AgenTwin: An End-to-End Framework for Building Agentic 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]
[1]Amazon [2]Stanford University
{saranvn,tdt,shunyl,devichen,yuningwu,kaiwe,jonallie,idorosen,cambrose}@amazon.com
imbens@stanford.edu

Abstract

Creating authentic agentic twins of mobile users through realistic user behavior simulation is critical to truly understand and anticipate customer needs at scale. Toward this, we introduce **Agen**tic **Twin**s of Mobile Users (AgenTwin), 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 agentic 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 agentic 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.

1 Introduction

User simulation is critical for understanding customers (Wang et al., 2025a), anticipating their needs (Bougie & Watanabe, 2025), driving product discovery (Füller & Matzler, 2007), and enhancing experience (Zhang et al., 2024b; Lin et al., 2025a) across domains like e-commerce (Čavoški & Marković, 2017), streaming (Li et al., 2016), social media (Lin et al., 2025b), and finance (Wenjing, 2025). It enables risk-free testing of promotional strategies, A/B experiments, and feature rollouts without disrupting live production systems. For large scale online retail stores, where millions of daily customers interact primarily via mobile ($\sim$70% of transactions (Pew Research Center, 2024)), realistic end-to-end user simulation, or *agentic twins of customers*, is key to enhancing customer experience, optimizing personalized recommendations, and testing marketing initiatives at scale.

However, creating authentic mobile agentic twins poses several challenges. First, the subjective nature of human decision-making makes it hard to model preferences consistently across customer segments (Aoyagui et al., 2025; De Vries et al., 2008; Dammu et al., 2025; Wang et al., 2025a). Second, existing approaches evaluate products independently, failing to capture how customers process competing options simultaneously within limited screen space (Hsee, 1996a; Hsee & Leclerc, 1998a; Park et al., 2023c). Third, granular mobile interactions such as touch gestures, scrolling, and micro-interactions demand high-fidelity device control that traditional methods miss (Hu et al., 2025; Shi et al., 2025; Liu et al., 2025).

Fourth, constraints on training throughput, memory, and data ingestion create scalability bottlenecks (Zhang et al., 2024b; Sapkota et al., 2025; Dong et al., 2024; Feng et al., 2024).

To address these, we introduce AgenTwin, an agentic end-to-end, high-fidelity, data-efficient user simulation framework for mobile agentic twins, comprising four integrated components: (1) an individualized persona generator that creates fine-grained customer profiles capturing subjective decision-making diversity; (2) an interpretable behavior simulation module that jointly evaluates multiple options rather than each in isolation; (3) a high-fidelity device control component that translates decisions into realistic touch gestures, scrolling, and navigation flows; and (4) a self-evolving contextualized memory module that dynamically stores and retrieves experiences from raw interaction data.

Our framework offers three key contributions: (1) *multi-faceted joint decision architecture*: agents simultaneously compare and rank multiple options, mirroring real mobile behavior and improving prediction accuracy by 10–20% across tasks; (2) *adaptive persona-memory co-evolution*: fine-grained personas integrate with contextual memories that adapt to interaction history, enabling authentic long-term consistency over static profiles; (3) *data-efficient modular pipeline*: an end-to-end framework balancing behavioral fidelity and operational efficiency.

We evaluate AgenTwin on public benchmarks and proprietary mobile e-commerce datasets, showing significant gains over baselines (RecAgent (Wang et al., 2025a), Generative Agent (Park et al., 2023c)) in prediction accuracy, robustness to positional bias, and efficiency. Our findings show agent design must shift from binary to joint processing, as real-world agency rarely operates in isolation. Beyond task benchmarks, dedicated emotion perception experiments show our simulated responses align closely with real human behavior. Our prototype has enabled high-fidelity mobile user simulations yielding rich insights across offline business metrics.

2 Related Works

Agentic twin simulation of users has been explored in recommender systems (Wang et al., 2025a), conversational agents (Sekulić et al., 2024), and agent-based modeling (Zhang et al., 2024b), yet existing approaches fail to capture the full complexity of user behavior in mobile contexts. Generative agent frameworks (Park et al., 2023b) and LLM-driven agent-based modeling (Gao et al., 2024) show the potential of lifelike simulacra but target open-ended or desktop settings rather than production-scale mobile environments. Interactive recommender agents like InteRecAgent (Huang et al., 2025) and RecMind (Wang et al., 2023d) integrate LLM reasoning with recommendation flows, but typically model single-option decisions, overlooking the multi-item, screen-limited comparisons that dominate mobile behavior. Work on persona conditioning (Tan et al., 2025; Afzoon et al., 2024) and dynamic memory (Packer et al., 2023; Xu et al., 2025b) highlights long-horizon consistency, yet most systems rely on static profiles or simplified buffer memories without self-evolving, persona-linked adaptation. From a decision-making perspective, behavioral economics underscores the contrast between joint and separate evaluations (Hsee, 1996b; 1998), motivating listwise and choice-to-rank formulations such as Plackett–Luce models (Plackett, 1975) and slate bandits (Swaminathan et al., 2017) that reason over sets rather than isolated items. Parallel work on mobile GUI control (Zhang et al., 2025; Wen et al., 2024; Chai et al., 2025) focuses on task completion via tap/scroll action spaces but rarely couples fine-grained device control with realistic cognitive joint decision processes. Finally, counterfactual learning-to-rank emphasizes robustness to exposure and position bias via inverse propensity scoring and slate-aware estimators (Joachims et al., 2017), while recent affective-computing benchmarks (Sabour et al., 2024; Chen et al., 2024) introduce protocols for emotional alignment and authenticity. Our work builds on these foundations, advancing them by integrating joint choice modeling, persona–memory co-evolution, and high-fidelity mobile interaction into a single end-to-end framework, with 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 agentic 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 agentic 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.

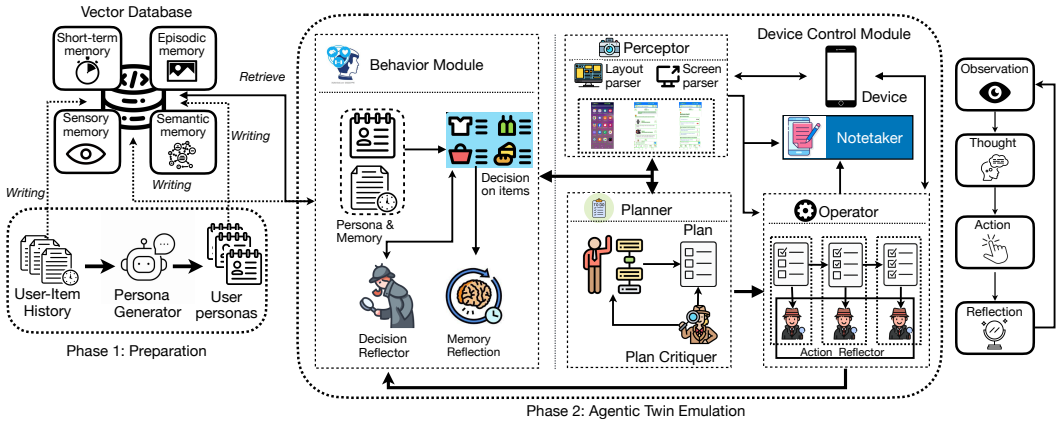

Figure 1: AgenTwin's overall architecture:In Phase 1, a Preparation stage constructs Personas and associated Memories for initialization. In Phase 2, the initialized twins interact with and simulate user behaviors on mobile devices.

4 AgenTwin Framework

AgenTwin is an end-to-end framework for constructing agentic twins of mobile users, capable of emulating human-like behavior by autonomously navigating mobile devices and making informed decisions guided by users' goals and contextual information retrieved from their historical interactions. Figure 1 presents AgenTwin, which operates in two phases. In Phase 1, a one-to-one mapping between each user and their persona is generated, along with initialization of associated memories. In Phase 2, a relevant or preselected persona together with relevant memories are retrieved to initialize a agentic twin, which then interacts with the mobile application, making decisions via product exploration to mimic human-like behavior. We describe components of Phase 1 (Persona Generation, Agent Memory) and Phase 2 (Behavior Module, Device Control Module) below.

4.1 Individualized and Fine-Grained Persona Generation

A persona is a profile of a user characterized by demographic attributes (e.g., age, occupation), emotional and motivational elements, and affective indicators that provide context about deep-rooted preferences and intrinsic motivations. To capture diversity in subjective decision-making, we develop an individualized persona generation method that creates realistic profiles with nuanced behavioral tendencies, serving as consistent seeds for agentic twin interactions (Zhang et al., 2024a; Xi et al., 2025; Wang et al., 2023a; Park et al., 2023a; Wang et al., 2023c). Since granular features for constructing personas at scale are usually unavailable, we use an Large Language Model (LLM) to infer personas from a user's past interaction history, given empirical evidence of their superior performance in generating nuanced persona representations (Ge et al., 2024; Zhang et al., 2024a). We use Claude 3.5 Sonnet v2 as the backbone LLM for all components of the pipeline.

For a user u_i with historical interaction data $\mathcal{I}_{T,i} = [e_{i,1}, \ldots, e_{i,K_i}]$, we prompt an LLM to generate a fine-grained profile that infers the user's preferences (e.g., "budget-conscious"), goals (e.g., "find natural beauty products"), and demographic attributes (e.g., "middle-aged"). To enrich preference boundaries, we additionally incorporate *impression items* as items similar to those a user engaged with but that received no engagement (Tran et al., 2018), retrieved via semantic similarity. Full details on impression item retrieval and privacy constraints are provided in Appendix A.1.

4.2 Self-Evolving Contextualized Agent Memory

Inspired by the hierarchy of human memory (Squire et al., 1993; Xu et al., 2025a), AgenTwin's memory is structured into four segments: *episodic memory* (past experiences and user-item interactions), *semantic memory* (general knowledge and persona facts), *short-term memory* (recent interactions, capped at 9 per Miller's Law (Miller, 1956)), and *sensory memory* (multimodal perceptual inputs such as product images and interface signals). We store generated personas in semantic memory and interaction data in short-term and episodic memories. Each segment is implemented as a collection in a Vector Database for fast retrieval.

To retrieve relevant personas/memories during simulation, we adopt a two-tower retrieval architecture (Huang et al., 2013) for its scalability. We encode textual and visual data using *multilingual-e5-large-instruct* and *CLIP-ViT-B/32*, respectively. To mitigate prompt-length limits from raw memory events (Zhang et al., 2024a), we implement a memory reflection mechanism that uses an LLM to summarize each memory event before encoding. Detailed memory definitions and retrieval design are provided in Appendix A.2.

4.3 Joint Multi-Option Decision Processing Behavior Module

The *Behavior Module* is the brain of the agentic twin, synthesizing user preferences and predicting next-step actions. Our joint multi-option processing is inspired by Hsee's Evaluability Hypothesis (Hsee, 1996b; 1998; Hsee & Leclerc, 1998a), which shows that easily comparable attributes disproportionately influence decisions when options are evaluated simultaneously, particularly relevant for mobile e-commerce, where users rapidly scroll and compare products. Unlike binary decision-making, where options are evaluated in isolation, joint processing alters preference structures (preference reversal (Hsee & Leclerc, 1998b)). Accordingly, our chain-of-thought architecture anchors on directly comparable numeric attributes (e.g., prices, review ratings), mirroring how shoppers gravitate toward quantitative comparisons ("$15 cheaper", "2000 more reviews") on a small screen. The Behavior Module prompt contains: persona, relevant episodic memory, relevant semantic memory, short-term memory, and current ongoing memory or the active search mission and content displayed in the target app $\mathcal{T}$ (Complete template in Appendix Figure 5). The module operates in four steps:

a. Initialization: The agentic twin is assigned a persona (retrieved or prespecified), and its memory is populated. The current ongoing memory is initialized with the search mission and displayed product list, extracted by the Device Control Module (Section 4.4).

b. Multi-criteria Item Assessment: Following least-to-most prompting (Zhou et al., 2022), the agentic twin reasons about each item's alignment with (1) its preferences, (2) historical behavior, and (3) current search intent.

c. Decision Refinement with Feedback Loop: AgenTwin defines a granular, interest-descending decision schema per domain (e.g., [BUY-NOW], [ADD-TO-CART], [CLICK], [SKIP] for shopping; [WATCH-NOW], [ADD-TO-WATCH-LIST], [WATCH-TRAILER], [SKIP] for streaming). Inspired by Cai et al. (2025); Bodhwani et al. (2025), a separate feedback agent critiques whether the decision aligns with the stated reasoning and user preferences; if inconsistent, the agent self-reflects and revises.

d. Memory Reflection: To avoid prompt overflow and noise from raw action events, we prompt an LLM to summarize actions into higher-level inferences over time, which are fed back into the memory stream to influence future behavior.

4.4 High-Granularity Device Control Module

The *Device Control Module* manages low-level interactions with the target app $\mathcal{T}$. Using a layout understanding model, it identifies GUI components and performs human-device interactions (tapping, scrolling, swiping, typing). Inspired by Kahneman's dual-process theory (Kahneman, 2011), we combine a customized *ReAct* agent (Yao et al., 2023) (System 1) for rapid, specialized UI interactions with Mobile-Agent-E (Wang et al., 2025b) (System 2) for deliberate processing. System 1 cycles through *Observation*, *Thought*, *Action*, *Reflection*, and *Reiteration*; when reflection detects failures or complex tasks, AgenTwin switches to System 2. Architectural details of System 2's submodules and our Perceptor implementation are provided in Appendix A.3.

5 Experimental Results

We now present our experimental settings and results on the application and comparative evaluation of AgenTwin-based agentic twins for 2 real-world ([Internal] Mobile Shopping Datasets) as well as a public benchmark (MovieLens)-based scenarios against baseline agents for user simulation (RecAgent (Wang et al., 2025a) and Generative Agents (Park et al., 2023c)).

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

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] Mobile Shopping Dataset2 are shown in increments as per data sharing policies.

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 (Wang et al., 2025a; Tran et al., 2019b;a), 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** (Harper & Konstan, 2015) and a proprietary **[Internal] Mobile Shopping Dataset1** for this multi-item decision making setting. The former comprises movie ratings, while the latter consists of user interactions across multiple product domains. 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).

5.1.2 Real World Joint Decision Dataset

To validate our agentic twins in practice, we evaluate on a proprietary real-world joint decision dataset denoted as [Internal] Mobile Shopping Dataset2. This dataset captures authentic customer-product interactions from production search and recommendation system, 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.

5.2 Multi-Faceted Decision Making

5.2.1 Click Prediction Task

Using the hard negative setting from benchmark adaptation ([Internal] Mobile Shopping 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] Mobile Shopping 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 AgenTwin with 2 other agents: Generative Agent (Park et al., 2023a) and RecAgent (Wang et al., 2025a). [Internal] Mobile Shopping Dataset1 contains customer interactions with products from multiple domains. We present results for two such domains that have high volume of interactions per user. From Table 1, we see that overall (averaged across values of k) AgenTwin performs better than RecAgent and Generative Agent for both [Internal] Mobile Shopping 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 AgenTwin's F1-score by 1.2% for $k = 1$ setting. For MovieLens data, for $k >= 3$, RecAgent performs best, while AgenTwin 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] Mobile Shopping Dataset1 and MovieLens, respectively at $k = 1$. For [Internal] Mobile Shopping 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 AgenTwin outperform the baseline Generative Agent across all metrics and values of k, with AgenTwin performing better at lower values of $k < 4$, and an overall increase of 19% F1 from the baseline. Our findings demonstrate the effectiveness and efficiency (Appendix A.4) of targeted design adaptations in advancing agent capabilities, while underscoring the necessity of evaluation frameworks that reflect real-world decision complexity.

[Internal] Mobile Shopping Dataset1 (Select Any from 5)			
Avg. F1@K	Generative Agent	RecAgent	AgenTwin
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	AgenTwin
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 → 4] or last [5].

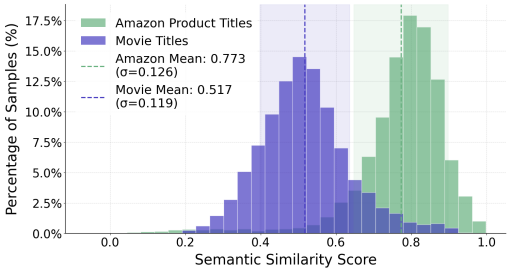

Figure 2: Distribution of cosine similarities between positive (ground truth) and negative product/movie titles derived from [Internal] Mobile Shopping Dataset1 and MovieLens datasets, respectively. The higher mean similarity in shopping product titles ($\mu = 0.77$) compared to movie titles ($\mu = 0.51$) suggests that movie titles are more distinct than product titles.

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 (Shi et al., 2024). To quantify the impact of this bias on AgenTwin's decision-making process, we conduct a two-fold analysis:

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 → 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

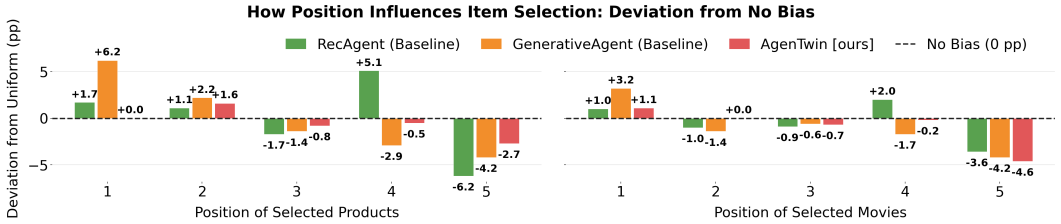

Figure 3: **Position Bias is Agent's chosen items. Are agents biased to picking items shown first?** Y-axis show each agent's selection frequency of products/movies shown at different positions in a list, from 1 to 5 [l→r] on the X-axis, expressed as deviation (in percentage points) from the uniform 20% baseline (dashed line). Agents tend to pick the item in position 1 more often than those further down the list.

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 shopping 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 AgenTwin demonstrates superior robustness with minimal degradation (1.86% total loss), further supporting evidence for position-invariant agent decision-making.

Distribution of Agent Clicks by Position We analyze the distribution of selection frequencies across input item positions relative to the unbiased 20% baseline. Our analysis in Figure 3 shows that all agents exhibit position bias, though with varying intensities. GenerativeAgent demonstrates the most pronounced bias, with the largest single swing among all agents (+6.2pp at position 1 for products), while AgenTwin shows the smallest average deviation from the 0pp line across most positions, though it is not uniformly closest at every point. For movies, its position-5 deviation (−4.6pp) is marginally larger than either baseline's. This overall pattern suggests AgenTwin's list-based decision making, which considers all items at once rather than scoring them independently, offers some built-in resistance to position bias without eliminating it entirely. Domain-specific analysis further reveals that position bias tends to be more severe in product recommendations as movie domain shows a narrower spread of deviations across positions (−4.6pp to +3.2pp) than products (−6.2pp to +6.2pp), a disparity that may stem from differences in negative sampling, where product titles tend to share more semantic similarity than the more distinct movie titles, as shown in Figure 2. While the bias patterns and their magnitudes vary by agent and domain, the consistency of the effect across all three agents points to a systematic challenge in LLM-driven agentic frameworks, with AgenTwin showing promising bias mitigation capabilities.

5.3 Device Perception Success Rate

We collected 72 trajectories through screenshots of AgenTwin'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 AgenTwin'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 4).

5.4 Reasoning Analysis

We evaluate agent reasoning strategies by using an LLM-as-a-judge evaluation (Chiang & Lee, 2023) followed by human-in-the-loop verification for a random subset. We obtained binary (Yes/No) judgments and single-sentence justifications about 1k samples of each agent's 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 AgenTwin's reasoning is the most user-centric, given the high occurrence (96.4%) of mentions of the user's past behaviors and preferences ($\delta = +34.6\%$ compared to Generative Agent, $\delta = +44.5\%$ compared to RecAgent). While the Generative Agent tends to rely on price-value analysis much more frequently (86.9%) than AgenTwin (43.2%), our empirical click decision prediction results lend support to AgenTwin's multi-criteria strategy that considers multiple product attributes, user preferences, and intent instead of focusing on one aspect (such as price). Lower occurrences of alternatives consideration (24.87%) and negative preference incorporation (21.17%) across all agents' reasoning may indicate an opportunity for future architectural improvements to help agents better leverage negative signals from user behaviors.

5.5 Affective Realism

As a means to evaluate if AgenTwin 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. (2019). We initialized AgenTwin 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 (full results in Appendix Table 4). Both humans and our agentic 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 their behavioral validity.

6 Conclusion

In this work, we introduced an end-to-end framework for building agentic twins of mobile users through realistic behavior simulation. By unifying persona generation, self-evolving contextual memory, joint multi-option decision-making, and fine-grained device control, our framework addresses four limitations of prior simulation agents: subjective decision diversity, scalable experience retention, simultaneous multi-option processing, and granular interaction fidelity. Experiments across multiple datasets show that our approach improves prediction accuracy, robustness to positional bias, and computational efficiency, while producing agents whose responses more closely match authentic human behavioral and emotional patterns, advancing benchmark adaptation for holistic evaluation through contextually grounded settings that mirror real-world user cognition and interaction.

References

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

Paula Akemi Aoyagui, Kelsey Stemmler, Sharon A Ferguson, Young-Ho Kim, and Anastasia Kuzminykh. 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*, pp. 1–16, 2025.

Umesh Bodhwani, Yuan Ling, Shujing Dong, Yarong Feng, Hongfei Li, and Ayush Goyal. A calibrated reflection approach for enhancing confidence estimation in llms. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pp. 399–411, 2025.

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

Shihao Cai, Jizhi Zhang, Keqin Bao, Chongming Gao, Qifan Wang, Fuli Feng, and Xiangnan He. 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*, pp. 2235–2244, 2025.

Sava Čavoški and Aleksandar Marković. Agent-based modelling and simulation in the analysis of customer behaviour on b2c e-commerce sites. *Journal of Simulation*, 11(4): 335–345, 2017.

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

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

Cheng-Han Chiang and Hung-Yi Lee. 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)*, pp. 15607–15631, 2023.

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

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

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

Xiaofei Dong, Xueqiang Zhang, Weixin Bu, Dan Zhang, and Feng Cao. 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)*, pp. 407–413. IEEE, 2024.

Zhichao Feng, JunJie Xie, Kaiyuan Li, Yu Qin, Pengfei Wang, Qianzhong Li, Bin Yin, Xiang Li, Wei Lin, and Shangguang Wang. Context-based fast recommendation strategy for long user behavior sequence in meituan waimai. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 355–363, 2024.

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

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

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

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

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

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 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*, pp. 639–648, 2020.

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

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

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

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

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

Xueyu Hu, Tao Xiong, Biao Yi, Zishu Wei, Ruixuan Xiao, Yurun Chen, Jiasheng Ye, Meiling Tao, Xiangxin Zhou, Ziyu Zhao, et al. 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)*, pp. 7436–7465, 2025.

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

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

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

Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011.

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

Xinyu Lin, Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, and Fuli Feng. 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*, pp. 4134–4137, 2025a.

Ziyue Lin, Yi Shan, Lin Gao, Xinghua Jia, and Siming Chen. Simspark: Interactive simulation of social media behaviors. *Proceedings of the ACM on Human-Computer Interaction*, 9(2): 1–32, 2025b.

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

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

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

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

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

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

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

Pew Research Center. Mobile fact sheet. `https://www.pewresearch.org/internet/fact-sheet/mobile/`, 2024. Accessed: 2025-07-22.

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

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

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

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

Ivan Sekulić, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, Andre Ferreira Manso, and Roland Mathis. 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)*, pp. 19–35, 2024.

Lin Shi, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *arXiv e-prints*, pp. arXiv–2406, 2024.

Yucheng Shi, Wenhao Yu, Zaitang Li, Yonglin Wang, Hongming Zhang, Ninghao Liu, Haitao Mi, and Dong Yu. Mobilegui-rl: Advancing mobile gui agent through reinforcement learning in online environment. *arXiv preprint arXiv:2507.05720*, 2025.

Larry R Squire, Barbara Knowlton, and Gail Musen. The structure and organization of memory. *Annual review of psychology*, 44(1):453–495, 1993.

Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems*, 30, 2017.

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. PersonaBench: Evaluating AI Models on Understanding Personal Information through Accessing (Synthetic) Private User Data. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 878–893, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.49. URL https://aclanthology.org/2025.findings-acl.49/.

Thanh Tran, Kyumin Lee, Yiming Liao, and Dongwon Lee. Regularizing matrix factorization with user and item embeddings for recommendation. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pp. 687–696, 2018.

Thanh Tran, Xinyue Liu, Kyumin Lee, and Xiangnan Kong. Signed distance-based deep memory recommender. In *The world wide web conference*, pp. 1841–1852, 2019a.

Thanh Tran, Renee Sweeney, and Kyumin Lee. Adversarial mahalanobis distance-based attentive song recommender for automatic playlist continuation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 245–254, 2019b.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023a.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, et al. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(2):1–37, 2025a.

Siyu Wang, Xiaocong Chen, Quan Z Sheng, Yihong Zhang, and Lina Yao. 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*, pp. 1874–1878, 2023b.

Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu, and Yingzhen Yang. Recmind: Large language model powered agent for recommendation. *arXiv preprint arXiv:2308.14296*, 2023c.

Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu, and Yingzhen Yang. Recmind: Large language model powered agent for recommendation. *arXiv preprint arXiv:2308.14296*, 2023d.

Zhenhailong Wang, Haiyang Xu, Junyang Wang, Xi Zhang, Ming Yan, Ji Zhang, Fei Huang, and Heng Ji. Mobile-agent-e: Self-evolving mobile assistant for complex tasks. *arXiv preprint arXiv:2501.11733*, 2025b.

Kai Wei, Jaime Booth, and Rachel Fusco. 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):213–236, 2019.

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

Chen Wenjing. Simulation application of virtual robots and artificial intelligence based on deep learning in enterprise financial systems. *Entertainment Computing*, 52:100772, 2025.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 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*, pp. 4818–4829, 2024.

Tianyu Xiong and Xiaohan Yu. Multi-tower multi-interest recommendation with user representation repel. *arXiv preprint arXiv:2403.05122*, 2024.

Wujiang Xu, Kai Mei, Hang Gao, Juntao Tan, Zujie Liang, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*, 2025a.

Wujiang Xu, Kai Mei, Hang Gao, Juntao Tan, Zujie Liang, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*, 2025b.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Wenwen Yu, Zhibo Yang, Jianqiang Wan, Sibo Song, Jun Tang, Wenqing Cheng, Yuliang Liu, and Xiang Bai. 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.

An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. On generative agents in recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in Information Retrieval*, pp. 1807–1817, 2024a.

Chi Zhang, Zhao Yang, Jiaxuan Liu, Yanda Li, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. Appagent: Multimodal agents as smartphone users. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–20, 2025.

Wei Zhang, Dai Li, Chen Liang, Fang Zhou, Zhongke Zhang, Xuewei Wang, Ru Li, Yi Zhou, Yaning Huang, Dong Liang, et al. Scaling user modeling: Large-scale online user representations for ads personalization in meta. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 47–55, 2024b.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

A Appendix

A.1 Persona Generation Details

For user u_i with historical interaction data $\mathcal{I}T, i = [ei, 1, e_{i,2}, \ldots, e_{i,K_i}]$, where $e_{i,j}$ is the j-th chronological interaction event and K_i is the total number of interactions on application T, we prompt an LLM to generate a detailed fine-grained profile summary. Following Bougie & Watanabe (2025), we sample the most recent 50 interactions to construct this summary.

In line with findings that non-preferred items provide valuable insight into user preference boundaries Tran et al. (2018), we simulate a realistic browsing scenario in which users are exposed to relevant products during search sessions. Items aligning with u_i's preferences (e.g., cost, color, brand) are captured in $\mathcal{I}_{T,i}$, while *impression items* i.e. similar items receiving no engagement, offer more granular preference signals but are often overlooked. To identify impression items, we use a text encoder (*all-MiniLM-L6-v2*) to embed each item, using item p_k's embedding as the query and the remaining items as candidates. We retrieve the *top-k* candidates with the highest cosine similarity as impression items, setting *top-k*$= 5$ to match typical UI patterns (4–6 products per search in apps like Rufus and Prime Video).

To ensure privacy, we incorporate explicit policy constraints within persona generation prompts, directing the model to generalize outputs based on behavioral patterns rather than stereotypes or surface-level details, and to avoid personally identifiable information. This prompt-level control maintains privacy while enabling meaningful simulation.

A.2 Agent Memory Details

Memory Structure. *Episodic memory* stores past experiences or events with temporal context, enabling AgenTwin to record and later retrieve *episodes* of interactions, analogous to a personal diary (e.g., remembering *the purchase of Nike white shoes at $59*). It supports temporal reasoning, planning, and learning from past outcomes DeChant (2025). *Semantic memory* holds general knowledge, facts, and concepts regardless of when learned (e.g., *Nike is a popular shoe brand*); unlike episodic memory, it is a dynamic knowledge base that can be shared across users. The two interact: episodic experiences can be abstracted into semantic knowledge, while semantic memories provide context for new episodes. *Short-term* (working) memory is a limited-capacity segment for active tasks; inspired by Miller's Law Miller (1956) (7 ± 2 items), we store the latest 9 interactions. *Sensory memory* contains multi-modal perceptual inputs (product images, videos, interface signals) for real-time grounding.

Persona and Memory Initialization via Retrieval. Among retrieval techniques including graph neural networks He et al. (2020), graph embeddings Perozzi et al. (2014); Grover & Leskovec (2016), multi-interest methods Xiong & Yu (2024), autoencoder-based methods Wang et al. (2023b), and collaborative filtering, we adopt a two-tower architecture Huang et al. (2013) for scalability. We encode text and images using *multilingual-e5-large-instruct* and *CLIP-ViT-B/32*, storing persona embeddings and interaction data $\mathcal{I}_{\mathcal{T},\rangle}$ in corresponding Vector Database collections. To mitigate prompt overflow from raw memory events Zhang et al. (2024a), our memory reflection mechanism uses an LLM to generate concise event summaries, which are encoded and stored for compact, contextual retrieval. This enables robust multi-modal retrieval across languages and content types (text, image, text+image), enhancing AgenTwin's ability to access relevant memories accurately and efficiently.

A.3 Device Control Module Details

When the reflection module detects potential failures or complex tasks, AgenTwin switches to System 2 (Mobile-Agent-E Wang et al. (2025b)), comprising five submodules: (1) **Planner**, which decomposes tasks into tractable sub-goals for step-by-step execution and verification; (2) **Perceptor**, which interprets visual elements (icons, images, text) and their spatial relationships; (3) **Operator**, which executes actions per the Planner's sub-goals; (4) **Action Reflector**, which evaluates whether actions yield expected outcomes by comparing current and historical screenshots and updating internal state (e.g., confirming a product was added to cart); and (5) **Notetaker**, which maintains session-level information such as selected products, filled fields, and click history, optimizing memory usage across modalities.

We adopt a two-stage Perceptor architecture combining grounding and captioning. For layout detection, we employ a fine-tuned Faster R-CNN Ren et al. (2016) trained on human-annotated data to identify GUI component bounding boxes, chosen for its lightweight, efficient performance. The captioning stage uses the pre-trained Florence model Xiao et al. (2024) and OCR models from OmniParser-v2 Yu et al. (2025) combined with XML data to generate textual metadata, enabling rapid experimentation in real-device environments.

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.

	Negative Emotion		**Positive Emotion**	
	Human Responses	AgenTwin Responses	Human Responses	AgenTwin 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 (Wei et al., 2019) between Human Subjects and AgenTwin. $^{***}p < 0.001$. 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.

A.4 Efficiency Analysis

Runtime analysis shows that Generative Agent takes approximately 7 minutes per test iteration, RecAgent takes 3.53 minutes whereas AgenTwin 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.

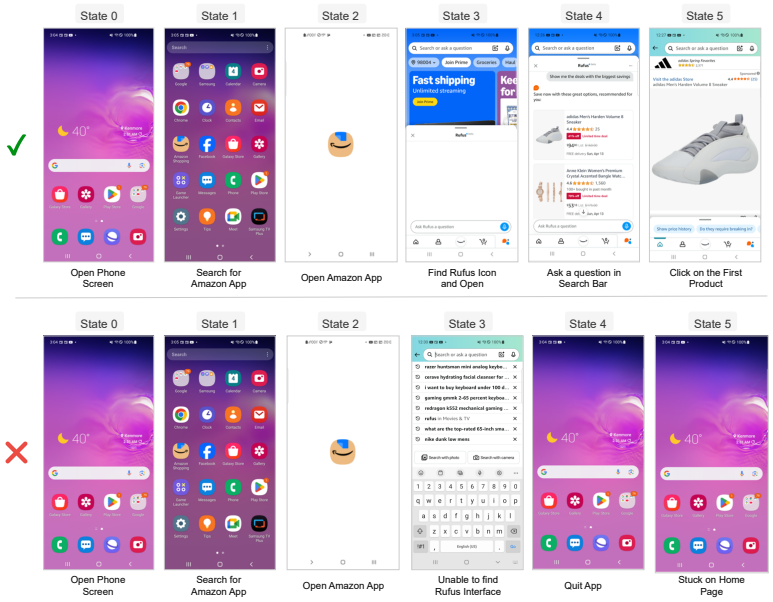

Figure 4: Examples of screenshots from device interaction trajectories used to evaluate AgenTwin's Device Control Module's perception capabilities. Screenshots capture key states during task execution to "open and search products using a shopping assistant on a retail website, and click on a product". 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. AgenTwin's device control module exhibits fine-grained high-fidelity device interactions with an overall success rate of 91.7%.

Aspect	Question	Gen	Rec	AgenTwin
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 & 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 the percentage of positive responses in evaluation ($n = 1000$); best per row in **bold**. Gen = Generative Agent, Rec = RecAgent. AgenTwin excels at user-preference integration and logical transparency, while Generative Agent relies more on price considerations and alternatives. RecAgent scores lower across metrics, indicating limited reasoning ability.

AgenTwin Behavior Module Template

Input Parameters

- **persona:** Customer 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 5: AgenTwin 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 it to simulate realistic multi-item decision making.