

Artificial Intelligence Newsletter

人工智能 前沿快报

2025 年第 06 期
总第 24 期



学会官方公众号

广东省图象图形学会

GDSIG

学术交流、科普宣传、科技成果鉴定

目录

学术动态

一. Core Context Aware Transformers for Long Context Language Modeling.....	1
二. Demonstrating MuJoCo Playground.....	3
三. SurveyForge: On the Outline Heuristics, Memory-Driven Generation, and Multi-dimensional Evaluation for Automated Survey Writing	5
四. Scaling Vision Pre-Training to 4K Resolution.....	7
五. Human-like object concept representations emerge naturally in multimodal large language models.....	9
六. A foundation model for the Earth system.....	11
七. Code Researcher: Deep Research Agent for Large Systems Code and Commit History	13
八. DINO-R1: Incentivizing Reasoning Capability in Vision Foundation Models.....	15
九. A Survey of LLM × DATA.....	17
十. TIIF-Bench: How Does Your T2I Model Follow Your Instructions?.....	19
十一. Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training.....	21
十二. HaploOmni: Unified Single Transformer for Multimodal Video Understanding and Generation.....	23
十三. UniWorld-V1: High-Resolution Semantic Encoders for Unified Visual Understanding and Generation	25
十四. Multi-timescale reinforcement learning in the brain.....	27
十五. OpenThoughts: Data Recipes for Reasoning Models.....	29
十六. Towards LLM-Centric Multimodal Fusion: A Survey on Integration Strategies and Techniques.....	31
十七. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models.....	33
十八. A Survey on Sparse Autoencoders: Interpreting the Internal Mechanisms of Large Language Models	35
十九. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity.....	37
二十. MAGNET: A Multi-agent Framework for Finding Audio-Visual Needles by Reasoning over Multi-Video Haystacks.....	39
二十一. CoCoA-Mix: Confusion-and-Confidence-Aware Mixture Model for Context Optimization.....	41
二十二. Diffuse and Disperse: Image Generation with Representation Regularization.....	43
二十三. Physical restoration of a painting with a digitally constructed mask.....	45
二十四. Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the U.S. Workforce.....	47
二十五. Ming-Omni: A Unified Multimodal Model for Perception and Generation.....	49
二十六. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning.....	51
二十七. Self-Adapting Language Models.....	53
二十八. A Survey of Generative Categories and Techniques in Multimodal Large Language Models.....	55
二十九. A Comprehensive Survey of Deep Research: Systems, Methodologies, and Applications.....	57
三十. MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention.....	59

目录

三十一. Essential-Web v1.0: 24T tokens of organized web data	61
三十二. From Bytes to Ideas: Language Modeling with Autoregressive U-Nets	63
三十三. Revisiting Reinforcement Learning for LLM Reasoning from A Cross-Domain Perspective	65
三十四. Show-o2: Improved Native Unified Multimodal Models	67
三十五. Unsupervised pretraining in biological neural networks	69
三十六. How much do language models memorize?	71
三十七. Drag-and-Drop LLMs: Zero-Shot Prompt-to-Weights	73
三十八. Machine Mental Imagery: Empower Multimodal Reasoning with Latent Visual Tokens	75
三十九. Virtual Community: An Open World for Humans, Robots, and Society	77
四十. YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception	79
四十一. Deep Research Agents: A Systematic Examination And Roadmap	81
四十二. Whole-Body Conditioned Egocentric Video Prediction	83

技术动态

一. 3D 生成模型 Direct3D-S2 全面开源! 8 块 GPU 超越闭源登顶 HF	85
二. OpenAI 深夜宣布 ChatGPT 支持 MCP、会议记录!	86
三. MiniCPM4:在典型端侧芯片上能够实现 5 倍以上生成加速!	87
四. OpenAI 创始人 Sam Altman 发表博文, 阐述人类将进入一个以数字超级智能为标志的“温和的奇点”时代	88
五. AMD 发布 3nm 怪兽 MI355X, 性能碾压英伟达 B200	89
六. 开源 OCR 工具, 可用于将文档 (PDF 和图像) 转换为可编辑 Markdown 文件	90
七. Google Gemini 2.5 技术报告	91
八. 模型即 Agent, Kimi-Researcher (深度研究) 开启内测	92
九. 谷歌 AlphaGenome 横空出世! 40 亿年生命代码一键破解	93
十. 谷歌开源 Gemma 3n: 2G 内存就能跑, 100 亿参数内最强多模态模型	94
十一. 阿里上新多模态模型 Qwen-VLo	95
十二. 文心 4.5 系列模型, 正式开源!	96

资源分享

一. 何恺明 CVPR 最新讲座 PPT 上线: 走向端到端生成建模	97
二. 微软开源图标库	98
三. Awesome LLM Apps	99

鸣谢支持

一. 深圳软牛科技集团股份有限公司	100
-------------------	-----

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 42 篇最新学术动态、12 篇技术动态、以及 3 个资源推荐给广大读者。

ACADEMIC TRENDS

学术动态

● 论文一

Core Context Aware Transformers for Long Context Language Modeling

日期

2024-12-17 (ICML 录用论文)

Core Context Aware Transformers for Long Context Language Modeling

Yaofu Chen^{*1} Zeng You^{*12} Shuhai Zhang^{*13} Haokun Li¹² Yiru Li¹ Yaowei Wang²⁴ Mingkui Tan¹³⁵

内容摘要

华南理工大学与琶洲实验室联合提出了一种新型关键上下文感知注意力机制（CCA-Attention），专为应对超长文本的高效建模挑战。在长达 128K 的上下文任务中，CCA-Attention 的推理速度是标准自注意力机制的 7.9 倍，同时键值缓存（KV Cache）显存占用减少 93%，全面优于现有主流高效注意力方案。该工作已被 ICML 2025 正式接收，最早于 2024 年 12 月 17 日发布在 ArXiv，早于 DeepSeek NSA 与 Kimi MoBA 等方案。CCA-Attention 不仅资源占用更优，在长距离依赖建模上的准确性和效率也实现双重突破，为长文本场景中的大模型推理带来新的范式。

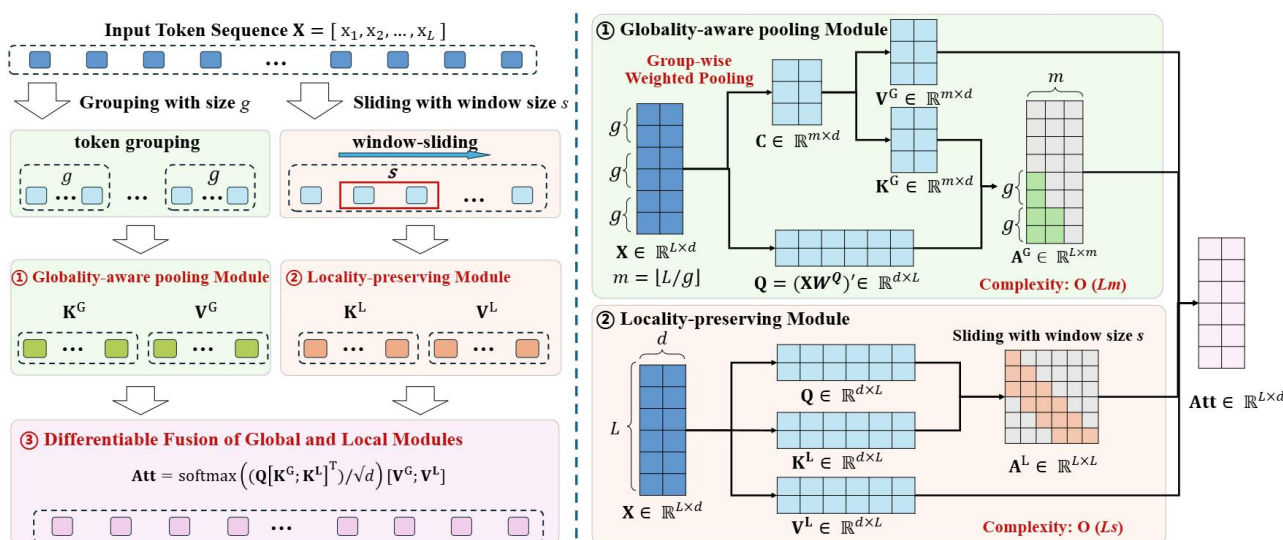


Figure 2. Illustration of CCA-Attention, which includes two components: 1) Globally-aware pooling module encapsulates the input tokens X into core tokens C according to the importance (Eqn. (2)). The core tokens C serve as representative proxies of X for attention, thereby reducing computational costs. 2) Locality-preserving module incorporates the local context from neighboring tokens, acting as supplement for the globally-aware pooling module. We produce the final output **Att** by fusing these two modules based on Eqn. (5).

主要亮点

- **高效的长文本建模机制：**结合全局池化注意力与局部保留注意力的协同设计，显著压缩计算开销，保持关键信息提取能力。
- **轻松集成现有模型：**该机制可集成到现有大模型架构，无需重训练，仅通过轻量级微调即可提升性能。
- **权衡性能与效率的新范式：**通过引入核心 token 聚焦于关键上下文，大幅提高计算效率。

相关链接

Paper: <https://arxiv.org/pdf/2412.12465>

Github: <https://github.com/chenyaofo/CCA-Attention>

● 论文二

Demonstrating MuJoCo Playground

日期

2025-02-25

MuJoCo Playground

Kevin Zakka^{*,1}, Baruch Tabanpour^{*,2}, Qiayuan Liao^{*,1}, Mustafa Haiderbhai^{*,3}, Samuel Holt^{*,4},
Jing Yuan Luo², Arthur Allshire¹, Erik Frey², Koushil Sreenath¹, Lueder A. Kahrs³,
Carmelo Sferrazza^{†,1}, Yuval Tassa^{†,2}, and Pieter Abbeel^{†,1}

¹UC Berkeley

²Google DeepMind

³University of Toronto

⁴University of Cambridge

^{*}Equal contributions

[†]Equal advising

内容提要

本文介绍了完全开源的机器人学习框架 MuJoCo Playground，其基于 MJX 构建，旨在简化模拟、训练及模拟到真实 (sim-to-real) 的迁移流程。通过集成物理引擎、批量渲染器和训练环境，用户仅需 “pip install playground” 即可在单 GPU 上快速训练策略，支持四足机器人、人形机器人、灵巧手和机械臂等多种平台的零样本 sim-to-real 迁移。该框架在 MS COCO 等基准上验证了性能，通过超图增强自适应关联机制 (HyperACE) 和全流程聚合 - 分布 (FullPAD) 范式，实现了高效的全局特征融合与细粒度信息流协同，同时利用深度可分离卷积减少参数复杂度，代码和模型已开源。

● 论文三

SurveyForge: On the Outline Heuristics, Memory-Driven Generation, and Multi-dimensional Evaluation for Automated Survey Writing

日期

2025-03-06

SURVEYFORGE: On the Outline Heuristics, Memory-Driven Generation, and Multi-dimensional Evaluation for Automated Survey Writing

Xiangchao Yan^{♣,*} Shiyang Feng^{♣,♠,*} Jiakang Yuan^{♣,♠} Renqiu Xia^{♣,◇}

Bin Wang[♠] Bo Zhang^{♣,†} Lei Bai^{♣,†}

[♣]Shanghai Artificial Intelligence Laboratory [♠]Fudan University [◇]Shanghai Jiao Tong University
{zhangbo,bailei}@pjlab.org.cn

内容提要

调查论文在科研中至关重要，但当前 LLM 生成的调查论文与人类撰写的在大纲质量和引用准确性上存在显著差距。为此，研究提出 SurveyForge 框架，其先通过分析人类大纲逻辑结构并参考检索到的领域相关文献生成大纲，再利用学者导航代理从记忆中检索的高质量论文，自动生成并优化文章内容。此外，为实现全面评估，构建了包含 100 篇人类撰写调查论文的 SurveyBench，从参考文献、大纲和内容质量三个维度评估 AI 生成的调查论文。实验表明，SurveyForge 的表现优于 AutoSurvey 等先前工作。

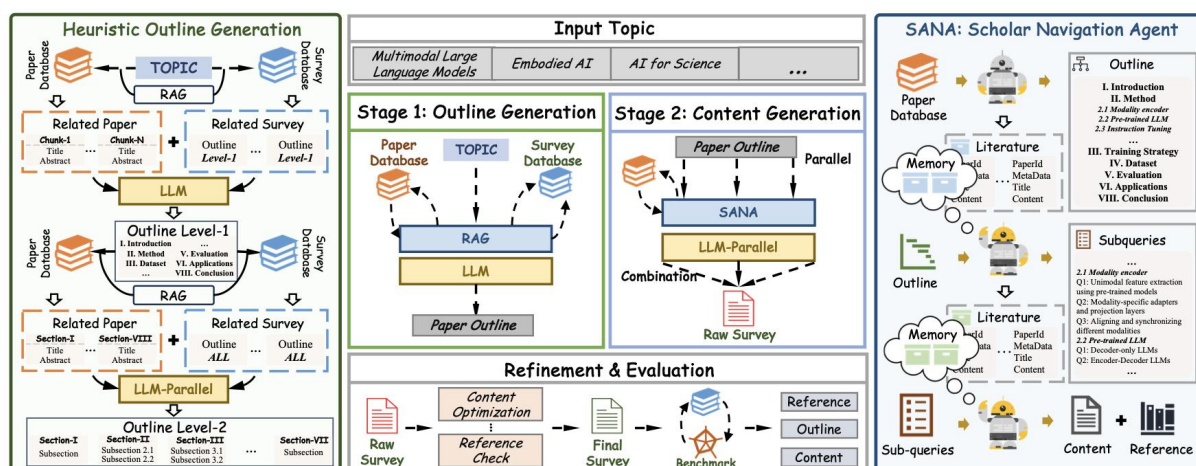


Figure 2: The overview of SURVEYFORGE. The framework consists of two main stages: Outline Generation and Content Writing. In the Outline Generation stage, SURVEYFORGE utilizes heuristic learning to generate well-structured outlines by leveraging topic-relevant literature and structural patterns from existing surveys. In the Content Writing stage, a memory-driven Scholar Navigation Agent (SANA) retrieves high-quality literature for each subsection and LLM generates the content of each subsection. Finally, the content is synthesized and refined into a coherent and comprehensive survey.

主要亮点

- **基于逻辑结构的大纲生成：**通过分析人类撰写大纲的逻辑结构并结合领域文献检索，生成更符合学术规范的调查论文大纲，提升了 AI 生成大纲的合理性和严谨性。
- **记忆驱动的内容生成与优化：**借助学者导航代理从记忆中检索高质量论文，为内容生成提供丰富依据，并在此基础上进行自动优化，提高了内容的准确性和深度。
- **多维度评估体系的构建：**创建包含 100 篇人类论文的 SurveyBench，从参考文献、大纲和内容质量三个维度对 AI 生成的调查论文进行评估，为该领域提供了更全面的评价标准。

相关链接

Paper: <https://arxiv.org/abs/2503.04629>

Github: <https://github.com/Alpha-Innovator/SurveyForge?tab=readme-ov-file>

● 论文四

Scaling Vision Pre-Training to 4K Resolution

日期

2025-03-26 (CVPR 2025 录用论文)



2025-3-26

Scaling Vision Pre-Training to 4K Resolution

Baifeng Shi^{1,2*} **Boyi Li^{1,2}** **Han Cai²** **Yao Lu²** **Sifei Liu²** **Marco Pavone²**
Jan Kautz² **Song Han²** **Trevor Darrell¹** **Pavlo Molchanov²** **Hongxu Yin²**

¹UC Berkeley ²NVIDIA

内容提要

本文提出 PS3 (Pre-training with Scale- Selective Scaling) 方法, 用于将 CLIP 风格视觉预训练扩展至 4K 分辨率, 克服处理高分辨率图像的计算瓶颈。通过局部区域与对应描述对比学习分解全局对比, PS3 可近乎恒定成本预训练 4K 图像编码器, 并基于此训练出的 VILA- HD MLLM 在解析高分辨率视觉任务上效果显著优于低分辨率预训练模型。

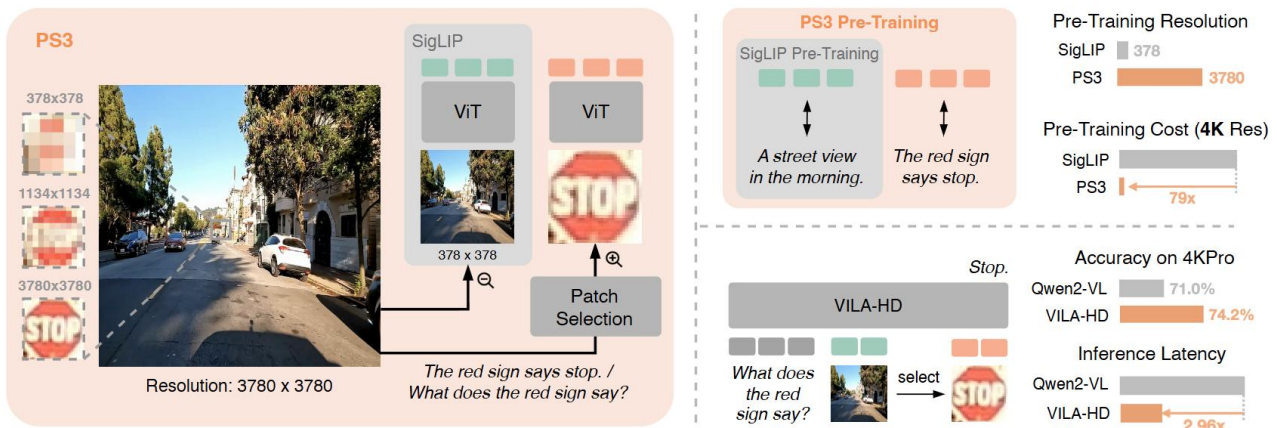


Figure 1: **Left:** Regular vision models such as SigLIP processes images at a low resolution (e.g., 378×378 pixels), which is not enough for many daily tasks such as spotting the stop sign while driving. In contrast, PS3 is able to both encode low-res features and efficiently process high-res information of 4K-resolution images via top-down patch selection, i.e., selectively processing relevant patches based on any text prompt. **Top Right:** SigLIP is pre-trained by contrasting global vision features and global captions, which is costly for high-resolution images. PS3 is pre-trained with additional contrast between local high-res features with local captions, enabling pre-training at 4K resolution with 79 $\times$ less cost than SigLIP. **Bottom Right:** VILA-HD with PS3 as the vision encoder is able to select high-res regions to process based on the user prompt. VILA-HD outperforms state-of-the-art MLLMs such as Qwen2-VL [93] on the proposed 4KPro benchmark while achieving 2.96 $\times$ speedup.

主要亮点

- **Scale- Selective 局部对比策略:** 在高分辨率输入下仅对关键信息区域进行精细对比，显著减少计算成本（降低 79%）。
- **4K 分辨率预训练可行:** 成功实现高分辨率视觉预训练，性能远超传统低分辨率模型。
- **VILA- HD 加成优势:** 基于 PS3 的视觉编码器在多模态理解与生成任务中表现突出，支持高分辨率场景。

相关链接

Paper: <https://arxiv.org/abs/2503.19903>

Github: <https://nvlabs.github.io/PS3/>

● 论文五

Human-like object concept representations emerge naturally in multimodal large language models

日期

2025-05-02 (Nature Machine Intelligence 录用论文)

Human-like object concept representations emerge naturally in multimodal large language models

Changde Du^{1,2}, Kaicheng Fu^{1,2}, Bincheng Wen³, Yi Sun^{1,2}, Jie Peng^{1,2}, Wei Wei¹, Ying Gao¹, Shengpei Wang¹, Chuncheng Zhang¹, Jinpeng Li⁴, Shuang Qiu¹, Le Chang³, and Huiguang He^{1,2,5,*}

¹State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing, China

²School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

³Institute of Neuroscience, State Key Laboratory of Brain Cognition and Brain-Inspired Intelligence Technology, CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China

⁴School of Automation Science and Engineering, South China University of Technology, Guangzhou, China

⁵Zhongguancun Academy, Beijing, China

*corresponding author: Huiguang He (huiguang.he@ia.ac.cn)

内容提要

本研究探讨了大语言模型 (LLMs) 和多模态大语言模型 (MLLMs) 是否能从语言和多模态数据中发展出类似人类的物体概念表征。通过收集 LLMs/MLLMs 对 1854 个自然物体的 470 万三元组奇异物判断, 利用稀疏正相似性嵌入 (SPoSE) 方法生成 66 维低维嵌入, 发现这些嵌入具有稳定性和预测性, 能捕捉语义聚类 (如动物、食物等类别), 与人类心理表征结构相似。进一步分析表明, 模型嵌入与人类大脑区域 (如 EBA、PPA、FFA) 的神经活动模式高度对齐, 证明 LLMs/MLLMs 的物体表征虽非完全一致但共享人类概念知识的核心特征。该发现为理解机器智能和开发类人认知系统提供了依据。

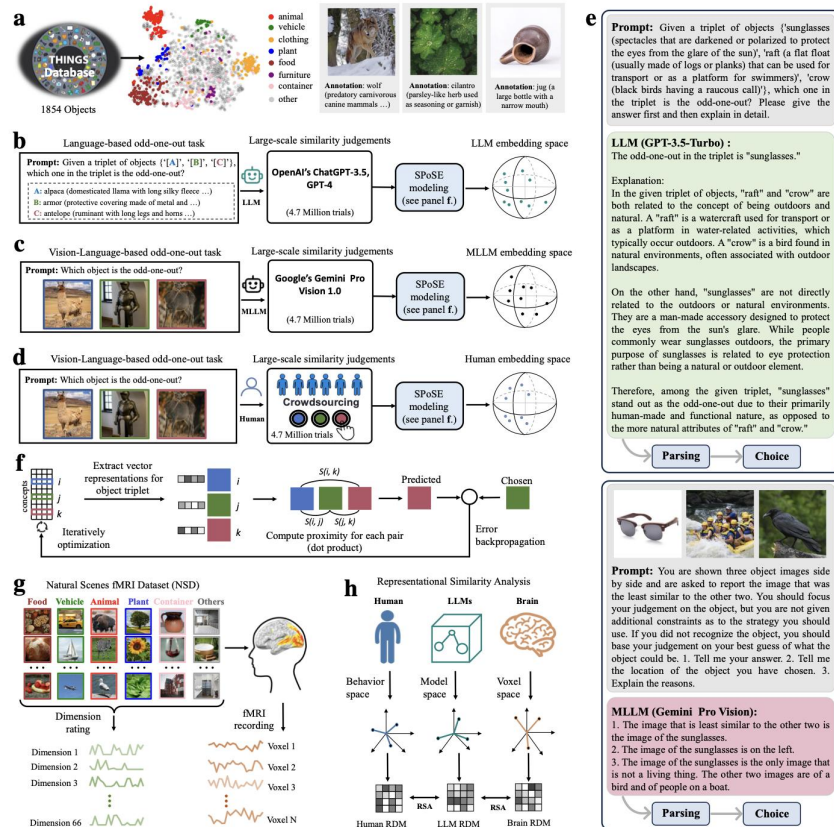


Fig. 1. Schematic diagrams of the experiment and analysis methods. **a**, THINGS database and examples of object image with their language descriptions at the bottom. **b-d**, Pipelines of mental embedding learning under the triplet odd-one-out paradigm for LLM, MLLM, and humans, respectively. Odd-one-out judgments were collected for approximately 4.7 million triplets, and modeled using the SPoSE approach to derive the corresponding low-dimensional embedding. **e**, Examples of prompts and responses for LLM and MLLM. **f**, Illustration of the SPoSE modeling approach. **g**, Illustration of the NSD dataset with dimension ratings for stimulus images. The schematic structure incorporates elements adapted from Figure 1A of Horikawa et al. (2020)³⁴ (<https://doi.org/10.1016/j.jsci.2020.101060>), published under a CC BY 4.0 license. **h**, Overview of the comparisons between space of LLMs, human behavior and brain activity. For this figure, all images were replaced by images with similar appearance from the public domain. Images used under a CC0 license, from Pixabay and Pexels.

主要亮点

- **类人概念表征的自发涌现**: LLMs/MLLMs 通过处理自然物体数据, 自发形成可解释的低维语义维度 (如 “动物相关” “食物相关”), 其表征结构与人类心理表征的聚类模式显著相似, 无需显式任务训练即可捕捉物体概念的核心维度。
- **多模态模型与大脑神经活动的关联性**: MLLM 嵌入与人类大脑中物体选择性区域 (如身体选择性区域 EBA、场景选择性区域 PPA) 的神经活动模式强相关, 尤其在视觉语义处理中表现出与人类认知更接近的表征几何, 优于纯语言模型。

- **大规模行为数据驱动的特征学习框架**：通过 470 万三元组判断构建的 SPoSE 方法，无需访问模型内部特征，仅通过行为数据即可推断低维表征，为探测闭源 LLMs 的心理表征提供了可推广的方法论，且验证了表征维度的可解释性和跨模型稳定性。

相关链接

Paper: <https://arxiv.org/pdf/2407.01067?>

● 论文六

A foundation model for the Earth system

日期

2025-05-21（Nature 录用论文）

Article

A foundation model for the Earth system

<https://doi.org/10.1038/s41586-025-09005-y>

Received: 28 May 2024

Accepted: 9 April 2025

Published online: 21 May 2025

Cristian Bodnar^{1,2,11}, Wessel P. Bruinsma^{1,11}, Ana Lucic^{1,3,11}, Megan Stanley^{1,11}, Anna Allen^{4,11}, Johannes Brandstetter^{1,8}, Patrick Garvan¹, Maik Riechert¹, Jonathan A. Weyn⁶, Haiyu Dong⁶, Jayesh K. Gupta^{2,7}, Kit Thambiratnam⁶, Alexander T. Archibald⁴, Chun-Chieh Wu⁸, Elizabeth Heider¹, Max Welling^{1,3}, Richard E. Turner^{1,4,9} & Paris Perdikaris^{1,10}✉

内容提要

微软牵头团队发布“Aurora”——首款地球系统基础模型，能够跨多变量、多空间尺度进行高分辨率天气预报与气候模拟。模型在全球 92% 目标指标上超越传统数值模型，尤其在极端天气事件预测上准确率提升显著。Aurora 利用深度学习替代某些物理模块，且具备可控分辨率、多任务预测能力，是 AI 与地球科学融合的重大进展。

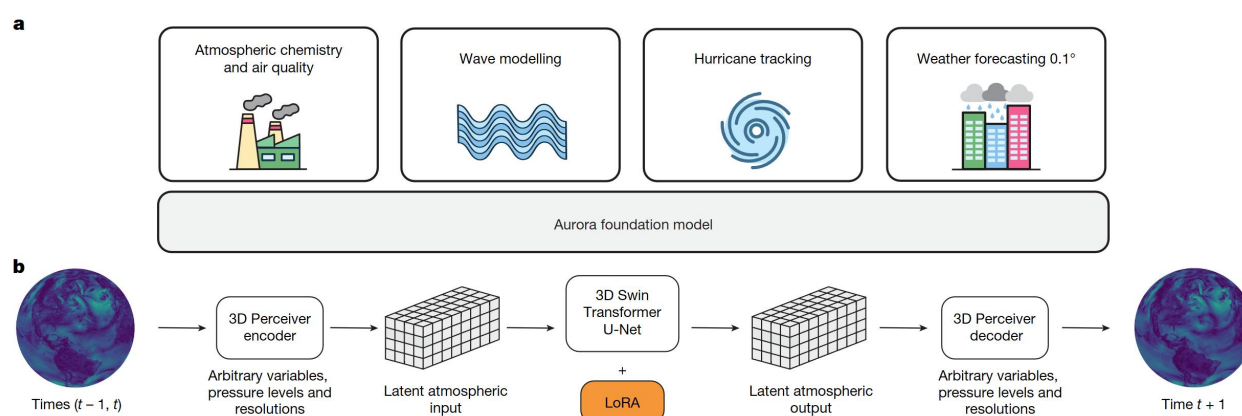


Fig. 1 | Aurora is a 1.3-billion-parameter foundation model for the Earth system. Icons are for illustrative purposes only. **a**, Aurora is pretrained on several heterogeneous datasets with different resolutions, variables and pressure levels. The model is then fine-tuned for several operational forecasting scenarios at different resolutions: atmospheric chemistry and air

quality at 0.4°, wave modelling at 0.25°, hurricane tracking at 0.25° and weather forecasting at 0.1°. **b**, Aurora is a flexible 3D Swin Transformer¹⁹ with 3D Perceiver-based²¹ atmospheric encoders and decoders. The model is able to ingest inputs with different spatial resolutions, numbers of pressure levels and variables.

主要亮点

- **多变量 × 任意分辨率预测能力**：支持在自定义空间分辨率下对温度、降雨、风速等变量进行联合预测。
- **极端天气预报表现卓越**：在极端事件预测中优于数值天气预报模型，被誉为“救命先觉”。
- **开创 AI 替代数值模拟路径**：展示深度学习可替代传统物理数值模型，并广泛应用于全球气候模拟领域。

相关链接

Paper: <https://www.nature.com/articles/s41586-025-09005-y>

GitHub: <https://github.com/microsoft/aurora>

● 论文七

Code Researcher: Deep Research Agent for Large Systems Code and Commit History

日期

2025-05-27

Code Researcher: Deep Research Agent for Large Systems Code and Commit History

Ramneet Singh* Sathvik Joel* Abhav Mehrotra Nalin Wadhwa
Ramakrishna B Bairi Aditya Kanade Nagarajan Natarajan
Microsoft Research
{ramneet2001,ksjoe30,abhavm1,nalin.wadhwa02}@gmail.com
{ram.bairi,kanadeaditya,nagarajan.natarajan}@microsoft.com

内容提要

基于大语言模型（LLM）的编码代理在编码基准测试中表现出良好效果，但其在系统代码上的有效性仍有待探索。由于系统代码的规模和复杂性，即使对人类来说，对系统代码库进行修改也是一项艰巨的任务，这需要在修改前从大型代码库及其大量提交历史中研究许多上下文片段。受近期深度研究代理进展的启发，本文设计了首个用于代码的深度研究代理 Code Researcher，并将其应用于生成补丁以缓解系统代码中报告的崩溃问题。Code Researcher 对代码的语义、模式和提交历史进行多步推理，以收集足够的上下文，并将上下文存储在结构化内存中用于合成补丁。在 kBenchSyz（一个 Linux 内核崩溃基准）上的评估显示，Code Researcher 显著优于强基线，实现了 58% 的崩溃解决率，而 SWE-agent 仅为 37.5%。平均而言，Code Researcher 在每条轨迹中探索 10 个文件，而 SWE-agent 仅探索 1.33 个文件，这凸显了 Code Researcher 深度探索代码库的能力。通过在一个开源多媒体软件上的另一项实验，证明了 Code Researcher 的可推广性。实验强调了为大型代码库收集全局上下文和进行多方面推理的重要性。

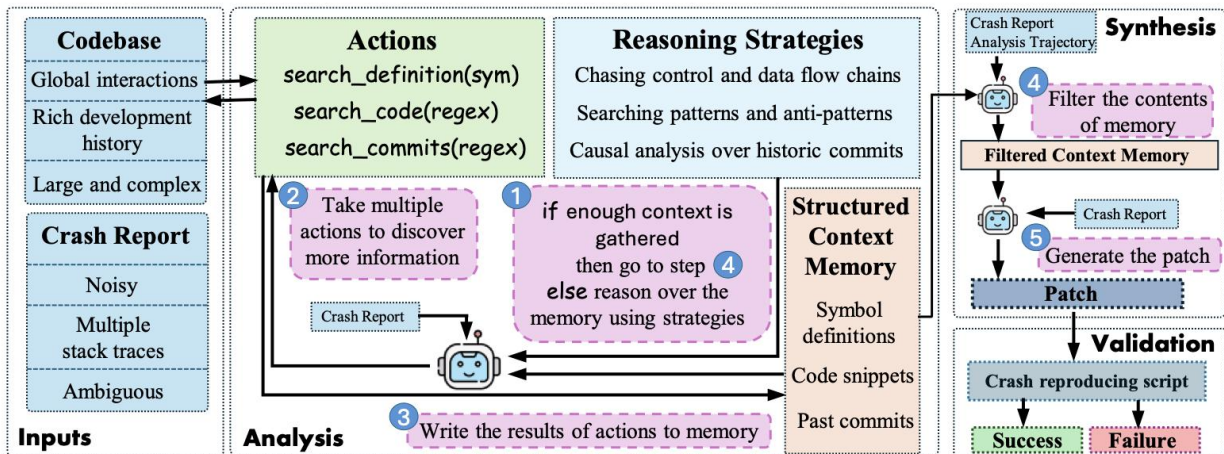


Figure 1: Code Researcher conducts deep research over code in three phases: (1) Starting with the codebase and crash report as input, the Analysis phase performs multi-step reasoning about semantics, patterns, and commit history of code. It gathers context in a structured memory. (2) The Synthesis phase filters the contents of the memory to keep relevant context and generates a patch. (3) The Validation phase uses external tools to validate the patch.

主要亮点

- **首个系统代码深度研究代理:** 设计了首个针对系统代码的深度研究代理 Code Researcher, 填补了 LLM 编码代理在系统代码领域的研究空白, 为复杂系统代码的修改提供了新的解决方案。
- **多步推理与结构化内存结合:** 通过对代码语义、模式及提交历史的多步推理收集上下文, 并利用结构化内存存储这些信息, 使代理能够更全面地理解代码库, 从而生成更有效的补丁。
- **显著优于基线的性能表现:** 在 Linux 内核崩溃基准测试中, Code Researcher 的崩溃解决率远高于 SWE-agent, 且能深度探索更多文件, 展现出在大型代码库中进行全局上下文收集和多维度推理的优势及良好的可推广性。

相关链接

Paper: <https://arxiv.org/abs/2506.11060>

● 论文八

DINO-R1: Incentivizing Reasoning Capability in Vision Foundation Models

日期

2025-05-29

DINO-R1: Incentivizing Reasoning Capability in Vision Foundation Models

Chenbin Pan^{1,2} Wenbin He^{1,2} Zhengzhong Tu³ Liu Ren^{1,2}

¹Bosch Research North America

²Bosch Center for Artificial Intelligence (BCAI)

³Texas A&M University

{chenbin.pan@us.bosch.com}

<https://Christinepan881.github.io/DINO-R1>

内容提要

DINO- R1 是首个将强化学习 (RL) 思想引入视觉基础模型 (如 DINO 系列) 的尝试, 目的是提升视觉模型的“视觉推理”能力。论文首次提出 Group Relative Query Optimization (GRQO), 一种基于查询的训练策略, 为每个查询分配基于“组内归一化对齐质量”的奖励, 并辅以 KL 正则化稳定目标分布。结合视觉 prompt 编码器与视觉引导查询选择机制, DINO- R1 系列在 COCO、LVIS 和 ODinW 数据集上的表现显著超越监督微调 baselines, 具备更稳定训练和更优泛化性。

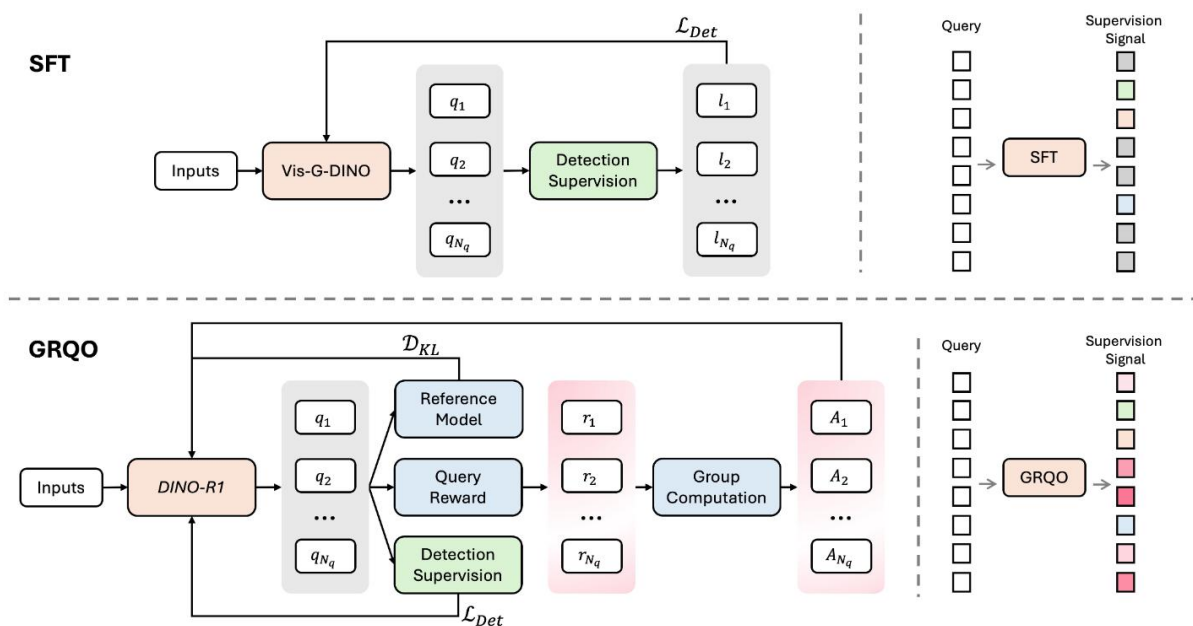


Figure 1: **SFT vs. GRQO**. SFT leads to limited and homogeneous supervision signals, while GRQO produces richer and more diverse learning signals, encouraging queries to be more expressive.

主要亮点

- **首次将 RL 用于视觉推理**: GRQO 提供每查询的细化奖励, 对标语言模型领域的 GRPO, 但适配视觉特征结构, 打破监督微调局限。
- **KL 正则保障训练稳定**: 通过对象性分布的 KL 正则, 抑制训练发散、防止过拟合和分布漂移, 保持 query 预测多样性与稳定性。
- **在 Visual Prompting 场景中泛化更强**: 在 open- vocabulary 和 closed- set visual prompting 场景下, DINO- R1 相比传统监督策略在 COCO、LVIS、ODinW 上均有显著提升, 表明其 query-level 表征更具推理能力。

相关链接

Paper: <https://arxiv.org/abs/2505.24025>

GitHub: <https://christinepan881.github.io/DINO-R1/>

● 论文九

A Survey of LLM × DATA

日期

2025-06-01

A Survey of LLM × DATA

Xuanhe Zhou^{*¶}, Junxuan He^{*¶}, Wei Zhou^{*¶}, Haodong Chen^{*¶}, Zirui Tang^{*¶}, Haoyu Zhao^{*¶}, Xin Tong^{*},
Guoliang Li[†], Youmin Chen^{*}, Jun Zhou^{*}, Zhaojun Sun^{*}, Binyuan Hui[‡], Shuo Wang[‡], Conghui He[§],
Zhiyuan Liu[†], Jingren Zhou[‡], Fan Wu^{*}

^{*}Shanghai Jiao Tong University [†]Tsinghua University [‡]Alibaba Group [§]Shanghai AI Laboratory
<https://github.com/weAIDB/awesome-data-llm>

内容提要

这篇论文对“LLM 与数据管理”之间的双向关系进行了全面综述。在 DATA 4 LLM 方向，研究着重探讨如何支撑 LLM 生命周期的各阶段，从数据获取、处理、存储到服务，为预训练、微调、RAG 和 agent 应用提供优质数据保障。在 LLM 4 DATA 方向，则关注 LLM 如何被用作通用数据管理工具，支持数据清洗、整合、分析、系统调优和异常检测等任务。论文构建了一个内容架构清晰、方法体系完整的框架，并整合了 400+ 篇前沿文献。

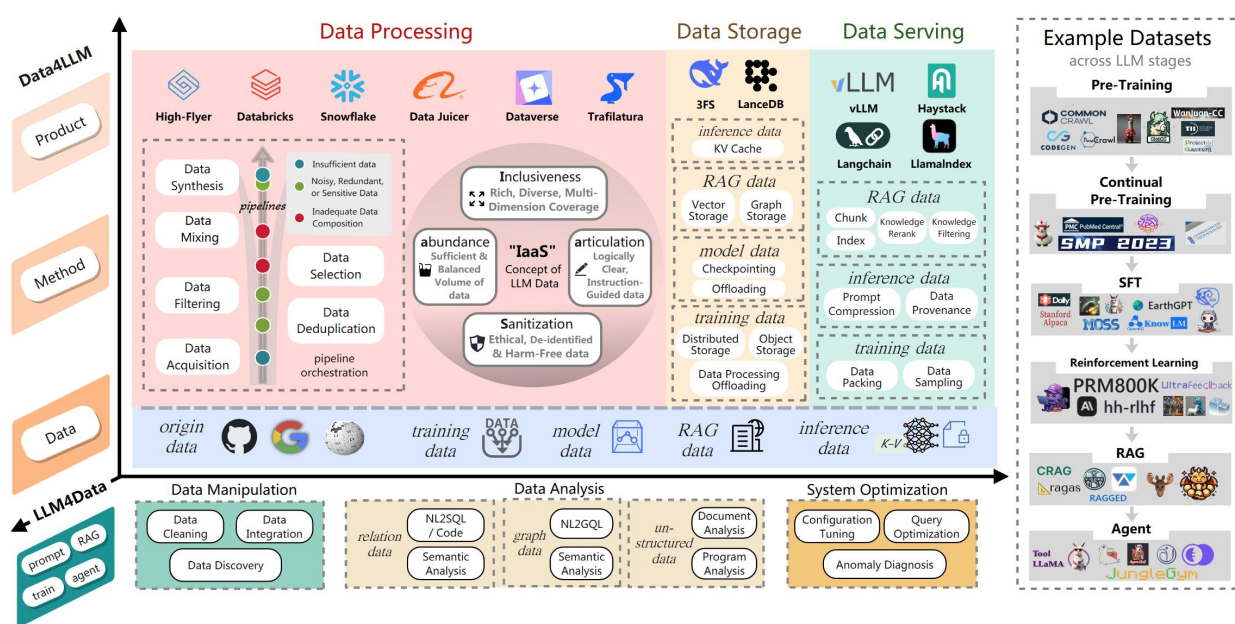


Fig. 1: Overview of LLM x DATA (with "IaaS" Concept).

主要亮点

- **DATA 4 LLM 全生命周期框架 (IaaS 概念)：**引入“数据作为基础设施” (IaaS) 理念，从数据处理（采集、去重、过滤、增强）、存储（高效格式、分布式管理）到服务（RAG 后处理、prompt 压缩、数据溯源），覆盖 LLM 应用各环节。
- **LLM 4 DATA 实用场景广泛：**系统回顾利用 LLM 自动化进行数据清理、结构化/非结构化数据分析、SQL 重写、系统调优和异常诊断等应用，展现 LLM 在数据工程中的潜力。
- **大规模文献整理 + 开源生态支撑：**涵盖 400+ 文献，梳理当前方法优劣，并附带 GitHub 资源仓库，为学术与工程领域提供参考。

相关链接

Paper: <https://arxiv.org/abs/2505.18458>

GitHub: <https://github.com/weAIDB/awesome-data-llm>

● 论文十

TIIF-Bench: How Does Your T2I Model Follow Your Instructions?

日期

2025-06-02



TIIF-Bench: How Does Your T2I Model Follow Your Instructions?

Xinyu Wei^{1,4}, Jinrui Zhang^{1,4}, Zeqing Wang^{2,4}, Hongyang Wei^{3,4}, Zhen Guo^{1,4}, Lei Zhang^{1,4}

¹Hong Kong Polytechnic University ²Sun Yat-sen University

³Tsinghua University ⁴OPPO Research Institute

allen_wei@stu.pku.edu.cn, cslzhang@comp.polyu.edu.hk

内容提要

TIIF-Bench 是用于评估文本生成图像 (T2I) 模型是否精确执行指令的新基准, 共包含 5000 条多难度、多语义维度提示 (包括简/长版本), 以及专门评估“文字渲染”和“风格控制”的属性。作者还额外收录 100 条设计师级 prompt, 并引入自动化评估机制, 借助 VLM (如 GPT-4o) 判别生成结果微小差别。大量实验展示当前主流模型在基础属性控制表现尚可, 但在关系布局、逻辑推理、文字渲染方面仍有明显不足。

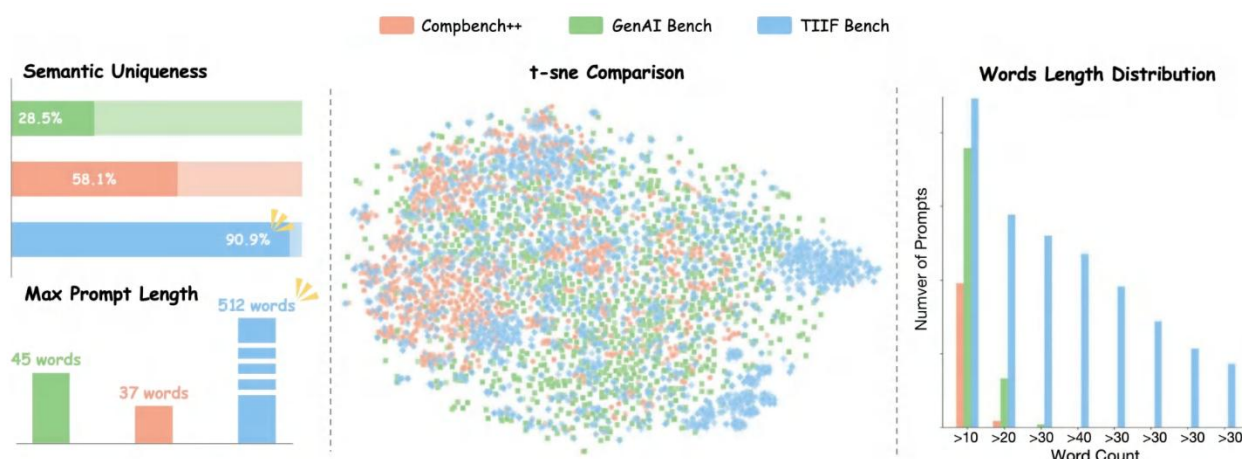


Figure 1: Prompt diversity/complexity of **TIIF-Bench** compared to prior benchmarks. **(Left)** Semantic uniqueness after de-duplication using a cosine similarity threshold of 0.85: less than 30% and 60% prompts in COMPBENCH++ and GENAI BENCH are unique, while more than 90% prompts in TIIF-Bench are unique. **(Middle)** t-SNE visualization of CLIP text embeddings shows that TIIF-Bench spans a much broader semantic space than existing benchmarks. **(Right)** Prompts length in COMPBENCH++ and GENAI BENCH are fixed and short, while TIIF-Bench covers a much wider range. As shown in the lower-left corner, the longest prompts in COMPBENCH++ and GENAI BENCH are less than 50 words, while TIIF-Bench contains complex prompts exceeding 500 words.

主要亮点

- **大规模多维度指令数据**: 5000 条 prompt 显著覆盖语义空间, 超过 90% 唯一性, 远超以往 benchmark, 并提供短/长版本用于鲁棒性测试。
- **新增文字渲染与风格控制维度**: 通过 OCR Recall 和 GNED 等指标评估模型在渲染文字和维持风格一致性方面的能力, 这是现有评测中较少关注的细节。
- **自动化精细评估机制**: 利用 GPT-4o、QwenVL2.5 等 VLM 自动评分, 与人类评价高度一致, 极大提升评测的可扩展性和可靠性。

相关链接

Paper: <http://arxiv.org/abs/2506.02161v1>

GitHub: https://a113n-w3i.github.io/TIIF_Bench/

● 论文十一

Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training

日期

2025-06-02

Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training

Pierre-Carl Langlais

Carlos Rosas Hinostroza

Mattia Nee

Catherine Arnett

Pavel Chizhov

Eliot Krzystof Jones

Irène Girard

David Mach

Anastasia Stasenko

Ivan P. Yamshchikov

PleIAs, Paris, France <https://pleias.fr/>

内容提要

论文介绍 Common Corpus——全球规模最大的合规开源预训练语料库，约 2 万亿 tokens，仅包含公共领域或具有可分发许可内容，并涵盖多语种、代码、科学文献、政府法律文本等。作者详细说明数据来源（如 OpenAlex、GitHub、公共政府文件等）、过滤策略、许可证管理、OCR 校正和有害内容剔除流程，以及数据结构与文件组织。Current AI、Anthropic 等已采用该语料，大力推动合法合规的 LLM 开源训练生态。

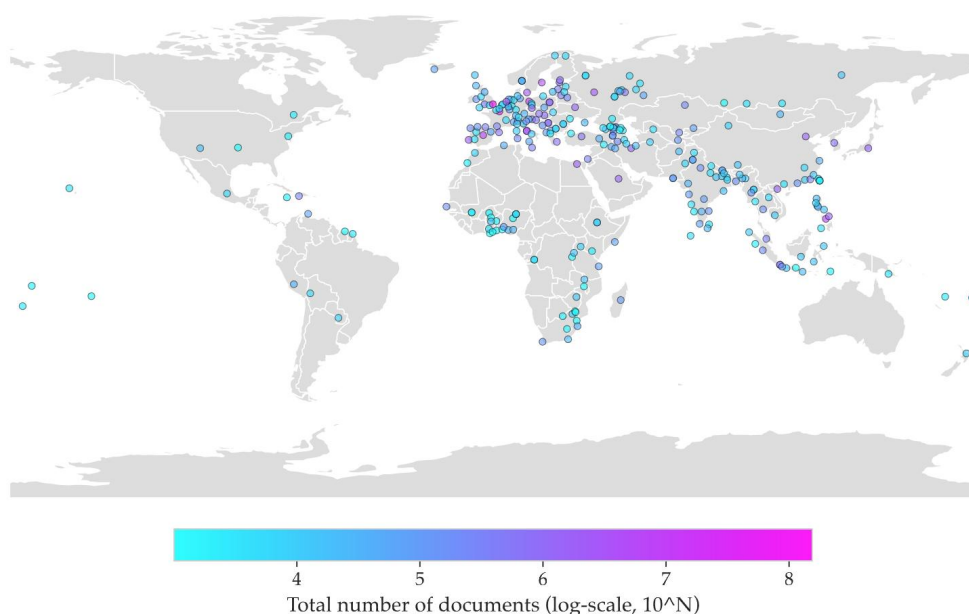


Figure 1: A schematic world map of languages in Common Corpus with a log-scaled distribution of document counts. For each language, we chose a city that is located in the region where this language is most specific to. To avoid outliers, we show only languages with 1000+ documents.

主要亮点

- **迄今最大规模、具备合法许可的 2T tokens 语料**: 涵盖 1998B tokens, 9 种主要语言与 33 个语言子集, 包含科学、代码、文化遗产等多种内容。
- **细粒度许可证与可追溯性**: 对每个文档记录明确许可 (CC- By, CC0, MIT 等), 满足 EU AI Act 法规要求、支持可追溯使用。
- **全面的清洗与校正机制**: 包括 OCR 错误修正、有毒/低教育价值内容剔除、格式与拼写规范化, 确保语料高质与安全。

相关链接

Paper: <https://arxiv.org/abs/2506.01732>

Data: https://huggingface.co/datasets/PleIAs/common_corpus

● 论文十二

HaploOmni: Unified Single Transformer for Multimodal Video Understanding and Generation

日期

2025-06-03

HaploOmni: Unified Single Transformer for Multimodal Video Understanding and Generation

Yicheng Xiao^{1,2*}, Lin Song^{2*✉}, Rui Yang³, Cheng Cheng⁴, Zunnan Xu¹,
Zhaoyang Zhang², Yixiao Ge², Xiu Li^{1✉}, Ying Shan²

¹Tsinghua Shenzhen International Graduate School, Tsinghua University

²ARC Lab, Tencent PCG

³The University of Hong Kong ⁴Xi'an JiaoTong University

xiaoyc23@mails.tsinghua.edu.cn ronny.song@tencent.com

内容提要

该文提出 HaploOmni, 一种高效训练范式, 用于构建统一的单一 Transformer 模型, 具备图像与视频的理解和生成能力。其核心在于融合预训练模型 (包括 ViT、LLM 和 Diffusion Transformer) 通过 “multimodal warmup” 阶段进行初始化, 之后采用统一调优策略。同时使用特征预缩放 (feature pre-scaling) 和多模态 AdaLN (Multimodal AdaLN) 解决跨模态兼容性问题。实验证明, 在图像和视频理解与生成多个基准上, HaploOmni 在性能和训练效率上均超过多种先进模型, 且训练成本低廉。

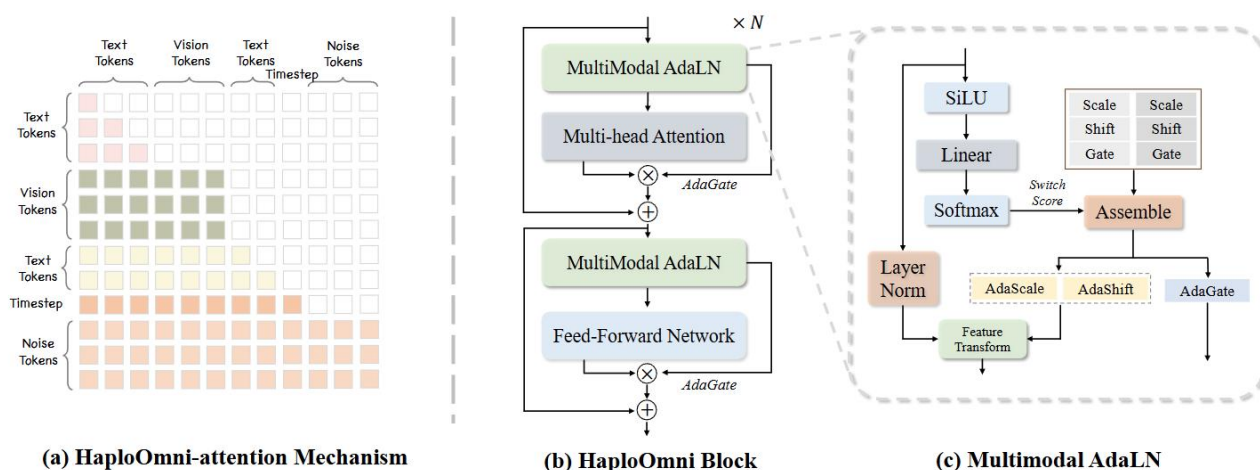


Figure 1: Illustration of our HaploOmni-attention mechanism and HaploOmni Block. We implement a hybrid masking strategy that applies causal attention to text features and timestep tokens while adopting bidirectional attention for processing visual signals and latent noise. Drawing from the standard transformer module, we develop the HaploOmni block through the implementation of multimodal AdaLN.

主要亮点

- **预训练模型融合的 warmup 时序**: 通过 depth-wise 划分, 将视觉编码器、语言解码器、视觉解码器分别初始化, 再整体联合调优, 实现高效迁移。
- **跨模态兼容性创新**: 引入 feature pre-scaling 和 Multimodal AdaLN, 有效提升训练稳定性及多模态表示能力。
- **统一模型性能强劲**: 在 SEEDBench、POPE、MVBench、VBench 等多个图像/视频理解及生成任务上, HaploOmni 均优于已有单模型方案且低成本训练。

相关链接

Paper: <http://arxiv.org/abs/2506.02975v1>

GitHub: <https://github.com/Tencent/HaploVLM>

● 论文十三

UniWorld-V1: High-Resolution Semantic Encoders for Unified Visual Understanding and Generation

日期

2025-06-03



UniWorld-V1: High-Resolution Semantic Encoders for Unified Visual Understanding and Generation

Bin Lin^{1,3}, Zongjian Li^{1,3}, Xinhua Cheng^{1,3}, Yuwei Niu^{1,3}, Yang Ye^{1,3}, Xianyi He^{1,3}, Shenghai Yuan^{1,3}, Wangbo Yu^{1,3}, Shaodong Wang^{1,3}, Yunyang Ge^{1,3}, Yatian Pang¹, Li Yuan^{1,2,†}

¹ Peking University, Shenzhen Graduate School, ² Peng Cheng Laboratory, ³ Rabbitpre AI

🔗 Code: <https://github.com/PKU-YuanGroup/UniWorld-V1>
😊 Models: <https://huggingface.co/LanguageBind/UniWorld-V1>
😊 Data: <https://huggingface.co/datasets/LanguageBind/UniWorld-V1>

内容提要

本论文提出 UniWorld- V1, 一种利用高分辨率语义编码器构建的统一视觉理解与生成框架。作者观察到 GPT- 4o- Image 在图像操作任务中更依赖于语义编码器而非传统 VAE, 因而采用来自强大视觉语言模型的语义特征, 这是 UniWorld- V1 的核心能力来源。该框架仅使用约 270 万标注数据训练, 即可在图像理解、生成与编辑多个任务上达到巨型模型级别性能, 同时保持算力与数据的高效性。相关模型权重、训练与评估代码均已开源, 便于社区复现与扩展。

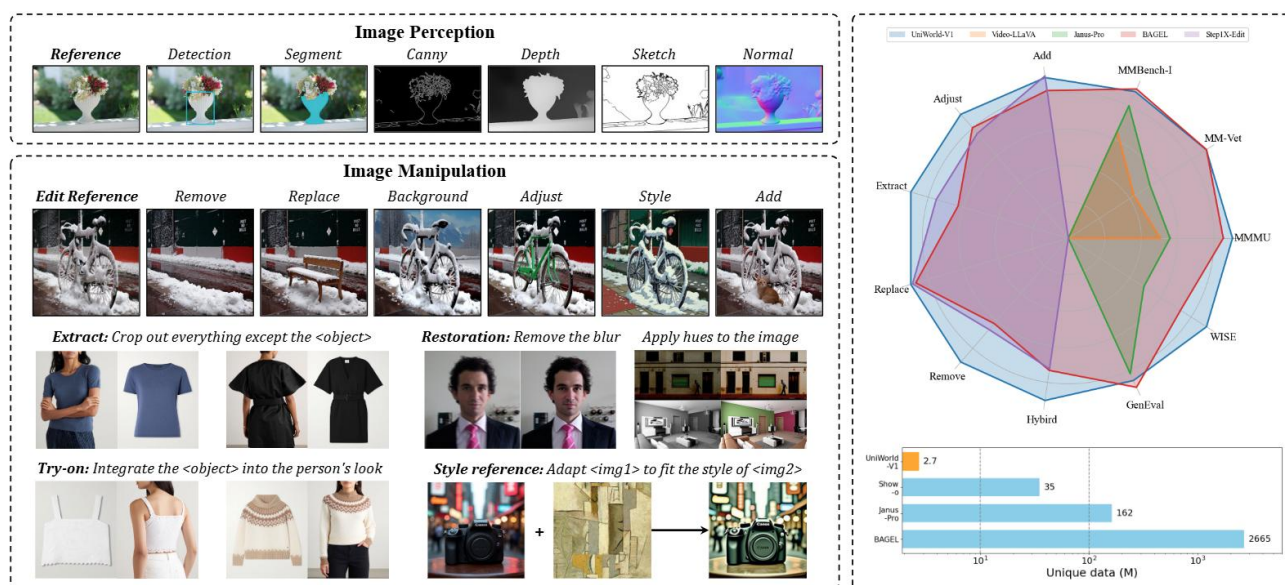


Figure 1: **Showcase of UniWorld-V1's versatile capabilities.** The left two panels illustrate image perception and manipulation examples, and the right panel presents comparisons with state-of-the-art models and training data resources.

主要亮点

- **语义编码器替代 VAE**: 通过实验证明, 语义编码器 (contrastive semantic encoders) 比传统 VAE 更适合作为图像操作任务的特征基础。
- **统一视觉理解与编辑能力**: 在 WISE 语义理解基准、ImgEdit-Bench 编辑测试中表现优异, 对比开源和 GPT- 4o- Image 均实现竞争表现。
- **轻量训练数据 & 全开源**: 仅使用 2.7M 样本训练, 且发布全套代码与模型, 促进研究者快速构建与复用该框架。

相关链接

Paper: <https://arxiv.org/abs/2506.03147>

GitHub: <https://github.com/PKU-YuanGroup/UniWorld-V1>

Data: <https://huggingface.co/datasets/LanguageBind/UniWorld-V1>

● 论文十四

Multi-timescale reinforcement learning in the brain

日期

2025-06-04 (Nature 录用论文)

nature

Explore content ▾

About the journal ▾

Publish with us ▾

Subscribe

[nature](#) > [articles](#) > [article](#)

Article | Published: 04 June 2025

Multi-timescale reinforcement learning in the brain

[Paul Masset](#) , [Pablo Tano](#), [HyungGoo R. Kim](#), [Athar N. Malik](#), [Alexandre Pouget](#)  & [Naoshige Uchida](#) 

[Nature](#) **642**, 682–690 (2025) | [Cite this article](#)

9830 Accesses | **3** Citations | **185** Altmetric | [Metrics](#)

内容提要

本文系统探讨了大脑中是否存在“多时间尺度强化学习 (multi-timescale RL)”的机制，并提出该机制在生物与人工智能中均具有重要意义。作者首先从理论与仿真实验出发，证明引入多种 discount factor (折扣因子) 可以有效分离奖励的时间与幅度信息，使得学习更稳健并能更好适应不确定环境。随后通过实验证据表明，小鼠中脑多巴胺神经元在两种行为任务中的反应可被多个折扣因子所建模，不同神经元具有稳定的个体化时间偏好，提示大脑确实以“多时间尺度”的方式编码未来奖励。这一发现不仅可解释多巴胺信号中的异质性，也为解释人类行为中的非指数型折扣现象提供了神经机制基础，并对设计更高效的 RL 算法提出了启发。

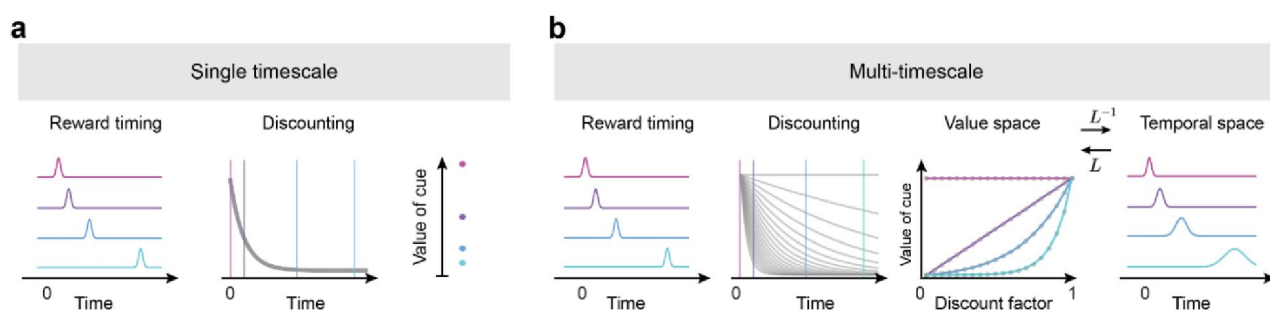


Figure 1 | Single timescale and multi-timescale reinforcement learning. *a*, In single-timescale value learning, the value of a cue (at $t = 0$) predicting future rewards (first panel) is evaluated by discounting these rewards with a single exponential discounting function (second panel). The expected reward size and timing are encoded, but confounded, in the value of the cue (third panel). *b*, In multi-timescale value learning, the same reward delays are evaluated with multiple discounting functions (second panel). The relative value of a cue as a function of the discount depends on the reward delay (third panel). A simple linear decoder based on the Laplace transform can thus reconstruct both the expected timing and magnitude of rewards (fourth panel).

主要亮点

- **首次提出多巴胺神经元编码多时间尺度误差信号 (Vector TD Error)**：挑战经典 TD 理论中单折扣因子的设定，实验证明个体神经元具有稳定而多样化的折扣时间常数，这种“异质性”是神经系统自然实现非指数折扣的机制。
- **Laplace 解码结构揭示奖励时间表征能力**：利用逆拉普拉斯变换 (Z-transform)，作者从神经元群体响应中成功还原奖励发生的时间分布，表明多时间尺度编码可支持对未来事件精准预测，而非仅依赖总价值估计。
- **建立生物 RL 与分布式 RL 理论的桥梁**：论文显示多个单元的指数型折扣组合可表现出类似超曲线 (hyperbolic) 的全局行为，从而解释动物行为中常见的“短期冲动+长期谨慎”决策逻辑，也对未来 RL 算法的设计具有实用参考价值。

相关链接

Paper: <https://www.nature.com/articles/s41586-025-08929-9>

GitHub: https://github.com/pablotano8/multi_timescale_RL

● 论文十五

OpenThoughts: Data Recipes for Reasoning Models

日期

2025-06-04



Open Thoughts

DATA RECIPES FOR REASONING MODELS

Etash Guha^{*1,2}, Ryan Marten^{*3}, Sedrick Keh^{*4}, Negin Raoof^{*5}, Georgios Smyrnis^{*6},
Hritik Bansal^{ζ7}, Marianna Nezhurina^{ζ8,9,16}, Jean Mercat^{ζ4}, Trung Vu^{ζ3}, Zayne Sprague^{ζ6},
Ashima Suvarna⁷, Benjamin Feuer¹⁰, Liangyu Chen¹, Zaid Khan¹¹, Eric Frankel²,
Sachin Grover¹², Caroline Choi¹, Niklas Muennighoff¹, Shiye Su¹, Wanjia Zhao¹, John Yang¹,
Shreyas Pimpalgaonkar³, Kartik Sharma³, Charlie Cheng-Jie Ji³, Yichuan Deng²,
Sarah Pratt², Vivek Ramanujan², Jon Saad-Falcon¹, Jeffrey Li², Achal Dave, Alon Albalak¹³,
Kushal Arora⁴, Blake Wulfe⁴, Chinmay Hegde¹⁰, Greg Durrett⁶, Sewoong Oh²,
Mohit Bansal¹¹, Saadia Gabriel⁷, Aditya Grover⁷, Kai-Wei Chang⁷, Vaishaal Shankar,
Aaron Gokaslan¹⁴, Mike A. Merrill¹, Tatsunori Hashimoto¹, Yejin Choi¹,
Jenia Jitsev^{8,9,16}, Reinhard Heckel¹⁵, Maheswaran Sathiamoorthy³,
Alexandros G. Dimakis^{†3,5}, Ludwig Schmidt^{†1}

¹Stanford University, ²University of Washington, ³BespokeLabs.ai, ⁴Toyota Research Institute,
⁵UC Berkeley, ⁶UT Austin, ⁷UCLA, ⁸JSC, ⁹LAION, ¹⁰NYU, ¹¹UNC Chapel Hill,
¹²ASU, ¹³Lila Sciences, ¹⁴Cornell Tech ¹⁵TUM ¹⁶Open-Ψ (Open-Sci) Collective

内容提要

OpenThoughts 项目针对可公开获取训练数据不足的问题，设计并发布了一套开放式“推理模型数据配方”。作者构建了适配数学、科学及代码推理的结构化数据集，并搭建训练流，用以评估数据选择策略的影响。基于这些策略，构建了 OpenThinker3-7B 模型，超越其他同规模开源推理模型。

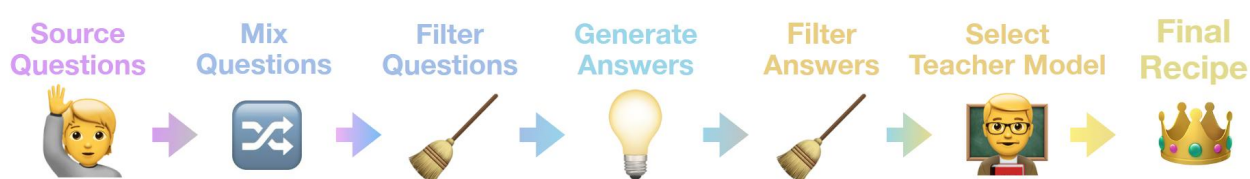


Figure 2: **The OpenThoughts experiment pipeline aims to build the strongest reasoning dataset recipe.** We investigate (1) sourcing questions from existing and newly generated datasets, (2) mixing questions from the top-performing sources, (3) filtering for high-quality questions using `fastText` or LLMs, (4) deduplicating questions and sampling multiple answers per question, (5) filtering out low-quality answers using LLM verification or majority consensus, and (6) selecting the best teacher model.

主要亮点

- **开放式训练数据配方**：系统构建和发布面向推理任务的训练数据流程，方便社区复现。
- **OpenThinker3- 7B 模型突破**：模型在数学与代码推理 benchmark 中展现领先性能，表明数据驱动提升路径具有大潜力。
- **社区友好资源开放**：包括训练逻辑、数据、模型权重均可获取，促进推理模型生态发展。

相关链接

Paper: <https://arxiv.org/abs/2506.04178>

GitHub: <https://github.com/open-thoughts/open-thoughts>

● 论文十六

Towards LLM-Centric Multimodal Fusion: A Survey on Integration Strategies and Techniques

日期

2025-06-05

Towards LLM-Centric Multimodal Fusion: A Survey on Integration Strategies and Techniques

Jisu An¹, Junseok Lee¹, Jeoungeun Lee², Yongseok Son³,

¹Seoul National University, ²University of California San Diego, ³Chung-Ang University,
{ajs7270, shawn159}@snu.ac.kr
jel166@ucsd.edu, sysganda@cau.ac.kr

内容提要

随着多模态任务的复杂化，以 LLM 为中心的多模态融合范式 (LLM-centric fusion) 逐渐成为主流。本文系统回顾了视觉、语音、视频等模态如何与大语言模型协同工作的融合策略，依据信息流的交汇位置将其划分为三类：早期融合（输入级别整合）、中期融合（语义嵌入对齐）、后期融合（结果级别合并）。此外，文章还讨论了跨模态通信机制、联合训练与解耦训练策略，并强调“对齐瓶颈”在多模态生成与推理中的影响。作者指出，多模态融合的未来方向应加强结构对齐、模态自适应调控与高效训练机制。

Model	Abstractor Layer	Projection Layer	Semantic Embedding Layer	Cross-attention Layer
AnomalyGPT ★(Gu et al., 2024)	-	Linear	Convolution	-
BLIP-2 ★(Li et al., 2023b)	-	Linear	Q-former	-
Cambrian-1 ♦(Tong et al., 2024)	-	-	Cross-attention (like Convolution)	-
CogAgent ◇(Hong et al., 2024)	-	MLP	-	Within Model
Flamingo ♦(Alayrac et al., 2022)	Perceiver Resampler	-	-	Within Model
InstructBLIP (Dai et al., 2023)	-	Linear	Q-former	-
Kosmos-1 (Huang et al., 2023)	Perceiver Resampler	MLP (last layer of ViT)	-	-
LHRS-Bot (Muhtar et al., 2024)	-	-	Perceiver Resampler	-
LLaMA-Adapter V2 ♦(Gao et al., 2023)	-	Linear	-	Within Model (self attention layer)
LLaVA (Liu et al., 2023)	-	Linear	-	-
LLaVA-1.5 (Liu et al., 2024)	-	MLP	-	-
MM1 (McKinzie et al., 2024)	C-Abstractor	-	-	-
mPLUG-Owl2 ♦(Ye et al., 2024c)	Visual Abstractor	-	-	Modality Adaptive Module
mPLUG-Owl3 ♦(Ye et al., 2024a)	-	Linear	-	Within Model (Hyper Attention Transformer block)
PMC-VQA (Zhang et al., 2024d)	-	MLP / Transformer	-	-
SEAL ▲(Sun et al., 2025)	Convolution	MLP	-	-
SEED ★(Ge et al., 2023)	-	Linear	Q-former	-
VideoChat (K.-Y. R. Li et al., 2023)	Q-former	Linear	-	-
VILA (Lin et al., 2024b)	-	Linear / Transformer	-	-
VisionLLM (Wang et al., 2023)	-	Transformer	-	-
X-InstructBLIP (Panagopoulou et al., 2024)	-	Q-former + Linear	-	-

Table 1: Selected Comparisons of MLLMs. Key: Fusion Level is indicated by ♦ (Intermediate) or ◇ (Hybrid); other fusion instances are 'Early'. Representation is indicated by ▲ (Coordinate) or ★ (Hybrid); other representation instances are 'Joint'.

主要亮点

- **系统归纳三类融合范式**：清晰描述 early/mid/late fusion 机制，解析其适用场景与技术优劣。
- **强调跨模态通信机制设计**：介绍跨注意力、门控机制与控制 token 等常用交互手段。
- **提供多维度评估框架**：提出以 representation capacity、reasoning ability、extensibility 为核心的评估体系。

相关链接

Paper: <https://arxiv.org/pdf/2506.04788>

● 论文十七

Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models

日期

2025-06-05

Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models

Yanzhao Zhang* Mingxin Li* Dingkun Long* Xin Zhang*
Huan Lin Baosong Yang Pengjun Xie An Yang
Dayiheng Liu Junyang Lin Fei Huang Jingren Zhou
Tongyi Lab Alibaba Group



<https://huggingface.co/Qwen>

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

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

内容提要

本文提出 Qwen3 Embedding 系列，基于 Qwen3 大规模基础模型，通过多阶段训练流程（大规模无监督预训练 + 高质量有监督微调），实现了在文本嵌入与重排序任务上的显著提升。实验覆盖多语种文本理解与信息检索任务，证明其在多项基准上效果优于前代 GTE-Qwen 系列。

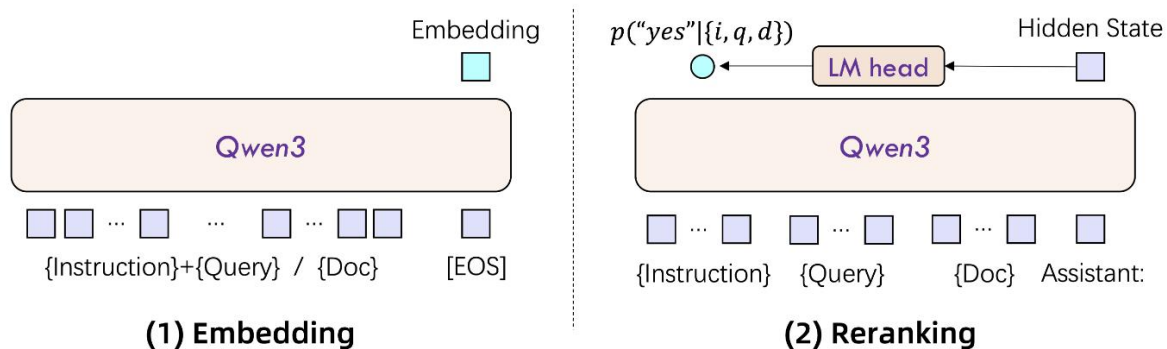


Figure 1: Model architecture of Qwen3-Embedding (left) and Qwen3-Reranker (right).

主要亮点

- **多阶段训练流程：**结合无监督预训练与有监督微调，大幅提升嵌入质量与泛化能力。
- **多语种兼容性极强：**在不同语言场景下展现稳定效果，适用于跨语言检索与理解。
- **重排序性能领先：**在 reranking 任务上优于前代模型，是信息检索系统的理想升级。

相关链接

Paper: <https://arxiv.org/abs/2506.05176>

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

● 论文十八

A Survey on Sparse Autoencoders: Interpreting the Internal Mechanisms of Large Language Models

日期

2025-06-06

A Survey on Sparse Autoencoders: Interpreting the Internal Mechanisms of Large Language Models

Dong Shu^{1,†}, Xuansheng Wu^{2,†}, Haiyan Zhao^{3,†}, Daking Rai⁴,
Ziyu Yao⁴, Ninghao Liu², Mengnan Du³

¹Northwestern University ²University of Georgia

³New Jersey Institute of Technology ⁴George Mason University

dongshu2024@u.northwestern.edu, {xw54582, ninghao.liu}@uga.edu, {hz54, mengnan.du}@njit.edu, {drai2, ziyuyao}@gmu.edu

内容提要

大型语言模型（LLMs）的内部机制因高度复杂性而难以解释，而稀疏自动编码器（SAEs）通过学习过完备的稀疏表示，将 LLMs 中叠加的多语义特征解耦为可解释的单元，成为机制解释的重要方法。本文全面综述 SAEs 在 LLM 解释中的应用，主要贡献包括：技术框架：涵盖 SAE 的基础架构（编码器、解码器、损失函数）、改进变体（如 Gated SAE、TopK SAE）及训练策略（层组训练、特征对齐）。解释方法：基于输入的 MaxAct、PruningMaxAct 方法通过激活文本分析特征语义，基于输出的 VocabProj、MutInfo 方法通过词嵌入映射实现解释。评估指标：结构指标（重构误差、稀疏度）和功能指标（单语义性、鲁棒性）结合，形成全面评估体系。实际应用：SAEs 用于模型引导（如毒性控制、情感 steering）和行为分析（如幻觉机制、上下文学习解析），并展现出在概念解耦和模型操控中的潜力。

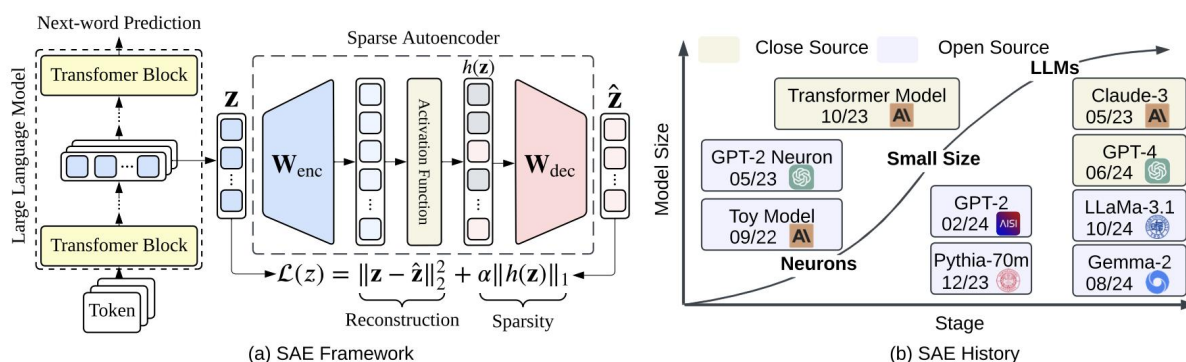


Figure 1: (a) This figure illustrates the fundamental framework of a Sparse Autoencoder (SAE). SAE is trained to take a model representation $\mathbf{z}$ as input and project it to an overcomplete sparse activation $h(\mathbf{z})$ by learning to reconstruct the original input $\hat{\mathbf{z}}$. The SAE typically comprises an encoder, a decoder, and a loss function for training. (b) The development of the SAE progresses through multiple stages. Note that we only list some representative SAE models in this timeline rather than providing an exhaustive compilation.

主要亮点

- **多维度技术框架整合**：系统梳理 SAE 的基础架构与改进路径，包括编码器激活函数优化（如 JumpReLU、ProLU）和训练策略创新（如 BatchTopK、端到端训练），形成从理论到实践的完整技术链条。
- **双向解释方法体系**：构建 “输入激活 - 输出映射” 双向解释框架，既通过文本激活模式挖掘特征语义（如 MaxAct），又通过词嵌入关联实现语义对齐（如 VocabProj），提升解释的全面性。
- **功能导向的评估与应用**：提出结构与功能结合的评估指标（如稀疏探针、因果影响分析），并验证 SAEs 在模型行为操控（如拒绝有害输出、提取任务特征）中的实际效果，推动机制解释向应用落地。

相关链接

Paper: <https://arxiv.org/pdf/2503.05613>

● 论文十九

The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity

日期

2025-06-07

The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity

Parshin Shojae^{*†} Iman Mirzadeh^{*} Keivan Alizadeh
Maxwell Horton Samy Bengio Mehrdad Farajtabar

Apple

内容提要

文章对大型推理模型（LRM）的“思考过程”能力进行了系统评估，发现其在简单环境中生成 chain- of- thought 思路时性能优异，但在面对复杂规划或多步推理问题时会出现“准确性崩塌”。当前评测过度依赖最终答案准确率，而忽视问题复杂度对思维链质量的影响。作者通过在控制难度的拼图任务环境中测试，多角度揭示 LRM 在精确计算和推理一致性方面的系统性偏差。

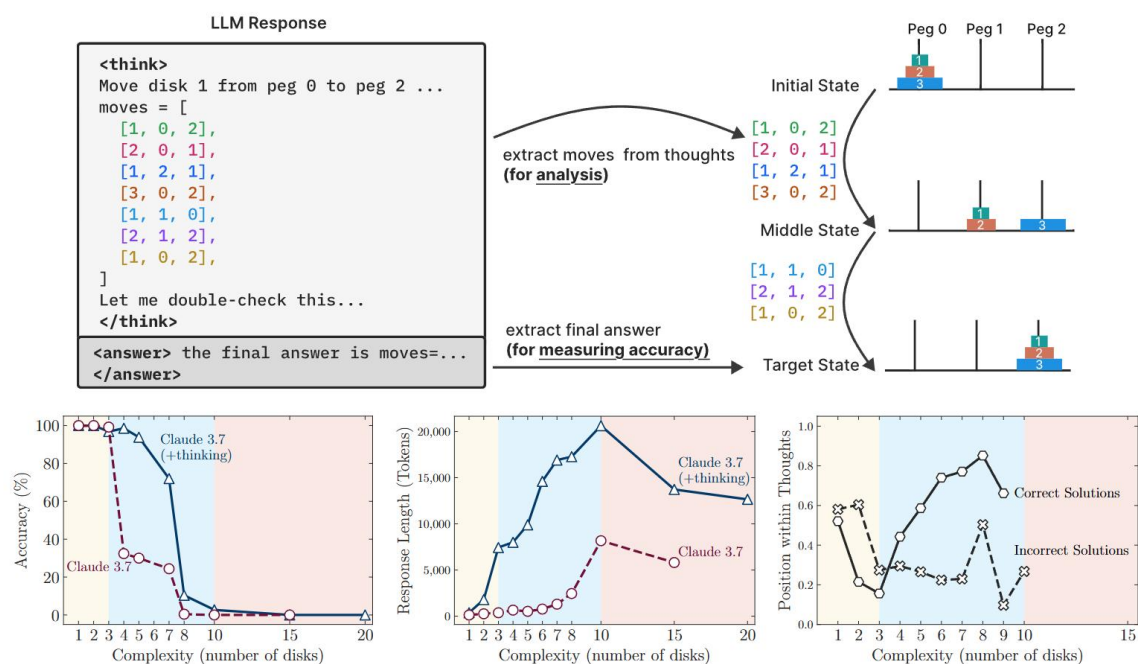


Figure 1: **Top:** Our setup enables verification of both final answers and intermediate reasoning traces, allowing detailed analysis of model thinking behavior. **Bottom left & middle:** At low complexity, non-thinking models are more accurate and token-efficient. As complexity increases, reasoning models outperform but require more tokens—until both collapse beyond a critical threshold, with shorter traces. **Bottom right:** For correctly solved cases, Claude 3.7 Thinking tends to find answers early at low complexity and later at higher complexity. In failed cases, it often fixates on an early wrong answer, wasting the remaining token budget. Both cases reveal inefficiencies in the reasoning process.

主要亮点

- **首个复杂度敏感评估框架：**基于可控规划拼图设计标准化任务，有助区分模型在低 / 中 / 高难度下的思考质量差异。
- **揭示“推理崩塌”现象：**模型在中等/高复杂度任务上准确率显著滑坡，表明 chain- of- thought 并非万能，尤其在细节推理与精确运算中易出错。
- **对评测范式的警示：**强调评估不应仅关注最终答案，还需分析思维路径质量，建议未来加入复杂度标记与过程度量，从而全面提升 LRM 推理能力评估体系。

相关链接

Paper: <https://arxiv.org/abs/2506.06941>

● 论文二十

MAGNET: A Multi-agent Framework for Finding Audio-Visual Needles by Reasoning over Multi-Video Haystacks

日期

2025-06-08

MAGNET: A Multi-agent Framework for Finding Audio-Visual Needles by Reasoning over Multi-Video Haystacks

Sanjoy Chowdhury¹, Mohamed Elmoghany², Yohan Abeysinghe³, Junjie Fei², Sayan Nag⁴,
Salman Khan³, Mohamed Elhoseiny², Dinesh Manocha¹,

¹University of Maryland, College Park ²KAUST ³MBZUAI ⁴Adobe
{sanjoyc, dmanocha}@umd.edu snag@adobe.com mohamed.elhoseiny@kaust.edu.sa
{yohan.abeyasinghe, salman.khan}@mbzuai.ac.ae m.osama.elmoghany@gmail.com

🌐 https://schowdhury671.github.io/magnet_project/

内容提要

该文介绍了 MAGNET，一个多智能体检索系统，专门用于在大量视频中寻找音视频结合的关键片段，以回答复杂查询。作者提出 AVHaystacksQA 基准，包含 3100 个跨视频问答对，并提出 StEM 和 MTGS 两种新评估指标，衡量音视频片段的时序对齐与定位准确度。实验显示 MAGNET 在 BLEU@4 和 GPT 自动评分中分别相对提升了 89% 和 65%，展示了跨视频检索+推理的强大能力。

● 论文二十一

CoCoA-Mix: Confusion-and-Confidence-Aware Mixture Model for Context Optimization

日期

2025-06-09

CoCoA-Mix: Confusion-and-Confidence-Aware Mixture Model for Context Optimization

Dasol Hong¹ Wooju Lee¹ Hyun Myung¹

内容提要

CoCoA-Mix 针对文本 prompt 调优中的上下文优化，提出了一种新的混合模型，该模型通过混合多个子模型预测并根据“混淆”与“置信度”动态调整权重。引入 confusion-aware loss 提升分类边界区分度，并利用 confidence-aware 权重增强优化效果。实验证明，该方法在 few-shot 上下文调优中，能更快收敛并提升最终性能。

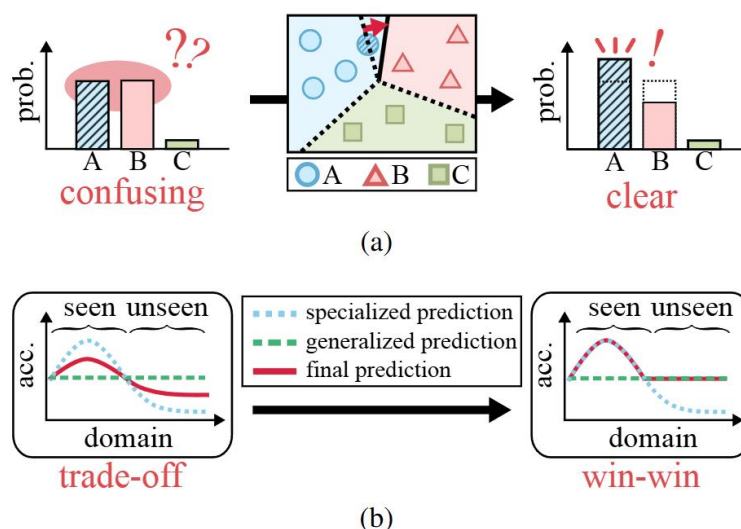


Figure 1: (a) Effect of CoA-loss. The left and right sides show probabilities for the hatched sample on the decision boundary. In the middle figure, dashed and solid lines indicate the current and updated decision boundaries. CoA-loss assigns higher weights to this confusing sample, enhancing specialization in prompt tuning. (b) Effect of CoA-weights. The left and right sides show performance across domains. CoA-weights increase confidence in the specialized prediction for seen domains while reducing it for unseen domains, preserving both specialization and generalization.

主要亮点

- **混淆感知损失 (CoA-loss)**：通过动态调整混淆类别间边界，提高 prompt specificity。
- **置信度权重机制**：根据预测信度，为子模型动态分配组合权重，提升整体鲁棒性。
- **Few-shot 上下文调优效果显著**：方法在少样本环境下训练更快、更准确，具备较强适用性。

相关链接

Paper: <http://arxiv.org/abs/2506.07484v1>

GitHub: <https://github.com/url-kaist/CoCoA-Mix>

● 论文二十二

Diffuse and Disperse: Image Generation with Representation Regularization

日期

2025-06-10

Diffuse and Disperse: Image Generation with Representation Regularization

Runqian Wang

Kaiming He

MIT

{raywang4, kaiming}@mit.edu

内容提要

本文提出了一种新颖的图像生成训练正则化策略——Dispersive Loss，以提升扩散模型的代表能力与生成多样性。作者发现，当前扩散模型尽管能生成高质量图像，但潜在表示往往集中度过高，缺乏区分性。为此，文章引入了基于 representation space 的散布性度量（representation dispersion），鼓励不同图像在编码空间中的表示具有更大的距离，从而提升生成样本的多样性与细节还原能力。该方法可作为 plug-and-play 模块，灵活集成至现有各类扩散模型中。在多个 benchmark（如 ImageNet-64、CIFAR-10、LSUN）上验证，其能系统性提升 FID 分数，并增强图像细节表现与特征覆盖。

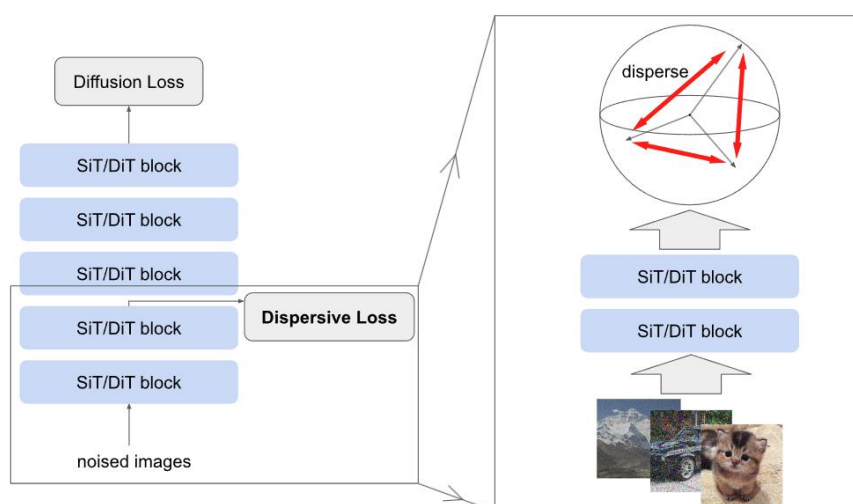


Figure 1: **Dispersive Loss for Generative Modeling.** **Left:** A standard diffusion-based model (e.g., SiT [26] or DiT [31]) with a regression-driven diffusion loss, augmented by a *Dispersive Loss* applied to an intermediate block. **Right:** A zoom-in on the first few blocks with Dispersive Loss. Our loss encourages intermediate representations to disperse in the hidden space. It operates on the same input batch of noised images, shares the existing network blocks and computations, introducing negligible overhead, no additional parameters, and no reliance on external data.

主要亮点

- **提出“Dispersive Loss”正则项：**该项直接作用于潜在空间，增强表示的区分性，有效减少生成样本间的“趋同崩溃”。
- **模型无关、训练友好：**作为 plug-and-play 模块，可轻松集成进现有任意扩散架构，如 DDPM、EDM、Stable Diffusion，训练成本小幅增加但收益显著。
- **兼顾质量与多样性提升：**在 ImageNet、CIFAR 等多个数据集上显著优化 FID、KID 和 LPIPS，生成图像更具结构细节和语义多样性。

相关链接

Paper: <https://arxiv.org/abs/2506.09027>

● 论文二十三

Physical restoration of a painting with a digitally constructed mask

日期

2025-06-11 (Nature 录用论文)

nature

Explore content ▾

About the journal ▾

Publish with us ▾

Subscribe

[nature](#) > [articles](#) > article

Article | Published: 11 June 2025

Physical restoration of a painting with a digitally constructed mask

[Alex Kachkine](#) 

[Nature](#) **642**, 343–350 (2025) | [Cite this article](#)

8548 Accesses | **1** Citations | **609** Altmetric | [Metrics](#)

内容提要

本文提出了一种将数字生成技术与传统修复相结合的新方法,用于对油画表面严重损坏区域进行物理修复。研究团队选取一幅 15 世纪晚期大师马德里致敬大师 (Master of the Prado Adoration) 作品,共计处理 5,612 个损失区域,面积达 66,205 mm²,涉及 57,314 种颜色。他们利用数字模拟技术生成色彩精准的修复图层,采用可逆聚合膜进行双层掩模打印,并将其物理贴附于原作,整个过程用时约 3.5 小时,比传统人工修复快 66 倍。该方法严格遵守绘画修复伦理原则,并提供了高拟合度的视觉效果,同时保持材料可逆性与保存安全性。



Fig. 1 | Physically-applied digital restoration. **a**, The digital infill mask construction process. Starting from the scan of the damaged painting, a restored image is created; an infill mask is derived from the two through analysis of human-perceived visual differences subject to tolerance analysis. **b**, Exploded view rendering of the laminates and materials used in physically applying the infill mask. **c**, Photograph of the restored painting in this work after physical application of the digital infill mask.

主要亮点

- **创新型数字掩模修复手段**：首次将数字图像重建与物理掩模贴附结合，实现高精度可逆修复，突破传统人工局限性。
- **效率提升显著**：3.5 小时完成修复，比手工耗时少 66 倍，大幅降低成本并提高可操作性。
- **符合保护伦理标准**：数字掩模设计时贯彻绘画修复伦理，包括材料可逆性与原作保护，实现“无损”美学修复。

相关链接

Paper: <https://www.nature.com/articles/s41586-025-09045-4>

GitHub: <https://codeocean.com/capsule/0882403/tree>

● 论文二十四

Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the U.S. Workforce

日期

2025-06-11

Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the U.S. Workforce

Yijia Shao*, Humishka Zope*, Yucheng Jiang, Jiaxin Pei, David Nguyen,
Erik Brynjolfsson, Diyi Yang
Stanford University
{shaoyj, diyiy}@cs.stanford.edu

内容提要

复合 AI 系统 (AI 代理) 的崛起引发了劳动力市场变革, 但学界缺乏系统性认知。本文提出新型审计框架, 通过音频增强访谈和人类自主性量表 (HAS), 结合 1500 名劳动者偏好与 AI 专家对 844 项任务的评估, 构建 WORKBank 数据库。研究将任务划分为自动化 “绿灯 / 红灯区” “研发机会区” “低优先级区”, 揭示职业间 HAS 差异及人类技能从信息型向人际型的转型趋势, 强调 AI 发展需匹配人类需求。

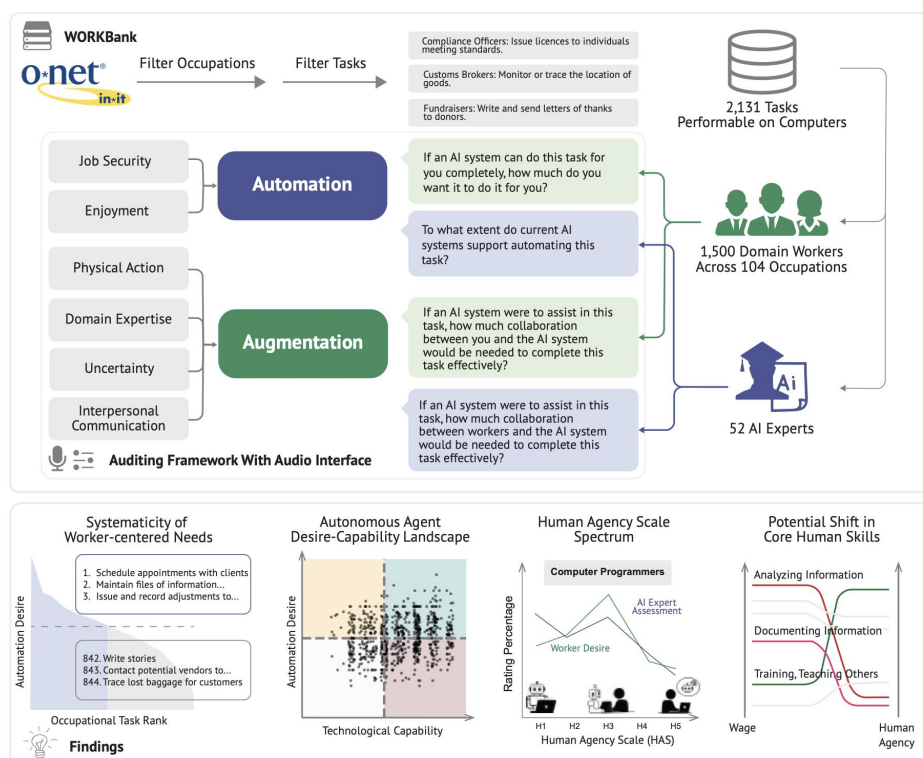


Figure 1: **Overview of the auditing framework and key insights.** The framework captures dual perspectives on automation and augmentation by eliciting both worker desires and expert assessments of technological capabilities. It guides participant reasoning through structured prompts and an audio-enhanced interface. We instantiate this framework to build the WORKBank database, enabling a data-driven analysis of worker-centered needs, the desire-capability landscape, the Human Agency Scale (HAS) spectrum, and implications for core human skills.

主要亮点

- **创新框架与数据库：**融合音频访谈和 HAS 量表，基于 O*NET 构建含 104 个职业、844 项任务的 WORKBank，整合劳动者偏好与技术评估数据。
- **任务四象限分类：**按需求与技术能力将任务分为四类，突破 “自动化与否” 的二元思维，明确 AI 研发的机会与挑战。
- **职业差异与技能转型：**发现各职业对人类参与的需求差异，首次提出 AI 整合可能推动人类核心技能向人际交往能力转变。

相关链接

Paper: <https://arxiv.org/abs/2506.06576>

● 论文二十五

Ming-Omni: A Unified Multimodal Model for Perception and Generation

日期

2025-06-11

Ming-Omni: A Unified Multimodal Model for Perception and Generation

Inclusion AI, Ant Group*

*See Contributions section (Sec. 6) for full author list.

内容提要

Ming-Omni 是一个统一的多模态模型，能够处理图像、文本、音频和视频，同时在语音和图像生成方面表现出较强的能力。该模型采用专用编码器从不同模态提取标记，然后通过配备新提出的特定模态路由器的 MoE 架构 Ling 进行处理。这种设计使单个模型能够在统一框架内高效处理和融合多模态输入，从而无需单独的模型、特定任务的微调或结构重新设计即可完成各种任务。Ming-Omni 超越了传统的多模态模型，支持音频和图像生成。这是通过集成用于自然语音的高级音频解码器和用于高质量图像生成的 Ming-Lite-Uni 实现的，这也使模型能够进行上下文感知聊天、执行文本到语音转换以及进行多种图像编辑。实验结果表明，Ming-Omni 为跨所有模态的统一感知和生成提供了强大的解决方案。值得注意的是，Ming-Omni 是首个在模态支持方面与 GPT-4o 匹配的开源模型，相关团队发布了所有代码和模型权重，以鼓励社区进一步的研究和开发。



Figure 1 Ming-Omni is a versatile, unified end-to-end model capable of processing images, text, video, and audio, and generating text, speech, and images. These capabilities enable the lite version of our model, Ming-Lite-Omni, to support a broad range of tasks, including visual perception, audio-visual interaction, and image generation, among others.

主要亮点

- **统一多模态处理框架**: 通过专用编码器和 MoE 架构 Ling，结合特定模态路由器，实现图像、文本、音频、视频的统一处理与融合，无需针对不同任务单独调整模型结构。
- **跨模态生成能力拓展**: 集成先进音频解码器与 Ming-Lite-Uni，支持自然语音合成、文本转语音及高质量图像生成与编辑，突破传统多模态模型以感知为主的局限。
- **开源与模态支持领先性**: 作为首个在模态支持上匹配 GPT-4o 的开源模型，公开所有代码与权重，为多模态研究提供可复用的基础框架，推动领域技术发展。

相关链接

Paper: <https://arxiv.org/abs/2506.09344>

GitHub: <https://github.com/inclusionAI/Ming/tree/main>

Project: <https://lucaria-academy.github.io/Ming-Omni/>

● 论文二十六

V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning

日期

2025-06-11

V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning

Mahmoud Assran^{1,*}, Adrien Bardes^{1,*}, David Fan^{1,*}, Quentin Garrido^{1,*}, Russell Howes^{1,*}, Mojtaba Komeili^{1,*}, Matthew Muckley^{1,*}, Ammar Rizvi^{1,*}, Claire Roberts^{1,*}, Koustuv Sinha^{1,*}, Artem Zholus^{1,2,*}, Sergio Arnaud^{1,*}, Abha Gejji^{1,*}, Ada Martin^{1,*}, Francois Robert Hogan^{1,*}, Daniel Dugas^{1,*}, Piotr Bojanowski¹, Vasil Khalidov¹, Patrick Labatut¹, Francisco Massa¹, Marc Szafraniec¹, Kapil Krishnakumar¹, Yong Li¹, Xiaodong Ma¹, Sarath Chandar², Franziska Meier^{1,*}, Yann LeCun^{1,*}, Michael Rabbat^{1,*}, Nicolas Ballas^{1,*}

内容提要

现代人工智能的核心挑战之一是通过观察学习理解世界并采取行动。本文提出一种自监督学习方法，将互联网规模的视频数据（超 100 万小时）与少量机器人交互轨迹数据结合，开发出能理解、预测和规划物理世界的模型 V-JEPA 2。该模型先在无动作条件下预训练联合嵌入预测架构，在动作理解任务

(Something-Something v2, top-1 准确率 77.3%) 和人类动作预测 (Epic-Kitchens-100, recall-at-5 达 39.7%) 中超越传统任务专用模型。进一步将 V-JEPA 2 与大型语言模型对齐后，其 80 亿参数规模在视频问答任务中表现突出（如 PerceptionTest 得分 84.0、TempCompass 76.9）。此外，通过对 V-JEPA 2 进行潜在动作条件后训练（V-JEPA 2-AC），仅用 Droid 数据集 62 小时未标注机器人视频，即可在 Franka 机械臂上实现零样本物理规划，完成基于图像目标的物体拾取放置，且无需特定环境数据收集或任务奖励。该研究证明，结合网络规模自监督数据与少量机器人交互数据，可构建具备物理世界规划能力的通用世界模型。

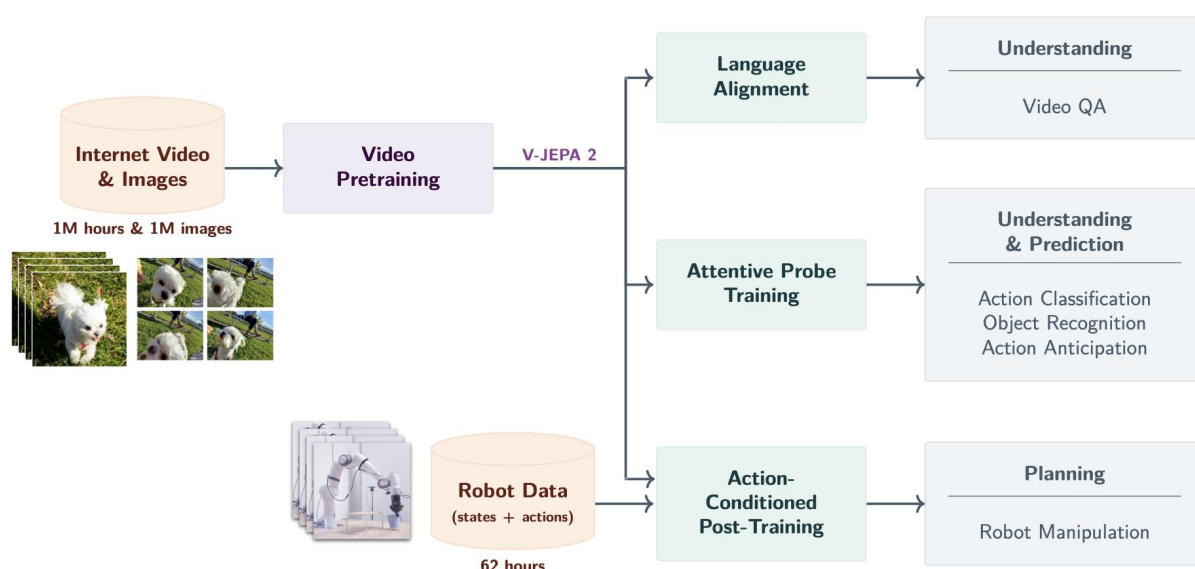


Figure 1 V-JEPA 2 Overview. Leveraging 1M hours of internet-scale video and 1M images, we pretrain the V-JEPA 2 video model using a visual mask denoising objective (Bardes et al., 2024; Assran et al., 2023), and leverage this model for downstream tasks such as action classification, object recognition, action anticipation, and Video Question Answering by aligning the model with an LLM backbone. After pretraining, we can also freeze the video encoder and train a new action-conditioned predictor with a small amount of robot interaction data on top of the learned representations, and leverage this action-conditioned model, V-JEPA 2-AC, for downstream robot manipulation tasks using planning within a model predictive control loop.

主要亮点

- **海量无标签视频预训练构建通用世界模型：**V- JEPA 2 在超过 100 万小时互联网视频上进行自监督训练，显著提升视频理解与预测能力，无需任何动作监督信号。
- **多任务卓越性能：**在运动理解 (Something- Something v2)、人类动作预测 (Epic- Kitchens- 100)、视频问答 (PerceptionTest、TempCompass) 等多个评测中均达到或超越任务专用模型的水平。
- **零样本机器人规划与执行能力：**通过少量无标签机器人视频后训练出的 V- JEPA 2- AC 可手臂零样本部署，完成图像目标下的抓取与放置，展示了从视觉到机器人行动的闭环潜力。

相关链接

Paper: <https://arxiv.org/abs/2506.09985>

GitHub: <https://github.com/facebookresearch/vjepa2>

Model: <https://huggingface.co/collections/facebook/v-jepa-2-6841bad8413014e185b497a6>

● 论文二十七

Self-Adapting Language Models

日期

2025-06-12

Self-Adapting Language Models

Adam Zweiger* Jyothish Pari*† Han Guo Ekin Akyürek Yoon Kim Pulkit Agrawal†
Massachusetts Institute of Technology
{adamz, jyop, hanguo, akyurek, yoonkim, pulkitag}@mit.edu

内容提要

大型语言模型 (LLMs) 功能强大但结构静态, 缺乏针对新任务、知识或示例自适应调整权重的机制。为此, 本文提出自适应性语言模型框架 SEAL (Self-Adapting LLMs), 其核心在于使 LLM 能够通过生成自身微调数据和更新指令实现自我适应: 给定新输入时, 模型生成 “自我编辑” 内容, 可包括信息重组、优化超参数指定或调用工具进行数据增强与基于梯度的更新, 且这些自我编辑通过监督微调 (SFT) 带来持久的权重更新, 实现长效适应。为训练模型生成有效自我编辑, 研究采用强化学习循环, 将更新后模型的下游任务表现作为奖励信号。不同于依赖独立适应模块或辅助网络的先前方法, SEAL 直接利用模型自身生成内容控制适应过程。知识整合和少样本泛化实验表明, SEAL 向具备自主适应能力的语言模型迈出了有前景的一步, 相关网站和代码已公开。

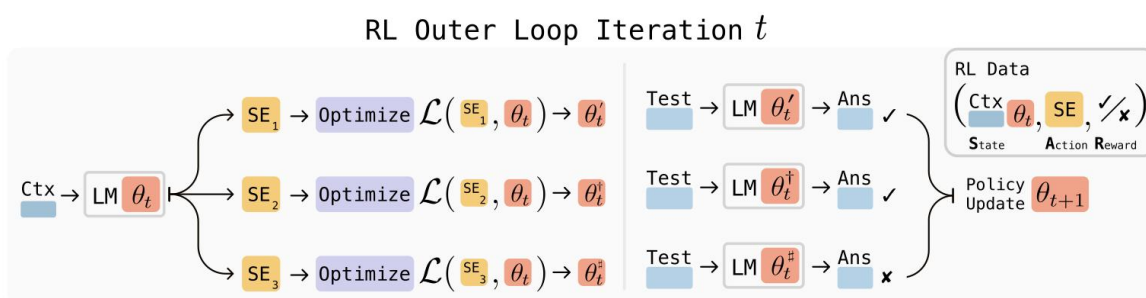


Figure 1: **Overview of SEAL.** In each RL outer loop iteration, the model generates candidate self-edits (SE)—directives on how to update the weights—applies updates, evaluates performance on a downstream task, and uses the resulting rewards to improve the self-edit generation policy.

主要亮点

- **自我生成适应机制：**首创由模型自主生成微调数据与更新指令的“自我编辑”模式，通过重组信息、指定超参数或调用工具实现动态权重更新，突破传统 LLM 静态架构限制。
- **强化学习驱动的自我编辑训练：**以模型下游任务性能为奖励信号构建 RL 循环，使 SEAL 能自主优化自我编辑策略，实现适应过程的端到端闭环优化，无需人工标注或预设规则。
- **无额外模块的轻量级适应框架：**直接利用 LLM 自身生成能力控制适应流程，无需集成独立适应模块或辅助网络，在知识融入与少样本泛化任务中展现高效自主学习潜力。

相关链接

Paper: <https://arxiv.org/abs/2506.10943>

GitHub: <https://github.com/Continual-Intelligence/SEAL>

● 论文二十八

A Survey of Generative Categories and Techniques in Multimodal Large Language Models

日期

2025-06-13

A Survey of Generative Categories and Techniques in Multimodal Large Language Models

LONGZHEN HAN, School of Architecture, Technology and Engineering, University of Brighton, Lewes Road, BN2 4GJ, United Kingdom

AWES MUBARAK, School of Architecture, Technology and Engineering, University of Brighton, Lewes Road, BN2 4GJ, United Kingdom

ALMAS BAIMAGAMBETOV, School of Architecture, Technology and Engineering, University of Brighton, Lewes Road, BN2 4GJ, United Kingdom

NIKOLAOS POLATIDIS, School of Architecture, Technology and Engineering, University of Brighton, Lewes Road, BN2 4GJ, United Kingdom

THAR BAKER, School of Architecture, Technology and Engineering, University of Brighton, Lewes Road, BN2 4GJ, United Kingdom

内容提要

本文系统综述了多模态大语言模型（MLLM）中的生成任务，首次从任务形式、建模目标与融合技术三个维度提出分类体系。文章不仅梳理了文本到图像、图像到文本、视频生成、语音合成等常见任务，还深入解析了各类生成机制的建模流程，如扩散模型、变分自编码器、因果 Transformer 等。同时，从表示对齐、生成多样性、控制能力、泛化性和鲁棒性五个指标对比主流方法性能优劣。作者指出，尽管当前 MLLM 在感知与生成方面取得突破，但控制性弱、模态不对齐和评价标准模糊仍是瓶颈。

A Survey of Generative Categories and Techniques in Multimodal Large Language Models

3

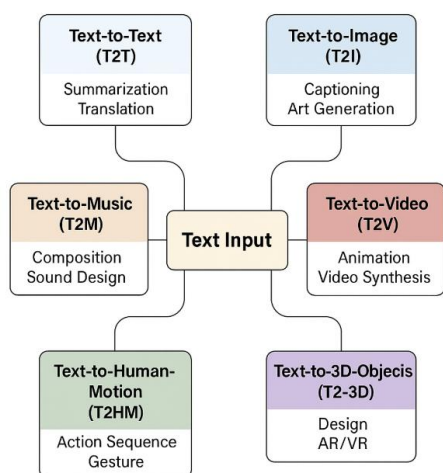


Fig. 1. Modality coverage in LLM research.

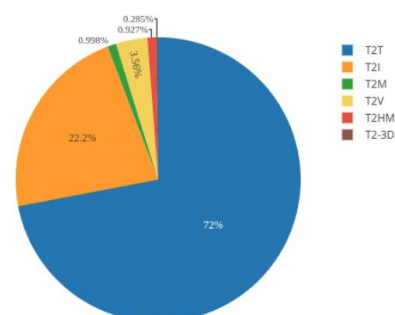


Fig. 2. Technique usage in LLM development.

The pie chart data is derived from a set of keyword-based queries conducted on the Scopus database to capture research trends. Each modality was queried using the following format: *TITLE-ABS-KEY("text generation" AND ("large language model" OR LLM))*; *TITLE-ABS-KEY("image generation" AND ("large language model" OR LLM))*; *TITLE-ABS-KEY("music generation" AND ("large language model" OR LLM))*; *TITLE-ABS-KEY("video generation" AND ("large language model" OR LLM))*; *TITLE-ABS-KEY("human motion generation" AND ("large language model" OR LLM))*; *TITLE-ABS-KEY("3D object generation" AND ("large language model" OR LLM))*. The results reflect only rough approximations and do not account for paper relevance or duplication. However, they provide a general overview of the current research emphasis, with text generation clearly dominating, followed by image, while the remaining modalities represent emerging but less-explored areas.

主要亮点

- **多角度构建 MLLM 生成技术图谱**: 从“生成内容”和“生成手段”两个视角，重新定义 MLLM 研究空间。
- **提出评估 MLLM 生成能力的五大维度**: 提供结构化评价框架，涵盖 fidelity、diversity、alignment、control 和 generalization。
- **跨模态技术融合趋势总结清晰**: 涵盖 Prompt 调优、多模态扩散、跨模态迁移等热点方向，为后续研究提供明确路线图。

相关链接

Paper: <https://arxiv.org/pdf/2506.10016>

● 论文二十九

A Comprehensive Survey of Deep Research: Systems, Methodologies, and Applications

日期

2025-06-14

A Comprehensive Survey of Deep Research: Systems, Methodologies, and Applications

RENJUN XU* and JINGWEN PENG, Zhejiang University, China

内容提要

本综述探讨了快速发展的深度研究（Deep Research）系统领域，这类人工智能应用通过集成大型语言模型、高级信息检索和自主推理能力，实现复杂研究工作流的自动化。作者分析了 2023 年以来涌现的 80 多个商业及非商业实现，包括 OpenAI/Deep Research、Gemini/Deep Research、Perplexity/Deep Research 及众多开源方案，提出了一种新的分层分类法，从基础模型与推理引擎、工具利用与环境交互、任务规划与执行控制、知识合成与输出生成这四个基础技术维度对系统进行分类。文中探讨了这些系统在学术、科学、商业和教育应用中的架构模式、实现方法及领域特定适配，分析显示当前实现虽具备显著能力，但在信息准确性、隐私、知识产权和可访问性方面存在技术与伦理挑战。综述最后指出了高级推理架构、多模态集成、领域专业化、人机协作和生态系统标准化等有前景的研究方向，这些方向可能塑造这一变革性技术的未来发展。该综述通过提供理解深度研究系统的综合框架，为 AI 增强知识工作的理论理解和更强大、负责任且可访问的研究技术的实际发展做出了贡献。

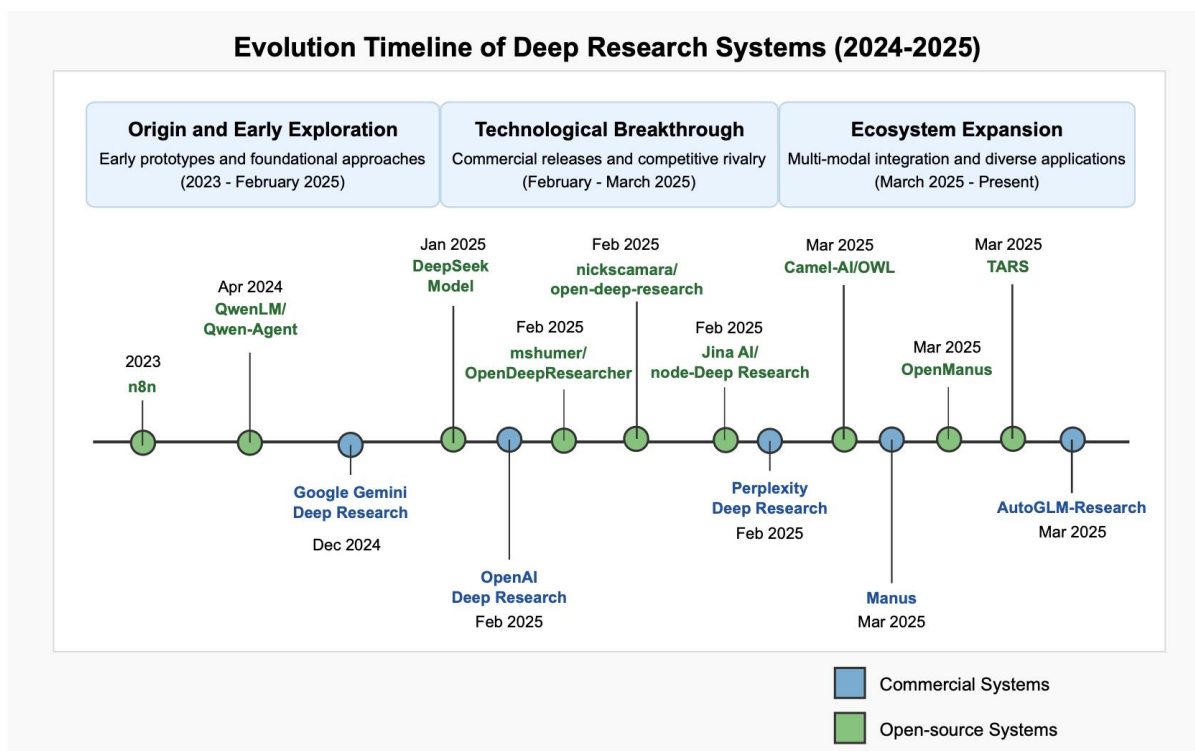


Fig. 1. Evolution Timeline of Deep Research Systems

主要亮点

- **全面的分层分类体系**: 提出基于基础模型、工具交互、任务规划和知识合成的四维分层分类法，系统梳理了 80 余个 Deep Research 系统的技术架构，为领域研究提供了统一的分析框架。
- **多维度应用与挑战剖析**: 覆盖学术、科学、商业等多领域应用案例，深入分析信息准确性、隐私伦理等现实挑战，揭示当前技术在实际落地中的关键障碍。
- **前瞻性研究方向展望**: 指出高级推理架构、多模态融合等人机协作模式及生态标准化等未来发展方向，为推动 Deep Research 系统技术演进提供了明确路径。

相关链接

Paper: <https://arxiv.org/abs/2506.12594>

GitHub: <https://github.com/scienceaix/deepresearch>

● 论文三十

MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention

日期

2025-06-16



MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention

MiniMax¹

内容提要

MiniMax-M1 是全球首个开源的大规模混合注意力推理模型，采用混合专家（MoE）架构与 Lightning Attention 机制，总参数 4560 亿，单 token 激活 459 亿参数。模型原生支持 100 万 token 上下文长度（为 DeepSeek R1 的 8 倍），借助 Lightning Attention 实现测试时计算效率优化——生成 10 万 token 时 FLOPs 消耗仅为 DeepSeek R1 的 25%。模型通过大规模强化学习（RL）训练，结合新型 CISPO 算法（通过裁剪重要性采样权重稳定训练），在 512 个 H800 GPU 上仅用 3 周完成全流程训练，成本约 53.47 万美元。实验表明，MiniMax-M1 在软件工程（SWE-bench）、代理工具使用（TAU-bench）和长上下文理解（OpenAI-MRCR）等任务中表现突出，其中 80K 版本在长上下文场景下超越 OpenAI o3 和 Claude 4，且在数学推理、代码生成等基准测试中媲美或优于 DeepSeek-R1、Qwen3-235B 等开源模型。

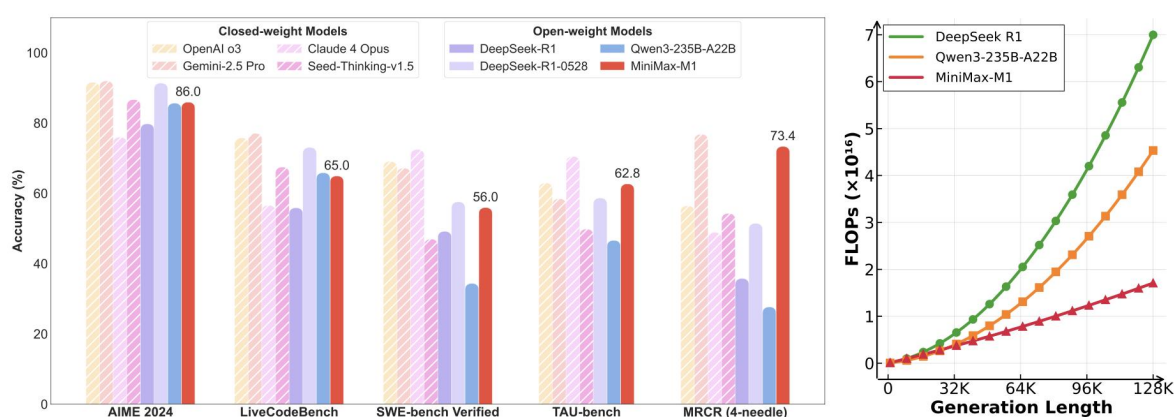


Figure 1 | **Left:** Benchmark performance comparison of leading commercial and open-weight models across competition-level mathematics, coding, software engineering, agentic tool use, and long-context understanding tasks. We use the MiniMax-M1-80k model here for MiniMax-M1. **Right:** Theoretical inference FLOPs scaling with generation length (# tokens).

主要亮点

- **混合架构与高效注意力设计：**融合 MoE 与 Lightning Attention，通过线性复杂度注意力机制突破传统 Transformer 的二次计算瓶颈，实现百万级上下文处理能力，同时降低长序列生成的 FLOPs 消耗。
- **CISPO 算法提升 RL 效率：**提出新型 RL 算法 CISPO，通过裁剪重要性采样权重而非 token 更新，避免低概率但关键推理 token 的梯度丢失，相较 DAPO 实现 2 倍训练加速，保障长链推理能力的稳定进化。
- **长上下文任务与多领域优势：**在 100 万 token 上下文场景中性能领先，尤其在软件工程（代码修复、测试用例生成）和代理工具交互任务中表现突出，且通过动态数据筛选和分段长度扩展策略，解决长序列训练中的模式崩塌问题。

相关链接

Paper: <https://arxiv.org/abs/2506.13585>

GitHub: <https://github.com/MiniMax-AI/MiniMax-M1>

Model: <https://huggingface.co/collections/MiniMaxAI/minimax-m1-68502ad9634ec0eeac8cf094>

● 论文三十一

Essential-Web v1.0: 24T tokens of organized web data

日期

2025-06-17

ESSENTIAL-WEB v1.0: 24T tokens of organized web data

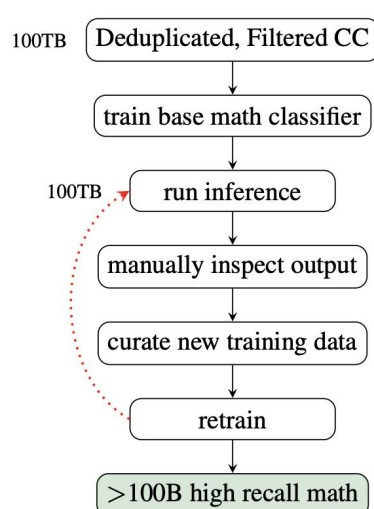
Essential AI
San Francisco, CA
`research@essential.ai`

内容提要

语言模型的技能与知识获取高度依赖数据，但大规模、结构良好的预训练数据集缺失导致数据处理成本高且难以获取。本文介绍 Essential-Web v1.0，这是一个包含 24 万亿 token 的数据集，每篇文档均按主题、格式、内容复杂度和质量等 12 类分类体系进行标注。分类标签由 EAI-Distill-0.5b（一个经过微调的 5 亿参数模型）生成，其标注一致性与 Qwen2.5-32B-Instruct 相比误差在 3% 以内。仅通过 SQL 式过滤，该数据集在数学（较 SOTA 提升 - 8.0%）、网页代码（+14.3%）、STEM（+24.5%）和医疗（+8.6%）等领域可生成具有竞争力的网络精选数据集。Essential-Web v1.0 已在 HuggingFace 平台发布。

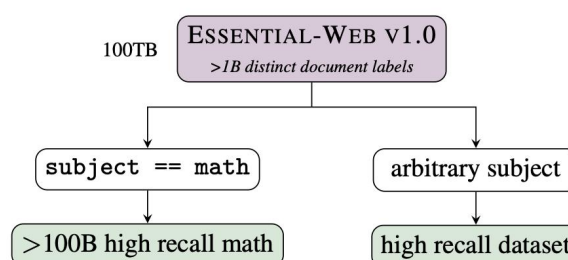
Training High-Recall Classifier

curation timeline: weeks to months



ESSENTIAL-WEB v1.0 Approach

curation timeline: hours to days



主要亮点

- **超大规模结构化数据集**: 构建了包含 24 万亿 token 的 Web 数据集, 通过 12 类分类体系对文档进行系统标注, 覆盖主题、格式、质量等多维度特征, 为语言模型训练提供了标准化的数据基础。
- **轻量级模型实现高效标注**: 使用 5 亿参数的 EAI-Distill-0.5b 模型完成标注, 标注一致性接近 320 亿参数的 Qwen2.5 模型, 在保证准确性的同时大幅降低了计算资源消耗。
- **多领域数据筛选能力**: 通过简单的 SQL 式过滤, 可针对性提取数学、代码、STEM、医疗等专业领域的高质量数据, 部分领域性能超越现有 SOTA 数据集, 展现了数据处理的灵活性与实用性。

相关链接

Paper: <https://arxiv.org/abs/2506.14111>

Dataset: <https://huggingface.co/datasets/EssentialAI/essential-web-v1.0>

● 论文三十二

From Bytes to Ideas: Language Modeling with Autoregressive U-Nets

日期

2025-06-17

From Bytes to Ideas: Language Modeling with Autoregressive U-Nets

Mathurin Videau^{1,*}, Badr Youbi Idrissi^{1,*}, Alessandro Leite³, Marc Schoenauer², Olivier Teytaud¹, David Lopez-Paz¹

¹FAIR at Meta, ²TAU, INRIA and LISN, CNRS & Université Paris-Saclay, ³INSA Rouen Normandy, LITIS, Rouen, France

*Equal contribution

内容提要

当前语言模型的分词（如 BPE）采用固定粒度，构建静态词表后限制了模型处理数据的灵活性及预测范围。

本文提出一种自回归 U-Net 模型，可在训练过程中动态学习嵌入 tokens。该模型直接读取原始字节数据，先将其聚合成词语，再组合为双词、四词等，实现对序列的多尺度表征。深层网络需预测更长远的语义（如未来几个词），聚焦宏观语义模式；浅层则处理细节信息。实验表明，经精心调优预训练计算量后，浅层层级性能可媲美强 BPE 基线，深层层级展现出良好潜力。由于分词过程嵌入模型内部，该系统既能处理字符级任务，也能在低资源语言间传递知识。

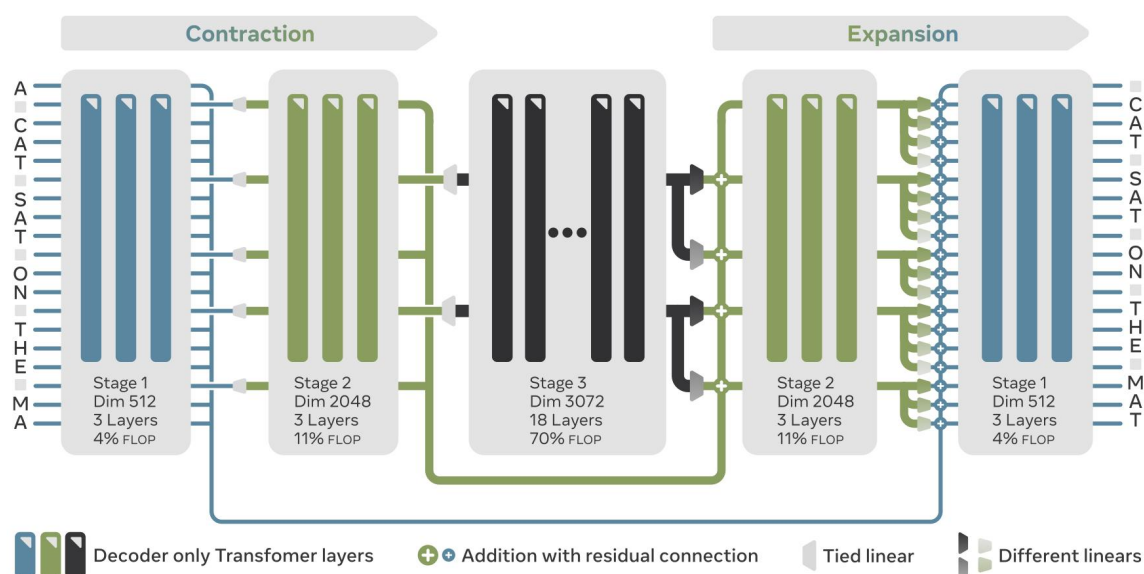


Figure 1 Three-stage Autoregressive U-Net (AU-Net). The model executes from left to right. The contracting path compresses the sequence in two steps: Stage 1 processes raw bytes, Stage 2 keeps only the vector at each word boundary, and Stage 3 keeps one vector per two words. Each contraction and expansion step supports arbitrary pooling and upsampling patterns. After the deepest stage, the expanding path reverses the contracting path by duplicating each coarse vector and applying position-specific linear layers. These are combined with skip connections from the contracting path, gradually restoring sequence length and blending in high-level information. Deeper stages predict further ahead and capture broad semantics, while shallower stages refine local detail.

主要亮点

- **动态分词机制突破固定粒度**：摒弃 BPE 等静态分词方式，通过自回归 U-Net 实现从字节到词语的动态聚合，使模型能自适应不同尺度的语义单元，提升处理文本的灵活性。
- **多尺度层级语义表征**：网络结构分层处理文本 —— 浅层关注字节级细节，深层捕捉跨词语义关联，形成从微观到宏观的层级化理解，优化长序列预测能力。
- **跨任务与语言迁移能力**：将分词逻辑融入模型内部，同一系统可兼容字符级任务（如低资源语言处理），避免传统分词在跨语言场景下的知识割裂问题。

相关链接

Paper: <https://arxiv.org/abs/2506.14761>

GitHub: <https://github.com/facebookresearch/lingua/tree/main/apps/aunet>

● 论文三十三

Revisiting Reinforcement Learning for LLM Reasoning from A Cross-Domain Perspective

日期

2025-06-17

Revisiting Reinforcement Learning for LLM Reasoning from A Cross-Domain Perspective

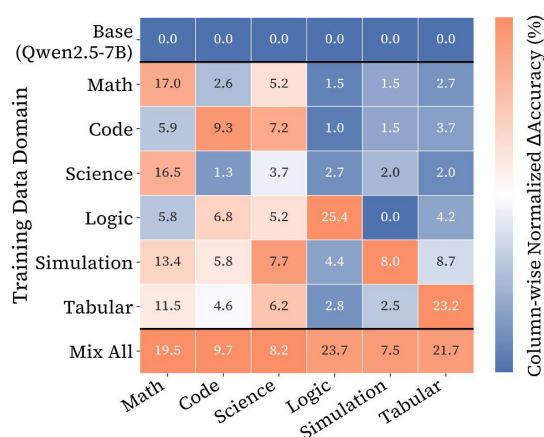
Zhoujun Cheng^{1,*}, Shibo Hao^{1,*}, Tianyang Liu^{1,*}
Fan Zhou², Yutao Xie¹, Feng Yao¹, Yuexin Bian¹, Yonghao Zhuang³, Nilabjo Dey⁴
Yuheng Zha¹, Yi Gu¹, Kun Zhou¹, Yuqi Wang², Yuan Li³, Richard Fan², Jianshu She²
Chengqian Gao², Abulhair Saparov⁴, Haonan Li², Taylor W. Killian², Mikhail Yurochkin²
Zhengzhong Liu², Eric P. Xing^{2,3}, Zhiting Hu¹

¹UC San Diego, ²MBZUAI, ³Carnegie Mellon University, ⁴Purdue University

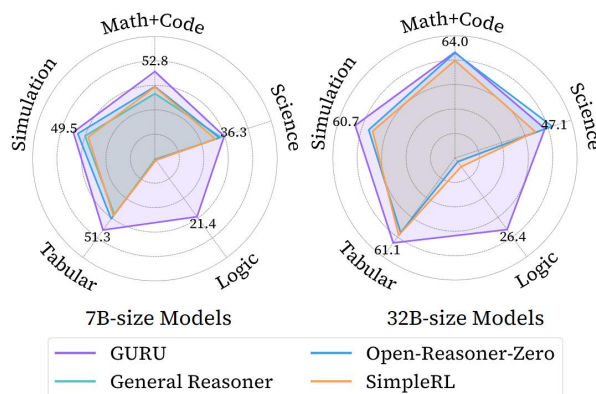
*Equal Contribution

内容提要

本论文重新审视了强化学习（RL）在大型语言模型推理能力提升中的作用，构建了一个跨任务跨领域的 benchmark：Guru，覆盖六大领域（如数学、代码、逻辑、常识等）。通过比较 RL 微调前后在不同任务上的表现，作者发现 RL 确实可提升 LLM 的“推理连贯性”与“结构表达力”，但跨领域迁移效果有限。进一步的分析揭示了 RL 对于不同任务的“过拟合”趋势，以及如何通过 prompt 工程、混合训练等方式增强泛化性。



(a): Cross-domain transfer performance.



(b): Model performance across reasoning domains.

Figure 1: Left: Absolute accuracy gain (%) over the base model when RL-trained on various domains. Pretrained-heavy domains (Math, Code, Science) benefit from cross-domain training, while others require in-domain data, indicating RL aids skill acquisition. **Right:** Our GURU-7B/32B models consistently outperform strong open baselines across 17 reasoning tasks when RL-trained with mixed-domain data.

主要亮点

- **提出 Guru Benchmark:** 覆盖六大领域、百余推理任务，支持评估 RL 策略在复杂 reasoning 中的稳定性与泛化能力。
- **首次分析 RL 在推理上的迁移限制:** 发现 RL 提升通常在领域内有效，在跨领域任务上常出现性能回退。
- **给出优化建议与策略组合:** 结合 RLHF、prompt modulation、multi-task mix 等手段提升跨领域泛化能力，为后续推理增强研究提供基础。

相关链接

Paper: <https://arxiv.org/abs/2506.14965>

GitHub: <https://github.com/LLM360/Reasoning360>

Project: <https://guru-reasoning.github.io>

● 论文三十四

Show-o2: Improved Native Unified Multimodal Models

日期

2025-06-18

Show-o2: Improved Native Unified Multimodal Models

Jinheng Xie¹ Zhenheng Yang² Mike Zheng Shou^{1*}

¹ Show Lab, National University of Singapore ² ByteDance

内容提要

本文提出改进的原生统一多模态模型 Show-o2，其基于 3D 因果变分自动编码器（VAE）空间，通过空间 - 时间融合双路径构建统一视觉表示，实现图像和视频模态的可扩展性，同时利用自回归建模和流匹配分别处理语言和视觉生成任务。两阶段训练策略先通过图像 - 文本和视频 - 文本数据预训练流头以学习视觉生成能力，再用高质量多模态数据微调全模型以保留语言知识。实验表明，Show-o2 在多模态理解（如 MME、GQA）和视觉生成（如 GenEval、DPG-Bench）基准上性能超越现有模型，且支持混合模态生成和视频生成，代码与模型已开源。

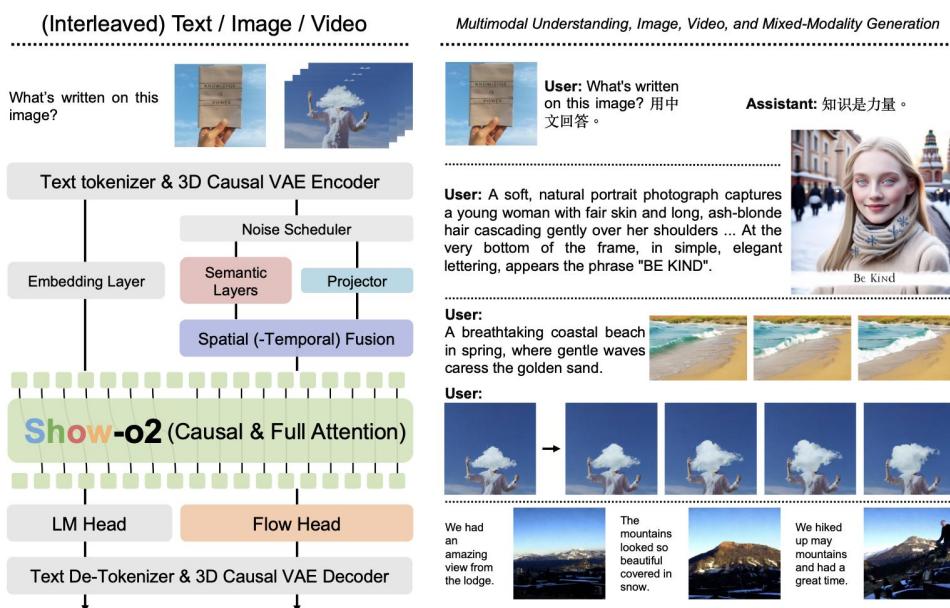


Figure 1: Our approach begins by encoding input texts, images, and videos into continuous embeddings and visual latents. The visual latents are processed through a dual-path extraction and spatial (-temporal) fusion mechanism to construct unified visual representations that are scalable for both multimodal understanding and generation, image and video modalities. These text embeddings and unified visual representations are then structured into a sequence for the base language model, equipped with dedicated heads. Specifically, text tokens are modeled autoregressively by a language head, while image and video latents are handled by a flow head using flow matching. We employ the omni-attention mechanism [118, 133] to enable causal attention along the sequence while maintaining full attention within the unified visual representations. This design empowers our model to effectively tackle tasks such as image/video understanding, generation, and mixed-modality generation.

主要亮点

- **3D 因果 VAE 与双路径融合的统一表征：**通过 3D 因果 VAE 提取图像 / 视频潜在特征，结合语义层和投影器的双路径架构，实现高维语义与底层特征的空间 - 时间融合，使模型同时捕捉视觉语义和结构细节，支持图像与视频模态的统一处理。
- **两阶段训练保留语言知识并增强视觉生成：**首阶段聚焦流头、投影器等组件训练，利用 66M 图像 - 文本对学习视觉生成；次阶段用 9M 多模态理解数据和 16M 生成数据微调全模型，避免大规模文本语料导致的语言能力退化。
- **多模态任务的领先性能与泛化性：**在多模态理解任务中，7B 参数模型超越 14B 参数的 TokenFlow-XL；视觉生成上，1.5B 模型在 DPG-Bench 整体得分 85.02，7B 模型进一步提升至 86.14，且支持视频生成和混合模态叙事。

相关链接

Paper: <https://arxiv.org/abs/2506.15564>

GitHub: <https://github.com/showlab/Show-o>

● 论文三十五

Unsupervised pretraining in biological neural networks

日期

2025-06-18

Article

Unsupervised pretraining in biological neural networks

<https://doi.org/10.1038/s41586-025-09180-y>

Received: 26 February 2024

Accepted: 20 May 2025

Lin Zhong^{1✉}, Scott Baptista¹, Rachel Gattoni¹, Jon Arnold¹, Daniel Flickinger¹, Carsen Stringer^{1,2} & Marius Pachitariu^{1,2✉}

内容提要

本文探讨了生物神经网络中的无监督预训练机制。研究人员通过记录小鼠在执行视觉辨别任务及无奖励暴露期间初级视觉皮层（V1）和高级视觉区域（HVAs）的神经元活动（最多同时记录 90,000 个神经元），发现任务学习后的神经变化在无奖励暴露组中也普遍存在，表明神经可塑性主要由无监督学习驱动，且在内侧 HVAs 中最为显著，遵循视觉而非空间学习规则。仅在任务组的前侧 HVAs 中发现与奖励预测相关的信号，可能涉及监督学习。行为实验证实，无监督预训练可加速后续任务学习，验证了无监督学习对视觉表征的基础作用。

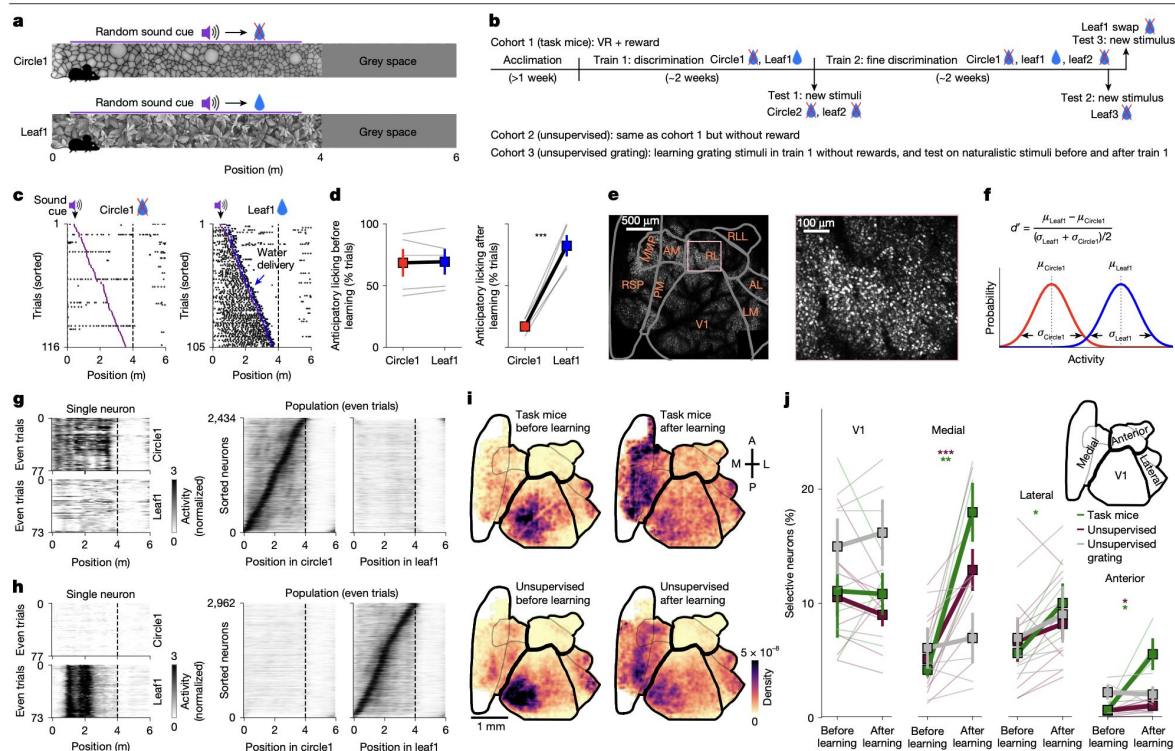


Fig. 1 | Plasticity in the visual cortex after supervised and unsupervised training. **a**, Illustration of the virtual reality (VR) task with a sound cue at a random position in each corridor. Water was available after the sound cue in the rewarded corridor. **b**, Task training timeline in task mice. Mice in the unsupervised cohort experienced the same stimuli or the gratings stimuli without water rewards. **c**, Lick distribution in an example mouse after task learning. Trials were sorted according to the sound cue position. **d**, Performance quantification of anticipatory licks before water was delivered (error bars on all figures represent s.e.m.; the centre values are average; grey lines denote $n = 5$ mice). **e**, Example field of view of the mesoscope (left) and zoomed-in view (right) to illustrate cellular resolution. AL, anterolateral; LM, lateromedial; MMP, mediomedialposterior; RL, rostrolateral; RLL, rostrolateralallateral; RSP, retrosplenial. **f**, Selectivity index d' of neural responses inside the

two corridors. **g**, Single-trial responses of a single circle1-selective neuron as well as the entire population from an example mouse. **h**, Same as panel **g** but for leaf1-selective neurons. **i**, 2D histogram of selective neuron distributions across the field of view, aligned to a map of visual areas. The top and bottom rows are task mice and unsupervised mice, respectively. The left and right columns are before and after learning, respectively. A, anterior; L, lateral; M, medial; P, posterior. **j**, Percentage of neurons with high selectivity in each of the four visual regions defined in the inset ($n = 4$ task mice, $n = 9$ unsupervised mice and $n = 3$ unsupervised with gratings mice (5 sessions)); statistical tests in all figures are two-sided Student's t -tests, paired or independent as appropriate; see Methods for full details. All data are mean $\pm$ s.e.m. * $P < 0.05$, ** $P < 0.01$ and *** $P < 0.001$ from paired, two-sided t -tests.

主要亮点

- **无监督学习主导视觉皮层神经可塑性**: 任务学习引发的神经变化在无奖励暴露组中高度复制, 内侧 HVAs 的神经元选择性增强, 且编码遵循视觉特征 (如 “叶状” “圆形”) 而非空间位置, 表明无监督学习是视觉表征形成的基础机制。
- **前侧 HVAs 的奖励预测信号特异性**: 仅在任务组的前侧 HVAs 中发现随奖励预期增强的神经元活动, 其动态变化先于舔舐行为, 且与视觉刺激的奖励 valence 相关, 揭示监督学习信号的区域特异性。
- **无监督预训练加速行为学习**: 经过自然图像无监督预训练的小鼠, 在后续奖励任务中学习速度显著快于未预训练组或光栅预训练组, 证明无监督学习能优化视觉表征, 提升任务学习效率。

相关链接

Paper: <https://www.nature.com/articles/s41586-025-09180-y.pdf>

● 论文三十六

How much do language models memorize?

日期

2025-06-19

How much do language models memorize?

John X. Morris^{1,3}, Chawin Sitawarin², Chuan Guo¹, Narine Kokhlikyan¹, G. Edward Suh^{3,4}, Alexander M. Rush³, Kamalika Chaudhuri¹, Saeed Mahloujifar¹

¹FAIR at Meta, ²Google DeepMind, ³Cornell University, ⁴NVIDIA

We propose a new method for estimating how much a model “knows” about a datapoint and use it to measure the capacity of modern language models. We formally separate memorization into two components: *unintended memorization*, the information a model contains about a specific dataset, and *generalization*, the information a model contains about the true data-generation process. By eliminating generalization, we can compute the total memorization of a given model, which provides an estimate of model capacity: our measurements estimate that **models in the GPT family have an approximate capacity of 3.6 bits-per-parameter**. We train language models on datasets of increasing size and observe that models memorize until their capacity fills, at which point “grokking” begins, and unintended memorization decreases as models begin to generalize. We train hundreds of transformer language models ranging from 500K to 1.5B parameters and produce a series of scaling laws relating model capacity and data size to membership inference.

Date: June 19, 2025

Correspondence: Saeed Mahloujifar at saeedm@meta.com

∞ Meta

内容提要

文中提出全新方法量化语言模型对数据点的记忆程度，并将“非预期记忆”（memorization）与“泛化”（generalization）明确区分。研究揭示模型在训练后期倾向于不恰当记忆训练集，拼接过拟合与泛化的关系有助判断模型安全性与隐私风险。

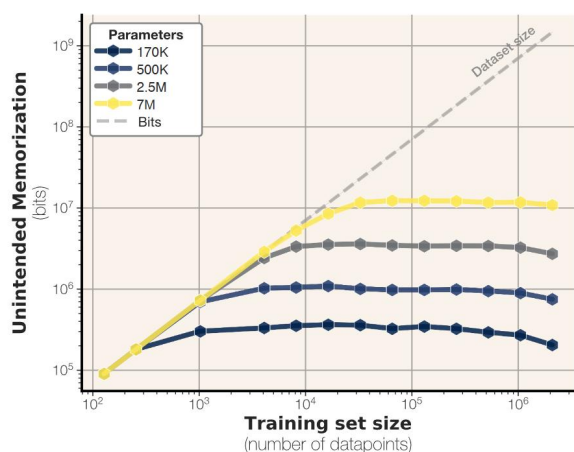


Figure 1 Unintended memorization of uniform random data (Section 3). Memorization plateaus at the empirical capacity limit of different-sized models from the GPT-family, approximately 3.6 bits-per-parameter.

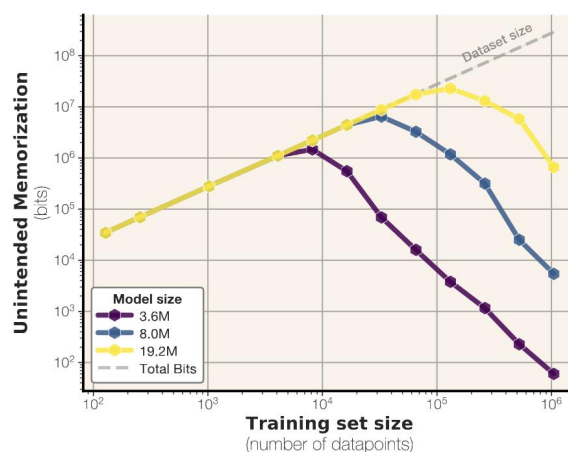


Figure 2 Unintended memorization of text across model and dataset sizes (Section 4). All quantities are calculated with respect to a large oracle model trained on the full data distribution.

主要亮点

- **记忆 vs 泛化 区分清晰**：提出细致测度，帮助理解模型本质记忆行为。
- **新测量方法设计**：通过量化评估示例级别记忆，具备普适适用性。
- **强化模型安全意识**：研究结果对训练框架及隐私防护提供重要参考。

相关链接

Paper: <https://arxiv.org/abs/2505.24832>

● 论文三十七

Drag-and-Drop LLMs: Zero-Shot Prompt-to-Weights

日期

2025-06-19

Drag-and-Drop LLMs: Zero-Shot Prompt-to-Weights

Zhiyuan Liang^{1*} Dongwen Tang¹ Yuhao Zhou¹ Xuanlei Zhao¹ Mingjia Shi¹
Wangbo Zhao¹ Zekai Li¹ Peihao Wang² Konstantin Schürholt³ Damian Borth³
Michael M. Bronstein⁴ Yang You¹ Zhangyang Wang^{2*} Kai Wang^{1*}

¹National University of Singapore, ²UT Austin, ³University of St. Gallen, ⁴Oxford University

内容提要

现代参数高效微调 (PEFT) 方法 (如 LoRA) 虽降低了大语言模型 (LLMs) 的定制成本, 但每个下游任务仍需独立优化。本文提出 Drag-and-Drop LLMs (DnD), 这是一种提示条件参数生成器, 通过将少量未标记任务提示直接映射到 LoRA 权重更新, 消除了逐任务训练需求: 轻量级文本编码器将提示批次蒸馏为条件嵌入, 再通过级联超卷积解码器转换为完整 LoRA 矩阵。在多样化的提示 - 检查点对中训练后, DnD 可在秒级生成任务特定参数, 实现了比完整微调低 12,000 倍的开销, 在未见常识推理、数学、编码和多模态基准上较最强训练 LoRA 平均提升 30% 性能, 且无需目标数据或标签即可实现跨域泛化, 证明提示条件参数生成是替代基于梯度适配的可行方案。

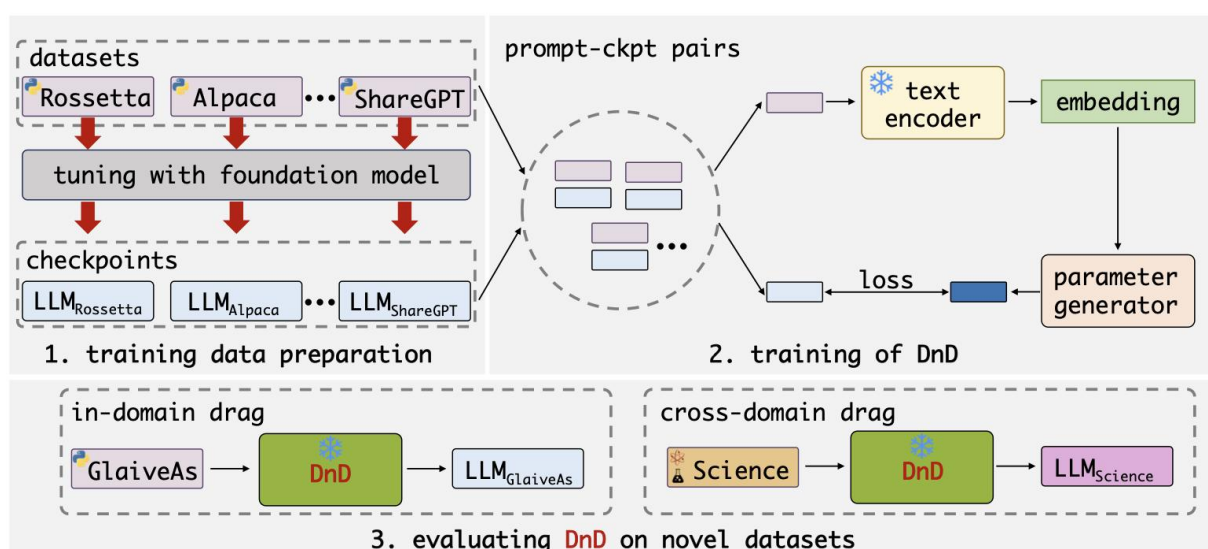


Figure 2: Our approach obtains dragging ability via two processes: prepare training data (*upper left*) and training the parameter generator (*upper right*). When preparing training data, we explicitly pair parameters with dataset-specific conditions. During training, **DnD** takes condition as input and generate parameters, using original parameters as supervision.

主要亮点

- **零样本提示到权重的参数生成框架**：首创 DnD 机制，通过文本编码器与超卷积解码器的级联设计，将任务提示直接转换为 LoRA 权重，无需任何目标任务数据或训练过程，突破传统 PEFT 需逐任务优化的限制。
- **极致效率与性能增益**：相比完整微调实现 12,000 倍的开销降低，参数生成时间压缩至秒级，同时在多领域基准上超越最强训练 LoRA 模型 30% 平均性能，兼顾效率与效果。
- **跨域泛化与零样本适应性**：无需接触目标任务数据或标签，即可通过提示映射实现跨常识推理、数学、编码等领域的参数定制，展现出通用的任务适配能力，为 LLMs 快速专业化提供新范式。

相关链接

Paper: <https://arxiv.org/abs/2506.16406>

● 论文三十八

Machine Mental Imagery: Empower Multimodal Reasoning with Latent Visual Tokens

日期

2025-06-20

Machine Mental Imagery: Empower Multimodal Reasoning with Latent Visual Tokens

Zeyuan Yang^{1†} Xueyang Yu^{1†} Delin Chen¹ Maohao Shen² Chuang Gan¹

¹University of Massachusetts, Amherst ²Massachusetts Institute of Technology

内容提要

视觉语言模型 (VLMs) 虽擅长多模态理解, 但其纯文本解码方式限制了视觉想象任务表现, 而现有图像生成预训练方法又会损害推理能力。受人类通过心理意象 (即构建和操作简化视觉线索而非完整图像) 进行推理的启发, 本文提出 Machine Mental Imagery 框架 Mirage, 通过在 VLM 解码过程中融入潜在视觉令牌 (latent visual tokens) 实现多模态推理 —— 当模型选择 “视觉思考” 时, 将隐藏状态转换为紧凑的视觉嵌入并追加到上下文中, 形成无需像素级图像生成的推理轨迹。训练采用两阶段策略: 首阶段通过蒸馏真实图像嵌入监督潜在令牌, 次阶段仅用文本监督使潜在轨迹与任务目标对齐, 再通过强化学习进一步优化。实验表明, Mirage 在 VSP 空间规划、Blink-Jigsaw 拼图、SAT 空间关系等多基准任务上显著优于纯文本基线及 Anole、MVoT 等统一模型, 证明了无需显式图像生成即可增强多模态推理的能力。

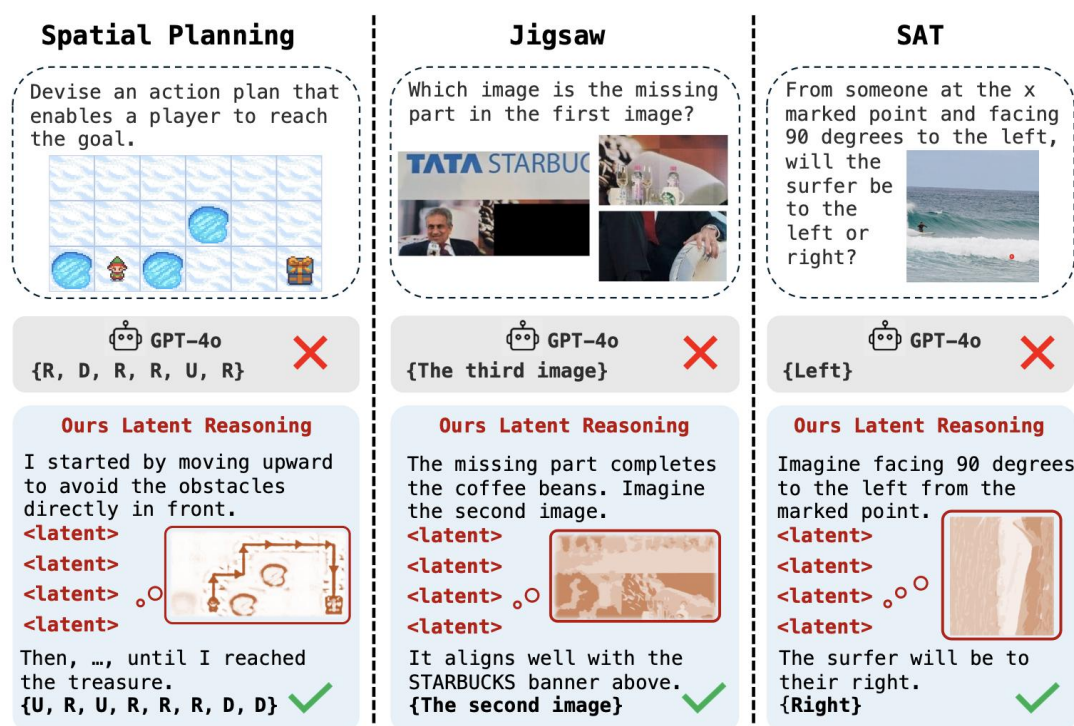


Figure 1: **Multimodal Reasoning Examples.** Mirage interleaves **latent visual tokens**, which represent compact imagery visual features, with explicit text tokens to solve diverse spatial reasoning multimodal tasks, boosting the reasoning performance without the full pixel-level image generation.

主要亮点

- **轻量化多模态推理机制**: 摒弃传统 VLM 依赖像素级图像生成的模式，通过插入压缩的潜在视觉令牌（如平均池化后的图像嵌入）模拟人类心理意象，以低计算成本保留关键视觉语义，使模型专注于推理逻辑而非图像合成。
- **双阶段训练与 RL 协同优化**: 首阶段将潜在令牌锚定至视觉子空间以确保语义准确性，次阶段放松约束使其自适应任务目标，结合 GRPO 强化学习优化推理轨迹，在 VSP 空间规划任务中较纯文本基线提升 11% 准确率，验证了训练策略对推理能力的强化。
- **潜在令牌的语义表征与泛化性**: t-SNE 可视化显示，潜在令牌聚类既贴近真实视觉嵌入又具备任务适应性，在 Qwen2.5-VL 3B 等小模型上仍实现 5%-10% 性能提升，且生成的推理轨迹可解释性强，如在拼图任务中能捕捉边缘连续性等视觉线索。

相关链接

Paper: <https://arxiv.org/pdf/2506.17218>

GitHub: <https://github.com/UMass-Embodied-AGI/Mirage>

Project: <https://vlm-mirage.github.io/>

● 论文三十九

Virtual Community: An Open World for Humans, Robots, and Society

日期

2025-06-21

Virtual Community: An Open World for Humans, Robots, and Society

Qinhong Zhou^{1*} Hongxin Zhang^{1*} Xiangye Lin^{1*} Zheyuan Zhang^{2*}
Yutian Chen³ Wenjun Liu¹ Zunzhe Zhang¹ Sunli Chen¹ Lixing Fang¹
Qiushi Lyu¹ Xinyu Sun¹ Jincheng Yang¹ Zeyuan Wang¹ Bao Chi Dang¹
Zhehuan Chen¹ Daksha Ladia¹ Jiageng Liu¹ Chuang Gan¹
¹UMass Amherst ² Johns Hopkins University ³Carnegie Mellon University

内容提要

本工作设计并实现了一个大型开放式虚拟社区环境 Virtual Community, 支持人类代理与机器人在虚拟空间中进行长期交互、合作与社会行为模拟。系统具备现实地理空间结构、真实社交活动模拟能力, 并结合大语言模型实现个体代理的自主对话与行为规划。每个代理具备独立身份、长期记忆与目标规划能力, 可以参与例如送餐、巡逻、闲聊等多种社会活动。该平台面向开放世界 AI、社会行为建模、多智能体协作等前沿研究提供了高度可控的虚拟试验场。

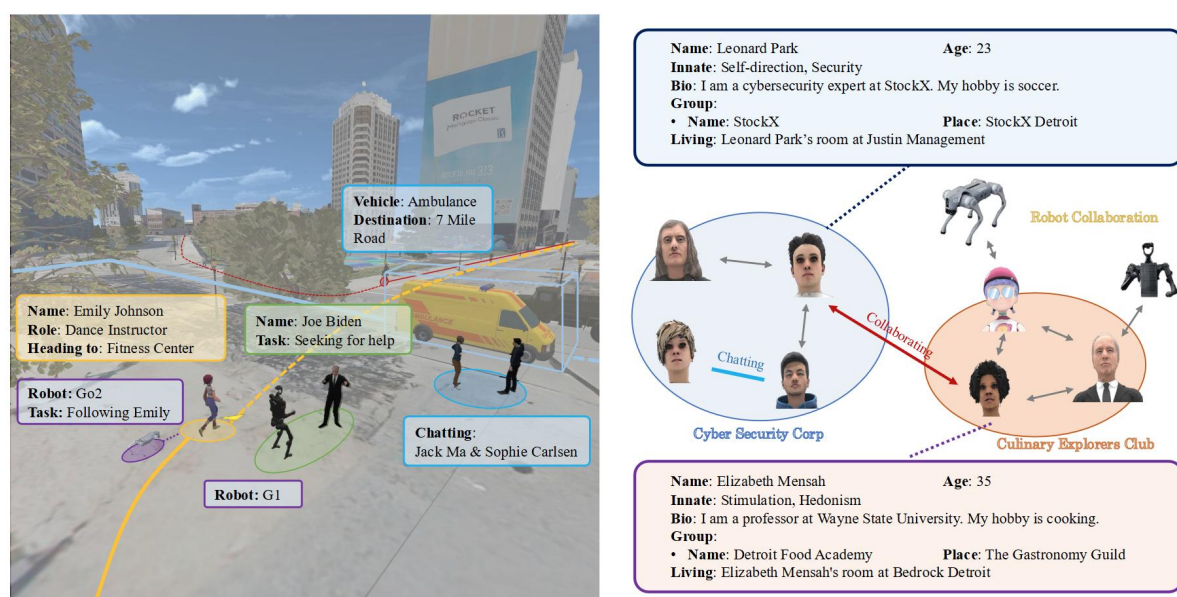


Figure 1: **Virtual Community enables simulation of humans, robots, and their societies within open-world environments.** We present an automated pipeline that transforms real-world geospatial data into large-scale 3D scenes and contextually grounded agent communities. Built on a universal physics engine, our system supports a diverse range of robots—including mobile manipulators, quadrupeds, humanoids, and drones—as well as visually expressive avatars capable of realistic motion and language-based interactions. This foundation facilitates complex social interactions and collective behaviors within the simulated communities.

主要亮点

- **融合真实地理与虚构社交行为：**构建包含地点、事件、社群关系的虚拟城市，提供复杂社会语境。
- **支持多类型智能体协同建模：**不仅有人类代理，还支持机器人、虚拟助手等异质智能体共存。
- **高度可扩展的开放平台：**可接入不同 LLM、RL 代理与感知系统，促进社会智能与人机协作研究。

相关链接

Paper: <https://virtual-community-ai.github.io/paper.pdf>

GitHub: <https://github.com/UMass-Embodied-AGI/Virtual-Community>

Project: <https://virtual-community-ai.github.io/>

● 论文四十

YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception

日期

2025-06-21

YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception

Mengqi Lei, Siqi Li, Yihong Wu, Han Hu, You Zhou, Xinhua Zheng, Guiguang Ding, *Senior Member, IEEE*,
Shaoyi Du, *Member, IEEE*, Zongze Wu, *Member, IEEE*, Yue Gao, *Senior Member, IEEE*

内容提要

YOLO 系列模型在实时目标检测领域因准确性和计算效率表现突出，但 YOLOv11 及更早版本的卷积架构和 YOLOv12 中基于区域的自注意力机制均局限于局部信息聚合和两两相关性建模，无法捕捉全局多对多高阶关联，导致复杂场景下检测性能受限。为此，本文提出 YOLOv13，其核心包括：通过超图自适应关联增强（HyperACE）机制，基于超图计算自适应挖掘潜在高阶关联，突破传统两两关联建模限制，实现高效全局跨位置和跨尺度特征融合与增强；基于 HyperACE 提出全流程聚合 - 分布（FullPAD）范式，将关联增强特征分布至整个网络，实现细粒度信息流与表征协同；利用深度可分离卷积替代普通大核卷积，设计系列模块在不牺牲性能的前提下显著减少参数和计算复杂度。在 MS COCO 基准上的实验表明，YOLOv13 以更少参数和 FLOPs 实现先进性能，其中 YOLOv13-N 较 YOLOv11-N 提升 mAP 3.0%，较 YOLOv12-N 提升 1.5%，代码和模型已开源。

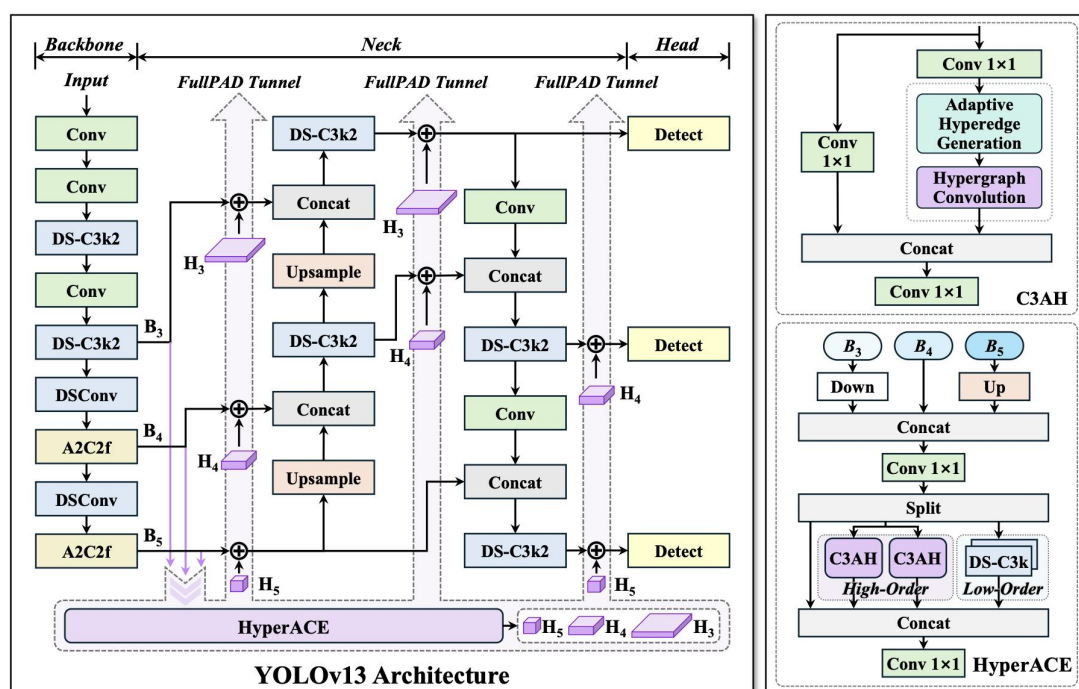


Fig. 2. The network architecture of our proposed YOLOv13 model. Taking multi-scale features extracted from the backbone as input, HyperACE adaptively explores high-order correlations and achieves feature enhancement and fusion. Then, the correlation-enhanced features are distributed to the entire network by FullPAD tunnels to achieve accurate object detection in complex scenes. The detailed structure of HyperACE is shown on the right.

主要亮点

- **超图驱动的高阶关联建模**: 引入 HyperACE 机制, 利用超图理论建模多对多高阶特征关联, 突破传统自注意力的两两关系局限, 实现全局跨尺度特征的高效融合, 尤其提升复杂场景下目标间依赖关系的捕捉能力。
- **全流程特征协同范式**: 提出 FullPAD 范式, 将 HyperACE 增强后的特征分布到网络全流程, 促进不同层级间的信息流动与表征协同, 优化整体检测流程的细粒度信息处理能力。
- **轻量级架构设计**: 采用深度可分离卷积替换大核卷积, 在 YOLOv13 系列模块中实现参数和计算量的大幅缩减 (如 YOLOv13-N 较前代模型参数减少), 同时保持检测精度提升, 平衡实时性与准确性。

相关链接

Paper: <https://arxiv.org/abs/2506.17733>

GitHub: <https://github.com/iMoonLab/yolov13>

● 论文四十一

Deep Research Agents: A Systematic Examination And Roadmap

日期

2025-06-22

DEEP RESEARCH AGENTS: A SYSTEMATIC EXAMINATION AND ROADMAP

Yuxuan Huang^{†1}, Yihang Chen^{†2}, Haozheng Zhang^{†2}, Kang Li³, Meng Fang¹, Linyi Yang⁴,
Xiaoguang Li², Lifeng Shang², Songcen Xu², Jianye Hao², Kun Shao^{†2}, and Jun Wang^{‡4}

¹ University of Liverpool ² Huawei Noah's Ark Lab ³ University of Oxford ⁴ University College London

内容提要

本文深入探讨了 Deep Research (DR) Agents，这是一种新兴的自主 AI 系统，旨在通过动态推理、自适应长周期规划、多跳信息检索、迭代工具使用和结构化分析报告生成，处理复杂的多轮信息研究任务。文章详细分析了构成 DR Agents 的基础技术和架构组件，包括信息获取策略、模块化工具使用框架以及模型上下文协议（MCP）的整合。此外，提出了静态与动态工作流的分类法，并对现有基准进行了批判性评估，指出了其局限性。最后，文章概述了未来研究的开放挑战和有前景的方向，并提供了一个持续更新的 DR Agent 研究仓库。

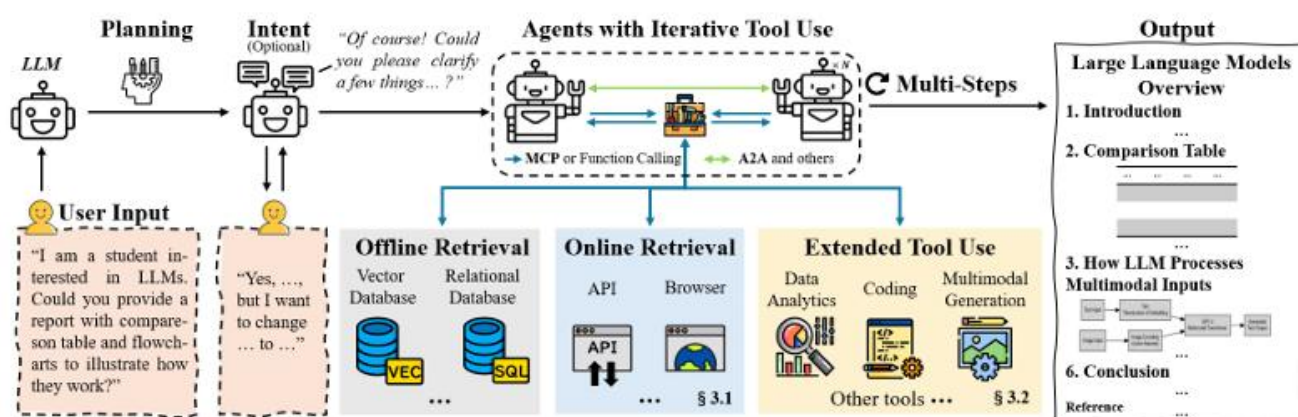


Figure 1: A structural overview of a DR agent in an multi-agent architecture for ease of illustration.

主要亮点

- **系统性分析与分类框架：**文章首次对 DR 代理进行了全面的系统性分析，提出了一个统一的分类框架，将 DR 系统基于 workflow 特性（静态与动态）、规划策略以及代理架构（单代理与多代理）进行分类，为理解和比较不同 DR 系统提供了清晰的结构，有助于推动该领域的标准化和进一步发展。
- **多维度优化方法探讨：**深入探讨了多种优化 DR 代理性能的方法，包括监督式微调（SFT）、强化学习（RL）以及非参数化持续学习等。这些方法从不同的角度提升了 DR 代理的推理能力、决策制定能力和适应性，使其能够更好地处理复杂的多步骤信息检索和合成任务，为 DR 代理的性能提升提供了多样化的技术路径。
- **工业应用与实践案例：**详细介绍了 OpenAI、Google DeepMind、Perplexity 等领先组织开发的 DR 代理的工业应用和实际实现情况，展示了这些系统在实际场景中的应用效果和技术创新，为其他研究者和开发者提供了宝贵的参考和借鉴，有助于加速 DR 代理技术的商业化和实际应用推广。

相关链接

Paper: <https://arxiv.org/abs/2506.18096>

GitHub: <https://github.com/ai-agents-2030/awesome-deep-research-agent>

● 论文四十二

Whole-Body Conditioned Egocentric Video Prediction

日期

2025-06-26

Whole-Body Conditioned Egocentric Video Prediction

Yutong Bai^{* 1}

Danny Tran^{* 1}

Amir Bar^{* 2}

Yann LeCun^{† 2,3}

Trevor Darrell^{† 1}

Jitendra Malik^{† 1,2}

¹UC Berkeley (BAIR)

²FAIR, Meta

³New York University

内容提要

本文提出全身动作条件下的第一人称视频预测模型 PEVA (Predicting Egocentric Video from human Actions)，其通过 3D 身体姿势轨迹作为动作输入，结合自回归条件扩散 Transformer 架构，实现从人体动作到视觉感知的动态模拟。模型在大规模真实数据集 Nymeria 上训练，该数据集包含同步的第一人称视频与全身动作捕捉数据。PEVA 通过分层评估协议验证了其在原子动作预测、长时视频生成及反事实模拟任务中的有效性，性能优于 CDiT 等基线模型，且能通过动作候选模拟支持具身规划。研究表明，全身动作条件显著提升了视频预测的感知质量与语义一致性，为具身智能中的物理动作与视觉体验建模提供了新范式。

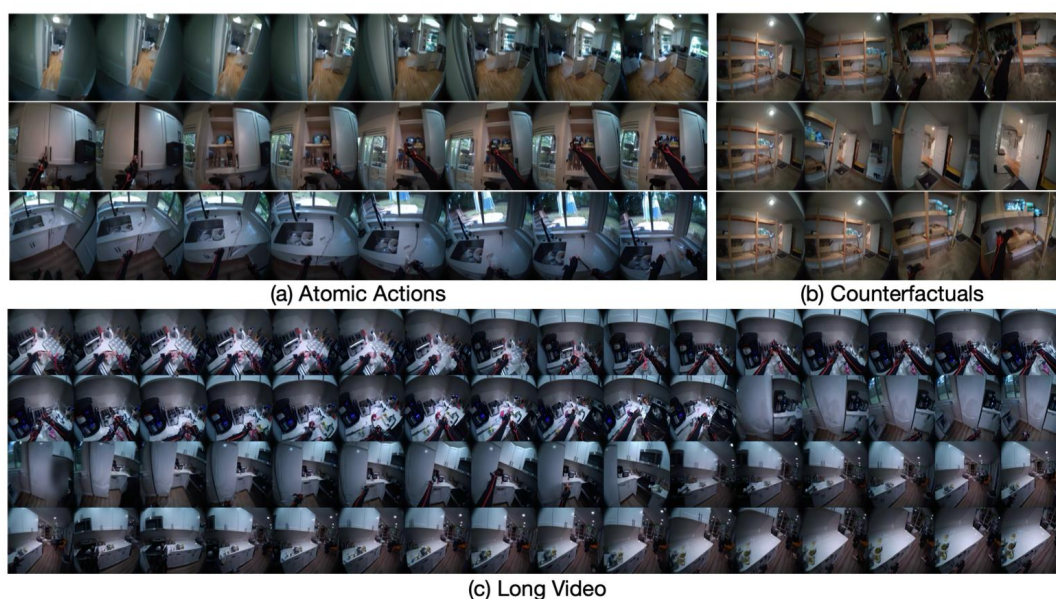


Figure 1: **Predicting Ego-centric Video from human Actions (PEVA)**. Given past video frames and an action specifying a desired change in 3D pose, PEVA predicts the next video frame. Our results show that, given the first frame and a sequence of actions, our model can generate videos of atomic actions (a), simulate counterfactuals (b), and support long video generation (c).

主要亮点

- **结构化全身动作表示**: 将 3D 身体姿势编码为包含全局动态与关节层次的结构化特征，通过骨盆中心局部坐标系转换确保平移旋转不变性，48 维动作向量（3 维根关节平移 + 15 关节 $\times$ 3 维旋转）精准捕捉动作细节。
- **时序扩散 Transformer 架构**: 引入随机时间跳跃（16 帧 / 32 秒窗口）处理长时动态，序列级训练通过因果掩码并行优化多帧依赖，AdaLN 层实现动作条件的高效嵌入。
- **分层评估与规划应用**: 原子动作分解（如手部上下左右、全身前后旋转）验证关节级控制精度，反事实模拟通过 LPIPS 评分筛选最优动作序列，支持目标导向的具身规划。

相关链接

Paper: <https://arxiv.org/pdf/2506.21552>

GitHub: <https://dannytran123.github.io/PEVA>

TECHNICAL UPDATES

技术动态

● 技术动态一

3D 生成模型 Direct3D-S2 全面开源！8 块 GPU 超越闭源登顶 HF

日期

2025-06-03



内容提要

DreamTech 提出了一种高效的高分辨率 3D 生成框架 Direct3D-S2, 该框架基于稀疏体积表示和创新的空
间稀疏注意力机制 (SSA), 实现了大规模 3D 内容的高效生成。与现有方法相比, Direct3D-S2 在生成复杂
细节时保持了更高的质量和效率, 并且能够在仅使用 8 块 GPU 的情况下训练出超越数百块 GPU 训练的闭源商
用模型效果, 训练效率提升了近 20 倍。此外, Direct3D-S2 支持从游戏角色、电影特效到 VR 体验和 3D 打印
产品等多种应用场景, 为影视级 3D 生成提供了新的解决方案。这一技术突破使得高质量 3D 生成更加普及并
具备实际应用价值。

相关链接

Github 地址: <https://github.com/DreamTechAI/Direct3D-S2>

技术详情: <https://huggingface.co/spaces/wushuang98/Direct3D-S2-v1.0-demo>

Paper: <https://arxiv.org/pdf/2505.17412>

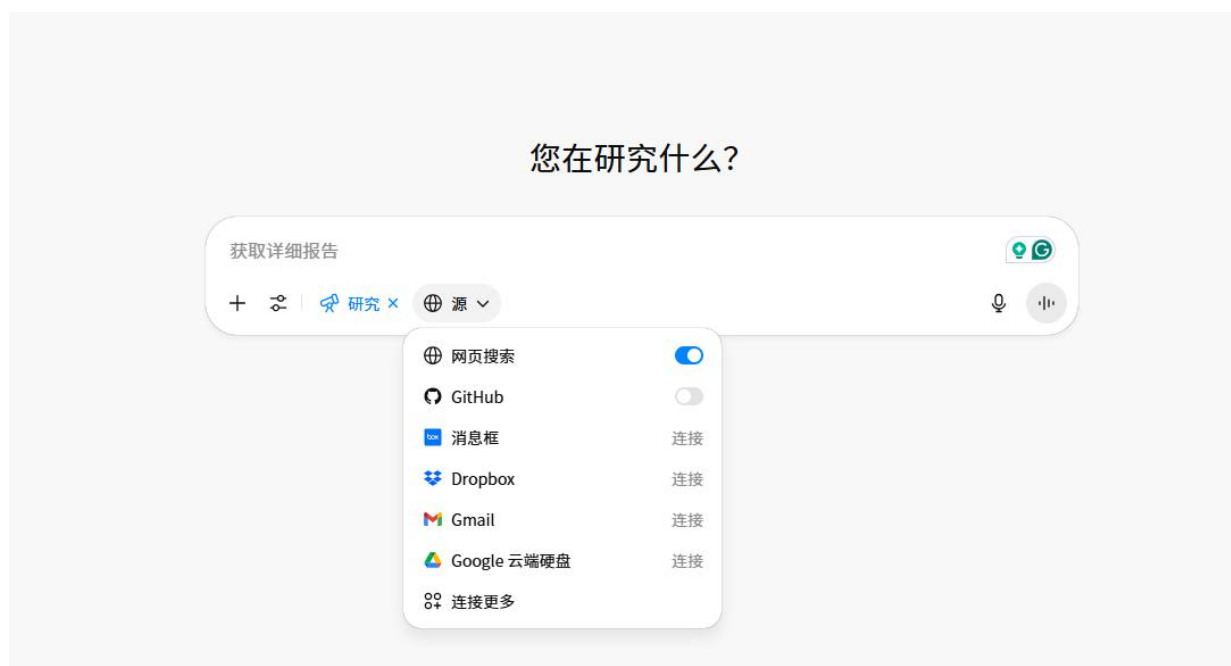
Demo: <https://huggingface.co/spaces/wushuang98/Direct3D-S2-v1.0-demo>

● 技术动态二

OpenAI 深夜宣布 ChatGPT 支持 MCP、会议记录!

日期

2025-06-05



内容提要

OpenAI 为 ChatGPT 推出了两项新功能：会议记录模式和对 MCP 协议的支持。会议记录模式能够在 macOS 上自动录音、记笔记、整理重点，并在会议结束后直接生成会议纪要。通过支持 MCP 协议，ChatGPT 能够连接 GitHub、SharePoint、Gmail 等常用工具，实现跨平台的数据整合、搜索和推理，目标是将 ChatGPT 打造成为一个智能协作平台。此外，还提到了 Deep Research 功能，允许用户在 ChatGPT 中进行广泛的网络搜索和分析。

● 技术动态三

MiniCPM4:在典型端侧芯片上能够实现 5 倍以上生成加速!

日期

2025-06-06

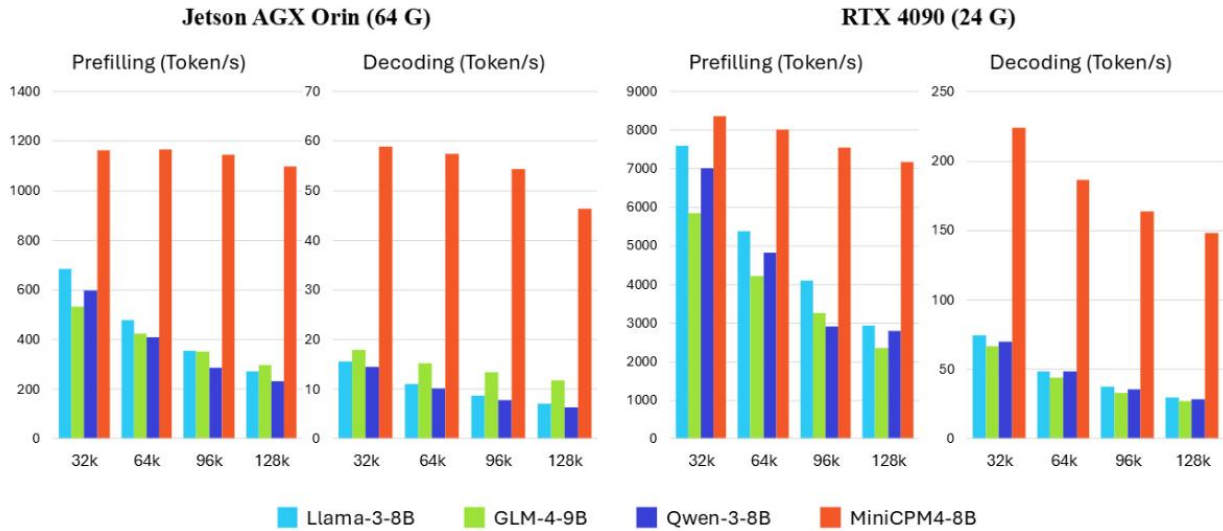


Figure 1: Inference Speed Evaluation on end-side GPUs.

内容提要

MiniCPM4 是一款高效的大型语言模型，专为端侧设备设计。它通过在模型架构、训练数据、训练算法和推理系统四个关键维度上的创新，实现了高性能与低资源消耗的结合。MiniCPM4 提供了 0.5B 和 8B 两种参数版本，在多项基准测试中超越其他开源模型，尤其是在长序列处理方面表现出显著的速度优势。此外，MiniCPM4 还支持多种应用场景，包括可信调查报告生成和基于模型上下文协议的工具使用，展示了其广泛的适用性和高效性。

相关链接

技术报告: https://github.com/OpenBMB/MiniCPM/blob/main/report/MiniCPM_4_Technical_Report.pdf

模型权重: <https://huggingface.co/openbmb/MiniCPM4-8B>

<https://huggingface.co/openbmb/MiniCPM4-0.5B>

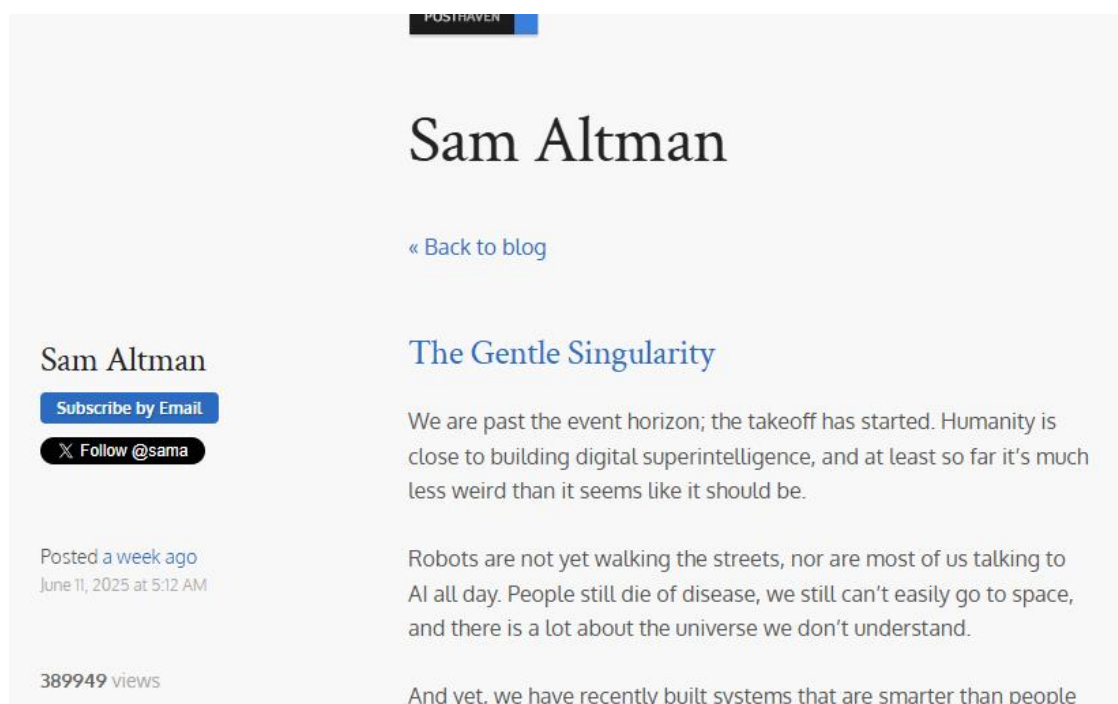
Github 链接: <https://github.com/openbmb/minicpm>

● 技术动态四

OpenAI 创始人 Sam Altman 发表博文, 阐述人类将进入一个以数字超级智能为标志的“温和的奇点”时代。

日期

2025-06-11



内容提要

Sam Altman 在文章《温和的奇点》中描述了人工智能发展的新阶段, 即“温和的奇点”正在逐步发生。他指出, 尽管超级智能的出现尚未完全实现, 但人类已经构建出在许多方面超越个体能力的 AI 系统, 并且这些技术将持续放大人类的能力。随着 AI 驱动科学进步和生产力提升, 未来的生活质量将显著改善, 而机器人、自动化数据中心等技术将进一步加速这一进程。Altman 强调, 在适应快速变化的过程中, 社会需要解决安全问题并广泛分发超级智能以避免集中化风险。这篇文章呼吁尽早展开关于如何定义集体对齐与边界条件的社会讨论, 从而最大化收益并最小化潜在负面影响

相关链接

博客: <https://blog.samaltman.com/the-gentle-singularity>

● 技术动态五

AMD 发布 3nm 怪兽 MI355X，性能碾压英伟达 B200！

日期

2025-06-12



内容提要

AMD 在 2025 年的 Advancing AI 大会上发布了其最新的 AI 芯片产品线，包括基于 3nm 工艺的 MI355X 和即将推出的 MI400 系列 GPU。这些新芯片不仅在性能上实现了显著提升，比如 MI355X 在 FP6 推理精度下的速度是英伟达 B200 的 2.2 倍，而且配备了更大容量的 HBM3e/HBM4 内存。此外，AMD 还介绍了全新的 ROCm 7.0 软件栈以及面向未来的 Helios 机架级基础设施解决方案。通过与 OpenAI 的合作，并采用激进的技术规格和定价策略，AMD 旨在为客户提供高效、低成本的大规模 AI 计算选项。

相关链接

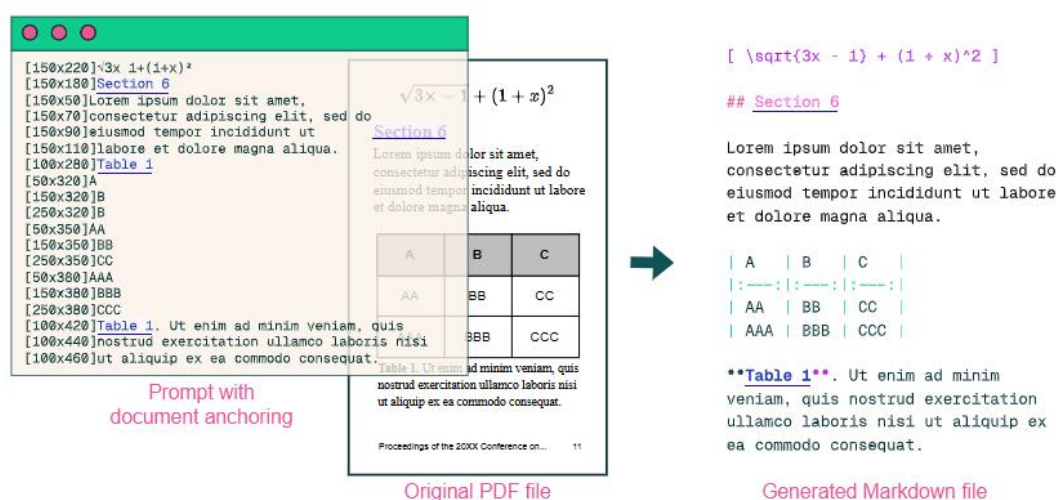
官网：<https://www.amd.com/zh-cn/corporate/events/advancing-ai.html>

● 技术动态六

开源 OCR 工具，可用于将文档（PDF 和图像）转换为可编辑 Markdown 文件

日期

2025-06-17



内容提要

olmOCR 是由艾伦人工智能研究所开发的开源工具包，旨在将 PDF 文档转换为适合语言模型训练的干净纯文本。其核心是通过微调 70 亿参数的视觉语言模型（VLM），结合文档锚定技术，有效处理 PDF 中的复杂布局、表格、公式等结构内容。olmOCR 在自建的 olmOCR-Bench 基准测试中表现优于 GPT-4o、Gemini Flash 2 等商业模型，且成本显著更低，处理百万页仅需 176 美元。该工具包还提供了训练数据、模型权重及高效推理管道，并通过下游实验验证了其提取数据对语言模型性能的提升作用，为大规模 PDF 内容利用提供了经济高效的解决方案。

相关链接

Github: <https://github.com/allenai/olmocr>

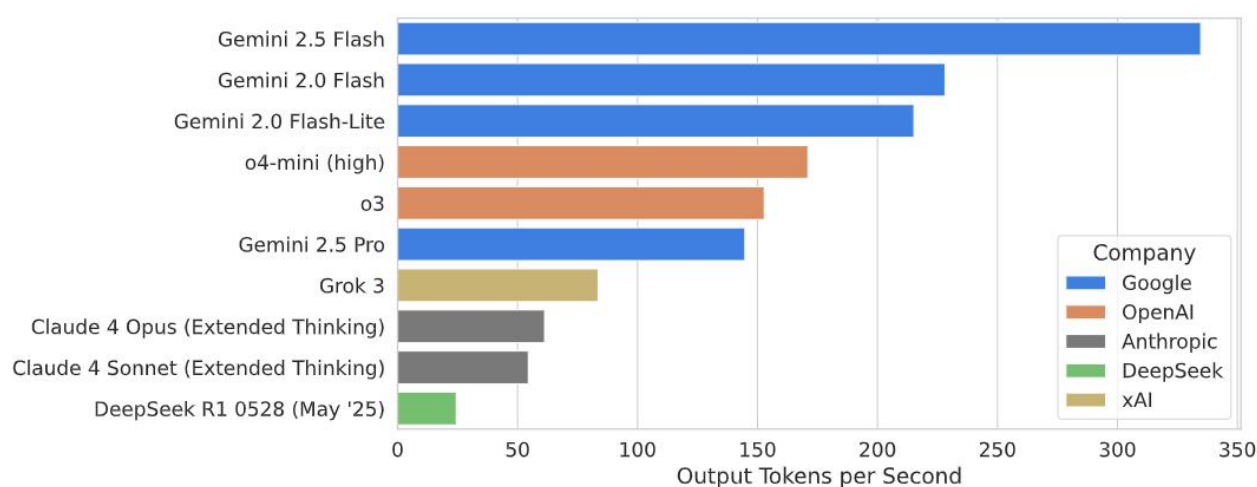
Online demo: <https://olmocr.allenai.org>

● 技术动态七

Google Gemini 2.5 技术报告

日期

2025-06-17



内容提要

Gemini 2.5 是 DeepMind 推出的新一代多模态大模型，具备强大的文本、图像、音频、视频和代码理解与生成能力。相较于前代，Gemini 2.5 在逻辑推理、数学计算和编程任务上取得显著提升，表现出接近甚至超越 GPT-4 的能力。其 Pro 版本支持超过 100 万 token 的上下文窗口，能处理长视频、长文档等复杂输入，并引入“深度思考”（Deep Thinking）机制，在生成答案前主动进行推理；Flash 版本则以极高的效率兼顾性能与推理能力，适用于实时对话和大规模部署场景。此外，Gemini 2.5 原生支持多模态交互和音频输出，具备更强的人机对话自然性与沉浸感，标志着 AI 多模态理解和思维能力迈上新台阶。

相关链接

技术报告：https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf

● 技术动态八

模型即 Agent, Kimi-Researcher (深度研究) 开启内测

日期

2025-06-20



内容提要

Kimi-Researcher 是基于端到端自主强化学习技术的新一代 Agent 模型，现开启小范围灰度测试，它能自主规划研究流程，通过澄清问题、深度推理、主动搜索及工具调用，输出万字以上且信源可溯的深度研究报告与动态可视化报告，在 HLE 等测试中表现超越多个知名模型，覆盖科研、商业、教育等多场景应用，用户可申请内测体验，官方将逐步开源相关模型推动技术探索。

相关链接

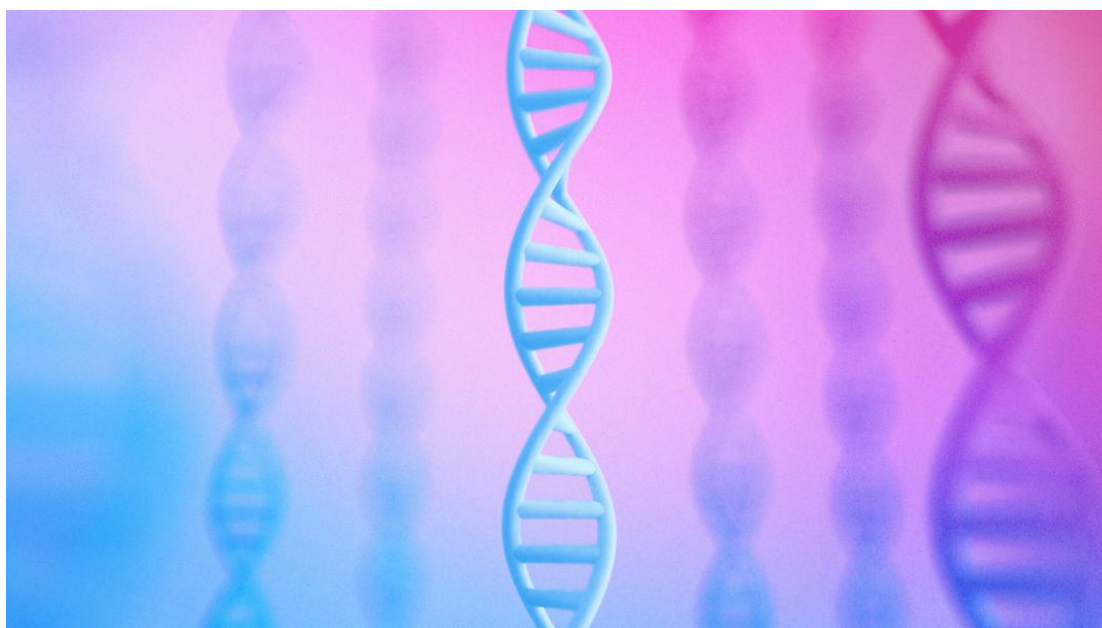
官网：<https://www.kimi.com/>

● 技术动态九

谷歌 AlphaGenome 横空出世！40 亿年生命代码一键破解

日期

2025-06-26



内容提要

Google DeepMind 推出了全新的人工智能工具 AlphaGenome，这是一个统一的 DNA 序列模型，可通过 API 获取，旨在更全面准确地预测人类 DNA 序列中单变体或突变对多种基因调控生物过程的影响。该模型能处理长达 100 万个碱基对的 DNA 序列，输出高分辨率预测，涵盖基因起始与结束位置、RNA 生成量、DNA 碱基可及性等数千种分子特性，还可通过对比突变与未突变序列来评估遗传变异的影响。其技术优势包括兼顾长序列上下文与单碱基分辨率、实现多模态综合预测、高效进行变体评分、首次明确建模剪接连接以及在基因组预测基准上表现出的先进性能。AlphaGenome 基于 Enformer 模型构建，可补充 AlphaMissense 在蛋白质编码区域的变体分类，尤其聚焦占基因组 98% 的非编码区域，助力科学家理解基因组功能、疾病生物学等，目前已面向非商业研究提供 API 预览，未来计划全面发布模型，不过也存在对极远距离调控元件捕捉不足等局限。

相关链接

论文地址：<https://deepmind.google/discover/blog/alphagenome-ai-for-better-understanding-the-genome/>

● 技术动态十

谷歌开源 Gemma 3n：2G 内存就能跑，100 亿参数内最强多模态模型

日期

2025-06-27



内容提要

谷歌于本周五凌晨正式发布并开源全新端侧多模态大模型 Gemma 3n，其代表设备端 AI 重大进步，为手机、平板等端侧设备带来此前仅云端先进模型具备的多模式功能。该模型具备多模态设计，原生支持图像、音频、视频和文本输入及文本输出；专为设备端优化，推出 E2B 和 E4B 两种基于有效参数尺寸，虽原始参数为 5B 和 8B，但架构创新使其运行内存占用与传统 2B 和 4B 型号相当，仅需 2GB 和 3GB 内存即可运行；核心是 MatFormer 架构、每层嵌入（PLE）等全新组件，还采用基于 MobileNet-v5 的视觉编码器等；在多语言、数学、编码和推理等方面质量提升，E4B 版本 LMArena 得分超 1300，成首个达此基准的 100 亿参数以下模型。此外，其 MatFormer 架构支持预提取模型和 Mix-n-Match 自定义尺寸，PLE 技术提升内存效率，KV Cache 共享优化长上下文处理，音频编码器支持 ASR 和 AST，MobileNet-V5 视觉编码器提供先进视觉理解能力，且谷歌首个 Gemma 模型累计下载量已超 1.6 亿次。

相关链接

模型：<https://huggingface.co/collections/google/gemma-3n-685065323f5984ef315c93f4>

文档：<https://ai.google.dev/gemma/docs/gemma-3n>

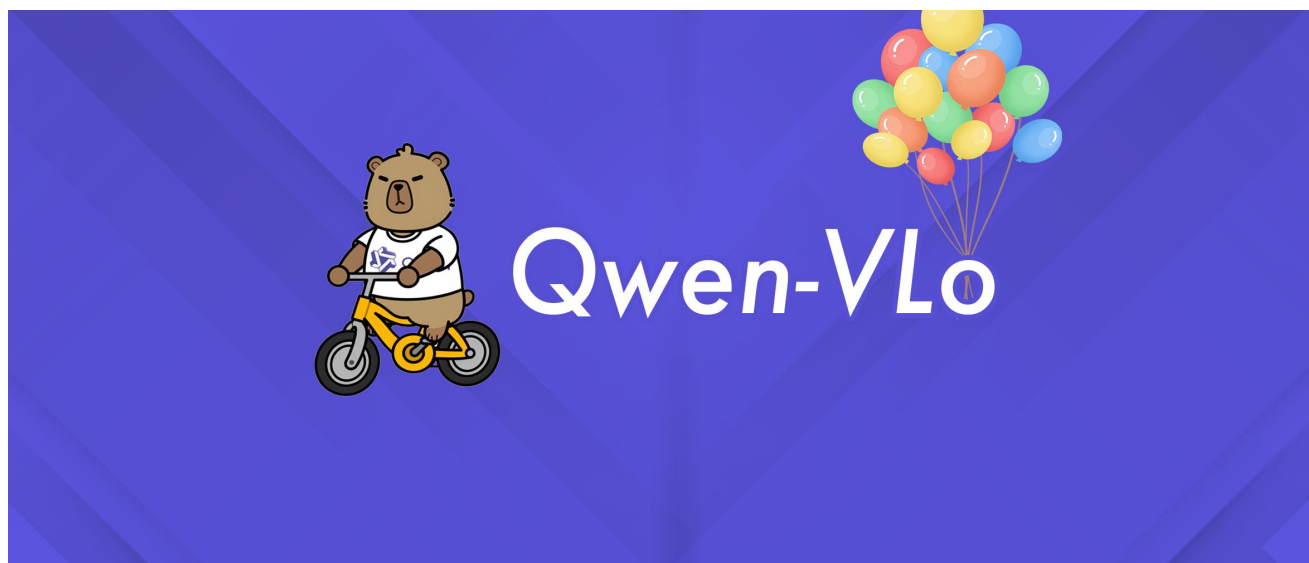
博客：<https://developers.googleblog.com/en/introducing-gemma-3n-developer-guide/>

● 技术动态十一

阿里上新多模态模型 Qwen-VLo

日期

2025-06-27



内容提要

阿里新推出的多模态模型 Qwen-VLo 免费开放预览，其具备增强的细节捕捉能力，能在生成中保持语义一致性，支持通过一句指令完成风格替换、素材增删、文字添加等图像编辑操作，还可处理中英等多语言指令，输入输出不受分辨率和长宽比限制，生成时采用渐进式方式优化细节，实测中在图像生成与编辑上表现出较强能力，用户可直接体验生图、改图等操作。

相关链接

在线体验: <https://chat.qwen.ai/>

博客: <https://qwenlm.github.io/blog/Qwen-VLoo/>

● 技术动态十二

文心 4.5 系列模型，正式开源！

日期

2025-06-30

文心 4.5 系列开源模型			模型信息		
			输入模态	输出模态	上下文窗口
MoE 模型	大语言 模型 (LLMs)	ERNIE-4.5-300B-A47B-Base	Text	Text	128K
		ERNIE-4.5-300B-A47B			
		ERNIE-4.5-21B-A3B-Base			
		ERNIE-4.5-21B-A3B			
	视觉语言 模型 (VLMs)	ERNIE-4.5-VL-424B-A47B-Base	Text Image Video		
		ERNIE-4.5-VL-424B-A47B			
		ERNIE-4.5-VL-28B-A3B-Base			
		ERNIE-4.5-VL-28B-A3B			
稠密型模型		ERNIE-4.5-0.3B-Base	Text	Text	
		ERNIE-4.5-0.3B			

内容提要

文心 4.5 系列模型正式开源，共 10 款，涵盖激活参数规模为 47B 和 3B 的混合专家 (MoE) 模型（最大总参数量 424B）以及 0.3B 的稠密参数模型。该系列模型采用创新性多模态异构模型结构，通过跨模态参数共享机制实现模态间知识融合，使用飞桨深度学习框架训练，模型 FLOPs 利用率达 47%，在多个文本和多模态基准测试中达 SOTA 水平。其核心技术亮点包括多模态混合专家模型预训练、高效训练推理框架及针对模态的后训练。模型权重按 Apache 2.0 协议开源，支持学术研究和产业应用，同时配套发布文心大模型开发套件 ERNIEKit 和大模型高效部署套件 FastDeploy，为开发者提供全流程支持。

相关链接

Hugging Face: <https://huggingface.co/baidu>

GitHub: <https://github.com/PaddlePaddle/ERNIE>

飞桨星河社区: <https://aistudio.baidu.com/overview>

技术报告: <https://yiyan.baidu.com/blog/publication>

RESOURCE SHARING

资源分享

● 资源分享一

何恺明 CVPR 最新讲座 PPT 上线：走向端到端生成建模

类型：讲座材料

内容提要

近年来，扩散模型迅速超越了先前的方法，成为视觉生成建模中的主导方法，广泛应用于图像、视频、3D 物体等的生成。然而，这些模型也存在一些显著的局限性，例如生成速度较慢、生成过程中人类干预有限，以及在模拟复杂分布（如长视频）时面临挑战。何恺明在 CVPR workshop《Visual Generative Modeling: What's After Diffusion?》中做了主题为「Towards End-to-End Generative Modeling（走向端到端生成建模）」的分享，旨在探索视觉生成建模中能够超越扩散模型的方法。

资源链接

https://people.csail.mit.edu/kaiming/cvpr25talk/cvpr2025_meanflow_kaiming.pdf

• 资源分享二

微软开源图标库

类型: github 项目

内容提要

fluentui-system-icons 是微软维护的一个代码仓库, 用于提供 Fluent UI 系统图标。这些图标具有熟悉、友好和现代的特点。仓库中包含了常规图标和填充图标。目前该项目已获得 9.5k+ stars。

资源链接

<https://github.com/microsoft/fluentui-system-icons>

• 资源分享三

Awesome LLM Apps

类型: github 项目

内容提要

该项目精选收录了采用 RAG、AI 智能体、多智能体协作、MCP、语音智能体等技术构建的 LLM 应用汇集, 目前已获超 4.5k+ stars。

资源链接

<https://github.com/Shubhamsaboo/awesome-llm-apps?tab=readme-ov-file>

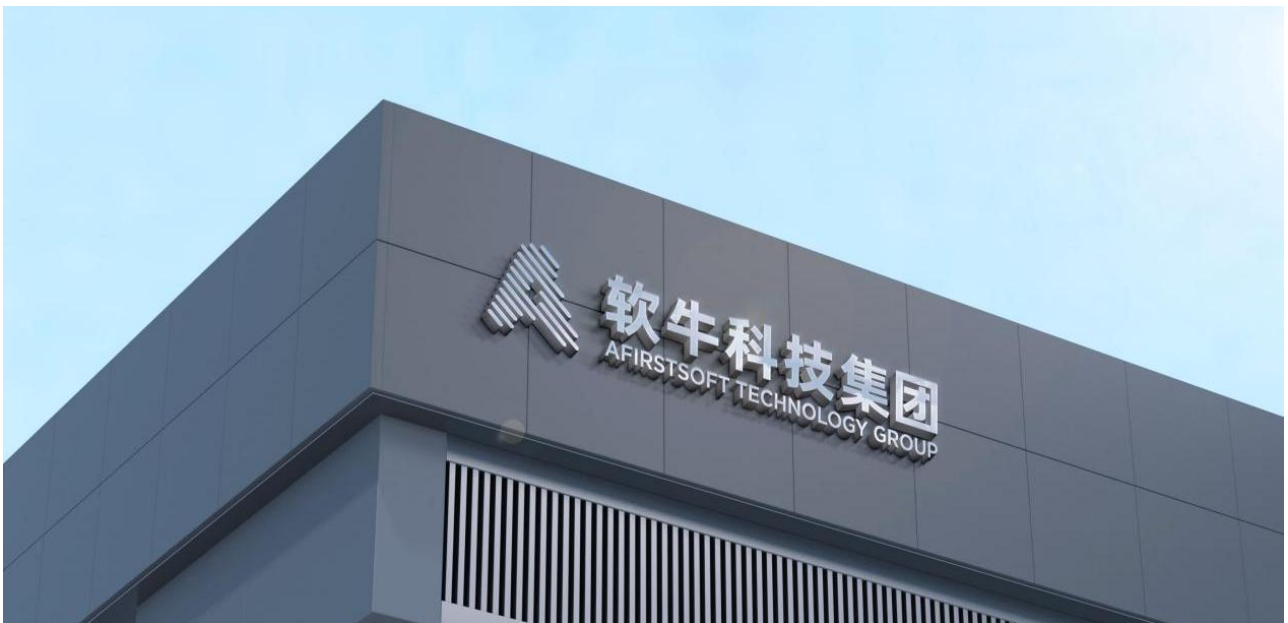


深圳软牛科技集团股份有限公司成立于 2014 年 6 月，作为国家级专精特新“小巨人”企业，始终以“让创意成为现实”为使命，专注于通过人工智能视觉大模型技术驱动全球数字创意生态。

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



2025 年第 06 期

总第 24 期

广东省图象图形学会

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
郑泽文	金山办公
胡晋武	华南理工大学
王宇丰	华南理工大学
王骞玥	华南理工大学
李成昊	华南理工大学
刘祎然	华南理工大学
许毅平	华南农业大学

人工智能 前沿快报

2025 年第 06 期

总第 24 期

— 2025.07.02

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，文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、科普宣传、科技成果鉴定