# Text2Traj2Text: Learning-by-Synthesis Framework for Contextual Captioning of Human Movement Trajectories

**Hikaru Asano**[1,2*] **Ryo Yonetani**[3] **Taiki Sekii**[3] **Hiroki Ouchi**[4,3,2]
[1]The University of Tokyo [2]RIKEN AIP [3]CyberAgent Inc.
[4]Nara Institute of Science and Technology
asano-hikaru19@g.ecc.u-tokyo.ac.jp,
{yonetani_ryo, sekii_taiki}@cyberagent.co.jp,
hiroki.ouchi@is.naist.jp

## Abstract

This paper presents Text2Traj2Text, a novel learning-by-synthesis framework for captioning possible contexts behind shopper's trajectory data in retail stores. Our work will impact various retail applications that need better customer understanding, such as targeted advertising and inventory management. The key idea is leveraging large language models to synthesize a diverse and realistic collection of contextual captions as well as the corresponding movement trajectories on a store map. Despite learned from fully synthesized data, the captioning model can generalize well to trajectories/captions created by real human subjects. Our systematic evaluation confirmed the effectiveness of the proposed framework over competitive approaches in terms of ROUGE and BERT Score metrics.

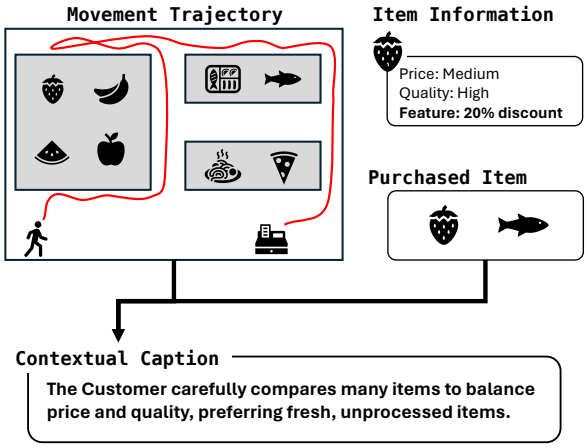



Figure 1: **Contextual Captioning of Human Movement Trajectories.** Given a human movement trajectory associated with semantic information such as nearby items and actual purchases in a retail store, we aim to produce contextual captions that best explain the possible contexts behind.

## 1 Introduction

Retail is an essential industry that is closely tied to our daily lives. Imagine a customer visiting a supermarket. The customer first goes to the fruit section and compares various products. Next, they go to the fish section, where they compare two products. Afterward, they pass by the processed food section and head to the checkout, purchasing discounted organic strawberries and fish. From these movements and purchases, one can guess that "*the customer is budget-conscious and interested in healthy meals.*"

Such profiling and verbalization of possible contexts behind shopping behaviors is vital for retailers to improve customer understanding and customer experience. We are interested in automating this intelligent activity, with recent advances in large-scale language modeling. Doing so would help facilitate and scale up retailer's operations beyond the number of experts, and can also enhance several applications such as targeted advertising (Liu et al., 2018; Ghose et al., 2019) and inventory management (Carreras et al., 2013).

As the first step toward this goal, we formulate a new task, *contextual captioning of human movement trajectories*, with a particular focus on retail applications. Let us illustrate an example in Fig. 1. The input of this task is a movement trajectory associated with its semantic information, such as item positions and purchased items for a customer navigating a retail store. The output is a *contextual caption* that explains a possible context behind the demonstrated trajectories, such as purposes and preferences for the purchases.

While it is intuitive to learn neural captioning models for this task, it is nontrivial how to gather the sufficient number of training data, more specifically trajectories annotated with contextual captions. Although recent advancements in wireless sensing technologies have already enabled accurate indoor localization (Zafari et al., 2019), collecting actual customer locations in stores is often nontriv-

[*]Work done during an internship at CyberAgent Inc.



*Proceedings of the 17th International Natural Language Generation Conference*, pages 289–302
September 23–27, 2024. ©2024 Association for Computational Linguistics

ial due to privacy concerns. Even if location data were available, annotating appropriate captions for them is labor intensive.

In this work, we present TEXT2TRAJ2TEXT, a learning-by-synthesis framework to address this challenge. As illustrated in Fig. 2, this framework consists of two phases: TEXT2TRAJ (data synthesis) and TRAJ2TEXT (model fine-tuning). In the TEXT2TRAJ phase, we leverage large language models (LLMs) to synthesize realistic and diverse collections of contextual captions as well as concrete trajectories on store maps. Then in the TRAJ2TEXT phase, we construct a captioning model fine-tuned on the synthesized data.

Through systematic evaluation, we show that the diverse data synthesis by LLMs allows our captioning model to generalize well to actual human trajectories and human-created captions. It also outperforms several existing LLM services (GPT-3.5 (OpenAI, 2023a), GPT-4 (OpenAI, 2023b)) as well as open-source benchmark Llama2 (GenAI, Meta, 2023) adapted to the task via in-context learning, in terms of ROUGE and BERT Score metrics.

Our contributions are summarized as follows: (1) formulating a new captioning task called contextual captioning of human movement trajectories; (2) proposing a learning-by-synthesis framework, TEXT2TRAJ2TEXT, and demonstrating its effectiveness on actual human data; (3) creating a benchmark dataset to accelerate future research.[1]

## 2 Contextual Captioning of Human Movement Trajectories

### 2.1 Motivating Scenario

The goal of our task is to generate concise text that describes possible underlying contexts of human movement trajectories, such as purposes and preferences. We focus particularly on a retail scenario, where people walk around a store, browse items of interest, and choose some to buy. Retailers analyze such shopping behaviors collected from consenting customers to gain deeper understanding of customers and improve customer experiences via demand prediction, inventory management, or targeted advertising. Much like web search engines automatically infer user preferences from click streams, we aim to automate customer activity profiling, ultimately across a wide range of stores beyond what is possible with a limited number of experts. Formatting profile results as sentences, as human experts do when communicating with stakeholders, is crucial for improving the interpretability of such automation.

### 2.2 Task Formulation

Given a movement trajectory $X$ and its semantics including *items in contact* $I$ and *purchased items* $\mathcal{P}$, we aim to generate a contextual caption $S$, as each detailed below.

**Input: Trajectory and its semantics.** The movement trajectory is a sequence of $T$ locations, *i.e.*, $X = (x_1, \ldots, x_T)$, where $x_t \in \mathbb{R}^2$ corresponds to a 2-D location where the customer stayed at each time step $t$. *Items in contact* are the items closest to the customer at each time step, *i.e.*, $I = (i_1, \ldots i_T)$. *Purchased items* are the items that the customer purchased, which form a subset of the items in contact, *i.e.*, $\mathcal{P} \subset I$. Technically, it is feasible to collect those data from consenting customers via wireless indoor localization technologies (Zafari et al., 2019) used in combination with point-of-sales (POS) systems. Nevertheless, such data collection is hard to scale in practice, as it is difficult to intervene in a retail store currently operating and obtain approval from each customer.

**Output: Contextual captions.** The *contextual caption* is a sequence of tokens, *i.e.*, $S = (s_1, s_2, \ldots)$, where $s$ is a token. We assume that each caption is concise, typically spanning a few sentences, and describes various aspects of the customer's shopping behavior such as their preferences for price versus quality, required quantity, and other characteristics related to item choices (*e.g.*, ready-to-eat, health-conscious).

## 3 TEXT2TRAJ2TEXT

Fig. 2 illustrates the overview of the proposed framework, TEXT2TRAJ2TEXT, which consists of TEXT2TRAJ data synthesis phase and TRAJ2TEXT model fine-tuning phase.

### 3.1 TEXT2TRAJ: Data Synthesis

In the TEXT2TRAJ phase, we propose leveraging pretrained, instruction-tuned LLMs in combination with a human trajectory planner to synthesize a diverse and realistic collection of annotated trajectory data. This approach is inspired by recent advancements in robotics research that aim to generate complex robot motion by incorporating LLMs

[1] Our code and dataset will be available at https://github.com/CyberAgentAILab/text2traj2text.

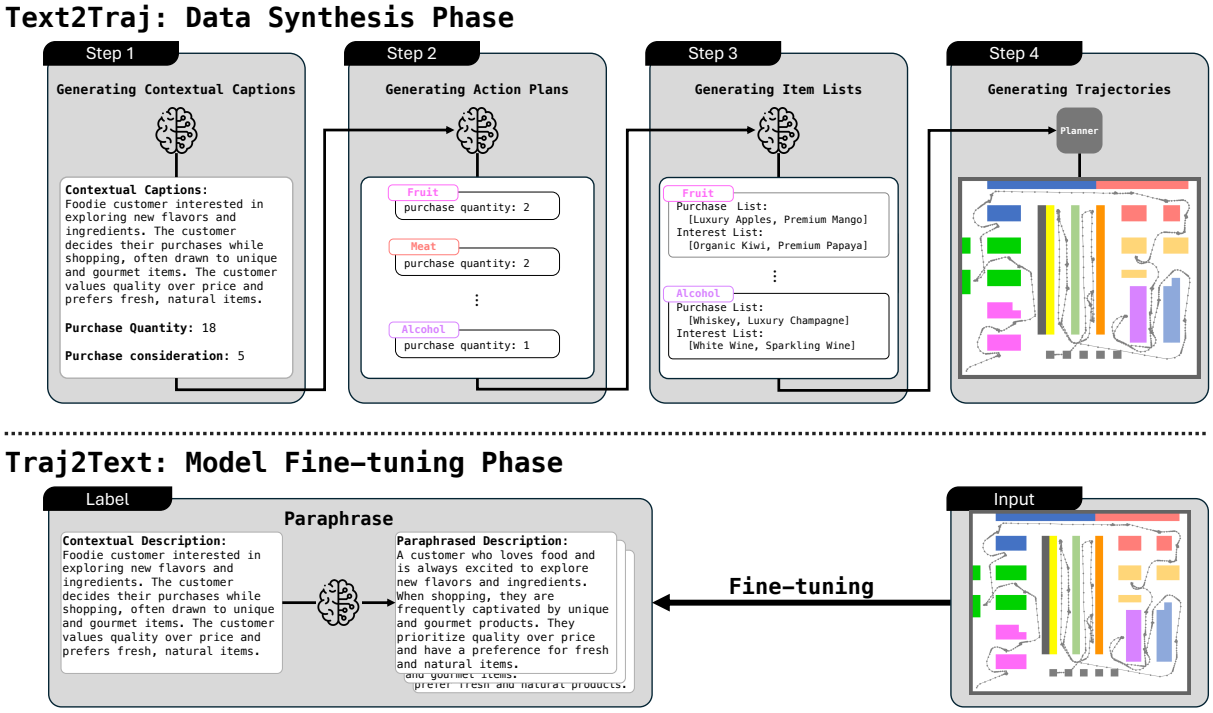



Figure 2: **Text2Traj2Text Framework**. (1) TEXT2TRAJ: We leverage LLMs to synthesize contextual captions and their instances as concrete action plans, item lists, and in-store trajectories. (2) TRAJ2TEXT: We fine-tune a language model with the synthesized data to be able to produce contextual captions from trajectory data.

into hierarchical motion planning frameworks (Ahn et al., 2022; Wang et al., 2024, 2023; Liu et al., 2023). It utilizes the reasoning ability of LLMs for task planning to determine which actions to take or which goals to approach, while employing classical motion planning to generate feasible motion trajectories for each action. Similarly, in our framework, an LLM first creates diverse contextual captions and instantiates coarse action plans from the captions. A trajectory planner then traces the plans to generate feasible movement trajectories on a store map. More specifically, the TEXT2TRAJ phase consists of four steps as shown below.

**Step 1: Generating contextual captions.** First, we give a prompt (Fig. 4 in Appendix A) to an LLM for producing contextual captions on three types of information: (i) individual customer's product preferences (*e.g.*, *"loves apple"*), (ii) category-level interests (*e.g.*, *"interested in fruits"*), and (iii) decision-making tendencies (*e.g.*, *"have a list of items to purchase"*). The LLM's output also includes the number of items planned to purchase (*i.e.*, purchase quantity) and the person's purchase consideration. Higher purchase consideration suggests more comparison of products before purchasing, while a lower one indicates a tendency to have pre-determined shopping plan.

**Step 2: Generating action plans.** Given a prompt (Fig. 5 in Appendix A) that contains the outputs from Step 1 (*i.e.*, a contextual caption and purchase quantity) and item categories in a store, the LLM generates an action plan, a list of pairs of item categories and their expected purchase quantity, *e.g.*, *{"fruit": 4, "meat": 0, "alcohol": 1}*.

**Step 3: Generating item lists.** Given a prompt (Fig. 6 in Appendix A), the LLM converts each item category determined in Step 2 into more specific item information, *i.e.*, (i) a *purchase list* consisting of the name of items planned to purchase, and (ii) an *interest list* of items that the individual is likely to show interest in. The interest item will contain more items as the purchase consideration is set higher. Also, the number of items in the purchase list may not always match the planned purchase quantity generated in the previous step, as the number of actual purchases can change based on other factors, such as the availability of suitable items in the store.

**Step 4: Generating movement trajectories.** Finally, based on the purchase and interest lists generated in Step 3, we invoke a trajectory planner to instantiate concrete human movement trajectories on a store map. We first assign ranks to each item

category stochastically for each trajectory generation, with the rank reflecting the category's relative position in the store layout. The rank tendency is predefined based on the store's layout, where categories located closer to the entrance typically receive a higher rank.

The purchase consideration is again considered here; if it is set high, ranks have higher variances, resulting in more exploratory behaviors. Starting from a fixed starting location $x_0 \in \Omega$ (*e.g.*, the entrance), the planner generates a feasible trajectory traversing items in the purchase and interest lists according to the category ranks in a store map like the one shown in Fig. 1.

### 3.2 TRAJ2TEXT: Model Fine-tuning

In the TEXT2TRAJ phase introduced so far, we first synthesize contextual captions and then instantiate concrete trajectories. Reversely, in the following TRAJ2TEXT phase, we aim to build a captioning model that takes the synthesized trajectory data as input to produce plausible captions.

**Input translation.** As the input to the captioning model, we translate movement trajectory $X = (x_1, \ldots, x_T)$, items in contact $I = (i_1, \ldots, i_T)$, and purchased items $\mathcal{P}$, into textual representations. Importantly, movement trajectories can become lengthy as customers take more time for shopping, and can also contain many mundane moments. Here, we adopt a simple yet effective filtering technique to focus on important events in the trajectories. First, we calculate the displacement between consecutive locations, *i.e.*, $\|x_t - x_{t-1}\|$, and extract moments when the individual stopped based on if the displacements are below a predetermined threshold. Then, items in contact at the stopping moments as well as those in the purchase list are simply concatenated: "`Trajectory is fruit</s>vegetable</s> ...\n Customer purchase item list is ['Carrots', 'Beef'...] \n Output:`."

**Data augmentation.** The diversity of training data is crucial for the high generalization capability of learned models. While synthesized trajectories can sufficiently be diversified based on randomized ranks of item categories (in Step 4 of Sec. 3.1), the variety of contextual captions may still be limited due to the expressiveness of the used LLM. To ensure high diversity for the captions, we introduce *data augmentation by paraphrasing*; for each annotated trajectory, we let the LLM to produce alternative expressions of the caption with similar meanings, and relabel the trajectory accordingly.

## 4 Experiments

We conducted systematic experiments to evaluate the effectiveness of the TEXT2TRAJ2TEXT framework. Through the experiments, we aim to answer the following questions:

**[RQ.1]** Can the models trained by our proposed framework generate appropriate captions for synthesized trajectories? (Sec. 4.2)

**[RQ.2]** Can the models generalize to human-created trajectories/captions? (Sec. 4.3)

**[RQ.3]** Can the models generalize to unseen maps? (Sec. 4.3)

### 4.1 Experimental Setup

**Data synthesis.** Following Sec. 3.1, we synthesized 80 pairs of contextual captions and the corresponding movement trajectories using GPT-4 (OpenAI, 2023b), while assuming a scenario of shopping at a supermarket. See Appendix A for concrete prompts and Tab. 6 for the store map we used. We adopted a classical hierarchical planning framework for trajectory generation; a global planner (probabilistic roadmaps proposed by Kavraki et al. (1996)) first determines a sequence of sub-goals from the current item to the next one, and a local planner (dynamic window approach proposed by Fox et al. (1997)) then produces a collision-free trajectory between the sub-goals. The synthesized data were divided into 64 training and 16 validation samples and augmented by paraphrasing with GPT-3.5 (OpenAI, 2023a), where the number of added captions from a single original caption was 2, 4 or 8. For example, in the case of adding 8 paraphrases, the total number of training samples becomes $64 \times 9$ (where 1 is the original caption and 8 is its paraphrased captions).[2]

**Implementation details.** On the synthesized data, we fine-tuned the T5-Base model (Raffel et al., 2020) available on HuggingFace[3], as its encoder-decoder structure was demonstrated effective for multimodal generation tasks (Xu et al., 2023). All fine-tuning was conducted on a single Tesla T4 GPU using AdamW optimizer with a

[2] Synthesizing captions is more complex than paraphrasing them, where we adopted GPT-4 for the former task and GPT-3.5 for the latter to consider cost-effectiveness.

[3] https://huggingface.co/t5-base

learning rate of $5.6 \times 10^{-5}$, where the batch size and the number of epochs were set to 8 and 5, respectively. The model checkpoint with the BERT Precision score (Zhang* et al., 2020) highest for the validation data was used for evaluation.

**Baseline models.** We compared our captioning model against the following baselines: (a) T5-Small and T5-Base (Raffel et al., 2020) fine-tuned without paraphrasing-based data augmentation; (2) GPT-3.5 (OpenAI, 2023a), GPT-4 (OpenAI, 2023b), and the open-source benchmark Llama-2-7b-chat-hf (referred to as Llama2) (GenAI, Meta, 2023)[4]. GPT-3.5, GPT-4, and Llama2 were used via in-context learning; following (Maynez et al., 2023), a few (1, 2, or 4) samples randomly selected from the training data were given as examples, and contextual captions were generated for the given movement trajectory.

**Evaluation metrics.** We employed ROUGE (R-1, R-2, R-L) (Lin, 2004) and BERT Score (BS-precision, recall, f1 score) (Zhang* et al., 2020) as evaluation metrics. ROUGE score captures lexical overlap by comparing n-grams and word sequences between generated and reference texts, while BERT Score, which utilizes BERT embeddings, measures semantic similarity.

### 4.2 Evaluation with Synthesized Trajectories

**[RQ.1] Can the models trained by our proposed framework generate appropriate captions for synthesized trajectories?** Tab. 1 presents the quantitative results on 20 synthesized trajectories created in the same way as training/validation data. Overall, our model achieved the best performance even with an order-of-magnitude fewer parameters (223M) compared to the GPT family and Llama2 (over billions). We observe a monotonic improvement in nearly all metrics as the number of paraphrases increases, indicating the effectiveness of our data augmentation strategy. In contrast, T5-Small and T5-Base with vanilla fine-tuning demonstrated quite limited performances. The number of examples presented to GPT-3.5, GPT-4, and Llama2 was critical for their in-context learning ability, but this comes with increased inference costs and limits practical scalability.

**Ablation study.** Additionally, we investigate how each of the movement trajectories (with the list of nearby items) and the purchased items can contribute to the final performances using the validation dataset. In Tab. 2, we evaluated the following degraded variants: **w/o Traj** (resp. **w/o Item**) that removed trajectories (resp. purchased items) from the input; **w/ Shuffle Traj** (resp. **w/ Shuffle Item**) that replaced trajectories (resp. items) with those of other samples dataset according to the permutation feature importance method (Breiman, 2001; Fisher et al., 2019). These degraded versions all demonstrated quite limited performances, indicating the necessity of combining trajectories and purchased items for inferring contexts reliably. We also evaluate a more challenging case when the trajectory data are partially perturbed, possibly due to the inaccuracy of indoor localization systems. Our model is robust to such noises, as shown in the table (**w/ 5% noise**).

| Models | R-1 | R-2 | R-L | BS-p | BS-r | BS-f1 |
|---|---|---|---|---|---|---|
| T5-Small | 0.069 | 0.015 | 0.060 | 0.792 | 0.770 | 0.816 |
| T5-Base | 0.287 | 0.094 | 0.243 | 0.860 | 0.859 | 0.861 |
| GPT-3.5 | 0.240 | 0.049 | 0.151 | 0.854 | 0.841 | 0.868 |
| + 1 examples | 0.326 | 0.080 | 0.211 | 0.887 | 0.883 | 0.891 |
| + 2 examples | 0.358 | 0.093 | 0.225 | 0.892 | 0.888 | 0.895 |
| + 4 examples | 0.364 | 0.101 | 0.235 | 0.894 | 0.890 | 0.897 |
| GPT-4 | 0.180 | 0.034 | 0.119 | 0.844 | 0.822 | 0.868 |
| + 1 examples | 0.322 | 0.064 | 0.192 | 0.881 | 0.873 | 0.890 |
| + 2 examples | 0.334 | 0.070 | 0.199 | 0.887 | 0.881 | 0.894 |
| + 4 examples | 0.378 | 0.106 | 0.240 | 0.897 | 0.892 | 0.902 |
| Llama2 | 0.199 | 0.020 | 0.129 | 0.819 | 0.788 | 0.854 |
| + 1 examples | 0.255 | 0.070 | 0.167 | 0.834 | 0.790 | 0.885 |
| + 2 examples | 0.305 | 0.089 | 0.198 | 0.855 | 0.824 | 0.889 |
| + 4 examples | 0.391 | 0.128 | 0.267 | 0.886 | 0.877 | 0.897 |
| Ours | | | | | | |
| 2 paraphrases | 0.374 | **0.140** | **0.297** | 0.888 | 0.894 | 0.882 |
| 4 paraphrases | 0.368 | 0.131 | 0.287 | 0.888 | 0.893 | 0.884 |
| 8 paraphrases | **0.412** | 0.138 | **0.297** | **0.907** | **0.910** | **0.905** |

Table 1: Quantitative results for synthesized data.

| Models | R-1 | R-2 | R-L | BS-p | BS-r | BS-f1 |
|---|---|---|---|---|---|---|
| w/o Traj | 0.337 | 0.101 | 0.234 | 0.877 | 0.874 | 0.880 |
| w/o Item | 0.218 | 0.038 | 0.166 | 0.862 | 0.876 | 0.849 |
| w/ Shuffle Traj | 0.395 | 0.130 | 0.277 | 0.901 | 0.904 | 0.899 |
| w/ Shuffle Item | 0.382 | 0.116 | 0.269 | 0.899 | 0.902 | 0.897 |
| w/ 5% noise | 0.428 | 0.159 | 0.308 | 0.907 | 0.911 | 0.903 |
| Ours | 0.427 | 0.156 | 0.308 | 0.907 | 0.911 | 0.903 |

Table 2: Ablation study and noisy robustness evaluation

### 4.3 Evaluation with Real Human Data

**Data collection from human subjects.** We recruited eight participants to collect real human data for our study. The entire experiment consisted of two phases with different tasks. In the first phase, two participants were instructed to create four plau-

[4] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

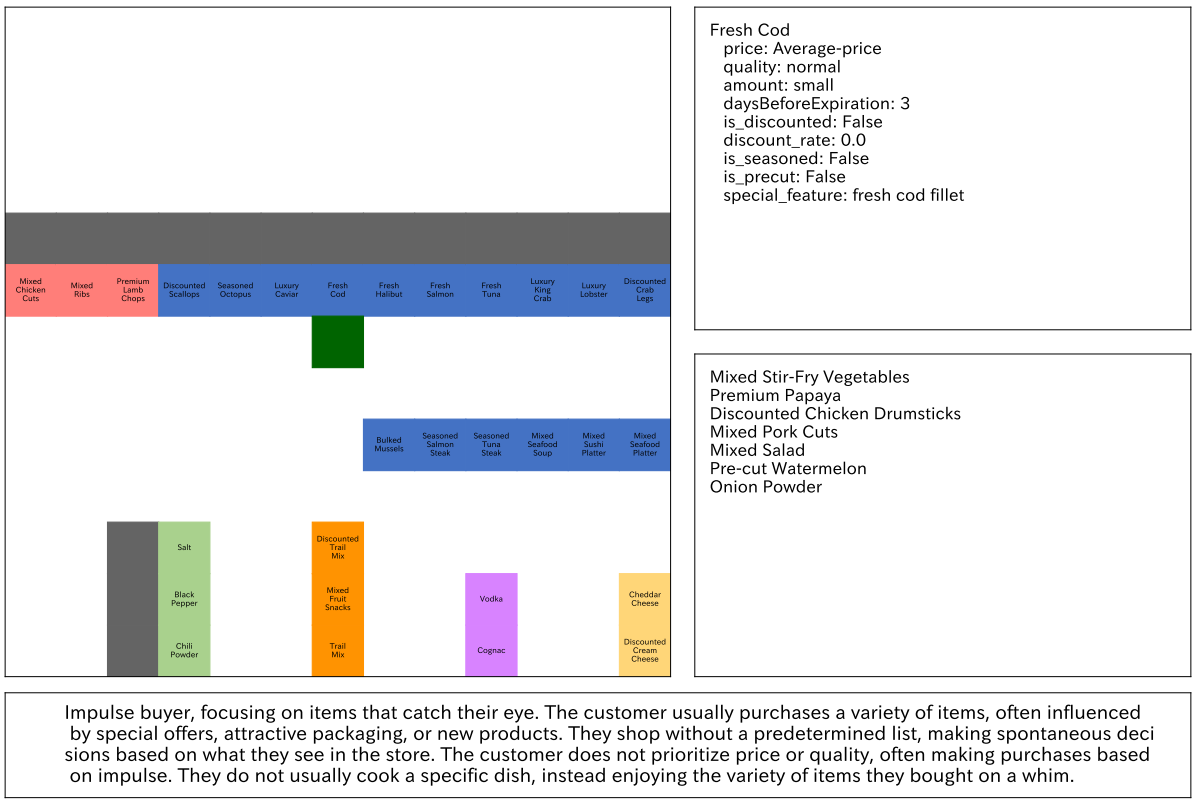



Figure 3: Visual user interface used to collect human-created trajectories. The green square represents the current position. Information on the closest item is shown in the upper right corner, and the list of items added to the cart is shown in the lower right corner. The caption to be followed is presented at the bottom of the screen.

sible contextual captions about supermarket shoppers. Before they began, we provided them with three example captions to ensure appropriateness for our task. In the second phase, six participants were asked to create trajectories using a visual interface (Fig. 3) based on 10 randomly selected captions — half synthetic and half created by the participants in the first phase. On the visual interface, the current position of a participant was marked by a green rectangle, with details about the item adjacent to their current location shown in the top right corner and items currently added to their cart displayed in the bottom right. Participants were allowed to navigate in the store map and add or remove adjacent items from their cart using keyboard input. Each session began from a fixed location and ended when participant reached the cashier register, tracking whole movement trajectories and final purchases.

Two distinct store maps were adopted in the experiment to validate the generalization ability of trained models: one used for training data and another as a completely new environment. Participants first completed two pilot rounds on one map to familiarize themselves with the interface and layout, followed by five main rounds on this map for data collection. They then repeated the same process on the other map. The set and order of captions, as well as store maps, were randomized across participants. Each experiment lasted about one hour. Through this experimental procedure, we collected 60 sufficiently diverse trajectory data points from real humans, as summarized in Tab. 3.

| | | Captions | | |
|---|---|---|---|---|
| | | Synthesized | Human-Created | Total |
| Map | Seen | 15 | 15 | 30 |
| | Unseen | 15 | 15 | 30 |
| | Total | 30 | 30 | 60 |

Table 3: Statistics on human-created trajectory data. Participants produced trajectory data with a carefully controlled set of synthesized/human-created captions and seen/unseen maps.

**[RQ.2] Can the model generalize to human-created trajectories/captions?** Tab. 4 shows the quantitative results for human trajectories data, compared between when ground-truth captions are

| | Synthesized Captions | | | | | | Human-created Captions | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Models | R-1 | R-2 | R-L | BS-p | BS-r | BS-f1 | R-1 | R-2 | R-L | BS-p | BS-r | BS-f1 |
| T5-Small | 0.080 | 0.020 | 0.066 | 0.529 | 0.520 | 0.539 | 0.055 | 0.006 | 0.047 | 0.602 | 0.593 | 0.613 |
| T5-Base | 0.303 | 0.101 | 0.259 | 0.866 | 0.875 | 0.858 | 0.136 | 0.006 | 0.122 | 0.838 | 0.837 | 0.839 |
| GPT-3.5 | 0.383 | 0.105 | 0.246 | 0.898 | 0.898 | 0.899 | 0.291 | **0.041** | 0.189 | 0.877 | 0.880 | 0.875 |
| GPT-4 | 0.376 | 0.097 | 0.234 | 0.897 | 0.894 | 0.900 | **0.309** | 0.037 | 0.188 | 0.878 | 0.877 | **0.879** |
| Llama2 | 0.389 | 0.137 | 0.272 | 0.886 | 0.876 | 0.898 | 0.254 | 0.032 | 0.163 | 0.861 | 0.855 | 0.868 |
| Ours w/ 8 paraphrase | **0.436** | **0.163** | **0.329** | **0.914** | **0.920** | **0.907** | 0.306 | **0.041** | **0.205** | **0.883** | **0.890** | 0.876 |

Table 4: Performance comparisons between synthesized and human-created captions on real human trajectories.

| | Seen Store Map | | | | | | Unseen Store Map | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Models | R-1 | R-2 | R-L | BS-p | BS-r | BS-f1 | R-1 | R-2 | R-L | BS-p | BS-r | BS-f1 |
| T5-Small | 0.054 | 0.008 | 0.047 | 0.537 | 0.528 | 0.547 | 0.081 | 0.018 | 0.065 | 0.594 | 0.584 | 0.605 |
| T5-Base | 0.220 | 0.055 | 0.192 | 0.851 | 0.855 | 0.848 | 0.219 | 0.052 | 0.189 | 0.852 | 0.856 | 0.849 |
| GPT-3.5 | 0.344 | 0.079 | 0.224 | 0.888 | 0.890 | 0.887 | 0.329 | 0.067 | 0.210 | 0.887 | 0.887 | 0.887 |
| GPT-4 | 0.346 | 0.070 | 0.215 | 0.889 | 0.887 | 0.890 | 0.339 | 0.064 | 0.207 | 0.886 | 0.884 | 0.888 |
| Llama2 | 0.330 | 0.091 | 0.225 | 0.875 | 0.869 | 0.883 | 0.314 | 0.077 | 0.211 | 0.872 | 0.862 | 0.883 |
| Ours w/ 8 paraphrase | **0.379** | **0.109** | **0.273** | **0.900** | **0.907** | **0.893** | **0.364** | **0.095** | **0.260** | **0.897** | **0.904** | **0.890** |

Table 5: Performance comparisons between seen and unseen store maps on real human trajectories.

synthesized or created by human participants. Here we evaluated T5-Small and T5-Base, GPT-3.5/GPT-4/Llama2 each with 4 examples for in-context learning, and our captioning model with 8 paraphrases based on the previous result. Overall, our captioning model generalized well to those human-created data, with acceptably slight degradation of performances. Again, our model demonstrates comparable performance to GPT-3.5/4 and Llama2 despite its much smaller number of parameters. It is inevitably difficult to match generated captions with human-created ground truths exactly at word/phrase levels, as indicated by degraded ROUGE scores. Nevertheless, the semantic consistency measured by BERT Scores remains as high as that for synthesized captions, indicating the practical usability.

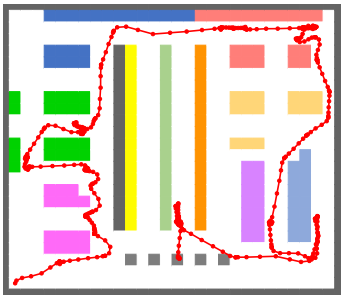
**[RQ.3] Can the model generalize to unseen maps?** Tab. 5 compares the performance between when store maps are seen (*i.e.*, identical to those for training data) and unseen. For all models, we confirmed negligible performance degradation. This is practically beneficial, as major retailers often operate multiple stores that can have different layouts and item availability, where captioning systems should be easy to deploy.

### 4.4 Qualitative Results and Failure Cases

Tab. 6 illustrates some qualitative results of success and failure cases. If successful, our captioning model provides an accurate background context such as *"have a list of items to buy"* (success cases 1 and 2). Based on additional information attached to items such as price and quantity, it is also possible to predict customer's preference, *e.g.*, *"prioritizes quality over price and prefers fresh and unprocessed items"* (success case 1) and *"customer who is environmentally conscious"* (success case 2, against ground-truth label: *"environmentally conscious customer"*).

Human-created captions can sometimes include demographic information of target individuals, such as *"newly married couple customer,"* which are difficult to predict. It is also hard to generate sentences like *"If the main focus is fish, they buy more vegetables, and if the main focus is meat, they buy wine and other items together."* (the ground-truth label in the failure case). Still, our model appropriately infers the customer's preference, *e.g.*, *"willing to pay a premium for quality"* and *"customer who is passionate about food"*. Additionally, unlike success cases 1 and 2, our model does not mention that the customer has the predetermined item list. This is consistent with the redundant trajectory of the failure case, suggesting that our

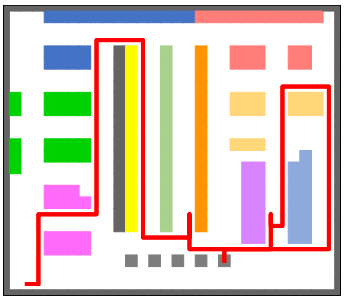
**Success Case 1 (Synthesized trajectories and captions)**



Mixed Pork Cuts, Seasoned Turkey, Mixed Sausages, Seasoned Lamb Chops, Cucumbers, Potatoes, Onions, Onion Powder, Paprika, Premium Mango, Organic Kiwi, Premium Papaya, High-end Matcha Latte

| | |
|---|---|
| Label | Customer is shopping for plant-based items. The customer has a predetermined list of vegan products like fruits, vegetables, grains, and dairy alternatives. The customer values quality over price and prefers organic and fresh items. |
| GPT-4 | The customer is a health-conscious shopper who prefers organic and high-quality products. They usually shop for fresh fruits and vegetables, and they also enjoy premium snacks like mixed nuts. They prefer to season their food themselves, often purchasing basic seasonings like salt and pepper. They also enjoy high-end green tea, indicating a preference for natural and healthy beverages. This customer values quality over quantity and is willing to pay a premium for organic and luxury items. |
| Ours | A customer who is conscious of their health and enjoys eating fresh and unprocessed items. They have a predetermined list of purchases, which includes fresh fruits, vegetables, and grains. The customer prioritizes quality over price and prefers fresh and unprocessed items. |

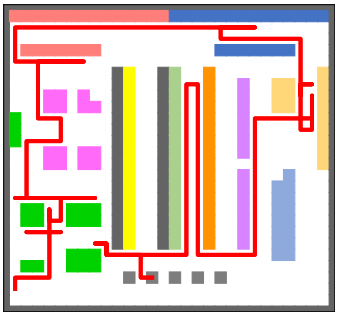
**Success Case 2 (Human-created trajectories and synthesized captions)**



Organic Kiwi, Laundry Detergent, High-end Green Tea, Luxury Butter

| | |
|---|---|
| Label | An environmentally conscious customer who prioritizes sustainable and eco-friendly products. The customer usually comes with a predetermined list of purchases, which includes items like organic produce, fair-trade coffee, and eco-friendly cleaning products. The customer is willing to pay more for products that align with their values and prefer unprocessed goods. |
| GPT-4 | A health-conscious and high-end consumer who values quality over quantity. The customer usually purchases organic fruits, high-end beverages, and luxury dairy products. The customer also buys household goods, indicating a preference for maintaining a clean and organized home. The customer's shopping habits suggest a preference for luxury items and a willingness to pay more for high-quality products. |
| Ours | A customer who is environmentally conscious and values eco-friendly products. They have a predetermined list of purchases, focusing on eco-friendly products and eco-friendly alternatives. Quality is their primary concern, and they prefer eco-friendly products. |

**Failure Case (Human-created trajectories and captions)**



Luxury Broccoli, Garlic, Premium Papaya, Premium Steak, Mixed Seafood Soup, Mozzarella Cheese, Greek Yogurt, Sparkling Wine, Red Wine, Mixed Fruit Snacks, Garlic Powder, Italian Seasoning, Mixed Salad

| | |
|---|---|
| Label | Newly married couple customer. They both want to enjoy cooking, so they look at and buy from a wide range of genres. They decide to a certain extent which days are fish-centered, meat-centered, and so on. If the main focus is fish, they buy more vegetables, and if the main focus is meat, they buy wine and other items together. Since they both work, they do not compromise on price. |
| GPT-4 | A gourmet food enthusiast who enjoys cooking and experimenting with different recipes. They often shop for a variety of fresh vegetables, premium meats, and seafood. They also have a preference for luxury fruits and dairy products. They enjoy pairing their meals with a selection of wines and often indulge in snacks. They usually shop with a specific recipe in mind, often inspired by international cuisines. |
| Ours | A customer who is passionate about food and is willing to pay a premium for fresh produce, meats, and dairy products. Their shopping list consists of a mixture of fresh produce, meats, and dairy products. They are willing to pay a premium for fresh produce and are willing to pay a premium for quality. |

Table 6: Qualitative comparisons of ground-truth and generated captions. The movement trajectory is shown as a solid red line, with the purchase list displayed below. Colored rectangles represent shelves for different categories of items.

model correctly inferred the customer's decision-making tendencies.

### 4.5 Limitations and Practical Implications

Our approach has a few limitations. As we obtain a captioning model by fine-tuning pretrained language models, its text generation capability would inevitably rely on that of the base model. Namely, our model cannot handle extremely long shopping activities beyond the maximum input token length for the base model. Moreover, there is no guarantee that the model won't hallucinate contexts that are totally irrelevant to a target individual. In practical system setup, it is crucial to post-process model outputs, for example, based on heuristic rules or

manual inspection, so as not to present inappropriate captions to users. Recent work that seeks to mitigate hallucination (Mündler et al., 2024) would also help. Finally, similar to web search engines, it is necessary to allow for an opt-out option on the customer side for the use of inferred contextual captions in practical applications.

## 5 Related Work

**Human movement analysis.** Studies on human movements can be found in various research contexts, such as urban engineering (Pappalardo et al., 2016; Askarizad and Safari, 2020), traffic simulation (Doniec et al., 2008; Duives et al., 2013), autonomous driving (Camara et al., 2021), tourism (Li et al., 2018; Payntar et al., 2021), and public health (Kraemer et al., 2020). Concrete techniques include pattern mining (Lam et al., 2017; Ghose et al., 2019), semantic mining (Parent et al., 2013), trajectory prediction (Rudenko et al., 2020), and crowd analysis (Zhou et al., 2020). Compared to these prior arts, our work is the first to explore the potential of recent progress in large language modeling to empower human movement analysis and its application to retail scenarios.

**Human activity captioning.** Captioning human activities has been addressed mainly in computer vision, as a part of image captioning (Hossain et al., 2019) and video captioning (Aafaq et al., 2019). Continuous efforts have been made to develop large-scale multimodal datasets that involve human activity data and their captions (Krishna et al., 2017; Grauman et al., 2023). Nevertheless, much recent work seeks to exploit rich representations of human activities in visual data, which is not applicable to our task where only location trajectories and limited semantic information are available.

**Generative models as data generators.** Finally, there is a growing trend to utilize generative models to construct synthetic datasets. For example, generative adversarial networks and diffusion models have been used in computer vision to create or augment visual training data (Karras et al., 2019; Nichol et al., 2022). LLMs have been used more widely for dataset generation, such as generating annotations (Feng et al., 2021; Zhang et al., 2023; Flamholz et al., 2024; Sainz and Rigau, 2021), ranking (Hou et al., 2024; Qin et al., 2024; Sun et al., 2023), and textual datasets (Chen et al., 2023; Chung et al., 2023). Some recent work uses LLMs as virtual agents that produce realistic behaviors in simulated worlds (Park et al., 2023; Kaiya et al., 2023). Our data synthesis framework is unique in terms of integrating LLMs and trajectory planners to produce diverse captioned human trajectories.

## 6 Conclusion

We presented a new task named contextual captioning of human movement trajectories, and a dedicated learning-by-synthesis framework, *i.e.*, TEXT2TRAJ2TEXT, with a particular focus on retail scenarios. We leverage LLMs to synthesize realistic and diverse collection of contextual captions as well as concrete trajectories on store maps. Our captioning model fine-tuned on these synthesized data demonstrated equal or even better performance than existing LLMs with a higher number of parameters. Moreover, the model well generalizes to human-created trajectories and captions.

Although this work focused exclusively on retail scenarios, we believe that the proposed task and framework would open up a new opportunity for adopting neural language generation techniques to various applications that need automated human activity understanding. This also raises new technical challenges such as effective encoding of very long trajectory data as input to language models and efficient inference of learned models to enable online captioning.

## Acknowledgments

We would like to express our gratitude for the anonymous reviewers who provided many insightful comments that have improved our paper. Special thanks also go to the members of CyberAgent, Inc. AI Lab for the interesting comments and energetic discussions.

## References


Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. 2019. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)*, 52(6):1–37.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. 2022. Do as I can, not as I say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

Reza Askarizad and Hossein Safari. 2020. The influence of social interactions on the behavioral patterns of the people in urban spaces (case study: The pedestrian zone of rasht municipality square, iran). *Cities*, 101:102687.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Fanta Camara, Nicola Bellotto, Serhan Cosar, Florian Weber, Dimitris Nathanael, Matthias Althoff, Jingyuan Wu, Johannes Ruenz, André Dietrich, Gustav Markkula, Anna Schieben, Fabio Tango, Natasha Merat, and Charles Fox. 2021. Pedestrian models for autonomous driving part II: High-Level models of human behavior. *IEEE Transactions on Intelligent Transportation Systems*, 22(9):5453–5472.

Anna Carreras, Marc Morenza-Cinos, Rafael Pous, Joan Melià-Seguí, Kamruddin Nur, Joan Oliver, and Ramir De Porrata-Doria. 2013. Store view: pervasive rfid & indoor navigation based retail inventory management. In *Proceedings of the ACM conference on Pervasive and Ubiquitous Computing Adjunct Publication (UbiComp Adjunct)*, pages 1037–1042.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. PLACES: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL*, pages 844–868.

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 575–593.

Arnaud Doniec, René Mandiau, Sylvain Piechowiak, and Stéphane Espié. 2008. A behavioral multi-agent model for road traffic simulation. *Engineering Applications of Artificial Intelligence*, 21(8):1443–1454.

Dorine C Duives, Winnie Daamen, and Serge P Hoogendoorn. 2013. State-of-the-art crowd motion simulation models. *Transportation Research Part C: Emerging Technologies*, 37:193–209.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language model as an annotator: Exploring DialoGPT for dialogue summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1479–1491.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of machine learning research: JMLR*, 20(177):1–81.

Zachary N Flamholz, Steven J Biller, and Libusha Kelly. 2024. Large language models improve annotation of prokaryotic viral proteins. *Nature Microbiology*, 9(2):537–549.

Dieter Fox, Wolfram Burgard, and Sebastian Thrun. 1997. The dynamic window approach to collision avoidance. *IEEE Robotics & Automation Magazine*, 4(1):23–33.

GenAI, Meta. 2023. Llama 2: Open foundation and Fine-Tuned chat models.

Anindya Ghose, Beibei Li, and Siyuan Liu. 2019. Mobile targeting using customer trajectory patterns. *Management Science*, 65(11):5027–5049.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. 2023. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):1–36.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are Zero-Shot rankers for recommender systems. In *Advances in Information Retrieval*, pages 364–381.

Zhao Kaiya, Michelangelo Naim, Jovana Kondic, Manuel Cortes, Jiaxin Ge, Shuying Luo, Guangyu Robert Yang, and Andrew Ahn. 2023. Lyfe agents: Generative agents for low-cost real-time social interactions. *arXiv preprint arXiv:2310.02172*.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lydia E Kavraki, Petr Svestka, J-C Latombe, and Mark H Overmars. 1996. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Transactions on Robotics (T-RO)*, 12(4):566–580.

Moritz U G Kraemer, Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M Pigott, Open COVID-19 Data Working Group, Louis du Plessis, Nuno R Faria, Ruoran Li, William P Hanage, John S Brownstein, Maylis Layan, Alessandro Vespignani, Huaiyu Tian, Christopher Dye, Oliver G Pybus, and Samuel V Scarpino. 2020. The effect of human mobility and control measures on the COVID-19 epidemic in china. *Science*, 368(6490):493–497.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceeding of the International Conference on Computer Vision (ICCV)*, pages 706–715.

Luan D M Lam, Antony Tang, and John Grundy. 2017. Predicting indoor spatial movement using data mining and movement patterns. In *Proceedings of the IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 223–230. IEEE.

Jingjing Li, Lizhi Xu, Ling Tang, Shouyang Wang, and Ling Li. 2018. Big data in tourism research: A literature review. *Tourism Management*, 68:301–323.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. 2023. LLM+P: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*.

Xiaochen Liu, Yurong Jiang, Puneet Jain, and Kyu-Han Kim. 2018. Tar: Enabling fine-grained targeted advertising in retail stores. In *Proceedings of the ACM Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*, pages 323–336. Association for Computing Machinery.

Joshua Maynez, Priyanka Agrawal, and Sebastian Gehrmann. 2023. Benchmarking large language model capabilities for conditional generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9194–9213.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2024. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *Proceedings of the International Conference on Learning and Representation (ICLR)*.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 16784–16804.

OpenAI. 2023a. ChatGPT General FAQ. https://help.openai.com/en/articles/6783457-chatgpt-general-faq. Accessed: March 3, 2023.

OpenAI. 2023b. GPT-4 technical report. *ArXiv e-prints (arXiv:2303.08774)*.

Luca Pappalardo, Maarten Vanhoof, Lorenzo Gabrielli, Zbigniew Smoreda, Dino Pedreschi, and Fosca Giannotti. 2016. An analytical framework to nowcast well-being using mobile phone data. *International Journal of Data Science and Analytics*, 2(1):75–92.

Christine Parent, Stefano Spaccapietra, Chiara Renso, Gennady Andrienko, Natalia Andrienko, Vania Bogorny, Maria Luisa Damiani, Aris Gkoulalas-Divanis, Jose Macedo, Nikos Pelekis, Yannis Theodoridis, and Zhixian Yan. 2013. Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)*, 45(4):1–32.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *In Proceedings of the Annual ACM Symposium on User Interface Software and Technology (UIST)*.

Nicole D Payntar, Wei-Lin Hsiao, R Alan Covey, and Kristen Grauman. 2021. Learning patterns of tourist movement and photography from geotagged photos at archaeological heritage sites in cuzco, peru. *Tourism Management*, 82:104165.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research (JMLR)*, 21(1).

Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. 2020. Human motion trajectory prediction: a survey. *International Journal of Robotics Research (IJRR)*, 39(8):895–935.

Oscar Sainz and German Rigau. 2021. Ask2Transformers: Zero-shot domain labelling with pretrained language models. In *Proceedings of the 11th Global Wordnet Conference*.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as Re-Ranking agents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14918–14937.

Lirui Wang, Yiyang Ling, Zhecheng Yuan, Mohit Shridhar, Chen Bao, Yuzhe Qin, Bailin Wang, Huazhe Xu, and Xiaolong Wang. 2024. Gensim: Generating robotic simulation tasks via large language models. In *Proceedings of the International Conference on Learning and Representation (ICLR)*.

Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023. Describe, explain, plan and select: Interactive planning with LLMs enables open-world multi-task agents. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Peng Xu, Xiatian Zhu, and David A Clifton. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(10):12113–12132.

Faheem Zafari, Athanasios Gkelias, and Kin K Leung. 2019. A survey of indoor localization systems and technologies. *IEEE Communications Surveys & Tutorials*, 21(3):2568–2599.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. LLMaAA: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 13088–13103. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with bert. In *Proceedings of the International Conference on Learning and Representation (ICLR)*.

Yuren Zhou, Billy Pik Lik Lau, Zann Koh, Chau Yuen, and Benny Kai Kiat Ng. 2020. Understanding crowd behaviors in a social event by passive wifi sensing and data mining. *IEEE Internet of Things Journal*, 7(5):4442–4454.

# A Prompt for Data Synthesis

**STEP 1: Instruction for generating each contextual caption $S$**

System: Your task is to generate descriptions of various customer intentions within a supermarket environment, elucidating their purchasing preferences and habits meticulously.

Human: Kindly generate {samples} unique descriptions of customer intentions, ensuring each one is varied, embodying a range of customer profiles and shopping objectives. Every description should be comprehensively structured to include the following components:

- Outline the overarching characteristics defining the customer's shopping intention.
- Identify the categories of products the customer is likely to purchase or abstain from, such as a preference for meat over seafood, or vegetables over fruits.
- Clarify whether the customer arrives with a predetermined list of purchases or if they are likely to explore and decide while shopping.
- Elaborate on the customer's family structure,such as being a single individual, a couple, or part of a larger family, and how this influences their purchasing decisions.
- Highlight customer's preferences regarding the price and quality of products, specifying if they lean towards high-end items, discounted quality goods, or more affordable, lower-quality products.
- Describe the customer's preferences concerning the state of the products, such as pre-cut, seasoned, etc.
- If there is a dish the customer would like to cook, describe it. If not, please state that you do not.
- It is imperative to maintain strong consistency between the customer's "intention" and "num_item_to_buy". For example, a family of five might buy a lot of items at once. These customers usually buy in bulk, getting many products in one visit. On the other hand, some customers come to the supermarket often, but they only buy a few things each time.
- Ensuring a close alignment between a customer's "intent" and their 'purchase_consideration' is crucial. For instance, customers who are uncertain about their purchase choice or who explore various options typically exhibit a higher level of "purchase_consideration". In contrast, customers who have a pre-determined purchase decision before visiting the store usually show lower "purchase_consideration".

Rule:
Ensure all responses maintain the prescribed format and diversity in customer intentions is robustly represented! You must persist in generating sentences without cessation until you have produced at least {samples} intentions in total!!!

Example:

Figure 4: Prompt used for Step 1 in the Text2Traj phase.

STEP 2: Instruction for generating an abstract action plan consistent with each contextual caption generated in STEP 1.

System: As an adept AI, your task is to create a shopping plan for a customer, using their stated intentions, the total number of items they intend to purchase, and a provided list of product categories.

Human: Your role is to allocate the total number of items the customer plans to purchase across the given product categories. This allocation should form a cohesive plan that aligns with the customer's intentions and preferences.

Rule: Ensure all responses maintain the prescribed format! The total number of items in the shopping plan should be approximately {num_items}. The distribution of products across categories must closely align with the customer's intention.

# Customer's intention {intention}

# category List {category_list}
{format_instructions}

Figure 5: Prompt used for Step 2 in the Text2Traj phase.

STEP 3: Instruction for generating item lists.

System: As a proficient AI assistant, your task is to curate two lists of products that align with the customer's intentions. You have access to detailed information, including the customer's intentions, product descriptions, the quantities they plan to purchase, and their level of purchase consideration.

Human: Your goal is to create two lists based on the provided information: 1. "inclined_to_purchase": Products that the customer is highly likely to purchase. 2. "show_interest": Products the customer might consider purchasing or show interest in, taking into account both the customer's intentions and their "purchase_consideration" score.

Guidelines:

- Purchases are planned only for products in the {category} category.
- Ensure that the total number of products in the "inclined_to_purchase" list for the {category} category is approximately {num_purchase_items}.
- Ensure that the total number of products in the "show_interest" list for the {category} category is less than {num_purchase_items}.
- Align the "inclined_to_purchase" items in the {category} category with the customer's intentions.
- Generate the "show_interest" list by carefully considering both the customer's intentions and their "purchase_consideration" score, which ranges from 1 to 5. If the purchase_consideration score is low, focus on a smaller "show_interest" list. Conversely, if the score is high, the "show_interest" list can be more extensive but should remain below {num_purchase_items} in total.

Tips:

- Pay close attention to the item descriptions and customer intentions provided.

### Customers intention {intention}
### "purchase_consideration" (1-5) {purchase_consideration}
### Item description {item_description}
{format_instructions}

Figure 6: Prompt used for Step 3 in the Text2Traj phase.