



GraphPilot: GUI Task Automation with One-Step LLM Reasoning Powered by Knowledge Graph

Mingxian Yu¹, Siqu Luo¹, Xu Chen^{1,†}

¹School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China

[†]E-mail: chenxu35@mail.sysu.edu.cn

Received: November 2, 2025 / Revised: December 25, 2025 / Accepted: January 20, 2026 / Published online: February 4, 2026

Abstract: Mobile graphical user interface (GUI) agents are designed to automate everyday tasks on smartphones. Recent advances in large language models (LLMs) have significantly enhanced the capabilities of mobile GUI agents. However, most LLM-powered mobile GUI agents operate in stepwise query-act loops, which incur high latency due to repeated LLM queries. We present GraphPilot, a mobile GUI agent that leverages knowledge graphs of the target apps to complete user tasks in almost one LLM query. GraphPilot operates in two complementary phases to enable efficient and reliable LLM-powered GUI task automation. In the offline phase, it explores target apps, records and analyzes interaction history, and constructs an app-specific knowledge graph that encodes functions of pages and elements as well as transition rules for each app. In the online phase, given an app and a user task, it leverages the knowledge graph of the given app to guide the reasoning process of LLM. When the reasoning process encounters uncertainty, GraphPilot dynamically requests the HTML representation of the current interface to refine subsequent reasoning. Finally, a validator checks the generated sequence of actions against the transition rules in the knowledge graph, performing iterative corrections to ensure it is valid. The structured, informative information in the knowledge graph allows the LLM to plan the complete sequence of actions required to complete the user task. On the DroidTask benchmark, GraphPilot improves task completion rate over Mind2Web and AutoDroid, while substantially reducing latency and the number of LLM queries.

Keywords: Mobile GUI Agents; Mobile Task Automation; Large Language Models; App Analysis

<https://doi.org/10.64509/jicn.12.45>

1 Introduction

Mobile graphical user interface (GUI) agents have emerged as a promising paradigm for automating smartphone tasks through natural language instructions. Powered by recent advances in large language models (LLMs), these agents can understand user intents and interact with apps in a generalizable way. In contrast to traditional solutions that require predefined scripts for each app and scenario, LLM-powered mobile GUI agents leverage the understanding and reasoning abilities of LLMs to interact dynamically across different interfaces [1].

Despite these advances, most contemporary LLM-powered mobile GUI agents follow a stepwise interaction paradigm. In this approach, the agent iterates through query-act loops: at each interface, it queries the LLM with the current GUI (often a screenshot [2] or view hierarchy [3]) and the task description, receives a single action in return (such as

“click on button X”), performs that action, and then repeats for the next interface until task completion. A simple example of this process is shown in Figure 1. Although straightforward, this iterative process is highly inefficient, as repeated LLM queries at each interaction accumulate substantial latency. Such a stepwise design inherently trades computational efficiency for incremental reasoning, leading to considerable overhead in multi-step tasks.

Currently, GUI agents such as Mind2web [4] and AutoDroid [3] must issue multiple LLM queries per user task, leading to significant network latency. So we introduce GraphPilot, a mobile GUI agent that improves task completion rate while reducing latency by almost one LLM query with the help of knowledge graphs. First, GraphPilot constructs app-specific knowledge graphs that encode the functions of pages and elements, as well as the transition rules about element-to-page transitions for each app. Such a representation of a knowledge graph naturally captures the

[†] Corresponding author: Xu Chen

* Academic Editor: Mengwei Xu

© 2026 The authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

complex relationships between pages and elements, providing structured contextual knowledge for reasoning. Second, GraphPilot enables almost one LLM query to generate a complete sequence of actions for a user task by leveraging the previously constructed knowledge graph, thereby significantly reducing the need for repetitive query–act loops.

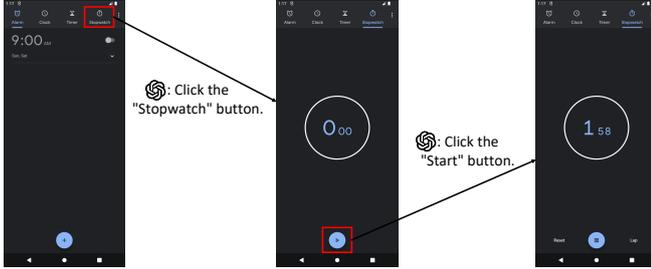


Figure 1: A simple example of the process of the stepwise GUI agents handling the task “start the stopwatch” in the Clock app.

GraphPilot operates in two phases. In the offline phase, it explores target apps, records exploration history, and extracts the functions of pages and elements, along with their transition rules, by querying the LLM. The extracted knowledge is then organized into app-specific knowledge graphs, which serve as a structured representation of the app interaction logic. In the online phase, GraphPilot leverages this knowledge graph to efficiently generate a complete sequence of actions for a given app and a user task with almost one LLM query. When uncertainty arises during reasoning, GraphPilot dynamically extracts the current HTML representation of the interface to guide subsequent reasoning. Finally, a validator checks the generated sequence of actions against the transition rules encoded in the knowledge graph, ensuring it is valid. By integrating a structured app-specific knowledge graph into LLM reasoning, GraphPilot bridges static understanding and dynamic task execution, achieving both higher efficiency and reliability in mobile GUI automation.

Essence: In the stepwise paradigm, each LLM query contains only limited contextual information. This requires multiple incremental queries to process user tasks step by step, which increases network latency. In contrast, GraphPilot incorporates rich contextual information into LLM queries, enabling the LLM to generate a more complete sequence of actions. This approach significantly reduces the number of LLM queries, effectively lowering network latency.

Experiments on DroidTask demonstrate that GraphPilot outperforms both Mind2Web and AutoDroid. It achieves the highest task completion rate of 74.1%, compared to 65.2% for Mind2Web and 62.0% for AutoDroid, while reducing the latency of LLM queries by 70.4% and 66.7%, respectively.

Our main contributions are summarized as follows:

1. **The construction of app-specific knowledge graphs.** GraphPilot constructs a knowledge graph for each app, encoding page and element functions, as well as transition rules. This structured representation enables a comprehensive understanding of the app’s interaction logic, providing accurate, informative knowledge for task execution.
2. **Almost one LLM query to generate a complete sequence of actions.** GraphPilot departs from the prevalent

stepwise query–act paradigm, which suffers from high latency due to repeated LLM queries. By leveraging the informative knowledge graph for each app, GraphPilot generates a complete sequence of actions in almost one LLM query, substantially reducing the latency of LLM queries while maintaining reasoning quality.

2 Related Work

Traditional mobile GUI task automation has predominantly relied on scripted approaches and Robotic Process Automation (RPA) tools such as Monkey [5] and Sikuli [6], which execute predefined sequences of GUI actions. However, these conventional methods suffer from brittleness, necessitating manual maintenance whenever interface layouts evolve [1]. In contrast, recent advances in large language model (LLM)-powered agents demonstrate the capability to interpret high-level natural language instructions and dynamically adapt to evolving application interfaces. This section surveys the key developments in LLM-powered mobile GUI automation.

2.1 Interface Representation

Effective GUI automation requires a comprehensive representation of interface elements and visual content. Contemporary approaches can be broadly categorized into two primary representation paradigms: text-based and multimodal (text and image) categories. Text-based methods leverage structured GUI data, including accessibility trees and Document Object Model (DOM)-like structures, as demonstrated in DroidBot-GPT [7] and AutoDroid [3]. Conversely, multimodal approaches integrate visual analysis through Vision-Language Models (VLMs): ScreenAI [8] represents a dedicated vision-language architecture for screenshot interpretation, while VisionDroid [9] and GUI Narrator [10] augment textual input with OCR-extracted text and icon annotations via Set-of-Mark (SoM) prompting techniques. Advanced multimodal agents such as Mobile-Agent-v2 [11] and OmniParser [12] incorporate sophisticated icon recognition and layout analysis capabilities. Our proposed GraphPilot adopts a text-based paradigm via an HTML-based interface representation, achieving high computational efficiency while preserving semantic richness and avoiding visual processing overhead.

2.2 Knowledge and Memory Representations

To handle diverse app functionality, recent agents augment LLMs with memory or knowledge structures. AutoDroid [3] builds a GUI Transition Graph (UTG) by exploring the app and injects this app-specific memory into prompts. MobileGPT [13] stores a hierarchical memory of tasks, subtasks, and actions for recall. AppAgent [2] uses autonomous exploration and human demonstrations to build a multimodal knowledge base for each app. RAG-based updating appears in AppAgentX [14] and AdaptAgent [15]. These approaches show the benefit of app-level context. While GraphPilot is novel in organizing app-specific knowledge into knowledge graphs that capture both the behaviors and structures of apps, enabling the reasoning process in nearly a single LLM query.

2.3 Task Planning and Action Execution

Most LLM-driven GUI agents operate in an iterative query-act cycle, where a single agent repeatedly observes the interface state and then determines the next action. For example, DroidBot-GPT [7] and AutoDroid [3] both use an LLM to choose the next UI element to interact with at each step, continuing this loop until the task is completed. This stepwise strategy offers adaptability to dynamic interfaces but incurs the overhead of multiple LLM queries. To improve efficiency, MobileGPT [13] introduces a hierarchical “explore-select-derive-recall” workflow and a multi-level memory (tracking tasks, subtasks, and actions) that significantly boosts automation accuracy and speed through better context management. Likewise, CogAgent [16] follows a similar iterative paradigm, leveraging a visual-language model to reason about the GUI for each incremental action.

To handle more complex tasks, some agents focus on making plans described in natural language executable on real interfaces or even employ multiple specialized agents. SeeAct [17] first drafts a tentative sequence of actions, then performs action grounding and element alignment to locate the correct target element and translate each step into a concrete click or input action. The emphasis is on this grounding and alignment procedure, which reliably turns high-level plans into runnable operations. Alternatively, multi-agent systems have been proposed to divide responsibilities among cooperative agents. For instance, MMAC-Copilot [18] coordinates multiple modalities through role-specific agents, and MobileExperts [19] uses a dynamic team of tool-driven agents to accomplish tasks jointly. Other works integrate specialized modules into the agent framework. For example, ClickAgent [20] combines an MLLM-based reasoner with a dedicated GUI locator model to better pinpoint interface elements, while Ponder & Press [21] adopts a divide-and-conquer design that splits the problem into high-level

instruction interpretation followed by low-level visual element localization. By partitioning the task in these ways, such approaches aim to reduce error propagation and improve success rate.

Despite these advances, the majority of LLM-driven agents still rely on incremental action generation, which accumulates latency over multiple iteration cycles. In contrast, our proposed GraphPilot agent generates a complete action sequence in almost a single interaction with the LLM by leveraging knowledge graphs, rather than looping through each step individually. This one-step planning capability eliminates the need for repeated LLM queries, significantly reducing network latency.

3 Design of GraphPilot

This section provides a detailed design of the two-phase architecture of GraphPilot and the functions of its components.

3.1 Overview

The overview of GraphPilot is shown in Figure 2. GraphPilot operates in two phases: an offline phase and an online phase.

The offline phase focuses on constructing knowledge graphs for each app. These knowledge graphs provide enough and accurate information for the online phase of GraphPilot. During this phase, GraphPilot first explores mobile apps to build a comprehensive understanding of their functions and structures. It maintains detailed *exploration history* ❶ about the sequence of actions and corresponding interface changes. Then GraphPilot analyzes the functionality of each page and its constituent elements using an LLM. Additionally, GraphPilot records transition information in the history, such as “interacting with element X on page A will lead to page B.” This *element-to-page* transition information provides directional guidance for the online phase of GraphPilot. Finally, all such information is consolidated into a *knowledge graph* ❷

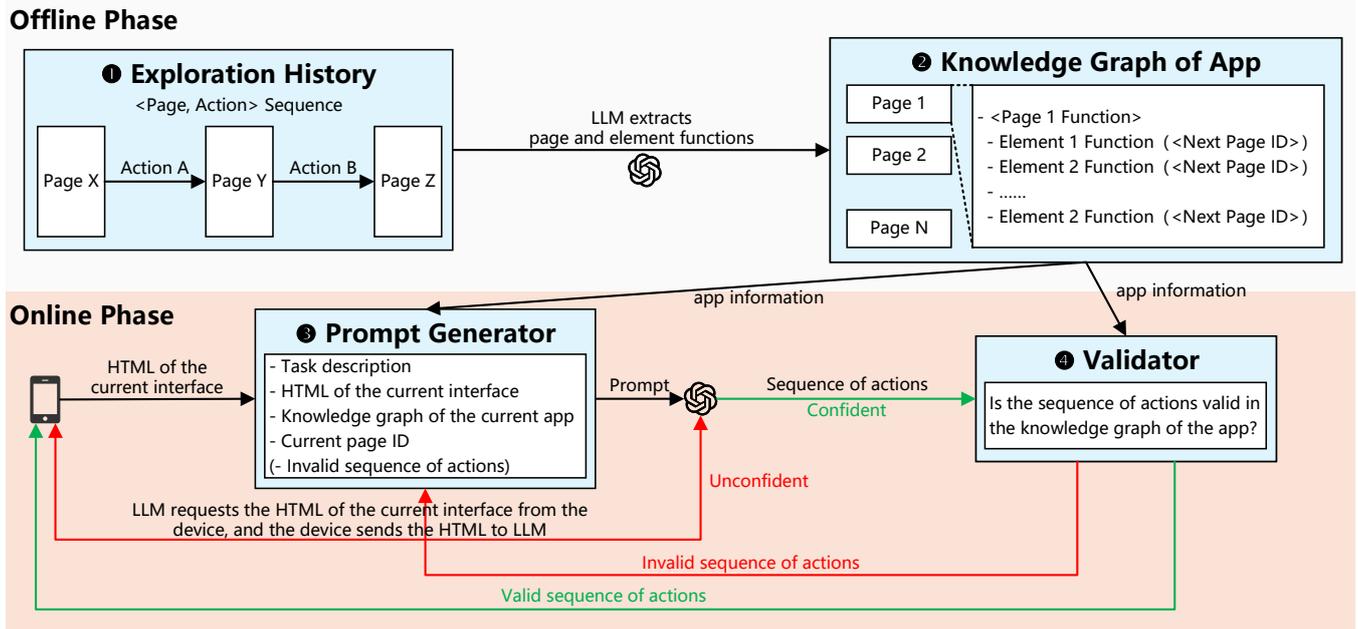


Figure 2: Overview of GraphPilot.

that captures both the behaviors (functions of pages and elements) and structures (element-to-page transition rules) of the apps. An example of the knowledge graph for an app is shown in Figure 3. The nodes in the knowledge graph represent the functions of pages and elements, and the edges represent the element-to-page transition rules. Finally, in the online phase, GraphPilot will use the knowledge graph as an informative context to support its task planning.

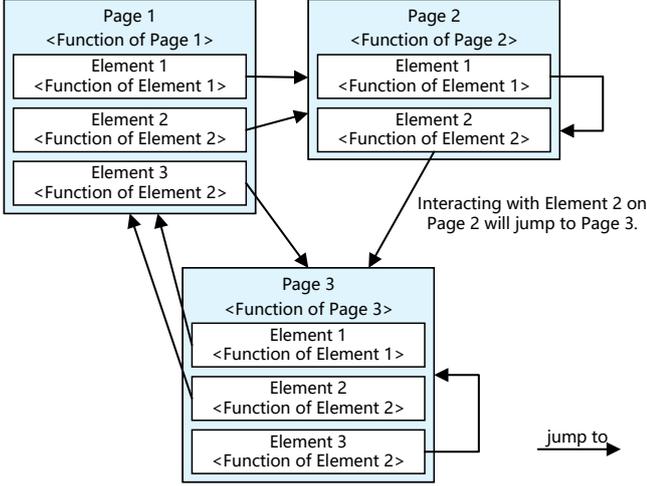


Figure 3: An example of the knowledge graph for an app.

The online phase leverages the knowledge graph constructed in the offline phase to perform task planning. Given a user-specified task, a *prompt generator* ③ generates a prompt that includes the task description and the knowledge graph (erroneous actions from the validator may also be included, which will be discussed later). The generated prompt is sent to the LLM to generate a candidate sequence of actions. This approach differs from traditional stepwise GUI agents that query the LLM at each interaction step. GraphPilot may encounter uncertainty during its reasoning. When this happens, it can dynamically request the HTML of the current interface. This provides additional context to proceed when the knowledge graph is insufficient. After GraphPilot generates a candidate sequence of actions, a *validator* ④ examines its correctness by checking the transition rules in the knowledge graph. If any rule violations occur, the validator detects the specific erroneous actions and sends them back to the prompt generator for resubmission, guaranteeing that only valid sequences are executed.

This two-phase architecture enables GraphPilot to achieve both high accuracy and low latency. The offline phase ensures comprehensive coverage of behaviors and structures of apps, while the online phase enables rapid task completion through almost one LLM query with prompt refinement when necessary.

3.2 Formulation of Offline Phase

The offline phase constructs a knowledge graph for each app by exploring the target app, analyzing the functions of its pages and elements, and extracting element-to-page transition rules from the recorded interaction history. We formally

define the knowledge graph construction process as shown in Algorithm 1.

Algorithm 1 Offline Knowledge Graph Construction for app A_i

Require: Exploration history $S_i = \{(H^{(t)}, s^{(t)}, H^{(t+1)})\}_{t=1}^T$
Ensure: Knowledge graph K_i

- 1: $K_i \leftarrow \emptyset$
- 2: **for** $t = 1$ to T **do**
- 3: $p_{\text{prev}} \leftarrow \text{PAGEID}(A_i, H^{(t-1)})$ (set to null if $t = 1$)
- 4: $p_{\text{curr}} \leftarrow \text{PAGEID}(A_i, H^{(t)})$
- 5: $p_{\text{next}} \leftarrow \text{PAGEID}(A_i, H^{(t+1)})$ (set to p_{curr} if $t = T$)
- 6: **if** $F_{p_{\text{curr}}}^{\text{page}}$ is undefined **then**
- 7: $F_{p_{\text{curr}}}^{\text{page}} \leftarrow \text{QUERYPAGEFUNCTION}(H^{(t-1)}, s^{(t-1)}, H^{(t)}, s^{(t)}, H^{(t+1)})$
- 8: $K_i \leftarrow K_i \cup \{(\text{"Page"}, p_{\text{curr}}, F_{p_{\text{curr}}}^{\text{page}})\}$
- 9: **end if**
- 10: $(a^{(t)}, e^{(t)}) \leftarrow s^{(t)}$ (set $e^{(t)} = 0$ if $t = T$)
- 11: **if** $e^{(t)} \neq 0$ and $F_{p_{\text{curr}}, e^{(t)}}^{\text{elem}}$ is undefined **then**
- 12: $F_{p_{\text{curr}}, e^{(t)}}^{\text{elem}} \leftarrow \text{QUERYELEMENTFUNCTION}(H^{(t)}, s^{(t)}, H^{(t+1)})$
- 13: $K_i \leftarrow K_i \cup \{(\text{"Element"}, p_{\text{curr}}, e^{(t)}, F_{p_{\text{curr}}, e^{(t)}}^{\text{elem}})\}$
- 14: $K_i \leftarrow K_i \cup \{(\text{"Transition"}, p_{\text{curr}}, e^{(t)}, p_{\text{next}})\}$
- 15: **end if**
- 16: **end for**
- 17: **return** K_i

App and Exploration History. Consider a target mobile app denoted as A_i , where i is the app index. During exploration, we record the exploration history $S_i = \{(H^{(t)}, s^{(t)}, H^{(t+1)})\}_{t=1}^T$, which consists of T sequential interaction steps. Each step is represented by a tuple containing: (1) $H^{(t)}$, the HTML representation of the current interface state at step t ; (2) $s^{(t)}$, the action performed at step t ; and (3) $H^{(t+1)}$, the resulting HTML representation after the action.

Actions and Elements. Each action $s^{(t)}$ is decomposed as $s^{(t)} = (a^{(t)}, e^{(t)})$, where $a^{(t)}$ represents the specific action type and $e^{(t)}$ represents the target element at step t . The action type $a^{(t)}$ is selected from a predefined action space $\{\text{click}, \text{text}, \text{stop}\}$. The element $e^{(t)}$ identifies element being interacted with on page $p^{(t)}$, while $e^{(t)} = 0$ indicates no element interaction (used for the **stop** action). For the terminal step, we enforce $a^{(T)} = \text{stop}$ and $e^{(T)} = 0$, indicating that the exploration sequence finishes without targeting any specific element. Function $\text{PAGEID}(\cdot)$ extracts the page ID based on the given HTML representation. When $t = 1$, p_{curr} denotes the initial page ID (line 4), and no preceding page p_{prev} exists. Accordingly, we set p_{prev} to null in this case (line 3).

Knowledge Graph Structure. The constructed knowledge graph for app A_i is denoted as K_i . For each page p_j , the knowledge graph stores three types of information: (1) $F_{p_j}^{\text{page}}$, a natural language description of the function of page p_j , which is obtained by function $\text{QUERYPAGEFUNCTION}(\cdot)$ (line 7); (2) $F_{p_j, e_k}^{\text{elem}}$, a natural language description of the functionality of element e_k on page p_j , which is obtained by function $\text{QUERYELEMENTFUNCTION}(\cdot)$ (line 12); and (3)

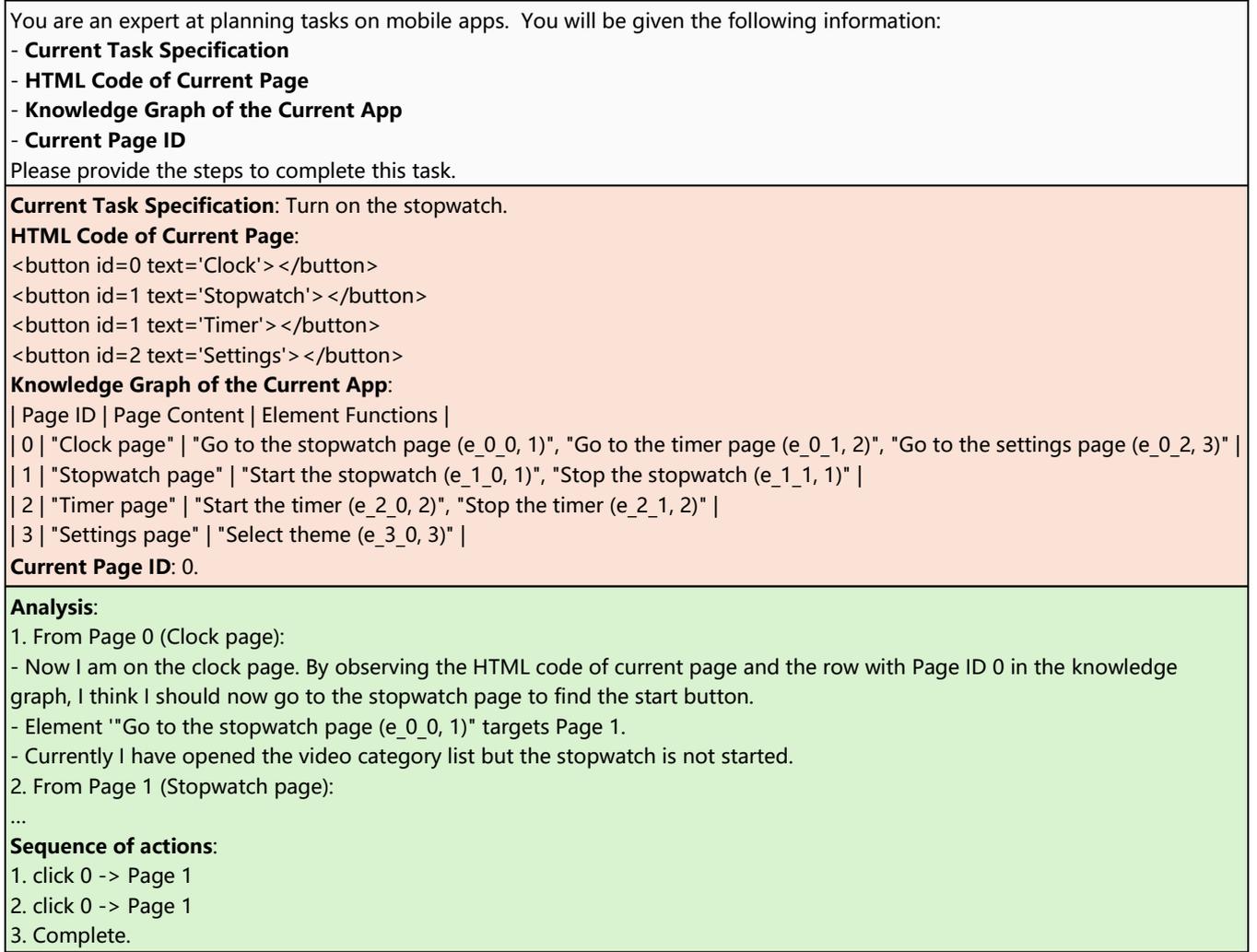


Figure 4: A simple example of the prompt generated by the prompt generator. The gray part represents a fixed background description. The red part represents dynamic content. The green part shows an example output generated by the LLM. (X, A) in the knowledge graph represents that interacting with element X will jump to page A.

the tuple starting with “Transition” in K_i represents the transition rules, and the following three elements represent the starting page, the element interacted with, and the arrival page, respectively (line 14).

3.3 Formulation of Online Phase

During the online phase, GraphPilot generates a sequence of actions utilizing the knowledge graph constructed in the offline phase. GraphPilot iteratively proposes a sequence of actions and refines it according to the transition rules through the validator. As illustrated in Figure 4, while generating the Analysis part, the LLM may reach an intermediate step where the knowledge graph no longer provides enough information to proceed. When this happens, the LLM is unconfident to continue the inference and therefore requests the HTML representation of the current interface to obtain the missing context. Conversely, if the LLM can complete the reasoning process in the Analysis part without encountering informational insufficiency, it is considered to be confident. This process is shown in Algorithm 2.

Algorithm 2 Online Action Sequence Generation for app A_i

Require: User task description d , knowledge graph K_i , current page p_{curr}

Ensure: A validated action sequence S

```
1:  $S \leftarrow \emptyset$ 
2:  $I \leftarrow \emptyset$ 
3: while  $S = \emptyset$  and maximum number of iterations is not
   reached do
4:    $prompt \leftarrow \text{PROMPTGENERATOR}(d, K_i, I, p_{curr})$ 
5:    $S \leftarrow \text{QUERYACTIONSEQUENCE}(prompt)$ 
6:    $I \leftarrow \text{VALIDATOR}(S, K_i)$ 
7:   if  $I \neq \emptyset$  then
8:      $S \leftarrow \emptyset$ 
9:   end if
10: end while
11: return  $S$ 
```

Given app A_i and user task description d , GraphPilot generates a sequence of actions $S = \{(p^{(t)}, a^{(t)}, e^{(t)}, p^{(t+1)})\}_{t=1}^T$, where each step t contains: (1) $p^{(t)}$, the page at step t ; (2)

$a^{(t)}$, the action type from $\{\text{click}, \text{text}, \text{stop}\}$; (3) $e^{(t)}$, the target element; and (4) $p^{(t+1)}$, the page at step $t+1$, which is the next page. We enforce $a^{(T)} = \text{stop}$ and $e^{(T)} = 0$ for the terminal step. Let p_{curr} denote the current page at the beginning of the action sequence generation, which provides the starting context (i.e., the HTML representation of the current interface and the current page ID) for task planning. The on-line action sequence generation process follows an iterative refinement approach. The `PROMPTGENERATOR(.)` function constructs a prompt from the task description d , knowledge graph K_i , invalid step indices I , and the current page p_{curr} .

Figure 4 illustrates a concrete example of the prompt generated by the prompt generator. The prompt consists of both a fixed background description and dynamic content based on the current scenario. The knowledge graph transitions are represented as (X, A) , indicating that interacting with element X will navigate to page A . The notation e_{i_j} within the brackets represents the element with id j on page p_i .

The function `QUERYACTIONSEQUENCE(.)` generates a candidate sequence of actions from the prompt, while the function `VALIDATOR(.)` examines the sequence against transition rules in K_i and returns invalid steps, if they exist. This process repeats until a valid sequence is produced or the maximum number of iterations is reached.

4 Implementation and Evaluation

4.1 Experimental Setups

We implement GraphPilot using GPT-4o [22] to evaluate its performance. We conduct comprehensive experiments on the DroidTask dataset [3], a benchmark specifically designed to assess the performance of end-to-end mobile task automation agents. DroidTask dataset comprises 158 high-level tasks extracted from 13 popular Android apps, providing a comprehensive evaluation suite that encompasses diverse mobile interaction scenarios and varying task complexities. We employ AutoDroid [3] to extract the HTML representations of mobile app interfaces. Mind2Web [4] and AutoDroid [3] are selected as our baseline methods.

4.2 Task Completion Rate

We evaluate task completion performance using the task completion rate (TCR) metric, which is the percentage of tasks successfully completed out of the total number of evaluated tasks. A task is considered successfully completed when its generated sequence of actions exactly matches the ground truth sequence provided in the DroidTask dataset. This metric directly reflects the accuracy and effectiveness of a GUI agent.

We evaluate TCR for Mind2Web, AutoDroid, and GraphPilot. As summarized in Table 1, GraphPilot achieves the highest overall TCR of 74.1% and consistently outperforms both baselines, leading Mind2Web by 8.9% and AutoDroid by 12.1%. The per-app breakdown in Figure 5 further shows that GraphPilot attains superior TCR on the majority of apps, indicating robust generalization across diverse tasks and

interfaces. While performance varies across apps due to differences in task structure and GUI complexity, the advantage of GraphPilot remains stable.

Table 1: TCR comparison across different GUI agents.

Agent	Mind2Web	AutoDroid	GraphPilot
TCR (%)	65.2	62.0	74.1

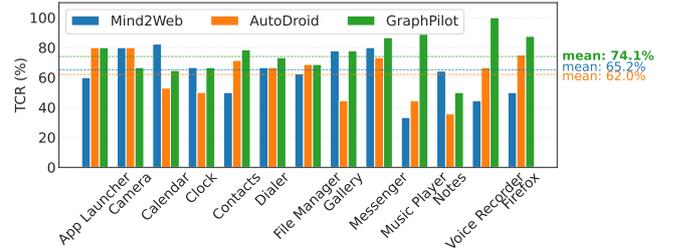


Figure 5: TCR by apps and average TCR across different GUI agents.

To analyze the relationship between task complexity and performance, we measure task complexity by the number of actions required to complete the task. Figure 6 presents the TCR across different numbers of actions, revealing a downward trend as the number of actions increases for all GUI agents. Notably, GraphPilot almost always maintains a higher TCR than Mind2Web and AutoDroid across virtually all complexity levels, demonstrating its robustness in handling tasks of varying difficulty.

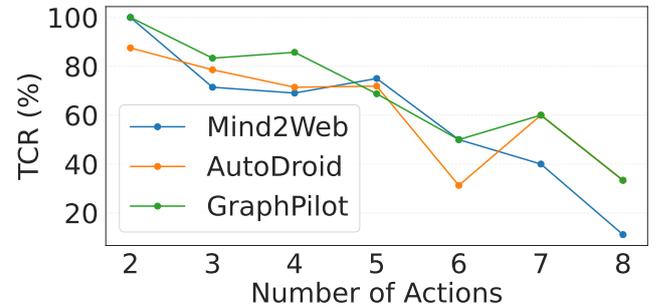


Figure 6: TCR in different numbers of actions required to complete the task across different GUI agents.

4.3 Latency

We measure both the latency of LLM queries and the number of queries required for each task, as these represent the primary time costs for mobile GUI agents. We compare the performance among Mind2Web, AutoDroid, and GraphPilot. The results are presented in Figure 7 and Table 2.

The results demonstrate that GraphPilot achieves significantly lower latency than Mind2Web and AutoDroid across all apps. In particular, GraphPilot reduces latency by 70.4% compared to Mind2Web and by 66.7% compared to AutoDroid. The reason is that GraphPilot requires substantially fewer LLM queries, representing a 77.3% reduction in query

frequency. This improvement stems from GraphPilot’s architectural design. This design typically requires only a single LLM query for each task to produce the complete sequence of actions. In contrast, agents like Mind2Web and AutoDroid must query the LLM for each action, which incurs additional latency.

Table 2: Comparison of the number of queries for each task across different GUI agents.

Agent	Number of Queries Each Task
Mind2Web	4.54
AutoDroid	4.54
GraphPilot	1.03

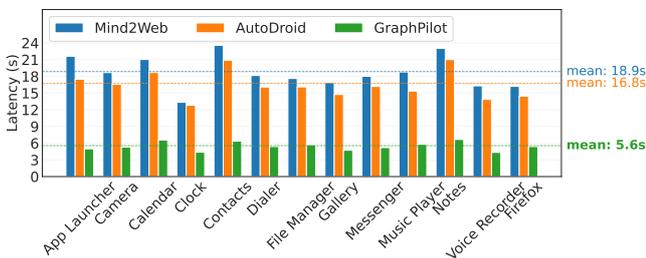


Figure 7: Latency comparison across different GUI agents.

4.4 Ablation Study

To analyze the contribution of different components in GraphPilot, we conduct ablation experiments by removing components and evaluating their impact on TCR. The TCR drops by removing transition rules, validator, and dynamic HTML representation requests, shown in Table 3.

Table 3: TCR of GraphPilot with different components removed.

	TCR (%)
w/o transition rules	60.8 (13.3 ↓)
w/o validator	72.8 (1.3 ↓)
w/o HTML request	71.5 (2.5 ↓)

4.4.1 Transition Rules

The removal of transition rules underscores their critical importance. The numbers in parentheses represent the performance drop compared to the complete GraphPilot. Without transition rules, GraphPilot experiences substantial performance degradation (13.3% drop). This significant decline highlights the key role of transition rules. They provide explicit navigation guidance during LLM reasoning, thereby preventing invalid interaction paths and ensuring more reliable task execution.

4.4.2 Validator

Removing the validator results in a slight performance degradation. While these drops are substantially smaller (1.3% drop) than those observed when removing transition rules,

they demonstrate that correcting invalid sequences generated by the LLM provides meaningful improvements to TCR. This suggests that, although the transition rules are clearly specified in the prompt, even an advanced LLM may occasionally produce action sequences that violate these rules. The validator serves as a quality assurance mechanism to ensure the reliability of execution.

4.4.3 Dynamic HTML Representation Request

Removing the dynamic HTML representation request causes a slight performance degradation (2.5% drop). Because the information in the knowledge graph is rich enough for the LLM to provide the complete sequence of actions directly in most cases, dynamic requests are rarely made. This component serves as a fallback mechanism when uncertainty arises during reasoning and provides additional context when the knowledge graph is insufficient to support decision-making.

Overall, the experimental results demonstrate that transition rules, which represent the edges in the knowledge graph, contribute most significantly to TCR improvement among all evaluated components. This importance stems from their fundamental role in constraining the action space and providing explicit navigation guidance during LLM reasoning.

5 Discussion

The knowledge graph can be naturally extended to multi-app collaborative scenarios. In such cases, nodes can represent functions of apps, pages, and elements, while edges can capture inter-app (element-to-page) and intra-app (element-to-app) transitions. This extension would enable GraphPilot to coordinate complex tasks that span multiple apps.

Currently, GraphPilot supports a limited action space consisting of click and text operations. Future work could expand this action space by supporting additional interaction types, such as drag, pinch, and long-press. This enhancement would enable GraphPilot to handle more complex mobile interactions and access a broader range of app functionalities.

Tasks can often be completed through multiple valid pathways. However, current evaluation methods require the output sequence to match the ground truth exactly, which may incorrectly classify valid alternative approaches as failures. Future research could explore more flexible evaluation approaches that assess task completion success rather than strict sequence matching, thereby allowing for recognition of alternative valid solution paths.

In the experiments, we assume that the information GraphPilot needs to perform tasks is available in the knowledge graph. GraphPilot currently relies on the integrity of the knowledge graph. If the knowledge graph lacks corresponding information, such as transition rules, the TCR of GraphPilot will be affected. Making GraphPilot more robust is our subsequent goal.

6 Conclusion

We present GraphPilot, a mobile GUI automation agent that addresses the accuracy and latency limitations in conventional stepwise approaches by constructing a knowledge graph for each app. Evaluation on DroidTask demonstrates

that GraphPilot outperforms Mind2Web and AutoDroid with improvements in TCR while dramatically reducing query latency and LLM queries. We believe that incorporating domain knowledge of apps into LLM prompts in an appropriate form will allow LLMs to handle GUI task automation more accurately and efficiently. This will enable the creation of practical personal AI assistants.

Funding

This work was supported in part by Guangdong S&T Programme (Grant No. 2024B0101040007); in part by the Guangdong Basic and Applied Basic Research Foundation (No. 2023B1515120058); and the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (No. 2017ZT07X355).

Author Contributions

Conceptualization, Mingxian YU and Xu CHEN; methodology, Mingxian YU; software, Mingxian YU; validation, Mingxian YU; formal analysis, Mingxian YU and Xu CHEN; investigation, Mingxian YU; resources, Xu CHEN; data curation, Mingxian YU; writing—original draft preparation, Mingxian YU; writing—review and editing, Siqi LUO; visualization, Mingxian YU and Siqi LUO; supervision, Siqi LUO and Xu CHEN; project administration, Xu CHEN; funding acquisition, Xu CHEN. All authors have read and agreed to the published version of the manuscript.

Conflict of Interest

All the authors declare that they have no conflict of interest.

Data Available

The data and materials used in this study are available upon request from the corresponding author.

References

- [1] Liu, G., Zhao, P., Liang, Y., Liu, L., Guo, Y., Xiao, H., Lin, W., Chai, Y., Han, Y., Ren, S., *et al.*: LLM-Powered GUI Agents in Phone Automation: Surveying Progress and Prospects. *Transactions on Machine Learning Research*, 1–75 (2025)
- [2] Zhang, C., Yang, Z., Liu, J., Li, Y., Han, Y., Chen, X., Huang, Z., Fu, B., Yu, G.: AppAgent: Multimodal Agents as Smartphone Users. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–20 (2025). <https://doi.org/10.1145/3706598.3713600>
- [3] Wen, H., Li, Y., Liu, G., Zhao, S., Yu, T., Li, T.J.-J., Jiang, S., Liu, Y., Zhang, Y., Liu, Y.: AutoDroid: LLM-powered Task Automation in Android. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pp. 543–557 (2024). <https://doi.org/10.1145/3636534.3649379>
- [4] Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., Sun, H., Su, Y.: Mind2Web: Towards a generalist agent for the web. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 28091–28114 (2023)
- [5] UI/Application Exerciser Monkey. <https://developer.android.com/studio/test/other-testing-tools/monkey> Access September 12, 2025
- [6] SikuliX by RaiMan. <http://www.sikulix.com> Access September 12, 2025
- [7] Wen, H., Wang, H., Liu, J., Li, Y.: DroidBot-GPT: GPT-powered UI automation for Android. *arXiv preprint arXiv:2304.07061* (2023). <https://doi.org/10.48550/arXiv.2304.07061>
- [8] Baechler, G., Sunkara, S., Wang, M., Zubach, F., Mansoor, H., Etter, V., Cărbune, V., Lin, J., Chen, J., Sharma, A.: ScreenAI: A vision-language model for UI and infographics understanding. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 3045–3053 (2024). <https://doi.org/10.24963/ijcai.2024/339>
- [9] Liu, Z., Li, C., Chen, C., Wang, J., Wu, B., Wang, Y., Hu, J., Wang, Q.: Vision-driven automated mobile GUI testing via multimodal large language model. *arXiv preprint arXiv:2407.03037* (2024). <https://doi.org/10.48550/arXiv.2407.03037>
- [10] Wu, Q., Gao, D., Lin, K.Q., Wu, Z., Guo, X., Li, P., Zhang, W., Wang, H., Shou, M.Z.: GUI action narrator: Where and when did that action take place? *arXiv preprint arXiv:2406.13719* (2024). <https://doi.org/10.48550/arXiv.2406.13719>
- [11] Huang, F., Jia, H., Sang, J., Shen, W., Wang, J., Xu, H., Yan, M., Zhang, J., Zhang, X.: Mobile-Agent-v2: Mobile Device Operation Assistant with Effective Navigation via Multi-Agent Collaboration. In *Advances in Neural Information Processing Systems 37*, pp. 2686–2710 (2024). <https://doi.org/10.52202/079017-0088>
- [12] Wan, J., Song, S., Yu, W., Liu, Y., Cheng, W., Huang, F., Bai, X., Yao, C., Yang, Z.: OmniParser: A Unified Framework for Text Spotting, Key Information Extraction and Table Recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15641–15653 (2024). <https://doi.org/10.1109/cvpr52733.2024.01481>
- [13] Lee, S., Choi, J., Lee, J., Wasi, M.H., Choi, H., Ko, S., Oh, S., Shin, I.: MobileGPT: Augmenting LLM with Human-like App Memory for Mobile Task Automation. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pp. 1119–1133 (2024). <https://doi.org/10.1145/3636534.3690682>
- [14] Jiang, W., Zhuang, Y., Song, C., Yang, X.,

- Zhou, J.T., Zhang, C.: AppAgentX: Evolving GUI agents as proficient smartphone users. arXiv preprint arXiv:2503.02268 (2025). <https://doi.org/10.48550/arXiv.2503.02268>
- [15] Verma, G., Kaur, R., Srishankar, N., Zeng, Z., Balch, T., Veloso, M.: AdaptAgent: Adapting Multimodal Web Agents with Few-Shot Learning from Human Demonstrations. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 20635–20651 (2025). <https://doi.org/10.18653/v1/2025.acl-long.1008>
- [16] Hong, W., Wang, W., Lv, Q., Xu, J., Yu, W., Ji, J., Wang, Y., Wang, Z., Dong, Y., Ding, M., *et al.*: CogAgent: A Visual Language Model for GUI Agents. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14281–14290 (2024). <https://doi.org/10.1109/cvpr52733.2024.01354>
- [17] Zheng, B., Gou, B., Kil, J., Sun, H., Su, Y.: GPT-4V(ision) is a generalist web agent, if grounded. In Proceedings of the 41st International Conference on Machine Learning, pp. 61349–61385 (2024)
- [18] Song, Z., Li, Y., Fang, M., Li, Y., Chen, Z., Shi, Z., Huang, Y., Chen, X., Chen, L.: MMAC-Copilot: Multimodal Agent Collaboration Operating Copilot. arXiv preprint arXiv:2404.18074 (2024). <https://doi.org/10.48550/arXiv.2404.18074>
- [19] Zhang, J., Zhao, C., Zhao, Y., Yu, Z., He, M., Fan, J.: MobileExperts: A dynamic tool-enabled agent team in mobile devices. arXiv preprint arXiv:2407.03913 (2024). <https://doi.org/10.48550/arXiv.2407.03913>
- [20] Hoscilowicz, J., Janicki, A.: ClickAgent: Enhancing UI Location Capabilities of Autonomous Agents. In Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 471–476 (2025)
- [21] Wang, Y., Zhang, H., Tian, J., Tang, Y.: Ponder & Press: Advancing Visual GUI Agent towards General Computer Control. In Findings of the Association for Computational Linguistics: ACL 2025, pp. 1461–1473 (2025). <https://doi.org/10.18653/v1/2025.findings-acl.76>
- [22] Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., *et al.*: GPT-4o system card. arXiv preprint arXiv:2410.21276 (2024). <https://doi.org/10.48550/arXiv.2410.21276>