ScaleTrack: 缩放与回溯自动化 GUI 代理

Jing Huang[1*], **Zhixiong Zeng**[1*†], **Wenkang Han**[1,2], **Yufeng Zhong**[1], **Liming Zheng**[1], **Shuai Fu**[3], **Jingyuan Chen**[2], **Lin Ma**[1†]

[1]Meituan

[2]Zhejiang University

[3]University of Adelaide

Abstract

自动化GUI代理旨在通过在数字环境中自动执行复杂任务来简化用户交互，例如网页、移动设备和桌面设备。它接收文本任务指令和 GUI 描述以生成可执行操作（例如，点击）并逐步操作框。训练一个 GUI 代理主要涉及基础设置和规划阶段，在这个过程中，GUI 基础设置集中在根据任务找到执行坐标，而规划阶段则旨在基于历史动作预测下一个动作。然而，先前的工作在 GUI 基础设置方面的训练数据不足，并且忽略了回溯历史行为对 GUI 规划的影响。为了解决上述挑战，我们提出了 ScaleTrack，一个通过扩展基础设置和回溯规划来训练自动化 GUI 代理的框架。我们从广泛的来源精心收集了不同合成标准的 GUI 样本，并将它们统一到同一模板中进行训练 GUI 基础设置模型。此外，我们设计了一种新的培训策略，该策略预测当前 GUI 图像的下一个动作，同时回溯导致该 GUI 图像的历史动作。通过这种方式，ScaleTrack 解释了 GUI 图像与动作之间的对应关系，有效地描述了 GUI 环境的发展规则。广泛的实验结果证明了 ScaleTrack 的有效性。数据和代码将在 url 上提供。

1 介绍

GUI 代理旨在为图形用户界面（GUI）开发原生自动化代理，并满足用户在数字环境中自动执行复杂任务的日益增长的需求。它在移动/计算使用领域引起了广泛关注，可以提供准确的任务完成和便捷的用户交互。得益于大型语言模型 Team et al. (2024); Achiam et al. (2023) 新兴的推理能力，GUI 代理能够从任务指令中进行推理并规划多步可执行操作。

早期的 GUI 代理 Kim et al. (2023); Zheng et al. (2023) 提取描述 GUI 环境的结构化文本（例如，网站 HTML），然后利用大型语言模型进行推理和规划以生成可执行的操作。实际上，GUI 环境的结构化文本信息往往冗长且可能无法访问，并且伴随着位置和布局等属性信息不足的问题。因此，最近的研究 Cheng et al. (2024); Wu et al. (2024) 集中于基于多模态大型语言模型（MLLMs）的视觉 GUI 代理，这些代理仅依赖于 GUI 环境的截图来进行用户交互。

*Equal contribution.

†Corresponding author.

Preprint. Under review.

为了基于视觉截图训练 GUI 代理，现有方法 Xu et al. (2024); Qin et al. (2025) 通常遵循将多模态知识从 MLLMs（例如, QWen2-VL）转移到 GUI 环境的范式。通常，之前的工作通常会在两种类型的 GUI 数据上微调 MLLMs：一种是用于预测坐标位置的 GUI 定位，另一种是用于预测动作的 GUI 规划。前者侧重于根据任务指令预测相关的坐标位置，而后者则专注于预测多步可执行的动作，并最终实现自动化任务执行。

然而，现有的 GUI 定位方法主要依赖于孤立的数据合成标准，从包括移动、网络和桌面的大批元数据中生成定位数据。具体来说，它们采用了一种强大的 LLM（例如，GPT-4o），该模型利用不同平台上的用户界面元数据（例如，所有文本/图标/部件的元数据）来合成元素描述。几种方法强调了孤立的数据合成标准，在这种方法中，UgroundGou et al. (2024) 使用多种参考模式合成定位数据，Aria-UIYang et al. (2024) 合成具有上下文感知能力的定位数据，而 AugvisXu et al. (2024) 则合成模板增强型定位数据。不幸的是，先前的工作忽略了使用不同合成方法生成的定位数据之间的互补性，因此很少将它们整合起来以扩展 GUI 定位的训练过程。

此外，现有的研究仅考虑基于任务指令和 GUI 环境预测下一步行动的正向规划。通常依赖于人工标注来收集各种任务指令的行动轨迹 Deng et al. (2023)，或者提示 MLLMs 生成详细的推理思路 Xu et al. (2024)。事实上，GUI 环境遵循与任务执行内在相关的固有模式，例如在当前状态下规划下一步行动，或回溯导致当前状态的历史行动。然而，现有的工作仅在训练过程中收集正向规划数据，缺乏对回溯数据的收集以及相应训练策略的探索。

在本文中，我们介绍了 ScaleTrack，一个通过扩展 GUI 定位和回溯 GUI 规划来应对上述挑战的新训练框架。首先，我们利用几种基于数据的 GUI 元素增强方法来扩展 GUI 定位的训练过程，包括元素引用、上下文感知和功能描述。为此，我们的工作集成了由孤立数据合成标准生成的各种定位样本，并将它们统一到一个固定的训练模板中。然后，我们设计了一个数据模板，整合了多个连续 GUI 图像之间的动作标注，并收集向前规划的下一个动作的真实值以及用于回溯的历史动作真实值（如图 1 所示）。最后，为了学习任务执行中的内在模式，我们设计了一种混合训练策略，包括向前规划和回溯，在这种策略中，GUI 代理需要同时预测当前状态下接下来的动作和历史动作。实验中的实证结果表明，回溯有效提高了 GUI 代理的任务执行能力。

本工作的主要贡献如下：

- 我们提出了第一个具有回溯能力的 GUI 代理，并设计了有效的数据构建和训练策略。
- 我们将多个数据综合准则整合到 GUI 元素描述中，显著扩大了训练过程，从而带来了性能的一致性提升。
- 我们在多个基准数据集上进行了广泛的实验，包括接地评估、离线和在线评估，实验证明了我们提出的 ScaleTrack 的有效性。

2 相关工作

2.1 多模态大语言模型

近期闭源和开源的多模态大语言模型（MLLMs）显著提升了非自然图像理解和 GUI 任务规划的能力。像 GPT-4V 202 (2023) 和 GPT-4o Hurst et al. (2024) 这样的闭源模型，具备强大的视觉和任务规划能力，通常作为 “大脑” 服务于规划与接地分离的 GUI 代理框架中。开源模型使 GUI 代理能够在特定领域进行技能精进。Qwen-VL Bai et al. (2023); Wang et al. (2024a); Bai et al. (2025) 系列通过精细的视觉理解和多模态能力脱颖而出，其中 Qwen2-VL Wang et al.

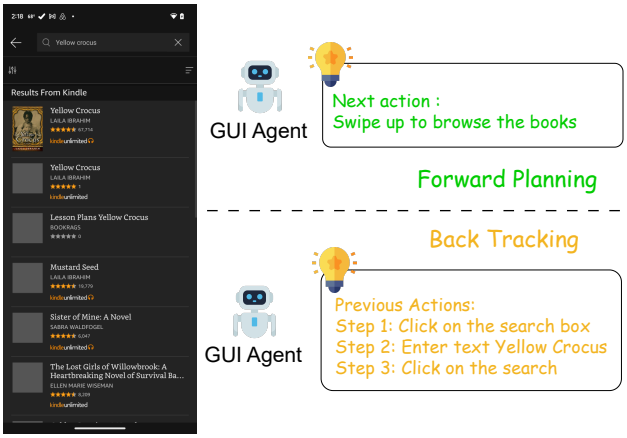

图 1: 前向规划与回溯的区别。

(2024a) 特别作为大多数之前 GUI 代理实现的基础 MLLM 骨干。其他开源模型在 GUI 代理领域具有独特的技术优势。InternVL-2 Chen et al. (2024) 通过逐步对齐提升多模态任务性能。CogVLM Wang et al. (2024b) 使用视觉专家模块融合视觉和语言特征。Ferret You et al. (2023) 通过增强的空间理解能力提升人机交互精度。LLAVA Liu et al. (2023, 2024); Li et al. (2024a) 系列因其轻量级投影层而被使用，实现了快速训练和多模态理解。

2.2 图形用户界面代理

近期构建于 MLLMs 的 GUI 代理可以按照人类指令在手机或计算机上执行自主操作，适用于网络、桌面和移动场景。它们的操作包括规划和接地阶段。根据这两个阶段是否由不同的模型处理，现有的 GUI 代理工作可分为纯粹的 GUI 接地模型如 Aria-UI Yang et al. (2024) 和 Uground Gou et al. (2024)，以及统一的 GUI 代理框架如 Aguvis Xu et al. (2024)、OS-ATLAS Wu et al. (2024) 和 UI-TARS Qin et al. (2025)。

在感知内容方面，早期的作品如 WebPilot Zhang et al. (2025), Hybrid Agent Song et al. (2024), WebDreamer Gu et al. (2024), Agent-Q Putta et al. (2024), LASER Ma et al. (2023) 和 SeeAct Zheng et al. (2024) 集中在 web 平台上，以自动化网络交互和导航。它们将结构化文本（例如、可访问性树、HTML-DOM）与环境视觉结构（截图）结合起来，在 GUI 代理领域奠定了基础。然而，获取桌面和 iOS 应用程序等现实世界设置中的结构化文本的难度以及这种文本在代币效率上的不足影响了 MLLM 的表现，这促使转向仅基于视觉的方法。Claude 3.5 十四行诗（计算机使用）Hu et al. (2024) 领先提出了这一范式，在桌面任务自动化中整合了任务规划和环境交互动作预测到一个系统中。

仅使用视觉的方法的兴起也引发了跨平台统一建模。近期的方法如 Aguvis Xu et al. (2024)、UI-TARS Qin et al. (2025) 和 OS-ATLAS Wu et al. (2024) 使用了统一的动作空间，并在多平台数据集上进行了训练。Aguvis 在其框架中整合了显式的规划和推理，增强了与复杂数字环境的交互。OS-ATLAS 提出了一个标准化跨平台数据集的统一动作空间，涵盖了基本和自定义动作。UI-TARS 在模型的规划阶段应用了多样化的推理模式，包括任务分解、反思和里程碑识别。

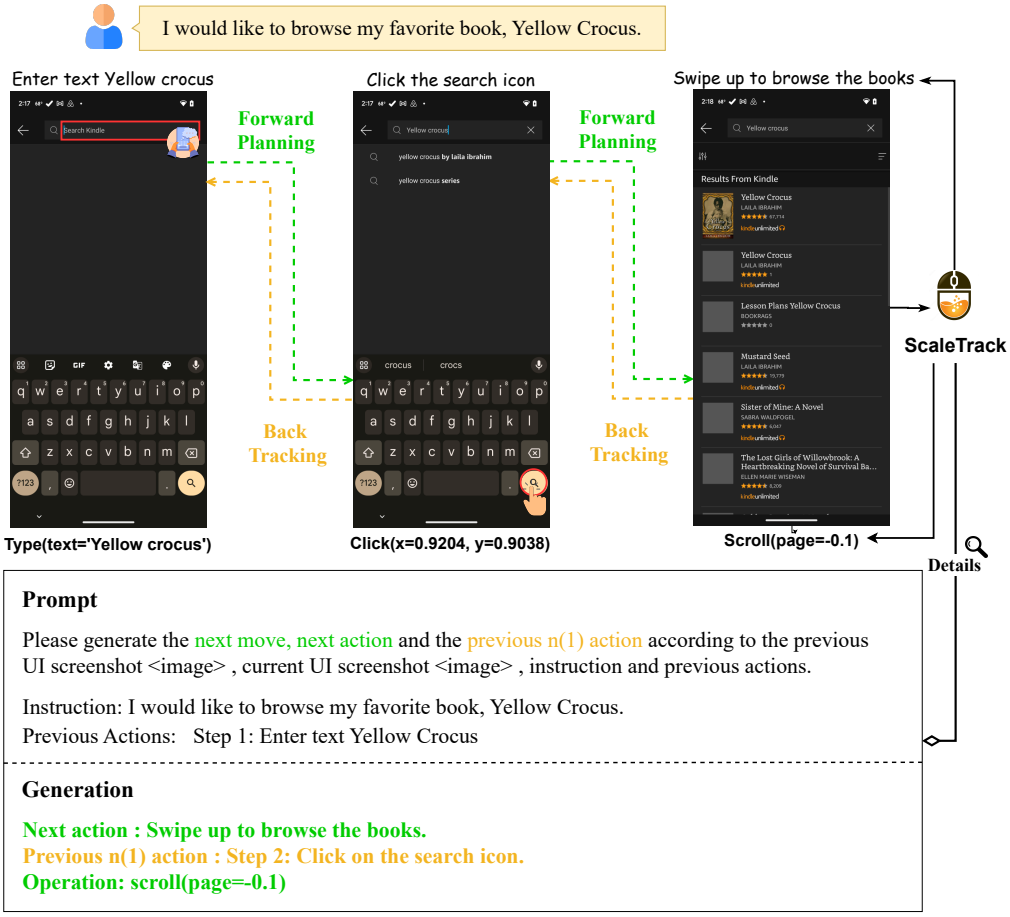

图 2: 我们提出的 ScaleTrack 在处理任务指令、通过前向规划和回溯生成动作以及训练数据格式方面的总体描述。

3 方法

ScaleTrack 的优化分为两个阶段：定位和规划。在第 3.2 节中，我们介绍了定位阶段的数据规模。在第 3.3 节中，我们讨论了规划阶段的正向规划和回溯。

3.1 公式推导

给定任务描述和初始环境观察 o_1，自主 GUI 进行规划并预测一个属于动作空间的动作，即 $a_1 \in \mathbb{A}$。随后，客户端环境在接收到该动作后更新，并提供新的观察 o_2。上述过程持续重复直至预测动作为终止，这标志着任务的完成。整个过程可以表示为：

$$a_n = \mathcal{M}_\theta(\text{task}, (o_1, a_1), ..., (o_{n-1}, a_{n-1}), o_n), \tag{1}$$

其中 $\mathcal{M}$ 是策略，相当于 GUI 代理模型，而 θ 表示其相关参数。

3.2 不同领域和类型的数据规模

3.2.1 统一数据格式

如何表示 GUI 截图中文本、图标和控件的空间信息是 GUI 代理需要解决的主要问题。所提出的 ScaleTrack 将实际坐标映射到 0 到 1000 之间的相对坐标，然后对其进行缩放以获得一个介于 0 和 1 之间的值，以此来表示相对于图像大小的距离。使用相对坐标的优点在于它们为不同分辨率的图像提供了一种统一的表示方法，并允许自由调整大小同时保持相同的纵横比，从而满足具有不同输入令牌长度约束的模型需求。

在之前的 GUI 接地数据集标注中，目标元素的坐标通常以框的形式存在：$(x1, y1, x2, y2)$，其中 $(x1, y1)$ 和 $(x2, y2)$ 表示包围该元素的最小边界框的左上角和右下角的坐标。然而，在许多 GUI 代理操作中，基于点击的交互更为普遍。例如，当用户在 GUI 上点击按钮或链接时，本质上是一种点操作。因此，我们将坐标表示统一为点格式，以便更好地满足 GUI 代理任务的操作要求，从而提高元素接地和互动的准确性和效率。

3.2.2 合并数据源

早期的工作中，智能体通过提取的结构化数据（例如，HTML-DOM）与环境进行交互，并使用纯文本中的大型语言模型进行任务规划。自 SeeClickCheng et al. (2024) 以来，许多工作尝试将截图的纯粹视觉信息作为输入来实现不同界面和平台之间的泛化。然而，与计算机视觉领域积累的大量通用图像数据相比，GUI 智能体领域涉及的基础数据仍存在很大局限性：来自不同平台和设备的数据难以泛化，并且不同的任务使用孤立的数据合成标准，这使得彼此之间难以互补。

为了解决 GUI 定位中的泛化问题，我们将来自不同来源且具有不同合成标准的数据融合在了一起。具体而言，借鉴 Uground Gou et al. (2024)，我们引入了通过整合多种参考模式构建的大量网站定位数据。跟随 Aria-UI Yang et al. (2024) 的思路，我们引入了通过集成上下文感知构建的具有情景意识的定位数据。受到 Aguvis Xu et al. (2024) 的启发，我们引入了通过统一动作空间构建的模板增强型定位数据。如表 1 所示，我们的工作统一了这些数据，并研究不同的数据合成标准是否可以互相补充，从而从各个方面提高 Agent 定位的泛化能力。

在现实世界的情景中，一个复杂的截图可能包含数百个 UI 元素，现有的开源基础数据集自然会在单张图像内包括多个对象。为了避免重复加载图像的冗余操作并减少训练开销，我们将来自同一截图的不同指令/描述-答案对合并到单个对话中，从而创建多轮训练数据。

3.3 增强动作理解与回溯联合

3.3.1 从预测未来到回溯其中

如第 3.1 节所述，GUI 代理的训练规划阶段通常被建模为部分可观测马尔可夫决策过程。它向模型提供过去的行为和状态感知，只需要模型进行前向规划，并基于这些预测未来行动的概率。然而，这种方法忽视了模型反思和回溯历史决策的能力。也就是说，模型知道自己到达的状态，但不知道是如何达到这个状态的。为了克服这一局限性，我们通过引入回溯来扩展 GUI 代理与其环境之间的交互。具体来说，在每个时间步骤 t，ScaleTrack 不仅预测在当前总体目标下的下一个行动，还预测导致当前状态的历史行动。这可以表述为：

$$a_{n-1}, a_n = \mathcal{M}_\theta(\text{task}, (o_1, a_1), ..., (o_{n-1}, a_{n-1}), o_n) \tag{2}$$

表 1: 开源接地和规划数据的统计。

数据源	数据类型	接地		多步	
		元素	截图	轨迹	平均步骤
地表	Web	9M	773K	/	/
操作系统图谱	Web	7.8M	1.6M	/	/
	Mobile	1.1M	107K	/	/
	Desktop	11.3M	54K	/	/
Aria-UI	Web	-	173K	/	/
	Mobile	-	104K	/	/
	Desktop	-	7.8K	/	/
阿古维斯	Web	723K	-	6.3k	6.7
	Mobile	232K	-	28.7K	8.5
	Desktop	7K	-	/	/
Ours	Ours	*	7.5M	*	8.2

在前向规划方面，类似于传统方法，ScaleTrack 将当前状态和任务指令作为输入，并生成下一步行动的概率分布。这使代理能够确定在当前状态下最有可能执行的下一步行动，从而实现逐步的任务执行。

关于回溯，ScaleTrack 引入了一种逆向预测机制。根据当前状态和任务指令，它预测在到达当前状态之前可能发生的行为。通过这种方式，智能体可以更清晰地了解其达到当前状态的路径，从而更好地评估其先前行为的合理性并相应调整后续计划。

3.3.2 先前动作的不同表达方式

状态观测实现了从结构化数据（如 HTML）到截图的转换。之前的动作也有两种不同的表达形式：一种是自然语言中的低级指令，另一种是实际的操作动作序列（例如，图 2 中每个帧上下方的文字标注）。为了研究之前动作的不同表达方式对 VLM 理解历史动作的影响，我们在训练和测试阶段都进行了实验。

4 实验

在本节中，我们分别进行了 GUI 接地评估和离线/在线 GUI 代理评估的实验。我们选择了 Qwen2-VL-7BWang et al. (2024a) 作为基础模型进行训练，并使用第 4.1.1 节介绍的数据对其进行微调。训练分为两个步骤：接地和带有回溯的规划。

4.1 训练详情

4.1.1 训练数据

对于基础阶段的训练，我们整合了开源数据：OS-Atlas Wu et al. (2024)，Uground Gou et al. (2024)，Aguvis Xu et al. (2024) 和 Aria-UI Yang et al. (2024)，并将它们标准化为统一格式。我们在表 1 中提供了训练的基本数据统计。在规划阶段的训练中，我们选择了带有观察、思考和低级指令标注的数据集 Aguvis Xu et al. (2024) 作为数据源，并进行了回溯变换。此外，

为了研究先前动作格式对模型理解动作序列的影响，我们将数据的先前动作替换以获得一个副本。

4.1.2 训练设置

我们选择 Qwen2-VL(Wang et al., 2024a) 作为训练的基础模型，并使用 AdamW 优化器，学习率为 $1e-5$，采用余弦学习率调度器，预热比例为 0.03 步。我们在基础阶段利用全局批量大小为 128，在规划阶段为 64，并采用了 DeepSpeed ZERO3 风格的数据并行。我们遵循两阶段过程训练 ScaleTrack。首先，使用所有基础数据来训练 ScaleTrack 的基本 GUI 定位能力。然后，基于在基础阶段训练的模型，将带有前向规划和历史回溯的规划数据输入模型以进一步增强模型的规划能力。我们在一个由 4 节点 V100-80G GPU 组成的集群上训练 ScaleTrack。

4.2 接地能力评估

为了评估数据缩放对代理的接地能力的影响，我们在 ScreenSpot 数据集上进行了实验。我们首先比较了我们的模型与其他基线的准确性，然后评估了在不同数据规模下模型接地能力的变化。

屏幕点。表 2 报告了我们提出的 ScaleTrack 以及各种基线模型的准确率得分，其中包括 UGroundGou et al. (2024)、Aria-UI Yang et al. (2024)、OS-AtlasWu et al. (2024)、Aguvis Xu et al. (2024) 和 UI-TARS Qin et al. (2025) 等开源模型，这些模型使用了内部数据。我们可以从表格中看出，ScaleTrack 的性能明显超越了那些使用开源数据的基线方法，并且在平均得分上比之前的最先进 (SOTA) 模型 Xu et al. (2024) 高出 1.2%。此外，在更难概括的图标/小部件子项上，ScaleTrack 获得了更加明显的优点，分别提升了 4.4%，8.6%和 7.8%。结果表明了通过数据扩展获得的泛化能力。ScaleTrack 使用的综合数据合成策略比任何孤立的数据合成标准都取得了更好的效果，我们的方法的优势也显示了结合来自不同来源数据的优点。

表 2: ScreenSpot 上各种规划器和基础方法的比较。

方法		数据来源	移动		桌面		网络		平均值
			文本	图标/小部件	文本	图标/小部件	文本	图标/小部件	
代理框架									
GPT-4	SeeClick	Public	76.6	55.5	68.0	28.6	40.9	23.3	48.8
	OmniParser	In-house	93.9	57.0	91.3	63.6	81.3	51.0	73.0
	UGround-7B	Public	90.1	70.3	87.1	55.7	85.7	64.6	75.6
GPT-4o	SeeClick	Public	81.0	59.6	69.6	33.6	43.9	26.2	52.3
	UGround-7B	Public	93.4	76.9	92.8	67.9	88.7	68.9	81.4
代理模型									
GPT-4o		In-house	20.2	24.9	21.1	23.6	12.2	7.8	18.3
克劳德计算机使用		In-house	-	-	-	-	-	-	83.0
双子座 2.0		In-house	-	-	-	-	-	-	84.0
UI-TARS-7B		In-house	94.5	85.2	95.9	85.7	90.0	83.5	89.5
认知代理		Public	67.0	24.0	74.2	20.0	70.4	28.6	47.4
见点击		Public	78.0	52.0	72.2	30.0	55.7	32.5	53.4
Qwen2-VL		Public	75.5	60.7	76.3	54.3	35.2	25.7	55.3
UGround-7B		Public	82.8	60.3	82.5	63.6	80.4	70.4	73.3
OS-Atlas-7B		Public	93.0	72.9	91.8	62.9	90.9	74.3	82.5
AGUVIS-7B		Public	95.6	77.7	93.8	67.1	88.3	75.2	84.4
ScaleTrack-7B		Public	93.8	82.1	91.7	75.7	87.4	83.0	86.8

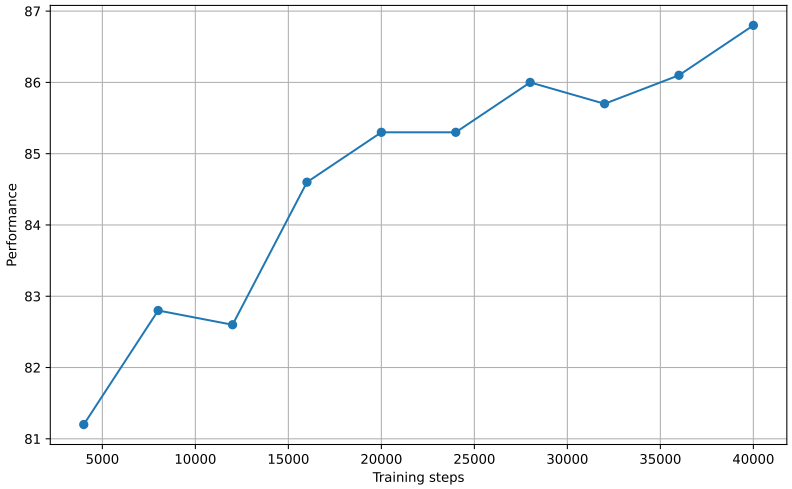

图 3: ScaleTrack-7B 在 ScreenSpot 上的缩放曲线。

接地数据缩放的影响。为进一步分析接地数据扩展的有效性，我们绘制了 ScaleTrack 在 ScreenSpot 不同训练步骤中的准确率得分。如图 3 所示，随着数据规模的增加，平均准确率得分会出现波动，但总体上会逐渐提高。结果表明，持续扩大接地数据以提升性能具有巨大潜力。

4.3 离线代理能力评估

Android 控制。我们在移动设备上使用 Android 自动化数据集 AndroidControl 评估 ScaleTrack，该数据集包含来自人类评分者在 833 个不同应用（跨越 Android 设备上的 40 个应用程序类别）执行各种任务的 15,000 个演示。根据 Li et al. (2024b); Xu et al. (2024)，我们随机抽取 800 步来创建一个子集。我们在高阶和低阶任务中的域外数据上报告动作类型准确性、定位准确性和步骤成功率。AndroidControl-High 的评估过程依赖于历史操作输入，我们严格遵循 OS-AtlasWu et al. (2024) 使用低级自然语言描述历史操作。

如表 3 所示，ScaleTrack 在 Low 和 High 两个级别上都超越了使用现有数据的强大基线，实现了 Low 级别的步成功率为 86.6%，High 级别的步成功率为 77.9%。值得注意的是，ScaleTrack 的规划阶段的数据来源与 Aguvis 相同，并未添加额外的规划数据标注，这证明了我们的回溯策略的有效性和可扩展性。

GUI-奥德赛。GUI Odyssey 是一个用于评估跨应用导航代理的综合数据集。它由来自 6 台移动设备的 7,735 个片段组成。根据 Xu et al. (2024)，我们随机抽取 500 个片段创建了一个子集，并报告了动作类型准确性、定位准确性和步成功率。表 4 报告了 ScaleTrack-7B 在步成功率方面在公开数据模型中表现最佳。

AndroidControl 中动作描述的影响。值得注意的是，在之前的工作中，我们在对 AndroidControl 的高层次子集进行测试时发现了输入组织方式上的差异。一些研究工作，如 Aguvis Xu et al. (2024)，似乎使用动作序列（即“点击”和“类型”）作为历史操作的描述。相比之下，其他研究工作，如 OS-Atlas Wu et al. (2024)，则使用低级指令（例如“点击问题如何取消或更改我的预订?”）作为历史操作的描述。这种差异影响了模型对历史操作序列的理解。为了

表 3: 在 AndroidControl 数据集上两种设置（AndroidControl-Low 和- High）的比较结果。

代理模型	数据源	Android 控制-低			Android 控制-高级		
		类型	接地	SR	类型	接地	样本率
Claude	In-house	74.3	0.0	19.4	63.7	0.0	12.5
GPT-4o	In-house	74.3	0.0	19.4	66.3	0.0	20.8
InternVL-2-4B	In-house	90.9	84.1	80.1	84.1	72.7	66.7
Qwen-2VL-7B	In-house	91.9	86.5	82.6	83.8	77.7	69.7
UI-TARS-7B	In-house	98.0	89.3	90.8	83.7	80.5	72.5
SeeClick	Public	93.0	73.4	75.0	82.9	62.9	59.1
Aria-UI	Public	-	87.7	67.3	-	43.2	10.2
OS-Atlas-4B	Public	91.9	83.8	80.6	84.7	73.8	67.5
OS-Atlas-7B	Public	93.6	88.0	85.2	85.2	78.5	71.2
Aguvis-7B	Public	-	-	80.5	-	-	61.5
ScaleTrack-7B	Public	93.9	84.9	86.6	89.2	72.8	77.9

表 4: GUI Odyssey 数据集上的比较结果。

代理模型	数据来源	图形用户界面之旅		
		类型	接地	SR
Claude*	In-house	60.9	0.0	3.1
GPT-4o	In-house	34.3	0.0	3.3
InternVL-2-4B	In-house	82.1	55.5	51.5
Qwen-2VL-7B	In-house	83.5	65.9	60.2
UI-TARS-7B	In-house	94.6	90.1	87.0
SeeClick	Public	71.0	52.4	53.9
Aria-UI	Public	-	86.8	36.5
OS-Atlas-4B	Public	83.5	61.4	56.4
OS-Atlas-7B	Public	84.5	67.8	62.0
Aguvis-7B	Public	-	-	-
ScaleTrack-7B	Public	85.6	69.3	65.3

促进评估过程和该领域研究的一致性，我们使用这两种描述进行了测试，并提供了结果的详细分析与比较。

我们在没有回溯的情况下对数据集进行了实验。如表 5 所示，当训练和测试分别使用指令和动作形式来描述之前的动作时，保持训练和测试设置相同将获得更好的性能，否则准确率会下降。这表明对于 GUI 代理而言，以一致的方式表达上下文中的动作是非常重要的，这对规划阶段的数据规模和数据标注具有关键的指导意义。

4.4 在线代理能力评估

为了更好地在实际环境中测试 ScaleTrack 的性能，我们也对 ScaleTrack 进行了实时交互基准测试。我们在 Android 模拟器环境下使用 AndroidWorld Rawles et al. (2024) 和 MobileMini-

表 5: 不同动作描述在 AndroidControl-High 数据集中的影响。

训练-测试	安卓控制-高级		
	类型	接地	样品率
Instruction-Action	81.8	71.6	66.7
Action-Action	89.2	75.2	77.4
Action-Instruction	82.3	66.2	68.3
Instruction-Instruction	88.3	73.7	76.1

WobRawles et al. (2023) 进行在线移动代理评估。我们使用 GPT-4o 作为规划器，并用 CaleTrack-7B 来定位元素和指令。

安卓世界 Rawles et al. (2024) 包含在 20 个应用中手工制作的 116 项高度可重复任务的基准测试，并通过检查最终系统状态来计算设备上的最终状态成功率。如表 6 所示，ScaleTrack 实现了最高的平均任务成功率为 44%，超过了基线模型，这进一步突显了数据扩展和回溯有助于模型处理现实环境中的多样化元素描述和指令。

移动迷你工作表 ~ Rawles et al. (2023) 包含 92 个来自 MiniWob++ Zheng et al. (2023) 的任务。如表 6 所示，在使用 GPT-4o 作为规划器时，ScaleTrack 在 MobileMiniWobs 上的任务成功率超过了现有工作，分别在 AndroidWorld 和 MobileMiniWob 上达到了平均 SR 为 44.0% 和 61%。此比较特别突出了我们在 GUI 代理模型中的数据扩展和回溯的有效性。

表 6: 任务成功率（SR）在 AndroidWorld 和 MobileMiniWob 上的表现。

规划者	接地	Android 世界_样本率	MobileMiniWob_样本率
GPT-4-Turbo	UGround	31.0	-
GPT-4o	UGround	32.8	-
GPT-4o	AGUVIS-7B	37.1	55.0
GPT-4o	ScaleTrack-7B	44	61.0

4.5 消融研究

我们进一步研究了 ScaleTrack 通过追踪导致当前状态的动作历史记录如何提高模型的规划能力。我们在三个离线评估数据集上比较了回溯数据分析训练前后模型的表现。

我们在表 7 中展示了消融实验结果。在回溯数据训练后，模型在 AndroidControl-High 和 GUI-Odyssey 数据集上的准确率分别提高了 1.8% 和 0.7%。此外，在回溯训练后的动作类型识别准确率分别提高了 1.8%、0.9% 和 0.6%。结果证明了通过回溯训练获得的动作理解和预测能力。

5 结论

在这项工作中，我们介绍了 ScaleTrack 用于扩展和回溯自动 GUI 代理。我们发现两个普遍存在的问题包括孤立的数据合成标准和未考虑的回溯能力。为缓解这些问题，我们首先提出整合几种数据驱动的 GUI 元素增强方法以扩展 GUI 接地的训练过程，并将广泛的接地数据统一到一个固定的训练模板中。然后，我们提出了一种混合培训策略同时学习前向规划和回溯

表 7: ScaleTrack 的消融研究结果。

代理模型	安卓控制-低			Android 控制-高级			GUI 奇遇记		
	类型	接地	样品率	类型	接地	SR	类型	接地	样本率
ScaleTrack-7B	93.9	84.9	86.6	89.2	72.8	77.9	85.6	69.3	65.3
w/o back-tracking	92.1	85.6	86.6	88.3	73.7	76.1	85	69.4	64.6

能力。我们在几个基准数据集上进行了广泛的实验，如接地评估、离线和在线评估，结果证明了我们提出的 ScaleTrack 的有效性。

参考文献

Gpt-4v(ision) system card. 2023.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents. *arXiv preprint arXiv:2401.10935*, 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.

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.

Yu Gu, Kai Zhang, Yuting Ning, Boyuan Zheng, Boyu Gou, Tianci Xue, Cheng Chang, Sanjari Srivastava, Yanan Xie, Peng Qi, et al. Is your llm secretly a world model of the internet? model-based planning for web agents. *arXiv preprint arXiv:2411.06559*, 2024.

Siyuan Hu, Mingyu Ouyang, Difei Gao, and Mike Zheng Shou. The dawn of gui agent: A preliminary case study with claude 3.5 computer use. *arXiv preprint arXiv:2411.10323*, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. Language models can solve computer tasks. *Advances in Neural Information Processing Systems*, 36:39648–39677, 2023.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

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, 2024b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

Kaixin Ma, Hongming Zhang, Hongwei Wang, Xiaoman Pan, Wenhao Yu, and Dong Yu. Laser: Llm agent with state-space exploration for web navigation. *arXiv preprint arXiv:2309.08172*, 2023.

Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. Agent q: Advanced reasoning and learning for autonomous ai agents. *arXiv preprint arXiv:2408.07199*, 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.

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 Clinckemaillie, 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.

Yueqi Song, Frank Xu, Shuyan Zhou, and Graham Neubig. Beyond browsing: Api-based web agents. *arXiv preprint arXiv:2410.16464*, 2024.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024b.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, et al. Os-atlas: A foundation action model for generalist gui agents. *arXiv preprint arXiv:2410.23218*, 2024.

Yiheng Xu, Zekun Wang, Junli Wang, Dunjie Lu, Tianbao Xie, Amrita Saha, Doyen Sahoo, Tao Yu, and Caiming Xiong. Aguvis: Unified pure vision agents for autonomous gui interaction. *arXiv preprint arXiv:2412.04454*, 2024.

Yuhao Yang, Yue Wang, Dongxu Li, Ziyang Luo, Bei Chen, Chao Huang, and Junnan Li. Aria-ui: Visual grounding for gui instructions. *arXiv preprint arXiv:2412.16256*, 2024.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

Yao Zhang, Zijian Ma, Yunpu Ma, Zhen Han, Yu Wu, and Volker Tresp. Webpilot: A versatile and autonomous multi-agent system for web task execution with strategic exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 23378–23386, 2025.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. Gpt-4v (ision) is a generalist web agent, if grounded. *arXiv preprint arXiv:2401.01614*, 2024.

Longtao Zheng, Rundong Wang, Xinrun Wang, and Bo An. Synapse: Trajectory-as-exemplar prompting with memory for computer control. *arXiv preprint arXiv:2306.07863*, 2023.