User Consumption Intention Prediction in Meituan

Yukun Ping[1,2], Chen Gao[1], Taichi Liu[1,2], Xiaoyi Du[2], Hengliang Luo[2], Depeng Jin[1], Yong Li[1]
[1]Beijing National Research Center for Information Science and Technology (BNRist)
Department of Electronic Engineering, Tsinghua University, Beijing, China
[2]Meituan Inc., Beijing, China
liyong07@tsinghua.edu.cn

ABSTRACT

For online life service platforms, such as Meituan, user consumption intention, as the internal driving force of consumption behaviors, plays a significant role in understanding and predicting users' demand and purchase. However, user consumption intention prediction is quite challenging. Different from consumption behaviors, consumption intention is implicit and always not reflected by behavioral data. Moreover, it is affected by both user intrinsic preference and spatio-temporal context. To overcome these challenges, in Meituan, we design a real-world system consisting of two stages, intention detection and prediction. Specifically, at the intention-detection stage, we combine the knowledge of human experts and consumption information to obtain explicit intentions and match consumption with intentions based on user review data. At the intention-prediction stage, to collectively exploit the rich heterogeneous influencing factors, we design a graph neural network-based intention prediction model GRIP, which can capture user intrinsic preference and spatio-temporal context. Extensive offline evaluations demonstrate that our prediction model outperforms the best baseline by 10.26% and 33.28% for two metrics and online A/B tests on millions of users validate the effectiveness of our system.

CCS CONCEPTS

• **Information systems** → **Information systems applications**.

KEYWORDS

Consumption Intention Prediction, Graph Neural Networks

ACM Reference Format:
Yukun Ping[1,2], Chen Gao[1], Taichi Liu[1,2], Xiaoyi Du[2], Hengliang Luo[2], Depeng Jin[1], Yong Li[1]. 2021. User Consumption Intention Prediction in Meituan. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3447548.3467178

1 INTRODUCTION

Meituan, similar as Yelp, is one of the largest consumption portal-platforms around the world connecting more than 100 million of daily consumers and various local services, including food, entertainment, shopping, etc. Usersfi nd and browse local services in Meituan before consumption, and complete digital payment in Meituan after consumption. The platform collects users' consumption behaviors and provides personalized experiences, including personalized recommendation, itinerary planning, etc., supported by understanding and predicting of users' consumption behaviors. The current solutions for consumption prediction is to directly learn from users' historical behaviors, such as collaborativefiltering [11, 17, 25] based on the similarity of historical behaviors.

However, directly and merely learning from consumption behaviors is suffering from two limitations at two perspectives. First, the consumption behavior on a specific service may happen by coincidence. In other words, there are many replaceable services and the consumed one may be chosen accidentally. On a deeper level, the consumption behavior is internally driven by the user's demand. For example, a user wants to *have a rest* at a coffee bar and coincidentally choose Starbucks, although Costa is also a promising choice. A model directly learning from consumption behaviors may think the user dislike Costa. Second, the same service may satisfy different types of demands for different users. For example, a user consumes at Starbucks to *have a rest*, while another user is due to *hold a remote meeting*. Thus, predicting users' future consumption behaviors merely and directly based on historical behaviors may cause biases, leading to sub-optimal prediction results.

As a result, understanding and predicting users' demand, which can be formally named as *user consumption intention prediction*, is a convincing and reasonable approach to aid the prediction of user consumption behavior. In general, the problem of consumption intention prediction can be defined as predicting a user's desire or demand that driven him/her to consume a kind of service or product, such as to *have a rest* or *hold a remote meeting*. Nevertheless, developing prediction models of user consumption intention for real-world scenarios is quite challenging due to three aspects.

- **First, the consumption intention is always implicit and hard to capture.** Compared with consumption behaviors, the consumption intention reflects a morefi ne-grained user preference. As mentioned above, it serves as the internal driving force, and the same consumption behavior may be driven by different consumption intentions. As a result, consumption intention is not explicitly reflected by the behavioral data, which is the most critical issue in intention prediction.
- **Second, the consumption intention is highly affected by user intrinsic preference and spatio-temporal context.** In the real world, the consumption intention is dynamic and can be largely affected by the spatio-temporal context, besides user intrinsic preference such as taste style, consumption level, etc. For example, a consumption intention of having a rest is more

KDD '21, August 14–18, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467178

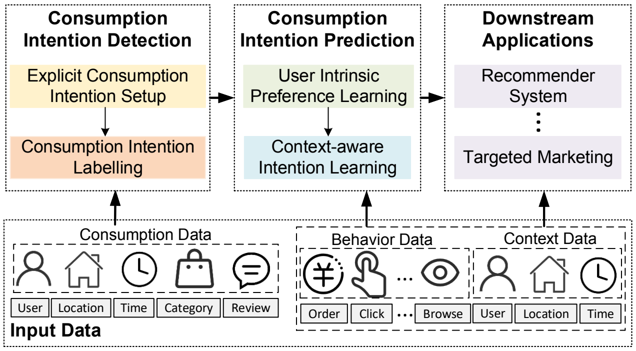

Consumption Intention Detection
Explicit Consumption Intention Setup
Consumption Intention Labelling
Consumption Intention Prediction
User Intrinsic Preference Learning
Context-aware Intention Learning
Downstream Applications
Recommender System
Targeted Marketing
Consumption Data
User
Location
Time
Category
Review
Behavior Data
Order
Click
Browse
Context Data
User
Location
Time
Input Data

Figure 1: The framework of our user consumption intention detection and prediction system.

likely to happen at weekends and entertainment areas. Thus, combining the modeling of user intrinsic preference and spatio-temporal context is essential but challenging.

- **Third, the internal structure of consumption intention is complex.** Consumption intention itself is in a complex structure, making it difficult to learn. Some kinds of consumption intentions, although not the same, are highly correlated, such as to *have lunch* and *have dinner*. More specifically, these intentions can be classified into the clustering structure, of which the intentions belonging to one same cluster are highly correlated and similar, while there are great contrasts among the intentions belonging to different clusters.

There is seldom work well exploring the problem of consumption intention prediction due to these above-mentioned challenges. There are some works using the term, intention, to refer to the behavior [22, 32, 33], product category [5] or query word [8], which do not conform to the precise definition of consumption intention. There are some other works [21, 31, 35] using the term, intention, to refer to user latent factor in behavior-prediction models, while the concept of intention in these works is still vague and unseen.

We develop the Meituan's intention prediction system in a two-stage manner, as shown in Figure 1, to address these challenges.

- The first stage is ***consumption intention detection***. Although the consumption intention is not explicitly reflected by user behavior , users tend to write reviews mentioning why they consume the product or service. Thus, taking advantage of the user review data, we propose a simple yet effective keyword matching method for detecting and classifying consumption intentions.
- The second stage is ***consumption intention prediction***. To model both the user intrinsic preferences and spatio-temporal context, we build a dual-stage graph neural network model ***GRIP***, namely ***GRaph neural network for Intention Prediction***, which can not only learn from users' long-term interactions but also capture the effect of spatio-temporal context. To be more specific, with shared embedding layers between two graph neural networks, we design two graph convolutional layers to extract two kinds of effects, respectively. As for the third challenge of the complex internal structure of intention, we propose to disentangle the latent representations of intentions, with the disentangling loss that supervises the learning to meet the clustering structure.

The contribution of this work can be summarized as follows.

- To the best of our knowledge, we take the first step into the problem of consumption intention prediction, which is a real-world problem with high value and wide applications. The consumption intention serves as the driving force of consumption behavior, and we detect and define 19 types of intentions in Meituan.
- We develop a system in Meituan consisting of consumption intention detection and prediction. We first propose to explicitly reveal consumption intentions from users' review data. We then design a dual graph convolutional network model GRIP for capturing both the users' intrinsic preferences and the effect of spatio-temporal context. We develop a disentangling loss function for modeling the clustering-structure of intentions to enhance the intention's representation learning.
- We evaluate our system via both offline experiments and online A/B tests. The experimental results show that our consumption intention prediction model, GRIP, effectively outperforms a group of state-of-the-art prediction methods and shows vigorous contribution on 6%-66% improvements in prediction performance. The online industrial results reveal that even with a simple prediction method, our system achieves average relative improvements of about 1.55% on *Page View* and 5.15% on *Page View for Cold-start Customers*. Moreover, compared with baseline prediction method, our GRIP achieves 44.72% improvements on prediction precision and 84.96% improvements on *Customer Growth*.

.

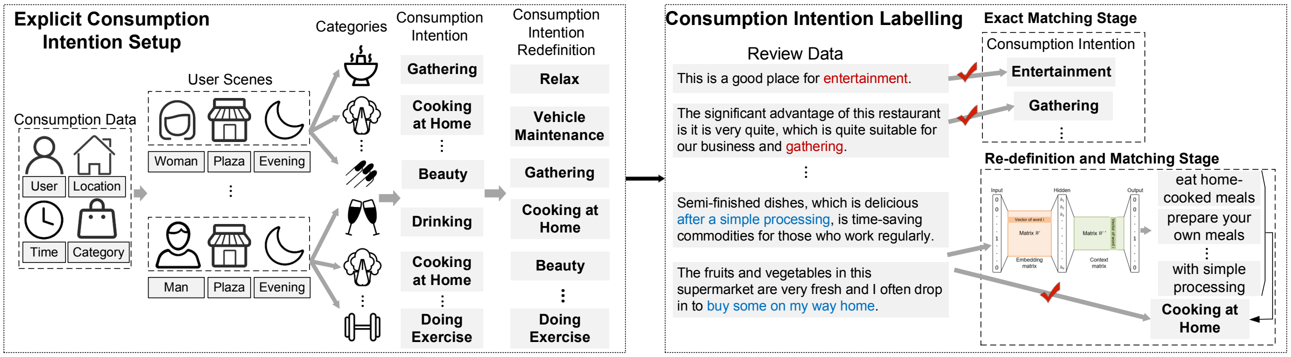
2 CONSUMPTION INTENTION DETECTION

To discover users' explicit desire or demand from their consumption data, we propose an effective method to detect consumption intentions and match consumption with intentions based on reviews. It has been shown that reviews, as a natural source of data to exploit, can be intended to 'explain' the underlying dimensions behind users' decisions [2, 23, 35, 41]. Reviews are much more expressive with text and useful to infer the consumption intention of users, compared with interaction data, which may be sparse or noisy. Specifically, we first set a series of classifications of consumption intentions with the help of human experts and historical consumption data. Then, we propose a keyword matching method to match each consumption with a intention. The architecture of this process is illustrated in Figure 2, and we elaborate on the design of these two stages as following.

2.1 Explicit Consumption Intention Setup

We aim to set explicit consumption intention utilizing consumption data on the life service platform. For different kinds of users, their consumption intentions are highly related to their preference or habits. For example, white-collar workers may have higher demand for cooking at home than students. In addition, the consumption intentions for life service are strongly relevant to location and time of users' visits. For example, users tend not to have the desire to do car maintenance in the morning of a working day. So to consider the influence of all these factors, we use a tuple of (user attributes, time, location) to objectively describe the conditions of each consumption, and we name it as, *user scene*.

As showing in Figure 2, with the consumption data, including a user purchased a certain category of products at a certain time and location, we obtain a set of user scenes, for example, (female, plaza, weekend afternoon). Then for each user scene, we combine the



This is a good place for entertainment.

The significant advantage of this restaurant is it is very quite, which is quite suitable for our business and gathering.

Semi-finished dishes, which is delicious after a simple processing, is time-saving commodities for those who work regularly.

The fruits and vegetables in this supermarket are very fresh and I often drop in to buy some on my way home.

Figure 2: The architecture of our proposed user consumption intention detection method.

Table 1: 19 types of explicit consumption intention.

Explicit Consumption Intention	Gathering, Cooking at home, Entertainment, Drinking, Business trip, Beauty, Housewares, Doing exercise, Pets, Snack, Family life, Vehicle maintenance, Leisure activities, Massage, Traveling, Grab a bite, Childcare, Dainty morsel, Training and learning

knowledge of human experts and consumption information (e.g. category, price, etc.) to set the detailed consumption intentions. For example, the consumption on cosmetic makeover is related to the intentions of `Beauty` while board games is related to `Leisure Activities`. We further redefine the consumption intentions via merging and differentiating them in different user scenes and obtain the final 19 classes of consumption intentions, showing in Table 1.

It can be observed that the consumption intentions we set include all kinds of life service consumption. Moreover, for the consumption of some categories with complex desires, e.g. dining, there are multiple related intentions in detail, each of which focuses on different aspects, e.g. `Grab a bite` relates to food that can be eaten quickly while `Dainty morsel` relates to food that is greatly pleasing to the taste. But for those relatively limited consumption categories, where the desire is simple and unified, we summarize a generalized intention for each kind, e.g. `Pets`, `Vehicle maintenance`.

2.2 Consumption Intention Labelling

After obtaining the classifications of intentions, we use reviews written by consumers based on their experience to match each consumption with a intention, i.e., user consumption intention labelling. It can be treated as a keywords matching task, where we need to search whether each review contains the keywords of the intentions. If the keywords of an intention is included, the consumption data related to the review can be matched with the intention. Thus, we design a two-stage matching method for consumption labelling, showing in Figure 2. In the first stage, *Exact Matching Stage*, we directly use each intention, which is a word or a list of words, as the keywords, to do exact matching with the reviews. For most intentions, such as `Entertainment` and `Gathering`, we can directly find the words in the review for matching. However, for the rest of the intentions where the keywords rarely appear and hard to be found directly in reviews, we need the second stage, *Re-definition and Matching Stage*. In this stage, we utilize Word2Vec [18, 24] to obtain the re-defined keywords for each intention. The training set of Word2Vec also comes from reviews. For each sentence in reviews, we use a sliding window with a length of $2 \times c + 1$, which select c words before and after, to obtain phrases. Here we use $c = 2$ since our intentions are no more than 3 words. We put the phrase obtained each time in the training set and use Skip-gram [18, 24] to find the most related phrases for each intention. For example, reviews like `eat home-cooked meals`, `prepare your own meals` and `with simple process` are all corresponding to the intention of `Cooking at home`. We utilize these phrases as keywords and do the matching operation again.

With the above proposed consumption intention labelling method, we can not only label the consumption data related to common intentions, but further explore those intentions with relatively rare occurrences or more diverse expressions.

By introducing the objective formulation of user scene as (user attributes, time, location) and incorporating it with categories and reviews of consumption data, we enrich the semantics and rationality of consumption intentions which sets up the dataset for following prediction task. This method will be further evaluated by online evaluation in Section 5.

3 CONSUMPTION INTENTION PREDICTION

To predict the consumption intention of users, we need to comprehensively model all the influencing factors. In general, there are two major types of factors. a) *User intrinsic preference* implies higher attention or interests of users for certain aspects of the item, such as taste style, consumption level, etc. It can be represented in the interactions between users and items on the platform. b) *Spatio-temporal context* implies the current spatio-temporal factor of the scene and users' intention could change greatly in different scenes, as introduced in Section 2.1. Such context information can be very helpful because most users go to a place with a certain degree of intention rather than completely wandering around.

To fully utilize all these heterogeneous factors as well as their attributes for consumption intention prediction, we design a dual-stage GCN framework, ***GRIP*** for short, to first model user intrinsic preference according to users' historical interactions with different categories, and then learn a context-aware intention embedding with spatio-temporal information, as showing in Figure 3. In GRIP, the attributes of users are used in both graphs and thus these embeddings are shared in two graphs. Now we introduce the detailed design of the prediction model.

3.1 User Intrinsic Preference Modeling

In the first GCN model, we aim to model user intrinsic preference for consumption intention prediction, which is represented in their

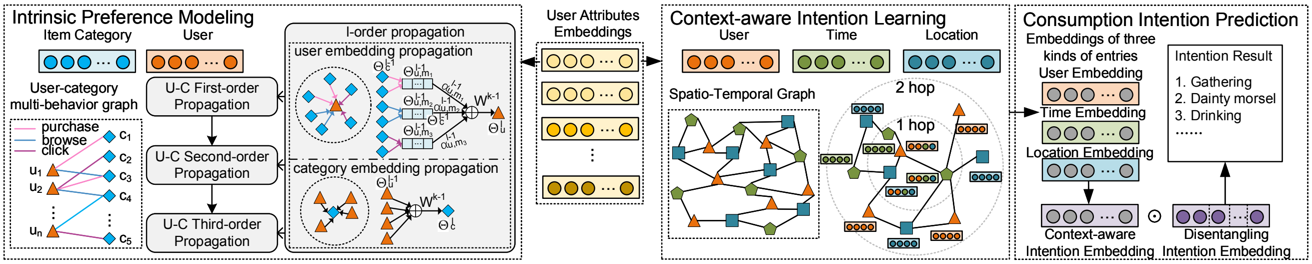


Figure 3: The architecture of our proposed user consumption intention prediction method GRIP.

historical behaviors, especially interactions with different items. For example, if a user only has browsed and purchased high-priced items on the platform, it means that (s)he is serious about the quality so may tend not to have the consumption intention of `Grab a bite`.

However, item-wise preference may be difficult to fully generalize because of the sparsity of observable user-item pairs. In addition, different behaviors, e.g., clicking, collecting,purchasing, etc., help to exploit different types of feedback of users intrinsic preference. To capture the category-wise users intrinsic preference based on multi-behaviors, we build a heterogeneous graph consisting of two kinds of nodes of users and categories and multiple types of edges, where an edge connecting user node and category node refers to a specific type of user-category interaction behavior.

As showing in Figure 3, we utilize three types of behavior: clicking, browsing and purchasing and three interaction matrices Y^1, Y^2, Y^3 to denote if user has interacted with categories under these three behaviors, respectively. Specifically, each interaction matrice Y^m is in the binary form, of which each entry has value 1 or 0, defined as follows for user u and category c,

$$y_{uc}^m = \begin{cases} 1, \text{if } u \text{ has interacted with } c \text{ under behavior } m, \\ 0, \text{otherwise.} \end{cases} \tag{1}$$

We represent the interaction data by an undirected graph $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{X})$, where nodes are $\mathbf{V}$ consisting of user nodes $u \in U$ and category nodes $c \in C$. The edges in $\mathbf{E}$ contain different user-category interaction edges of different behaviors, namely $(u,c)_m, m \in N_r$, where N_r is the set of all behavior types. $X = \left(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{|V|}\right)$ represents the feature vector of each nodes. For users, the feature information is indicated by the profiles of users (e.g., age, gender, marital, etc.). We represent each profile with a set of embeddings for different values and denote each user node as $\mathbf{u}_q = [u_{q1}, u_{q2}, ..., u_{qP}]$ with P profiles. For categories, the feature information is the ID of the category. There will be an edge $(u,c)_m$ built when $y_{uc}^m = 1$. Similar to the message-passing architecture in [12], we utilize a behavior-aware user-category propagation method for better utilization of multi-behavior information. In the category-to-user embedding propagation module, we learn the importance of different behaviors automatically and assign different weights to categories interacted under different behaviors. In the user-to-category embedding propagation module, we aggregate the information of neighbor user for each category directly.

3.1.1 User Embedding Propagation. To consider categories' influence on users according to the behavior type, we assign a weight for each behavior, namely w_m for behavior m and the propagation weight for a particular behavior m for user u denoted as α_{um}:

$$\alpha_{um} = \frac{w_m \cdot n_{um}}{\sum_{t \in N_r} w_t \cdot n_{um}}, \tag{2}$$

where w_m is a behavior-wised importance weight of behavior m, and n_{um} is the total number of behavior m operated by user u. To be specific, behavior with larger w will be of higher importance comparing to behavior with smaller w. α_{um} is the final propagation weight of behavior m for user u, which will be used in propagation layer and $\sum_m \alpha_{um} = 1$. w_t is learned by the model so that the importance of each behavior can be learned automatically.

For each user u, we aggregate the categories that have the same behavior interaction together to obtain one embedding for each behavior, namely $\Theta_{u,m}^{(l)}$ under behavior m, which is defined as follows,

$$\Theta_{u,m}^{(l)} = \text{aggregate}\left(\Theta_c^{(l)} \mid c \in N_m^Y(u)\right), \tag{3}$$

where l refers to the l-th layer, $N_m^Y(u)$ is the set of categories that user u has interacted under behavior m. $\Theta_c^{(l)}$ is the embedding of category c in the l-th layer and $\Theta_{u,m}^{(l)}$ is the category aggregated embedding for user u under behavior m in l-th layer. They are aggregated according to weight α_{um} and an encoder matrix is used to obtain the final neighbor aggregation for user u, as follows,

$$\Theta_u^{(l+1)} = W^{(l)} \cdot \left(\sum_{m \in N_r} \alpha_{um} \Theta_{u,m}^{(l)}\right), \tag{4}$$

where $\Theta_u^{(l+1)}$ is the embedding of use u in the $(l+1)$-th layer and $W^{(l)}$ is the encoding matrix for aggregation in the l-th layer.

3.1.2 Category Embedding Propagation. Generally different users have the same contribution to one category. Thus we aggregate user embedding $\Theta_{uj}^{(l)}$ for the next layer category embedding $\Theta_{ci}^{(l+1)}$ as follows,

$$\Theta_c^{(l+1)} = W^{(l)} \cdot \text{aggregate}\left(\Theta_j^{(l)} \mid j \in N^U(c)\right), \tag{5}$$

where $N^U(c)$ is the set of users that category c has connected with, $\Theta_{uj}^{(l)}$ is the embedding for user j and $W^{(l)}$ is the encoding matrix for information aggregation on the l-th layer.

After propagating through L layers, we obtain multiple representations, namely $\left\{\Theta_u^{(0)}, \ldots, \Theta_u^{(L)}\right\}$ for user u and $\left\{\Theta_c^{(0)}, \ldots, \Theta_c^{(L)}\right\}$ for category c. We utilize the representations obtained in the last layer as the final embedding for users and categories. To learn the parameters in this graph, we employ inner product to calculate the score based on user-category propagation, as follows,

$$y_(u,c) = \Theta_u^L \cdot \Theta_c^L, \tag{6}$$

Then, we use BPR loss [25] to emphasize that the observed interaction, which is instructional for user intrinsic preference learning, should be assigned a higher score than unobserved ones for this graph. The loss function is defined as follows,

$$\text{loss}_{\text{bpr}} = \sum_{(u,c_j,c_k) \in O} -\ln \sigma(y(u,c_j) - y(u,c_k)), \tag{7}$$

where $O = \left\{(u,c_j,c_k) \mid (u,c_j) \in R^+, (u,c_k) \in R^-\right\}$ denotes the set of training data, R^+ represents observed links, R^- represents unobserved links, and $\sigma(\cdot)$ is the sigmoid function.

3.2 Spatio-temporal Context Modeling

Although the consumption intention is affected by the intrinsic preference of users, it is also strongly relevant to location and time of the scene. Thus we aim to model these factors together to obtain a context-aware intention representation for intention prediction.

Tofi gure out the interior relation between intrinsic preference, time and location, we construct a graph $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{X})$ to encode the connections. With the user feature vector from Section 3.1, which reflects the user intrinsic preference, the nodes $\mathbf{V}$ in the graph consist of user nodes $u \in U$, time nodes $t \in T$ and location nodes $l \in L$. Edges $\mathbf{E}$ representing the co-occurrence relationship between different nodes which can be generalized as three edge types in the graph: a)*time-location* edge, b)*user-time* edge and c)*user-location* edge. To effectively model the strength of different connections among nodes, each edge is assigned with a weight. Here we employ min-max normalization to calculate weights between nodes. we denote the co-occurrence matrix as C and the maximum weight for *Time-Location* edge as C^{tl}_{max}, *User-Time* edge as C^{ut}_{max} and *User-Location* edge as C^{ul}_{max} respectively. Then, the normalized weight of edge between node p and q is defined as:

$$A_{pq} = \begin{cases} C_{pq}/C^{tl}_{max} & (p,q) \text{ is time-location edge,} \\ C_{pq}/C^{ut}_{max} & (p,q) \text{ is user-time edge,} \\ C_{pq}/C^{ul}_{max} & (p,q) \text{ is user-location edge,} \\ 1 & p = q, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

$X = \left(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{|V|}\right)$ represents the feature vector of each nodes and consists of three types of vector: user feature vector $\boldsymbol{u}$, location feature vector $\boldsymbol{l}$, and time feature vector $\boldsymbol{t}$. For users, the feature information is indicated by the profiles of users (e.g., age, gender, marital, etc.) and each user node is denoted as $\mathbf{u}_q = [u_{q1}, u_{q2}, ..., u_{qP}]$ for P profiles. The feature vector of these user profiles are shared in user intrinsic preference modeling graph and this spatio-temporal modeling graph. For times, we categorize the time slots as working days and non-working days since people's lifestyles differ a lot between these two types of days. Then, we discretize the time in one day into N units for better modeling. For locations, the feature information is indicated by the location id. Since each node possesses only one type (user/time/location), for types which are different from the node type, the corresponding feature is set to be zero vectors and the overall feature vector for each node is $\boldsymbol{x} = [\boldsymbol{u}, \boldsymbol{l}, \boldsymbol{t}] \in \mathbb{R}^D$

Based on the built graph, the representation learning method is based on Graph Convolutional Network (GCN)[16]. Let $\Theta^{(l)} = \left(\boldsymbol{\theta}^{(l)}_1, \boldsymbol{\theta}^{(l)}_2, \ldots, \boldsymbol{\theta}^{(l)}_{|V|}\right)$ be the matrix of all node embedding vectors at step l, where $\Theta^{(0)} = \mathbf{X}$, the aggregation function is calculated as:

$$\Theta^{(l)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}\Theta^{(l-1)}W^{(l)}\right), \tag{9}$$

where $\tilde{A} = A + I_N$, and A is the adjacency matrix with the identity matrix, D is a diagonal matrix with $D_{ii} = \sum_j \tilde{A}_{ij}$ and σ represents the non-linear activation function. Input node feature X serves as the initial embedding $\Theta^{(0)}$ and $W^{(l)}$ is the trainable parameters in the l-th layer. In this way, for each node, the embedding is updated to the weighted average of itself and its neighbors in the graph. Since the neighborhood relationship in the graph indicates co-occurrence in records, via aggregating the high-order neighborhood information for each node, we can utilize the contextualized information for each consumption.

After propagating through L layers, we obtain the representations for users Θ_u, locations Θ_l and time Θ_t. We utilize the representations obtained in the last layer as thefi nal embedding and the context-aware intention representation for user scene v can be computes as follows,

$$\mathbf{r}_v = \frac{(\boldsymbol{\theta}(u_v) + \boldsymbol{\theta}(t_v) + \boldsymbol{\theta}(l_v))}{3}. \tag{10}$$

3.3 Consumption Intention Disentangling

For different intentions, they are not isolated but have certain relations with each other. For example, the consumption intention of *Gathering* and *Drinking* although not the same, are highly correlated, however, the correlation between *Gathering* and *Training and learning* is relatively weaker. Thus, the consumption intentions can be classified into the clustering structure, where intentions belonging to one same cluster are similar, while there are great contrasts among the intentions belonging to different clusters. To capture the inter-cluster difference and inner-cluster closeness of consumption intentions, we introduce latent representations for 19 types of intentions $\mathbf{e}_i = \mathbf{e}^1_i, \mathbf{e}^2_i, \ldots, \mathbf{e}^{19}_i$ and generate disentangled representations. More formally, for the j-th consumption intention, we disentangle its latent representations $\mathbf{e}^j_i$ into K independent parts as follows,

$$\mathbf{e}^j_i = \left(\mathbf{e}^j_{1i}, \mathbf{e}^j_{2i}, \ldots, \mathbf{e}^j_{Ki}\right), \tag{11}$$

where $\mathbf{e}_{ki}$ is the k-th latent intent influence for consumption intention. All these K parts are in the same dimension $\frac{D}{K}$ where D is the dimension of each context-aware intention representation $\mathbf{r}_v$. To make the different parts of intention representations to be independent, we utilize the distance correlation [27, 28], which is able to characterize independence of any two vectors and is zero if and only if these vectors are independent. So the loss function can be computed as follows,

$$\text{loss}_{\text{dis}} = \sum_{k=1}^{K} \sum_{k'=k+1}^{K} d_{Cor}\left(\mathbf{E}_k, \mathbf{E}_{k'}\right), \tag{12}$$

where $\mathbf{E}_k = \left[\mathbf{e}^1_{ki}, \mathbf{e}^2_{ki}, ;\mathbf{e}^{19}_{ki}\right] \in N \times \mathbb{R}^{\frac{D}{K}}$ is the k-th disentangled embedding of the 19 consumption intention representations and d_{Cor} is the function of distance correlation defined as:

$$d_{Cor}\left(\mathbf{E}_k, \mathbf{E}_{k'}\right) = \frac{d_{Cov}\left(\mathbf{E}_k, \mathbf{E}_{k'}\right)}{\sqrt{d_{Var}\left(\mathbf{E}_k\right) \cdot d_{Var}\left(\mathbf{E}_{k'}\right)}}, \tag{13}$$

where d_{Cov} represents the distance covariance between two matrices; d_{Var} is the distance variance of each matrix.

3.4 Consumption Intention Prediction

To predict the consumption intention of a user with the spatio-temporal context, we compare the context-aware intention representation r_v with 19 disentangled embeddings of consumption intentions. The scores for different consumption intentions $\mathbf{e}^j_i$ can be computed as follows,

$$s_{vj} = \frac{\boldsymbol{r}_v \cdot \mathbf{e}_i^j}{\|\boldsymbol{r}_v\| \cdot \left\|\mathbf{e}_i^j\right\|}. \tag{14}$$

We take the consumption intention with the maximum scores as the prediction result.

3.5 Training and Model Optimization

To overcome the class imbalance of different consumption intentions, we design a multi-class focal loss [19] to down-weight the data from those consumption intentions of a large amounts of training data in the total prediction loss. Thus the prediction loss is defined as follows:

$$\text{loss}_{\text{pred}} = -\sum_{v \in O} \left(\sum_{j=1}^{19} \left(1 - p_{vj}\right)^{\gamma} q_{vj} \log\left(p_{vj}\right) \right), \tag{15}$$

where O denotes the set of training data, $p_{vj} = \text{Softmax}\left(s_{vj}\right) = e^{s_{vj}} / \sum_k e^{s_{vk}}$, γ is a adjustable parameter and $q_{vj} = 1$ if $\boldsymbol{r}_v$ belongs to the true label, else it is 0.

The entire model with the aforementioned four parts is trained with standard backpropagation as an end-to-end framework. Thus, the entire loss function is:

$$\text{Loss} = \lambda_1 \, \text{loss}_{\text{bpr}} + \lambda_2 \, \text{loss}_{\text{pred}} + \lambda_3 \, \text{loss}_{\text{dis}}, \tag{16}$$

where λ_1, λ_2 and λ_3 are three defined hyper-parameters. During the training, we alternatively optimize the parameters in user intrinsic preference modeling GCN and that in context-aware intention learning GCN.

To evaluate the effectiveness of the system, including consumption intention detection and prediction, we conduct both offline experiments and online A/B test. Wefi rst evaluate the prediction performance of GRIP on offline dataset in Section 4. Then we deploy the system based on Meituan mobile App’s Homepage for online evaluation in Section 5.

4 EVALUATION OF INTENTION PREDICTION

4.1 Experimental Settings

Dataset We collect large-scale user behavior logs and consumption data from Meituan ranging from Jun. 30 to Sep. 24 2020. There are three types of behaviors, including clicking, browsing and purchasing in user behavior logs. With the consumption intention labelling method, we label each piece of consumption data with review to a type of consumption intentions listed in Table 1. We utilize the *user scene*, including user attributes, location and time currently as the input to predict which type the user’s consumption intention belongs to. Then, to evaluate the performance of our intention prediction model GRIP, we divide the data based on the time span into training set and test set. More statistics are provided in Table 2.

Baselines We compare our GRIP with three categories of baseline methods: a) two most representative feature-based prediction models: XGBoost [4] and DeepFM [9], b) two graph embedding based models: ReconEmbed [39] and GraphEmbed [39], and c) four graph neural network based models:SA-GCN [37], SA-LightGCN [10], HAN [34] and HetGNN [38]. The details of these baseline methods are provided in Appendix A.1.

Metrics To evaluate the performance of prediction model, we use two widely used metrics, Precision@K (Prec@K) and NDCG@K [12, 14, 36]. The details of the definitions are provided in Appendix A.1. In our experiments, we use the popular setting of $K = 1, 3, 5$ [14, 36] for evaluation.

Table 2: Summary of statistics of the dataset. ‘# Locations’ indicates the number of location scene we use in spatio-temporal graph. ‘# Category’ indicates the number of item categories we use in user intrinsic preference modeling graph. ‘# Users’ indicates the number of users in the training or test data.

	Time Span	#Users	#Location	#Category
Training set	6/30-8/31/2020	339,982	12	824
Test set	9/1-9/24/2020	97,947		

Table 3: Comparison of overall intention prediction performance of GRIP and baselines.

	Prec@1	Prec@3	Prec@5	NDCG@3	NDCG@5
XGBoost	0.1120	0.2887	0.3182	0.2103	0.2457
DeepFM	0.1157	0.3010	0.3402	0.2257	0.2816
ReconEmbed	0.1213	0.3125	0.4103	0.2473	0.3019
GraphEmbed	0.1274	0.3306	0.4405	0.2556	0.3228
SA-GCN	0.1336	0.3219	0.4302	0.2157	0.2916
SA-LightGCN	0.1573	0.3527	0.4457	0.2331	0.3005
HAN	0.2620	0.3471	0.5391	0.2411	0.3039
HetGNN	0.1747	0.3836	0.6443	0.2651	0.3861
GRIP	**0.2903**	**0.4356**	**0.6857**	**0.3809**	**0.4744**

4.2 Overall Performance

We start by comparing the performance of our proposed GRIP with all other baselines. The result is reported in Table 3. From the result, we have the following observations.

- **Our proposed GRIP outperforms all the baselines significantly.** Benefiting from the interaction graph with multi-behaviors and spatio-temporal graph for dynamic context, GRIP is capable of modeling user intrinsic preference and context-aware consumption intentions. Compared to all baseline methods, GRIP obtains the best performance in terms of Prec@K and NDCG@K. Specifically, its relative improvements over the strongest baselines are 10.80% for Prec@1, 13.56% for Prec@3, 6.43% for Prec@5, 43.68% for NDCG@3 and 22.87% for NDCG@5. On average, it outperforms the best baseline by 10.26% for Prec and 33.28% for NDCG. Such improvements on various metrics demonstrate the effectiveness of GRIP.
- **Traditional feature-based prediction models perform worse than graph-based models.** We can observe that those feature-based methods achieve a worse performance compared with graph embedding based models and graph neural network based models. The average performance of graph-based models is higher than the best performance of feature-based models in terms of 6.52% for Prec@1, 5.39% for Prec@3, 17.35% for Prec@5, 4.13% for NDCG@3 and 6.14% for NDCG@5, wihch indicates the graph’s power to model the complex relationship among different influencing factors in our problem.
- **Graph convolutional methods achieve better performance compared with normal graph embedding models.** SA-GCN

and SA-LightGCN, which derive the node representation by aggregating the features appearing in the neighborhood, perform better than ReconEmbed and GraphEmbed, which learn representations for nodes only to preserve their correlations. This indicates that structure information of the graphs can benefit the intention prediction.

- **User intrinsic preference modeling is essential.** The performance of the three models, ReconEmbed, GraphEmbed and SA-GCN, which are not able to make use of the interaction between users and item categories and only take spatio-temporal context information into account, drops significantly compared with other graph-based models. This demonstrates that modeling users intrinsic preference from interaction-data is essential for consumption intention prediction.

In general, the prediction performance not only justifies the effectiveness of our model, but also demonstrates the rationality of our design to model both user intrinsic preference and spatio-temporal context information for consumption intention prediction.

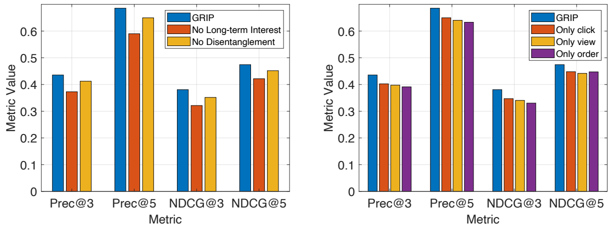
4.3 Ablation Study

In GRIP, we utilize a dual-stage graph neural network to combine intrinsic preference of users and spatio-temporal context influence to predict users' consumption intentions. An intuitive question is whether the designed component really help in our model?

To answer it, we conduct experiments on two types of degenerative models of GRIP. The first is to remove the utilized components from the full model, which is called model ablation, while the second is to remove some types of data we use, which is called data ablation. We adopt the same evaluation methods with above experiments, and the performance comparison is shown in Figure 4.

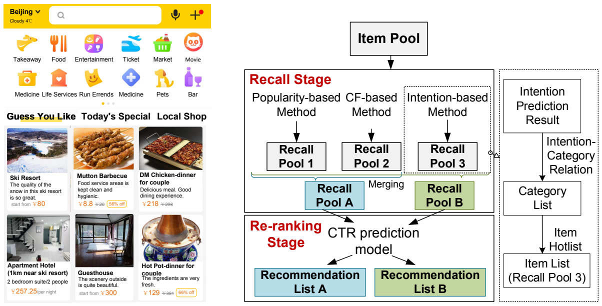
Model Ablation Study. In order to evaluate the effect of user intrinsic preference learning and disentangling intention embedding, we remove these two parts from the full prediction model, respectively. The result is shown in Figure 4(a). It is shown that the complete model outperforms the model without user intrinsic preference learning and the model without disentangling intention embedding by 16.85% and 5.6% on Prec@3 and 18.51% and 6.67% on NDCG@3, respectively. This demonstrates that both the intrinsic preference learning and disentangling intention embedding are essential in users' consumption intention prediction.

Data Ablation Study. We use multi-behavior data in user intrinsic preference learning, including clicking, browsing and purchasing. In order to evaluate the effect of the importance of different behaviors and the effectiveness of multi-behavior information, we use only one type of behavior in the interaction graph each time and the performance is shown in Figure 4(b). It is shown that multi-behavior data performs better than any single-behavior data by at least 8.25% on Prec@3 and 9.71% on NDCG@3. Besides, comparing the performance of single-behavior data, we can find that clicking data leads to the best performance of learning user intrinsic preference while purchasing data performs worst. This demonstrates that single-behavior data leads to the loss of some of the information in user intrinsic preference modeling. Moreover, compared with the sparse purchasing data, clicking and browsing data can better reflect the intrinsic preference of users.



(a) Model Ablation (b) Data Ablation

Figure 4: Performance comparison of ablation study.


Beijing
Cloudy 4℃

Guess You Like Today's Special Local Shop

Ski Resort
The quality of the snow in this ski resort is so great.
start from ￥80

Mutton Barbecue
Food service areas is kept clean and hygienic.
￥8.8 ~~￥20~~ 56% off

DM Chicken-dinner for couple
Delicious meal. Good dining experience.
￥218 ~~￥298~~

Apartment Hotel (1km near ski resort)
2 bedroom suite/2 people
￥257.25/per night

Guesthouse
The scenery outside is quite beautiful.
start from ￥300

Hot Pot-dinner for couple
The ingredients are very fresh.
￥129 ~~￥384~~ 66% off

(a) Meituan App's homepage. (b) Online recommendation engine

Figure 5: Illustration of Meituan App's homepage and its recommendation engine.

To further analyze the influence of hyper-parameters and effectiveness of disentangling, we perform hyper-parameter study and visualization for intention embeddings, which is represented in Appendix A.3 and Appendix A.4, respectively.

5 INDUSTRIAL RESULTS

We conduct online A/B tests to evaluate the effectiveness of our designed system's online performance, including consumption intention detection and prediction. Specifically, we deploy it at the recommendation engine for Meituan mobile App's Homepage, which is the page with the most user behaviors. At the App's Homepage, as shown in Figure 5(a), a personalized list of local services can be exposed to user. As we have mentioned above, consumption intention, as the internal driving force of users' consumption behaviors, can well support predicting the behaviors, which is consistent with the goal of the recommendation scenario at the Home Page. The user intention can be correlated to a list of services by matching with services' profiles. Specifically, we first get the category list related to the intentions with top prediction scores and then for each category, we select the items with top real-time sales to generate the recall pool. That is, our consumption intention system can generate a list of services (Recall Pool B in Figure 5(b)). This pool, together with other two pools generated by two other strategies of popularity and collaborative filtering (CF), can be fed into the re-ranking stage for obtaining the final recommendation list. The online A/B tests are conducted from October 31st to November 7th, 2020, involving about 8 millions of users.

We first evaluate the intention prediction module of our system. To ensure the results can reflect the effect of intention detection, for intention prediction we only adopt a very simple DeepFM model [9]. We randomly divide users into two buckets, each of which has similar amount of users, and assign them different recall strategies.

Specifically, for the first bucket we adopt popularity-based and CF-based recall. For the second bucket, there is an additional intention-based recall. Note that we keep all the other modules exactly the same for fair comparison. Two important metrics measuring the quality of recommendation list, *Page View* (PV) and *Page View for Cold-start Customer* (PV-CC) are used, of which PV-CC emphasizes whether the recommendation list can attract users to consume a new kind of service. The relative improvements in different Business Units (BUs) of the second bucket (of which our system are introduced) compared with the first one are shown in Table 4.

From the result, we can observe that for all BUs there are average relative improvements of about 1.55% on PV and 5.15% on PV-CC. Such improvement can lead to great increase in consumption amount of the platform. Among these BUs, PV improves most in `Holidays` by 4.01%, while PV-CC improves most in `Flash Sale`, which is a new BU launched by Meituan in 2018, by 11.79%. This shows the intention-based recall can help customers explore types of services that has not consumed before greatly. This further validates that directly learning from the behavior data may cause the model highly constrained by the historical behaviors, as we mentioned in Introduction. Moreover, the slightly decreases of 0.91% on PV and 0.51% on PV-CC for `Catering` and 0.83% on PV and 0.92% on PV-CC for `Hotel`, are also within the normal fluctuation range of the market.

We further evaluate our proposed intention prediction model GRIP. We randomly divide users into three buckets, each of which has similar amount of users, and assign them different intention prediction models. Specifically, we deploy MLP (Multi-layer perceptrons), DeepFM, and our proposed GRIP as the intention-prediction model, respectively. We also keep all the other modules the exactly same. Two metrics, Prec@3 (precision of intention prediction) and *Customer Growth of All Business Units, shorted as CG* are used. The reason of using CG, instead of PV or PV-CC, is that CG is a more critical metric that can better distinguish good or ordinary intention prediction models. The results of different buckets are shown in Figure 6. We can observe that our proposed GRIP can bring about relative improvement of about 44.7% on Prec@3 and 83.9% on CG compared with the DeepFM model. This demonstrates the effectiveness of our proposed GRIP in accurately predicting user consumption intention and then generating higher-quality recommendation results.

In short, we conduct extensive online A/B tests on million-level users to evaluate the real-world performance of our system. The results demonstrate that a) intention prediction can largely benefit the personalized recommendation, especially in recommending cold-start service types, and b) our proposed intention prediction model, GRIP, can achieve promising results of customer growth.

6 RELATED WORK

Consumption Intention Prediction Consumption intention of users is an important factor to be considered nowadays, especially for recommender systems. The focus of some recent studies has turned to what users want (user intention) from what users like (user behavior) and these studies can be divided into two types. The first type is to use the term intention to refer to the behavior [5, 32, 33], product category [5] or query words [8]. Wang *et al.* [32, 33] indicate the complex relations between user behaviour and intentions that intentions are the driving factors of actions and one intention requires a series of relevant actions to accomplish with psychological theories. So they use the series of relevant actions to represent the intention. Chen *et al.* [5] claim that user intention carries two aspects of information: the category of products that a user tends to interact with and the way the user wants to perform the interaction. Thus, they formulate user intention as a tuple of action type (e.g., click, purchase, etc.) and product category tag, i.e., (action, category). Fan *et al.* [8] present several words describing the product in the search box without any input from users as the result of their intent recommendation.

BU	PV	PV-CC
Holidays	+4.01%	+7.33%
Catering	-0.91%	-0.51%
Flash Sale	+1.73%	+11.79%
Kids	+3.16%	+10.10%
Beauty	+2.17%	+3.11%
Hotel	-0.83%	-0.92%

Table 4: Result of online A/B experiments for our recall model.

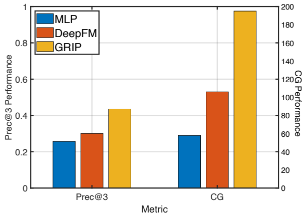


Figure 6: Result of online A/B experiments for intention prediction model.

In another type of researches, they utilize latent representation for intentions [21, 31, 35] to improve the accuracy of recommender systems. Wang *et al.* [35] establish as set of intent-aware graphs, based on the fact that a user generally has multiple intents to adopt certain items and different intents could motivate different user behaviors, to learn latent representations for intents and distill the signals of user intents. Liu *et al.* [21] set an intent embedding for the instant intent of users' current browsing and employ the intent embedding for online recommender systems.

Different from the researches above, in this work, we propose to explicitly express the desire or demand of users to consume a kind of service or product as the consumption intention.

Heterogeneous Graph Recently, the heterogeneous graph, consisting of multiple types of nodes and/or links, has been leveraged as a powerful modeling method to fuse complex information and applied to many tasks such as recommender system [1, 8, 13, 26, 34, 40] and natural language processing [3, 20]. The usage of semi-supervised models [16, 30, 34, 38], where there exist some labeled vertices for representation learning enriches the representation ability in heterogeneous graphs. For example, GCN [16] proposed a localized graph convolutions to improve the performance in a classification task. GAT [30] used self-attention network for information propagation, which leverages a multi-head attention mechanism. GCN and GAT are popular architectures of the general graph networks and can be regarded as plug-in graph representation modules in heterogeneous graph, such as HetGNN [38], HAN [34]. In this work, we take advantage of heterogeneous graph to model different kinds of influencing factors for consumption intention prediction and learn representations for them.

Disentangled Representation Learning In consumption intention prediction task, the internal structure of consumption intention is complex, which makes it difficult to learn. Recently, some works [6, 35, 42] utilize the disentangled representation learning method for capturing different influencing factors, respectively. In this work, we target at exploiting the relations and structure of intentions rather than the influencing factors to learn disentangled representations for consumption intentions. The result demonstrates that there is indeed a clustering structure of which the intentions belonging to one cluster are highly correlated and similar, while there are great contrasts among the intentions belonging to different clusters for consumption intentions. Moreover, the disentangled representation is able to capture the inter-cluster difference and inner-cluster closeness.

7 CONCLUSION

In this work, we build a system consisting of consumption intention detection and prediction to understand the internal driven force of user consumption behaviors. The consumption intention detection method aims to explicitly reveal intentions of users' consumption from their reviews. For prediction, we design a dual-stage graph convolutional network model. Offline experimental results demonstrate that our prediction model can improve the performance by 6%-43% over state-of-the-art prediction models. Our system has also been successfully deployed on the Meituan's mobile App's Home Page. Results on the billion-scale industrial dataset further confirm the effectiveness of our system in practice.

8 ACKNOWLEDGMENTS

This work was supported in part by The National Key Research and Development Program of China under grant 2020AAA0106000.

REFERENCES

[1] Jianxin Chang, Chen Gao, Xiangnan He, Depeng Jin, and Yong Li. 2020. Bundle recommendation with graph convolutional networks. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1673–1676.

[2] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. In *WWW*. 1583–1592.

[3] Kunlong Chen, Weidi Xu, Xingyi Cheng, Zou Xiaochuan, Yuyu Zhang, Le Song, Taifeng Wang, Yuan Qi, and Wei Chu. 2020. Question Directed Graph Attention Network for Numerical Reasoning over Text. *arXiv preprint arXiv:2009.07448* (2020).

[4] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. *SIGKDD* (2016), 785–794.

[5] Tong Chen, Hongzhi Yin, Hongxu Chen, Rui Yan, Quoc Viet Hung Nguyen, and Xue Li. 2019. Air: Attentional intention-aware recommender systems. In *ICDE*. IEEE, 304–315.

[6] Weiguang Chen, Wenjun Jiang, Xueqi Li, Kenli Li, Albert Zomaya, and Guojun Wang. 2020. Semi-Disentangled Representation Learning in Recommendation System. *arXiv preprint arXiv:2010.13282* (2020).

[7] Zhiyong Cheng, Ying Ding, Xiangnan He, Lei Zhu, Xuemeng Song, and Mohan S Kankanhalli. 2018. Aˆ 3NCF: An Adaptive Aspect Attention Model for Rating Prediction.. In *IJCAI*. 3748–3754.

[8] Shaohua Fan, Junxiong Zhu, Xiaotian Han, Chuan Shi, Linmei Hu, Biyu Ma, and Yongliang Li. 2019. Metapath-guided heterogeneous graph neural network for intent recommendation. In *KDD*. 2478–2486.

[9] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247* (2017).

[10] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. *arXiv preprint arXiv:2002.02126* (2020).

[11] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *WWW*. 173–182.

[12] Bowen Jin, Chen Gao, Xiangnan He, Depeng Jin, and Yong Li. 2020. Multi-behavior recommendation with graph convolutional networks. In *SIGIR*. 659–668.

[13] Bowen Jin, Chen Gao, Xiangnan He, Depeng Jin, and Yong Li. 2020. Multi-behavior recommendation with graph convolutional networks. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 659–668.

[14] Manas R Joglekar, Cong Li, Mei Chen, Taibai Xu, Xiaoming Wang, Jay K Adams, Pranav Khaitan, Jiahui Liu, and Quoc V Le. 2020. Neural input search for large scale recommendation models. In *SIGKDD*. 2387–2397.

[15] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[16] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[17] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[18] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*. 1188–1196.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *ICCV*. 2980–2988.

[20] Hu Linmei, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. 2019. Heterogeneous graph attention networks for semi-supervised short text classification. In *(EMNLP-IJCNLP*. 4823–4832.

[21] Zhaoyang Liu, Haokun Chen, Fei Sun, Xu Xie, Jinyang Gao, Bolin Ding, and Yanyan Shen. 2020. Intent Preference Decoupling for User Representation on Online Recommender System. In *IJCAI*.

[22] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019. Learning disentangled representations for recommendation. *arXiv preprint arXiv:1910.14238* (2019).

[23] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *RecSys*. 165–172.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[25] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[26] Chuan Shi, Binbin Hu, Wayne Xin Zhao, and S Yu Philip. 2018. Heterogeneous information network embedding for recommendation. *TKDE* 31, 2 (2018), 357–370.

[27] Gábor J Székely, Maria L Rizzo, et al. 2009. Brownian distance covariance. *The annals of applied statistics* 3, 4 (2009), 1236–1265.

[28] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. 2007. Measuring and testing dependence by correlation of distances. *The annals of statistics* 35, 6 (2007), 2769–2794.

[29] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[30] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[31] Chenyang Wang, Weizhi Ma, Min Zhang, Chong Chen, Yiqun Liu, and Shaoping Ma. 2020. Toward Dynamic User Intention: Temporal Evolutionary Effects of Item Relations in Sequential Recommendation. *TOIS* 39, 2 (2020), 1–33.

[32] Shoujin Wang, Liang Hu, Yan Wang, Quan Z Sheng, Mehmet Orgun, and Longbing Cao. 2020. Intention nets: psychology-inspired user choice behavior modeling for next-basket prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 6259–6266.

[33] Shoujin Wang, Liang Hu, Yan Wang, Quan Z Sheng, Mehmet Orgun, and Longbing Cao. 2020. Intention2Basket: A Neural Intention-driven Approach for Dynamic Next-basket Planning. In *IJCAI*. 2333–2339.

[34] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous graph attention network. In *WWW*. 2022–2032.

[35] Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tong Xu, and Tat-Seng Chua. 2020. Disentangled Graph Collaborative Filtering. In *SIGIR*. 1001–1010.

[36] Su Yan, Xin Chen, Ran Huo, Xu Zhang, and Leyu Lin. 2020. Learning to Build User-tag Profile in Recommendation System. In *CIKM*. 2877–2884.

[37] Yue Yu, Tong Xia, Huandong Wang, Jie Feng, and Yong Li. 2020. Semantic-aware spatio-temporal app usage representation via graph convolutional network. *IMWUT* 4, 3 (2020), 1–24.

[38] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. 2019. Heterogeneous graph neural network. In *KDD*. 793–803.

[39] Chao Zhang, Keyang Zhang, Quan Yuan, Haoruo Peng, Yu Zheng, Tim Hanratty, Shaowen Wang, and Jiawei Han. 2017. Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In *WWW*. 361–370.

[40] Jun Zhang, Chen Gao, Depeng Jin, and Yong Li. 2021. Group-Buying Recommendation for Social E-Commerce. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE.

[41] Lei Zheng, Vahid Noroozi, and Philip S Yu. 2017. Joint deep modeling of users and items using reviews for recommendation. In *ACM WSDM*. 425–434.

[42] Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. 2021. Disentangling user interest and popularity bias for recommendation with causal embedding. In *TheWebConf*.

A APPENDIX FOR REPRODUCIBILITY

In the appendix, we provide the detailed descriptions of baseline models and metrics for the comparison of intention prediction. Then we introduce the parameter settings in the offline experiments. The hyper-parameter study and case study of our proposed consumption intention prediction model, GRIP is then given.

A.1 Baseline Model and Metric

Baseline Models We give the implementation details about all compared methods as follows.

- **XGBoost** [4]: It is a gradient boosted classification tree approach based on the gradient boosting machine. The weight for each factor is updated during training and the forecasting result is calculated as the weighted sum of all factor values.
- **DeepFM** [9]: It combines FM and deep neural network to low and high order feature combinations simultaneously. We utilize a fully-connected layer for prediction with the concatenation of all feature combinations as the input.
- **ReconEmbed** [39]: It is a general representation learning model for spatio-temporal activity prediction and samples random-walks from the graph to learn embedding vectors for nodes. We adapt this method to a graph with three kinds of nodes: time, location and intention. We utilize time and location embedding to generate context-aware intention representation and the prediction function of inner product to calculate the score for different consumption intentions.
- **GraphEmbed** [39]: It is a general representation learning model for spatio-temporal activity prediction and learns low-dimensional representations by preservingfirst and second-order proximities in the graph. We adapt this method to a graph with three kinds of nodes: time, location and intention and three types of edge types for the co-occurrence relations. We utilize time and location embedding to generate context-aware intention representation and the prediction function of inner product to calculate the score for different consumption intentions.
- **SA-GCN** [37]: It is a state-of-the-art method in spatio-temporal prediction and adopts Graph Convolutional Network (GCN) with meta path-based objective function to combine the structure of graph and the attribute of nodes. We adapt this method to the graph with three kinds of nodes: time, location and intention and three types of edge for the co-occurrence relations. We utilize time and location embedding to generate context-aware intention representation and the prediction function of inner product to calculate the score for different consumption intentions.
- **SA-LightGCN** [10]: It is a state-of-the-art method to learn node representations and capture interactions between nodes. We adapt this method to the graph with four kinds of nodes: user, item category, time and location and four types of edge for the co-occurrence relations and use pair-wise interaction for each single-typed edge with the transformation methods in LightGCN. We utilize user, time and location embedding to generate context-aware intention representation and the prediction function of inner product to calculate the score for different consumption intentions.
- **HAN** [34]: It is a general heterogeneous graph learning method to obtain representations of the nodes by aggregating information from the features of their neighboring nodes with different type-level attention and node-level attention. We adapt this method to the graph with four kinds of nodes: user, item category, time and location. We utilize user, time and location embedding to generate context-aware intention representation and the prediction function of inner product to calculate the score for different consumption intentions.
- **HetGNN** [38]: It is a stat-of-the-art heterogeneous graph learning method. For each node, itfirst aggregate the features of its neighbors of each type, and then aggregate the neighbors information of different types with attention. We adapt this method to the graph with four kinds of nodes: user, item category, time and location. We utilize user, time and location embedding to generate context-aware intention representation and the prediction function of inner product to calculate the score for different consumption intentions.

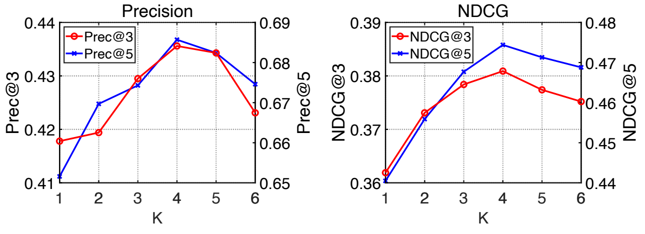
Metrics To evaluate the performance of prediction model, we use two widely used metrics, Precision and NDCG [12, 14, 36]. Specifically, for each case in test set, with the prediction scores for different consumption intentions, we form a ranked list by sorting these 19 consumption intentions according to their scores in a descending order, where we use $rank(s_g)$ to denote the position of the ground-truth intention in the ranking list. Then, we form a top-K intention list by picking the K top ranked intentions from the list. With this ranking list, the two metrics are defined as follows,

- **Prec@K.** If $rank(s_g) \leq K$, we have a *True Positive* case. To avoid biases from consumption intentions with a large amount, we computes the precision for each intention type and then compute the mean for all intention types.
- **NDCG@K.** If $rank(s_g) \leq K$, we have a hit and otherwise, we have a miss. It assigns higher scores to hits at a higher position in the ranking list, which emphasizes that for each test case, the right intention should be ranked as higher as possible.

A.2 Parameter Setting

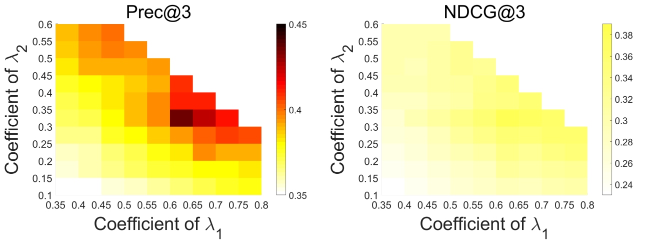
For above baselines, we explore hyper-parameters similarly as the original paper. For our dual graph neural network consumption intention prediction model, GRIP, we list the parameters in the user intrinsic preference modeling graph and the spatio-temporal context modeling graph respectively in following. For user intrinsic preference modeling graph, the node embedding size is set to 32, which is suitable to learn a strong representation embedding for users and item category [7, 12]. We try several layers for this graph andfind $L = 3$ to be the best. For spatio-temporal context modeling graph, the node embedding size is set to be 96 (where the embedding size for each type of feature vector is 32). The number of discrete time units in one day is set to 24. We try several layers for this graph andfind $L = 2$ to be the best.

We optimize all deep learning models with Adam optmize [15], having batch sizefixed to 2048. We search L_2 regularizer and learning rate in [10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}] and [0.001, 0.005, 0.01, 0.05, 0.1], respectively. Furthermore, we use early stop to detect over-fitting, and the training process will be stopped if Prec@3 on the validation set does not increase for 20 epochs.



(a) Precission (b) NDCG

Figure 7: Impact of number of disentangled parts of intention embeddings.



(a) Prec@3 (b) NDCG@3

Figure 8: Impact of coefficient in the joint loss.

A.3 Hyper-Parameter Study

We analyze the effects of two hyper-parameters used in our consumption intention prediction model, GRIP, including the number of disentangled independent parts of intention representations and the coefficient in the joint loss.

- **Number of disentangled independent parts of intention representations.** We vary it in [1,6] (where 1 means) and compare the intention performance of all metrics, which is shown in Figure 7. We can observe that even only one selected series to be used in the decoding stage, the forecasting performance can have a huge improvement, compared with not using OMN. Furthermore, within a certain range [1-6] (where K = 1 means not restricting the latent representation of intentions to be disentangled), the forecasting precision continuously increase as number of independent components gets larger but tends to get worse when it exceeds 4.
- **Coefficient in the joint loss.** It controls the weight of each component in GRIP and is a very important parameter of our method. Note that $\lambda_1 + \lambda_2 + \lambda_3 = 1$, we tune λ_1 in $[0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0, 45, 0.5, 0.55, 0.6]$ and λ_2 in $[0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8]$. Then we plot the performance of Acc@3 and NDCG@3 in Figure 8. When λ_1 and λ_2 are given, then value of λ_3 is determined. Therefore each block represents a setting of $(\lambda_1, \lambda_2, \lambda_3)$. And in these threefi gures, deeper blocks means better performance. The best performance of GRIP is achieved at the same setting, $(0.3, 0.6, 0.1)$ for both metrics, which represent a relatively large λ_2 which is the coefficient of prediction loss and a relatively low coefficient of disentangling loss. This is also consistent with thefi ndings in our ablation study. The result is shown in Figure 8.

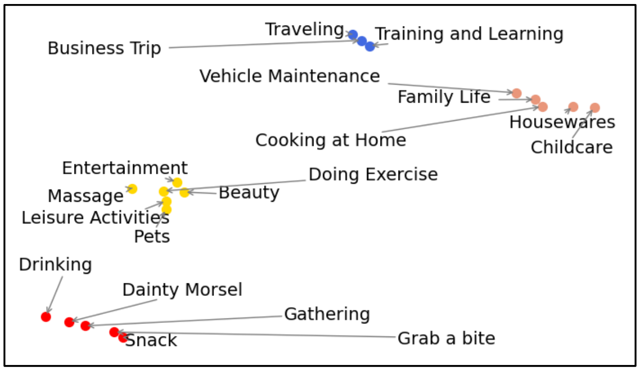


Figure 9: Visualization of the learned t-SNE transformed representations derived for disentangled embedding of consumption intentions.

A.4 Case Study

To analyze the effectiveness of disentangling and the representation of consumption intentions, we perform their visualization for embeddings showing in Figure 9. We set the number of disentangled independent parts $K = 4$ and record the output embeddings. Then, we project them to 2D for visualization using t-SNE [29] where the 2D result will be more clustered if the embedding vectors are more similar. As we can observe, the intentions are well separated into four clusters, which demonstrates that the intentions are not isolated but have certain relations, specifically, clustering structure. For example, thefi rst category, which contains Snack, Grab a bite, Gathering , Dainty Morsel and Drinking are intentions related to eating or drinking outside. This verifies that although the internal structure of consumption intention is complex, GRIP is capable of capturing their relations and structure with the help of disentangled representations of consumption intentions.