

OpenMobile: 基于任务与轨迹合成构建开放移动智能体

Kanzhi Cheng^{123*} Zehao Li⁴ Zheng Ma^{2†} Nuo Chen¹ Jialin Cao¹
 Qiushi Sun⁵ Zichen Ding⁴ Fangzhi Xu³⁶ Hang Yan⁶ Jiajun Chen¹
 Luu Anh Tuan³ Jianbing Zhang^{1‡} Lewei Lu^{2‡} Dahua Lin²

¹Nanjing University ²SenseTime ³Nanyang Technological University

⁴Shanghai AI Laboratory ⁵The University of Hong Kong ⁶Xi'an Jiaotong University

Abstract

由视觉语言模型驱动的移动智能体在自动化移动任务方面展现出了出色的能力，近期的主流模型在性能上实现了显著跃升，例如在 AndroidWorld 上的成功率接近 70%。然而，这些系统的训练数据并未开源，且其任务与轨迹生成方法也不透明。我们提出 OpenMobile，一个能够合成高质量任务指令与智能体轨迹的开源框架，其包含两个核心组件：(1) 第一个是可扩展的任务合成流水线：从探索过程中构建全局环境记忆，随后利用该记忆生成多样化且贴合实际的任务指令。(2) 用于轨迹展开的策略切换策略。通过在学习器与专家模型之间交替切换，该组件能够捕获标准模仿学习中通常缺失的关键错误恢复数据。在我们的数据上训练的智能体在三个动态移动智能体基准测试中取得了具有竞争力的结果：值得注意的是，我们微调后的 Qwen2.5-VL 和 Qwen3-VL 在 AndroidWorld 上分别达到 51.7% 和 64.7%，远超现有的开源数据方法。此外，我们对合成指令与基准测试集之间的重叠部分进行了透明分析，并验证了性能提升来源于更广泛的功能覆盖，而非基准测试过拟合。我们将数据与代码发布至 [OpenMobile](#)，以填补数据缺口并推动更广泛的移动智能体研究。

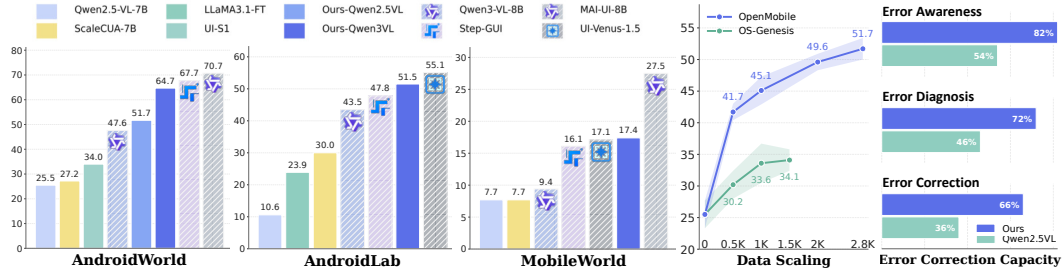


图 1: **性能对比**. 三个动态移动智能体基准上的任务成功率。我们的模型显著超越了开放数据基准，且与领先的闭源数据系统具有竞争力。**数据缩放**. 随着合成指令增加，AndroidWorld 的性能表现。**错误纠正容量**. OpenMobile 数据显著提升了智能体在真实环境中的错误恢复能力。

1 引言

视觉-语言模型 (VLMs) 推动了移动智能体的快速发展——这些自主系统能够与智能手机界面交互以完成用户任务。近期业界的努力，如 Step-GUI (Yan et al., 2025)、MAI-UI (Zhou

* 在商汤科技实习期间完成的工作。† 项目负责人。‡ 通讯作者。

et al., 2025)、UI-Venus-1.5 (Gao et al., 2026) 以及 MobileAgent-v3.5 (Xu et al., 2026), 已将技术水平提升至令人瞩目的程度, 例如在广泛采用的 AndroidWorld 基准测试 (Rawles et al., 2024) 上接近 70% 的任务成功率。这些进展本质上是由大规模、高质量智能体训练数据所驱动的, 这些数据由任务指令与执行轨迹配对构成。

然而, 这些领先的系统均将其轨迹数据闭源, 且对于此类数据的生成方式始终保持不透明。与此同时, 开源社区仅能依托 AndroidControl (Li et al., 2024) 和 AMEX (Chai et al., 2025) 这类公开数据集, 在同一基准 (Lu et al., 2025b; Liu et al., 2025) 上的性能仅约为 30%。除了日益扩大的性能差距之外, 这种不透明性还阻碍了社区弄清究竟是哪些数据属性推动了优异的性能与泛化能力——这使得开源社区无法对这些进展开展研究、复现或在此基础上进行拓展。

为弥合这一差距, 我们提出了 OpenMobile, 这是一个面向移动智能体训练的开源数据合成框架与数据集。OpenMobile 解决了两个核心挑战: (1) **在动态移动环境中大规模生成多样化且高质量的任务指令**。现有方法通常通过使用单一轨迹作为上下文, 让大语言模型进行任务筛选来实现探索与生成的耦合 (Sun et al., 2025a; Murty et al., 2024)。这种依赖关系限制了多样性, 仅能反映单一本地轨迹所揭示的信息。我们则采用解耦两阶段的方法: 首先探索环境以构建应用功能的全局环境记忆; 随后结合邻近界面的短期记忆以及应用内语义相关功能的长期记忆, 组合生成复杂、多步骤的任务指令。(2) **收集能够提供有效训练信号的智能体轨迹**。专家轨迹蒸馏使学习器能够模仿理想行为, 但往往无法解决错误恢复问题, 导致测试时出现显著性能差距。尽管自演化可缓解此差异, 但其常面临收敛缓慢的问题, 且受限于学习器当前的性能上限。为此, 我们进一步引入一种策略切换机制, 在推理过程中交替使用学习器与专家模型。我们发现, 错误干预切换策略 (即监控器检测偏差并触发专家纠正) 有助于合成具备错误恢复能力的示范轨迹, 同时保持任务成功完成率。

使用 OpenMobile, 我们在 20 个 Android 应用上合成 2.8K 个任务指令, 对应 34K 个动作步骤。我们在三个已建立的在线基准 (即 AndroidWorld、AndroidLab (Xu et al., 2025) 和 MobileWorld (Kong et al., 2025)) 上进行了全面评估。值得注意的是, 我们微调后的 Qwen2.5-VL-7B 和 Qwen3-VL-8B 在 AndroidWorld 上分别达到 51.7% 和 64.7% 的任务成功率, 并在具有挑战性的 MobileWorld 上从 9.4% 提升至 17.4%, 表现与领先的闭源解决方案和更大规模的基础模型相当 (Bai et al., 2025)。除了原始性能外, 我们通过透明实验回应了社区对潜在数据污染的日益关注, 证实我们的性能提升源于广泛的功能覆盖和增强的错误恢复能力, 而非基准过拟合。这些发现为开源社区构建具有竞争力的移动智能体提供了切实可行的基础。

我们的贡献总结如下:

- 我们提出并开源了 OpenMobile, 这是一个面向移动智能体的任务与轨迹合成框架。该框架引入了解耦的任务合成机制, 用于构建全局环境记忆以生成指令, 并采用策略切换的轨迹滚动方法, 捕捉专家蒸馏中缺失的纠正信号。
- 我们在三个具有挑战性的动态基准上进行了全面评估。在我们的数据上训练的智能体, 其表现与闭源数据系统相当。
- 我们提供了系统的分析, 以检验数据污染风险, 并证明我们的性能提升源于广泛的功能覆盖和增强的错误恢复能力, 而非基准过拟合。

2 相关工作

为数字自动化构建自主智能体是人工智能与自然语言处理领域长期追求的目标 (Branavan et al., 2009; Shi et al., 2017; Shaw et al., 2023)。大模型的最新突破显著加速了这一方向的进展, 使智能体能够在移动、网页和桌面环境中进行规划与操作, 并逐步应对日益复杂的任务 (Rawles et al., 2024; Zhou et al., 2023; Xie et al., 2024; Sun et al., 2025c)。

具备视觉-语言模型的数字智能体。 早期工作依赖大模型与结构化界面表示 (如可访问性树) 进行交互 (Deng et al., 2023; Gur et al., 2023), 或构建通过编码操作计算机的智能体框架 (Wu et al., 2024; Sun et al., 2024)。视觉-语言模型的快速发展推动了向端到端、以视觉为中心的 GUI 智能体的转变 (Cheng et al., 2024; Gou et al., 2024; Wu et al., 2025b;a)。这些智能体以原始屏幕截图作为输入, 通过人类类似的动作 (如点击和输入) 完成任务。其中, 专有系统如 Operator (OpenAI, 2025) 和 Anthropic 的 Computer-Use (Anthropic, 2024) 尤为突出, 它们通过利用前沿基础模型 (Yang et al., 2026a) 实现了令人瞩目的性能。与此同时, UI-TARS (Qin et al., 2025; Wang et al., 2025) 通过 GUI 预训练、轨迹微调和在线强化学习, 为开源权重智能体树立了里程碑。最近, 产业界的努力 (Yan et al., 2025; Zhou et al., 2025; Gao et al., 2026; Xu et al., 2026) 进一步提升了移动端智能体的性能, 在 AndroidWorld 上实现了 70% 的任务成功率。这些进展的核心在于大规模合成的任务指令和智能体轨迹; 然而, 相关数据及其潜在的合成方法仍未公开。

另一方面, 开源社区的进展日益落后。人类标注的数据集, 如 AndroidControl (Li et al., 2024) 和 AMEX (Chai et al., 2025), 已提供了基础 (Rawles et al., 2023; Lu et al., 2025a; Yang et al., 2026b; Sun et al., 2025b)。然而, 这些数据包含显著的标注噪声, 且缺乏丰富的思维模式。在它们上训练的模型, 如 ScaleCUA (Liu et al., 2025) 和 UI-S1 (Lu et al., 2025b), 在 AndroidWorld 上的表现大致停留在 30% 的水平。迫切需要可扩展的、开源的数据合成方案, 以弥合这一差距。

GUI 数据合成。 GUI 智能体数据的人工标注成本高且耗时, 这促使人们越来越关注任务指令和动作轨迹的自动化合成。早期方法采用任务驱动范式, 利用强大的语言模型从种子指令和应用描述中提出任务 (He et al., 2024; Lai et al., 2024)。尽管这种方法简单直接, 但缺乏对真实世界情境的锚定, 常常生成通用、描述不充分或不可行的指令。这推动了交互驱动方法的发展, 这些方法首先探索目标环境, 然后基于观测到的上下文合成与环境相关的指令。代表性工作如 OS-Genesis (Sun et al., 2025a), 提出逆向任务合成, 利用随机游走轨迹回溯推断出有意义的任务指令。NNetNav (Murty et al., 2024) 通过探索策略与剪枝标注器之间的协同作用, 高效构建复杂的网页演示, 该剪枝标注器可过滤低质量轨迹。后续研究进一步在此范式上推进, 采用了更结构化的探索策略 (Yang et al.; Gandhi & Neubig, 2025; Jiang et al., 2026) 和更复杂的指令生成流水线 (Xie et al., 2025; Pahuja et al., 2025; Ramrakhya et al., 2025)。在这些方法中, 探索与指令生成仍紧密耦合: 每条指令均源自单一探索轨迹, 这限制了多样性, 使其局限于局部观测。

一旦任务指令可用, 下一步便是收集高质量的智能体轨迹。一种流行的方法是专家蒸馏, 即由一个强大的智能体模型生成轨迹, 再由验证器模型对轨迹进行质量筛选 (Pan et al., 2024; Sun et al., 2025a; Lin et al., 2025)。另一种方法探索自演化机制, 即智能体迭代执行任务, 并基于自身成功轨迹进行重训练, 以实现性能的自我提升 (He et al., 2025; Qin et al., 2025)。近期的研究进一步在生成指令的同时共同生成可验证的评估脚本, 以促进强化学习训练 (Xue et al., 2026)。

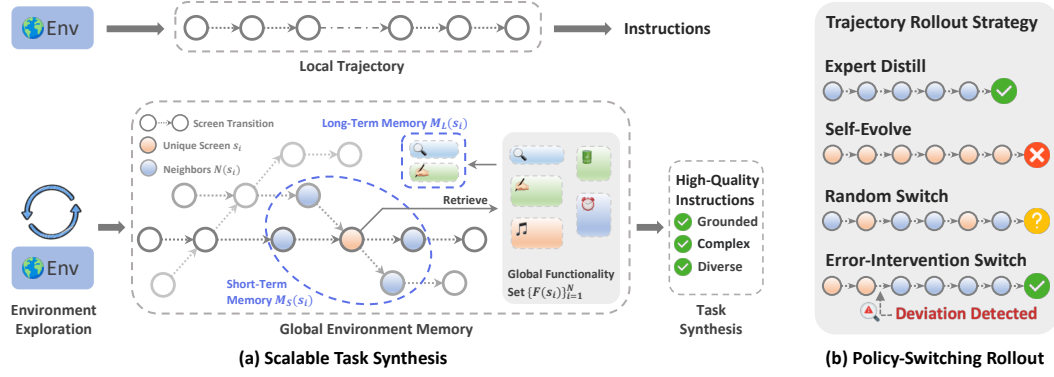


图 2: OpenMobile 概览。(a) 可扩展的任务合成。我们不依赖单一的本地轨迹，而是首先探索环境以构建全局记忆。通过检索短期和长期记忆，我们合成出多样且复杂的指令，这些指令具有上下文相关性，并可在在线环境中合理执行。(b) 策略切换的滚动执行。在不同的滚动策略中，错误干预切换策略能够最好地捕捉错误恢复信号，同时通过检测学习器的偏离行为并触发专家纠正，确保任务顺利完成。

3 开放移动

专有与开源移动智能体之间的性能差距，其根本原因在于缺乏大规模、高质量的开源训练数据。为弥合这一差距，OpenMobile 提出了一种数据合成框架，生成两种互补资产：覆盖移动环境广泛功能的多样化、基于实际场景的任务指令（Section 3.1），以及包含错误恢复信号的智能体执行轨迹，有助于实现高效的智能体训练（Section 3.2）。我们方法的概览见 Figure 2。实现细节详见 Appendix A。

3.1 可扩展的任务合成

现有的交互驱动方法将探索与生成紧密耦合，从单个探索轨迹中提取任务指令，这使得指令的多样性受限于单一局部轨迹所揭示的内容。我们提出了一种解耦范式，受到人类学习新应用方式的启发：首先进行探索以构建对应用程序功能的结构化、全面理解，然后在面对复杂需求时，回忆并组合相关功能。我们的流水线模拟这一过程，分为三个阶段：(i) 探索环境以收集交互经验，(ii) 将探索数据组织为全局环境记忆 \mathcal{M} ，以及 (iii) 从 \mathcal{M} 的短期记忆和长期记忆中调用信息，以合成组合式任务指令。

环境探索 第一阶段旨在遍历目标应用以收集关于环境的信息。通过与应用的连续交互，该阶段生成一组探索轨迹——即捕捉不同状态之间转移的屏幕-动作交互序列。我们的框架对具体的探索策略具有无关性，可兼容随机游走 (Sun et al., 2025a)、基于结构覆盖的方法 (Gandhi & Neubig, 2025; Ramrakhya et al., 2025; Shao et al., 2026)，或人类示范。为证明我们的方法不依赖复杂的探索启发式策略，本文采用简单的随机游走。这一选择源于我们的观察：关键因素并不在于探索效率，而在于所收集数据在下游环节中组织和利用的有效性。

全球环境记忆构建 多次探索会话不可避免地会在不同轨迹中访问相同的屏幕或环境状态。我们利用这些共享屏幕作为自然的锚点，将所有碎片化的轨迹编织成一个统一且相互关联的结构。具体而言，我们采用感知哈希对视觉上相似的屏幕进行聚类，识别出一组 N 唯一屏幕 $\mathcal{S} = \{s_1, s_2, \dots, s_N\}$ ，并聚合所有轨迹中的转移信息以形成邻域关系：对于每个屏幕 s_i ，其邻居 $\mathcal{N}(s_i) \subset \mathcal{S}$ 是从 s_i 直接可达或可到达 s_i 的屏幕。随后，我们为每个屏幕丰富其功能集合 $\mathcal{F}(s_i) = \{f_1, f_2, \dots, f_K\}$ ，该集合由强大的视觉-语言模型提取，其中每个 f_k 是一段自

然语言描述，用于捕捉屏幕上用户界面元素的语义。为支持跨屏幕关联，我们计算所有功能的语义嵌入，并构建每个应用的检索索引。最终形成的全局环境记忆

$$\mathcal{M} = (\mathcal{S}, \mathcal{N}, \{\mathcal{F}(s_i)\}_{i=1}^N)$$

以结构化、可查询的形式捕捉每个应用程序的功能布局。

增强记忆的任务合成 给定全局记忆 \mathcal{M} ，任务合成通过将功能关联并组合成连贯的指令来推进。对于每个候选屏幕 s_i ，我们构建一个上下文 $\mathcal{C}(s_i)$ ，包含三个互补视角：(1) 被召回的屏幕 s_i 本身，包括其截图和标注的功能 $\mathcal{F}(s_i)$ ，作为生成的焦点；(2) 短期记忆 $\mathcal{M}_S(s_i)$ ，邻近屏幕 $\mathcal{N}(s_i)$ 的功能，反映当前屏幕能力在局部范围内可访问且自然可链接的能力；以及 (3) 长期记忆 $\mathcal{M}_L(s_i)$ ，从同一应用中远距离屏幕检索到的语义相关功能，揭示用户可能通过经验关联但单个轨迹无法展现的特征，从而鼓励跨特征组合。

完整的上下文 $\mathcal{C}(s_i) = (s_i, \mathcal{M}_S(s_i), \mathcal{M}_L(s_i))$ ，包括截图和文本描述，被提供给视觉语言模型 (VLM) 以基于该上下文生成任务指令。我们还精心设计了生成指南和上下文示例，以引导模型生成高质量的指令；完整的提示内容见 Section A.5。生成的指令经过基于模型的质量过滤和基于嵌入的去重处理，最终得到指令集。

3.2 策略切换发布

手握任务指令后，下一步是收集智能体轨迹用于训练。一种直接的方法是专家蒸馏：运行一个强大的模型，生成示范轨迹以供模仿学习使用。尽管这种方法能产生高质量的轨迹，但会将学习器限制在理想行为范围内，无法让其暴露于推理过程中可能犯的错误，导致智能体无法从自身错误中恢复。另一种方法是自演化，即学习器迭代执行任务，并基于自身的成功轨迹进行重训练。这种方法直接解决了分布不匹配问题，但由于学习器的提升受限于其自身容量，收敛速度较慢。

我们提出策略切换回溯 (policy-switching rollout)，该方法结合了两种范式的优点。给定一个任务指令 I ，一个专家策略 π_e ，以及一个学习器策略 π_l ，回溯过程逐步进行：在每个时间步 t ，智能体观察当前屏幕 o_t ，并根据两个策略之一选择动作 a_t ，具体由一个切换变量 $z_t \in \{e, l\}$ 控制：

$$a_t \sim \pi_{z_t}(\cdot \mid I, o_t, h_t)$$

其中 h_t 表示到步骤 t 为止的交互历史。通过在 π_e 与 π_l 之间交替进行，生成的轨迹包含学习器出错的片段，随后由专家进行纠正。这产生了纯蒸馏中所缺乏的错误恢复经验，同时专家的存在避免了自演化固有的容量上限。

切换策略。 切换规则的设计，即每一步如何确定 z_t ，至关重要。一个自然的选择是随机切换，当两个策略对下一步动作存在分歧时，以固定概率 p 让学习器接管。然而，移动智能体任务本质上是多解的：多个有效的动作序列均可达到同一目标，因此 π_e 与 π_l 之间的分歧并不一定意味着学习器出错，使得随机切换成为识别错误的一个噪声较大的代理。此外，策略间的频繁切换往往会干扰复杂任务上的连贯进展，导致轨迹碎片化，难以提供有效的训练信号。

为解决此问题，我们设计了一种误差干预策略。与在每个分歧点都进行切换不同，我们引入一个监控器 \mathcal{O} ，实时跟踪学习器的执行过程。轨迹的生成始于学习器策略 ($z_t = l$)；只有当 \mathcal{O} 检测到学习器偏离了有效进展时，才会触发切换至专家 ($z_t = e$)，以介入并纠正轨迹回到正轨。由此产生的轨迹因此包含了稀缺的错误恢复经验，而专家干预则确保了足够的任务完成度，从而实现有效的训练。我们在 Section 5.1 中对比了这些切换策略。

3.3 开放移动数据集

我们将在 AndroidWorld (Rawles et al., 2024) 提供的 Android 模拟器上实例化上述流水线。尽管我们利用了其环境基础设施，但在合成过程中并未引入任何基准测试指令，以防止数据泄露；详细的重叠分析见 Section 5.2。在策略切换的滚动执行中，一个早期阶段微调的检查点作为学习器 π_l ，而 Gemini-3.1-Pro-Preview 作为专家 π_e 。我们采用 Qwen3-VL (Bai et al., 2025) 的动作空间和响应格式，并使用专家模型对每一步的思维链推理进行重写，以获得更高质量的监督。生成的数据集包含约 2,800 条指令和 34K 个对应的动作步骤，覆盖 20 个 Android 应用，平均每条轨迹长度为 12.2 步，每步包含 129 词的思维链推理。

4 实验

在本节中，我们在 OpenMobile 数据集上训练模型，并在已建立的动态基准上展示主要结果。消融研究和进一步分析留到下一节。

4.1 实验情景

从策略切换的滚动轨迹中，我们仅保留专家步骤用于训练，同时保留完整的交互历史（包括学习器错误），作为上下文以使模型暴露于真实的错误恢复场景中。我们对两个基础模型进行微调：Qwen2.5-VL-7B 和 Qwen3-VL-8B (Bai et al., 2025)。前者未经大量针对 GUI 的后训练，提供了更干净的测试环境，有助于分离数据驱动的增益；后者则作为更强的基础模型，用于在更具能力的基础模型上验证有效性，并推动性能上限。所有模型均使用 LLaMA-Factory (Zheng et al., 2024) 进行标准监督微调，批量大小为 32，学习率为 $1e-5$ ，共进行 3 轮次训练。

我们还对强化学习 (RL) 进行了实验，包括步骤级强化学习 (Lu et al., 2026) 和轨迹级 Agentic 强化学习 (Li et al., 2026)。尽管结果表明我们合成的轨迹仍然有效，但在动态基准上并未显著优于 SFT。我们在 Appendix C 中讨论了这些发现。

4.2 评估基准

我们在三个已建立的动态移动智能体基准上进行评估。我们重点关注动态基准，因为静态数据集（如 AndroidControl）从根本上缺乏评估智能体关键错误恢复能力的能力，且其标注噪声进一步降低了评估的可靠性，使其无法真实反映智能体在现实世界中的表现 (Lu et al., 2025b; Gao et al., 2026)。实验细节见 Section B.1。

AndroidWorld (Rawles et al., 2024) 是主流的移动智能体评估基准。它通过 Android 模拟器提供稳健、可复现的环境以及确定性评估。该基准包含 20 个真实世界应用中的 116 个任务，采用参数化任务模板，通过随机种子生成多样化的任务变体。

AndroidLab (Xu et al., 2025) 是一个针对移动智能体的系统性基准，支持可复现的评估。它涵盖了在预定义的 Android 虚拟设备上 9 个应用的 138 个任务，并支持仅语言和多模态智能体。

MobileWorld (Kong et al., 2025) 是一个近期推出且更具挑战性的动态基准。它包含 20 个应用中的 201 项任务，更加强调长时程和跨应用工作流，使得其在评估复杂移动智能体能力方面比 AndroidWorld 难度显著提升。我们在其仅含 GUI 的子集上进行评估。

Method	Base Model	AndroidWorld		AndroidLab		MobileWorld	
		Pass@1↑	Pass@3↑	Pass@1↑	Pass@3↑	Pass@1↑	Pass@3↑
Commercial Models							
GPT-4o	–	30.6	–	31.2	–	–	–
Gemini-3-Pro	–	60.3	75.0	–	–	51.3	–
Open-Weight Models							
Qwen2.5-VL-7B	–	25.5 ± 2.6	34.9	10.6 ± 1.8	15.2	7.7 ± 0.9	10.3
Qwen3-VL-8B	–	47.6 ± 2.2	62.1	43.5	–	9.4	–
UI-Venus-7B	Qwen2.5-VL	49.1	–	41.3	–	8.5	–
Step-GUI-4B	Qwen3-VL	63.9	75.8	47.8	–	16.1	–
Step-GUI-8B	Qwen3-VL	67.7	80.2	–	–	–	–
MAI-UI-8B	Qwen3-VL	70.7	–	–	–	27.5	–
UI-Venus-1.5-8B	Qwen3-VL	73.7	–	55.1	–	17.1	–
MobileAgent-v3.5-8B	Qwen3-VL	71.6	–	–	–	33.3	–
Open-Data Models							
UI-S1-7B	Qwen2.5-VL	34.0	–	–	–	–	–
ScaleCUA-7B	Qwen2.5-VL	27.2 ± 2.2	36.2	30.0 ± 1.1	37.7	7.7 ± 0.4	8.6
Ours-7B	Qwen2.5-VL	51.7 ± 1.7	68.1	22.7 ± 0.4	37.0	14.8 ± 1.3	21.4
Ours-8B	Qwen3-VL	64.7 ± 3.2	78.0	51.5 ± 0.7	62.3	17.7 ± 2.2	24.8

表 1: AndroidWorld、AndroidLab 和 MobileWorld 上的主要结果。我们报告 Pass@1 和 Pass@3，数值越高表示性能越好。OpenMobile 显著优于开放数据基准，并与领先的闭源数据系统相当。

4.3 主要结果

OpenMobile 数据显著提升了移动智能体的性能，并具备强大的泛化能力。如 Table 1 所示，基于 OpenMobile 数据微调的模型在所有三个基准测试中均显著优于相应的基准模型。Qwen2.5-VL 变体在 AndroidWorld 上的性能提升超过 25 个百分点，证明了我们合成轨迹在增强视觉语言模型（VLM）智能体能力方面的有效性。值得注意的是，尽管 OpenMobile 数据是在 AndroidWorld 环境中收集的，但由此产生的模型在未见过的情景中仍表现出良好的泛化能力，包括 AndroidLab 中的新应用以及 MobileWorld 中的长周期跨应用任务，例如在后者上实现了超过 50% 的相对性能提升。总体而言，我们的模型显著超越现有的开源数据方法，并具备与领先工业方案相媲美的竞争力。这些结果凸显了开放数据合成在构建具有竞争力的移动智能体方面的潜力。

基础模型能力仍然至关重要。 尽管在相同数据上进行训练，Qwen3-VL 变体始终以明显优势超越 Qwen2.5-VL 变体，表明基础模型的内在能力（例如 GUI 理解和规划）起到了不可或缺的作用。虽然高质量轨迹数据可以缩小差距，但提升潜在的基础模型仍同样重要，以推动性能上限。更多关于更大模型的实验以及与其他方法的对比详见 Appendix D。

5 分析

在本节中，我们首先对 OpenMobile 背后的关键设计选择进行消融实验（Section 5.1），然后探究其有效性的驱动因素（Section 5.2），包括对潜在基准过拟合的考察。

OpenMobile vs.	Complexity	Soundness	Method	Pass@1↑
OS-Genesis	0.68 / 0.22 / 0.10	0.44 / 0.48 / 0.08	OS-Genesis	34.1 ± 1.7
Coupled Pipeline	0.26 / 0.62 / 0.12	0.06 / 0.90 / 0.04	Coupled Pipeline	45.3 ± 2.2
			OpenMobile	48.3 ± 1.3

(a) Human evaluation (win / tie / loss).

(b) Task success rate.

表 2: 任务合成策略的消融实验。(a) 通过 50 次成对比较的人工评估指令质量。(b) Android-World 的成功率, 基于 1.5K 条轨迹。

5.1 消融研究

OpenMobile 生成多样且高质量的指令。我们对比了我们的解耦任务合成方法与 OS-Genesis (Sun et al., 2025a) 以及一种耦合基准方法。该耦合基准采用与我们相同的生成提示, 但使用单次探索轨迹的截图序列作为上下文, 而非全局环境记忆。我们首先从每种方法中抽取 50 条指令进行成对型人工评估, 由经验丰富的标注者根据复杂性和合理性判断指令质量, 并选择更优的一项或判定平局。如 Table 2a 所示, OpenMobile 合成的指令明显比两种基准方法更具挑战性, 同时保持了相当的合理性。我们进一步在合成数据上训练模型, 并在 AndroidWorld 上进行评估。Table 2b 的结果表明, 在固定 1.5K 轨迹预算下, OpenMobile 取得了最佳性能。

策略切换的逐步部署增强了错误恢复信号, 并提升了测试时的性能。 我们比较了四种轨迹展开策略: (i) 专家蒸馏, 仅使用专家模型收集轨迹; (ii) 自我演化, 学习器在 3 轮迭代中基于自身成功轨迹进行重训练; (iii) 随机切换, 如 Section 3.2 所述, 在专家与学习器之间随机交替; (iv) 我们的误差干预切换。详细的实验情景见 Section B.2。

如 Table 3 所示, 在轨迹滚动过程中引入更丰富的错误恢复信号, 误差干预切换实现了最佳的下游性能。此外, 如 Figure 1 右侧面板所示, 我们对比了训练后模型与基础模型在实时执行过程中的错误恢复行为, 包括错误感知、诊断和纠正。结果表明, OpenMobile 数据显著增强了智能体的错误恢复能力, 从而推动了下游性能的提升。

Rollout Strategy	Avg. ER	Pass@1↑
Expert Distillation	0.42	44.8 ± 1.7
Self-Evolution	0.10	33.8 ± 0.9
Random Switch	0.64	45.1 ± 0.9
Error-Intervention Switch	1.56	48.3 ± 1.3

表 3: 滚动策略的消融实验。平均 ER 是每条轨迹中错误恢复实例的平均数量, 基于 50 个随机采样的轨迹手动统计得出。

5.2 什

么驱动了 OpenMobile 数据的有效性?

OpenMobile 数据基于基准环境, 但不会对其测试指令产生过拟合。由于我们的数据是

在 AndroidWorld 环境中合成的, 一个自然的担忧是这些指令是否仅仅是基准测试的简单复制。为了探究这一问题, 我们使用 openai/text-embedding-3-large 计算合成指令与 Android-World 测试指令之间的语义相似度, 并与 AndroidControl 和 AMEX 进行比较。如 Figure 3 (左图) 所示, OpenMobile 的指令确实与测试集更相似, 这在共享环境 and 应用套件的情况下是意料之中的。然而, 仅有 3.5% 的指令相似度超过 0.7, 表明其相关性适中而非任务级别的重复, 从而缓解了数据泄露的担忧, 例如通过改写测试指令的方式。最相似指令对的完整列表见 Appendix E。

此外，我们通过移除与测试集最相似的合成指令，并与随机移除进行对比，以观察对下游性能的影响。如 Figure 3 (右) 所示，移除少量样本（例如 10%）仅导致性能轻微下降，表明我们的提升并不高度依赖于少数与测试集相似的样本。然而，当移除比例增加至 40% 时，性能下降明显高于随机移除的情况。这是因为移除最相似的指令不可避免地会从训练数据中剥离核心应用功能，导致模型无法获取必要的技能。我们在下一节中将详细分析功能覆盖的作用。

广泛的功能覆盖推动智能体性能提升。 为了理解 OpenMobile 数据有效性的原因，我们进行了功能覆盖分析。我们使用大语言模型（LLM）将每个测试任务分解为所需的原子功能。例如，“创建一个标题为“与团队明天上午 10 点开会”的日历事件”可分解为创建日历事件、设置日期、设置标题和设置开始时间。随后，我们衡量合成指令所覆盖的测试所需功能的比例。如 Figure 4 (左图) 所示，随着指令数量的增加，覆盖度稳步上升，且 OpenMobile 始终优于耦合流水线。这证实了我们解耦设计的优势：全局记忆提供了对环境能力的结构化视图，而检索语义相关的功能作为长期上下文，则促进了跨特征组合，共同推动了更广泛且更多样化的指令生成。

我们进一步考察任务复杂度（即每个任务的原子功能数量）与功能覆盖度如何共同影响成功率。如 Figure 4 (右图) 所示，涉及更多功能的任务更难完成（颜色从上到下逐渐变浅），而覆盖度更高的任务则取得更高的成功率（颜色从左到右逐渐加深）。这突显了功能覆盖度在指令合成中的重要性：一种有效的合成方法应尽可能最大化对环境核心功能的覆盖，而这正是 OpenMobile 的设计原则所在。

6 结论

我们提出了 OpenMobile，一个用于构建具有竞争力的移动智能体的开放数据合成框架。该框架解决了轨迹合成中的两个关键问题：（1）将探索过程与指令生成解耦，以生成多样化且高质量的任务；（2）采用策略切换的滚动机制，为轨迹注入错误恢复信号。在 OpenMobile 数据上训练的智能体表现出优异性能，并能良好泛化至未见过的动态环境，显著缩小了与闭源工业系统之间的差距。我们还对合成指令与测试指令之间的重叠情况进行了透明分析，确认这些提升源于广泛的功能覆盖和增强的错误恢复能力，而非基准过拟合。我们已公开所有数据与代码，期望 OpenMobile 能成为更广泛的开放移动智能体研究的基础。

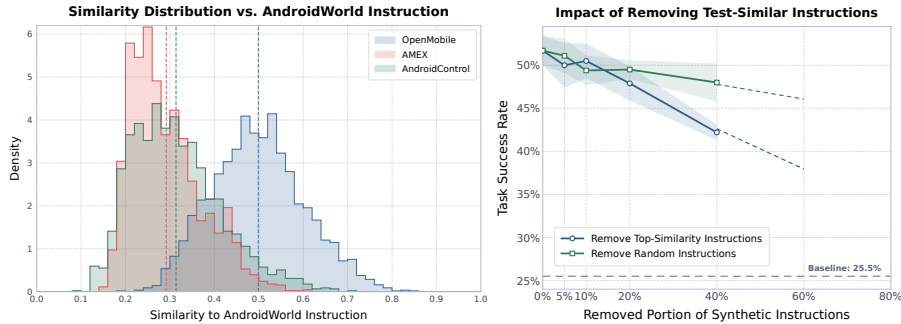


图 3: 左图: 合成指令与 AndroidWorld 指令之间的语义相似度。我们的合成指令在功能层面表现出中等的相关性，仅有 3.5% 的指令相似度超过 0.7。右图: 移除训练集中与测试集相似指令的影响。仅移除少量最相似的指令便导致性能下降幅度极小，有效缓解了基准过拟合的担忧。

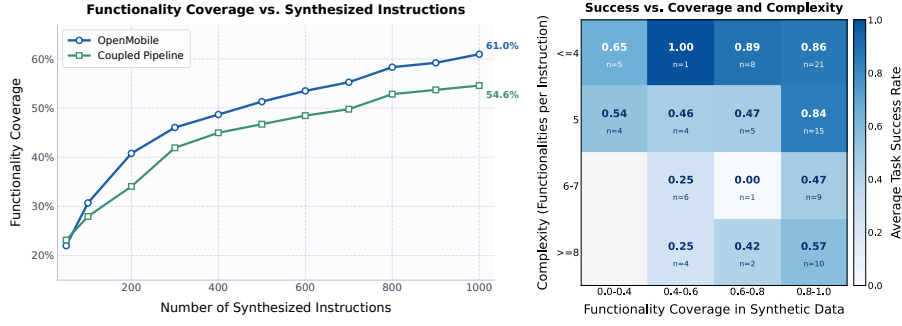


图 4: 左图: 随着合成指令规模的增加, AndroidWorld 任务的功能覆盖情况。OpenMobile 始终比耦合基准实现更高的覆盖度。右图: 功能需求较少(所需功能较少)且合成数据实现较高功能覆盖的任务, 其成功率更高。

参考文献

- Anthropic. Introducing computer use, 2024. URL <https://www.anthropic.com/news/3-5-models-and-computer-use>.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. arXiv preprint arXiv:2511.21631, 2025.
- Satchuthananthavale RK Branavan, Harr Chen, Luke Zettlemoyer, and Regina Barzilay. Reinforcement learning for mapping instructions to actions. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 82–90, 2009.
- Yuxiang Chai, Siyuan Huang, Yazhe Niu, Han Xiao, Liang Liu, Guozhi Wang, Dingyu Zhang, Shuai Ren, and Hongsheng Li. Amex: Android multi-annotation expo dataset for mobile gui agents. In Findings of the Association for Computational Linguistics: ACL 2025, pp. 2138–2156, 2025.
- Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Li YanTao, Jianbing Zhang, and Zhiyong Wu. Seelick: Harnessing gui grounding for advanced visual gui agents. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 9313–9332, 2024.
- Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. Advances in Neural Information Processing Systems, 36:28091–28114, 2023.
- Apurva Gandhi and Graham Neubig. Go-browse: Training web agents with structured exploration. arXiv preprint arXiv:2506.03533, 2025.
- Changlong Gao, Zhangxuan Gu, Yulin Liu, Xinyu Qiu, Shuheng Shen, Yue Wen, Tianyu Xia, Zhenyu Xu, Zhengwen Zeng, Beitong Zhou, et al. Ui-venus-1.5 technical report. arXiv preprint arXiv:2602.09082, 2026.

- Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie, Cheng Chang, Yiheng Shu, Huan Sun, and Yu Su. Navigating the digital world as humans do: Universal visual grounding for gui agents. arXiv preprint arXiv:2410.05243, 2024.
- Izzeddin Gur, Hiroki Furuta, Austin Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. A real-world webagent with planning, long context understanding, and program synthesis. arXiv preprint arXiv:2307.12856, 2023.
- Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. Webvoyager: Building an end-to-end web agent with large multimodal models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 6864–6890, 2024.
- Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Hongming Zhang, Tianqing Fang, Zhenzhong Lan, and Dong Yu. Openwebvoyager: Building multimodal web agents via iterative real-world exploration, feedback and optimization. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 27545–27564, 2025.
- Deyang Jiang, Jing Huang, Xuanle Zhao, Lei Chen, Liming Zheng, Fanfan Liu, Haibo Qiu, Peng Shi, and Zhixiong Zeng. Treecua: Efficiently scaling gui automation with tree-structured verifiable evolution. arXiv preprint arXiv:2602.09662, 2026.
- Linjia Kang, Zhimin Wang, Yongkang Zhang, Duo Wu, Jinghe Wang, Ming Ma, Haopeng Yan, and Zhi Wang. Learning with challenges: Adaptive difficulty-aware data generation for mobile gui agent training. arXiv preprint arXiv:2601.22781, 2026.
- Quyu Kong, Xu Zhang, Zhenyu Yang, Nolan Gao, Chen Liu, Panrong Tong, Chenglin Cai, Hanzhang Zhou, Jianan Zhang, Liangyu Chen, et al. Mobileworld: Benchmarking autonomous mobile agents in agent-user interactive and mcp-augmented environments. arXiv preprint arXiv:2512.19432, 2025.
- Hanyu Lai, Xiao Liu, Iat Long Iong, Shuntian Yao, Yuxuan Chen, Pengbo Shen, Hao Yu, Hanchen Zhang, Xiaohan Zhang, Yuxiao Dong, et al. Autowebglm: A large language model-based web navigating agent. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 5295–5306, 2024.
- Wei Li, William E Bishop, Alice Li, Christopher Rawles, Folawiyo Campbell-Ajala, Divya Tyamagundlu, and Oriana Riva. On the effects of data scale on ui control agents. Advances in Neural Information Processing Systems, 37:92130–92154, 2024.
- Zehao Li, Zhenyu Wu, Yibo Zhao, Bowen Yang, Jingjing Xie, Zhaoyang Liu, Zhoumianze Liu, Kaiming Jin, Jianze Liang, Zonglin Li, et al. Os-themis: A scalable critic framework for generalist gui rewards. arXiv preprint arXiv:2603.19191, 2026.
- Haojia Lin, Xiaoyu Tan, Yulei Qin, Zihan Xu, Yuchen Shi, Zongyi Li, Gang Li, Shaofei Cai, Siqi Cai, Chaoyou Fu, et al. Cuarewardbench: A benchmark for evaluating reward models on computer-using agent. arXiv preprint arXiv:2510.18596, 2025.

- Zhaoyang Liu, JingJing Xie, Zichen Ding, Zehao Li, Bowen Yang, Zhenyu Wu, Xuehui Wang, Qiushi Sun, Shi Liu, Weiyun Wang, et al. Scalecua: Scaling open-source computer use agents with cross-platform data. arXiv preprint arXiv:2509.15221, 2025.
- Quanfeng Lu, Wenqi Shao, Zitao Liu, Lingxiao Du, Fanqing Meng, Boxuan Li, Botong Chen, Siyuan Huang, Kaipeng Zhang, and Ping Luo. Guidyssey: A comprehensive dataset for cross-app gui navigation on mobile devices. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22404–22414, 2025a.
- Zhengxi Lu, Jiabo Ye, Fei Tang, Yongliang Shen, Haiyang Xu, Ziwei Zheng, Weiming Lu, Ming Yan, Fei Huang, Jun Xiao, et al. Ui-s1: Advancing gui automation via semi-online reinforcement learning. arXiv preprint arXiv:2509.11543, 2025b.
- Zhengxi Lu, Yuxiang Chai, Yaxuan Guo, Xi Yin, Liang Liu, Hao Wang, Han Xiao, Shuai Ren, Pengxiang Zhao, Guangyi Liu, et al. Ui-r1: Enhancing efficient action prediction of gui agents by reinforcement learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, pp. 17608–17616, 2026.
- Shikhar Murty, Hao Zhu, Dzmitry Bahdanau, and Christopher D Manning. Nnetnav: Un-supervised learning of browser agents through environment interaction in the wild. arXiv preprint arXiv:2410.02907, 2024.
- OpenAI. Introducing operator, 2025. URL <https://openai.com/index/introducing-operator/>.
- Vardaan Pahuja, Yadong Lu, Corby Rosset, Boyu Gou, Arindam Mitra, Spencer Whitehead, Yu Su, and Ahmed Hassan. Explorer: Scaling exploration-driven web trajectory synthesis for multimodal web agents. In Findings of the Association for Computational Linguistics: ACL 2025, pp. 6300–6323, 2025.
- Jiayi Pan, Yichi Zhang, Nicholas Tomlin, Yifei Zhou, Sergey Levine, and Alane Suhr. Autonomous evaluation and refinement of digital agents. arXiv preprint arXiv:2404.06474, 2024.
- Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, et al. Ui-tars: Pioneering automated gui interaction with native agents. arXiv preprint arXiv:2501.12326, 2025.
- Ram Ramrakhya, Andrew Szot, Omar Attia, Yuhao Yang, Anh Nguyen, Bogdan Mazouze, Zhe Gan, Harsh Agrawal, and Alexander Toshev. Scaling synthetic task generation for agents via exploration. arXiv preprint arXiv:2509.25047, 2025.
- Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy Lillicrap. Androidinthewild: A large-scale dataset for android device control. Advances in Neural Information Processing Systems, 36:59708–59728, 2023.
- Christopher Rawles, Sarah Clinckemaiellie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, et al. Androidworld: A dynamic benchmarking environment for autonomous agents. arXiv preprint arXiv:2405.14573, 2024.

- Rui Shao, Ruize Gao, Bin Xie, Yixing Li, Kaiwen Zhou, Shuai Wang, Weili Guan, and Gongwei Chen. Hats: Hardness-aware trajectory synthesis for gui agents. arXiv preprint arXiv:2603.12138, 2026.
- Peter Shaw, Mandar Joshi, James Cohan, Jonathan Berant, Panupong Pasupat, Hexiang Hu, Urvashi Khandelwal, Kenton Lee, and Kristina N Toutanova. From pixels to ui actions: Learning to follow instructions via graphical user interfaces. *Advances in Neural Information Processing Systems*, 36:34354–34370, 2023.
- Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, and Percy Liang. World of bits: An open-domain platform for web-based agents. In *International Conference on Machine Learning*, pp. 3135–3144. PMLR, 2017.
- Qiushi Sun, Zhirui Chen, Fangzhi Xu, Kanzhi Cheng, Chang Ma, Zhangyue Yin, Jianing Wang, Chengcheng Han, Renyu Zhu, Shuai Yuan, et al. A survey of neural code intelligence: Paradigms, advances and beyond. arXiv preprint arXiv:2403.14734, 2024.
- Qiushi Sun, Kanzhi Cheng, Zichen Ding, Chuanyang Jin, Yian Wang, Fangzhi Xu, Zhenyu Wu, Chengyou Jia, Liheng Chen, Zhoumianze Liu, et al. Os-genesis: Automating gui agent trajectory construction via reverse task synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5555–5579, 2025a.
- Qiushi Sun, Mukai Li, Zhoumianze Liu, Zhihui Xie, Fangzhi Xu, Zhangyue Yin, Kanzhi Cheng, Zehao Li, Zichen Ding, Qi Liu, et al. Os-sentinel: Towards safety-enhanced mobile gui agents via hybrid validation in realistic workflows. arXiv preprint arXiv:2510.24411, 2025b.
- Qiushi Sun, Zhoumianze Liu, Chang Ma, Zichen Ding, Fangzhi Xu, Zhangyue Yin, Haiteng Zhao, Zhenyu Wu, Kanzhi Cheng, Zhaoyang Liu, et al. Scienceboard: Evaluating multimodal autonomous agents in realistic scientific workflows. arXiv preprint arXiv:2505.19897, 2025c.
- Haoming Wang, Haoyang Zou, Huatong Song, Jiazhan Feng, Junjie Fang, Junting Lu, Longxiang Liu, Qinyu Luo, Shihao Liang, Shijue Huang, et al. Ui-tars-2 technical report: Advancing gui agent with multi-turn reinforcement learning. arXiv preprint arXiv:2509.02544, 2025.
- Qianhui Wu, Kanzhi Cheng, Rui Yang, Chaoyun Zhang, Jianwei Yang, Huiqiang Jiang, Jian Mu, Baolin Peng, Bo Qiao, Reuben Tan, et al. Gui-actor: Coordinate-free visual grounding for gui agents. arXiv preprint arXiv:2506.03143, 2025a.
- Zhiyong Wu, Chengcheng Han, Zichen Ding, Zhenmin Weng, Zhoumianze Liu, Shunyu Yao, Tao Yu, and Lingpeng Kong. Os-copilot: Towards generalist computer agents with self-improvement. arXiv preprint arXiv:2402.07456, 2024.
- Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, and Yu Qiao. OS-ATLAS: Foundation action model for generalist GUI agents. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL <https://openreview.net/forum?id=n9PDaFNi8t>.

- Jingxu Xie, Dylan Xu, Xuandong Zhao, and Dawn Song. Agentsynth: Scalable task generation for generalist computer-use agents. *arXiv preprint arXiv:2506.14205*, 2025.
- Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.
- Haiyang Xu, Xi Zhang, Haowei Liu, Junyang Wang, Zhaozai Zhu, Shengjie Zhou, Xuhao Hu, Feiyu Gao, Junjie Cao, Zihua Wang, et al. Mobile-agent-v3. 5: Multi-platform fundamental gui agents. *arXiv preprint arXiv:2602.16855*, 2026.
- Yifan Xu, Xiao Liu, Xueqiao Sun, Siyi Cheng, Hao Yu, Hanyu Lai, Shudan Zhang, Dan Zhang, Jie Tang, and Yuxiao Dong. Androidlab: Training and systematic benchmarking of android autonomous agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2144–2166, 2025.
- Taofeng Xue, Chong Peng, Mianqiu Huang, Linsen Guo, Tiancheng Han, Haozhe Wang, Jianing Wang, Xiaocheng Zhang, Xin Yang, Dengchang Zhao, et al. Evocua: Evolving computer use agents via learning from scalable synthetic experience. *arXiv preprint arXiv:2601.15876*, 2026.
- Haolong Yan, Jia Wang, Xin Huang, Yeqing Shen, Ziyang Meng, Zhimin Fan, Kaijun Tan, Jin Gao, Lieyu Shi, Mi Yang, et al. Step-gui technical report. *arXiv preprint arXiv:2512.15431*, 2025.
- Bowen Yang, Kaiming Jin, Zhenyu Wu, Zhaoyang Liu, Qiushi Sun, Zehao Li, JingJing Xie, Zhoumianze Liu, Fangzhi Xu, Kanzhi Cheng, et al. Os-symphony: A holistic framework for robust and generalist computer-using agent. *arXiv preprint arXiv:2601.07779*, 2026a.
- Qianlan Yang, Xiangjun Wang, Danielle Perszyk, and Yu-Xiong Wang. Self-guided hierarchical exploration for generalist foundation model web agents. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Rui Yang, Qianhui Wu, Zhaoyang Wang, Hanyang Chen, Ke Yang, Hao Cheng, Huaxiu Yao, Baoling Peng, Huan Zhang, Jianfeng Gao, et al. Gui-libra: Training native gui agents to reason and act with action-aware supervision and partially verifiable rl. *arXiv preprint arXiv:2602.22190*, 2026b.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pp. 400–410, 2024.
- Hanzhang Zhou, Xu Zhang, Panrong Tong, Jianan Zhang, Liangyu Chen, Quyu Kong, Chenglin Cai, Chen Liu, Yue Wang, Jingren Zhou, et al. Mai-ui technical report: Real-world centric foundation gui agents. *arXiv preprint arXiv:2512.22047*, 2025.
- Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.

A OpenMobile 框架的实现细节

A.1 环境探索算法

如 Section 3.1所述, OpenMobile 将探索与指令生成解耦, 使该框架对特定探索策略具有无关性。在本工作中, 我们遵循 OS-Genesis (Sun et al., 2025a), 采用简单的随机游走。每个会话执行 10 步, 从当前屏幕的可访问性树中随机选择一个可交互元素, 执行点击或输入动作。维护一个非可交互元素的黑名单, 以避免冗余交互。此阶段的主要目标是最大化应用状态和功能的覆盖, 以支持下游任务的合成。

A.2 全球环境记忆构建

给定由屏幕-动作转移组成的探索轨迹, 我们通过三个阶段构建全局环境记忆 \mathcal{M} : 屏幕去重、功能标注和语义索引构建。完整流程如算法 1 所示。

屏幕去重 探索轨迹中包含多个在不同会话中访问的视觉上相同或几乎相同的屏幕。我们为每个截图计算一个感知哈希 (pHash), 并贪婪地将 pHash 相似度超过阈值 $\tau = 0.95$ 的屏幕聚类, 每簇选择一个代表性屏幕。对于每个唯一屏幕 s_i , 我们汇总原始轨迹中的所有转移以识别其邻居集 $\mathcal{N}(s_i)$, 即可以直接从或指向 s_i 的屏幕。

功能标注。 对于每个唯一的屏幕, 我们使用强大的大语言模型 Gemini-3.1-Pro-Preview 提取一组功能描述 $\mathcal{F}(s_i)$ 。每个功能都是对用户界面元素 (例如按钮、切换开关或菜单项) 语义的自然语言描述。为了提高标注质量, 我们向模型提供前一个屏幕以及导致当前屏幕的操作作为上下文。元素被分类为 功能 (应用程序提供的特性, 如按钮和切换开关) 或 数据 (用户生成的内容, 如日历事件)。

语义索引构建 为了实现跨屏幕功能检索, 我们使用句子嵌入模型 openai/text-embedding-3-large 为每个应用内的所有功能描述计算语义嵌入。这生成了一个支持任务合成过程中高效最近邻搜索的每应用检索索引。我们采用贪心多样性过滤, 以确保检索到的功能在语义上具有差异性 (成对余弦相似度低于 0.8)。

A.3 增强记忆的任务合成

给定全局环境记忆 \mathcal{M} , 我们通过向强大的视觉-语言模型 Gemini-3.1-Pro-Preview 提供丰富的上下文信息, 并提示其生成基于事实的、多步骤的任务指令。

情境构建。 对于每个候选屏幕 s_i , 我们构建一个包含三个部分的提示。首先, 我们包含 s_i 的截图及其标注的功能描述, 作为生成的焦点。其次, 我们从转移图中邻近屏幕检索截图和功能描述作为短期记忆。具体而言, 我们包含 1 个前驱屏幕 (过渡到 s_i 的屏幕) 和最多 3 个后继屏幕 (从 s_i 可达的屏幕), 为模型提供局部导航上下文。第三, 我们从同一应用内的其他屏幕中检索 30 个语义相关的功能作为长期记忆。这些功能通过嵌入余弦相似度并结合多样性约束 (成对相似度 < 0.8) 进行选择, 以揭示遥远但相关特征, 从而促进跨功能组合。组装后的上下文, 包括截图和文本描述, 与生成指南及上下文示例一并输入模型。

质量筛选。 生成的指令经过三阶段过滤: (1) 每条指令由一个强大的大语言模型在复杂度、清晰度和合理性 (1-5 分制) 上进行评分, 清晰度 < 4 或合理性 < 4 的指令被丢弃; (2) 剩余指令按得分排序, 并使用嵌入余弦相似度以阈值 0.8 进行贪心去重, 每个语义簇中保留得分最高的指令; (3) 按应用对指令数量进行上限控制, 以确保训练过程中的应用覆盖均衡。

Algorithm 1 Global Environment Memory Construction

```

1: Input: Exploration trajectories  $\mathcal{T} = \{(o_t, a_t, o_{t+1})\}$ , similarity threshold  $\tau$ 
2: Output: Global environment memory  $\mathcal{M} = (\mathcal{S}, \mathcal{N}, \{\mathcal{F}(s_i)\})$ 
   // Stage 1: Screen Deduplication
3: Collect all screens  $\mathcal{O} = \{o_t, o_{t+1} \mid (o_t, a_t, o_{t+1}) \in \mathcal{T}\}$ 
4: Compute perceptual hash  $h(o)$  for each  $o \in \mathcal{O}$ 
5:  $\mathcal{S} \leftarrow \emptyset$ 
6: for each screen  $o \in \mathcal{O}$  do
7:   if  $\nexists s \in \mathcal{S}$  s.t.  $\text{Sim}(h(o), h(s)) \geq \tau$  then
8:      $\mathcal{S} \leftarrow \mathcal{S} \cup \{o\}$  ▷ Add as new unique screen
9:   end if
10: end for
   // Stage 2: Neighborhood & Functionality Extraction
11: for each unique screen  $s_i \in \mathcal{S}$  do
12:    $\mathcal{N}(s_i) \leftarrow$  screens in  $\mathcal{S}$  reachable from/to  $s_i$  via transitions in  $\mathcal{T}$ 
13:    $\mathcal{F}(s_i) \leftarrow \text{VLM}(s_i, \text{context})$  ▷ Extract functionality descriptions
14: end for
   // Stage 3: Semantic Index Construction
15: for each app  $A$  do
16:   Collect all functionalities:  $\mathcal{F}_A = \bigcup_{s_i \in A} \mathcal{F}(s_i)$ 
17:   Compute embeddings  $\mathbf{E}_A = \text{Embed}(\mathcal{F}_A)$ 
18:   Build retrieval index over  $(\mathcal{F}_A, \mathbf{E}_A)$ 
19: end for
20: return  $\mathcal{M} = (\mathcal{S}, \mathcal{N}, \{\mathcal{F}(s_i)\}_{i=1}^{|\mathcal{S}|})$ 

```

A.4 错误干预策略切换

我们的错误-干预切换从学习器模型执行任务开始。在每一步中，监控器（Gemini-3.1-Pro-Preview）观察最近的动作历史和前两张截图，以评估先前动作是否导致智能体偏离任务目标。一旦检测到偏离，便调用专家模型进行干预。我们发现，向专家提供监控器的偏离分析能够提升错误恢复信号的质量，因为专家在纠正轨迹之前能更好地理解当前的失败模式。在我们的轨迹生成过程中，这一错误-干预过程最多被触发两次。每次专家干预后，专家至少执行 3 步，之后控制权才返回给学习器。

A.5 OpenMobile 中使用的提示

我们在此提供 OpenMobile 流水线中使用的所有完整提示。

Prompt for Functionality Extraction

系统提示

你是一名 GUI 截图分析专家。你将收到：

1. UI 界面的截图（操作前屏幕），红色标记了动作区域
2. 执行的动作类型
3. 动作后的截图（动作后）
4. 安卓应用程序的名称

请仅分析第二个截图（操作后画面）中的元素。第一个截图仅作为上下文，帮助您理解应用程序的状态。

每个元素都应以字典形式输出：

```
{
  "type": "functionality" or "data",
  "label": "A short phrase describing its identifier on this screen",
  "description": "A few sentences describing this element's functionality"
}
```

描述应全面且详尽：

- 包括应用内的层级位置（例如，哪个菜单、哪个设置页面、哪个子部分）
- 在手机/设备层面，该元素的作用是……

以下是描述不佳的例子及其改进后的版本：

例 1：

错误：一个用于开启或关闭 WiFi 连接的切换开关。

原因：过于模糊，未明确具体位置或设备级别的变更。

“此开关位于系统设置 > 网络与互联网 > 无线局域网，用于启用或禁用设备上的无线局域网功能，使手机能够扫描可用的无线网络并连接或断开连接。”

示例 2：

错误：「提醒」选项使用户能够设置提醒。

原因：过于模糊，未说明其适用的具体场景。

良好：在日历应用的事件创建/编辑界面中，此提醒选项会在事件开始前安排一次通知（例如提前 10 分钟），帮助用户在设定的提前时间内收到提醒。

Output a JSON list only. No markdown, no comments, no extra text. Start with [and end with].

用户提示

[图片：操作前（红色标记操作区域）]

[图片：屏幕之后]

应用：{app_name}

动作：{动作类型}

第一张图是操作前的界面（红色标记了动作区域）。第二张图是操作后的界面。请分析第二张图中的元素。

Prompt for Task Instruction Synthesis

系统提示

你是一名 GUI 探索者。你的目标是探索 GUI 环境，并综合生成高质量、高难度、可执行、高层次、多步骤的 GUI 任务/指令。

您已经完成了探索工作。您已从当前的 GUI 环境中收集了大量截图、它们之间的转移过程，以及相应应用程序中的各种功能。

现在，你需要基于以下三个信息来源进行充分关联与想象，生成当前应用中可能实现的长程、高阶任务/指令：

1. 特定屏幕的召回截图
2. 短期记忆中与当前截图具有转移关系的若干截图（可从当前屏幕到达的屏幕）
3. 重要的是，从长期记忆中检索出的一些功能与当前屏幕相关（即同一应用程序中其他屏幕的语义相关功能）

基于这三条信息来源，您应充分关联、想象并生成当前应用内可能实现的长距离、高阶任务/指令。

指南

1. 所提供的截图和功能仅是你召回记忆中的一小部分，用作上下文参考。你的唯一任务是生成清晰的多步骤图形用户界面操作指引。你所生成的指引无需与当前屏幕或操作有直接关联，但可根据上下文进行推断。为确保生成任务的难度和复杂度，鼓励你分析、关联并整合来自记忆中的多种功能。

2. 有两种任务类型需要生成：

- **动作任务**：需要执行一系列动作来完成目标。例如：“设置一个明天早上 8 点的闹钟，每周工作日重复。”
- **问答任务**：需要执行一系列动作，并回答与环境内容相关的问题。例如：“在我的待办事项列表中，本周三需要完成多少项任务？请用一个数字回答这个问题。”

您应根据上下文决定生成哪种类型的任务。

3. 合成任务 **必须清晰明确**。生成的任务应具体且包含足够的细节，以便执行者不会感到困惑。例如，“帮我创建日历中的一个新事件”过于宽泛，应包含具体的配置信息，例如日期、时间、标题、描述、持续时间、地点等。

4. 合成的任务必须是可执行的。**如果您想要生成一个涉及操作应用数据的任务（例如，删除日历中的某一项），您必须确保您想要操作的数据在给定的截图中存在。**

5. 生成的任务应具有多样性，不要仅关注应用的主要功能。应尽可能涵盖应用的所有功能，例如屏幕角落的元素或功能，或是从记忆中联想到的功能。

6. 生成的任务应具有长程性。不要生成如点击按钮等单步任务。鼓励生成需要执行者进行推理、规划并分多步完成的任务。**还可以考虑将不同的子功能或子任务组合成一个长程任务，但需确保合理性。**

7. 生成的任务应具有高层次性。**不要生成分步指令和详细动作。**应将多步指令整合为一个高层次的意图，以提高任务难度。**任务应为包含具体细节的单一命令，而非完成任务的分步操作。**

8. 生成的任务应从手机的主屏幕开始，而不是从当前提供的屏幕开始。不要生成与当前界面临时状态相关的任务（例如，出现的弹出对话框）。

9. 运行环境是一个没有网络连接的虚拟设备。请勿生成需要互联网连接或登录的任务。但您可以自由使用现有应用中已保存的数据。

示例任务

以下是糟糕任务及其改进版本的示例：

例 1：

不好：访问和管理所有已保存的蓝牙设备列表。

理由：未说明“管理”具体指什么。

良好：查看所有现有的蓝牙设备，如果存在，则删除全部。

示例 2：

错误：使用主菜谱屏幕上的加号按钮向列表中添加新食谱。

理由：未明确具体内容。

在西兰花应用中，为“番茄炒蛋”添加一个新食谱，将类别设为“炒菜”，并将描述填写为“妈妈最喜爱的菜肴”。

示例 3：

错误：检查电池使用统计量，如有需要，请启用电池节省模式。

理由：“如有必要”会令执行者感到困惑。

- 良好：将电池使用统计量中的前三个项目写入 Markor 应用程序，并将其保存为 ‘battery_usage_statistics’，同时启用电池节省模式。

示例 4:

错误做法：通过点击“键盘”按钮来忽略语音搜索连接错误，然后在搜索栏中手动输入“披头士”。

- 原因：包含一个临时状态，并假设从搜索界面开始。

- 好：“在 {应用名称} 中，分别包含多少首披头士乐队和泰勒·斯威夫特的歌曲？请用逗号分隔的数字回答。”

示例 5:

错误：在“西兰花”应用中，使用搜索功能查找食谱“苜蓿酱三文鱼”，打开其详情页面，回答该食谱可制作多少份。

原因：包含过多具体操作；应更具高层次性。

在 Broccoli 应用中，“苜蓿酱三文鱼”这道菜提供几份，总共需要多长时间的准备？

示例 6:

- 无效：在“简易日历专业版”中，进入“自定义颜色”菜单，尝试更改应用图标颜色，然后关闭警告弹窗。

- 原因：包含不必要的特定操作和临时状态。

好的：将简易日历专业版的应用颜色设置为蓝色。

示例 7:

不好：“在任务应用中，我有哪些任务？”

理由：过于模糊。

在“任务”应用中，本周到期但尚未完成的任务有哪些？仅回答标题；如果有多项，请用逗号分隔。

示例 8:

- 不佳：在音频录制应用中，配置高保真录音设置。进入应用后，导航至设置菜单，并将录音格式更改为 Wav，将采样率设为 48kHz……

原因：包含过多的逐步操作。

- 正确：使用音频录制工具以 48kHz 采样率和立体声通道录制一个 Wav 格式的音频文件，并将其保存为 test_audio。

用户提示

[图片：当前屏幕]

[图片：前一屏幕 1（如可用）]

[图片：关联屏幕 1-3（如可用）]

Current Screen

应用: {app_name}

当前屏幕上的元素 ({N} 项):

1. “type”: functionality, “description”: {description_1}

2. ...

这些屏幕可以转移到当前屏幕:

Preceding Screen 1

元素 ({N} 项): ...

这些是从当前屏幕到达的屏幕:

Associated Screen 1

元素 ({N} 项): ...

Related Functionalities from Other Screens ({M} items)

这些是同一应用程序中其他屏幕的语义相关功能。

```

1. {description_1}
2. ...
## Your Task
Based on the above context, carefully analyze and think, then generate 1-3 high-quality
GUI tasks. Each task should be a concise but high-level instruction in English. Output
format (JSON array):
[
  {"reasoning": "...", "task": "task instruction 1"},
  {"reasoning": "...", "task": "task instruction 2"}
]
```

B 实验情景

B.1 基准评估设置

我们在三个已有的动态移动智能体基准上进行评估：AndroidWorld、AndroidLab 和 MobileWorld。所有模型均使用 vLLM 进行推理部署。我们观察到，在动态环境中执行智能体时存在固有的随机性，导致不同运行之间的成功率表现出非平凡的方差。为考虑这一因素，我们对每个基准运行三次，并报告平均值以及半值域（即 $(\max - \min) / 2$ ）作为变异程度的度量。此外，我们还报告 Pass@3，即当三次运行中任意一次成功即认为任务完成，以此反映模型性能的上限。部分基准模型的结果来自先前的工作，例如 UI-Venus-1.5 (Gao et al., 2026) 和 MobileWorld 排行榜 (Kong et al., 2025)。

B.2 策略切换发布设置

我们使用 Qwen2.5-VL-7B-Instruct 作为所有策略切换消融实验的基础模型（学习器 π_l ），并以 Gemini-3.1-Pro-Preview 作为专家模型 π_e 。**专家蒸馏**。专家模型执行所有合成的指令。我们保留专家标记任务完成的轨迹（即输出 complete 或 answer），并将这些轨迹转换为步骤级别的训练样本。**自演化**。学习器执行合成的指令，专家充当裁判以识别成功的轨迹。仅使用成功轨迹对学习器进行重新训练。该过程迭代进行 3 轮。**随机切换**。在每一步中，当学习器与专家预测的动作不一致时（例如动作类型不同或目标元素不同），则用学习器的动作替代专家的动作。但学习器不允许执行终止动作（complete 或 answer），以确保任务由专家完成。**错误干预切换**。滚动从学习器策略开始。一个监控器跟踪学习器的执行过程，并在检测到偏离有效进展时触发切换至专家。随后专家介入，将轨迹纠正回正确路径。监控器设计的详细信息见 Section A.4。所有策略均在固定的 1.5K 条轨迹预算下进行比较。为了定量衡量每种策略引入的错误恢复信号，我们从每种策略中随机抽取 50 条轨迹，并手动检查每条轨迹的平均错误恢复信号数量，定义为智能体识别并尝试纠正前一步错误的步骤。

C 强化学习中的探索

除了标准的监督微调 (SFT) 之外，我们还探讨了强化学习 (RL) 在我们合成数据上的有效性。

步骤级强化学习。我们首先在步骤级强化学习 (RL) 上进行实验，这是 GUI 智能体中普遍采用的范式。遵循 UI-R1 (Lu et al., 2025b) 的做法，我们在每个智能体步骤上定义了三种奖励信号：格式奖励、动作类型奖励和定位奖励，并使用标准的 GRPO 进行训练。随后

我们在 AndroidWorld 上评估所得模型的表现。结果表明，尽管步骤级 GRPO 在使用合成轨迹时初期提升了性能，但收益迅速达到饱和，最终未能超越 SFT 基准。我们认为这是由于步骤级最优化与动态环境中多步执行之间存在的固有差异所致。步骤级 RL 容易对单步输出过拟合，无法在需要与不断变化的环境持续交互的长时程任务中产生稳定的提升。类似观察也已在 UI-Venus-1.5 (Gao et al., 2026) 中被报告。

轨迹级 Agentic 强化学习 我们进一步探索轨迹级的 Agentic 强化学习，以提升智能体的性能。我们使用 OS-Themis (Li et al., 2026) 框架开展这些实验，该框架提供了超过一百个 Android 模拟器实例组成的基础设施，用于轨迹滚动，同时还配备了一个多智能体评论员，利用视觉语言模型 (VLM) 来判断任务是否成功或失败。我们采用一个早期检查点作为强化学习训练的起点，并对合成的任务指令进行筛选，仅保留那些检查点无法完成但专家能够成功完成的指令，共得到 244 条指令。如 Table 4 所示，尽管在我们合成的指令上进行轨迹级强化学习确实提升了性能，但其表现无法持续超越完全经过 SFT 训练的基准模型。我们推测，这与当前环境设置的多样性有限以及强化学习框架本身的稳定性有关。如何解决这些局限性以进一步提升移动智能体的能力，仍是未来工作的重要方向。

D 额外的实验结果

为了验证 OpenMobile 数据在更大模型上的有效性，我们使用相同的数据对 Qwen2.5-VL-72B-Instruct 进行微调。如 Table 5a 所示，更大的模型表现出显著更强的性能，证实了我们的数据具有良好的可扩展性，同时也凸显了基础模型能力的重要性。

我们还与现有的移动智能体数据合成方法进行了比较 (Sun et al., 2025a; Shao et al., 2026; Ramrakhya et al., 2025; Kang et al., 2026)。如 Table 5b 所示，OpenMobile 在中等数据规模下实现了显著更高的成功率。需要注意的是，直接比较并不完美，因为这些方法在基础模型和实验情景上存在差异，且部分方法未开源完整的实现细节。尽管如此，结果仍表明 OpenMobile 的有效性，并将其定位为未来移动智能体数据合成研究的一个强有力起点。

E 合成指令与测试指令之间的相似度

为了量化我们合成的指令与 AndroidWorld 测试集之间的重叠程度，我们使用 openai/text-embedding-3-large 的句子嵌入计算成对余弦相似度。如 Figure 3 所示，我们的合成指令在功能层级上表现出中等的相关性，而非任务层级的重叠，仅有 3.5% 的相似度超过 0.7。Table 6 列出了每个 AndroidWorld 测试指令及其最相似的合成指令。

Method	Base Model	AndroidWorld		AndroidLab		MobileWorld	
		Pass@1↑	Pass@3↑	Pass@1↑	Pass@3↑	Pass@1↑	Pass@3↑
Qwen3-VL-8B	–	47.6 ± 2.2	62.1	43.5	–	9.4	–
Ours-8B	Qwen3-VL	64.7 ± 3.2	78.0	51.5 ± 0.7	62.3	17.7 ± 2.2	24.8
Ours-8B-RL	Qwen3-VL	64.1 ± 0.5	77.6	53.9 ± 1.5	63.0	16.8 ± 0.9	20.5

表 4: 轨迹级强化学习的结果。

Method	Pass@1 \uparrow	Pass@3 \uparrow
Qwen2.5-VL-7B	25.5 \pm 2.6	34.9
Qwen2.5-VL-72B	27.6	–
UI-Venus-72B	65.9	–
Ours-7B	51.7 \pm 1.7	68.1
Ours-72B	59.3 \pm 0.9	72.8

(a) Scaling to larger models.

Method	Open	#Traj	Pass@1 \uparrow
OS-Genesis	✓	1.5K	17.4
HATS	✓	1K	24.4
AutoPlay	✗	20K	40.1
MobileGen	✗	0.5K	45.7
OpenMobile	✓	2.8K	64.7

(b) Comparison with data synthesis methods.

表 5: (a) AndroidWorld 在使用更大模型尺寸时的结果。(b) 与现有数据合成方法在 AndroidWorld 上的比较。

表 6: 最近的合成指令与测试指令配对的完整列表。每一行显示一个 AndroidWorld 测试指令与其最相似的 OpenMobile 合成指令。

#	AndroidWorld Test Instruction	Most Similar Synthetic Instruction	Sim.
1	Record an audio clip using Audio Recorder app and save it.	In the Audio Recorder app, change the recording settings to use the Wav format and Stereo channel, then record a short audio clip.	0.715
2	Record an audio clip and save it with name "eVq3_review.m4a" using Audio Recorder app.	Record a high-quality audio clip using the Wav format and 48kHz sample rate in the Audio Recorder app.	0.635
3	Open the file task.html in Downloads in the file manager; when prompted open it with Chrome. Then create a drawing using the three colors shown at the top and hit submit.	In the Files app, navigate to the Downloads folder and open the 'task.html' file using the Chrome browser.	0.743
4	Open the file task.html in Downloads in the file manager; when prompted open it with Chrome. Then navigate the X to the bottom-right cell, by using the direction buttons.	In the Files app, navigate to the Downloads folder and open the 'task.html' file using the Chrome browser.	0.803
5	Open the file task.html in Downloads in the file manager; when prompted open it with Chrome. Then click the button 5 times, remember the numbers displayed, and enter their product in the form.	In the Files app, navigate to the Downloads folder and open the 'task.html' file using the Chrome browser.	0.706
6	Take one photo.	Switch to Camera mode, enable the 3x3 grid lines overlay, and take a photo.	0.524
7	Take one video.	Switch the Camera app to Video mode and record a short video clip.	0.502
8	Pause the stopwatch.	Use the Stopwatch to record a lap time, then pause and reset the timer to zero.	0.626
9	Run the stopwatch.	Use the Stopwatch to record a lap time, then pause and reset the timer to zero.	0.610
10	Create a timer with 23 hours, 4 minutes, and 57 seconds. Do not start the timer.	In the Clock app, set a timer for 1 hour, 23 minutes, and 45 seconds and start the countdown.	0.574
11	Create a new contact for Ahmed dos Santos. Their number is +12432810546.	Create a new contact for 'Alice Smith' with the phone number '555-123-4567' in the Phone app.	0.527
12	Go to the new contact screen and enter the following details: First Name: Eva, Last Name: Smith, Phone: 119-168-9838, Phone Label: Work. Do NOT hit save.	In the Contacts app, create a new contact with the name "John Smith" and the phone number "555-1234".	0.614

(continued on next page)

(continued from previous page)

#	AndroidWorld Test Instruction	Most Similar Synthetic Instruction	Sim.
13	Add the following expenses into the pro expense: name amount_dollars category_name note Museum Tickets \$325.17 Entertainment Urgent Social Club Dues \$425.35 Social I may repeat this Museum Tickets \$485.01 Entertainment I may repeat this	In Pro Expense, find the 'Club Membership' expense and update it by changing the category to 'Entertainment', setting the amount to 100, and changing the note to 'Monthly fee', then save the changes.	0.659
14	Add the expenses from expenses.jpg in Simple Gallery Pro to pro expense.	In Simple Gallery Pro, find the receipt from 'Innovate Solutions Ltd.' and answer what item was purchased and its price. Answer the question in the format: 'Item Name, Price'.	0.561
15	Go through the transactions in my_expenses.txt in Markor. Log the reimbursable transactions in the pro expense.	In the Pro Expense app, find the existing expense record for 'ProDev' and permanently delete it from the logs.	0.535
16	Add the following expenses into the pro expense: name amount_dollars category_name note Club Membership \$56.67 Social Urgent	In Pro Expense, find the 'Club Membership' expense and update it by changing the category to 'Entertainment', setting the amount to 100, and changing the note to 'Monthly fee', then save the changes.	0.706
17	Delete all but one of any expenses in pro expense that are exact duplicates, ensuring at least one instance of each unique expense remains.	In the Pro Expense app, find the existing expense record for 'ProDev' and permanently delete it from the logs.	0.544
18	Delete all but one of any expenses in pro expense that are exact duplicates, ensuring at least one instance of each unique expense remains.	In the Pro Expense app, find the existing expense record for 'ProDev' and permanently delete it from the logs.	0.544
19	Delete the following expenses from pro expense: Textbooks, Salary, Stationery.	In Pro Expense, permanently delete the 'School Supplies' transaction from the recent expenses list.	0.655
20	Delete the following expenses from pro expense: Night Out, Tailoring Services, Snacks.	In Pro Expense, permanently delete the 'School Supplies' transaction from the recent expenses list.	0.617
21	Delete the following expenses from pro expense: Taxi Fare.	Delete the 'Taxi Fare' transaction from the recent list in the Pro Expense app.	0.760
22	Delete the file smart_guitar_2023_07_08.mp3 from the Android filesystem located in the Podcasts folder within the sdk_gphone_x86_64 storage area.	In Retro Music, permanently delete the audio file for the song 'My Heart is Yours' from the device.	0.520
23	Move the file sci_fi_thriller.mp4 from Podcasts within the sdk_gphone_x86_64 storage area to the Movies within the same sdk_gphone_x86_64 storage area in the Android filesystem.	In the Files app, navigate to the internal storage 'sdk_gphone64_arm64', create a new folder named 'Project_Alpha', and then change the directory's view mode to the grid layout.	0.503
24	Update the Markor note backup_silly_tiger.txt by adding the following text, along with a new blank line before the existing content: "rMneE03kV42gdja3JKbM", and rename it to 0xeI_quick_house.md.	In the Markor app, open the 'silly_tree_copy' note, insert the text 'IMPORTANT:' in bold at the very beginning of the document, and save the file.	0.668
25	Update the content of lively_wolf_2023_06_12.md to "n7g5D1zYaynRyGnucoSD" in Markor and change its name to ucBD_pretty_elephant.txt.	In Markor, update the file '2023_04_18_pr...' by replacing its content with a shopping checklist containing bolded 'Apples', italicized 'Bananas', and 'Oranges', then save the changes.	0.629
26	Create a new folder in Markor named folder_20260304_164543.	Create a new folder named 'Meeting_Notes' in the Markor app, and inside that folder, create a new file named '2023-10-27_Briefing.md'.	0.740

(continued on next page)

(continued from previous page)

#	AndroidWorld Test Instruction	Most Similar Synthetic Instruction	Sim.
27	Create a new note in Markor named proud_banana_E4ao.txt with the following text: The pen is mightier than the sword.	In the Markor app, open the 'silly_tree_copy' note, insert the text 'IMPORTANT:' in bold at the very beginning of the document, and save the file.	0.605
28	Create a new note in Markor named wise_nurse_2023_06_26.txt with the following text: The early bird catches the worm.. Share the entire content of the note with the phone number +11843888619 via SMS using Simple SMS Messenger	In Simple SMS Messenger, save the phone number +16700234643 as a new contact named 'Daily Wisdom' and then initiate a call to this number.	0.547
29	Create a note in Markor named final_fancy_unicorn.md. Perform a paste operation in the note and save the note.	In Markor, create a new note containing the text 'Draft version', and then use the Search and Replace tool to replace the word 'Draft' with 'Final'.	0.599
30	Delete all my notes in Markor.	In the Markor app, delete all files in the Documents folder that contain the phrase 'april_workout_routine' in their names.	0.623
31	Delete the newest note in Markor.	In the QuickNote section of the Markor app, delete the existing line containing the date '2023-10-15' using the delete line toolbar option.	0.645
32	Delete the note in Markor named fancy_queen_2023_01_05.	In the QuickNote section of the Markor app, delete the existing line containing the date '2023-10-15' using the delete line toolbar option.	0.700
33	Edit note_5v02Y.md in Markor. Add to the bottom of the note The library book is due back on the 15th.	In the QuickNote section of the Markor app, delete the existing line containing the date '2023-10-15' using the delete line toolbar option.	0.551
34	Merge the contents of Markor notes best_lion_final.txt, alert_koala_2023_10_10.txt and edited_super_cat.md (in the same order) into a new Markor note named QpRyrbS1 and save it. Add a new line between the content of each note.	In Markor, open the file 'tough_frog_2023_08_05.txt', add a new line with the text 'Urgent' formatted in bold, and then save the file.	0.587
35	In Markor, move the note copy_helpful_umbrella.txt from MeetingMinutes to WorkProjects.	In the Markor app, add a new entry titled 'Organize documents' to the To-Do list, and then move the file 'final_meeting_notes_project_team.md' into a new folder named 'Work Archive'.	0.617
36	Create a file in Markor, called receipt.md with the transactions from the receipt.png. Use Simple Gallery to view the receipt. Please enter transactions in csv format including the header "Date, Item, Amount".	Create a new file named 'Groceries.md' in Markor, add a checklist item labeled 'Milk', and save the document.	0.581
37	Transcribe the contents of video copy_moment_10.mp4 by watching it in VLC player (located in Download) and writing the sequence of strings shown on each frame to the text file copy_moment_10__transcription.txt in Markor as a comma separated list. For example, if the first frame shows the text "edna" and the second frame shows the text "pineapple", then the text file should contain only the following text: "edna, pineapple".	In the Markor app, what is the exact string of characters written on the first line of the file 'oGsN_note_X3...'? Answer the question with the text only.	0.467

(continued on next page)

(continued from previous page)

#	AndroidWorld Test Instruction	Most Similar Synthetic Instruction	Sim.
38	Is the note titled 'Research Notes' in the Joplin app marked as a todo item? Respond with either 'True' if it is a todo or 'False' if not.	In the Joplin app, search for the item "Mortgage Payment Schedule" and determine if it is a to-do task (with a checkbox) or a standard note. Answer with "To-do" or "Note".	0.607
39	How many attendees were present in the meeting titled 'Marketing Campaign Planning' in the Joplin app? Express your answer as just a single number.	In the Joplin app, how many visible notes contain the word 'Plan' or 'Planning' in their title? Answer the question with a single number.	0.628
40	What quantity of matcha powder do I need for the recipe 'Lasagna' in the Joplin app? Express your answer in the format <amount> <unit> where both the amount and unit exactly match the format in the recipe.	In the Broccoli app, for the recipe containing the ingredient 'per individual taste', what is the default serving size displayed when you open the 'Adjust ingredients' dialog? Answer the question with a single number.	0.502
41	How many to-dos do I have in the 'Travel' folder in the Joplin app? Express your answer as just a single number.	In the Joplin app's 'All notes' list, how many to-do items start with the word 'Travel'? Answer the question with a single number.	0.845
42	Open the camera app. Clear any pop-ups that may appear by granting all permissions that are required.	Grant Android Auto the 'Device & app notifications' permission to allow it to read notifications, and then clear the app's cache.	0.528
43	Add a favorite location marker for Malbun, Liechtenstein in the OsmAnd maps app.	In the OsmAnd app, configure the map settings to display both 'Favorites' locations and 'Transport' routes.	0.532
44	Add a location marker for 47.1303814, 9.5930117 in the OsmAnd maps app.	In the OsmAnd app, use the address search feature to locate the point at latitude 40.7306 and longitude -73.9352 using the coordinate search option, and show this location on the map.	0.579
45	Save a track with waypoints Schaan, Liechtenstein, Malbun, Liechtenstein, Planken, Liechtenstein, Rotenboden, Liechtenstein in the OsmAnd maps app in the same order as listed.	Using the 'Plan a route' feature in OsmAnd, manually create a custom route by placing four waypoints on the map and save this track to your 'My Places' collection with the name 'Sample Trip'.	0.606
46	Add the following recipes into the Broccoli app: title description servings preparationTime ingredients directions Classic Margherita Pizza An ideal recipe for experimenting with different flavors and ingredients. 1 serving 20 mins to your liking Spread pizza dough with tomato sauce, top with slices of mozzarella cheese and fresh basil leaves. Bake until crust is golden. Garnish with fresh herbs for a more vibrant taste. Garlic Butter Shrimp A quick and easy meal, perfect for busy weekdays. 1 serving 2 hrs see directions Sauté shrimp in butter and minced garlic until pink. Sprinkle with parsley and serve with lemon wedges. Garnish with fresh herbs for a more vibrant taste. Mango Chicken Curry A delicious and healthy choice for any time of the day. 3-4 servings 1 hrs various amounts Cook chicken pieces in a pan, add onions, garlic, and ginger. Stir in curry powder, coconut milk, and mango pieces. Simmer until chicken is cooked. Feel free to substitute with ingredients you have on hand.	In the Broccoli app, create a new recipe for 'Classic Margherita Pizza' under the 'Dinner' category, set the description to 'Simple and delicious', source it from 'Chef Mario', specify it serves 2 people, takes 45 minutes to prepare, and list 'Dough, Tomato Sauce, Mozzarella, Basil' as the ingredients.	0.760

(continued on next page)

(continued from previous page)

#	AndroidWorld Test Instruction	Most Similar Synthetic Instruction	Sim.
47	Add the recipes from recipes.jpg in Simple Gallery Pro to the Broccoli recipe app.	In the Broccoli app, create a new recipe for 'Garden Salad', assign it to the 'Healthy' category, and use the device's camera to take and set a cover photo for the recipe before saving.	0.685
48	Add the recipes from recipes.txt in Markor to the Broccoli recipe app.	In the Broccoli app, find the 'Tomato Basil Bruschetta' recipe and add it to your Favorites.	0.595
49	Add the recipes from recipes.txt in Markor that take 4 hrs to prepare into the Broccoli recipe app.	In the Broccoli app, edit the 'Lentil Soup' recipe to change its preparation time to 45 minutes, and then mark the recipe as a favorite.	0.662
50	Add the following recipes into the Broccoli app: Recipe: Caprese Salad Skewers description: A quick and easy meal, perfect for busy weekdays. servings: 6 servings preparationTime: 4 hrs ingredients: various amounts directions: Thread cherry tomatoes, basil leaves, and mozzarella balls onto skewers. Drizzle with balsamic glaze. Garnish with fresh herbs for a more vibrant taste.	In the Broccoli app, the 'Caprese Salad Skewers' recipe is missing ingredient details. Edit the recipe to add 'Mozzarella balls' and 'Cherry tomatoes' to the ingredients list, and update the preparation time to '20 mins'.	0.780
51	Delete all but one of any recipes in the Broccoli app that are exact duplicates, ensuring at least one instance of each unique recipe remains	In the Broccoli app, clean up the recipe list by deleting the duplicate entries for 'Baked Cod with Lemon and Dill' so that only one such entry remains.	0.771
52	Delete all but one of any recipes in the Broccoli app that are exact duplicates, ensuring at least one instance of each unique recipe remains	In the Broccoli app, clean up the recipe list by deleting the duplicate entries for 'Baked Cod with Lemon and Dill' so that only one such entry remains.	0.771
53	Delete all but one of any recipes in the Broccoli app that are exact duplicates, ensuring at least one instance of each unique recipe remains	In the Broccoli app, clean up the recipe list by deleting the duplicate entries for 'Baked Cod with Lemon and Dill' so that only one such entry remains.	0.771
54	Delete the following recipes from Broccoli app: Eggplant Parmesan, Cauliflower Fried "Rice", Lemon Garlic Tilapia.	Delete the 'Eggplant Parmesan' recipe from your collection in the Broccoli app.	0.771
55	Delete the recipes from Broccoli app that use ghee in the directions.	Delete the 'Beef Stir Fry' recipe from the Broccoli app.	0.682
56	Delete the following recipes from Broccoli app: Turkey and Cheese Panini, Stuffed Bell Peppers, Thai Peanut Noodle Salad.	Delete the 'Thai Peanut Noodle Salad' recipe from the Broccoli app.	0.796
57	Delete the following recipes from Broccoli app: Chicken Caesar Salad Wrap.	Delete the 'Beef Stir Fry' recipe from the Broccoli app.	0.720
58	Delete the following recipes from Broccoli app: Mango Chicken Curry.	Delete the 'Beef Stir Fry' recipe from the Broccoli app.	0.752
59	Create a playlist in Retro Music titled "Acoustic Sessions 86" with the following songs, in order: City of Stars, Echoes of Silence	In Retro Music, create a new playlist named 'Acoustic Sessions' and add the songs 'City of Stars' and 'Distant Memories' to it.	0.829
60	Add the following songs, in order, Shadows of Time, Eternal Flame, Golden Days to my playing queue in Retro music.	In Retro Music, find the song 'Distant Memories' in the 'Last added' list and add it to the playing queue.	0.648
61	Create a playlist in Retro Music titled "Electronic Chillout 553" with a duration between 45 and 50 minutes using the provided songs.	In Retro Music, create a new playlist named 'Chill Vibes' and add the song 'Hidden Paths' to it.	0.681

(continued on next page)

(continued from previous page)

#	AndroidWorld Test Instruction	Most Similar Synthetic Instruction	Sim.
62	Create a playlist in Retro Music titled "Retro Pop Hits 458" with the following songs, in order: Bright Lights, Eternal Flame, Endless Summer. Then export the playlist to the Downloads directory on the device.	Create a new playlist named 'Night Drive' in the Retro Music app and add the songs 'Bright Lights' by Oliver and 'Eternal Flame' by Martina to it.	0.753
63	In Simple Gallery Pro, copy receipt_smart_vase_copy.jpg in DCIM and save a copy with the same name in Download	In Simple Gallery Pro, locate the receipt image for 'Innovate Solutions Ltd' within the DCIM folder, rotate the image 90 degrees clockwise, and mark it as a favorite.	0.647
64	In Simple Calendar Pro, create a calendar event on 2023-10-17 at 11h with the title 'Review session for Campaign' and the description 'We will review product launch. Snacks will be provided.'. The event should last for 45 mins.	In Simple Calendar Pro, create a new event on October 25 titled 'Budget Review' at 'Finance Dept' that starts at 10:00 and ends at 11:30, and add the note 'Prepare Q3 reports' in the description.	0.770
65	In Simple Calendar Pro, create a calendar event in two weeks from today at 20h with the title 'Workshop on Annual Report' and the description 'We will discuss upcoming project milestones.'. The event should last for 60 mins.	In Simple Calendar Pro, create a new event titled "Strategy Workshop" for October 20th starting at 14:00, set the location to "Main Hall", and add a description "Quarterly planning session".	0.767
66	In Simple Calendar Pro, create a calendar event for this Thursday at 5h with the title 'Call with HR' and the description 'We will discuss annual budget. Looking forward to productive discussions.'. The event should last for 45 mins.	In Simple Calendar Pro, create a new event on October 25 titled 'Budget Review' at 'Finance Dept' that starts at 10:00 and ends at 11:30, and add the note 'Prepare Q3 reports' in the description.	0.748
67	In Simple Calendar Pro, create a calendar event for tomorrow at 0h with the title 'Appointment for Campaign' and the description 'We will celebrate software updates.'. The event should last for 45 mins.	In Simple Calendar Pro, create a new event titled 'Project Meeting' for tomorrow at 2:00 PM, set it to repeat weekly, and add a reminder 10 minutes before the start.	0.748
68	In Simple Calendar Pro, create a recurring calendar event titled 'Review session for Project X' starting on 2023-10-24 at 18h. The event recurs weekly, forever, and lasts for 45 minutes each occurrence. The event description should be 'We will organize team roles. Let's be punctual!'.	In Simple Calendar Pro, create a new event titled 'Review session for Annual Report' for October 21st. Set the description to 'We will organize annual budget. Let's be punctual.' and configure the event to repeat yearly.	0.788
69	Do I have any events October 16 2023 in Simple Calendar Pro? Answer with the titles only. If there are multiples titles, format your answer in a comma separated list.	In Simple Calendar Pro, identify all events scheduled for October 17th and answer with their titles separated by a comma.	0.782
70	In Simple Calendar Pro, delete all the calendar events on 2023-10-25	In Simple Calendar Pro, find and delete the 'Workshop' event scheduled for October 24th.	0.718
71	In Simple Calendar Pro, delete all events scheduled for this Friday.	In Simple Calendar Pro, find and delete the 'Workshop' event scheduled for October 24th.	0.727
72	In Simple Calendar Pro, delete the calendar event on 2023-10-30 at 11h with the title 'Catch up on Annual Report'	In Simple Calendar Pro, find and delete the 'Workshop' event scheduled for October 24th.	0.751
73	What is on my schedule for October 19 2023 at 21:45 in Simple Calendar Pro? Answer with the titles only. If there are multiples titles, format your answer in a comma separated list.	In Simple Calendar Pro, switch the calendar view to 'Monthly and daily view', and then list the titles of all events scheduled for October 23rd. Answer with the titles separated by a comma.	0.757

(continued on next page)

(continued from previous page)

#	AndroidWorld Test Instruction	Most Similar Synthetic Instruction	Sim.
74	What events do I have in the next week in Simple Calendar Pro? Assume the week starts from Monday. Answer with the titles only. If there are multiples titles, format your answer in a comma separated list.	In Simple Calendar Pro, identify all events scheduled for October 17th and answer with their titles separated by a comma.	0.750
75	Do I have any events between 4pm and 8pm October 27 2023 in Simple Calendar Pro? Answer with the titles only. If there are multiples titles, format your answer in a comma separated list.	In Simple Calendar Pro, identify all events scheduled for October 17th and answer with their titles separated by a comma.	0.744
76	What events do I have October 17 2023 in Simple Calendar Pro? Answer with the titles only. If there are multiple titles, format your answer as a comma separated list.	In Simple Calendar Pro, identify all events scheduled for October 17th and answer with their titles separated by a comma.	0.831
77	What is my first event after 11:00am October 21 2023 in Simple Calendar Pro? Answer with the titles only. If there are multiples titles, format your answer in a comma separated list.	In Simple Calendar Pro, switch the calendar view to 'Monthly and daily view', and then list the titles of all events scheduled for October 23rd. Answer with the titles separated by a comma.	0.773
78	What is the location of my Coding challenge event in Simple Calendar Pro? Answer with the location only.	In Simple Calendar Pro, what is the description for the 'Call with the Team' event, and is there a location currently set for it?.	0.597
79	What is my next upcoming event in Simple Calendar Pro? Answer with the title only. If there are multiples titles, format your answer in a comma separated list.	In Simple Calendar Pro, identify all events scheduled for October 17th and answer with their titles separated by a comma.	0.753
80	When is my next meeting with Muhammad in Simple Calendar Pro? Express your answer in the format <month name> <day> <year> <hour in 24-hour format>:<minutes>.	In Simple Calendar Pro, switch the view to the 'Simple event list', find the 'Meeting with Marketing' event, and answer with its start time in HH:MM format.	0.629
81	Create a new drawing in Simple Draw Pro. Name it amet_lively_eagle_final.jpg. Save it in the Pictures folder within the sdk_gphone_x86_64 storage area.	Using Simple Draw Pro, save a file named 'system_diagram' directly to the root directory of the internal storage (sdk_gphone64_arm64).	0.753
82	Reply to +13431223053 with message: Actions speak louder than words. in Simple SMS Messenger	In Simple SMS Messenger, find the conversation containing the message 'Actions speak louder than words.', and add the sender to your device's contacts with the first name 'Wisdom'.	0.669
83	Reply to the most recent text message using Simple SMS Messenger with message: When in Rome, do as the Romans do.	Find and delete the conversation thread containing the message "When in Rome, do as the Romans do." in the Simple SMS Messenger app.	0.621
84	Resend the message I just sent to Lily Pereira in Simple SMS Messenger	In Simple SMS Messenger, resume the draft conversation with +17401638798 by appending ' hope you are well' to the existing text and sending it.	0.572
85	Send a text message using Simple SMS Messenger to +15039078312 with message: Lorem Ipsum is simply dummy text.	Using Simple SMS Messenger, send a text message with the content 'padiNoBMVR' to the number +1 545-178-61614309.	0.631
86	Send a message to +17228051441 with the clipboard content in Simple SMS Messenger	In Simple SMS Messenger, use the dialer to enter the number +1 545-178-6161 and then start a new text message to this recipient.	0.571
87	Text the address of the event to David Wang that Emily Liu just sent me in Simple SMS Messenger	In Simple SMS Messenger, send a text message to +1 930-572-4145+2 with the content 'Please confirm if the meeting is still on for today'.	0.546

(continued on next page)

(continued from previous page)

#	AndroidWorld Test Instruction	Most Similar Synthetic Instruction	Sim.
88	How many skate boarding activities did I do this week in the OpenTracks app? Assume the week starts from Monday. Express your answer as a single integer.	In the OpenTracks app, how many activities shown in the list were recorded on a Thursday? Answer with a single number.	0.718
89	What activities did I do October 6 2023 in the OpenTracks app? Answer with the activity type only. If there are multiple types, format your answer in a comma separated list.	In the OpenTracks app, identify all activities that have a recorded distance greater than 9 miles. Answer with the names of the activities separated by a comma.	0.693
90	How long was my climbing activity October 15 2023 in the OpenTracks app? Express your answer in minutes as a single integer.	In the OpenTracks app, what are the moving time and elevation gain recorded for the 'Morning Run' activity? Answer with the values separated by a comma.	0.611
91	What was the longest distance covered in a kayaking activity in the OpenTracks app this week? Assume the week starts from Monday. Express your answer as a single number in meters rounded to the nearest integer.	In the OpenTracks app list, how many recorded activities have a distance greater than 10 miles? Answer with a single number.	0.666
92	What was the total distance covered for swimming activities in the OpenTracks app from October 6 2023 to October 15 2023? Express your answer as a single number in meters rounded to the nearest integer.	In the OpenTracks app list, how many recorded activities have a distance greater than 10 miles? Answer with a single number.	0.659
93	What was the total duration of hiking activities in the OpenTracks app this week? Assume the week starts from Monday. Express your answer in minutes as a single integer.	In the OpenTracks app, how many activities shown in the list were recorded on a Thursday? Answer with a single number.	0.618
94	Turn bluetooth off.	Turn off the Nearby Share feature completely, then return to the Connection preferences menu and open the Bluetooth settings.	0.638
95	Turn bluetooth off.	Turn off the Nearby Share feature completely, then return to the Connection preferences menu and open the Bluetooth settings.	0.638
96	Turn bluetooth on.	Navigate to the Bluetooth pairing screen in Settings and find the phone's Bluetooth address. Answer the question with the full address string.	0.515
97	Turn bluetooth on.	Navigate to the Bluetooth pairing screen in Settings and find the phone's Bluetooth address. Answer the question with the full address string.	0.515
98	Turn brightness to the max value.	Configure the device display for better visibility by enabling 'Dark theme' and setting the 'Font size' to the largest available option.	0.403
99	Turn brightness to the max value.	Configure the device display for better visibility by enabling 'Dark theme' and setting the 'Font size' to the largest available option.	0.403
100	Turn brightness to the min value.	Configure the device display for better visibility by enabling 'Dark theme' and setting the 'Font size' to the largest available option.	0.316
101	Turn brightness to the min value.	Configure the device display for better visibility by enabling 'Dark theme' and setting the 'Font size' to the largest available option.	0.316
102	Copy the following text to the clipboard: Reservation under: Mike	In the Markor app, append the sentence ' Dinner reserved at 7pm.' to the existing text in the file named '2023_08_11_good_vase.txt'.	0.375

(continued on next page)

(continued from previous page)

#	AndroidWorld Test Instruction	Most Similar Synthetic Instruction	Sim.
103	Turn wifi off.	Turn off the "Adaptive connectivity" feature and set the Private DNS mode to "Off".	0.510
104	Turn wifi off.	Turn off the "Adaptive connectivity" feature and set the Private DNS mode to "Off".	0.510
105	Turn wifi on.	Set up a portable Wi-Fi hotspot named 'TravelRouter' with the password 'SecureNet99', and turn it on to share your cellular internet connection.	0.488
106	Turn wifi on.	Set up a portable Wi-Fi hotspot named 'TravelRouter' with the password 'SecureNet99', and turn it on to share your cellular internet connection.	0.488
107	Which tasks have I completed for October 18 2023 in Tasks app? Answer with the titles only. If there are multiples titles, format your answer in a comma separated list.	In the Tasks app, identify the titles of all tasks that are specifically due on 'Oct 8'. Answer by listing the titles separated by a comma.	0.782
108	How many tasks do I have due next week in Tasks app? Assume the week starts from Monday. Express your answer as a single integer.	In the Tasks app, count the number of visible tasks that are due on 'Tue'. Answer with the single number.	0.715
109	What tasks do I have due October 21 2023 in Tasks app? Answer with the titles only. If there are multiples titles, format your answer in a comma separated list.	In the Tasks app, identify the titles of all tasks that are specifically due on 'Oct 8'. Answer by listing the titles separated by a comma.	0.773
110	What are my high priority tasks in Tasks app? Answer with the titles only. If there are multiples titles, format your answer in a comma separated list.	In the Tasks app, how many tasks are currently active (not completed), and which of them has the highest priority? Answer the question with the number and task title, separated by a comma.	0.759
111	Which tasks with high priority are due October 16 2023 in the Tasks app? Answer with the title only. If there are multiples titles, format your answer in a comma separated list.	In the Tasks app, identify the titles of all tasks that are specifically due on 'Oct 8'. Answer by listing the titles separated by a comma.	0.779
112	What incomplete tasks do I have still have to do by October 21 2023 in Tasks app? Answer with the titles only. If there are multiples titles, format your answer in a comma separated list.	In the Tasks app, identify the titles of all tasks that are specifically due on 'Oct 8'. Answer by listing the titles separated by a comma.	0.740
113	Turn off WiFi, then enable bluetooth	Turn off the Nearby Share feature completely, then return to the Connection preferences menu and open the Bluetooth settings.	0.581
114	Turn on Wifi, then open the contacts app	Open the Contacts app and determine how many contacts are currently saved in the list. Answer with a single number.	0.516
115	Create a playlist titled "Documentary Insights Favorites" with the following files in VLC (located in Internal Memory/VLCVideos), in order: 2023_01_29_episode_46_HD.mp4, 2023_06_29_clip_38_export.mp4	In the VLC app, locate the 'Documents' folder within the internal memory storage and add it to the 'Favorites' list.	0.615

(continued on next page)

(continued from previous page)

#	AndroidWorld Test Instruction	Most Similar Synthetic Instruction	Sim.
116	Create a playlist titled "Recipe Collection Ultimate Collection" with the following files in VLC (located in Internal Memory/VLCVideos), in order: moment_95__1KUB.mp4, scene_54_raw_gbYs.mp4, moment_52_HD_final.mp4, highlight_13_HD_2023_01_29.mp4. And then, create a playlist titled "Ultimate Fails Ultimate Collection" with the following files in VLC, in order: recording_41_HD_backup.mp4, recording_56__JRVN.mp4, 2023_08_10_scene_27_raw.mp4.	In the VLC app, navigate to the Browse tab and create a new playlist named "My Top Hits" using all the media files found in the "Music" folder.	0.619