SmartAgent: Chain-of-User-Thought for Embodied Personalized Agent in Cyber World

Jiaqi Zhang[1], **Chen Gao**[2*], **Liyuan Zhang**[3], **Quoc Viet Hung Nguyen**[4], **Hongzhi Yin**[1*]

[1]School of Electrical Engineering and Computer Science, The University of Queensland, Brisbane, Australia
[2]BNRist, Tsinghua University, Beijing, China
[3]Yale University, New Haven, CT, USA
[4]Griffith University, Gold Coast, Australia
jiaqi.zhang@uq.edu.au, chgao96@gmail.com, quocviethung.nguyen@griffith.edu.au, h.yin1@uq.edu.au

Abstract

Recent advances in embodied agents with multimodal perception and reasoning capabilities based on large vision-language models (LVLMs), excel in autonomously interacting either real or cyber worlds, helping people make intelligent decisions in complex environments. However, the current works are normally optimized by golden action trajectories or ideal task-oriented solutions toward a definitive goal. This paradigm considers limited user-oriented factors, which could be the reason for their performance reduction in a wide range of personal assistant applications. To address this, we propose **Chain-of-User-Thought (COUT)**, a novel embodied reasoning paradigm that takes a chain of thought from basic action thinking to explicit and implicit personalized preference thought to incorporate personalized factors into autonomous agent learning. The main challenges of achieving COUT include: 1) the definition of embodied personalized tasks, 2) the embodied environment epitomizes personalized preference, and 3) the way to model embodied personalized actions. To target COUT, we introduce **SmartAgent**, an agent framework perceiving cyber environments and reasoning personalized requirements as: 1) interacting with GUI to access an item pool, 2) generating users' explicit requirements implied by previous actions, and 3) recommending items to fulfill users' implicit requirements. To demonstrate SmartAgent's capabilities, we also create a brand-new dataset **SmartSpot** that offers a full-stage personalized action-involved environment. To our best knowledge, our work is the first to formulate the COUT process, serving as a preliminary attempt towards embodied personalized agent learning. Our extensive experiments on SmartSpot illuminate SmartAgent's functionality among a series of embodied and personalized sub-tasks.

Code&Datasets —
https://github.com/tsinghua-fib-lab/SmartAgent

Introduction

Embodied artificial intelligence (Duan et al. 2022) is considered as a crucial stride toward Artificial General Intelligence (AGI) (Duéñez-Guzmán et al. 2023). Powered by the recent advances in large multi-modal models, embodied agents have been built upon to behave like real humans, capable of perceiving, reasoning, and acting with their surroundings in both real and cyber worlds. The enthusiasm for deploying such humanoid capabilities is evident in various tasks, including autonomic robotics (Driess et al. 2023; Barreiros et al. 2022), game AI (Yang et al. 2025; Tan et al. 2024), smart device assistants (Rawles et al. 2024a; Hong et al. 2024), and smart city (Gao et al. 2024a; Xu et al. 2023). Many of these scenarios require embodied agents to do more than follow instructions and execute actions like emotionless robots. At the same time, they are expected to serve as personal assistants attuned to human preferences in the meantime. For instance, for smart device assistance, agents struggle to personalize responses to ambiguous user queries such as providing music recommendations, though they fully understand the operation logic of a music player. Generally speaking, a fully functional embodied agent necessitates personalized perceptual capabilities. Such capabilities enable a comprehensive agent-environment-user triadic perception of the world , much like JARVIS[1], Iron Man's fictional assistant which can deliver a wide range of personalized, intelligent decisions as situations evolve.

However, personalized considerations are largely absent among the current embodied agent works, where the optimization normally relies on golden action trajectories (Rawles et al. 2024b; Deng et al. 2024b) or ideal task-oriented solutions (Drouin et al. 2024; Kim, Baldi, and McAleer 2024). Although these fixed paths can accomplish task goals effectively, they tend to train embodied agents as rigid, task-oriented problem solvers. Such training overlooks the fact that multiple valid paths often exist as indicators of user preference. Furthermore, practical environments may exhibit behaviors that are more difficult to predict, such as when new functions are involved or unseen scenes appear. In these cases, task-oriented agents are insufficiently flexible to capture dynamic changes even in basic task goals (Kim et al. 2022; Srivastava et al. 2022), let alone shifts in user preference. As a result, such training paradigms restrict the learning of user-oriented perceptual

*Corresponding Authors.

[1]Stands for 'Just a Rather Very Intelligent System', a fictional AI character in the Marvel Comics

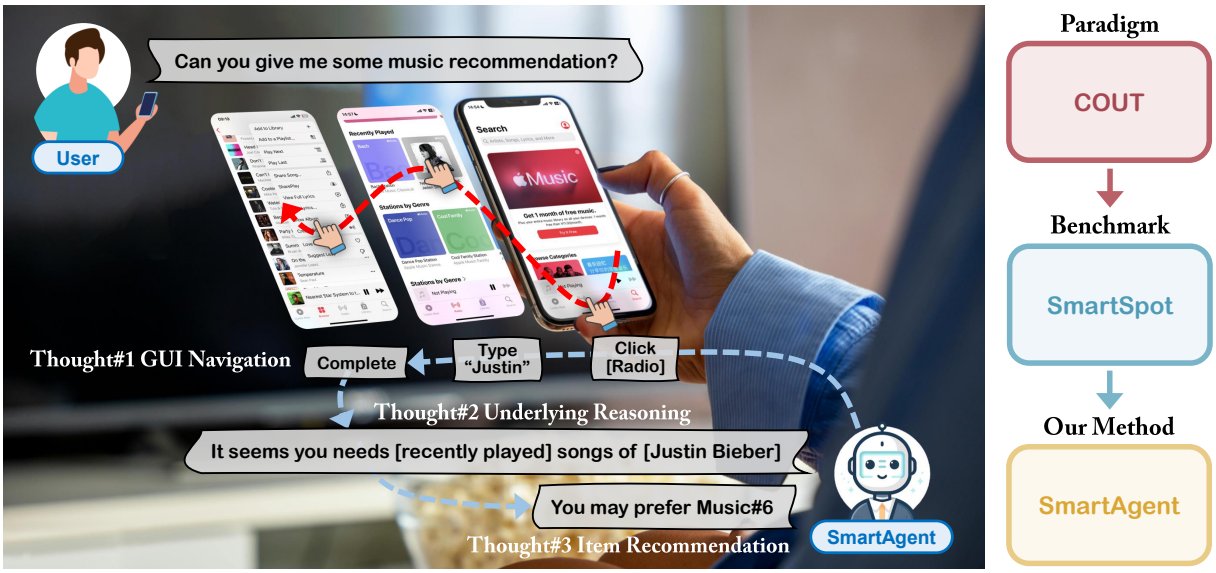

Figure 1: **The Chain-of-User-Thought (COUT) reasoning paradigm.** The red line shows a sequence of GUI actions, while the blue line illustrates our three-step thought process. In Thought #1, according to a user's instruction, an agent performs GUI actions to search for an item pool. In Thought #2, by seeing the pool, the agent reasons underlying requirements behind the original instruction, as implied by the previous actions. In Thought #3, based on the underlying thought, the agent recommends items within the pool to complete the user's instruction. By leveraging user-oriented thoughts, this COUT could enable full-stage embodied personalized capabilities across various information systems.

capabilities, which likely contributes to the lower performance in many embodied personalized scenarios. For example, in the daily usage of smart devices as shown in Figure 1, there is no golden line of actions to access a music collection, i.e., item pools. Instead, users may follow diverse paths that imply different intents (Rawles et al. 2024a). That is, existing embodied agents are typically designed only to complete product-level operations (e.g., controlling a mobile music app). They lack the ability to uncover users' intentions behind their usage habits and, in turn, infer deeper preferences to provide personalized services.

In this work, we propose a novel reasoning paradigm **Chain-of-User-Thought (COUT)**, which *extends embodied agents from task-only optimization to personalization-oriented optimization*. We summarize COUT process as: training agents in **a progression of thinking from basic embodied behaviors, gradually to explicit requirements reasoning, and finally to high-level implicit personalization understanding**, as shown in Figure 1. Notably, COUT could be a general paradigm that can be flexibly adapted to various cyber environments to enable personalized reasoning. However, several critical challenges remain in achieving COUT, as follows:

- First, the learning task of embodied personalized agents has not been systematically defined, because the task goals are often ambiguous queries from users, which go beyond the existing works that take explicit task instructions.
- Second, there is a lack of suitable datasets and benchmarks to support research on COUT. The commonly used datasets generally do not include personalized features.
- Third, the modeling of personalized features is underexplored, largely due to the absence of clear task definition and supportive training environments.

To address these challenges, we construct **SmartSpot**, the first embodied AI benchmark with explicit personalization-related evaluations. SmartSpot comprises five single channels and two multi-channel scenarios to simulate complex real-world environments, featuring a total of 144 episodes and over 1,400 steps. Building upon SmartSpot, we further design **SmartAgent**, the first embodied personalized agent. It takes visual observations of GUI screenshots and textual instructions as input and generates multi-step thoughts. Specifically, the SmartAgent undergoes a two-stage training process: *embodiment stage* and *personalization stage*, as illustrated in Figure 3. In the embodiment stage, the agent receives GUI actions and item pool screenshots as visual inputs, together with other textual contexts. After encoding these multimodal tokens, a *Perceiver* module outputs specific GUI actions, regarded as Thought #1. Based on this initial GUI-action thought, a *Reasoner* module infers Thought #2, identifying the user's potential underlying requirements in a concise textual output. Finally, in the personalization stage, powered by the deeper-level Thought #2, the same *Perceiver* outputs the recommendation result either "Yes" or "No" as Thought #3.

We quantitatively evaluate SmartAgent's functionality in comprehensive embodied and personalized sub-tasks, including GUI Grounding, Autonomous GUI Operation, Underlying Reasoning, Personalized Recommendation, and Zero-shot Reasoning, as illustrated in Figure 2. The results indicate (i) through three-step of COUT process with efficient LoRA tuning, SmartAgent achieves the first full-

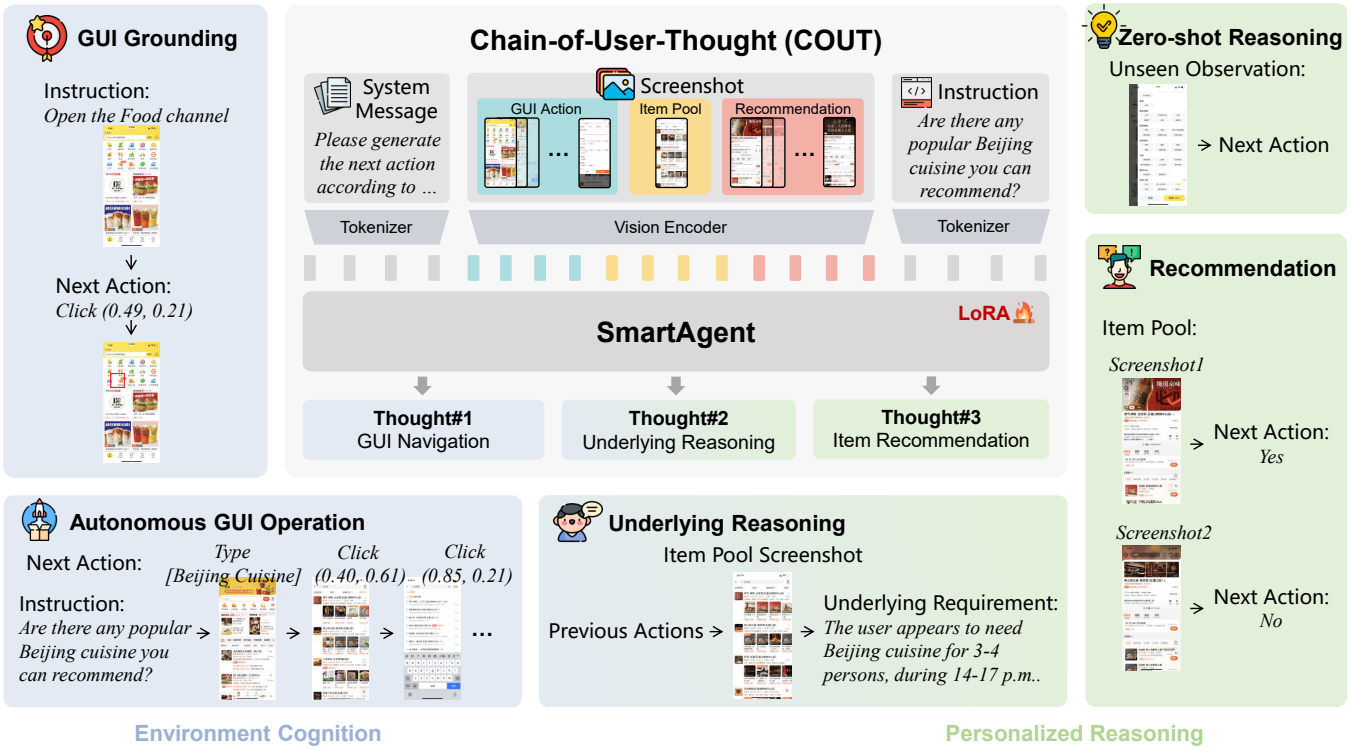

Figure 2: The full-stage embodied personalized capabilities of our proposed SmartAgent, range from basic environment cognition to advanced user personal intention reasoning.

stage embodied personalized reasoning; (ii) SmartAgent delivers comparable performances to state-of-the-art task-specific models on fundamental GUI Grounding and advanced Autonomous GUI Operation tasks, capable of generating accurate action commands; (iii) SmartAgent shows proficiency in reasoning explicit underlying intentions, effectively bridging surface-level operations with implicit-level user needs; (iv) SmartAgent excels at leveraging these preceding thoughts to uncover users' implicit items requirements; (v) SmartAgent manifests zero-shot reasoning capabilities across new channels, a hallmark of a well-established embodied agent. We further present qualitative evaluations to illustrate SmartAgent's effectiveness and proficiency in personal assistant scenarios. The contributions of this work can be summarized as follows:

- We take a pioneering step to formulate a novel reasoning paradigm COUT for embodied AI, which introduces personalized concerns that have not been addressed in previous embodied AI research. To support COUT, we construct the SmartSpot benchmark as the first case environment for further research.
- We develop the first embodied personalized agent SmartAgent, built upon a meticulous two-stage training process that effectively implements the three-step COUT thinking process. SmartAgent is capable of performing a range of tasks, from basic embodied functions to both explicit and implicit-level personalized reasoning.
- Our rigorous experiments demonstrate the outstanding performance of SmartAgent on both overall evaluations and sub-tasks. SmartAgent demonstrates notable embodied personalized reasoning and zero-shot reasoning capabilities.

Chain-of-User-Thought (COUT)

Definition

Chain-of-User-Thought (COUT) is a reasoning paradigm where an agent controls smart devices based on both task goals and user personalized preference as follows:

$$\mathcal{A}_{\text{COUT}} = \{Action_i \mid \textit{Task goal}, \textit{User preference}\}, \quad (1)$$

where $\mathcal{A}_{\text{COUT}}$ is the action space, which differs from the existing work that relies solely on task goal as:

$$\mathcal{A}_{\text{existing work}} = \{Action_i \mid \textit{Task goal}\}. \quad (2)$$

Specifically, an agent is required to generate $Action_i$ through a progressive reasoning chain from *basic embodied behavior level*, gradually to *deeper explicit personalized reasoning level*, and finally to *high implicit personalized reasoning level*. This process requires embodied agents the ability of a multi-faceted understanding of user personalized preferences.

Components

As illustrated in Figure 2, we formulate the COUT process in a common cyber world case, personal assistance of smart devices, in terms of three stages of thoughts: a) Thought #1 GUI Navigation, b) Thought #2 Underlying Reasoning, and c) Thought #3 Personalized Recommendation. Thought

Scenario	Channel	#Episodes	#Steps			Instruction Mean	Underlying Mean
			GUI	Item Pool	RS		
Single-channel	FOOD	20	7.00	1.00	4.00	12.90	27.50
	HOTEL	20	11.0	1.00	4.00	14.80	35.45
	FLIGHT	20	9.20	1.00	4.50	19.95	24.25
	MOVIE	20	8.00	1.00	7.00	13.45	34.15
	MEDICINE	20	6.55	1.00	4.20	13.15	19.00
Multi-channel	TRAVEL1	12	17.30	1.00	4.00	23.10	48.30
	TRAVEL2	10	21.33	1.00	4.00	23.67	55.25

Table 1: Datasets statistics. Instruction Mean and Underlying Mean denote the mean length of user instructions and corresponding underlying requirements.

#1 denotes the surface-level thought for basic embodied behavior on the device GUI, e.g., *This action clicks a button named [Button_name]*. Thought #2 denotes the deeper-level thought for explicit user preference, e.g., *It seems the user needs items with [Restriction1], [Restriction2], and [Restriction3]*. Thought #3 denotes the high-level thought for implicit user preference, e.g., *I recommend [item#1] from the item pool*.

Challenges. There are several key challenges to achieving COUT. The key challenge starts with the fact that task goals are often user's ambiguous queries, lacking definitive goals matched to the observations. For example, the existing embodied agents are trained to either touch a specific object using robotic arms in 3D space or click a particular button on 2D screens. However, these golden targets are not presented in user queries as personalized preference is typically subjective and nonverbal. How to define the personalized task in embodied environments and evaluate it remains unknown and challenging. Second, the deficiency of supportive data for COUT research poses a considerable challenge. The commonly used datasets are typically collected from task-oriented demonstrations. This gap highlights the need for more comprehensive datasets that can better facilitate advancements in COUT research. Third, due to the lack of clear task definitions and the absence of supportive training environments, the methods for analyzing personalized features have not been thoroughly explored.

The SmartSpot Benchmark

Given the scarcity of training data for embodied agents that explicitly captures the personalized analysis highlighted in the COUT paradigm, we propose to construct a novel benchmark named SmartSpot.

Dataset Summary

We collect SmartSpot data from Meituan[2], a well-known Internet service platform in China that offers a variety of life services, including food recommendations, hotel bookings, online flight ticket sales, etc. To create more practical personal scenes, we develop SmartSpot with two scenarios: the single-channel scenario which focuses on one type of service, and the multi-channel scenario which combines two single channels. We select five of the most daily used single channels: Food, Hotel, Flight, Movie, and Medicine. For the multi-channel scenarios, we pair Flight and Food as Travel1, and Flight and Hotel as Travel2. These combinations reflect more complicated and practical situations, such as traveling to a destination and booking hotels or restaurants. The data in all scenarios are GUI action episodes that include several steps to complete an instruction. Specifically, each episode consists of three groups of steps, as shown in different colors in Table 1. The blue "GUI" steps denote a series of GUI actions, like entering and completing a search bar, to access an item pool, the yellow "Item Pool" step denotes the page for the found item pool, and the "RS" steps denote the details page of each item awaiting recommendations. Each step contains a GUI screenshot, a ground-truth action (with possibly a bounding box or textual value), a list of previous actions, and an episodic instruction with a corresponding underlying requirement. Finally, SmartSpot covers seven scenarios that present a wealth of personal assistant tasks, supporting more embodied personalized research. The data statistics are illustrated in Table 1.

SmartSpot offers a set of common human-GUI interaction actions. We provide a detailed description of the full action space in Appendix Table 3.

Dataset Collection & Analysis

We curate SmartSpot following the real-life usage of this platform. The process begins with generating user instruction and underlying requirement pairs. To ensure consistency, we establish intention seeds to generate them simultaneously. Specifically, we recruit 10 annotators experienced on this platform to identify and select 2-3 significant attributes as seeds for searching specific item pools in each channel. For example, the "[recently played]" in Figure 1 presents the result of one seed, which could be other choices that users can click on. By creating multiple permutations of intention seeds, we collect a diverse batch of instruction pairs while ensuring there are no duplicates. Then, according to these instruction pairs, the annotators executed episodic GUI operations, capturing screenshots and ground-truth actions, including their bounding boxes at each step. All episodes are completed based on the annotators' personal usage habits on the platform. In total, we gathered 144 episodes with over 1,400 steps.

[2]https://www.meituan.com/en-US/about-us

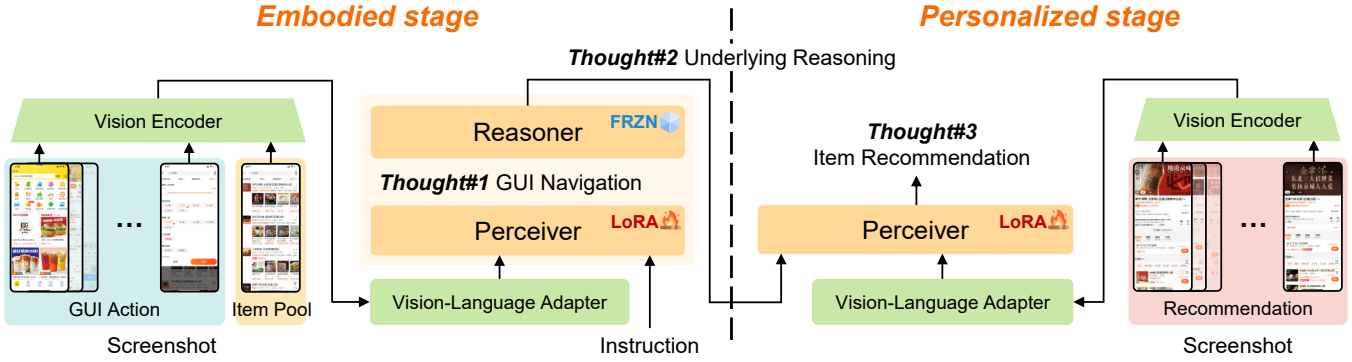

Figure 3: Two-stage training paradigm of SmartAgent. Thought #2 serves as a key intermediate result, connecting the Embodied and Personalized stages. This summarized user underlying requirement in the GUI operations greatly narrows the scope of the item pool for the next stage of personalized reasoning and item recommendation.

SmartAgent Approach

Overview

Given a user instruction $i \in \mathcal{I}$ to complete, the agent will navigate through an episode comprising three groups of steps: searching for an item pool, finding the item pool, and making item recommendations. At time step t, the agent receives a screenshot observation $o_t \in \mathcal{O}$. Then, the agent should take an action $a_t \in \mathcal{A}$ according to its current assets $\{o_t, i, h_{t-1}\}$, where $h_{t-1} = (o_1, a_1, ..., o_{t-1}, a_{t-1})$ is the historical actions. The chosen action signifies either a GUI command, a signal that an item pool has been found, or a recommendation result for items within that pool.

The primary design principles of SmartAgent are twofold: i) It should process multimodal input of high-resolution screenshot images and textual instructions, and the output of embodied action, along with textual thoughts; ii) It should handle ambiguous instructions as reasoning goals at all the stages of COUT. We therefore transform all data of different modalities into a token sequence, illustrated below:

$$\underbrace{\texttt{Generate...}}_{\text{system message}} \underbrace{T^{(1)}_{\text{image}}, ..., T^{(M)}_{\text{image}}}_{\text{screenshot image tokens}} \underbrace{\texttt{step1:...}}_{\text{previous actions}}, \\ \underbrace{\texttt{User:I want... Agent:}}_{\text{instruction}} \underbrace{T^{(1)}_{\text{res}}, ...T^{(N)}_{\text{res}}}_{\text{response}}. \quad (3)$$

Using this representation, we formulate the learning of SmartAgent as GPT-style auto-regressive approach, in line with (Huang et al. 2023). For instance, in Figure 1, given a smartphone screenshot and user's instruction "Can you give me some music recommendations?", we craft a query prompt as: "Please generate the next action according to the <screenshot>and <instruction>".

Agent Training

To achieve COUT reasoning, we divide each episode into the following embodied and personalized stages to train SmartAgent successively. Specifically, the SmartAgent backbone LVLM functions in two roles: a Perceiver trained in our environment to predict actions or a Reasoner utilizes the original LVLMs to generate thoughts.

Embodied Stage. This stage aims to complete ambiguous instructions to find the item pool. The agent takes only GUI action and item pool screenshots as visual input. These multimodal assets first feed into the Perceiver to predict the specific embodied actions, referred to as Thought #1 in COUT. Subsequently, the Reasoner infers step-wise action thoughts and summaries as an underlying requirement as Thought #2. The Thought #2 indicates the user's intention explicitly reflected in this stage. For instance, as illustrated in Figure 1, for user inquiries about music, Thought #2 may include specific constraints not present in the original instruction, such as "Justin Bieber". As a result, the underlying requirement serves as a key intermediate output, offering a clearer representation of user intentions for the next personalized stage.

Personalized Stage. With the fine-grained underlying requirement in Thought #2, this stage focuses on making personalized recommendations. The same Perceiver model takes item screenshots as visual input and determines if a recommendation is warranted by responding with either textual "Yes" or "No", which is designated as Thought #3.

Implementation Details

We choose Qwen-VL (Bai et al. 2023) as our backbone LVLM, which encodes visual inputs with a high resolution of 448*448. The training of SmartAgent starts from the continual pre-training on SeeClick base model (Cheng et al. 2024a) for basic GUI grounding abilities. Following the SeeClick's approach, we intuitively present numerical coordinates as natural languages without additional tokenization. We take 8 historical actions with screenshots during training considering the GPU memory limitation. We train SmartAgent 15 epochs for both the embodied and personalized stages. All baselines are trained for 15 rounds. Results on the ScreenSpot (Cheng et al. 2024a) and Mind2Web (Deng et al. 2024b) benchmarks are evaluated via direct testing and 10 epochs of training, as in SeeClick. We utilize AdamW as the optimizer, exploring learning rates of [1e-5, 2e-5, 3e-5, 4e-5, 1e-4] and global batch size of [10, 12, 14]. The optimal combination is a learning rate of 3e-5 and a global batch size of 14. All training is conducted on two NVIDIA A100 GPUs with LoRA fine-tuning.

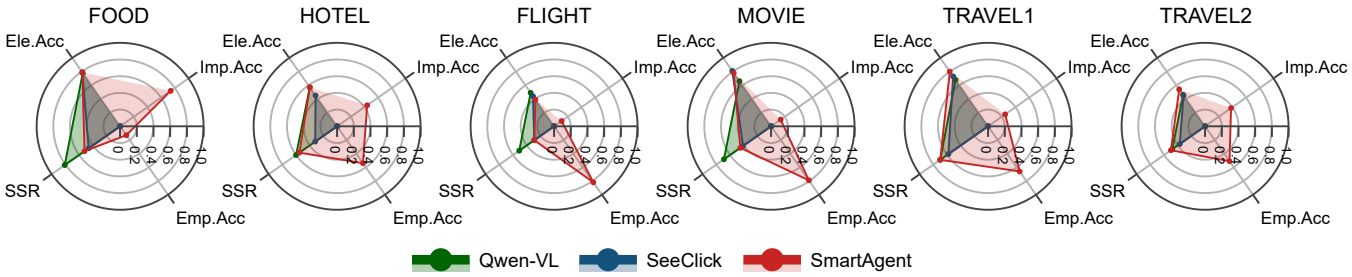

Figure 4: **Main comparisons on SmartSpot.** Notably, in the more complex scenarios, TRAVEL1 and TRAVEL2 that involve longer episodes stimulating the common practice, SmartAgent consistently shows excellent performance across all metrics.

Evaluation

In this section, we demonstrate SmartAgent's capabilities by comprehensive evaluations of its overall abilities and a full spectrum of sub-tasks encompassing GUI Grounding, Autonomous GUI Operation, Explicit & Implicit Personalized Reasoning, and Zero-shot Reasoning. We also report more insightful discovers in the Appendix.

Metrics. Following most of the settings in SeeClick (Cheng et al. 2024a), we compute the cross-entropy loss for Thought #1 and Thought #3 with their ground-truth actions, and semantic similarity for Thought #2 with the underlying requirement. A bounding box may be contained in ground-truth action to verify if a click action is hit. We therefore evaluate SmartAgent using the below metrics, in terms of embodied action metrics (following Mind2web) for Thought #1, and personalized preference metrics for Thought #2 and Thought #3.

Embodied action metrics:

- **Element Accuracy (Ele.Acc)**: The accuracy of predicted coordinate with ground-truth for click and type action.
- **Step Successful Rate (SSR)**: The rate of steps that both the action type and value are successfully predicted.

Personalized preference metrics:

- **Explicit Preference Accuracy (Exp.Acc)**: The semantic similarity between the predicted underlying requirement and ground truth.
- **Implicit Preference Accuracy (Imp.Acc)**: The accuracy of predicted items with the ground truth.

Embodied Personalized Reasoning

We first investigate the comprehensive capabilities of SmartAgent in handling embodied tasks and personalized reasoning, primarily focusing on SmartSpot for validation. Specifically, we categorize this experiment into simpler single-domain tasks and more complicated scenarios that mimic real multi-channel interactions. We exclude the MEDICINE channel to evaluate the zero-shot ability later. We select SeeClick (a widely used GUI agent backbone using pure visual observation), along with the well-known LLM foundation model, Qwen-VL series, as our baselines. We keep the original prompt reasoning setting for all general LLM and GUI Specialist baselines to show their features and deficiencies on the COUT reasoning task. Thus, only embodied action results are reported.

Results & analysis. Figure 4 shows their comparisons on all channels of SmartSpot. Generally, SmartAgent performs comparably and sometimes surpasses the GUI specialist and general LLM in GUI-related metrics. Based on this, we are excited to find that, mostly, even a modest accuracy in predicting users' underlying requirements can yield effective recommendations. This trend is evident in both single-channel scenarios, such as FOOD and HOTEL, and multi-channel scenarios like TRAVEL1 and TRAVEL2. This supports the COUT's idea that a solid cognition of basic embodied actions can facilitate meaningful predictions of users' underlying requirements, ultimately leading to competent personalized recommendations. More importantly, in multi-channel scenarios, SmartAgent consistently performs the best. This achievement highlights its superior potential in the real-world practical scene. Some anomalies are observed in the FLIGHT and MOVIE, where a high Exp.Acc score followed by less favorable outcomes. This suggests that overfitting to users' underlying needs may accumulate errors within the chain. A trade-off between reliance on intermediate results and long-chain reasoning requires further attention. We'll discuss this issue in more detail later. Overall, SmartAgent greatly extends the capabilities of LLMs through impressive personalized embodied reasoning.

GUI Grounding

An important consideration is whether training with personalized capabilities leads to catastrophic forgetting of pre-trained embodied abilities. In this section, we evaluate SmartAgent on a renowned GUI Grounding benchmark ScreenSpot (Cheng et al. 2024a) to assess its foundational perception of raw GUI data. The comparative baselines are divided into two main categories: GUI specialist models and general LLMs.

Results & analysis. As shown in Appendix Table 4, SmartAgent achieves second-best results in most metrics, even securing the top position in Desktop-Icon/Widget. This indicates that training with personalized capabilities not only preserves foundational embodied abilities but also enhances proficiency in operations involving user intent. This is also the primary achievement of the COUT paradigm. Note that we do not overly require SmartAgent to maintain a SOTA level on such traditional test environment that do not include

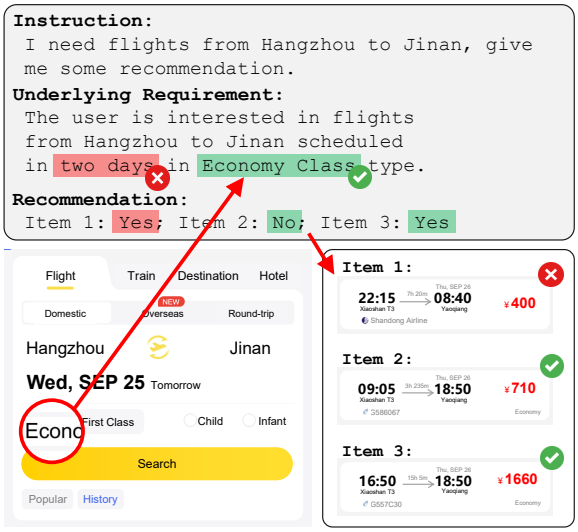

Figure 5: Case study for embodied personalized reasoning.

personalized factors. Our primary goal in this section is to ensure that it retains sufficient basic capabilities in tasks after incorporating personalized considerations.

Autonomous GUI Operation

Automated execution of human instructions is a foundational capability of embodied agents. In this section, we validate SmartAgent's basic abilities to handle GUI action episodes autonomously using the classic GUI agent benchmark Mind2Web (Deng et al. 2024b).

Results & analysis. We transfer the well-trained SmartAgent on SmartSpot to Mind2Web. As shown in Appendix Table 5, it achieves second place generally in the pure vision-based baselines across three downstream scenarios. This indicates that SmartAgent possesses strong embodied generalization capabilities, enabling seamless integration into various downstream tasks. Similarly, we do not expect SmartAgent to surpass the SOTA model in this basic task. The first-tier autonomous agent capabilities that have not been forgotten in the backbone model have laid the foundation for subsequent personalized reasoning tasks.

Explicit & Implicit Personalized Reasoning

In this section, we present a case study to demonstrate SmartAgent's performance in explicit user underlying reasoning and implicit item recommendation tasks. As shown in Figure 5, SmartAgent correctly predicted the action of clicking on the economy class option in the visual observation. This prediction is reflected in the summarization of the user's underlying requirements. Leveraging this textual representation of explicit needs, SmartAgent made recommendations for the last two flight options, effectively addressing the user's implicit requirements.

Zero-shot Reasoning

Zero-shot perception is the ultimate goal for embodied agents, enabling them truly to learn from interactions with their environment. In this section, we utilize the MEDICINE channel in SmartSpot to evaluate SmartAgent's zero-shot performance in unseen scenarios.

Type	MEDICINE			
	Ele.Acc	SSR	Exp.Acc	Imp.Acc
Zero-shot	0.04	0.08	**0.77**	0.14
Full-stage	0.64	0.50	0.71	0.24

Table 2: Zero-shot results on channel MEDICINE.

Results & analysis. As shown in Table 2, SmartAgent surprisingly exceeds its average performance achieved through fine-tuning on SmartSpot in the Exp.Acc metric. We argue that in cold-start scenarios like MEDICINE, the training data may be very limited, rendering fine-tuning ineffective or even leading to overfitting. SmartAgent's zero-shot capability can compensate for this limitation by leveraging general GUI knowledge to interact directly with new scenarios. This enhancement is also evident in the subsequent item recommendation task, which approaches the performance of full-stage fine-tuning. Overall, this demonstrates SmartAgent's robustness in interpreting users' explicit intentions and also indicates a preliminary zero-shot reasoning capability.

More Insights into COUT

In this section, we provide deeper insights into the COUT process. We begin by providing a detailed analysis of the aforementioned anomaly case through comparisons between the two-stage training and end-to-end training settings (without underlying thought generation). A general trade-off is observed between these two paradigms, which we will go through in detail in Appendix Table 6.

Finally, we will explore a broader view of SmartAgent's overall performance through a comparison of baseline average scores across all channels. This detailed analysis will be presented in Appendix Table 7.

Conclusion and Future Work

In this paper, we introduce a novel embodied reasoning paradigm, COUT, which defines an embodied personalized task for the first time. We establish a clear definition and components of the COUT paradigm and analyze its challenges. To address them, we propose SmartAgent to instantiate COUT through a two-stage training from essential GUI reasoning to high-level user thought reasoning. To evaluate this progress, we created SmartSpot, the first embodied AI benchmark featuring personalization evaluations. The results demonstrate the effectiveness and proficiency of SmartAgent over full-stage embodied personalized reasoning tasks. Furthermore, SmartAgent showcases the key capability of zero-shot embodied reasoning, highlighting its potential for efficient adaptation in new scenarios.

As for the future work, we plan to further extend the scale of the benchmark, with more extensive experiments and analysis. We also plan to apply the proposed method to the real user devices.

Acknowledgments

This work was supported in part by Australian Research Council under the streams of Future Fellowship (Grant No. FT210100624), in part by the Discovery Project (Grant No. DP240101108 and DP260100326), in part by the Linkage Projects (Grant No. LP230200892 and LP240200546), in part by the National Key Research and Development Program of China 2022YFB3104702, and in part by the National Natural Science Foundation of China 62272262, 72442026, 72342032.

References

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Barreiros, J. A.; Xu, A.; Pugach, S.; Iyengar, N.; Troxell, G.; Cornwell, A.; Hong, S.; Selman, B.; and Shepherd, R. F. 2022. Haptic perception using optoelectronic robotic flesh for embodied artificially intelligent agents. *Science Robotics*, 7(67): eabi6745.

Cheng, K.; Sun, Q.; Chu, Y.; Xu, F.; Li, Y.; Zhang, J.; and Wu, Z. 2024a. Seeclick: Harnessing gui grounding for advanced visual gui agents. *arXiv preprint arXiv:2401.10935*.

Cheng, Y.; Pan, Y.; Zhang, J.; Ni, Y.; Sun, A.; and Yuan, F. 2024b. An Image Dataset for Benchmarking Recommender Systems with Raw Pixels. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, 418–426. SIAM.

Deng, S.; Xu, W.; Sun, H.; Liu, W.; Tan, T.; Liu, J.; Li, A.; Luan, J.; Wang, B.; Yan, R.; et al. 2024a. Mobile-Bench: An Evaluation Benchmark for LLM-based Mobile Agents. *arXiv preprint arXiv:2407.00993*.

Deng, X.; Gu, Y.; Zheng, B.; Chen, S.; Stevens, S.; Wang, B.; Sun, H.; and Su, Y. 2024b. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36.

Driess, D.; Xia, F.; Sajjadi, M. S.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Drouin, A.; Gasse, M.; Caccia, M.; Laradji, I. H.; Del Verme, M.; Marty, T.; Boisvert, L.; Thakkar, M.; Cappart, Q.; Vazquez, D.; et al. 2024. WorkArena: How Capable are Web Agents at Solving Common Knowledge Work Tasks? *arXiv preprint arXiv:2403.07718*.

Duan, J.; Yu, S.; Tan, H. L.; Zhu, H.; and Tan, C. 2022. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2): 230–244.

Duéñez-Guzmán, E. A.; Sadedin, S.; Wang, J. X.; McKee, K. R.; and Leibo, J. Z. 2023. A social path to human-like artificial intelligence. *Nature Machine Intelligence*, 5(11): 1181–1188.

Fu, J.; Yuan, F.; Song, Y.; Yuan, Z.; Cheng, M.; Cheng, S.; Zhang, J.; Wang, J.; and Pan, Y. 2024. Exploring Adapter-based Transfer Learning for Recommender Systems: Empirical Studies and Practical Insights. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24. ACM.

Gao, C.; Xu, F.; Chen, X.; Wang, X.; He, X.; and Li, Y. 2024a. Simulating Human Society with Large Language Model Agents: City, Social Media, and Economic System. In *Companion Proceedings of the ACM on Web Conference 2024*, 1290–1293.

Gao, C.; Zhao, B.; Zhang, W.; Mao, J.; Zhang, J.; Zheng, Z.; Man, F.; Fang, J.; Zhou, Z.; Cui, J.; et al. 2024b. EmbodiedCity: A Benchmark Platform for Embodied Agent in Real-world City Environment. *arXiv preprint arXiv:2410.09604*.

Geng, S.; Liu, S.; Fu, Z.; Ge, Y.; and Zhang, Y. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*, 299–315.

He, H.; Yao, W.; Ma, K.; Yu, W.; Dai, Y.; Zhang, H.; Lan, Z.; and Yu, D. 2024. WebVoyager: Building an End-to-End Web Agent with Large Multimodal Models. *arXiv preprint arXiv:2401.13919*.

Hong, W.; Wang, W.; Lv, Q.; Xu, J.; Yu, W.; Ji, J.; Wang, Y.; Wang, Z.; Dong, Y.; Ding, M.; et al. 2024. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14281–14290.

Hou, Y.; Li, J.; He, Z.; Yan, A.; Chen, X.; and McAuley, J. 2024. Bridging Language and Items for Retrieval and Recommendation. *arXiv preprint arXiv:2403.03952*.

Huang, J.; Yong, S.; Ma, X.; Linghu, X.; Li, P.; Wang, Y.; Li, Q.; Zhu, S.-C.; Jia, B.; and Huang, S. 2023. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*.

Kang, W.-C.; and McAuley, J. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, 197–206. IEEE.

Kim, G.; Baldi, P.; and McAleer, S. 2024. Language models can solve computer tasks. *Advances in Neural Information Processing Systems*, 36.

Kim, H.; Padmakumar, A.; Jin, D.; Bansal, M.; and Hakkani-Tur, D. 2022. On the Limits of Evaluating Embodied Agent Model Generalization Using Validation Sets. *arXiv preprint arXiv:2205.09249*.

Li, R.; Deng, W.; Cheng, Y.; Yuan, Z.; Zhang, J.; and Yuan, F. 2023. Exploring the upper limits of text-based collaborative filtering using large language models: Discoveries and insights. *arXiv preprint arXiv:2305.11700*.

Li, Y.; He, J.; Zhou, X.; Zhang, Y.; and Baldridge, J. 2020. Mapping natural language instructions to mobile UI action sequences. *arXiv preprint arXiv:2005.03776*.

Ma, X.; Zhang, Z.; and Zhao, H. 2024. CoCo-Agent: A Comprehensive Cognitive MLLM Agent for Smartphone GUI Automation. *arXiv preprint arXiv:2402.11941v3*.

Mialon, G.; Fourrier, C.; Swift, C.; Wolf, T.; LeCun, Y.; and Scialom, T. 2023. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*.

Moskalenko, O.; Parra, D.; and Saez-Trumper, D. 2020. Scalable recommendation of wikipedia articles to editors using representation learning. *arXiv preprint arXiv:2009.11771*.

Rawles, C.; Clinckemaillie, S.; Chang, Y.; Waltz, J.; Lau, G.; Fair, M.; Li, A.; Bishop, W.; Li, W.; Campbell-Ajala, F.; et al. 2024a. AndroidWorld: A dynamic benchmarking environment for autonomous agents. *arXiv preprint arXiv:2405.14573*.

Rawles, C.; Li, A.; Rodriguez, D.; Riva, O.; and Lillicrap, T. 2024b. Androidinthewild: A large-scale dataset for android device control. *Advances in Neural Information Processing Systems*, 36.

Ren, X.; Chen, T.; Nguyen, Q. V. H.; Cui, L.; Huang, Z.; and Yin, H. 2024. Explicit knowledge graph reasoning for conversational recommendation. *ACM Transactions on Intelligent Systems and Technology*, 15(4): 1–21.

Ren, X.; Yin, H.; Chen, T.; Wang, H.; Huang, Z.; and Zheng, K. 2021. Learning to ask appropriate questions in conversational recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 808–817.

Shi, T.; Karpathy, A.; Fan, L.; Hernandez, J.; and Liang, P. 2017. World of bits: An open-domain platform for web-based agents. In *International Conference on Machine Learning*, 3135–3144. PMLR.

Srivastava, S.; Li, C.; Lingelbach, M.; Martín-Martín, R.; Xia, F.; Vainio, K. E.; Lian, Z.; Gokmen, C.; Buch, S.; Liu, K.; et al. 2022. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Conference on robot learning*, 477–490. PMLR.

Tan, W.; Ding, Z.; Zhang, W.; Li, B.; Zhou, B.; Yue, J.; Xia, H.; Jiang, J.; Zheng, L.; Xu, X.; et al. 2024. Towards general computer control: A multimodal agent for red dead redemption ii as a case study. *arXiv preprint arXiv:2403.03186*.

Wang, B.; Li, G.; and Li, Y. 2023. Enabling conversational interaction with mobile ui using large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17.

Wang, J.; Xu, H.; Ye, J.; Yan, M.; Shen, W.; Zhang, J.; Huang, F.; and Sang, J. 2024. Mobile-agent: Autonomous multi-modal mobile device agent with visual perception. *arXiv preprint arXiv:2401.16158*.

Wei, W.; Ren, X.; Tang, J.; Wang, Q.; Su, L.; Cheng, S.; Wang, J.; Yin, D.; and Huang, C. 2024. Llmrec: Large language models with graph augmentation for recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 806–815.

Xu, F.; Zhang, J.; Gao, C.; Feng, J.; and Li, Y. 2023. Urban generative intelligence (ugi): A foundational platform for agents in embodied city environment. *arXiv preprint arXiv:2312.11813*.

Yan, A.; Yang, Z.; Zhu, W.; Lin, K.; Li, L.; Wang, J.; Yang, J.; Zhong, Y.; McAuley, J.; Gao, J.; et al. 2023. Gpt-4v in wonderland: Large multimodal models for zero-shot smartphone gui navigation. *arXiv preprint arXiv:2311.07562*.

Yang, J.; Dong, Y.; Liu, S.; Li, B.; Wang, Z.; Tan, H.; Jiang, C.; Kang, J.; Zhang, Y.; Zhou, K.; et al. 2025. Octopus: Embodied vision-language programmer from environmental feedback. In *European Conference on Computer Vision*, 20–38. Springer.

Yang, Z.; Liu, J.; Han, Y.; Chen, X.; Huang, Z.; Fu, B.; and Yu, G. 2023. Appagent: Multimodal agents as smartphone users. *arXiv preprint arXiv:2312.13771*.

Yin, H.; Qu, L.; Chen, T.; Yuan, W.; Zheng, R.; Long, J.; Xia, X.; Shi, Y.; and Zhang, C. 2024. On-device recommender systems: A comprehensive survey. *arXiv preprint arXiv:2401.11441*.

Yuan, F.; Karatzoglou, A.; Arapakis, I.; Jose, J. M.; and He, X. 2019. A simple convolutional generative network for next item recommendation. In *Proceedings of the twelfth ACM international conference on web search and data mining*, 582–590.

Yuan, W.; Yin, H.; Wu, F.; Zhang, S.; He, T.; and Wang, H. 2023. Federated unlearning for on-device recommendation. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, 393–401.

Zhan, Z.; and Zhang, A. 2023. You only look at screens: Multimodal chain-of-action agents. *arXiv preprint arXiv:2309.11436*.

Zhang, J.; Cheng, Y.; Ni, Y.; Pan, Y.; Yuan, Z.; Fu, J.; Li, Y.; Wang, J.; and Yuan, F. 2024a. Ninerec: A benchmark dataset suite for evaluating transferable recommendation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, J.; Wu, J.; Teng, Y.; Liao, M.; Xu, N.; Xiao, X.; Wei, Z.; and Tang, D. 2024b. Android in the zoo: Chain-of-action-thought for gui agents. *arXiv preprint arXiv:2403.02713*.

Zhang, J.; Yu, J.; Wang, Z.; Yuan, W.; Chen, T.; Nguyen, Q. V. H.; Cui, B.; and Yin, H. 2025a. Towards Reasoning-Aware Recommender Systems: A Survey in the LLM Era. *ResearchGate preprint: 10.13140/RG.2.2.21786.30405*.

Zhang, W.; Zhou, Z.; Zheng, Z.; Gao, C.; Cui, J.; Li, Y.; Chen, X.; and Zhang, X.-P. 2025b. Open3dvqa: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 12784–12791.

Zhao, B.; Wang, Z.; Fang, J.; Gao, C.; Man, F.; Cui, J.; Wang, X.; Chen, X.; Li, Y.; and Zhu, W. 2025. Embodied-r: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 11071–11080.

Zheng, L.; Wang, R.; Wang, X.; and An, B. 2023. Synapse: Trajectory-as-exemplar prompting with memory for computer control. In *The Twelfth International Conference on Learning Representations*.