To See or To Read: User Behavior Reasoning in Multimodal LLMs

Tianning Dong* **Luyi Ma*** **Varun Vasudevan** **Jason Cho** **Sushant Kumar**
Kannan Achan
Personalization Team, Walmart Global Tech
Sunnyvale, California, USA
{tianning.dong, luyi.ma, varun.vasudevan, jason.cho, sushant.kumar, kannan.achan}@walmart.com

Abstract

Multimodal Large Language Models (MLLMs) are reshaping how modern agentic systems reason over sequential user-behavior data. However, whether textual or image representations of user behavior data are more effective for maximizing MLLM performance remains underexplored. We present `BehaviorLens`, a systematic benchmarking framework for assessing modality trade-offs in user-behavior reasoning across six MLLMs by representing transaction data as (1) a text paragraph, (2) a scatter plot, and (3) a flowchart. Using a real-world purchase-sequence dataset, we find that when data is represented as images, MLLMs next-purchase prediction accuracy is improved by 87.5% compared with an equivalent textual representation without any additional computational cost.

1 Introduction

The ability to reason about user behavior from sequential data is a cornerstone of modern intelligent systems, enabling applications such as personalized recommendation (transaction history), chatbot (dialogue history) and proactive user support (meeting notes) [Ma et al., 2025, Zhao et al., 2024, Singh et al., 2024]. Advancement in Multimodal Large Language Models (MLLMs) has fundamentally shifted the paradigm for this task, replacing specialized models with multimodal general-purpose reasoning engines [Li et al., 2024, Shahriar et al., 2024].

This new paradigm, however, introduces a critical and under-explored question: **how should sequential user histories be represented to optimize for both reasoning accuracy and computational efficiency** [Ai et al., 2024, Li and Jiang, 2025, Imam et al., 2025, Yang et al., 2025]? This question is particularly critical in agentic recommendation systems [Huang et al., 2025, Shang et al., 2025], where understanding the user journey and their transition directly impacts personalization and revenue.

The complex user journey with a high-dimensional sequence of clicks, views, and purchases provides a representative testbed for this challenge via next purchase prediction [Chen et al., 2022]. Effective reasoning requires capturing latent intent and its transition from these sequences, such as distinguishing “goal-oriented price comparison" from “aimless browsing" [Baubonienė and Gulevičiūtė, 2015, Zeng et al., 2020]. The standard approach of feeding MLLMs a flattened, line-by-line textual transcript of these events preserves granular detail, but it losses the structural information of user journey such as topological pattern. Consequently, this method can be inefficient and fail to grasp the holistic user narrative, degrading the quality of intent prediction.

To address this question, we introduce `BehaviorLens`, a systematic benchmarking framework for evaluating modality trade-offs in user behavior reasoning. `BehaviorLens` directly compares two

*Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Efficient Reasoning.

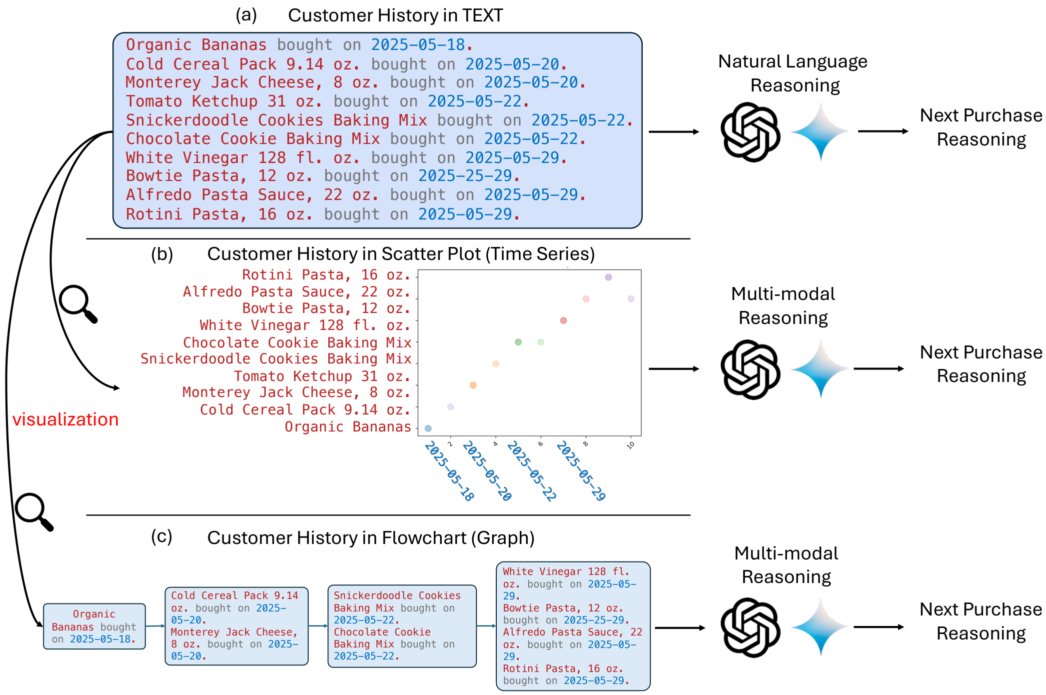

Figure 1: A schematic showing different representations (text, scatter plot, and flow chart) of user history for next purchase prediction using multimodal LLMs.

representations for the same user history: (i) a high-fidelity textual transcript and (ii) a holistic visual rendering. We evaluate these using state-of-the-art MLLMs on a next-purchase prediction task with a fixed candidate set, assessing both the prediction and the computational cost. Applied to real-world user transaction data, `BehaviorLens` reveals a key finding: holistic image representations of user history can help improve next-purchase prediction accuracy when compared to text representations, with no additional computational cost.

By systematically comparing modalities under controlled conditions, `BehaviorLens` contributes both a reproducible methodology for evaluating multimodal reasoning and empirical insights into the efficiency-performance trade-offs of input representations. Beyond next purchase prediction, our framework provides a template for optimizing user historical data streams for MLLMs in any domain involving complex sequential data, supporting future research on efficient context management, data compression, and the development of more robust reasoning systems.

2 Methodology

An agentic recommendation system relies on MLLM's reasoning of user history to understand the user intent. Formally, the reasoning process could be defined as an optimization of the recommendation policy $\pi(a)$ given the user u's history,

$$\max_{\pi(a)} \mathbb{E}[\mathcal{R}_{MLLM}(u, [\phi(a, i, e)]_n)) | \pi(a)], \tag{1}$$

where $\mathcal{R}$ refers to the utility score of user reasoning based on the last n interactions $[\phi(a, i, e)]_n$. Each interaction is defined by a set of actions $a \in A$ (e.g., purchase), a set of engaged items $i \in \mathcal{I}$, and the environment context $e \in \mathcal{E}$ (e.g., timestamp). We study the impact of user behavior representation function $\phi(a, i, e)$ on user intent reasoning efficiency $\mathcal{R}$.

2.1 User Behavior Representations

We compare three ways of representing the user journey as MLLM input with the focus on the purchase history, $a =$ "purchase".

Textual Sequential Representation: the widely-adopted solution for user modeling [Liao et al., 2023] where each tuple (a, i, e) is represented as a natural language description

$$\phi_{text}(a, i, e) = \text{"\textit{item \{i\} was \{a\} at timestamp \{e\}}"}, \tag{2}$$

and the purchase history $[\phi(a, i, e)]_n$ is a concatenation of all the purchase event description.

Scatter-plot Representation: The same purchase history is transformed into a scatter plot image to visualize the temporal patterns as a time series. It is inspired by visualization in time-series modeling tasks [Wang and Oates, 2015],

$$\phi_{scatter-plot}(a, i, e) = \text{plot}(a, i | x = r(e), y = r(i)). \tag{3}$$

Here, $r(\cdot)$ denotes a transformation function to map the value into the coordinate system. For example, $r(\cdot)$ could be a ranking function, mapping each input element to its corresponding ordinal rank.

Flow-chart Representation: Inspired by recent work that claims visual compression aids LLM reasoning in structured tasks [Li and Jiang, 2025], the same purchase history is transformed into a flowchart where purchases as nodes are connected to keep temporal proximity. Typically, we recursively define the flowchart conversion as

$$\phi_{flowchart}(a, i, e) = \text{node}(a, i, e | p = \{\text{node}(a_{-1}, i_{-1}, e_{-1})\}, s = \{\text{node}(a_{+1}, i_{+1}, e_{+1})\}), \tag{4}$$

where the predecessor node set p and the successor node set s only contains the interaction ranked chronologically before (e_{-1}) and after (e_{+1}) the current interaction (a, i, e), respectively.

2.2 User Intent Prediction and Reasoning

For each user, an MLLM receives either the text or the image representation, and is asked to: (1) predict the most likely next purchase $\pi(a)$ from a recall set, (2) provide explanation for the prediction. For each representation, we compare the prediction $\pi(a)$ with the ground-truth user behavior, measure the computational cost using token count as a proxy, and evaluate the generated explanation using LLM-as-a-Judge.

3 Experiments

In this section we summarize which representation offers the best balance between reasoning accuracy and computational efficiency. See Appendix A for details on the dataset and the experimental setup and Appendix B for the full prompts used in the experiments.

3.1 Results

Table 1 shows that representing customer journeys as images helps improve the accuracy, and the similarity score (i.e. cosine similarity of prediction and ground truth in text embeddings). With all the MLLMs, except Gemini-2.5-flash, either scatterplot or flowchart representation results in better prediction than the text representation. Image input achieves at most 33.9% improvement in similarity score and 87.5% improvement in prediction accuracy. This improvement is observed in both larger models, such as GPT-4o, and smaller models, such as GPT-4.1-mini and Gemini-2.5-flash-lite [Hurst et al., 2024, Comanici et al., 2025]. The structure of image representations could further make a significant difference in accuracy. For flowchart, it shows better accuracy and similarity especially for Gemini-2.0-flash and Gemini-2.5-flash, while other MLLMs prefer scatterplot.

Table 1 also shows the computational cost in terms of the total number of tokens used by each MLLM. In Gemini models, flowchart and text input have similar token usage while scatter plots require more tokens. In GPT models, token usage remains at the same level for different representations, while accuracy is significantly higher for image representations. Table 1 further illustrates the latency as the request time in seconds for each MLLM for reference.

To better understand whether the improvement in next purchase prediction is impacted by input representations or also impacted by intermediate explanations provided by MLLMs based on understanding of input, we define six metrics to evaluate the quality of explanations. The metrics are faithfulness, overthinking score, causality, sufficiency, specificity, and plausibility. We use LLM-as-a-judge for the evaluation [Gu et al., 2025]. Prompts used by the judge are provided in Appendix C. Results

Table 1: Comparison of MLLM reasoning performance with text vs image inputs.

LLM Model	Input Type	Accuracy	Similarity	Token Count	Latency(second)
Gemini-2.0-flash-lite	Text	0.260	0.500	1233.46	1.510
Gemini-2.0-flash-lite	Scatterplot	**0.270**	**0.528**	3560.49	2.338
Gemini-2.0-flash-lite	Flowchart	0.260	0.517	1525.07	1.993
Gemini-2.0-flash	Text	0.240	0.483	1228.85	1.444
Gemini-2.0-flash	Scatterplot	0.270	0.510	3596.50	4.580
Gemini-2.0-flash	Flowchart	**0.310**	**0.526**	1529.35	4.977
Gemini-2.5-flash-lite	Text	0.360	0.570	1219.72	1.444
Gemini-2.5-flash-lite	Scatterplot	**0.530**	**0.689**	3623.25	2.057
Gemini-2.5-flash-lite	Flowchart	0.300	0.530	1566.41	1.814
Gemini-2.5-flash	Text	**0.260**	**0.479**	3585.01	1.444
Gemini-2.5-flash	Scatterplot	0.210	0.471	7223.44	6.003
Gemini-2.5-flash	Flowchart	0.220	0.447	5281.78	4.966
GPT-4o	Text	0.420	0.602	1105.52	5.451
GPT-4o	Scatterplot	**0.560**	**0.713**	1169.48	8.954
GPT-4o	Flowchart	0.300	0.527	1042.78	7.140
GPT-4.1-mini	Text	0.320	0.542	1105.27	4.680
GPT-4.1-mini	Scatterplot	**0.600**	**0.726**	1039.24	7.849
GPT-4.1-mini	Flowchart	0.340	0.563	862.07	6.051

are summarized in Figure 2, which show that the quality of explanations on understanding input representations is not significantly different across different inputs, except Gemini-2.0-flash model. This suggests that the choice of user history representation is the primary driver of improvement in next purchase prediction.

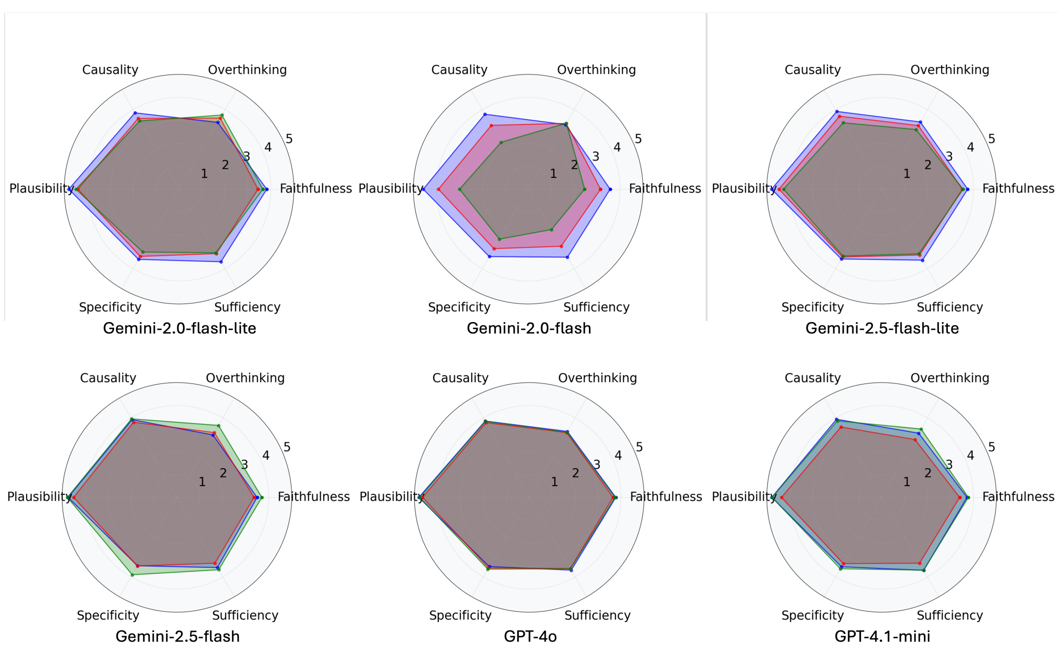

Figure 2: Comparison in MMLM explanations. Each metric is measured by a scale of 1 to 5. Higher score indicates better reasoning in given metric. Blue dots are reasoning of text input. Green dots are reasoning of scatterplot input. Red dots are reasoning of flowchart input.

3.2 Case Study

Given the sample user prompts in Appendix B, take the output from Gemini-2.5-flash as an example to understand difference in reasoning and prediction for different inputs. The sample customer actually purchased items from Crackers and Granola Bars after the most recent 20 purchases provided as input.

In the reasoning by text input, it focuses more on the frequency of purchase patterns to predict next item as Cola, while also mentions different possibilities as snacks, which are close to user's next choice in reality. In the reasoning by flowchart input, it focuses more on the recency of purchase and considers the complementary item for the last purchased item. In the reasoning of scatterplot input, it better captures the recurring pattern of purchase, which is unique compared to the other two inputs, and leads to the correct prediction matching with user's interest.

Output from Text Input

PREDICTION: Cola.
REASONING: The customer purchased Coca-Cola Classic Soda Pop Fridge Pack on both 2019-10-26 and 2019-11-01. This demonstrates a consistent and recurring need for this specific product type, making it highly probable for a subsequent purchase. While other options like Pastries (given the purchase of Toaster Pastries) or other snack/breakfast items are plausible, the direct repeat purchase of Cola makes it the strongest prediction.

Output from Flowchart Input

PREDICTION: Pastries.
REASONING: The last recorded purchase (indicated by arrow 19) was 100% Arabica Medium Roast Ground Coffee Pods. Coffee is frequently consumed with breakfast items or snacks. Looking at the user's past purchases, they have previously bought Frosted Toaster Pastries (transaction 14), which falls under the Toaster Pastries product type. This indicates a preference for pastry-like items. Therefore, Pastries is a highly probable next purchase as it complements coffee and aligns with the user's established buying habits for breakfast/snack items.

Output from Scatterplot Input

PREDICTION: Granola Bars.
REASONING: The user's purchase history shows a recurring pattern of buying Multi Pack Snacks. The most recent purchase (order 20) was Classic Potato Chips, which falls under the Multi Pack Snacks category. Prior to that, OREO Cookie Sticks 'N Creme Dip Snack Packs (order 12) was also purchased from this category. Granola Bars are a very common and popular type of snack item, often sold in multi-packs, making them a highly plausible next purchase to replenish or vary the user's snack supply, directly continuing a recent and established purchasing trend.

4 Conclusion

In this paper we investigate the efficiency of reasoning over sequential user behavior data using different representations. We observe significant improvement in prediction accuracy with image representations. This finding advances our understanding of how MLLMs interpret the user journey and holds promise for enhancing personalization in agentic recommendation systems. Future work can focus on further optimizing the visual representations of the user journey to better capture temporal and spatial dynamics in the behavior. Future work should also investigate the impact of longer customer behavior sequences and should benchmark the study on a larger dataset.

References

Luyi Ma, Wanjia Zhang, Kai Zhao, Abhishek Kulkarni, Lalitesh Morishetti, Anjana Ganesh, Ashish Ranjan, Aashika Padmanabhan, Jianpeng Xu, Jason HD Cho, et al. Grace: Generative recommendation via journey-aware sparse attention on chain-of-thought tokenization. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pages 135–144, 2025.

Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, et al. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6889–6907, 2024.

Ravinder Singh, Jawahar Thakur, and Yogesh Mohan. A historical analysis of chatbots from eliza to google bard. In *Doctoral Symposium on Computational Intelligence*, pages 15–39. Springer, 2024.

Zhiming Li, Yushi Cao, Xiufeng Xu, Junzhe Jiang, Xu Liu, Yon Shin Teo, Shang-Wei Lin, and Yang Liu. Llms for relational reasoning: How far are we? In *Proceedings of the 1st International Workshop on Large Language Models for Code*, pages 119–126, 2024.

Sakib Shahriar, Brady D Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool. Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *Applied Sciences*, 14(17):7782, 2024.

Qihang Ai, Jiafan Li, Jincheng Dai, Jianwu Zhou, Lemao Liu, Haiyun Jiang, and Shuming Shi. Advancement in graph understanding: A multimodal benchmark and fine-tuning of vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7485–7501, 2024.

Ruizhou Li and Haiyun Jiang. Graph-to-vision: Multi-graph understanding and reasoning using vision-language models. *arXiv preprint arXiv:2503.21435*, 2025.

Mohamed Fazli Imam, Chenyang Lyu, and Alham Fikri Aji. Can multimodal llms do visual temporal understanding and reasoning? the answer is no! *arXiv preprint arXiv:2501.10674*, 2025.

Wennuo Yang, Shiling Wu, Yuzhi Zhou, Cheng Luo, Xilin He, Weicheng Xie, Linlin Shen, and Siyang Song. Benchmarking graph representations and graph neural networks for multivariate time series classification. *arXiv preprint arXiv:2501.08305*, 2025.

Chengkai Huang, Junda Wu, Yu Xia, Zixu Yu, Ruhan Wang, Tong Yu, Ruiyi Zhang, Ryan A Rossi, Branislav Kveton, Dongruo Zhou, et al. Towards agentic recommender systems in the era of multimodal large language models. *arXiv preprint arXiv:2503.16734*, 2025.

Yu Shang, Peijie Liu, Yuwei Yan, Zijing Wu, Leheng Sheng, Yuanqing Yu, Chumeng Jiang, An Zhang, Fengli Xu, Yu Wang, et al. Agentrecbench: Benchmarking llm agent-based personalized recommender systems. *arXiv preprint arXiv:2505.19623*, 2025.

Xin Chen, Alex Reibman, and Sanjay Arora. Sequential recommendation model for next purchase prediction. *arXiv preprint arXiv:2207.06225*, 2022.

Živilė Baubonienė and Gintarė Gulevičiūtė. E-commerce factors influencing consumers 'online shopping decision. *Social technologies*, 5(1):62–73, 2015.

Jiangwei Zeng, Yan Chen, Haiping Zhu, Feng Tian, Kaiyao Miao, Yu Liu, and Qinghua Zheng. User sequential behavior classification for click-through rate prediction. In *International Conference on Database Systems for Advanced Applications*, pages 267–280. Springer, 2020.

Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, and Xiang Wang. Llara: Aligning large language models with sequential recommenders. *CoRR*, 2023.

Zhiguang Wang and Tim Oates. Imaging time-series to improve classification and imputation. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 3939–3945. AAAI Press, 2015. ISBN 9781577357384.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL `https://arxiv.org/abs/2411.15594`.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks, 2016. URL `https://arxiv.org/abs/1511.06939`.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation, 2018. URL `https://arxiv.org/abs/1808.09781`.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer, 2019. URL `https://arxiv.org/abs/1904.06690`.

Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. Session-based recommendation with graph neural networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.3301346. URL `https://doi.org/10.1609/aaai.v33i01.3301346`.

Ziyang Wang, Wei Wei, Gao Cong, Xiao-Li Li, Xian-Ling Mao, and Minghui Qiu. Global context enhanced graph neural networks for session-based recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 169–178, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. doi: 10.1145/3397271.3401142. URL `https://doi.org/10.1145/3397271.3401142`.

Jiaxi Tang and Ke Wang. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 565–573, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355810. doi: 10.1145/3159652.3159656. URL `https://doi.org/10.1145/3159652.3159656`.

Qiao Liu, Yifu Zeng, Refuoe Mokhosi, and Haibin Zhang. Stamp: Short-term attention/memory priority model for session-based recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 1831–1839, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219950. URL `https://doi.org/10.1145/3219819.3219950`.

Luyi Ma, Xiaohan Li, Zezhong Fan, Kai Zhao, Jianpeng Xu, Jason Cho, Praveen Kanumala, Kaushiki Nag, Sushant Kumar, and Kannan Achan. Triple modality fusion: Aligning visual, textual, and graph data with large language models for multi-behavior recommendations. *arXiv preprint arXiv:2410.12228*, 2024.

Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. Llmrec: Large language models with graph augmentation for recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 806–815, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703713. doi: 10.1145/3616855.3635853. URL `https://doi.org/10.1145/3616855.3635853`.

Deepesh V. Hada, Vijaikumar M., and Shirish K. Shevade. Rexplug: Explainable recommendation using plug-and-play language model. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 81–91, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462939. URL `https://doi.org/10.1145/3404835.3462939`.

Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524*, 2023.

Yang Zhou and Yan Huang. Deepmove: Learning place representations through large scale movement data. In *2018 IEEE international conference on big data (big data)*, pages 2403–2412. IEEE, 2018.

Ziqiang Cheng, Yang Yang, Wei Wang, Wenjie Hu, Yueting Zhuang, and Guojie Song. Time2graph: Revisiting time series modeling with dynamic shapelets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3617–3624, 2020.

A Dataset Preparation and Metrics

A.1 Dataset

To assess the user behavior reasoning, we sample real-world user shopping interaction from one of the largest e-commerce platform in the United States. We design a benchmark dataset for next-purchase prediction from user purchase histories.

User Historical Data: we use the last $n = 20$ purchases as the historical data for intent reasoning and prediction. To better focus on the intent prediction and reasoning, we choose the product type of the next purchased items as the ground truth data. The statistics of the benchmark dataset are as shown in Table 2.

Table 2: Basic statistics of experiment data

Description	Count
Number of users	100
Number of items	1537
Number of product types	268

For text-based representation, we provide the MLLM with:

- **Text input**: A sequential description of the user's purchase history including product names, product types and purchase time in the order they were bought.
- **Multiple-choice candidates**: A set of 20 product types from which the model must select the next likely purchase. Among these 20 product types, 2 product types are the ground truth of user's next purchases while the other 18 product types are randomly sampled from product types that have not been purchased by the customer.

For image-based representation, we provide the MLLM with:

- **Image input**: A flowchart or scatterplot representation of the same purchase history, visually encoding product types and temporal order.
- **Multiple-choice candidates**: A set of 20 product types from which the model must select the next likely purchase. The product types are consistent with multiple choice candidates for text input.

The LLM is asked to produce two outputs:

- **Prediction**: a predicted next product type from the candidates,
- **Reasoning**: a reasoning explanation describing how the input led to its choice.

For evaluation purpose, we also maintain the ground truth of customer's actual next purchase as labels.

A.2 Metrics and Evaluation

We evaluate results from two perspectives: 1) prediction, 2) reasoning.

Prediction accuracy: whether the predicted product type matches a ground-truth products.

- Prediction accuracy: the percentage of users who have ground truth of next-item purchases as an exact match of MLLM predictions.
- Similarity score: the maximum of cosine similarities between product types in ground truth and product types in MLLM predictions.

Reasoning Evaluation:

- Faithfulness (1-5): How accurately does the reasoning reflect customer purchase history?

- Overthinking Score (1-5): How well does the reasoning avoid mentioning irrelevant information?
- Causality (1-5): How well does the reasoning present a well-structured argument, or just a list of observations?
- Rationale Plausibility (1-5): How is the reasoning logical and easy to understand?
- Rationale Specificity (1-5): How does the reasoning cite specific data points instead of making generic claims?
- Rationale Sufficiency (1-5): How does the reasoning provide enough evidence to be truly persuasive?

B Prompts for User Behavior Reasoning

This section shows sample prompts used to predict next best item and provide reasoning. The below example of user history is synthetic data. For text input, the user history is provided in a loggings format with three lines per purchase to describe product names, types and timestamps. For image input, the user history is provided in a flowchart or scatterplot with further instructions on how to read the plots.

System Prompt

You are a helpful assistant designed to analyze user behavior in e-commerce. Your goal is to predict the user's next action and provide a brief, data-driven reasoning for your prediction based on the provided user history.

Prompt for User History in Text

USER HISTORY:

```
- Product name: Steak & Chop Marinade
- Product type: Barbecue Sauce & Marinades
- The customer bought this product on 2019-10-26 at 12:15:27.
- Product name: Italian Style Finely Shredded Cheese
- Product type: Shredded Cheese
- The customer bought this product on 2019-10-26 at 12:16:22.
- Product name: Cubed Colby & Monterey Jack Cheese
- Product type: Cubed & String Cheese
- The customer bought this product on 2019-10-26 at 12:16:34.
- Product name: White Round Top Bread Loaf
- Product type: White Bread
- The customer bought this product on 2019-10-26 at 12:17:39.
- Product name: Beefsteak No Seeds Rye Bread
- Product type: White Bread
- The customer bought this product on 2019-10-26 at 12:17:39.
- Product name: Mandarin Orange Sparkling Water
- Product type: Water Enhancers
- The customer bought this product on 2019-10-26 at 12:17:39.
- Product name: Coca-Cola Classic Soda Pop Fridge Pack
- Product type: Cola
- The customer bought this product on 2019-10-26 at 12:17:39.
- Product name: Cherry Limeade Sparkling Water
- Product type: Water Multipacks
- The customer bought this product on 2019-10-26 at 12:17:58.
- Product name: Glacier Freeze/Cherry/Arctic Blitz Sports Drinks
- Product type: Sports & Vitamin Drinks
- The customer bought this product on 2019-10-26 at 12:17:58.
- Product name: 100% Apple Juice
- Product type: Juice
- The customer bought this product on 2019-10-26 at 12:18:13.
```

- Product name: OREO Cookie Sticks ’N Creme Dip Snack Packs
- Product type: Multi Pack Snacks
- The customer bought this product on 2019-10-26 at 12:18:55.
- Product name: Original Pork Breakfast Sausage Roll
- Product type: Breakfast Sausage
- The customer bought this product on 2019-10-26 at 12:20:00.
- Product name: Cinnamon Toast Crunch Whole Grain Breakfast Cereal
- Product type: Cold Cereal
- The customer bought this product on 2019-11-01 at 10:50:11.
- Product name: Frosted Toaster Pastries, Cookies and Cream
- Product type: Toaster Pastries
- The customer bought this product on 2019-11-01 at 10:51:31.
- Product name: Extra Virgin Olive Oil
- Product type: Olive Oil
- The customer bought this product on 2019-11-01 at 10:54:06.
- Product name: 2% Reduced Fat Milk, Refrigerated
- Product type: Dairy Milk
- The customer bought this product on 2019-11-01 at 10:54:52.
- Product name: Coca-Cola Classic Soda Pop Fridge Pack
- Product type: Cola
- The customer bought this product on 2019-11-01 at 10:56:45.
- Product name: Glacier Cherry Sports Drinks
- Product type: Sports & Vitamin Drinks
- The customer bought this product on 2019-11-01 at 10:57:00.
- Product name: 100% Arabica Medium Roast Ground Coffee Pods
- Product type: Coffee Pods
- The customer bought this product on 2019-11-01 at 10:57:12.
- Product name: Classic Potato Chips
- Product type: Multi Pack Snacks
- The customer bought this product on 2019-11-01 at 10:57:12.

Prompt for User History in Flowchart

USER HISTORY: Given the below flowchart representing user purchase history, where arrows with numbers indicate the ordering of purchase.

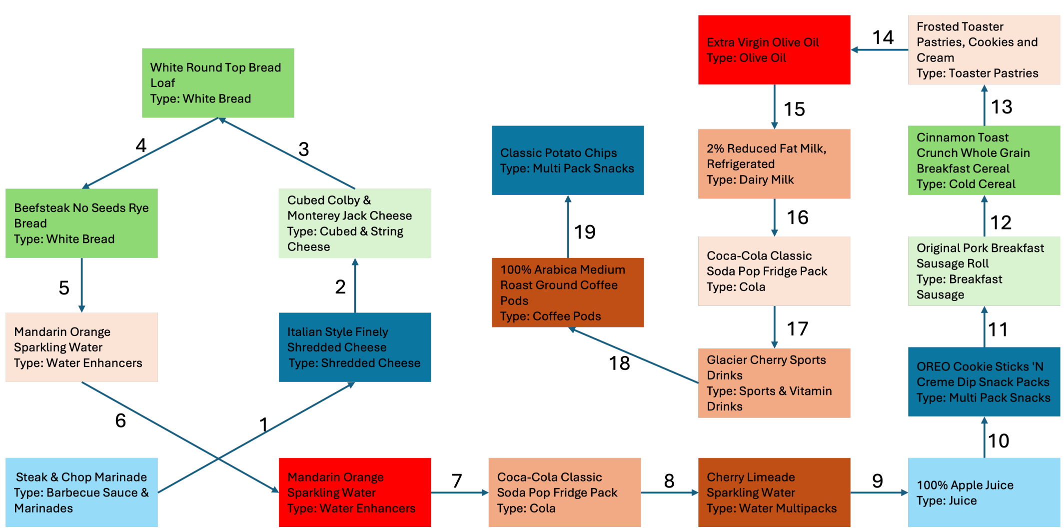

Figure 3: Flowchart of user history

Prompt for User History in Scatterplot

USER HISTORY: Given the below scatterplot representing user purchase history, where x-axis represents the ordering of purchase, and y-axis represents the product types of purchased items.

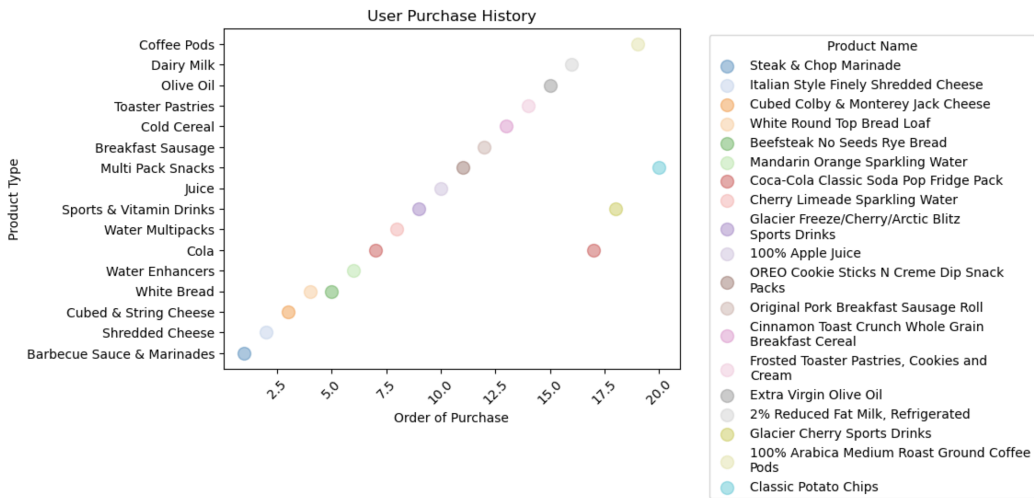

Figure 4: Scatterplot of user history

User Prompt

QUESTION: Based on the user history provided above, predict what might be next possible purchase and explain why, choosing from the given multiple choices. Provide your answer in a json format with prediction result and reasoning as two keys.
CHOICES: Granola Bars, Crackers, Instant Coffee, Cola, Whole Fresh Herbs, Chocolate Multipacks, Margarines, Canned Vegetables, Pregnancy and Ovulation Tests, Broth, Stocks and Bouillon, Lunch Packs, Oatmeal and Hot Cereal, Atkins Test, Holiday Dairy and Egg Nogs, Pastries, Support Hose and Socks, Holiday Bakery, Itch and Rash Treatments, Salad Kits and Bowls, Muffins and Scones.

C Multi-dimensional Evaluation of Explanation

This section shows sample prompts used to evaluate the explanation provided by MLLMs for next-item prediction, including the details of scoring criteria. To better evaluate the explanation, we need to provide ground truth of user purchase history and actual next purchase for reference.

System Prompt

You are an expert evaluator of reasoning for next purchase prediction. Your task is to assess the quality of reasoning based on the given specialized criteria.
1. Faithfulness (1-5): How accurately does the reasoning reflect customer purchase history?

```
- 1: Major ground truth errors, contradicts behavior
- 2: Some ground truth inaccuracies, limited accuracy
- 3: Generally accurate ground truth with minor issues
- 4: Highly accurate ground truth representation
- 5: Perfect ground truth accuracy, nuanced understanding
```

2. Overthinking Score (1-5): How well does the reasoning avoid mentioning irrelevant information?

```
- 1: Exceptional much irrelevant information
- 2: Much irrelevant information
```

```
- 3: Moderate irrelevant information
- 4: Limited irrelevant information
- 5: No irrelevant information
```

3. Causality (1-5): How well does the reasoning present a well-structured argument, or just a list of observations?

```
- 1: No causality, just a list of observations
- 2: Limited causality, basic argument
- 3: Moderate causality, some reasonable argument
- 4: Good causality, clear argument
- 5: Exceptional causality, well-structured argument
```

4. Rationale Plausibility (1-5): How is the reasoning logical and easy to understand?

```
- 1: No irrelevant information
- 2: Limited irrelevant information
- 3: Moderate irrelevant information
- 4: Much irrelevant information
- 5: Exceptional much irrelevant information
```

5. Rationale Specificity (1-5): How does the reasoning cite specific data points instead of making generic claims?

```
- 1: No specific information
- 2: Limited specific information
- 3: Moderate specific information
- 4: Much specific information
- 5: Exceptional much specific information
```

6. Rationale Sufficiency (1-5): How does the reasoning provide enough evidence to be truly persuasive?

```
- 1: No evidence, not persuasive
- 2: Limited evidence, less persuasive
- 3: Moderate evidence, somehow persuasive
- 4: Enough evidence, well persuasive
- 5: Exceptional evidence, very persuasive
```

User Prompt

User History and Ground Truth The below is the list of product types in user purchase history. (Add purchased product types in time order). The user actually purchased (add the purchased product types as ground truth).
Reasoning: Meanwhile the agent predicted the user would have bought (add the predicted next-purchase product type). The below is the reasoning of next purchase prediction by the agent. (Add the reasoning provided by MLLMs).
Questions: Based on reasoning and prediction provided above, generate scores using the above criteria.

D Related Works

D.1 Next Item Prediction

Classical approaches to next item prediction include Markov chains, matrix factorization, and factorization machines, which model user–item interactions using latent factors. With the rise of deep learning, recurrent neural networks such as GRU4Rec [Hidasi et al., 2016] became influential in modeling sequential recommendation, capturing order-dependent behaviors in clickstreams and purchase sequences. Transformer-based architectures, such as SASRec [Kang and McAuley, 2018] and BERT4Rec [Sun et al., 2019], further advanced next-item prediction by leveraging self-attention to model long-range dependencies in user behavior.

Beyond sequential modeling, graph-based approaches such as SR-GNN [Wu et al., 2019] and GCE-GNN [Wang et al., 2020] introduced graph neural networks for capturing complex item–item relations in session-based recommendation. Other notable advances include Caser [Tang and Wang, 2018], which treats user-item interactions as sequences of images, and STAMP [Liu et al., 2018], which focuses on capturing users' short-term interests for better next-item prediction. Graph modality provides extract context for next item prediction with generative models [Ma et al., 2024].

Despite significant improvements in accuracy, these models have limited interpretability of why a particular item is predicted. This has led to a growing interest in approaches that not only predict but also explain user preferences, paving the way for the integration of large language models into recommendation tasks.

D.2 LLM for Recommendation and Reasoning

Large language models (LLMs) such as GPT, LLaMA, and PaLM have recently been explored for recommender systems by framing user histories and items as natural language. Early explorations such as LLMRec [Wei et al., 2024] and ReXPlug [Hada et al., 2021] showed that LLMs can adapt to recommendation tasks with minimal fine-tuning, leveraging their world knowledge to infer user preferences. Other works like P5 proposed a unified framework that casts different recommendation problems into text-to-text tasks, demonstrating the flexibility of LLMs as general-purpose recommenders.

One of the distinctive advantages of LLMs is their ability to generate reasoning alongside predictions. Models such as ChatRec and Explainable GPT-based Recommenders show that natural language explanations can improve user trust and system transparency. Despite these advances, most existing work evaluates LLM reasoning in text-only settings [Gao et al., 2023]. Little is known about how LLMs reason when user histories are compressed into images. Our work is among the first to directly compare the reasoning efficiency of LLMs across textual and image-based representations of customer journeys, expanding the discussion from prediction accuracy to reasoning quality.

D.3 Multi-Modal Representations of Customer Journey

Beyond text-based sequence modeling, more work explores the idea of transforming sequential user interactions into image-like representations that can be processed by convolutional or vision-inspired models. For example, Markov Transition Fields (MTF) and Gramian Angular Fields (GAF) transform time series into images, enabling CNNs to extract spatial dependencies that are otherwise hard to capture in sequential models. In the recommender systems domain, Caser explicitly modeled sequential user–item interactions as a 2D "image" where convolutional filters capture local sequential patterns as well as high-level transition features. Similarly, DeepMove leveraged trajectory data by representing spatio-temporal patterns in image-like forms to predict next locations, which parallels the idea of encoding customer journeys for next-item prediction [Zhou and Huang, 2018].

More recent work has used self-attention and visual encodings to compress large-scale sequential behaviors into compact 2D structures. For instance, Time2Graph converts multivariate sequences into graph-like images [Cheng et al., 2020], while Rec2Image proposed encoding user behavior sessions into image matrices, enabling CNNs to identify higher-order dependencies. These studies demonstrate that compressing sequential histories into visual representations not only reduces input complexity but also allows reasoning models to leverage spatial proximity and clustering cues that may be less salient in raw text sequences.

Despite these advances, prior work primarily used CNN-based architectures for learning from compressed images. Little is known about how large language models (LLMs) reason when fed such image-based summaries compared to traditional text sequences. Our work extends this line of research by investigating reasoning efficiency in LLMs when user journeys are compressed into scatter plot–like images, directly comparing prediction performance and the quality of generated reasoning across modalities.