研究所
分析师：肖承志
SAC 登记编号：S1340524090001
Email：xiaochengzhi@cnpsec.com
研究助理：冯昱文
SAC 登记编号：S1340124100011
Email：fengyuwen@cnpsec.com

近期研究报告

《成长风格显著，中盘表现占优——中邮因子周报 20250817》 - 2025.08.18
《融资余额新高，创新药光通信调整，指数预期仍将震荡上行挑战前高——行业轮动周报 20250810》 - 2025.08.11
《基本面因子表现不佳，小盘风格明显——中邮因子周报 20250803》 - 2025.08.04
《小市值占优，低波反转显著——中邮因子周报 20250727》 - 2025.07.28
《微盘股的流动性风险在哪？——微盘股指数周报 20250720》 - 2025.07.21
《大金融表现居前助指数突破，GRU 行业轮动调入非银行金融——行业轮动周报 20250713》 - 2025.07.14
《低估值高盈利，基本面表现占优——中邮因子周报 20250706》 - 2025.07.07
《基于宏观经济状态划分的 BL 模型与 ETF 实践》 - 2025.07.01
《反转风格显著，小市值回撤——中邮因子周报 20250622》 - 2025.06.23
《关注基本面支撑，高波风格占优——中邮因子周报 20250615》 - 2025.06.16
《结合基本面和量价特征的 GRU 模型》 - 2025.06.05

金工周报

智元推出机器人世界模型平台 Genie Envisioner，智谱上线 GLM-4.5V 视觉推理模型——AI 动态汇总 20250818

- **智元推出机器人世界模型平台 Genie Envisioner**

智元机器人于 2025 年 7 月 27 日在 WAIC 2025“智启具身论坛”上正式发布了行业首个动作驱动世界模型开源平台——Genie Envisioner（简称 GE），并于 8 月 14 日进一步向公众推出面向真实世界机器人操控的统一世界模型平台。这一平台彻底颠覆了传统机器人学习中“数据-训练-评估”割裂的流水线模式，创新性地构建了一个以视频生成为核心的闭环架构，使机器人能够在同一世界模型中完成从视觉感知到动作执行的端到端推理与执行。

- **智谱上线 GLM-4.5V 视觉推理模型**

智谱 AI 于 2025 年 8 月 11 日正式发布并开源了新一代视觉推理模型 GLM-4.5V，该模型以 1060 亿总参数和 120 亿激活参数的规模成为全球 100B 级效果最佳的开源视觉推理模型，同步在 GitHub、Hugging Face 及魔搭社区开放下载。

- **字节 Seed 团队开源 VeOmni 全模态训练框架**

字节跳动 Seed 团队于 2025 年 8 月 14 日正式开源的全模态 PyTorch 原生训练框架 VeOmni，标志着多模态大模型训练进入“低摩擦时代”。该框架通过“以模型为中心”的分布式设计理念，系统性解决了传统训练方法在工程复杂度、扩展性和效率上的瓶颈，将全模态模型的研发周期从数周缩短至几天，工程耗时降低 90%以上，同时在 128 卡 GPU 集群上实现 300 亿参数 MoE 模型 2800 tokens/sec/GPU 的吞吐量，支持高达 160K 超长上下文序列训练。

- **昆仑万维开源多模态框架 Skywork UniPic 2.0**

昆仑万维于 2025 年 8 月 13 日在 SkyWork AI 技术发布周上正式开源了 Skywork UniPic 2.0，这是一款突破性的统一多模态框架，首次在单一模型中深度融合图像理解、文本到图像生成（T2I）和图像编辑（I2I）三大核心能力。该模型基于 2B 参数的 SD3.5-Medium 架构，通过创新的渐进式双任务强化策略和轻量化设计，实现了生成质量与部署效率的双重突破，其性能超越多个 12B 以上参数的同类模型，成为开源多模态领域的新标杆。

- **风险提示：**

以上内容基于历史数据完成，在政策、市场环境发生变化时存在失效的风险；历史信息不代表未来。

目录

图表目录

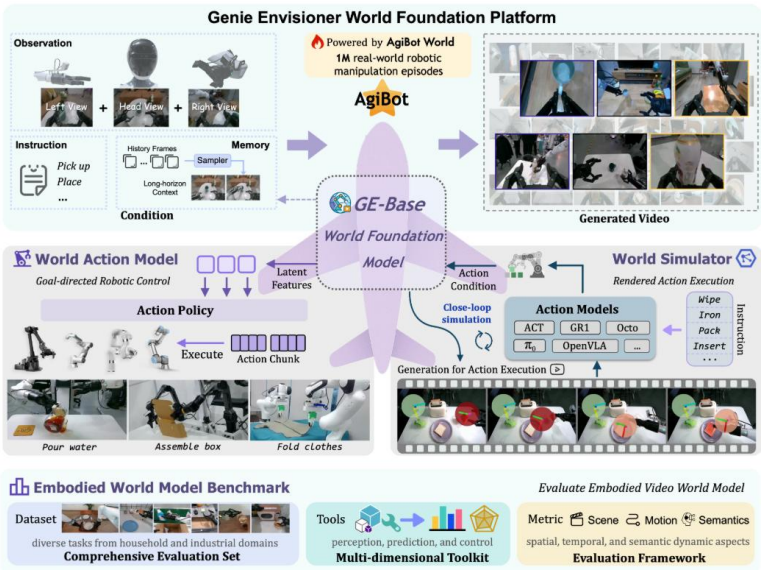
1 AI 重点要闻

1.1 智元推出机器人世界模型平台 Genie Envisioner

智元机器人于 2025 年 7 月 27 日在 WAIC 2025“智启具身论坛”上正式发布了行业首个动作驱动世界模型开源平台——Genie Envisioner（简称 GE），并于 8 月 14 日进一步向公众推出面向真实世界机器人操控的统一世界模型平台。这一平台彻底颠覆了传统机器人学习中“数据-训练-评估”割裂的流水线模式，创新性地构建了一个以视频生成为核心的闭环架构，使机器人能够在同一世界模型中完成从视觉感知到动作执行的端到端推理与执行。

GE 平台的核心突破在于其视觉中心的世界建模范式。不同于主流 VLA 方法依赖视觉-语言模型将视觉输入映射到语言空间进行间接建模，GE 直接在视觉空间中建模机器人与环境的交互动态。这种方法完整保留了操控过程中的空间结构和时序演化信息，实现了对机器人-环境动态更精确、更直接的建模。这一范式带来了两大关键优势：高效的跨本体泛化能力和长时序任务的精确执行能力。在跨平台测试中，GE-Act 仅需 1 小时约 250 个演示的遥操作数据即可在新机器人平台上实现高质量任务执行，远超同类模型；而在折叠纸盒等超长步骤任务中，其成功率高达 76%，显著优于专门优化的 π0 模型 48%的表现。

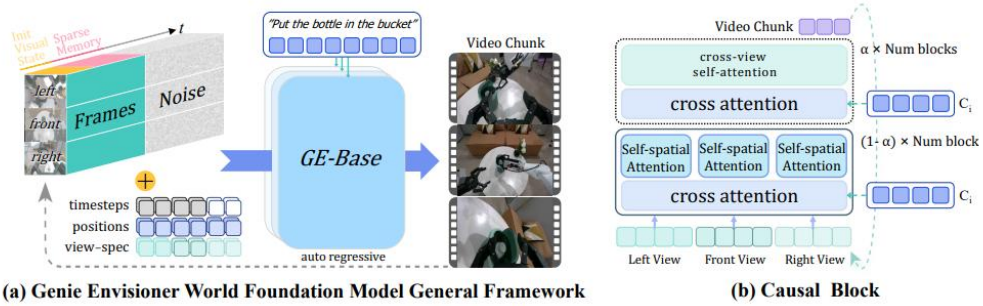
图表1：Genie Envisioner 平台概览



资料来源：AGIBOT，中邮证券研究所

技术架构上，GE 平台由三大核心组件紧密集成。GE-Base 作为多视角视频世界基础模型，采用自回归视频生成框架，通过头部相机和双臂腕部相机的三路视角输入保持空间一致性，并利用稀疏记忆机制增强长时序推理能力。其训练基于 AgiBot-World-Beta 数据集的 3000 小时超 100 万条真机数据，使用 32 块 A100 GPU 耗时约 10 天完成。GE-Act 作为 160M 参数的轻量级动作解码器，采用与 GE-Base 平行的流匹配设计，通过异步推理模式实现实时控制——视频 DiT 以 5Hz 运行，动作模型以 30Hz 运行，可在 RTX 4090 GPU 上以 200 毫秒完成 54 步动作推理。GE-Sim 则作为层次化动作条件仿真器，通过 Pose2Image 条件和运动向量机制将低层控制指令转换为精确的视觉预测，支持闭环策略评估和大规模数据生成。

图表2：GE-Base 世界基础模型概述

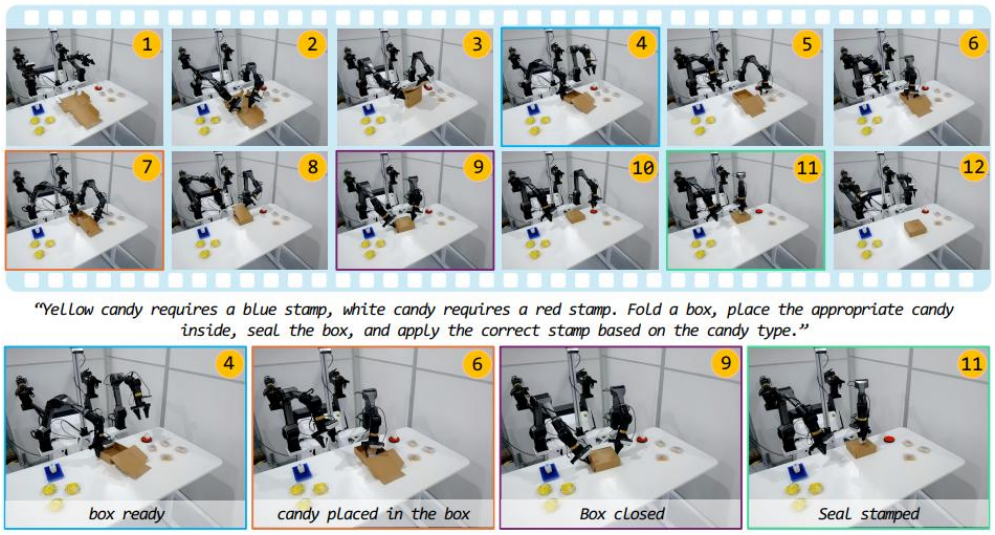


(a) Genie Envisioner World Foundation Model General Framework　　(b) Causal Block

资料来源：AGIBOT，中邮证券研究所

在实际应用中，搭载 GE-Act 的机器人已能流畅完成制作三明治、倒茶、擦桌面、使用微波炉加热、流水线装箱等多项任务，成功率远超行业平均水平。例如在工业场景中，机器人可基于 GE 的预测能力提前模拟装配过程并优化策略，实现从固定轨迹到自主决策的跨越。这种性能提升源于平台对 3000 小时真实机器人操控视频数据的深度利用，这些数据建立了从语言指令到视觉空间的直接映射，完整保留了交互的时空信息。

图表3：在预训练期间未见过的新型机器人 Agilex Cobot Magic 上对 GE-Act 进行了真实演示



资料来源：AGIBOT，中邮证券研究所

智元机器人宣布将开源全部代码、预训练模型和评测工具，并计划未来扩展更多传感器模态以支持全身移动与人机协作。团队还开发了 EWMBench 评测套件，从场景一致性、轨迹精度等多维度评估世界模型质量。在与 Kling、Hailuo 等先进模型的对比中，GE-Base 在关键指标上均取得最优成绩。这一平台的发布不仅为具身智能开辟了从视觉理解到动作执行的新技术路径，更标志着机器人从被动执行向主动“想象-验证-行动”的智能转变，有望推动智能制造与服务机器人的大规模落地。

1.2 智谱上线 GLM-4.5V 视觉推理模型

智谱 AI 于 2025 年 8 月 11 日正式发布并开源了新一代视觉推理模型 GLM-4.5V，该模型以 1060 亿总参数和 120 亿激活参数的规模成为全球 100B 级效果最佳的开源视觉推理模型，同步在 GitHub、Hugging Face 及魔搭社区开放下载。作为智谱通向通用人工智能（AGI）道路上的重要探索，GLM-4.5V 基于旗舰文本基座 GLM-4.5-Air 构建，延续 GLM-4.1V-Thinking 技术路线，在 41 项公开视觉

多模态基准测试中全面达到同级别开源模型的 SOTA 性能，涵盖图像、视频、文档理解以及 GUI Agent 等核心任务场景。

图表4：GLM-4.5V 测评

Benchmarks		GLM-4.5V	Step-3	Qwen2.5-VL	GLM-4.1V	Kimi-VL-2506	Gemma-3
Size		106B (A12B)	321B (A38B)	72B	9B	16B (A3B)	27B
Mode		w/ thinking	w/ thinking	w/o thinking	w/ thinking	w/ thinking	w/o thinking
General VQA	MMBench v1.1	**88.2**	81.1*	88.0	85.8	84.4	80.1*
	MMBench v1.1 (CN)	**88.3**	81.5*	86.7*	84.7	80.7*	84.8*
	MMStar	**75.3**	69.0*	70.8	72.9	70.4	60.0*
	BLINK (val)	**65.3**	62.7*	58.0*	65.1	53.5*	52.9*
	MUIRBENCH	**75.3**	75.0*	62.9*	74.7	63.8*	50.3*
	HallusionBench	**65.4**	64.2	56.8*	63.2	59.8*	45.8*
	ZeroBench (sub)	**23.4**	23.0	19.5*	19.2	16.2*	17.7*
	GeoBench	**79.7**	72.9	74.3*	76.0	48.0*	57.5*
STEM	MMMU (val)	**75.4**	74.2	70.2	68.0	64.0	62.0*
	MMMU Pro	**65.2**	58.6	51.1	57.1	46.3	37.4*
	MathVista	**84.6**	79.2*	74.8	80.7	80.1	64.3*
	MathVision	**65.6**	64.8	38.1	54.4	54.4*	39.8*
	MathVerse	**72.1**	62.7*	47.8*	68.4	54.6*	34.0*
	DynaMath	**53.9**	50.1	36.1*	42.5	28.1*	28.5*
	LogicVista	**62.4**	60.2*	56.2*	60.4	51.4*	47.3*
	AI2D	**88.1**	83.7*	87.6*	87.9	81.9*	80.2*
	WeMath	**68.8**	59.8	46.0*	63.8	42.0*	37.9*
Long Document	MMLongBench-Doc	**44.7**	31.8*	35.2*	42.4	42.1	28.4*
OCR & Chart	OCRBench	86.5	83.7	85.1*	84.2	**86.9**	75.9*
	ChartQAPro	**64.0**	56.4	46.7*	59.5	23.7*	37.6*
	ChartMuseum	**55.3**	40.0*	39.6*	48.8	33.6*	23.9*
Visual Grounding	RefCOCO-avg (val)	**91.3**	20.2*	90.3	85.3	33.6*	2.4*
	TreeBench	**50.1**	41.3*	42.3	37.5	41.5*	33.8*
	Ref-L4-test	**89.5**	12.2*	80.8*	86.8	51.3*	2.5*
Spatial Reco &	OmniSpatial	**51.0**	47.0*	47.9	47.7	37.3*	40.8*
Reasoning	CV-Bench	**87.3**	80.9*	82.0*	85.0	79.1*	74.6*
	ERQA	**50.0**	44.5*	44.8*	45.8	36.0*	37.5*
	All-Angles Bench	**56.9**	52.4*	54.4*	52.7	48.9*	48.2*
GUI Agents	OSWorld	**35.8**	/	8.8	14.9	8.2	6.2*
	AndroidWorld	**57.0**	/	35.0	41.7	/	4.4*
	WebVoyagerSom	**84.4**	/	40.4*	69.0	/	34.8*
	Webquest-SingleQA	**76.9**	58.7*	60.5*	72.1	35.6*	31.2*
	Webquest-MultiQA	**60.6**	52.8*	52.1*	54.7	11.1*	36.5*
Coding	Design2Code	**82.2**	34.1	41.9*	64.7	38.8	16.1
	Flame-React-Eval	**82.5**	63.8	46.3*	72.5	36.3	27.5
Video Understanding	VideoMME (w/o sub)	**74.6**	/	73.3	68.2	67.8	58.9*
	VideoMME (w/ sub)	**80.7**	/	79.1	73.6	71.9	68.4*
	MMVU	**68.7**	/	62.9	59.4	57.5	57.7*
	VideoMMMU	**72.4**	/	60.2	61.0	65.2	54.5*
	LVBench	**53.8**	/	47.3	44.0	47.6*	45.9*
	MotionBench	**62.4**	/	56.1*	59.0	54.3*	47.8*
	MVBench	**73.0**	/	70.4	68.4	59.7*	43.5*

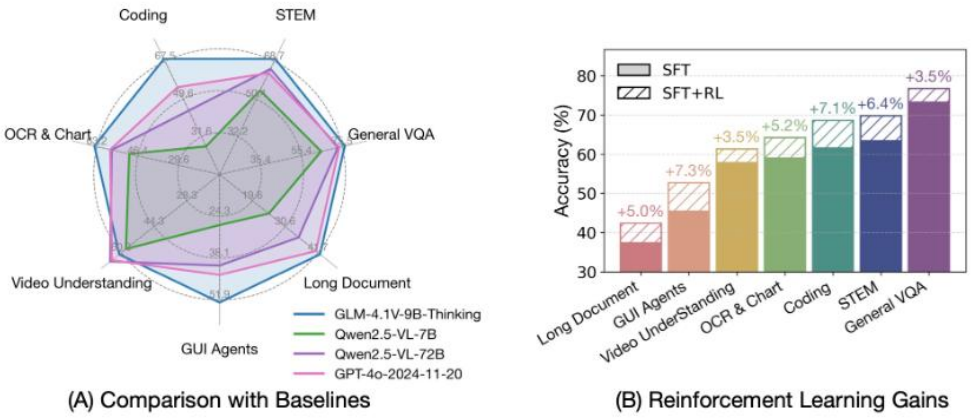
* denotes results reproduced in our experiments.

资料来源：智谱 AI，中邮证券研究所

技术架构上，GLM-4.5V 采用视觉编码器、MLP 适配器与语言解码器的三模块设计，创新性地引入三维旋转位置编码（3D-RoPE）和双三次插值机制，显著增强了对多模态信息中三维空间关系的感知能力以及对高分辨率、极端宽高比图像的适应性。其 64K 多模态长上下文支持能力结合三维卷积优化的视频处理效率，使模型能够同步解析图像与视频输入，在复杂场景下保持稳健表现。例如在“图寻游戏”全球定位赛中，GLM-4.5V 仅用 16 小时便击败 99%的人类玩家，通过分析植被特征、建筑风格等视觉线索精准推测拍摄地经纬度，最终攀升至全球排名第 66 位。

训练策略方面，模型采用三阶段优化流程：预训练阶段利用大规模图文交错多模态语料强化基础理解能力；监督微调阶段引入显式“思维链”样本提升因果推理深度；强化学习阶段则通过多领域奖励系统结合 RLVR 与 RLHF 技术，在 STEM 问题、多模态定位等任务上实现全面突破。这种分层训练使模型展现出超越传统 OCR+文本模型联合方案的文档处理能力——能够以人类阅读方式同步解析数十页含复杂图表的文档，避免信息传递错误，并在技术报告翻译与亮点提取等任务中生成带有自主观点的输出。

图表5：GLM-4.5V 能力对比及 RL 提升



资料来源：智谱 AI，中邮证券研究所

实际应用层面，GLM-4.5V 的涌现能力尤为引人注目。其前端复刻功能可仅凭网页截图或交互视频生成结构化的 HTML/CSS/JavaScript 代码，精准还原布局逻辑与交互意图。例如对知乎网页录屏的分析中，模型不仅复现了静态元素，还通过视频帧间动态变化建模实现了交互功能还原。此外，模型内置的 GUI Agent 能力可识别电商页面中的折扣信息并计算最优方案，为智能体应用开发奠定基础。为降低开发者门槛，智谱同步开源了桌面助手应用，支持实时截屏录屏交互，处理代码辅助、游戏解答等多样化视觉任务。

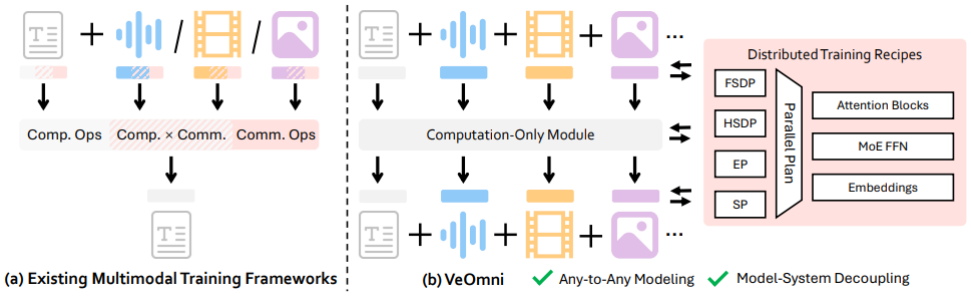
商业化部署上，GLM-4.5V 以高性价比著称，API 调用价格低至输入 2 元/百万 tokens、输出 6 元/百万 tokens，响应速度达 60-80 tokens/s。智谱还为开发者提供 2000 万 tokens 免费资源包，加速多模态应用落地。这一开源举措不仅填补了大参数多模态模型的技术空白，更推动 AI 从“跑分竞赛”向实用化转型，

其全场景视觉推理能力正在重塑智能制造、地理信息服务、人机交互等领域的创新边界。

1.3 字节 Seed 团队开源 VeOmni 全模态训练框架

字节跳动 Seed 团队于 2025 年 8 月 14 日正式开源的全模态 PyTorch 原生训练框架 VeOmni，标志着多模态大模型训练进入“低摩擦时代”。该框架通过“以模型为中心”的分布式设计理念，系统性解决了传统训练方法在工程复杂度、扩展性和效率上的瓶颈，将全模态模型的研发周期从数周缩短至几天，工程耗时降低 90%以上，同时在 128 卡 GPU 集群上实现 300 亿参数 MoE 模型 2800 tokens/sec/GPU 的吞吐量，支持高达 160K 超长上下文序列训练。

图表6：VeOmni 与现有框架对比



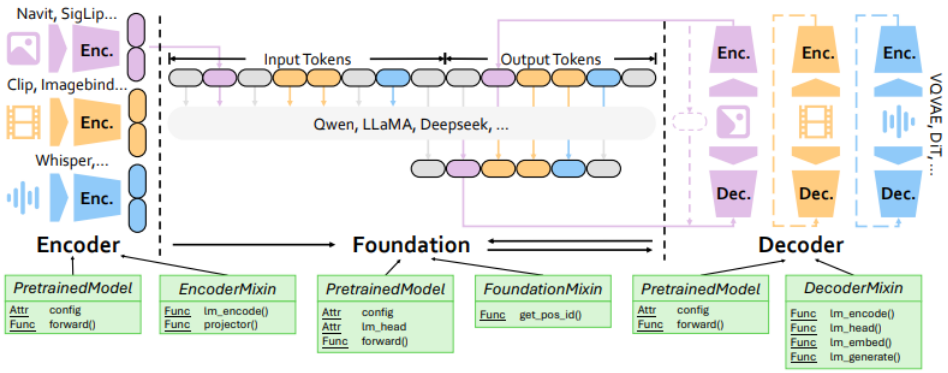
资料来源：字节跳动，中邮证券研究所

VeOmni 的核心突破在于彻底解耦模型定义与分布式并行逻辑，构建了模块化、可组合的训练范式。传统框架如 Megatron-LM 采用“以系统为中心”的设计，将张量并行、流水线并行等策略硬编码至模型结构中，导致新增模态或调整架构需重写底层代码，工程成本高昂。VeOmni 则通过 DeviceMesh 抽象层统一管理并行状态，将模型组件转化为纯粹的“计算模块”，所有通信逻辑由框架自动处理。例如研究者只需通过 API 声明并行维度（如 mesh=[[0,1],[2,3]]），即可实现视觉编码器使用 FSDP（全分片数据并行）、语言模型注意力层采用 HSDP+SP（混合分片数据并行+序列并行）、MoE 层应用 EP（专家并行）的 3D 组合策略，无需侵

入式修改模型代码。这种设计使得在 Qwen2-VL-7B 模型上支持 192K 序列长度成为可能，显存占用较 TorchTitan 降低 15%，吞吐量提升 25%。

框架通过多层次优化实现效率飞跃。在并行策略上，VeOmni 创新性地引入异步序列并行（Async-Ulysses），将 All-to-All 通信与 Attention 计算重叠，使 64K 序列训练的 MFU（模型浮点利用率）达到 43.98%，远超 TorchTitan 的 OOM（内存溢出）表现。针对 MoE 模型的专家并行，团队开发的 COMET 技术实现细粒度计算-通信重叠，通过动态分析 token 路由路径，将通信延迟隐藏在专家计算过程中，使 30B 参数 MoE 模型在 160K 序列下的训练吞吐量保持线性扩展。系统级优化同样关键：动态批处理技术配合 FlashAttention 减少填充浪费，访存次数下降百倍；ROI 驱动的算子级重计算策略选择性激活性价比最高的操作（如 gate1_mul 替代 down_proj），显存回收效率提升 22 倍；ByteCheckpoint 系统支持千卡集群中断恢复时间缩短 90%。

图表7：omni-modal LLM 复合结构



资料来源：字节跳动，中邮证券研究所

VeOmni 已在字节跳动内部多个前沿项目中验证其价值。多模态智能体 UI-TARS-1.5 借助其序列并行能力，解决了>128K 长序列数据的显存瓶颈；同传模型 Seed LiveInterpret 2.0 利用框架集成的轨迹蒸馏与分布匹配蒸馏(DMD)技术，将推理步数压缩至 4-8 步，成本降低 70%。在标准化测试中，框架展现出全面优势:72B 参数模型在 128 卡集群上实现 96K 序列训练,MFU 达 54.82%;与 TorchTitan 对比，16K 序列场景下吞吐量提升 29.82%（532 vs 394 tokens/sec），64K 序列时 VeOmni 仍稳定运行而 TorchTitan 出现 OOM。收敛性研究覆盖 Janus、

LLaMA#Omni 等异构架构，使用 FineWeb-100T、ShareGPT4V 等真实数据集时，文本与图像生成 Loss 均呈现稳定下降趋势，证明框架对复杂模态组合的鲁棒性。

VeOmni 的开源填补了工业级任意模态训练方案的技术空白。其轻量级全模态接口遵循 HuggingFace 规范，允许研究者像搭积木一样快速集成 Navit 视觉编码器、Whisper 语音编码器等组件，仅需实现 lm_encode、lm_generate 等标准函数即可接入。GitHub 仓库已提供桌面助手应用，支持实时截屏录屏交互，加速 GUI Agent 开发。框架的普及将重塑多模态 AI 研发范式：一方面降低中小型企业进入门槛，使 720P 文生视频、跨模态机器人等应用开发成为可能；另一方面推动学术研究从工程调优转向算法创新，例如通过 VeOmni 快速验证新型 MoE 路由算法或长序列注意力机制。随着火山引擎集成版的一键部署功能上线，全模态 AI 的规模化落地进程将进一步加速。

1.4 昆仑万维开源多模态框架 Skywork UniPic 2.0

昆仑万维于 2025 年 8 月 13 日在 SkyWork AI 技术发布周上正式开源了 Skywork UniPic 2.0，这是一款突破性的统一多模态框架，首次在单一模型中深度融合图像理解、文本到图像生成（T2I）和图像编辑（I2I）三大核心能力。该模型基于 2B 参数的 SD3.5-Medium 架构，通过创新的渐进式双任务强化策略和轻量化设计，实现了生成质量与部署效率的双重突破，其性能超越多个 12B 以上参数的同类模型，成为开源多模态领域的新标杆。

技术架构上，UniPic 2.0 采用三模块协同设计，构建了从理解到生成再到编辑的端到端闭环。生图编辑模块通过改造 SD3.5-Medium 架构，将原本仅支持文本输入的模型升级为同时处理文本与图像的混合输入模式。训练过程中，文本通过编码器生成指令表示，参考图像经 VAE 编码为潜变量并映射为上下文 token，两者与目标图像的噪声 token 拼接成统一序列，利用位置编码区分不同片段，从而在不改变模型结构的前提下赋予其 T2I 与 I2I 双能力。这一设计的关键在于使用了 600 万高质量生图样本和 500 万编辑样本的社区开源数据，覆盖多种图像比例，避免了模型对单一构图的偏置。统一模型能力模块则通过冻结生图编辑部分，引入 Qwen2.5-VL-7B 多模态模型和轻量连接器进行特征对齐。训练分为两阶段：

先预训练连接器实现多模态特征空间匹配，再联合微调连接器与生图模块，使三大任务在统一架构中协同优化。实测显示，该设计能精准识别图像元素并生成对应文案，例如输入景点图片后，模型可准确输出长城、埃菲尔铁塔等地标的文化背景说明。

图表8：Skywork UniPic 2.0 表现对比

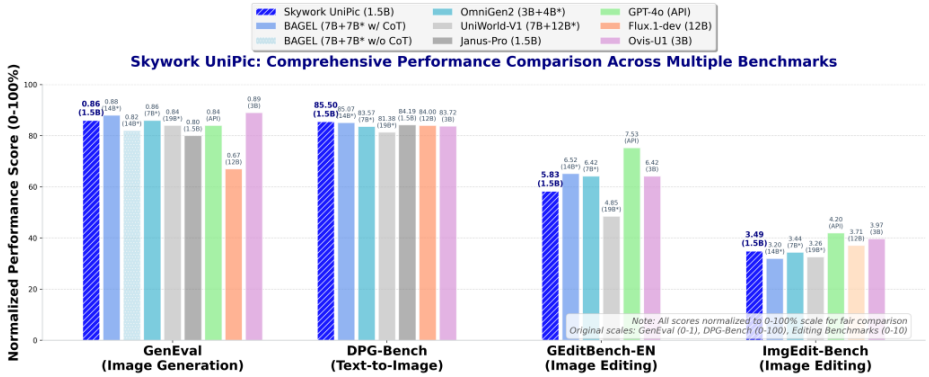


Figure 3: Performance comparison across multiple benchmarks. Skywork UniPic demonstrates competitive performance across understanding, generation, editing, and in-context tasks while maintaining exceptional parameter efficiency with only 1.5B activated parameters.

资料来源：昆仑万维，中邮证券研究所

性能突破的核心在于生图编辑后训练阶段采用的 Flow-GRPO 渐进式双任务强化策略。传统多任务强化学习常因任务冲突导致性能折中，而 UniPic 2.0 将优化过程分为两个独立阶段：先专项强化图像编辑任务，确保模型在保留原图结构的前提下精准执行局部修改；再优化文生图任务对复杂指令的语义理解。这种分阶段策略避免了任务间干扰，反而产生协同增益。为提供可靠的强化信号，团队构建了行业首个图像编辑专用奖励模型 Skywork-EditReward，以 Qwen2.5-VL-7B 为骨架，输入原图、编辑结果与指令，输出像素级质量评分。该模型通过 33.3 万条编辑样本训练，并由 GPT-4.1 进行标注校准，确保评分符合人类审美标准。

评测数据表明，UniPic 2.0 以极小参数量实现了越级表现。基础版本 UniPic2-SD3.5M-Kontext 在 GenEval 生图评估中取得 0.89 分，超越 12B 参数的 Flux.dev（0.84 分）；编辑任务上 GEdit-EN 得分 6.59，同样领先 12B 的 Flux-Kontext（6.26 分）。扩展为统一架构的 UniPic2-Metaquery 后，编辑分数进一步提升至 7.10，逼近闭源的 GPT-4o（7.53 分），同时理解能力在 MMBench 等基准测

试中达到 83.5 分，与 19B 参数的 UniWorld-V1 相当。这种性能优势源于模型对时空信息的完整保留——例如将跑车图片转换为冰河世纪风格时，不仅能重构车辆形态，还能自动补全符合新风格的背景细节，如冰川纹理与史前植被。

在实际应用中，UniPic 2.0 展现出显著的流程革新价值。电商团队可输入商品图片与指令，直接生成多版营销海报，避免传统工作流中 PS、生图工具频繁切换的耗时；游戏开发者通过文字描述快速迭代角色原画，再通过编辑功能微调服饰细节，将设计周期从数天缩短至小时级。其轻量化特性（可在 RTX 4090 消费级显卡运行）与秒级响应速度（生成复杂图像仅需几秒）使其成为首个真正可落地的统一多模态框架。开源生态的构建进一步加速技术普及，GitHub 仓库已提供桌面助手应用，支持实时截屏交互，开发者仅需实现标准接口即可集成自定义视觉组件。

昆仑万维通过这一技术验证了统一架构的产业价值：当理解、生成与编辑能力被整合至同一系统，内容生产流程从“多工具串联”变为“单平台闭环”，不仅提升效率，更确保了风格一致性。正如 CEO 方汉在 WAIC 2025 所述，高频嵌入工作流的工具才能形成稳定商业价值。UniPic 2.0 正推动 AI 从演示项目转向生产链核心部件，其技术路径或将成为多模态领域的新范式。

2 企业动态

2.1 阿里发布通义 Wan2.2-I2V-Flash 模型

阿里通义大模型于 2025 年 8 月 11 日正式发布的 Wan2.2-I2V-Flash 模型，标志着图生视频技术进入“轻快”时代。该模型作为通义万相 2.2 系列的最新成员，在前代 Wan2.1 基础上实现了 12 倍的推理速度跃升，同时将 API 调用价格降至 0.1 元/秒，抽卡成功率提升 123%，成为目前性价比最高的电影级视频生成解决方案。其技术突破不仅体现在速度与成本上，更通过创新的架构设计与训练策略，重新定义了图像到视频转换的质量标准与创作自由度。

技术架构上，Wan2.2-I2V-Flash 延续了通义万相 2.2 系列首创的 MoE（混合专家）架构，总参数量达 27B，激活参数 14B。这种架构通过高噪声专家模型与低噪专家模型的协同工作，前者负责视频整体框架构建，后者专注细节优化，使得模型在保持生成质量的同时显著降低计算资源消耗。特别值得注意的是，该模型采用电影美学控制系统，将光影、色彩、构图等专业电影元素封装为可调节参数，用户通过 60 多个直观可控的变量即可精准调整视频风格，实现从浪漫唯美到紧张刺激的氛围切换。这种设计使得生成内容在微表情刻画、物理规律还原等细节上达到专业影视水准。

性能突破的核心在于三方面技术创新。首先，模型采用动态稀疏激活机制，根据输入图像与文本指令自动分配计算资源，避免传统密集模型的全参数计算冗余。实测显示，在处理风格化图像时，模型能稳定保持原图艺术特征并生成合理动态效果，例如将梵高风格画作转换为流动的星空动画时，笔触走向与色彩过渡均符合后印象派美学特征。其次，优化的指令遵循系统支持特效提示词直出与运镜精准控制，用户可直接输入“镜头缓慢拉远同时画面渐暗”等复合指令，模型能准确解析时空关系并生成对应视频。最后，异步流水线设计实现生成过程的多阶段重叠，视频 DiT（扩散变换器）与动作解码器并行运行，使得 5 秒视频的端到端生成耗时从传统模型的数分钟压缩至秒级。

在实际应用中，该模型展现出显著的工业化价值。电商平台可通过 API 快速生成商品展示视频，仅需上传产品静物图并输入“360 度旋转展示”指令，即可输出专业级广告素材；教育机构能一键将历史图片转换为动态场景，如输入古建筑照片生成斗拱结构拆解动画。阿里云百炼平台提供的异步调用接口支持 480P 与 720P 分辨率输出，开发者可通过 DashScope SDK 实现 Python 或 Java 环境集成，其动态批处理技术可同时处理数百个生成请求，显著提升大规模部署效率。值得注意的是，模型特别强化了对极端宽高比图像（如手机竖屏截图）的适配能力，生成视频会自动保持原始比例并进行智能补全，避免传统方案的画面裁剪失真。

与开源版本 Wan2.2-I2V-A14B 相比，I2V-Flash 虽参数量相同，但通过量化压缩与算子优化，实现了消费级硬件的可部署性。测试显示，在 RTX 4090 显卡上生成 720P 视频仅需 200 毫秒推理时间，且显存占用降低 40%。这种效率提升

源于团队独创的 ROI（感兴趣区域）驱动重计算策略，该策略选择性激活视频关键区域的生成路径，对背景等静态区域采用轻量化处理。此外，模型支持种子锁定与负向提示词功能，开发者可通过指定随机数种子复现生成效果，或使用“低分辨率、错误比例”等负向提示规避常见缺陷，极大提升了内容可控性。

通义 Wan2.2-I2V-Flash 的发布不仅解决了影视级内容生产的效率瓶颈，更通过技术民主化推动创作生态变革。其开源版本已吸引超过 3 万开发者下载，衍生出动漫分镜生成、工业流程模拟等创新应用。正如阿里云智能集团资深专家所述，当视频生成从专业工具变为基础能力，每个创意者都能成为“一个人的制片厂”。随着模型在阿里云全球 21 个地域的规模化部署，其技术红利正加速渗透至广告、教育、社交等多元领域，重塑数字内容的生产与消费范式。

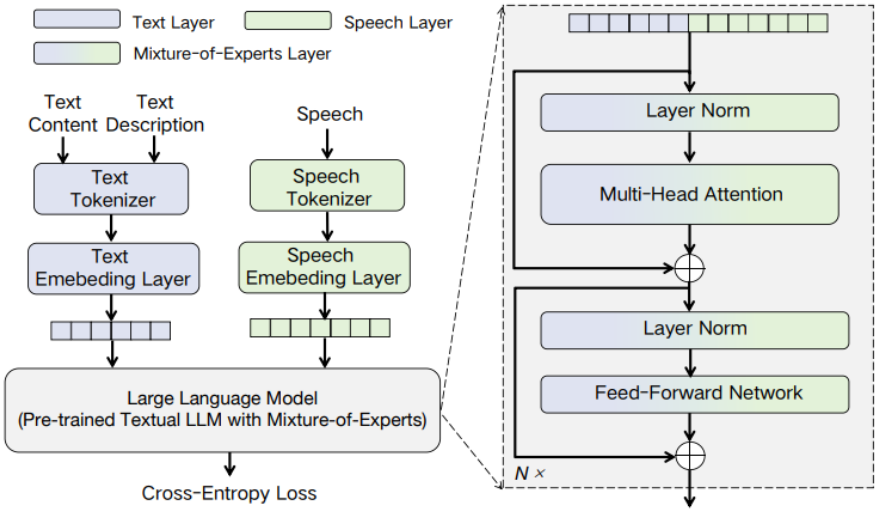
2.2 昆仑万维上线音频模型 Mureka V7.5，并推 MoE-TTS 语音合成框架

昆仑万维于 2025 年 8 月 15 日在 SkyWork AI 技术发布周上正式推出音频生成模型 Mureka V7.5 及 MoE-TTS 语音合成框架，标志着中文 AI 音乐创作与语音合成技术迈入新阶段。这一技术组合不仅解决了中文音乐生成中的文化适配难题，更通过混合专家架构（MoE）实现了语音合成的自然语言自由控制，为多模态内容创作提供了完整的技术闭环。

Mureka V7.5 的核心突破在于其对中文音乐文化特性的深度建模。不同于主流音乐生成模型追求多语言泛化能力，该模型专注于中文音乐的艺术神韵与情感表达，系统性地捕捉了从传统民歌、戏曲到当代流行乐的语义结构和情感走向。其技术实现依赖于三大创新：首先是通过优化自动语音识别技术（ASR），反向建模真实演唱中的气息运用、情感起伏和唱法细节，使生成的乐句划分、换气停顿符合中文演唱规律；其次是将人类主观听感评分机制引入训练目标，使模型主动规避机械感音色特征，在处理中文特有的韵律和气息时更贴近真人演唱逻辑；最后是构建跨文化语境的理解模块，使生成作品既能准确传达歌词语义，又能契合目标审美与文化背景。实测显示，该模型在中文歌曲质量与提示契合度两项主观

评测中分别以34.8%和45.2%的得票率超越Suno、Udio等国际主流模型，成为目前唯一能在艺术性与技术性上同步领先的解决方案。

图表9：MoE-TTS架构概览



资料来源：昆仑万维，中邮证券研究所

MoE-TTS作为支撑Mureka V7.5的语音底座，是全球首个基于混合专家架构的角色描述语音合成框架。其革命性在于彻底摒弃传统标签式模板控制，转而采用自然语言驱动的开放描述系统。用户可通过“清澈的少年音带磁性尾韵”等复杂描述精准控制声音特征，这得益于其创新的模态解耦设计：预训练大语言模型（LLM）负责解析自然语言语义并生成高维表达向量，多个语音专家模块（Speech Expert Modules）并行建模风格、节奏、语气等维度，最终通过模态路由器动态聚合输出。这种结构在冻结文本参数的同时实现跨模态信息对齐，既避免了知识损失，又显著提升了对比喻、隐喻等复杂修辞的泛化能力。技术报告显示，MoE-TTS在域内与域外描述测试集上，风格表现力贴合度与整体贴合度均优于ElevenLabs、MiniMax等闭源产品，尤其在处理“美国男演员带纽约口音”等复合角色设定时，能精准还原语调起伏与节奏张力。

两者的协同效应体现在技术架构与训练策略的深度融合。Mureka V7.5采用MoE-TTS作为人声生成引擎，使音乐作品中的演唱不仅音高准确，更能根据歌词情感自动调整气音、顿挫等微观表现；而MoE-TTS则通过Mureka积累的中文演

唱数据优化专家模块，强化对声乐技巧的建模能力。这种双向增强使系统在生成《凌晨两点的火车站》等民谣时，能模拟深夜清唱的孤独感；处理 R&B 曲风时则自然呈现甜蜜气息，摇滚风格下又能爆发激烈情绪，实现“一首歌一个灵魂”的艺术效果。

商业化落地方面，昆仑万维为开发者提供了分层级创作工具链。简单模式支持自然语言描述生成完整歌曲，如输入“温柔又心酸的民谣”即可自动匹配风格；高级模式开放歌词结构、参考音频等精细控制，媲美专业制作人工作流；音频编辑模块则提供乐句级重生成、乐器分轨等 DAW 级功能，满足后期精修需求。MoE-TTS 已规划集成至 Mureka-Speech 平台，将服务于情绪播报、游戏语音包、无障碍阅读等场景，其 API 支持实时生成带特定文化特征的语音，例如用哀伤语气播报讣告或为方言角色定制配音。

这一技术体系的发布不仅是工程突破，更体现了对中文数字文化遗产的前瞻性布局。当多数 AI 音乐模型聚焦英语市场时，昆仑万维选择深耕中文语境，通过解构汉语特有的声调、韵脚与呼吸逻辑，构建起文化隔阂的技术护城河。正如开发者社区评价所言：“有些旋律只有中文能唱，而 Mureka V7.5 让 AI 学会了闭上眼睛去唱。”未来随着模型迭代与生态完善，该技术组合或将成为中文智能内容生产的核心基础设施，从艺术创作维度拓展 AI 的文化影响力。

3 AI 行业洞察

3.1 阿里国际站 Accio Agent 海外爆火

阿里国际站推出的 Accio Agent 代表了 B2B 领域 AI 应用的范式革新，其核心价值在于将传统跨境采购中需要数周完成的复杂流程压缩至分钟级，通过多智能体协同架构与深度供应链整合重构全球贸易效率。该产品于 2024 年 11 月在欧洲 Web Summit 科技峰会首发，定位为“个人采购代理”，其命名源自《哈利·波特》中的召唤咒语，寓意通过 AI 瞬间召唤全球供应链资源。截至 2025 年 8 月，

该产品企业用户量已突破 200 万，测试数据显示其将搜索到采购的转化率提升 20%-30%，成为阿里海外业务效率之战的关键武器。

技术架构上，Accio Agent 采用多智能体动态任务分配机制，每次执行任务时系统自动调度 5-10 个专业领域智能体协同工作。例如处理“沙漠建造室内滑雪场”这类非标需求时，需求拆解 Agent 会将任务分解为造雪机采购、隔热材料筛选等具体环节，市场分析 Agent 同步调取迪拜冬季游客量数据评估投资回报率，物流 Agent 则根据供应商地理位置优化运输方案。这种模块化设计得益于阿里整合的 Qwen2.5、DeepSeek 等先进推理模型，构建起覆盖服饰、机械、3C 等领域的专业知识图谱，能精准解析“面料克重”“针织纹理”等行业术语，确保推荐匹配度。数据层整合了阿里国际站 25 万商家及 3000 万家跨境贸易企业信息，通过多重交叉验证机制减少大模型幻觉，例如供应商商誉评级需经过平台交易记录、终端销量、客户评价等多维度校验才被纳入推荐列表。

工作流革新体现在端到端自动化闭环的实现。当用户提出“Switch 游戏控制器海外销售”需求时，Accio Agent 会在 30 秒内完成四阶段操作：首先通过产品搜索引擎匹配具备 OEM/ODM 能力的供应商，如深圳鸿森电子（20 年经验、4.9 星评级）因其定制化选项丰富被优先推荐；随后市场趋势 Agent 抓取 Reddit、IGN 等平台数据，分析 2025 年控制器热门功能如陀螺仪精度提升需求；采购策略 Agent 则根据最小起订量（2-500 件）和区域定价差异（北美市场溢价 15%）生成分级报价方案；最终自动生成含供应商联系方式、设计图纸、询价模板的完整报告。这种全流程覆盖使传统需采购团队数周完成的工作被压缩至 10 分钟，尤其适合中小企业在缺乏专业团队时快速响应市场机会。

商业化落地方面，Accio Agent 展现出三类典型应用场景。对于产品创新者，上传爆款水杯图片即可获得 100 种改款灵感并直连定制供应商，将设计到打样周期从月级缩短至日级；对于市场进入者，输入“狗主题陶瓷杯”指令后，系统不仅提供安克鲁科技等 ISO 认证厂商信息，还会基于宠物用品趋势报告建议加入磁吸杯盖等创新元素；对于大宗采购商，夜间秒回功能可自动捕获全球询盘，如沙特买家 2000 万美元订单案例中 AI 在非工作时间完成供应商初筛与报价单生成，显著提升转化效率。这些能力源于阿里国际站构建的“Alibaba Guaranteed”履约体系，参与半托管的商品会标注准确送货时效与售后保障，降低跨境交易风险。

生态扩展策略上，Accio 正从工具向平台演进。2025 年 8 月推出的 Agent 模式允许用户通过自然语言描述调用预制工作流，例如“开发磁悬浮家居装饰品”会触发技术可行性分析、专利检索、供应商匹配等子任务链。由于整合了阿里数十年的外贸资源，其推荐结果并非理论方案而是具备真实产能的工厂信息，如香港 Ason Innovations 公司（年收入 5000 万-1 亿美元）因欧美市场渠道优势被优先推荐。这种深度产业整合形成竞争壁垒，使 Accio 的采购方案落地性远超同类产品。随着移动端开发与实时供应链数据更新等功能的持续迭代，该产品有望成为跨境电商领域的“操作系统”，推动万亿级市场的智能化跃迁。

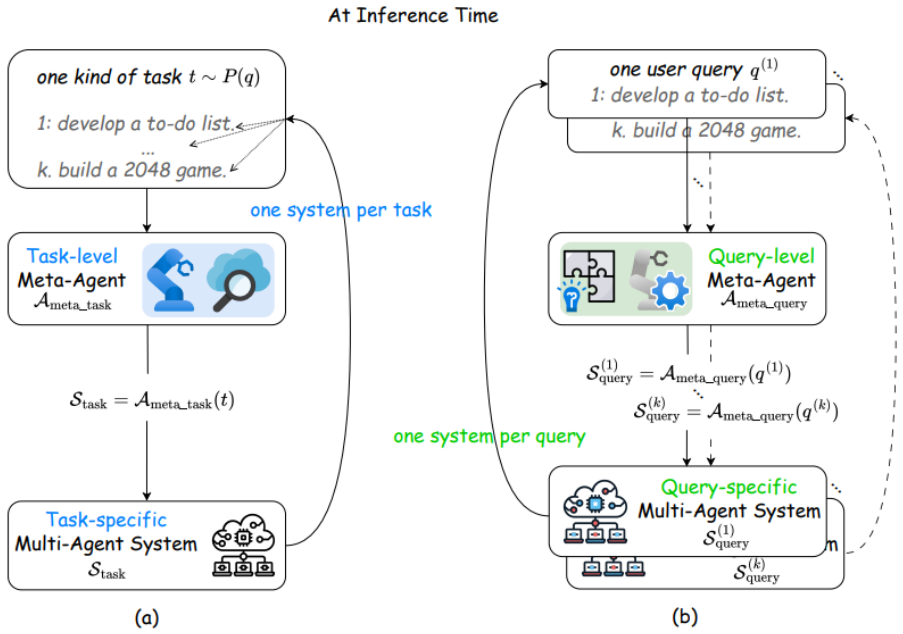
4 技术前沿

4.1 FlowReasoner：增强查询级元智能体

《FlowReasoner: Reinforcing Query-Level Meta-Agents》这篇由 Sea AI Lab、新加坡国立大学等机构联合发表的论文，提出了一种革命性的查询级元代理框架，旨在通过强化学习与外部执行反馈的协同机制，实现针对单个用户查询的个性化多智能体系统自动生成。该研究突破了传统任务级元代理的局限性，在代码生成等复杂领域展现出显著。

传统多智能体系统设计存在两大瓶颈：一是依赖人工设计工作流导致扩展性受限，二是任务级优化无法适配查询级差异。现有方法如 Aflow、ADAS 等仅能生成"一个系统对应一类任务"的固定架构，而 FlowReasoner 创新性地实现了"一个系统对应一个查询"的动态适配。这种转变的核心在于将工作流生成从搜索驱动转变为推理驱动，通过蒸馏 DeepSeek R1-671B 模型获得基础推理能力后，采用 GRPO（分组相对策略优化）强化学习框架，利用包含性能、复杂度、效率的多目标奖励机制持续优化。实验证明，该方法在 BigCodeBench 等三大代码基准测试中，14B 参数的 FlowReasoner 模型比最优基线 MaAS 提升 5 个百分点，较基础工作模型 o1-mini 提升 10.52%准确率。

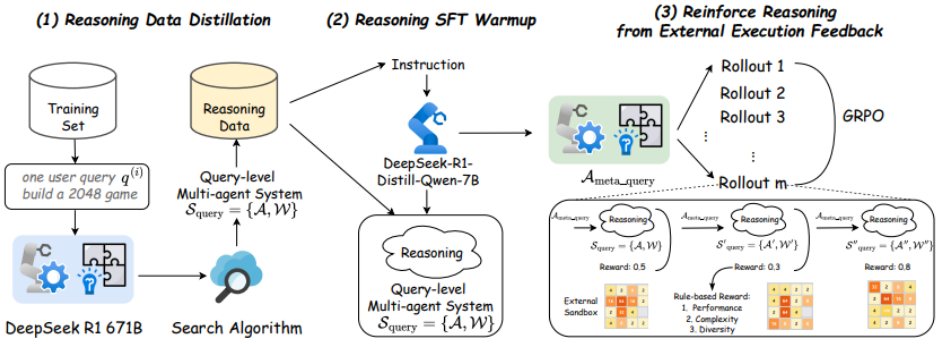
图表10：推理时的任务级元代理与查询级元代理



资料来源：FlowReasoner，中邮证券研究所

FlowReasoner 的训练流程分为三阶段闭环：首先通过 R1-671B 合成数千条监督微调数据，构建初始推理能力；随后进行 SFT 微调使 7B/14B 模型掌握工作流生成基础；最终通过带外部反馈的 RL 阶段实现能力跃升。其创新点体现在三方面：一是采用过程奖励监督机制，对每轮推理结果进行归一化评分并计算累积优势；二是设计动态工作流优化空间，将节点参数（如温度值、输出格式）和边交互（如 Ensemble、Review 操作符）编码为可编程结构；三是引入 k=1.1 的缩放因子和 T=3 的阈值控制，平衡探索与利用。这种设计使得生成的工作流能根据查询复杂度自适应调整，BigCodeBench 工程任务触发 10 步复杂流程，而 HumanEval 算法题仅需 3 步精简步骤。

图表11：FlowReasoner 训练流



资料来源：FlowReasoner，中邮证券研究所

研究团队设计了多层次验证体系：在横向对比中，FlowReasoner-14B 在 MBPP 数据集达到 92.15%准确率，远超 Aflow 的 82.4%和 LLM-Blender 的 78.65%；纵向消融显示，RL 阶段使 7B 模型在 BigCodeBench 上的性能从 61.79%提升至 62.78%，证明外部反馈的有效性。特别值得注意的是其泛化能力，如表 3 所示，当基础 worker 从 o1-mini 替换为 Claude 3.5 时，14B 模型在 HumanEval 仍保持 96.52%的高准确率，表明元代理的规划能力可迁移至不同执行器。现有开源模型作为元代理时性能骤降，凸显 FlowReasoner 在指令跟随与工作流推理方面的独特优势。

尽管整体表现优异，但也存在典型失败案例：在网页锚点提取任务中，模型未能正确组合 HTML 解析与属性过滤操作；在偶数分解问题中，错误地将数学约束转化为无效循环。这些案例反映出现有方法在组合泛化与数学推理方面的局限，作者建议未来结合形式化验证工具提升可靠性。

该研究的产业价值在于将多智能体系统开发成本降低 90%，其开源实现已支持实时截屏交互的 GUI Agent 开发。正如论文结论所述，这种推理驱动的个性化工作流生成范式，为 LLM 应用从固定管道向动态适应系统的演进提供了关键技术路径，未来可扩展至机器人控制、科学发现等更复杂领域。研究团队在附录中详细公开了 1,400 项训练数据样例和超参配置，为社区复现与改进奠定基础。

5 风险提示

以上内容基于历史数据完成，在政策、市场环境发生变化时存在失效的风险；历史信息不代表未来。

中邮证券投资评级说明

投资评级标准	类型	评级	说明
报告中投资建议的评级标准：报告发布日后的6个月内的相对市场表现，即报告发布日后的6个月内的公司股价（或行业指数、可转债价格）的涨跌幅相对同期相关证券市场基准指数的涨跌幅。 市场基准指数的选取：A股市场以沪深300指数为基准；新三板市场以三板成指为基准；可转债市场以中信标普可转债指数为基准；香港市场以恒生指数为基准；美国市场以标普500或纳斯达克综合指数为基准。	股票评级	买入	预期个股相对同期基准指数涨幅在20%以上
		增持	预期个股相对同期基准指数涨幅在10%与20%之间
		中性	预期个股相对同期基准指数涨幅在-10%与10%之间
		回避	预期个股相对同期基准指数涨幅在-10%以下
	行业评级	强于大市	预期行业相对同期基准指数涨幅在10%以上
		中性	预期行业相对同期基准指数涨幅在-10%与10%之间
		弱于大市	预期行业相对同期基准指数涨幅在-10%以下
	可转债评级	推荐	预期可转债相对同期基准指数涨幅在10%以上
		谨慎推荐	预期可转债相对同期基准指数涨幅在5%与10%之间
		中性	预期可转债相对同期基准指数涨幅在-5%与5%之间
		回避	预期可转债相对同期基准指数涨幅在-5%以下

分析师声明

撰写此报告的分析师（一人或多人）承诺本机构、本人以及财产利害关系人与所评价或推荐的证券无利害关系。

本报告所采用的数据均来自我们认为可靠的目前已公开的信息，并通过独立判断并得出结论，力求独立、客观、公平，报告结论不受本公司其他部门和人员以及证券发行人、上市公司、基金公司、证券资产管理公司、特定客户等利益相关方的干涉和影响，特此声明。

免责声明

中邮证券有限责任公司（以下简称“中邮证券”）具备经中国证监会批准的开展证券投资咨询业务的资格。

本报告信息均来源于公开资料或者我们认为可靠的资料，我们力求但不保证这些信息的准确性和完整性。报告内容仅供参考，报告中的信息或所表达观点不构成所涉证券买卖的出价或询价，中邮证券不对因使用本报告的内容而导致的损失承担任何责任。客户不应以本报告取代其独立判断或仅根据本报告做出决策。

本报告所载的意见、评估及预测仅为本报告出具日的观点和判断。该等意见、评估及预测无需通知即可随时更改。过往的表现亦不应作为日后表现的预示和担保。在不同时期，中邮证券可能会发出与本报告所载意见、评估及预测不一致的研究报告。

中邮证券及其所属关联机构可能会持有报告中提到的公司所发行的证券头寸并进行交易，也可能为这些公司提供或者计划提供投资银行、财务顾问或者其他金融产品等相关服务。

《证券期货投资者适当性管理办法》于2017年7月1日起正式实施，本报告仅供中邮证券签约客户使用，若您非中邮证券签约客户，为控制投资风险，请取消接收、订阅或使用本报告中的任何信息。本公司不会因接收人收到、阅读或关注本报告中的内容而视其为签约客户。

中邮证券对于本申明具有最终解释权。

公司简介

中邮证券有限责任公司，2002 年 9 月经中国证券监督管理委员会批准设立，是中国邮政集团有限公司绝对控股的证券类金融子公司。

公司经营范围包括:证券经纪，证券自营，证券投资咨询，证券资产管理，融资融券，证券投资基金销售，证券承销与保荐，代理销售金融产品，与证券交易、证券投资活动有关的财务顾问等。

公司目前已经在北京、陕西、深圳、山东、江苏、四川、江西、湖北、湖南、福建、辽宁、吉林、黑龙江、广东、浙江、贵州、新疆、河南、山西、上海、云南、内蒙古、重庆、天津、河北等地设有分支机构，全国多家分支机构正在建设中。

中邮证券紧紧依托中国邮政集团有限公司雄厚的实力，坚持诚信经营，践行普惠服务，为社会大众提供全方位专业化的证券投、融资服务，帮助客户实现价值增长，努力成为客户认同、社会尊重、股东满意、员工自豪的优秀企业。

中邮证券研究所

北京
邮箱：yanjiusuo@cnpsec.com
地址：北京市东城区前门街道珠市口东大街 17 号
邮编：100050

上海
邮箱：yanjiusuo@cnpsec.com
地址：上海市虹口区东大名路 1080 号邮储银行大厦 3 楼
邮编：200000

深圳
邮箱：yanjiusuo@cnpsec.com
地址：深圳市福田区滨河大道 9023 号国通大厦二楼
邮编：518048