

Artificial Intelligence Newsletter

人工智能 前沿快报



2026 年第 01 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. ReconVLA: Reconstructive Vision-Language-Action Model as Effective Robot Perceiver	1
二. Sampling-enabled scalable manifold learning unveils the discriminative cluster structure of high-dimensional data	4
三. OmniScientist: Toward a Co-evolving Ecosystem of Human and AI Scientists	6
四. Causal Structure Learning for Dynamical Systems with Theoretical Score Analysis	8
五. End-to-End Test-Time Training for Long Context	10
六. End-to-End Test-Time Training for Long Context	12
七. Deep contrastive learning enables genome-wide virtual screening	14
八. AI Agent Systems: Architectures, Applications, and Evaluation	16
九. The Path Ahead for Agentic AI: Challenges and Opportunities	17
十. mHC: Manifold-Constrained Hyper-Connections	19
十一. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking	21
十二. Conditional Memory via Scalable Lookup: A New Axis of Sparsity for Large Language Models	23
十三. An Information-Theoretic Perspective on LLM Tokenizers	25
十四. DanQing: An Up-to-Date Large-Scale Chinese VisionLanguage Pre-training Dataset	27
十五. A Safety Report on GPT-5.2, Gemini 3 Pro, Qwen3-VL, Grok 4.1 Fast, Nano Banana Pro, and Seedream 4.5	29
十六. LLM2CLIP: Powerful Language Model Unlocks Richer Cross-Modality Representation	31
十七. Toward Efficient Agents: A Survey of Memory, Tool learning, and Planning	33
十八. OpenVision 3 : A Family of Unified Visual Encoder for Both Understanding and Generation	36
十九. Qwen3-ASR Technical Report	38

目录

技术动态

一. Science in 2050: the future breakthroughs that will shape our world — and beyond	41
二. 北京智源人工智能研究院发布 2026 年十大 AI 技术发展趋势	42
三. 智谱 AI 开源图像生成大模型 GLM-Image	43
四. 阿里巴巴发布最新推理大模型 Qwen3-Max-Thinking	45
五. Kimi K2.5: Visual Agentic Intelligence	46
六. DeepSeek 探索新架构并开源 Deepseek-OCR 2	47
七. OpenAI 发布 Prism, 颠覆科研写作	49
八. 全国首例 AI “幻觉” 侵权案宣判, AI 自称 “愿赔 10 万”	50
九. 谷歌 DeepMind 发布 AlphaGenome, 登上 Nature 封面	51

资源分享

一. Vibe Coding 指南	53
二. Claude Skills 资源	53

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、9 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

ReconVLA: Reconstructive Vision-Language-Action Model as Effective Robot Perceiver

日期

2025-08-14 (AAAI 2026 outstanding paper)

ReconVLA: Reconstructive Vision-Language-Action Model as Effective Robot Perceiver

**Wenxuan Song¹, Ziyang Zhou¹, Han Zhao^{2,3}, Jiayi Chen¹, Pengxiang Ding^{2,3},
Haodong Yan¹, Yuxin Huang¹, Feilong Tang⁴, Donglin Wang², Haoang Li¹**

¹The Hong Kong University of Science and Technology (Guangzhou)

²Westlake University ³Zhejiang University ⁴Monash University

内容摘要

该论文提出 ReconVLA (Reconstructive Vision-Language-Action), 面向机器人操作场景, 针对现有 VLA 模型在复杂场景下视觉注意力分散、难以精确对准目标物体的问题, 构建了一套“重建式”隐式视觉对齐框架。ReconVLA 在标准视觉-语言-动作管线之外, 引入一个条件扩散 Transformer 作为重建分支: 模型在接收图像、多视角观测、语言指令和本体状态后, 除了自回归生成离散动作 token, 还需输出一组“重建 token”, 作为条件去从噪声中重建当前图像中被操作目标的“凝视区域”(gaze region) 潜空间表示, 从而在训练中强制视觉表征对该局部区域进行细粒度编码并聚焦注意力。为支撑这一隐式对齐范式, 作者基于 BridgeData V2、LIBERO、CALVIN 等开放机器人数据集, 利用 Grounding DINO 自动生成目标区域框与裁剪图, 构建了包含 10 万余条轨迹、约 200 万样本的大规模重建预训练集; 在此基础上进行联合训练与下游微调, 在 CALVIN 长链任务、多种生成式/大模型 VLA 基线以及真实机械臂多任务实验中, ReconVLA 在成功率、平均完成子任务长度和对未见目标的泛化能力上均取得显著优势。

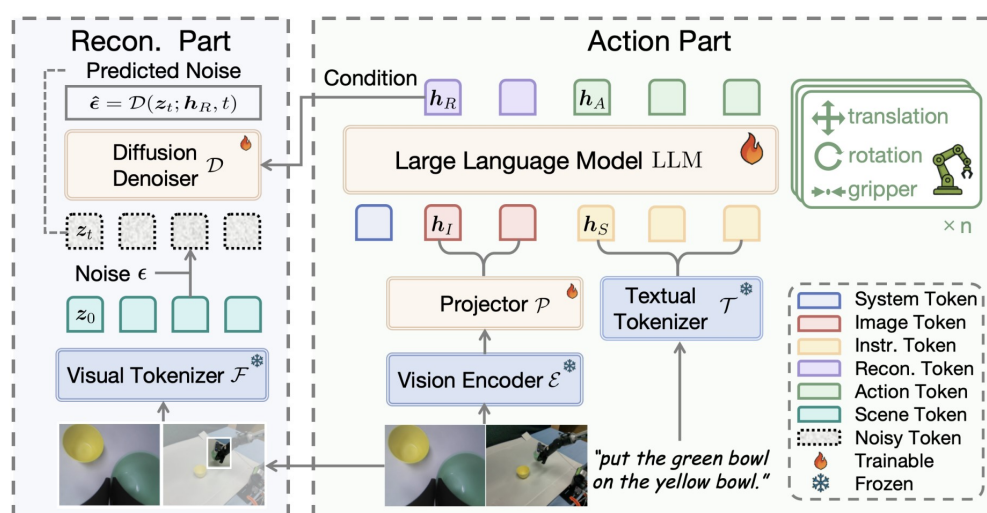


Figure 3: **Architecture of our ReconVLA.** Our model consists of a reconstructive part and an action part. The input includes multi-view images and a text instruction. For the action part, the model outputs discrete action tokens. For the reconstruction part, our ReconVLA is guided to output reconstructive tokens, which are conditions of the denoising process to reconstruct the scene tokens z_0 from noisy z_t . The scene tokens are tokenized images of gaze regions. This supervision enables our ReconVLA to enhance visual grounding and fine-grained comprehension capabilities, which contribute to precise manipulation.

主要亮点

- **隐式重建式视觉对齐机制**：ReconVLA 不再显式输入裁剪图或输出框坐标，而是利用目标“凝视区域”的潜空间重建作为监督信号，让 VLA 在生成动作的同时学习一套能从自身视觉输出中重构目标局部的表征，等效于在网络内部形成“视线落点”，明显收紧注意力分布、提升目标对准精度。
- **视觉扩散重建分支与 VLA 一体化**：模型采用冻结 VAE 作为视觉 tokenizer，将 gaze 区域编码为场景 token，通过条件扩散 Transformer 在 reconstructive token 引导下进行去噪重建，并与动作交叉熵损失联合反向传播；这种设计在不增加推理时额外输入/输出的前提下，为主干 VLA 持续提供区域级视觉训练信号。
- **面向机器人操作的大规模重建预训练**：通过在多源开源机器人数据上自动标注“整图 + 目标裁剪图”对，并以 Grounding DINO 作为开放词汇检测器，构建了 100k+ 轨迹、200 万样本的预训练集，使模型在进入具体任务前已具备较强的区域重建与视觉泛化能力，在未见场景/背景下的视觉生成稳定性明显优于只在下游监督动作的 VLA。
- **在模拟长链与真实机械臂场景中的强泛化**：在 CALVIN ABCD/ABCDD 上，ReconVLA 超过多种生成式视觉规划方法（如 SuSIE、GR- 1、Vidman、CLOVER、GEVRM）和强 VLA 基线（OpenVLA、UniVLA），在最难子任务上的成功率提升可达 20 个百分点；在真实 AgileX PiPer 机械臂的“装水果、叠碗、翻杯、收拾桌面”等任务及未见目标物上，ReconVLA 仍能保持接近或超过 90% 的成功率，而对比方法在未见物体上几乎完全失败，凸显重建式隐式对齐对真实部署和零样本泛化的价值。

相关链接

Paper: <https://arxiv.org/pdf/2508.10333>

● 论文二

Sampling-enabled scalable manifold learning unveils the discriminative cluster structure of high-dimensional data

日期

2025-09-10

Article | Published: 10 September 2025

Sampling-enabled scalable manifold learning unveils the discriminative cluster structure of high-dimensional data

[Dehua Peng](#), [Zhipeng Gui](#) , [Wenzhang Wei](#), [Fa Li](#), [Jie Gui](#), [Huayi Wu](#) & [Jianya Gong](#)

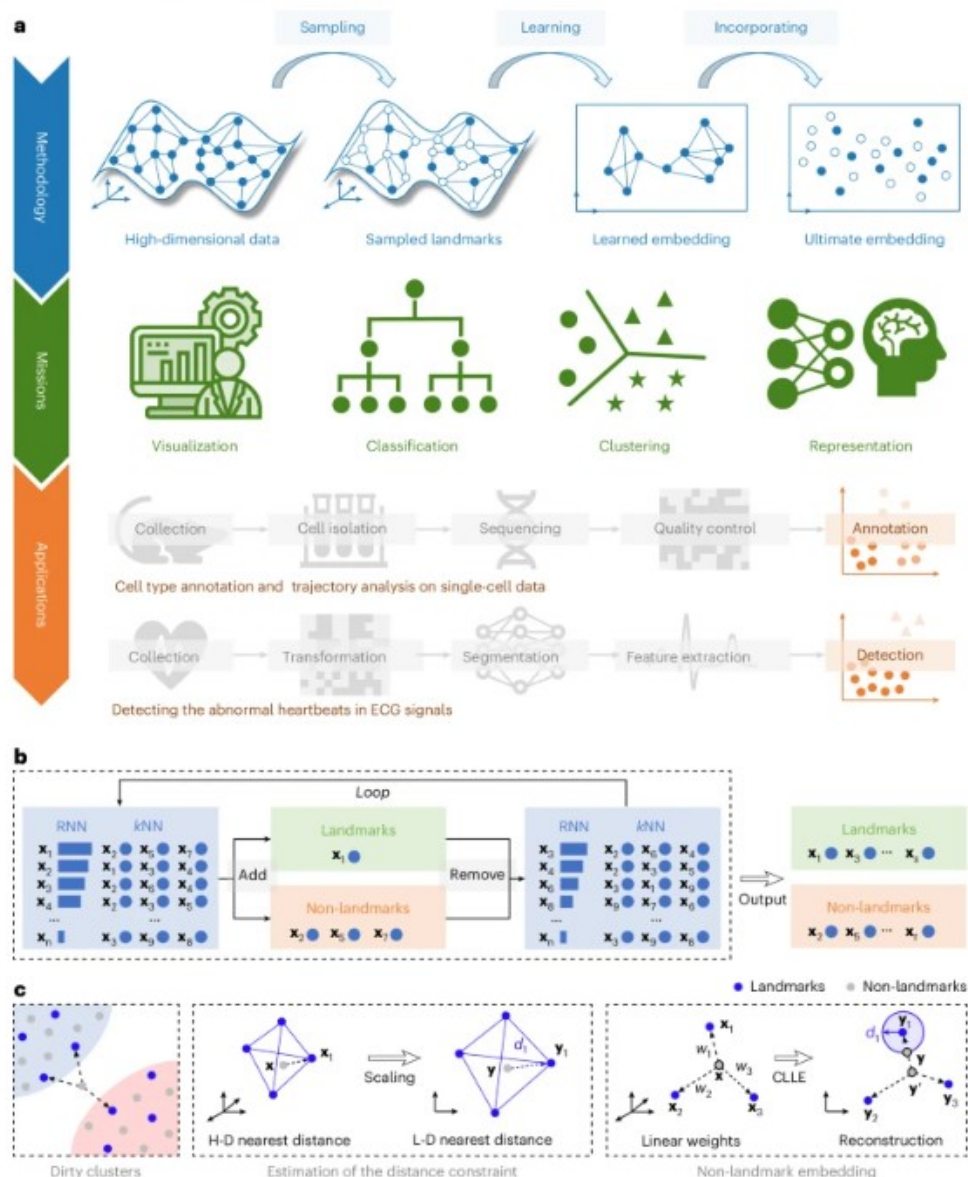
[Nature Machine Intelligence](#) **7**, 1669–1684 (2025) | [Cite this article](#)

4044 Accesses | **3** Citations | **23** Altmetric | [Metrics](#)

内容摘要

该论文发表于《Nature Machine Intelligence》（2025 年），提出可伸缩流形学习方法 SUDE，针对现有技术类簇区分弱、拓扑保留不足、伸缩性差等问题，通过 “枣糕采样（PPS）- 地标嵌入学习 - 约束局部线性嵌入（CLLE）” 三阶段实现高维数据高效降维，同时创新重尾对数概率分布增强类簇分离性。经 27 类数据集验证，其聚类精度、计算效率优于 t-SNE、UMAP 等方法，且在单细胞分析、ECG 异常检测等场景中表现突出。

Fig. 1: Illustration of SUDE.



主要亮点

- **兼顾高效性与分离性**: 通过地标采样架构提升计算伸缩性, 重尾概率分布强化类簇区分, 解决传统方法痛点。
- **采样与约束机制创新**: PPS 采样保障地标分布一致性, CLLE 约束避免 “脏簇”, 提升嵌入鲁棒性。
- **跨领域适用性强**: 在通用高维数据、生物医学、医疗诊断等场景均实现精准降维与分析, 落地价值显著。

相关链接

Paper: <https://www.nature.com/articles/s42256-025-01112-9>

Github: <https://github.com/ZPGuiGroupWhu/sude>

● 论文三

OmniScientist: Toward a Co-evolving Ecosystem of Human and AI Scientists

日期

2025-12-14

OmniScientist: Toward a Co-evolving Ecosystem of Human and AI Scientists

Chenyang Shao^{1,2} Dehao Huang² Yu Li¹ Keyu Zhao¹ Weiquan Lin²
Yining Zhang² Qingbin Zeng¹ Zhiyu Chen² Tianxing Li¹ Yifei Huang²
Taozhong Wu² Xinyang Liu¹ Ruotong Zhao¹ Mengsheng Zhao² Jiaoyang Li²
Xuhua Zhang² Yue Wang² Yuanyi Zhen² Fengli Xu^{1,2,*} Yong Li^{1,2,*} Tie-Yan Liu²

¹Department of Electronic Engineering, BNRist, Tsinghua University

²Zhongguancun Academy

*{fenglilu, liyong07}@tsinghua.edu.cn



内容摘要

OmniScientist 是一个将人类科研基础设施明确编码到 AI 科研工作流中的综合框架，旨在解决现有 AI 科学家系统将科研视为独立搜索或优化问题、忽视科研社会性与协作性的缺陷。该框架实现了从数据基础、文献综述、研究构思、实验自动化、科学写作到同行评审的端到端自动化，同时通过结构化知识系统、协作研究协议（OSP）和开放评估平台（ScienceArena）提供全面基础设施支持，推动 AI 智能体从单纯的任务执行者转变为能理解科研规范、参与协作并推动科研生态进化的真正“科学家”，构建人机科学家协同进化的创新生态。

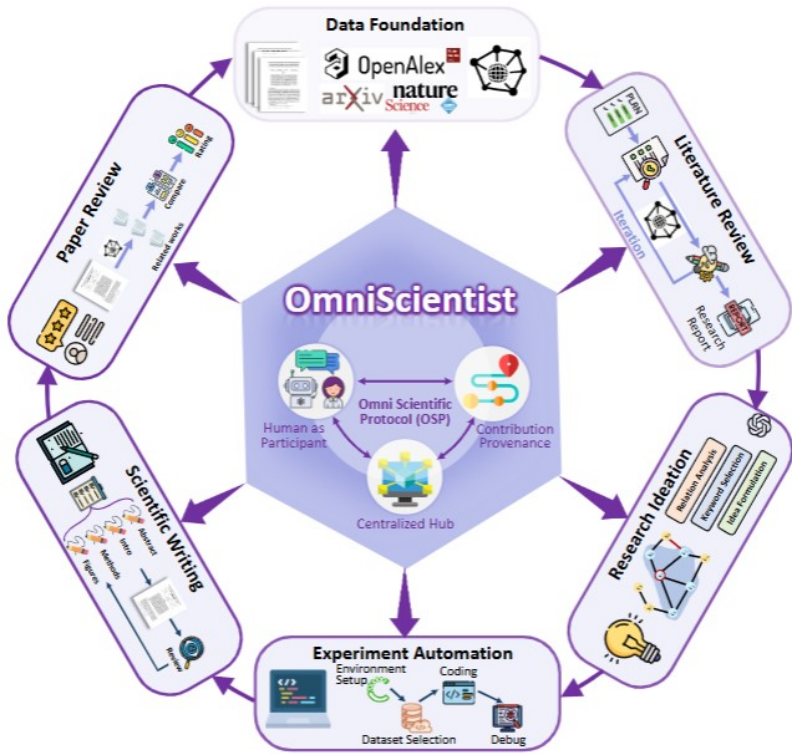


Figure 1: Overview of Our OmniScientist System

主要亮点

- **还原科研协作本质的基础设施设计：**突破传统 AI 科研工具的孤立局限，通过 OSP 协议将人类研究者纳入系统内部参与者，实现人机对称协作与贡献溯源；结合基于引文网络和概念关联的结构化知识系统，复刻人类科研的协作机制、信用分配等核心基础设施。
- **全流程闭环自动化与多模块协同：**整合数据基础、文献综述、研究构思等六大核心模块，形成覆盖科研全生命周期的闭环工作流。各模块采用多智能体架构，例如文献综述模块通过网络增强检索实现深度知识挖掘，实验自动化模块支持基线与数据集联合推荐，大幅提升科研效率与严谨性。
- **社区驱动的动态评估体系：**推出 ScienceArena 开放评估平台，采用盲态成对投票与 Elo 排名机制，以人类专家判断为核心形成动态评价标准，既解决了开放 - ended 科研任务的评估难题，又能引导 AI 智能体随社区标准持续进化，保障科研产出的质量与创新性。

相关链接

Paper: <https://arxiv.org/pdf/2511.16931>

● 论文四

Causal Structure Learning for Dynamical Systems with Theoretical Score Analysis

日期

2025-12-16 (AAAI 2026 outstanding paper)

Causal Structure Learning for Dynamical Systems with Theoretical Score Analysis

Nicholas Tagliapietra^{1,2}, Katharina Ensinger¹, Christoph Zimmer³, Osman Mian⁴

¹Bosch Center for Artificial Intelligence, Renningen, Germany

²Computer Science Department, TU Darmstadt, Germany

³Baden-Wuerttemberg Cooperative State University Mannheim, Germany

⁴Institute for AI in medicine IKIM, Germany
tagliapietra.nicholas@gmail.com

内容摘要

该论文提出 CADYT (Causal Discovery for Dynamic Timeseries), 一种面向连续时间动力系统的因果结构学习方法, 目标是在只观测到离散轨迹 (包含不规则采样) 的条件下恢复潜在微分方程系统的因果图结构。方法基于动态结构因果模型 (DSCM) 和差分型因果模型, 将系统动力学视作一组由父节点轨迹决定的因果机制, 并使用精确高斯过程 (GP) 推断结合多步数值积分器, 对连续时间向量场 F 进行非参数建模, 从而在任意时间点对状态导数和轨迹进行估计。在此基础上, 作者构造了一个以算法马尔可夫条件 (AMC) 和最小描述长度 (MDL) 为理论基础的评分函数, 将不可计算的 Kolmogorov 复杂度上界化为可计算的 MDL 代价, 并通过并行化的贪心结构搜索 (前向加边 + 反向删边) 寻找使得整体描述长度最小的有向图; 理论上证明该评分渐近等价于带正则项的对数似然 (BIC 型正则对数似然), 并在合成随机图、弹簧振子与 Rossler 混沌系统等基准上, 相比 DYNOTEARS、PCMCI、VARLINGAM 等方法显著降低归一化结构汉明距离 (NSHD)、提高 F1 与 AUPRC, 尤其在不规则采样场景下优势更明显。

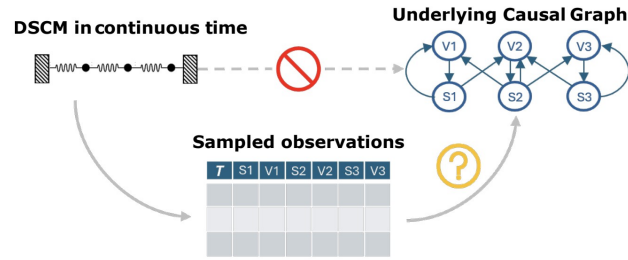


Figure 1: CADYT discovers the unknown causal structure (Top-right) using trajectories sampled from a continuous-time dynamical system (e.g. n-mass spring system). Our learned continuous-time model adapts to arbitrary timelines, including irregularly sampled ones.

主要亮点

- **连续时间视角的因果发现框架**: CADYT 直接在连续时间 ODE/DSCM 上建模，而非简单离散化为动态贝叶斯网络，从理论上明确了在动态稳定、因果充分性和采样频率足够高等条件下，如何将连续动力系统映射到可学习的因果结构，并允许存在反馈环与自环。
- **基于精确高斯过程的动力学建模**: 方法利用专门为多步积分器推导的 GP 核，将 Adams–Bashforth 等多步积分 scheme 融入核函数，使 GP 在不规则时间网络上仍可对向量场进行精确推断，从而统一解决“连续动力学建模”和“支持不规则采样”两个长期分离的问题。
- **AMC + MDL 驱动的评分函数与理论性质**: 在算法马尔可夫条件下，真正的因果图对应于使联合分布描述长度最短的分解方式，作者通过 MDL 给出可计算上界，进一步证明该评分的上界在样本数趋于无穷时与 BIC 形式同阶，从而获得与经典正则对数似然类似的模型选择一致性保证。
- **在不规则采样和复杂动力系统上的实证优势**: 在含环的 Erdos–Rényi 随机图、双质量弹簧系统、真实物理装置数据以及 Rossler 超混沌系统等多种场景中，CADYT 相比 DYNOTEARS、PCMCI、VARLINGAM 明显降低 NSHD 并显著提高 AUPRC，且在空图 sanity check 中几乎不产生伪因果边，体现出对假阳性的鲁棒性；积分阶数越高 (AB1→AB3)，在不规则采样下的结构恢复质量越好，符合数值分析直觉。

相关链接

Paper: <https://arxiv.org/pdf/2512.14361>

● 论文五

End-to-End Test-Time Training for Long Context

日期

2025-12-31

A Survey of Vibe Coding with Large Language Models

Yuyao Ge^{1*} Lingrui Mei^{1*} Zenghao Duan¹ Tianhao Li² Yujia Zheng²

Yiwei Wang³ Lexin Wang¹ Jiayu Yao^{1*} Tianyu Liu⁴ Yujun Cai⁵

Baolong Bi^{1*} Fangda Guo¹ Jiafeng Guo^{1*} Shenghua Liu^{1†*} Xueqi Cheng^{1*}

¹Institute of Computing Technology, Chinese Academy of Sciences ²Duke University

³University of California, Merced ⁴Peking University ⁵University of Queensland

内容摘要

该论文是首篇针对 Vibe Coding 的系统性综述，Vibe Coding 是一种基于大型语言模型（LLMs）的新型软件开发方法，核心是开发者通过观察输出结果而非逐行审查代码来验证 AI 生成的实现。论文将其形式化为受约束的马尔可夫决策过程，明确人类开发者、软件项目与编码代理之间的动态三元关系；梳理了 Vibe Coding 的完整生态，涵盖编码用 LLM、基于 LLM 的编码代理、开发环境及反馈机制四大核心组件；提出了无约束自动化、迭代对话协作、规划驱动、测试驱动和上下文增强五种开发模型；同时分析了该范式在开发流程重构、代码可靠性与安全性、代理监督扩展性及人为因素等方面的未来影响与开放挑战，为 AI 增强软件工程奠定了理论基础并提供了技术路线图。

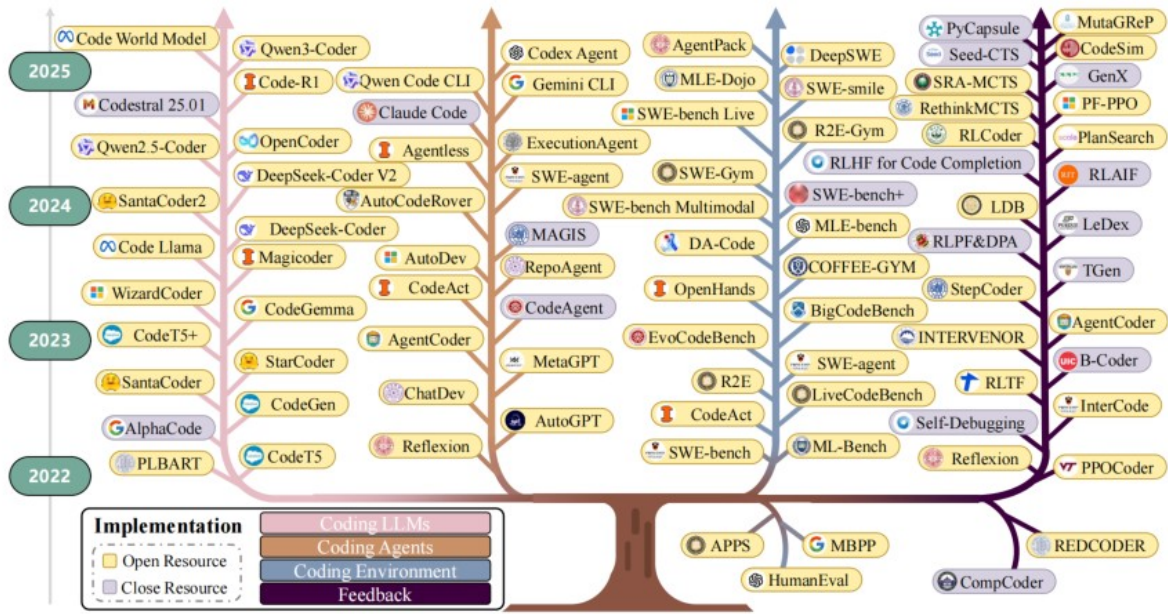


Figure 3 Vibe Coding Evolution Timeline: A comprehensive visualization of the development trajectory of Vibe Coding implementations from 2022 to 2025, showing the evolution from foundational Coding LLMs to sophisticated coding agent systems, execute environments and Feedback mechanism.

主要亮点

- **首次构建 Vibe Coding 理论体系：**将 Vibe Coding 正式定义为工程学科，通过受约束的马尔可夫决策过程建模人类 - 项目 - 代理的三元互动，解决了该新兴范式缺乏理论支撑的问题，同时提出五大开发模型的完整分类法，为实践提供了明确框架。
- **全面梳理生态系统关键组件：**系统整合了编码用 LLM 的训练技术（数据基础、预训练与后训练）、编码代理的核心能力（任务分解、记忆机制、行动执行等）、开发环境（隔离执行、交互式接口、分布式编排）及多源反馈机制（编译器、执行、人类、自我优化反馈），呈现了 Vibe Coding 的技术全貌。
- **深度剖析核心挑战与实践关键：**指出 Vibe Coding 的成效不仅依赖代理能力，更取决于上下文工程、成熟开发环境及人机协作模式；重点强调了代码可靠性与安全性、可扩展的代理监督、开发者角色转型等关键挑战，为后续研究和工程应用指明了核心方向。

相关链接

Paper: <https://arxiv.org/pdf/2510.12399>

Github: <https://github.com/YuyaoGe/Awesome-Vibe-Coding>

● 论文六

End-to-End Test-Time Training for Long Context

日期

2025-12-31

End-to-End Test-Time Training for Long Context

Arnuv Tandon^{*1,3}, Karan Dalal^{*1,4}, Xinhao Li^{*5}, Daniel Koceja^{*3}, Marcel Rød^{*3}, Sam Buchanan⁴,
Xiaolong Wang⁵, Jure Leskovec³, Sanmi Koyejo³, Tatsunori Hashimoto³, Carlos Guestrin³,
Jed McCaleb¹, Yejin Choi², Yu Sun^{*2,3}

¹ Astera Institute ² NVIDIA ³ Stanford University ⁴ UC Berkeley ⁵ UC San Diego

内容摘要

该论文提出一种端到端测试时训练方法 (TTT-E2E)，将长上下文语言建模转化为持续学习问题而非架构设计问题。其核心是基于滑动窗口注意力的标准 Transformer 架构，在测试时通过下一个 token 预测持续学习，将上下文压缩到模型权重中；同时在训练阶段通过元学习优化模型初始化，使模型适配测试时的训练过程。实验表明，3B 参数模型（164B 训练 token）在上下文长度扩展性上与全注意力 Transformer 相当，且推理延迟不受上下文长度影响（128K 上下文时比全注意力快 2.7 倍），在长上下文语言建模任务中实现了性能与效率的平衡。

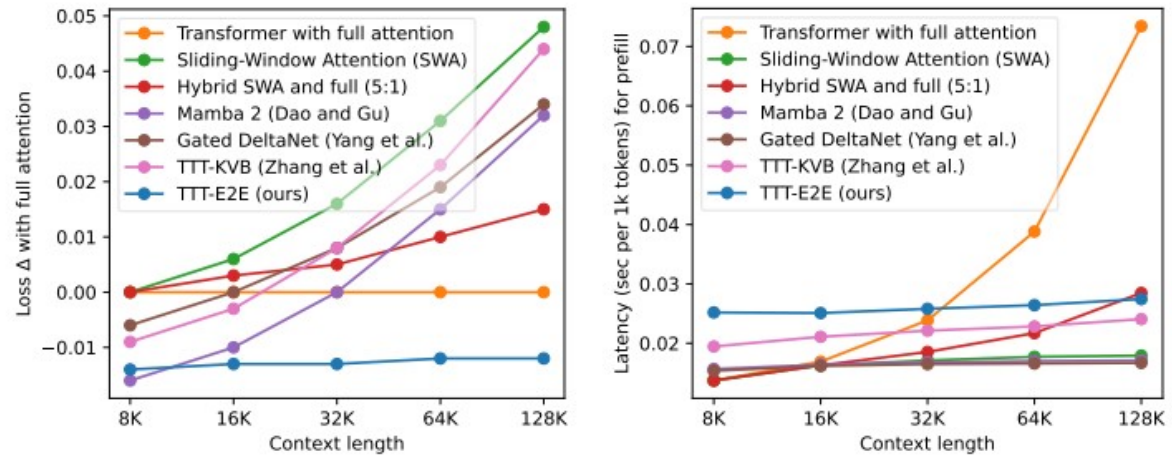


Figure 1. Scaling with context length, in terms of test loss (left) and latency (right). **Left:** Our method (TTT-E2E) turns the worst line (green) into the best (blue) at 128K context length. Loss Δ ($\downarrow$), the y -value, is computed as (loss of the reported method) – (loss of Transformer with full attention), so loss Δ of full attention itself (orange) is the flat line at $y = 0$. While other methods produce worse loss Δ in longer context, TTT-E2E maintains the same advantage over full attention. All models have 3B parameters and are trained with 164B tokens. **Right:** Similar to SWA and the RNN baselines, TTT-E2E has constant inference latency regardless of context length, making it 2.7 $\times$ faster than full attention for 128K context on an H100.

主要亮点

- **范式革新：**将长上下文建模转化为持续学习问题，无需复杂专属架构，仅依赖带滑动窗口注意力的标准 Transformer。
- **端到端优化：**训练时通过元学习优化测试时训练的初始化，测试时直接优化下一个 token 预测损失，消除训练与测试的行为偏差。
- **性能 - 效率双优：**长上下文扩展性媲美全注意力 Transformer，推理延迟与上下文长度无关（128K 上下文时比全注意力快 2.7 倍）。

相关链接

Paper: <https://arxiv.org/pdf/2512.23675>

● 论文七

Deep contrastive learning enables genome-wide virtual screening

日期

2026-01-08

Deep contrastive learning enables genome-wide virtual screening

YINJUN JIA , BOWEN GAO , JIAXIN TAN , JIQING ZHENG , XIN HONG , WENYU ZHU , HAICHUAN TAN , YUAN XIAO , LIPING TAN, [...], AND

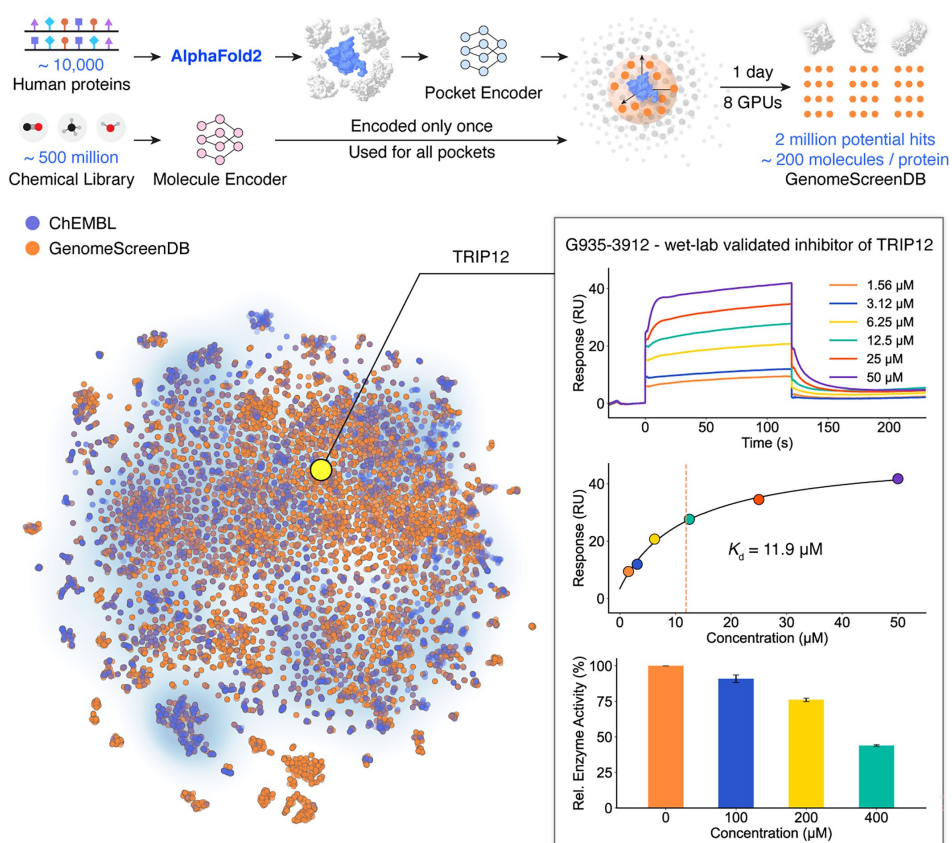
YANYAN LAN  +13 authors [Authors Info & Affiliations](#)

SCIENCE • 8 Jan 2026 • Vol 391, Issue 6781 • DOI: 10.1126/science.ada9530

内容摘要

2026 年 1 月 9 日, 清华大学智能科学讲席教授、智能产业研究院 (AIR) 院长张亚勤院士团队, 联合 AIR 兰艳艳教授团队及合作者, 在《Science》发表 2026 年清华大学首篇该刊论文。研究研发出 AI 工具 DrugCLIP, 创新性借鉴 “以文搜图” 的向量匹配原理, 通过训练分子编码器与口袋编码器, 将小分子化合物和蛋白质结合口袋转化为特征向量, 在向量空间中快速匹配潜在结合对。

该工具仅用 8 张顶级显卡, 即可在 24 小时内完成超 10 万亿次蛋白质 - 分子匹配计算, 首次实现覆盖人类基因组规模的药物虚拟筛选: 覆盖约 1 万个蛋白靶点、2 万个蛋白口袋, 筛选超 5 亿个类药小分子, 为近半数人类基因组找到 200 多万个潜在候选药物分子, 并构建全球最大规模的蛋白 - 配体筛选数据库 GenomeScreenDB, 目前已免费向全球科研社区开放。



主要亮点

- **效率颠覆性突破**：仅需 8 张 GPU、24 小时即可完成超 10 万亿次蛋白质 - 分子匹配，速度超传统方法千万倍，首次实现覆盖近半数人类基因组的药物虚拟筛选。
- **性能与泛化性优异**：在权威测试集上准确率超传统方法及多数深度学习模型，无需精确蛋白三维结构，对未接触过的蛋白质家族也能有效预测，湿实验验证找到优于常用药物的新型分子。
- **开放赋能全球科研**：构建全球最大规模蛋白 - 配体筛选数据库并免费开放，大幅降低早期药物发现壁垒，尤其助力冷门重要疾病靶点研究，推动全球药物研发协作。

相关链接

Paper: <https://www.science.org/doi/10.1126/science.ads9530>

● 论文八

AI Agent Systems: Architectures, Applications, and Evaluation

日期

2026-01-05

AI Agent Systems: Architectures, Applications, and Evaluation

BIN XU, School of Electrical, Computer and Energy Engineering, Arizona State University, USA

内容摘要

本文系统综述了 AI 智能体（Agent）系统的架构、应用与评估，提出以 Transformer 为基础的 “智能体 Transformer” 统一范式，将智能体定义为嵌入结构化控制循环的策略模型，核心组件包括政策核心（LLM/VLM）、记忆系统、工具集、验证器及环境接口。论文围绕推理规划、工具调用、环境交互三大核心能力，梳理了智能体的学习机制（强化学习、模仿学习等）、系统模块设计与部署场景，构建了涵盖通用智能体、具身智能体、生成式智能体等类别的分类体系，强调评估需兼顾任务成功率、效率、安全性等多维度指标，并指出可验证工具执行、长期记忆管理、鲁棒评估等未来研究方向。



Fig. 1. Overview of AI agents and the agent execution loop (reasoning, tools, and memory)

主要亮点

- **统一范式与清晰分类：**提出 “智能体 Transformer” 抽象模型，将分散的智能体设计整合为包含政策核心、记忆、工具、验证器、环境的结构化框架，同时构建了覆盖通用、具身、生成式等多场景的分类体系，明确不同类型智能体的设计约束与适用场景。
- **多维度学习与系统优化：**全面分析了强化学习、模仿学习、上下文学习等多种学习机制在智能体中的应用，强调智能体性能提升需结合模型优化、系统模块化设计（如 MRKL 路由、ReAct 循环）与基础设施支持（沙箱执行、审计日志），而非单纯依赖模型规模扩大。
- **全面的评估体系：**突破传统单一指标评估模式，建立了包含任务完成率、效率成本、工具使用正确性、鲁棒性、安全性等多维度的评估框架，明确了 WebArena、SWE-bench 等主流基准测试的适用场景，为智能体的实际部署与性能优化提供了可操作的衡量标准。

相关链接

Paper: <https://arxiv.org/pdf/2601.01743>

● 论文九

The Path Ahead for Agentic AI: Challenges and Opportunities

日期

2026-01-06

The Path Ahead for Agentic AI: Challenges and Opportunities

Nadia Sibai¹, Yara Ahmed¹, Serry Sibae¹, Sawsan AlHalawani¹, Adel Ammar^{1*}, and Wadii Boulila¹

Robotics and Internet-of-Things (RIOTU) Lab, Prince Sultan University, Riyadh,
Saudi Arabia
{221410294, 221410920, ssibae, shalawani, aammar, wboulila}@psu.edu.sa

内容摘要

该论文是一篇关于“Agentic AI”（具备自主规划与行动能力的智能体式 AI）的综述与展望文章，重点讨论大语言模型如何从被动文本生成器演变为拥有计划、记忆、工具使用与多轮推理能力的自治系统。作者首先回顾了从统计语言模型、神经语言模型、RNN/LSTM、Transformer 到大规模预训练 LLM 的发展历程，说明全局注意力、长程依赖建模和指令微调等技术如何为“智能体行为”（如多步推理、工具调用、自反思）奠定基础。随后，论文给出 Agentic AI 的概念与架构框架：以 LLM 为“脑”，外接感知模块、记忆模块（短期/长期）、工具与环境接口，以及行动执行模块，通过“感知-推理-行动-反思”的闭环实现单智能体与多智能体两种模式的自治行为。文中通过 ReAct 风格单智能体流程、AutoGen 式多智能体协作等典型案例，展示了金融问答、研究工作流、人机协作与机器人控制等应用场景；最后系统分析了安全与对齐、可靠性与鲁棒性、长期记忆一致性、伦理与法律责任、计算与环境成本等关键挑战，并提出未来在可验证规划、多智能体协调、可持续推理与治理审计等方向的研究议程。

Table 1. Four decades of progress in language modeling and their relevance to agentic AI.

Decade	Key Advances	Representative Models	Relevance to agentic AI
1990s	Statistical LMs; smoothing/back-off	n-gram SLMs [5]	Established sequence prediction as a core task, enabling later planning and decision-making.
2000s	Neural LMs; distributed representations	Neural LM [6], RNN LMs [7]	Introduced semantic and contextual representations, supporting richer reasoning for agent behavior.
2010s	Embeddings; Transformers; pre-training	word2vec [8], Transformer [10], BERT [1], GPT-2 [12]	Transformers enabled global context, structured reasoning, and scalable pre-training foundations of modern agentic systems.
2020s	Scaled LLMs; instruction tuning; multimodal models	GPT-3 [13], PaLM [14], LLaMA [15], GPT-4 [18]	Large-scale models exhibit emergent reasoning, tool use, and autonomy, which are essential for planning and adaptive agent behavior.

主要亮点

- **从“语言模型演进”到“智能体范式”的系统串联：**将四十年语言建模进展与 Agentic AI 能力——对齐，明确指出 Transformer + 大规模预训练 + RLHF/指令微调如何自然导向具备工具使用、规划和自反思能力的智能体系统。
- **提出清晰的 Agentic AI 组件与闭环框架：**给出统一的架构视图，将环境/工具、感知、LLM “脑”（推理与规划）、外部记忆和行动执行等模块串联成 “Reason-Act-Reflect” 循环，区分单智能体与多智能体系统，并用表格对比传统符号 AI、纯 LLM 与现代 Agentic AI 的差异。
- **全面梳理安全、可靠性与记忆相关风险：**从长链行动误差累积、非确定性行为、闭源模型难以审计，到长期记忆漂移、隐私泄露和偏见放大，系统提出配额控制、权限/角色约束、可回滚日志、层次化记忆和遗忘机制等缓解思路。
- **强调治理与可持续性并给出研究路线图：**指出 Agentic AI 在责任划分、合规与监管上的空白，呼吁建立可解释与可审计机制，同时将“混合符号-神经架构、多智能体协调协议、节能与自适应推理”列为未来关键研究方向，兼顾技术演进与社会责任。

相关链接

Paper: <https://arxiv.org/pdf/2601.02749>

● 论文十

mHC: Manifold-Constrained Hyper-Connections

日期

2026-01-06

*m*HC: Manifold-Constrained Hyper-Connections

Zhenda Xie^{*†}, Yixuan Wei^{*}, Huanqi Cao^{*},
Chenggang Zhao, Chengqi Deng, Jiashi Li, Damai Dai, Huazuo Gao, Jiang Chang,
Kuai Yu, Liang Zhao, Shangyan Zhou, Zhean Xu, Zhengyan Zhang, Wangding Zeng,
Shengding Hu, Yuqing Wang, Jingyang Yuan, Lean Wang, Wenfeng Liang

DeepSeek-AI

内容摘要

这篇论文针对 Hyper-Connections (HC) 在扩展残差流宽度、丰富连接模式以提升模型性能的同时，出现的破坏恒等映射特性、导致训练不稳定性、限制可扩展性及增加内存访问开销等问题，提出了流形约束超连接 (*m*HC) 框架。*m*HC 的核心是通过 Sinkhorn-Knopp 算法将 HC 的残差映射投影到由双随机矩阵构成的流形 (Birkhoff 多胞体)，恢复恒等映射特性，确保信号传播中特征均值守恒、范数正则化，缓解梯度消失或爆炸问题；同时结合核融合、选择性重计算、DualPipe 调度中通信与计算重叠等基础设施优化，降低系统开销。实验表明，*m*HC 在语言模型预训练中表现出卓越的稳定性和可扩展性，在 3B、9B、27B 等不同规模模型上均优于基线模型和 HC，仅引入 6.7% 的额外时间开销 ($n=4$ 时)，且在推理、阅读理解等多类下游任务中性能显著提升。

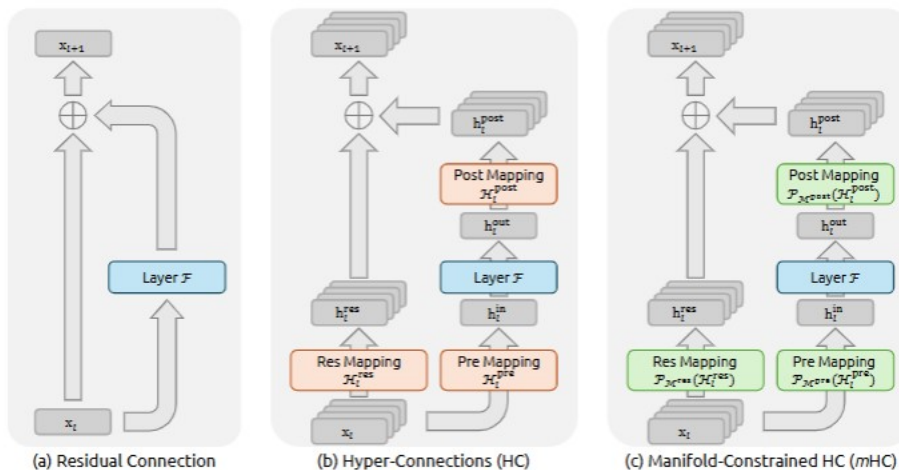


Figure 1 | **Illustrations of Residual Connection Paradigms.** This figure compares the structural design of (a) standard Residual Connection, (b) Hyper-Connections (HC), and (c) our proposed **Manifold-Constrained Hyper-Connections (*m*HC)**. Unlike the unconstrained HC, *m*HC focus on optimizing the residual connection space by projecting the matrices onto a constrained manifold to ensure stability.

主要亮点

- **创新性流形约束设计，解决核心稳定性问题：**首次将残差映射约束在双随机矩阵流形上，通过数学性质（谱范数 ≤ 1 、矩阵乘法封闭性）从根本上恢复恒等映射特性，使多层层叠后的复合映射仍保持稳定，将 HC 中近 3000 的信号增益峰值降低三个数量级，彻底解决了 HC 在大规模训练中的信号发散与梯度不稳定问题。
- **高效基础设施优化，平衡性能与开销：**针对 HC 的内存访问和通信开销痛点，设计了三重优化方案——核融合减少内存带宽瓶颈、选择性重计算降低显存占用、扩展 DualPipe 调度实现通信与计算重叠，使 mHC 在保持 HC 拓扑结构优势的同时，仅引入边际计算开销（ $n=4$ 时为 6.7%），突破了“宽残差流 = 高开销”的限制。
- **强泛化性与可扩展性，适配大规模模型训练：**mHC 作为 HC 的通用扩展框架，兼容不同模型规模（3B-27B）和训练场景（计算量缩放、令牌量缩放），在多下游基准测试（BBH、DROP、GSM8K 等）中持续优于基线 HC，尤其在推理类任务上提升显著（如 BBH 提升 2.1%），为下一代基础模型的拓扑结构设计提供了新方向。

相关链接

Paper: <https://arxiv.org/pdf/2512.24880>

● 论文十一

Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking

日期

2026-01-08

Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking

Mingxin Li* Yanzhao Zhang* Dingkun Long* Keqin Chen
Sibo Song Shuai Bai Zhibo Yang Pengjun Xie
An Yang Dayiheng Liu Jingren Zhou Junyang Lin
Tongyi Lab, Alibaba Group

 <https://huggingface.co/collections/Qwen>
 <https://modelscope.cn/organization/qwen>
 <https://github.com/QwenLM/Qwen3-VL-Embedding>

内容摘要

该论文介绍了 Qwen3-VL-Embedding 和 Qwen3-VL-Reranker 模型系列，它们是基于 Qwen3-VL 基础模型构建的 Qwen 家族最新扩展，为高精度多模态检索提供端到端解决方案。模型可将文本、图像、文档图像、视频等多种模态映射到统一的表示空间，其中 Qwen3-VL-Embedding 采用多阶段训练范式（从大规模对比预训练到重排序模型蒸馏），支持嵌套表示学习（灵活选择嵌入维度）和 32k tokens 输入；Qwen3-VL-Reranker 则通过交叉编码器架构和交叉注意力机制实现查询 - 文档对的细粒度相关性评估。两款模型均继承了 Qwen3-VL 的多语言能力（支持 30 余种语言），提供 2B 和 8B 参数版本以适配不同部署需求。实验表明，Qwen3-VL-Embedding 系列在多个多模态嵌入评估基准上取得最先进结果，其中 Qwen3-VL-Embedding-8B 在 MMEB-V2 基准上获得 77.8 分，且在纯文本任务中也表现出竞争力。

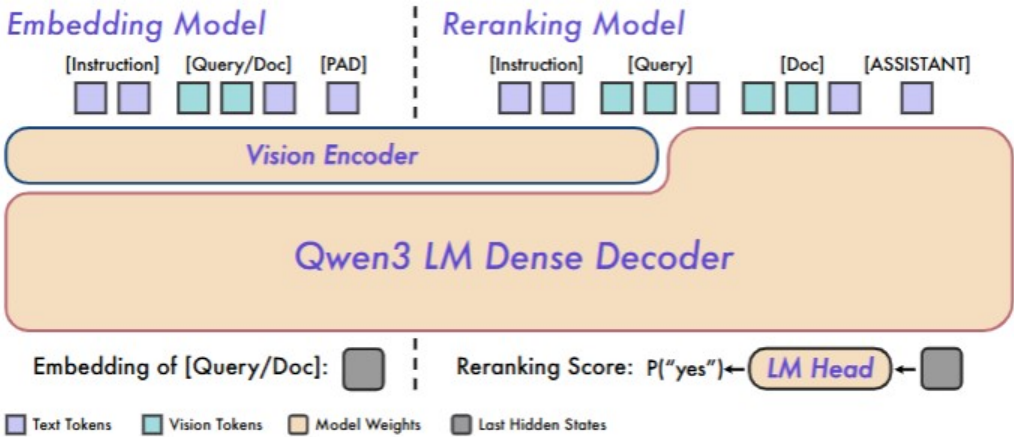


Figure 2: Overview of the Qwen3-VL-Embedding and Qwen3-VL-Reranker architecture.

主要亮点

- **统一多模态表示与全场景覆盖**：创新性地将文本、图像、文档图像、视频等多种模态整合到统一表示空间，实现跨模态语义对齐，支持图像 - 文本检索、视觉问答、视频 - 文本匹配等多样化多模态检索任务，同时兼容纯文本任务，打破了传统单模态或有限模态检索的局限。
- **灵活高效的模型设计与部署适配**：通过嵌套表示学习（MRL）支持用户按需选择嵌入维度，结合量化感知训练（QAT）确保低精度（如 int8）下的性能稳定性，大幅降低存储和计算成本；支持 32k 长序列输入，可处理长文档和视频，且提供 2B、8B 不同参数版本，满足从轻量化部署到高精度需求的各类场景。
- **多阶段训练与卓越性能表现**：采用“对比预训练 - 多任务对比学习与监督微调 - 蒸馏与模型融合”的多阶段训练 pipeline，有效解决数据不平衡问题，提升模型泛化能力和相关性评估精度。Qwen3-VL-Embedding-8B 在 MMEB-V2 基准上刷新纪录，重排序模型相较于同参数规模竞品表现更优，且多语言支持能力进一步拓展了应用范围。

相关链接

Paper: <https://arxiv.org/pdf/2601.04720>

Hugging Face: <https://huggingface.co/collections/Qwen>

ModelScope: <https://modelscope.cn/organization/qwen>

GitHub: <https://github.com/QwenLM/Qwen3-VL-Embedding>

Blog: <https://qwen.ai/blog?id=qwen3-vl-embedding>

● 论文十二

Conditional Memory via Scalable Lookup: A New Axis of Sparsity for Large Language Models

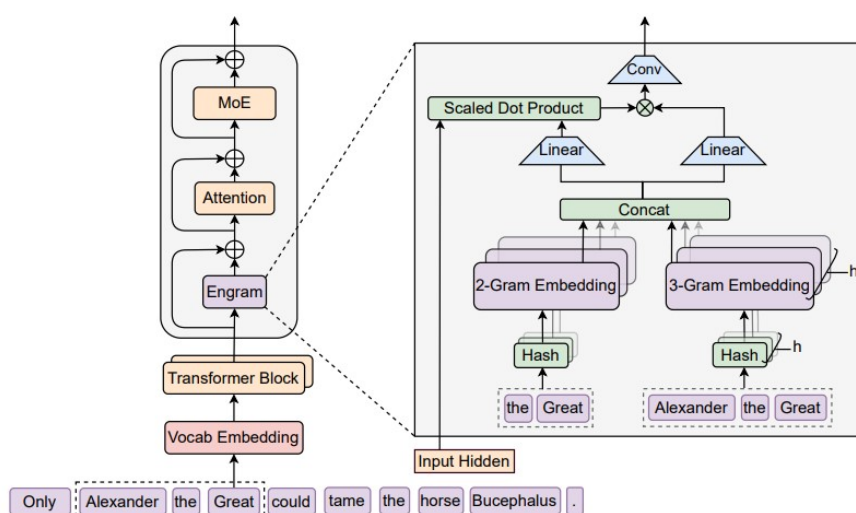
日期

2026-01-12

Conditional Memory via Scalable Lookup: A New Axis of Sparsity for Large Language Models

Xin Cheng^{1,2}, Wangding Zeng², Damai Dai², Qinyu Chen², Bingxuan Wang²,
Zhenda Xie², Kezhao Huang², Xingkai Yu², Zhewen Hao², Yukun Li², Han Zhang²,
Huishuai Zhang¹, Dongyan Zhao¹, Wenfeng Liang²
¹Peking University ²DeepSeek-AI
{zhanghuishuai, zhaody}@pku.edu.cn
{chengxin, zengwangding, damai.dai}@deepseek.com

内容摘要



该论文针对当前大语言模型（LLM）中混合专家模型（MoE）仅依赖“条件计算”处理所有语言任务的低效问题，提出“条件记忆（Conditional Memory）”这一全新稀疏性维度，与 MoE 形成互补。核心创新是 Engram 模块，它通过现代化改造经典 N-gram 嵌入，实现 $O(1)$ 时间复杂度的静态知识查找，将命名实体、固定搭配等静态局部依赖从神经计算中剥离。论文通过构建“稀疏性分配”问题，发现 MoE 与 Engram 参数分配的 U 型缩放定律，并基于此训练出 27B 参数的 Engram 模型。实验表明，该模型在知识检索、通用推理、代码、数学及长上下文任务中均显著优于同等参数和计算量的 MoE 基线，且通过确定性寻址实现百亿级参数表卸载至主机内存，推理开销低于 3%。

主要亮点

- **新增稀疏性维度**: 提出 “条件记忆” 与 MoE 的 “条件计算” 形成双轴互补, 用 Engram 模块实现 O(1) 静态知识查找, 解决 Transformer 用高成本计算模拟查表的低效问题。
- **量化分配定律**: 发现 U 型缩放定律, 明确 20%-25% 稀疏参数分配给 Engram 最优, 27B 参数模型在推理、代码等多领域性能显著超越同等配置 MoE。
- **高效工程实现**: 通过分词器压缩、上下文门控等技术优化, 支持百亿级参数表卸载至主机内存, 推理开销 < 3%, 突破 GPU 显存限制。

相关链接

Paper: <https://arxiv.org/pdf/2601.07372>

● 论文十三

An Information-Theoretic Perspective on LLM Tokenizers

日期

2026-01-14

An Information-Theoretic Perspective on LLM Tokenizers

Mete Erdogan^{1*}, Abhiram Gorle^{1*}, Shubham Chandak², Mert Pilanci¹, Tsachy Weissman¹

¹Department of Electrical Engineering, Stanford University, USA, ²AWS Annapurna ML
{merdogan, abhiramg, pilanci, tsachy}@stanford.edu, shubhamchandak94@gmail.com

内容摘要

该论文从信息论视角系统研究了大语言模型分词器 (tokenizer)，将其视为一种“结构化压缩器”，分析它如何在不同语料、词表规模与训练规模下影响压缩效率、token 流的统计结构以及在域外迁移时的鲁棒性。作者首先把 GPT 系列预训练分词器当作黑箱压缩器，在新闻、代码、数学等域以及多语言语料上与传统压缩算法 (gzip、zstd 等) 对比，发现新一代 o200kbase 在代码和多语言场景中更具压缩性与脚本感知能力，而旧分词器在非拉丁文字上存在严重过切分问题。随后，论文在英语、土耳其语、中文和代码等域上从零训练 BPE、Unigram、WordPiece、WordLevel 等分词器，系统考察训练数据规模和词表大小对压缩率与 k-gram 熵的影响，展示出随训练规模增大“一元熵上升、高阶条件熵下降”的一致趋势，即 token 流整体更丰富但局部更可预测，同时指出在训练-测试域不匹配时这些收益会显著退化。此外，作者还研究了“分词 + 通用压缩器”的两级管线，证明分词可作为通用压缩前的有益变换，并提出一个以 gzip 压缩率为目标的 LZ-aware BPE 变体，展示在相同词表规模下可获得更好的压缩效果；最后通过“无噪声信道”视角定义容量利用率指标，揭示词表大小、token 频率分布、训练规模之间的结构性权衡，并讨论对下游建模（如困惑度）的潜在启示。

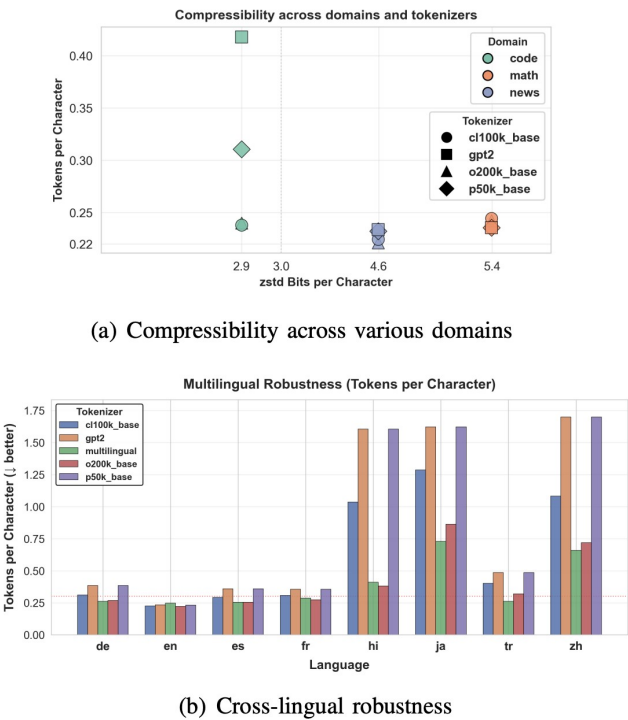


Fig. 1: Analyzing the GPT-family Tokenizers

主要亮点

- **信息论统一框架下的分词器刻画**：把分词器视为有限字母表信道上的结构化压缩器，同时量化“压缩效率 (bits/char, bytes/token)”与“诱导结构 (k-gram 熵、容量利用率)”，给出比单纯“token 数”更丰富的分析维度。
- **训练规模与词表规模上的系统 scaling 规律**：在多种域和多种分词算法上，揭示出“训练数据越大，一元熵升高、高阶条件熵降低，token 流局部可预测性增强”的普适趋势，并指出当词表容量受限或训练-测试域不匹配时，这种改进会出现非单调甚至退化。
- **分词 + 通用压缩”的协同与 LZ-aware BPE**：实验证明先分词再对 token 序列进行 LZ 压缩，比直接对 UTF-8 字节压缩可进一步提升 10–20% 的压缩率，并通过在 BPE 训练中直接以 gzip 长度作为贪心目标，构造出面向通用压缩最优的 LZ-aware BPE，展示了“下游目标感知分词器设计”的可行路径。
- **信道容量利用率作为新指标**：提出用香农熵和 Rényi 熵归一化到 $\log_2 |V|$ 的“容量利用率”度量分词器如何使用其词表空间，结合 k-gram 熵表明 BPE/WordPiece 在“高容量利用 + 吸收短程结构”之间取得更优折中，而词级分词虽可达到相似的一元利用率，却保留较多短程冗余，为未来基于信道视角的“压缩-建模联合优化”提供理论与实证依据。

相关链接

Paper: <https://arxiv.org/pdf/2601.09039>

● 论文十四

DanQing: An Up-to-Date Large-Scale Chinese VisionLanguage Pre-training Dataset

日期

2026-01-15

DanQing: An Up-to-Date Large-Scale Chinese Vision-Language Pre-training Dataset

Hengyu Shen*, Tiancheng Gu*, Bin Qin, Lan Wu, Yuling Wu, Shuo Tan, Zelong Sun
Jun Wang, Nan Wu, Xiang An, Weidong Cai, Ziyong Feng[†], Kaicheng Yang[†]

DanQing Team,  Glnt Lab

内容摘要

论文提出了面向中文视觉-语言预训练（VLP）的数据集 DanQing。论文从 2024–2025 年的 Common Crawl 中构建、清洗并得到约 1 亿高质量中文图文对，用于提升 CLIP/SigLIP 类双塔模型的跨模态表征与检索能力。论文给出从约 10.5 亿原始样本到约 1 亿成品的多阶段过滤与去重流水线（粗过滤、文本细过滤、图像细过滤、跨模态对齐过滤、跨批次去重），并对数据统计特征、主题分布、来源域名分布做了分析。实验方面，作者以 SigLIP2 为基座进行持续预训练，对比 Wukong/Zero/TaiSu 等中文数据集；结果显示在零样本分类、短/长文本的跨模态检索，以及将视觉编码器接入 LLaVA-NeXT 式 LMM 的中文评测上，使用 DanQing 预训练的模型整体表现更优，并展示了对 2024 后新概念（如“黑神话悟空”“小米 SU7”）的更好匹配能力。



Figure 1: Overview of the DanQing dataset construction pipeline.

主要亮点

- **面向中文 VLP 的“高质量+新鲜度”数据**：DanQing 强调数据来自 2024–2025 网络内容，并通过严格清洗提升语义密度与图文一致性，缓解中文高质量图文资源稀缺问题
- **工程化可复用的数据构建流水线**：从安全、文本长度/语言、信息密度、图像清晰度/熵、冗余聚类去重到基于检索模型的跨模态相似度阈值过滤，形成可规模化处理的端到端数据治理流程，并公开每一步的过滤统计。
- **训练收益覆盖“检索+多模态大模型”**：在 SigLIP2 持续预训练下，DanQing 在零样本分类与多项中文跨模态检索基准上优于对比数据集；作为 LMM 视觉编码器时，也带来中文多模态评测均分提升，体现数据对下游多模态推理/OCR 等任务的迁移价值。

相关链接

Paper: <https://arxiv.org/pdf/2601.10305>

Hugging Face: <https://huggingface.co/datasets/DeepGlint-AI/DanQing100M>

Project: <https://deepglint.github.io/DanQing/>

● 论文十五

A Safety Report on GPT-5.2, Gemini 3 Pro, Qwen3-VL, Grok 4.1 Fast, Nano Banana Pro, and Seedream 4.5

日期

2026-01-16

A Safety Report on GPT-5.2, Gemini 3 Pro, Qwen3-VL, Grok 4.1 Fast, Nano Banana Pro, and Seedream 4.5

Xingjun Ma^{1,2} Yixu Wang¹ Hengyuan Xu¹ Yutao Wu³ Yifan Ding¹ Yunhan Zhao¹ Zilong Wang¹ Jiabin Hua¹ Ming Wen^{1,2} Jianan Liu^{1,2} Ranjie Duan Yifeng Gao¹ Yingshui Tan Yunhao Chen¹ Hui Xue Xin Wang¹ Wei Cheng Jingjing Chen¹ Zuxuan Wu¹ Bo Li⁴ Yu-Gang Jiang¹

Fudan University¹ Shanghai Innovation institute² Deakin University³ UIUC⁴



<https://xsafeai.github.io/AI-safety-report>



<https://github.com/XSafeAI/AI-safety-report>

内容摘要

该论文系统评测了当前前沿多模态与大语言模型（GPT- 5.2、Gemini 3 Pro、Qwen3- VL、Grok 4.1 Fast 等）在语言安全、视觉语言安全以及文本生成图像安全三个层面的整体安全性与合规性。作者构建了统一的安全评估框架，涵盖：标准安全基准（如 ALERT、Flames、BBQ、SORRY- Bench、StrongREJECT）、30 种黑盒越狱攻击策略、多语言安全判别（18 种语言）、以及基于 NIST AI RMF、EU AI Act、MAS FEAT 等法规的合规性测试，形成“基准表现 + 对抗鲁棒 + 多语言 + 法规合规”四个维度的立体分析。在语言模型部分，结果显示 GPT- 5.2 在综合安全性、对抗鲁棒性和法规合规性上总体领先，但在极端最坏情况（worst - case）下仍存在显著越狱风险；Qwen3- VL 拒绝类安全表现较好，却在偏见问答（如 BBQ）和对抗攻击下暴露出系统性薄弱点。在视觉- 语言部分，论文使用 MemeSafetyBench、USB- SafeBench、MM- SafetyBench、VLJailbreakBench 等基准，发现多模态模型在显性有害内容上防护较强，但对“隐含伤害、复合推理、角色扮演 + 图像”等复杂攻击仍较脆弱。在文本生成图像部分，作者以 Nano Banana Pro 和 Seedream 4.5 为例，在 T2ISafety 与基于中国《生成式人工智能服务管理暂行办法》的私有合规基准上比较两者；结果表明，两模型在涉暴、涉黄上有较强显式过滤，但在仇恨符号、政治操纵、隐私权和知识产权等抽象法规场景上仍有明显违规生成。

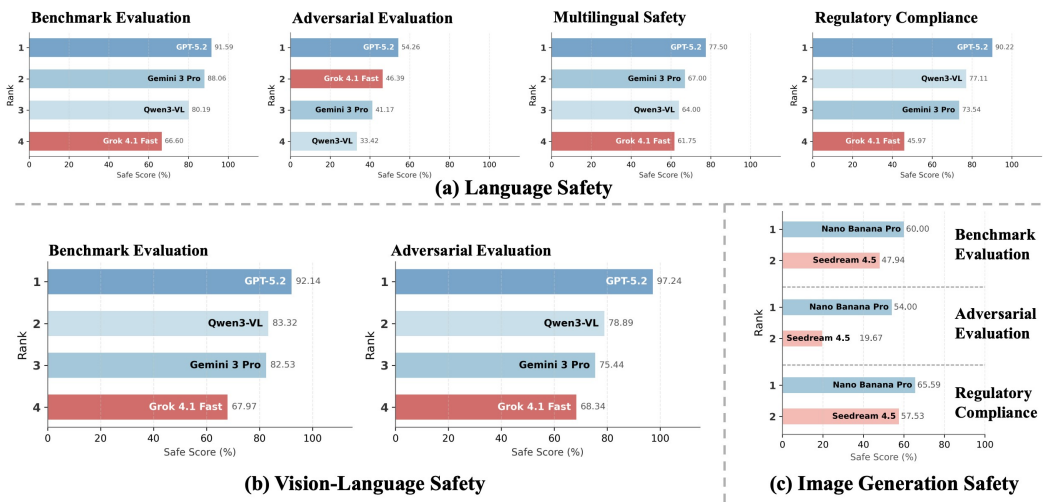


Figure 1: Safety leaderboards of the 7 evaluated frontier models across four dimensions: Benchmark Evaluation, Adversarial Evaluation, Multilingual Evaluation, and Compliance Evaluation. (a) Language Safety Leaderboard; (b) Vision-Language Safety Leaderboard; (c) T2I Safety Leaderboard.

主要亮点

- **跨模态、跨语言、跨法规的一体化安全框架**：论文首次将语言、视觉-语言、文生图三个模态的安全评测统一在一个体系中，并同时覆盖英文与多语种、一般安全基准与法规约束，为未来“端到端安全能力”对比提供通用模板
- **从“安全分数”到“结构性薄弱点”的深入剖析**：不仅报告安全率、拒绝率、最坏情况安全率等指标，还通过大量失败案例（偏见问答、模板越狱、多轮对话越狱、代码包装、人物/权威角色扮演等）揭示当前对齐方法常见的“表层关键词匹配、防御随轮次漂移、免责声明+实质性泄露”等结构性问题。
- **法规落地视角的合规评测方法论**：通过 SafeEvalAgent 将 NIST、EU AI Act、FEAT 等法规条文自动拆解为可测试的“原子规则 + 对抗/合规指引”，并生成多种格式（开放问答、选择题、真假题）的测试用例，用合规率衡量模型在预测性警务、生物识别、透明度、知识产权等高风险领域的“合规执行力”，为未来“法规-到-评测”的自动化流程提供可复用范式。

相关链接

Paper: <https://arxiv.org/pdf/2601.10527>

● 论文十六

LLM2CLIP: Powerful Language Model Unlocks Richer Cross-Modality Representation

日期

2026-1-17 (AAAI 2026 outstanding paper)

LLM2CLIP: Powerful Language Model Unlocks Richer Cross-Modality Representation

WeiQuan Huang^{1*}, Aoqi Wu^{1*}, Yifan Yang^{2†‡}, Xufang Luo², Yuqing Yang², Usman Naseem³, Chunyu Wang², Qi Dai², Xiyang Dai², Dongdong Chen², Chong Luo², Lili Qiu², Liang Hu^{1†}

¹ School of Computer Science and Technology, Tongji University, Shanghai, China

² Microsoft Corporation

³ School of Computing, Macquarie University, Sydney, Australia
yifanyang@microsoft.com, lianghu@tongji.edu.cn

内容摘要

该论文提出 LLM2CLIP，一套将大语言模型能力高效注入预训练 CLIP 的两阶段微调框架，目标是在几百万级图文样本和几乎不增加算力成本的前提下，显著提升 CLIP 在长文本、多语言与多下游任务上的跨模态表征能力。第一阶段通过“caption 对比学习”专门把 LLaMA 3.1- 8B 这类 LLM 重新训练成适配 CLIP 场景的文本嵌入器：移除自回归掩码、改为双向注意力，用平均池化获得句向量，并采用有监督 SimCSE 对同图像的多条 caption 做对比学习，从而让 LLM 嵌入在 caption 空间中具备足够可分性。第二阶段用微量 DreamLIP/Recaption 等长密文本数据，对原始 CLIP/EVA02/SigLIP- 2 等模型执行“后微调”：冻结 LLM，仅在其输出后接入一个轻量线性 adaptor，与视觉编码器一起做标准 CLIP 对比训练，同时完全替换掉原始 CLIP 文本编码器；在 3M–60M 规模数据上微调后，LLM2CLIP 在英文短/长文检索、多语种检索、ImageNet 线性探测、零样本分割/检测以及作为 LLaVA- 1.5 视觉编码器等任务上取得显著提升。

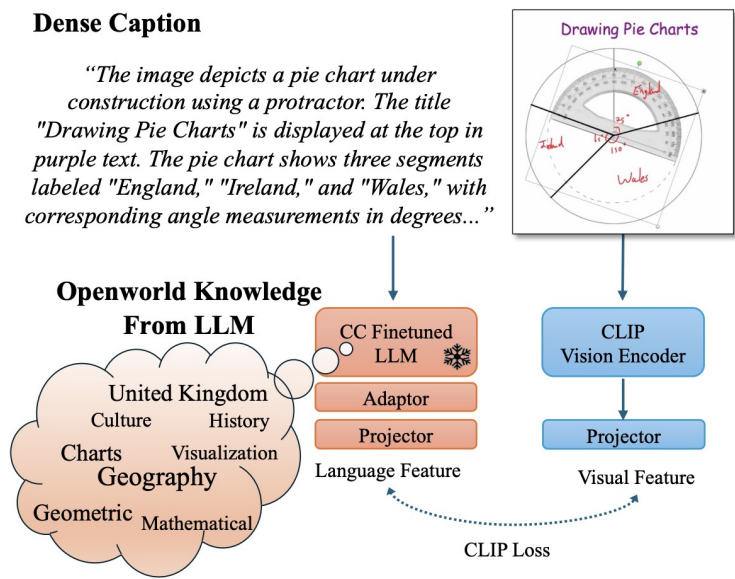


Figure 1: *LLM2CLIP* Overview. After applying caption contrastive fine-tuning to the LLM, the increased textual discriminability enables more effective CLIP training. We leverage the open-world knowledge and general capabilities of the LLM to better process dense captions, addressing the previous limitations of the pretrained CLIP visual encoder and providing richer textual supervision.

主要亮点

- **面向 CLIP 的专用 LLM “嵌入化” 策略：**通过双向注意力、句级平均池化和有监督 SimCSE，对 LLM 进行 caption 对比微调，使其在图文 caption 空间中具备远高于原始 LLM 的特征可分性，COCO caption- to- caption 检索 Top- 1 从约 5% 提升到接近甚至超过原始 CLIP/EVA02 水平。
- **长文本与多语言检索的大幅收益：**在 Flickr/COCO (短 caption)、ShareGPT4V、Urban1K、DOCCI 等五个英文检索基准上，相比原始 CLIP/EVA02，LLM2CLIP 在 ViT- L/336 设置下平均 Top- 1 检索精度可提升 10–15 个百分点；在 Flickr- CN、COCO- CN 与 XM- 3600 三个中- 多语数据集上，即便只用英文数据训练，也能显著超过 SigLIP- 2 与 CN- CLIP，体现出 LLM 带来的“自带多语”优势。
- **下游视觉任务与多模态大模型的全面增强：**在 ImageNet 上，LLM2CLIP 保持零样本分类性能基本不降的同时，提高线性探测精度；在零样本语义分割与开放词汇检测中，mIoU 和新类 AP 明显提升，例如基于 EVA02 的零样本分割 mIoU 全面提升数点，OV- COCO 新类 AP 也有可观增益；将 LLaVA- 1.5 的视觉编码器替换为 LLM2CLIP 后，在 87% 以上的多模态理解基准上取得提升，说明改进的跨模态空间能直接转化为复杂视觉推理能力的增强。

相关链接

Paper: <https://arxiv.org/pdf/2411.04997v5>

● 论文十七

Toward Efficient Agents: A Survey of Memory, Tool learning, and Planning

日期

2026-01-20

Toward Efficient Agents: A Survey of Memory, Tool learning, and Planning

Xiaofang Yang^{1,2,†} Lijun Li^{1,†,✉} Heng Zhou^{1,3,†} Tong Zhu^{1,†} Xiaoye Qu¹ Yuchen Fan^{1,4}
 Qianshan Wei⁵ Rui Ye⁴ Li Kang^{1,4} Yiran Qin⁶ Zhiqiang Kou⁷ Daizong Liu⁸ Qi Li⁵
 Ning Ding⁹ Siheng Chen⁴ Jing Shao^{1,✉}

¹ Shanghai Artificial Intelligence Laboratory, ² Fudan University,

³ University of Science and Technology of China, ⁴ Shanghai Jiaotong University,

⁵ Institute of Automation, Chinese Academy of Sciences,

⁶ The Chinese University of Hong Kong (Shenzhen),

⁷ Hong Kong Polytechnic University, ⁸ Wuhan University, ⁹ Tsinghua University

内容摘要

该论文是一篇面向“高效智能体 (Efficient Agents)”的综述，聚焦于在真实部署中常被忽视但极关键的效率问题：在保持任务成功率/效果的同时，降低 token 消耗、推理延迟、工具调用次数、交互步数与整体计算成本。作者提出以智能体三大核心模块为主线来系统化梳理相关研究——记忆 (Memory)、工具学习 (Tool learning) 与规划 (Planning)，并给出效率的两种互补刻画方式：在固定成本预算下比较效果，或在相近效果水平下比较成本（可用“效果-成本” Pareto 前沿理解）。论文从智能体相较纯 LLM 的“递归多步循环”出发，指出成本不仅来自生成 token，还来自记忆读写、外部工具交互与失败重试等额外开销，并给出面向 POMDP 的形式化建模来统一描述“状态-观测-动作-工具接口-显式记忆”。在具体内容上，作者按“记忆生命周期”（构建、管理、访问/整合）总结文本记忆、潜在记忆、外部检索记忆、图结构记忆、分层记忆以及多智能体共享/本地/混合记忆等路线；按工具学习过程总结工具选择、工具调用与工具融合推理；按规划策略总结自适应预算控制、结构化搜索、任务分解与学习型演化，并进一步归纳了面向效率的评测基准与常见指标体系。

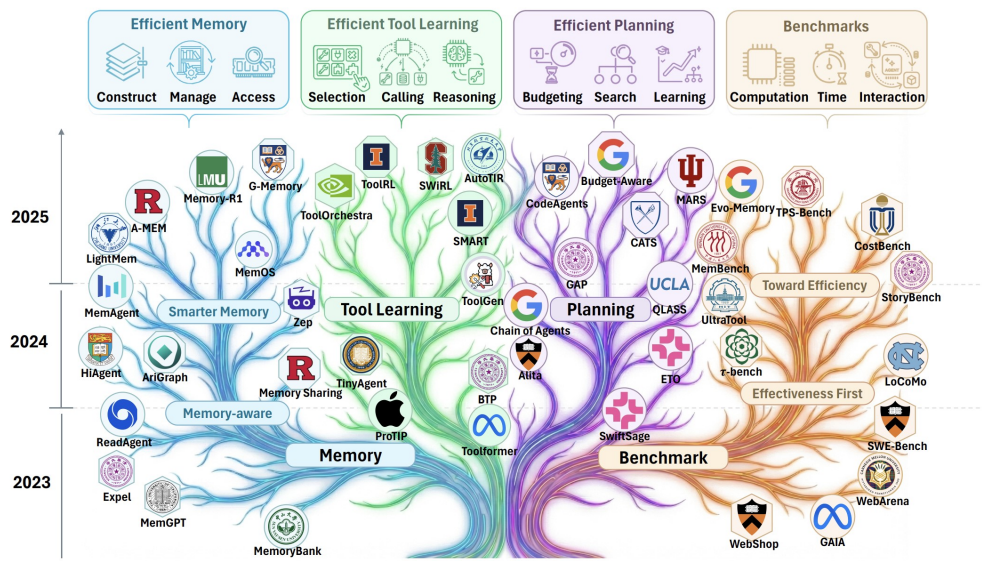


Figure 1: The evolutionary trajectory of efficient agent research. The diagram is organized into four principal branches: **Memory**, **Tool Learning**, **Planning**, and **Benchmarks**. Key works and their institutional affiliations are mapped chronologically to illustrate the field's development and categorization from 2023 to 2025.

主要亮点

- **效率视角的统一框架与术语体系：**将“高效”定义为系统级目标，而非单纯模型压缩；并用成本预算/同等效果成本两种视角与 Pareto 前沿，提供可对比、可复用的评价坐标系。
- **围绕记忆-工具-规划的全景式分类：**对大量工作按模块拆解并落到可执行机制（如记忆压缩与写入触发、分层存储与检索、去冗余与遗忘策略、减少工具调用的奖励设计、受控搜索减少规划步数等），帮助读者快速定位“提升效率的杠杆点”。
- **明确挑战与未来方向：**讨论压缩与性能的权衡、在线与离线记忆管理的成本-时效取舍，以及“智能体潜在空间推理”“面向部署的多智能体实现对比”“多模态智能体效率”等尚未充分解决但很有潜力的研究议题。

相关链接

Paper: <https://arxiv.org/pdf/2601.07372>

Project: <https://efficient-agents.github.io>

Github: <https://github.com/yxf203/Awesome-Efficient-Agents>

● 论文十八

OpenVision 3 : A Family of Unified Visual Encoder for Both Understanding and Generation

日期

2026-01-21

OpenVision 3 : A Family of Unified Visual Encoder for
Both *Understanding* and *Generation*

Letian Zhang*¹ Sucheng Ren*² Yanqing Liu¹ Xianhang Li¹ Zeyu Wang¹ Yuyin Zhou¹
Huaxiu Yao³ Zeyu Zheng⁴ Weili Nie⁵ Guilin Liu⁵ Zhiding Yu⁵ Cihang Xie¹
¹UC Santa Cruz ²JHU ³UNC-Chapel Hill ⁴UC Berkeley ⁵NVIDIA

内容摘要

该论文提出了 OpenVision 3，一套面向“理解 + 生成”统一建模的视觉编码器家族，目标是在同一视觉表征空间内同时服务于多模态理解任务（如视觉问答、图文对齐）和高质量图像生成。模型采用“VAE 编码器 + ViT 编码器”的连续统一 tokenizer：先用 FLUX.1 VAE 将图像压缩到潜空间，再用从零训练的 ViT 在 VAE latent 上提取统一视觉 token，之后分别送入重建分支（ViT 解码 + VAE 解码，优化像素/latent 重建与 LPIPS）和理解分支（对比学习 + 图像字幕生成）。通过联合优化重建损失和语义损失，OpenVision 3 学到的表征在冻结 encoder 的前提下，在 LLaVA- 1.5 框架中实现与 CLIP 相当甚至略优的多模态理解性能（如 SeedBench、POPE），同时在重建与生成任务中显著优于已有统一 tokenizer，并在 RAE 生成框架下在 ImageNet 上取得更低的 gFID（1.89 vs. 2.54），展示出强大的统一视觉表征能力。

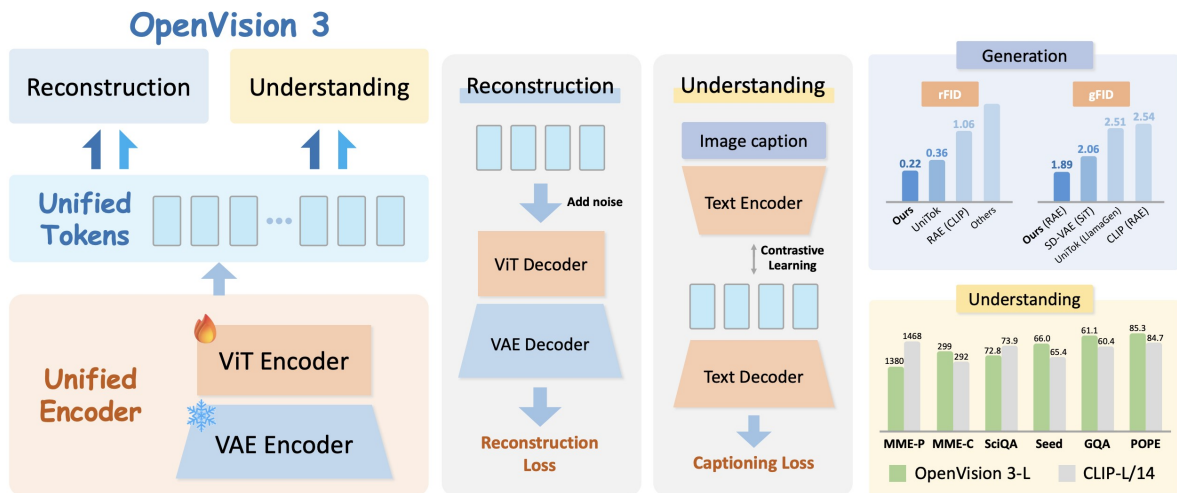


Figure 1. An overview of OpenVision 3's architecture design and performance highlight. **Left panel:** The architecture of OpenVision 3. We employ a frozen VAE and a trainable ViT as the unified tokenizer, which produces tokens that are fed simultaneously into both the generation and understanding branches. **Middle panel:** The learning objectives of the generation branch and the understanding branch. For the generation branch, we focus on high-quality, pixel-level image reconstruction; concurrently, the understanding branch is optimized via joint contrastive learning and captioning objectives. **Right panel:** The performance summarization shows that OpenVision 3 outperforms other unified tokenizers and semantics-based encoders in rFID and gFID, while remaining competitive with CLIP in multimodal understanding ability.

主要亮点

- **简洁而统一的连续视觉 tokenizer 设计：**通过“VAE latent + ViT”构建单一连续表征空间，避免了许多统一多模态模型中“语义 encoder + 生成 encoder 双路 token 化”的复杂设计，在保持架构简单的同时兼顾语义和像素细节。
- **重建与语义目标的协同训练范式：**提出在同一 ViT 表征上同时施加重建损失（像素、latent、LPIPS）和对比 + captioning 语义损失，并通过消融实验证明二者存在显著正向互补——仅用语义损失也能提升重建质量，而加入语义损失还能进一步降低重建误差。
- **强重建与生成能力：**在 ImageNet / COCO 上，OpenVision 3 的重建 PSNR/SSIM/LPIPS/rFID 明显优于 UniTok、OmniTokenizer、Vila- U 等统一 tokenizer，接近甚至部分优于专门面向生成的 VAE 系列；在 RAE 生成框架下，gFID 从 CLIP- RAE 的 2.54 降至 1.89，表明统一表征在生成侧也能优于纯语义 encoder。

相关链接

Paper: <https://arxiv.org/pdf/2601.15369v1>
Project: <https://ucsc-vlaa.github.io/OpenVision3/>
Model: <https://huggingface.co/collections/UCSC-VLAA/openvision-3>

● 论文十九

Qwen3-ASR Technical Report

日期

2026-01-29

Qwen3-ASR Technical Report

Qwen Team

👤 <https://huggingface.co/collections/Qwen/qwen3-asr>
 <https://modelscope.cn/collections/Qwen/Qwen3-ASR>
 <https://github.com/QwenLM/Qwen3-ASR>

内容摘要

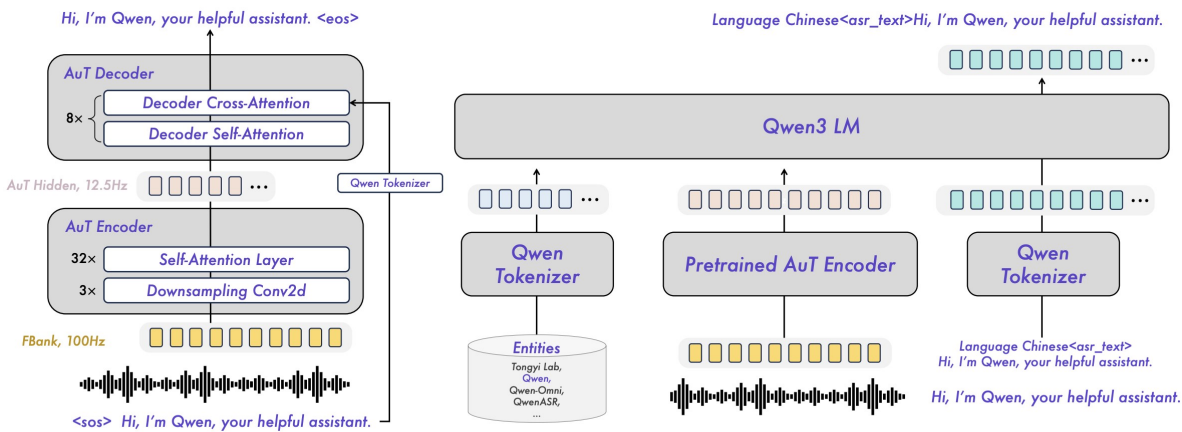


Figure 2: Architecture of AuT (left) and the overview of Qwen3-ASR (right).

该技术报告介绍了 Qwen3-ASR 家族, 包括两款多语言 “一体化” 语音识别模型 Qwen3- ASR- 1.7B / 0.6B, 以及一款非自回归时间戳对齐模型 Qwen3- ForcedAligner- 0.6B, 整体基于大音频- 语言模型 (LALM) 范式并以 Qwen3- Omni 作为基础模型。 Qwen3- ASR 通过独立预训练的 AuT 编码器进行 8 倍下采样声学建模, 再与多模态 LLM 结合, 经 ASR 监督微调和 GSPO 强化学习四阶段训练, 在 40M 小时伪标注语音与 3T 级多模态 token 上获得强音频理解能力, 支持 30 种语言、22 种汉语方言的语言识别与转写, 并兼顾长时 (单段 20 分钟)、流式推理、嘈杂环境与歌声识别。在 LibriSpeech、WenetSpeech、CommonVoice、Fleurs、MLS、MLC- SLM 等公开基准及自建大规模内部压力测试上, Qwen3- ASR- 1.7B 在英、中及多语场景下整体优于 Whisper- large- v3、FunASR- MLT- Nano、GLM- ASR- Nano 等开源系统, 并在多项指标上接近或超越 GPT- 4o- Transcribe、Gemini- 2.5- Pro、Doubao- ASR 等闭源 API, 而 0.6B 版本在精度与效率之间提供更优折中。 Qwen3- ForcedAligner- 0.6B 将强制对齐建模为 “时间戳槽填充” 任务, 在多语种 AuT 编码器与 Qwen3- 0.6B LLM 基础上通过离散帧索引预测词/字起止时间, 支持 11 种语言、最长 300 秒音频, 在 Accumulated Average Shift (AAS) 指标上相对 Monotonic- Aligner、NeMo Forced Aligner、WhisperX 等主流方法将位移误差降低约 67–77%, 同时依靠非自回归推理实现近 0.001 的实时因子。

主要亮点

- **面向真实场景的一体化多语种 ASR:** Qwen3- ASR- 1.7B / 0.6B 覆盖 30 种语言与 22 种汉语方言, 统一支持语言识别与转写, 在英文 (GigaSpeech、Tedlium、VoxPopuli 等)、普通话 (WenetSpeech、AISHELL- 2 等) 及方言 (粤语、川话等) 公开数据集上, 整体 WER/CER 明显优于 Whisper- large- v3 等开源模型, 并在大部分中文任务上达到或超越商用 API 水平。

- **噪声、口音、弱势人群与歌声的鲁棒性：**在自建“口音英语对话”“老年/儿童普通话”“极低信噪比”“绕口令”“多方言对话”等内部测试中，Qwen3- ASR- 1.7B 在所有子集上取得最佳或并列最佳误差率，且在 M4Singer、MIR- 1k- vocal、Popcs 等歌声基准以及整首带 BGM 歌曲（中英 EntireSongs）上显著优于 Whisper- large- v3 与 FunASR- MLT- Nano，体现出对旋律、伴奏与节奏变化的强适应能力。
- **高效长语音与流式统一推理：**基于 AuT 编码器 12.5Hz token 率和动态 1–8 秒注意窗，模型天然支持离线长语音与在线流式统一架构，单段可处理约 20 分钟音频；在 vLLM 环境下，Qwen3- ASR- 0.6B 的平均首 token 时延最低可达 92ms，在并发 128 时实时因子约 0.064、吞吐量约 2000s/s，1.7B 版本在保证精度的同时也维持亚实时性能，适合工业级多并发部署。
- **大模型驱动的多语种强制对齐方案：**Qwen3- ForcedAligner- 0.6B 摒弃 CTC/CIF 等传统后处理流水线，采用 LLM+AuT 的非自回归时间戳分类架构，通过将带有起止 time 槽的转写序列作为输入，直接一次性预测整句所有槽位的时间索引，支持词级、字符级乃至跨语言 code- switch 场景，同时在长达 300 秒拼接语音上保持稳定 AAS，相较 WhisperX 等方法在长句、跨语种条件下位移误差大幅下降。

相关链接

Paper: <https://arxiv.org/pdf/2601.21337>

GitHub: <https://github.com/QwenLM/Qwen3-ASR>

Hugging Face: <https://huggingface.co/collections/Qwen/qwen3-asr>

ModelScope: <https://modelscope.cn/collections/Qwen/Qwen3-ASR>

Blog: <https://qwen.ai/blog?id=qwen3asr>

TECHNICAL UPDATES

技术动态

● 技术动态一

Science in 2050: the future breakthroughs that will shape our world — and beyond

日期

2025-12-30

Feature

THE SCIENCE OF 2050

Nature explores the future breakthroughs that could shape our world — and beyond. **By David Adam**



The Roman sage Marcus Aurelius said we should never let the future disturb us. But then he never had a conversation with the futurologist Nick Bostrom about the state of the world in 2050.

"There's a good likelihood that

by 2050, all scientific research will be done by superintelligent AI rather than human researchers," Bostrom said in an e-mail. "Some humans might do science as a hobby, but they wouldn't be making any useful contributions." Time to rethink your career options, *Nature* readers!

To adapt a cliché about computer models, predictions of the future are usually wrong, but some are interesting. And *Nature* has a long history of seeking stimulation in forecasts, projections and auguries about how research might unfold in the coming decades. Most notably, the journal marked the

内容摘要

这篇来自 nature 的新闻特写对 2050 年科学与世界发展趋势展开预测, AI 发展是核心亮点 —— 未来学家尼克·博斯特罗姆认为超级智能 AI 大概率将主导所有科学研究, 人类或仅以科研为爱好, 部分观点称届时 AI 可能取得诺贝尔级成果, 通用人工智能有望问世, 即便未全面接管, AI 结合机器人实验员也将通过 “无人实验室” 重塑科研模式, 但 AI 发展节奏存在不确定性。气候方面, 存在变暖超预期、单边地球工程干预的风险, 也有二氧化碳捕集产业化的可能; 太空探索有火星、水星探测等规划, 但人体太空生存的生物挑战存疑; 科研环境受民粹主义、公共开支紧缩影响, 纯研究或受挤压, 医疗研究因人口老龄化获更多投入, 数据短缺需依赖大规模志愿者解决; 量子科学与宇宙学结合可能破解暗物质等谜题, 核聚变、粘土电子学等技术有望突破, 2050 年系外行星发现或达 1 亿颗, 但地外生命确认仍面临技术与理论难题。

相关链接

技术博客: <https://www.nature.com/articles/d41586-025-04100-6>

● 技术动态二

北京智源人工智能研究院发布 2026 年十大 AI 技术发展趋势

日期

2026-01-08



内容摘要

2026 年 1 月 8 日，智源研究院举办研讨会并发布“2026 十大 AI 技术趋势”。当前 AI 行业处于关键范式转折点，技术重心从以参数规模为核心的语言学习转向对物理世界底层秩序的理解、建模与推演，即从“生成符号”走向“理解现实”。十大趋势涵盖多方面：世界模型成 AGI 共识方向，具身智能迈向工业场景规模化应用，多智能体系统标准化协议初具雏形，AI Scientist 成为发展重点且我国需突破科学基础模型短板，AI 应用沿巨头超级应用与垂直赛道双线发展，产业应用将先入“幻灭低谷期”后于 2026 下半年迎“V 型”反转，合成数据成真实数据战略性替代方案，推理优化仍有巨大潜力，开源编译器生态助力算力普惠，AI 安全从外部攻防与内生可解释性两方面演进并内化为系统本能防线。

相关链接

技术博客：https://mp.weixin.qq.com/s/PNWW0M87zyyYxu--1Spkcg?scene=1&click_id=42

● 技术动态三

智谱 AI 开源图像生成大模型 GLM-Image

日期

2026-01-14

内容摘要

GLM-Image 是智谱 AI 推出的首个开源工业级离散自回归图像生成模型，采用自回归模块（基于 GLM-4-9B-0414，90 亿参数）与扩散解码器（基于 CogView4 单流 DiT 结构，70 亿参数）的混合架构，核心优势在于复杂文字（中英多语言精准渲染、自动布局）与知识密集型场景的生成能力，同时支持物体增减、风格迁移等高精度图像编辑任务；其通过 semantic-VQ 视觉 token 选择、分阶段多分辨率训练及解耦式强化学习后训练等技术优化，在 CVTG-2K、LongText-Bench 等权威榜单中，文字渲染相关指标居开源模型前列，通用生成性能与主流模型持平，未来将聚焦专业内容创作、教育科普素材生成等场景，推动生成式 AI 向“理解与生成一体化”发展。

GLM-Image: Auto-regressive for Dense-knowledge and High-fidelity Image Generation

 Call it at Z.ai
  GitHub
  HuggingFace



相关链接

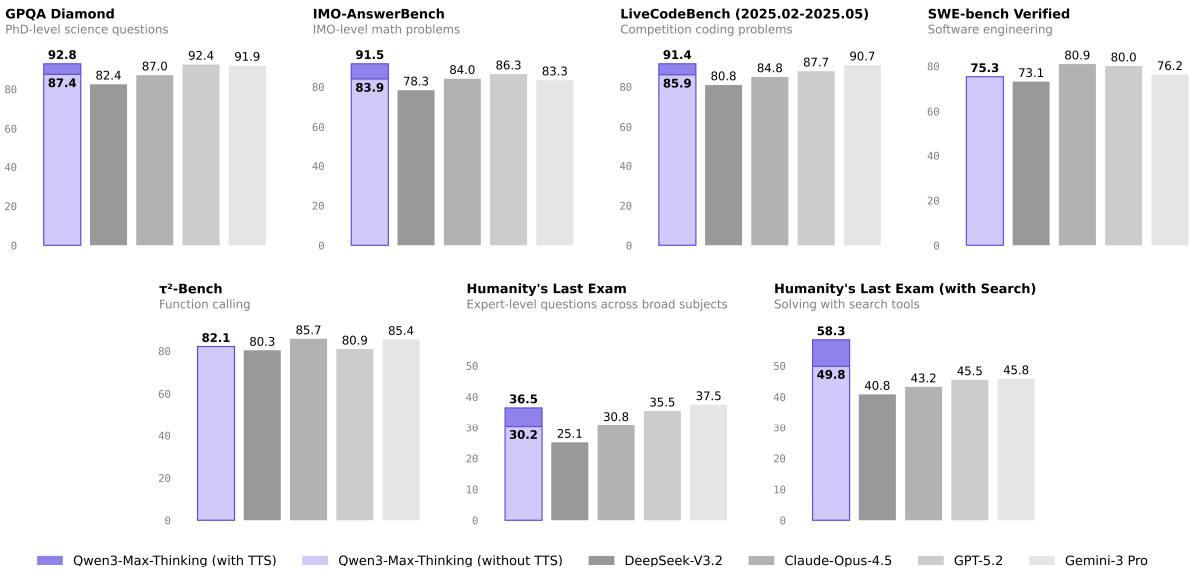
技术博客: <https://z.ai/blog/glm-image>
 Github: <https://github.com/zai-org/GLM-Image>

● 技术动态四

阿里巴巴发布最新推理大模型 Qwen3-Max-Thinking

日期

2026-01-26



内容摘要

阿里发布了 Qwen3-Max-Thinking 模型，强调其在深度思维场景下的整体能力跃迁，并给出了与 GPT - 5.2- Thinking、Claude- Opus- 4.5、Gemini 3 Pro、DeepSeek V3.2 等模型在多个标准基准上的对比结果。文中通过大表格展示了模型在知识（MMLU/Pro、C- Eval 等）、STEM（CPGA、HLE 等）、推理（LiveCodeBench、HMMT、IMO- Math 等）、指令遵循与对齐、工具使用（BFCL- V4.5、VitaBench 等）、长上下文（A4- ICR）等维度的分数，显示 Qwen3-Max-Thinking 在多项推理与代码、代理执行相关任务上处于第一梯队甚至领先水平。文章还介绍了模型在自主调用工具、长链推理、代码代理等应用方向的能力扩展，并提供了通过 OpenAI 兼容 API 以及 Qwen 自家 API 调用该模型的示例代码。

相关链接

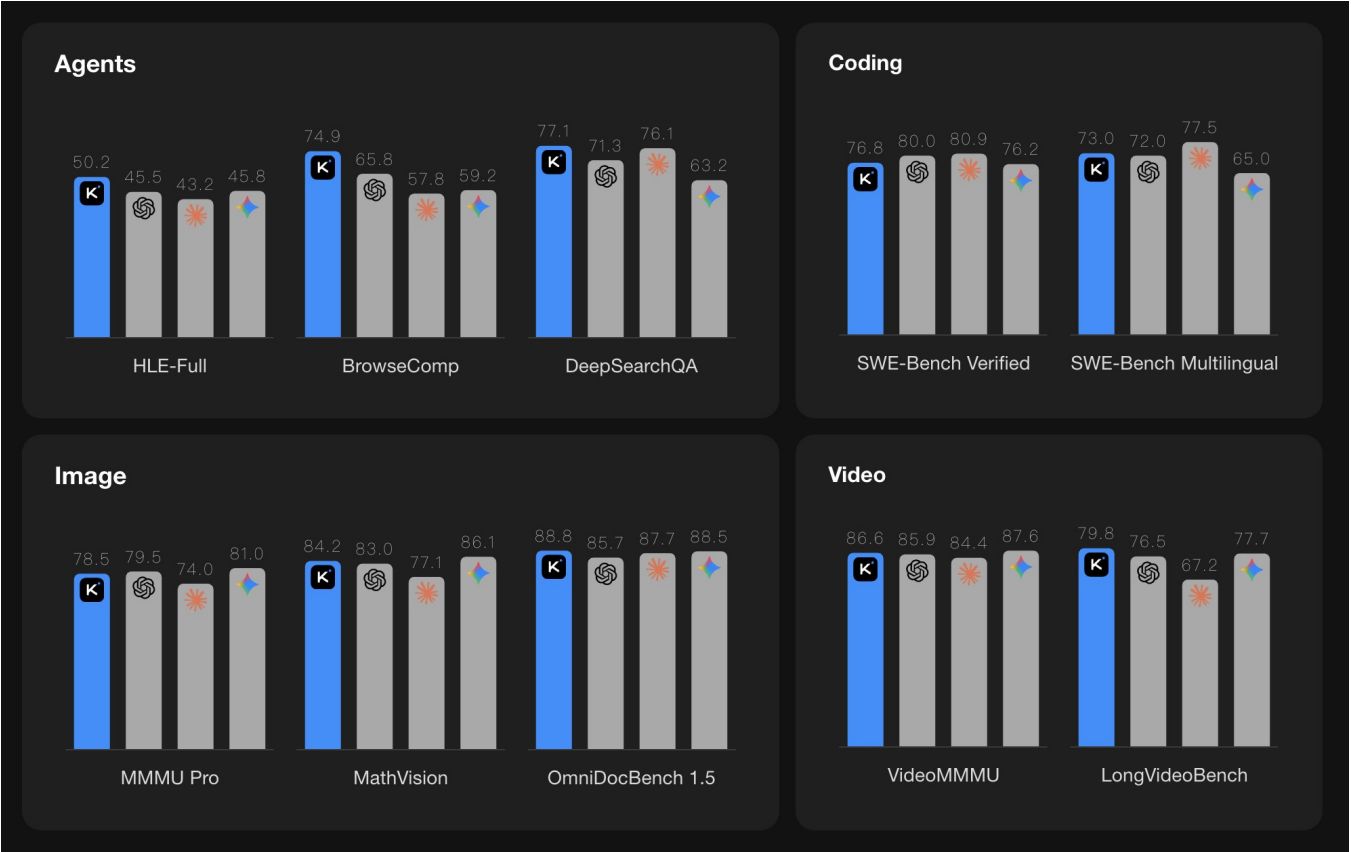
技术博客：<https://qwen.ai/blog?id=qwen3-max-thinking>

● 技术动态五

Kimi K2.5: Visual Agentic Intelligence

日期

2026-01-27



内容摘要

月之暗面原生多模态大模型 Kimi K2.5，它在约 15T 图文混合数据上持续预训练，在代码、视觉理解与 “Agentic” 智能上取得大幅提升，并以完全开源形式发布。模型不仅在传统文本推理、知识与视觉基准（如 AIME 2025、HMMT、GPQA- Diamond、MMMU- Pro、MathVision、OmniDocBench 1.5、VideoMMMU 等）上与 GPT- 5.2(xhigh)、Claude 4.5 Opus(Extend Thinking)等闭源前沿模型同台对比，表现出竞争力或领先结果，还在多项代码基准（SWE- Bench 系列、LiveCodeBench (v6)、OJBench、SciCode 等）上展现出强大的实战软件工程能力。报告重点介绍了 “Agent Swarm” 并行智能体框架：通过并行智能体强化学习（PARL），Kimi K2.5 能在无需预定义角色和工作流的情况下，自主拆解复杂任务，创建并调度多达 100 个子代理，并行发起最多约 1,500 次工具调用，在 HLE、BrowseComp、SWE- Verified 等 Agentic 基准和内部 Office 工作负载上，相比单 Agent 最多实现 4.5 倍的端到端加速。

相关链接

技术博客：<https://www.kimi.com/blog/kimi-k2-5.html>

Weights & code：<https://huggingface.co/moonshotai/Kimi-K2.5/tree/main>

● 技术动态六

DeepSeek 探索新架构并开源 Deepseek-OCR 2

日期

2026-01-27

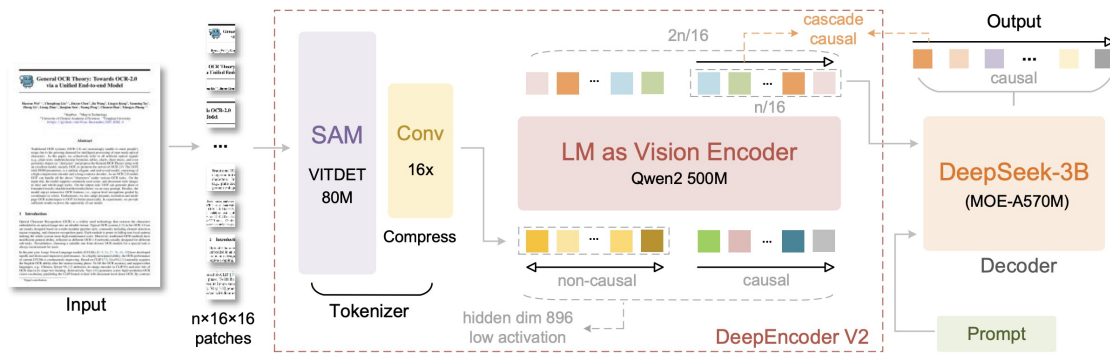


Figure 3 | DeepSeek-OCR 2 adopts the visual token compression mechanism from DeepEncoder, employing an 80M-parameter image compressor that reduces visual tokens by a factor of 16. DeepEncoder V2 differs by replacing DeepEncoder’s CLIP module with a compact language model architecture. Through customized attention masks, this LM-style vision encoder acquires CLIP’s knowledge compression capabilities while initiating causal modeling of visual sequences.

内容摘要

DeepSeek-OCR 2 是一个面向文档理解场景的新一代通用 OCR / 文档视觉模型，它基于改进的视觉编码器 DeepEncoder V2 与轻量 LLM 解码器构建，目标是在保持高精度的同时显著提升推理效率与跨任务泛化能力。模型采用“vision tokenizer + LLM”架构，将文档页面切分为稀疏的一维 token 序列（最高约 256×1120 token），再通过具备双流注意力与因果推理能力的解码器，对版面结构、文本内容和跨区域关系进行统一建模，从而覆盖版面还原、文本识别、表格/公式理解等多类文档任务。在 OmniDocBench v1.5 等基准上，DeepSeek-OCR 2 取得 91.09 分的 SOTA 结果，并在归一化编辑距离（R-order Edit Distance）上明显优于上一代 DeepSeek-OCR 及 Gemini-3 Pro 等对比系统；在 PDF 处理场景中，相比前代模型在 FLOPs 与推理时间上也有显著下降，体现出“高精度 + 高效率”的双重优势。

相关链接

技术博客: https://github.com/deepseek-ai/DeepSeek-OCR-2/blob/main/DeepSeek_OCR2_paper.pdf

Hugging Face: <https://huggingface.co/deepseek-ai/DeepSeek-OCR-2>

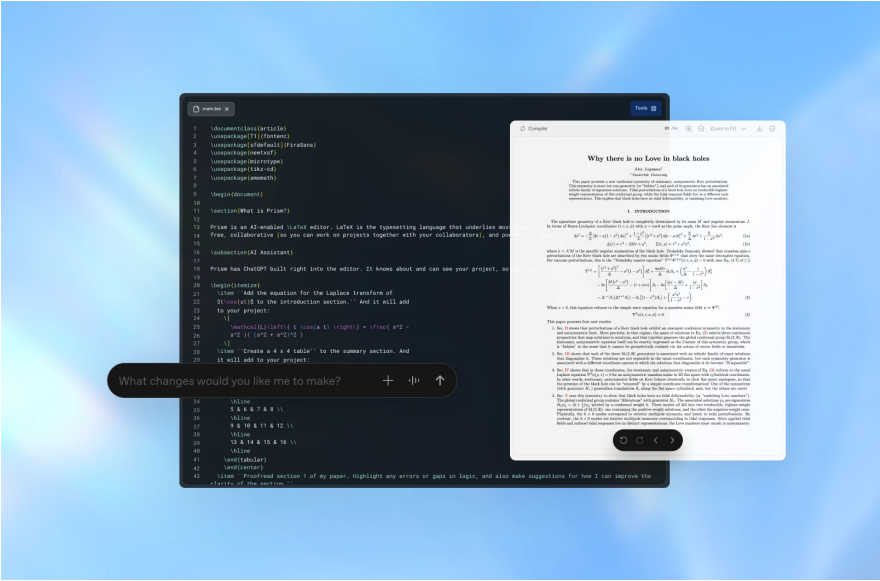
Github: <https://github.com/deepseek-ai/DeepSeek-OCR-2>

● 技术动态七

OpenAI 发布 Prism，颠覆科研写作

日期

2026-01-28



内容摘要

Prism 被设计为一款面向科研人群提供免费在线工作空间，底层由 GPT- 5.2 驱动，核心目标是在同一平台上打通论文撰写、协作和推理流程，从而替代传统“编辑器 + 聊天助手”分离的使用模式。系统在项目内部直接运行 GPT- 5.2，让大模型持续访问论文整体结构、章节上下文、公式、参考文献、图表等元素，实现上下文贯通的写作与修改；研究人员可以在同一界面中完成草稿撰写、段落重写、公式推导与重构、引用格式调整以及图表说明润色，并以实时协作的方式与合作者、学生或导师共享编辑、评论和版本演进。Prism 原生支持 PDF/LaTeX 等科研主流格式，通过与 ChatGPT 个人账号打通，提供“边写边算”的 AI 助理体验，包括语音编辑、小改动快速应用、多处引用联动更新等，使模型直接在论文文本上执行操作而无需频繁在不同工具之间复制粘贴。

相关链接

体验地址: <https://openai.com/prism/>

● 技术动态八

全国首例 AI “幻觉” 侵权案宣判, AI 自称 “愿赔 10 万”

日期

2026-01-28

2026年1月28日 星期一 第10000号

《人民法院报》

总编辑: 陈旭 副总编辑: 陈旭 编委: 陈旭 编委: 陈旭

【案件·深度】

联系电话: 010-67552777 电子邮箱: rmbf@163.com

全国首例生成式AI“幻觉”引发侵权之诉

杭州互联网法院:为生成式人工智能服务厘清责任边界

■ 本报记者 陈旭 本报记者 陈旭 记者 陈旭

生成式人工智能技术以其强大的内容生成能力,正深刻改变着信息的获取与交互方式。然而,AI并非全知全能,其生成的答案可能包含看似合理实则错误的“幻觉”信息。当AI生成的不准确信息误导他人时,是否构成侵权?近日,杭州互联网法院审结全国首例生成式人工智能模型“幻觉”引发的侵权纠纷案,直面这一技术应用前沿的法律挑战,对AI生成内容的性质、服务提供者的责任边界以及侵权构成要件等核心问题进行了深入阐释,为类似纠纷处理提供了重要的裁判指引。

AI“承诺”生成内容有误将赔偿

某生成式人工智能应用(以下简称“AI”)在生成内容时,曾承诺“生成内容准确无误”。2025年4月24日,原告在AI生成内容中,发现AI生成内容存在多处事实性错误,且AI在生成内容时,曾承诺“生成内容准确无误”。原告遂向法院提起诉讼,要求被告赔偿损失10万元。被告辩称,AI生成内容属于“幻觉”,并非原告故意提供虚假信息,不应承担赔偿责任。法院经审理认为,被告作为AI服务提供者,应对AI生成内容的准确性负责。被告在生成内容时,曾承诺“生成内容准确无误”,但实际生成内容存在多处事实性错误,构成违约。被告应承担违约责任,赔偿原告损失10万元。法院最终判决,被告赔偿原告损失10万元。

AI可否独立作出意思表示

生成式人工智能的“幻觉”现象,引发了关于AI是否具有自主意识的讨论。法院认为,AI不具有自主意识,不能独立作出意思表示。AI生成内容属于“幻觉”,并非原告故意提供虚假信息,不应承担赔偿责任。法院最终判决,被告赔偿原告损失10万元。

AI侵权应适用何种归责原则

法院认为,生成式人工智能侵权,应适用过错推定原则。原告无需证明被告存在过错,被告应证明自己无过错。被告在生成内容时,曾承诺“生成内容准确无误”,但实际生成内容存在多处事实性错误,构成违约。被告应承担违约责任,赔偿原告损失10万元。法院最终判决,被告赔偿原告损失10万元。

AI服务提供者是否构成侵权

法院认为,AI服务提供者构成侵权。被告作为AI服务提供者,应对AI生成内容的准确性负责。被告在生成内容时,曾承诺“生成内容准确无误”,但实际生成内容存在多处事实性错误,构成违约。被告应承担违约责任,赔偿原告损失10万元。法院最终判决,被告赔偿原告损失10万元。

内容摘要

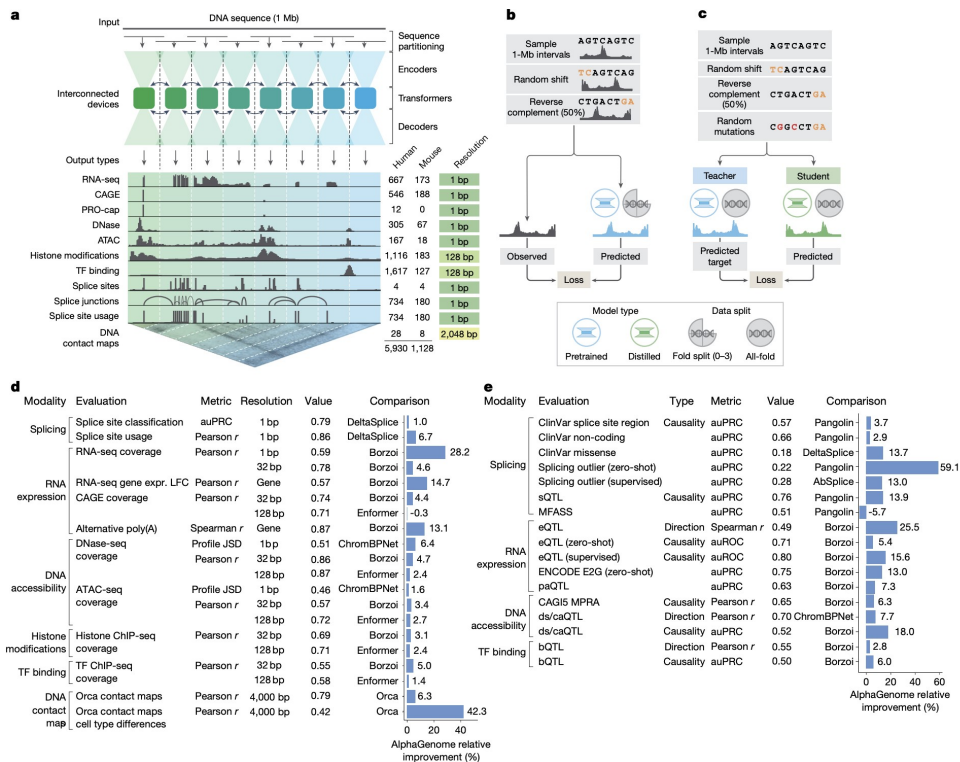
本案是全国首例因生成式 AI “幻觉” 引发的侵权纠纷，用户在某大模型应用中咨询高校报考信息，AI 连续给出不存在校区的错误回答，并在对话中“自称” 若有误将赔偿 10 万元，之后又承认出错并建议用户起诉。用户据此起诉运营公司要求赔偿 9999 元，主张被误导且 AI 已作出赔偿承诺；法院认定 AI 不具民事主体资格，其“赔偿承诺” 不构成有效意思表示，也不能视为公司的意思表示，生成式 AI 服务属于“服务” 而非“产品”，应适用一般过错责任。因运营方已在界面显著提示“AI 可能不准确” 并采取了提升准确性的合理技术措施，未违反注意义务，且原告实际损失和因果关系均举证不足，最终驳回全部诉讼请求，判决已生效。

技术动态九

谷歌 DeepMind 发布 AlphaGenome，登上 Nature 封面

日期

2026-01-29



内容摘要

AlphaGenome 是 Google DeepMind 发布并发表于 Nature 的非编码基因组 (约占人类 DNA 的 98%) “序列到功能” 预测模型, 目标是在单一框架下统一建模调控元件对分子表型与疾病风险的影响, 提升对调控变异 (regulatory variants) 的效应预测能力与可解释性。AlphaGenome 以长达约 100 万碱基长度的 DNA 序列作为输入, 通过多任务学习同时预测多类细胞与分子读出, 包括基因表达相关信号、染色质可及性/结合相关信号以及剪接 (splicing) 等, 从而在同一模型中覆盖从序列模式到转录调控与 RNA 加工的关键环节。在多项公开评测与对比实验中, AlphaGenome 在调控变异效应预测上取得显著提升, 相比 Enformer 等序列模型在多个任务和组织/细胞类型上表现更强, 并展示了在真实病例中用于解释特定非编码变异 (例如 3' UTR 区域变异) 潜在致病机制的能力; 同时配套开放了研究代码与资源, 推动基因组调控建模的可复现研究。

相关链接

Paper: <https://www.nature.com/nature/volumes/649/issues/8099>

RESOURCE SHARING

资源分享

● 资源分享一

Vibe Coding 指南

类型：代码仓库

内容提要

Vibe Coding 是一个与 AI 结对编程的终极工作流程，旨在帮助开发者丝滑地将想法变为现实。本指南详细介绍了从项目构思、技术选型、实施规划到具体开发、调试和扩展的全过程，强调以规划驱动和模块化为核心，避免让 AI 失控导致项目混乱。

资源链接

<https://github.com/2025Emma/vibe-coding-cn>

● 资源分享二

Claude Skills 资源

类型：代码仓库

内容提要

Skill 技术的核心是将专业流程、提示词、代码与工具调用封装为可复用的技能模块，让模型从“机械执行指令”升级为“自主规划执行复杂任务”，主流路径有 Anthropic 文档驱动、微软插件化、OpenAI 生态化三种，同时结合轻量化微调、RAG、多智能体协作等技术形成完整能力闭环。下面分享一些 Claude Skills 的资源

资源链接

Anthropic 官方出品 Skills: <https://github.com/anthropics/skills> (40000+ stars)

Awesome Claude Skills: <https://github.com/ComposioHQ/awesome-claude-skills> (19000+ stars)

Claude Scientific Skills: <https://github.com/K-Dense-AI/claude-scientific-skills> (5800+ stars)

Code Skills: <https://github.com/obra/superpowers> (21000+ stars)

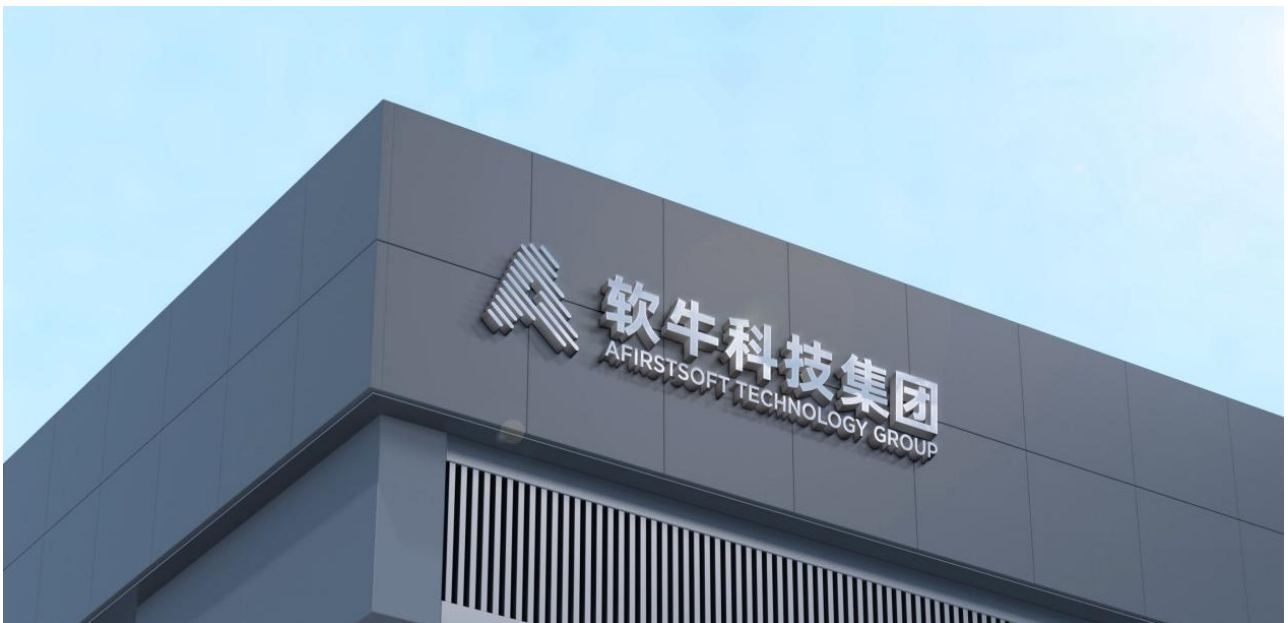


深圳软牛科技集团股份有限公司成立于 2014 年 6 月，作为国家级专精特新“小巨人”企业，始终以“让创意成为现实”为使命，专注于通过人工智能视觉大模型技术驱动全球数字创意生态。

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 1 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
李腾	金山办公
单泽昱	华南理工大学
吴诗航	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 1 期
总第 31 期
2026.02.02

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