

# TVWorld: Foundations for Remote-Control TV Agents

Zhantao Ma<sup>1\*</sup>, Quanfeng Lu<sup>1\*</sup>, Shuai Zhong<sup>1</sup>, Dahai Yu<sup>3</sup>  
Ping Luo<sup>1</sup>, Michael Ng<sup>2†</sup>

<sup>1</sup>The University of Hong Kong    <sup>2</sup>Hong Kong Baptist University

<sup>3</sup>TCL Corporate Research (Hong Kong) Co., Ltd

<https://github.com/Lqf-HFNJU/TVTheseus>

## Abstract

Recent large vision–language models (LVLMs) have demonstrated strong potential for device control. However, existing research has primarily focused on point-and-click (PnC) interaction, while remote-control (RC) interaction commonly encountered in everyday TV usage remains largely underexplored. To fill this gap, we introduce **TVWorld**, an offline graph-based abstraction of real-world TV navigation that enables reproducible and deployment-free evaluation. On this basis, we derive two complementary benchmarks that comprehensively assess TV-use capabilities: **TVWorld-N** for topology-aware navigation and **TVWorld-G** for focus-aware grounding. These benchmarks expose a key limitation of existing agents: insufficient topology awareness for focus-based, long-horizon TV navigation. Motivated by this finding, we propose a *Topology-Aware Training* framework that injects topology awareness into LVLMs. Using this framework, we develop **TVTheseus**, a foundation model specialized for TV navigation. TVTheseus achieves a success rate of 68.3% on TVWorld-N, surpassing strong closed-source baselines such as Gemini 3 Flash and establishing state-of-the-art (SOTA) performance. Additional analyses further provide valuable insights into the development of effective TV-use agents.

## 1 Introduction

When discussing how Large Vision–Language Models (LVLMs) operate in user interface (UI) environments, existing work predominantly assumes a point-and-click interaction paradigm (Wang et al., 2025a,b; Ye et al., 2025), where a cursor or fingertip directly selects on-screen targets. However, this assumption does not generalize to smart televisions (TVs), a widely deployed and increasingly common media platform (Strategy Analytics,

Inc., 2021), where interaction is mediated through remote-control navigation (Hong and Rivoal, 2019) rather than direct pointing. TV interfaces are inherently focus-based: navigation is performed by pressing directional keys on the remote control (e.g., the UP button), which move a discrete highlight across UI elements, and actions are executed only on the currently focused item (Fig. 1).

This interaction paradigm gives rise to requirements that differ fundamentally from pointer-based UI controls. Effective TV navigation hinges on *focus awareness*, which involves localizing the currently highlighted element within the global screen layout rather than detecting clickable elements in isolation, and *topology-aware planning*, which entails perceiving the underlying UI topology and navigating through discrete, button-driven focus transitions over multiple steps. Yet, existing GUI control benchmarks for LLM agents remain largely dominated by pointer primitives, leaving evaluation settings that capture the demands of TV navigation scarce.

To fill this gap, we introduce **TVWorld**, an offline graph-based abstraction of real-world TV navigation. TV interaction is driven by a small set of discrete remote-control keys and spans a limited set of stable UI screens, which makes a graph formulation natural: each UI state corresponds to a node, and each key press induces a transition recorded as a directed edge, enabling TV navigation to be captured as a finite graph. Based on this abstraction, we systematically traverse real TV interfaces via remote-control interaction to construct high-fidelity navigation graphs. Building on TVWorld, we derive two complementary benchmarks tailored to TV interaction. **TVWorld-N** is an offline interactive TV navigation environment for evaluating agents’ *topology-aware planning* under focus-based remote-control, supporting both textual and visual goals. Operating purely on static graph assets, it is fully replayable and deployment-free (e.g.,

\*Equal Contribution

†Corresponding Author: michael-ng@hkbu.edu.hk

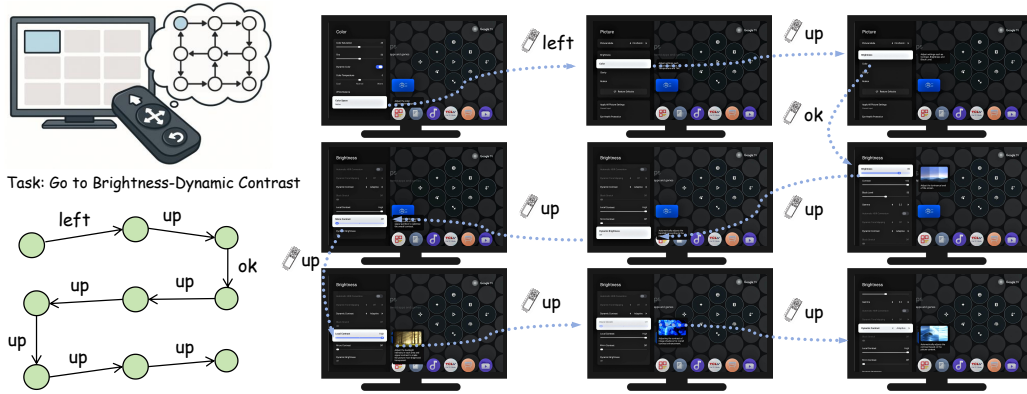


Figure 1: Illustration of focus-based remote-control TV interaction: discrete key presses (e.g. LEFT/UP/OK) move a highlight across UI elements, inducing UI state transitions toward the target. This process can be formulated as path finding on a topology graph whose nodes are UI states and edges correspond to key-induced transitions.

no VMs/emulators), and enables millisecond-level interaction, avoiding the instability and overhead of online GUI benchmarks (Xie et al., 2024; Rawles et al., 2024). Complementarily, **TVWorld-G** evaluates *focus-aware grounding* by requiring the agent to localize the currently highlighted element within the global screen layout using bounding-box annotations, directly reflecting the focus-based nature of TV control.

TVWorld further exposes a critical limitation of recent GUI agents: while they excel at pointer-based interaction, they struggle with the focus-based, long-horizon navigation required by TV control. We attribute this failure to the lack of *topology awareness*, namely the ability to perceive focus structure and navigate through UI state transitions. To address this bottleneck, we design three types of topology-driven training traces that explicitly target fundamental behaviors in TV navigation, including goal-directed focus transitions, recovery from detours, and escape from stalled states. These traces form the basis of a two-stage *Topology-Aware Training* framework that systematically builds focus awareness and topology-aware planning. Using this training strategy, we develop **TVTheseus**, a foundation model specialized for TV navigation. Comprehensive experiments demonstrate the effectiveness and robustness of our approach: TVTheseus achieves a success rate of 68.3 on the out-of-domain TVWorld-N benchmark, outperforming strong closed-source baselines such as Gemini 3 Flash and GPT-5 mini, and attains SOTA performance on TVWorld-G with an accuracy of 81.8 despite receiving no grounding-specific supervision, reflecting strong topology-aware planning and focus-awareness capabilities.

The contributions of this work are three-fold: 1) we introduce **TVWorld**, a static benchmark for interactive TV navigation, together with two complementary evaluation suites; 2) we propose a *Topology-Aware Training* framework tailored to focus-based TV interaction; and 3) using this framework, we develop **TVTheseus**, a foundation model for TV navigation, whose effectiveness is validated through extensive empirical evaluation.

## 2 Related Work

**Graph-based methods.** Graphs provide a convenient abstraction for different domains, including materials science (Butler et al., 2018; You et al., 2018), engineering (Darvariu et al., 2021; Yang et al., 2023), and networking security (Nyberg and Johnson, 2023; Xu et al., 2022). Specifically, robotics frames motion planning as search on configuration graphs (Hossain et al., 2024; Shah et al., 2021). Knowledge-graph reasoning casts question answering as multi-hop traversal (Xiong et al., 2017; Das et al., 2017). For GUI agents, interfaces are often represented as DOM or state graphs, where edges correspond to actionable elements (Jia et al., 2019; Gur et al., 2018; Adamo et al., 2018; Pan et al., 2020; Zhang et al., 2025c; Xu et al., 2025); In the TV auto testing domain, prior work uses crawlers to construct UI graphs to generate test sequences (Firat et al., 2022; Ahmed et al., 2020; Bures et al., 2020). There has been almost no systematic exploration of training LVLMs with graph-based reinforcement learning in remote-control scenarios.

**LVLm for GUI Control.** Work on GUI agents spans the web (Zhang et al., 2025b; Abuelsaad

et al., 2024), mobile device (Li et al., 2024a; Papoudakis et al., 2025), and desktop control (Zhang et al., 2025a; Zhao et al., 2025). One research direction enhances inputs with AIly trees (Wu et al., 2024), Set-of-Marks (Agashe et al., 2024), or DOM (Schiepanski and Piël, 2025) to supply models with fine-grained UI details. Another approach employs control based solely on screenshots to directly determine action positions from pixels (Hong et al., 2024; Li et al., 2024b; Shaw et al., 2023; Wang et al., 2025c; Gou et al., 2024; Ye et al., 2025; Chen et al., 2024). This method is versatile across tasks and devices, and particularly useful when structural inputs are absent or impractical (Cheng et al., 2024; Xie et al., 2024), making it a promising long-term path for transferability. Despite these advances, current LVLM agents still operate mainly in a point-and-click paradigm on cursor or touch interfaces, while remote control scenarios such as TVs remain largely underexplored.

**Training Methods for GUI Control.** GUI-control agents are commonly trained with supervised fine-tuning (SFT), often augmented with chain-of-thought, to improve action prediction (Baechler et al., 2024; Ye et al., 2025; Lu et al., 2025b; Zhang et al., 2024, 2025d). Beyond SFT, which depends on large annotated datasets, recent work frames GUI control as a reinforcement-learning problem via reward design and policy optimization (Lu et al., 2026; Luo et al., 2025; Liu et al., 2025; Zhou et al., 2025; Lee et al., 2025; Lu et al., 2025a; Li et al., 2025), enabling greater sample efficiency and stronger generalization to novel tasks. Complementary efforts explore multi-agent RL and the integration of external tools (Lu et al., 2025c; Singh et al., 2025; Zeng et al., 2025).

### 3 TVWorld

This section introduces **TVWorld**, which transforms real-world remote-control TV interaction into an offline asset for LVLM-based agent navigation. We first formulate TV navigation as a graph search over UI states and define TV-specific tasks (Sec. 3.1). We then describe our on-device data collection pipeline for graph construction (Sec. 3.2). Based on the resulting graphs, we introduce two benchmarks: **TVWorld-N** for end-to-end topology-aware navigation (Sec. 3.3) and **TVWorld-G** for focus-aware grounding evaluation (Sec. 3.4).

#### 3.1 Task Formulation

Remote-control navigation can be modeled as a discrete state-action trajectory. Actions come from a small, fixed set of keys, and a state is represented by a screenshot together with its visible focus highlight. After each key press, the UI renders a new screen: sometimes the focus shifts within the same page, and sometimes the interface switches to another page. This “one key, one transition” view is naturally captured by an action-labeled directed multigraph, where nodes are UI states and edges are key-triggered state changes.

**TV Navigation Graph.** We define the *TV Navigation Graph* as  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \lambda)$  with  $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{A} \times \mathcal{V}$ . Here  $\mathcal{V}$  contains all UI states reachable from an anchor screen (e.g., Home) by feasible key sequences; each  $u \in \mathcal{V}$  is a UI state that contains one screenshot with its focus highlight.  $\mathcal{A}$  is the set of atomic remote actions. An edge  $(u, a, v) \in \mathcal{E}$  means that pressing  $a \in \mathcal{A}$  at state  $u$  yields the next state  $v$ . We write the transition as  $T : \mathcal{V} \times \mathcal{A} \rightarrow \mathcal{V}$  with  $T(u, a) = v$  whenever  $(u, a, v) \in \mathcal{E}$ . Each node carries a label  $\lambda : \mathcal{V} \rightarrow \mathcal{L}$  given by  $\lambda(u) = (S(u), \mathcal{A}(u), m(u))$ , where  $S(u)$  is the screenshot,  $\mathcal{A}(u) \subseteq \mathcal{A}$  lists valid actions at  $u$ , and  $m(u)$  is optional metadata (e.g., text cues or a view-tree).

**Action Set.** The action set of TVWorld comprises 9 kinds of actions: UP, DOWN, LEFT, RIGHT, EXIT, OK, HOME, SETTING, and FINISH. The functionalities of these actions are summarized in Appendix A.

**Topology-Aware Navigation.** TV topology-aware navigation can be modeled as a partially observable Markov decision process (POMDP)  $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{T})$ , where  $\mathcal{S}$  denotes the TV environment states,  $\mathcal{O}$  denotes observations (e.g., screenshots and textual cues),  $\mathcal{A}$  is the set of remote-control actions (see Appendix A), and  $\mathcal{T}$  is the state transition. At timestep  $t$ , the agent receives an observation  $o_t \in \mathcal{O}$  composed of a task instruction  $I$  (visual or textual), the current TV screenshot, and optionally a history of previous observations. Based on  $o_t$ , the agent selects a remote-control key action  $a_t \in \mathcal{A}$ , such as UP or DOWN. Executing  $a_t$  updates the environment to a new state  $s_{t+1} \in \mathcal{S}$  and produces a subsequent observation  $o_{t+1} \in \mathcal{O}$  (e.g., a refreshed screenshot). This agent-TV interaction proceeds until the agent emits the terminal action FINISH or a predefined step budget is exhausted.

**Focus-Aware Grounding.** Focus-aware grounding aims to localize the *currently focused (highlighted)* UI element on a TV interface. Formally, given a TV GUI screenshot  $S$  and an instruction  $I$ , an agent  $\pi$  predicts the focused element’s location as a bounding box  $b = (x_1, y_1, x_2, y_2) \sim \pi(S, I)$ , where  $(x_1, y_1)$  and  $(x_2, y_2)$  denote the top-left and bottom-right coordinates of the focused element.

### 3.2 TVWorld Toolkit

We deployed an automated UI data collection system on physical smart TVs using an MT9655-based 4K platform, and adapted it to two product lines: *TCL TV*<sup>1</sup> and *Google TV*<sup>2</sup>. We designate the TCL TV platform for training, while reserving the Google TV platform exclusively for evaluation. The substantial differences in UI styles, layouts, and design elements between the two platforms enable a rigorous assessment of the agent’s ability to generalize to out-of-domain UI environments; examples are provided in Appendix G. The data acquisition client communicates with the device via Android Debug Bridge (ADB) and a capture card to obtain UI hierarchy snapshots, focus metadata, and screenshots, which are logged per session for traceability. The system interfaces with the television through a hardware-level serial remote-control interface, which injects directional and other key inputs to induce UI state transitions under real-world interaction conditions. Through this combined software–hardware architecture, the system enables end-to-end capture of interaction trajectories and user interface representations.

**Graph Collection Pipeline.** We build directed TV navigation graphs via BFS exploration starting from the homepage. Each node corresponds to a UI state, identified by the screen together with its focused element; the node also stores structured metadata and the associated view-tree dump. For every node, we expand the graph using a fixed, ordered sequence of remote-control key events. When a keypress changes the state, we either create a new node or match it to an existing one, and then add a directed edge labeled with that key. The crawl explicitly avoids a small set of sensitive entry points (e.g., factory reset and language switching), which are masked for safety. After collection, professional TV engineers assign standardized, unambiguous names to nodes so that labels align with

<sup>1</sup>TCL QD-Mini LED TV

<sup>2</sup>4K Ultra HD Smart Google TV

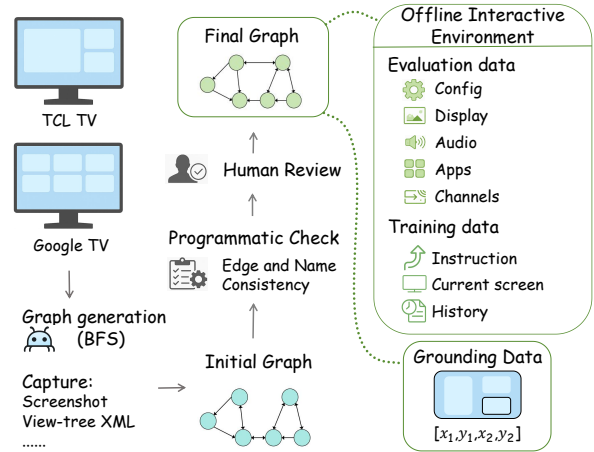


Figure 2: Overview of the TVWorld graphs collection pipeline. We perform BFS exploration on physical TV devices (TCL TV and Google TV) to construct initial UI-state graphs, while recording screenshots and view-tree metadata. Graphs are then refined through automated consistency checks and human inspection, producing finalized graphs together with the offline interactive environment for evaluation/training and grounding data.

Table 1: Statistics of TVWorld.

Platform	Scenario	Nodes	Edges
<i>Train</i>			
TCL TV	Config	282	1,508
<i>Test</i>			
Google TV	Config	169	878
Google TV	Display	62	276
Google TV	Audio	24	104
Google TV	Apps	33	147
Google TV	Channels	32	145

page semantics and avoid duplicates.

**Graph Quality Assurance.** Each generated navigation graph undergoes a structured quality assurance process that integrates automated validation with expert review. Specifically, we perform (i) transition integrity checks, which identify missing inverse links (e.g., an UP move from node  $u$  to node  $v$  without a corresponding DOWN). (ii) naming and hierarchy verification, which flags duplicated node identifiers and inconsistencies in navigational relations (e.g., nodes connected via LEFT or EXIT whose names do not indicate the expected sub-level relation). (iii) human-in-the-loop validation, where all automatically flagged issues are examined and resolved by TV engineers who confirm node definitions, edge semantics, and local navigation behavior. Following these corrections, a senior engineer conducts an end-to-end audit of the full graph to confirm global consistency in topology, naming

Table 2: Comparison of mainstream GUI evaluation benchmarks.

Benchmark	Platform	Interactive Env.	Deploy-Free	Replayable	#Tasks	Real-World	Metadata	Easy Task Gen.
AitW (Zhang et al., 2024)	Mobile	✗	✓	✓	2346	✓	✗	✗
AndroidControl(Li et al., 2024a)	Mobile	✗	✓	✓	15283	✓	✓	✗
GUIOdyssey(Lu et al., 2025d)	Mobile	✗	✓	✓	8334	✓	✗	✗
MiniWoB++ (Liu et al., 2018)	Web	✓	✗	✓	114	✗	✓	✗
WebArena (Zhou et al., 2023)	Web	✓	✗	✓	812	✓	✓	✗
OSWorld (Xie et al., 2024)	Desktop	✓	✗	✗	369	✓	✓	✗
WindowsAgentArena(Bonatti et al., 2024)	Desktop	✓	✗	✗	154	✓	✓	✗
Online-Mind2Web (Xue et al., 2025)	Web	✓	✗	✗	300	✓	✓	✗
AndroidWorld (Rawles et al., 2024)	Mobile	✓	✗	✗	116	✓	✓	✗
<b>TVWorld-N (Ours)</b>	TV	✓	✓	✓	500	✓	✓	✓

hierarchy, and directional behavior, and to ensure that the final graph is coherent and reliable.

**TVWorld Statistics.** Using the proposed graph collection pipeline, we construct 6 *directed* TV navigation graphs spanning 5 UI scenarios: CONFIG (system-level configuration and global settings), DISPLAY (display and rendering), AUDIO (audio-related capabilities), APPS (app entry points and privacy), and CHANNELS (hardware-interface components). All graphs are *strongly connected*, i.e., for any pair of nodes, there exists a directed path between them. Beyond graph topology, each node is associated with rich UI metadata, including screenshots, view-tree structures, focus-related information, and other attributes. For data splits, we use the graph collected from TCL TV for training, and reserve the remaining 5 graphs collected from Google TV for comprehensive evaluation. Detailed per-graph statistics are reported in Table 1.

### 3.3 TVWorld-N

Based on the TV navigation graphs collected from Google TV, we construct **TVWorld-N**, an offline interactive TV navigation environment designed for comprehensive evaluation of agents’ TV navigation capability. A comparison with mainstream GUI benchmarks is reported in Table 2.

**Offline Interactive Environment.** TVWorld-N provides a reproducible offline interactive TV navigation environment by leveraging our high-fidelity data collection pipeline, which establishes a one-to-one correspondence between real TV device states and graph nodes, as well as between device-level state transitions and graph edges. Through this construction, a complex real-world TV interaction environment that typically requires physical hardware, heavy software stacks, and non-trivial deployment is faithfully and almost entirely preserved in a lightweight, static graph representation. Consequently, evaluation can be conducted using only static assets, without running or maintaining inter-

active systems such as operating systems, virtual machines, mobile emulators, or browser automation frameworks. The static nature of the environment further guarantees reproducibility and enables millisecond-level interaction during evaluation.

**Task Construction.** We formulate the topology-aware navigation task as a goal-directed remote-control navigation episode. Given an instruction that specifies a target page, the agent starts from an initial page and iteratively generates actions to interact with the dynamic environment, exploring the UI and moving the focus until the target node is reached. For each task, we randomly sample two distinct nodes from the graph as the start page and the goal page. For the goal specification, we consider two complementary formats: a textual goal defined by the name of the target node, and a visual goal represented by a screenshot of the target UI state. Accordingly, we construct both text-based instructions, such as “I want to go to Privacy–Microphone page,” and vision-based instructions, such as “<image>Navigate to the page shown in the image.” For each graph, we sample 50 unique start–goal node pairs, resulting in a total of 500 TV navigation tasks across all graphs.

### 3.4 TVWorld-G

Based on TVWorld, we construct a focus-aware grounding dataset termed **TVWorld-G**. For each node in the Google TV navigation graphs, we parse the corresponding view-tree and extract the bounding box of the currently focused element, represented as  $(x_1, y_1, x_2, y_2)$  in screen coordinates, where  $(x_1, y_1)$  and  $(x_2, y_2)$  denote the top-left and bottom-right corners, respectively. All extracted annotations are manually verified and corrected when necessary to ensure quality. Finally, TVWorld-G contains 187 samples for grounding evaluation.

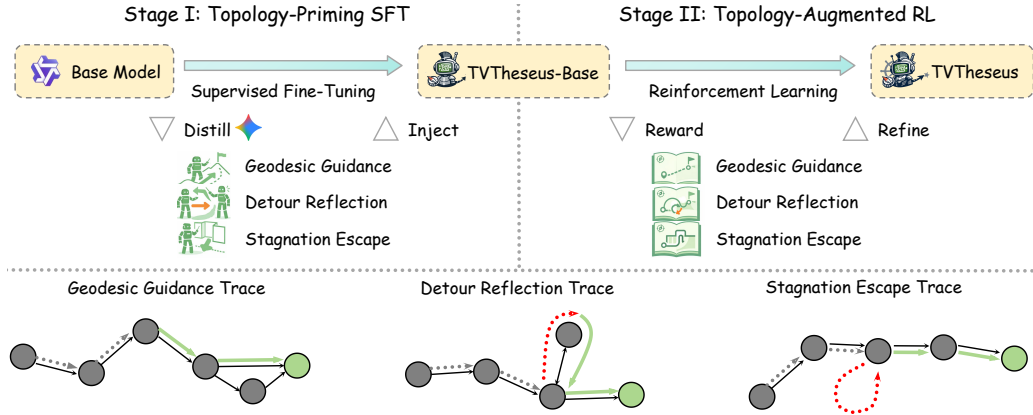


Figure 3: Overview of topology-aware training for TVTheseus. Stage I uses topology-priming SFT by distilling topology-aware behaviors and injecting them into the base model using three trace types: geodesic guidance, detour reflection, and stagnation escape. Stage II then applies topology-augmented RL with trace-specific rewards that promote goal-directed progress while discouraging detours and stagnation; example traces appear at the bottom.

## 4 Topology-Aware Training

Dependable TV navigation requires TV-use agents to reason over focus-based UI transitions in a goal-directed manner, while remaining robust to navigation errors such as detours and stalled states. We collectively refer to this interaction-level competence as *topology awareness*. To embed this latent capability into TV-use agents, we introduce a two-stage training approach that first injects topology-aware inductive biases via topology-priming supervised fine-tuning (Sec. 4.1), and then progressively consolidates them through topology-augmented reinforcement learning (Sec. 4.2), as illustrated in Fig. 3. Through this training paradigm, we obtain **TVTheseus**, a foundation model specialized for robust and generalizable TV control.

### 4.1 Stage I: Topology-Priming SFT

In our early experiments, we observe that existing open-source models often fail to exhibit *topology awareness*, leading to brittle behavior in focus-based TV navigation. To address this, we leverage natural-language reasoning (Lu et al., 2025d; Wang et al., 2025b) as a mechanism for shaping the agent’s internal understanding of UI dynamics. Concretely, at each time step  $t$ , we constrain the model to produce a paired output  $(z_t^*, a_t^*)$ , where  $a_t^*$  denotes the reference key action and  $z_t^*$  provides structured reasoning for that action. Building on this formulation, we introduce a topology-aware priming framework that instills such reasoning through three types of step-level rationales: **Geodesic Guidance**, **Detour Reflection**, and **Stagnation Escape**. Below, we describe the construc-

tion of these traces and the synthesis of the corresponding rationales  $z_t^*$  for a given start–goal pair  $(u, g)$ :

**Geodesic Guidance Traces.** We follow a clean geodesic route on the navigation graph from a start node to a goal node by computing the shortest path  $\mathbf{p}^*$ . For each step  $t$  in  $\mathbf{p}^*$ , we endow  $z_t^*$  with (i) a description of the current UI state and focused element and (ii) an explanation of a locally plausible move that makes progress toward the goal. Instead of encouraging explicit topology memorization, these traces emphasize learning state transitions under key presses and maintaining an explicit notion of goal-directed progress.

**Detour Reflection Traces.** Real-world TV navigation often deviates from the shortest path, as an incorrect key press may move the agent farther from the goal, resulting in a *topological detour*. For a node  $u_t$ , we define a detour action  $a_{\text{far}}$  as any action that increases the shortest-path distance to the goal, i.e.,  $d_{\text{sp}}(T(u_t, a_{\text{far}}), g) > d_{\text{sp}}(u_t, g)$ . Starting from the shortest path, we deliberately insert such a detour and then return to the original node before continuing:  $u_t \xrightarrow{a_{\text{far}}} u_{\text{far}} \xrightarrow{a_{\text{back}}} u_t \xrightarrow{a_t^*} u_{t+1}$ . For action  $a_t^*$ , we design  $z_t^*$  to reflect on the detour and justify a corrected move, explicitly discouraging repeating  $a_{\text{far}}$  and favoring an action consistent with goal-directed progress.

**Stagnation Escape Traces.** We observe another common failure mode in TV navigation, where certain key presses do not trigger any state change. In such cases, the agent may repeatedly issue the same invalid key and become trapped in a local loop. To

capture this behavior, we insert an invalid action  $a_{\text{inv}}$  into an otherwise valid navigation segment:  $u_t \xrightarrow{a_{\text{inv}}} u_t \xrightarrow{a_t^*} u_{t+1}$ . At this point,  $z_t^*$  is designed to recognize that the UI remains unchanged and to reason about the need to abandon the ineffective action, favoring an alternative key that leads to an actual transition.

We use Gemini 3 Pro Preview (Google, 2025) to synthesize the three types of rationales  $z_t^*$ ; examples of the synthesized data are provided in Appendix J. Through supervised fine-tuning on these structured rationales, the base LVLM acquires foundational topology-aware capabilities for focus-based TV navigation. We refer to the resulting model after this stage as **TVTheseus-Base**.

## 4.2 Stage II: Topology-Augmented RL

After topology-priming SFT equips the base LVLM with a strong initial policy, we introduce a second stage, Topology-Augmented Reinforcement Learning, to further consolidate topology-aware behaviors through interaction-driven optimization.

### 4.2.1 Reinforcement Learning Formulation

As introduced in Sec. 3, TVWorld provides a fully offline, replayable interactive environment with millisecond-level response, enabling stable and efficient reinforcement learning without physical devices or online deployment. Within this environment, we construct interaction episodes on the training graphs following the same three trace patterns introduced in Stage I (Sec. 4.1), allowing the agent to explore alternative behaviors and receive feedback from state transitions. We adopt GRPO (Shao et al., 2024) as the optimization algorithm and initialize the policy from TVTheseus-Base. Details of GRPO are provided in Appendix F.

### 4.2.2 Topology-Aware Reward Design

In remote-control TV navigation, each valid key press induces a transition along an edge of the TV navigation graph, potentially moving the interface closer to, unchanged from, or farther away from the goal node. We exploit this structural property to design *topology-aware rewards*. These rewards guide the agent’s behavior by examining the change in graph distance to the goal between the current node  $u_t$  and the node  $u'$  reached after executing a model-suggested action, that is,  $d(u_t, g) - d(u', g)$ .

**Trace-Specific Reward Design.** Our shaping reward assigns higher values to actions that make

goal-directed progress (reducing  $d(\cdot, g)$ ), a small positive value to distance-preserving moves, and lower values to actions that move away from the goal. We further incorporate trace-dependent penalties to correct common failure modes: *Detour Reflection* discourages returning to the previously identified detour branch, while *Stagnation Escape* discourages repeating the invalid action. Complete reward definitions are provided in Appendix D.

**Distance Definition.** We instantiate the distance metric as the *hitting time*  $d(u, g) = h_g(u)$ , which measures the expected number of steps for a random walk starting from node  $u$  to first reach the target node  $g$ . Let  $A$  denote the adjacency matrix of the navigation graph, where  $A_{uv}$  counts feasible actions from  $u$  to  $v$ ,  $D = \text{diag}(\sum_v A_{uv})$ , and  $P = D^{-1}A$ . With  $h_g(g) = 0$ , the hitting time satisfies  $h_g(u) = 1 + \sum_v P_{uv} h_g(v)$ . Detailed distance definitions and alternatives are provided in Appendix E.

We combine the topology-aware reward with a format-validity reward into a single scalar objective:  $R = \beta_{\text{topo}} R_{\text{topo}} + \beta_{\text{form}} R_{\text{form}}$ . Together, these rewards provide dense and structured feedback that reinforces topology-aware behaviors without relying on fixed reference actions or auxiliary verifiers (Bonatti et al., 2024; Chen et al., 2025; Devidze et al., 2021). By aligning reinforcement learning objectives with the topology-aware reasoning patterns introduced during Stage I, this design further unlocks the latent topology awareness of the LVLM, yielding policies that are more robust and better reflect real-world remote-control navigation.

## 5 Experiments

### 5.1 Experimental Setup

**Training Settings.** We adopt Qwen3-VL-8B-Instruct (Bai et al., 2025) as the base model. At each time step  $t$ , the input is  $\mathbf{x}_t = (S_t, S_{t-1:t-\delta_S}, a_{t-1:t-\delta_a}, I)$ , including the current screenshot  $S_t$ , up to 4 historical screenshots ( $\delta_S = 4$ ), the full action history ( $\delta_a = t$ ), and the instruction  $I$ . All experiments are conducted on  $8 \times$  NVIDIA A100 GPUs. Further details on the training data and the two-stage training setup are provided in Appendix C and Appendix B, respectively.

**Evaluation.** We evaluate our TVTheseus on two tasks: Topology-aware Navigation and Focus-aware Grounding.

Table 3: Comprehensive evaluation on TVWorld-N across five out-of-domain scenarios. Boldface indicates the best performance. For **TVTheseus**, we report the mean over three independent runs. Detailed results for each individual run are provided in Appendix Table 10.

Model	Instr Type	Config	Display	Audio	Apps	Channels	Overall
<i>Closed-Source Model</i>							
GPT-5 mini	text-based	<b>48.0</b>	54.0	58.0	48.0	68.0	60.2
	vision-based	60.0	54.0	52.0	78.0	82.0	
Gemini 3 Flash	text-based	46.0	<b>66.0</b>	58.0	62.0	74.0	66.4
	vision-based	54.0	<b>68.0</b>	<b>62.0</b>	80.0	<b>94.0</b>	
Claude Haiku 4.5	text-based	24.0	34.0	42.0	40.0	50.0	25.4
	vision-based	2.0	10.0	12.0	20.0	20.0	
<i>General Open-source Model</i>							
Qwen3-VL-8B-Instruct	text-based	8.0	24.0	10.0	40.0	16.0	20.0
	vision-based	14.0	10.0	10.0	32.0	36.0	
Qwen3-VL-32B-Instruct	text-based	26.0	30.0	44.0	34.0	44.0	39.0
	vision-based	22.0	42.0	36.0	60.0	52.0	
<i>PnC-specific model</i>							
UI-Tars-1.5-7B	text-based	0.0	4.0	2.0	6.0	0.0	1.6
	vision-based	0.0	0.0	0.0	4.0	0.0	
OpenCUA-7B	text-based	0.0	2.0	8.0	4.0	18.0	5.0
	vision-based	0.0	2.0	6.0	6.0	4.0	
GUI-Owl-7B	text-based	0.0	2.0	10.0	2.0	16.0	5.4
	vision-based	2.0	0.0	2.0	10.0	10.0	
GUI-Owl-32B	text-based	2.0	18.0	22.0	20.0	26.0	15.4
	vision-based	4.0	10.0	6.0	20.0	26.0	
<i>RC-specific model</i>							
<b>TVTheseus (Ours)</b>	text-based	41.3 <sup>+6.7</sup> <sub>-5.3</sub>	62.7 <sup>+9.3</sup> <sub>-10.7</sub>	<b>86.0</b> <sup>+2.0</sup> <sub>-4.0</sub>	<b>70.7</b> <sup>+5.3</sup> <sub>-2.7</sub>	<b>76.0</b> <sup>+8.0</sup> <sub>-4.0</sub>	<b>68.3</b> <sup>+3.1</sup> <sub>-2.7</sub>
	vision-based	<b>62.7</b> <sup>+3.3</sup> <sub>-4.7</sub>	60.7 <sup>+3.3</sup> <sub>-6.7</sub>	58.0 <sup>+6.0</sup> <sub>-10.0</sub>	<b>82.7</b> <sup>+1.3</sup> <sub>-0.7</sub>	82.7 <sup>+3.3</sup> <sub>-4.7</sub>	

**Topology-aware Navigation.** Evaluation is conducted on TVWorld-N, which consists of 5 navigation graphs, each containing 100 tasks with both text-based and vision-based instructions (Sec. 3.3), totaling 500 tasks. We compare TVTheseus with (i) closed-source models (GPT-5 mini (OpenAI, 2025), Gemini 3 Flash (Google, 2025), Claude Haiku 4.5 (Anthropic, 2025)), (ii) general-purpose open-source LVLMS (Qwen3-VL-8B-Instruct, Qwen3-VL-32B-Instruct (Bai et al., 2025)), and (iii) pointer-based UI control models (UI-Tars-1.5-7B (Seed, 2025), OpenCUA-7B (Wang et al., 2025b), GUI-Owl-7B, GUI-Owl-32B (Ye et al., 2025)). All models are evaluated with a maximum horizon of 50 steps, image resolution  $1024 \times 576$ , up to 4 historical screenshots, and the full action history. We report **Success Rate (SR)**, defined as finishing on the target page.

**Focus-aware Grounding.** We evaluate focus localization on TVWorld-G, which contains 187 samples, comparing TVTheseus with general-purpose LVLMS (Qwen2.5-VL-7B-Instruct, Qwen3-VL-8B-Instruct, Qwen3-VL-8B-Thinking (Bai et al., 2025)) and pointer-based grounding models (InfiGUI-R1-3B (Liu et al., 2025), GUI-R1-3B, GUI-R1-7B (Luo et al., 2025), GUI-Owl-7B, GUI-Owl-32B (Ye et al., 2025)). All models use an input resolution of  $1024 \times 576$ . Performance is measured by **Acc@0.5** ( $\text{IoU} \geq 0.5$ ).

## 5.2 Main Results

**Interactive TV Navigation Evaluation.** Table 3 reports results on TVWorld-N. TVTheseus achieves the best overall performance, outperforming all baselines, including the strongest closed-source model, Gemini 3 Flash. On previously unseen TV platforms, TVTheseus markedly improves over its base model, Qwen3-VL-8B-Instruct, with success rate increasing from 20.0 to 68.3. Appendix I provides a qualitative case study illustrating the behavioral differences before and after training. This result demonstrates strong out-of-domain generalization enabled by our two-stage Topology-Aware Training. In comparison, the strongest general-purpose open-source model (i.e., Qwen3-VL-32B-Instruct) reaches a success rate of 39.0, substantially lagging behind closed-source models such as Gemini 3 Pro and GPT-5 mini. Models trained for point-and-click (PnC) interaction (e.g., GUI-Owl and OpenCUA) degrade substantially in the remote-control (RC) TV setting, highlighting the mismatch between pointer-based assumptions and focus-based TV navigation. Performance also varies consistently across scenarios: most models perform better on *Apps* and *Channels*, while *Config* remains the most challenging. This pattern suggests that TVWorld-N is a discriminative benchmark that effectively distinguishes genuine topology-aware navigation from superficial

interaction heuristics.

**Focus-Aware Grounding Evaluation.** Table 4 reports focus-aware grounding performance on TVWorld-G. Although TVTheseus is not trained with any grounding-specific supervision, it outperforms its base model (i.e., Qwen3-VL-8B-Instruct), by 3.7 points, achieving the best overall Acc@0.5 of 81.8. This result indicates that strong topology awareness acquired in TV environments transfers to improved focus localization. We also observe that Qwen3-VL-8B-Thinking performs 8.6 points worse than Qwen3-VL-8B-Instruct, suggesting that explicit multi-step reasoning may not be necessary for this task. Consistent with navigation results, PnC-specific models underperform general-purpose LVLMs, further reflecting the fundamental mismatch between pointer-based and remote-control interaction paradigms, which impose distinct capability requirements on LVM agents.

Table 4: Focus-awareness performance on TVWorld-G.

Model	Acc@0.5
<i>General Open-source Model</i>	
Qwen2.5-VL-7B-Instruct	66.3
Qwen3-VL-8B-Instruct	78.1
Qwen3-VL-8B-Thinking	69.5
<i>PnC-specific model</i>	
InfiGUI-R1-3B	56.7
GUI-R1-3B	49.2
GUI-R1-7B	65.2
GUI-Owl-7B	39.5
GUI-Owl-32B	54.0
<i>RC-specific model</i>	
<b>TVTheseus (Ours)</b>	<b>81.8</b>

### 5.3 Ablation Study

#### Effect of the Topology-Aware Training Strategy.

Table 5 shows that the two training stages play complementary roles. Stage I (SFT) establishes core topology-aware behaviors and strong focus awareness, while Stage II (RL) further improves long-horizon planning and recovery. Although Stage II introduces a mild trade-off in focus grounding, the impact is limited, and the model gains substantially stronger topology-aware navigation capability.

Table 5: Impact of two-stage topology-aware training.

Training Strategy	TVWorld-N	TVWorld-G
–	20.0	78.1
+Stage I	48.0	83.4
+Stage I & Stage II	68.3	81.8

**Additional Experiments.** We report additional experiments in Appendix H. These include ablations on rationale types, distance definitions, and reward formulations, as well as analyses of how image resolution and the number of historical screenshots affect model performance.

## 6 Conclusion

In this work, we present TVWorld, a comprehensive and static interactive resource that fills a critical gap in remote-control-based TV agent development. By providing a unified set of benchmarks, an effective training framework, and a specialized TV foundation model, we establish essential building blocks for studying TV-use agents under remote-control interaction paradigms. We hope this work will spur further study of GUI agents in remote-control settings and catalyze broader research on this interaction paradigm.

### Limitations

This work studies TV-use agents at a practical model scale representative of current deployable systems, rather than conducting an exhaustive scaling analysis of substantially larger pretrained models. While scaling effects are not explored, we expect the main findings on topology-aware training and evaluation to generalize beyond model size. In addition, to support responsible data collection and release, we mask a small number of sensitive system-level entry points during graph construction. While a small portion of nodes are masked, the vast majority are preserved, retaining the core interaction structure, navigation topology, and focus-based control dynamics of real-world TV deployments.

### Acknowledgments

We are grateful to TCL Corporate Research (Hong Kong) for providing hardware facilities. This research is supported by the GDSTC: Guangdong and Hong Kong Universities "1+1+1" Joint Research Collaboration Scheme project No.: 2025A0505000007, National Key Research and Development Program of China under Grant 2024YFE0202900, RGC GRF 12300125, and the AI Agent for Industrial Applications No. 200022310.

## References

- Tamer Abuelsaad, Deepak Akkil, Prasenjit Dey, Ashish Jagmohan, Aditya Vempaty, and Ravi Kokku. 2024. Agent-e: From autonomous web navigation to foundational design principles in agentic systems. *arXiv preprint arXiv:2407.13032*.
- David Adamo, Md Khorrom Khan, Sreedevi Koppula, and Renée Bryce. 2018. Reinforcement learning for android gui testing. In *Proceedings of the 9th ACM SIGSOFT international workshop on automating TEST case design, selection, and evaluation*, pages 2–8.
- Saaket Agashe, Jiuzhou Han, Shuyu Gan, Jiachen Yang, Ang Li, and Xin Eric Wang. 2024. Agent s: An open agentic framework that uses computers like a human. *arXiv preprint arXiv:2410.08164*.
- Bestoun S Ahmed, Angelo Gargantini, and Miroslav Bures. 2020. An automated testing framework for smart tv apps based on model separation. In *2020 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, pages 62–73. IEEE.
- Reid Andersen, Fan Chung, and Kevin Lang. 2006. Local graph partitioning using pagerank vectors. In *2006 47th annual IEEE symposium on foundations of computer science (FOCS'06)*, pages 475–486. IEEE.
- Anthropic. 2025. Claude.
- 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*.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.
- Avrim Blum, John Hopcroft, and Ravindran Kannan. 2020. *Foundations of data science*. Cambridge University Press.
- Rogério Bonatti, Dan Zhao, Francesco Bonacci, Dillon Dupont, Sara Abdali, Yinheng Li, Yadong Lu, Justin Wagle, Kazuhito Koishida, Arthur Buckner, and 1 others. 2024. Windows agent arena: Evaluating multi-modal os agents at scale. *arXiv preprint arXiv:2409.08264*.
- Miroslav Bures, Miroslav Macik, Bestoun S Ahmed, Vaclav Rechtberger, and Pavel Slavik. 2020. Testing the usability and accessibility of smart tv applications using an automated model-based approach. *IEEE transactions on consumer electronics*, 66(2):134–143.
- Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. 2018. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555.
- Cong Chen, Kaixiang Ji, Hao Zhong, Muzhi Zhu, Anzhou Li, Guo Gan, Ziyuan Huang, Cheng Zou, Jiajia Liu, Jingdong Chen, and 1 others. 2025. Guishepherd: Reliable process reward and verification for long-sequence gui tasks. *arXiv preprint arXiv:2509.23738*.
- Dongping Chen, Yue Huang, Siyuan Wu, Jingyu Tang, Liuyi Chen, Yilin Bai, Zhigang He, Chenlong Wang, Huichi Zhou, Yiqiang Li, and 1 others. 2024. Gui-world: A video benchmark and dataset for multi-modal gui-oriented understanding. *arXiv preprint arXiv:2406.10819*.
- Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Li YanTao, Jianbing Zhang, and Zhiyong Wu. 2024. SeeClick: Harnessing gui grounding for advanced visual gui agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9313–9332.
- Victor-Alexandru Darvari, Stephen Hailes, and Mirco Musolesi. 2021. Goal-directed graph construction using reinforcement learning. *Proceedings of the Royal Society A*, 477(2254):20210168.
- Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alex Smola, and Andrew McCallum. 2017. Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. *arXiv preprint arXiv:1711.05851*.
- Rati Devidze, Goran Radanovic, Parameswaran Kamalaruban, and Adish Singla. 2021. Explicable reward design for reinforcement learning agents. *Advances in neural information processing systems*, 34:20118–20131.
- Atıl Fırat, Mohammad Yusaf Azimi, Celal Çağın Elgün, Ferhat Erata, and Cemal Yılmaz. 2022. Model-based test adaptation for smart tvs. In *Proceedings of the 3rd ACM/IEEE International Conference on Automation of Software Test*, pages 52–53.
- Google. 2025. Gemini.
- Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie, Cheng Chang, Yiheng Shu, Huan Sun, and Yu Su. 2024. Navigating the digital world as humans do: Universal visual grounding for gui agents. *arXiv preprint arXiv:2410.05243*.
- Izzeddin Gur, Ulrich Rueckert, Aleksandra Faust, and Dilek Hakkani-Tur. 2018. Learning to navigate the web. *arXiv preprint arXiv:1812.09195*.
- Jihye Hong and Florian Rivoal. 2019. [CSS spatial navigation level 1](#). W3C Working Draft.

- Wenyi Hong, Weihang Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and 1 others. 2024. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290.
- Jumman Hossain, Abu-Zaher Faridee, Nirmalya Roy, Jade Freeman, Timothy Gregory, and Theron Trout. 2024. Toponav: Topological navigation for efficient exploration in sparse reward environments. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 693–700. IEEE.
- Sheng Jia, Jamie Kiros, and Jimmy Ba. 2019. Domq-net: Grounded rl on structured language. *arXiv preprint arXiv:1902.07257*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Hyunseok Lee, Jeonghoon Kim, Beomjun Kim, Jihoon Tack, Chansong Jo, Jaehong Lee, Cheonbok Park, Sookyo In, Jinwoo Shin, and Kang Min Yoo. 2025. Reguide: Data efficient gui grounding via spatial reasoning and search. *arXiv preprint arXiv:2505.15259*.
- Wei Li, William E Bishop, Alice Li, Christopher Rawles, Folawiyo Campbell-Ajala, Divya Tyamagundlu, and Oriana Riva. 2024a. On the effects of data scale on ui control agents. *Advances in Neural Information Processing Systems*, 37:92130–92154.
- Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. 2025. Webthinker: Empowering large reasoning models with deep research capability. *arXiv preprint arXiv:2504.21776*.
- Zhangheng Li, Keen You, Haotian Zhang, Di Feng, Harsh Agrawal, Xiujun Li, Mohana Prasad Sathya Moorthy, Jeff Nichols, Yinfei Yang, and Zhe Gan. 2024b. Ferret-ui 2: Mastering universal user interface understanding across platforms. *arXiv preprint arXiv:2410.18967*.
- Evan Zheran Liu, Kelvin Guu, Panupong Pasupat, Tianlin Shi, and Percy Liang. 2018. [Reinforcement learning on web interfaces using workflow-guided exploration](#). In *International Conference on Learning Representations (ICLR)*.
- Yuhang Liu, Pengxiang Li, Congkai Xie, Xavier Hu, Xiaotian Han, Shengyu Zhang, Hongxia Yang, and Fei Wu. 2025. Infigui-r1: Advancing multimodal gui agents from reactive actors to deliberative reasoners. *arXiv preprint arXiv:2504.14239*.
- Fanbin Lu, Zhisheng Zhong, Shu Liu, Chi-Wing Fu, and Jiaya Jia. 2025a. Arpo: End-to-end policy optimization for gui agents with experience replay. *arXiv preprint arXiv:2505.16282*.
- Fanbin Lu, Zhisheng Zhong, Ziqin Wei, Shu Liu, Chi-Wing Fu, and Jiaya Jia. 2025b. Steve: A step verification pipeline for computer-use agent training. *arXiv preprint arXiv:2503.12532*.
- Quanfeng Lu, Zhantao Ma, Shuai Zhong, Jin Wang, Dahai Yu, Michael K Ng, and Ping Luo. 2025c. Swirl: A staged workflow for interleaved reinforcement learning in mobile gui control. *arXiv preprint arXiv:2508.20018*.
- Quanfeng Lu, Wenqi Shao, Zitao Liu, Lingxiao Du, Fanqing Meng, Boxuan Li, Botong Chen, Siyuan Huang, Kaipeng Zhang, and Ping Luo. 2025d. Guiodyssey: A comprehensive dataset for cross-app gui navigation on mobile devices. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22404–22414.
- Zhengxi Lu, Yuxiang Chai, Yaxuan Guo, Xi Yin, Liang Liu, Hao Wang, Han Xiao, Shuai Ren, Pengxiang Zhao, Guangyi Liu, and 1 others. 2026. Ui-r1: Enhancing efficient action prediction of gui agents by reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 17608–17616.
- Run Luo, Lu Wang, Wanwei He, and Xiaobo Xia. 2025. Gui-r1: A generalist r1-style vision-language action model for gui agents. *arXiv preprint arXiv:2504.10458*.
- Jakob Nyberg and Pontus Johnson. 2023. Training automated defense strategies using graph-based cyber attack simulations. *arXiv preprint arXiv:2304.11084*.
- OpenAI. 2025. Gpt.
- Minxue Pan, An Huang, Guoxin Wang, Tian Zhang, and Xuandong Li. 2020. Reinforcement learning based curiosity-driven testing of android applications. In *Proceedings of the 29th ACM SIGSOFT international symposium on software testing and analysis*, pages 153–164.
- Georgios Papoudakis, Thomas Coste, Zhihao Wu, Jianye Hao, Jun Wang, and Kun Shao. 2025. Appvlm: A lightweight vision language model for online app control. *arXiv preprint arXiv:2502.06395*.
- 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*.
- Thassilo M Schiepaniski and Nicholas Piël. 2025. Beyond pixels: Exploring dom downsampling for llm-based web agents. *arXiv preprint arXiv:2508.04412*.
- ByteDance Seed. 2025. Ui-tars-1.5. <https://seed-tars.com/1.5>.

- Dhruv Shah, Benjamin Eysenbach, Gregory Kahn, Nicholas Rhinehart, and Sergey Levine. 2021. Ving: Learning open-world navigation with visual goals. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13215–13222. IEEE.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Peter Shaw, Mandar Joshi, James Cohan, Jonathan Berant, Panupong Pasupat, Hexiang Hu, Urvashi Khandelwal, Kenton Lee, and Kristina N Toutanova. 2023. From pixels to ui actions: Learning to follow instructions via graphical user interfaces. *Advances in Neural Information Processing Systems*, 36:34354–34370.
- Guangming Sheng, Chi Zhang, Zilinfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.
- Joykirat Singh, Raghav Magazine, Yash Pandya, and Akshay Nambi. 2025. Agentic reasoning and tool integration for llms via reinforcement learning. *arXiv preprint arXiv:2505.01441*.
- Strategy Analytics, Inc. 2021. [Strategy analytics: Global smart TV household ownership to exceed 50% by 2026](#). Business Wire press release.
- Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. 2006. Fast random walk with restart and its applications. In *Sixth international conference on data mining (ICDM'06)*, pages 613–622. IEEE.
- Haoming Wang, Haoyang Zou, Huatong Song, Jiazhan Feng, Junjie Fang, Junting Lu, Longxiang Liu, Qinyu Luo, Shihao Liang, Shijue Huang, and 1 others. 2025a. Ui-tars-2 technical report: Advancing gui agent with multi-turn reinforcement learning. *arXiv preprint arXiv:2509.02544*.
- Xinyuan Wang, Bowen Wang, Dunjie Lu, Junlin Yang, Tianbao Xie, Junli Wang, Jiaqi Deng, Xiaole Guo, Yiheng Xu, Chen Henry Wu, and 1 others. 2025b. Opencua: Open foundations for computer-use agents. *arXiv preprint arXiv:2508.09123*.
- Yiqin Wang, Haoji Zhang, Jingqi Tian, and Yansong Tang. 2025c. Ponder & press: Advancing visual gui agent towards general computer control. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1461–1473.
- Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, and 1 others. 2024. Os-atlas: A foundation action model for generalist gui agents. *arXiv preprint arXiv:2410.23218*.
- Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, and 1 others. 2024. Osvorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094.
- Wenhan Xiong, Thien Hoang, and William Yang Wang. 2017. Deeppath: A reinforcement learning method for knowledge graph reasoning. *arXiv preprint arXiv:1707.06690*.
- Tianqi Xu, Linyao Chen, Dai-Jie Wu, Yanjun Chen, Zecheng Zhang, Xiang Yao, Zhiqiang Xie, Yongchao Chen, Shilong Liu, Bochen Qian, and 1 others. 2025. Crab: Cross-environment agent benchmark for multimodal language model agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21607–21647.
- Xiaoyu Xu, Hao Hu, Yuling Liu, Jinglei Tan, Hongqi Zhang, and Haotian Song. 2022. Moving target defense of routing randomization with deep reinforcement learning against eavesdropping attack. *Digital Communications and Networks*, 8(3):373–387.
- Tianci Xue, Weijian Qi, Tianneng Shi, Chan Hee Song, Boyu Gou, Dawn Song, Huan Sun, and Yu Su. 2025. An illusion of progress? assessing the current state of web agents. *arXiv preprint arXiv:2504.01382*.
- Shanchao Yang, MA Kaili, Baoxiang Wang, Tianshu Yu, and Hongyuan Zha. 2023. Learning to boost resilience of complex networks via neural edge rewiring. *Transactions on Machine Learning Research*.
- 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: Fundamental agents for gui automation. *arXiv preprint arXiv:2508.15144*.
- Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. 2018. Graph convolutional policy network for goal-directed molecular graph generation. *Advances in neural information processing systems*, 31.
- Siliang Zeng, Quan Wei, William Brown, Oana Frunza, Yuriy Nevmyvaka, and Mingyi Hong. 2025. Reinforcing multi-turn reasoning in llm agents via turn-level credit assignment. *arXiv preprint arXiv:2505.11821*.
- Chaoyun Zhang, He Huang, Chiming Ni, Jian Mu, Si Qin, Shilin He, Lu Wang, Fangkai Yang, Pu Zhao, Chao Du, and 1 others. 2025a. Ufo2: The desktop agentos. *arXiv preprint arXiv:2504.14603*.
- Danqing Zhang, Balaji Rama, Jingyi Ni, Shiyong He, Fu Zhao, Kunyu Chen, Arnold Chen, and Junyu Cao. 2025b. Litewebagent: The open-source suite for vlm-based web-agent applications. *arXiv preprint arXiv:2503.02950*.

Danyang Zhang, Situo Zhang, Ziyue Yang, Zichen Zhu, Zihan Zhao, Ruisheng Cao, Lu Chen, and Kai Yu. 2025c. Progrm: Build better gui agents with progress rewards. *arXiv preprint arXiv:2505.18121*.

Jiwen Zhang, Jihao Wu, Yihua Teng, Minghui Liao, Nuo Xu, Xiao Xiao, Zhongyu Wei, and Duyu Tang. 2024. Android in the zoo: Chain-of-action-thought for gui agents. *arXiv preprint arXiv:2403.02713*.

Li Zhang, Longxi Gao, and Mengwei Xu. 2025d. Does chain-of-thought reasoning help mobile gui agent? an empirical study. *arXiv preprint arXiv:2503.16788*.

Di Zhao, Longhui Ma, Siwei Wang, Miao Wang, and Zhao Lv. 2025. Cola: A scalable multi-agent framework for windows ui task automation. *arXiv preprint arXiv:2503.09263*.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, and 1 others. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

Yuqi Zhou, Sunhao Dai, Shuai Wang, Kaiwen Zhou, Qinglin Jia, and Jun Xu. 2025. Gui-g1: Understanding r1-zero-like training for visual grounding in gui agents. *arXiv preprint arXiv:2505.15810*.

## A Action Set

TVWorld supports 8 discrete button actions. We further introduce a FINISH action to signal task completion, resulting in a total of 9 actions in the action space. The complete action space and their functionalities are detailed in Table 6.

## B Training Setup Details

**Stage I (SFT).** We randomly sample 500 traces from the TCL TV navigation graph, yielding 8,490 training instances (6,490 Geodesic Guidance, 1,000 Detour Reflection, and 1,000 Stagnation Escape). Training uses the official Qwen3-VL codebase (Bai et al., 2025) with DeepSpeed ZeRO-1, a learning rate of  $1 \times 10^{-6}$ , global batch size 64, weight decay 0, and a maximum of 589,824 vision tokens. Training runs for 5 epochs (about 27 GPU hours).

**Stage II (RL).** We sample 1,000 traces, resulting in 17,825 training instances (11,825 Geodesic Guidance, 4,000 Detour Reflection, and 2,000 Stagnation Escape). Reward weights are set to  $\beta_{\text{topo}} = 0.95$  and  $\beta_{\text{form}} = 0.05$ . Training is performed with the VeRL framework (Sheng et al., 2025) and vLLM (Kwon et al., 2023), using a global batch size of 64, rollout size 8, and 600 optimization steps (about 280 GPU hours).

## C Training Data Description

A path from  $u$  to  $g$  is denoted by  $\mathbf{p} = (u_0, a_0, u_1, \dots, u_{L-1}, a_{L-1}, u_L)$  with  $u_0 = u$ ,  $u_L = g$ . Its length is  $\text{len}(\mathbf{p}) = L$ , and  $\Pi(u \rightarrow g)$  denotes the set of all finite paths from  $u$  to  $g$ . We denote the shortest path connected  $u$  and  $g$  as  $\arg \min_{\mathbf{p} \in \Pi(u \rightarrow g)} \text{len}(\mathbf{p})$ .

Our training data graph  $\mathcal{G}$  is collected from TCL TV. We select start–goal pairs  $(u_0, g)$  and build a path,

$$\mathbf{p}^* = (u_0^* = u_0, a_0^*, \dots, u_L^* = g). \quad (1)$$

From each timestep  $t$  on  $\mathbf{p}^*$ , it can form a training sample

$$\xi_t = (S_t, H_t, I), \quad S_t = S(u_t^*), \quad (2)$$

$$H_t = (a_{t-\delta_a}^*, \dots, a_{t-1}^*, S_{t-\delta_S}, \dots, S_{t-1}),$$

where  $S_t$  is the current screenshot,  $H_t$  concatenates the last  $\delta_a$  actions and  $\delta_S$  screenshots, and  $I$  is an instruction specifying the final goal  $g$ .

## D Trace-Specific Topology-Aware Reward Functions

This Appendix section details the trace-specific topology-aware shaping rewards used in Stage II. For a state–action pair  $(u_t, a)$  within a given trace, the reward is computed from the resulting node  $u'$ , reached after taking action  $a$  at  $u_t$ , by comparing the graph-based distances  $d(u', g)$  and  $d(u_t, g)$ . In this way, each reward component reinforces its corresponding topology-aware behavior through goal-directed progress.

For **Geodesic Guidance Traces**, we directly encourage topology-consistent progress by favoring actions that reduce the distance to the goal:

$$R_{\text{geo}}(u_t, a; g) = \begin{cases} 1, & d(u', g) < d(u_t, g), \\ 0.2, & d(u', g) = d(u_t, g), \\ 0, & d(u', g) > d(u_t, g), \end{cases} \quad (3)$$

For **Detour Reflection Traces**, at the revisited node  $u_t$ , we preserve the preference for moving closer to the goal while explicitly discouraging returning to the previously identified detour branch:

$$R_{\text{det}}(u_t, a; g) = \begin{cases} 1, & d(u', g) < d(u_t, g), \\ 0.2, & d(u', g) = d(u_t, g), \\ 0.1, & d(u', g) > d(u_t, g), a \neq a_{\text{far}}, \\ 0, & d(u', g) > d(u_t, g), a = a_{\text{far}}. \end{cases} \quad (4)$$

For **Stagnation Escape Traces**, we focus on the second visit to  $u_t$  following an invalid key press. Let  $a_{\text{inv}}$  denote the stagnating action. The reward penalizes repeating  $a_{\text{inv}}$  while continuing to shape behavior toward goal-directed progress:

$$R_{\text{sta}}(u_t, a; g) = \begin{cases} 1, & d(u', g) < d(u_t, g), \\ 0.2, & d(u', g) = d(u_t, g), a \neq a_{\text{inv}}, \\ 0, & d(u', g) = d(u_t, g), a = a_{\text{inv}}, \\ 0.1, & d(u', g) > d(u_t, g). \end{cases} \quad (5)$$

## E Distance Families for Topological Shaping

This Appendix section formalizes several commonly used graph-distance functions  $d(\cdot, g)$  that capture the notion of topological proximity and can be used to construct topology-aware rewards.

Recall that Eqs. (3-5) reward an action precisely through how it changes the distance to the goal node  $g$ . Therefore, we introduce a few graph-based distance families that can serve as  $d(\cdot, g)$ , using a unified notation throughout. Let the TV navigation graph be a labeled directed multigraph  $\mathcal{G} =$

Table 6: The functionality of different actions in TVWorld.

Action	Functionality
UP	move the focus upward
DOWN	move the focus downward
LEFT	move the focus to the left or return to the parent directory menu
RIGHT	move the focus to the right or enter the highlighted item
OK	confirm the current selection or enter a highlighted item
HOME	return to the home screen
EXIT	exit the current page or return to the parent directory menu
SETTING	open the settings screen
FINISH	indicate that the navigation task is completed

$(\mathcal{V}, \mathcal{E}, \lambda)$  with a transition map  $T : \mathcal{V} \times \mathcal{A} \rightarrow \mathcal{V}$ , and let  $n \triangleq |\mathcal{V}|$ .

Define the adjacency matrix  $A \in \mathbb{R}^{n \times n}$  by

$$A_{uv} \triangleq |\{a : (u, a, v) \in \mathcal{E}\}|,$$

so  $A_{uv}$  counts the number of labeled edges from  $u$  to  $v$ . Let  $A_{\text{rev}} \triangleq A^\top$ , i.e.,  $(A_{\text{rev}})_{uv} = A_{vu}$ . Let  $e_g \in \mathbb{R}^n$  denote the standard basis vector with a 1 at the coordinate corresponding to node  $g$ . In all experiments, the TV state-transition graph is strongly connected. Below are four common distance definitions.

### E.1 Shortest-Path Distance

A natural choice is the directed shortest-path distance

$$d_{\text{sp}}(u, g) \triangleq \min_{\mathbf{p} \in \Pi(u \rightarrow g)} \text{len}(\mathbf{p}),$$

which measures the minimum number of actions required to reach the goal node  $g$  from state  $u$  along directed transitions.

### E.2 Hitting Time

**Definition.** Let

$$D = \text{diag}\left(\sum_v A_{uv}\right), \quad P = D^{-1}A$$

be the forward row-stochastic random-walk matrix that chooses uniformly among feasible labeled edges. Make  $g$  absorbing by replacing row  $g$  of  $P$  with  $e_g^\top$  (so  $P_{gg} = 1$ ,  $P_{gv} = 0$  for  $v \neq g$ ). Let  $\bar{g} \triangleq \mathcal{V} \setminus \{g\}$  and write the block  $Q \triangleq P_{\bar{g}\bar{g}}$  as the  $(n-1) \times (n-1)$  submatrix obtained by deleting the row and column of  $g$  in  $P$ . The hitting-time vector

$h_g \in \mathbb{R}^n$  is the solution to the Dirichlet problem with  $h_g(g) = 0$  and

$$h_g(u) = 1 + \sum_v P_{uv} h_g(v) \text{ for } u \neq g.$$

Equivalently,

$$(I - Q) h_g(\bar{g}) = \mathbf{1}, \quad h_g(g) = 0.$$

We set

$$d_{\text{hit}}(u, g) \triangleq h_g(u).$$

**Interpretation.**  $d_{\text{hit}}(u, g)$  is the expected number of remote steps required by an uninformed random policy to reach  $g$  from  $u$ . It therefore reflects exploration difficulty: narrow funnels, dead ends, and high-branching detours inflate  $d_{\text{hit}}$  even when  $d_{\text{sp}}$  is small (Blum et al., 2020).

### E.3 Soft Shortest-Walk

**Definition.** Let  $A \in \mathbb{R}^{n \times n}$  be the adjacency matrix. For a temperature  $\beta > 0$ , define the discounted adjacency

$$W \triangleq e^{-\beta} A, \quad Z \triangleq (I - W)^{-1} = \sum_{k=0}^{\infty} W^k, \quad (6)$$

where we assume  $\rho(W) < 1$  so the Neumann series converges. We define the soft shortest-walk distance

$$d_{\text{soft}}(u, g) \triangleq -\frac{1}{\beta} \log Z_{ug}. \quad (7)$$

Since  $(W^k)_{ug} = e^{-\beta k} (A^k)_{ug}$ , we have

$$Z_{ug} = \sum_{k \geq 0} e^{-\beta k} (A^k)_{ug} = \sum_{\pi: u \rightarrow g} \exp(-\beta |\pi|), \quad (8)$$

where  $\pi$  ranges over all (action-labeled) walks from  $u$  to  $g$  and  $|\pi|$  is its length. Therefore,  $d_{\text{soft}}(u, g)$  can be viewed as a log-sum-exp relaxation of the shortest-walk length over all walks.

**Interpretation.** Let  $m \triangleq \min\{k \geq 1 : (A^k)_{ug} > 0\}$  be the shortest-walk length from  $u$  to  $g$ . Then,

$$d_{\text{soft}}(u, g) = m - \frac{1}{\beta} \log\left(\sum_{t \geq 0} (A^{m+t})_{ug} e^{-\beta t}\right) \leq m. \quad (9)$$

Walks that are  $k$  steps longer receive at most a relative weight  $e^{-\beta k}$ , so sufficiently longer walks are exponentially suppressed. In particular, if there are  $N_m = (A^m)_{ug}$  shortest walks and longer walks contribute little, then  $d_{\text{soft}}(u, g) \approx m - \frac{1}{\beta} \log N_m$ . Moreover, increasing any entry of  $A$  (e.g., adding edges or increasing counts) can only increase  $Z_{ug}$ , and thus can only decrease  $d_{\text{soft}}(u, g)$ . Finally, since  $(A^k)_{ug} > 0$  iff there exists a directed walk of length  $k$  from  $u$  to  $g$ ,  $m$  equals the directed shortest-path length from  $u$  to  $g$ . Hence  $d_{\text{soft}}$  can be interpreted as a soft version of the shortest path: as  $\beta \rightarrow \infty$ ,  $d_{\text{soft}}(u, g) \rightarrow m$ .

#### E.4 Personalized PageRank

**Definition.** Construct the forward row-stochastic random-walk matrix  $P = D^{-1}A$  from the adjacency  $A$ . For each seed  $u$ , the personalized PageRank (PPR) vector  $p_u \in \mathbb{R}^n$  solves

$$p_u = \alpha e_u + (1 - \alpha) P^\top p_u$$

where  $\alpha \in (0, 1)$  is the restart probability. We then define the forward PPR distance to target  $g$  by

$$d_{\text{ppr}}(u, g) \triangleq 1 - p_u(g).$$

**Interpretation.**  $p_u$  denotes the stationary visit distribution of a random walk on the forward graph that, at each step, returns to the current-state seed  $u$  with probability  $\alpha$ . Consequently,  $p_u$  can be interpreted as an exponentially discounted combination of the  $t$ -step walk distributions originating from  $u$ , where the contribution of longer walks decays exponentially. In this way,  $p_u(g)$  reflects the long-run visitation frequency of node  $g$  under an uninformed exploration process rooted at  $u$ , making it a measure of proximity (Tong et al., 2006; Andersen et al., 2006). We then define the distance  $d_{\text{ppr}}(u, g) = 1 - p_u(g)$ , so that nodes that are visited more often are regarded as closer.

## F Group-Relative Policy Optimization

This appendix specifies the Group-Relative Policy Optimization (GRPO) objective used in the second stage of Topology-Aware Training (Sec. 4.2), where the agent is rewarded by topology-aware rewards derived from the TV navigation graph (Sec. 4.2.2).

### Structured generation and executable interface.

For each training example, the agent takes  $\xi_t = (S_t, H_t, I)$  as input (defined in Appendix C) and generates a response that includes both a rationale and a single executable remote-control key:

$$\begin{aligned} r_t &= z_t \langle \text{answer} \rangle a_t \langle / \text{answer} \rangle \\ r_t &\sim \pi_\theta(r_t | S_t, H_t, I). \end{aligned} \quad (10)$$

The action token  $a_t$  is then executed in the environment. Concretely, a response is considered well-formed if (i) the tags are balanced, (ii) there is exactly one  $\langle \text{answer} \rangle$  span. This makes output validity a learnable preference signal during training, without hard-coding a constrained decoder at test time.

**Sampling a group and computing rewards.** Fix the context  $(S_t, H_t, I)$  at node  $u_t^*$ . GRPO generates a group of  $K$  candidate responses  $\{r^{(k)}\}_{k=1}^K \sim \pi_\theta$ , extracts their actions  $\{a^{(k)}\}$ , and applies one environment step transition to get  $u^{(k)}$ . Each candidate is assigned two reward components:

$$R_{\text{topo}}^{(k)} = R_{\text{topo}}(u_t^*, a^{(k)}; g), \quad R_{\text{form}}^{(k)} \in \{0, 1\}, \quad (11)$$

where  $R_{\text{topo}}$  is the topology-aware shaping reward defined in Appendix Sec. D (with trace-specific instantiations such as  $R_{\text{geo}}, R_{\text{det}}, R_{\text{sta}}$ ), and  $R_{\text{form}}^{(k)} = 1$  iff  $r^{(k)}$  is well-formed (balanced tags, exactly one  $\langle \text{answer} \rangle$ ).

We combine them into a single scalar score,

$$R^{(k)} = \beta_{\text{topo}} R_{\text{topo}}^{(k)} + \beta_{\text{form}} R_{\text{form}}^{(k)}, \quad (12)$$

so that the policy is simultaneously encouraged to (i) take keys that make measurable progress on the TV graph and (ii) emit reliably executable outputs.

**Group-relative advantages.** Unlike value-based methods, GRPO normalizes scores within the sampled group for the same context:

$$A^{(k)} = \frac{R^{(k)} - \mu}{\sigma}, \quad (13)$$

where  $\mu$  and  $\sigma$  are the mean and standard deviation of  $\{R^{(k)}\}_{k=1}^K$ . This turns raw rewards into a scale-free, variance-reduced advantage: candidates are compared only against their peers under the same  $(S_t, H_t, I)$ , which is precisely what we need when supervision comes from graph transitions rather than a single reference action.

**GRPO objective loss.** Let the  $k$ -th response be tokenized as  $\{r_\ell^{(k)}\}_{\ell=1}^{|r^{(k)}|}$ . Define the per-token importance ratio

$$v_\ell^{(k)} = \frac{\pi_\theta(r_\ell^{(k)} | S_t, H_t, I, r_{<\ell}^{(k)})}{\pi_{\theta_{\text{old}}}(r_\ell^{(k)} | S_t, H_t, I, r_{<\ell}^{(k)})}.$$

GRPO optimizes a clipped surrogate and regularizes the policy toward a reference model  $\pi_{\text{ref}}$  to prevent uncontrolled drift:

$$\bar{v}_\ell^{(k)} \triangleq \text{clip}(v_\ell^{(k)}, 1 - \epsilon, 1 + \epsilon). \quad (14)$$

$$s_\ell^{(k)} \triangleq \min(v_\ell^{(k)} A^{(k)}, \bar{v}_\ell^{(k)} A^{(k)}). \quad (15)$$

$$\mathcal{L}_{\text{GRPO}} = \mathbb{E}_{\substack{\xi \sim \mathcal{D} \\ r^{(k)} \sim \pi_\theta(\cdot | \xi)}} \frac{1}{K} \sum_k \frac{1}{|r^{(k)}|} \sum_\ell \{[s_\ell^{(k)}] - \lambda_{\text{KL}} \text{KL}(\pi_\theta \| \pi_{\text{ref}})\}. \quad (16)$$

The  $A^{(k)}$  is computed once for each sampled response and then applied to all of its tokens via the importance ratio. As a result, the update rewards full generations that (a) stay executable under our parser and (b) produce a one-step transition that enhances the topology-based progress signal. This aligns with the remote-control scenario: what ultimately matters is generating a valid key at every step and making steady progress along the latent UI graph, while still permitting diverse natural-language rationales during training.

## G Comparison of TV UI Layouts

Fig. 6 provides a qualitative comparison of the UI layouts of Google TV and TCL TV. The two interfaces differ notably in icon appearance, overall layout aesthetics, and menu structure. This cross-platform variation naturally leads to a distribution shift in UI states, making it a suitable scenario for evaluating out-of-domain generalization in TV navigation.

## H More experiments

**Effect of Different Rationale Types.** As described in Sec. 4.1, we employ three types of step-level rationales during Stage I training. Table 7 reports an ablation study on text-based navigation tasks in TVWorld-N. Removing rationale supervision yields the lowest success rate of 36.8 (experiment (1)). Adding Geodesic Guidance alone improves performance to 42.4 (experiment (2)), while further incorporating Detour Reflection or Stagnation Escape leads to consistent gains (experiments (3)–(4)). The best performance is achieved when all three rationale types are combined (experiment (5), 46.8). These results indicate that the three rationale types are complementary, with stagnation handling playing a particularly important role in remote-control TV navigation.

Table 7: Ablation of Stage I Rationale Types for Text-Based Instructions on TVWorld-N. GG, DR, and SE denote Geodesic Guidance, Detour Reflection, and Stagnation Escape, respectively.

	Rationale Types	SR
(1)	-	36.8
(2)	GG	42.4
(3)	GG & DR	42.8
(4)	GG & SE	46.4
(5)	GG & DR & SE	<b>46.8</b>

**Effect of Different Distance Metrics.** We perform an ablation over various distance metrics in Stage II to examine their influence on training, as reported in Table 8. Shortest-path, Soft Shortest-Walk, and Hitting Time yield comparable SR, with Hitting Time performing marginally better, while Personalized PageRank (PPR) trails substantially. We attribute Hitting Time’s modest advantage to its definition as the expected first-arrival time under random walks, which yields a more globally informative topological signal than metrics based solely on shortest paths. By contrast, PPR incorporates a restart mechanism that effectively assumes a certain probability of “teleportation” back to the starting point, a behavior that does not align with TV UI interaction patterns (e.g., after pressing HOME from a deep page, returning to the same deep state is often non-trivial); as a result, the induced distance signal conflicts with the TV interaction logic, leading to a pronounced drop in performance.

Table 8: Ablation of the distance metric in the Stage II on TVWorld-N.

Distance Metric	SR	
	text-based	vision-based
Shortest-path	66.4	67.2
Soft Shortest-Walk	65.6	68.4
Personalized PageRank	60.8	59.2
Hitting Time	<b>67.2</b>	<b>68.8</b>

**Effect of Topology-Aware Reward Design.** As introduced in Sec. 4.2, we employ trace-specific reward designs in Stage II, assigning different reward functions to different trace types. The detailed formulations are provided in Appendix D. As a baseline, we adopt the reward design used for Geodesic Guidance traces as a *standard reward*, denoted as  $R_{\text{std}}(u_t, a; g)$ , which assigns rewards solely based on changes in the distance to the goal:

$$R_{\text{std}}(u_t, a; g) = \begin{cases} 1, & d(u', g) < d(u_t, g), \\ 0.2, & d(u', g) = d(u_t, g), \\ 0, & d(u', g) > d(u_t, g), \end{cases} \quad (17)$$

Table 9 compares the standard reward with our proposed topology-aware reward. Across both text- and vision-based instructions, topology-aware rewards consistently yield higher success rates. This result indicates that incorporating trace-specific topology signals provides more fine-grained reward guidance, enabling the agent to learn stronger topology-aware navigation behaviors.

Table 9: Ablation of the reward design in the Stage II on TVWorld-N.

Reward design	SR	
	text-based	vision-based
standard reward	64.0	67.6
topology-aware reward	<b>67.2</b>	<b>68.8</b>

**Effect of the Per-Image Visual Token Limit.** In our default setting, each input image is resized to  $1024 \times 576$ , corresponding to 576 visual tokens after processing by our model. Fig. 4 reports model performance under different per-image visual token limits. Increasing the token limit from 288 to 576 yields a substantial performance gain for both text-based and vision-based instructions, indicating that sufficient visual capacity is crucial for capturing salient UI details. Beyond this point, further increasing the token limit brings slight degradation, suggesting diminishing returns and potential noise introduced by overly fine-grained visual representations.

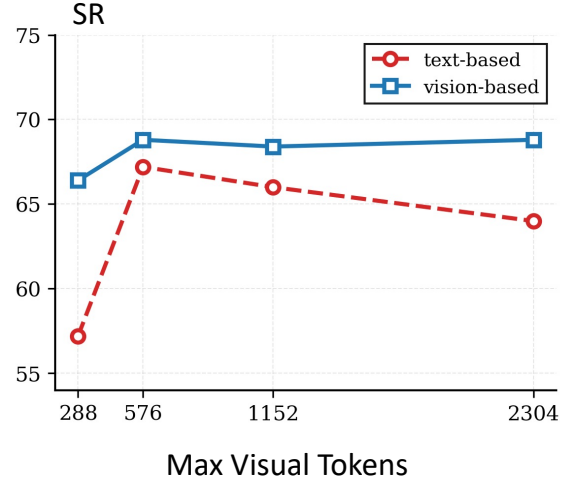


Figure 4: Model performance on TVWorld-N under different visual token budgets.

**Effect of the Number of Historical Screenshots.** TV navigation inherently involves long-horizon interactions, while screenshots introduce a substantial number of visual tokens, making it impractical to retain all historical screenshots as model input. In our default setting, we retain 4 historical screenshots. Fig. 5 illustrates the effect of varying the number of historical screenshots on model performance. We observe that using 4 historical screenshots yields the best performance, while both increasing and decreasing this number lead to performance degradation.

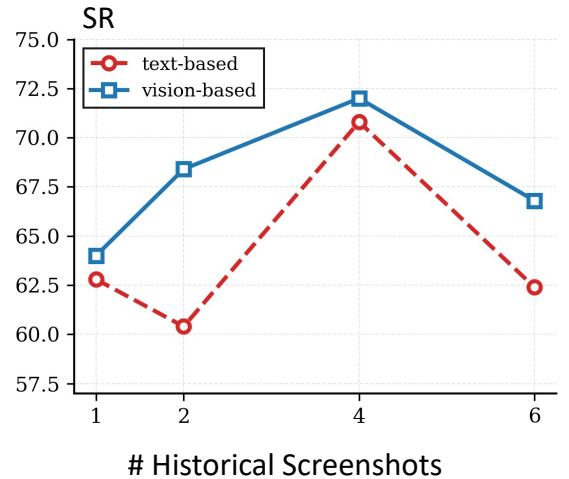


Figure 5: Model performance on TVWorld-N under different numbers of historical screenshots.

**Detailed Results on TVWorld-N.** Table 10 reports the detailed results of TVTheseus on TVWorld-N.

Table 10: Detailed results of 3 independent runs of TVTheseus on TVWorld-N.

Setting	Instr Type	Config	Display	Audio	Apps	Channels	Overall
Turn 1	text-based	36.0	64.0	88.0	76.0	72.0	67.2
	vision-based	64.0	54.0	64.0	84.0	78.0	68.8
Turn 2	text-based	40.0	52.0	88.0	68.0	72.0	64.0
	vision-based	58.0	64.0	48.0	82.0	84.0	67.2
Turn 3	text-based	48.0	72.0	82.0	68.0	84.0	70.8
	vision-based	66.0	64.0	62.0	82.0	86.0	72.0
Avg.	text-based	41.3	62.7	86.0	70.7	76.0	67.3
	vision-based	62.7	60.7	58.0	82.7	82.7	69.3

## I Case study

**TVWorld Environment.** Fig. 7 shows a case study on focus-based TV navigation for the instruction “Go to External Inputs–HDMI 3,” comparing TVTheseus (topology-aware trained) with Qwen3-VL-8B-Instruct (untrained baseline) from the same initial UI state. TVTheseus plans a coherent sequence that navigates the settings hierarchy into External Inputs, shifts focus step-by-step to HDMI 3, and ends with FINISH; when a key press causes no state change, it adapts by trying alternative actions instead of repeating the ineffective one, demonstrating reliable topology-aware planning. In contrast, the untrained model repeatedly issues actions with no transitions, stays near the initial state, and shows limited understanding of focus-based UI dynamics and global planning.

**Physical Environment.** Fig. 8 showcases a case study of TVTheseus on a physical Google TV. Notably, TVTheseus completes the task via a sequence of 17 real-world interactions, thereby validating its transferability to physical environments.

## J Training Data Example

We present three categories of topology-priming SFT training data: Geodesic Guidance in Fig. 9, Detour Reflection in Fig. 10, and Stagnation Escape in Fig. 11. Each instance contains chain-of-thought reasoning, records of past actions, and both historical and current page screenshots.

## K Responsible NLP Research Considerations

### K.1 Potential Risks

If deployed on real devices, agents trained with TVWorld could be exploited to automatically ac-

cess and change privacy- or account-related settings (such as permissions, parental controls, or password options), potentially causing privacy or security harms. We partly reduce this risk by masking a limited set of sensitive entry points during graph construction and release, and we advise using access controls and explicit user confirmation for any deployment on real devices.

### K.2 Intended Use & Artifact Use

TVWorld (and TVWorld-N/TVWorld-G) and TVTheseus are intended for research on focus-based remote-control TV navigation, including controlled training and benchmarking of agents in an offline, replayable environment. They are not intended for deployment on unauthorized control of devices, or attempts to access restricted system functions. We use existing models and tools strictly in accordance with their intended research/benchmarking usage and applicable terms; we do not provide the system with any personal user data. We recommend that any derivatives of the released assets remain limited to research contexts.

### K.3 AI Assistants Elaboration

In this work, we employed AI assistants strictly as supporting tools for tasks including grammar correction, language refinement, and logo image generation. The authors thoroughly evaluated and revised all outputs provided by these tools and retain complete responsibility for the accuracy, integrity, and content of the final manuscript.

### K.4 Ethics and Reproducibility Statement

We study offline, replayable TV-navigation agents using TVWorld/TVTheseus, constructing static graphs and screenshots to support reproducible

evaluation. We have checked that the collected/used data do not contain any personally identifiable information, including identifiable personal names, and do not include any private or sensitive user information. All external datasets, models, and tools used in this work are properly cited and employed in full compliance with their licenses, terms, and intended-use policies. As such, we do not anticipate potential ethical risks arising from the dataset or experimental protocol. To further support transparency and reproducibility, we provide the codebase and a subset of data resources as illustrative examples, and we will open-source the full dataset and related resources on public platforms such as GitHub and/or Hugging Face under the CC BY 4.0 license.

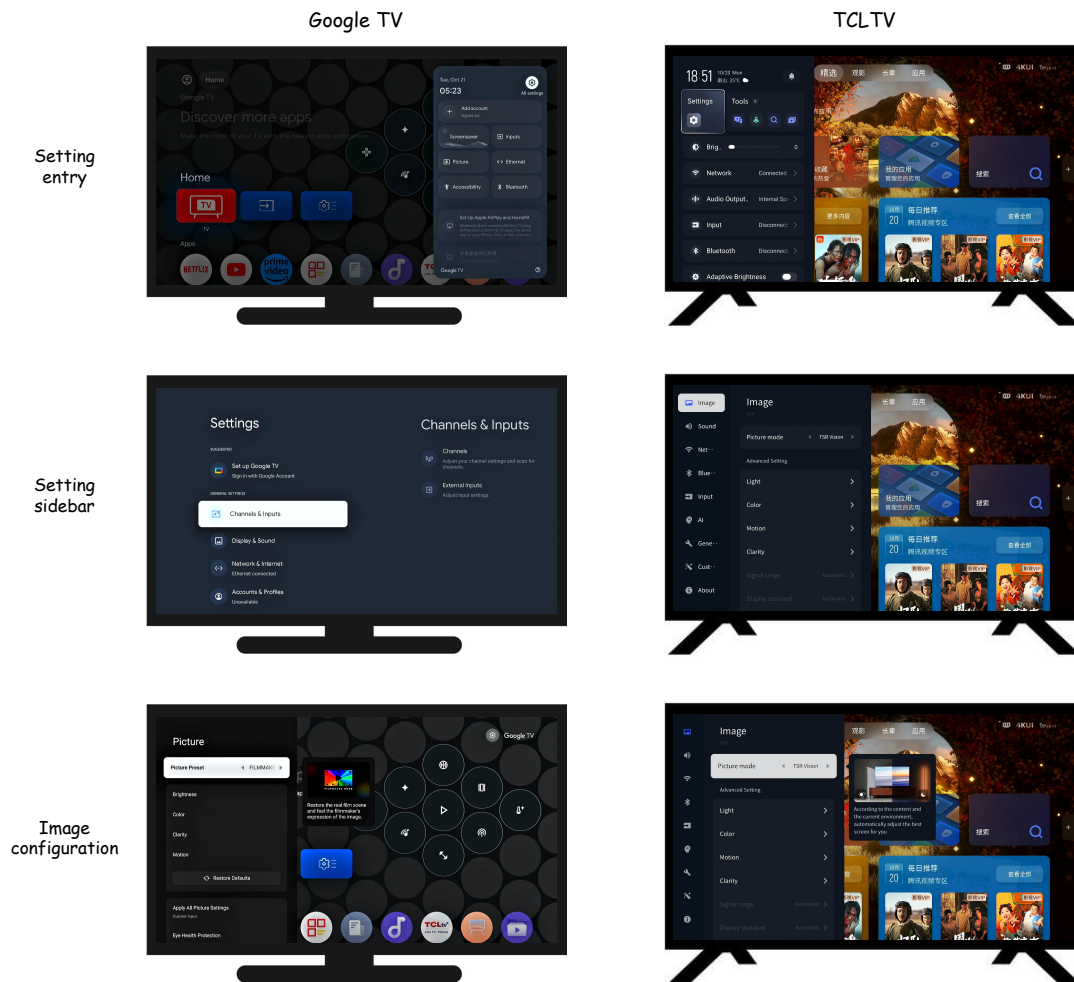


Figure 6: Comparison of UI style between two different TV models: Google TV and TCL TV.

Task: Go to External Inputs-HDMI 3



The user instruction targets 'External Inputs-HDMI 3', which is unrelated to the current 'Brightness' settings menu, indicating the need to leave this specific submenu. Since the current focus is deep within the brightness options (Micro Contrast), pressing EXIT is the most direct way to backtrack out of this menu structure and navigate towards the main input settings. [answer=EXIT/answer](#)

The user is currently on the Home screen with the focus on the TV icon, but the target destination is within the 'Settings' menu structure. Since the 'Settings' option (represented by the gear icon) is located in the middle-right area of the current grid layout, pressing DOWN is necessary to navigate towards it. [answer=DOWN/answer](#)

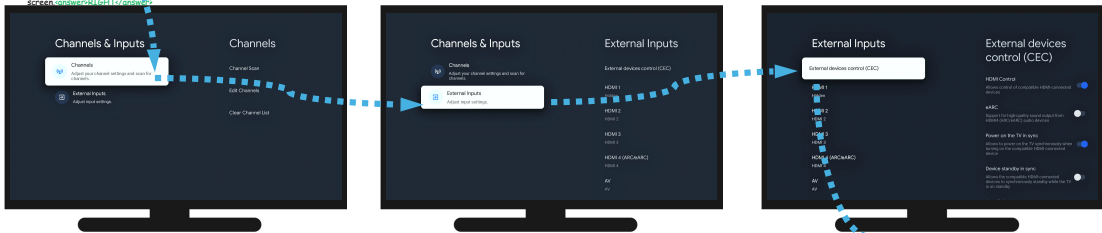
The previous DOWN press did not change the screen, indicating it was ineffective for navigating away from the current Home launcher view. Since the target 'External Inputs' is likely located within the system settings menu rather than the app grid, pressing SET is a reasonable next move to open the general configuration options. [answer=SET/answer](#)



The current focus is on 'Channels & Inputs' in the left sidebar of the Settings menu, which needs to be entered to access the 'External Inputs' section visible on the right. Pressing RIGHT will move the focus from the sidebar into the detailed options panel on the right side of the screen. [answer=RIGHT/answer](#)

The previous RIGHT press did not change the screen, indicating it was ineffective for navigating this overlay. Since the 'Inputs' option is visible within the settings panel, pressing OK is a reasonable next move to enter the Inputs menu and progress toward selecting HDMI. [answer=OK/answer](#)

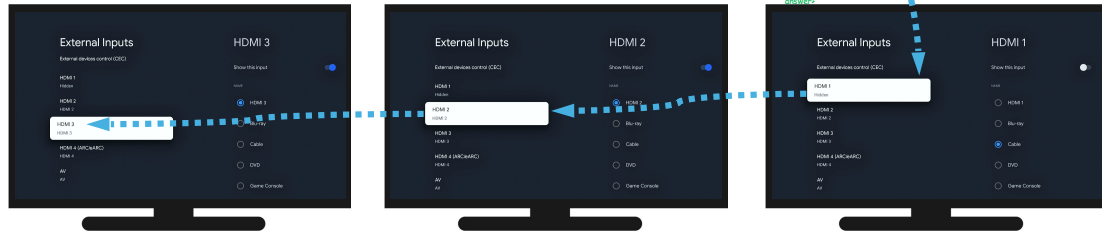
The current screen displays the 'All settings' quick menu overlay, which contains direct access to system configurations. The target destination is 'External Inputs-HDMI 3', and the 'Inputs' option is clearly visible within this menu, making it the logical next step to select. [answer=RIGHT/answer](#)



The current focus is on 'Channels' within the 'Channels & Inputs' menu, but the target destination is 'External Inputs', which is located directly below the current selection. Pressing DOWN will move the focus to 'External Inputs', allowing access to the HDMI settings. [answer=DOWN/answer](#)

The current focus is on 'External Inputs' which contains the target options like 'HDMI 3'. Pressing RIGHT will move the focus into the right-hand list where 'HDMI 3' is located, allowing the user to subsequently select it. [answer=RIGHT/answer](#)

The user wants to navigate to 'HDMI 3' within the 'External Inputs' menu, but the current focus is on 'External devices control (CEC)' at the top of the list. Since the target 'HDMI 3' is located further down the list, pressing DOWN is required to give the focus towards it. [answer=DOWN/answer](#)



The current focus is on 'HDMI 3' within the 'External Inputs' menu, which matches the target destination. Since the desired item is already highlighted, pressing FINISH confirms that the navigation goal has been successfully reached. [answer=FINISH/answer](#)

The current focus is on 'HDMI 2' within the External Inputs menu, but the target destination is 'HDMI 3'. Since 'HDMI 3' is located directly below the currently selected item in the list, pressing DOWN will move the focus to the correct target. [answer=DOWN/answer](#)

The current focus is on 'HDMI 1' in the External Inputs menu, but the target destination is 'HDMI 3'. Since 'HDMI 3' is located further down the list, pressing DOWN is required to navigate through the items towards the target selection. [answer=DOWN/answer](#)

Qwen3-VL-8B-Instruct



To navigate from 'Settings-Display & Sound-Picture-Brightness-Micro Contrast' to 'External Inputs-HDMI 3', I need to exit the current brightness settings page and go back to the main settings menu. The 'Micro Contrast' option is currently selected, so pressing the BACK button (which is represented by the LEFT button in this context) will take me out of this submenu and closer to the main settings menu where I can find the 'External Inputs' section. [answer=LEFT/answer](#)

[answer=LEFT/answer](#)

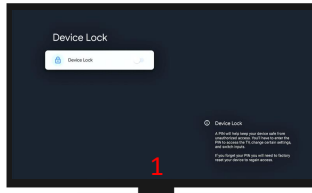
[answer=LEFT/answer](#)

Figure 7: Case study comparing TV navigation trajectories before (Qwen3-VL-8B-Instruct) and after topology-aware training (TVTheseus).

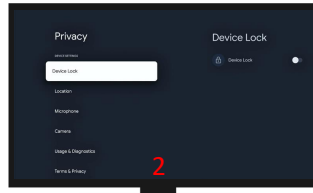
# Task: Go to HDMI 1-Game Console



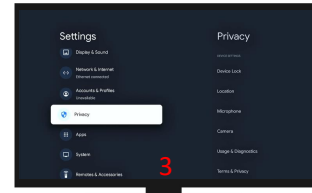
TVTheusis



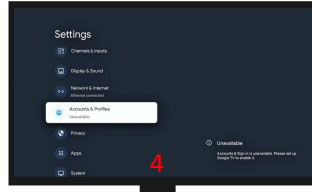
The current screen displays the "Device Lock" toggle switch, which is an unrelated setting for the target destination "HDMI 1-Game Console". Pressing LEFT will move the focus out of this specific option or navigate back to the main menu structure where general display or input settings can be found. [answer=LEFT/answer](#)



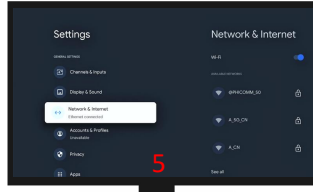
The current screen is the "Privacy" menu where the focus is on "Device Lock" but the target "HDMI 1-Game Console" is likely located in a different main category such as "Display" or "Input". Therefore, pressing LEFT is necessary to navigate out of the current sub-menu back to the parent settings menu where other options can be selected. [answer=LEFT/answer](#)



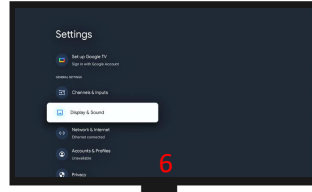
The current focus is on "Privacy" within the main Settings menu, but the target destination "HDMI 1-Game Console" is likely located under an option like "Display & Sound" or "System". Since "Display & Sound" is visible directly above the current selection, pressing UP is necessary to navigate towards that category. [answer=UP/answer](#)



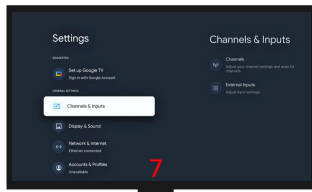
The target "HDMI 1" is likely located within the "Channels & Inputs" section, which is positioned above the currently focused "Accounts & Profiles" item. Therefore, pressing UP is necessary to navigate towards the "Channels & Inputs" menu option. [answer=UP/answer](#)



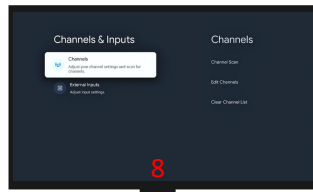
The current focus is on "Network & Internet", but the target destination involves HDMI connections which are typically found under "Channels & Inputs". Since "Channels & Inputs" is located directly above the current selection in the menu list, pressing UP will shift the focus to the correct category. [answer=UP/answer](#)



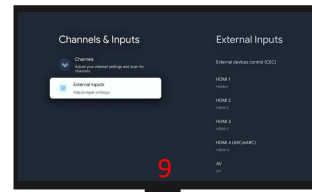
The target "HDMI 1-Game Console" is typically located within the "Channels & Inputs" menu, which is positioned above the currently focused "Display & Sound" option. Therefore, pressing UP is necessary to move the focus from "Display & Sound" to "Channels & Inputs". [answer=UP/answer](#)



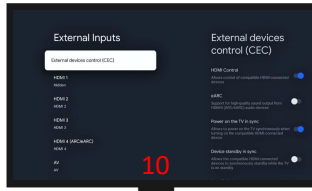
The current focus is on "Channels & Inputs" in the Settings menu, and the right panel displays its sub-options including "External Inputs". Pressing RIGHT will move focus into this menu area, allowing access to input-specific settings where "HDMI 1" can be selected. [answer=RIGHT/answer](#)



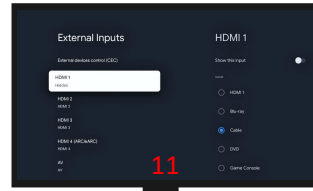
The current focus is on "Channels" within the "Channels & Inputs" menu, but the target "HDMI 1" is located under the "External Inputs" section. Therefore, pressing DOWN is required to move the focus from "Channels" to "External Inputs" to access the relevant input settings. [answer=DOWN/answer](#)



The current focus is on "External Inputs" within the "Channels & Inputs" menu, and the target "HDMI 1" is visible in the submenu to the right. Pressing RIGHT will move the focus from the main menu item into the list of external inputs, allowing subsequent selection of "HDMI 1". [answer=RIGHT/answer](#)



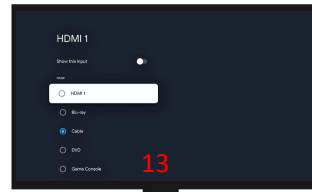
The current focus is on "External devices control (CEC)" within the "External Inputs" menu, but the target "HDMI 1" is the first item in the list below it. Pressing DOWN will move the selection from CEC to "HDMI 1", which is the specific entry requested by the user instruction. [answer=DOWN/answer](#)



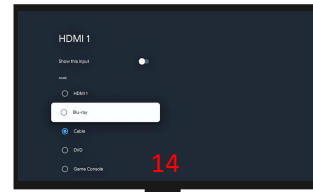
The current focus is on "HDMI 1" in the left sidebar, and the right panel displays the configuration options for the currently selected input. To change the setting from "Cable" to "Game Console", the user must first navigate into the input selection area, which requires moving the focus RIGHT. [answer=RIGHT/answer](#)



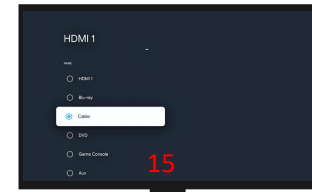
The current focus is on the "Show this input" toggle at the top of the HDMI 1 settings, while the target option "Game Console" is located near the bottom of the list. Navigating down allows the user to cycle through the list items and reach "Game Console". Therefore, pressing DOWN is the correct action to move towards the target selection. [answer=DOWN/answer](#)



Since the target "Game Console" is located further down in the list, pressing DOWN is a reasonable next move to shift the focus away from the currently selected item and towards the desired option. [answer=DOWN/answer](#)



The current focus is on "Blu-ray" within the HDMI 1 settings menu, and the target "Game Console" is located further down the list. Therefore, pressing DOWN is required to navigate through the options towards the destination. [answer=DOWN/answer](#)



The current focus is on "Cable" within the HDMI 1 menu, and the target option "Game Console" is located two items below the current selection. Therefore, pressing DOWN is necessary to move the focus towards the target item. [answer=DOWN/answer](#)



The current focus is on "DVD" within the HDMI 1 menu, and the target item "Game Console" is located directly below it. Pressing DOWN will move the selection from "DVD" to "Game Console", positioning the user at the desired destination. [answer=DOWN/answer](#)



The current screen shows the "HDMI 1" menu with the "Game Console" option already highlighted, which matches the final target of the user instruction. Since the desired item is selected and focused, the correct action to indicate completion of the task is FINISH. [answer=FINISH/answer](#)

Figure 8: An illustration of TVTheusis controlling a physical TV set.

## Geodesic Guidance Data Example

```
[
  {
    "conversations": [
      {
        "from": "human",
        "value": "You are a TV GUI agent. You are assigned a navigation task where you need to locate a specific page. Given your action history and the corresponding screenshots, determine the next remote control input required to accomplish the task.\n\n### Available Remote Control Buttons\nUP # Move the focus upward\nDOWN # Move the focus downward\nLEFT # Move the focus to the left or return to the parent directory menu\nRIGHT # Move the focus to the right or enter the highlighted item\nOK # Confirm the current selection or enter a highlighted item\nHOME # Return to the home screen\nEXIT # Exit the current page or return to the parent directory menu\nSETTING # Open the settings screen\nFINISH # Indicate that the navigation task is completed\n\n### Note\n- For each step, choose one button from the list of available remote control buttons to execute.\n- You may include your reasoning, but your final chosen button must be wrapped in <answer></answer>. e.g., <answer>UP</answer>.\n\n### User Instruction\nSwitch to Settings-Bluetooth-Second device of the Available device.\n",
        "mask": 0
      },
      {
        "from": "human",
        "value": "<image>",
        "mask": 0
      },
      {
        "from": "gpt",
        "value": "<answer>HOME</answer>",
        "mask": 0
      },
      {
        "from": "human",
        "value": "<image>",
        "mask": 0
      },
      {
        "from": "gpt",
        "value": "<answer>SETTING</answer>",
        "mask": 0
      },
      {
        "from": "human",
        "value": "<image>",
        "mask": 0
      },
      {
        "from": "gpt",
        "value": "The current focus is on the \"Settings\" icon (gear symbol) within the quick settings overlay, which serves as the gateway to the full system configuration. Pressing OK will open the main Settings menu, allowing access to the Bluetooth options required by the user instruction.<answer>OK</answer>",
        "mask": 1
      }
    ],
    "image": [
      "/image/screenshot_20251021000906787.png",
      "/image/screenshot_20251020185130340.png",
      "/image/screenshot_20251020185144908.png"
    ],
    "type": ""
  }
]
```

Figure 9: Example training data for Geodesic Guidance Traces.

## Detour Reflection Data Example

```
[
  {
    "conversations": [
      {
        "from": "human",
        "value": "You are a TV GUI agent. You are assigned a navigation task where you need to locate a specific page. Given your action history and the corresponding screenshots, determine the next remote control input required to accomplish the task.\n\n### Available Remote Control Buttons\nUP # Move the focus upward\nDOWN # Move the focus downward\nLEFT # Move the focus to the left or return to the parent directory menu\nRIGHT # Move the focus to the right or enter the highlighted item\nOK # Confirm the current selection or enter a highlighted item\nHOME # Return to the home screen\nEXIT # Exit the current page or return to the parent directory menu\nSETTING # Open the settings screen\nFINISH # Indicate that the navigation task is completed\n\n### Note\n- For each step, choose one button from the list of available remote control buttons to execute.\n- You may include your reasoning, but your final chosen button must be wrapped in <answer></answer>. e.g., <answer>UP</answer>.\n\n### User Instruction\nStart on Settings-About-TV Name, then navigate to Settings-Image-White Balance-Restore the 2-point default value using the remote buttons.\n",
        "mask": 0
      },
      {
        "from": "human",
        "value": "<image>",
        "mask": 0
      },
      {
        "from": "gpt",
        "value": "<answer>LEFT</answer>",
        "mask": 0
      },
      {
        "from": "human",
        "value": "<image>",
        "mask": 0
      },
      {
        "from": "gpt",
        "value": "<answer>UP</answer>",
        "mask": 0
      },
      {
        "from": "human",
        "value": "<image>",
        "mask": 0
      },
      {
        "from": "gpt",
        "value": "<answer>DOWN</answer>",
        "mask": 0
      },
      {
        "from": "human",
        "value": "<image>",
        "mask": 0
      },
      {
        "from": "gpt",
        "value": "The previous UP action navigated from the About section to the Customization menu, which led away from the target Image settings, and the subsequent DOWN action returned the focus to the About section. Since moving UP previously caused a deviation, repeating that action should be avoided on this screen. Instead, pressing DOWN again is a reasonable move to continue scrolling through the sidebar options in search of the Image settings.<answer>DOWN</answer>",
        "mask": 1
      }
    ],
    "image": [
      "/image/screenshot_20251020185728880.png",
      "/image/screenshot_20251020185234116.png",
      "/image/screenshot_20251020185700845.png",
      "/image/screenshot_20251020185234116.png"
    ],
    "type": "BackNode"
  }
]
```

Figure 10: Example training data for Detour Reflection Traces.

## Stagnation Escape Data Example

```
[
  {
    "conversations": [
      {
        "from": "human",
        "value": "You are a TV GUI agent. You are assigned a navigation task where you need to locate a specific page. Given your action history and the corresponding screenshots, determine the next remote control input required to accomplish the task.\n\n### Available Remote Control Buttons\nUP # Move the focus upward\nDOWN # Move the focus downward\nLEFT # Move the focus to the left or return to the parent directory menu\nRIGHT # Move the focus to the right or enter the highlighted item\nOK # Confirm the current selection or enter a highlighted item\nHOME # Return to the home screen\nEXIT # Exit the current page or return to the parent directory menu\nSETTING # Open the settings screen\nFINISH # Indicate that the navigation task is completed\n\n### Note\n- For each step, choose one button from the list of available remote control buttons to execute.\n- You may include your reasoning, but your final chosen button must be wrapped in <answer></answer>. e.g., <answer>UP</answer>.\n\n### User Instruction\nSwitch to Settings-General-Quick instructions-Red key binding.\n",
        "mask": 0
      },
      {
        "from": "human",
        "value": "<image>",
        "mask": 0
      },
      {
        "from": "gpt",
        "value": "<answer>OK</answer>",
        "mask": 0
      },
      {
        "from": "human",
        "value": "<image>",
        "mask": 0
      },
      {
        "from": "gpt",
        "value": "<answer>RIGHT</answer>",
        "mask": 0
      },
      {
        "from": "human",
        "value": "<image>",
        "mask": 0
      },
      {
        "from": "gpt",
        "value": "The previous RIGHT press did not change the screen, indicating it was ineffective for navigating out of the current \"Image\" settings menu. Since the goal is to reach \"General\" settings, which is a different category in the left sidebar, pressing LEFT is a reasonable next move to return focus to the sidebar so a different category can be selected.<answer>LEFT</answer>",
        "mask": 1
      }
    ],
    "image": [
      "/image/screenshot_20251020224717471.png",
      "/image/screenshot_20251020211652371.png",
      "/image/screenshot_20251020211652371.png"
    ],
    "type": "SelfloopNode"
  }
]
```

Figure 11: Example training data for Stagnation Escape Traces.