Agent+P: Guiding UI Agents via Symbolic Planning

Anonymous ACL submission

Abstract

Large Language Model (LLM)-based UI agents show great promise for UI automation but often hallucinate in long-horizon tasks due to their lack of understanding of the global UI transition structure. To address this, we introduce AGENT+P, a novel framework that leverages symbolic planning to guide LLM-based UI agents. Specifically, we model an app's UI transition structure as a UI Transition Graph (UTG), which allows us to reformulate the UI automation task as a pathfinding problem on the UTG. This further enables an off-the-shelf symbolic planner to generate a provably correct and optimal high-level plan, preventing the agent from redundant exploration and guiding the agent to achieve the automation goals. AGENT+P is designed as a plug-and-play framework to enhance existing UI agents. Evaluation on the AndroidWorld benchmark demonstrates that AGENT+P improves the success rates of state-of-the-art UI agents by up to 14% and reduces the action steps by 37.7%. Our code is available at: https://anonymous.4open.science/r/agentp-F7AF.

1 Introduction

With mobile applications (apps) woven into all parts of our daily life, it is critically important to ensure the high quality of apps. User interface (UI) automation, the process of programmatically executing sequences of UI interactions, has become an essential method for improving app quality by enabling automated testing for bug and vulnerability detection (Lai and Rubin, 2019; Ma et al., 2025) and supporting user task automation (Orrù et al., 2023; Rawles et al., 2024; Li et al., 2025b; Zhang et al., 2023).

While recent advances in UI automation, notably the integration of LLM-based UI agents (Liu et al., 2025; Zhao et al., 2024; Ran et al., 2024) that explicitly model the available actions in each UI screen, have demonstrated encouraging results, the ever-growing complexity of modern UIs continues to hinder effective and efficient automation. In particular, existing approaches struggle in **long-horizon planning** tasks that require navigating via multiple UIs since such multi-step planning often lead to increased hallucination rates (Liu et al., 2023; Wei et al., 2025; Xie et al., 2025; Wu et al.). For example, the LLM-based agents employed by these approaches typically follow a depth-first strategy to find valid action sequences, making decisions based on local UI states without understanding the global transition structure, where different UIs can lead to distinct subsequent actions. Consequently, they often fail to derive valid sequences that accomplish task goals and repeatedly waste effort on actions that diverge from those goals.

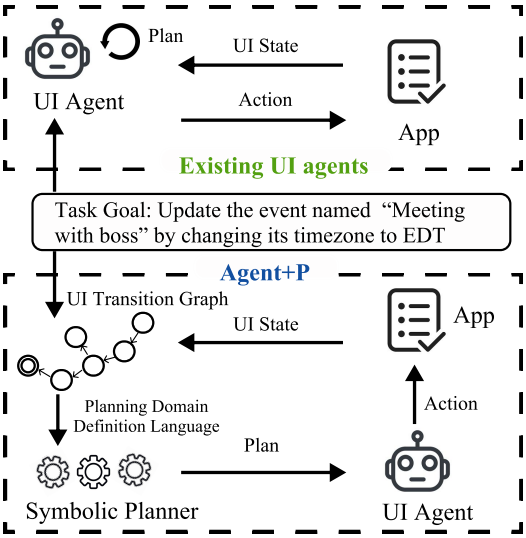


Figure 1: AGENT+P compared with existing UI agents. AGENT+P improves performance by constructing a UI Transition Graph via program analysis and leveraging an external symbolic planner to generate a high-level, globally aware transition plan, thereby guide the agent towards the automation goal.

Key Insights. To address this fundamental limitation, *we introduce an external planner module that*

provides LLM-based UI agents with global transition knowledge extracted through program analysis. This module leverages established planning algorithms to prevent redundant exploration and guide the agent toward diverse action sequences that are more likely to achieve the automation goals. Specifically, we model an app's global transition structure via a UI Transition Graph (UTG) (Sun et al., 2025; Wen et al., 2024), with nodes representing UIs and edges representing user-triggered UI transitions. A UI automation problem, navigating from a start UI to a target UI, can thus be formulated as a *pathfinding problem* on the UTG, where the objective is to find an optimal path from the start node to the target node. This formulation enables the use of off-the-shelf symbolic planners to derive provably correct and optimal plans, thereby fundamentally eliminating high-level planning hallucinations and enhancing the reliability of LLM-based UI agents.

Our Method. Building upon these insights, in this paper, we introduce AGENT+P, an agentic framework that leverages symbolic planning to guide UI agents. Given a natural language UI automation goal and the app for automation, AGENT+P operates iteratively in four stages, each executed by an LLM-based module, until the goal is achieved: the **UTG Builder** constructs a static UTG of the app and dynamically updates it during automation. The **Node Selector** maps the natural language goal to a targeted node in the UTG. The **Plan Generator** translates the UTG into a problem problem using Planning Domain Definition Language (PDDL) (Aeronautiques et al., 1998), which is then solved by an external symbolic planner. The resulting symbolic plan is subsequently converted into natural language instructions. Finally, the **GUI Explorer** interacts with the app to execute the translated instructions to navigate to the goal.

AGENT+P is designed as a plug-and-play planning framework that can be incorporated with and enhance existing UI agents. We evaluate AGENT+P by integrating it with four state-of-the-art agents on the AndroidWorld benchmark. Our results demonstrate that by leveraging symbolic planning on the UTG, AGENT+P increases the success rates of baseline agents by up to 14.0% and reduces the action steps by 37.7%.

Our primary contributions are as follows:

- We propose AGENT+P, a novel framework that leverages symbolic planning to provide LLM-based UI agents with global transition information derived from program analysis, mitigating the long-horizon planning failures commonly faced by these agents.
- We present a novel formulation that maps the problem of UI automation into a pathfinding problem in the UTG, making it solvable with provably correct symbolic planners.
- We conduct extensive evaluation on the AndroidWorld benchmark, demonstrating that AGENT+P substantially improves the success rate and efficiency of three state-of-the-art UI agents.

2 Motivation

2.1 UI and UI Transition Graph

To motivate our method, we begin by introducing the concepts of widget, UI, and the modeling of UI transitions (i.e., UTG).

Definition 1 (Widget, Action)**.** A *widget,* denoted as w, is a basic interactive element on a UI screen. An *action,* denoted as a, is a 2-tuple $a = (w, e)$, where w is a widget, e is the user event (e.g., `click`, `input`).

Following existing UI agents that represent a UI as a sequence of widgets and supported actions (Android World, 2025; Ye et al., 2025; Dai et al., 2025; Li et al., 2025b), we define the UI state (shortened as UI) as follows:

Definition 2 (UI)**.** A *UI,* denoted as u, is an n-tuple of all unique actions available on the screen, $u = (a_1, a_2, \ldots, a_n)$, where n is the total number of available actions.

This definition is sufficient to represent UI transitions while avoiding state explosion (Valmari, 1996).

Definition 3 (UI Transition Graph)**.** *A UTG for an app is a directed graph $G = (\mathscr{U}, \mathscr{T}, \varepsilon)$ that models the transition structure of the app.*

- $\mathscr{U}$ is a finite set of nodes, where each node $u \in \mathscr{U}$ represents a UI u in the app.
- $\mathscr{T} \subseteq \mathscr{U} \times \mathscr{U}$ is a set of directed edges. An edge $(u_i, u_j) \in \mathscr{T}$ represents a transition from UI u_i to UI u_j.
- $\varepsilon : \mathscr{T} \to \mathscr{A}$ is an edge-labeling function. It maps each transition (u_i, u_j) to the action $a = (w, e)$ that triggers it, where the widget w is an element of the source UI u_i.

Figure 3 shows an example of UTG of an Android app named Simple Calendar Pro in AndroidWorld benchmark.

Table 1: Average number of UTG nodes and edges for apps where existing UI agents succeeded vs. failed.

	Nodes		Edges	
Agent	**Success**	**Failed**	**Success**	**Failed**
DroidRun	27.0	71.2	62.7	168.0
LX-GUIAgent	29.3	53.8	65.9	129.2
AutoGLM	39.4	42.0	90.6	100.3
Finalrun	23.8	55.0	59.5	125.6
UI-Venus	18.0	50.7	66.0	108.0

2.2 Motivational Study

To investigate how UI complexity affects UI agent performance, we conduct a motivational study using the AndroidWorld benchmark (Rawles et al., 2024). This benchmark has 116 programmatic tasks across 20 Android apps, where an agent is given a natural language instruction and must navigate the app to achieve the goal. Performance is evaluated using a task-level "success rate". It is worth noting that published results typically report only an overall average for all tasks. To obtain per-app success rates, we download and analyze the trajectories from five top-performed UI agents. Specifically, we classify an app as "successful" for a given agent if its performance on that app exceeds its overall average, and as a "failed" app otherwise.

Table 1 compares the UI complexity measured by the number of UTG nodes and edges between successful and failed apps. Our Wilcoxon Signed-Rank test (Wilcoxon, 1945) shows that failed apps have significantly more UTG nodes and edges than successful ones ($p = 0.03$). This finding indicates that existing UI agents struggle with navigating apps with complex UIs, which motivates our work to leverage an app's transition structure to enhance agent performance.

3 Problem Formulation

In this section, we first define the problem of targeted UI automation, and how to convert it into a equivalent classical planning problem.

3.1 Problem Definition

With the structure of the app modeled as a UTG, we can now formally define the task of UI automation.

Definition 4 (UI Automation)**.** *UI automation is the process of programmatically executing a sequence of actions $\pi = \langle a_1, a_2, \ldots, a_N \rangle$, where each action $a_i \in \mathscr{A}$.*

In this work, we focus on a specific, goal-oriented variant of this task.

Listing 1: Domain PDDL for targeted UI automation.

```
(define (domain utg-automation)
  (:requirements :strips :typing)

  (:types
    node - object
  )

  (:predicates
    (at ?n - node)
    (connected ?from - node ?to - node)
    (visited ?n - node)
    (goal-node ?n - node)
    (goal-achieved ?n - node)
  )

  (:action navigate
    :parameters (?from - node ?to - node)
    :precondition (and
      (at ?from)
      (connected ?from ?to)
    )
    :effect (and
      (not (at ?from))
      (at ?to)
      (visited ?to)
      (when (goal-node ?to) (goal-achieved ?to))
    )
  )
)
```

Definition 5 (Targeted UI Automation)**.** *Given an app with an initial UI u_{init} and a target UI u_{target}, the objective is to find a valid sequence of actions $\pi = \langle a_1, a_2, \ldots, a_N \rangle$, where $a_i = (w_i, e_i)$, that navigates the app from u_{init} to u_{target}.*

We formulate targeted UI automation as a pathfinding problem on the UTG. Let $\kappa : \mathscr{A} \to \mathbb{R}^+$ be a cost function that assigns a positive cost to each action, representing computational resources, execution time, or other relevant metrics. The problem is then to find a path from u_{init} to u_{target} that minimizes the total execution cost:

$$\pi^* = \arg\min_{\pi} \sum_{i=1}^{N} \kappa(a_i).$$

To simplify the formulation, in this work, we adopt a uniform action cost. This reduces the cost-minimization task to the classical shortest path problem, where the objective is to find the path from u_{init} to u_{target} with the fewest actions.

3.2 Symbolic and Classical Planning

Symbolic planning is a long-standing area of AI concerned with finding a sequence of actions to achieve a predefined goal. The most fundamental and widely studied form is **classical planning** where a planning problem instance, P, is formally described as a tuple $P = \langle \mathscr{D}, s_{init}, \mathscr{G} \rangle$, where $\mathscr{D} = \langle \mathscr{F}, \mathscr{A} \rangle$ is the planning domain. These components are defined as follows:

- **States**: $\mathscr{F}$ is a set of fluents or predicates that describe the properties of the world. A state s is a

Table 2: Mapping of UI automation notation to classical planning equivalent.

UI Automation Notation	Classical Planning Equivalent
A UI state (UI) $u \in \mathscr{U}$	A state $s \in \mathscr{S}$ where the predicate $at(u)$ is true
The set of all UIs $\mathscr{U}$	The state space $\mathscr{S}$
The initial UI u_{init}	The initial state s_{init}, defined by $at(u_{init})$
The target UI u_{target}	The goal $\mathscr{G}$, specified by the condition $at(u_{target})$
A UI transition via the edge (u_i, u_j) with label a	A planning action
Precondition: the app is on UI u_i	$pre(a) : at(u_i)$
Effect: the app moves to UI u_j	$eff(a) : \neg at(u_i) \wedge at(u_j)$
A path from u_{init} to u_{target} via the sequence of action π	A plan π

complete assignment of truth values to all fluents in $\mathscr{F}$. The set of all possible states is the state space $\mathscr{S}$.

- **Initial States**: $s_{init} \in \mathscr{S}$ is the initial state of the world.
- **Goal**: $\mathscr{G}$ is the goal specification, a set of conditions on states. Any state $s \in \mathscr{S}$ that satisfies all conditions in $\mathscr{G}$ is a goal state.
- **Actions**: $\mathscr{A}$ is a set of actions. Each action $a \in \mathscr{A}$ is defined by its preconditions, $\text{pre}(a)$, and its effects, $\text{eff}(a)$. An action can only be executed in a state where its preconditions are met, and its effects describe how the state changes after its execution.

A solution, or plan, π, is a sequence of actions $\langle a_1, a_2, \ldots, a_N \rangle$ that transforms the initial state s_{init} into a goal state. This is achieved by applying the actions sequentially, where each action a_i is applicable in the state resulting from the execution of a_{i-1}, and the final state after executing a_N satisfies $\mathscr{G}$.

The Planning Domain Definition Language (PDDL) (Aeronautiques et al., 1998) is the standard language for representing such planning problems, typically using two files: a domain file defining $\mathscr{F}$ and $\mathscr{A}$, and a problem file defining s_{init} and $\mathscr{G}$.

3.3 UI Automation to Classical Planning

The pathfinding problem can be naturally cast into a classical planning problem, allowing us to leverage classical planners to compute the solution, i.e., the shortest path. Table 2 illustrates the mapping from the UI automation domain to the classical planning domain, which is elaborated as follows:

- **States**: A planning state s corresponds to the app being at a specific UI u. We can define a predicate $at(u)$ which is true if the app is currently on UI $u \in \mathscr{U}$. The state space $\mathscr{S}$ is the set of all possible UIs, $\mathscr{U}$.
- **Initial State**: The initial state s_{init} is defined by the predicate $at(u_{init})$ being true.
- **Goal**: The goal $\mathscr{G}$ is specified by the condition that the predicate $at(u_{target})$ must be true.
- **Actions**: For each UI transition (u_i, u_j) in the UTG triggered by an action $a = (w, e)$, we define a planning operator with precondition $at(u_i)$ and effects $\neg at(u_i)$ and $at(u_j)$.

Following this formulation, we define a general PDDL domain file template applicable to any targeted UI automation task, as shown in Listing 1. Any specific user task, represented by UTG with a specified target UI, u_{target}, can be translated into a corresponding PDDL problem file via this template. For instance, to solve the "Change the time zone" task in the Simple Calendar Pro app, the UTG from Figure 3 is converted into the problem file presented in Listing 2. This symbolic representation allows us to employ a classical planner to efficiently compute a valid and optimal sequence of actions to navigate the app from the initial UI, u_{init}, to the target, u_{target}.

4 AGENT+P

Algorithm 1 shows the workflow of AGENT+P. Given a natural language goal specified by the user, AGENT+P operates through four primary modules that interact in a continuous loop until the goal is achieved: the **UTG Builder**, the **Node Selector**, the **Plan Generator**, and the **UI Explorer**. In the following subsections, we elaborate on the design rationale and functionality of each module.

4.1 UTG Builder

Existing methods for constructing a UTG rely on either dynamic analysis (Wen et al., 2024; Sun et al., 2025), which is accurate but often suffers from high cost and incomplete coverage, or static analysis (Azim and Neamtiu, 2013; Yang et al., 2018), which is more comprehensive but can introduce infeasible transitions (Liu et al., 2022).

To overcome these limitations, AGENT+P utilizes a hybrid approach that synergizes both techniques to build UTG. AGENT+P begins by performing static analysis to construct an initial UTG (Line 1 of Algorithm 1), following established practices that track API calls responsible for UI transitions in the app's source code (Yang et al., 2018; Liu et al., 2022).

This initial UTG is then dynamically verified and refined as the UI Explorer explores the app (Line 10 and Line 16). Specifically, let the current UTG be $G = (\mathscr{U}, \mathscr{T}, \varepsilon)$. During UI automation, an observed transition (u_i, a_{obs}, u_j), where $u_i \in \mathscr{U}$, updates the graph to $G' = (\mathscr{U}', \mathscr{T}', \varepsilon')$ in one of three ways:

- **Update Edge:** An action leads from an existing source node to a target node, but the recorded action does not match the corresponding one in the UTG. Formally, if $(u_i, u_j) \in \mathscr{T}$ and $\varepsilon((u_i, u_j)) \neq a_{obs}$, the labeling function is updated such that $\varepsilon'((u_i, u_j)) = a_{obs}$, while $\mathscr{U}' = \mathscr{U}$ and $\mathscr{T}' = \mathscr{T}$.
- **Add Edge:** An action connects two existing UI nodes, but no corresponding edge exists in the UTG. This occurs when $u_j \in \mathscr{U}$ and $(u_i, u_j) \notin \mathscr{T}$. A new transition is added by setting $\mathscr{T}' = \mathscr{T} \cup \{(u_i, u_j)\}$ and extending ε' with $\varepsilon'((u_i, u_j)) = a_{obs}$.
- **Add Node:** An action leads to a UI that is not yet in the UTG. If $u_j \notin \mathscr{U}$, a new node and edge are added to the graph: $\mathscr{U}' = \mathscr{U} \cup \{u_j\}$, $\mathscr{T}' = \mathscr{T} \cup \{(u_i, u_j)\}$, and ε' is extended with $\varepsilon'((u_i, u_j)) = a_{obs}$.

This approach allows AGENT+P to maintain a UTG that is both comprehensive and dynamically accurate, combining the breadth of static analysis with the precision of real-time exploration.

4.2 Node Selector

A key challenge in AGENT+P is mapping an unstructured, natural language user goal (e.g., "Create a playlist") to a specific, concrete UI state within the app, i.e., u_{target} (Line 5 of Algorithm 1). Based on existing techniques that represent UI data from source code, screenshots, or text and convert it into structured formats such as text (Baechler et al., 2024), binary vectors (Li et al., 2021), or graphs (Li et al., 2025a), in AGENT+P, we implement two complementary strategies for this mapping process:

1. **Embedding-based Matching:** We compute semantic embeddings for the user's input and for the textual representations of all UI nodes in the UTG. The node with the highest cosine similarity to the input query is selected as the target.
2. **MLLM-based Identification:** We query a Multimodal Large Language Model (MLLM) with the user's goal and representations of candidate UIs. The MLLM is prompted to directly identify which UI screen best corresponds to the user's objective. The prompt template used for this process is detailed in Listing 3.

Algorithm 1: Workflow of AGENT+P

Input: Natural language goal g_{nl}, App $\mathscr{A}$, Max running steps *maxStep*
Output: Automation outcome: *Success* or *Failure*
Aliases: ***UB*** $\leftarrow$ *UtgBuilder;* ***NS*** $\leftarrow$ *NodeSelector;* ***PG*** $\leftarrow$ *PlanGenerator;* ***UE*** $\leftarrow$ *UiExplorer*

1 $G \leftarrow$ UB.buildStaticUTG($\mathscr{A}$);
2 $u_{cur} \leftarrow$ getCurrentUI($\mathscr{A}$);
3 $steps \leftarrow 0$;
4 **while** $steps \leq maxStep$ **do**
5 $\quad u_{target} \leftarrow$ NS.selectTargetNode(g_{nl}, G);
6 $\quad Plan \leftarrow$ PG.generatePlan(u_{cur}, u_{target}, G);
7 $\quad$ **if** *Plan is valid* **then**
8 $\quad\quad$ **for** *each action* a_i *in Plan* **do**
9 $\quad\quad\quad u_{next} \leftarrow$ UE.act(a_i);
10 $\quad\quad\quad G \leftarrow$ UB.update($G, u_{cur}, a_i, u_{next}$);
11 $\quad\quad\quad u_{cur} \leftarrow u_{next}$;
12 $\quad$ **else**
13 $\quad\quad neighbors \leftarrow$ PG.getNeighbors(u_{cur}, G);
14 $\quad\quad a \leftarrow$ UE.decideAction($neighbors, u_{cur}$);
15 $\quad\quad u_{next} \leftarrow$ UE.act(a);
16 $\quad\quad G \leftarrow$ UB.update($G, (u_{cur}, a, u_{next})$);
17 $\quad\quad u_{cur} \leftarrow u_{next}$;
18 $\quad$ **if** *UE.evaluate*(g_{nl})) **then**
19 $\quad\quad$ **return** *Success*;
20 $\quad steps \leftarrow steps + 1$;
21 **return** *Failure*;

4.3 Plan Generator

Once the target node u_{target} is identified, the task becomes a classical planning problem: finding the shortest sequence of actions from the current UI state to the target state. The Plan Generator orchestrates this by first converting the UTG into the PDDL format. It then invokes a classical planner to solve for an optimal PDDL plan (Line 6).

This symbolic plan is subsequently translated back into a sequence of clear, natural language instructions for the UI Explorer to execute. An example of a raw PDDL plan and its corresponding natural language translation are shown in Appendix in Listing 4 and Table 5, respectively. In cases where the classical planner fails to find a valid path (e.g., if the target is unreachable), AGENT+P implements a fallback strategy. Instead of a plan, it generates a textual summary of the k-hop neighboring nodes from the current UI, providing contextual information to help the agent decide on its next steps (Line

Table 3: Statistics of apps used in evaluation.

App	Tasks	Nodes	Edges
VLC	3	85	190
Simple Calendar Pro	17	24	25
Tasks	6	86	127
Markor	14	14	17
OsmAnd	3	152	508

12-14).

4.4 UI Explorer

The UI Explorer acts as the AGENT+P's execution engine, emulating user interactions with the app (Line 9, Line 14-15, and Line 18-19). It takes the natural language plan and executes each step programmatically. After each action, it evaluates whether the resulting UI state matches the expected goal.

A key design feature of the UI Explorer is its modularity. It is a plug-and-play component, allowing AGENT+P to be integrated with various types of UI agents, including those based on LLMs, MLLMs, or other specialized models that incorporate capabilities like reflection and grounding. This flexibility is demonstrated in our evaluation (Section 5), where we integrate AGENT+P with different UI agents to showcase its broad applicability.

5 Evaluation

5.1 Evaluation Setup

Dataset. We evaluate our AGENT+P on the AndroidWorld benchmark (Rawles et al., 2024). As introduced in Subsection 2.2, AndroidWorld is a widely acknowledged benchmark for evaluating UI agents on real-world tasks and provides automatic metrics to assess an agent's success rate.

Baselines. We evaluate AGENT+P with agents from the official AndroidWorld leaderboard (Android World, 2025): Droidrun, MobileUse, and the T3A agent included with the benchmark. We select these agents as others are either not open-sourced or their released code is not functional. We integrate each of these agents as the GUI Explorer module within AGENT+P for our evaluation.

Implementation. At the time of our experiments, MobileUse claimed a success rate of 0.84, and DroidRun claimed 0.78 on the official AndroidWorld leaderboard. However, the exact configurations used to obtain these results (e.g., backbone LLMs, enabling reasoning or vision capabilities) were not publicly disclosed. To ensure a fair comparison, we rerun each agent using three state-of-the-art LLMs: GPT-5, Gemini-2.5 Pro, and Grok 4, under multiple parameter configurations. We adopt the configuration with the best performance for our evaluation.

For the classical planner, we employ Fast Downward with an A^* search algorithm to find optimal paths. The static UTG is constructed using ICCBot (Yan et al., 2022) and FlowDroid (Arzt et al., 2014). To ensure the reliability of our findings, all experiments were conducted three times, and we report the average values.

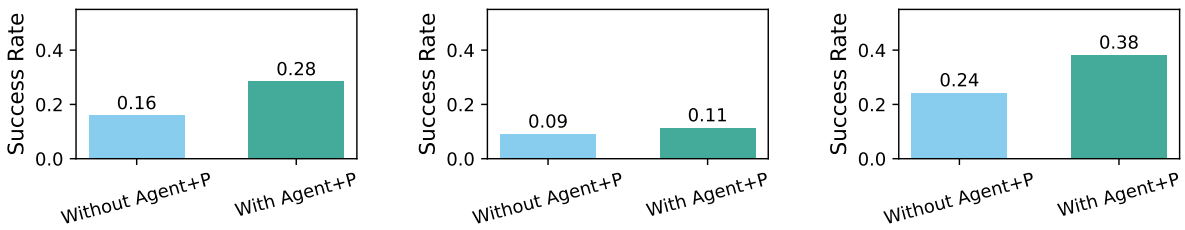
5.2 Effectiveness

To demonstrate the effectiveness of AGENT+P in helping UI agents navigate complex UIs, we conduct an evaluation on a targeted subset of the AndroidWorld benchmark. This subset consists of five apps where baseline agents most frequently failed, as identified in our motivational study in Subsection 2.2. Table 3 summarizes the characteristics of these apps, including the number of tasks and the complexity of their UTGs (nodes and edges).

As illustrated in Figure 2, all baseline agents achieve higher success rates when integrated with AGENT+P compared to running independently. Specifically, AGENT+P enhances the success rate of Droidrun by 12%, MobileUse by 2%, and T3A by mostly 14%. These gains confirm that AGENT+P effectively augments existing UI agents by providing them with global transition information of an app's UIs.

5.3 Efficiency

To evaluate whether AGENT+P enables UI agents to reach target UIs more quickly, we design a specific navigation task: "Go to the privacy policy page". We select this task for two primary reasons. (1) It is simple enough for most agents to accomplish within a reasonable step budget, usually three successive correct actions (e.g., from the main UI, click the "About" or "Settings" widget, scroll down, and then click "Privacy Policy"). (2) the app design, such as visual implementation of the necessary widgets (e.g., settings icons, menu buttons) varies significantly across apps, posing a challenge for LLM-based agents navigating to the target UI (i.e., the privacy policy page) without errors. Therefore, this task provides a fair benchmark to compare agents' abilities to understand diverse



(a) DroidRun (b) MobileUse (c) T3A

Figure 2: Comparison of agent performance across three baselines.

Table 4: Comparison of agent performance on the specific task "Go to the privacy policy page of the current app".

	Success Rate	Actions	Time (s)
w/ AGENT+P	0.60	3.80	16.90
w/o AGENT+P	0.60	6.10	19.90

UI structures and plan action sequences.

For this experiment, we use the Droidrun agent, which has a higher success rate compared to MobileUse, while T3A is coupled only to the AndroidWorld benchmark. We set a maximum limit of 10 steps per task. We compare the performance of the standalone Droidrun agent against Droidrun integrated with AGENT+P on all the apps in AndroidWorld benchmark.

As shown in Table 4, both configurations achieve the same success rate, indicating that the task is generally solvable for a capable agent within the step limit. However, the efficiency gains with AGENT+P are substantial: the agent integrated with AGENT+P required fewer steps on average (3.8) and less time (16.9s) to complete the task compared to the standalone version. The reduction in steps (37.7% fewer) is more pronounced than the reduction in time, which is expected since AGENT+P introduces additional computational overhead for planning. Overall, these results demonstrate that by incorporating global transition information, AGENT+P enables agents to navigate more directly and complete tasks with higher efficiency.

6 Discussion

Symbolic Planning as a Remedy for LLM Hallucination. Our study demonstrates that integrating classical symbolic planning with LLM-based agents substantially mitigates long-horizon planning failures. Unlike end-to-end LLM reasoning, which often suffers from hallucination and myopic exploration, symbolic planners provide *global guarantees* on correctness and optimality. This synergy leverages the complementary strengths of both paradigms: the interpretability and reliability of symbolic reasoning, and the perception and linguistic versatility of LLMs. We believe that such hybrid architectures represent a promising direction for future UI automation and broader embodied AI research.

Broader Applicability and Generalization. Although AGENT+P is designed for Android UI automation, its methodology is domain-agnostic. Any environment that can be abstracted as a state–transition graph (e.g., desktop applications, web navigation, robotic task planning) could potentially benefit from our symbolic-agentic design. In particular, the UTG can be replaced by other forms of structured knowledge, such as the Document Object Model (DOM) in web automation or state machines in robotic control. We anticipate that future systems can extend AGENT+P to operate across platforms and modalities, forming a unified symbolic–neural framework for general user-interface reasoning. web-side utg techniques such as GUITAR (Nguyen et al., 2014).

Multi-Goal Automation. Our current implementation simplifies UI automation tasks into single-goal navigation problems, where the objective is to reach a single target UI. However, many practical user tasks are inherently multifaceted and require achieving a sequence of sub-goals. For instance, a task like *"Add an item to the shopping cart and then proceed to checkout"* involves successfully reaching the item's page (goal 1) and subsequently navigating to the checkout screen (goal 2).

Extending AGENT+P to handle multi-goal scenarios would involve evolving the Node Selector into a more sophisticated *"Goal Decomposer"* capable of parsing a complex natural language instruction into an ordered list of target UI nodes $\{u_{target_1}, u_{target_2}, \ldots, u_{target_n}\}$. Subsequently, the Plan Generator would leverage the native ability

of PDDL to support multiple goal predicates, enabling the generation of a single, cohesive plan that traverses the UTG to satisfy all sub-goals. This effectively extends AGENT+P to a hierarchical planning paradigm. However, this approach also raises new research questions regarding goal ordering and dependency resolution, especially when goals become infeasible at runtime.

7 Related Work

7.1 UI Automation

Traditionally, research in UI automation has centered on *automated testing*, where the primary objective is to systematically explore an app's UIs to discover bugs (Ran et al., 2024; Hu and Neamtiu, 2011; Lai and Rubin, 2019), or security vulnerabilities (Shahriar and Zulkernine, 2009; Moura et al., 2023; Liu et al., 2020). With the recent advent of LLMs, the focus has expanded to *task-driven GUI agents*, which aim to complete specific real-world tasks rather than maximizing the exploration coverage (Wen et al., 2024; Rawles et al., 2024).

7.2 LLM+Classical Planner

Studies using classical planners to enable more reliable planning for LLMs by leveraging LLMs to either translate natural language problems into a formal language like PDDL for a classical planner to solve (Liu et al., 2023; Dagan et al., 2023; Guan et al., 2023) or to generate an initial plan that the planner then refines (Valmeekam et al., 2023). Borrowing from this insight, AGENT+P uses a classical planner to enhance an LLM's capability in performing complex GUI tasks.

7.3 UI Transition Modeling

Existing works construct UTGs (or similar concepts) using various methods across multiple platforms. One primary approach is *dynamic analysis*, where the application is executed to observe real transitions. This is seen in early, multi-platform (e.g., web, mobile) systems like GUITAR (Nguyen et al., 2014) and more recent tools such as GUI-Xplore (Sun et al., 2025) and Autodroid (Wen et al., 2024). The other approach is *static analysis*, which infers control flow from the app's code, such as A^3E's activity transition graph (Azim and Neamtiu, 2013) and Gator's window transition graph (Yang et al., 2018). AGENT+P constructs the UTG to model UI transitions by synergistically combining both static and dynamic techniques.

8 Conclusion

In this paper, we introduced AGENT+P, a novel agentic framework designed to address the critical challenge of long-horizon planning in UI automation. By modeling an app's transition structure as a UTG and leveraging an external symbolic planner, AGENT+P provides LLM-based agents with a globally optimal, high-level plan, effectively mitigating hallucination that causes common automation failure. Our evaluation on the AndroidWorld benchmark demonstrates that AGENT+P can be integrated as a plug-and-play module to substantially improve the success rates of state-of-the-art UI agents. We believe AGENT+P lays the foundation for future research into neuro-symbolic planning paradigms for creating more robust and reliable UI agents for complex UI automation tasks.

Limitations

While AGENT+P effectively bridges global planning with local reasoning, several challenges remain. First, constructing accurate UTGs still depends on program analysis quality. Static analysis may include infeasible edges, while dynamic exploration may yield incomplete coverage. Second, translating between natural language and PDDL representations introduces an additional layer of abstraction; errors in this translation can propagate through the pipeline. Finally, classical planners assume deterministic transitions, yet real-world GUIs often contain stochastic behaviors (e.g., pop-ups or async UI updates), which may lead to plan divergence during execution. Addressing these issues requires tighter integration between the planner and the environment to support real-time re-planning and uncertainty handling.

Ethical Considerations

Automated UI agents that can execute complex tasks on real applications must be deployed with care. Potential misuse, such as automating sensitive or privacy-related operations without explicit consent, highlights the need for transparent auditing and permission control. In our implementation, all experiments are conducted on benign benchmark apps under controlled environments.

References

Constructions Aeronautiques, Adele Howe, Craig Knoblock, ISI Drew McDermott, Ashwin Ram,

Manuela Veloso, Daniel Weld, David Wilkins Sri, Anthony Barrett, Dave Christianson, and 1 others. 1998. Pddl—the planning domain definition language. *Technical Report, Tech. Rep.*

Android World. 2025. Leaderboard of AndroidWorld.

Steven Arzt, Siegfried Rasthofer, Christian Fritz, Eric Bodden, Alexandre Bartel, Jacques Klein, Yves Le Traon, Damien Octeau, and Patrick McDaniel. 2014. Flowdroid: Precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for android apps. *ACM sigplan notices*, 49(6):259–269.

Tanzirul Azim and Iulian Neamtiu. 2013. Targeted and depth-first exploration for systematic testing of android apps. In *Proceedings of the 2013 ACM SIGPLAN International Conference on Object Oriented Programming Systems Languages & Applications*, OOPSLA '13, page 641–660, New York, NY, USA. Association for Computing Machinery.

Gilles Baechler, Srinivas Sunkara, Maria Wang, Fedir Zubach, Hassan Mansoor, Vincent Etter, Victor Cărbune, Jason Lin, Jindong Chen, and Abhanshu Sharma. 2024. Screenai: A vision-language model for ui and infographics understanding. *arXiv preprint arXiv:2402.04615*.

Gautier Dagan, Frank Keller, and Alex Lascarides. 2023. Dynamic planning with a llm. *arXiv preprint arXiv:2308.06391*.

Gaole Dai, Shiqi Jiang, Ting Cao, Yuanchun Li, Yuqing Yang, Rui Tan, Mo Li, and Lili Qiu. 2025. Advancing mobile gui agents: A verifier-driven approach to practical deployment. *arXiv preprint arXiv:2503.15937*.

Google Developers. 2025. Introduction to activities.

Lin Guan, Karthik Valmeekam, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Leveraging pretrained large language models to construct and utilize world models for model-based task planning. *Advances in Neural Information Processing Systems*, 36:79081–79094.

Cuixiong Hu and Iulian Neamtiu. 2011. Automating gui testing for android applications. In *Proceedings of the 6th International Workshop on Automation of Software Test*, pages 77–83.

Duling Lai and Julia Rubin. 2019. Goal-driven exploration for android applications. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 115–127. IEEE.

Jiawei Li, Jiahao Liu, Jian Mao, Jun Zeng, and Zhenkai Liang. 2025a. Ui-ctx: Understanding ui behaviors with code contexts for mobile applications. In *NDSS*.

Ning Li, Xiangmou Qu, Jiamu Zhou, Jun Wang, Muning Wen, Kounianhua Du, Xingyu Lou, Qiuying Peng, and Weinan Zhang. 2025b. Mobileuse: A gui agent with hierarchical reflection for autonomous mobile operation. *arXiv preprint arXiv:2507.16853*.

Toby Jia-Jun Li, Lindsay Popowski, Tom Mitchell, and Brad A Myers. 2021. Screen2vec: Semantic embedding of gui screens and gui components. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. 2023. Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*.

Changlin Liu, Hanlin Wang, Tianming Liu, Diandian Gu, Yun Ma, Haoyu Wang, and Xusheng Xiao. 2022. Promal: precise window transition graphs for android via synergy of program analysis and machine learning. In *Proceedings of the 44th International Conference on Software Engineering*, pages 1755–1767.

Chenxu Liu, Zhiyu Gu, Guoquan Wu, Ying Zhang, Jun Wei, and Tao Xie. 2025. Temac: Multi-agent collaboration for automated web gui testing. *arXiv preprint arXiv:2506.00520*.

Tianming Liu, Haoyu Wang, Li Li, Xiapu Luo, Feng Dong, Yao Guo, Liu Wang, Tegawendé Bissyandé, and Jacques Klein. 2020. MadDroid: Characterizing and detecting devious ad contents for android apps. In *Proceedings of The Web Conference 2020*, pages 1715–1726.

Shang Ma, Chaoran Chen, Shao Yang, Shifu Hou, Toby Jia-Jun Li, Xusheng Xiao, Tao Xie, and Yanfang Ye. 2025. Careful about what app promotion ads recommend! detecting and explaining malware promotion via app promotion graph. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*.

Thiago Santos de Moura, Everton LG Alves, Hugo Feitosa de Figueirêdo, and Cláudio de Souza Baptista. 2023. Cytestion: Automated gui testing for web applications. In *Proceedings of the XXXVII Brazilian Symposium on Software Engineering*, pages 388–397.

Bao N Nguyen, Bryan Robbins, Ishan Banerjee, and Atif Memon. 2014. Guitar: an innovative tool for automated testing of gui-driven software. *Automated software engineering*, 21(1):65–105.

Graziella Orrù, Andrea Piarulli, Ciro Conversano, and Angelo Gemignani. 2023. Human-like problem-solving abilities in large language models using chatgpt. *Frontiers in artificial intelligence*, 6:1199350.

Dezhi Ran, Hao Wang, Zihe Song, Mengzhou Wu, Yuan Cao, Ying Zhang, Wei Yang, and Tao Xie. 2024. Guardian: A runtime framework for llm-based ui exploration. In *ISSTA*.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, and 1 others. 2024. Androidworld: A dynamic

benchmarking environment for autonomous agents. *arXiv preprint arXiv:2405.14573*.

Hossain Shahriar and Mohammad Zulkernine. 2009. Automatic testing of program security vulnerabilities. In *2009 33rd Annual IEEE International Computer Software and Applications Conference*, volume 2, pages 550–555. IEEE.

Yuchen Sun, Shanhui Zhao, Tao Yu, Hao Wen, Samith Va, Mengwei Xu, Yuanchun Li, and Chongyang Zhang. 2025. Gui-xplore: Empowering generalizable gui agents with one exploration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19477–19486.

Antti Valmari. 1996. The state explosion problem. In *Advanced Course on Petri Nets*, pages 429–528. Springer.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. On the planning abilities of large language models-a critical investigation. *Advances in Neural Information Processing Systems*, 36:75993–76005.

Hui Wei, Zihao Zhang, Shenghua He, Tian Xia, Shijia Pan, and Fei Liu. 2025. Plangenllms: A modern survey of llm planning capabilities. *arXiv preprint arXiv:2502.11221*.

Hao Wen, Yuanchun Li, Guohong Liu, Shanhui Zhao, Tao Yu, Toby Jia-Jun Li, Shiqi Jiang, Yunhao Liu, Yaqin Zhang, and Yunxin Liu. 2024. Autodroid: Llm-powered task automation in android. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 543–557.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, and 1 others. Osatlas: A foundation action model for generalist gui agents, 2024c.

Yuquan Xie, Zaijing Li, Rui Shao, Gongwei Chen, Kaiwen Zhou, Yinchuan Li, Dongmei Jiang, and Liqiang Nie. 2025. Mirage-1: Augmenting and updating gui agent with hierarchical multimodal skills. *arXiv preprint arXiv:2506.10387*.

Jiwei Yan, Shixin Zhang, Yepang Liu, Jun Yan, and Jian Zhang. 2022. Iccbot: fragment-aware and context-sensitive icc resolution for android applications. In *Proceedings of the ACM/IEEE 44th international conference on software engineering: companion proceedings*, pages 105–109.

Shengqian Yang, Haowei Wu, Hailong Zhang, Yan Wang, Chandrasekar Swaminathan, Dacong Yan, and Atanas Rountev. 2018. Static window transition graphs for android. *Automated Software Engineering*, 25:833–873.

Jiabo Ye, Xi Zhang, Haiyang Xu, Haowei Liu, Junyang Wang, Zhaoqing Zhu, Ziwei Zheng, Feiyu Gao, Junjie Cao, Zhengxi Lu, and 1 others. 2025. Mobile-agent-v3: Foundamental agents for gui automation. *arXiv preprint arXiv:2508.15144*.

Chi Zhang, Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. 2023. Appagent: Multimodal agents as smartphone users. *arXiv preprint arXiv:2312.13771*.

Kangjia Zhao, Jiahui Song, Leigang Sha, Haozhan Shen, Zhi Chen, Tiancheng Zhao, Xiubo Liang, and Jianwei Yin. 2024. Gui testing arena: A unified benchmark for advancing autonomous gui testing agent. *arXiv preprint arXiv:2412.18426*.

A Additional Figures and Tables

Listing 2: Example problem PDDL for the Simple Calendar Pro app with the task *Change the time zone*.

```
(define (problem change-time-zone)
  (:domain utg-automation)

  (:objects
    SplashActivity - node
    MainActivity - node
    EventActivity - node
    SettingsActivity - node
    AboutActivity - node
    TaskActivity - node
    SelectTimeZoneActivity - node
    ManageEventTypesActivity - node
    WidgetListConfigureActivity - node
    ContributorsActivity - node
    FAQActivity - node
    LicenseActivity - node
  )

  (:init
    (at SplashActivity)

    (goal-node SelectTimeZoneActivity)

    (connected SplashActivity MainActivity)
    (connected MainActivity EventActivity)
    (connected MainActivity SettingsActivity)
    (connected MainActivity AboutActivity)
    (connected MainActivity TaskActivity)
    (connected EventActivity EventActivity)
    (connected EventActivity SelectTimeZoneActivity)
    (connected SettingsActivity
     ManageEventTypesActivity)
    (connected SettingsActivity
     WidgetListConfigureActivity)
    (connected AboutActivity ContributorsActivity)
    (connected AboutActivity FAQActivity)
    (connected AboutActivity LicenseActivity)
    (connected TaskActivity TaskActivity)
  )

  (:goal
    (goal-achieved SelectTimeZoneActivity)
  )
)
```

Listing 3: Prompt template of Node Selector.

```
Given the user goal: "{user_goal}".

Select the most relevant UTG nodes to achieve this
     goal.

Available UTG nodes:
{chr(10).join(nodes_information)}

Return your response in JSON format:
{{
    "nodes": ["node1", "node2", ...],
    "confidence": 0.8,
    "reasoning": "brief explanation"
}}

For example,
Available UTG nodes are:


Return ONLY the JSON, no other text.
```

Listing 4: The plan generated by Fast Downward A^* for the Simple Calendar Pro app with the task *Add a new event type named '1-on-1 meeting'*.

```
(navigate  SplashActivity MainActivity)
(navigate MainActivity SettingsActivity)
(navigate SettingsActivity ManageEventTypesActivity)
; cost = 3 (unit cost)
```

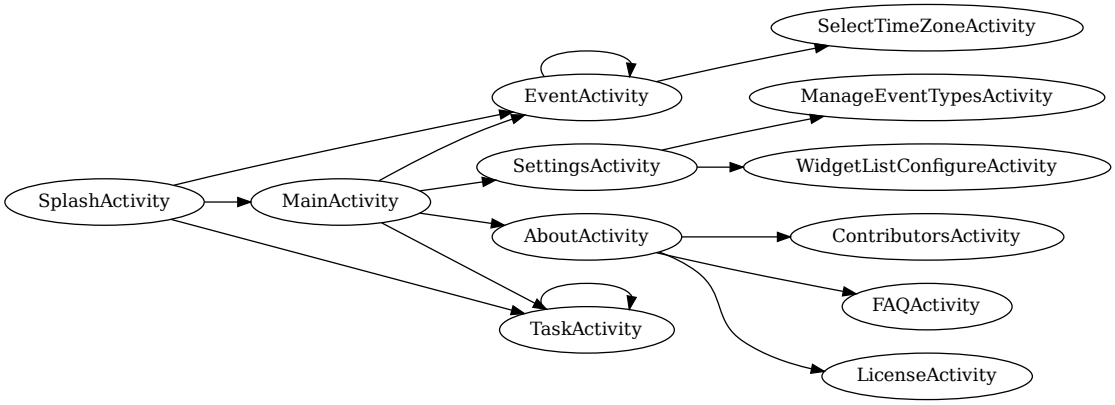


Figure 3: Graphviz visualization of the UTG of Simple Calendar Pro in AndroidWorld. An activity is an unit of Android UI (Google Developers, 2025). Edge labels are removed for visual clarity.

Table 5: Example natural language prompt for the Simple Calendar Pro app with the task *Add a new event type named '1-on-1 meeting'*.

Category	Content
Natural Language Instructions	— UTG Navigation Guide — Current UI: MainActivity — NAVIGATION PLAN FOR YOUR GOAL — Goal Analysis: Based on your goal, the system identified these target destinations: • ManageEventTypesActivity Confidence: 95% OPTIMAL PATH (Follow these steps in order): Step 1: Navigate from MainActivity to SettingsActivity Step 2: Navigate from SettingsActivity to ManageEventTypesActivity IMMEDIATE NEXT ACTION: → on click the ImageView widget with content-description "more options" via API call "virtualinvoke r0.<android.content.Context: void startActivity(android.content.Intent)>(r1)' ()" This will take you to: SettingsActivity Total steps in optimal path: 2
Usage Tips	This path was computed using PDDL planning for guaranteed optimality — USAGE TIPS — • If a NAVIGATION PLAN is shown above, follow it step by step for optimal path • The plan was computed using formal PDDL planning algorithms • If no plan exists, the target is unreachable from current location • Some UI elements may not be visible - scroll if needed
Fallback	— ALL AVAILABLE NAVIGATION OPTIONS FROM HERE — • TO REACH: EventActivity → on click ImageButton widget with context-description "New Event" via API call "virtualinvoke r0.<android.content.Context: void startActivity(android.content.Intent)>(r1)' ()" • TO REACH: SettingsActivity → on click Button widget with content-description "Settings" via API call "virtualinvoke r0.<android.content.Context: void startActivity(android.content.Intent)>(r1)' ()" • TO REACH: TaskActivity → on click ImageButton widget via api call "'virtualinvoke r2.<android.content.Context: void startActivity(android.content.Intent)>(r3)' ()" — End Navigation Guide —