A Foundation Model for Consumption, Transactions, and Actions: The Inception of BehaviorGPT

Rickard Brüel Gabrielsson,* **et al.**
Unbox AI

Abstract

We present BehaviorGPT-v1, a foundation model for grocery consumption built by applying language modeling techniques to large-scale consumer data. By treating each user's purchase history as a sequence of tokens—"the language of grocery consumption"—we trained a Transformer capable of predicting future consumption patterns. Our dataset spans approximately 600M online actions and 15B offline grocery purchases. The resulting 150M-parameter model incorporates architectural modifications tailored to the unique challenges of this tokenization. We position this work as a step toward a broader foundation model for payments, retail, and ultimately human behavior—what we call BehaviorGPT.

The results were notable:

- 10× improvement in recommendations over baseline,
- +9.4% conversion against RichRelevance search and +5.7% over Algolia,
- +2.2% sales in physical stores after using dense vectors of physical stores to dynamically assign assortments based on regional behavioral patterns.
- Several qualitative demonstrations of substantial performance gains.

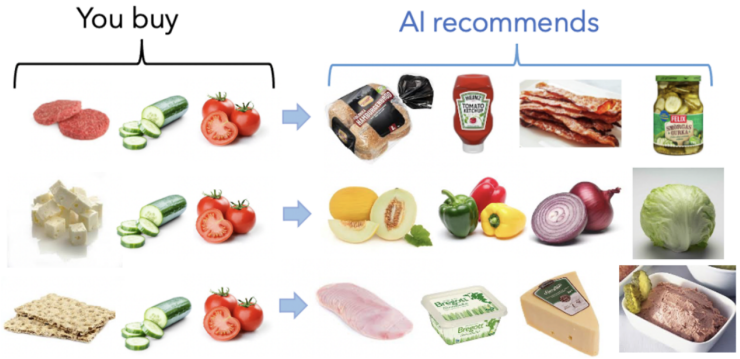

Figure 1: Personalized Recommendations. Changing the third-to-last item in the user's purchase history can steer the model's interpretation of user intent.

1 Introduction

Foundation models—large models trained on massive datasets to learn general-purpose representations—have shown remarkable success in text, vision, and genomics. Their power lies in extracting rich abstractions from diverse data, before adapting to specific downstream tasks.

*Corresponding author: rickard@unboxai.com

Original post (published May 15, 2025): https://research.unboxai.com/foundation-model-for-consumption-transactions-and-actions.html

However, a major slice of modern human-generated data isn't textual or visual at all—it's behavioral. Every day, billions of people create chronological "trails" of discrete actions: transactions, searches, clicks, workforce logs, etc. We see a clear opportunity to model these sequences of actions at scale as the next frontier for foundational AI.

In language modeling, a model learns the probabilities of words (tokens) in a sequence, capturing linguistic and contextual relationships to predict future words. We extend this approach to behavioral data: for instance, user sessions on a retail platform or sequences of payment transactions. By adopting the "next event prediction" paradigm, we can learn powerful representations of both agents (the users) and items or services (the objects of interaction).

Personalization has always been key to higher customer engagement. Accurately predicting the next product or interaction is akin to how large language models "keep track of your conversation's context." In Figure 1, we see examples of how a single tweak in the third-to-last product drastically changes a grocery recommendation, pivoting from hamburgers to salads or to a Swedish breakfast. This ability to present the right product to the right customer at the right time is the core intelligence for anyone selling products. This intelligence also supports search, recommendations, fraud detection, and business intelligence.

We chose groceries as our starting point because grocery data is incredibly rich: people purchase groceries 10× more frequently (roughly ten times as many pieces as clothing and twenty times as many personal-care products, per the BLS Consumer Expenditure Survey) than other product categories, and they typically return to the same stores. This frequency and consistency provide unusually thick transaction sequences. It gives us deep insight into an individual's preferences, spanning myriad product categories. Our analysis even uncovered telling behavior clusters—like a growing health-conscious group and a frequent "feta cheese" cluster—both of which a leading consulting firm missed in a study for the same retailer.

2 Related work

Our work situates a foundation model directly in the domain of *consumption, transactions, and actions*, leveraging language-modeling ideas to learn from long-horizon consumption behavior. We review (i) foundation models and their transfer to behavioral sequences, (ii) multimodal, behavior-grounded item understanding, (iii) sequential recommendation and generative retrieval at retail scale, and (iv) large behavioral models—clarifying how our setting differs.

2.1 Foundation models for behavioral sequences

From language modeling to consumption sequences. Early representation learning for text (e.g., Word2vec) [Mikolov et al., 2013, Firth, 1957] and pre-train–then–fine-tune paradigms (e.g., BERT) [Devlin et al., 2019] paved the way for large auto-regressive models that unify many tasks via next-token prediction and in-context learning [McCann et al., 2018, Brown et al., 2020]. Instruction tuning and preference optimization (e.g., RLHF and GRPO) [Ouyang et al., 2022, DeepSeek-AI et al., 2025] further improved alignment. We adopt the same causal modeling principle, but the "tokens" are discrete human actions (purchases, clicks, searches) and their context. Treating a user's chronological behavioral history as a sequence enables general next-event prediction that powers recommendation, search, fraud detection, and behavioral analytics.

Autoregression and masking in our setting. In language modeling, both autoregressive and masked training objectives have been widely explored; we adopt the same perspective for behavioral modeling. Specifically, we compare causal training over action sequences with masked objectives applied to item descriptors (text and, when available, images), enabling us to compress millions of SKUs into behavior-aware embeddings. This approach parallels—and extends—recent behavior-focused work [Gabrielsson et al., 2025d,b,c].

Self-supervision beyond language. Foundation-style self-supervision has flourished across domains including time series and structured records [Bardes et al., 2022, Baevski et al., 2022, Chen et al., 2020, Zaheer et al., 2021, Wornow et al., 2023, Das et al., 2024, Woo et al., 2024, Gabrielsson et al., 2025d, Brüel-Gabrielsson and Scarvelis, 2022]. Compared to vision/text, transactional streams are heterogeneous, sparse, and evolve quickly; grocery in particular exhibits sharp temporal regime

shifts (holidays, seasons, promotions). We incorporate *domain* cues (date/time, region, device) directly as tokens and use augmentation strategies that promote invariances under such shifts [Chen et al., 2020, Gabrielsson et al., 2025a, Srivastava et al., 2014].

2.2 Behavior-grounded multimodal item understanding

From VFMs to behavior-native semantics. Vision–language foundation models target semantic alignment from pixels to words [Radford et al., 2021, Awais et al., 2023]. In retail, however, the meaning of an item is best revealed by *co-occurrence and substitution* in baskets over time. Our embedder ingests text (and images where available) but is trained end-to-end on *sequential consumption* rather than stand-alone captions, yielding descriptors that align with purchase intent and substitution patterns, and that unify offline and online catalogs.

Descriptor prediction vs. item ID classification. To bridge enormous SKU spaces and frequent catalog churn, we pair two complementary targets: (i) next-item prediction over a large candidate set (for retrieval/ranking), and (ii) next-item *descriptor* generation (for generalization to unseen items and richer reasoning). This stands between pure retrieval and pure text generation, and is tailored to retail dynamics. We further extend this approach with a staged pipeline: Stage 1 focuses on learning a strong embedder, while Stage 2 maximizes coverage of the candidate set, providing a robust foundation for the subsequent prediction and generation tasks.

2.3 Sequential recommendation and generative retrieval at retail scale

From two-tower retrieval to sequential Transformers. Deep recommendation began with dual-encoder retrieval (e.g., YouTube's two-tower) [Covington et al., 2016] and moved toward sequence models [Kang and McAuley, 2018]. Recent surveys document the entrance of LLMs into recommendation pipelines [Wu et al., 2024]. In parallel, large-scale behavior models show promising scaling laws yet often remain confined to single-domain retrieval/ranking [Zhai et al., 2024].

Unified retrieval and ranking via causal sequences. We follow a unified, causal formulation where the same sequence model supports both retrieval and ranking by conditioning on interleaved actions and items, rather than training disjoint systems. Flexible task formatting for recommendation has been explored [Geng et al., 2023], but we find that straightforward auto-regressive training—mirroring language modeling [Raffel et al., 2023, Brown et al., 2020]—remains the most efficient backbone in grocery, especially when augmented with behavior-informed negatives and candidate sampling.

Generative recommenders and quantisation. To handle billion-scale catalogs, generative recommenders compress item IDs via learned quantization, trading embedding tables for compact codes [Lee et al., 2022]. We instead combine efficient GPU/CPU candidate generation with behavior-trained embedders and descriptor prediction, avoiding heavy decoding (e.g., beam search) at inference while retaining generalization to new products.

2.4 Large Behavioral Models

LBMs across domains. Large Behavioral Models have largely emphasized short-horizon control (e.g., robotics, app UIs) [Team et al., 2025, Zhou et al., 2024], whereas we target *long-horizon, high-frequency* consumption with strong seasonal signals and regionality. Our objective is not only to rank within a single app domain, but to learn behavior-native representations that transfer across offline and online retail touchpoints.

Summary of differences. Relative to prior work, our contribution is a *foundation model for consumption* that (1) treats behavior as a language with domain tokens for time and place, (2) learns behavior-grounded multimodal item descriptors end-to-end, (3) unifies offline and online catalogs while scaling to million-plus SKUs, and (4) supports both retrieval/ranking and descriptor generation for cold-start and catalog churn—providing a practical base model for retail search, recommendation, and transaction intelligence.

3 Definitions

3.1 CartMetric

In basket–completion, the exact order of forthcoming items is often immaterial—predicting *Fanta* instead of *Coca-Cola* is still far more useful than predicting *Salmon*, yet standard cross-entropy penalises both mistakes equally. Unlike language modelling, where positional fidelity is indispensable, cart prediction benefits from a loss that tolerates such permutation slack. While a sufficiently large model could, in theory, discover this nuance from data alone, we make it explicit with **CartMetric**. Given past carts $\big(c_1, \ldots, c_{k-1}\big)$ and the subset of items already placed in the current cart $c'_k \subset c_k$, $\mathrm{CartMetric}\big(n, (c_1, \ldots, c_{k-1}), c'_k, c_k\big)$ measures the probability that the model's top-n predictions intersects the unseen set $c_k \setminus c'_k$. By rewarding any correct item regardless of position, the metric better aligns evaluation with real-world recommendation quality.

Finally, we ensure that *each ground-truth product can contribute to the score at most once per cart*. Let $H_t \subseteq \widehat{c}_t$ be the set of correct hits among the model's top-n predictions at step t. We maintain a global hit set $H = \varnothing$ for the entire session and update $H \leftarrow H \cup (H_t \setminus H)$ after every step; only the $|H_t \setminus H|$ *new* hits are credited. Consequently, predicting *Fanta* correctly at multiple steps yields a single reward, forcing the model to surface *novel* items to improve its overall score and preventing precision inflation through repeated mentions of the same popular product.

4 Iterations

4.1 One-hot embeddings and domains

Our initial approach was straightforward: treat each product as a unique token and apply a causal language modeling scheme to predict the next product based on past products in the sequence—see Figure 2 for an architectural sketch. However, since we also get multiple pieces of contextual information when a user visits (e.g., region, date, device, demographics), we introduced domain embeddings for these factors. We concatenated these domain tokens with each product token embedding in the sequence.

We measured performance using "CartMetric"—the accuracy of predicting an item that actually appeared in the user's next cart, given previous carts. Adding domain tokens improved our test-set CartMetric from 35% to 38.6% (a 10.3% improvement), especially for shorter sequences and cold-start situations. Further analysis revealed date/time (month, day of week) was most impactful, followed by region, then device.

Compared to a company baseline that simply recommended each user's most frequently clicked item, our model delivered a 10.5× lift in recommendation conversion. We also leveraged it for search by combining our model's probability scores with a simple Elasticsearch mechanism, which outperformed RichRelevance by 9.4%.

In Figure 3, you can see the learned embedding projected onto a 2D space using UMAP. The dense embedding vectors were semantically meaningful and useful for understanding assortment, finding errors in categorization, and improving categorization and assortment.

We also analyzed and embedded users to understand user groups; see Figure 4. We uncovered a growing health-conscious group and another uniquely characterized by frequent feta cheese purchases—both of which a leading consultancy firm's consumer group study overlooked for the same company.

4.2 MLM on text descriptions

As we moved on to include 15B offline transactions spanning 1M unique events, having a 1M-sized output softmax layer became expensive. Moreover, offline and online product IDs didn't match across regions. So, we shifted to using product descriptions for identification, resulting in a reduced vocabulary of approximately 50K tokens.

We tried a Masked Language Modeling (MLM) approach, illustrating with a tiny two-item example:

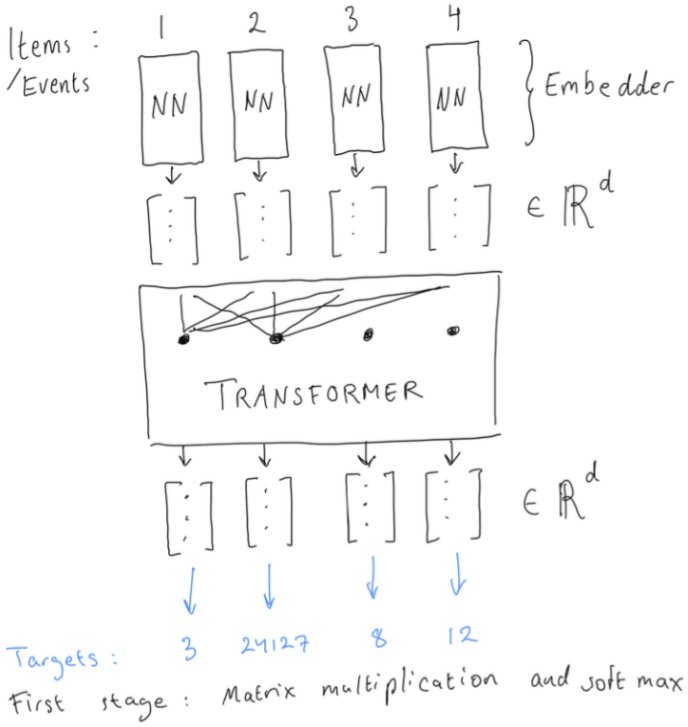

Figure 2: Architectural sketch showing embedder with transformer core (contextualizer).

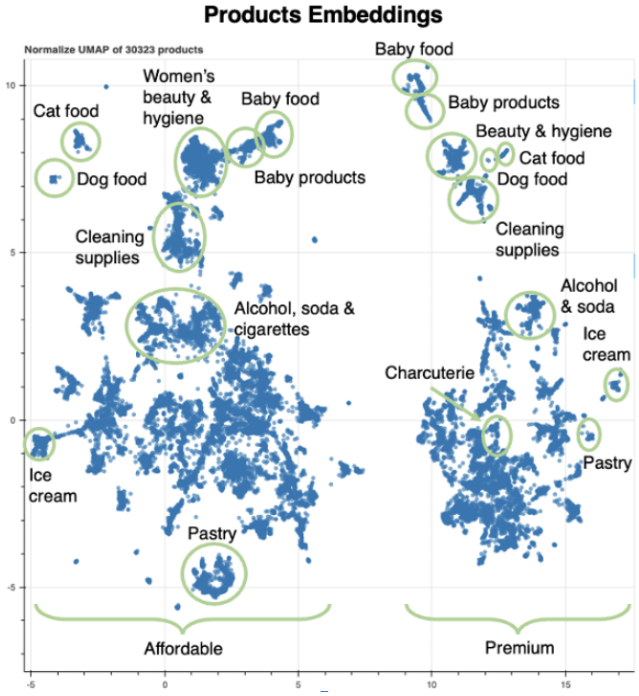

Figure 3: A 2D UMAP projection of product embeddings from this first model. Similar products cluster tightly, guiding improvements in categorization and assortment.

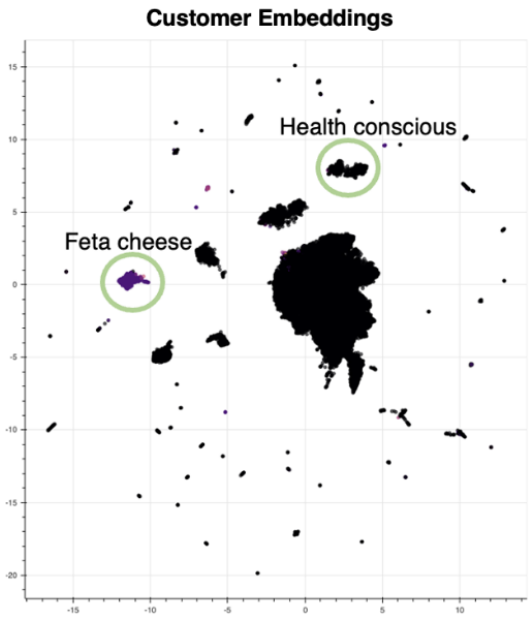

Figure 4: Customer embeddings revealing interesting shopper cohorts.

```
pasta 1kg, barilla, $7.3 <sep> rao's tomato basil sauce, $6.8
<sep>
```

We tested:

1. Random Masking: Standard MLM.
2. Mask All Text for a Single Product: Instead of partially masking a product, we masked its entire description at once.

The second strategy worked better, leading to behavior-driven embeddings that grouped together products commonly bought together, even if their descriptions varied. The idea was that "meaning is defined by the company it keeps," and here we want the "company" to be co-purchased items rather than lexical similarity in the product name.

These learned embeddings were used to cluster stores and enable dynamic assortment assignments based on regional behavioral patterns, boosting physical store sales by 2.2%. See Figure 5, for the colored clusters of store embeddings across Sweden.

4.3 End-to-end image, text, and product embeddings

We wanted to handle multiple data sources (images, text, and more) seamlessly, including user searches. Since we can't enumerate every possible user query as a token, we needed a flexible model that understands items from their text, visuals, and historical context—and also generalizes to unseen products.

In Figure 6, you can see the full data with images, text, etc. We want to use vision to interpret images and language processing to understand text, ideally end-to-end so we learn these features based on behavior rather than other semantic aspects.

Our solution was to replace the simple one-hot embedder with a feature-embedder that ingests text and image features and outputs a dense vector. We then trained in two stages:

1. Stage 1: Train a product embedder for the input events, together with the Transformer "contextualizer" in the middle, but with a large output matrix for the 1M+ products as tokens.

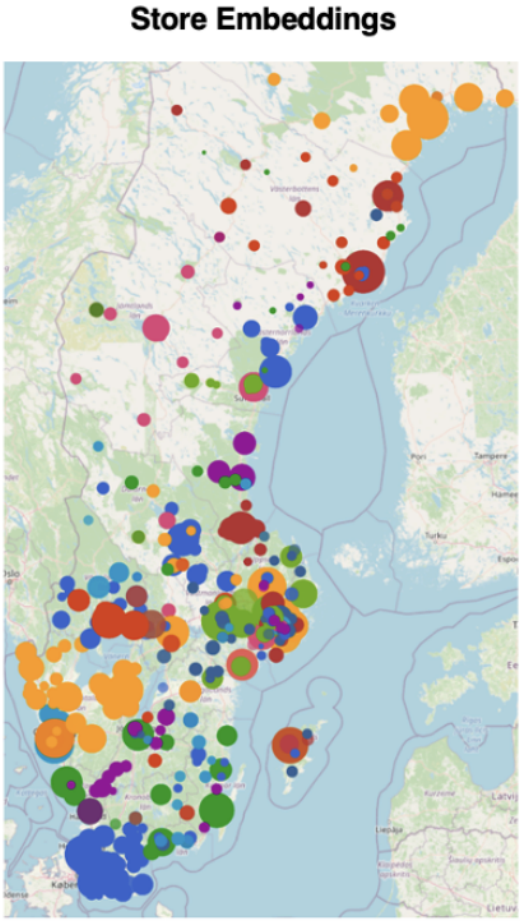

Figure 5: Store embeddings derived from the product embeddings, illustrating how store-level demand patterns can be clustered and optimized.

2. Stage 2: Freeze or partially freeze that embedder, cache the embeddings, and train the main Transformer (the “contextualizer”) to predict which product comes next, given the stacked cached embeddings as targets. See Figure 7.

For input features, this approach is efficient because there are fewer events in the input than candidate events—particularly when using the traditional method of computing distributions over an entire assortment (e.g., 1M possible events). However, we shifted from training embeddings exclusively on inputs to jointly training an embedder on both inputs and outputs, employing smart sampling of candidates. Subsequently, we fine-tuned the model with the embedder frozen. Additionally, we introduced a decoder approach that predicts the text (and other features) of the next event instead of merely retrieving it, enabling more generative capabilities—see Figure 8.

Including text and images allowed the model to discover behavioral similarities more efficiently. It proved especially critical for search, where user queries are short and brand-specific. We found that training the tokenizer and text embedder directly on the sequence task (rather than a general-purpose autoencoder) was vital—likely because grocery consumption has a very domain-specific "language."

This approach handily beat Algolia by 5.7% and Loop54 by 7.5% (two commonly used search-tools) in conversion, despite both solutions relying on considerable manual tagging. Our new model also produced search term suggestions that increased interactions by 49.1%. As a recommendation bar, it achieved a 54% conversion rate—reflecting the system’s deep personalization. Additionally, we benchmarked this search against text and sentence embedders, which simply did not stand a chance.

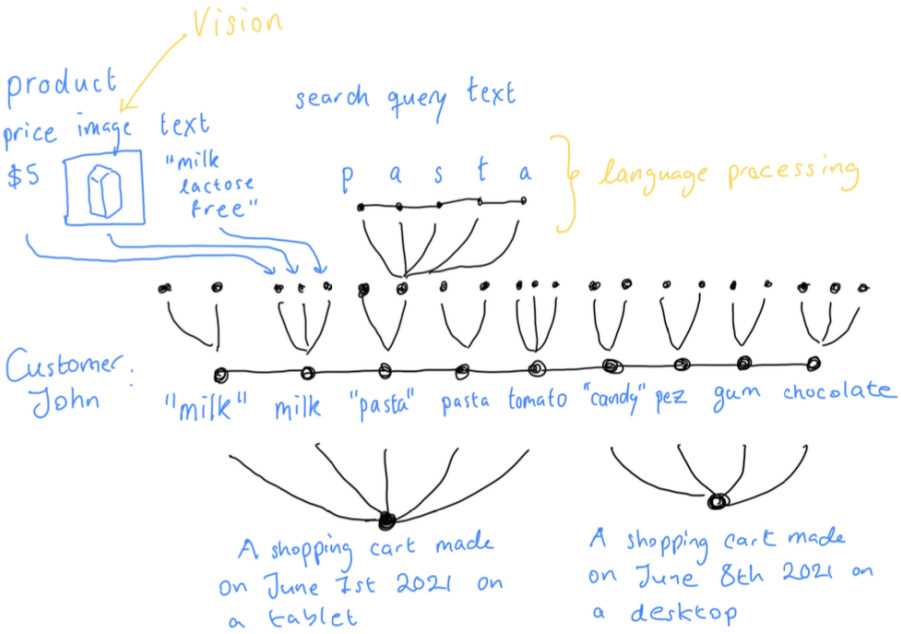

Figure 6: Data flow diagram, showing how images, text, and product features feed into one main chronological action sequence. "Cart" here refers to a session of actions, including purchases, clicks, and searches.

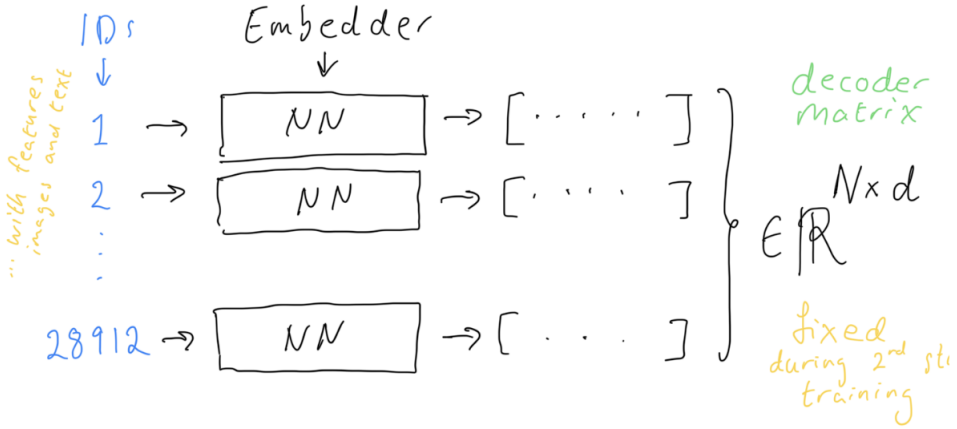

Figure 7: Learning to embed product/event features into dense vectors (embeddings) first as a source, then as the target matrix, and ultimately at the same time.

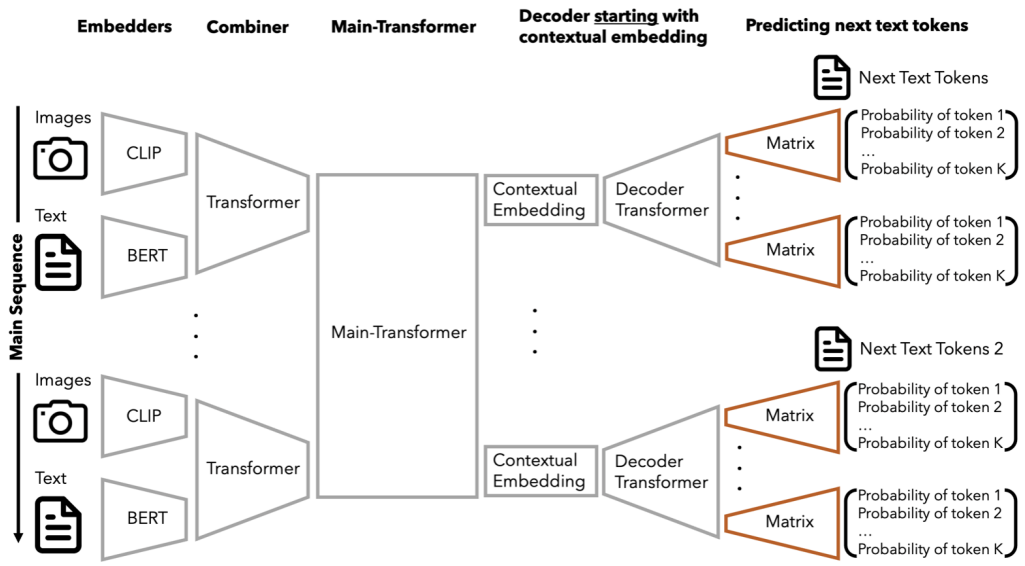

Figure 8: Diagram of model that learns to embed and decode product features, such as text descriptions, end-to-end, enabling enhanced generative capabilities like producing optimal product descriptions and copy.

It's also worth noting that we're modeling a probability distribution p(sequence of behaviors), representing how likely a given sequence of user behaviors is. This enables us to detect fraud by identifying sequences of behavior that are highly improbable. For instance, during self-checkout, if someone buys almost all the ingredients for a dish but omits just one, we can flag this as potentially suspicious—suggesting that item may have been intentionally unpaid. We found this approach more effective than simply using dense semantic transaction vectors or embeddings.

5 Qualitative intelligence: taco, "fredagsmys," and seasonality

The intelligence of this model was astounding. Trained mainly on the Swedish market, it captured Swedish consumption patterns with surprising fidelity.

In Figure 9, you can see the output from simply searching "taco" and then clicking "next" on the recommended item each time. The model assembles a full Swedish taco meal, complete with chips and Coke. If you already had organic items in your basket, it would swap in organic versions seamlessly.

The model also captures abstract and idiosyncratic expressions, like "fredagsmys" (a Swedish phrase for a cozy Friday night at home, typically with candy, soda, and a movie). In Figure 10, you see results that perfectly match this tradition—setting a new standard for the industry.

Domains proved highly useful for time-dependent nuances. In Figure 11, you can see how the model's outputs change from just before Christmas to the Swedish summer holiday, Midsummer. Of course, it adapts daily throughout the year, but these occasions highlight clear seasonal shifts. Previously, manual teams had to configure such seasonal results—costly and time-intensive. Now, the data itself handles these transitions automatically, shaping not only the landing pages but all subsequent search and recommendation results.

6 Discussion

We built this Foundation Model for Grocery Consumption—the inception of BehaviorGPT, now a leading foundation model for payments and retail. Long before "foundation models" and "ChatGPT" became common terms, we showed that scaling laws and transformer-based architectures apply

Figure 9: Typing "taco," then clicking each next recommended product, yields the typical "Swedish taco night" with tortilla bread, sauce, cheese, chips, and soda. If your basket already had organic (called "ecological" in Sweden) items, it swaps in organic versions automatically.

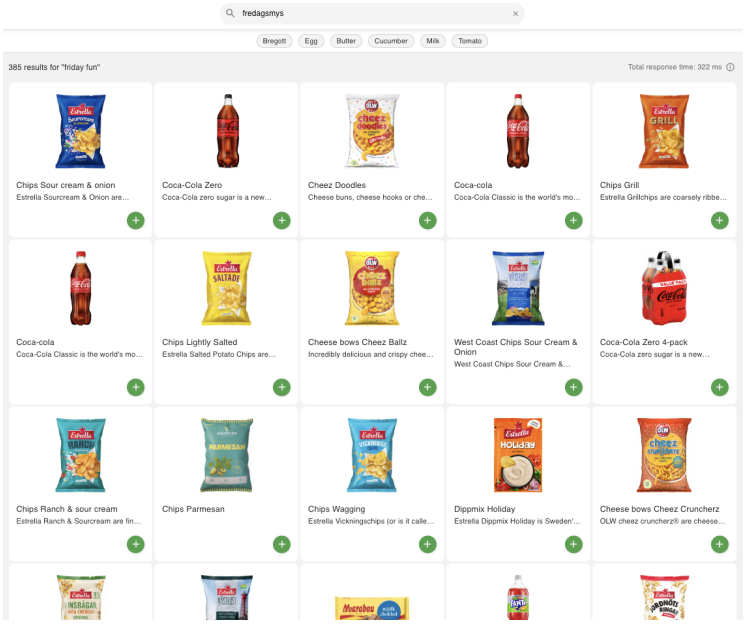

Figure 10: Searching for "fredagsmys"—a Swedish phrase for a cozy Friday evening with snacks—yields a perfect set of candy, chips, and drinks.

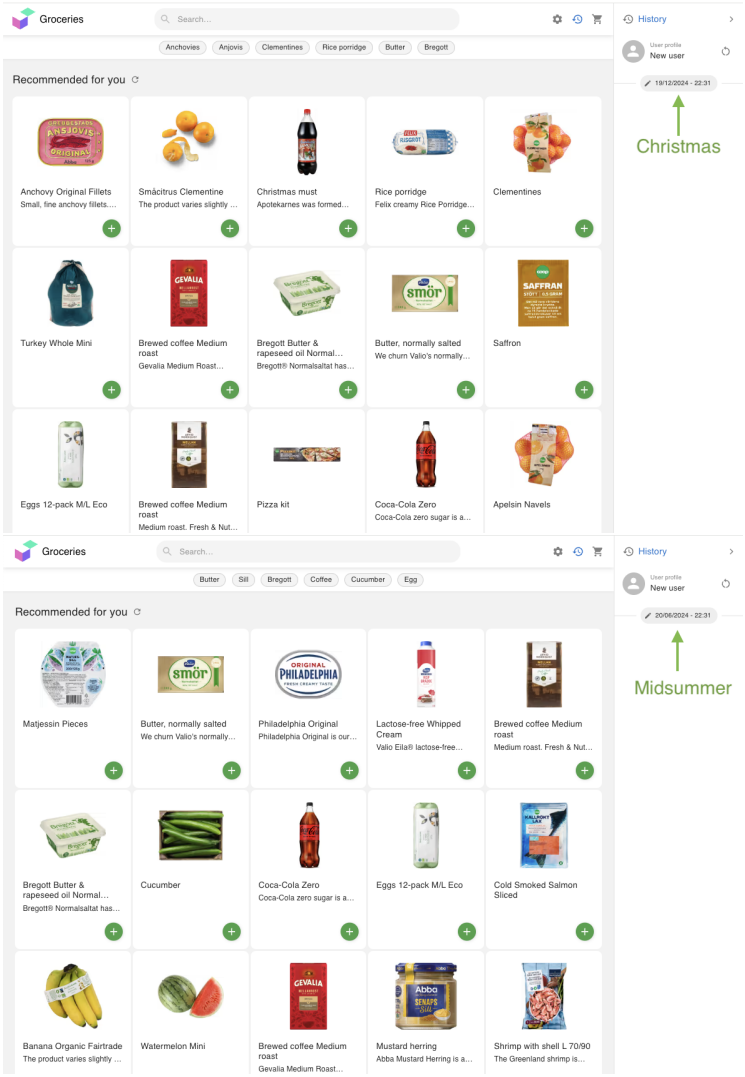

Figure 11: Model outputs before Christmas vs. Midsummer, demonstrating how date embeddings capture holiday-specific preferences.

powerfully to behavioral data. The results outperformed incumbent solutions and added real value to retailers.

Key takeaways include:

- Treating user actions like tokens in a language model offers strong performance on recommendation, search, and beyond.
- Domain embeddings (date/time, region, device) help with seasonality and localization.
- Text/image-based embeddings unify offline, online, and newly introduced products under the same model, and should be trained end-to-end on the sequential task.

BehaviorGPT not only improves user experience (recommendations, search) but also helps with fraud detection, store assortment, and deeper business intelligence.

Acknowledgements

October 2020 pitch deck — our original vision for modelling behaviour, transactions, and retail with language-model techniques.

June 2021 data & model sketches — a outline of the data pipeline and modelling approach.

How to cite

```
@article{unbox2025behaviorgpt,
  author = {Br{\"u}el Gabrielsson, Rickard and others},
  title = {A Foundation Model for Consumption, Transactions, and Actions: The
      Inception of BehaviorGPT},
  journal = {Unbox AI Blog},
  year = {2025},
  month = may,
  url = {https://research.unboxai.com/foundation-model-for-consumption-transactions-
      and-actions.html}
}
```

References

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundational models defining a new era in vision: A survey and outlook, 2023. URL https://arxiv.org/abs/2307.13721.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language, 2022. URL https://arxiv.org/abs/2202.03555.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2022. URL https://arxiv.org/abs/2105.04906.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Rickard Brüel-Gabrielsson and Chris Scarvelis. Relative position prediction as pre-training for text encoders, 2022. URL https://arxiv.org/abs/2202.01145.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. URL https://arxiv.org/abs/2002.05709.

Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys)*, Boston, MA, USA, 2016. ACM, ACM. URL https://storage.googleapis.com/pub-tools-public-publication-data/pdf/45530.pdf. Google, Mountain View, CA.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting, 2024. URL https://arxiv.org/abs/2310.10688.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong,

Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

J. R. Firth. *Papers in Linguistics 1934–1951*. Oxford University Press, London, 1957. "You shall know a word by the company it keeps.".

Rickard Brüel Gabrielsson, Tongzhou Wang, Manel Baradad, and Justin Solomon. Deep augmentation: Dropout as augmentation for self-supervised learning. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856. URL https://openreview.net/forum?id=OjWB2671AR.

Rickard Brüel Gabrielsson, Vasudev Gupta, et al. Behaviorgpt at work: A foundation model for workforce actions & dynamics through large behavioral modeling. *Unbox AI Blog*, jun 2025b. URL https://research.unboxai.com/behaviorgpt-foundation-model-workforce.

Rickard Brüel Gabrielsson, Vasudev Gupta, et al. Behaviorgpt for visual art: A foundation model for aesthetics. *Unbox AI Blog*, jul 2025c. URL https://research.unboxai.com/behaviorgpt-visual-art-and-aesthetics.html.

Rickard Brüel Gabrielsson et al. A foundation model for consumption, transactions, and actions: The inception of behaviorgpt. *Unbox AI Blog*, may 2025d. URL https://research.unboxai.com/foundation-model-for-consumption-transactions-and-actions.html.

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5), 2023. URL https://arxiv.org/abs/2203.13366.

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

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization, 2022. URL https://arxiv.org/abs/2203.01941.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering, 2018. URL https://arxiv.org/abs/1806.08730.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL https://arxiv.org/abs/1910.10683.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1): 1929–1958, January 2014. ISSN 1532-4435.

TRI LBM Team, Jose Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, Masha Itkina, Naveen Kuppuswamy, Kuan-Hui Lee, Katherine Liu, Dale McConachie, Ian McMahon, Haruki Nishimura, Calder Phillips-Grafflin, Charles Richter, Paarth Shah, Krishnan Srinivasan, Blake Wulfe, Chen Xu, Mengchao Zhang, Alex Alspach, Maya Angeles, Kushal Arora, Vitor Campagnolo Guizilini, Alejandro Castro, Dian Chen, Ting-Sheng Chu, Sam Creasey, Sean Curtis, Richard Denitto, Emma Dixon, Eric Dusel, Matthew Ferreira, Aimee Goncalves, Grant Gould, Damrong Guoy, Swati Gupta, Xuchen Han, Kyle Hatch, Brendan Hathaway, Allison Henry, Hillel Hochsztein, Phoebe Horgan, Shun Iwase, Donovon Jackson, Siddharth Karamcheti, Sedrick Keh, Joseph Masterjohn, Jean Mercat, Patrick Miller, Paul Mitiguy, Tony Nguyen, Jeremy Nimmer, Yuki Noguchi, Reko Ong, Aykut Onol, Owen Pfannenstiehl, Richard Poyner, Leticia Priebe Mendes Rocha, Gordon Richardson, Christopher Rodriguez, Derick Seale, Michael Sherman, Mariah Smith-Jones, David Tago, Pavel Tokmakov, Matthew Tran, Basile Van Hoorick, Igor Vasiljevic, Sergey Zakharov, Mark Zolotas, Rares Ambrus, Kerri Fetzer-Borelli, Benjamin Burchfiel, Hadas Kress-Gazit, Siyuan Feng, Stacie Ford, and Russ Tedrake. A careful examination of large behavior models for multitask dexterous manipulation, 2025. URL https://arxiv.org/abs/2507.05331.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers, 2024. URL https://arxiv.org/abs/2402.02592.

Max Wornow, Yikuan Xu, Rishav Thapa, et al. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6:135, 2023. doi: 10.1038/s41746-023-00879-8. URL https://doi.org/10.1038/s41746-023-00879-8.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. A survey on large language models for recommendation, 2024. URL https://arxiv.org/abs/2305.19860.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences, 2021. URL https://arxiv.org/abs/2007.14062.

Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Jiayuan He, Yinghai Lu, and Yu Shi. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp,

editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 58484–58509. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhai24a.html.

Zikang Zhou, Haibo Hu, Xinhong Chen, Jianping Wang, Nan Guan, Kui Wu, Yung-Hui Li, Yu-Kai Huang, and Chun Jason Xue. Behaviorgpt: Smart agent simulation for autonomous driving with next-patch prediction, 2024. URL https://arxiv.org/abs/2405.17372.