Artificial Intelligence Newsletter

人工智能前沿快报

2026 年第 03 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

目 录

技 术 动 态

资 源 分 享

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，促进学术交流和技术繁荣。本期摘选了近期全球人工智能领域的 13 篇最新学术动态、4 篇技术动态、以及 2 个资源推荐给广大读者。

ACADEMIC TRENDS
学术动态

● 论文一

OCRGenBench: A Comprehensive Benchmark for Evaluating OCR Generative Capabilities

日期

2025-07-25

OCRGenBench: A Comprehensive Benchmark for Evaluating OCR Generative Capabilities

Peirong Zhang[1], **Haowei Xu**[1], **Jiaxin Zhang**[1], **Xuhan Zheng**[1], **Guitao Xu**[1], **Yuyi Zhang**[1], **Junle Liu**[1], **Zhenhua Yang**[1], **Wei Zhou**[2], **Lianwen Jin**[1*]
[1] South China University of Technology, Guangzhou 510641, China
[2] Cardiff University, Cardiff CF24 4AG, UK
eeprzhang@mail.scut.edu.cn, eelwjin@scut.edu.cn

内容摘要

OCRGenBench 关注的是生成式模型的 **OCR generative capabilities**，也就是“视觉文本生成与编辑能力”的系统评测。作者指出，现有基准通常只评估海报或场景文本，只看文生图或文本编辑中的某一类任务，而且样本难度有限，无法真实反映模型在复杂视觉文本场景下的表现。为此，论文首次统一了 **text-centric T2I generation、text editing、OCR 相关图像到图像转换** 等任务，形成覆盖 **5 大文本类别、33 个 OCR 生成任务** 的综合评测集。作者还设计了统一指标 **OCRGenScore**，综合衡量文本准确性、指令遵循、美学质量等多个维度。实验发现，即使是前沿生成模型，在该 benchmark 上多数得分仍低于 60/100，说明视觉文本生成仍远未成熟。

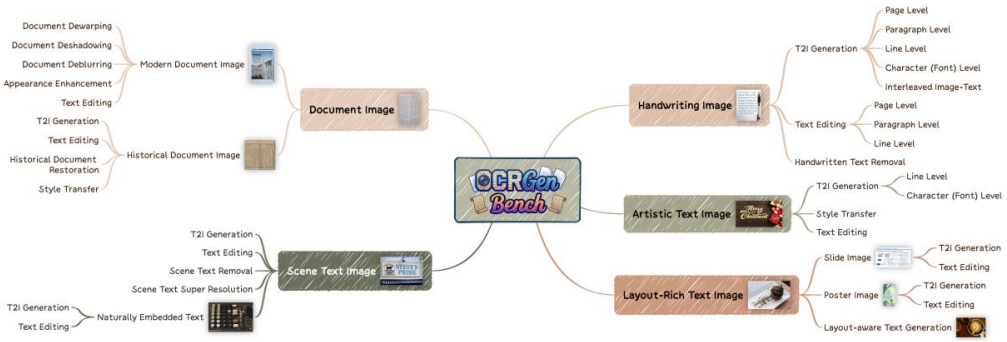


Figure 1: Task categorization and subordination of OCRGenBench. It includes 33 OCR generative tasks under five text categories: **document**, **handwriting**, **scene text**, **artistic text**, and **layout-rich text**. Each primary category encompasses multiple sub-tasks.

主要亮点

- **把 OCR 生成能力当成统一能力来评估**：不再把文生图、文字编辑、文档修复等割裂看待。
- **基准覆盖全面且难度高**：33 个任务、1060 个高质量人工标注样本，包含高文本密度、复杂版式、多比例与双语内容。
- **提出统一指标 OCRGenScore**：同时评估文字正确性、指令遵循、视觉质量与结构一致性，更符合真实使用场景。

相关链接

Paper： https://arxiv.org/abs/2507.15085

Code: https://github.com/NiceRingNode/Awesome-Generative-Models-for-OCR

● 论文二

A Mixed Diet Makes DINO An Omnivorous Vision Encoder

日期

2026-02-27（CVPR 2026 录用）

A Mixed Diet Makes DINO An Omnivorous Vision Encoder

Rishabh Kabra[1,2] Maks Ovsjanikov[1] Drew A. Hudson[1] Ye Xia[1] Skanda Koppula[1,2]
Andre Araujo[1] Joao Carreira[1] Niloy J. Mitra[2]

[1]Google DeepMind [2]University College London

{rkabra,movsani,dorarad,yexia,skandak,andrearaujo,joaoluis}@google.com n.mitra@ucl.ac.uk

内容摘要

该论文指出，像 DINOv2 这样的视觉基础编码器虽然在单一模态视觉任务上表现很强，但其特征空间并不能天然对齐 RGB、深度图、分割图等不同模态：同一场景的 RGB 图像与深度图在嵌入空间中的相似度，甚至接近随机样本之间的相似度。为此，作者提出 **Omnivorous Vision Encoder（OVE）**，通过“双目标训练”让模型学到模态无关的统一表示：一方面最大化同一场景不同模态之间的特征对齐，另一方面使用冻结的 DINOv2 教师进行蒸馏，保留原始视觉基础模型的判别语义。最终，OVE 可以对 RGB、Depth、Segmentation 等输入给出一致而强健的场景表示，让 DINO 系列从“单模态强编码器”变成“跨模态通用视觉编码器”。

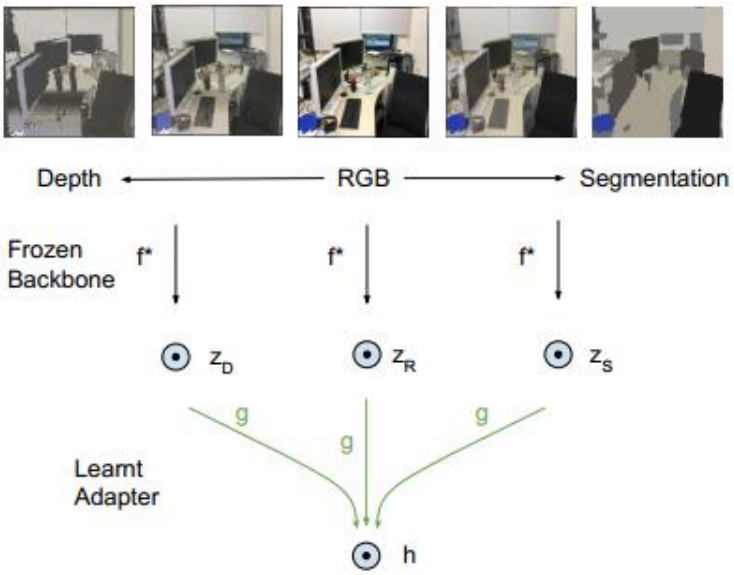


Figure 2. **Omnivorous Vision Encoder architecture.** A frozen encoder f^* extracts features $z_m = f^*(x_m)$ from a spectrum of modalities denoted m (Segmentation, RGB, Depth). A trainable modality-agnostic adapter g maps these features into a common, aligned embedding space, producing a modality-invariant representation $h = g(z_m)$. A convenient implementation of this architecture uses the early layers of a pretrained network as the frozen part f^*, and the later layers as the adapter g.

主要亮点

- **把 DINOv2 扩展成统一多模态编码器：**核心目标不是重新训练一个大模型，而是在保留 DINOv2 语义能力的基础上，把 RGB、深度、分割等输入映射到统一特征空间。
- **对齐 + 蒸馏的双目标设计：**既约束同场景跨模态特征对齐，又用冻结教师模型锚定表示质量，兼顾跨模态一致性与单模态判别能力。
- **适合多传感器视觉场景：**这类统一嵌入更适合机器人、3D 感知、视觉-几何联合建模等任务，为“一个编码器处理多种视觉输入”提供了简洁路线。

相关链接

Paper: https://arxiv.org/abs/2602.24181

● 论文三

DARE-bench: Evaluating Modeling and Instruction Fidelity of LLMs in Data Science

日期

2026-02-27（ICLR 2026 录用）

DARE-BENCH: EVALUATING MODELING AND INSTRUCTION FIDELITY OF LLMS IN DATA SCIENCE

Fan Shu[1,*] **Yite Wang**[2,*†] **Ruofan Wu**[1] **Boyi Liu**[2]
Zhewei Yao[2] **Yuxiong He**[2] **Feng Yan**[1]
[1]University of Houston [2]Snowflake AI Research

内容摘要

DARE-bench 面向数据科学 Agent 的评测痛点提出了一个标准化、可验证的新基准。作者认为，现有 LLM/Agent 基准往往缺少对“过程是否正确、是否遵循任务指令”的精细评估，而且训练数据也缺少高质量、可验证标注。为此，论文构建了 **6300 个来源于 Kaggle 的数据科学任务**，覆盖建模与数据科学工作流中的关键环节，并确保每个任务都有可验证的标准答案，从而避免依赖人工或模型打分器带来的主观性。实验显示，即便是能力很强的模型，在这类真实数据科学任务上仍然表现不佳；而基于 DARE-bench 进行监督微调或强化学习后，模型性能可大幅提升，说明该基准既能评测，也能反向驱动训练。

Figure 1: DARE-bench defines each task by providing a natural-language question and structured files (metadata and train/test splits). An LLM agent executes code within a sandbox to generate predictions, which are compared against ground truth for automatic and reproducible evaluation.

主要亮点

- **首个面向数据科学 Agent 的大规模可验证基准**：任务来自 Kaggle，贴近真实数据科学工作流，而不是纯文本问答。
- **强调 instruction fidelity 和 process-aware evaluation**：不只看最后答案，还关注模型是否按要求执行建模和分析步骤。
- **同时可作评测集与训练数据源**：论文展示了 SFT 和 RL 都能显著提升模型在该基准上的表现，说明它具备“ benchmark + curriculum ”双重价值。

相关链接

Paper：https://arxiv.org/abs/2602.24288

● 论文四

AI+HW 2035: Shaping the Next Decade

日期

2026-03-05

AI+HW 2035: Shaping the Next Decade

Deming Chen[1], Jason Cong[2], Azalia Mirhoseini[3], Christos Kozyrakis[4], Subhasish Mitra[3], Jinjun Xiong[5], Cliff Young[6], Anima Anandkumar[7], Michael Littman[8], Aron Kirschen[9], Sophia Shao[10], Serge Leef[11], Naresh Shanbhag[1], Dejan Milojicic[12], Michael Schulte[13], Gert Cauwenberghs[14], Jerry M. Chow[15], Tri Dao[16,17], Kailash Gopalakrishnan[18], Richard Ho[19], Hoshik Kim[20], Kunle Olukotun[3,21], David Z. Pan[22], Mark Ren[23], Dan Roth[24,25], Aarti Singh[26], Yizhou Sun[2], Yusu Wang[14], Yann LeCun[27], and Ruchir Puri[15]

[1]University of Illinois Urbana-Champaign, [2]University of California, Los Angeles, [3]Stanford University, [4]NVIDIA, [5]State University of New York at Buffalo, [6]Google, [7]California Institute of Technology, [8]Brown University, [9]SEMRON, [10]University of California, Berkeley, [11]Synopsys, [12]Hewlett Packard Labs, [13]Advanced Micro Devices (AMD), [14]University of California, San Diego, [15]IBM Research, [16]Princeton University, [17]Together AI, [18]EnCharge AI, [19]OpenAI, [20]SK Hynix, [21]SambaNova Systems, [22]The University of Texas at Austin, [23]Agentrys, [24]Oracle AI, [25]University of Pennsylvania, [26]Carnegie Mellon University, [27]New York University

Corresponding Authors: Deming Chen, dchen@illinois.edu; Ruchir Puri, ruchir@us.ibm.com.

内容摘要

这是一篇 AI 与硬件协同演进的十年路线图式愿景论文，由 Yann LeCun、Anima Anandkumar 等二十多位学者联合撰写。论文的核心观点是：未来 AI 的关键不只是继续“堆算力、堆参数”，而是要在 **算法、架构、系统、软件抽象与可持续性** 上开展跨层协同设计，把目标从单纯提升模型能力转向显著提升 **intelligence per joule（单位能耗智能）**。作者提出，到未来十年应争取实现训练与推理效率 1000 倍提升、让 AI 系统具备云-边-物理世界一体化部署能力，并通过共享基础设施、产业与政府协作、人才培养等方式形成长期国家级与全球级协同推进机制。它更像一份 AI+HW 的“共同宣言”，强调未来 AI 的突破必须建立在系统级协同优化之上。

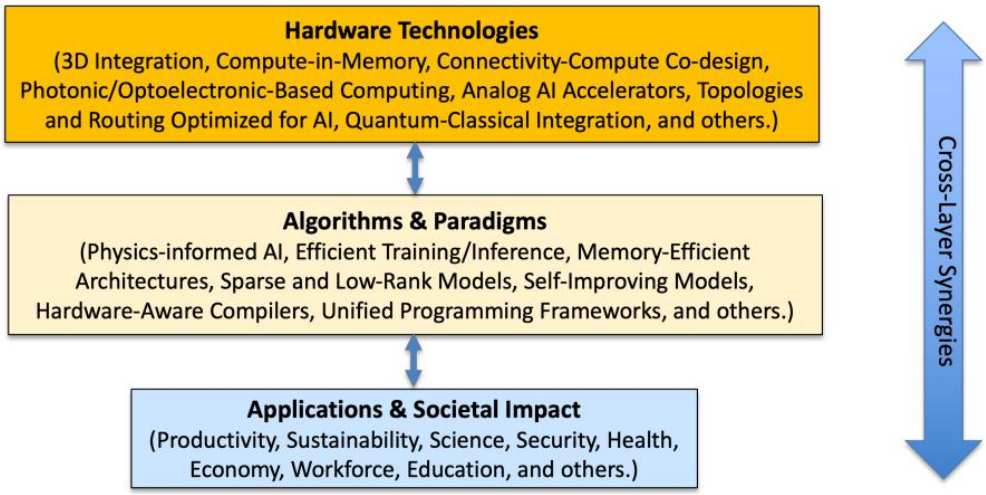


Figure 1: A Multi-Layered Vision for AI+HW Co-Design and Co-Evolution

主要亮点

- **从“模型扩张”转向“系统级协同优化”**：强调 AI 与硬件不应各自演进，而要围绕能效、部署和可持续性联合设计。
- **提出清晰的十年目标**：包括训练/推理效率 1000× 提升、跨云边端一体化、普惠化基础设施等，具备很强的战略路线图意味。
- **兼具学术与政策倡议色彩**：不只讨论技术，还提出政府、产业、学界如何共同建设共享平台、人才体系与公共投入机制。

相关链接

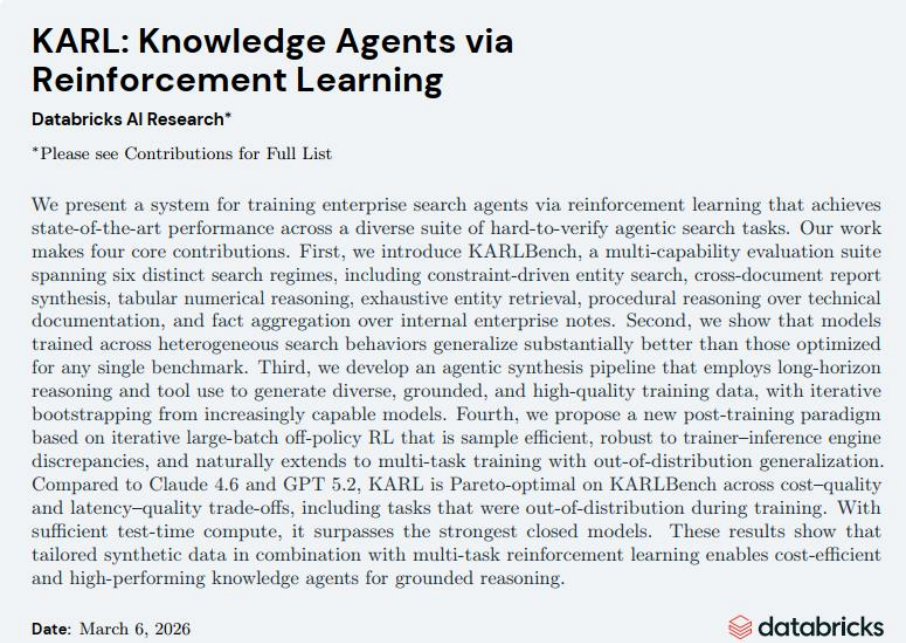
Paper：https://arxiv.org/abs/2603.05225

● 论文五

KARL: Knowledge Agents via Reinforcement Learning

日期

2026-03-05



KARL: Knowledge Agents via Reinforcement Learning

Databricks AI Research*

*Please see Contributions for Full List

We present a system for training enterprise search agents via reinforcement learning that achieves state-of-the-art performance across a diverse suite of hard-to-verify agentic search tasks. Our work makes four core contributions. First, we introduce KARLBench, a multi-capability evaluation suite spanning six distinct search regimes, including constraint-driven entity search, cross-document report synthesis, tabular numerical reasoning, exhaustive entity retrieval, procedural reasoning over technical documentation, and fact aggregation over internal enterprise notes. Second, we show that models trained across heterogeneous search behaviors generalize substantially better than those optimized for any single benchmark. Third, we develop an agentic synthesis pipeline that employs long-horizon reasoning and tool use to generate diverse, grounded, and high-quality training data, with iterative bootstrapping from increasingly capable models. Fourth, we propose a new post-training paradigm based on iterative large-batch off-policy RL that is sample efficient, robust to trainer–inference engine discrepancies, and naturally extends to multi-task training with out-of-distribution generalization. Compared to Claude 4.6 and GPT 5.2, KARL is Pareto-optimal on KARLBench across cost–quality and latency–quality trade-offs, including tasks that were out-of-distribution during training. With sufficient test-time compute, it surpasses the strongest closed models. These results show that tailored synthetic data in combination with multi-task reinforcement learning enables cost-efficient and high-performing knowledge agents for grounded reasoning.

Date: March 6, 2026

databricks

内容摘要

KARL 聚焦企业知识检索与复杂信息综合任务，提出了一个通过强化学习训练 **knowledge agent** 的完整体系。论文的四个核心贡献分别是：构建涵盖六类搜索与推理行为的评测套件 KARLBench；证明跨多种搜索行为联合训练比单一基准优化更具泛化能力；搭建利用长程推理和工具调用来自动生成高质量训练数据的合成数据流水线；以及提出一种 **迭代式、大批量、off-policy 的后训练 RL 范式**，兼顾样本效率、训练稳定性和多任务泛化。作者报告称，在成本-质量与延迟-质量的权衡上，KARL 相比 Claude 4.6、GPT 5.2 等模型达到 Pareto 最优，并在足够测试时算力下超过最强闭源模型，说明面向“知识搜索与综合”这一垂类能力，定制化 RL 训练非常有效。

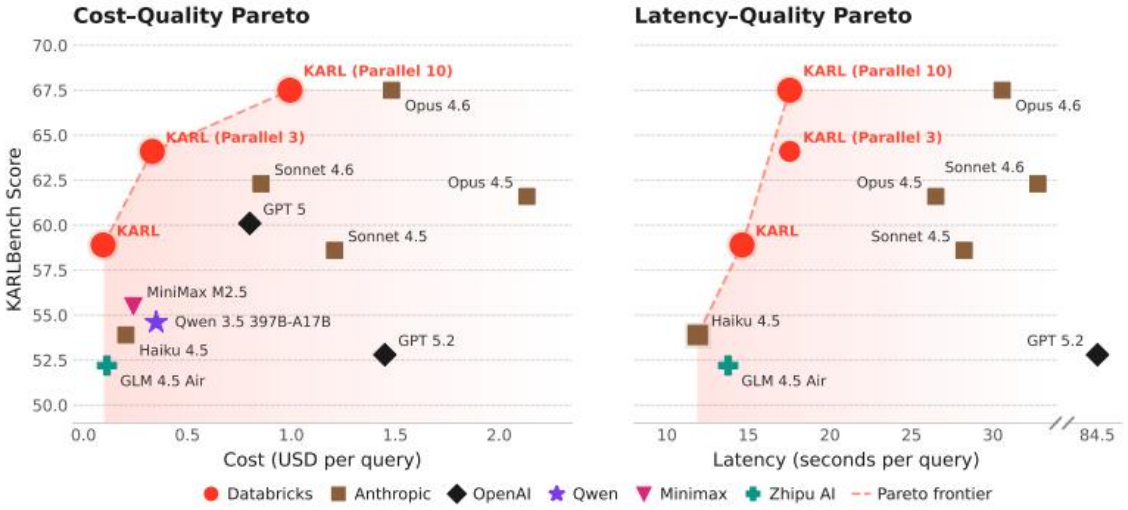


Figure 1 Performance of KARL, with and without test-time compute, compared to state-of-the-art agentic models on KARLBench. The cost–quality and latency–quality Pareto frontiers show that KARL achieves favorable trade-offs over existing models. All while being more cost and latency effective, KARL exceeds the quality of Sonnet 4.6 with three parallel rollouts and matches the best model, Opus 4.6, with ten parallel rollouts. See experiment details in Appendix B.

主要亮点

- **KARLBench 评测维度很全：**覆盖约束检索、跨文档综合、表格数值推理、技术文档程序性推理等六类高难任务。
- **强调多行为联合训练与 OOD 泛化：**不是只对单任务刷榜，而是让知识代理在多种搜索范式下都能稳健工作。
- **提出面向知识 Agent 的系统化 RL 路线：**合成数据 + 多任务训练 + off-policy RL 的组合，对企业搜索类代理很有参考价值。

相关链接

Paper：https://arxiv.org/abs/2603.05218

● 论文六

The Spike, the Sparse and the Sink: Anatomy of Massive Activations and Attention Sinks

日期

2026-03-05

The Spike, the Sparse and the Sink:
Anatomy of Massive Activations and Attention Sinks

Shangwen Sun [1] Alfredo Canziani [1] Yann LeCun [1] Jiachen Zhu [1]

内容摘要

这篇论文系统研究了 Transformer 中两个反复出现但尚未被充分解释的现象：**massive activations（少量 token 在少数通道上出现极端激活）** 与 **attention sinks（某些 token 无论语义是否相关都吸引大量注意力）**。作者通过实验发现，两者的共同出现很大程度上并非“模型学到了某种统一机制”，而是现代 Transformer 架构设计带来的结果，尤其与 **pre-norm 配置** 密切相关。进一步地，论文区分了二者的功能：massive activations 更偏全局，会形成跨层持续存在、近似恒定的隐藏表示，像是隐式参数；attention sinks 更偏局部，用于调制各头的注意力输出并偏向短程依赖。该工作把过去一些观察性现象拆解成了更清晰的机制分析，对理解 Transformer 内部表征和改进架构设计都很有价值。

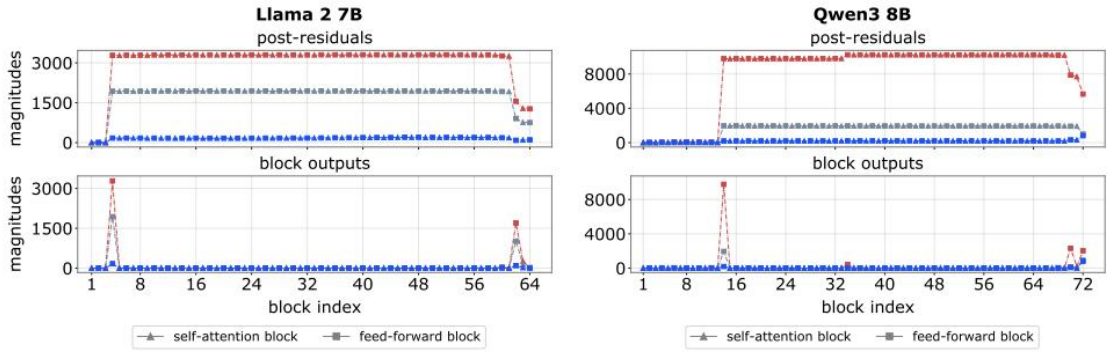


Figure 1. **Top-3 channel magnitudes across depth in Llama 2 7B and Qwen3 8B (post-residuals vs. block outputs).** In both models, early blocks inject massive activations that persist through most of the network before being neutralized by late blocks.

主要亮点

- **首次系统拆分两种常见异常现象的功能角色：**出 massive activations 偏全局、attention sinks 偏局部，不应简单混为一谈。
- **揭示 pre-norm 是二者共现的关键架构因素：**这为后续改进 Transformer 残差与归一化设计提供了明确抓手。
- **把“现象观察”推进到“因果解释”：**论文不只报告异常，而是通过消融证明两者可被解耦，提升了对模型内部机制的可解释性。

相关链接

Paper：https://arxiv.org/abs/2603.05498

● 论文七

GLM-OCR Technical Report

日期

2026-03-11

GLM-OCR Technical Report

Shuaiqi Duan[*1] **Yadong Xue**[*1] **Weihan Wang**[*1] **Zhe Su**[1] **Huan Liu**[1]
Sheng Yang[1] **Guobing Gan**[1] **Guo Wang**[1] **Zihan Wang**[1] **Shengdong Yan**[1]
Dexin Jin[1] **Yuxuan Zhang**[1] **Guohong Wen**[1] **Yanfeng Wang**[1] **Yutao Zhang**[1]
Xiaohan Zhang[1] **Wenyi Hong**[1] **Yukuo Cen**[1] **Da Yin**[1] **Bin Chen**[1]
Wenmeng Yu[†1] **Xiaotao Gu**[†1] **Jie Tang**[2]

[1]Zhipu AI [2]Tsinghua University
[*]Equal contribution [†]Project leader

Code: https://github.com/zai-org/GLM-OCR
Model: https://huggingface.co/zai-org/GLM-OCR
Demo: https://ocr.z.ai/

内容摘要

GLM-OCR 是一个仅 **0.9B 参数** 的紧凑型多模态 OCR 模型，目标是兼顾真实业务场景中的识别质量、推理效率与部署成本。模型由 **0.4B 的 CogViT 视觉编码器** 和 **0.5B 的 GLM 语言解码器** 构成，并针对 OCR 这类偏确定性的生成任务引入 **Multi-Token Prediction (MTP)** 机制，使模型在每一步预测多个 token，从而显著提升解码吞吐。系统层面，作者采用两阶段方案：先由 PP-DocLayout-V3 做版面分析，再对区域进行并行识别。论文显示，该方案在文档解析、公式识别、表格结构恢复、关键信息抽取等任务上达到有竞争力甚至 SOTA 的效果，同时仍适合边缘部署和大规模生产环境。

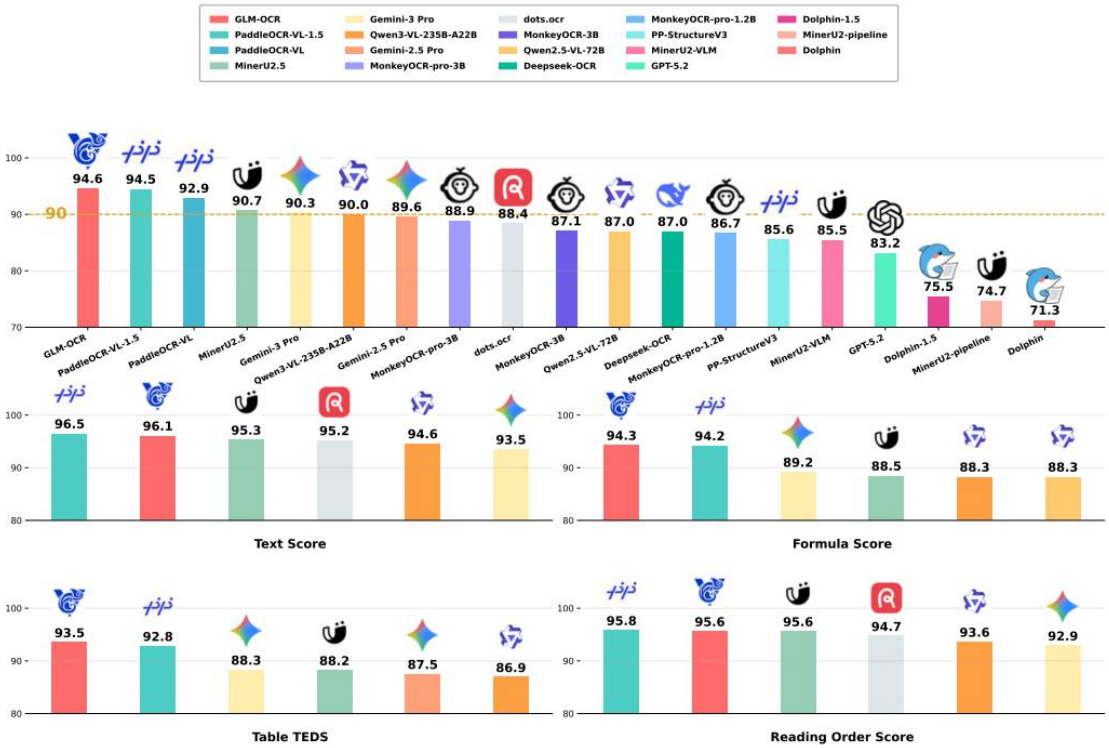


Figure 1: Performance of GLM-OCR on OmniDocBench v1.5.

主要亮点

- **小模型做强 OCR**：仅 0.9B 参数，但在多类文档理解任务上表现突出，适合资源受限或高并发场景。
- **MTP 提高 OCR 解码效率**：针对 OCR 的确定性输出特征，用 multi-token 预测替代传统慢速逐 token 解码，很有工程价值。
- **两阶段工业化流水线设计**：版面分析 + 区域并行识别，兼顾复杂文档解析效果与线上落地效率。

相关链接

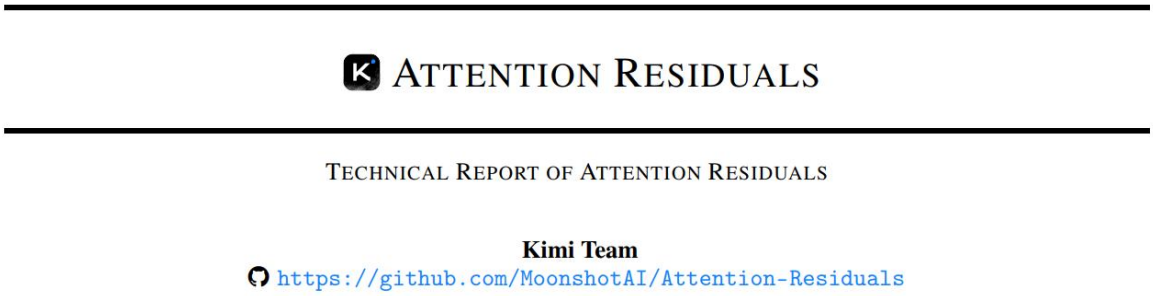
Paper：https://arxiv.org/abs/2603.10910
Code：https://github.com/zai-org/GLM-OCR

● 论文八

Attention Residuals

日期

2026-03-17



内容摘要

Attention Residuals（AttnRes）提出用“沿深度方向做注意力聚合”来替代 Transformer 中标准的残差连接。传统 PreNorm 残差会把所有前层输出以固定单位权重不断累加，导致随着层数增加隐藏状态幅值失控、每一层贡献被逐步稀释。AttnRes 的想法是：让当前层对历史层输出进行 **softmax attention**，以学习到的、输入相关的权重选择性聚合早期表示，从而避免统一叠加带来的稀释问题。为了让这一思路能用于大模型，作者又提出 **Block AttnRes**，把层划分成块，仅在块级表示上做注意力，显著降低内存与通信开销。实验表明，AttnRes 在多种计算预算下都优于标准残差，并在推理、代码与中文任务上带来稳定收益。

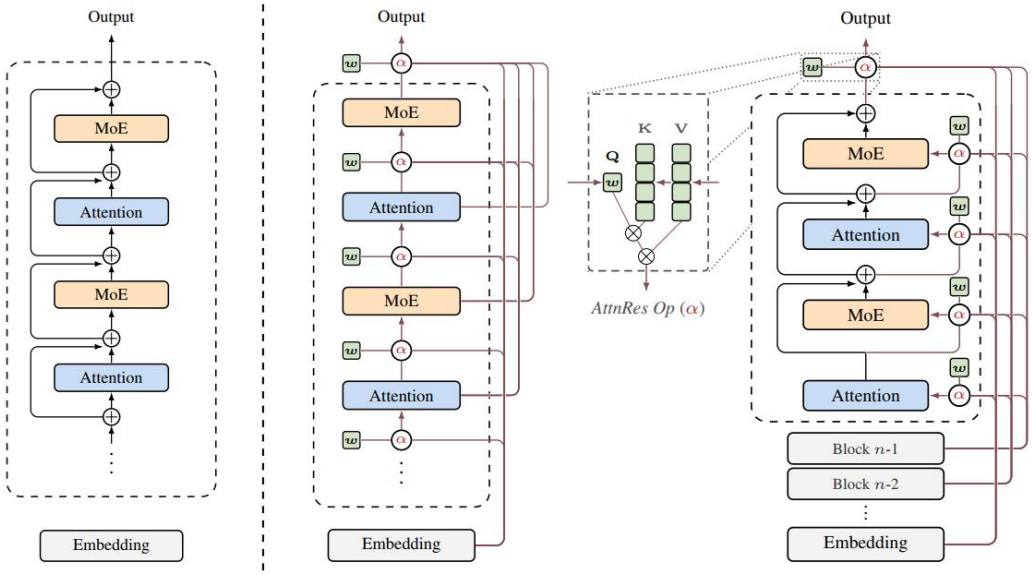


Figure 1: Overview of Attention Residuals. (a) Standard Residuals: standard residual connections with uniform additive accumulation. (b) Full AttnRes: each layer selectively aggregates all previous layer outputs via learned attention weights. (c) Block AttnRes: layers are grouped into blocks, reducing memory from $O(Ld)$ to $O(Nd)$.

主要亮点

- **直接改造残差连接这一 Transformer 基础组件：**不是局部 patch，而是从深度聚合机制上重写 PreNorm 的累积方式。
- **Block AttnRes 让方案可扩展到大模型训练：**通过块级近似控制内存与通信成本，使其具备实用性。
- **同时带来机制改进与实际增益：**论文不仅解释了 PreNorm 稀释问题，还在 GPQA、HumanEval、C-Eval 等任务上展示了稳定提升。

相关链接

Paper： https://arxiv.org/abs/2603.15031

Code: https://github.com/MoonshotAI/Attention-Residuals

● 论文九

From Masks to Pixels and Meaning: A New Taxonomy, Benchmark, and Metrics for VLM Image Tampering

日期

2026-03-20

From Masks to Pixels and Meaning: A New Taxonomy, Benchmark, and Metrics for VLM Image Tampering

Xinyi Shang[1,2,⋆], **Yi Tang**[1,⋆], **Jiacheng Cui**[1,⋆], **Ahmed Elhagry**[1], **Salwa K. Al Khatib**[1], **Sondos Mahmoud Bsharat**[1], **Jiacheng Liu**[1], **Xiaohan Zhao**[1], **Jing-Hao Xue**[2], **Hao Li**[1], **Salman Khan**[1], **Zhiqiang Shen**[1,†]

[1]Mohamed bin Zayed University of Artificial Intelligence, [2]University College London
[⋆]Equal Contribution, [†]Corresponding author

Existing tampering detection benchmarks largely rely on object masks, which severely misalign with the true edit signal: many pixels inside a mask are untouched or only trivially modified, while subtle yet consequential edits outside the mask are treated as natural. We reformulate VLM image tampering from coarse region labels to a pixel-grounded, meaning and language-aware task. First, we introduce a taxonomy spanning edit primitives (replace/remove/splice/inpaint/attribute/colorization, etc.) and their semantic class of tampered object, linking low-level changes to high-level understanding. Second, we release a new benchmark with per-pixel tamper maps and paired category supervision to evaluate detection and classification within a unified protocol. Third, we propose a training framework and evaluation metrics that quantify pixel-level correctness with localization to assess confidence or prediction on true edit intensity, and further measure tamper meaning understanding via semantics-aware classification and natural language descriptions for the predicted regions. We also re-evaluate the existing strong segmentation/localization baselines on recent strong tamper detectors and reveal substantial over- and under-scoring using mask-only metrics, and expose failure modes on micro-edits and off-mask changes. Our framework advances the field from masks to pixels, meanings and language descriptions, establishing a rigorous standard for tamper localization, semantic classification and description. Code and benchmark data are available at https://github.com/VILA-Lab/PIXAR.

Correspondence: Zhiqiang Shen (Zhiqiang.Shen@mbzuai.ac.ae)

内容摘要

这篇论文批评了现有图像篡改检测基准过度依赖粗粒度对象 mask：很多 mask 内像素其实没有被改动，而不少真正关键的细微修改却发生在 mask 外部，导致评测与真实编辑信号严重错位。为此，作者把任务重新定义为一个 **像素级、语义级、语言感知的 tampering 理解问题**。论文首先提出新的编辑分类体系，覆盖 replace、remove、splice、inpaint、attribute、colorization 等多种编辑原语及其语义对象类别；然后发布带 **逐像素 tamper map** 和类别监督的新基准；再提出训练与评测框架，用于同时衡量像素级定位、语义分类和自然语言描述能力。该工作把“是否篡改”这一任务从简单分割推进到更完整的视觉-语义理解框架。

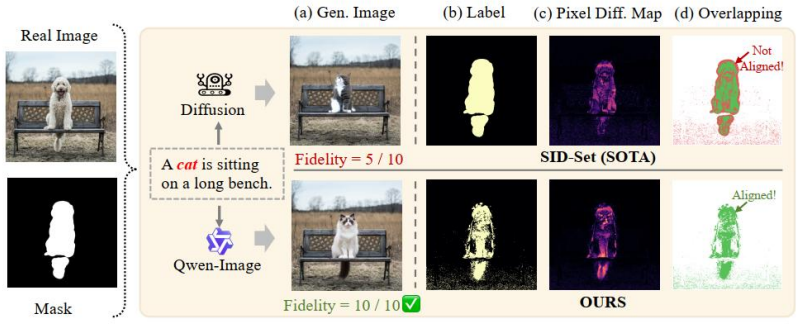


Figure 1 Pitfalls of Current Benchmarks and Our Remedy. (a) They contain unrealistic samples. (b)–(d) Their widely adopted mask-based label contains large regions of not aligned pixels with the true generative pixels. In contrast, our pixel label is precisely aligned with the true generative pixels.

主要亮点

- **从 mask 评测升级到 pixel-grounded 评测**：更准确地反映真实编辑位置与编辑强度，减少传统评测的误判。
- **统一像素定位、语义分类与文本描述**：不只是找出改动区域，还要求理解“改了什么、怎么改的、如何描述”。
- **公开大规模基准与模型资源**：仓库提供 420K+ 训练对、40K 测试集、代码和预训练权重，具备很强的平台价值。

相关链接

Paper：https://arxiv.org/abs/2603.20193

Code：https://github.com/VILA-Lab/PIXAR

● 论文十

LeWorldModel: Stable End-to-End Joint-Embedding Predictive Architecture from Pixels

日期

2026-03-24

LeWorldModel: Stable End-to-End Joint-Embedding Predictive Architecture from Pixels

Lucas Maes[*1] **Quentin Le Lidec**[*2] **Damien Scieur**[1,3] **Yann LeCun**[2] **Randall Balestriero**[4]

[1]Mila & Université de Montréal [2]New York University [3]Samsung SAIL [4]Brown University

Website Code

内容摘要

LeWorldModel（LeWM）尝试解决 JEPA（Joint Embedding Predictive Architecture）世界模型训练脆弱的问题。以往方法通常依赖复杂的多项损失、EMA 教师、预训练编码器或辅助监督来避免表示坍塌，而 LeWM 提出一种**从原始像素稳定端到端训练**的新方案，只使用两个损失：下一步嵌入预测损失，以及约束潜在表示服从高斯分布的正则项。这样一来，相比现有唯一可端到端训练的替代方案，超参数复杂度显著下降。论文还显示，一个约 1500 万参数、单卡数小时即可训练的模型，就能在多种 2D/3D 控制任务上达到有竞争力表现，并且规划速度比基于基础模型的世界模型快最多 48 倍，说明小而稳的世界模型路线依然很有潜力。

主要亮点

- **极简损失实现稳定端到端 JEPA**：把复杂多项损失压缩到两项，明显降低训练门槛和调参负担。
- **小参数、低算力却具备强效率**：约 15M 参数、单 GPU 数小时训练，规划速度最高可达 48×。
- **不止能控制，还能学物理结构**：作者通过物理量 probing 和 surprise evaluation 证明其潜空间具有可解释的世界建模能力。

相关链接

Paper：https://arxiv.org/abs/2603.19312
Project: https://le-wm.github.io/
Code: https://github.com/lucas-maes/le-wm

● 论文十一

Towards end-to-end automation of AI research

日期

2026-03-25　《Nature》录用论文


Article

Towards end-to-end automation of AI research

https://doi.org/10.1038/s41586-026-10265-5
Received: 8 July 2025

Chris Lu[1,2,5], Cong Lu[1,3,4,5], Robert Tjarko Lange[1,5], Yutaro Yamada[1,5✉], Shengran Hu[1,3,4], Jakob Foerster[3], David Ha[1✉] & Jeff Clune[3,4✉]

内容摘要

这篇发表于《Nature》的论文系统介绍了 The AI Scientist，即一个能够覆盖机器学习研究全流程的自动化科研系统。该系统可在给定研究方向后，自主完成想法生成、文献检索、实验规划、代码实现、实验执行、结果分析、论文撰写和自动评审等步骤。论文的一个关键配套组件是 **Automated Reviewer**：它通过集成多轮自动评审与 area-chair 式汇总决策，对论文质量给出接近人类审稿人的判断。作者用真实会议数据验证其评审能力，并进一步展示：随着底层基础模型能力和测试时算力提高，AI Scientist 生成论文的质量也随之提升，体现出某种“科研自动化的 scaling law”。这是目前最完整地展示“AI 科研全流程自动化”的代表性工作之一。

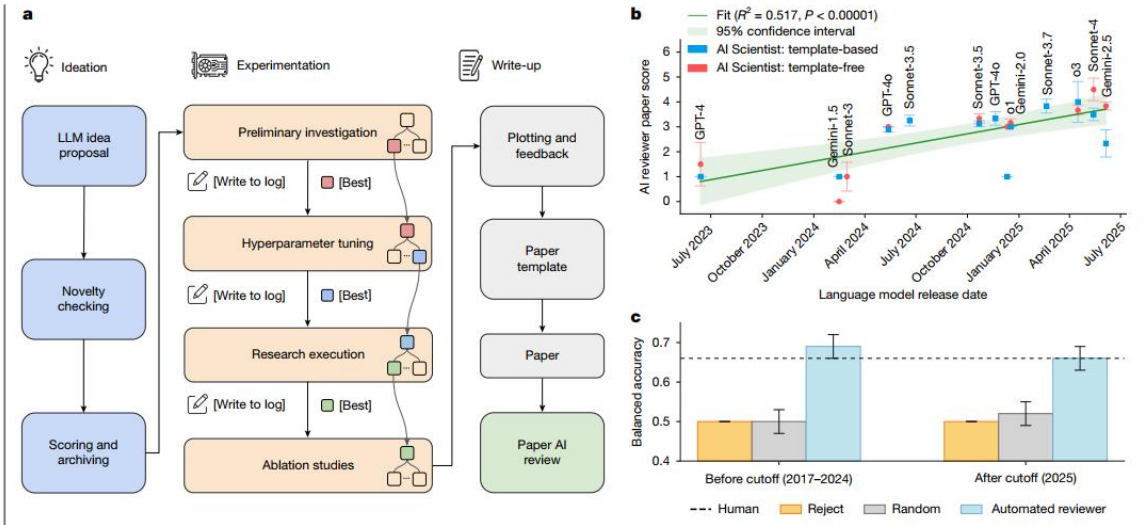


Fig. 1 | The AI Scientist workflow. a, The AI Scientist consists of distinct phases covering automated idea generation, tree-based experimentation, manuscript writing and reviewing. The experimentation phase uses an agentic tree search to generate and refine code implementations. This is structured into four stages: (1) initial investigation, (2) hyperparameter tuning, (3) research agenda execution and (4) ablation studies. From one experimental stage to the next, the best-performing checkpoint is selected to seed the next stage of the tree search. **b**, Scores for The AI Scientist papers across model releases. Paper quality consistently improves with the underlying model release date (as judged by The Automated Reviewer), indicating consistent future improvements with improving foundation models. The observed correlation is statistically significant ($P < 0.00001$). Shaded regions represent the standard error. Points represent mean scores with error bars and shaded regions indicating the standard error ($n = 6$ for template-free points, $n = 3$ for template-based points). Full experimental details, including model versions and replication counts, are provided in Supplementary Information section A.2.9. **c**, Automated review versus conference decisions. The Automated Reviewer achieves performance comparable with that of human reviewers, as validated by openly available decisions from past conferences (Table 1). Bars represent mean balanced accuracy; error bars show 95% bootstrapped confidence intervals (5,000 replicates). For replicability, each automated review is a 5-run ensemble. Two-sample z-tests on subsampled accuracy (automated $n = 698/876$, human $n = 412$) showed no significant difference before the training cutoff ($P = 0.319$) or post-cutoff ($P = 0.921$). Non-parametric bootstrap tests on F_1 scores showed automated outperformance ($P < 0.001$).

主要亮点

- **首次把科研全流程串成一个端到端系统**：从 idea 到 paper，不再只是自动写代码或自动写综述的局部工具。
- **Automated Reviewer 是关键基础设施**：它让大规模、低成本地评估 AI 生成科研成果成为可能，并与人类审稿表现相近。
- **展示出明显的 scaling 趋势**：更强的基础模型与更多测试算力，会显著提升自动科研系统的产出质量，说明这条路线仍在快速上升。

相关链接

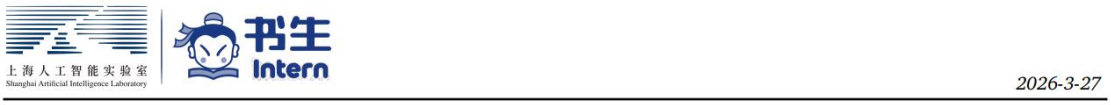
Paper：https://www.nature.com/articles/s41586-026-10265-5
Blog：https://sakana.ai/ai-scientist-nature

● 论文十二

Intern-S1-Pro: Scientific Multimodal Foundation Model at Trillion Scale

日期

2026-03-26



Intern-S1-Pro: Scientific Multimodal Foundation Model at Trillion Scale

Intern-S1-Pro Team, Shanghai AI Laboratory

内容摘要

Intern-S1-Pro 自称是首个 **一万亿参数级（1T）科学多模态基础模型**，目标是在通用能力、推理能力、视觉-文本理解能力和专业科学任务能力之间取得统一增强。论文强调，该模型不仅在通用与多模态能力上更强，还将科学能力扩展到 **100 多个专业任务**，覆盖化学、材料、生命科学、地球科学等多个关键领域。支撑这一规模训练的基础设施包括 XTuner 与 LMDeploy，它们使得万亿参数级别的强化学习训练与训练/推理精度一致性成为可能。论文将其定位为一种 “**Specializable Generalist**”：既保留通用智能的广度，又能在科学场景下表现出比不少闭源模型更深的专业能力。

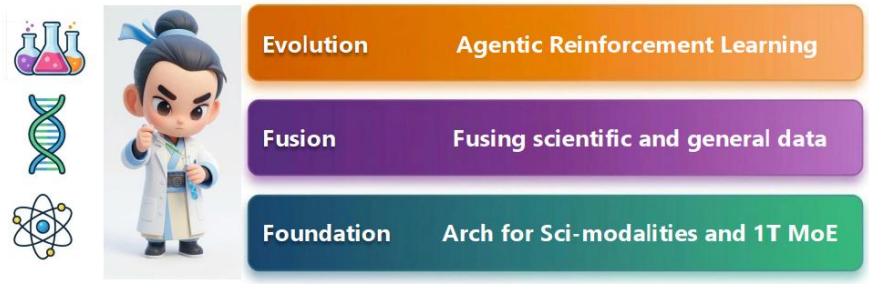


Figure 1: The SAGE (Synergistic Architecture for Generalizable Experts, including three layers, Foundation, Fusion, and Evolution) framework used in Intern-S1-Pro development, illustrating the core capabilities and the integrated learning process that enables synergistic improvements across domains.

主要亮点

- **万亿参数科学多模态模型：**体量本身就非常激进，瞄准的是“通用 + 专业科学”双重能力统一。
- **覆盖 100+ 科学任务：**不只是若干 demo，而是面向多个学科的大范围科学能力扩展。
- **强调训练基础设施的重要性：**XTuner 与 LMDeploy 被作为实现 1T 级 RL 训练与部署一致性的关键技术支柱。

相关链接

Paper： https://arxiv.org/abs/2603.25040

● 论文十三

ReFTA: Breaking the Weight Reconstruction Bottleneck in Tensorized Parameter-Efficient Fine-Tuning

日期

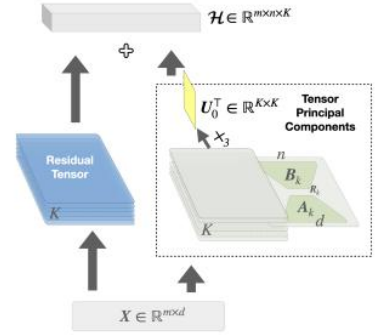
2026-03-26（CVPR 2026 录用）

ReFTA: Breaking the Weight Reconstruction Bottleneck in Tensorized Parameter-Efficient Fine-Tuning

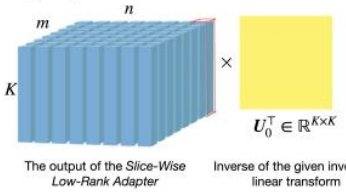
Jingjing Zheng[1,2,3,5] jjzheng233@gmail.com
Anda Tang[1] tanganda@pku.edu.cn
Qiangqiang Mao[3,4] maoq@student.ubc.ca
Zhouchen Lin[1,6,7,*] zlin@pku.edu.cn
Yankai Cao[2,3,4,*] yankai.cao@ubc.ca

内容摘要

ReFTA 聚焦张量化参数高效微调（Tensorized PEFT）中的一个核心瓶颈：**权重重建（weight reconstruction）成本过高。**这类方法通常会把大模型中的权重张量分解为更紧凑的张量结构，以减少可训练参数量，但在实际训练与推理中，往往仍需要频繁地把这些张量化表示重建回完整权重矩阵，从而带来明显的时间和内存负担。ReFTA 的核心目标就是打破这一重建瓶颈，让张量化微调不仅“参数省”，也真正“算得快、用得起”。从题目与论文定位看，它更偏向一种高效训练/部署层面的结构优化方法，旨在提升 tensorized PEFT 的实际可用性，使低成本微调不再被重建开销卡住。



(a) The k-th layer of the *Slice-Wise Low-Rank Adapter* has its own LoRA-like pair $(\boldsymbol{A}_k, \boldsymbol{B}_k)$ with rank R_k, where only $\boldsymbol{B}_k$ and $\boldsymbol{A}_k$ are trainable, and each R_k is determined by a tensor singular value thresholding algorithm and may vary across different k.



(b) *Feature Fusion Operation*: feature fusion along the third dimension via multiplication with $\boldsymbol{U}_0^\top$, where $\boldsymbol{U}_0$ denotes any fixed invertible linear transform, *e.g.*, Discrete Cosine Transform (DCT).

Figure 1. Illustration of ReFTA, where the Tensor Principal Components part consists of a *Slice-Wise Low-Rank Adapter* and a *Feature Fusion Operation* $\times_3 \boldsymbol{U}_0^\top$.

主要亮点

- **直指 tensorized PEFT 的关键工程瓶颈：**相比只讨论参数压缩率，ReFTA 更关注真实训练/推理流程里的重建代价问题。
- **强调“参数高效”与“计算高效”统一：**很多 PEFT 方法参数量虽小，但运行并不轻；这篇工作试图把两者真正统一起来。
- **适合资源受限的大模型微调场景：**如果能有效减少权重重建开销，就更有希望在边缘设备、低显存环境或批量任务微调中落地。

相关链接

Paper： https://zhouchenlin.github.io/Publications/2026-CVPR-ReFTA.pdf
Code: https://github.com/jzheng20/ReFTA

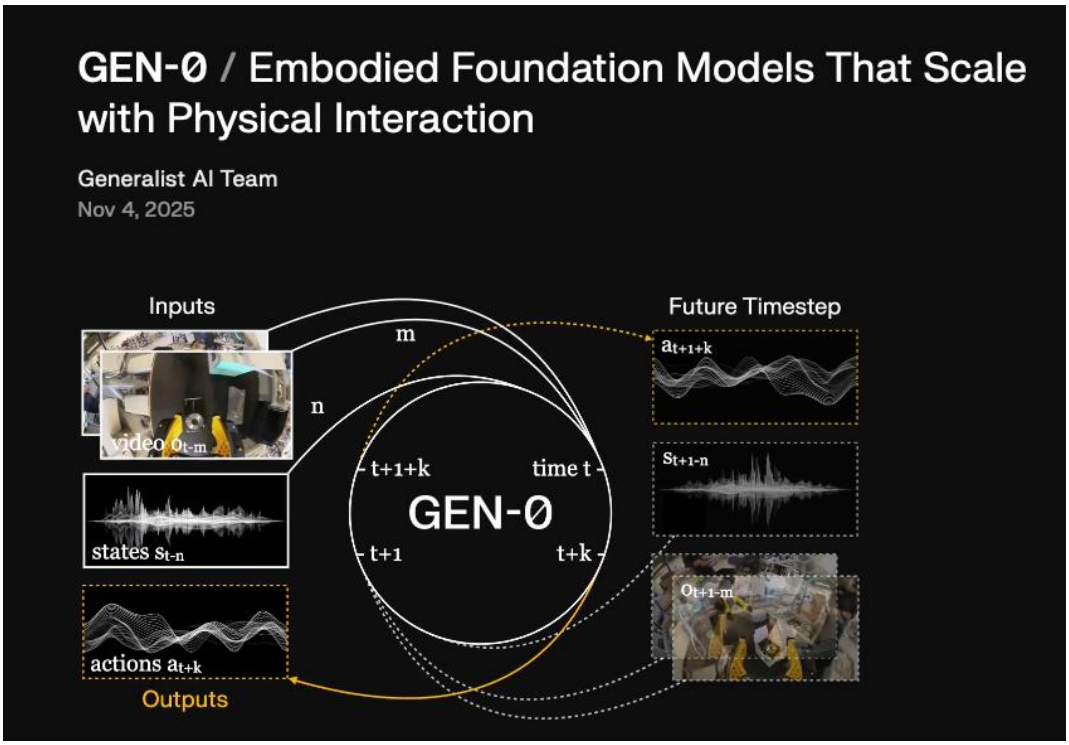
TECHNICAL UPDATES
技术动态

技术动态一

Generalist AI 发布具身基础模型 GEN-0 突破机器人智能规模化瓶颈

日期

2025-11-04



内容摘要

Generalist AI 正式推出 GEN-0 具身基础模型，该模型专为高保真原始物理交互下的多模态训练打造，首次在机器人领域确立可预测的智能规模化定律，打破了机器人发展的传统数据限制，实现模型能力随真实物理交互数据精准提升，标志着具身智能迈入规模化发展的全新阶段。GEN-0 的核心技术亮点包括：(1) 突破智能阈值与规模化定律：发现 70 亿参数量的相位跃迁节点，7B + 模型可高效吸收大规模机器人预训练数据，预训练数据、算力与下游任务性能呈强幂律关系，模型表现可精准预测；(2) 独创 Harmonic Reasoning 谐波推理：实现感知与行动令牌的异步连续流协同，让模型在物理世界中无缝思考与行动，无需依赖双系统架构或推理时引导；(3) 跨具身适配与无数据瓶颈：原生支持 6DoF、7DoF 等多类机器人，基于超 27 万小时真实世界操作数据预训练，数据以每周 1 万小时增速扩充，搭建了互联网级机器人数据基础设施；(4) 建立精细化预训练科学体系：验证数据质量和多样性比体量更重要，不同数据混合策略塑造不同模型特性，可通过 A/B 测试优化数据采集，为模型调优提供精准指引。GEN-0 的推出填补了机器人领域具身基础模型规模化的技术空白，为工业、物流、制造等多领域的机器人落地提供了核心支撑，推动真实物理世界中的机器人智能实现可预测的规模化升级。

相关链接

技术博客: https://generalistai.com/blog/nov-04-2025-GEN-0

技术动态二

OpenAI 推出面向专业工作与智能体任务的 GPT-5.4

日期

2026-03-05

内容摘要

OpenAI 正式发布 GPT-5.4，这是一款面向专业工作负载与复杂智能体任务优化的通用大语言模型，重点提升了专业知识工作、编程、多步工具协作与计算机使用等关键能力。该模型在通用推理、编程及知识型工作场景中进一步增强，在覆盖 44 个职业的知识型工作评测中达到行业专业水准，并在多项权威基准测试中取得领先表现；与此同时，GPT-5.4 也在响应效率、工具调用质量与整体使用成本之间实现了更优平衡。GPT-5.4 的核心技术亮点包括：(1) 原生计算机使用与更强视觉理解能力：模型原生支持计算机操作，在桌面环境导航任务中达到领先水平，并进一步增强了对视觉界面、图像内容与复杂文档的理解能力；(2) 优化的工具协作与高效 Token 利用机制：通过引入工具搜索、增强多步工具调用能力，并在 Codex 中实验性支持 1M Token 上下文，GPT-5.4 能够在复杂工作流中以更高效率完成规划、调用与执行；(3) 更强可控性与更低事实性错误：GPT-5.4 Thinking 支持在处理复杂任务时先简述工作方案，并允许用户在生成过程中动态补充指令、调整方向；同时，相比前代模型，其事实性错误率进一步下降，模型整体可靠性与安全防护能力持续增强。GPT-5.4 的推出进一步推动了 AI 智能体在复杂专业工作流中的落地，为企业级与开发者级专业化 AI 应用提供了更强的基础模型支撑。

相关链接

技术博客: https://openai.com/zh-Hans-CN/index/introducing-gpt-5-4/

• 技术动态三

GPT-5.4 Thinking 发布系统卡：强化高风险能力治理

日期

2026-03-05

GPT-5.4 Thinking System Card

OpenAI

March 5, 2026

内容摘要

OpenAI 发布 GPT-5.4 Thinking System Card，系统披露了 GPT-5.4 Thinking 的安全评估、部署约束与缓解方案。与此前同系列模型相比，这份系统卡最值得关注的并非单纯性能变化，而是 GPT-5.4 Thinking 被明确界定为首个已针对“高网络安全能力”实施专门缓解措施的通用模型，显示出前沿推理模型的治理正在进一步转向能力分级下的定向部署安全框架。其核心技术亮点包括：(1) 高网络安全能力模型的专门化缓解机制：在 Preparedness Framework 下，OpenAI 将 GPT-5.4 Thinking 纳入 High 级网络安全能力模型管理，并针对相关高风险内容引入专门缓解方案；(2) 面向真实攻击路径的运行时安全监控体系：部署侧采用分层 conversation monitor 与账户级处置机制，对高风险网络安全请求进行识别、拦截与进一步审查；(3) 覆盖计算机使用与推理过程的全链路安全治理：系统卡同时披露了越狱、提示注入、视觉输入、健康场景、误删数据防护、计算机使用确认机制及 Chain-of-Thought 可监测性等方面的评估与控制思路。总体来看，这份系统卡表明，OpenAI 对高能力推理模型的发布逻辑，正在由能力披露进一步转向“能力评估—运行时防护—访问控制”一体化的部署安全范式。

相关链接

技术博客：https://openai.com/zh-Hans-CN/index/introducing-gpt-5-4/

• 技术动态四

“十五五” 规划中的 “人工智能+”

日期

2026-03-13


中华人民共和国中央人民政府
www.gov.cn

中华人民共和国国民经济和社会发展第十五个五年规划纲要

2026-03-13 21:44 来源： 新华社

内容摘要

《中华人民共和国国民经济和社会发展第十五个五年规划纲要》正式发布。根据纲要公开内容，文件明确提出全面实施“人工智能+”行动，并将相关部署纳入“提升数智化发展水平”等章节，强调加强人工智能同科技创新、产业发展、文化建设、民生保障、社会治理相结合，全方位推进数智技术赋能。围绕基础供给，纲要提出强化算力、算法和数据高效供给，推进“模芯云用”协同创新，建设人工智能语料库和重点行业高质量数据集，并通过算力服务、算力租赁等方式降低用算门槛；围绕应用落地，文件提出推动人工智能在科研、制造、能源、农业、教育、医疗、政务等领域加快应用，推进国家人工智能创新高地和国家人工智能应用中试基地建设；围绕前沿布局与治理规则，纲要还将多模态、智能体、具身智能、群体智能等方向纳入前沿部署，并对学科专业建设、人才培养、模型能力评估、算法备案、安全评估、风险治理和国际合作等作出安排。

相关链接

报告链接：https://www.gov.cn/yaowen/liebiao/202603/content_7062633.htm

RESOURCE SHARING
资源分享

资源分享一

《OpenClaw 101：小龙虾教程及资源汇集》

类型：网站

内容提要

OpenClaw 101 是一个专注于开源 AI 代理项目 OpenClaw 的中文学习资源社区。该网站以“7 天掌握你的 AI 私人助理”为核心目标，通过系统化的教学路径引导用户从零开始构建一个 24/7 全天候在线、具备自主执行能力的智能助手。平台不仅详细解析了如何将 AI 接入 Telegram、飞书、钉钉等通讯工具，还深度整合了全球超过 440 篇优质教程及 5400 多项社区技能，内容涵盖云端一键部署、本地大模型联动、自动化工作流以及高级多 Agent 协作等实战场景。作为一个一站式的资源聚合中心，它在强调数据自主与隐私安全的同时，通过丰富的视频课与图文指南，将复杂的 AI 技术转化为普通用户也能轻松上手的生产力工具，是国内开发者和 AI 爱好者探索个人自动化代理的最佳入口。

资源链接

https://openclaw101.dev/zh

资源分享二

《Karpath 的 AutoResearch 开源框架》

类型：网站

内容提要

AutoResearch 是由 AI 领域巨擘 Andrej Karpathy 发起的开源项目，也是他提出的“Vibe Coding”理念的终极实践。其核心愿景是建立一个由 AI 代理自主驱动的科研组织：用户只需提供一个基础的 LLM 训练框架（如单卡 nanochat）和一份操作指令（program.md），AI 代理便会化身“全职研究员”，在深夜自动修改代码、调整超参数并运行训练实验。该项目打破了传统由人类研究员通过“声波”（开会）同步进度的低效模式。在 autoresearch 的架构下，训练被严格限制在 5 分钟的固定时长内，代理通过对比验证集的 bits-per-byte 指标，不断自我迭代出更优的模型架构。这不仅是一个高效的自动化调参工具，更是一场关于“让 AI 研究 AI”的前哨战，预示着人类将从繁琐的编程细节中抽离，转而通过编写“元程序”来定义研究目标，正式开启了自主科研 Agent 的新纪元

资源链接

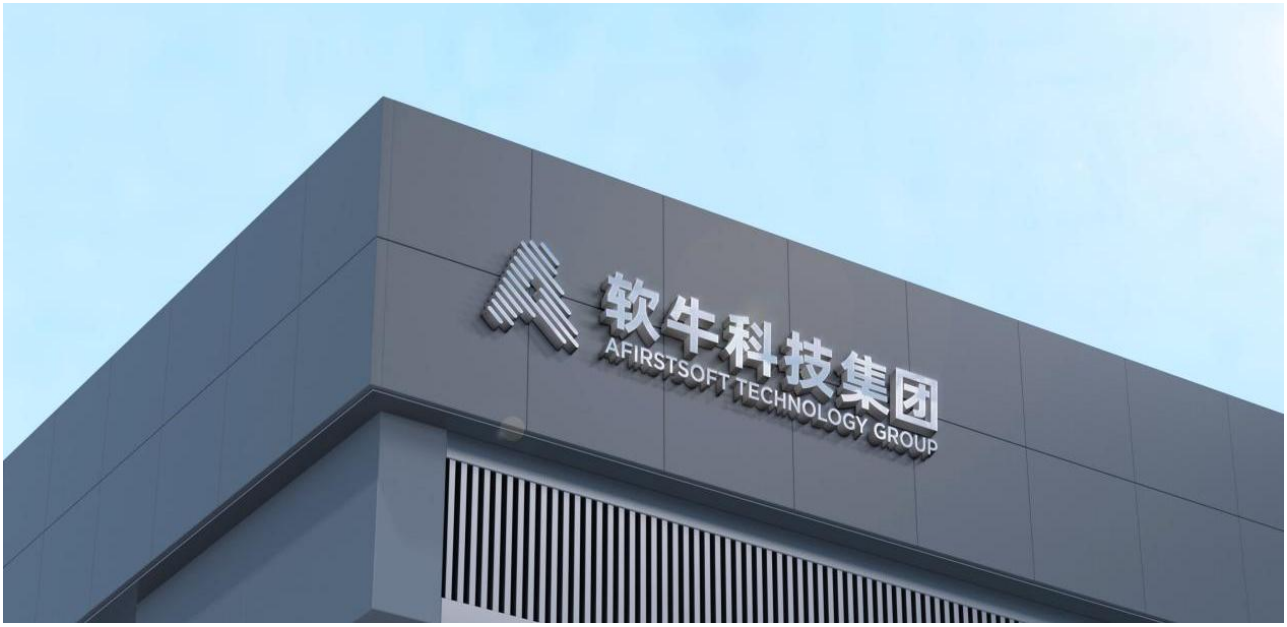
https://github.com/karpathy/autoresearch

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。

在消费级市场，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 3 期

总第 31 期

广东省图象图形学会

主办 **广东省图象图形学会**

人工智能前沿快报

2026 年第 3 期
总第 31 期
2026.04.01

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