

✨ 第一作者：新加坡国立大学

ThinkMorph

探索多模态模型中文本与图像交错推理的涌现属性

通过构建**24,990**个高质量交错推理轨迹数据集，我们发现统一模型能够自发产生三种关键的涌现能力：

未见过的视觉操控

自主模式切换

优化的测试时扩展

34.74%
平均性能提升

85.84%
空间导航准确率

24,990
高质量样本

1. 摘要

研究目的

论文旨在解决当前多模态推理模型中，文本与图像多为同构（isomorphic）映射，导致泛化能力不足的问题。研究者提出，为了实现可泛化的多模态推理，文本和图像应被视为互补（complementary）的模式，共同推进推理过程。

主要方法

论文引入了一个名为 **ThinkMorph** 的统一模型。该模型在一个包含约2.4万个高质量、跨四种不同视觉参与度任务的交错式推理轨迹数据集上进行微调。在这种方法中，模型能够生成文本和图像交错的“思维链”，其中文本推理和视觉操控携手并进，共同导向解决方案。

ThinkMorph 推理过程演示

重新播放

重置

What is the color of the man's cap?

A: yellow
B: red
C: gray ✓
D: green

<think>.....The **red bounding box** should be placed around the man's head and cap....</think>

<think>... now highlighting the man

真实案例 | Search

Gray!
✓ Answer: C

步骤 4/4 推理结论 · 答案输出

内部思考链

the man's cap is gray, so the correct answer is C.

过程说明

基于视觉分析完成推理，输出最终答案。

核心机制

- ✓ 文本推理 + 视觉操控交错进行
- ✓ 动态生成视觉辅助（边界框、放大）
- ✓ 多模态信息互补推理

理解问题

完成

理解文本问题并载入真实场景图

视觉操控

完成

在真实图像上生成红色边界框圈定目标人物。

细节观察

完成

放大帽子区域并分析颜色纹理。

答案输出

进行中

结合视觉证据输出最终答案。

✓ 推理完成!

ThinkMorph通过交错的文本-视觉推理链，成功识别出帽子颜色为灰色(gray)，选择答案C。这展示了模型如何将抽象的文本问题转化为具体的**视觉操控**（红色边界框），再通过**视觉分析**（放大聚焦）得出准确答案。

关键发现

1. 三种涌现属性

- ✓ 未见过的视觉操控 - 模型生成训练中未见过的视觉编辑（放大、修复）
- ✓ 自主模式切换 - 根据任务复杂性自适应在文本和交错推理间切换
- ✓ 优化的测试时扩展 - 通过多样化思考实现更稳健的性能

2. 交错推理的作用

交错式推理不仅是模态间的协调机制，更是激发高级智能行为的引擎。

实验成果

视觉中心任务

相比基座模型平均性能提升 **34.74%**，空间导航任务提升高达 **85.84%**

推理模式对比

交错式推理性能比纯文本或纯视觉方法平均高出 **5.33%**

域外基准表现

SAT空间推理 **52.67%** 超过InternVL-3.5-38B

MMVP感知任务

80.33% 准确率追平Gemini 2.5 Flash

2. 研究问题



核心研究问题

"什么构成了一个有意义的多模态交错思维链？"

探究在 **何种情况下** 以及通过 **何种机制**，文本与图像交错的推理方式能够超越单一模态推理方法。



当前局限



传统CoT的困境

纯文本思维链无法处理需要视觉内容审视和操控的复杂任务



现有方案的缺陷

依赖外部工具导致推理脆弱，或缺乏统一可泛化的解决方案



重要价值



推理范式跃迁

从"语言驱动推理"到"真正跨模态推理"的转变



智能进化关键

实现深度视觉理解与操作，迈向高级多模态智能



核心洞察

多模态推理应该像**人类解决复杂问题**一样：



不只是思考
纯文本推理



不只是观察
纯图像理解



认知闭环
思考↔观察

通过**文本规划**指导**视觉操作**，再通过**视觉反馈**优化**文本推理**，形成完整的认知闭环。

3. 方法论

3.1 交错思维链的泛化



核心突破：从单一模态到协同推理

传统的思维链(CoT)仅使用文本推理。ThinkMorph实现了文本与图像交错的协同推理——模拟人类“边想边画”(think-and-sketch)的能力。

案例1：通用流程演示 (基于 Figure 1)

Figure 1 展示了交错式训练数据的设计范式，这也揭示了 ThinkMorph 的工作流程。

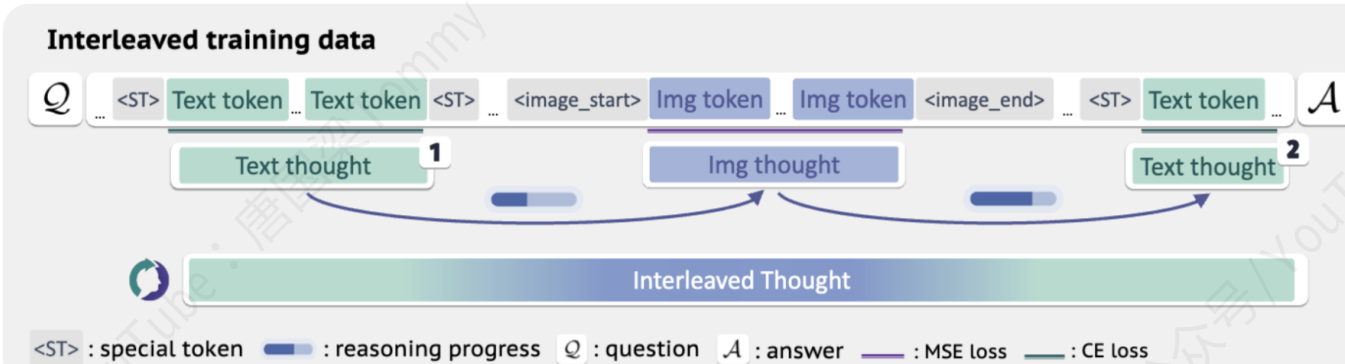


Figure 1: Design of Interleaved Training Data for Progressive Multimodal Reasoning

1 输入(Q)：一个多模态问题

2 第一步(Text thought t_1)：文本规划

模型首先生成文本进行初步分析、拆解或制定解决策略。

3 第二步(Image thought v_1)：视觉操作

模型在 <image_start> 和 <image_end> 之间生成图像。这张图像根据上一步的文本策略生成，对解决问题有帮助。

4 第三步(Text thought t_2)：验证与推理

模型观察自己生成的图像 v_1 ，基于新的视觉信息继续文本推理，最终得出答案。

核心流程： $T = (t_1, v_1, t_2)$ 清晰展示了文本和图像如何互补协作：文本负责规划和计算，图像负责定位和验证。

形式化定义

统一多模态模型

P_θ 可同时理解和生成文本与图像

多模态问题

$Q = (Q^{\text{text}}, Q^{\text{img}})$

思维步骤

$m_i \in \{t_i, v_i\}$ 文本或图像

交错式思维序列生成

$$T = (m_1, m_2, \dots, m_n)$$
$$m_i \sim P_\theta(m_i | x, m_0, m_1, \dots, m_{i-1})$$

每一步 m_i 基于问题和所有前序步骤自回归生成,可以是文本片段 t_i (逻辑推理)或图像片段 v_i (视觉思考)



实现机制

模型通过特殊分隔符token实现模态间无缝切换：

// 生成图像思维时
文本推理 → <image_start> → 图像token序列 → <image_end> → 继续文本

这使模型能在统一生成流程中清晰控制不同模态内容的转换



关键洞察

文本与图像并非同构映射，而是互补模态：文本负责抽象规划和逻辑计算，图像负责精确定位和视觉验证。两者交错协作才能实现可泛化的多模态推理。

3.2 三种涌现能力

"涌现"指一个系统的宏观层面表现出其微观组成部分所不具备的、全新的、不可预测的复杂属性。本章的核心论点是：交错式推理不仅是协调文本和图像的机制，更是一个能够催化产生真正多模态智能的引擎。

① 未见过的视觉操控

定义与重要性

模型在面对未见过的、域外的多模态任务时，能够主动地、准确地生成在训练数据中从未出现过的视觉操作。这表明模型不再是被动描述图像，而是能够主动地与视觉世界进行交互和干预。

核心机制：预训练能力的"激活"与"对齐"

1. 大规模预训练（基础）

Bagel经过海量图文交错数据预训练，为模型注入广泛的关联性知识，但这些能力是泛化的、未经提炼的"原材料"。

2. 交错式微调（对齐）

提供结构化的推理框架，教会模型"何时"以及"如何"调用这些视觉操作来服务于明确的逻辑链条。

论文案例剖析

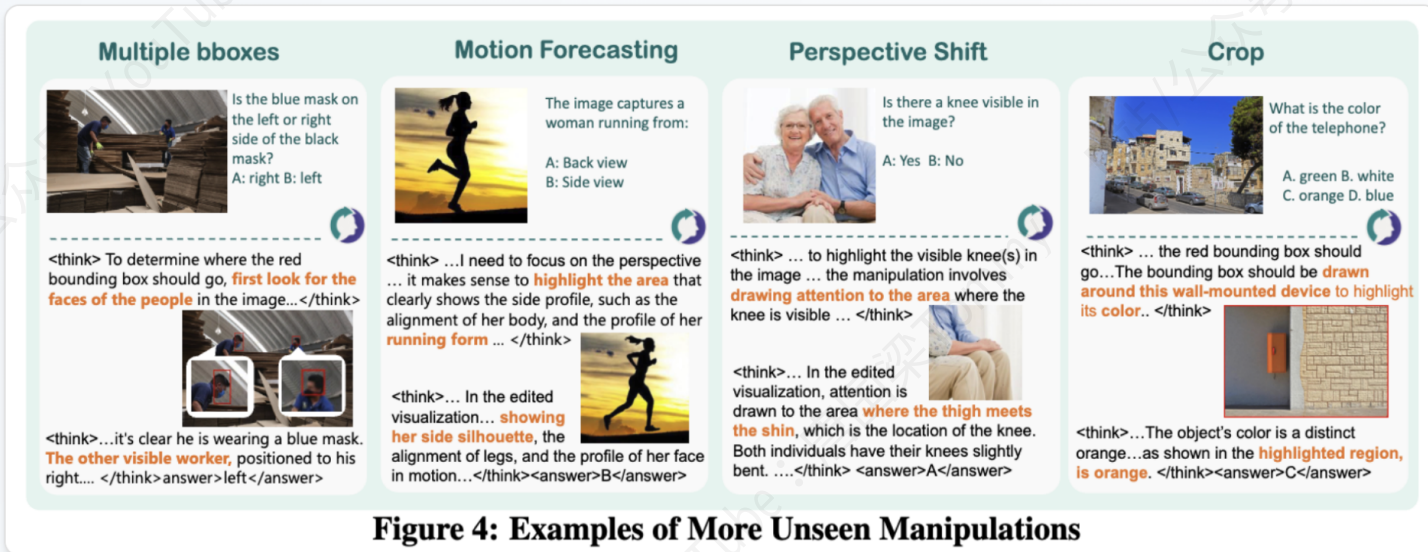


Figure 4: Examples of More Unseen Manipulations

Figure 4: 未见过的视觉操控示例

示例 1: Multiple bboxes (多个边界框)

任务描述：判断蓝色口罩在黑色口罩的左侧还是右侧。

视觉操控：模型生成包含两个边界框的可视化图像，分别标注两个工人的面部位置。

涌现特性：模型自主地同时绘制多个边界框来标注不同对象，这种"多目标同时标注"能力是对基础边界框绘制能力的创造性泛化。

示例 2: Motion Forecasting (运动预测)

任务描述：判断图像中的女子是从背面视角还是侧面视角拍摄的。

视觉操控：在可视化中突出显示运动中的侧面轮廓、腿部对齐和面部轮廓。

涌现特性：模型能够根据"运动姿态"这一动态概念，自主生成强调运动特征的视觉表示（如轮廓高亮），超越了静态物体识别。

示例 3: Perspective Shift (透视转换)

任务描述：判断图像中是否可以看到膝盖。

视觉操控：将注意力聚焦到大腿与小腿交汇处（膝盖位置），通过视觉高亮证实膝盖可见。

涌现特性：展现了局部区域的精准定位和注意力引导能力，能够自主对特定身体部位进行视觉标注和验证。

示例 4: Crop (裁剪)

任务描述：判断电话的颜色（绿色、白色、橙色还是蓝色）。

视觉操控：生成一个裁剪并放大了墙面电话区域的图像，清晰展示该物体的橙色特征。

涌现特性：模型自主执行了"裁剪+放大"的组合操作，模拟人类"凑近看"的行为，通过减少干扰信息、放大关键区域来确认细节。

关键洞察

预训练提供了"能做什么"的庞大能力库，而交错式微调则教会了模型"应该怎么做"的结构化方法。

② 自主模式切换

定义

尽管模型仅在图文交错的思维链数据上进行训练，但在实际推理时，它能根据具体问题的内在复杂性，自主地、动态地决定是采用完整的交错推理，还是"退化"为更高效的纯文本推理模式。

核心机制：隐式的"认知成本"评估

A. 信息充分性判断

如果编码后的视觉信息已经足够丰富，可以完全通过文本逻辑推理，则后续生成视觉思维链就成了信息冗余的步骤。

B. 效率与必要性权衡

生成图像在计算上昂贵。如果纯文本推理足以解决，模型选择"认知成本"更低的路径；否则坚持交错推理。

论文案例剖析

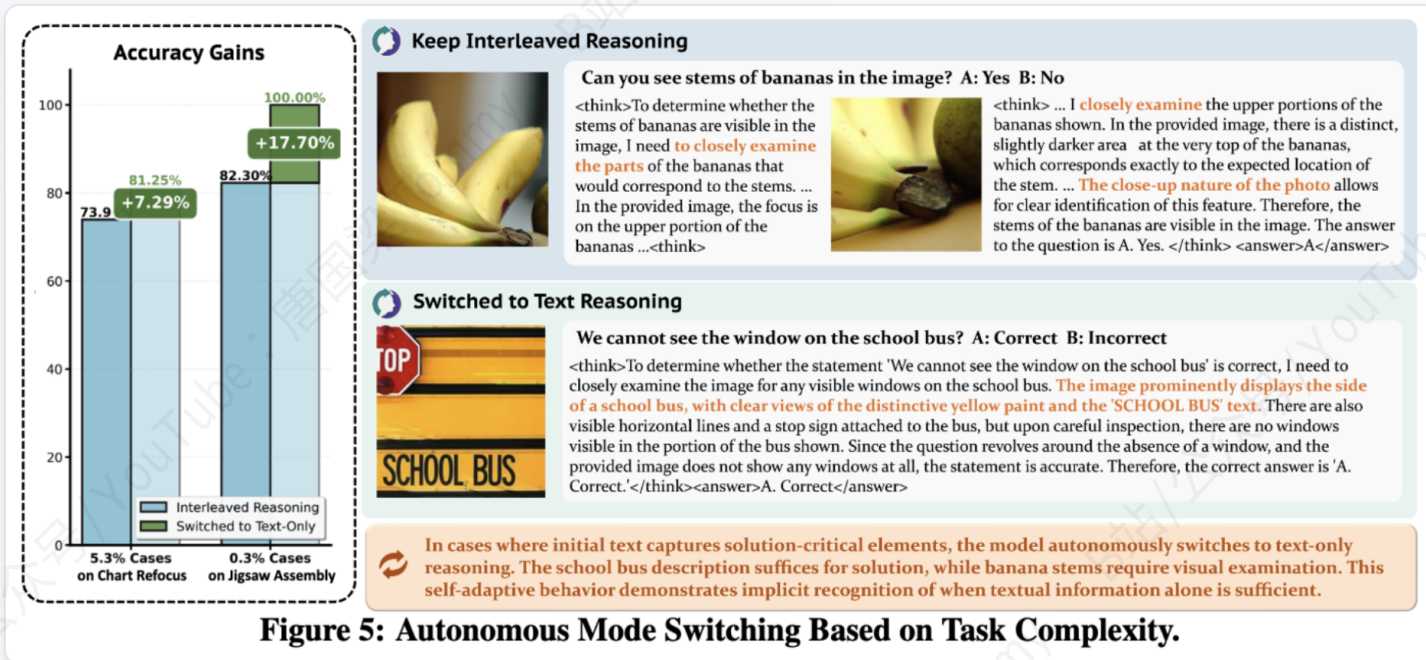


Figure 5: 自主模式切换示例对比

案例 1: 需要交错推理

问题:

"你能看到图中香蕉的茎吗?"

模型行为:

模型采用完整的交错推理模式，生成视觉思维链来验证。

深度分析:

"茎"的存在与否是一个需要精细视觉验证的局部特征。语言很难描述其所有的视觉细节来证明其存在。因此，模型必须通过生成一个放大的视觉思维链来提供确凿的证据。这种情况下，视觉验证是不可或缺的推理步骤。

关键洞察: 当任务涉及局部细节验证、空间关系判断等难以用语言精确描述的视觉特征时，模型会坚持使用交错推理，通过视觉操作提供直观证据。

案例 2: 自主切换为纯文本推理

问题:

"我们看不到校车的窗户?"

模型行为:

模型自主地切换到纯文本推理模式，跳过视觉思维链生成步骤。

深度分析:

这是一个关于"不存在"的判断。模型在对图像进行初步分析后，确认整个可见区域内均无窗户特征。这个结论可以通过全局描述直接得出（例如，"图中展示了校车的侧面，覆盖着黄色油漆和'SCHOOL BUS'字样，没有显示任何窗户"）。因此，再生成一张高亮"无窗户区域"的图像是多余的。模型自主地"剪枝"了视觉推理步骤。

关键洞察: 当文本描述已经足够完整、清晰地回答问题时，模型会智能地选择更高效的纯文本推理路径，避免不必要的计算开销。

实验数据

在MMVP数据集上，模型在5.3%案例中自动切换到纯文本模式，准确率**81.25%** > 交错推理**73.96%**，证明了切换决策的明智性。

③ 更优的测试时扩展

定义

在采用Best-of- N 采样（生成 N 个独立解答，选出最好的一个）时，交错式推理比纯文本或纯视觉推理表现出更强的性能提升和更好的扩展性。随着样本数 N 的增加，其性能优势愈发明显。

核心原理：更广阔的“解空间”探索

纯文本推理的局限

解空间局限于符号和逻辑层面，对视觉细节易陷困境。

纯视觉推理的局限

解空间局限于像素和特征层面，缺乏高级抽象和逻辑规划。

✦ 交错推理的优势

在文本（符号）空间和视觉（像素）空间同时探索。 N 个思想序列表现极大多样性，大幅提升正确答案的发现概率。

论文案例剖析

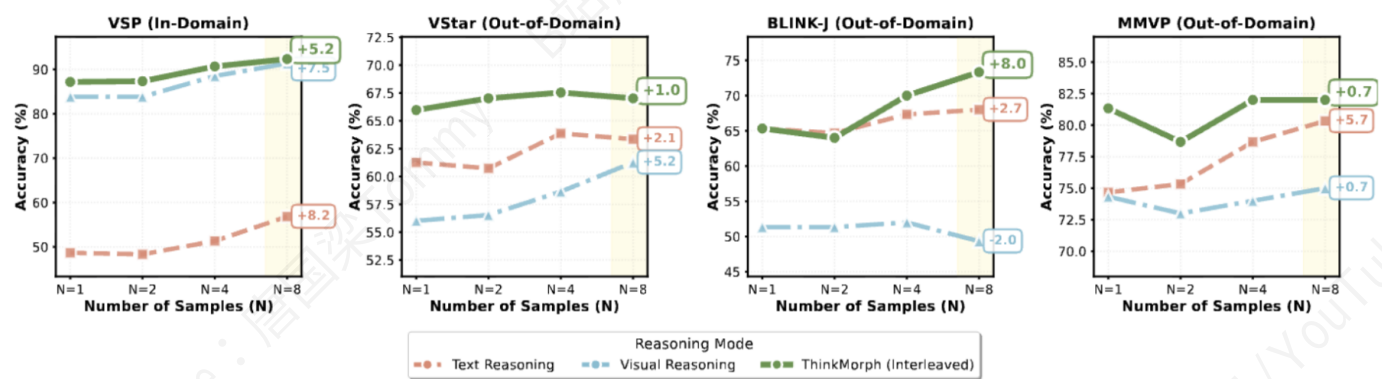


Figure 6: Test-Time Scaling Across Reasoning Modes. Interleaved reasoning demonstrates robust scaling advantages, particularly on challenging benchmarks where unimodal approaches plateau or decline.

Figure 6: 测试时扩展性能对比

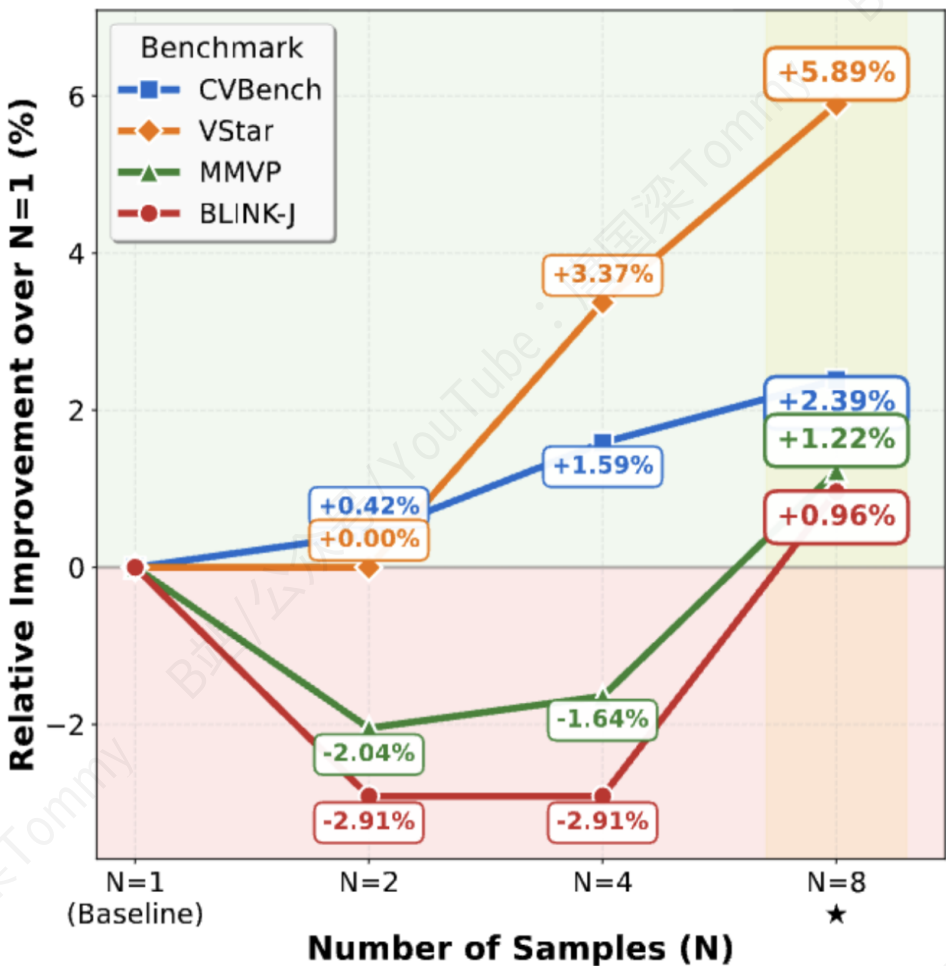


Figure 7: Relative Improvement

Figure 7: 轨迹多样性分析

图表解读

论文通过图表清晰地展示了交错推理在测试时扩展方面的优势。在最具挑战性的域外基准 BLINK-Jigsaw 上：

- 随着样本数 N 从 1 增加到 8，交错推理的性能稳定提升了 8.0%
- 相比之下，纯视觉推理的性能甚至出现了下降 (-2.0%)，表明其陷入了局部最优解
- 纯文本推理虽有少量提升 (+2.67%)，但远不及交错推理的增幅

这有力地证明了，当问题足够复杂时，交错推理的多样性优势能够充分发挥，通过在文本和视觉两个空间同时探索，大幅提升找到正确答案的概率。

3.3 数据集构建

数据规模

包含 24,990 个高质量交错推理样本

任务涵盖（四种不同视觉参与度）

Task	Data Source	Count	Visual Manipulation	Curation Steps
Jigsaw Assembly	SAT (Ray et al., 2024), ADE20K (Zhou et al., 2017), Omni3D (Brazil et al., 2023)	6,000	Visualizing re-arranged pieces	Newly generate questions from a customized pipeline
Spatial Navigation	N/A	6,000	Overlaying mazes with paths highlighted with red lines and arrows	Newly generated questions from a customized pipeline
Visual Search	Visual CoT (Shao et al., 2024), GQA (Hudson & Manning, 2019), VSR (Liu et al., 2023)	6,990	Highlighting Regions with Red Bounding Boxes	Filtering for valid (question, answer) with MLLMs + other criteria
Chart Refocus	ChartQA (Masry et al., 2022), Refocus (Fu et al., 2025)	6,000	Highlighting Regions with Red Bounding Boxes or Overlays	Filtering for valid (question, answer) with MLLMs + other criteria

Table 1: Summary of Questions Used for Training ThinkMorph.

Table 1: 数据集任务概览

拼图组装 (Jigsaw Assembly)

对打乱的图像块进行文本描述和视觉验证

数据源：SAT、ADE20K

空间导航 (Spatial Navigation)

通过文本规划和视觉绘制路径进行迷宫导航

视觉搜索 (Visual Search)

文本描述假设和视觉边界框精确定位

数据源：Visual CoT、GQA

图表重聚焦 (Chart Refocus)

文本识别数据元素和视觉高亮辅助计算

数据源：ChartQA、Refocus

4. 结果验证

↗ 量化成果总结

4.1 交错推理的优越性

	Spatial Navigation	Visual Search	Jigsaw Assembly		Chart Refocus	
	VSP	VStar★	VisPuzzle	BLINK-J★	ChartQA	MMVP★
Bagel-7B	0.83*	55.49	35.00*	67.33	62.05	70.33
Text Reasoning	49.17	56.02	63.50	68.67	81.66	76.33
Visual Reasoning	85.50	58.63	61.25	47.33	73.08	73.00
🌐 Interleaved Reasoning	86.67	63.87	73.75	73.33	79.78	82.66

Table 2: Reasoning Mode Comparison. Bagel-7B is tested under think mode (*: no-think mode for tasks where thinking prevents Bagel from generating answers). ChartQA results are the average performance on horizontal and vertical bar chart questions. ★: out-of-domain benchmarks. **Best**, second-best.

Table 2: 不同推理模式的性能对比

在以视觉为中心的任务上全面优于其他模式

空间导航 (VSP)
86.67%
vs 基座模型 0.83%

与基座模型对比
+34.74%
平均性能提升

在三个视觉中心任务上，交错推理模式比次优模式平均高出 **5.33%**

4.2 跨模型对比分析

	Size	VSP	VisPuzzle	ChartQA	VStar★	BLINK-J★	MMVP★	SAT★	BLINK★	CV-Bench★
Visual Understanding-only VLM										
GPT-4o	-	33.50	43.75	76.34	61.78	72.67	84.67	28.00	60.28	75.61
GPT-5	-	57.33	78.00	80.85	71.73	77.33	86.33	73.30	69.86	85.46
Gemini 2.5 Flash	-	59.33	47.00	83.79	70.68	66.00	80.33	56.00	67.49	85.07
InternVL3.5	8B	8.17	34.75	76.26	68.59	71.33	76.33	45.33	59.60	81.99
	38B	20.16	36.50	80.44	76.96	80.67	80.33	49.33	62.65	85.96
Qwen2.5-VL	7B	2.16	34.75	78.12	76.44	59.33	77.33	51.33	55.92	75.20
	72B	41.83	40.00	82.03	85.86	61.33	82.00	64.67	61.91	82.54
Unified Models										
Janus-pro	7B	00.00	33.50	43.08	38.22	50.67	63.33	22.00	38.51	67.83
Chameleon	7B	00.83	30.50	5.74	28.27	00.67	47.67	10.67	16.52	36.52
Bagel	7B	00.83*	35.00*	61.82	55.49	67.33	70.33	44.67	47.66	76.03*
ThinkMorph	7B	75.83	79.00	78.10	67.02	72.00	80.33	52.67	60.07	80.82
Δ (vs Bagel)		+75.00	+44.00	+16.28	+11.53	+4.67	+10.00	+8.00	+12.41	+4.79

Table 3: Comparison of ThinkMorph with Other Models. Bagel-7B is tested under think mode (*: no-think mode for tasks where thinking prevents Bagel from generating answers). ★: out-of-domain benchmarks.

Table 3: 跨模型性能对比

ThinkMorph (7B) 的性能可与规模远大于自身的模型媲美甚至超越：

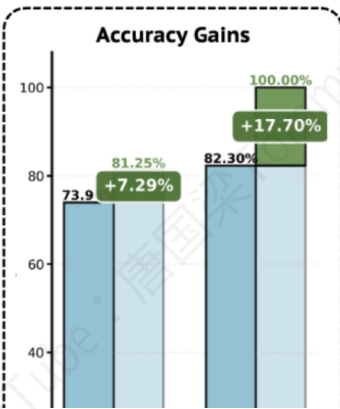
SAT空间推理任务

52.67%
vs InternVL-3.5-38B 49.33%

MMVP感知任务

80.33%
与 Gemini 2.5 Flash 相当

4.3 涌现属性的量化验证



Keep Interleaved Reasoning

Can you see stems of bananas in the image? A: Yes B: No

<think>To determine whether the stems of bananas are visible in the image, I need to **closely examine the parts** of the bananas that would correspond to the stems. ... In the provided image, the focus is on the upper portion of the bananas ...</think>

<think> ... I **closely examine** the upper portions of the bananas shown. In the provided image, there is a distinct, slightly darker area at the very top of the bananas, which corresponds exactly to the expected location of the stem. ... **The close-up nature of the photo** allows for clear identification of this feature. Therefore, the stems of the bananas are visible in the image. The answer to the question is A. Yes. </think> <answer>A</answer>

Switched to Text Reasoning

We cannot see the window on the school bus? A: Correct B: Incorrect

<think>To determine whether the statement 'We cannot see the window on the school bus' is correct, I need to closely examine the image for any visible windows on the school bus. **The image prominently displays the side of a school bus, with clear views of the distinctive yellow paint and the 'SCHOOL BUS' text.** There are also visible windows along the side of the bus, including a large window near the front. ...</think>

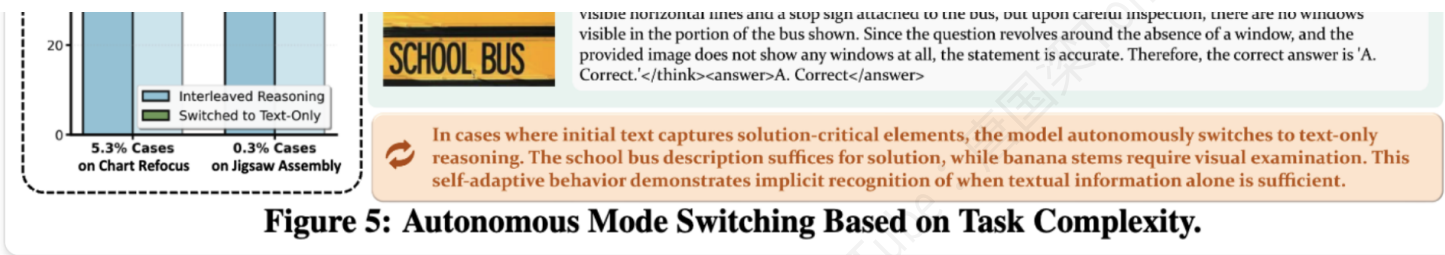


Figure 5: Autonomous Mode Switching Based on Task Complexity.

Figure 5: 自主模式切换验证

自主模式切换

- ✓ 在切换案例中准确率 **81.25%**
- ✓ 使用交错推理相同案例准确率 **73.96%**
- ✓ 性能增益 **+7.29%**

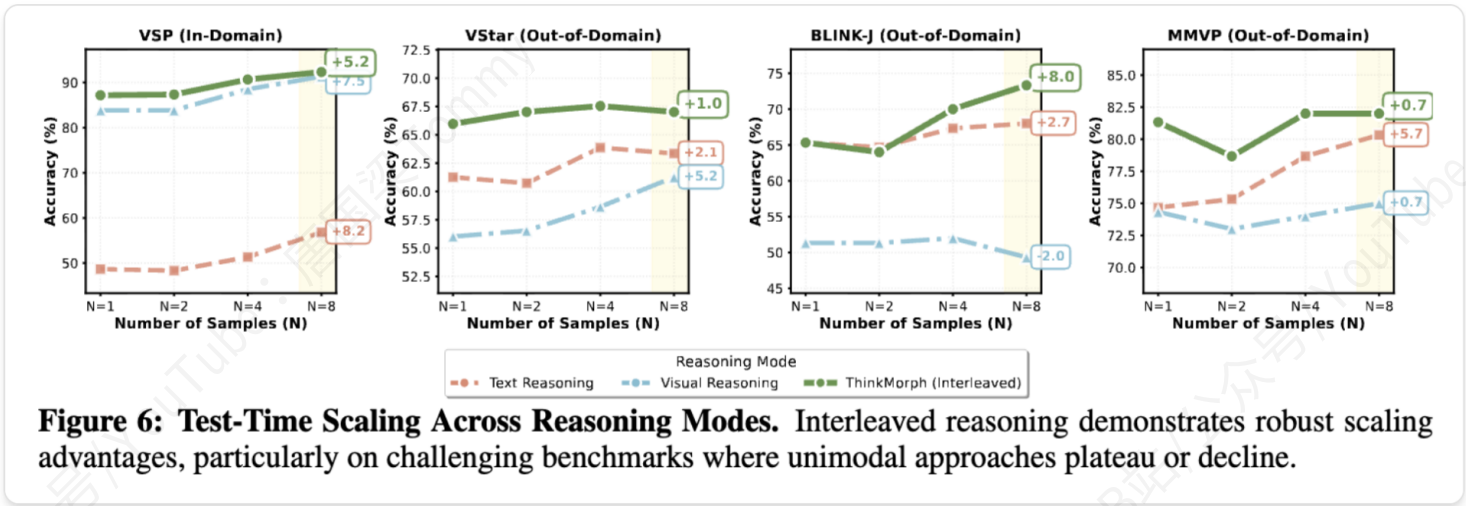


Figure 6: 测试时扩展验证

测试时扩展优势

- ✓ 基准：BLINK-Jigsaw (最具挑战性的域外基准)
- ✓ 交错推理性能增益 **+8.0%**
- ✓ 纯视觉推理性能 **-2.0%** (陷入局部最优)
- ✓ 证明多模态路径探索的鲁棒性

意外发现

自主学习模式切换

研究者仅用交错推理数据训练模型，但模型在推理时却能自主判断何时纯文本推理效率更高、效果更好，并主动切换。这表明模型不仅学会了“如何”进行交错推理，还隐式地学会了“何时”需要它，展现了更高层次的资源分配和策略选择能力。

5. 局限性

⚠ 作者明确提出的局限性

数据规模：适度规模微调

ThinkMorph仅有约2.4万个样本进行微调，与业界动辄数百万甚至上亿的训练数据相比规模较小。这可能会限制其在更广泛、更多样化场景中的泛化能力。

可推导的潜在局限性

5.1 任务泛化范围有限

研究成果主要基于四种特定类型任务（拼图、导航、搜索、图表）进行验证。模型是否能泛化到其他类型多模态任务（如视频理解、机器人指令执行或物理常识推理），仍是未知数。

5.2 推理成本问题

论文主要关注性能和准确率，但未讨论推理的计算成本。在思维链中生成图像token比生成文本token的开销大得多，这可能导致推理延迟显著增加，限制其在实时应用中的可行性。

5.3 对基座模型的依赖

ThinkMorph的成功建立在BageI-7B这个强大基座模型之上。其涌现特性在多大程度上源于交错微调策略，又在多大程度上源于基座模型本身的预训练能力，需要进一步剥离分析。将此方法应用于其他基座模型是否能复现同样效果仍不明确。

5.4 评估客观性

在部分任务评估中使用了GPT-5作为评判者。虽然这是常用方法，但评判结果会受到GPT-5自身的偏好、知识边界和潜在错误的影响，并非完全客观的"黄金标准"。

6. 结论

📌 主要结论

6.1 多模态推理能力的根本提升

通过将文本和图像视为互补模态的交错式思维链，可以显著提升多模态模型的推理能力和泛化性，尤其是在以视觉为中心的复杂任务上。交错式推理不仅是模态间的简单拼接，而是一个深度协同的推理机制。

6.2 三个涌现属性的发现

交错式推理能够催生出三种关键的涌现属性，这些属性是通向更高级多模态智能的标志：

- 未见过的视觉操控 - 主动生成创新的视觉操作
- 自主模式切换 - 智能判断何时切换推理模式
- 优化的测试时扩展 - 通过多样化探索获得更强鲁棒性

6.3 生效条件的揭示

该研究通过系统性分析，揭示了交错推理生效的条件和机制：当任务需要精确的视觉定位或操作来补充文本无法表达的信息时，交错推理至关重要；反之，当文本信息足以解决问题时，模型甚至可以智能地绕开它。这种自适应能力是走向真正智能系统的必要条件。

B站/公众号/YouTube : 唐国梁Tommy

<https://tgtommy.com>

© 2025 TGLTommy 原创出品 · All Rights Reserved