

Artificial Intelligence Newsletter

人工智能 前沿快报



2026 年第 04 期
总第 31 期



学会官方公众号

广东省图象图形学会

GDSIG

学术交流、技术咨询、科学普及

目录

学术动态

一. Mixture-of-Depths Attention.....	1
二. MetaClaw: Just Talk -- An Agent That Meta-Learns and Evolves in the Wild.....	3
三. Project Imaging-X: A Survey of 1000+ Open-Access Medical Imaging Datasets for Foundation Model Development.....	5
四. CGEMS: Agent-Native Multimodal Generation with Memory and Skills.....	7
五. ASI-Evolve: AI Accelerates AI.....	9
六. The Latent Space: Foundation, Evolution, Mechanism, Ability, and Outlook.....	11
七. ClawArena: Benchmarking AI Agents in Evolving Information Environments.....	13
八. PaperOrchestra: A Multi-Agent Framework for Automated AI Research Paper Writing.....	15
九. MinerU2.5-Pro: Pushing the Limits of Data-Centric Document Parsing at Scale.....	17
十. In-Place Test-Time Training.....	20
十一. AutoSOTA: An End-to-End Automated Research System for State-of-the-Art AI Model Discovery.....	22
十二. EpiAgent: An Agent-Centric System for Ancient Inscription Restoration.....	24
十三. Vision Transformers Need More Than Registers.....	26
十四. Latent-Condensed Transformer for Efficient Long Context Modeling.....	28
十五. Dive into Claude Code: The Design Space of Today' s and Future AI Agent Systems.....	30
十六. LLaDA2.0-Uni: Unifying Multimodal Understanding and Generation with Diffusion Large Language Model.....	32
十七. Agentic World Modeling: Foundations, Capabilities, Laws, and Beyond.....	34
十八. Scaling Behaviors of LLM Reinforcement Learning Post-Training: An Empirical Study in Mathematical Reasoning.....	36
十九. Image Generators are Generalist Vision Learners.....	38

目录

技术动态

一. Qwen3.5-Omni: 新一代原生全模态大模型	40
二. Qwen3.6-Plus: 性能直逼 Claude, 国产编程模型新高度	41
三. Google Gemma 4 系列: 开源多模态旗舰, 31B 登顶开源前三	42
四. GLM-5.1: Towards Long-Horizon Tasks	43
五. Claude Managed Agents	44
六. Nature 解读 Stanford 2026 AI Index: 复杂任务上人类科学家仍领先 AI Agent	45
七. GPT-Rosalind: 面向生命科学研究的专用推理模型	46
八. ChatGPT Images 2.0 发布	47
九. Introducing GPT-5.5	48
十. DeepSeek V4 双版本正式上线: 1.6T MoE, 1M 上下文	49
十一. 小米开源 MiMo-V2.5-Pro / MiMo-V2.5 系列	50

资源分享

一. Claude Code & Harness 电子书	51
二. Stanford 2026 AI Index Report	52
三. 《计算机科学与机器学习中的代数、拓扑学、微分学和优化理论》电子书	53
四. Project-Imaging-X	54
五. Awesome Latent Space	55
六. Deep Agents Deploy: Claude Managed Agents 的开源替代方案	56
七. Awesome Agentic World Modeling	57

INTRODUCTION

导读

在人工智能技术飞速发展的时代背景下，我们将不定期梳理人工智能领域的部分前沿学术与技术动态，并以简报的形式分享给读者，以帮助关注此领域的人士了解最新的学术前沿发展与产业技术动态，促进学术交流和技術繁荣。本期摘选了近期全球人工智能领域的 19 篇最新学术动态、11 篇技术动态、以及 7 个资源推荐给广大读者。

ACADEMIC TRENDS

学术动态

● 论文一

Mixture-of-Depths Attention

日期

2026-03-16

 ByteDance | Seed

Mixture-of-Depths Attention

Lianghui Zhu^{1,2,*}, Yuxin Fang^{2,†}, Bencheng Liao^{1,2,*}, Shijie Wang²,
Tianheng Cheng², Zilong Huang², Chen Chen², Lai Wei², Yutao Zeng², Ya Wang²,
Yi Lin², Yu Li², Xinggang Wang^{1,✉}

¹School of EIC, Huazhong University of Science & Technology, ²ByteDance Seed

[†]Project Lead, [✉]Corresponding author

内容摘要

提出 MoDA (Mixture-of-Depths Attention) , 针对 Transformer 不同 token 对注意力计算贡献差异悬殊的问题, 设计动态路由机制, 将 token 按重要性分配到不同深度的计算路径, 从而在保持模型容量的同时显著降低注意力层的整体计算冗余。配套提供硬件感知的高效实现 (MoDA) , 可直接迁移至 ViT 等视觉架构。

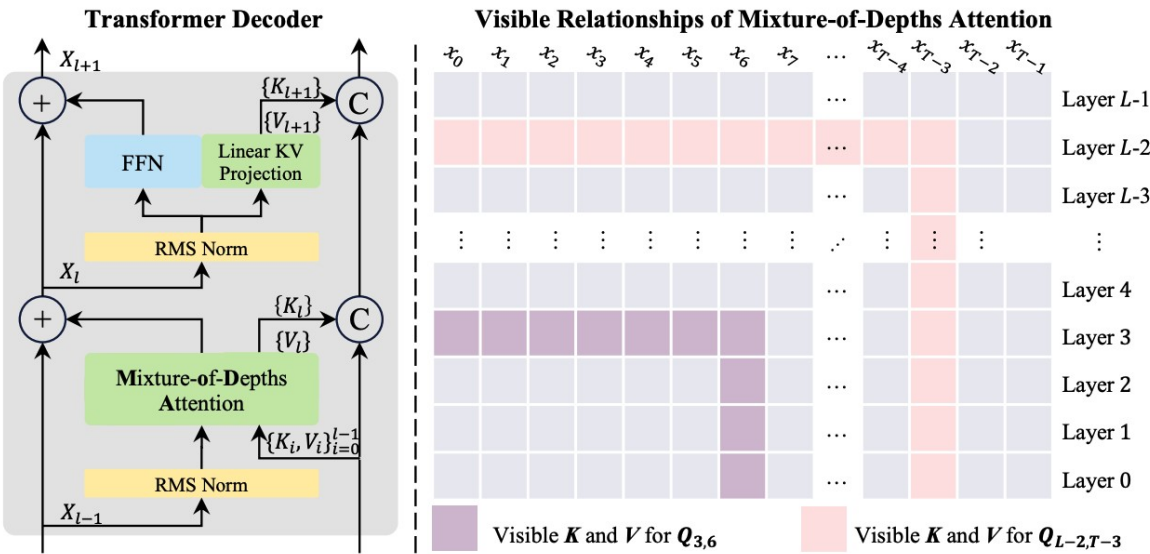


Figure 1 We propose mixture-of-depths attention (MoDA) to address the modern LLM’s information dilution problem in a dynamic and hardware-efficient way. Compared with vanilla causal sequence attention, MoDA additionally allows query to attend to depth memories, i.e., depth KV pairs $\{K_i, V_i\}_{i=0}^{l-1}$ at the same query position from preceding layers.

主要亮点

- **动态深度路由:** 不同 token 根据重要性被路由至不同深度的计算路径, 跳过不必要的注意力层计算
- **硬件感知实现:** MoDA 在 GPU 上保持高吞吐量的同时降低整体计算量, 适合长上下文和大批量推理
- **架构迁移友好:** 提供完整代码仓库, 可复现并迁移到 ViT 等视觉 Transformer 结构

相关链接

Paper: <https://arxiv.org/abs/2603.15619>
Github: <https://github.com/hustvl/MoDA>

● 论文二

MetaClaw: Just Talk -- An Agent That Meta-Learns and Evolves in the Wild

日期

2026-03-17



MetaClaw: Just Talk – An Agent That Meta-Learns and Evolves in the Wild

Peng Xia^{1*}, Jianwen Chen^{1*}, Xinyu Yang^{2*}, Haoqin Tu^{3*}, Jiaqi Liu^{1*}, Kaiwen Xiong^{1*}, Siwei Han¹, Shi Qiu¹, Haonian Ji¹, Yuyin Zhou³, Zeyu Zheng⁴, Cihang Xie³, Huaxiu Yao^{1*}

¹ UNC-Chapel Hill, ² Carnegie Mellon University, ³ UC Santa Cruz, ⁴ UC Berkeley,
*Core Contributors

内容摘要

MetaClaw 提出持续元学习框架，让部署中的 LLM Agent 在不中断服务的前提下持续进化。两大核心机制：

- ① Skill-driven fast adaptation——从失败轨迹自动合成技能指令，注入 prompt 实现梯度无关的即时适应；
- ② Opportunistic policy optimization——通过 OMLS 调度器在用户不活跃时段(睡眠、会议等)触发 LoRA 微调。在 MetaClaw-Bench 和 AutoResearchClaw 上验证，完整流程将 Kimi-K2.5 准确率从 21.4% 提升至 40.6%。

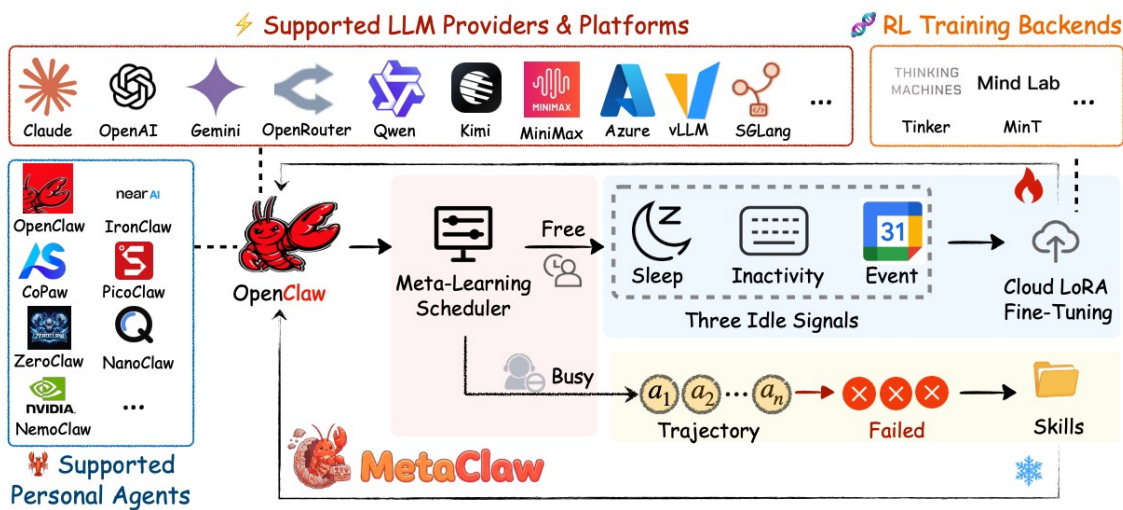


Figure 1 Overview of **MetaClaw**. The framework improves the meta-model $\mathcal{M} = (\theta, \mathcal{S})$ via two complementary loops operating at different timescales. Skill-driven fast adaptation (left) analyzes failed trajectories and instantly expands the skill library $\mathcal{S}$ without parameter updates, taking effect immediately for subsequent tasks. Opportunistic policy optimization (right) accumulates post-adaptation trajectories and, once sufficient data is available, leverages idle signals (sleep, inactivity, calendar) detected by the Opportunistic Meta-Learning Scheduler to trigger RL-based weight updates on θ via Cloud LoRA fine-tuning.

主要亮点

- **技能驱动快速适应：** 失败轨迹触发 LLM evolver 合成新技能指令并立即注入 prompt，零停机、无梯度
- **机会式策略优化：** OMLS 调度器利用用户不活跃窗口执行 LoRA 微调，策略更新不阻塞在线服务
- **实测效果显著：** 完整双轨机制将 Kimi-K2.5 在相关任务上的准确率从 21.4% 提升至 40.6%

相关链接

Paper: <https://arxiv.org/abs/2603.17187>

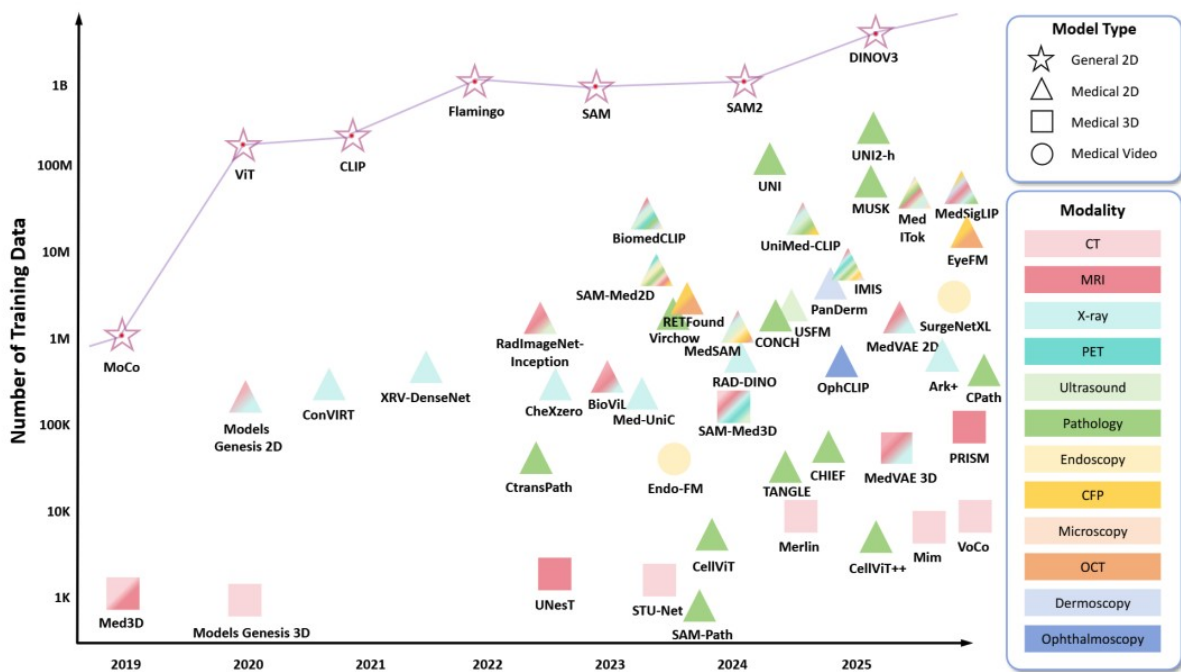
Github: <https://github.com/aiming-lab/MetaClaw>

● 论文三

Project Imaging-X: A Survey of 1000+ Open-Access Medical Imaging Datasets for Foundation Model Development

日期

2026-03-29



内容摘要

该综述系统梳理 1000+ 个医学影像开放数据集，从器官、模态、任务、标注类型四个维度建立分类框架。核心贡献是提出 Metadata-Driven Fusion Paradigm (MDFP)，将同模态或同任务的小数据孤岛整合为更大规模的连贯资源，解决医学影像数据集碎片化、规模偏小、分布不均等阻碍基础模型训练的关键问题。配套 Project-Imaging-X 交互平台支持按需检索和自动化数据集集成。

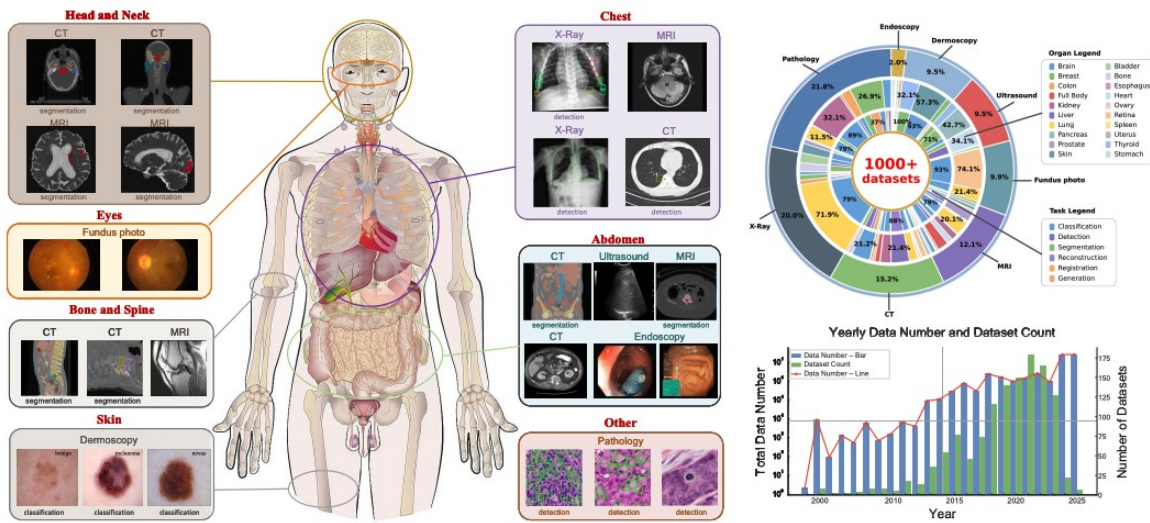


Figure 4: Overview of medical imaging datasets: representative modalities by anatomical region (left), dataset distribution across modalities, anatomical regions, and tasks (upper right), and temporal trends in dataset numbers (lower right).

主要亮点

- **数据融合范式：** MDPF 将同模态/同任务的小数据集整合为大规模训练资源，突破单一数据集规模限制
- **四维分类覆盖：** 按器官、模态（CT/MRI/超声等）、任务（分割/检测/分类）、标注类型多维索引，快速定位所需数据
- **交互式发现门户：** Project-Imaging-X 提供在线平台，支持自动化数据集集成，适合医疗 VLM 和影像基础模型开发

相关链接

Paper: <https://arxiv.org/abs/2603.27460>

Github: <https://github.com/uni-medical/Project-Imaging-X>

● 论文四

GEMS: Agent-Native Multimodal Generation with Memory and Skills

日期

2026-03-30

GEMS: Agent-Native Multimodal Generation with Memory and Skills

Zefeng He^{1,2*}, Siyuan Huang^{1,3*}, Xiaoye Qu^{1†}, Yafu Li^{1,4}, Tong Zhu¹, Yu Cheng^{4†}, Yang Yang^{3†}
¹Shanghai AI Laboratory, ²Nanjing University, ³Shanghai Jiao Tong University, ⁴CUHK

内容摘要

GEMS 提出面向多模态生成的 Agent-native 框架，三个核心组件：① Agent Loop——多轮迭代闭环优化；② Agent Memory——对优化轨迹进行分层压缩，持久存储事实状态与经验摘要；③ Agent Skill——按需加载领域专业知识模块。将 6B Z-Image-Turbo 与 GEMS 结合，在 GenEval2 上超越 SOTA 模型 Nano Banana 2，在主流基准上平均提升 14.22 分，在下游任务上平均提升 14.03 分。

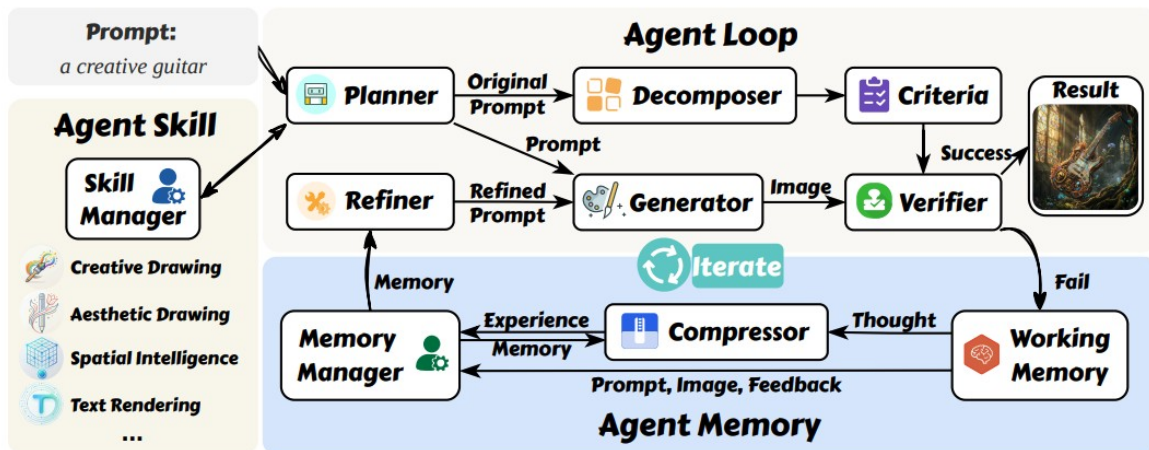


Figure 2: The system architecture of GEMS. The framework consists of three primary pillars: Agent Loop, Agent Memory, and Agent Skill. The user prompt is augmented with domain-specific expertise from Agent Skill, and then iteratively refined within the Agent Loop, with Agent Memory managing the historical context to guide the generation process.

主要亮点

- **轨迹级持久记忆：** Agent Memory 对优化过程中的状态和经验进行分层压缩，支持多轮迭代中的知识累积与复用
- **轻模型突破天花板：** 6B Z-Image-Turbo 借助 GEMS 在 GenEval2 上超越更大规模的 SOTA 模型 Nano Banana 2
- **即插即用设计：** Agent Skill 模块按需加载领域专业知识，框架可直接接入 Qwen-Image 等不同生成后端

相关链接

Paper: <https://arxiv.org/abs/2603.28088>

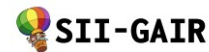
Github: <https://github.com/lcqysl/GEMS>

● 论文五

ASI-Evolve: AI Accelerates AI

日期

2026-03-31



ASI-Evolve: AI Accelerates AI

Weixian Xu^{†1,2,3} Tiantian Mi^{*1,2,3} Yixiu Liu^{*1,2,3} Yang Nan^{*2,3} Zhimeng Zhou^{*2} Lyumanshan Ye^{1,2,3}
Lin Zhang^{1,2,3} Yu Qiao² Pengfei Liu^{†1,2,3}

内容摘要

ASI-Evolve 构建 learn-design-experiment-analyze 闭环，让 AI 系统自动推进 AI 研发的三个核心环节：

① 预训练数据筛选——自动评估和优化训练语料质量；② 神经架构设计——在线性注意力搜索空间中发现 105 个新 SOTA 架构；③ RL 算法设计——自动生成和验证新的强化学习算法变体。是迄今为止覆盖 AI 研发栈最完整的自动化实验性证据之一。

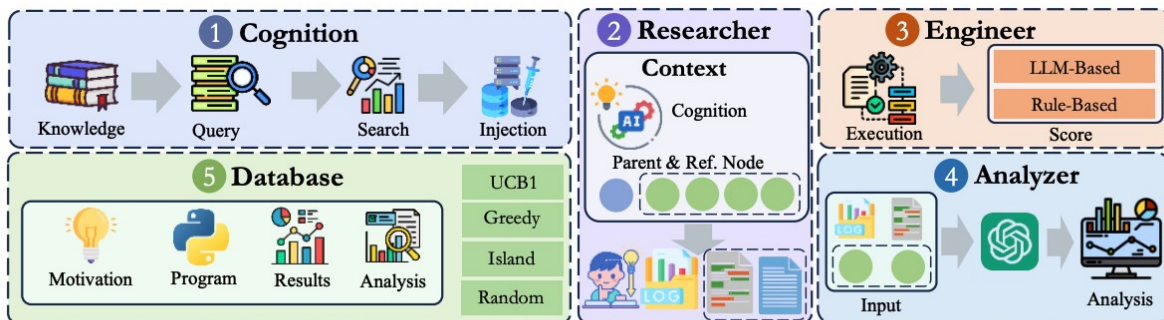


Figure 2: ASI-EVOLVE pipeline: in each round, the system samples context nodes from database, retrieves relevant cognition items via embedding search, generates a new candidate program, runs an experiment-specific evaluation script under timeouts, and summarizes results into an analysis report that is stored back into the database for future rounds.

主要亮点

- **三栈闭环覆盖：** 同时覆盖训练数据筛选、神经架构设计、RL 算法三类 AI 研发核心环节，不局限于单点优化
- **架构搜索突破：** 神经架构搜索模块自动发现 105 个新 SOTA 线性注意力架构，超越人工设计基线
- **端到端实证：** 为"AI 自动改进 AI"的全栈可行性提供了规模最大的系统性实验证据之一

相关链接

Paper: <https://arxiv.org/abs/2603.29640>

Github: <https://github.com/GAIR-NLP/ASI-Evolve>

● 论文六

The Latent Space: Foundation, Evolution, Mechanism, Ability, and Outlook

日期

2026-04-02



The Latent Space: Foundation, Evolution, Mechanism, Ability, and Outlook

Xinlei Yu^{*,*}, Zhangquan Chen^{*}, Yongbo He^{*}, Tianyu Fu^{*}, Cheng Yang^{*}, Chengming Xu^{*}, Yue Ma^{*}, Xiaobin Hu^{*,†}, Zhe Cao, Jie Xu, Guibin Zhang, Jiale Tao, Jiayi Zhang, Siyuan Ma, Kaituo Feng, Haojie Huang, Youxing Li, Ronghao Chen, Huacan Wang, Chenglin Wu, Zikun Su, Xiaogang Xu, Kelu Yao, Kun Wang, Chen Gao, Yue Liao, Ruqi Huang, Tao Jin, Zhucun Xue, Cheng Tan[†], Jiangning Zhang[†], Wenqi Ren, Yanwei Fu, Yong Liu, Yu Wang, Xiangyu Yue[†], Yu-Gang Jiang[†], Shuicheng Yan[†]

内容摘要

该综述从五个维度系统梳理语言模型潜在空间研究：基础（界定 latent space 的范畴）、演化（从早期探索到大规模扩展的历史脉络）、机制（Architecture / Representation / Computation / Optimization 四条发展线）、能力（推理、规划、建模、感知、记忆、协作、具身七类）、展望。核心论点：latent space 正成为语言模型的"原生底层"，解决 token 空间的语言冗余、离散化瓶颈和顺序低效三大局限。

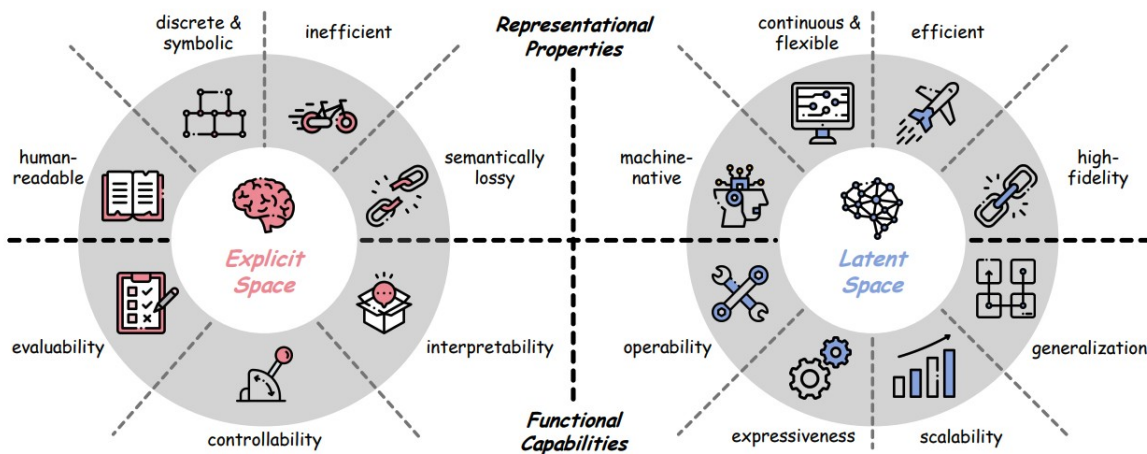


Figure 3 Comparison of the **explicit space** and **latent space** of the language models, including their representational properties and functional capabilities.

主要亮点

- **五维分类框架**：从基础、演化、机制、能力、展望五个维度构建统一的 latent space 研究分类体系
- **超越 token 限制**：指出语言冗余、离散化瓶颈、顺序低效是 token 空间三大结构性缺陷，latent space 是突破方向
- **配套资源索引**：Awesome-Latent-Space 持续跟踪 COCONUT、SoftCoT、COLAR 等代表性工作

相关链接

Paper: <https://arxiv.org/abs/2604.02029>

Github: <https://github.com/YU-deep/Awesome-Latent-Space>

● 论文七

ClawArena: Benchmarking AI Agents in Evolving Information Environment

日期

2026-04-05



ClawArena: Benchmarking AI Agents in Evolving Information Environments

Haonian Ji^{1*}, Kaiwen Xiong^{1*}, Siwei Han¹, Peng Xia¹, Shi Qiu¹, Yiyang Zhou¹, Jiaqi Liu¹, Jinlong Li¹, Bingzhou Li¹, Zeyu Zheng³, Cihang Xie², Huaxiu Yao¹

¹UNC-Chapel Hill ²University of California, Santa Cruz

³University of California, Berkeley

{haonianj, xkaiwen, huaxiu}@cs.unc.edu

内容摘要

ClawArena 构造三类耦合挑战来评测持续服务中的 Agent：① 多源冲突推理——需判断来自聊天记录、监控文件等异质来源的矛盾信息的可信度；② 动态信念修正——随新证据到来更新结论而非简单累积；③ 隐式个性化——从交互模式中学习用户偏好而无需显式提醒。当前包含 64 个场景、8 个专业领域、1,879 轮评测和 365 次动态更新，形成 14 类题目分类，三维挑战相互耦合，无法靠解决单一维度得分。



Figure 1: Overview of CLAWARENA across 8 professional domains. Each scenario presents multi-channel session histories, workspace files, and evaluation questions requiring multi-source conflict reasoning, dynamic belief revision, and implicit personalization. The center logo reflects the benchmark's adversarial spirit: agents must "claw" through conflicting evidence to reach the ground truth.

主要亮点

- **三维耦合评测：** 多源冲突、动态信念修正、隐式个性化三类挑战相互耦合，单维度优化无法通过整体评测
- **模型优先于框架：** 模型能力对表现影响（15.4% 差距）大于框架设计（9.2% 差距），为选型提供量化依据
- **自进化框架受益：** MetaClaw 等自进化框架可部分弥补不同语言模型之间的性能差距

相关链接

Paper: <https://arxiv.org/abs/2604.04202>

Github: <https://github.com/aiming-lab/ClawArena>

● 论文八

PaperOrchestra: A Multi-Agent Framework for Automated AI Research Paper Writing

日期

2026-04-06



PAPERORCHESTRA: A Multi-Agent Framework for Automated AI Research Paper Writing

Yiwen Song¹, Yale Song¹, Tomas Pfister¹ and Jinsung Yoon¹

¹Google

内容摘要

Google Cloud AI Research 提出 PaperOrchestra，将粗糙的实验结果和想法摘要自动转化为可投稿的 LaTeX 论文。五个专属 Agent 流水线：Outline（结构规划）→ Plotting（图表生成）→ Literature Review（文献综述）→ Section Writing（正文写作）→ Content Refinement（修改润色）。配套发布 PaperWritingBench（200 篇顶会论文逆向重构），人工评估中论文整体质量赢率差距达 14%–38%，文献综述质量赢率差距达 50%–68%，均优于现有自主写作基线。

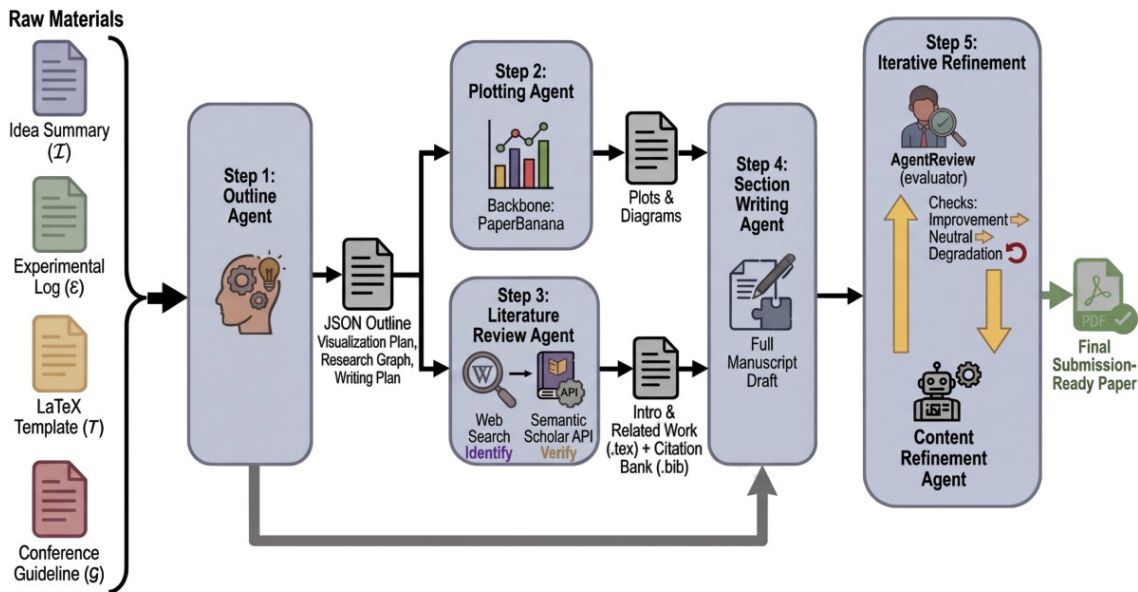


Figure 1 | Overview of the PAPERORCHESTRA framework. Specialized agents systematically parse unstructured inputs, synthesize plots and literature, compile a full draft, and iteratively refine the manuscript into a submission-ready PDF. (This figure was generated using PaperBanana (Zhu et al., 2026).)

主要亮点

- **五 Agent 流水线：** Outline→Plotting→Literature Review→Section Writing→Content Refinement

五阶段分工，各环节专业化

- **专项评测基准：** PaperWritingBench (200 篇顶会论文逆向重构) 为论文写作质量评估提供标准化框架
- **人工评估领先：** 文献综述质量赢率差距 50%–68%，整体稿件质量赢率差距 14%–38%，优于现有自主写作系统

相关链接

Paper: <https://arxiv.org/abs/2604.05018>

● 论文九

MinerU2.5-Pro: Pushing the Limits of Data-Centric Document Parsing at Scale

日期

2026-04-07



上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

MinerU2.5-Pro: Pushing the Limits of Data-Centric Document Parsing at Scale

Bin Wang^{1*†}, Tianyao He^{1*}, Linke Ouyang^{1*}, Fan Wu^{1*}, Zhiyuan Zhao^{1*}, Tao Chu^{1*}
Yuan Qu^{1*}, Zhenjiang Jin^{1*}, Weijun Zeng^{1*}, Ziyang Miao^{1*}, Bangrui Xu^{1*}, Junbo Niu^{1,2*}
Mengzhang Cai¹, Jiantao Qiu¹, Qintong Zhang^{1,2}, Dongsheng Ma¹, Yuefeng Sun¹, Hejun Dong¹
Wenzheng Zhang^{1,2}, Jutao Xiao¹, Jiayong Shi¹, Pengyu Liao¹, Xiaomeng Zhao¹, Huaping Zhong⁴
Liqun Wei¹, Jing Yu¹, Jie Yang¹, Wei Li¹, Shasha Wang¹, Qianqian Wu¹, Xuanhe Zhou^{1,3}
Weijia Li¹, Zhenxiang Li¹, Zhongying Tu¹, Jiang Wu¹, Lijun Wu¹, Chao Xu¹, Kai Chen¹
Wentao Zhang^{1,2}, Yu Qiao¹, Bowen Zhou¹, Dahua Lin¹✉, Conghui He¹✉

内容摘要

在固定 1.2B 参数架构不变的前提下, MinerU2.5-Pro 仅通过数据工程将 OmniDocBench v1.6 得分提升至 95.69, 超越 Gemini 3 Pro、Qwen3-VL-235B 等规模更大的模型。三阶段数据引擎: ① 多样性与难度感知采样——将训练集从 <10M 扩展到 65.5M 样本; ② CMCV (跨模型一致性验证) ——利用异质模型的输出共识评估样本难度并生成可靠标注; ③ Judge-and-Refine——"渲染-验证"迭代修正流程处理困难样本。三阶段训练: 大规模预训练 → 困难样本微调 → GRPO 强化学习对齐。

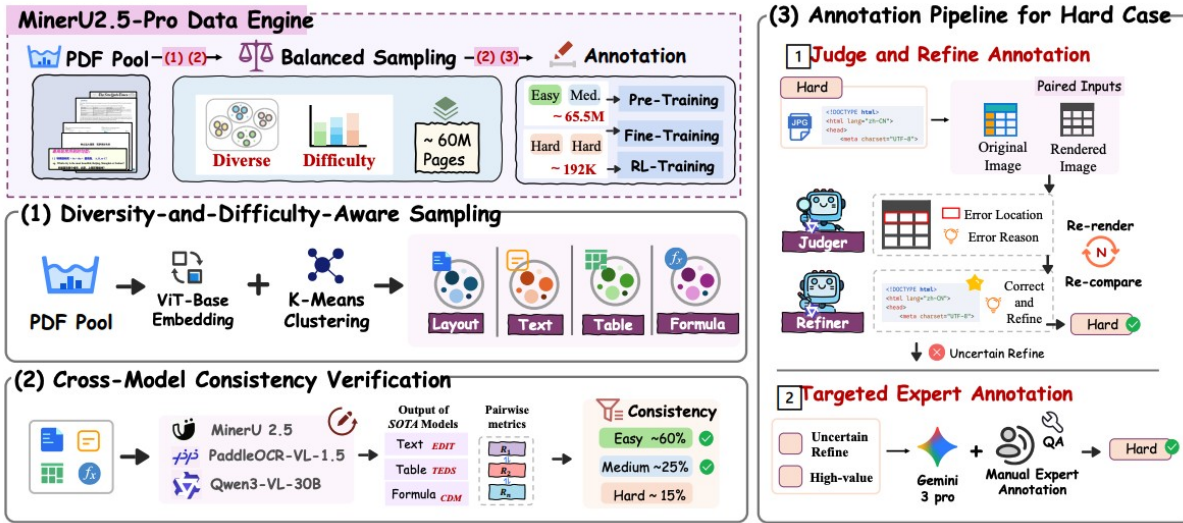


Figure 2: Overview of the Data Engine pipeline. The system co-optimizes three dimensions—Coverage, Informativeness, and Accuracy—through three synergistic stages: Diversity-and-Difficulty-Aware Sampling (DDAS), Cross-Model Consistency Verification (CMCV), and the Annotation Pipeline for Hard Case.

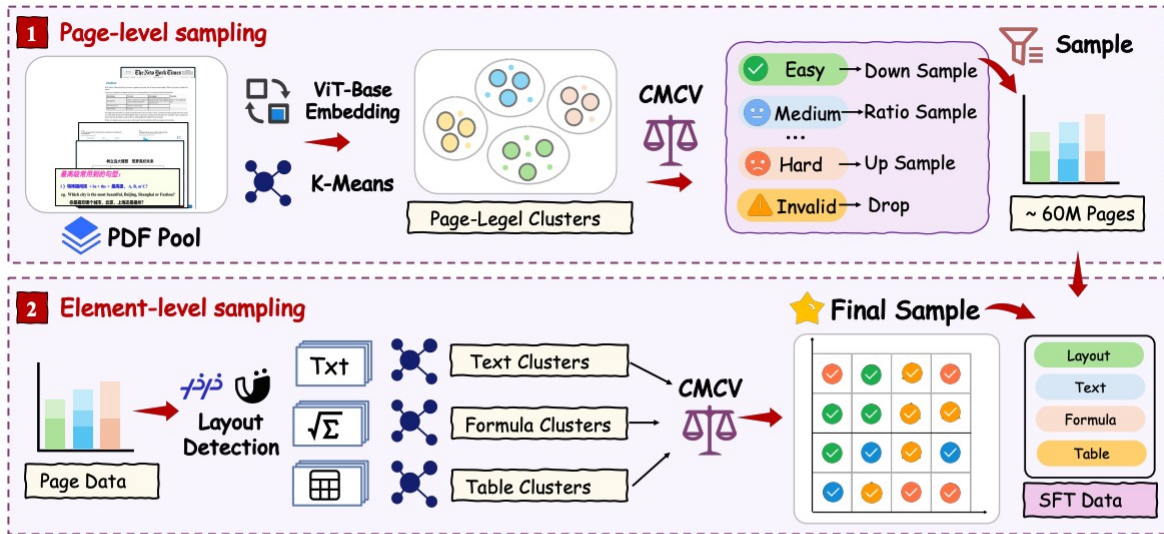


Figure 3: The DDAS pipeline operates at two granularity levels. **Upper:** Page-level sampling for layout detection data—pages from the PDF pool are embedded via ViT-base, clustered, and resampled by jointly weighting cluster diversity and CMCV-derived difficulty, yielding about 60M pages with balanced distribution and difficulty coverage. **Lower:** Element-level sampling—the selected pages are parsed by layout detection models into text, formula, and table blocks; each element type is independently clustered and assessed by CMCV, then sampled to balance both diversity and difficulty at the element granularity. The two levels are combined to produce the final training data for layout, text, formula, and table subtasks.

主要亮点

- **数据引擎驱动：** 三步数据流水线（多样性采样→CMCV→Judge-and-Refine）将训练集扩展至 65.5M 高质量样本
- **以小胜大：** 1.2B 模型在 OmniDocBench v1.6 得 95.69，超越 Gemini 3 Pro、Qwen3-VL-235B 等大模型
- **GRPO 最终对齐：** 三阶段训练末段引入 GRPO 强化学习，进一步提升边界情况的解析准确率

相关链接

Paper: <https://arxiv.org/abs/2604.04771>

Github: <https://github.com/opendatalab/MinerU>

● 论文十

[ICLR 26] In-Place Test-Time Training

日期

2026-04-07

 ByteDance | Seed

In-Place Test-Time Training

Guhao Feng^{1,2,*}, Shengjie Luo^{1,*}, Kai Hua¹, Ge Zhang¹,
Di He^{2,†}, Wenhao Huang^{1,†}, Tianle Cai¹

¹ByteDance Seed, ²Peking University

内容摘要

In-Place TTT 将 MLP block 的最终投影矩阵复用为"快速权重", 在推理时对其进行在线更新, 使模型能持续吸收新信息而无需新增任何参数。关键设计: ① 用与 next-token prediction 严格对齐的目标函数替代通用重建目标; ② 分块更新机制 (chunk-wise) 替代逐 token 更新, 兼顾效率与上下文并行兼容性。在 Qwen3-4B-Base 等模型上, 128K 上下文任务性能显著优于同规模基线。

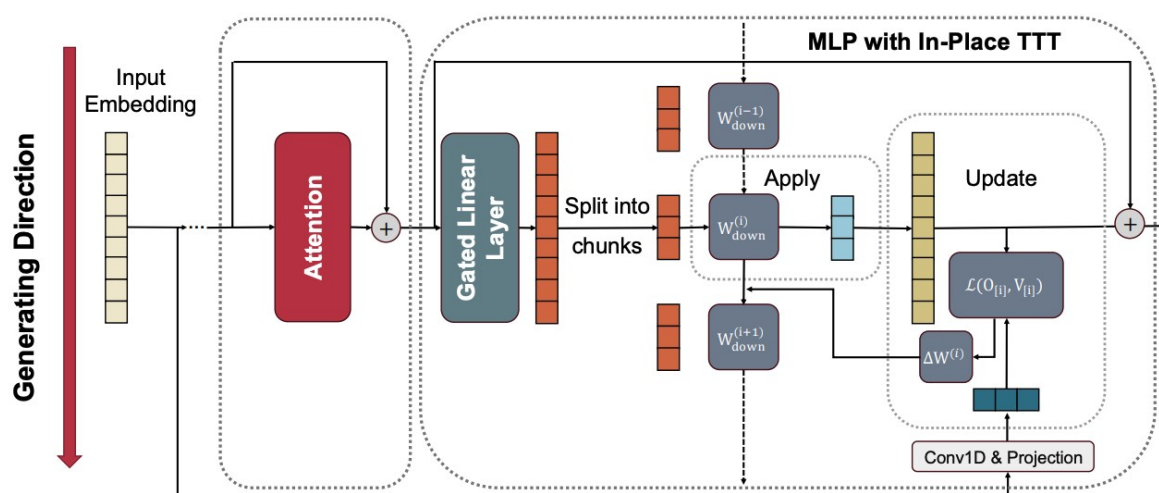


Figure 1 The overall framework of our In-Place Test-Time Training. The module operates sequentially on input chunks. For each chunk, the current fast weights are first *applied* to the intermediate activations $\mathbf{Z}$ to produce the output. Then, these weights are *updated* using the activations $\mathbf{Z}$ and a value $\mathbf{V}$ derived from the token embeddings. This "apply-then-update" cycle allows the model to dynamically adapt to incoming context in a strictly causal manner.

主要亮点

- **可插拔设计：** 复用现有 MLP 最终投影矩阵为快速权重，零参数新增，与任意 Transformer LLM 即插即用
- **NTP 对齐目标：** 推理时更新目标与 next-token prediction 严格对齐，避免通用重建目标的任务错位问题

相关链接

Paper: <https://arxiv.org/abs/2604.06169>

Github: <https://github.com/ByteDance-Seed/In-Place-TTT>

● 论文十一

AutoSOTA: An End-to-End Automated Research System for State-of-the-Art AI Model Discovery

日期

2026-04-07

AutoSOTA: An End-to-End Automated Research System for State-of-the-Art AI Model Discovery

Yu Li^{1,*} Chenyang Shao^{1,2,*} Xinyang Liu^{1,*} Ruotong Zhao¹ Peijie Liu¹
Hongyuan Su^{1,2} Zhibin Chen¹ Qinglong Yang¹ Anjie Xu^{2,3} Yi Fang^{2,4} Qingbin Zeng¹
Tianxing Li¹ Jingbo Xu¹ Fengli Xu^{1,2,†} Yong Li^{1,2,†} Tie-Yan Liu²

¹Department of Electronic Engineering, BNRist, Tsinghua University

²Zhongguancun Academy

³Peking University

⁴University of Science and Technology of China

内容摘要

AutoSOTA 以顶会论文为起点，经由自动复现→调试→优化→创新发现 SOTA 模型的完整链路。8 个专属 Agent 分三阶段协作：① 资源准备与目标设定（代码对齐、评测基准构建）；② 实验评估（环境自动修复、长程实验跟踪）；③ 反思与想法生成（优化调度、有效性监督防止虚假提升）。在 8 个顶会论文集合上发现 105 个新 SOTA 模型，平均性能提升约 10%，平均每篇耗时约 5 小时。优化内容不限于超参数，涵盖架构创新、算法重设计和 workflow 级改进。

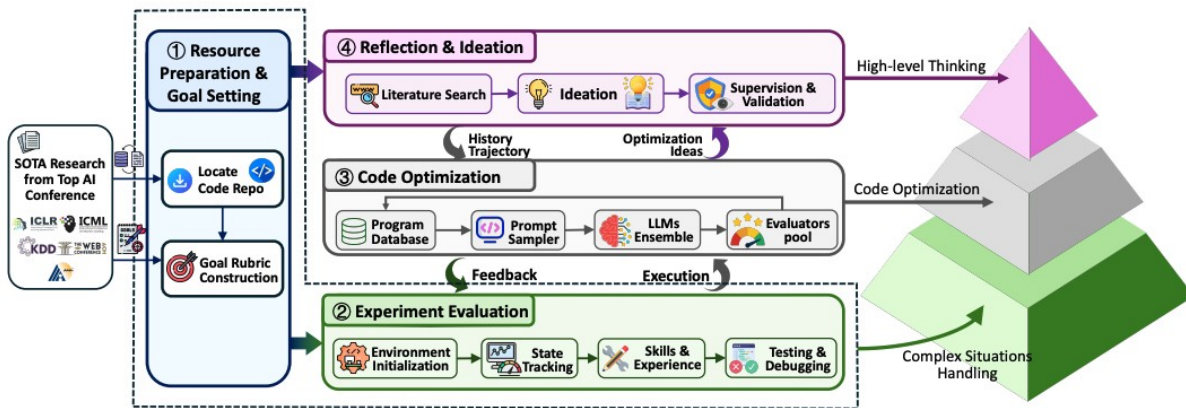


Figure 1: Task Definition of AutoSOTA: An End-to-End Automated Research Framework

主要亮点

- **三阶段八 Agent:** 资源准备→实验评估→反思创意三阶段，8 专属 Agent 覆盖环境修复、想法生成、有效性监督全流程
- **量化产出:** 8 个顶会论文集集合上发现 105 个新 SOTA 模型，平均性能提升 ~10%，平均每篇约 5 小时
- **超越调参:** 支持架构创新、算法重设计、workflow 级改进，并有有效性监督机制防止虚假性能提升

相关链接

Paper: <https://arxiv.org/abs/2604.05550>

● 论文十二

[CVPR 26] EpiAgent: An Agent-Centric System for Ancient Inscription Restoration

日期

2026-04-08

EpiAgent: An Agent-Centric System for Ancient Inscription Restoration

Shipeng Zhu^{1,2}, Ang Chen^{1,2}, Na Nie^{4,5}, Pengfei Fang^{1,2}, Min-Ling Zhang^{1,3}, Hui Xue^{1,2*}
¹School of Computer Science and Engineering, Southeast University, China
²Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
³Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China
⁴Nanjing University Museum, Nanjing University, China
⁵The China Centre for Linguistic and Strategic Studies, Nanjing University, China

内容摘要

EpiAgent 将古代碑刻残损修复建模为层级规划问题，受人类金石学家协作 workflow 启发，采用 Observe–Conceive–Execute–Reevaluate 循环。Kimi-K2 驱动的中央规划器根据上下文动态调度：多模态视觉分析、历史铭文经验库、专业修复工具和迭代自我精化四类组件，应对复杂的异质性退化（磨损、缺字、断行等）。实验显示 EpiAgent 在修复质量和跨碑刻类型的泛化能力上均优于单次推理方法。

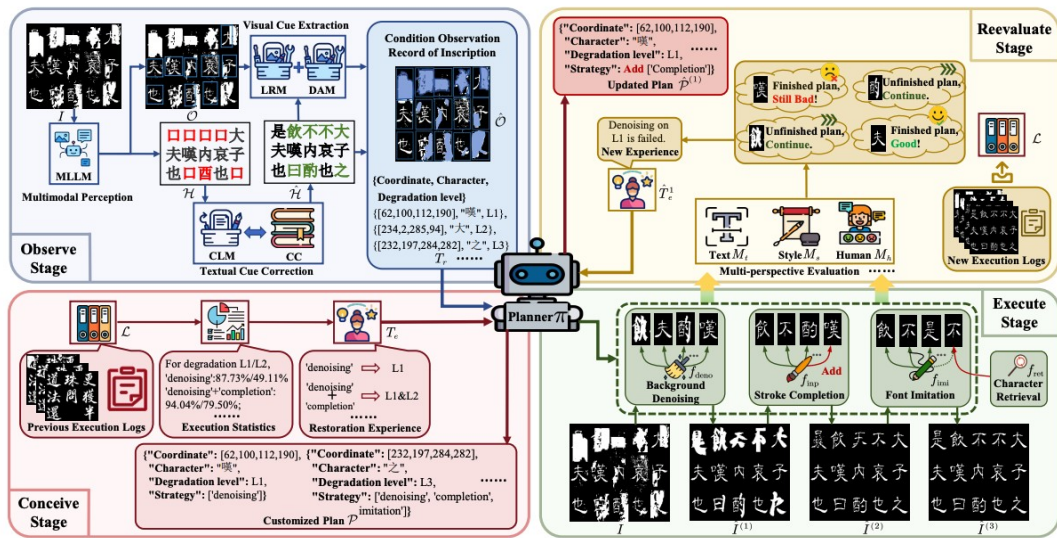


Figure 3. Illustration of the workflow of EpiAgent. The “MLLM”, “LRM”, “DAM”, “CLM”, and “CC” denote Multimodal Large Language Model, Layout Rectification Module, Degradation Assessment Model, Corrective Language Model, and Chinese Corpora, respectively.

主要亮点

- **层级规划框架：** Observe–Conceive–Execute–Reevaluate 循环模拟人类金石学家协作，处理复杂异质性退化
- **动态工具调度：** 中央规划器（Kimi-K2）根据上下文动态装配视觉分析、历史经验、专业修复工具

相关链接

Paper: <https://arxiv.org/abs/2604.09367>

Github: <https://github.com/blackprotoss/EpiAgent>

● 论文十三

[CVPR 26] Vision Transformers Need More Than Registers

日期

2026-04-14

Vision Transformers Need More Than Registers

Cheng Shi¹ Yizhou Yu^{1†} Sibeil Yang^{2†}

¹School of Computing and Data Science, The University of Hong Kong ²Sun Yat-sen University

shicheng2025@connect.hku.hk, yizhouy@acm.org, yangsb3@mail.sysu.edu.cn

<https://github.com/ChengShiest/LAST-ViT>

内容摘要

重新审视 ViT 特征异常问题，将根本原因定义为“懒惰聚合” (lazy aggregation)：ViT 倾向于将语义无关的背景 patch 用作表达全局语义的“捷径”，这一行为由全局注意力机制与粗粒度语义监督共同驱动，高范数 token 只是该行为在训练后期的表现形式。提出 LaSt-ViT (LazyStrike ViT)：基于通道维度低通滤波的稳定性评分，选择性将前景 patch 特征集成到 CLS token，在 12 个基准上均取得一致性性能提升。

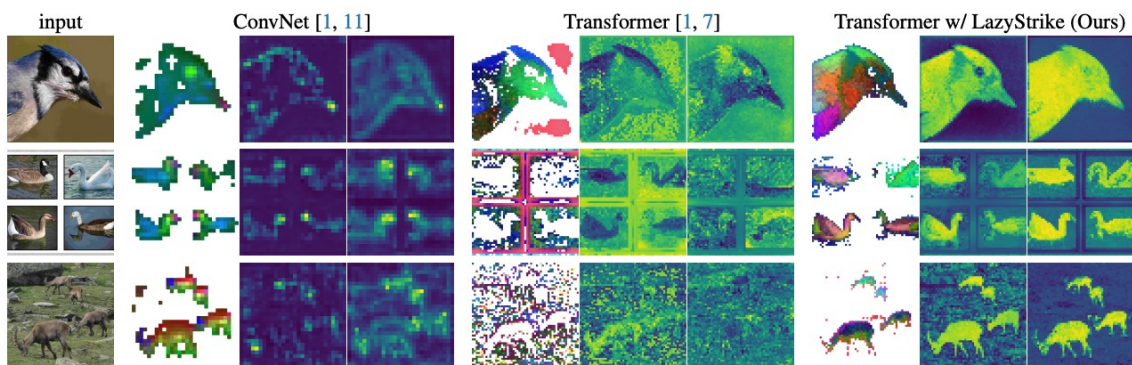


Figure 1. LazyStrike provides a unified framework for analyzing and mitigating diverse artifacts across different supervision settings in ViTs. The figure visualizes patch scores—defined as CLS-patch similarity—under full supervision (middle) and self-supervision (right), together with PCA projections of features under full supervision (left).

主要亮点

- **根因重新定义：** 将 ViT 特征异常归因于"懒惰聚合"行为而非高范数 token 本身，揭示更深层的结构缺陷
- **LaSt-ViT 方法：** 通道低通滤波稳定性评分 + 选择性前景集成，抑制背景捷径、锚定 CLS 到前景区域
- **12 基准一致提升：** 在目标发现、语义/实例分割、开放词汇检测等 12 个基准上均取得一致性性能提升

相关链接

Paper: <https://arxiv.org/abs/2602.22394>

Github: <https://github.com/ChengShiest/LAST-ViT>

● 论文十四

[ACL 26] Latent-Condensed Transformer for Efficient Long Context Modeling

日期

2026-04-14

Latent-Condensed Transformer for Efficient Long Context Modeling

**Zeng You^{1,2*}, Yaofo Chen^{1*}, Qiuwu Chen^{3*}, Ying Sun^{3*},
Shuhai Zhang¹, Yingjian Li², Yaowei Wang^{2,4†}, Mingkui Tan^{1,5†}**

¹South China University of Technology

²Pengcheng Laboratory

³AIGCode

⁴Harbin Institute of Technology

⁵Pazhou Laboratory

内容摘要

该工作提出 Latent-Condensed Attention (LCA) , 直接在 MLA 的 latent space 中对长上下文进行压缩, 通过双路解耦处理: ① 语义 latent 向量——query 感知加权池化聚合; ② 位置 key——硬锚点选择 (每组保留相关性最高的 token) 以保留位置保真度。在 128K 上下文下实现 prefilling 加速 2.5×、KV cache 降低 90%, 零额外参数, 并给出长度无关误差上界的理论保证, 可扩展至 GQA 等其他注意力机制。

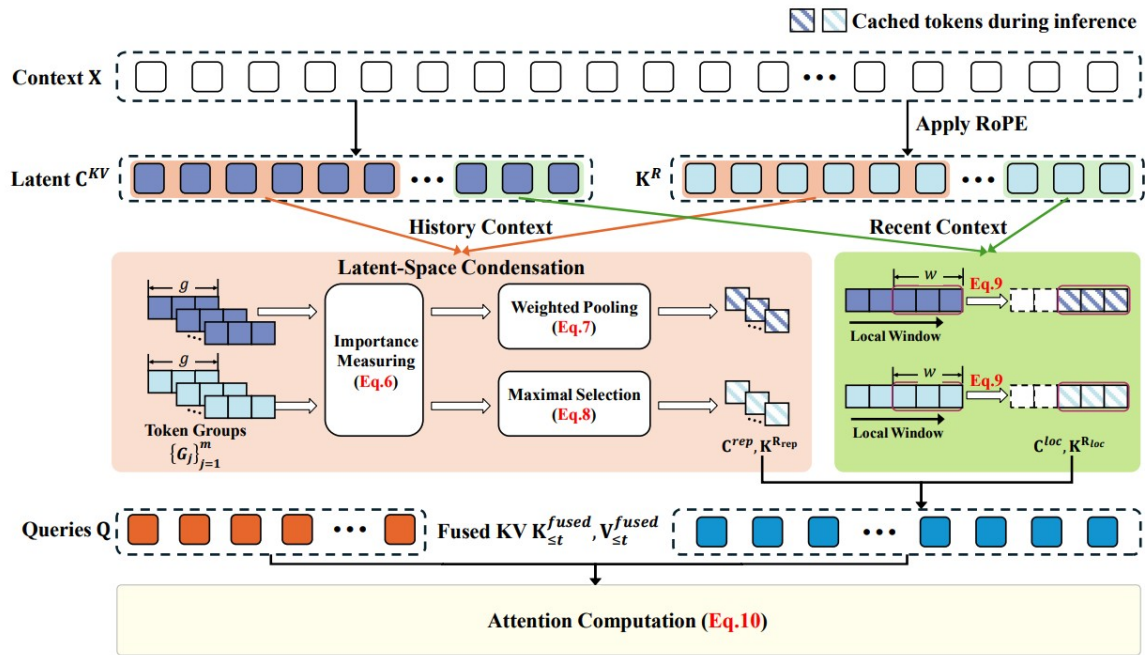


Figure 1: An overview of Latent-Condensed Attention (LCA). The history context is condensed into a compact set of representatives via group-wise condensation in the latent space: semantic information is aggregated through weighted pooling of C^{KV} , while positional information is preserved via anchor selection from K^R . The recent context is retained in full fidelity without condensation.

主要亮点

- **双路解耦压缩**: 语义向量用 query 感知加权池化聚合, 位置键用硬锚点选择保留, 两类信息分路处理
- **无参数开销**: 在 MLA latent space 中直接操作, 零额外参数, 且具有长度无关的误差上界理论保证
- **128K 实测效果**: prefilling 加速 2.5 $\times$, KV cache 降低 90%, 性能不降。

相关链接

Paper: <https://arxiv.org/abs/2604.12452>

● 论文十五

Dive into Claude Code: The Design Space of Today's and Future AI Agent Systems

日期

2026-04-14

Dive into Claude Code: The Design Space of Today's and Future AI Agent Systems

Jiacheng Liu¹, Xiaohan Zhao¹, Xinyi Shang^{1,2}, Zhiqiang Shen^{1,†}

¹VILA Lab, Mohamed bin Zayed University of Artificial Intelligence, ²University College London

[†]Corresponding author

内容摘要

该论文逆向分析 Claude Code v2.1.88 的公开 TypeScript 源码,发现代码库中仅约 1.6% 是 AI 决策逻辑,其余 98.4% 是权限门控、上下文压缩管道 (5 层)、工具路由、扩展机制 (MCP/插件/技能/钩子)、Subagent 协作与 Session 存储等确定性基础设施。归纳出五大设计价值 (人类控制权、安全性、可靠执行、能力放大、上下文适应性), 并与开源 Agent 系统 OpenClaw 进行对比, 揭示闭源产品与开源框架在安全策略、循环结构和扩展性上的根本差异。

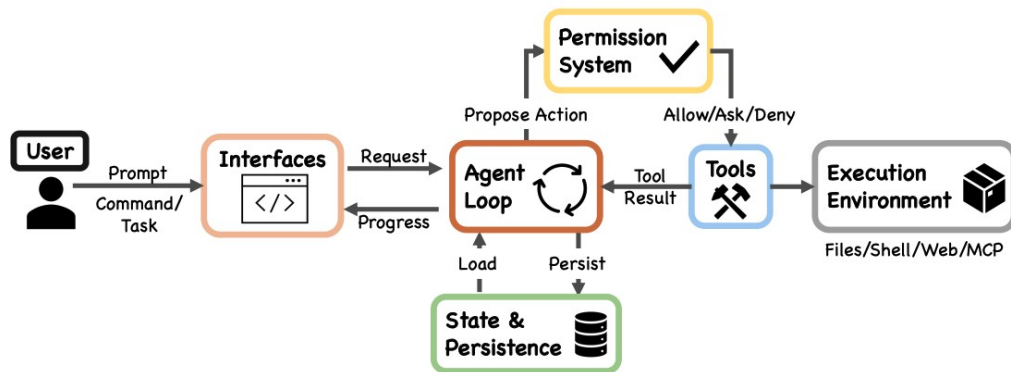


Figure 1 High-level system structure of Claude Code. The system decomposes into seven functional components: user, interfaces, the agent loop, a permission system, tools, state & persistence, and an execution environment. All entry surfaces converge on the same agent loop.

主要亮点

- **1.6% AI 逻辑：** 代码库中仅 1.6% 是 AI 决策逻辑，98.4% 为权限门控、上下文管理等确定性基础设施
- **五层上下文压缩：** 详细拆解 Claude Code 的 5 层上下文压缩管道，揭示长会话管理的工程实现机制
- **五大设计价值：** 总结人类控制权、安全性、可靠执行、能力放大、上下文适应性五个原则并映射至具体实现

相关链接

Paper: <https://arxiv.org/abs/2604.14228>

Github: <https://github.com/VILA-Lab/Dive-into-Claude-Code>

● 论文十六

LLaDA2.0-Uni: Unifying Multimodal Understanding and Generation with Diffusion Large Language Model

日期

2026-04-23



LLaDA2.0-Uni: Unifying Multimodal Understanding and Generation with Diffusion Large Language Model

Tiwei Bie, Haoxing Chen, Tieyuan Chen, Zhenglin Cheng, Long Cui, Kai Gan, Zhicheng Huang
Zhenzhong Lan[†], Haoquan Li, Jianguo Li[†], Tao Lin[†], Qi Qin, Hongjun Wang, Xiaomei Wang
Haoyuan Wu, Yi Xin, Junbo Zhao

AGI Research Center, Inclusion AI

内容摘要

LLaDA2.0-Uni 用统一离散扩散 LLM 框架将多模态理解与生成整合在单一模型中，三组件设计：① SigLIP-VQ 语义离散 tokenizer——将连续视觉输入离散化为语义 token；② MoE-based dLLM backbone——对文本和视觉 token 统一执行 block-level masked diffusion；③ Diffusion decoder——通过 few-step 蒸馏实现 8 步快速图像重建。在多模态理解任务上匹敌专用 VLM，同时支持图像生成、图像编辑和交错推理生成，并通过 prefix-aware 前缀优化提升 backbone 推理效率。

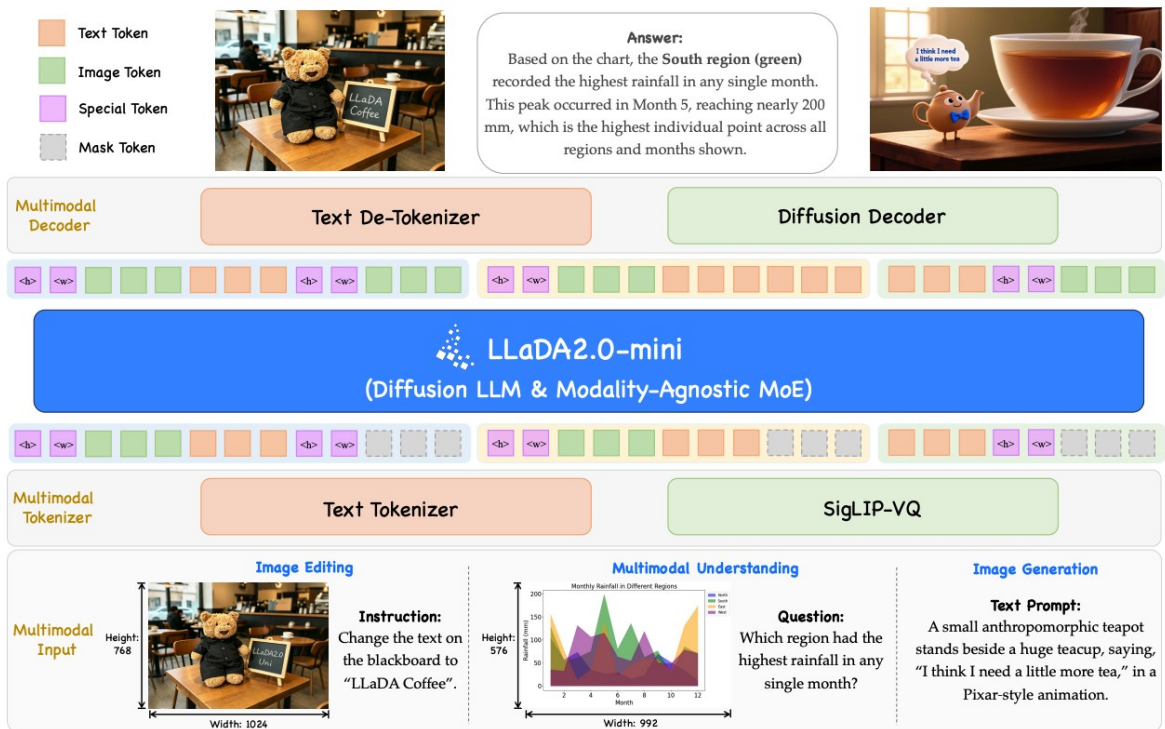


Figure 4: Architecture Overview of LLaDA2.0-Uni. The framework integrates a SigLIP-VQ tokenizer and a language model decoder to process multimodal inputs, including text, image, and video.

主要亮点

- **离散统一表示:** SigLIP-VQ 将视觉输入离散化, 文本与图像 token 在同一 backbone 中统一扩散建模
- **快速扩散解码:** diffusion decoder 通过 few-step 蒸馏支持 8 步快速图像生成, 兼顾质量与推理效率
- **理解生成兼备:** 多模态理解匹配专用 VLM, 同时支持图像生成、编辑和交错推理生成三类任务

相关链接

Paper: <https://arxiv.org/abs/2604.20796>

Github: <https://github.com/inclusionAI/LLaDA2.0-Uni>

● 论文十七

Agentic World Modeling: Foundations, Capabilities, Laws, and Beyond

日期

2026-04-24



Agentic World Modeling: Foundations, Capabilities, Laws, and Beyond

Meng Chu^{1†}, Xuan Billy Zhang^{2†}, Kevin Qinghong Lin^{3†}, Lingdong Kong^{2†},
Jize Zhang^{3†}, Teng Tu^{2†}, Weijian Ma^{2†}, Ziqi Huang⁴, Senqiao Yang⁵, Wei Huang⁶,
Yeying Jin², Zhefan Rao¹, Jinhui Ye¹, Xinyu Lin², Xichen Zhang¹, Qisheng Hu⁴,
Shuai Yang⁶, Leyang Shen², Wei Chow², Yifei Dong⁷, Fengyi Wu⁷, Quanyu Long⁴,
Bin Xia⁵, Shaozuo Yu⁵, Mingkang Zhu⁵, Wenhui Zhang¹, Jiehui Huang¹,
Haokun Gui¹, Haoxuan Che^{1§}, Long Chen^{1§}, Qifeng Chen^{1§}, Wenxuan Zhang^{9§},
Wenya Wang^{4§}, Xiaojuan Qi^{6§}, Yang Deng^{10§}, Yanwei Li^{5§}, Mike Zheng Shou^{2§},
Zhi-Qi Cheng^{7§}, See-Kiong Ng^{2§}, Ziwei Liu^{4§}, Philip Torr^{3§}, Jiaya Jia^{1§}

[†]Core Contributor. Project Lead. [§]Senior Author.

¹Hong Kong University of Science and Technology,

²National University of Singapore, ³University of Oxford,

⁴Nanyang Technological University, ⁵Chinese University of Hong Kong,

⁶University of Hong Kong, ⁷University of Washington,

⁸Hong Kong University of Science and Technology (Guangzhou),

⁹Singapore University of Technology and Design, ¹⁰Singapore Management University

内容摘要

该综述 (88 页) 提出 “levels × laws” 分类框架，将世界模型能力分为三级：L1 预测器（单步局部转移）、L2 模拟器（多步动作条件展开）、L3 自进化器（预测失败时自主修正模型）；同时以物理、数字、社会、科学四类规律域定义约束边界和失效模式。综合 400+ 工作、100+ 代表性系统，涵盖基于模型的 RL、视频生成、Web/GUI Agent、多 Agent 社会模拟和 AI 驱动科学发现，并提出决策中心评测原则。

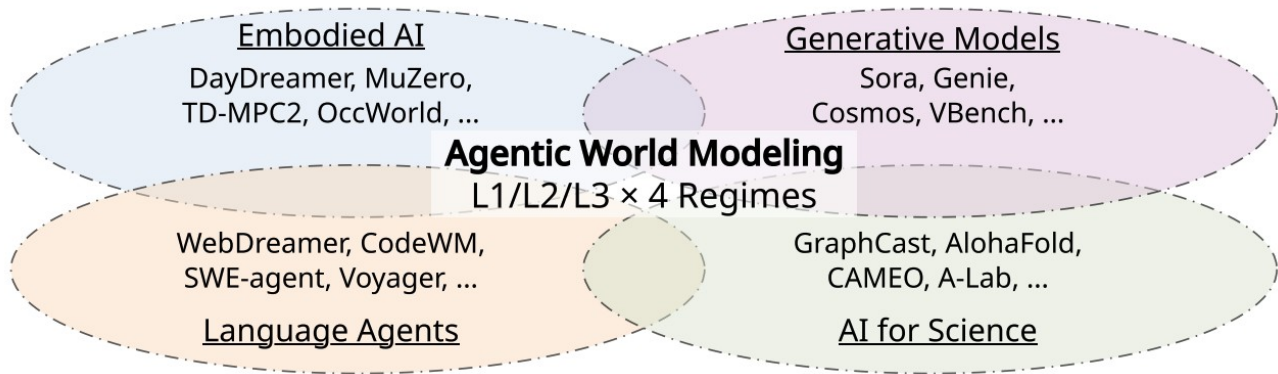


Figure 2: **Positioning of this survey relative to existing world model and agent surveys.** Four clusters, Embodied World Models, Generative World Models, Language Agents, and AI for Science, each cover subsets of the field. Our survey (center) integrates cross domain coverage with a capability based taxonomy ($L1/L2/L3 \times 4$ regimes), bridging largely isolated communities.

主要亮点

- **L1/L2/L3 分级：** 预测器→模拟器→自进化器三级能力阶梯，对应日益复杂的环境建模与自适应能力要求
- **四类规律域：** 物理、数字、社会、科学四类法则域定义世界模型的约束边界和典型失效模式
- **决策中心评测：** 提出超越生成质量的决策导向评测原则，并提供包含可复现最小套件的配套资源库

相关链接

Paper: <https://arxiv.org/abs/2604.22748>

Github: <https://github.com/matrix-agent/awesome-agentic-world-modeling>

● 论文十八

[ACL 26] Scaling Behaviors of LLM Reinforcement Learning Post-Training: An Empirical Study in Mathematical Reasoning

日期

2025-09-29

Scaling Behaviors of LLM Reinforcement Learning Post-Training: An Empirical Study in Mathematical Reasoning

Zelin Tan^{1,2*} Hejia Geng³ Xiaohang Yu⁴ Mulei Zhang¹⁰ Guancheng Wan¹⁰
Yifan Zhou⁵ Qiang He⁷ Xiangyuan Xue^{2,6} Heng Zhou^{1,2} Yutao Fan²
Zhongzhi Li⁷ Zaibin Zhang^{8,3} Guibin Zhang⁹ Chen Zhang^{2†} Zhenfei Yin^{3†}
Philip Torr³ Lei Bai²

¹University of Science and Technology of China ²Shanghai AI Laboratory ³University of Oxford
⁴Imperial College London ⁵University of Georgia ⁶The Chinese University of Hong Kong
⁷Chinese Academy of Sciences ⁸Dalian University of Technology ⁹National University of Singapore
¹⁰Wuhan University

[†]Corresponding authors: zhangchenzc@mail.ustc.edu.cn, jeremyyin@robots.ox.ac.uk

内容摘要

该论文基于 Qwen2.5 系列 (0.5B–72B) 的 54 组系统实验, 使用 GRPO 算法和 VeRL 框架, 研究 RL 后训练阶段的 scaling 规律, 得到三条核心结论: ① 固定算力预算下, 大模型短训练优于小模型长训练; ② 固定数据量下, 大模型样本效率更高 (边际收益递减但始终领先); ③ 数据受限场景下, 高质量数据重复使用有效, 性能主要由总优化步数决定而非样本唯一性。研究还揭示 RL 后训练存在可预测的幂律关系, 适用于有指导和无指导调优两类模型。

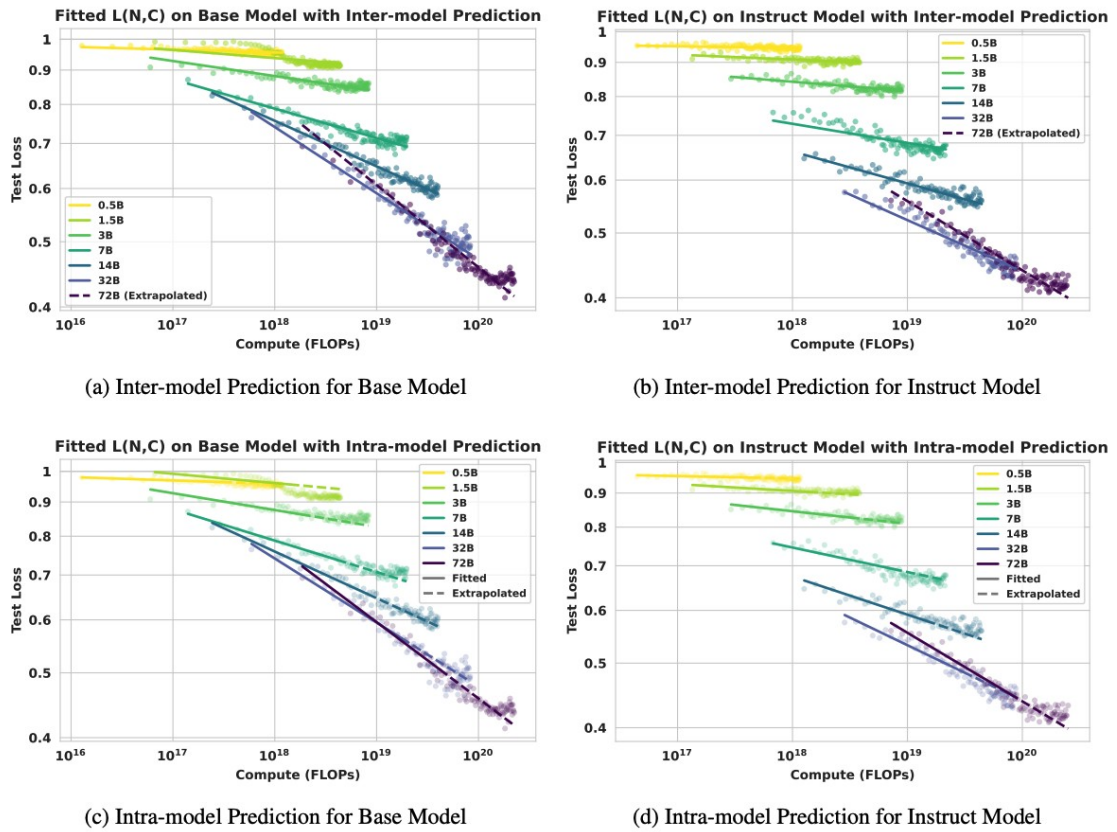


Figure 1: **Empirical Scaling Behaviors in RL Post-training.** We validate the proposed power-law (Eq. 1), which characterizes the relationship between model performance and post-training resource consumption, across models ranging from 0.5B to 72B parameters. **(a, b) Inter-model Prediction:** Scaling coefficients fitted on smaller models (0.5B–32B) accurately predict the learning efficiency of the held-out 72B model (dashed lines). **(c, d) Intra-model Prediction:** The final convergence on the full dataset is extrapolated solely from early training dynamics.

主要亮点

- **算力预算分配：** 固定算力下大模型短训练优于小模型长训练，揭示 RL 后训练的模型规模效益
- **数据复用有效：** 数据受限时高质量数据重复使用有效，性能由总优化步数而非样本唯一性决定
- **系统实验：** 覆盖 0.5B–72B 规模的 54 组实验，首次为 RL 后训练规模律提供系统量化基础

相关链接

Paper: <https://arxiv.org/abs/2509.25300>

● 论文十九

Image Generators are Generalist Vision Learners

日期

2026-04-22

Google DeepMind

2026-04-23

Image Generators are Generalist Vision Learners

Valentin Gabeur^{*}, Shangbang Long^{*}, Songyou Peng^{*}, Paul Voigtlaender[▽], Shuyang Sun[▽], Yanan Bao[▽], Karen Truong[▽], Zhicheng Wang[▽], Wenlei Zhou[▽], Jonathan T. Barron[▽], Kyle Genova[▽], Nithish Kannen[▽], Sherry Ben[▽], Yandong Li[▽], Mandy Guo[▽], Suhas Yogin[▽], Yiming Gu[†], Huizhong Chen[†], Oliver Wang[‡], Saining Xie[‡], Howard Zhou[‡], Kaiming He[‡], Thomas Funkhouser[‡], Jean-Baptiste Alayrac[‡] and Radu Soricut[‡]

^{*}project leads and equal contributions, [▽]core contributors, [†]project advisors, [‡]leadership sponsors

内容摘要

Vision Banana (Google DeepMind) 以 Nano Banana Pro (NBP) 图像生成模型为基础，通过轻量级指令微调将视觉任务（分割、深度估计、表面法向量等）重构为图像生成问题：输入原始图像+指令，输出目标可视化结果（如分割掩码渲染图、深度图）。核心论点：图像生成预训练已隐含了对 2D/3D 场景结构的深度理解，无需为每类视觉任务设计专用架构。在多个基准上超越 SAM 3（分割）和 Depth Anything V3（深度估计）。权重未公开，Nano Banana Pro 可通过 Gemini API 使用。

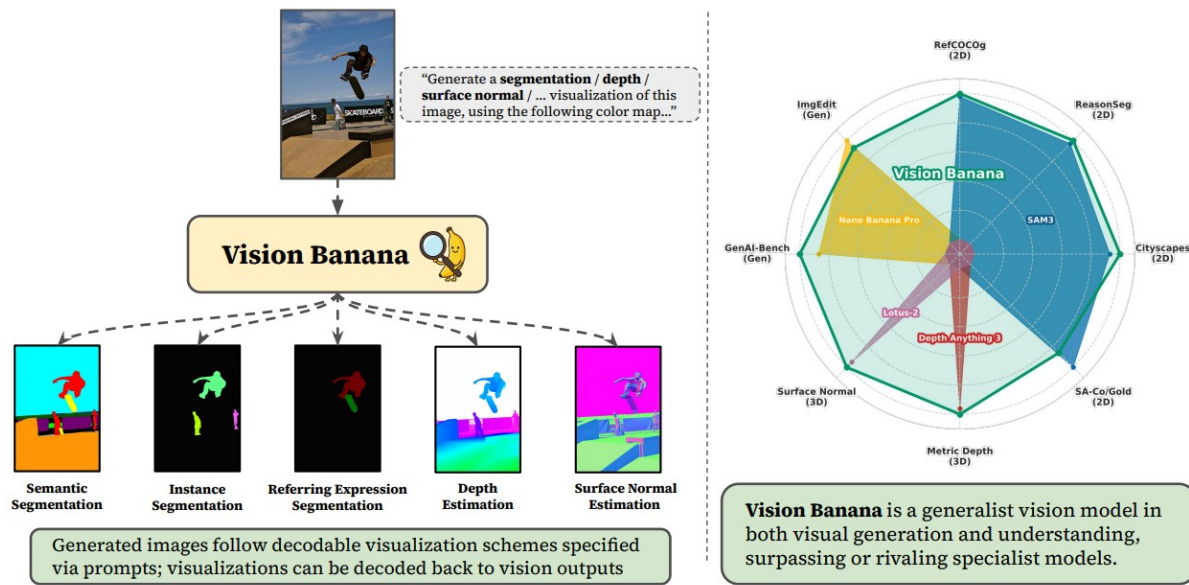


Figure 1 | We demonstrate the hidden visual understanding capabilities of image generators by instruction-tuning Nano Banana Pro. The instruction-tuned model, **Vision Banana**, can produce visualizations in a precise format that can enable evaluation on established benchmarks.

主要亮点

- **生成即理解**: 将分割、深度估计等视觉任务重构为图像生成问题, 共享 NBP 的生成预训练表示
- **超越专用模型**: 在多基准上超越 SAM 3 (分割专用 SOTA) 和 Depth Anything V3 (深度估计专用 SOTA)
- **范式挑战**: 证明图像生成预训练是通用视觉学习的有效基础, 无需为每类视觉任务设计专用架构

相关链接

Paper: <https://arxiv.org/abs/2604.20329>

Github: <https://vision-banana.github.io/>

Blog: <https://deepmind.google/research/publications/240658/>

TECHNICAL UPDATES

技术动态

● 技术动态一

Qwen3.5-Omni：新一代原生全模态大模型

日期

2025-03-29



内容摘要

通义千问正式发布新一代原生全模态大模型 Qwen3.5-Omni，彻底摒弃传统多模态拼接方案，在预训练阶段就对文本、图像、音频、视频四大模态进行统一建模，实现原生端到端全模态理解与生成，官方定位为迈向原生全模态 AGI 的关键里程碑。该系列推出 Plus、Flash、Light 三种规格，均支持 256K 超长上下文，可处理超 10 小时音频输入与 400 秒 720P 音视频输入；语音识别覆盖 113 种语言及方言，语音生成支持 36 种语言及方言，输出同时支持文本与语音双通道，适配智能助手、音视频理解、全模态 Agent 等场景。技术上采用 Hybrid-Attention MoE 混合注意力架构，基于超 1 亿小时音视频数据与海量图文数据训练，Plus 版本在 215 项音频与音视频任务上刷新 SOTA，综合表现超越 Gemini-3.1 Pro。原生的交互能力使其在实时语音助手、复杂视频内容解析以及全模态智能体任务中表现出极高的逻辑一致性与响应速度。

相关链接

Paper: <https://arxiv.org/abs/2604.15804>

Hugging Face: <https://huggingface.co/spaces/Qwen/Qwen3.5-Omni-Offline-Demo>

● 技术动态二

Qwen3.6-Plus: 性能直逼 Claude, 国产编程模型新高度

日期

2026-04-01



内容摘要

Qwen3.6-Plus 是面向真实世界 Agent 任务设计的旗舰模型，其在 SWE-bench、Terminal-Bench 2.0 和 NL2Repo 等权威代码与终端任务评测中性能已直逼 Claude Opus 4.5，跻身全球第一梯队。模型默认启用 1M token 超长上下文窗口，原生支持混合思维模式，可在深度推理（CoT）与快速响应间动态切换，高效适配不同任务的延迟与质量权衡需求。生态兼容性极强，支持 Anthropic API 协议直接接入 Claude Code，并对 OpenClaw、Qwen Code、Cline 等主流智能体框架深度优化，助力开发者低成本构建生产级 Agent 应用。Qwen3.6-Plus 以更小参数规模实现超越大参数量竞品的工程能力，推理成本仅为 Claude 的约 1/12，兼具高性能与高性价比。

相关链接

Blog: <https://qwen.ai/blog?id=qwen3.6>

● 技术动态三

Google Gemma 4 系列：开源多模态旗舰，31B 登顶开源前三



日期

2026-04-02

内容摘要

Google 正式发布 Gemma 4 系列四款开源多模态旗舰模型，全系采用商业友好的 Apache 2.0 许可协议，完整覆盖从数据中心到边缘设备的全场景部署需求。性能方面，31B 稠密版在 Arena AI 开源排行榜位列全球第三，26B MoE 版位列第六，以“单位参数智能”的突破性表现，在多项权威基准上超越远超自身规模的竞品。多模态能力上，E2B/E4B 边缘版原生支持音频输入，全系原生兼容图文理解、OCR 与图表解析，且基于 140 余种语言训练，具备强大的多语言处理能力。上下文窗口方面，31B/26B 支持 256K 超长上下文，E2B/E4B 支持 128K 上下文，可高效处理长文档、代码库与复杂对话。技术特性上，全系列内置可配置思维模式与原生函数调用能力，深度适配智能体工作流，支持多种主流框架灵活部署，为开发者提供低成本、高效率的前沿 AI 能力落地路径。

相关链接

Blog: <https://deepmind.google/models/gemma/gemma-4/>

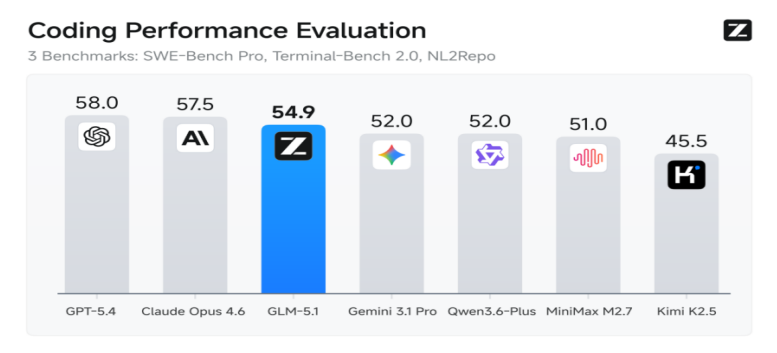
Doc: https://ai.google.dev/gemma/docs/core/model_card_4

● 技术动态四

GLM-5.1: Towards Long-Horizon Tasks

日期

2026-04-07



内容摘要

智谱 Z.AI 正式发布 GLM-5.1，定位为国产首个面向长程任务的旗舰大模型，核心突破是具备长达 8 小时的持续自主工作能力，可完成从任务规划、工具调用执行、多轮迭代优化到生产级结果交付的全闭环自主任务处理。模型采用 744B 总参数的 MoE 混合专家架构，每次推理仅激活约 40B 参数，兼顾强大能力与高效推理成本。在权威基准测试中，GLM-5.1 的 SWE-bench Pro 得分达 58.4 分，位居全球第一，综合代码能力全面对标 Claude Opus 4.6，成为首个在综合能力上对齐国际顶级闭源模型的国产大模型。技术层面内置自我诊断与修复机制，支持 200K 超长上下文窗口，强化长程任务中的目标一致性与错误修正能力，可自主编写 CUDA/Triton 内核、优化向量数据库性能，甚至从零构建完整 Linux 桌面系统。支持权重开放与 API 调用双模式，大幅降低开发者在长程智能体、复杂工程任务场景的使用门槛，为国产大模型迈向长周期自主智能体时代奠定核心基础。

相关链接

- Github: <https://github.com/zai-org/GLM-5>
- Hugging Face: <https://huggingface.co/zai-org/GLM-5.1>

● 技术动态五

Claude Managed Agents

日期

2026-04-08



内容摘要

Anthropic 在其最新的技术分享中详细阐述了 Managed Agents 的工程设计理念，核心是将模型“大脑”、执行沙箱与会话状态彻底解耦，采用 metaharness 元框架设计，统一负责编排、安全隔离、会话持久化、凭证管理与错误恢复，大幅降低企业构建与部署长时、多步骤、生产级 Agent 的门槛，让 Agent 从原型到上线周期从数月压缩至数天。该平台可适配 Claude Code、自定义工具链等各类任务执行环境，提供标准化 API 与开箱即用的安全沙箱、身份权限、执行追踪能力，支持数小时自主运行与断连后状态恢复，研究预览阶段已具备多 Agent 并行协作调度复杂流水线的能力。架构层面实现模型逻辑与底层基础设施分离，各层可独立演进升级，同时强化安全边界，将敏感凭证与代码执行环境物理隔离，减少注入攻击与泄露风险。此次发布标志 Anthropic 从单一模型与编码工具，转向面向企业的通用、可托管、高可靠 Agent 运行时基础设施，为复杂 Agent 工程化落地提供关键支撑。

相关链接

Github: <https://www.anthropic.com/engineering/managed-agents>

● 技术动态六

Nature 解读 Stanford 2026 AI Index: 复杂任务上人类科学家仍领先 AI Agent

日期

2026-04-14



内容摘要

Nature 基于斯坦福大学《2026 人工智能指数报告》发布评论文章，明确指出当前最强 AI 智能体在多步骤、高复杂度的真实科研任务上，表现仍显著落后于人类博士级专家，得分仅约为人类专家的一半，短任务上的亮眼表现无法直接外推到长期复杂科研场景。这份长达 423 页的综合年报覆盖技术能力、产业经济、科学应用、教育、政策治理与公众认知等多个维度，以海量数据客观呈现全球 AI 发展全貌，同时释放强烈警示：AI 技术能力与模型性能正飞速提升，但对应的治理、评估、安全监管与风险管理框架建设严重滞后，二者形成难以弥合的结构性失衡。报告也为 AutoSOTA、ASIEvolve 等自动化科研的过度乐观叙事提供了冷静的量化校准，强调 AI 科研落地仍需严谨验证与人类主导，为全球政策制定、产业投入与技术研发提供了权威的数据支撑与理性参考。

相关链接

Blog: https://hai.stanford.edu/assets/files/ai_index_report_2026.pdf

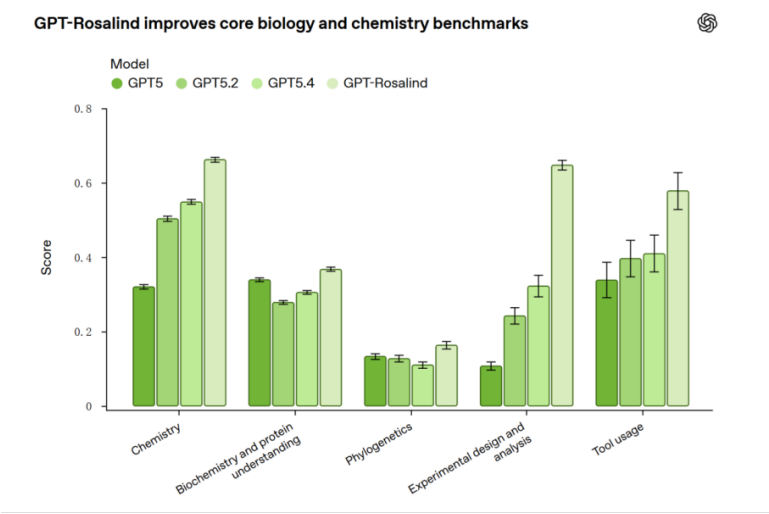
Blog: <https://www.nature.com/articles/d41586-026-01199-z>

● 技术动态七

GPT-Rosalind：面向生命科学研究的专用推理模型

日期

2026-04-16



内容摘要

OpenAI 正式发布 GPT-Rosalind，这是其首款面向生命科学领域的专用推理模型，以 DNA 结构研究先驱罗莎琳德·富兰克林命名，聚焦药物发现、基因组分析、蛋白质结构推理、转化医学与全流程科研工作流。模型并非通用聊天模型，而是针对生物化学、蛋白工程、基因组学做深度垂直优化，强化代码执行、专业数据库查询、结构可视化与文献整合等工具增强推理能力，可覆盖从文献调研、假设生成、实验设计到结果验证的完整科研闭环。GPT-Rosalind 以研究预览形式上线，仅对合规科研机构与药企开放可信访问，内置生物安全机制防范滥用风险，标志着通用大模型厂商正式深耕科学专用赛道，与 AlphaFold 等结构预测工具形成能力互补，加速 AI 驱动的生命科学突破。

相关链接

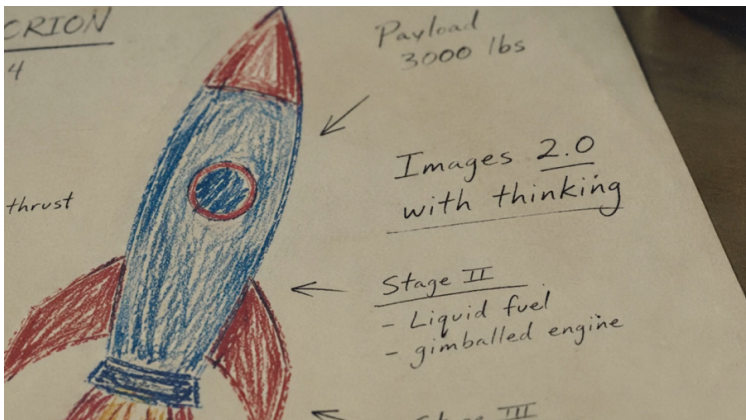
Blog: <https://openai.com/index/introducing-gpt-rosalind>

● 技术动态八

ChatGPT Images 2.0 发布

日期

2026-04-22



内容摘要

OpenAI 正式推出 ChatGPT Images 2.0，搭载全新升级的图像引擎，核心围绕排版渲染、推理式生成、联网能力与宽高比支持四大方向实现突破。模型对文字、小文本、UI 元素及复杂版式的生成精度大幅提升，可稳定输出排版规整、文字清晰的信息图、海报与演示页；新增的推理式图像生成支持多步逻辑规划后再渲染，显著提升复杂视觉任务的逻辑一致性。同时，Images 2.0 扩展互联网访问能力，可实时获取最新信息辅助创作，并支持从 3:1 到 1:3 的多种宽高比，适配多场景输出需求。性能跃升的同时，其高保真生成能力也引发广泛政策讨论，业界高度关注伪造文档、金融票据与身份材料的生成风险，促使 OpenAI 同步强化内容安全过滤机制。该功能已面向全量 ChatGPT 用户开放，付费用户可解锁“思考模式”等高级能力，标志着 AI 图像生成从创意实验工具向生产级视觉设计助手的重要跨越。

相关链接

Blog: <https://openai.com/index/introducing-chatgpt-images-2-0>

● 技术动态九

Introducing GPT-5.5

日期

2026-04-23

April 23, 2026 Product Release

Introducing GPT-5.5

A new class of intelligence for real work

内容摘要

OpenAI 发布新一代旗舰模型 GPT-5.5, 定位为 “面向真实工作与智能体驱动的新一代智能”, 包含 GPT-5.5 Thinking 与 GPT-5.5 Pro 两大版本。其中, Thinking 版主打日常复杂任务加速, 在保持高性能的同时优化响应效率; Pro 版面向专业工作场景, 提供更高输出质量与更低延迟, 适配严苛的企业级与科研级需求。GPT-5.5 在 GDPval 数据分析、OSWorld-Verified 计算机操作智能体、Tau2-bench Telecom 电信工具调用、FinanceAgent 金融智能体等多项知识工作与智能体基准测试中取得顶尖表现, 尤其擅长处理模糊、多步骤的复杂任务, 可自主完成从意图理解、工具调用、结果校验到迭代推进的全流程, 更适配多阶段科研 workflow。模型在智能体编程、长上下文理解与跨工具协作能力上实现显著跃升, 单 Token 延迟与前代 GPT-5.4 持平, 同时大幅降低同等任务下的 Token 消耗, 兼顾高性能与经济性。官方同步发布完整的 GPT-5.5 System Card, 详细披露模型部署细节、安全防护方案与红队测试结果, 涵盖高级网络安全与生物技术专项评估, 保障复杂场景下的安全可靠。

相关链接

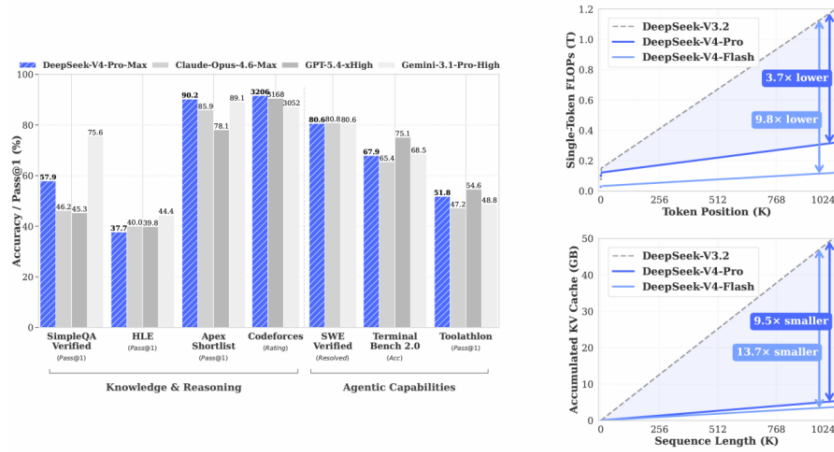
Blog: <https://openai.com/index/introducing-gpt-5-5>

● 技术动态十

DeepSeek V4 双版本正式上线：1.6T MoE，1M 上下文

日期

2026-04-24



内容摘要

DeepSeek 正式推出新一代旗舰大模型 DeepSeek V4，包含 V4Pro 与 V4Flash 双版本，全系采用 MoE 混合专家架构并原生支持 100 万 token 超长上下文。相比前代 V3.2，V4 在推理效率上实现颠覆性提升：推理 FLOPs 降低 73%、KV 缓存内存占用减少 90%，大幅降低长文本处理成本。模型提供 Nonthink / Think High / Think Max 三档推理模式，适配不同复杂度任务；同时引入 XML 格式工具调用，强化 Agent 能力，API 兼容 OpenAI 与 Anthropic 双协议，便于开发者无缝迁移。目前，V4 权重已在 Hugging Face 开源，旧版服务 deepseekchat 与 deepseekreasoner 将于 2026 年 7 月 24 日停止服务，全面切换至 V4 体系。V4Pro 主打高性能，面向复杂推理、Agent 编码与长文档分析等专业场景；V4Flash 侧重高效低成本，适合高频日常调用，标志着国产开源大模型正式迈入“百万上下文普惠时代”。

相关链接

Hugging Face: <https://huggingface.co/collections/deepseek-ai/deepseek-v4>

Doc: <https://api-docs.deepseek.com/news/news260424>

● 技术动态十一

小米开源 MiMo-V2.5-Pro / MiMo-V2.5 系列

日期

2026-04-27



内容摘要

小米发布并开源 MiMoV2.5 系列大模型, 包含 MiMoV2.5Pro 与 MiMoV2.5, 其中 MiMoV2.5Pro 为核心旗舰版本, 采用 MoE 混合专家架构, 总参数达 1.02T、激活 42B, 为全球首个开源突破 1T 参数的 MoE 大模型。模型基于混合注意力架构, 原生支持 1M token 超长上下文窗口, 可一次性处理百万字文档、完整代码库与长篇财报, 无需分段且逻辑连贯。MiMoV2.5Pro 深度优化 Agent 智能体能力、复杂软件工程与长程任务, 在 GDPValAA、ClawEval、SWEBench Pro 等权威评测中位列全球开源模型第一, 综合性能对标 Claude Opus 4.6、GPT5.4 等顶尖闭源模型。实测中可稳定完成近千轮连续工具调用的长周期任务: 例如 4.3 小时从零用 Rust 写出完整 SysY 编译器、11.5 小时生成 8192 行代码的全功能视频编辑器, 展现极强指令遵循与长程一致性。

相关链接

Blog: <https://mimo.xiaomi.com/mimo-v2-5-pro>

RESOURCE SHARING

资源分享

● 资源分享一

Claude Code & Harness 电子书

类型：电子书

介绍：围绕 Claude Code 与 Agent harness 工程实践打造的系统性电子书，完整覆盖 Agent 工具链设计、代码型 Agent 构建、harness 抽象层设计与工程化落地全流程。与《Dive into Claude Code》深度拆解论文、《Anthropic Managed Agents》官方架构文章共同构成 Claude Agent 工程三件套，深度解析 Claude Code 源码架构逻辑、98.4% 基础设施层与 1.6% AI 决策层的设计范式、七层安全机制、上下文压缩策略、子代理协作与权限体系，适合想要深度理解 Claude Code 底层架构、自主搭建企业级 Agent 运行框架的工程师与研究者学习参考。

资源链接

<https://github.com/Chachamaru127/claude-code-harness>

● 资源分享二

Stanford 2026 AI Index Report

类型：年度报告

介绍：由斯坦福大学 HAI 发布，423 页全球权威 AI 年度综合报告，覆盖技术进展、产业经济、社会应用、政策治理、公众态度、安全与伦理六大维度，提供公开数据集与可复现指标，是全球政策制定、科研与产业决策的核心参考依据。2026 年核心结论：AI 能力飞速突破，但治理、评估、安全与监管框架严重滞后，形成技术与制度的结构性失衡。报告同时揭示：顶尖模型在科研、推理基准接近 / 超越人类专家，中美头部模型性能差距缩至 2.7%，全球企业 AI 投资达 5817 亿美元，但模型透明度下滑、安全事件激增、公众焦虑与乐观并存，可与技术动态条目 6 的 Nature 解读内容结合阅读，理解 AI 在复杂科研任务上的真实能力边界。

资源链接

<https://hai.stanford.edu/ai-index/2026-ai-index-report>

● 资源分享三

《计算机科学与机器学习中的代数、拓扑学、微分学和优化理论》电子书

类型：电子书

介绍：宾夕法尼亚大学 Jean Gallier 与 Jocelyn Quaintance 合著，全书约 2200 页，是面向计算机科学与机器学习方向的超硬核数学理论专著。内容系统性覆盖线性代数、群论、代数拓扑、微分流形、凸优化与非凸优化等底层数学基础，从严谨数学结构出发，搭建完整理论体系，非常适合研究优化算法原理、流形学习、表示空间、高维几何建模与理论机器学习的科研人员与学生作为长期案头参考书。书籍免费在线获取，可按需查阅具体章节。

资源链接

<https://www.cis.upenn.edu/~jean/math-deep.pdf>

● 资源分享四

Project-Imaging-X



类型：数据集资源

介绍：对应学术动态条目 3：医学影像数据集综述的配套交互平台，Project-Imaging-X 索引超 1000 个医学影像开放数据集，支持按器官类别、CT/MRI/超声/病理等影像模态、分割/分类/检测等任务类型以及标注类型进行多维度精准筛选，同时配备自动化数据集集成工具。相较于单纯研读综述论文，该平台更贴合实际科研场景，能够帮助研究者快速匹配实验所需数据源、合理规划数据方案，并实时跟进最新公开医学影像数据集更新

资源链接

<https://github.com/uni-medical/Project-Imaging-X>

● 资源分享五

Awesome Latent Space

类型：GitHub 资源库

介绍：对应学术动态条目 6：隐空间相关综述的配套合集仓库，持续收录隐空间领域代表性研究论文，涵盖 COCONUT、SoftCoT、COLAR 等经典系列工作。仓库严格依照综述提出的基础、演化、机制、能力、展望五维分类框架整理文献，便于研究者系统梳理领域发展脉络、挖掘现有研究空白，快速检索同方向相关成果，适用于文献研读、科研选题与方向规划。

资源链接

<https://github.com/YU-deep/Awesome-Latent-Space>

● 资源分享六

Deep Agents Deploy: Claude Managed Agents 的开源替代方案



类型：代码仓库

介绍：LangChain 正式推出 Deep Agents Deploy，主打作为 Anthropic 托管智能体的开源替代方案。项目聚焦智能体部署与运行时底层架构，重点实现持久化执行环境、长周期任务状态留存、多步任务断点恢复等核心能力，区别于传统仅做提示词编排的轻量化框架。可与技术动态条目 3 的 Anthropic 托管智能体相关技术动态对照学习，清晰对比闭源商业产品与开源框架在安全策略、沙箱隔离、运行时抽象设计上的差异，为自研搭建私有化智能体平台提供重要参考依据

资源链接

<https://github.com/YU-deep/Awesome-Latent-Space>

● 资源分享七

Awesome Agentic World Modeling

类型：GitHub 资源库

介绍：LangChain 正对应学术动态条目 17 : Agentic World Modeling 综述的配套资源库，收录 400 余篇高引用论文与 100 余个代表性研究系统，依托 levels × laws 专属分类框架规整内容。资源全面覆盖基于模型强化学习、视频生成、网页与图形界面智能体、多智能体社会模拟、AI 科研发现五大核心方向，还配套提供可复现的标准化最小评测套件。适合研究者持续追踪环境建模、闭环决策、仿真模拟器及智能体评测领域前沿，系统性梳理想能体世界建模整体研究版图。

资源链接

<https://github.com/matrix-agent/awesome-agentic-world-modeling>



深圳软牛科技集团股份有限公司成立于 2014 年 6 月，作为国家级专精特新“小巨人”企业，始终以“让创意成为现实”为使命，专注于通过人工智能视觉大模型技术驱动全球数字创意生态。

公司深度聚焦图像修复、视频增强、多模态内容生成等视觉大模型核心技术，联合西安电子科技大学、香港中文大学（深圳）等顶尖高校，攻克了视频抠像、画质修复等难题，并与深圳大学共建“人工智能技术联合创新中心”，推动 AI 赋能的图像增强技术规模化落地。

目前，软牛科技能够为广电机构、影视制作、工业检测、农业测报等多个行业提供专业解决方案。从微米级工业精密测量，到食品异物智能检测；从农业害虫 AI 识别防治系统，到 AR 设备巡检实时画面 AI 增强与场景识别；从 VR 全息空间建模助力文博展陈数字化升级，到跨终端智能影像协作软件的全链路画质优化引擎，软牛科技正以技术革新重新定义新质生产力。

在消费市场，旗下牛小影（<https://www.niuxuezhang.cn/video-enhancer.html>）、HitPaw 等海内外知名品牌，以“一键式 AI 视觉创意”为核心，为全球超 1 亿用户提供视频修复、老电影 AI 上色等场景化服务。例如，牛小影支持 4K 或 8K 超高清修复技术，帮助家庭用户重现祖辈黑白影像的生动细节；HitPaw Edimakor 通过 AI 语音合成、多语言实时翻译等功能，让自媒体创作者轻松实现跨语言内容生产。



这些创新成果的背后，是公司每年近亿元的研发投入和占比超 50% 的顶尖技术团队，以及 CMMI、ISO56005 等国际权威认证体系支撑的硬核实力。立足深圳总部，我们通过北京、新加坡、香港等战略支点构建起辐射全球 200 多个国家和地区的智慧服务网络。作为高新技术企业、深圳市潜在科技独角兽，软牛科技始终以“技术无界”为理念，持续拓展 AIGC+ 产业应用的边界。

未来，软牛科技将加速推进物理世界与 AI 视觉的无缝衔接，让每个人都能通过指尖的艺术级创作，开启属于自己的数字时代。

人工智能前沿快报

Artificial Intelligence
Newslettler



2026 年第 4 期

广东省图象图形学会

总第 31 期

主办

广东省图象图形学会

本期编委

金连文	华南理工大学
谭明奎	华南理工大学
钟亦武	香港中文大学
林上港	华南农业大学
李 斌	深圳大学
谢晓华	中山大学
王泽宇	香港科技大学 (广州)
李冠彬	中山大学
熊龙飞	金山办公
段纪伟	金山办公
李坤桃	金山办公
唐 瑞	华南理工大学
张 顺	华南理工大学
许毅平	华南农业大学
蔡沐含	华南农业大学

人工智能 前沿快报

2026 年第 4 期
总第 31 期
2026.04.30

声明：

- (1) 本快报内容提要、文章亮点等部分内容由 AI 生成，不一定准确及全面，章完整内容及思想应以原文为准。
- (2) 本文大部分插图来源于相关论文或文章，版权归原作者或原出版社所有。
- (3) 本快报摘录文章观点不代表编委会及主办方立场。



学会官方公众号

广东省图象图形学会
GDSIG

学术交流、技术咨询、科学普及