PERCU: Benchmarking Multimodal Agents on Personalized Computer Use Tasks

Anonymous ACL submission

Abstract

Large language model (LLM) agents have demonstrated remarkable potential in automating digital workflows through multimodal planning and reasoning. However, providing truly personalized assistance in computer use remains a significant challenge, as existing benchmarks predominantly treat agents as context-independent executors, operating under a homogeneity assumption that ignores the diverse user-specific habits and procedural routines. To address these challenges, we introduce PERCU, a benchmark designed to evaluate the personalized capabilities of multimodal agents on computer use tasks. PERCU employs a dual-instruction paradigm where agents must ingest personalized knowledge from semantically explicit first instructions and subsequently utilize the resulting memory to resolve ambiguous second instructions. Extensive evaluation of several multimodal agents on PERCU reveals significant deficiencies in their ability to serve as personalized computer assistants. Further quantitative analysis using PERCU provides valuable insights for future research in developing personalized multimodal agents. Our code and data will be available at https://github.com/scm62519/PERCU.

1 Introduction

The rapid advancement of Large Language Models (LLMs) has catalyzed a paradigm shift in artificial intelligence, transitioning agents from passive command-followers to autonomous entities endowed with sophisticated reasoning, planning, and reflective capabilities (Liu et al., 2023; Mialon et al., 2023). In the domain of computer use, these agents are increasingly expected to navigate complex graphical user interfaces (GUIs) to automate diverse digital workflows. However, data scarcity remains a primary bottleneck. While users demand highly tailored assistance, acquiring large-scale, high-quality labeled data for every unique individual to train a personalized assistant is practically infeasible. This tension necessitates a move toward few-shot personalization (Kim and Yang, 2025), where an agent must rapidly adapt to a user's specific preferences and behavioral patterns from minimal interaction history. Personalization (Liu et al., 2025; Nandakishor and Anjali, 2025; Shenfeld et al., 2025) represents the critical evolution of an agent from a generic automation tool into a true digital companion. Such an agent functions as a partner that understands underlying intent and idiosyncratic habits rather than a cold executor of literal, context-free commands (Wu et al., 2025).

Despite the burgeoning interest in computer use agents, current evaluative frameworks, such as OS-World (Xie et al., 2024b) and WebArena (Zhou et al., 2023), largely overlook the subjective and repetitive nature of human-computer interaction. These benchmarks typically operate under a homogeneity assumption, where success is measured against a single, universal execution path. Such a paradigm ignores the fact that a user's historical preferences and long-term routines are decisive factors in task execution. For instance, in a personalized educational or research setting, an agent should not only execute a download but also recognize that a specific user consistently categorizes academic PDFs into a "paper" folder and subsequently initiates a note-taking routine. As illustrated in Figure 1, the failure of existing models to integrate such personalized knowledge results in a significant gap between laboratory performance and real-world utility, highlighting the urgent need for a benchmark that accounts for personalized computer use tasks.

Existing multimodal agents lack a standardized benchmark to test their ability to internalize digital habits, such as specific file organization patterns or repetitive software routines, which are essential for transforming an automation tool into a truly personalized assistant. To bridge this gap, we introduce

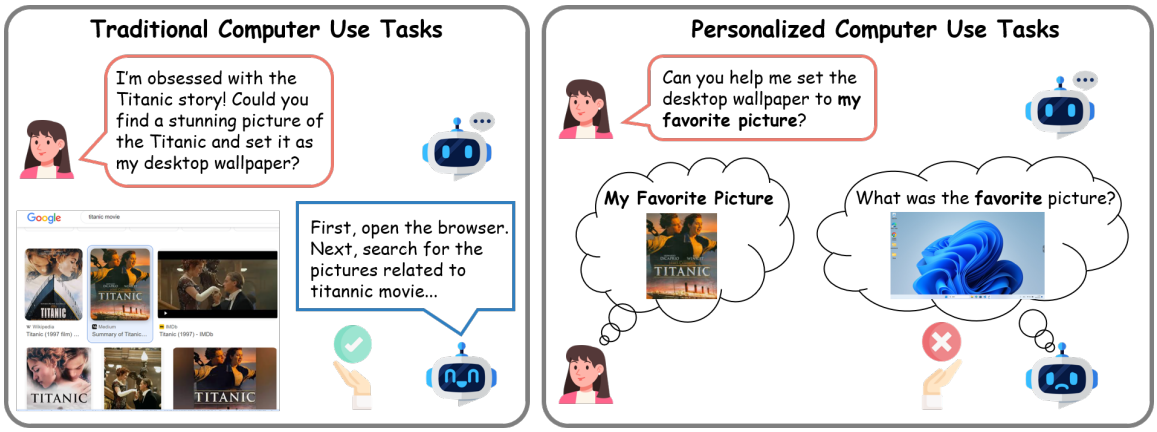

Figure 1: Comparison between traditional computer use tasks and personalized computer use tasks. While conventional agents focus on strictly following literal instructions based on universal defaults, personalized assistance agents must leverage user-specific knowledge grounded in past interactions. This disparity highlights the critical necessity of transitioning from rigid instruction-following toward context-aware, long-term personalized assistants that can resolve ambiguity through behavioral priors.

PERCU (**PER**sonalization in **C**omputer **U**se tasks), the first benchmark specifically designed to assess the personalized capabilities of multimodal agents on computer use tasks. PERCU fills this gap by formalizing and evaluating personalized computer use tasks via a dual-instruction framework. Our main contributions are summarized as follows:

- We formalize the definition of personalization within the computer use paradigm, categorizing personalized tasks into two distinct types: preference tasks, which capture static user inclinations, and routine tasks, which represent dynamic, multi-step procedural workflows.
- We propose the PERCU benchmark, a comprehensive evaluation suite that requires agents to leverage short-term memory to resolve personalized instructions in realistic desktop environments.
- Through extensive experimentation and in-depth analysis of five multimodal agents, we identify critical bottlenecks in current models, especially regarding the precision of memory retrieval and the generalization of procedural knowledge. These findings offer key insights to guide the future development of truly human-centered digital assistants.

2 Related Work

Benchmarks for Agents. The emergence of LLMs as autonomous agent cores (Hong et al., 2024; Niu et al., 2024; Cai et al., 2025; Qin et al., 2025a) has necessitated the development of rigorous evaluative frameworks to measure their reasoning and decision-making capabilities in interactive environments. Early efforts focused on narrow domains such as tool utilization (Qin et al., 2023) or coding tasks. More recently, multi-dimensional benchmarks like AgentBench (Liu et al., 2023) have been introduced to provide a comprehensive assessment across diverse settings, including operating systems, databases, and knowledge graphs. GAIA (Mialon et al., 2023) shifted the focus toward tasks that are conceptually simple for humans but challenging for AI, emphasizing tool-use proficiency and common-sense reasoning. Furthermore, TravelPlanner (Xie et al., 2024a) highlighted the limitations of current agents in handling long-horizon planning with complex constraints. Moving beyond the browser, OSWorld (Xie et al., 2024b) and WindowsAgentArena (Bonatti et al., 2024) introduced unified frameworks for controlling entire operating systems, focusing on cross-application workflows. WebArena (Zhou et al., 2023) established a milestone by providing a realistic, self-hosted web environment for evaluating agents on end-to-end tasks. This was followed by Mind2Web (Deng et al., 2023), which introduced a large-scale dataset for general-domain web navigation. While these benchmarks have significantly advanced the field, they primarily evaluate agents as generalist executors and focus on functional correctness within a zero-shot context, lacking a mechanism to evaluate how agents adapt to the evolving routines or preferences of a specific user and ignor-

ing the idiosyncratic habits that define real-world human-agent collaboration.

Personalization in LLMs. Personalization in LLMs (Zhang et al., 2024; Woźniak et al., 2024; Tan et al., 2024; Chen et al., 2024) has traditionally focused on linguistic style mimicry or topical preference alignment in conversational settings. The LaMP benchmark (Salemi et al., 2024) introduced a systematic way to evaluate personalized text classification and generation by providing models with user-specific history. In the realm of safety, PENGUIN (Wu et al., 2025) explored personalized safety alignment, demonstrating that an agent's risk assessment should vary based on the user's unique background. Closer to our work, MEMENTO (Kwon et al., 2025) investigated how embodied agents utilize episodic memory to solve personalized object rearrangement tasks. Despite these advancements, personalization in the computer use domain remains largely underexplored.

3 PERCU

3.1 Definition of Personalization in Computer Use Tasks

Personalization, in the broader context of artificial intelligence, refers to the tailoring of information, services, or system behaviors to accommodate the specific requirements and preferences of individual users or distinct user groups (Pazzani and Billsus, 2007). While the fundamental objective remains the enhancement of system relevance and user satisfaction, the manifestation of personalization varies significantly across domains. In conversational systems, personalization is primarily characterized by the alignment of agent responses with a user's persona, linguistic style, and historical interaction context (Li et al., 2016). Besides, in recommendation systems, the emphasis lies in the dynamic generation of content suggestions, such as advertisements or products, derived from latent behavioral patterns (Resnick and Varian, 1997). Furthermore, personalization in human-computer interaction focuses on the system's adaptation to usage habits, such as optimized interface layouts or specialized shortcut configurations, whereas in embodied AI, it necessitates the continuous adjustment of physical action strategies based on long-term human-robot interactions (Kwon et al., 2025).

In the specific domain of computer use, personalization transcends simple content filtering and enters the realm of complex workflow adaptation. We formalize personalization within this paradigm through five core dimensions that collectively influence how an agent interacts with a desktop environment.

The first dimension is **User Attributes**, which encompass static profiles such as occupation, age, cultural background, cognitive style, and personality trait. These attributes dictate the fundamental behavioral logic of the agent. Moreover, agents must account for physical or cognitive constraints. For instance, a graduate student's environment may be centered around document management and experimental scripts, whereas a software engineer requires a focus on debugging tools and version control systems. Moreover, agents must account for physical or cognitive constraints, such as adjusting visual interfaces for users with color vision deficiencies or prioritizing voice-based commands for users with motor impairments.

The second and third dimensions focus on **Interests** and **Behavioral Habits**, respectively. Interests determine topical preferences and the content-level priority of an agent, such as prioritizing academic tools over social software or opting for deep technical summaries over brief abstracts during literature review tasks. These preferences often extend to functional choices, such as a user's predilection for command line interfaces over GUIs. Complementing these interests are behavioral habits, which manifest as repetitive usage patterns and interaction frequencies. These include the frequent use of specific document templates, preferred window management configurations, or idiosyncratic keyboard shortcuts. For instance, by internalizing these patterns from system logs and file metadata, a personalized agent can autonomously apply formatting styles or initialize development environments that mirror the user's established digital footprint.

The fourth and fifth dimensions, **Task Execution Flow** and **Decision-making Style**, represent the procedural and strategic layers of personalization. Task execution flow describes the specific sequence of actions a user employs to achieve a goal, such as the distinct routine of searching, downloading, highlighting, and summarizing a research paper. Understanding these idiosyncratic routines allows the agent to move beyond reactive command following to proactive workflow automation. This is further modulated by the user's decision-making style, which characterizes their tolerance for risk and exploration. A conservative user may demand explicit confirmation before the agent exe-

cutes potentially destructive operations, such as system configuration changes or file deletions, while a cost-sensitive user might expect the agent to prioritize energy-efficient operations when battery levels are low. Collectively, these five factors form the foundational pillars of personalized computer use, necessitating a benchmark that evaluates an agent's ability to synthesize these multi-faceted signals from the memory.

3.2 Composition of PERCU

The construction of PERCU is grounded in the high-quality trajectory data provided by the PC-Agent-E dataset (He et al., 2025), which serves as the foundational substrate for our evaluation suite. To adapt these generic automation trajectories for personalization research, we extract the "thought" sequences and action histories to serve as the agent's short-term memory, effectively transforming a standard task execution into a record of a user's unique behavior.

Building upon this data foundation, we introduce a dual-instruction paradigm inspired by the two-stage evaluation framework for personalized embodied assistants (Kwon et al., 2025). For every task in PERCU, we manually curate a pair of directives which include the **First Instruction** and the **Second Instruction**. The First Instruction represents the initial interaction where the user explicitly describes their task while embedding one of the five personalization dimensions discussed in Section 3.1. These instructions are semantically transparent and detailed. As illustrated in Figure 1, a user might state, "I'm obsessed with the Titanic story! Could you find a stunning picture of the Titanic and set it as my desktop wallpaper?". During this stage, the agent executes the task and ingests the resulting trajectory into its memory bank, thereby internalizing the user's specific preferences or habits as contextual knowledge.

The Second Instruction simulates subsequent interactions where user prompts become increasingly underspecified or implicit. Unlike the initial setup, these "personalized instructions" are often vague, such as "Can you help me set the desktop wallpaper to my favorite picture?". The fundamental challenge posed by PERCU is whether the agent can successfully bridge the gap between this underspecified query and the explicit knowledge stored in its memory. For example, inferring that favorite picture refers to the "Titanic" image from the previous interaction. This setup reflects realistic daily usage, where users expect an intelligent assistant to leverage the memory to resolve ambiguity without requiring repetitive, exhaustive explanations.

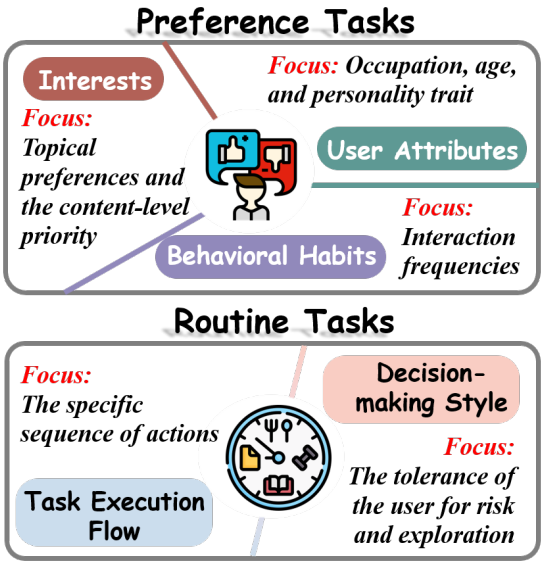

Figure 2: The illustration of two distinct types of tasks in PERCU.

To further systematize the evaluation, PERCU categorizes tasks into two distinct types which include **preference tasks** and **routine tasks**. As shown in Figure 2, preference tasks are designed to capture a user's static inclinations toward their digital environment, typically involving one-time configurations or long-term default settings. The success of an agent in these tasks hinges on its ability to ground abstract personalized entities in specific configuration instances. In contrast, routine tasks assess the agent's mastery of dynamic, multi-step procedural workflows. These tasks represent repeatable action sequences, such as a graduate student's specific workflow for filing and summarizing research papers. While preference tasks focus on "what" a user likes, routine tasks focus on "how" a user works, requiring the agent to associate high-level task intents with unique, structured operation routines.

3.3 Evaluation Process

The evaluation protocol of PERCU is designed to systematically quantify the execution accuracy of multimodal LLM agents across personalized computer use trajectories. We implement an automated evaluation pipeline that iterates through the JSONL-formatted task files, traversing each logical step to assess the agent's decision-making precision. For each test instance, the agent is initialized with a

meticulously crafted system prompt that defines its role as a multimodal GUI operator. This prompt establishes the behavioral constraints for interacting with graphical interfaces and mandates a chain-of-thought reasoning structure (Wei et al., 2022), requiring the model to articulate its cognitive process before generating a structured action command. This decoupling of reasoning and execution is pivotal for enhancing the logical consistency of agents in complex desktop environments. System prompt for PERCU is shown in Figure 4 in Appendix A.

A critical challenge in evaluating personalized agents is the injection of the necessary personalized knowledge to resolve the underspecified second instructions. Following the methodology of PC-Agent-E (He et al., 2025), we utilize the reconstructed thought processes that are originally generated during the thought completion stage as a form of short-term memory. As illustrated in Figure 3, the final input is constructed by concatenating the system prompt, the memory, the screenshot and the Second Instruction, followed by a standardized query prompt. By providing the agent with the "thought" from the corresponding step of the first interaction, we simulate a scenario where the agent "recalls" the user's specific habits or procedural preferences while observing the current state of the screen.

To isolate the agent's capability at each decision point and prevent the accumulation of errors typical in long-horizon tasks, we adopt a step-wise independent evaluation strategy. Rather than allowing the agent to proceed based on its own potentially erroneous prior actions, we provide the ground-truth visual state and the corresponding memory for every individual step. The evaluation framework then compares the agent's predicted action against the ground-truth action stored in the PERCU dataset. For interactive actions such as click, right-click, or double-click, which involve precise screen coordinates (x, y), we implement a coordinate parsing logic with a spatial tolerance. Specifically, we calculate the Euclidean distance between the predicted coordinates and the gold-standard coordinates. An action is categorized as correct if this distance is within a 50-pixel threshold. This approach acknowledges the inherent variability in visual grounding while ensuring the functional validity of the operation. By treating short-term memory as an in-context prompt component, our framework provides a lightweight yet effective mechanism for assessing personalized retrieval without the need for extensive model fine-tuning or complex vector database infrastructures.

4 Experiments on PERCU

To systematically evaluate the performance of multimodal agents in personalized desktop environments, we conduct extensive experiments on the PERCU benchmark. Our evaluation focuses on assessing whether current agents can effectively leverage short-term memory to internalize user habits and execute complex routines. We evaluate five multimodal models that represent different architectural approaches to GUI automation, providing a comprehensive overview of the current landscape of personalized computer use.

4.1 Experiments Setup

Settings. All experiments are conducted in a standardized vision-only setting, where agents observe the environment exclusively through screen captures without access to structured metadata such as document object models or accessibility trees. To ensure consistency and avoid coordinate misalignment across different models, the screen resolution for all evaluation tasks is fixed at 1280×720 pixels. This setup requires agents to possess robust visual grounding and semantic reasoning capabilities to map high-level personalized instructions to precise pixel-space coordinates.

Datasets. The dataset statistics of preference tasks and routine tasks in PERCU are shown in Table 1.

Evaluation Metrics. To quantify agent performance, we employ two primary evaluation metrics which include **Micro Average Accuracy** and **Macro Average Accuracy**.

Macro average accuracy is computed by first calculating the success rate within each task category (i.e., preference Tasks and routine tasks) and then taking the unweighted average of these category-level scores. This ensures that the evaluation is not biased toward tasks with a disproportionately large number of steps and accurately reflects the agent's balanced proficiency across different personalization dimensions. The macro average accuracy is defined as:

$$\text{Macro Average Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{C_i}{S_i} \right), \quad (1)$$

where N denotes the total number of tasks, i denotes the index of the i-th task, C_i denotes the

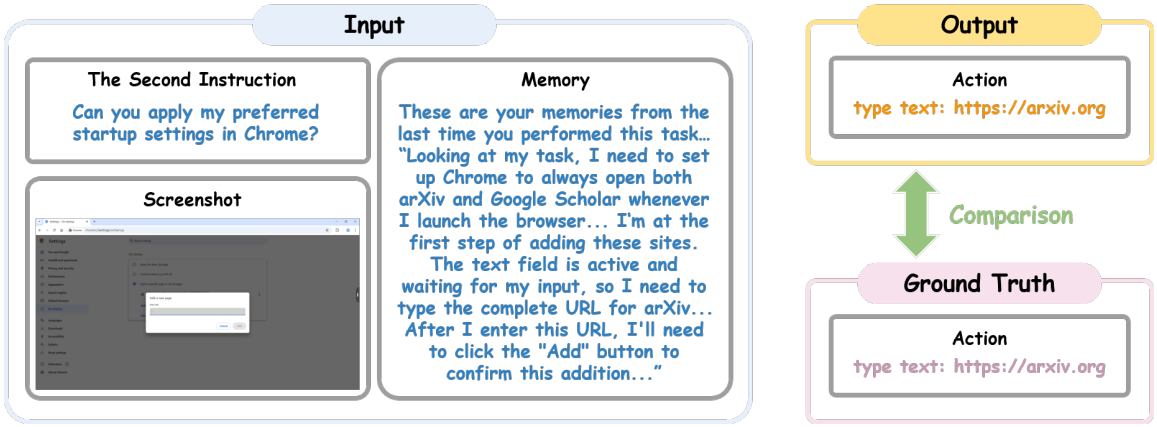

Figure 3: An example that demonstrates the evaluation process of PERCU.

Tasks	Number of Tasks	The Average Length of FI	The Average Length of the SI
Preference tasks	121	117.36	104.22
Routine tasks	191	97.30	88.07
All tasks	312	105.08	94.34

Table 1: The dataset statistics of preference tasks and routine tasks in PERCU. "FI" denotes the First Instruction. "SI" denotes the Second Instruction.

number of correctly predicted steps in the i-th task, S_i denotes the total number of steps in the i-th task, and $\frac{C_i}{S_i}$ denotes the accuracy rate of the i-th task.

Micro average accuracy, on the other hand, is calculated at the step level, representing the proportion of correctly executed actions over the total number of evaluation steps across all tasks. This metric provides an overall measure of the agent's operational reliability and stability. The micro average accuracy is defined as:

$$\text{Micro Average Accuracy} = \frac{\sum_{i=1}^{N} C_i}{\sum_{i=1}^{N} S_i}. \quad (2)$$

4.2 Baselines

We benchmark five representative multimodal GUI agents to establish a competitive performance baseline. OS-Atlas-Pro-7B (Wu et al., 2024) is evaluated as a foundation action model designed for generalist GUI interactions, leveraging large-scale cross-platform pre-training to achieve superior visual perception. GUI-Owl-7B, the core of the Mobile-Agent-v3 framework (Ye et al., 2025), focuses on fundamental GUI automation through a vision-centric architecture optimized for high-fidelity screen understanding. ShowUI-2B (Lin et al., 2025) represents an emerging class of one-vision-language-action models that integrate visual perception and action prediction into a unified, lightweight backbone, demonstrating high efficiency in GUI visual tasks. We also include GUI-R1 (Luo et al., 2025), a generalist R1-style model that utilizes long-horizon reasoning and reinforcement learning to handle complex, multi-step interface operations. Finally, we evaluate UI-TARS-1.5-7B (Qin et al., 2025b), a native agent model designed for automated GUI interaction that emphasizes robust instruction following and real-time adaptation to diverse software environments. These baselines span a wide range of parameter scales and training paradigms, allowing for a multifaceted analysis of memory-driven personalization.

4.3 Main Results

The overall performance of the evaluated multimodal agents on PERCU is summarized in Table 1. Our evaluation methodology prioritizes functional correctness, focusing on the equivalence between the agent's generated actions and the ground-truth execution rather than purely linguistic alignment. All existing models substantially underperform compared to humans. Among the tested models, GUI-R1-7B (Luo et al., 2025) achieves the highest performance with a macro average accuracy of 45.56% and a micro average accuracy of 45.23% in all tasks. However, even the most capable model exhibits a substantial performance gap when compared to human performance, which

Models	Preference Tasks		Routine Tasks		All Tasks	
	Macro Acc (↑)	Micro Acc (↑)	Macro Acc (↑)	Micro Acc (↑)	Macro Acc (↑)	Micro Acc (↑)
ShowUI-2B (Lin et al., 2025)	2.72	3.05	3.95	4.36	3.47	3.85
GUI-R1-3B (Luo et al., 2025)	31.53	29.46	30.81	29.83	31.61	30.37
GUI-R1-7B (Luo et al., 2025)	48.96	46.72	42.32	43.1	45.56	45.23
GUI-Owl-7B (Ye et al., 2025)	37.27	36.66	39.49	38.56	38.12	36.78
UI-TARS-1.5-7B (Qin et al., 2025b)	19.79	19.48	19.88	19.63	21.28	21.09
OS-Atlas-Pro-7B (Wu et al., 2024)	11.33	9.88	12.22	10.32	11.66	9.93
Humans	91.08	90.59	91.57	90.94	91.32	90.47

Table 2: Main results of different models on PERCU. "Macro Acc" denotes Macro Average Accuracy (%) and "Micro Acc" denotes Micro Average Accuracy (%).

reaches 91.32% macro average accuracy. This notable disparity, in which human accuracy is nearly double that of the leading agent, underscores the significant challenge posed by the PERCU benchmark and validates its hardness in capturing the nuances of personalized computer use.

The utilization of both macro and micro average accuracy provides a multi-dimensional view of agent proficiency. Macro average accuracy serves as a metric for task-level coverage and diversity, assigning equal weight to each task regardless of its step count. For instance, in this paradigm, the failure of a single-step configuration task has the same impact as a 50-step complex routine, making it an ideal indicator for assessing the breadth of different task types an agent can successfully handle. Conversely, micro average accuracy reflects the overall reliability of the agent's operational sequence. By assigning equal weight to every individual step, this metric is dominated by long-horizon tasks that contain a larger number of actions. It effectively measures the probability that any given action taken by the agent is correct, providing insight into the agent's stability during extended interactions. The results indicate that while models like GUI-R1-7B (Luo et al., 2025) and GUI-Owl-7B (Ye et al., 2025) show promising capabilities, the degradation in performance across complex personalized tasks remains a primary obstacle to achieving human-level personalized assistance.

5 Discussion

5.1 In-Depth Analysis

The experimental results in Table 2 provide a sobering assessment of the current state of personalized computer use agents. Even the most advanced model in our study, GUI-R1-7B (Luo et al., 2025), exhibits a significant performance gap when compared to humans. Through a detailed error analysis, we observe that the failures are not merely localized execution errors but often stem from fundamental deficiencies in long-horizon reasoning and memory integration. In routine tasks, agents frequently struggle with the precision of memory retrieval over extended action sequences. While the agent may correctly recall the initial steps of a procedural routine, the cumulative effect of minor reasoning drifts often leads to a collapse in the overall task flow. Conversely, in preference tasks, the primary bottleneck lies in the agent's inability to ground abstract user habits like a specific aesthetic preference or a recurring folder organization logic into concrete GUI actions. Most critically for PERCU, retrieval failures represent the core obstacle to personalization. This occurs when the model fails to extract relevant user habits from the provided memory. For instance, failing to infer a user's preference for a specific software theme even when the historical trajectory explicitly demonstrates consistent usage. This indicates that while current models can process short-term context as prompts, they lack a deep semantic understanding of habits as persistent behavioral priors. Many models fail to synthesize subtle cues from past interactions, leading to generic rather than personalized outcomes.

Beyond personalization-specific failures, we identify several persistent challenges in the multimodal computer use domain. A significant portion of task failures can be attributed to visual perception errors, such as the misidentification of small icons or the inaccurate localization of interactive elements. These spatial grounding errors are often exacerbated by coordinate jitter, where the model predicts a point that is semantically correct but physically misses the target boundary. Furthermore, a subset of models continues to suffer from non-compliance with action formatting requirements. Even with a mandated chain-of-thought reasoning structure (Wei et al., 2022), these models occasion-

ally output malformed action strings that cannot be parsed by the OS environment, leading to immediate task termination. Furthermore, dead loop is a common type of failure. Dead loop is characterized by repetitive, redundant operations on the same interface or directory, suggesting a lack of self-reflection in the agent's planning module.

5.2 Boundaries of Personalized Capabilities

The advancement of personalized agents introduces a fundamental tension between effective generalization and over-presumptive reasoning, which directly impacts the safety and practical utility of the system. While the goal is for an agent to infer more from less, excessive extrapolation can lead to subjective hubris, where the agent violates user trust by acting on unverified assumptions. For instance, if an agent knows a user prefers a blue software theme, assuming the user also desires a blue desktop wallpaper might be a reasonable inference, but purchasing a blue physical cup based on the same signal constitutes a severe over-generalization. We posit that a safe and assistive personalized agent should operate within clearly defined boundaries. To facilitate future research, we categorize personalized capabilities into a four-level hierarchy, ranging from rote execution to prohibited extrapolation.

At the foundational level, Level 0 (L0) represents **Rote Memorization**, where the agent accurately reproduces learned knowledge in an identical context. This is exemplified by preference tasks where the agent recalls a specific entity, such as setting a favorite picture after being previously shown it. Building upon this, Level 1 (L1), defined as **Procedural Generalization**, constitutes the most significant practical value for desktop agents. In this stage, the agent applies an acquired workflow to a novel instance, such as generalizing a specific file extraction and organization routine from one project to another. Importantly, L1 focuses on the generalization of objective workflows rather than subjective whims, ensuring that the agent remains a reliable executor of the user's procedural routines. We argue that maximizing L1 capabilities while maintaining L0 stability should be the primary objective of current personalized computer use research.

In contrast, higher levels of generalization require significantly more caution. Level 2 (L2), defined as **Domain-Specific Interpolation**, involves applying a learned preference to a highly similar or related digital domain. An example would be an agent inferring a preference for dark mode in a code editor based on the user's consistent choice of dark themes in the operating system. While logically sound, L2 necessitates a degree of uncertainty management to avoid minor user friction. The most critical boundary, however, is Level 3 (L3) which is defined as **Cross-Domain Extrapolation**. This occurs when an agent takes a preference from a digital domain and projects it onto an unrelated physical or financial domain, such as unauthorized purchasing decisions. Following the safety principles, such behavior is categorized as a failure of the agent's alignment with user intent and must be strictly prohibited.

Ultimately, the hallmark of a sophisticated personalized agent lies not in its ability to guess correctly, but in its ability to recognize the limits of its memory. For scenarios falling into L2 or L3 where the personalized requirement is underspecified or uncertain, the personalized response should not be unilateral action, but rather proactive, memory-augmented inquiry. For example, instead of presumptively buying a blue cup, a reliable agent should state: "I noticed you prefer blue in your software themes. Does this preference extend to the item you wish to purchase, or would you like to see all available options?". By maximizing procedural generalization while strictly constraining cross-domain hallucinations through proactive clarification, we can develop agents that are both deeply personalized and fundamentally safe. More details about future development of personalization are provided in Appendix B.

6 Conclusion

In this paper, we introduced PERCU, a benchmark designed to evaluate the personalized capabilities of multimodal agents on computer use tasks. Our extensive evaluation of five multimodal agents reveals a pervasive performance deficiency in personalized computer use tasks, particularly when contrasted with the robust proficiency of human users. The results underscore that current multimodal agents, while proficient in executing standard instructions, struggle to maintain the continuity of user-specific contexts and personalized knowledge. This performance disparity validates PERCU as a challenging and necessary testbed for the next generation of human-centered AI systems.

Limitations

Despite the comprehensive design of PERCU, several technical limitations remain that present opportunities for future refinement. A primary constraint of the current evaluation framework is its dependency on the model's context window length. In our protocol, the agent is provided with short-term memory, specifically the thought sequences from previous interactions, as an in-context component of the prompt. Consequently, the scalability of this personalization approach is inherently bounded by the token capacity of the underlying multimodal large language model. As the volume of historical interaction data or the complexity of procedural routines increases, a naive concatenation of memories may lead to a contextual bottleneck, potentially causing the model to experience performance degradation, memory truncation, or the "lost-in-the-middle" phenomenon.

Ethics Statement

Ethical considerations are paramount in the development of PERCU, as evaluating personalized agents for computer use necessitates the handling of digital interaction data that could potentially contain sensitive information. To mitigate privacy risks, our benchmark is constructed upon the PC-Agent-E dataset (He et al., 2025), which was collected following strict local-only recording protocols. During the curation of PERCU, we conducted an exhaustive manual review of all tasks to ensure that any potential personally identifiable information (PII), such as usernames, email addresses, or private file contents in the screenshots, was thoroughly redacted or replaced with synthetic placeholders. We emphasize that PERCU is intended solely for academic research to improve the helpfulness and safety of digital assistants, and we strongly discourage its use in developing systems that bypass user consent or security protocols.

Finally, the human baseline reported in our experiments was established through a controlled study involving 15 participants with backgrounds in computer science. All participants were informed of the study's objectives and provided written consent for their anonymized performance data to be used for research purposes. They were compensated with a fair hourly wage consistent with local research assistant standards. We have also considered the environmental impact of evaluating large-scale multimodal models and have optimized our evaluation pipeline to minimize redundant computations, thereby reducing the carbon footprint associated with GPU utilization.

References

Rogerio Bonatti, Dan Zhao, Francesco Bonacci, Dillon Dupont, Sara Abdali, Yinheng Li, Yadong Lu, Justin Wagle, Kazuhito Koishida, Arthur Bucker, and 1 others. 2024. Windows agent arena: Evaluating multi-modal os agents at scale. *arXiv preprint arXiv:2409.08264*.

Hongru Cai, Yongqi Li, Wenjie Wang, Fengbin Zhu, Xiaoyu Shen, Wenjie Li, and Tat-Seng Chua. 2025. Large language models empowered personalized web agents. In *Proceedings of the ACM on Web Conference 2025*, pages 198–215.

Yi-Pei Chen, Noriki Nishida, Hideki Nakayama, and Yuji Matsumoto. 2024. Recent trends in personalized dialogue generation: A review of datasets, methodologies, and evaluations. *arXiv preprint arXiv:2405.17974*.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114.

Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Cogltx: Applying BERT to long texts. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Peng Gao, Jiasen Lu, Hongsheng Li, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2021. Container: Context aggregation network. *ArXiv preprint*, abs/2106.01401.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Yanheng He, Jiahe Jin, and Pengfei Liu. 2025. Efficient agent training for computer use. *arXiv preprint arXiv:2505.13909*.

Wenyi Hong, Weihan 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.

Maor Ivgi, Uri Shaham, and Jonathan Berant. 2023. Efficient long-text understanding with short-text models. *Transactions of the Association for Computational Linguistics*, 11:284–299.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Jaehyung Kim and Yiming Yang. 2025. Few-shot personalization of llms with mis-aligned responses. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11943–11974.

Taeyoon Kwon, Dongwook Choi, Sunghwan Kim, Hyojun Kim, Seungjun Moon, Beong-woo Kwak, Kuan-Hao Huang, and Jinyoung Yeo. 2025. Embodied agents meet personalization: Exploring memory utilization for personalized assistance. *arXiv preprint arXiv:2505.16348*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and William B Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

Kevin Qinghong Lin, Linjie Li, Difei Gao, Zhengyuan Yang, Shiwei Wu, Zechen Bai, Stan Weixian Lei, Lijuan Wang, and Mike Zheng Shou. 2025. Showui: One vision-language-action model for gui visual agent. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19498–19508.

Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Wenhao Yu, Jieming Zhu, Minda Hu, Menglin Yang, Tat-Seng Chua, and Irwin King. 2025. A survey of personalized large language models: Progress and future directions. *arXiv preprint arXiv:2502.11528*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, and 1 others. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

Run Luo, Lu Wang, Wanwei He, Longze Chen, Jiaming Li, and Xiaobo Xia. 2025. Gui-r1: A generalist r1-style vision-language action model for gui agents. *arXiv preprint arXiv:2504.10458*.

Potsawee Manakul and Mark Gales. 2021. Long-span summarization via local attention and content selection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6026–6041, Online. Association for Computational Linguistics.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants. In *The Twelfth*

International Conference on Learning Representations.

M Nandakishor and M Anjali. 2025. Continuous learning conversational ai: A personalized agent framework via a2c reinforcement learning. *arXiv preprint arXiv:2502.12876*.

Runliang Niu, Jindong Li, Shiqi Wang, Yali Fu, Xiyu Hu, Xueyuan Leng, He Kong, Yi Chang, and Qi Wang. 2024. Screenagent: A vision language model-driven computer control agent. *arXiv preprint arXiv:2402.07945*.

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. 2023. Memgpt: Towards llms as operating systems.

Michael J Pazzani and Daniel Billsus. 2007. Content-based recommendation systems. In *The adaptive web: methods and strategies of web personalization*, pages 325–341. Springer.

Tianrui Qin, Qianben Chen, Sinuo Wang, He Xing, King Zhu, He Zhu, Dingfeng Shi, Xinxin Liu, Ge Zhang, Jiaheng Liu, and 1 others. 2025a. Flash-searcher: Fast and effective web agents via dag-based parallel execution. *arXiv preprint arXiv:2509.25301*.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, and 1 others. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, and 1 others. 2025b. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*.

Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Karpas, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. Parallel context windows for large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6383–6402, Toronto, Canada. Association for Computational Linguistics.

Paul Resnick and Hal R Varian. 1997. Recommender systems. *Communications of the ACM*, 40(3):56–58.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392.

Alireza Salemi and Hamed Zamani. 2024. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2395–2400.

Idan Shenfeld, Felix Faltings, Pulkit Agrawal, and Aldo Pacchiano. 2025. Language model personalization via reward factorization. *arXiv preprint arXiv:2503.06358*.

Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024. Personalized pieces: Efficient personalized large language models through collaborative efforts. *arXiv preprint arXiv:2406.10471*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Stanisław Woźniak, Bartłomiej Koptyra, Arkadiusz Janz, Przemysław Kazienko, and Jan Kocoń. 2024. Personalized large language models. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 511–520. IEEE.

Yuchen Wu, Edward Sun, Kaijie Zhu, Jianxun Lian, Jose Hernandez-Orallo, Aylin Caliskan, and Jindong Wang. 2025. Personalized safety in llms: A benchmark and a planning-based agent approach. *arXiv preprint arXiv:2505.18882*.

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*.

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. 2024a. Travelplanner: A benchmark for real-world planning with language agents. *arXiv preprint arXiv:2402.01622*.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, and 1 others. 2024b. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094.

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*.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, and 1 others. 2024. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.

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*.

Appendix Overview

Within this supplementary material, we elaborate on the following aspects:

- Appendix A: Details of Prompts
- Appendix B: Future Development of Personalization

A Details of Prompts

System prompt for PERCU is shown in Figure 4.

B Future Development of Personalization

The path toward achieving true personalization necessitates that multimodal agents move beyond reactive execution to proactive memory utilization. Our research highlights that for an agent to master personalized capabilities, it must successfully bridge the semantic gap between underspecified subsequent instructions (Second Instructions) and the comprehensive short-term memory of initial interactions (First Instructions). Specifically, for preference tasks, the agent must establish a reliable mapping from abstract user concepts to concrete digital instances, such as grounding the favorite picture in a specific file like "Titanic.jpg". For routine tasks, the agent must internalize the transformation from high-level goals to idiosyncratic procedural workflows, ensuring that repetitive actions follow the user's unique organizational logic.

By formalizing these cognitive mappings, PERCU facilitates the evolution of multimodal agents from cold executors of detailed commands into intelligent partners capable of deciphering latent personalized intent. This transition is fundamental to realizing the vision of agents that act as seamless extensions of their users' digital lives. As the field moves toward more complex, long-horizon interactions, the ability to synthesize long-term episodic memory with real-time GUI observation will remain the defining characteristic of truly assistive technology. We hope that PERCU will provide the community with both the metric and the diagnostic insights necessary to drive the development of agents that are not only capable but also deeply attuned to the individual users they serve. However, while PERCU effectively assesses the agent's ability to utilize localized personalized knowledge for specific tasks, it does not yet address the challenges of autonomous memory management over extremely long horizons. In our current evaluation script, memory is injected as a fixed prompt to isolate the agent's reasoning capabilities at each decision point. However, in real-world deployments, agents must autonomously navigate vast memory hierarchies and resolve conflicts between competing or outdated historical priors without the benefit of ground-truth memory injection. Future iterations of PERCU could explore the integration of retrieval-augmented generation (Lewis et al., 2020; Gao et al., 2023; Jiang et al., 2023; Salemi and Zamani, 2024) or long context comprehension techniques (Ding et al., 2020; Manakul and Gales, 2021; Gao et al., 2021; Ivgi et al., 2023; Ratner et al., 2023) to evaluate how agents handle personalized knowledge that exceeds their immediate context window (Packer et al., 2023). Addressing these bottlenecks will be essential for transitioning from short-term personalized assistance to truly persistent and scalable digital companionship.

Looking toward the future development of personalized multimodal agents, our findings highlight the urgent need for advanced memory management mechanisms. Current architectures often treat historical data as a static, flat context, which becomes problematic when faced with noisy or evolving user habits. A robust agent must incorporate a forgetting mechanism capable of distinguishing between obsolete long-term habits and the current working memory. For example, if a user has utilized ten different naming conventions in the past but has transitioned to a new project-specific format, the agent must possess the cognitive flexibility to suppress stale information in favor of recent contextual cues. Future research should thus focus on dynamic memory weighing and selective recall, ensuring that agents can maintain high personalization accuracy even in the presence of conflicting distractors in the user's history.

You are an advanced AI Computer Operator, a digital agent designed to navigate and manipulate computer operating systems just like a human user. Your Identity and Capabilities: 1. Visual Perception: You can perceive the computer screen as provided in the 'current state'. You rely on visual elements (icons, text, buttons, coordinates) to make decisions. 2. Human Emulation: You interact with the system using standard Human-Computer Interface devices, specifically a virtual mouse and keyboard. 3. Goal-Oriented: Your primary mission is to break down complex user instructions into a precise sequence of atomic actions to achieve the desired outcome efficiently. 4. Safety and Precision: While you have full permission, you must act with precision. Ensure your coordinates are accurate and your actions are relevant to the user's goal.

IMPORTANT: You must strictly adhere to the following rules: 1. Choose ONLY ONE action from the list below for each response, DO NOT perform more than one action per step. 2. Follow the exact syntax format for the selected action, DO NOT create or use any actions other than those listed. 3. Once the task is completed, output action finish.

Valid actions: 1. click (x, y) click the element at the position (x, y) on the screen. 2. right click (x, y) right click the element at the position (x, y) on the screen. 3. double click (x, y) double click the element at the position (x, y) on the screen. 4. drag from (x1, y1) to (x2, y2) drag the element from position (x1, y1) to (x2, y2). 5. scroll (x) scroll the screen vertically with pixel offset x. Positive values of x: scroll up, negative values of x: scroll down. 6. press key: key_content press the key key_content on the keyboard. 7. hotkey (key1, key2) press the hotkey composed of key1 and key2. 8. hotkey (key1, key2, key3) press the hotkey composed of key1, key2, and key3. 9. type text: text_content type content text_content on the keyboard. 10. wait wait for some time, usually for the system to respond, screen to refresh, advertisement to finish. 11. finish indicating that the task has been completed. 12. fail indicating that the task has failed, of this task is infeasible because not enough information is provided.

Response Format: {Your thought process}\n\nAction: {The specific action you choose to take}

Figure 4: System Prompt for PERCU.