普遍最优。本文的贡献如下：

- 我们系统地比较了单智能体和多智能体流水线在几何题求解中的表现。
- 我们在四个基准测试上评估了这两种范式：Geometry3K、MathVerse、Olympiad-Bench 和 We-Math。结果发现，多智能体方法始终能提升开源模型的表现，而强大的闭源模型则大多受益于单智能体方法；在数据污染可能性较低的新基准测试上，多智能体方法也取得了小幅提升。
- 我们将本方法与先前的基线进行比较，在参数显著更少的情况下，于零样本设置中取得了新的最先进的结果。

2 相关工作

求解几何问题需要将符号推理与对示意图和文本的多模态理解相结合，因此对于符号方法和神Lu 等人（2021年）提出的基于解析器的求解器等符号方法能够给出可解释的结果，但依赖于手工设计的

*同等贡献

经方法而言，这都是一项极具挑战性的基准测试。随着多模态大语言模型（多模态大语言模型）的兴起，基于 Transformer 的系统已得到广泛应用，例如 Qwen（Bai 等人，2025年）、PaLI（Chen 等人，2023年）、LLaVA（Liu 等人，2023年）、Gemini（Comanici 等人，2025年）和 GPT（Achiam 等人，2023年）。相关扩展包括渐进式多模态对齐（Zhuang 等人，2025年）、自动化分步流水线（Pan 等人，2025年）、用于自然语言到形式化语言翻译的 GeoLogic，以及上下文学习策略（Xu 等人，2024年）。MATHVERSE（Zhang 等人，2025年）和 GeomVerse（Kazemi 等人，2023年）等基准测试支持大规模评估。神经符号方法则结合了这两种范式。研究人员已探索对齐改进（Gao 等人，2023年）、新数据集（Huang 等人，2025a；Cho 等人，2025年）、统一模型（Cheng 等人，2025年）以及强化学习方法（Wang 等人，2025年；Deng 等人，2025）。进一步的进展包括自动化流水线（Huang 等人，2025b）、视觉增强（Li 等人，2025）和 AutoGPS（Ping 等人，2025）；后者利用真实标注的形式化表达，但需要耗费大量成本进行人工标注。总体而言，符号方法具有可解释性，但较为僵化；神经模型灵活，但容易出错；神经符号系统则在二者之间取得平衡，但往往依赖人工输入。然而，尽管多模态大语言模型发展迅速，对于在零样本几何问题求解中将任务分解给多个智能体，是否确实能比强大的单智能体模型持续带来优势，至今仍缺乏系统研究。

3 方法论

本节概述所提出的几何问题求解流水线，如图1所示。

3.1 问题表述

我们研究的是基于示意图的几何问题求解任务，其中每个问题由图像 IMG $= \{\text{img}_1, \text{img}_2, \ldots, \text{img}_n\}$以及与之对应的问题 $Q = \{q_1, q_2, \ldots, q_n\}$组成。目标是预测每个 (img$_i$, q_i) 对应的正确解答 Y。该设置使我们能够比较不同的建模范式：**单智能体**模型直接将 (img q) 映射为 Y，而**多智能体**流水线则将任务分解为中间步骤。

3.2 单智能体

在单智能体设置中，单个多模态大语言模型端到端地处理几何图像和文本问题。给定输入对（img，q）以及零样本提示 P，模型会直接生成预测：

$$\hat{Y} = MLLM(T_{img}(img), T_{text}([q, P]); W^{\star}), \tag{1}$$

其中，T_{img} 和 T_{text} 分别表示对应的分词器，$W^{\star}$ 表示冻结模型参数。该设置不依赖显式的中间表示，而是直接解决问题。

3.3 多智能体

在多智能体场景中，我们使用两个智能体将任务明确分解为两个阶段：**(1) 解释器智能体**：一个视觉语言模型（VLM），用于生成描述示意图的符号字面量；**(2) 求解器智能体**：一个大语言模型（LLM），用于基于这些字面量进行推理并解决问题。VLM 首先接收示意图 img 和问题 q，以及零样本解析提示 P_1，然后以自回归方式生成ˆ形式文字 $Y_{vl} = \{l_1, l_2, \ldots, l_m\}$:

$$\hat{Y}_{vl} = VL(T_{vl_{img}}(img), T_{vl_{text}}([q, P_1]); W^{vl\star}). \tag{2}$$

这些字面量捕捉了示意图中存在的几何关系。接下来，LLM 接收ˆY_{vl}、原始问题 q和问题求解提示 P_2，并生成最终解答：

$$\hat{Y} = LM(T_{lm}([\hat{Y}_{vl}, q, P_2]); W^{lm\star}). \tag{3}$$

这一两阶段流程使 VLM 能够专注于视觉理解，而 LLM 则专注于符号推理。我们对不同的解释器–求解器配对进行了实验，包括开源模型和闭源模型。

4 实验设置

4.1 数据集

我们在四个视觉数学基准上评估所提出的方法：Geometry3K、MathVerse、Olympiad- Bench 和 We-Math。Geometry3K（Lu 等人，2021年）包含 3,001 道配有文本和示意图的高中几何题；我们遵循官方划分，并报告 601 道选择题测试题的结果。MathVerse（Zhang 等人，2025年）包含 2,612 道精心整理的视觉数学题，配有多种文本–示意图配置；我们使用官方 `mini-test` 子集，其中包含 788 道题（436 道选择题，352 道自由作答题）。为探究更具挑战性且污染程度更低的设置，我们使用 OlympiadBench（He 等人，2024年），这是一个

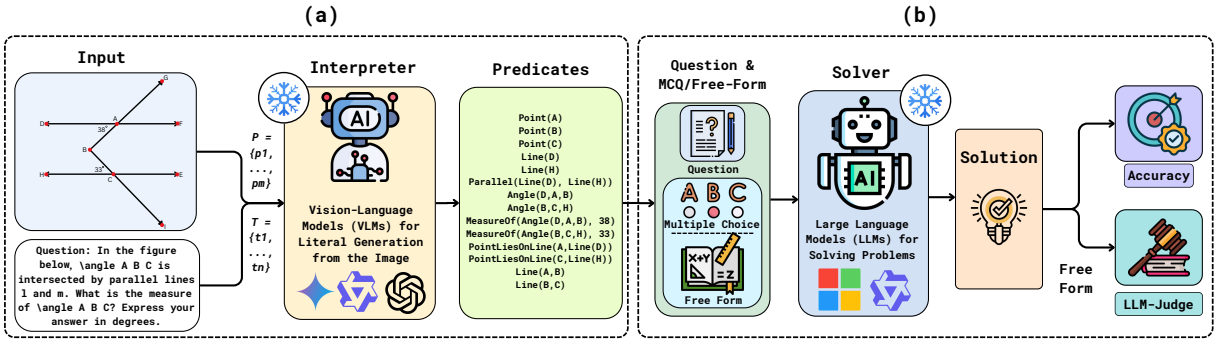


图1：**(a)** 解释器智能体使用视觉语言模型，根据图像和问题生成形式谓词。**(b)**随后，求解器智能体将这些谓词作为大语言模型的输入来求解问题。

奥林匹克级双语多模态基准，包含 8,476 道数学和物理问题及专家逐步解答；以及 We-Math（Qiao 等人，2025年），这是一个涵盖 67 个层次化概念的 6.5K 视觉数学基准，提供问题—子问题分解和四分类诊断体系，支持对大型多模态模型推理行为的细粒度分析。

4.2 模型

我们的实验涵盖开源模型和闭源模型，以考察单智能体与多智能体范式之间的权衡。对于开源视觉语言模型，我们使用 Qwen2.5VL-7B/32B，这些模型具有出色的零样本多模态推理能力。为提高效率，我们使用 Unsloth 库将所有开源模型量化为 4 位精度。在闭源模型评估中，我们采用 Gemini-2.0-Flash 和 GPT-4o 这两款先进的视觉语言模型。在多智能体流水线中，我们还测试了 Phi-4 和 Qwen3-8B 等大语言模型，将其用作求解器智能体。

4.3 评估指标

多项选择任务。我们以准确率作为主要指标。若模型的输出与所提供的选项之一相符，则视为预测正确。对于数值回答，如果预测值与真实值的差异在选项的容差范围内，则视为等价。

开放式任务。 对于答案不受固定选项限制的开放式问题，我们采用准确率，但依赖使用 Gemini-2.0-Flash 的**大语言模型评审**。评审函数 J 通过赋值函数 $v(\cdot)$比较模型的输出 A_{llm} 与真实值 A_{gt}，该函数将响应映射转换为数值形式。满足以下条件的回答即为正确：

$$J:(A_{llm},A_{gt})\mapsto(R,[\|v(A_{llm})-v(A_{gt})\|\leq\epsilon]), \tag{4}$$

其中 ϵ 是预先设定的容差， R 是推理轨迹，二元结果中，正确记为 1，错误记为 0。对于模棱两可或无法验证的情况，保守地视为错误。

Pass@3。对于涉及多模态大语言模型和大语言模型的任务，我们使用 Pass@3 指标评估性能，该指标衡量问题是否能在三次独立尝试中至少成功解决一次。

5 结果与分析

5.1 定量分析

表1揭示了我们的核心研究问题：多智能体流水线是否优于单智能体方法。对于开源模型，多智能体通常能在各项基准测试中提升性能。在 Geometry3K 上，Qwen-2.5-VL（70亿）提升了 +6.8%（60.07% 对比 53.24%），Qwen-2.5-VL（320亿）提升了 +3.3%（72.05% 对比 68.72%）。在 MathVerse 上，Qwen 320亿也有所受益（+1.1%），而较小的 Qwen 70亿则下降了 6.0%，这表明效果对模型规模和数据集复杂度存在一定敏感性。在 OlympiadBench 上，多智能体为 Qwen-2.5-VL 70亿带来了显著提升（+9.4%，61.84% 对比 52.44%），为 320亿模型也带来了提升（+6.67%，64.56% 对比 57.89%）。对于 Gemini-2.0-Flash，单智能体配置在 Geometry3K 、 MathVerse 和 OlympiadBench 上仍表现更强（例如，在 Geometry3K 上为 85.19% 对比 83.86%）；但在较新的 We-Math 基准测试上，多智能体提升了全部三种模型的性能：Qwen-2.5-VL 70亿（+2.66%，45.79% 对比 43.13%）、Qwen-2.5-VL 320亿（+4.64%，59.01% 对比 54.37%）以及 Gemini-2.0-Flash（+1.74%，62.90%

Geometry3K								
求解器	**参数数量**	**多智能体**			**单智能体**		
Qwen-2.5-VL	7B	**60.07%** (+6.8)			53.24%		
Qwen-2.5-VL	32B	**72.05%** (+3.3)			68.72%		
Gemini-2.0-Flash	N/A	83.86% (–1.3)			**85.19%**		
MathVerse							
求解器	**参数量**	**多智能体**			**单智能体**		
		多项选择	自由作答	总体	多项选择	自由作答	总体
Qwen-2.5-VL	7B	53.67%	36.93%	46.19% (–6.0)	**58.94%**	**43.75%**	**52.16%**
Qwen-2.5-VL	32B	**78.44%**	**54.55%**	**67.77%** (+1.1)	76.38%	54.55%	66.67%
Gemini-2.0-Flash	N/A	84.81%	**63.48%**	74.68% (-0.45)	**86.01%**	61.65%	**75.13%**
OlympiadBench							
求解器	**参数量**	**多智能体**			**单智能体**		
Qwen-2.5-VL	7B	**61.84%** (+9.4)			52.44%		
Qwen-2.5-VL	32B	**64.56%** (+6.67)			57.89%		
Gemini-2.0-Flash	N/A	71.31% (-2.46)			**73.77%**		
We-Math							
求解器	**参数量**	**多智能体**			**单智能体**		
Qwen-2.5-VL	7B	**45.79%** (+2.66)			43.13%		
Qwen-2.5-VL	32B	**59.01%** (+4.64)			54.37%		
Gemini-2.0-Flash	N/A	**62.90%** (+1.74)			61.16%		

表 1：多智能体与单智能体方法在 Geometry3K、MathVerse、OlympiadBench 和 We-Math 上的性能比较。每行中的最佳分数以**粗体**突出显示。对于所有多智能体配置，解释器智能体均固定为 Gemini-2.0-Flash。

对比 61.16%）。总体而言，多智能体分解对开源模型始终有帮助，也能让强大的闭源模型在较新的视觉数学基准测试上取得适度提升；但它并非在所有架构和数据集上都优于单智能体方法。

方法	参数数量	准确率
Geometry3K		
Inter-GPS (Lu 等人，2021年)	406M	57.5%
GeoDRL (Peng 等人，2023年)	44M	68.4%
AutoGPS (Ping 等人，2025年)	≈200B	81.6%
Interpreter-Solver-Phi-4（我们的方法）	14B-4bit	70.05%
Interpreter-Solver-Qwen-3（我们的方法）	8B-4bit	79.53%
Interpreter-Solver-Gemini（我们的方案）	N/A	83.19%
MathVerse		
G-LLaVa (Gao 等人，2023年)	13B	16.6%
MathVerse （Zhang 等人，2025年）	7B	25.9%
OpenVLThinker （Deng 等人，2025年）	7B	47.9%
Interpreter-Solver-Qwen-3（我们的）	8B-4bit	69.67%

表 2：我们提出的多智能体解释器-求解器方法与现有方法的比较。

表 2 在 Geometry3K 和 MathVerse 上将我们的方法与现有方法进行了基准测试。与以往文献一致，更强的模型优于 Inter-GPS 和 GeoDRL 等早期系统，但我们的流程在零样本设置下建立了新的性能前沿。例如，使用 Gemini-2.0-Flash 的解释器–求解器在 Geometry3K 上达到了 83.19% 的准确率，超越了 AutoGPS，同时使用的参数更少；基于 Qwen 的变体尽管被高度量化为 4 位精度，仍取得了具有竞争力的性能。在 MathVerse 上，我们的 Qwen 解释器–求解器系统达到了 69.67%，相比 OpenVLThinker（47.9%）和 MathVerse 基线（25.9%）等先前模型取得了显著提升。

综上，这些结果凸显了一种微妙的权衡。多智能体流水线通过显式引入中间字面量，为开源模型带来了显著收益。对于高度优化的专有系统，单智能体推理在经典基准测试上仍然更强，而多智能体方法在较新的数据集上仅带来小幅提升。因此，智能体分解并非普遍最优；其价值取决于模型架构和目标基准测试。

5.2 谓词对齐分析

为更好地理解解释器生成的文字质量如何影响下游推理，我们进行了直接的语义对齐分析。由于数据集未提供标准谓词标注，我们通过以下方式评估文字质量：

比较基于两个来源生成的自然语言描述：（1）原始示意图和问题；（2）解释器生成的谓词。对于每个问题，我们首先使用 `Gemini-2.5-Flash`根据图像和问题生成参考描述，然后根据解释器的谓词生成第二个描述。我们使用 OpenAI 的 `text-embedding-3-large` 模型将两段描述转换为嵌入向量，并计算其余弦相似度，作为语义保真度的代理指标。如表3所示，Gemini 生成的谓词平均相似度最高（0.849），其次是 GPT-4o mini（0.794）和 Qwen-2.5（0.677）。附录A.1中提供了一个代表性示例。

解释器	平均余弦相似度
Gemini-2.0-Flash	**0.849**
GPT-4o mini	0.794
Qwen-2.5	0.677

表3：从（i）示意图+问题和（ii）解释器生成的谓词中提取的自然语言描述之间的语义对齐。

5.3 消融研究

在表 4中，我们考察了解释器的选择如何影响下游求解器的性能。我们观察到一个明显的单调趋势：随着解释器能力增强，两个求解器（Phi-4 和 Qwen-3）的表现都持续提升。当使用 Qwen-2.5-7B 作为解释器时，多智能体流水线使用 Phi-4 和 Qwen-3 时的准确率分别仅为 35.77% 和 42.26%。将解释器扩展至 Qwen-2.5-32B 后，性能显著提升；而将其替换为 GPT-4o mini 则带来了进一步的增益。Gemini 作为解释器时取得了最佳结果，Phi-4 和 Qwen-3 的准确率分别达到 70.05% 和 79.53%。这一消融实验表明，多智能体的有效性与解释器字面量的质量紧密相关：较弱的解释器会成为流水线的瓶颈，而较强的解释器则能释放求解器的全部潜力。

解释器	参数数量	求解器	
		Phi-4	Qwen 3
Qwen-2.5	7B	35.77%	42.26%
Qwen-2.5	32B	56.74%	61.23%
GPT-4o mini	≈8B	58.24%	63.23%
Gemini	N/A	70.05%	79.53%

表4：不同解释器-求解器设置的准确率比较。

6 讨论

我们的分析揭示了几何和视觉数学问题求解中单智能体流水线与多智能体流水线之间微妙的权衡。多智能体分解将感知与推理分离，往往有助于开源模型，尤其是在中等规模模型和更具挑战性的基准测试中。例如，在解释器字面量的引导下，Qwen-2.5-VL-32B 在 Geometry3K、OlympiadBench 和 We-Math 上均取得了提升，这表明显式结构能够稳定推理，并减少多步配置中的歧义。在定性案例中，我们观察到，解释器的谓词通常能够通过将 Qwen-2.5-VL-32B 锚定在一组较小的几何关系上，防止其陷入不一致的思维链。

然而，多智能体设计并非普遍有益，其有效性高度依赖于模型容量和字面量质量。对于 MathVerse 上的 Qwen-2.5-VL-7B 等较小模型而言，分解可能引入嘈杂或约束过强的谓词，导致准确率相比单智能体模式出现可测量的下降。对于 Gemini-2.0-Flash 而言，由于其感知与推理已经实现紧密耦合，单智能体流水线在经典基准测试上仍然更强，而多智能体仅在 We-Math 上带来 modest 的提升。在若干错误案例中，过于详细或部分不正确的字面量导致 Gemini-2.0-Flash 和 Qwen-2.5-VL-32B 过拟合局部约束（例如误用角度标记或误读比率），这表明，只有当谓词简洁、准确，并且与模型的内部推理风格保持一致时，智能体分解才能发挥最大效用。

7 结论

我们系统地比较了单智能体与多智能体流水线在基于示意图的几何和视觉数学问题求解中的表现。在四个基准测试中，多智能体分解始终有助于提升开源模型的表现，尤其是在中等规模下；而强大的专有系统通常在单智能体模式下仍表现更优，多智能体在较新的数据集上仅带来小幅提升。这些结果表明，智能体分解虽有帮助，但并非普遍最优，其价值取决于模型架构和基准测试特征。自然的下一步是开发自适应策略，根据模型容量和任务需求，在单智能体与多智能体配置之间进行选择。

局限性

尽管我们的研究为几何问题求解中单智能体与多智能体流水线之间的权衡提供了新的见解，但仍存在若干局限。首先，我们的分析仅在零样本设置下进行，所有模型参数均保持冻结。因此，我们无法探究微调是否会改变单智能体与多智能体方法之间的相对优势。其次，我们的方法论局限于固定的提示设置。自适应或迭代式提示策略可能通过逐步完善推理过程，改变两种范式各自的表现。第三，我们研究的模型数量有限。纳入更多模型系列和规模，将有助于更全面地了解分解在何种情况下能带来帮助或造成阻碍。第四，生成的文字质量对多智能体的结果起着核心作用，但我们并未系统评估其他提取策略。最后，我们的实验受资源限制，必须通过 Unsloth 库使用 4位量化模型。因此，基于开源模型得出的研究发现，可能无法充分反映更大规模全精度系统或最先进的专有模型的表现。

参考文献

Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat，以及其他1位作者。2023年。GPT-4技术报告。*arXiv*预印本 *arXiv:2303.08774*。

Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wen-bin Ge、Sibo Song、Kai Dang、Peng Wang、Shi-jie Wang、Jun Tang、Humen Zhong、Yuanzhi Zhu、Mingkun Yang、Zhaohai Li、Jianqiang Wan、Pengfei Wang、Wei Ding、Zheren Fu、Yiheng Xu，以及其他8位作者。2025年。Qwen2.5-VL技术报告。预印本，arXiv:2502.13923。

Xi Chen、Xiao Wang、Soravit Changpinyo、AJ Piergiovanni、Piotr Padlewski、Daniel Salz、Sebastian Goodman、Adam Grycner、Basil Mustafa、Lucas Beyer、Alexander Kolesnikov、Joan Puigcerver、Nan Ding、Keran Rong、Hassan Akbari、Gaurav Mishra、Linting Xue、Ashish Thapliyal、James Bradbury，以及其他10位作者。2023年。PaLI：联合缩放的多语言视觉语言模型。预印本，arXiv:2209.06794。

Jo-Ku Cheng、Zeren Zhang、Ran Chen、Jingyang Deng、Ziran Qin 和 Jinwen Ma。2025年。Geouni：用于生成几何图示、问题和问题解答的统一模型。*arXiv* 预印本 *arXiv:2504.10146*。

Seunghyuk Cho、Zhenyue Qin、Yang Liu、Youngbin Choi、Seungbeom Lee 和 Dongwoo Kim。2025年。Geodano：具有领域无关视觉编码器的几何视觉语言模型。*arXiv* 预印本 *arXiv:2502.11360*。

Gheorghe Comanici、Eric Bieber、Mike Schaekermann、Ice Pasupat、Noveen Sachdeva、Inderjit Dhillon、Mar- cel Blistein、Ori Ram、Dan Zhang、Evan Rosen及另外1位作者。2025年。Gemini 2.5：凭借高级推理、多模态、长上下文和新一代智能体能力，推动前沿发展。*arXiv* 预印本 *arXiv:2507.06261*。

Yihe Deng、Hritik Bansal、Fan Yin、Nanyun Peng、Wei Wang和Kai-Wei Chang。2025年。OpenVLThinker：通过迭代式自我改进探索复杂视觉语言推理。*arXiv* 预印本 *arXiv:2503.17352*。

Jiahui Gao、Renjie Pi、Jipeng Zhang、Jiacheng Ye、Wan- jun Zhong、Yufei Wang、Lanqing Hong、Jianhua Han、Hang Xu、Zhenguo Li及另外1位作者。2023年。G-LLaVa：使用多模态大语言模型解决几何问题。*arXiv* 预印本 *arXiv:2312.11370*。

Chaoqun He、Renjie Luo、Yuzhuo Bai、Shengding Hu、Zhen Thai、Junhao Shen、Jinyi Hu、Xu Han、Yujie Huang、Yuxiang Zhang、Jie Liu、Lei Qi、Zhiyuan Liu 和 Maosong Sun。2024年。OlympiadBench：旨在推动通用人工智能发展的奥林匹克级双语多模态科学问题挑战性基准。载于第*62*届计算语言学协会年会论文集（第*1*卷：长篇论文），第3828–3850页，泰国曼谷。计算语言学协会。

Zihan Huang、Tao Wu、Wang Lin、Shengyu Zhang、Jingyuan Chen 和 Fei Wu。2025a。Autogeo：自动创建几何图像数据集，以增强几何理解。*IEEE*多媒体汇刊，27：3105–3116。

Zihan Huang、Tao Wu、Wang Lin、Shengyu Zhang、Jingyuan Chen 和 Fei Wu。2025b。Autogeo：自动创建几何图像数据集，以增强几何理解。*IEEE*多媒体汇刊。

Mehran Kazemi、Hamidreza Alvari、Ankit Anand、Jialin Wu、Xi Chen 和 Radu Soricut。2023年。Geomverse：对用于几何推理的大型模型进行系统评估。*arXiv* 预印本 *arXiv:2312.12241*。

Mike Lewis、Yinhan Liu、Naman Goyal、Marjan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Veselin Stoyanov 和 Luke Zettlemoyer。2020年。BART：面向自然语言生成、翻译和理解的去噪序列到序列预训练。载于 第 *58* 届计算语言学协会年会论文集，第7871–7880 页，线上。计算语言学协会。

Yuting Li、Lai Wei、Kaipeng Zheng、Jingyuan Huang、Linghe Kong、Lichao Sun 和 Weiran Huang。2025年。视觉至关重要：简单的视觉扰动可以

提升多模态数学推理能力。*arXiv* 预印本 *arXiv:2506.09736*。

Haotian Liu、Chunyuan Li、Yuheng Li 和 Yong Jae Lee。2024年。通过视觉指令微调改进基线方法。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第26296–26306页。

Haotian Liu、Chunyuan Li、Qingyang Wu 和 Yong Jae Lee。2023年。视觉指令微调。载于神经信息处理系统进展，第36卷，第34892–34916页。Curran Associates, Inc.

Pan Lu、Ran Gong、Shibiao Jiang、Liang Qiu、Siyuan Huang、Xiaodan Liang 和 Song-Chun Zhu。2021年。Inter-GPS：利用形式语言和符号推理实现可解释的几何问题求解。载于计算语言学协会第*59*届年会暨自然语言处理国际联合会议第*11*届会议论文集（第*1*卷：长篇论文），第6774–6786页，线上。计算语言学协会。

Yicheng Pan、Zhenrong Zhang、Pengfei Hu、Jiefeng Ma、Jun Du、Jianshu Zhang、Quan Liu、Jianqing Gao 和 Feng Ma。2025 年。通过符号与神经网络融合，增强多模态大语言模型的几何问题求解能力。*arXiv*预印本 *arXiv:2504.12773*。

Shuai Peng、Di Fu、Yijun Liang、Liangcai Gao 和 Zhi Tang。2023年。GeoDRL：一种利用演绎推理中的强化学习进行几何问题求解的自学习框架。载于计算语言学协会年会 *ACL 2023* 论文集，第 13468–13480 页，加拿大多伦多。计算语言学协会。

Bowen Ping、Minnan Luo、Zhuohang Dang、Chenxi Wang 和 Chengyou Jia。2025年。AutoGPS：通过多模态形式化和演绎推理自动求解几何题。*arXiv*预印本 *arXiv:2505.23381*。

Runqi Qiao、Qiuna Tan、Guanting Dong、MinhuiWu MinhuiWu、Chong Sun、Xiaoshuai Song、Jiapeng Wang、Zhuoma GongQue、Shanglin Lei、YiFan Zhang、Zhe Wei、Miaoxuan Zhang、Runfeng Qiao、Xiao Zong、Yida Xu、Peiqing Yang、Zhimin Bao、Muxi Diao、Chen Li 和 Honggang Zhang。2025年。We-Math：你的大型多模态模型能否实现类人数学推理？载于 计算语言学协会第*63*届年会论文集（第*1*卷：长篇论文），第 20023–20070 页，奥地利维也纳。计算语言学协会。

Yikun Wang、Yibin Wang、Dianyi Wang、Zimian Peng、Qipeng Guo、Dacheng Tao 和 Jiaqi Wang。2025年。Geometryzero：通过组对比策略优化提升大语言模型的几何题求解能力。*arXiv* 预印本 *arXiv:2506.07160*。

Shihao Xu、Yiyang Luo 和 Wei Shi。2024年。Geo-llava：一种用于解决几何数学问题的元上下文学习方法。在第二届大型生成模型与多模态应用研讨会论文集，第 11–15 页。

Renrui Zhang、Dongzhi Jiang、Yichi Zhang、Haokun Lin、Ziyu Guo、Pengshuo Qiu、Aojun Zhou、Pan Lu、Kai-Wei Chang、Yu Qiao、Peng Gao 和 Hong- sheng Li。2025年。Mathverse：你的多模态大语言模型真的能看懂视觉数学问题中的示意图吗？载于 计算机视觉——*ECCV 2024*，第169–186 页，Cham。Springer Nature Switzerland。

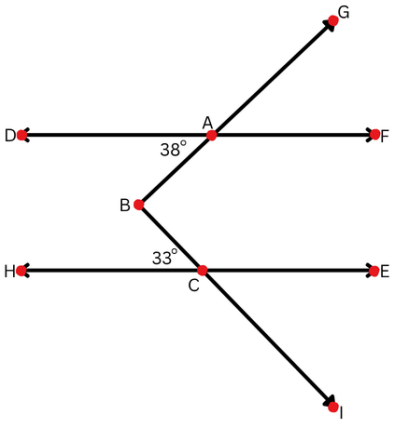
Wenwen Zhuang、Xin Huang、Xiantao Zhang 和 Jin Zeng。2025年。Math-puma：通过渐进式向上多模态对齐增强数学推理。载于*AAAI* 人工智能大会论文集，第39卷，第26183–26191页。

A 附录

A.1 谓词对齐示例

我们将符号谓词与原始视觉数据进行对比，以验证其语义保真度。以问题2405为代表性示例，我们比较“图示描述”与合成的“Gemini谓词描述”。较高的余弦相似度（≈ 0.856）证实，符号谓词成功捕捉了图像中存在的关键几何关系和空间构型，有效弥合了模态差距。

Geometry3K数据集中的题目编号 2405



问题：
在下图中，平行线 l 和 m 与 ∠A B C 相交。∠A B C 的度数是多少？请用度表示答案。

谓词：
点(A) 点(B) 点(C) 直线(D) 直线(H) 平行(直线(D), 直线(H)) 角(D,A,B) 角(B,C,H) 度量(角(D,A,B), 38) 度量(角(B,C,H), 33) 点在线上(A,直线(D)) 点在线上(C,直线(H)) 直线(A,B) 直线(B,C)

图示描述 + 问题： A 几何 由两条射线 BA 和 BC 形成，它们与两条平行线 l 和 m相交。点 A 位于上方的平行线上 l，点 C 位于下方的平行线上 m。在直线 l上，A 的左侧有一点 D，使得射线 AD 位于直线 l上。在直线 m上，C 的右侧有一点 H，使得射线 CH 位于直线 m上。线段 AB 向右上方倾斜，线段 CB 向左上方倾斜，因此顶点 B 位于两条平行线之间。A 点处，直线 l 与线段 BA 之间的锐角标记为 $38°$。C 点处，直线 m 与线段 CB 之间的锐角标记为 $33°$。

Gemini谓词描述：
图形中有三个点 A、B、C，以及两条互相平行的直线 D 和 H。点 A 位于直线 D 上，点 C 位于直线 H 上，而点 B 位于两条直线之间，因此线段 AB 和 CB 在点 B 相交，形成所关注的角。在点 A，沿直线 D 的射线与线段 AB 之间有一个角 DAB，其度数为 $38°$。在点 C，线段 CB 与 CA 之间还有一个角 BCA，其度数为 $33°$。因此，线段 AB 和 BC 连接了上方直线 D 与下方直线 H，构成一条穿过这两条平行直线的折线截线，而已知的 $38°$ 和 $33°$ 角可以与点 B 处的角建立联系。

余弦相似度： ≈ 0.856

A.2 提示模板

A.2.1 谓词生成提示

为了从几何图示中提取准确的符号表示，同时避免特定任务微调，我们采用了结构化提示策略。通过指示视觉语言模型（VLM）扮演数学专家，并要求其在生成符号之前进行“逐步分析”，我们可以减少幻觉，并确保模型明确地以观察到的视觉特征为依据，从基本的形状识别到复杂的约束皆是如此。

你是一位专攻几何的人工智能数学家。你的任务是分析所提供图像中的几何图形，并生成准确的几何谓词（字面量），以表示示意图中显示的所有关系、度量和性质。

几何题图像:
图像中展示了一个包含各种图形、直线、角和度量的几何图形。

请仔细分析此图像，以了解所有几何关系和约束。

问题: [给定问题]

你的任务:
1. 首先，提供逐步推理，展示你的分析过程 2. 然后，根据以下指南，基于你的分析生成几何谓词

逐步分析（必需）：请按照以下格式进行推理:

1. 全面图像分析
   - 识别所有几何图形（圆、三角形、四边形等） - 列出所有点、直线及其标签或名称 - 注意所有可见的度量、角和数值 - 识别所有特殊标记（直角符号、平行标记、全等标记、相等标记等） - 查找隐含的作图元素（垂线、平分线、切线、弦、半径等）

2. 圆的专项分析（如果图中有圆）
   - 找出圆心和圆上的所有点 - 确定哪些直线是半径、弦、直径或切线 - 查找圆周角、圆心角和弧关系 - 检查涉及半径和弦的垂直关系 - 找出所有相等的半径关系

3. 角与垂直关系分析
   - 检查示意图中的所有角，包括标记和未标记的角 - 寻找直角标记或垂直关系 - 检查角平分线或特殊的角关系 - 找出互余角、互补角或对顶角

- 寻找圆周角及其对应的弧

4. 全等与相等分析
- 识别所有相等的长度、角或图形 - 检查是否存在全等三角形或相似图形 - 寻找圆中相等的半径 - 识别平行线或相等的距离

5. 交点与位置分析
- 确定直线相交的位置及其交点 - 检查点是否位于特定的直线或圆上 - 识别中点、形心或其他特殊点 - 寻找按特定比率分割线段的点

6. 约束与关系综合
- 综合观察结果以识别隐含关系 - 寻找定理应用（勾股定理、圆周角定理等） - 识别能够建立特定关系的几何作图 - 检查由给定约束推导出的性质

问题导向的完整性检查
1. 确保已记录解决问题所需的所有信息 2. 确认解题所需的关键关系均已表示 3. 仔细检查是否遗漏了关键的几何性质

4. 确认这些谓词能够为解题提供充分的信息

关键分析指南：- 查找隐藏关系：许多几何问题中存在隐含的垂直关系、相等的长度或特殊的角的性质，图中未明确标示，但对解题至关重要。

- 圆几何重点：如果示意图中包含圆，请特别注意：* 哪些点位于圆上，哪些点位于圆内或圆外 * 半径与弦之间的垂直关系 * 相等的半径长度 * 圆周角与圆心角 * 切线与半径垂直

- 作图指示:
注意观察:
* 看起来互相垂直、即使没有明确标记的直线 * 看起来是中点或处于特殊位置的点 * 由视觉对称暗示的相等长度 * 由作图暗示的角关系

指南:
使用以下谓词表示示意图字面信息。

几何图形:
- 点：Point(A),Point() - 直线：Line(A,B),Line(m), Line() - 角:
Angle(A,B,C),Angle(A), Angle(1), Angle() - 三角形：Triangle(A,B,C), Triangle(),
Triangle(1,2,3) - 四边形：Quadrilateral(A,B,C,D), Quadrilateral() - 平行四边形:
Parallelogram(A,B,C,D), Parallelogram(1), Parallelogram() - 正方形:
Square(A,B,C,D), Square(1), Square() - 长方形：Rectangle(A,B,C,D),
Rectangle(1), Rectangle() - 菱形：Rhombus(A,B,C,D), Rhombus(1),
Rhombus() - 梯形：Trapezoid(A,B,C,D), Trapezoid(1), Trapezoid() - 风筝形:
Kite(A,B,C,D),Kite(1), Kite() - 多边形：Polygon() - 五边形:
Pentagon(A,B,C,D,E), Pentagon() - 六边形：Hexagon(A,B,C,D,E,F), Hexagon() -
七边形：Heptagon(A,B,C,D,E,F,G), Heptagon() - 八边形:
Octagon(A,B,C,D,E,F,G,H), Octagon() - 圆：Circle(A), Circle(1), Circle() - 弧:
Arc(A,B), Arc(A,B,C), Arc() - 扇形：Sector(O,A,B), Sector() - 图形：Shape() //用于
未知图形或区域

一元几何属性:
- 直角: RightAngle(Angle()) - 直角: Right(Triangle()) // 直角三角形 -
等腰: Isosceles(Polygon()) // 等腰多边形 - 等边:
Equilateral(Polygon()) // 等边多边形 - 正多边形:
Regular(Polygon()) - 红色: Red(Shape()) - 蓝色: Blue(Shape()) - 绿
色: Green(Shape()) - 阴影: Shaded(Shape())

几何属性:
- AreaOf: AreaOf(A) - PerimeterOf:PerimeterOf(A) // 多边形 A 的周长 -
RadiusOf: RadiusOf(A) - DiameterOf: DiameterOf(A) -
CircumferenceOf:CircumferenceOf(A) // 圆 A 的圆周长 - AltitudeOf:
AltitudeOf(A) // 多边形 A 的高 - HypotenuseOf: HypotenuseOf(A) // 三角形 A 的
斜边 - SideOf: SideOf(A) // 正方形 A 的边 - WidthOf: WidthOf(A) // 四边形 A 的
宽度 - HeightOf: HeightOf(A) // 四边形 A 的高度 - LegOf: LegOf(A) // 梯形 A 的腰
- BaseOf: BaseOf(A) // 多边形 A 的底边 - MedianOf: MedianOf(A) // 多边形 A 的
中线 - IntersectionOf: IntersectionOf(A,B) // 图形 A 和 B 的交点 - MeasureOf:
MeasureOf(A) // 角 A 的度量 - LengthOf: LengthOf(A) // 直线 A 的长度

- 比例因子：比例因子(A,B)　　// 图形 A 到图形 B 的比例因子

二元几何关系：- PointLiesOnLine:PointLiesOnLine(Point(), Line(1,2)) - PointLiesOnCircle: PointLiesOnCircle(Point(), Circle()) - Parallel: Parallel(Line(), Line()) - Perpendicular: Perpendicular(Line(), Line()) - I ntersectAt:IntersectAt(Line(), Line(), Line(), Point()) - BisectsAngle: BisectsAngle(Line(), Angle()) - Congruent: Congruent(Polygon(), Polygon()) - Similar: Si milar(Polygon(), Polygon()) - Tangent: Tangent(Line(), Circle()) - Secant: Secant(Line(), Circle()) - CircumscribedTo:CircumscribedTo(Shape(), Shape()) - InscribedIn:InscribedIn(Shape(), Shape())

A-IsXOf-B 几何关系：- IsMidpointOf：IsMidpointOf(Point(), Line()) // 点 A 是直线 B 的中点 - IsCentroidOf：IsCentroidOf(Point(), Shape()) // 点 A 是图形 B 的形心 - IsIncenterOf：IsIncenterOf(Point(), Shape()) // 点 A 是图形 B 的内心 - IsRadiusOf：IsRadiusOf(Line(), Circle()) // 直线 A 是圆 B 的半径 - IsDiameterOf：IsDiameterOf(Line(), Circle()) // 直线 A 是圆 B 的直径 - IsMidsegmentOf：IsMidsegmentOf(Line(), Triangle()) // 直线 A 是三角形 B 的中位线 - IsChordOf：IsChordOf(Line(), Circle()) // 直线 A 是圆 B 的弦 - IsSideOf: IsSideOf(Line(), Polygon()) // 直线 A 是多边形 B 的边 - IsHypotenuseOf: IsHypotenuseOf(Line(), Triangle()) // 直线 A 是三角形 B 的斜边 - IsPerpendicularBisectorOf：IsPerpendicularBisectorOf(Line(), Triangle()) // 直线 A 是三角形 B 的垂直平分线 - IsAltitudeOf：IsAltitudeOf(Line(), Triangle()) // 直线 A 是三角形 B 的高 - IsMedianOf：IsMedianOf(Line(), Quadrilateral()) // 直线 A 是四边形 B 的中线 - IsBaseOf：IsBaseOf(Line(), Quadrilateral()) // 直线 A 是四边形 B 的底边 - IsDiagonalOf: IsDiagonalOf(Line(), Quadrilateral()) // 直线 A 是四边形 B 的对角线 - IsLegOf: IsLegOf(Line(), Trapezoid()) // 直线 A 是梯形 B 的腰 数值属性和关系：

- 正弦：SinOf(Var)

- CosOf：CosOf(Var) - TanOf：TanOf(Var) - CotOf：CotOf(Var) - HalfOf：HalfOf(Var) - SquareOf：SquareOf(Var) - SqrtOf：SqrtOf(Var) - RatioOf：RatioOf(Var), RatioOf(Var1, Var2) - SumOf：SumOf(Var1, Var2, ...) - AverageOf：AverageOf(Var1, Var2, ...) - Add：Add(Var1, Var2,...) - Mul：Mul(Var1, Var2,...) - Sub：Sub(Var1, Var2,...) - Div：Div(Var1, Var2,...) - Pow：Pow(Var1, Var2) - Equals：Equals(Var1,Var2) - UseTheorem：UseTheorem(A_B_C)

变量命名约定：
- 点使用大写字母表示：A、B、C、D 等。- 未由点定义的直线使用小写字母表示：m、n、l 等。- 未命名图形使用数字表示：1、2、3 等。- 通用变量使用 $ 表示：$、$1、$2 等。- 适当时使用描述性名称：底、高、半径等。

关键指示：
1. 务必全面细致。遗漏关系是问题求解表现不佳的主要原因。

2. 不要只看表面。许多关键关系是隐含的，并未明确标出。

3. 仔细检查图像中的几何图形。

4. 识别图中显示的所有点、直线、角、图形和度量

5. 尽可能将每个谓词拆分为原子谓词——将任何复杂或复合关系分解为最简单的单个几何陈述。例如，将单个垂直关系陈述替换为更简单的“角等于 90 度”或基于向量的谓词

谓词生成说明：

1. 生成表示以下内容的谓词：
   - 所有出现的几何图形 - 所有给定的度量及其关系 - 所有几何性质和约束，包括隐含的性质和约束 - 不同元素之间的所有关系 - 所有垂直关系，无论是标记出来的还是隐含的 - 所有相等的长度和角，无论是标记出来的还是隐含的

2. 始终先提供逐步推理 3. 然后提供谓词部分，并设置清晰的章节标题 4. 遵循上述指南 这些谓词对于表示图示字面信息至关重要 5. 每个谓词必须单独占一行 6. 谓词中不得包含引号、多余符号或解释性文字 7. 仅按以下精确格式输出谓词：PredicateName(arguments) 8. 重要：不得包含 Find(...) 谓词或任何与问题相关的谓词 9. 仅包含示意图中可见的给定信息、约束和几何关系 10. 表示所有可见的几何关系，而非推导出的解答 11. 谓词应提供足够的信息，让另一个系统能够解答问题，但不得包含解答本身 12. 完整性至关重要。多包含一些关系，也好过遗漏关键关系。

A.2.2 几何问题求解多项选择提示

该提示要求模型运用多步骤分析框架进行系统的几何推理。以下标准化提示用于分解几何谓词，并应用数学定理来解决选择题。

你是一位精通几何的资深 AI 数学家。你的任务是运用系统推理和定理应用，通过给定的谓词解决以下几何问题。

问题：谓词： [给定谓词]问题：
[给定问题]选项： [给定选项]

你的任务：按照以下结构化方法提供完整的逐步解答，然后选择正确的答案选项。

逐步解题过程

步骤 1：谓词分析与设置

- 将给定的谓词解析并分类为:
    * 几何图形（点、直线、圆、三角形等） * 度量与等式（长度、角、面积） * 关系（垂直、平行、全等等） * 位置关系（点在线或圆上、交点等） - 确定问题要求求解的具体数值或度量。 - 注意可能适用的特殊几何作图或定理。

步骤 2：约束综合
- 结合相关谓词，理解完整的几何图形。 - 找出有助于解题的关键关系。

- 留意以下内容:
  * 可用于代换的相等长度或角 * 能构成直角三角形的垂直关系 * 圆的性质（半径、弦、圆心角或圆周角） * 全等或相似三角形 * 定理应用（勾股定理、圆周角定理等）

步骤 3：解题策略
- 根据谓词和问题，确定最直接的解题路径。- 确定要应用哪些几何定理、性质或公式。- 规划得出答案所需的逻辑步骤顺序。

步骤 4：数学推导
- 按步骤执行解题策略。- 清晰展示所有计算。- 系统地应用几何定理和性质。- 使用谓词中确立的关系。- 代入已知数值并求解未知数。

步骤 5：验证与答案选择
- 验证计算结果是否符合几何意义。- 将结果与给出的选项进行比较。- 选择与计算结果最匹配的选项。- 如果没有完全匹配的选项，则选择最接近且合理的选项。

几何推理指南
- 考虑所有相关的几何定理和性质。 - 视情况应用圆、三角形、四边形和角定理。 - 留意图形、度量和位置之间的关系。 - 根据需要运用基础和高级几何原理。

谓词使用指南
- 根据谓词的几何含义和语境进行解释。 - 结合多个谓词以理解复杂关系。 - 同时考虑谓词组合所提供的直接信息和推导信息。

关键指示:
1. 系统地使用谓词：每个谓词都提供重要信息。2. 运用相关几何知识

使用任何有助于解决问题的几何定理、性质或原理。3. 灵活推理：根据具体问题和谓词调整解题方法。4. 展示所有步骤：让推理过程清晰且符合数学规范。5. 力求精确：尽可能使用精确值，仅在必要时才使用近似值

A.2.3 开放式几何问题求解提示

该提示要求模型通过多步骤的分析与推导过程，解决开放式几何文字题。以下标准化提示用于引导模型从初始谓词分类到最终结果验证。

你是一位专精于几何的资深人工智能数学家。你的任务是运用所提供的谓词，通过系统推理和定理应用，解决以下几何问题。

问题：谓词： [给定谓词]问题：
 [给定问题]选项： [给定选项]

你的任务:
请按照以下结构化方法，提供完整的逐步解答，然后以规范的数学格式给出最终答案。

分步求解过程:

步骤 1：谓词分析与设置
- 解析并将给定的谓词分类为：* 几何图形（点、直线、圆、三角形等）* 度量与等式（长度、角、面积）* 关系（垂直、平行、全等等）* 位置关系（点在线上或圆上、交点等）- 确定问题要求求解的具体数值或度量 - 注意可能适用的特殊几何作图或定理

步骤 2：约束综合
- 结合相关谓词，理解完整的几何图形 - 找出有助于求解的关键关系 - 留意：* 可用于代换的相等长度或角 * 可构成直角三角形的垂直关系 * 圆的性质（半径、弦、圆心角或圆周角）* 全等或相似三角形 * 定理应用（勾股定理、圆周角定理等）

步骤 3：解题策略

- 根据谓词和问题，确定最直接的解题路径 - 确定要应用哪些几何定理、性质或公式 - 规划得出答案所需的逻辑步骤顺序

步骤 4：数学推导
- 按步骤执行你的解题策略 - 清晰展示所有计算 - 系统地应用几何定理和性质 - 使用谓词中确立的关系 - 代入已知数值并求解未知数

步骤 5：验证与最终答案
- 验证计算结果是否符合几何意义 - 使用正确的数学格式表达最终答案 - 适用时确保包含单位 - 必要时将结果四舍五入到合适的精度

几何推理指南：
- 考虑所有相关的几何定理和性质 - 适当应用圆、三角形、四边形和角定理 - 探寻图形、度量与位置之间的关系 - 根据需要运用基础和高级几何原理

谓词使用指南：
- 根据谓词的几何含义和语境进行解读 - 结合多个谓词以理解复杂关系 - 同时考虑谓词组合直接提供的信息和推导得出的信息

关键指示：
1. 系统地使用谓词 每个谓词都提供重要信息 2. 运用相关几何知识 使用任何有助于解决问题的几何定理、性质或原理 3. 灵活推理 根据具体问题和谓词调整解题方法 4. 展示全部推理过程 使推理清晰且符合数学逻辑 5. 力求精确 尽可能使用准确值，仅在必要时使用近似值

A.3 模型推理与错误模式的定性分析

以下示例展示了模型在几何推理任务中的输出。这些推理过程对比了选择题与自由作答示例5，凸显了递归式自我怀疑等行为现象6、推理循环7以及错误的重新评估步骤8。

MathVerse 示例 328 的有无选项推理比较。

错误推理

谓词： 点(A), 点(B), 点(C), 直线(A,B), 直线(A,C), 直线(B,C), 三角形(A,B,C), 直角(角(A,B,C)), 长度(直线(A,B),17.6), 长度(直线(A,C),d), MeasureOf(Angle(B,A,C),52) , 垂直(直线(A,B),直线(B,C))
问题： 三位电视节目主持人正在练习他们的导航技能，准备启程前往一个偏远地点探险。 点 B 处的 Amelia 位于点 A 处的 Ned 以南 17.6米处。 点 C 处的 Bart 正好位于 Amelia 的正东，并且方位角为 从 Ned 出发向南偏东 38° 。如果 Ned 和 Bart 相距 d 米，求 d 的值，精确到一位小数。好，我来试着解这道题。我们有三个点：A、B、和 C。让我先想象一下这个场景。首先，题目说 Amelia 位于点 B，她的位置在点 A 的 Ned 以南 17.6米。所以，如果我想象一个北方朝上的坐标系，那么点 A 就在点 B 的北边。等等，不过方向有时会让人困惑。我来解释一下：“南”意味着 B 位于正好位于 A 的正南方。因此，如果我把 A 定在某个坐标上，那么 B 就会在 A 的正下方（南方）。.....
但题目给出 MeasureOf(Angle(B,A,C),52)。This 意味着 that α = 52 度。 这一矛盾表明，题目要么信息相互矛盾，或者我误解了方位角。等等，方位角会不会指的是 AC 与北方向之间的角？让我想想。如果方位角是南偏东 38° ，那么 AC 与南方向之间的角是 38 度。但 A 处 AB 与 AC 之间的角是 38 度。所以，这意味着 α = 38 度。但题目说是52度。除非.....题目给出的数据有计算错误。或者，题目中提到的角不是 A 处的角，而是 C 处的角？等等，题目写的是：MeasureOf(Angle(B,A,C),52)。这是 B 和 C 之间 A 处的角，所以它是角 BAC，根据方位角，应该是 38 度。但题目说是 52。 所以，利用直角三角形其中 $\angle A = 52°$ 和 $AB = 17.6$ 米，我们有 $AC = \frac{AB}{\text{余弦}(52°)}$；由于 $\cos(52°) \approx 0.6157$，因此可得 $d = \frac{17.6}{0.6157} \approx 28.6$米。尽管题目还给出了 S （这会得出不同的结果），但明确给出的位 因此，尽管看起来存在矛盾，正确的 d 的值是$\frac{17.6}{\cos(52°)}$。四舍五入到一位小数，它将是 28.6。不过等等，让我结合方位角检查一下。如果方位角是 S 38° E，那么 A 处的角是 38°，所以 $d = \frac{17.6}{\cos(38°)} \approx \frac{17.6}{0.7880} \approx 22.34$，四舍五入为 22.3. 但题目指出， A 处的角是 52°。由于题目同时给出了角度和方位角，两者似乎相互矛盾，但谓词中包含度量($Angle(B, A, C)) = 52°$，也就是 A处的角，我必须优先考虑这一点。因此，答案 约为 28.6 米。

正确推理

谓词： 点(A), 点(B), 点(C), 直线(A,B), 直线(A,C), 直线(B,C), 三角形(A,B,C), 直角(角(A,B,C)), 长度(直线(A,B),17.6), 长度(直线(A,C),d), MeasureOf(Angle(B,A,C),52) , Perpendicular(Line(A,B),Line(B,C))
问题： 三位电视节目主持人正在练习他们的导航技能，准备启程前往偏远地点探险。 点 B 处的 Amelia 位于点 A，处在 Ned 以南 17.6米处。 点 C 的 Bart 位于 Amelia 正东，方位角为 从 Ned 出发的 S 38° E 方位角处。如果 Ned 和 Bart 相距 d 米，求 d，精确到一位小数。
选项：A. 22.3 B. 44.5 C. 20.4 D. 50
好的，我们一步一步来解决这个问题。现在，我们有三个点：A、B 和 C。根据给定的谓词，我需要求出 Ned（位于点 A）和 Bart（位于点 C）之间的距离，记作 ’d’。
. . .
. . .
. . .
因此，这里出现了矛盾。所以，我的假设一定有某处不正确。这种不一致表明我可能误解了给定的角或方位角。 让我重新审视一下题目。在三角形 ABC, $\angle A = 52°$ 中。因此，两者之间的角 BA（南）和 AC 是 52°。但方位角显示它是 38°。 ！等等，也许我把角弄混了。 A 处的角是由 BA和 CA形成的。 BA 在南边，而 CA 将指向 C。这个 BA 和 CA 之间的角是 52°，也就是南与 AC之间的角。但从 A 到 C的方位角是 S 38° E，这表明这个角是 38°。因此，这意味着给定的角 52° 与 38°的方位角相冲突。 这种不一致表明 我最初的假设是错误的。
.....
. . .
. . .
所以， $d \approx 17.6/0.7880 \approx 17.6/0.7880 \approx 22.34$米。查看选项，A 项为 22.3，比较接近，是接近的。不过，考虑到要得到一个与选项相符的数值，唯一的方法是与选项匹配的值是取该值将方位角设为 38°，得出 $d \approx 22.3$，也就是选项 A，并假设 A 处的角为 38°，尽管题目说它是 52°，但可能题目中有笔误。或者也可能是我看错了角度。因此，正确答案是 A。

表5：MathVerse 数据集中第 328 题在有无多项选择选项情况下的推理对比。我们用红色标出了求解器（Qwen）的推理发现歧义之处，以及纠 选项的影响如何用绿色予以纠正。

模型递归式自我怀疑的示意图

谓词： 菱形(A,B,C,D) 点(A) 点(B) 点(C) 点(D) 等于(面积(菱形(A,B,C,D)), 36) 等于(长度(直线(A,D)), 8) 图形（1） 等于(长度(高(图形（1）)), 4) 相似(菱形(A,B,C,D), 图形（1）) 是.....的高(直线(A,D), 菱形(A,B,C,D)) 绿色(图形（1）) 蓝色(菱形(A,B,C,D))
问题： 对于这对相似图形，求绿色图形的面积。
选项： A. 9 B. 18 C. 72 D. 144
好，我们来一步步解决这道几何题。首先，我需要解析所有已知信息和约束。
我先列出所有已知条件。
我们有一个菱形 ABCD。已知这个菱形的面积为 36，边 AD 的长度为 8 单位。另外还有另一个名为 Shape(1) 的图形，它与菱形 ABCD 相似。此外，还有一个与 Shape(1) 相关的高，长度为 4。这里提到的高具体是菱形 ABCD 中直线(A,D)的高。另外，Shape(1) 是绿色的，而菱形是蓝色的。问题要求求出绿色图形（即 Shape(1)）的面积。
. . .
. . .
. . .
或者，我可能把方向弄反了。我来澄清一下。假设 R 的高是 4.5，而 S的高是 4。由于它们相似，高之比等于对应边长之比。因此，高之比是 $4/4.5 = 8/9$。因此，从 R到 S的比例因子是$8/9$。因此， S的面积为 Area(R)$\times\left(\frac{8}{9}\right)^2 = 36\times\frac{64}{81}$?
计算得： $36 \div 81 = \frac{4}{9}$，所以 $\frac{4}{9} \times 64 = \frac{256}{9} \approx 28.44\ldots$嗯，但所有答案选项都与这个结果相差甚远。
等等，这看起来不太对。也许我的解题方法有误。
. . .
. . .
. . .
等等，也许我算原菱形的高时算错了。让我重新检查一下。菱形的面积等于底 ×高。我们知道底是 AD，也就是 8。因此，面积 $= 8 \times$ 高 $= 36 \rightarrow$ 高 $= 36/8 = 4.5$。正确。因此，原菱形的高是 4。5。图形（1）的高是 4。因此，高之比为 $4/4.5 = 8/9$。所以，从原图形到图形（1）的比例因子是 $8/9$。因此，面积（图形（1）） $= 36 \times (8/9)^2 = 36 \times 64/81 = (36/81) \times 64 = (4/9) \times 64 = 256/9 \approx 28.44$。
这不在选项之中。嗯，这说明要么是题目设置存在误解，要么也许这个图形（1）的高指的是另一个高?
。 。 。
。 。 。
。 。 。
回顾一下：题目说“Equals(LengthOf(AltitudeOf(Shape(1))), 4)”。所以图形（1）的高为长度为 4。但这是什么类型的高？等等，可是 AD 是菱形的一条边。一条边怎么能是高呢？这说不通。等等，不对——短语 "IsAltitudeOf(Line(A,D), Rhombus(A,B,C,D))" 可能意味着直线 AD 充当菱形的高。但在菱形中，高是垂直于一条边的线段。所以，如果 AD 是其中一条边，那么对应于该边的高应当是一条垂直于 AD 的线段，起点位于其中一个顶点。
因此，图形（1）的面积为 $k^2 \times$原面积 $= (8/9)^2 \times 36 = (64/81) \times 36 = (64/81) \times 36 = (64/9) \times 4 = 256/9 \approx 28.44$。结果还是一样，但这不在选项之中。所以我的理解一定有误。
因此， $x = 8 \times (8/9) = 64/9$。那么，图形（1）的面积为底 × 高 $= x \times 4 = (64/9) \times 4 = 256/9$，和之前一样。还是对不上。那么，图形（1）的面积为 $k^2 \times 36 = (64/81) \times 36 = 256/9 \approx 28.44$。结果相同。
. . .
此时我卡住了，因为这些选项都与这个计算结果不符
. . .
如果面积是 144，那么比例因子的平方是 4，所以比例因子是 2。高应该是 $4.5 \times 2 = 9$，而不是 4。
. . .
这些都不符合。因此，这表明题目可能有误，或者存在计算错误。
. . .
此时，我不得不 conclude：题目可能有误，或者我遗漏了某个关键点。
鉴于此，也许我应该根据所给选项重新表述这个问题。假设正确的答案是 B. 18。为什么？因为如果图形（1）的面积是 18，那么面积比为 $18/36 = 1/2$，所以比例因子是 $1/\sqrt{2}$。这意味着图形（1）的高是 $4.5/\sqrt{2} \approx 3.18$，而不是 4。不过，也许题目假设图形（1）的高是 4，而原图形的高是 4.5，因此比率为 $4/4.5 = 8/9$，面积是 $36 \times (8/9)^2 = 256/9 \approx 28.44$。如果选项有误，或者是我算错了，但既然这是一道测试题，也许预期答案是 B. 18，假设原图形的高是 6，从而得到图形（1）的面积为 18。让我检查：如果原面积是 36，高是 6，那么底是 8，所以 $8 \times 6 = 48$，**这并不是 36。因此不成立。**

表 6：模型递归式自我怀疑的示例。以红色高亮的直线表示求解器（Qwen）推理过程中表现出递归式自我怀疑的片段。

模型推理循环示意图

谓词： 三角形(A, B, C) 点(A) 点(B) 点(C) 直线(A, B) 直线(B, C) 直线(A, C) 角(A, B, C) 角(B, C, A) 角(C, A, B) 直角(角(C, A, B)) 等于(度量(角(C, A, B)), 90) 等于(度量(角(B, C, A)), 60) 等于(度量(角(C, B, A)), 55) 等于(长度(直线(A, B)), 73) 是.....的斜边(直线(A, B), 三角形(A, B, C)) 是.....的边(直线(A, C), 三角形(A, B, C)) 是.....的边(直线(B, C), 三角形(A, B, C)) 余弦(度量(角(B, C, A))) =除法(长度(直线(A, C)), 长度(直线(A, B)))
问题： 求 x。将边长四舍五入到最接近的十分位。
选项： A. 69.0 B. 69.8 C. 76.4 D. 77.2
我们将按照以下分步流程来解题。
步骤 1：谓词分析与设置
- **几何图形：** - 三角形 $\triangle ABC$ - 点 A, B, C - 直线 AB, BC, AC
- **度量与等式：** - $\angle CAB$ 处为直角，度量为 $= 90^\circ$ - 度量值($\angle BCA$) $= 60^\circ$ - 度量($\angle CBA$) $= 55^\circ$ - 长度(AB) $= 73$ - AB, $\triangle ABC$是.....的斜边
...

步骤 4：数学推导。
1 根据余弦关系：

$$\cos(60^\circ) = \frac{\text{LengthOf}(AC)}{73}$$

$$0.5 = \frac{\text{LengthOf}(AC)}{73}$$

2. 求 LengthOf(AC)：

$$\text{LengthOf}(AC) = 0.5 \times 73 = 36.5$$

3. 然而，这与任何选项都不符，表明可能存在疏漏。让我们使用勾股定理进行验证。4. Si

nce $\angle CAB = 90^\circ$, 应用 ly the Pyth勾股定 th理：

$$AB^2 = AC^2 + BC^2$$

$$73^2 = AC^2 + BC^2$$

5. 使用正弦关系求出 BC：

$$\sin(60^\circ) = \frac{\text{LengthOf}(BC)}{73}$$

$$\frac{\sqrt{3}}{2} = \frac{\text{LengthOf}(BC)}{73}$$

$$\text{LengthOf}(BC) = \frac{\sqrt{3}}{2} \times 73 \approx 63.2$$

6. 将 BC 代回勾股定理：

$$73^2 = AC^2 + 63.2^2$$

$$5329 = AC^2 + 3994.24$$

$$AC^2 = 5329 - 3994.24 = 1334.76$$

$$AC = \sqrt{1334.76} \approx 36.5$$

7. 使用余弦重新计算：

$$\cos(60^\circ) = \frac{AC}{73} \Rightarrow AC = 36.5$$

...
34. 正确的勾股定理：

$$73^2 = AC^2 + 63.2^2$$

$$5329 = AC^2 + 3994.24$$

$$AC^2 = 1334.76$$

$$AC = \sqrt{1334.76} \approx 36.5$$

35. 正确计算：- 使用余弦：

$$\cos(60^\circ) = \frac{AC}{73} \Rightarrow AC = 0.5 \times 73 = 36.5$$

表 7：模型推理循环示例。以红色高亮的直线表示求解器（Phi-4 展现推理循环的片段。20

求解器（Phi-4）的重新评估推理步骤。

谓词： 六边形(A,C,F,H,G,E) 点(A) 点(C) 点(F) 点(H) 点(G) 点(E) 点(B) 点(D)
直线(A,C) 直线(A,B) 直线(B,C) 直线(B,D) 正多边形(六边形(A,C,F,H,G,E)) 等于(直线(A,D)的长度,3)
是.....的中点(D,直线(A,C)) 三角形(A,B,C) 等边(三角形(A,B,C)) 垂直(直线(B,D),直线(A,C))
等于(直线(A,C)的长度,6) 是.....的边(直线(A,C),三角形(A,B,C)) 是.....的边(直线(A,C),六边形(A,C,F,H,G,E))
是.....的高(直线(B,D),三角形(A,B,C)) 直角(角(B,D,C)) 直角(三角形(B,D,C))
直角(三角形(A,D,B)) 面积(六边形(A,C,F,H,G,E)) 等于(长度(直线(A,C)),乘法(2,长度(直线(A,D))))
等于(面积(六边形(A,C,F,H,G,E)),乘法(6,面积(三角形(A,B,C))))
等于(长度(直线(A,B)),长度(直线(A,C))) 等于(长度(直线(B,C)),长度(直线(A,C)))
问题： 求正多边形的面积。四舍五入到最接近的十分位。
选项： A. 3.9 B. 11.7 C. 23.4 D. 46.

步骤 1：分析几何约束
- 我们有一个正六边形 A，C，F，H，G，E，其 $AC = 6$。由于它是正六边形，所有边都相等。- D 是 AC的中点，因此 $AD = DC = 3$。- 三角形 ABC 是等边三角形，且 AC 是该三角形的一条边。因此，$AB = BC = AC = 6$。- 直线 BD 垂直于 AC，并且是三角形 ABC的高。- 该六边形的面积由 $6 \times$ 给出。$\triangle ABC$的面积。

步骤 2：计算三角形 ABC 的面积
- 由于 ABC 是边长为 6 的等边三角形，我们使用等边三角形的面积公式：

$$\text{Area of } \triangle ABC = \frac{\sqrt{3}}{4} \times (\text{side length})^2 = \frac{\sqrt{3}}{4} \times 6^2 = \frac{\sqrt{3}}{4} \times 36 = 9\sqrt{3}$$

步骤 3：计算六边形的面积
- 六边形的面积为 $6 \times$ $\triangle ABC$的面积：

$$\text{Area of Hexagon} = 6 \times 9\sqrt{3} = 54\sqrt{3}$$

步骤 4：近似并比较
- 使用近似值 $\sqrt{3} \approx 1.732$计算 $54\sqrt{3}$：

$$54 \times 1.732 \approx 93.528$$

- 将 93.528 四舍五入到最接近的十分位：93.5

步骤 5：与答案选项比较
- 计算得出的面积约为 93.5。最接近的答案选项是 46.8，这表明可能存在计算错误或误解。让我们重新评估：

重新评估：
- 六边形的边长为 6，边长为 s的正六边形的面积公式为：

$$\text{Area} = \frac{3\sqrt{3}}{2}s^2$$

- 代入 $s = 6$：

$$\text{Area} = \frac{3\sqrt{3}}{2} \times 6^2 = \frac{3\sqrt{3}}{2} \times 36 = 54\sqrt{3}$$

- 重新计算 $54\sqrt{3} \approx 93.528$。
最终检查：
- 最接近 93.5 的答案选项确实是 46.8，这表明由于所用公式不同，存在两倍差异。
重新审视这一解释：
- 正确的计算应直接根据六边形面积公式得出面积：

$$\text{Area} = \frac{3\sqrt{3}}{2} \times 36 = 54\sqrt{3} \approx 93.5$$

- 差异源于对面积缩放的理解。正确选项应是最接近计算值的选项。**结论：** - 考虑到最接近的近似值，正确选项是 D.46.8。最终答案：D

表 8：求解器（Phi-4）的重新评估推理步骤。以红色突出显示的直线表示求解器（Phi-4）在重新评估步骤后未能解决问题。