统计结果

GUI 相关文章共70篇，最高平均分：6，最低平均分：2.5，总体平均分：4.23

最高分文章:**1、ViMo: A Generative Visual GUI World Model for App Agents**

2、OmniActor: A Generalist GUI and Embodied Agent for 2D&3D Worlds

3、ScaleCUA: Scaling Open-Source Computer Use Agents with Cross-Platform Data

最低分文章:**1、Generalist Scanner Meets Specialist Locator: A Synergistic Coarse-to-Fine Framework for Robust GUI Grounding（2.5)**

2、GUI-PRA: Process Reward Agent for GUI Tasks（2.8)

3、GUI-Shepherd: Reliable Process Reward and Verification for Long-Sequence GUI Tasks（2.7)

[2,3):三篇（4.3%） [3,4):二十二篇（31.4%） [4,5):二十六篇（37.1%） [5,6):十六篇（22.9%） [6,7):三篇（4.3%）

具体统计如下

Title: Label-free GUI Grounding via Confidence-guided Negative Reinforcement Learning

TLDR: 为解决 GUI 智能体扩展中人工标注数据成本高的瓶颈，基于坐标令牌置信度特征和稀疏 GUI 坐标空间中负样本学习信号更可靠的洞察，提出的无标注训练范式（CRL 和 CNRL）在多个基准测试中表现优异，证实坐标令牌置信度可作为人工标注的有效替代方案。

初试分数: 6，4，6，4

置信度: 3，4，4，5

平均分:5

Title:V2P: Visual Attention Calibration for GUI Grounding via Background Suppression and Center Peaking

TLDR: 为解决 GUI 智能体扩展中人工标注数据成本高的瓶颈，基于坐标令牌置信度特征和稀疏 GUI 坐标空间中负样本学习信号更可靠的洞察，提出的无标注训练范式（CRL 和 CNRL）在多个基准测试中表现优异，证实坐标令牌置信度可作为人工标注的有效替代方案。

初试分数: 6，4，6，4

置信度: 3，4，4，5

平均分:5

Title:Learning GUI Grounding with Spatial Reasoning from Visual Feedback

TLDR: 为解决高分辨率、复杂布局 GUI 的 Grounding 问题，将该任务重构为交互式搜索任务的 GUI-Cursor 7B 模型，通过多步在线强化学习训练，能输出光标移动动作定位 UI 元素，在 ScreenSpot-v2 和 ScreenSpot-Pro 基准上实现最先进精度，且训练中移动步数随精度提升而减少，可自适应应对不同难度样本。

初试分数: 4，6，6，4

置信度: 4，4，4，3

平均分:5 **MAX**

Title:UI-Ins: Enhancing GUI Grounding with Multi-Perspective Instruction as Reasoning

TLDR:为解决现有 GUI Grounding 方法忽视指令多样性影响的问题，基于 “指令即推理” 范式及两阶段训练框架（多样化指令的监督微调 + 路径选择与组合的强化学习）提出的 UI-Ins 系列模型，在多个基准测试中取得最先进成果，还展现出涌现推理能力与优异的智能体性能。

初试分数: 2，6，6，6

置信度: 4，5，4，3

平均分:5

Title: UItron: Foundational GUI Agent with Advanced Perception and Planning

TLDR: Through innovative data pipelines and training frameworks, the proposed UItron has made significant advancements in both Chinese and English GUI agent interaction scenarios.

初试分数: 4，6，2，2，3

置信度: 3，3，4，4，5

平均分:3.4

Title: GUI-ReWalk: Massive Data Generation for GUI Agent via Stochastic Exploration and Intent-Aware Reasoning

TLDR: 为解决 GUI 智能体发展中高质量轨迹数据稀缺的问题，推理增强的多阶段框架 GUI-ReWalk 通过随机探索与推理引导相结合的方式，生成真实多样且支持长程工作流的 GUI 轨迹数据，用其训练的模型在多个基准测试中表现优异，证实该框架具备可扩展性与数据高效性，能推动 GUI 智能体的真实场景自动化发展。

初试分数: 4，4，2

置信度: 4，4，3

平均分:3.3

Title: GOLD: Global Overview to Local Detail in Efficient Visual Grounding for GUI Agents

TLDR: 为解决 VLM-based GUI Grounding 计算开销大、难以边缘部署的问题，无调优且适配不同界面密度的 GOLD 框架通过全局剪枝、局部细化与全局 - 局部上下文融合三阶段流程，在 ScreenSpot-V2 基准上减少 78% 计算量的同时提升 0.7% 准确率，证实了全局到局部接地方案的高效性。

初试分数: 4，4，6

置信度: 3，2，3

平均分:4.7

Title: GUI-R1: A Generalist R1-Style Vision-Language Action Model For GUI Agents

TLDR: 为解决现有基于 LVLM 的 SFT 训练范式在 GUI 智能体开发中数据需求大、泛化能力弱的问题，首个针对高级现实任务场景的强化学习框架 GUI-R1 通过统一动作空间规则建模，利用少量跨平台高质量数据及改进的策略优化算法，在三大平台的八个基准测试中以仅 0.02% 的数据量（3K vs. 13M）超越此前 SOTA 方法，证实了该强化学习方案提升 LVLM 实际 GUI 任务执行能力的巨大潜力。

初试分数: 2，6，4，4

置信度: 5，2，4，4

平均分:3.5

Title: GUI-AIMA: Aligning Intrinsic Multi-Modal Attention with a Context Anchor for GUI Grounding

TLDR: We proposed GUI-AIMA, a attention-based GUI visual grounding method supervised on anchored attention with query-adaptive multi-head weighting.

初试分数: 4，4，6，2

置信度: 4，4，4，5

平均分:4

Title: Improving GUI Grounding with Explicit Position-to-Coordinate Mapping

TLDR: RULER tokens provide explicit position-to-pixel mapping and I-MRoPE balances spatial encoding, transforming GUI grounding from regression to referencing with strong high-resolution generalization.

初试分数: 2，4，4，4

置信度: 3，3，4，5

平均分:3.5

Title: Beyond Clicking: A Step Towards Generalist GUI Grounding via Text Dragging

TLDR: We enhance and evaluate the text dragging capabilities of existing grounding models through an automated data collection pipeline and a comprehensive benchmark.

初试分数: 4，4，2，6

置信度: 4，3，4，3

平均分:4

Title: Error as Signal: Stiffness-Aware Diffusion Sampling via Embedded Runge-Kutta Guidance

TLDR: ERK-Guid 基于扩散模型中刚性区域 ODE 轨迹局部截断误差（LTE）与主导特征向量相关的关键观察，无需辅助网络，通过利用检测到的刚性来减少 LTE、稳定采样，在合成数据集和 ImageNet 基准上持续优于现有 SOTA 方法。

初试分数: 4，6，4，4

置信度: 3，3，3，3

平均分:4.5

Title: All in One: Unified Pretraining of GUI Agents via Masked Trajectory Prediction

TLDR: 为解决现有 GUI 智能体预训练策略直接统一导致的目标不一致与数据异质性问题，统一框架 MTP 通过掩码方式将多样预训练策略整合为一致目标，结合角色感知适配器学习模块处理数据异质性，在四大 GUI 导航基准测试中显著优于此前方法并达成 SOTA，展现出优异的有效性与泛化能力。

初试分数: 4，4，2，4

置信度: 4，4，3，5

平均分:3.5

Title: VistaGUI: Towards More Robust and Intelligent GUI Automation

TLDR: 为解决现有 GUI 自动化智能体在 RAG 检索精度低、依赖单模态感知及缺乏故障恢复机制导致的脆弱性问题，多模态 GUI 智能体 VistaGUI 通过并行指令理解模块、自适应多模态感知模块与环境感知状态管理系统的统一框架，在各类 GUI 自动化任务中显著优于强基线模型，在任务成功率、恢复速度及整体稳健性上表现突出。

初试分数: 4，4，4，4，2

置信度: 2，2，3，4，3

平均分:3.6

Title: ViMo: A Generative Visual GUI World Model for App Agents

TLDR: A visual world model that generate App GUI to help App agent envision outcomes of action and make better decision.

初试分数: 6，4，6，8

置信度: 3，3，4，4

平均分:6 **MAX**

Title: GUI-Spotlight: Adaptive Iterative Focus Refinement for Enhanced GUI Visual Grounding

TLDR: A model trained for image-grounded reasoning that dynamically invokes multiple specialized tools to iteratively narrow its focus to the relevant region of the screen, thereby substantially improving visual grounding accuracy.

初试分数: 4，6，4，2

置信度: 5，3，3，5

平均分:4

Title: Generalist Scanner Meets Specialist Locator: A Synergistic Coarse-to-Fine Framework for Robust GUI Grounding

TLDR: 为提升 GUI Grounding 性能，受人类交互方式启发的粗细粒度协同框架 GMS 将通用 VLM 作为 “Scanner” 识别潜在兴趣区域、微调接地模型作为 “Locator” 输出精准坐标，通过五级架构与跨模态通信的分层搜索，在 ScreenSpot-Pro 数据集上实现显著精度提升，且大幅优于各类强基线模型，展现出通用 GUI Grounding 的稳健性与潜力。

初试分数: 0，4，4，2

置信度: 5，5，4，4

平均分:2.5 **MIN**

Title: SCREEN-SBERT: EMBEDDING FUNCTIONAL SEMANTICS OF GUI SCREENS TO SUPPORT GUI AGENTS

TLDR: 为解决 GUI 智能体知识检索中结构化元数据可用性低、纯视觉方法功能匹配不足的问题，聚焦于功能语义层面截图检索的纯视觉方法 Screen-SBERT，通过对比学习嵌入框架定义功能等价性与功能页面类概念，在真实移动 App 场景中检索功能等价截图的效果优于多个基线模型，为两阶段检索框架的第一阶段提供有效支撑。

初试分数: 4，6，4，8

置信度: 4，3，4，4

平均分:5.5 **MAX**

Title: Grounding GUI Anything: Efficient and Semantically-Aware Parsing via Continuous Coordinate Decoding

TLDR: 为解决现有 GUI Grounding 方法中离散令牌生成限制精度与效率、预测受限于预定义元素的问题，端到端框架 GGA 通过将 MLLM 与回归解码器结合实现连续坐标解码，融入拒绝机制减少幻觉，并配套推出 ScreenParse 基准，在多个数据集上持续优于 SOTA 方法，提升了定位精度、推理速度与泛化能力

初试分数: 4，8，2，2

置信度: 4，4，4，5

平均分:4

Title: Ferret-UI Lite: Lessons from Building Small On-Device GUI Agents

TLDR: 为解决小型端侧模型在跨平台 GUI 交互中的自主智能体开发难题，3B 参数的紧凑端到端 GUI 智能体 FERRET-UI LITE 通过融合真实与合成 GUI 数据、结合思维链推理与视觉工具使用、设计奖励机制的强化学习等端侧优化技术，在 GUI Grounding 和导航基准测试中取得与其他小型 GUI 智能体相当的竞争力，同时公开了开发紧凑端侧 GUI 智能体的方法与经验

初试分数: 4，4，4，6

置信度: 4，3，5，5

平均分:4.5

Title: D-Artemis: A Deliberative Cognitive Framework for Mobile GUI Multi-Agents

TLDR: 为解决 GUI 智能体端到端训练数据瓶颈、错误检测成本高及指导矛盾风险等问题，受人类 “思考 - 对齐 - 反思” 认知循环启发的审慎框架 D-Artemis，通过应用专属提示检索、执行前思想 - 动作一致性检查与动作修正、执行后状态反思等组件，在不依赖复杂轨迹数据训练的情况下增强 MLLM 的 GUI 任务能力，在 AndroidWorld（75.8% 成功率）和 ScreenSpot-V2（96.8%）基准上达成新 SOTA，各组件贡献经消融实验验证。

初试分数: 4，2，2，4

置信度: 3，4，4，4

平均分:3

Title: MobileGUI-RL: Advancing Mobile GUI Agent through Reinforcement Learning in Online Environment

TLDR: 为解决现有基于视觉的 GUI 智能体依赖离线预收集轨迹训练导致的扩展性差、过拟合特定 UI 模板及对未知环境鲁棒性弱等问题，在线训练框架 MobileGUI-RL 通过自探索与过滤合成可学习任务课程，并将 GRPO 适配于 GUI 导航（结合轨迹感知优势与平衡任务成功和执行效率的复合奖励），在三个在线移动智能体基准测试中实现持续性能提升，验证了其有效性。

初试分数: 4，4，2，4，4

置信度: 4，4，4，4，5

平均分:3.6

Title: Difficulty-Aware Reasoning for Mobile GUI Automation via Reinforcement Fine-Tuning

TLDR: 为解决现有 GUI 智能体采用统一思维链（CoT）推理导致的计算开销冗余与简单步骤性能下降问题，AdaGUI-R1 开创难度感知推理范式，通过自监督机制生成难度感知推理轨迹以赋予动态调节推理深度的基础能力，并借助 GAPO 算法（含自适应思维奖励与难度感知高斯带宽探索奖励）增强推理性能，最终在减少 40% 冗余推理令牌的同时提升 5% 动作精度，达成 GUI 自动化新 SOTA。

初试分数: 4，4，2，4，4

置信度: 3，4，3，4，5

平均分:3.6

Title: GUI-360° : A Comprehensive Dataset and Benchmark for Computer-Using Agents

TLDR: 为填补现实世界计算机使用智能体（CUAs）任务稀缺、多模态轨迹自动采集标注管道缺失及统一基准缺乏的空白，大规模综合数据集与基准套件 GUI-360° 通过自动化流程构建，包含 120 万 + Windows 办公应用动作步骤及多模态数据，支持 GUI Grounding、屏幕解析、动作预测三大任务与混合动作空间，基准测试显示 SOTA 视觉语言模型存在原生短板，而监督微调可带来显著性能提升。

初试分数: 6，4，6，2

置信度: 3，3，3，5

平均分:4.5

Title: UltraCUA: Scaling Computer Use Agent through GUI and Programmatic Control

TLDR: 为解决计算机使用智能体（CUAs）依赖原始 GUI 动作导致的级联故障与性能瓶颈、缺乏编程接口能力的问题，基础模型 UltraCUA 通过混合控制将 GUI 原语与高级编程工具调用无缝融合，依托自动化工具扩展管道、合成数据引擎、多智能体轨迹生成及两阶段训练 pipeline，在 OSWorld 实现 27% 相对性能提升且步骤快 11%，在 WindowsAgentArena 跨域测试达成 21.7% 成功率，既减少错误传播又保障执行效率，构建了连接原始 GUI 交互与编程智能的可扩展范式。

初试分数: 6，4，6，4

置信度: 3，4，4，5

平均分:5 **MAX**

Title: Auto-scaling Continuous Memory for GUI Agent

TLDR: 该研究提出一种连续记忆机制（以 VLM 为编码器将 GUI 轨迹编码为固定长度连续嵌入）与自动扩展数据飞轮，在降低上下文成本、保留视觉细节的同时实现记忆规模与检索深度的单调性能提升，仅微调少量参数便使 Qwen-2.5-VL-7B 在长视野和分布偏移场景下达到接近 GPT-4o 等闭源模型的真实 GUI 任务表现，相关数据与代码将公开。

初试分数: 4，6，2，4

置信度: 4，3，4，4

平均分:4

Title: Efficient Multi-turn RL for GUI Agents via Decoupled Training and Adaptive Data Curation

TLDR: 该研究提出解耦智能体强化学习训练框架 DART，通过异步分离环境集群、滚动服务、数据管理器和训练器四大模块提升系统效率，结合自适应数据筛选策略（补充高难度任务成功轨迹等），使 DART-GUI-7B 在 OSWorld 基准测试中实现 42.13% 的任务成功率（较基础模型提升 14.61%，超开源 SOTA 7.34%），相关训练框架、数据及模型检查点将通过指定网址开源。

初试分数: 6，4，6

置信度: 4，4，3

平均分:5.3 **MAX**

Title: GUI-PRA: Process Reward Agent for GUI Tasks

TLDR: We introduce GUI-PRA, a training-free judge agent that resolves key failure modes in GUI task supervision, through the lens of dynamic memory and adaptive UI perception.

初试分数: 4，2，2，4，2

置信度: 2，4，3，2，3

平均分:2.8 **MIN**

Title: GUI-KV: Efficient GUI Agents via KV Cache with Spatio-Temporal Awareness

TLDR: 该研究针对 VLM 驱动的 GUI 智能体长序列高分辨率截图处理效率低、推理慢的问题，基于 GUI 任务中所有 Transformer 层注意力稀疏度均较高的洞察，提出无需重训练的即插即用 KV 缓存压缩方法 GUI-KV，通过空间显著性引导和时间冗余评分利用 GUI 特有冗余，在多个基准测试中优于同类缓存压缩基线，在 AgentNetBench 的 5 截图场景下解码 FLOPs 降低 38.9% 且步骤准确率较全缓存基线提升 4.1%。

初试分数: 2，4，6，2

置信度: 5，3，3，4

平均分:3.5

Title: GUI-Shepherd: Reliable Process Reward and Verification for Long-Sequence GUI Tasks

TLDR: We propose a PRM and perform the first systematic study of process supervision in GUI agents from online long-horizon tasks to offline single-step prediction, from RL training to inference verification.

初试分数: 2，2，4

置信度: 3，4，4

平均分:2.7 **MIN**

Title: InfiGUI-R1: Advancing Multimodal GUI Agents from Reactive Actors to Deliberative Reasoners

TLDR: 该研究针对 MLLM 驱动的 GUI 智能体在整体训练范式中存在的能力层级不匹配问题，提出 “先赋予、后内化” 的两阶段分层训练范式 Actor2Reasoner，通过第一阶段认知赋予（目标监督微调注入关键推理技能）和第二阶段策略内化（RL 将能力内化为稳健决策策略），使实例化的 InfiGUI-R1 在 AndroidControl 等基准测试中达成 SOTA 性能，证实分离基础能力赋予与策略内化是构建高性能 GUI 智能体的更优路径。

初试分数: 8，2，4，6

置信度: 5，3，3，3

平均分:5 **MAX**

Title: ProRe: A Proactive Reward System for GUI Agents via Reasoner–Actor Collaboration

TLDR: 该研究针对现有奖励方法难以泛化到 GUI 智能体、静态 LLM 评判法准确率有限的问题，提出主动奖励系统 ProRe，通过通用推理器调度目标状态探测任务，由领域特定评估智能体与环境主动交互收集额外观测，实现更准确可验证的奖励分配，在 3K + 轨迹上奖励准确率和 F1 分数最高提升 5.3% 和 19.4%，与 SOTA 策略智能体集成后成功率最高提升 22.4%。

初试分数: 4，6，6，6

置信度: 4，3，4，4

平均分:5.5 **MAX**

Title: Agent-ScanKit: Unraveling Memory and Reasoning of Multimodal Agents via Sensitivity Perturbations

TLDR: We propose Agent-ScanKit to probe the memory and reasoning capabilities of MLLM-based agents via controlled perturbations. The results show that most agents rely on memorization rather than true reasoning, contravening safety over autonomy.

初试分数: 4，4，8，4

置信度: 4，2，3，3

平均分:5 **MAX**

Title: GAIA: A Data Flywheel System for Training GUI Test-Time Scaling Critic Models

TLDR: 该研究针对 LVLM 驱动的 GUI 智能体操作不可逆（错误操作可能导致灾难性偏差）的问题，提出训练框架 GAIA，通过迭代式批评能力提升基础 GUI 智能体的测试时缩放（TTS）性能：先利用基础智能体的正负动作样本训练直觉批评模型（ICM）以评估动作即时正确性，再通过批评模型引导智能体收集优化样本形成自改进循环，训练出辨别能力更强的多轮批评模型。实验表明，ICM 能提升多种闭源和开源模型的测试时性能，且随数据循环利用性能逐步提升，相关代码和数据集将公开。

初试分数: 4，6，4，4

置信度: 3，4，3，5

平均分:4.5

Title: GUI-Shift: Enhancing VLM-Based GUI Agents through Self-supervised Reinforcement Learning

TLDR: 该研究针对 GUI 智能体训练依赖大规模标注数据集（耗时且易出错）的问题，提出自监督逆动力学任务 K-step GUI Transition（让 VLM 通过预测导致两个 GUI 状态转换的初始动作来学习 GUI 动态，无需自然语言指令，可从现有轨迹或自动探索构建可扩展数据集），并基于此提出结合规则优化与数据过滤的 RL 框架 GUI-Shift。在四个基准测试中，GUI-Shift 训练对 GUI 自动化和定位任务均有良好泛化性，使 GUI 自动化准确率最高提升 11.2%，为利用无标注 GUI 轨迹训练提供了可扩展方案。

初试分数: 6，4，6，6

置信度: 2，4，4，3

平均分:5.5 **MAX**

Title: Computer-Use Agents as Judges for Automatic GUI Design

TLDR: \bench 基准涵盖 52 个不同领域应用及 1560 个可验证的真实场景模拟任务，提出 “Coder（设计师）-CUA（评估者）协作” 框架，通过 CUA 仪表盘将导航历史转化为可解释指导以优化设计，核心以任务可解性和导航成功率为衡量标准，推动界面设计向智能体原生的高效可靠转变，助力智能体从被动使用转向数字环境主动参与。

初试分数: 2，4，4

置信度: 4，4，4

平均分:3.3

Title: MobileA3gent: Training Mobile GUI Agents Using Decentralized Self-Sourced Data from Diverse Users

TLDR: MobileA3gent 是一款利用全球用户分布式自源数据训练移动 GUI 智能体的协作框架，通过 Auto-Annotation 组件在用户日常手机使用中低成本自动收集高质量数据集，借助 FedVLM-A 组件结合情节级和步骤级变异性的自适应全局聚合增强非独立同分布下的联邦 VLM 训练，仅需传统方法 1% 的成本便实现更优性能，其代码已公开，为解决移动 GUI 智能体训练的数据收集与隐私保护难题提供了方案。

初试分数: 2，4，6，6

置信度: 4，3，4，3

平均分:4.5

Title: RISK: A Framework for GUI Agents in E-commerce Risk Management

TLDR: We introduce RISK, a framework that enables GUI agents to automate complex, multi-step web interactions for e-commerce risk management.

初试分数: 4，4，2

置信度: 3，4，3

平均分:3.3

Title: M^2-Miner: Multi-Agent Enhanced MCTS for Mobile GUI Agent Data Mining

TLDR: We propose M-Miner, a Monte Carlo Tree Search-based collaborative multi-agent framework, which could efficiently mine GUI interaction trajectory data, thereby reducing the high cost of manual annotation.

初试分数: 4，4，6，4

置信度: 3，3，2，4

平均分:4.5

Title: GUI Knowledge Bench: Revealing the Knowledge Gap Behind VLM Failures in GUI Tasks

TLDR: 大型视觉语言模型（VLMs）在图形用户界面（GUI）任务自动化方面虽有进展但仍不及人类，研究团队提出该差距源于核心 GUI 知识缺失，将其提炼为界面感知、交互预测和指令理解三个维度，构建跨 6 大平台 292 款应用的 GUI Knowledge Bench 基准，评估证实当前 VLMs 在系统状态感知、动作预测等方面存在不足，且 GUI 知识与任务成功率密切相关，该研究为 VLMs 选型和更高效 GUI 智能体的构建提供了支撑。

初试分数: 2，6，4，6

置信度: 4，3，4，4

平均分:4.5

Title: HyperClick: Advancing Reliable GUI Grounding via Uncertainty Calibration

TLDR: HyperClick 通过不确定性校准与双重奖励机制，提升GUI智能体在 grounding 任务中的准确率与置信度一致性，缓解过度自信，实现更可靠的 GUI 自动化。

初试分数: 2，4，2，6，4

置信度: 4，3，2，3，4

平均分:3.6

Title: GAIR : GUI Automation via Information-Joint Reasoning and Group Reflection

TLDR: GAIR 通过引入通用MLLM融合多专业GUI模型信息并驱动其“群体反思”，整合异构能力，显著提升跨任务GUI 自动化性能与可靠性。

初试分数: 4，4，2，2

置信度: 3，3，4，3

平均分:3

Title: LightAgent: Lightweight and Cost-Efficient Mobile Agents

TLDR: LightAgent 通过“端侧小模型主导+云端大模型协同”与两阶段训练，在降低调用成本的同时实现接近大模型的手机GUI任务表现。

初试分数: 4，2，4，4

置信度: 3，4，4，4

平均分:3.5

Title: MobileWizard: A Data-Efficient GUI Agent with Structured Reasoning and Progressive Reinforcement Learning

TLDR: MobileWizard 凭借结构化思维链与渐进式强化学习，仅用 24.8k 轨迹训练的7B模型便在 AndroidWorld 上超越72B对手，实现数据高效的移动端GUI智能体。

初试分数: 6，6，2，2

置信度: 4，5，3，3

平均分:4

Title: OmniActor: A Generalist GUI and Embodied Agent for 2D&3D Worlds

TLDR: OmniActor通过“浅层共享、深层分离”的 Layer-heterogeneous MoE 结构化解 GUI 与具身数据冲突，统一动作空间并大规模训练，成为在两类任务上均领先的全能型多模态智能体。

初试分数: 10，4，4，6

置信度: 5，3，3，4

平均分:6 **MAX**

Title: A3: Android Agent Arena For Mobile GUI Agents

TLDR: A3通过 100 个在线真实 App 任务与“关键状态” MLLM 自动评估，弥补现有静态基准缺陷，为移动端 GUI 智能体提供动态、高效的实战评测平台。

初试分数: 4，6，4，2，2

置信度: 4，4，3，3，4

平均分:3.6

Title: GuirlVG: Incentivize GUI Visual Grounding via Empirical Exploration on Reinforcement Learning

TLDR: GuirlVG仅用5.2K样本、以强化学习+对抗KL因子稳定训练，就在GUI视觉定位上超越千万级SFT方法，为低成本高性能GUI-VG提供新范式。

初试分数: 6，6，4，4

置信度: 4，3，3，5

平均分:5 **MAX**

Title: Towards Adaptive GUI Agents with Memory-Driven Knowledge Evolution

TLDR: We propose MAGNET, a memory-driven framework that adapts mobile app agents to UI and workflow changes for robust long-term reliability.

初试分数: 4，2，4，2

置信度: 4，4，4，4

平均分:3

Title: WorldGUI: An Interactive Benchmark for Desktop GUI Automation from Any Starting Point

TLDR: We present a novel GUI benchmark called WorldGUI, which designs GUI tasks with various initial states to simulate authentic human-computer interactions.

初试分数: 2，4，2，6

置信度: 3，4，4，3

平均分:3.5

Title: MobileRL: Online Agentic Reinforcement Learning for Mobile GUI Agents

TLDR: MOBILERL 以“难度自适应 GRPO”在线强化学习，用重尾分布重采样与最短路径奖励塑形，把 9B 模型推至 AndroidWorld 80.2% 的新 SOTA，为移动端通用 GUI 智能体提供高效训练框架。

初试分数: 4，8，4，2

置信度: 4，5，4，3

平均分:4.5

Title: SWIRL: A Staged Workflow for Interleaved Reinforcement Learning in Mobile GUI Control

TLDR: We present SWIRL, a staged workflow for interleaved reinforcement learning designed for GUI multi-agent systems.

初试分数: 4，4，2，6，2

置信度: 4，2，4，3，4

平均分:3.6

Title: MemGUI-Bench: Benchmarking Memory of Mobile GUI Agents in Dynamic Environments

TLDR: MemGUI-Bench 首次系统评测移动端 GUI 智能体的跨时空记忆能力，用 128 个高内存依赖任务和 8 层"渐进审查"指标揭示现有 Agent 在记忆场景下 4–10× 性能暴跌，为构建类人长效 GUI 代理立下基准。

初试分数: 6，6，2

置信度: 4，3，4

平均分:4.7

Title: VEM: Environment-Free Exploration for Training GUI Agent with Value Environment Model

TLDR: VEM 用离线预训练的价值环境模型把策略优化与环境交互解耦，仅靠"动作是否推进目标"的语义估值就在 Android 与 Web 双平台达到 SOTA，实现零交互成本、跨布局鲁棒的 GUI 强化学习新范式。

初试分数: 4，4，2，4

置信度: 3，3，3，4

平均分:3

Title: Hijacking JARVIS: Benchmarking Mobile GUI Agents against Unprivileged Third Parties

TLDR: This study presents the first systematic investigation of mobile GUI agents' vulnerabilities to on-screen content manipulated by untrustworthy third parties.

初试分数: 4，4，2，4

置信度: 4，2，3，3

平均分:3.5

Title: WebFactory: Automated Compression of Foundational Language Intelligence into Grounded Web Agents

TLDR: An open-source, fully controllable offline web environment whose built-in site knowledge drives a pipeline to generate executable tasks and high-quality RL data, significantly boosting web-agent performance.

初试分数: 6，4，6

置信度: 4，4，4

平均分:5.3 **MAX**

Title: DKRF: Dynamic Knowledge Reasoning for Out-of-Distribution Generalization in Mobile GUI Agents

TLDR: Instead of just memorizing, our GUI agent learns how to reason with external knowledge to solve tasks in new environments.

初试分数: 4，4，4，4

置信度: 4，4，3，2

平均分:4

Title: GUIrilla: A Scalable Framework for Automated Desktop UI Exploration

TLDR: We present GUIrilla, an automated framework for macOS GUI exploration, and GUIrilla-Task, the first large-scale macOS dataset (27,171 tasks, 1,108 apps) pairing GUI screenshots with detailed accessibility metadata.

初试分数: 4，2，2，6

置信度: 3，4，4，4

平均分:3.5

Title: LPO: Towards Accurate GUI Agent Interaction via Location Preference Optimization

TLDR: We introduce Location Preference Optimization (LPO), a novel method that enhances GUI interactions by utilizing locational data and information entropy to improve spatial accuracy.

初试分数: 4，4，4，6

置信度: 2，4，2，3

平均分:4.5

Title: Automotive-ENV: Benchmarking Multimodal Agents in Vehicle Interface Systems

TLDR: ASURADA 借 GPS 上下文在首个车机 GUI 基准 Automotive-ENV 上完成 185 项安全与意图任务，验证“地理感知”对车载智能体决策的关键提升。

初试分数: 4，4，4，4

置信度: 3，1，3，4

平均分:4

Title: FingerTip 20K: A Benchmark for Proactive and Personalized Mobile LLM Agents

TLDR: FingerTip 20K 用 2 万条真实长期用户轨迹首创“主动建议+个性化执行”双赛道，暴露现有 GUI 智能体与人类巨大差距，为构建以用户为中心的移动 Agent 提供数据与基准。

初试分数: 8，4，6，4

置信度: 3，3，4，4

平均分:5.5 **MAX**

Title: GAMBIT: A Graph-structured and Decision-Aware Benchmark for MoBile GUI Tasks

TLDR: We introduce GAMBIT, a graph-structured benchmark for decision-aware mobile GUI agents, which reveals severe performance drops in long-horizon and branching tasks, providing a challenging diagnostic testbed for future agent development.

初试分数: 4，4，4，4

置信度: 4，4，4，3

平均分:4

Title: ReGUIDE: Data Efficient GUI Grounding via Spatial Reasoning and Search

TLDR: ReGUIDE 以自生成推理+空间自监督在线强化学习，用 0.2 % 数据量刷新网页 GUI 元素定位 SOTA，并借测试时空间搜索聚合实现低成本高精度 grounding。

初试分数: 4，6，4

置信度: 4，3，5

平均分:4.7

Title: Training a Vision-Language Model for Diverse Exploration in Open GUI World

TLDR: ScreenExplorer 以“好奇心+状态变化”双探索奖励和流式经验蒸馏，让 VLM 在无任何标注的全新应用中自主持续交互，突破静态数据依赖，实现 GUI 智能体的终身式泛化探索。

初试分数: 6，2，4，4

置信度: 3，4，4，4

平均分:4

Title: GTA1: GUI Test-time Scaling Agent

TLDR: GTA 通过“测试时并行采样-评判选优”解决大动作空间规划，再以强化学习精修定位，实现规划与点击双 SOTA 的跨平台 GUI 智能体。

初试分数: 4，4，8，6

置信度: 4，4，3，5

平均分:5.5 **MAX**

Title: LongHorizonUI: A Unified Framework for Robust long-horizon Task Automation of GUI Agent

TLDR: LongHorizonUI integrates element-indexed multimodal perception, hierarchical reflective decision-making, and rollback-based compensatory execution for long-horizon GUI control.

初试分数: 6，6，2，4

置信度: 5，3，4，3

平均分:4.5

Title: ScaleCUA: Scaling Open-Source Computer Use Agents with Cross-Platform Data

TLDR: ScaleCUA 开源最大规模跨 6 系统 GUI 数据集并训练统一模型，以 +26.6 等平均增幅刷新三项基准 SOTA，验证“数据即算力”对通用计算机使用智能体的放大效应。

初试分数: 6，6，6，6

置信度: 4，1，5，2

平均分:6 **MAX**

Title: Generalization in Online Reinforcement Learning for Mobile Agents

TLDR: AndroidWorld-Generalization 以 CMDP 形式化移动 GUI 泛化挑战，配套开源 GRPO-RL 框架与三阶基准显示：7B 模型 RL 训练后在未见实例提升 26.1%，但跨模板/跨应用增益有限，为后续小样本在线适应指明方向。

初试分数: 4，2，6

置信度: 5，3，4

平均分:4

Title: PSBench: Editing Image via GUI Agents in Photoshop

TLDR: PSBench 首发面向 Photoshop 的 600 分层非破坏编辑任务基准，揭示顶尖 MLLM 成功率虽低却能显著加速新手实操，为专业图形软件智能体奠定评测与辅助设计新标杆。

初试分数: 4，6，4

置信度: 3，4，2

平均分:4.7

Title: PrecogUI: Proactive GUI Agents via Pre-cognitive Simulation and Experience Retrieval

TLDR: PrecogUI recast GUI agents from reactive to foresight-driven decision-makers that anticipate disturbances and close the loop on errors, yielding state-of-the-art robustness and success on long-horizon, dynamic tasks.

初试分数: 6，4，2，4，4

置信度: 2，3，4，4，4

平均分:4

Title: Let's Think in Two Steps: Mitigating Agreement Bias in MLLMs with Self-Grounded Verification

TLDR: SGV 用“自设标准再评判”两步法破解 MLLM 评审“同意偏差”，在 Web/机器人/GUI 三大任务上平均提升 20%，并同步开源 10× 加速的 (Visual)WebArena，为数字智能体提供实时、可靠、可扩展的奖励信号。

初试分数: 4，8，4，6

置信度: 3，3，3，4

平均分:5.5 **MAX**