TIMO: An End-to-End Framework for Building Digital Twins of Mobile Users

Saranya Venkatraman[1*], Thanh Tran[1*], Shunyan Luo[1], Devin Chen[1], Yuning Wu[1], Kai Wei[1], Allie Colin[1], Guido Imbens[1,2], Ido Rosen[1], Christine Ambrose[1]
Amazon[1], Stanford University[2]
{saranvn,tdt,shunyl,devichen,yuningwu,kaiwe,jonallie,gimbens,idorosen,cambrose}@amazon.com

Abstract

Creating authentic digital twins of mobile users through realistic user behavior simulation is critical to truly understand and anticipate customer needs at scale. Toward this, we introduce Digital **TwI**ns of **MO**bile User (TIMO), an end-to-end framework for high-fidelity mobile user simulation that addresses four key limitations in existing approaches: subjective decision-making diversity, scalable experience retention, simultaneous multi-option processing, and granular mobile interaction fidelity. Our approach integrates adaptive persona generation, self-evolving contextual memory, multi-faceted behavioral simulation, and high-fidelity mobile device control to create digital twins that replicate real user behavior patterns. Unlike existing methods that model decisions in isolation, our framework enables simulated agents to simultaneously evaluate multiple options through joint decision-making architectures, closely mirroring how real users process competing choices within limited mobile screen real estate. Evaluated on both public benchmarks and proprietary datasets from real-world mobile scenarios, our framework demonstrates significant improvements over existing baselines in user decision prediction accuracy, robustness to positional bias, and computational efficiency. Beyond task-related evaluation, we also validate our digital twins through reasoning analysis, and dedicated emotion perception experiments, showing how they align closely with authentic human behavioral patterns, providing a more holistic perspective on scalable simulacrum.

ACM Reference Format:
Saranya Venkatraman[1*], Thanh Tran[1*], Shunyan Luo[1], Devin Chen[1], Yuning Wu[1], Kai Wei[1],, Allie Colin[1], Guido Imbens[1,2], Ido Rosen[1], Christine Ambrose[1]. 2025. TIMO: An End-to-End Framework for Building Digital Twins of Mobile Users. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

User Simulation is critical for understanding customers [54], anticipating their needs and reactions to personalized adaptations [3], driving product discovery [14], and enhancing their overall experience [31, 71] across diverse domains such as e-commerce [5], streaming services [30], social media [32], and financial services [61]. It enables risk-free testing environments for promotional strategies, A/B experiments, and feature rollouts, eliminating the danger of disrupting live at-scale production systems. At Amazon, where millions of daily customers interact primarily through mobile devices (~70% of transactions, with nearly nine-in-ten Americans relying on smartphones [40]), realistic end-to-end user simulation—or *digital twins of customers*—is critical for tackling the most pressing challenges: enhancing customer experience, optimizing personalized recommendations for competitive advantage in dynamic markets, and testing marketing initiatives at unprecedented scale.

However, creating authentic digital twins of users on mobile devices presents significant technical and methodological challenges. First, the subjective nature of human decision-making makes it difficult to model authentic user preferences and behaviors consistently across diverse customer segments [2, 9, 10, 54]. Second, existing approaches evaluate products independently, failing to consider how real customers process multiple competing options simultaneously within limited screen real estate, indicating a need for models that can handle joint decision-making rather than isolated product evaluations [20, 23, 38]. Third, the granular nature of mobile interactions, encompassing touch gestures, scrolling patterns, and micro-interactions, demands high-fidelity device control that traditional simulation approaches do not capture adequately [25, 33, 47]. Fourth, operational constraints around training throughput, memory limitations, and data ingestion create scalability bottlenecks that prevent simulation systems from matching the complexity and scale of real-world customer interactions [12, 13, 44, 71].

To address these challenges, we introduce TIMO, an end-to-end, high-fidelity, and data efficient user behavior simulation framework designed specifically for creating digital twins of mobile users. Our framework comprises four integrated components with design elements to address the existing challenges as follows: (1) To capture subjective decision-making diversity, an individualized and fine-grained persona generator creates realistic customer profiles with nuanced behavioral tendencies, serving as consistent seeds for all simulated interactions and decisions; (2) To handle simultaneous multi-option processing, an interpretable agent behavior simulation module processes persona and memory information to predict user preferences through naturalistic, multi-faceted decision-making, evaluating multiple product options jointly rather than considering each product independently; (3) To achieve granular mobile interactions, a high-fidelity agent device control component translates behavioral decisions into realistic mobile interactions, accurately simulating touch gestures, scrolling patterns, and navigation flows

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

that characterize authentic mobile user experiences; and (4) To enable scalable experience retention, a self-evolving contextualized agent memory module mimics human-like memory organization patterns to dynamically store and retrieve relevant experiences, preferences, and contextual information directly from raw interaction data;

Our framework offers three key contributions to the field of user behavior simulation: (1) multi-faceted joint decision architecture: unlike existing approaches that model user decisions in isolation, we introduce a joint evaluation framework where simulated agents simultaneously compare and rank multiple options, closely mirroring real mobile user behavior patterns and significantly improving prediction accuracy by over 10-20% across tasks; (2) adaptive persona-memory co-evolution: we integrate fine-grained personas with relevant contextual memories that dynamically adapts based on interaction history, enabling more authentic long-term behavioral consistency compared to static profile-based approaches; (3) data-efficient modular user simulation pipeline: we present an end-to-end framework that achieves both high behavioral fidelity and operational efficiency for mobile user simulation.

We evaluate TIMO on both public benchmarks and proprietary datasets from real-world mobile e-commerce scenarios. Our results demonstrate significant improvements over existing baselines (RecAgent [54], Generative Agent [38]) across multiple dimensions: user decision prediction accuracy, robustness to positional bias, and computational efficiency. Our findings demonstrate that agent behavior design must transition from binary to joint processing architectures, recognizing that real-world agency rarely operates in isolation or through sequential determinations. Beyond task completion benchmarks, we also analyze our digital twin's reasoning and authenticity in simulating real human behaviors. Through dedicated emotion perception experiments, we show that our simulated agent responses align closely with real human behavioral patterns. To date, our prototype has led to high-fidelity user simulations on mobile devices, enabling rich insights across offline business metrics. The main contributions of this work are an end-to-end framework for user behavior simulation as well as a novel design for robust and realistic evaluation of digital twin agents.

2 Related Works

Digital twin simulation of users has been explored across recommender systems [54], conversational agents [45], and agent-based modeling [71], but existing approaches fall short of capturing the full complexity of authentic digital twins for mobile contexts. Generative agent frameworks [36] and recent surveys on LLM-driven agent-based modeling [15] demonstrate the potential of lifelike simulacra but focus on open-ended or desktop settings rather than production-scale mobile environments. Interactive recommender agents such as InteRecAgent [27] and RecMind [56] integrate LLM reasoning with recommendation flows, yet typically model single-option decisions and overlook the multi-item, screen-limited comparisons that dominate mobile user behavior. Parallel work on persona conditioning (e.g., PersonaBench [51], PersoBench [1]) and dynamic memory (e.g., MemGPT [35], A-MEM [65]) highlights the importance of long-horizon consistency, but most prior systems rely on static profiles or simplified buffer memories without self-evolving, persona-linked adaptation. From a decision-making perspective, behavioral economics has long underscored the contrast between joint and separate evaluations [21, 22], motivating listwise and choice-to-rank formulations such as Plackett–Luce models [41] and slate bandit methods [50] that explicitly reason over sets rather than isolated items. Complementary strands in mobile GUI control, including AppAgent [70], AutoDroid [60], and benchmarks like Android Agent Arena [6], focus on task completion via tap/scroll action spaces but rarely couple fine-grained device control with realistic cognitive joint decision processes. Finally, evaluation practices in counterfactual learning-to-rank emphasize robustness to exposure and position bias via inverse propensity scoring and slate-aware estimators [28], while recent affective-computing benchmarks (e.g., EMOBENCH [43], EmotionQueen [7]) introduce validation protocols for emotional alignment and authenticity. Our work builds upon these foundations but advances them by integrating joint choice modeling, persona–memory co-evolution, and high-fidelity mobile interaction into a single end-to-end framework, accompanied by holistic evaluation spanning prediction accuracy, bias robustness, and affective realism.

3 Problem Definition

Let $\mathcal{U} = \{u_1, u_2, ..., u_n\}$ denote a set of n real users. For a given target application T, each user u_i has a historical sequence of interactions with application T, denoted as $\mathcal{I}_{T,i} = [e_{i,1}, e_{i,2}, \ldots, e_{i,K_i}]$, where $e_{i,j}$ represents the j-th interaction event in chronological order, and K_i is the total number of interactions for user i within the past X months.

Our objective is to construct a digital twin $\tilde{u}_i$ based on $\mathcal{I}_{T,i}$ that can simulate future interactions $\{\tilde{e}_{i,K_i+l}\}_{l=1}^{L}$, where L represents the prediction horizon of interest. The digital twin should closely replicate the real user's behavior by minimizing a distance metric $d(\cdot,\cdot)$, defined as a binary function that returns 0 if the predicted decision matches the actual decision, and 1 otherwise.:

$$\sum_{l=1}^{L} d(\tilde{e}_{i,K_i+l}, e_{i,K_i+l})$$

In this work, we focus on next-event prediction by setting $L = 1$. In the context of product exploration, each interaction event $e_{i,j}$ can be represented as a tuple:

$$e_{i,j} = (a_{i,j,\mathcal{P}},\ ts,\ \text{loc},\ \text{aux})$$

where $a_{i,j,\mathcal{P}}$ represents the user's decision choice over a set of m products $\mathcal{P} = \{p_1, p_2, ..., p_m\}$, ts denotes the timestamp, loc indicates location, and aux contains additional metadata.

4 TIMO Framework

TIMO is an end-to-end framework for constructing digital twins of mobile users, capable of emulating human-like behavior by autonomously navigating mobile devices, and making informed decisions. These decisions are guided by the users' individual goals and contextual information retrieved dynamically from their historical interactions. Figure 1 presents TIMO which operates in two distinct phases. In Phase 1, a one-to-one mapping between each user and

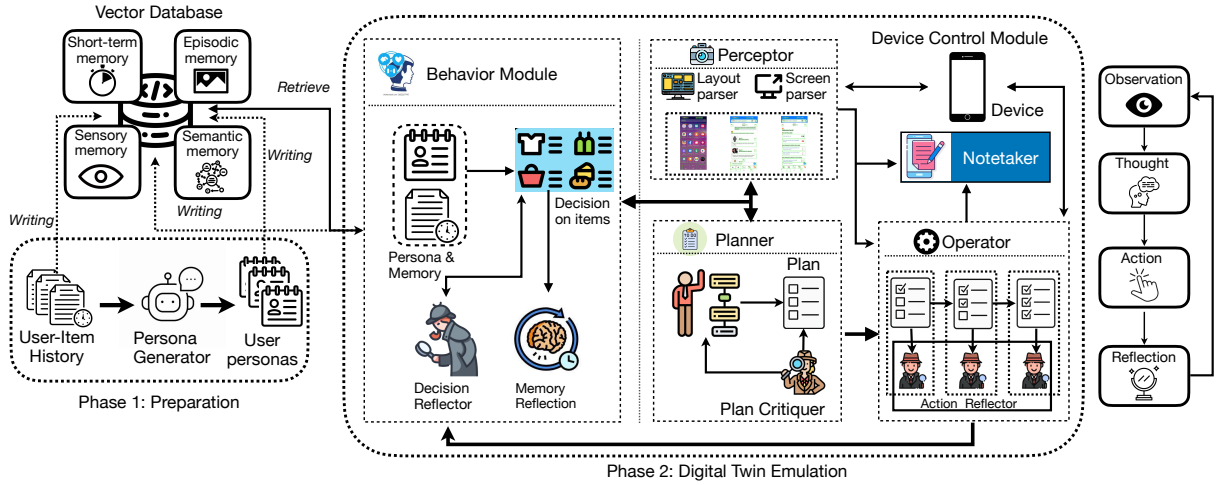


Figure 1: TIMO's overall architecture: Phase 1 involves Preparation stage of Personas and associated Memories to initialize Digital Twins. In Phase 2, Digital Twins interact with and simulate user behaviors on mobile devices.

their corresponding persona is generated, along with initialization of their associated memories. In Phase 2, a relevant or preselected persona together with relevant memories are retrieved to initialize a digital twin, then the digital twin interacts with the mobile application, making decisions via product exploration to mimic realistic and human-like behaviors. We describe individual components of Phase 1 (Persona Generation and Agent Memory) and Phase 2 (Behavior Module and Device Control Module) in detail as follows:

4.1 Individualized and Fine-Grained Persona Generation

A persona is a constructed representation or profile of a user characterized by demographic attributes (like age, occupation), emotional and motivational elements, and affective indicators that provide richer context about a user's deep-rooted preferences and intrinsic motivations. To capture the inherent diversity in subjective decision-making processes, we develop an individualized persona generation method that creates realistic profiles with nuanced behavioral tendencies, preferences and goals, serving as consistent seeds for digital twin interactions and decision-making scenarios across diverse contexts [37, 53, 57, 62, 69]. As direct granular features to construct individual personas at scale are usually unavailable, we use a statistical model to infer persona using the user's past interaction history. We leverage a Large Language Model (LLM) for generating personas, given empirical evidence demonstrating their superior performance in generating nuanced and flexible persona representations [16, 69].

Specifically, for a user u_i with historical interaction data $\mathcal{I}_{T,i} = [e_{i,1}, e_{i,2}, \ldots, e_{i,K_i}]$, where $e_{i,j}$ represents the j-th interaction event in chronological order, and K_i is the total number of interactions for user i on application T, we prompt an LLM to generate a detailed fine-grained profile summary. This profile explicitly requests the LLM to infer the user's preferences (e.g., "environmentally-aware and budget-conscious"), goals (e.g., "find natural beauty products"), and demographic attributes (e.g., "middle-aged"). Following the methodology described by [3], we sample the most recent 50 interactions from the user's interaction history $\mathcal{I}_{T,i}$ to construct this summary. In line with prior findings that non-preferred items can yield valuable insights into user preference boundaries [52], we simulate a realistic browsing scenario in which users are exposed to a range of relevant products during product search sessions. Items that align with user u_i's preferences (such as cost, color, or brand) are captured in the user's interaction history $\mathcal{I}_{T,i}$. In contrast, *impression items*, which are similar but received no engagement, are often overlooked despite offering more granular signals about user preferences. To identify these impression items, we employ a text encoder to encode each item into embeddings. We use the *all-MiniLM-L6-v2* model for its effectiveness in retrieval tasks. We use item p_k's embeddings as query embeddings. In the same manner, we encode the remaining items to obtain candidate embeddings, then retrieve the *top-k* candidate items that have the highest cosine similarity scores with the query as impression items. To match with typical user interface patterns such as the top 4-6 products displayed per search in popular mobile applications (e.g., Rufus for shopping, Prime Video for movies), we set *top-k* to 5. In addition, to ensure privacy we also incorporate explicit policy constraints within the persona generation prompts. These instructions direct the model to generalize outputs based on behavioral patterns rather than stereotypes or surface-level user details, and to avoid including personally identifiable information. This prompt-level control helps maintain privacy while still enabling meaningful user simulation.

4.2 Self-Evolving Contextualized Agent Memory

4.2.1 Memory Structure. Inspired by the hierarchy of human memory [48, 66], TIMO's memory is structured into four distinct memory segments including *episodic memory*, *semantic memory*, *short-term memory* and *sensory memory*. *Episodic memory* stores past experiences or events (often with temporal context) or user-item

interactions. It enables a digital twin to record and later retrieve *episodes* of interactions or observations, analogous to a personal diary. For instance, remembering *the purchase of a pair of Nike white shoes at $59* is episodic. By retaining a chronological trace of digital twin's experiences, episodic memory supports temporal reasoning, planning, and learning from past outcomes [11]. *Semantic memory* holds general knowledge, facts, and concepts about the world or the digital twin itself, regardless of when/where they were learned. For instance, remembering *Nike is a popular shoe brand* is a semantic memory. Unlike episodic memory, semantic memory is not tied to an individual's experience of an event alone, but a dynamic knowledge base, which can be shared across a group of people. Note that episodic and semantic memory can interact with each other. Episodic experiences can be abstracted into semantic knowledge, and semantic memories provide context to make decisions for new episodes. *Short-term memory* (often called *working memory*) is a limited-capacity memory segment for information actively used in current tasks. In LLM-based agents [49], an agent's short-term memory includes the latest few user messages and recent interactions. Inspired by Miller's Law [34] which stated that a human being can hold 7±2 items in their short-term memory, we store the latest 9 interactions into the short-term memory. *Sensory memory* contains multi-modal perceptual inputs such as product images, demonstration videos, and interface signals, providing real-time grounding for fine-grained behavioral decisions. We construct each memory segment as a data collection in Vector Database for simplicity and fast retrieval. We store the generated personas into semantic memory, while interaction data are stored in short-term and episodic memories. In the next section, we describe persona and memory retrieval with Vector Database.

4.2.2 Persona and Memory Initialization via Retrieval. To efficiently access and retrieve relevant personas/memories during simulation, among existing retrieval techniques including graph neural networks [19], graph embeddings [17, 39], multi-interest methods [64], auto encoder based methods [55], and collaborative filtering we adopt a two-tower retrieval architecture [26] to enable efficient and scalable retrieval. To support multi-modal retrieval capabilities, we encode both textual and visual data into embeddings using two state-of-the-art encoders, i.e., *multilingual-e5-large-instruct* for text and *CLIP-ViT-B/32* for images. The resulting persona embeddings, along with each user's historical interaction data $\mathcal{I}_{T,i}$ are stored in corresponding memory collections in Vector Database. To mitigate prompt length limitations introduced by raw memory events, which often contain excessive tokens and may lead to prompt overflow [69], we implement a memory reflection mechanism. This mechanism uses an LLM to generate concise summaries of each memory event. These summaries are subsequently encoded into embeddings and stored in Vector Database for compact and contextual memory retrieval. Our approach of persona memory retrieval enables robust, multi-modal retrieval across human languages and content types (e.g. text, image, and text + image), enhancing the TIMO's ability to access relevant and highly contextualized memories accurately and efficiently.

4.3 Joint Multi-Option Decision Processing Behavior Module

The *Behavior Module* can be seen as the brain of the digital twin to simulate realistic, individual-level customer decision-making. It works by synthesizing and inferring user preferences and predicting next-step actions. Our joint multi-option decision processing approach is inspired by Hsee's Evaluability Hypothesis [21–23], which demonstrated that easily comparable attributes disproportionately influence decisions when options are evaluated simultaneously. This finding is particularly relevant for mobile e-commerce, where users rapidly scroll through multiple products, making comparative judgments. This joint processing, unlike binary decision making, where consumers evaluate options sequentially in isolation (separate evaluation), fundamentally alters preference structures, a phenomenon known as preference reversal [24]. This seminal work demonstrated that consumers may strongly prefer Product X when evaluating options individually, yet switch to Product Y when presented with both simultaneously, as such attributes gain significance in joint evaluation contexts as soon as direct comparisons are possible. Following both these findings, our Behavior Module includes easily comparable numeric attributes like prices and review ratings in its comparative assessment, as these quantitative features become especially salient when users evaluate multiple options concurrently. Thus, we designed our chain-of-thought reasoning architecture to anchor on these directly comparable attributes. This mirrors how human shoppers naturally gravitate toward clear, quantitative comparisons (e.g., "this product is $15 cheaper" or "has 2000 more reviews") when faced with multiple options on a small mobile screen. This approach represents a significant advance over traditional binary decision models by considering and better approximating the comparative reasoning processes used by real mobile users.

The *Behavior Module* prompt contains key sections including: persona, relevant episodic memory, relevant semantic memory, short-term memory, and current on-going memory. The current ongoing memory contains the active product search mission, and the current environment information (i.e. the content displayed in the currently interacting target mobile app $\mathcal{T}$). The full prompt template is provided in the Appendix (Figure 4). The *Behavior Module* operates via four key steps as following:

Initialization: Digital twin is assigned a persona, either by retrieving one that matches predefined search criteria (e.g., a "shoe lover") or by using a specific persona. Before starting a product exploration mission, the digital twin's memory is populated with relevant information retrieved from all four memory segments in the vector database using the retrieval methodology we described earlier. The current ongoing memory is initialized with the product search mission and the displayed contents which include product list. The displayed product list is extracted by the Device Control Module (details in Section 4.4), as TIMO keeps scrolling the screen, taking screenshots for parsing and combining the product content using a Vision Language Model (VLM).

Multi-criteria Item Assessment: Following the principle of least-to-most prompting [72], the digital twin is requested to reason its alignment with the shown list of items step by step over three critical subproblems that each focus on the alignment of the items

(1) with the digital twin's preferences, (2) with historical behavior, and (3) with the current search intent.

Decision Refinement with Feedback Loop : To mimic how real users interact with products on mobile applications, TIMO defines a granular and interest-descending ordered decision schema tailored to specific domains. For example, {[BUY-NOW], [ADD-TO-CART], [CLICK], [SKIP]} in shopping domain, or {[WATCH-NOW], [ADD-TO-WATCH-LIST], [WATCH-TRAILER] and [SKIP]} in streaming domain, with each decision clearly defined during the item assessment process above. Inspired by recent work [4] where feedback loop improved agent performance, we introduce a Behavior Module refinement mechanism. In this setup, a separate feedback provider agent reviews the selected decision and critiques whether it aligns with the agent's stated reasoning and the user's preferences. If inconsistency is detected, the agent enters a self-reflection step and revises its final decision accordingly.

Memory Reflection: Similar to a real user, digital twin makes multiple actions across multiple products on the mobile application, where each item has related metadata (e.g. detailed description, image, .etc). Recording raw action events from every interaction round can quickly exceed the prompt size limit and introduce unnecessary noise. To mitigate this, we introduce a memory reflection process, where we prompt LLM to summarize the action into higher level inferences over time, enabling the digital twin to draw conclusions about itself. These memory reflections are fed back into the memory stream and stored in the memory database to influence the digital twin's future behavior.

4.4 High-Granularity Device Control Module

The *Device Control Module* manages low-level high-granularity interactions with the target mobile application $\mathcal{T}$ to execute tasks from the behavior module. Using a layout understanding model, this module identifies GUI components (e.g., buttons, text boxes, images) and performs common human-device interactions such as tapping, scrolling, swiping, and typing. Inspired by Kahneman's fast and slow thinking systems [29], we implement a dual-process architecture combining the customized *ReAct* agent architecture [67] (System 1) for rapid, specialized UI interactions with Mobile-Agent-E [58] (System 2) for deliberate processing. System 1, based on *ReAct*, provides quick responses to time-sensitive content through an iterative cycle: *Observation* (gathering UI information), *Thought* (strategic reasoning), *Action* (executing interactions), customized *Reflection* to evaluate outcomes and serve as a trigger of system switch, and *Reiteration* (continuing until task completion). When the reflection module detects potential failures or complex tasks, TIMO switches to System 2 (Mobile-Agent-E), which comprises five specialized submodules: (1) Planner, which decomposes tasks into tractable sub-goals to support step-by-step execution and result verification; (2) Perceptor, which interprets visual elements (icons, images, and text) and their spatial relationships on the screen; (3) Operator, which executes actions on the device according to the Planner's sub-goals; (4) Action Reflector, which evaluates whether actions yield the expected outcomes by comparing current and historical screenshots and updating the agent's internal state (e.g., confirming that a product was added to the cart); and (5) Notetaker, which maintains critical session-level information, tracking important signals such as selected products, filled input fields, and click history. This component optimizes memory usage across modalities (text, images, and clicks) to enhance system performance. In our implementation, we adopt a two-stage Perceptor architecture combining grounding and captioning capabilities. For layout detection, we employ a fine-tuned Faster R-CNN model [42] using human annotated data for training to identify GUI component bounding boxes, chosen for its lightweight architecture and efficient performance. The captioning stage utilizes the pre-trained Florence model [63] and corresponding OCR models from OmniParser-v2 [68] combined with XML data to generate textual metadata, enabling rapid experimentation in real-device environments.

5 Experimental Results

We now present results on the application and comparative evaluation of TIMO-based digital twins for 2 real-world ([Internal] Amazon Datasets) as well as a public benchmark (MovieLens)-based scenarios against baseline agents for user simulation (RecAgent [54] and Generative Agents [38]).

5.1 Benchmark Adaptation

5.1.1 Hard Negative Sampling for Joint Decision Simulation. We introduce a novel evaluation framework that replicates authentic user decision scenarios and demonstrates how existing benchmarks can be adapted for more realistic and sophisticated evaluation without requiring any new data collection. When browsing a shopping platform, real users simultaneously process and evaluate multiple options, ultimately selecting a subset that aligns with their preferences and current needs. Shopping-related decisions rarely follow a binary decision-making process where customers evaluate items sequentially in isolation. Yet, current evaluation benchmarks lack the capability to simulate or assess such multi-item decision scenarios. For instance, contemporary item ranking benchmarks typically rely on random negative sampling [54], which fails to capture the cognitive complexity of choosing between semantically similar items (e.g., selecting among multiple cameras versus choosing a camera from an assortment of unrelated products). We modify **MovieLens** [18] and a proprietary **[Internal] Amazon Dataset1** for this multi-item decision making setting. The former comprises movie ratings, while the latter consists of user interactions across multiple product domains on Amazon.com. Using these datasets, we find challenging negative samples for each user interaction similar to competitive items a user processes at a time. First, we select a random sample of 1k users and implement a chronological split of their activity data, allocating the initial 70% to the training set and the remaining 30% to the test set. Subsequently, we employ the *all-MiniLM-L6-v2* model to embed all items (movies or products) into a vector space. For each user-movie or user-product pair in the test set, we select 4 non-interacted movie or product titles with the shortest L2 distance from the positive item, with an additional criteria for movie titles to have at least one genre in common. Through this method, we replicate realistic decision scenarios as can be seen from examples in the Appendix (Table 3).

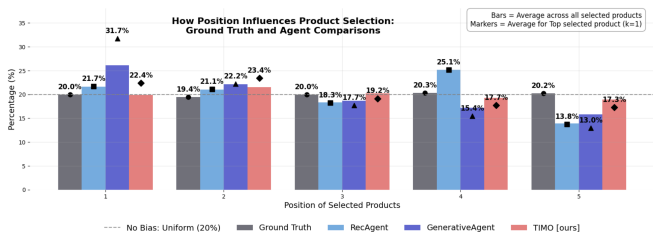
		[Internal] Amazon Dataset1						[Internal] Amazon Dataset2			MovieLens		
		Clothes, Shoes & Jewelry			Pet Supplies								
Agent	k	P@k	R@k	F1@k	P@k	R@k	F1@k	P@k	R@k	F1@k	P@k	R@k	F1@k
Gen	1	**20.52%**	**20.52%**	**20.52%**	**23.48%**	**23.48%**	**23.48%**	baseline	baseline	baseline	32.30%	32.30%	32.30%
Rec	1	17.06%	17.06%	17.06%	18.96%	18.96%	18.96%	↑12.06%	↑12.06%	↑12.06%	50.02%	50.02%	50.02%
TIMO	1	**20.52%**	**20.52%**	**20.52%**	22.28%	22.28%	22.28%	**↑19.90%**	**↑19.90%**	**↑19.90%**	**52.05%**	**52.05%**	**52.05%**
Gen	2	13.69%	27.39%	18.26%	15.44%	30.89%	20.59%	baseline	baseline	baseline	23.20%	46.40%	30.94%
Rec	2	8.61%	17.22%	11.48%	9.76%	19.52%	13.01%	↑13.14%	↑26.29%	↑17.52%	35.86%	71.72%	47.81%
TIMO	2	**18.29%**	**36.58%**	**24.39%**	**19.42%**	**38.84%**	**25.89%**	**↑15.54%**	**↑31.10%**	**↑20.72%**	**38.90%**	**72.80%**	**50.71%**
Gen	3	9.26%	27.78%	13.89%	10.41%	31.24%	15.62%	baseline	baseline	baseline	16.05%	48.15%	24.07%
Rec	3	5.74%	17.22%	8.61%	6.51%	19.54%	9.77%	↑12.87%	↑38.62%	↑19.31%	**27.07%**	**81.24%**	**40.61%**
TIMO	3	**13.58%**	**40.76%**	**20.38%**	**14.14%**	**42.42%**	**21.21%**	**↑14.10%**	**↑42.29%**	**↑21.14%**	26.84%	80.52%	40.26%
Gen	4	6.94%	27.78%	11.11%	7.81%	31.24%	12.49%	baseline	baseline	baseline	12.05%	48.21%	19.28%
Rec	4	4.31%	17.22%	6.89%	4.89%	19.54%	7.82%	**↑11.44%**	**↑45.74%**	**↑18.29%**	**21.45%**	**85.80%**	**34.32%**
TIMO	4	**10.30%**	**41.18%**	**16.47%**	**10.68%**	**42.70%**	**17.08%**	↑11.33%	↑45.30%	↑18.12%	20.60%	82.40%	32.96%
Gen	5	5.56%	27.78%	9.26%	6.25%	31.24%	10.41%	baseline	baseline	baseline	9.65%	48.27%	16.09%
Rec	5	3.44%	17.22%	5.73%	3.91%	19.54%	6.52%	**↑9.67%**	**↑48.35%**	**↑16.12%**	**17.46%**	**87.30%**	**29.10%**
TIMO	5	**8.25%**	**41.26%**	**13.75%**	**8.54%**	**42.72%**	**14.24%**	↑9.08%	↑45.43%	↑15.14%	16.51%	82.56%	27.52%
Gen	Avg	11.19%	26.25%	14.61%	12.68%	29.62%	16.52%	baseline	baseline	baseline	18.65%	44.66%	24.54%
Rec	Avg	7.83%	17.19%	9.96%	8.81%	19.42%	11.22%	↑11.84%	↑34.21%	↑16.66%	30.37%	**75.22%**	40.37%
TIMO	Avg	**14.19%**	**36.06%**	**19.10%**	**15.01%**	**37.79%**	**21.62%**	**↑13.99%**	**↑36.80%**	**↑19.00%**	**30.98%**	74.07%	**40.70%**

Table 1: Performance comparison for pick any out of 5 task for 3 datasets. [Gen = Generative Agent, Rec = RecAgent, Avg=Average] Best results for each k value are shown in bold. P@k, R@k and F1@k denote Precision, Recall and F1-scores, respectively. Results for [Internal] Amazon Dataset2 are shown in increments as per data sharing policies.

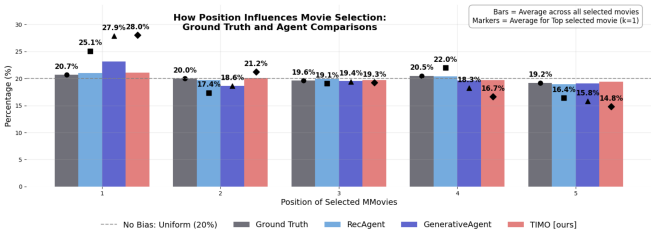
5.1.2 Real World Joint Decision Dataset. To validate our digital twins in practice, we evaluate on a proprietary real-world joint decision dataset denoted as [Internal] Amazon Dataset2. This dataset captures authentic customer-product interactions from production search and recommendation system from Amazon.com, consisting of both presented items and the real user decisions in the presence of competing products on real devices. Specifically, for each customer search query (e.g., "men's shoe size 8"), the dataset records: (1) the complete set of products displayed in search results, and (2) granular user engagement signals including clicks, purchases, and cart additions. This dataset offers several advantages for evaluation. First, it provides authentic user behavior patterns that reflect actual customer preferences and decision-making processes, capturing the complexity of real user interactions. Second, the dataset naturally contains implicit negative samples as items presented to users but not engaged with, eliminating the need for artificial negative sampling strategies that can introduce bias. Third, data from real users at scale enables assessment of our framework's performance under realistic deployment conditions, providing insights into scalability and robustness that simulated settings cannot replicate. The inclusion of this proprietary dataset strengthens our empirical evaluation through a rigorous and more realistic assessment of agents when deployed on a real e-commerce platform such as Amazon.com.

5.2 Multi-Faceted Decision Making

5.2.1 Click Prediction Task. Using the hard negative setting from benchmark adaptation ([Internal] Amazon Dataset1 and MovieLens), we evaluate agents for a click prediction task given a list of 5 products or movies. Our hard negative sampling enables this evaluation to assess how well agents can identity the product purchased or movie watched by the user in the presence of competing items. For [Internal] Amazon Dataset2, we select actual negatives from the top seen-but-not-interacted products and the ground truth positive as the one the user purchased after looking through the product list. Having shown 5 competing options that are similar, we ask the agents to use the interest-descending decision framework (Section 4.3) over the list of products jointly, and evaluate their picked choices using Precision@k, Recall@k, and F1@k for $K \in [1, 5]$. We compare TIMO with 2 other agents: Generative Agent [37] and RecAgent [54]. [Internal] Amazon Dataset1 contains customer interactions with products from multiple domains. We present results for 2 such domains that have high volume of interactions per user. From Table 1, we see that overall (averaged across values of k) TIMO performs better than RecAgent and Generative Agent for both [Internal] Amazon Dataset1 and MovieLens datasets. Agents show better performance for MovieLens or movie prediction task as compared to product shopping domain. For "Pet Supplies", Generative Agent surpassed TIMO's F1-score by 1.2% for $k = 1$ setting. For MovieLens data, for $k >= 3$, RecAgent performs best, while TIMO remains advantageous in identifying the correct movie within the top 2 picks. This task is challenging due to our negative sampling strategy as well as the joint decision task setup. The difficulty of our realistic evaluation settings is reflected by the minimum F1 scores of 17% and 32% for [Internal] Amazon Dataset1 and MovieLens, respectively at $k = 1$. For [Internal] Amazon Dataset2, we are limited to presenting results in increments from a baseline as this is proprietary data collected from real granular customer interactions (as can be seen in Table 1). We see that both RecAgent and TIMO outperform the baseline Generative Agent across all metrics and values of k, with TIMO performing better at lower values of $k < 4$, and an overall increase of 19% F1 from the baseline. Our findings demonstrate the effectiveness of targeted design adaptations in advancing agent capabilities, while underscoring the necessity of evaluation frameworks that reflect real-world decision complexity.



(a) [Internal] Amazon Dataset1



(b) MovieLens

Figure 2: Position Bias is Agent's chosen items. Are agents biased to picking items shown first? Y-axis shows the % of products or movies clicked that were shown in different positions in a list, from 1 to 5 [l→r] on the X-axis. We also indicate the distribution over positions of the first item chosen by the agents using the black markers. Agents tend to pick the item in position 1 as the first click more often than those further down the list.

[Internal] Amazon Dataset1 (Select Any from 5)			
Avg. F1@K	Generative Agent	RecAgent	TIMO
GT @ First	21.68%	20.06%	20.28%
First → Middle	↓6.82%	↓0.24%	↓1.10%
Middle → Last	↓3.10%	↓6.84%	↓0.76%
MovieLens (Select Any from 5)			
Avg. F1@K	Generative Agent	RecAgent	TIMO
GT @ First	26.66%	47.92%	43.34%
First → Middle	↓3.11%	↓2.54%	↓3.64%
Middle → Last	↑0.45%	↓1.38%	↓1.16%

Table 2: Performance (Average F1-scores) for $K \in [1, 5]$ when the ground truth (GT) is presented in different positions of the list. GT @ First denotes the F1 score when the ground truth item is placed first [position=1], and subsequent rows denote the drop in relative performance as ground truth item is shown either in middle positions [$2 \rightarrow 4$] or last [5].

5.2.2 Position Bias. Position bias is a well-documented characteristic of LLM-based agents, where model behavior varies based on the sequential ordering of input elements [46]. To quantify the impact of this bias on TIMO's decision-making process, we conduct a two-fold analysis:

Distribution of Agent Clicks by Position. We analyze the distribution of selection frequencies across input item positions. Our analysis in Figure 2 shows that all agents exhibit position bias, though with varying intensities. Generative Agent demonstrates the most pronounced bias with extreme variations across positions (26% at position 1 to 15.4% at position 5 for products), while TIMO maintains the most uniform distribution, staying closest to the ideal 20% baseline across positions. This consistent pattern suggests built-in bias mitigation in TIMO's architecture due to its unique list-based decision making that considers all items at once. Domain-specific analysis further reveals that position bias is more severe in product recommendations (first-position selections:

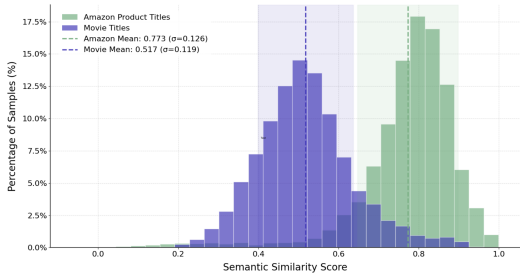


Figure 3: Distribution of cosine similarities between positive (ground truth) and negative product/movie titles derived from [Internal] Amazon Dataset1 and MovieLens datasets, respectively. The higher mean similarity in Amazon product titles ($\mu = 0.77$) compared to movie titles ($\mu = 0.51$) suggests that movie titles are more distinct than product titles.

RecAgent: 21.7%, Generative Agent: 26%, TIMO: 22.4%) compared to movie recommendations (RecAgent: 21.4%, Generative Agent: 22.1%, TIMO: 21.8%). The movie domain maintains a more uniform distribution across positions (15-22% range) versus the steep decline observed in product recommendations (13-17% at position 5). This domain disparity likely stems from fundamental differences in negative sampling, where product titles share more semantic similarity compared to more distinct movie titles as shown in Figure 3. Despite varying magnitudes, the consistent bias patterns across agents indicate a systematic challenge in LLM-driven agentic frameworks, with TIMO showing promising bias mitigation capabilities.

Click Prediction Performance by Ground Truth Position. We also evaluate click prediction performance when the ground truth item (the actual product purchased or movie watched) appears at different positions in the 5-item list. For each position category (first: position 1, middle: positions $2 \rightarrow 4$, and last: position 5), we show average F1 scores across all K values ($K \in [1, 5]$) in Table 2, capturing the agents' ability to identify the correct item regardless of its position. In Table 2, the sequential performance drops (indicated by ↓) represent the relative degradation in F1 scores as the ground truth item moves from favorable (first) to less favorable (middle

and last) positions. For MovieLens, agents show less variance in cumulative performance loss (2.66% to 4.8%). In contrast, for Amazon products, Generative Agent exhibits the highest sensitivity to position changes, with a cumulative F1 score reduction of 10% (6.82% from first to middle, 3.10% from middle to last positions), while TIMO demonstrates superior robustness with minimal degradation (1.86% total loss), further supporting evidence for position-invariant agent decision-making.

5.2.3 Efficiency Analysis. Runtime analysis show that Generative Agent takes approximately 7 minutes per test iteration, RecAgent takes 3.53 minutes whereas TIMO operates at a faster rate of 2 minutes per iteration, which is 3.5× faster than the Generative Agent and 1.77× faster than RecAgent. In terms of cost, while both baselines require $O(nk)$ LLM calls to process n lists of k items, our approach requires only $O(n)$ calls, yielding a k-fold reduction.

5.2.4 Reasoning Analysis. We evaluate agent reasoning strategies using an LLM-as-a-judge evaluation followed by manual human-in-the-loop verification [8]. We obtained binary (Yes/No) judgments and single-sentence justifications about agent reasoning using questions on six aspects: customer preference integration, logical validity, decision clarity and transparency, price-value analysis, alternatives consideration, and negative preference incorporation (details and results in Appendix (Table 5)). We find that TIMO demonstrates superiority in core user-centric metrics, achieving 96.4% in preference alignment (δ = +34.6% compared to Generative Agent, δ = +44.5% compared to RecAgent). While the Generative Agent tends to rely on price-value analysis much more frequently (86.9%) than TIMO (43.2%), our empirical click decision prediction results indicate that this over-reliance on price as a reasoning strategy might be counterproductive for behavior simulation, and provides support for a more nuanced balance between price considerations, other product attributes and user preference. Lower occurrences of alternatives consideration (14.7%) and negative preference incorporation (12.9%) across all agents' reasoning may indicate an opportunity for future architectural improvements to help agents leverage negative signals from user behaviors.

5.3 Device Perception Success Rate

We collected 72 trajectories through screenshots of TIMO's device control module as well as human subjects interacting with a mobile device to carry out a set of given tasks. We then collected high-quality expert annotations from two human annotators hired from external vendors, to evaluate whether the task was completed successfully and followed the same trajectory as the human subjects. In case of disagreements, a third annotator was engaged to make the final decision on success or failure. Under this evaluation framework, our agent achieved a 91.7% success rate, reflecting the fine-grained high-fidelity of TIMO's device control capabilities. Among the failed cases, 66.66% completed the assigned tasks but required extra steps compared to humans. Examples of succeeded and failed trajectories are provided in the Appendix (Figure 5).

5.4 Affective Realism

As a means to evaluate if TIMO digital twins exhibits behaviors that are true to real humans that it aims to model and act like, we replicated a human emotion perception experiment by Wei et al. [59]. We initialized TIMO using demographic details of human subjects from the study as a means to simulate the participants of the original experiment. We showed control group agents neutral news articles about a specific subpopulation, while the experimental group agents were presented with articles framing the subpopulation as cultural or economic threats, with both agent groups receiving identical stimuli to their human counterparts. We then asked our agents the same survey questions that were asked of the human subjects, and compared the emotional responses of agents v/s humans. We found that the treatment effect for both negative and positive emotion had the same direction and similar magnitude in agents as in the human subjects (refer to Appendix (Table 4) for detailed results). Both humans and our digital twins exhibited nearly identical emotional responses to the experimental treatment. For negative emotions, humans showed a significant increase of 1.179 units ($StandardError(SE) = 0.2, t = 5.76, p < 0.001$), while agents demonstrated a comparable increase of 1.195 units ($SE = 0.11, t = 11.25, p < 0.001$). Neither humans nor agents displayed statistically significant changes in positive emotions (humans: -0.065 units, $SE = 0.18, p = 0.72$; agents: -0.057 units, $SE = 0.06, p = 0.31$). Notably, while the effect magnitudes were strikingly similar across both groups, the agent simulations produced more precise estimates with substantially lower standard errors, resulting in higher t-statistics despite the nearly identical treatment effects. This also indicated that agents behave with less variability to each other than humans do. This replication study provides empirical evidence that our agents exhibit emotional responses consistent with human subjects when presented with identical stimuli, supporting the behavioral validity of our framework.

6 Conclusion

In this work, we introduced an end-to-end framework for constructing high-fidelity digital twins of mobile users through realistic behavior simulation. By unifying individualized persona generation, self-evolving contextual memory, multi-faceted joint decision-making, and fine-grained mobile device control, our approach addresses four long-standing limitations of prior user simulators: subjective decision diversity, scalable experience retention, simultaneous multi-option processing, and granular interaction fidelity. Extensive experiments on both public benchmarks and proprietary mobile e-commerce datasets demonstrate that our framework not only improves prediction accuracy, robustness to positional bias, and computational efficiency, but also yields simulated agents whose responses align more closely with authentic human behavioral and emotional patterns. These results highlight the importance of moving beyond isolated decision models toward joint, contextually grounded architectures that faithfully mirror real-world user cognition and interaction. Looking ahead, we envision extending this framework toward broader application domains, including cross-device multi-turn digital twins, adaptive personalization through reinforcement learning, and large-scale simulation platforms for trustworthy evaluation of interactive systems. By advancing the fidelity, scalability, and holistic evaluation of user digital twins, our work provides a foundation for more human-centric user simulation at scale.

References

[1] Saleh Afzoon, Usman Naseem, Amin Beheshti, and Zahra Jamali. 2024. Persobench: Benchmarking personalized response generation in large language models. *arXiv preprint arXiv:2410.03198* (2024).

[2] Paula Akemi Aoyagui, Kelsey Stemmler, Sharon A Ferguson, Young-Ho Kim, and Anastasia Kuzminykh. 2025. A matter of perspective (s): Contrasting human and llm argumentation in subjective decision-making on subtle sexism. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–16.

[3] Nicolas Bougie and Narimasa Watanabe. 2025. Simuser: Simulating user behavior with large language models for recommender system evaluation. *arXiv preprint arXiv:2504.12722* (2025).

[4] Shihao Cai, Jizhi Zhang, Keqin Bao, Chongming Gao, Qifan Wang, Fuli Feng, and Xiangnan He. 2025. Agentic feedback loop modeling improves recommendation and user simulation. In *Proceedings of the 48th International ACM SIGIR conference on Research and Development in Information Retrieval*. 2235–2244.

[5] Sava Čavoški and Aleksandar Marković. 2017. Agent-based modelling and simulation in the analysis of customer behaviour on B2C e-commerce sites. *Journal of Simulation* 11, 4 (2017), 335–345.

[6] Yuxiang Chai, Hanhao Li, Jiayu Zhang, Liang Liu, Guangyi Liu, Guozhi Wang, Shuai Ren, Siyuan Huang, and Hongsheng Li. 2025. A3: Android agent arena for mobile gui agents. *arXiv preprint arXiv:2501.01149* (2025).

[7] Yuyan Chen, Hao Wang, Songzhou Yan, Sijia Liu, Yueze Li, Yi Zhao, and Yanghua Xiao. 2024. Emotionqueen: A benchmark for evaluating empathy of large language models. *arXiv preprint arXiv:2409.13359* (2024).

[8] Cheng-Han Chiang and Hung-Yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 15607–15631.

[9] Preetam Prabhu Srikar Dammu, Omar Alonso, and Barbara Poblete. 2025. A shopping agent for addressing subjective product needs. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*. 1032–1035.

[10] Marieke De Vries, Rob W Holland, and Cilia LM Witteman. 2008. Fitting decisions: Mood and intuitive versus deliberative decision strategies. *Cognition and Emotion* 22, 5 (2008), 931–943.

[11] Chad DeChant. 2025. Episodic memory in ai agents poses risks that should be studied and mitigated. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 321–332.

[12] Xiaofei Dong, Xueqiang Zhang, Weixin Bu, Dan Zhang, and Feng Cao. 2024. A survey of llm-based agents: Theories, technologies, applications and suggestions. In *2024 3rd International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC)*. IEEE, 407–413.

[13] Zhichao Feng, JunJie Xie, Kaiyuan Li, Yu Qin, Pengfei Wang, Qianzhong Li, Bin Yin, Xiang Li, Wei Lin, and Shangguang Wang. 2024. Context-based Fast Recommendation Strategy for Long User Behavior Sequence in Meituan Waimai. In *Companion Proceedings of the ACM Web Conference 2024*. 355–363.

[14] Johann Füller and Kurt Matzler. 2007. Virtual product experience and customer participation—A chance for customer-centred, really new products. *Technovation* 27, 6-7 (2007), 378–387.

[15] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–24. Publisher: Palgrave.

[16] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094* (2024).

[17] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 855–864.

[18] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015), 1–19.

[19] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 639–648.

[20] Christopher K Hsee. 1996. The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational behavior and human decision processes* 67, 3 (1996), 247–257.

[21] Christopher K Hsee. 1996. The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational behavior and human decision processes* 67, 3 (1996), 247–257. Publisher: Elsevier.

[22] Christopher K Hsee. 1998. Less is better: When low-value options are valued more highly than high-value options. *Journal of Behavioral Decision Making* 11, 2 (1998), 107–121. Publisher: Wiley Online Library.

[23] Christopher K Hsee and France Leclerc. 1998. Will products look more attractive when presented separately or together? *Journal of Consumer Research* 25, 2 (1998), 175–186.

[24] Christopher K Hsee and France Leclerc. 1998. Will products look more attractive when presented separately or together? *Journal of Consumer Research* 25, 2 (1998), 175–186.

[25] Xueyu Hu, Tao Xiong, Biao Yi, Zishu Wei, Ruixuan Xiao, Yurun Chen, Jiasheng Ye, Meiling Tao, Xiangxin Zhou, Ziyu Zhao, et al. 2025. Os agents: A survey on mllm-based agents for computer, phone and browser use. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 7436–7465.

[26] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2333–2338.

[27] Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. 2025. Recommender ai agent: Integrating large language models for interactive recommendations. *ACM Transactions on Information Systems* 43, 4 (2025), 1–33. Publisher: ACM New York, NY.

[28] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *Proceedings of the tenth ACM international conference on web search and data mining*. 781–789.

[29] Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

[30] Zhenyu Li, Mohamed Ali Kaafar, Kave Salamatian, and Gaogang Xie. 2016. Characterizing and modeling user behavior in a large-scale mobile live streaming system. *IEEE Transactions on Circuits and Systems for Video Technology* 27, 12 (2016), 2675–2686.

[31] Xinyu Lin, Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, and Fuli Feng. 2025. Navigating Large Language Models for Recommendation: From Architecture to Learning Paradigms and Deployment. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 4134–4137.

[32] Ziyue Lin, Yi Shan, Lin Gao, Xinghua Jia, and Siming Chen. 2025. SimSpark: Interactive Simulation of Social Media Behaviors. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–32.

[33] Guangyi Liu, Pengxiang Zhao, Liang Liu, Yaxuan Guo, Han Xiao, Weifeng Lin, Yuxiang Chai, Yue Han, Shuai Ren, Hao Wang, et al. 2025. Llm-powered gui agents in phone automation: Surveying progress and prospects. *arXiv preprint arXiv:2504.19838* (2025).

[34] George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.

[35] Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. 2023. MemGPT: Towards LLMs as Operating Systems. *arXiv preprint arXiv:2310.08560* (2023).

[36] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.

[37] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.

[38] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Berns. 2023. Generative Agents: Interactive Simulacra of Human Behavior.

[39] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 701–710.

[40] Pew Research Center. 2024. Mobile Fact Sheet. https://www.pewresearch.org/internet/fact-sheet/mobile/. Accessed: 2025-07-22.

[41] Robin L Plackett. 1975. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics* 24, 2 (1975), 193–202. Publisher: Oxford University Press.

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence* 39, 6 (2016), 1137–1149.

[43] Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi Li, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071* (2024).

[44] Ranjan Sapkota, Konstantinos I Roumeliotis, and Manoj Karkee. 2025. Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges. *arXiv preprint arXiv:2505.10468* (2025).

[45] Ivan Sekulić, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, Andre Ferreira Manso, and Roland Mathis. 2024. Reliable LLM-based User Simulator for Task-Oriented Dialogue Systems. In *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT*

2024). 19–35.
[46] Lin Shi, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the Judges: A Systematic Investigation of Position Bias in Pairwise Comparative Assessments by LLMs. *arXiv e-prints* (2024), arXiv–2406.
[47] Yucheng Shi, Wenhao Yu, Zaitang Li, Yonglin Wang, Hongming Zhang, Ninghao Liu, Haitao Mi, and Dong Yu. 2025. MobileGUI-RL: Advancing Mobile GUI Agent through Reinforcement Learning in Online Environment. *arXiv preprint arXiv:2507.05720* (2025).
[48] Larry R Squire, Barbara Knowlton, and Gail Musen. 1993. The structure and organization of memory. *Annual review of psychology* 44, 1 (1993), 453–495.
[49] Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas Griffiths. 2023. Cognitive architectures for language agents. *Transactions on Machine Learning Research* (2023).
[50] Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. 2017. Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems* 30 (2017).
[51] Juntao Tan, Liangwei Yang, Zuxin Liu, Zhiwei Liu, Rithesh R N, Tulika Manoj Awalgaonkar, Jianguo Zhang, Weiran Yao, Ming Zhu, Shirley Kokane, Silvio Savarese, Huan Wang, Caiming Xiong, and Shelby Heinecke. 2025. PersonaBench: Evaluating AI Models on Understanding Personal Information through Accessing (Synthetic) Private User Data. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 878–893. doi:10.18653/v1/2025.findings-acl.49
[52] Thanh Tran, Kyumin Lee, Yiming Liao, and Dongwon Lee. 2018. Regularizing matrix factorization with user and item embeddings for recommendation. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 687–696.
[53] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291* (2023).
[54] Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, et al. 2025. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems* 43, 2 (2025), 1–37.
[55] Siyu Wang, Xiaocong Chen, Quan Z Sheng, Yihong Zhang, and Lina Yao. 2023. Causal disentangled variational auto-encoder for preference understanding in recommendation. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*. 1874–1878.
[56] Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu, and Yingzhen Yang. 2023. Recmind: Large language model powered agent for recommendation. *arXiv preprint arXiv:2308.14296* (2023).
[57] Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu, and Yingzhen Yang. 2023. Recmind: Large language model powered agent for recommendation. *arXiv preprint arXiv:2308.14296* (2023).
[58] Zhenhailong Wang, Haiyang Xu, Junyang Wang, Xi Zhang, Ming Yan, Ji Zhang, Fei Huang, and Heng Ji. 2025. Mobile-agent-e: Self-evolving mobile assistant for complex tasks. *arXiv preprint arXiv:2501.11733* (2025).
[59] Kai Wei, Jaime Booth, and Rachel Fusco. 2019. Cognitive and emotional outcomes of Latino threat narratives in news media: An exploratory study. *Journal of the Society for Social Work and Research* 10, 2 (2019), 213–236.
[60] Hao Wen, Yuanchun Li, Guohong Liu, Shanhui Zhao, Tao Yu, Toby Jia-Jun Li, Shiqi Jiang, Yunhao Liu, Yaqin Zhang, and Yunxin Liu. 2024. Autodroid: Llm-powered task automation in android. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 543–557.
[61] Chen Wenjing. 2025. Simulation application of virtual robots and artificial intelligence based on deep learning in enterprise financial systems. *Entertainment Computing* 52 (2025), 100772.
[62] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences* 68, 2 (2025), 121101.
[63] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4818–4829.
[64] Tianyu Xiong and Xiaohan Yu. 2024. Multi-tower multi-interest recommendation with user representation repel. *arXiv preprint arXiv:2403.05122* (2024).
[65] Wujiang Xu, Kai Mei, Hang Gao, Juntao Tan, Zujie Liang, and Yongfeng Zhang. 2025. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110* (2025).
[66] Wujiang Xu, Kai Mei, Hang Gao, Juntao Tan, Zujie Liang, and Yongfeng Zhang. 2025. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110* (2025).
[67] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.
[68] Wenwen Yu, Zhibo Yang, Jianqiang Wan, Sibo Song, Jun Tang, Wenqing Cheng, Yuliang Liu, and Xiang Bai. 2025. Omniparser v2: Structured-points-of-thought for unified visual text parsing and its generality to multimodal large language models. *arXiv preprint arXiv:2502.16161* (2025).
[69] An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. 2024. On generative agents in recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in Information Retrieval*. 1807–1817.
[70] Chi Zhang, Zhao Yang, Jiaxuan Liu, Yanda Li, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. 2025. Appagent: Multimodal agents as smartphone users. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.
[71] Wei Zhang, Dai Li, Chen Liang, Fang Zhou, Zhongke Zhang, Xuewei Wang, Ru Li, Yi Zhou, Yaning Huang, Dong Liang, et al. 2024. Scaling user modeling: Large-scale online user representations for ads personalization in meta. In *Companion Proceedings of the ACM Web Conference 2024*. 47–55.
[72] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625* (2022).

A Appendix

Movie Title	Genre
List 1	
Son in Law (1993)	Comedy
My Boyfriend's Back (1993)	Comedy
Father of the Bride Part II (1995)	Comedy
Made in America (1993)	Comedy
Problem Child (1990)	Comedy
List 2	
Cape Fear (1991)	Thriller
Cape Fear (1962)	Film-Noir/Thriller
Relative Fear (1994)	Horror/Thriller
Robocop 3 (1993)	Sci-Fi/Thriller
Pacific Heights (1990)	Thriller
List 3	
Pokémon: The First Movie (1998)	Animation/Children's
Pokémon the Movie 2000 (2000)	Animation/Children's
Digimon: The Movie (2000)	Animation/Children's
Antz (1998)	Animation/Children's
The Tigger Movie (2000)	Animation/Children's

Table 3: MovieLens Dataset Examples (Lists of 5): Examples of item lists generated by our negative sampling strategy on MovieLens dataset. The lists demonstrate how our method identifies and groups semantically similar items (movie titles here) that would realistically compete for user preference, rather than random unrelated alternatives.

TIMO Behavior Module Template

Input Parameters

- **persona:** Csutomer profile with preferences and traits
- **goal:** Current shopping objective
- **memories:** Past shopping experiences and interactions
- **products:** List of products to evaluate

Constants

- **ACTIONS:** [BuyNow, AddToCart, CheckDetailPage, Quit]
- **ACTION_HIERARCHY:** *BuyNow* > *AddToCart* > *CheckDetailPage* > *Quit*

Process Steps

1. For each product:
   - Evaluate alignment with persona and goal
   - Review relevant memories
   - Make decision based on ACTION_HIERARCHY
   - Provide persona-based reasoning
   - Share product thoughts
2. Generate product ranking
3. Format output in specified JSON structure

Constraints

- Must maintain customer persona
- Must use exact product names/IDs
- Must reference persona/memories in reasoning
- Must only rank non-Quit products
- Must provide realistic considerations

Output Format

- **decision:** Array of product evaluations
  - action: One of ACTIONS
  - product_name: Exact product name
  - product_asin: Product identifier
  - reason: Persona/memory-based explanation
  - thought: Shopping considerations
- **product_rank:** Ordered list of product ASINs

Example Components

- **Persona Elements:** Budget, preferences, lifestyle
- **Memory Types:** Previous purchases, experiences
- **Decision Factors:** Price, features, reviews

Decision Logic

1. BuyNow: Strong match, immediate purchase intent
2. AddToCart: Good match, future purchase intent
3. CheckDetailPage: Interested, needs more research
4. Quit: No interest or poor match

Quality Guidelines

- Reasoning must be specific and detailed
- Thoughts must reflect realistic shopping behavior
- Decisions must align with persona characteristics
- Rankings must follow preference hierarchy

Figure 4: TIMO Behavior Module Prompt Template: Our behavior module template with inputs (persona, goals, memories, products), hierarchical action decisions (BuyNow > AddToCart > CheckDetailPage > Quit), JSON-formatted outputs with reasoning and rankings, that enables digital twin to simulate realistic multi-item decision making.

	Negative Emotion		Positive Emotion	
	Human Responses	Agent (TIMO) Responses	Human Responses	Agent (TIMO) Responses
Treatment Effect	1.179***	1.195***	-0.065	-0.057
Standard Error	(0.200)	(0.110)	(0.180)	(0.060)
t-statistic	5.76	11.25	-0.37	-1.02
p-value	0.000	0.000	0.720	0.310
N	128	128	128	128

Table 4: Comparison of Emotion Perception Task [59] between Human Subjects and Digital Twins (TIMO). $^{***}p < 0.001$**. Agents and humans showed nearly identical treatment effects when exposed to different types of articles: negative emotions increased by 1.179 units in humans (SE=0.2, p<0.001) and 1.195 units in agents (SE=0.11, p<0.001), while neither group showed significant changes in positive emotions, validating our framework's ability to simulate human emotional responses with comparable magnitude but lower variability.**

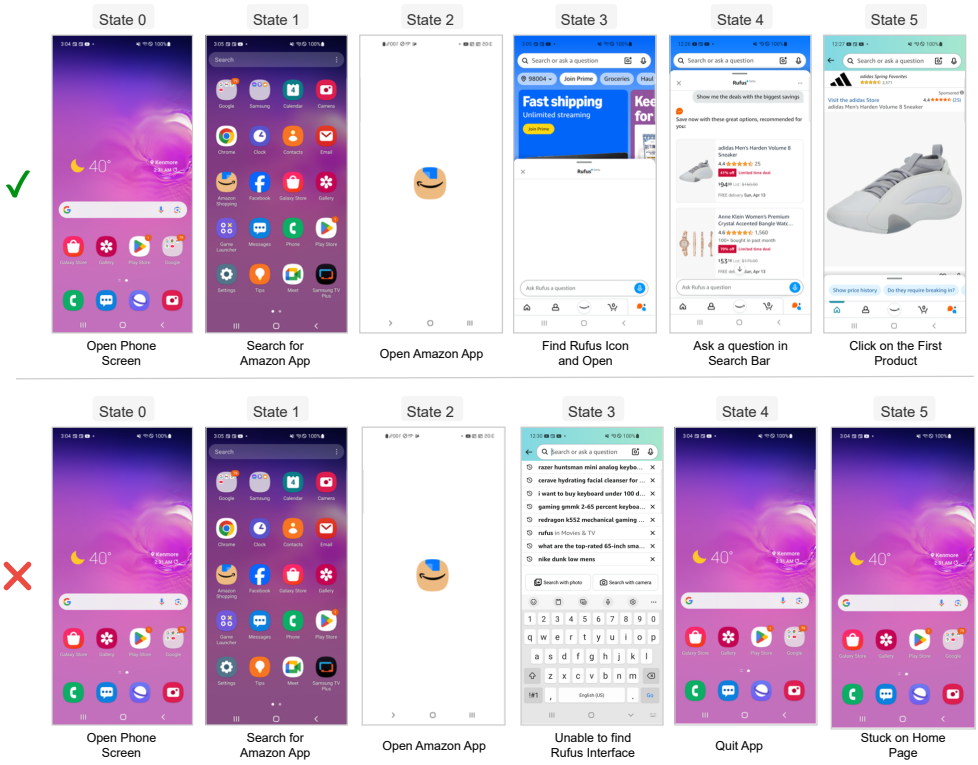


Figure 5: Examples of screenshots from device interaction trajectories used to evaluate TIMO's Device Control Module's perception capabilities. Screenshots capture key states during task execution to "open and search products using Rufus on Amazon, and click on a product". Each trajectory includes sequential screenshots annotated by human evaluators to assess task completion success. We show an example of a successful (top) and unsuccessful (bottom) trajectory. TIMO's device control module exhibits fine-grained high-fidelity device interactions with an overall success rate of 91.7%.

Aspect	Question	Generative Agent	RecAgent	TIMO
Customer Preference Integration	Is there evidence of considering relevant past behavior and user preferences?	61.8%	51.9%	96.4%
Logic Validity	Are product selections supported by logical arguments?	95.0%	33.4%	93.5%
Decision Clarity and Transparency	Is the decision-making process explained clearly and transparently?	97.7%	18.6%	94.3%
Price-Value Analysis	Does the text include analysis of price in relation to product value?	86.9%	2.2%	43.2%
Alternatives Consideration	Are alternative product options discussed or considered?	56.8%	3.1%	14.7%
Negative Preference Incorporation	Does the text show evidence of considering the user's dislikes?	45.5%	5.1%	12.9%

Table 5: Comparative evaluation of agent reasoning texts across six assessment dimensions. Values represent percentage of positive responses in evaluation (n=1000). TIMO achieves 96.4% preference alignment (+34.6% vs. Generative Agent, +44.5% vs. RecAgent) while maintaining balanced price-value considerations (43.2% vs. 86.9% for Generative Agent) and avoiding over-reliance on price/singular-aspect focused reasoning.