InfoQ 极客传媒

突破多模态大模型的效率瓶颈：结构、数据与训练优化

演讲人：余天予

清华大学 / 博士生

AiCon

全球人工智能开发与应用大会

AiCon
全球人工智能开发与应用大会

目录

极客邦科技 2026 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

InfoQ 极客传媒

01 多模态大模型

多模态大模型

- 传统语言大模型仅能处理文本模态信息
- 多模态大模型拓展大模型能力边界和应用场景，已成为人工智能前沿趋势和发展焦点

主要特点

基于大数据和人类反馈进行学习，与用户通过对话进行交互，能够处理多模态信息及多种任务，深层推理与常识运用能力大幅度提升

里程碑事件

2023/03/15: OpenAI 发布**多模态对话模型 GPT-4**
2023/12/06: Google 发布**多模态模型 Gemini**
2024/03/04 : Anthropic发布**多模态模型 Claude 3**
2024/05/13: OpenAI发布原生**多模态模型 GPT-4o**
2025/03/25: Google发布**多模态模型 Gemini 2.5**
2025/08/07: OpenAI发布**多模态模型 GPT-5**

多模态能力

支持多种模态建模，包括文本、图像、视频等

多模态大模型

- 多模态大模型的研究具有科学意义与实用价值
- 科学意义
  - 从多模态数据中学习为智能突破带来巨大潜力，是**智能跃迁的下一个关键引擎**
- 实用价值
  - 现实世界许多任务都需要**理解多模态输入**，例如具身智能、自动驾驶和视障群体辅助技术

Yann LeCun
图灵奖获得者

大部分的**人类知识**（以及几乎所有动物的知识） 都是通过**视觉、听觉、触觉、味觉**和**嗅觉**等**感官体验**，通过与物理世界的互动而获得的。

Ilya Sutskever
OpenAI前首席科学家

人类是多模态的动物。**没有多模态，神经网络的作用会远不及其上限**。通过多模态学习，人类可以更好地了解世界。

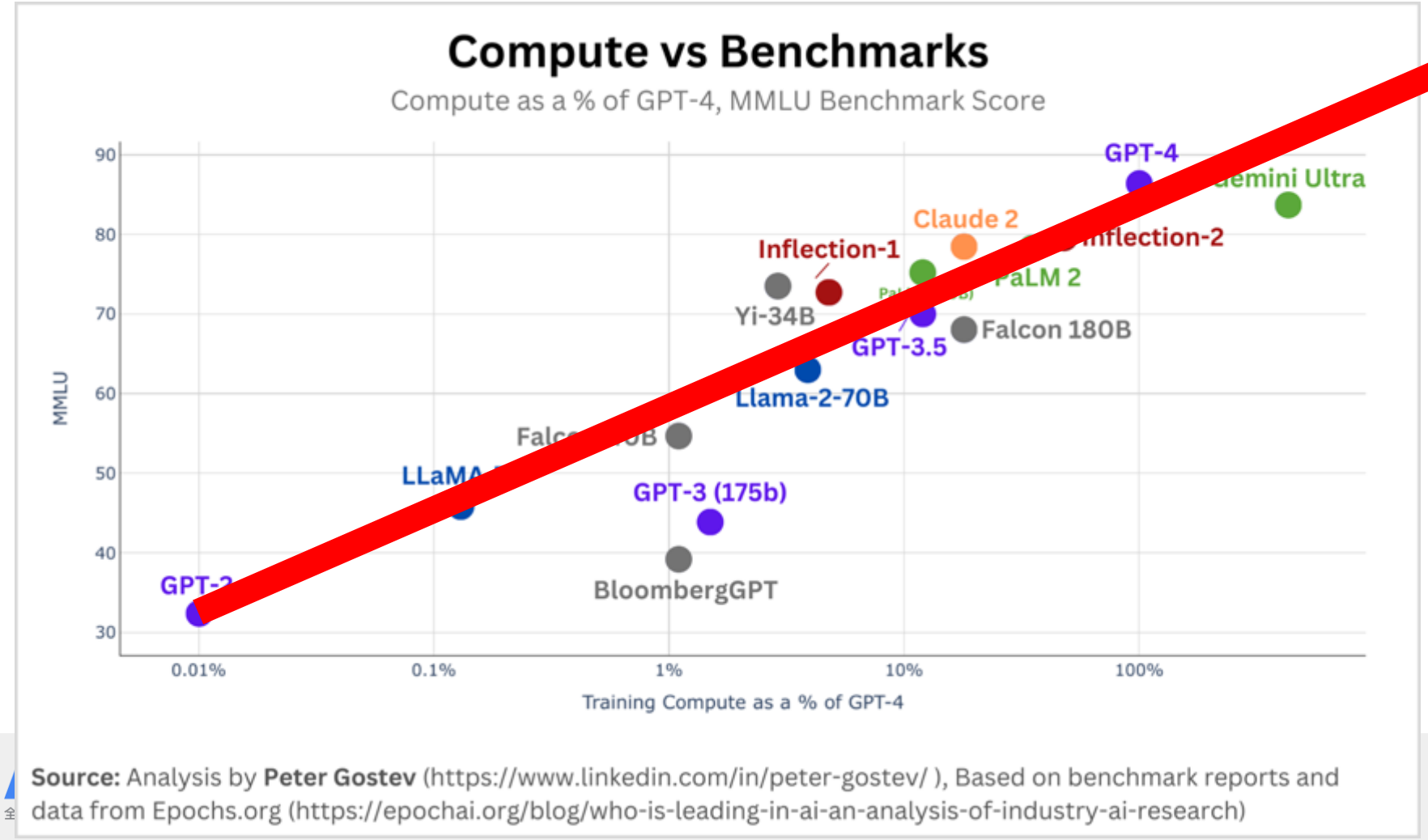
02

多模态大模型的效率瓶颈

多模态大模型的效率瓶颈

传统 Scaling Law：高资源低能效的粗犷式增长

- 更多智能 = 更多参数 + 更多数据 + 更多计算

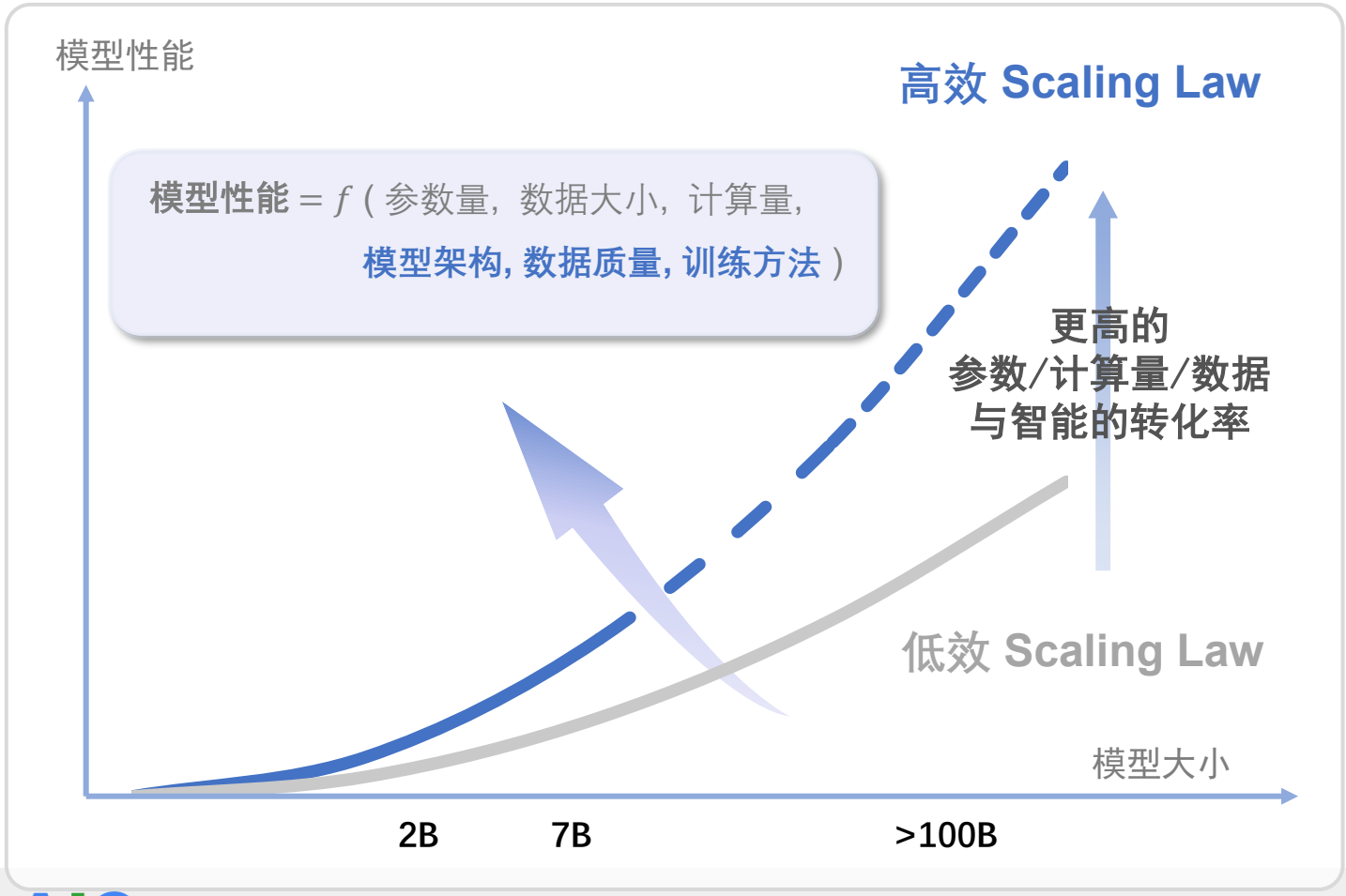


Source: Analysis by **Peter Gostev** (https://www.linkedin.com/in/peter-gostev/), Based on benchmark reports and data from Epochs.org (https://epochai.org/blog/who-is-leading-in-ai-an-analysis-of-industry-ai-research)

Scaling Law

更多的**智能**
=
更多的**参数**
+
更多的**数据**
+
更多的**计算**

高效 Scaling Law



	当前范式	我们的方案
模型	增加参数量	**高效率**结构
数据	增加数据量	**高质量**数据
训练	增加计算量	**高效率**训练

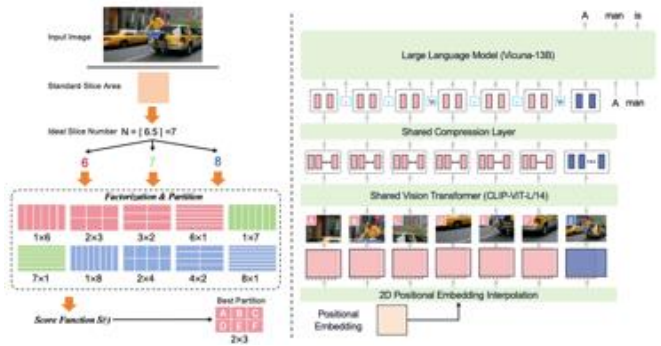
关键技术研究

高效模型结构

低清图，少细节

统一高分辨率视觉编码框架

- 支持原生长宽比
- 高效视觉 token 压缩
- 统一的单、多图、视频建模



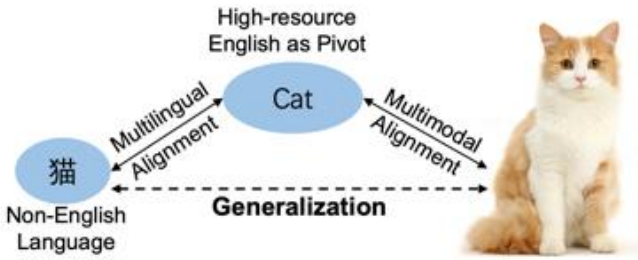
LLaVA-UHD: an LMM Perceiving Any Aspect Ratio and High-Resolution Images. ECCV 2024.

高效训练方法

英文强，中文弱

多语言多模态泛化

- 仅使用英文文本-图像数据进行预训练
- 中文跨语言多模态能力泛化



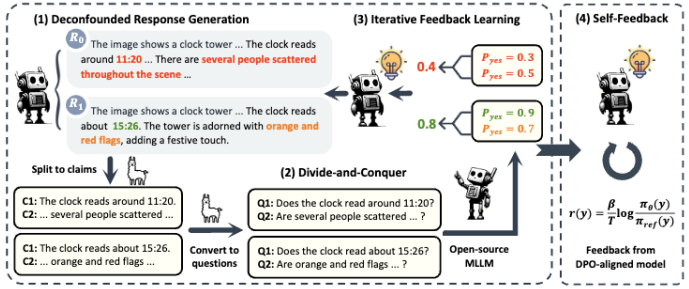
Large Multilingual Models Pivot Zero-Shot Multimodal Learning across Languages. ICLR 2024. Spotlight.

高质量数据构建

幻觉多，难置信

多模态反馈数据构建

- 通过细粒度的人类反馈 / AI自动反馈数据对齐模型行为
- 显著减少多模态幻觉



RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback. CVPR 2024.
RLAIF-V: Open-Source AI Feedback Leads to Super GPT-4V Trustworthiness. CVPR 2025. Highlights.

效率瓶颈：结构

多模态大模型的一个主要效率瓶颈就是巨大的视觉特征表征开销

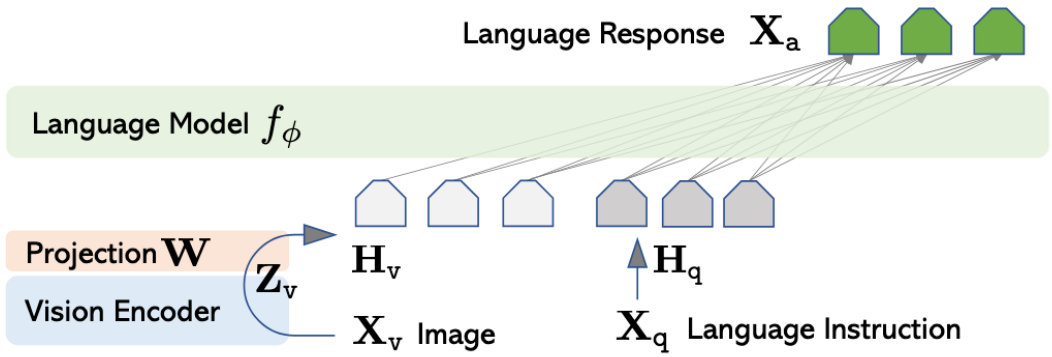
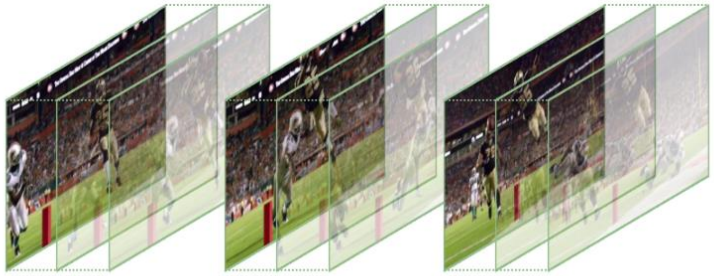


Figure 1: LLaVA network architecture.

主流视觉编码结构



448x448, 6s, 2fps → InternVL3: 3072 tokens
Qwen2.5 VL： 1536 tokens

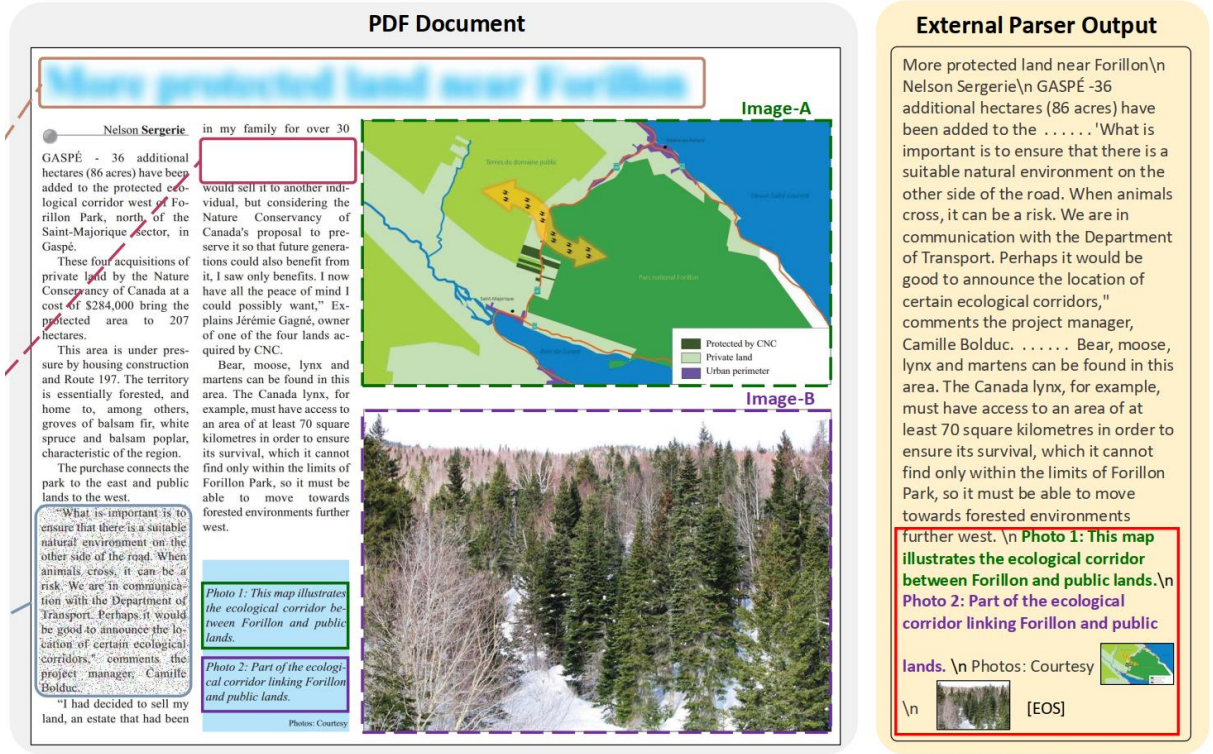
1920x1080, 6s,24fps → InternVL3: 380K tokens
Qwen2.5 VL： 190K tokens

1. Visual Instruction Tuning
2. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models
3. Qwen2.5-VL Technical Report

效率瓶颈：数据

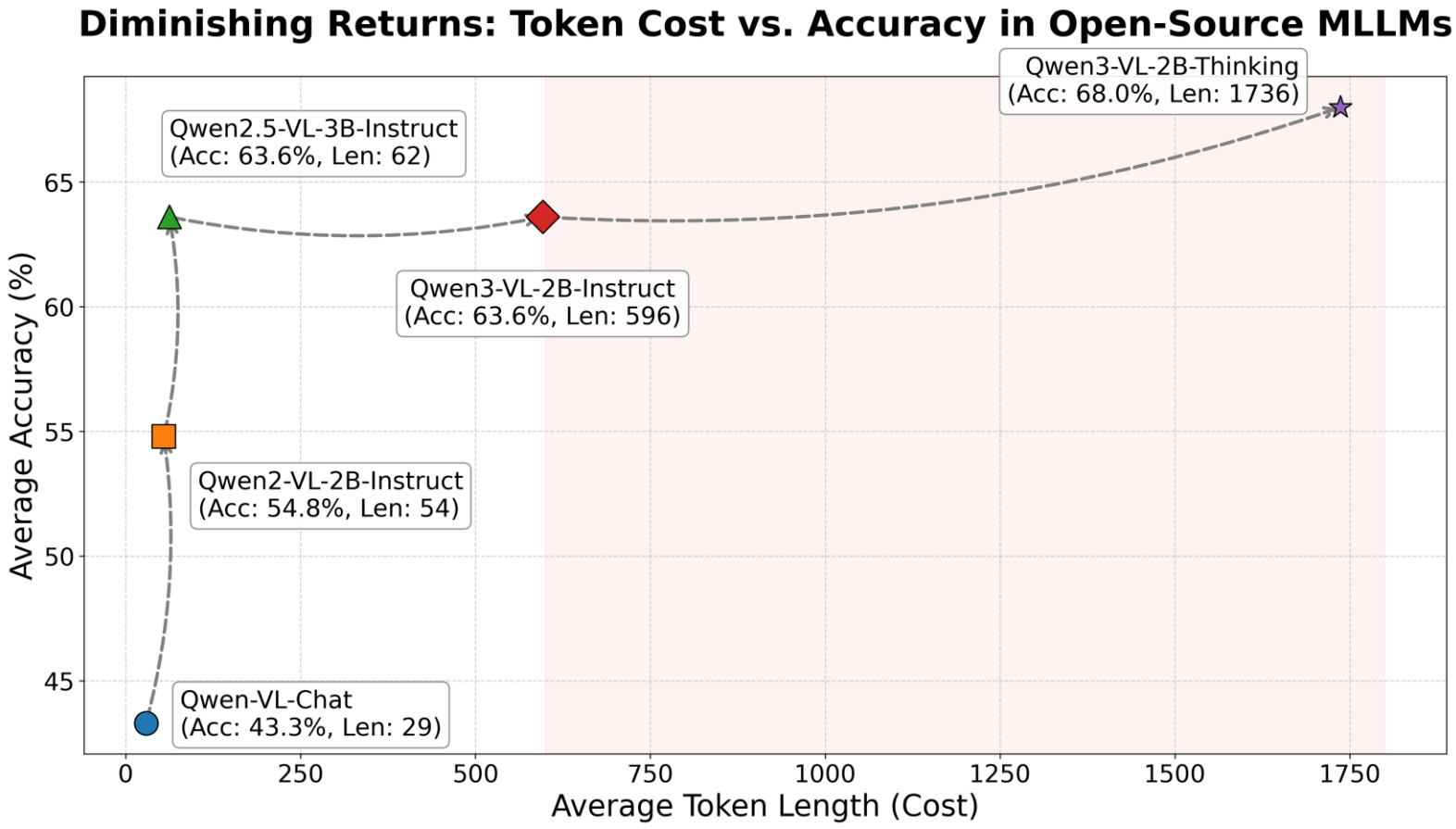
文档成为多模态大模型能力进一步增长的重要数据来源

但其数据处理和质量保障都带来了效率问题，广泛使用的图文解析工具频繁引入解析噪声



效率瓶颈：训练

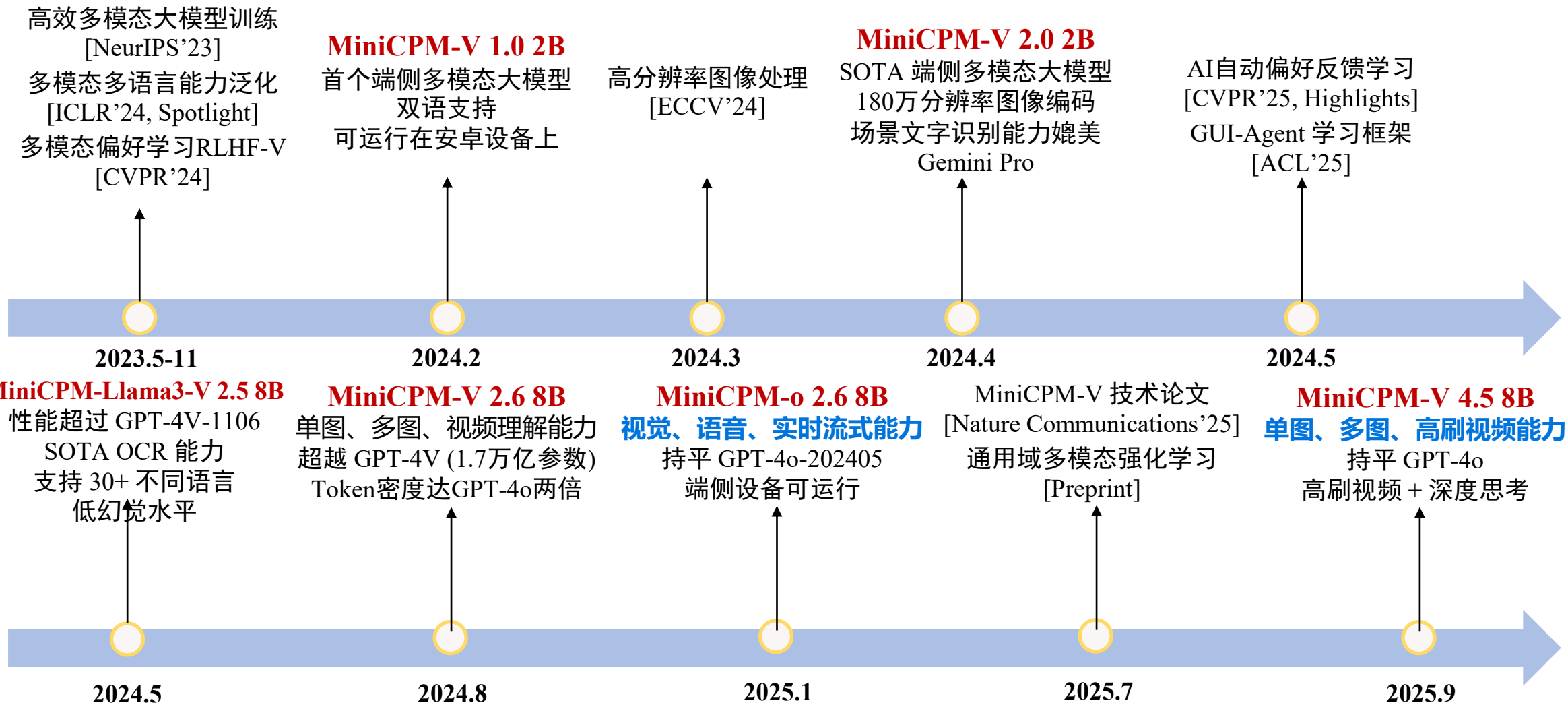
Token 的“通货膨胀”：深度推理范式提高了多模态大模型的推理能力，但于此同时显著增加了训练和推理过程的计算规模和时间开销。



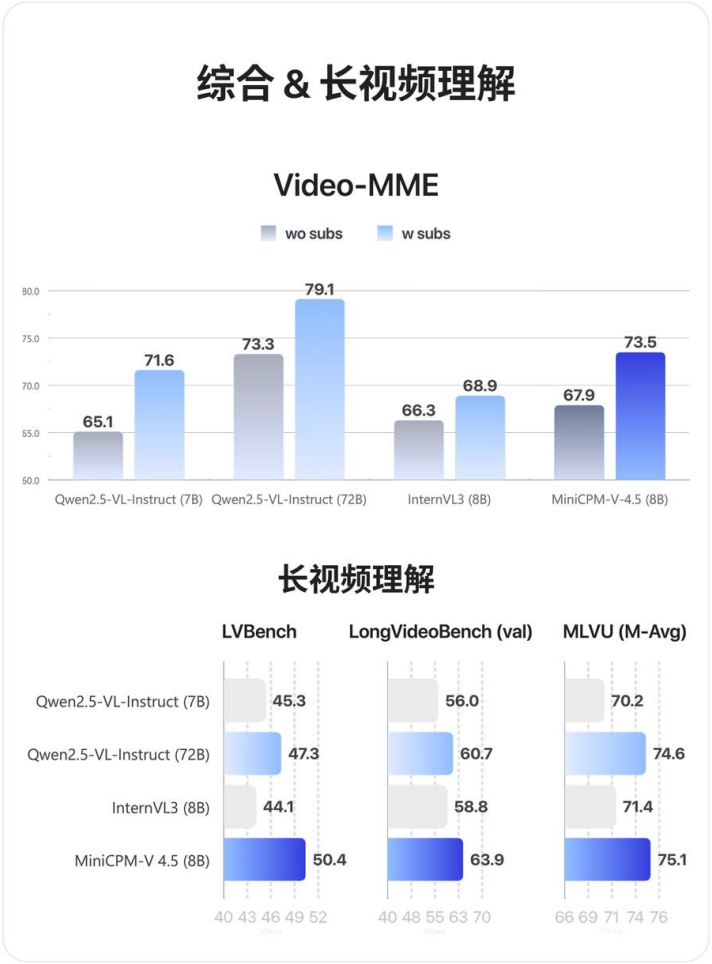
03

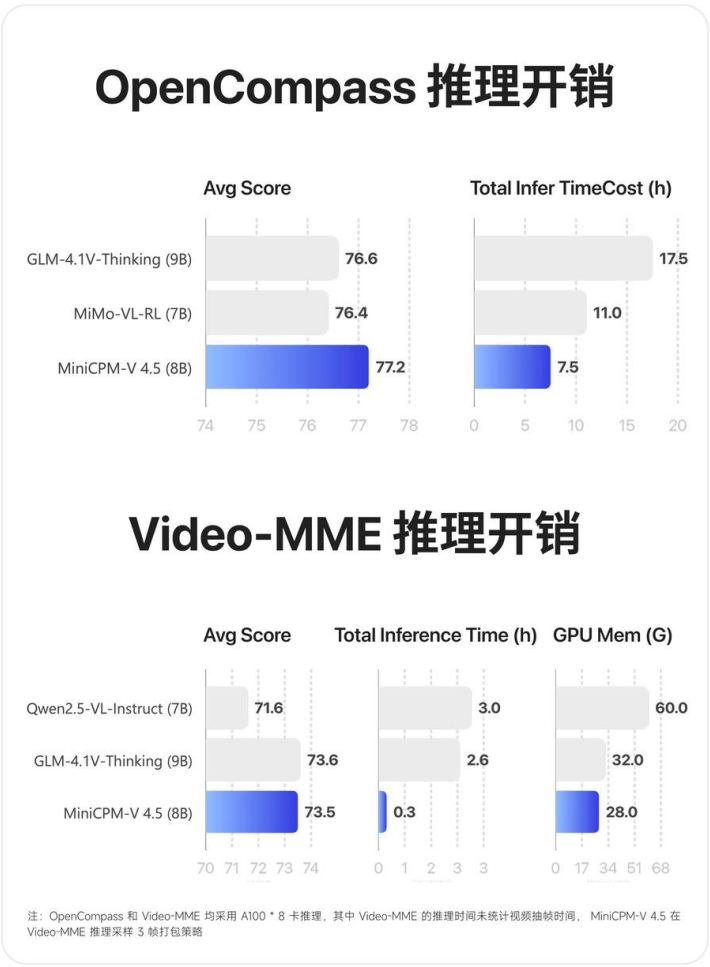
MiniCPM-V 4.5
高效多模态大模型

MiniCPM-V: 高效端侧多模态大模型



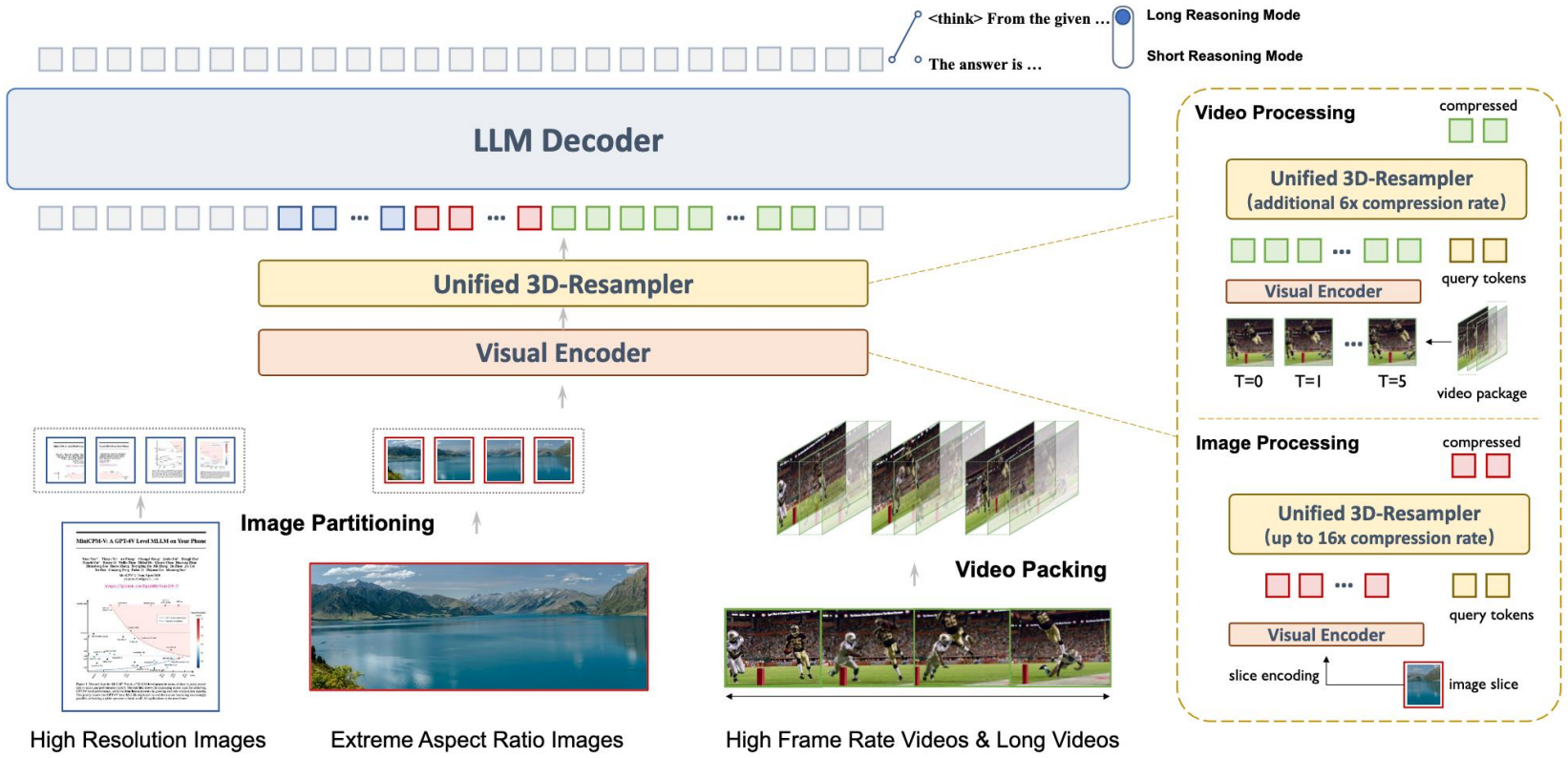
MiniCPM-V 4.5 高效多模态大模型

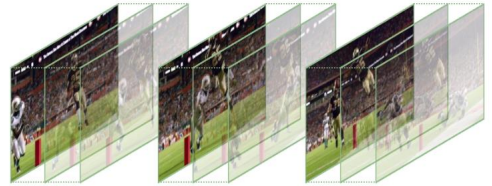




结构：高效统一视觉编码

3D-Resampler 结构：时空间统一压缩，显著降低图像、视频的视觉编码开销达 4~96 倍





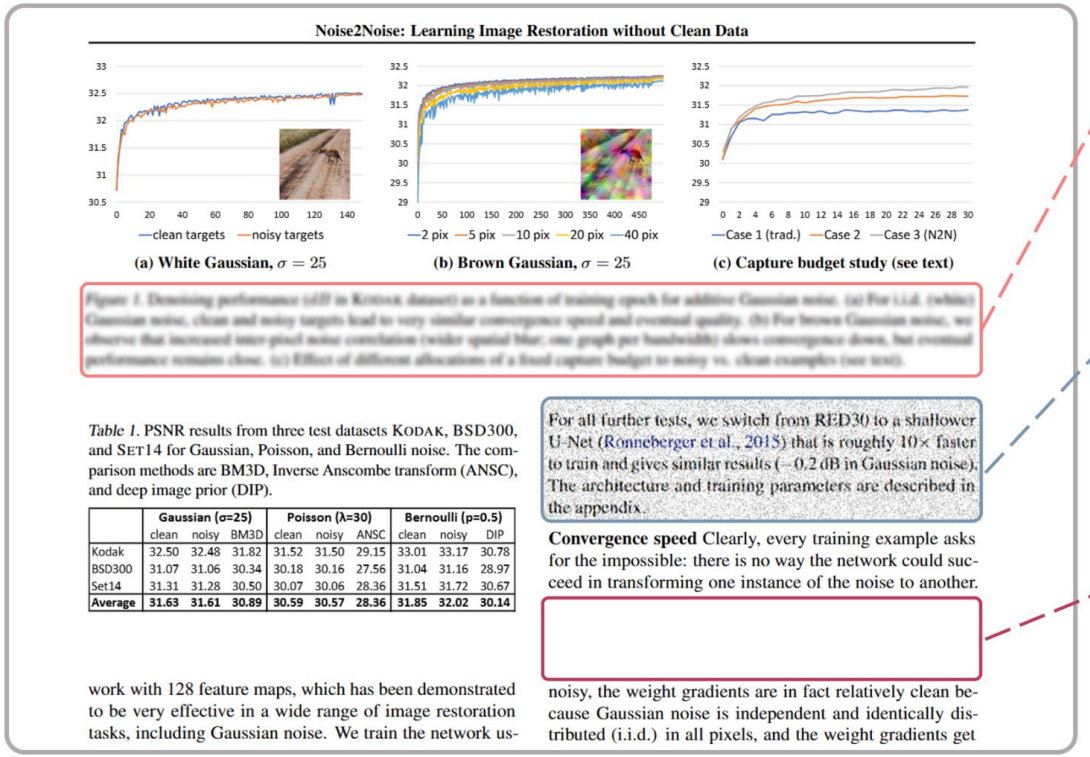
448x448, 6s, 2fps

1536 tokens
Qwen2.5 VL

128 tokens
MiniCPM-V 4.5

数据：高效文档知识统一学习格式

通过利用简单的动态视觉遮蔽，多模态文档图像可以直接用于OCR、文档知识学习等多种任务的训练，避免复杂解析器引入的数据错误问题



训练：高效混合思考后训练

在 RL 过程中混合长思考模式和短思考模式采样策略，在显著降低训练开销的同时进一步提升了模型的回答效率

Method	OpenCompass	Training Tokens
Short reasoning only	76.0	**1.6B**
Long reasoning only	77.0	4.4B
Hybrid	**77.1**	3.1B

Table 3: Ablation of hybrid reinforcement learning. We report training token cost and performance on OpenCompass.

Model	Size	Avg Score ↑	Time ↓
GLM-4.1V-9B-thinking	10.3B	76.6	17.5h
MiMo-VL-7B-RL	8.3B	76.4	11.0h
MiniCPM-V 4.5	8.7B	**77.0**	**7.5h**

极客邦科技 2026 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

InfoQ 极客传媒

InfoQ 极客传媒

THANKS

探索 AI 应用边界

Explore the limits of AI applications

AiCon

全球人工智能开发与应用大会