Probabilistic Attention for Sequential Recommendation

Yuli Liu
Qinghai University
Qinghai Provincial Key Laboratory of Media Integration Technology and Communication
Xining 810016, China
liuyuli012@gmail.com

Christian Walder
Google Research, Brain Team
Montreal, Canada
cwalder@google.com

Lexing Xie
Australian National University
Data61 CSIRO,
Canberra, Australia
lexing.xie@anu.edu.au

Yiqun Liu
Department of Computer Science and Technology, Tsinghua University
Zhongguancun Laboratory
Beijing 100084, China
yiqunliu@tsinghua.edu.cn

Abstract

Sequential Recommendation (SR) navigates users' dynamic preferences through modeling their historical interactions. The incorporation of the popular Transformer framework, which captures long relationships through pairwise dot products, has notably benefited SR. However, prevailing research in this domain faces three significant challenges: (i) Existing studies directly adopt the primary component of Transformer (*i.e.*, the self-attention mechanism), without a clear explanation or tailored definition for its specific role in SR; (ii) The predominant focus on pairwise computations overlooks the global context or relative prevalence of item pairs within the overall sequence; (iii) Transformer primarily pursues relevance-dominated relationships, neglecting another essential objective in recommendation, *i.e.*, diversity. In response, this work introduces a fresh perspective to elucidate the attention mechanism in SR. Here, attention is defined as dependency interactions among items, quantitatively determined under a global probabilistic model by observing the probabilities of corresponding item subsets. This viewpoint offers a precise and context-specific definition of attention, leading to the design of a distinctive attention mechanism tailored for SR. Specifically, we transmute the well-formulated global, repulsive interactions in Determinantal Point Processes (DPPs) to effectively model dependency interactions. Guided by the repulsive interactions, a theoretically and practically feasible DPP kernel is designed, enabling our attention mechanism to directly consider category/topic distribution for enhancing diversity. Consequently, the Probabilistic Attention mechanism (PAtt) for sequential recommendation is developed. Experimental results demonstrate the excellent scalability and adaptability of our attention mechanism, which significantly improves recommendation performance in terms of both relevance and diversity.

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671733

CCS Concepts

• **Information systems** → **Personalization**; **Learning to rank**.

Keywords

Attention Mechanism, Sequential Recommendation, DPPs

ACM Reference Format:
Yuli Liu, Christian Walder, Lexing Xie, and Yiqun Liu. 2024. Probabilistic Attention for Sequential Recommendation . In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671733

1 Introduction

Sequential recommendation (SR) has garnered considerable attention due to its alignment with real-world recommendation scenarios, where its primary objective is to provide users with content that caters to their interests through modeling their historical behavior. This heightened emphasis on SR has been substantially propelled by the success of the Transformer framework, which aims to capture long-range relationships utilizing pairwise dot products.

However, challenges still persist in this domain: (**Cl1**) While prior studies have explored various interpretations of attention concept in SR, including notions like influence [18, 23, 30], dependency [54, 62, 66], correlations [24, 67], and context [26, 50]. While these ambiguous interpretations are believed to significantly advance sequential recommendation (SR), there remains a lack of specific SR-tailored definitions and formalization. This lack of clarity impedes a profound understanding and potential improvements; (**Cl2**) Figure 1 reveals that the self-attention mechanism relies on pairwise dot products between a single item and all other items within the sequence to connect the entire sequence. This approach, in practice, fails to consider the relationships across different item pairs and how they are distributed in a specific manner within the sequence, *i.e.*, the **pairwise context**; (**Cl3**) In the field of recommender systems, it is common for data to exhibit sparsity [20, 31], often leading to short sequences. Traditional self-attention excels in modeling long-range relationships, which may not have a competitive advantage in addressing this inherent sparsity issue [65]; (**Cl4**) Existing research often allocates attention within a sequence based on item similarity, where higher similarity implies greater

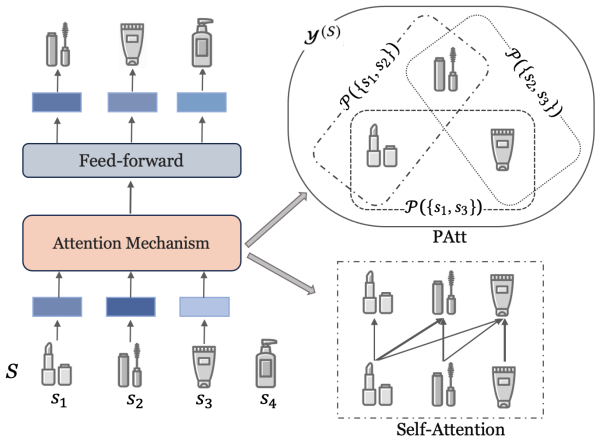


Figure 1: Comparative visualization of self-attention and probabilistic attention for sequential recommendation.

values of mutual attention weight. However, this approach could result in a system that tends to suggest monotonous items, thereby compromising diversity performance.

The recognition of challenges (**Cl1-Cl3**) has led us to explore a novel perspective, namely, the use of global probabilistic models to assign attention distribution within a sequence. This perspective offers several advantages: (i) It enables a formal definition of attention as a probability distribution, providing a robust foundation for the formulization and interpretation of attention concept; (ii) Global probabilistic models that derive probabilities (attentions) from a holistic viewpoint, extending beyond individual item pairs (**global context**); (iii) Probability modeling accommodates distributions over subsets of varying sizes. This enhances the ability of attention-based SR to effectively leverage limited information, which is crucial in addressing the frequently encountered challenge of sparse data in recommendation systems.

To explore this perspective, we consider the potential utility of Determinantal Point Processes (DPPs), which offer promising techniques for modeling global, repulsive interactions. By transmuting the well-established repulsive interactions, the **dependency interactions** among items can be readily derived, thus contributing to the explicit definition of attention in SR. In this definition, item attentions represent dependency interactions among items, where higher attention weights correspond to stronger dependencies across items. The degree of these interactions is qualified by the probability of the corresponding item subset being distributed as a DPP. Furthermore, in line with the critical characteristic of DPPs in measuring diversity, we can bring about the fourth advantage by introducing the concept of category/topic scope-related diversity into the attention distribution, thus mitigating the risk of monotonous recommendations (**Cl4**).

Based on this perspective, we develop Probabilistic Attention (PAtt) tailored for sequential recommendation. As Figure 1 depicts, the **mutual attention** between two items within a sequence S is formulated as the probability of drawing both of them from ground set $\mathcal{Y}^{(S)}$. This work aims to explore four research questions based on the novel perspective: (**RQ1**) Given that the attention degree is quantified based on the probability of a subset, and there are no restrictions on the number of items within a subset, can we infer that as the number of items involved in subsets increases, which implies more intricate dependency patterns are considered, the final recommendation performance improves? (**RQ2**) Considering our proposal of a substantially different attention definition from the original one, is it necessary for us to adhere to the complex setting of Query-Key-Value structures? (**RQ3**) What impact does existing self-attention of pairwise context have on diversity, and can we enhance recommendation diversity through the design of global probabilistic attention distribution? (**RQ4**) Can our PAtt be considered a more suitable attention mechanism for sparse environments?

The main contributions of this work are:

- We introduce a novel perspective of employing global probabilistic models to allocate attention distribution within a sequence. This approach allows us to mathematically and empirically define and elucidate the attention mechanism within the specific domain of sequential recommendation.
- Unlike conventional self-attention that requires the Query-Key-Value setting, our probabilistic attention mechanism relies solely on a **Sampler**. Remarkably, it achieves superior performance with significantly fewer parameters.
- Two crucial properties of our dependency-defined attention mechanism are validated: (i) Dependencies are mutual, without inherent ordering; (ii) As the number of items involved in dependency interactions increases, the recommendation performance improves.
- We incorporate the standard and category scope-related DPP kernels to formulate a novel kernel whose capability to promote recommendation diversity has been empirically and mathematically demonstrated.

2 Related Work

Self-attention Mechanism. The rational allocation of attention flow plays a crucial role in modeling long-range relationships for self-attention. Efforts in existing literature have primarily focused on four key approaches to achieve such rational allocation: (i) Some studies compute Wasserstein distances [11] or relative distances [16] between elements within a sequence, and utilize these distances as attention weights; (ii) Some approaches treat the attention graph as a flow network, allowing for the derivation of attention flow [1, 63]; (iii) In certain methods, softmax is considered a kernel and connected with other kernels, such as the Gaussian kernel, as a means to calculate attention weight matrices; (iv) Hybrid attention [42, 57] and sharing attention mechanisms [52] are employed to more efficiently compute the attention flow. Furthermore, given that self-attention inherently lacks explicit modeling of relative or absolute position information, the introduction of additional information during attention weight computation is a common practice. This additional information includes factors such as edge information [41], time intervals [30], and contextual signals [37, 59].

It is worth noting that the current approaches for calculating attention weights predominantly operate in a pairwise context. To date, there is a lack of research employing global probabilistic models to allocate attention flow in a more holistic manner.

Sequential Recommendation Models. The implementation of SR primarily relies on sequential models to predict the relevance between the previous sequence and the target items. An early method used for SR is the Markov Chain (MC) [8, 20], which operates under

the assumption that a user's next action can be predicted based on their recent actions. SR models that employ Convolutional neural network (CNN) treat the sequence embedding matrix as an image and apply convolutional operators to it [44, 58]. On the other hand, Recurrent neural network (RNN)-based methods enable the capture of longer-term semantics and take the order of the sequence into account, leading to performance improvements [3, 40, 56]. Currently, self-attention based models have taken a dominant role in the field of SR. The pioneering work that introduces self-attention into sequential recommendation is [26]. Subsequently, numerous efforts have been made to incorporate sequence order information [30, 33] and context [23, 48] into the attention mechanism to enhance performance. Furthermore, the integration of self-attention with other sequential models, including contrastive learning [39, 53, 60], graph neural networks [22, 55, 61], CNN [6, 25], and RNN [32, 51], has been frequently adopted to utilize each other's strengths in the pursuit of improved SR performance.

DPP for Recommendation. DPPs are elegant probabilistic models known for efficiently capturing global negative correlations through algorithms for tasks like sampling and marginalization. Leveraging DPPs to diversify recommendations often involves designing efficient maximum a posteriori (MAP) methods. Subsets of items, with the potential to spark user interest, are generated based on the quality *vs.* diversity decomposition of the DPP kernel through these methods [17, 49]. Learning DPP kernel parameters from historical interactions and using them as item representations is also a common approach [13, 47]. Furthermore, DPPs are frequently combined with other popular models, such as Generative Adversarial Networks (GANs) [49], bilateral branch networks [34], reinforcement learning [35], and CNN [2], among others. These collaborations aim to enhance the quality of recommendations by incorporating diverse perspectives and methodologies.

Currently, there has been no attempt in the literature to interpret and apply the DPP probability distribution as attention flow.

3 Methodology

In this section, we first provide relevant preliminaries and then introduce our probabilistic attention.

3.1 Preliminaries

The relevant preliminaries include: SR problem formulation, SR-tailored self-attention mechanism, and the standard DPP model.

3.1.1 Problem Formulation. Let sets of users and items be denoted by $\mathcal{U}$ and $\mathcal{V}$, with $|\mathcal{U}|$ and $|\mathcal{V}|$ indicating their respective sizes. In sequential recommendation, user u's chronologically ordered interaction sequence is denoted $\mathcal{S}_u = [v_1^u, v_2^u, \ldots, v_{|\mathcal{S}_u|}^u]$, where v_t^u specifies the item interacted with at time step t and $|\mathcal{S}_u|$ indicates the sequence length. The objective of SR is to match the next item in a sequence, formally, $p(v_{|\mathcal{S}^u|+1}^u = v|\mathcal{S}^u)$, by recommending a top-N list of items.

3.1.2 Self-attention Mechanism. Since a new attention mechanism is proposed as our sequence encoder, we introduce traditional self-attention before presenting ours. The self-attention mechanism inherently presumes that while items in sequences are correlated, they bear varied significance to items at different sequence positions. For a user's action sequence $\mathcal{S}^u$ and a maximal sequence length T, the sequence is either truncated by eliminating initial items if $|\mathcal{S}^u| > n$ or padded with zeros, resulting in $S = (s_1, s_2, \ldots, s_T)$. We define an item embedding matrix $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}|\times d}$, where d indicates the latent dimensions number. A trainable positional embedding $\mathbf{P} \in \mathbb{R}^{T\times d}$ is appended to the sequence embedding matrix as:

$$\mathbf{M}_{\mathcal{S}^u} = [\mathbf{e}_{s_1} + \mathbf{p}_{s_1}, \mathbf{e}_{s_2} + \mathbf{p}_{s_2}, \ldots, \mathbf{e}_{s_n} + \mathbf{p}_{s_n}]. \tag{1}$$

Utilizing dot-product calculations, self-attention discerns correlations between sequence items as follows:

$$\mathrm{SA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}, \tag{2}$$

with $\mathbf{Q} = \mathbf{M}_{\mathcal{S}^u}\mathbf{W}^{\mathbf{Q}}$, $\mathbf{K} = \mathbf{M}_{\mathcal{S}^u}\mathbf{W}^{\mathbf{K}}$, and $\mathbf{V} = \mathbf{M}_{\mathcal{S}^u}\mathbf{W}^{\mathbf{V}}$. In this Query-Key-Value attention mechanism, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are derived from the input, through learnable weight matrices. They serve to establish attention scores by comparing the query with all keys, and then determine the weighting and subsequent aggregation of values, to form a contextually enriched output. Other Transformer components, such as the point-wise feed-forward network, residual connection, and layer normalization, are also usually leveraged in self-attention-based SR models [11, 26].

3.1.3 Standard DPP. DPP measures set diversity by describing the probability for all subsets of the ground set [28]. Given a discrete set $\mathcal{Y} = \{1, 2, \ldots, M\}$ (*i.e.*, item set $\mathcal{V}$ in this work), a DPP $\mathcal{P}$ is a probability measure on $2^{\mathcal{Y}}$, the set of all subsets of $\mathcal{Y}$. When $\mathcal{P}$ gives nonzero probability to the empty set, there exists a kernel $\mathbf{L} \in \mathbb{R}^{M\times M}$ such that for every subset $Y \subseteq \mathcal{Y}$, the probability of Y is given by

$$\mathcal{P}(Y) = \frac{\det(\mathbf{L}_Y)}{\det(\mathbf{L} + \mathbf{I})}, \tag{3}$$

where $\mathbf{L}$ is a real, positive semi-definite matrix indexed by elements of Y, and $\mathbf{I}$ is an $M \times M$ identity matrix.

3.2 Probabilistic Attention

In a standard DPP (Equation 3), every subset of $\mathcal{Y}$ is assigned a probability, even empty set and the entire $\mathcal{Y}$. This standard DPP is not desirable in practical scenarios that need fixed-size result arrays [28]. Given this situation, an extension of DPP, *i.e.*, k-DPP [27], has been proposed, which conditions a DPP on the cardinality k of the random set.

3.2.1 Sequence-specified k-DPP. Compared to a standard DPP, where every k-set Y competes with all other subsets of $\mathcal{Y}$, in a k-DPP it competes only with sets of the same cardinality. In this work, k-DPP is explored to shape attention distribution, primarily for the following reasons: (i) As depicted in Figure 1, we interpret attention as the intensity of dependency among items. Therefore, considering the probability of a single item or an empty set is not meaningful and, conversely, would distract the global allocation of probabilities; (ii) Measuring the dependency among items within subsets of the same size and utilizing this to allocate attention is a rational and equitable approach.

If we directly model a k-DPP model over the ground set of the entire item set (*i.e.*, $\mathcal{V}$), and then learn to distribute probabilistic attention, two issues will arise: (i) The computational burden is pronounced due to the necessity of a DPP kernel $\mathbf{L} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$; (ii) The attention flow is incapable of reflecting the uniqueness embedded in varied user behavior sequences. Considering that the

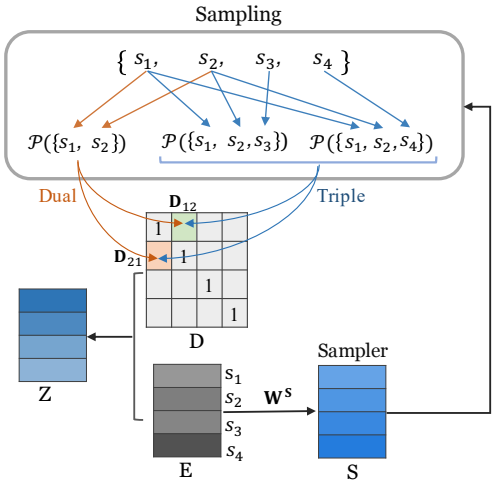


Figure 2: Illustration of the proposed attention mechanism.

global context of attention involves allocating attention in alignment with the pattern throughout the entire sequence, we therefore build a k-DPP for each sequence on the sequence-specified kernel, denoted as $\mathbf{L}^{(S)}$, which is defined over ground set $\mathcal{Y}^{(S)}$ composed of sequence items. It means that our global probabilistic k-DPP model selectively learns the unique attention distribution for each sequence. Formally, the k-DPP $\mathcal{P}^k_{\mathbf{L}^{(S)}}$ on $\mathbf{L}^{(S)}$ gives probabilities

$$\mathcal{P}^k_{\mathbf{L}^{(S)}}(Y) = \frac{\det\left(\mathbf{L}^{(S)}_Y\right)}{\sum_{|Y'|=k}\det\left(\mathbf{L}^{(S)}_{Y'}\right)}, \tag{4}$$

where $|Y| = k$. The normalization constant in the denominator indicates the summation over all subsets of size k, presenting a distinct contrast to the standard DPP where the denominator (of Equation 3) sums over all possible subsets, regardless of their size. This normalization constant enables us to associate the subset Y and all other k-subsets within a k-DPP distribution (global context), which is formulated as

$$Z_k = \sum_{|Y'|=k}\det\left(\mathbf{L}^{(S)}_{Y'}\right) = e_k\left(\lambda_1, \lambda_2, \ldots, \lambda_{|\mathcal{S}_u|}\right). \tag{5}$$

Here $\lambda_1, \lambda_2, \ldots, \lambda_{|\mathcal{S}_u|}$ are the eigenvalues of $\mathbf{L}^{(S)}$ with $|\mathcal{S}_u|$ items. The recursive algorithm given in [14, 28] can be used directly to calculate the kth elementary symmetric polynomial e_k on $\lambda_1, \lambda_2 \ldots, \lambda_{|\mathcal{S}_u|}$, which runs in $O(|\mathcal{S}_u|k)$ time. Thus, we can efficiently normalize a tailored k-DPP with a sequence ground set in $O(|s_u|k)$ time to model global repulsion interactions.

3.2.2 Dependency Interactions. This k-DPP probability serves as a quantitative measure of repulsion interactions (diversity) among items in subset Y [15, 27, 28]. In a novel approach, transmuting this repulsion measure by taking its negative can mathematically symbolize a converse concept; that is, it suggests a degree of dependency among items within a set. For instance, given two subsets, A and B, each containing k items, we examine their respective k-DPP probabilities, $\mathcal{P}^k_{\mathbf{L}^{(S)}}(A)$ and $\mathcal{P}^k_{\mathbf{L}^{(S)}}(B)$. A condition $\mathcal{P}^k_{\mathbf{L}^{(S)}}(A) > \mathcal{P}^k_{\mathbf{L}^{(S)}}(B)$ suggests that subset A demonstrates stronger repulsion interactions among its items compared to subset B. In contrast, considering the negation of probabilities, $-\mathcal{P}^k_{\mathbf{L}^{(S)}}(B) > -\mathcal{P}^k_{\mathbf{L}^{(S)}}(A)$, we infer that items in subset B exhibit stronger **dependency interactions** than those in subset A. To articulate the distinction between repulsion and dependency, we introduce $\mathcal{D}^{(S)}_A$, signifying the dependency degree among items in subset A under sequence context S, defined as $\mathcal{D}^{(S)}_A = -\mathcal{P}^k_{\mathbf{L}^{(S)}}(A)$. When k is set to 2, the relative strength of repulsion between any two items under the global measure $\mathbf{L}^{(S)}$ can be determined, which corresponds to the probability of being sampled by the *2*-DPP model. Thus, the necessity to quantify the dependency interactions between any paired items in a sequence becomes a matter of the sampling process. Specifically, the transmuted probability of sampling any two items is quantified as dependency interactions between them, as illustrated in the dual sampling of Figure 2. In such a scenario, we obtain a **dual dependency** defined within a global probability context. Similarly, we can obtain triple, or even quadruple, dependency.

3.2.3 Probabilistic Attention Mechanism. Reflecting upon Equation 2 within the self-attention mechanism, the attention weight matrix is obtained by performing a softmax transformation on the attention weight vector of an item with respect to all other items in the sequence. Fundamentally, this transformation is still in pairwise context, as it allocates attentions of one item onto all others, rather than deriving the relative relationships of all pairs of items from a global measure. According to Equation 2, the attention weight of traditional self-attention is calculated as:

$$\mathbf{A}_{rt} = \mathbf{Q}_r\mathbf{K}_t^\top/\sqrt{d}, \tag{6}$$

where $\mathbf{A}_{rt}$ is typically interpreted as the influence (weight) of position r on position t in the sequence [46] or the relationship (similarity) between positions r and t [41]. However, this pairwise context approach is not optimized for sequential recommendation tasks. In the previous discussion, we drew upon the motivations and algorithms of DPP to define the dependency interactions. Here, we formulate the **dependency attention** weight based on the degree of dual dependency interactions using a simple log-linear model,

$$\mathbf{D}_{rt} = \exp\left(-\lambda\mathcal{P}^2_{\mathbf{L}^{(S)}}(\{v_r, v_t\})\right), \tag{7}$$

where λ governs the range of the dependency degree. $\mathcal{P}^2_{\mathbf{L}^{(S)}}$ denotes a k-DPP distribution ($k = 2$) with kernel $\mathbf{L}^{(S)}$ over 2-sized subsets of a sequence. $\mathcal{P}^2_{\mathbf{L}^{(S)}}(\{v_r, v_t\})$ signifies the probability of items v_r and v_t ($r \neq t$) being selected from $\mathcal{Y}^{(S)}$ as a 2-DPP under global measure (global context). This dependency attention approach we proposed differs notably from the self-attention weight, which embodies the mutual dependency between any two items. This is driven by the desire to model dependency interactions from training sequences, hence capturing dependency relationships between previous actions and the next item, which is paramount to the realization of sequential recommendation [4, 7, 56].

While we have defined this dependency attention measurement, an important remaining question is how to devise our DPP kernel $\mathbf{L}^{(S)} \in \mathbb{R}^{T\times T}$, where T is the maximum sequence length. To bridge the input from the input layer with the learnable kernel parameters, we employ a low-rank factorization of the $T \times T$ $\mathbf{L}^{(S)}$ matrix:

$$\mathbf{L}^{(S)} = \mathbf{S}\mathbf{S}^\top. \tag{8}$$

Here, $\mathbf{S} \in \mathbb{R}^{T\times d}$ is transformed by $\mathbf{S} = \mathbf{M}_{\mathcal{S}^u}\mathbf{W}^{\mathbf{S}}$. $\mathbf{W}^{\mathbf{S}}$ is a trainable projection matrix, serving essentially as the learnable parameters

in kernel $\mathbf{L}^{(S)}$. Because our attention intuitively measures the dependency interactions between items, we seek to preserve these relationships as much as possible. Consequently, we discard the transformed Value matrix from the self-attention mechanism, directly applying the attention to the item's input representation $\mathbf{M}_{\mathcal{S}^u}$. In this setting, only a transformed representation matrix $\mathbf{S}$ is required to capture contextual information in a sequence. In view of the DPP sampling operations, we denote $\mathbf{S}$ as **Sampler**. Thus, our probabilistic attention based on the dependency measurement is formulated as

$$\mathbf{Z} = \mathrm{PAtt}(\mathbf{S}) = \mathbf{D}\mathbf{M}_{\mathcal{S}^u}. \tag{9}$$

$\mathbf{D}$ denotes the attention weight matrix of our proposed PAtt for sequence $\mathcal{S}^u$, derived by (i) computing the exponential values of the dependency interaction degree across entire 2-sized subsets, and (ii) allocating all these values to corresponding matrix entries, based on the positions of paired items in $\mathcal{S}^u$. $\mathbf{Z}$ is the output of our attention. It is worth noting that within the DPP subset, there is inherently no ordering, implying that the dependency interactions are mutual. That is, corresponding dependency attention weight matrix exhibits a symmetrical relationship as illustrated in Figure 2. However, due to the causality nature of sequences and to avoid considering future information [26, 46] in prediction, we only consider the lower triangular part of $\mathbf{D}$ matrix in practice. For the dependency of an item with itself, we consider it absolute and hence set the diagonal elements of the attention weight matrix to 1. We can find that PAtt breaks away from the intricate Query-Key-Value mechanism of self-attention, necessitating only a **Sampler**.

3.2.4 Diversity-aware Attention. In recommendation systems, the diversity metric is utilized to assess the scope of categories or topics within a recommended sequence; a broader scope indicates more diverse recommendations [49, 64]. However, PAtt introduced above and existing self-attention based models [11, 26, 53] do not account for diversity, which might affect the exhibition of diversified results under the guidance of relevance sequences. Therefore, in this section, we endeavor to directly integrate the concept of diversity into the attention mechanism, developing a diversity-aware probabilistic attention model, termed DPAtt.

To achieve this, a new DPP kernel form, denoted as $\mathbf{T}^{(S)}$, integrates the standard DPP kernel $\mathbf{L}^{(S)}$ and a category scope-related kernel $\mathbf{C}$. As previously discussed, $\mathbf{L}^{(S)}$ is given over the sequence-specified ground set because we aim to model the dependency interactions of user sequences, thus inheriting relevance. The concept of category scope-related diversity does not bear relevance or personalization; therefore, it is necessary to derive $\mathbf{C} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ over the entire ground set $\mathcal{V}$. To reduce the computational complexity of calculating a $|\mathcal{V}| \times |\mathcal{V}|$ matrix, the $\mathbf{C}$ kernel is represented using a low-rank factorization $\mathbf{C} = \mathbf{V}^\top\mathbf{V}$ following [13]. We use diversified item sets (subsets that have a broad coverage) from users' historical interactions as ground truth sets for learning $\mathbf{C}$ [36, 49]. In this way, $\mathbf{C}$ is associated with the item category scope, as a subset with more categories has a higher possibility of being selected under a DPP compared to the less diverse ones. Formally, $\mathbf{C}$ is learned following the objective function,

$$\mathcal{J} = \sum_{(A^+,A^-)\in\mathcal{A}} \log\det\left(\mathbf{C}_{A^+}\right) - \log\det\left(\mathbf{C}_{A^-}\right). \tag{10}$$

Here, $A^{(+)}$ is an observed diverse set and $A^{(-)}$ represents the set that contains negative items, and $\mathcal{A}$ denotes the set of paired sets used for training. $\mathbf{C}$ is not related to sequence, as it is only the category scope-related kernel.

$\mathbf{T}^{(S)}$ is then given making use of the following relationship:

$$\mathbf{T}^{(S)} = \mathbf{L}^{(S)}\mathbf{C}^{-1}_{\mathcal{S}_u}. \tag{11}$$

$\mathbf{C}^{-1}_{\mathcal{S}_u}$ denotes the reverse of matrix $\mathbf{C}_{\mathcal{S}_u}$ that is a sub-matrix of $\mathbf{C}$ indexed by items in sequence $\mathcal{S}_u$. Given a DPP distribution formulated with kernel $\mathbf{T}^{(S)}$, a crucial property is introduced: an increase in category scope reduces the probabilities of related subsets, and subsequently enhances attention weights derived from the negation of these probabilities Equation 7. This implies that the composite kernel $\mathbf{T}^{(S)}$, by influencing the DPP probability distribution, can strike a balance between dependency interactions and diversity, ensuring subsets with strong dependency relationships and ample diversity receive notable attention weights. The proof of this proposed proposition is provided in Appendix A.1, and its validation is discussed in Section 4.2.

3.2.5 Triple Dependency. Given that we define the attention weight as the dependency interactions within a subset, an intuitive concept emerges: the dependency interactions, when formulated to involve more items, should enable our attention mechanism to harness better modeling capability, thereby enhancing SR performance. This is because, in the field of SR, a model enhances its performance precisely by augmenting its capacity to capture and model complex dependency relationships among items. [10, 36]. This perspective also reveals a limitation of self-attention, in that the calculation of attention weight only involves two items, neglecting the consideration of combinations involving three or more items. The previously introduced PAtt only considers dependency between two items, which may not be sufficient. In this section, we analyze from a probabilistic perspective and design a method that introduces richer dependencies into the attention distribution.

Initially, we consider the situation of triple items. By setting $k = 3$, we can easily obtain the DPP probabilities of subsets with three items, denoted as $\mathcal{P}^3_{\mathbf{L}^{(S)}}$. Subsequently, the core challenge lies in how to apply these probabilities to the computation of dependency attention weight matrix. By perceiving the marginal probability of dual items as a marginalization over a joint probability, we can derive a richer dependency mapping, formally,

$$\mathcal{P}^2_{\mathbf{L}^{(S)}}\left(\{v_r, v_t\}\right) = \sum_x \mathcal{P}^3_{\mathbf{L}^{(S)}}\left(\{v_r, v_t, v_x\}\right). \tag{12}$$

Here, $(\{v_r, v_t, v_x\})$ represents any 3-subset in a sequence, which includes v_r and v_t. As demonstrated in Figure 2, since this approach formulates DPP of 3-sized item subsets and ultimately models **triple dependency** interactions, we refer to it as triple PAtt, abbreviated as PAtt^3. For consistency in representation, the previously discussed dual PAtt is denoted as PAtt^2. We can also derive the diversity-aware attention corresponding to triple PAtt, represented as DPAtt^3. With this explicit formulation of triple dependency, we can introduce more dependency learning to capture more complex behavior patterns. Moreover, referring to this method, we can also introduce dependency interactions among even more items, illustrating the extensibility of our attention method.

It is noteworthy that, analogous to the self-attention mechanism, our probabilistic attention can also be configured with multiple heads and attention blocks.

3.3 PAtt Calculation

3.3.1 Matrix Form Calculation. We can directly enumerate all combinations of T items of a sequence taken k at a time, resulting in a total of $C(T,k)$ combinations. These combinations, which store positions of items within a sequence and thus remain invariant throughout all training sequences, signify complete subsets with cardinality k. Every combination comprising k sequence positions forms a k-subset. Given a batch of training sequence inputs, we can acquire a batch of DPP kernels with shape $(\text{batch\_size}, T, T)$ referring to Equation 8. The subsequent step involves slicing corresponding sub-matrices from each kernel according to k positions in the combinations, *i.e.*, obtaining each $\mathbf{L}_Y^{(S)}$ associated with all subsets within every sequence ground set $\mathcal{Y}^{(S)}$, with shape $(\text{batch\_size}, C(T,k), k, k)$. Calculating the determinant of each sub-matrix in the batch-sized square matrices yields the corresponding $\det(\mathbf{L}_Y)$ values for all subsets, with the resultant shape being $(\text{batch\_size}, C(T,k))$. As we have selected all k-subsets from each sequence ground set, directly summing the $\det(\cdot)$ results for the $C(T,k)$ subsets provides the normalization constant. Consequently, it is not required to employ recursive algorithms from Equation 5 to ascertain the normalization constant. Upon obtaining the normalized DPP probabilities, we can arrange the batch-sized $C(T,k)$ probabilities into a batch of identity matrices whose diagonal entries are 1 and the other entries are 0. Firstly, we secure the lower triangular matrix indices of the identity matrix, and subsequently populate the relevant entries with corresponding probabilities. Specifically, the positions of paired items (within each subset) in a sequence directly correspond to the entry indices in the identity matrix. Through the aforementioned methodology, we can accomplish the derivation of the probabilistic attention matrix $\mathbf{D}$ at the batch level.

3.3.2 Complexity Analysis. Our PAtt, analogous to existing self-attention based SR models like SASRec [26] and STOSA [11], also employs transformer structures, including the embedding layer, point-wise feed-forward network, residual connection, and layer normalization. This means that PAtt overall requires fewer parameters (reducing the parameter count by $2 \times d^2$) compared to other related models, due to our exclusive need for a single transformed Sampler instead of using all Query, Key, and Value. In regards to time complexity, while the dominant term in SASRec and STOSA resides in the softmax computation, *i.e.*, $O(T^2 d)$, our method is predominantly influenced by the complexity associated with the calculation of DPP probabilities for subsets. The complexity for a k-sized subset equates to $O(k^3)$. Considering that we typically require a minor k value, such as 2 or 3, the computational complexity is not high. However, variations in sequence length can have a significant impact. If the maximum length, T, is set relatively large, the number of k-subsets could potentially prevail the dominant position in complexity calculation. Considering the state-of-the-art SR work and accounting for the sparsity of SR, the maximum length is usually set to 20, 30, or 50 [30, 31, 53]. Thus, globally, our method exhibits complexity comparable to traditional self-attention mechanisms.

Furthermore, when $k = 3$ and presuming $T = 30$, we identify $C(30,3)$ (4060) 3-subsets. Each attention weight $\mathbf{D}_{rt}$ is defined by the summation of joint probabilities across 28 3-sized subsets that contain items v_r and v_t, thus presenting a steep complexity of $C(30,3) \times 3^3$, surpassing $30^2 \times 32$ (with $d = 32$). To mitigate this issue in practical computation, we do not compute all the 3-subsets corresponding to an attention weight $\mathbf{D}_{rt}$. Rather, we randomly select 4 of the 28 combinations, meaning that the marginalization of an attention weight transpires over the joint probabilities of 4 arbitrary subsets containing v_r and v_t. This culminates in a computational complexity of: $C(30,2) \times 4 \times 3^3$, implies that, among the total $C(30,2)$ 2-subsets, each one is composed of the probabilities of four 3-subsets. This approach yields a time complexity comparable to the dominant term of the traditional self-attention-based SR model SASRec.

Table 1: Statistics of the datasets.

Dataset	#Users	#Items	#Interactions	#Categories	Density
Beauty	8,159	5,862	0.1M	213	0.02%
CDs	17,052	35,118	0.5M	340	0.08%
Anime	73,516	12,294	1.0M	43	0.1%
ML-1M	6,040	3,416	1.0M	18	4.8%

4 Experiments

In this section, comprehensive experimental comparisons and analyses are conducted to address the four research questions (**RQ1**-**4**) posed in Section 1. The code and datasets relevant to this study are available at https://github.com/l-lyl/PAtt.

4.1 Experimental Settings

4.1.1 Datasets. We select four widely-utilized real-world datasets to evaluate the proposed probabilistic attention, which is offered in two forms: one relevance-focused version (PAtt^2 and PAtt^3) and the other being diversity-aware (DPAtt^2 and DPAtt^3). Each dataset presents notably different category numbers and matrix densities concerning implicit interactions. The **Amazon-review**[1] dataset [21], encompasses a vast corpus of product ratings, from which we specifically select **Beauty** and **CDs** products. Both **Anime**[2] and **ML-1M**[3] [19] datasets offer ratings on anime and movies, containing 43 and 18 categories, respectively. In alignment with conventional practices as seen in [26, 53, 67], all numeric ratings are transmuted into implicit feedback marked as 1. To refine the data, we exclude users and items that have fewer than 10 interactions. Statistics for all datasets are provided in Table 1.

4.1.2 Evaluation Metrics. We employ the widely-utilized leave-one-out evaluation strategy [26, 53]. For every user, the last item with which they interacted is held out for testing, and the penultimate item is utilized for validation, with the remainder used for training. Our evaluation of PAtt and baselines hinges on three strands of metrics: (*i*) two accuracy-oriented metrics, namely, Normalized Discounted Cumulative Gain, *i.e.*, NDCG (**ND**) and Recall (**Re**), (*ii*) two prevalent and intuitive **diversity** metrics, Category Coverage (**CC**) [38, 49] and intra-list distance (**ILD**) [38, 64], and (*iii*) a **harmonic** F-score metric (**F1**) which harmonizes quality (accuracy) and diversity [9, 34]. We select $N \in \{5, 20\}$ for evaluation, allowing us to assess the model's robustness through both small and large retrieval sizes, covering varied retrieval contexts.

[1] http://jmcauley.ucsd.edu/data/amazon/
[2] https://www.kaggle.com/CooperUnion/anime-recommendations-database
[3] https://grouplens.org/datasets/movielens/1m/

4.1.3 Baselines. Nine prevalent state-of-the-art baselines in the SR field are selected for comparison with four PAtt based models. They fall into four groups: (i) CNN based model, **Caser** [2]; (ii) GNN based model, **GC-SAN** [55]; (iii) models that utilize contrastive learning (CL), **CL4SRec** [53] and **DuoRec** [39]; (iv) self-attention based SR models, **SASRec** [26], BERT4Rec (**BERT**) [43], **STOSA** [11], MOJITO (**MOJ**) [45], and **STRec** [29]; Further information on baselines and implementation details is given in Appendix A.2.

4.2 Comparisons

The overall performance comparison between four PAtt based models and nine baselines *w.r.t.* three types of metrics on four datasets is presented in Table 2, in which the best performance among all methods is highlighted in **bold**, while the highest-performing baseline is distinguished with an underline. Additionally, the percentage (denoted by %) in parentheses () represents the improvement of our specific model over the top-performing baseline. Based on the comparison, we can make the following main observations:

- The recently proposed methods based on self-attention (STRec and MOJ), as well as DuoRec based on CL, dominate the accuracy-related metrics among all baselines. Four methods based on PAtt we proposed show significant advantages compared to these three competitive baselines, with improvements exceeding 5% in most cases. This demonstrates the uniqueness and effectiveness of our probabilistic attention.
- The PAtt^3 method, which introduces triple dependency interactions into attention, performs better than PAtt^2, which directly considers the paired items' dependency degree as attention weight. This result can answer two research questions (**RQ1** and **RQ2**), because: (i) We define attention as dependency interactions, making the attention weight not limited to calculations involving only two items, but can introduce more items into the dependency attention measurement. Combined with corresponding results, it can be concluded that the more comprehensive the dependency interactions considered during the learning of attention distribution, the stronger the predictive capacity of the model (**RQ1**); (ii) Both PAtt^3 and PAtt^2 utilize the Sampler setting and yield favorable outcomes, signifying that our probabilistic attention can deviate from the Query-Key-Value mechanism (**RQ2**). In addition, the improvement brought about by triple dependency also demonstrates the rationality of our definition of attention as dependency interactions, as in extensive SR studies [10, 36], capturing more dependencies means better performance.
- Compared to SASRec and STRec baselines, which apply the traditional self-attention mechanism, our basic PAtt^3 and PAtt^2 approaches have shown obvious advantages in terms of both diversity and accuracy. This can answer **RQ3** and **RQ2** to some extent. The deficiency of SASRec and STRec in diversity metrics suggests that employing a similarity metric within a relevance-guided pairwise context prompts the learned attention distribution to allocate higher attention to analogous future items. Consequently, this amplifies the prevalence of similar items in predictions, thereby diminishing diversity (**RQ3**). Although the diversity concept is not considered in PAtt^3 and PAtt^2, they exhibit notable performance in diversity metrics. This can be attributed to our probabilistic attention, which is formulated by articulating mutual dependency and coordinating the probability distribution across all sequence subsets. This approach facilitates a more thoughtful allocation of mutual attention at the global context level. In addition, this intuitive comparison also validates the effectiveness of our unique Sampler setting (**RQ2**).
- After incorporating the diversity-aware kernel, DPAtt^2 and DPAtt^3 display a substantial improvement in diversity metrics while also delivering strong performance in relevance. This outcome provides a comprehensive answer to **RQ3**, demonstrating that our proposed model is capable of embedding the concept of category scope-related diversity into the attention distribution, thus effectively augmenting diversity. This also demonstrates the efficacy of our proposed method of constructing a new type of DPP kernel, $\mathbf{T}^{(S)}$, by integrating standard and category scope-related kernels, further validating the proof of the crucial property of the integrated kernel.
- When comparing the improvements achieved by PAtt methods across various datasets, the improvements are notably more pronounced on datasets with higher sparsity (*e.g.*, Beauty and CDs). This indicates that PAtt, through the incorporation of dependency interactions in global context, is capable of thoroughly modeling previous information in the presence of data sparsity. By extracting dependency interactions that fundamentally impact SR, PAtt demonstrates its adaptability to situations characterized by sparse data (**RQ4**).

As we focus on proposing a new attentive mechanism for SR from a probabilistic perspective, instead of balancing the diversity and accuracy of recommendations employing the quality *vs.* diversity decomposition of DPP kernel. We therefore omit the DPP-based diversity-promoting recommendation models in Table 2. To further demonstrate our proposed model's performance in balancing diversity and quality, three DPP-based SR models are employed for comparison. Due to space constraints, the details of the implementations and results comparison are provided in Appendix Table 3, where the advantages of our newly proposed kernel $\mathbf{T}^{(S)}$, which, at the model level, comprehensively considers diversity and quality (proved in Appendix A.1), are further examined.

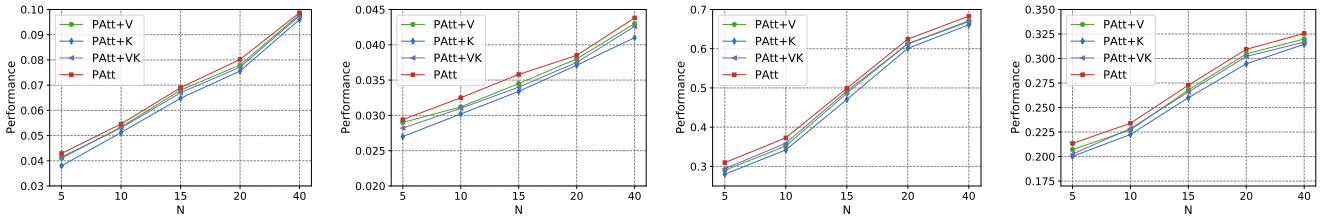
4.3 Property Analysis

In this section, we delve further into research questions intimately related to the characteristics of our probabilistic attention, particularly *w.r.t.* attention setting (**RQ2**) and sparse data problem (**RQ4**).

Figure 3 is utilized to further respond to **RQ2**. Three basic variants of PAtt (dual dependency) are designed in reference to the traditional Query-Key-Value attention mechanism, where the symbols "+K", "+V" and "+VK", respectively denote the addition of transformed key, value and both. Within these variants, we regard Sampler as Query. The newly introduced Key is intended to enrich the learning of the DPP kernel by considering the Key as a new low-rank representation of the DPP kernel. Subsequently, the Key-based and Sampler-based kernels are averaged. "+Value" implies that the probabilistic attention is not directly applied to the sequential items' input but to their transformed representations. Comparative results of recall and NDCG *w.r.t.* different Top-N on two datasets reveal that our succinct Sampler setting, well aligned with the DPP sampling process, is the most prominent. This provides a more intuitive answer to **RQ2**: utilizing Sampler to the implementation of PAtt instead of pursuing a complex Query-Key-Value mechanism. From

Table 2: Overall performance comparison.

Dataset	Metric	Caser	SASRec	BERT	STOSA	GC-SAN	CL4SRec	DuoRec	MOJ	STRec	**PAtt^2 (%)**	**DPAtt^2 (%)**	**PAtt^3 (%)**	**DPAtt^3 (%)**
Beauty	Re@5	0.1096	0.1293	0.1148	0.1302	0.1204	0.1312	0.1329	0.1296	0.1334	0.1415 (6.1)	0.1427 (7.0)	**0.1463** (9.7)	0.1459 (9.4)
	Re@20	0.2631	0.2789	0.2646	0.2817	0.2711	0.2834	0.2849	0.2835	0.2800	0.2965 (4.1)	0.2968 (4.2)	**0.2978** (4.5)	0.2973 (4.4)
	ND@5	0.0776	0.0803	0.0762	0.0839	0.0795	0.0830	0.0843	0.0813	0.0857	0.0960 (12.0)	0.0954 (11.3)	**0.0967** (12.8)	0.0962 (12.3)
	ND@20	0.1158	0.1228	0.1179	0.1219	0.1215	0.1261	0.1305	0.1280	0.1268	0.1386 (6.2)	0.1380 (5.7)	0.1392 (6.7)	**0.1395** (6.9)
	CC@5	0.0530	0.0510	0.0523	0.0512	0.0504	0.0517	0.0530	0.0522	0.0512	0.0538 (1.5)	0.0550 (3.8)	0.0535 (0.9)	**0.0554** (4.5)
	CC@20	0.1176	0.1170	0.1183	0.1095	0.1054	0.1157	0.1150	0.1104	0.1148	0.1180 (-0.3)	0.1209 (2.2)	0.1196 (1.1)	**0.1213** (2.5)
	ILD@5	0.7359	0.7363	0.7288	0.7400	0.7178	0.7402	0.7371	0.7295	0.7219	0.7464 (0.8)	0.7502 (1.4)	0.7429 (0.4)	**0.7507** (1.4)
	ILD@20	0.7729	0.7681	0.7594	0.7754	0.7469	0.7715	0.7576	0.7680	0.7726	0.7862 (1.7)	**0.7910** (2.3)	0.7802 (0.9)	0.7906 (2.3)
	F1@5	0.1513	0.1655	0.1535	0.1684	0.1586	0.1686	0.1704	0.1661	0.1707	0.1831 (7.3)	0.1838 (7.6)	**0.1862** (9.1)	**0.1862** (9.1)
	F1@20	0.2658	0.2763	0.2664	0.2772	0.2688	0.2802	0.2814	0.2802	0.2789	0.2937 (4.4)	0.2944 (4.6)	0.2941 (4.5)	**0.2953** (4.9)
CDs	Re@5	0.0313	0.0371	0.0365	0.0385	0.0349	0.0376	0.0408	0.0387	0.0392	0.0429 (5.1)	0.0424 (3.9)	**0.0436** (6.9)	0.0430 (5.4)
	Re@20	0.0734	0.0742	0.0740	0.0759	0.0745	0.0751	0.0762	0.0758	0.0753	0.0802 (5.2)	0.0794 (4.2)	**0.0820** (7.6)	0.0806 (5.8)
	ND@5	0.0206	0.0237	0.0253	0.0241	0.0230	0.0237	0.0262	0.0256	0.0267	0.0294 (10.1)	0.0289 (8.2)	0.0298 (11.6)	**0.0301** (12.7)
	ND@20	0.0328	0.0339	0.0345	0.0362	0.0337	0.0346	0.0354	0.0358	0.0363	0.0385 (6.1)	0.0382 (5.2)	**0.0391** (7.7)	0.0388 (6.9)
	CC@5	0.0569	0.0544	0.0548	0.0533	0.0556	0.0559	0.0547	0.0550	0.0560	0.0571 (0.4)	0.0580 (1.9)	0.0564 (-0.8)	**0.0587** (3.2)
	CC@20	0.1134	0.1146	0.1157	0.1150	0.1162	0.1195	0.1143	0.1207	0.1126	0.1216 (0.7)	0.1233 (2.2)	0.1206 (-0.1)	**0.1249** (3.5)
	ILD@5	0.7637	0.7469	0.7601	0.7648	0.7625	0.7651	0.7436	0.7460	0.7238	0.7672 (0.3)	0.7679 (0.4)	0.7668 (0.2)	**0.7681** (0.4)
	ILD@20	0.7925	0.7906	0.7917	0.7860	0.7954	0.8052	0.7846	0.7922	0.8014	0.8040 (-0.1)	0.8125 (0.9)	0.8079 (0.3)	**0.8134** (1.0)
	F1@5	0.0488	0.0565	0.0574	0.0582	0.0541	0.0570	0.0618	0.0595	0.0608	0.0665 (7.5)	0.0656 (6.2)	**0.0674** (9.0)	0.0672 (8.7)
	F1@20	0.0951	0.0966	0.0969	0.0997	0.0967	0.0981	0.0993	0.0994	0.0995	0.1052 (5.5)	0.1045 (4.8)	**0.1071** (7.5)	0.1059 (6.2)
Anime	Re@5	0.2672	0.2902	0.2847	0.2860	0.2898	0.2909	0.2920	0.2914	0.2873	0.3096 (6.0)	0.3106 (6.4)	**0.3203** (9.7)	0.3170 (8.6)
	Re@20	0.5899	0.6010	0.5873	0.5914	0.5910	0.5927	0.6002	0.6076	0.5912	0.6245 (2.8)	0.6250 (2.9)	**0.6318** (4.0)	0.6306 (3.8)
	ND@5	0.1758	0.2035	0.1895	0.2040	0.2006	0.2023	0.2051	0.2049	0.2060	0.2134 (3.6)	0.2146 (4.2)	**0.2190** (6.3)	0.2172 (5.4)
	ND@20	0.2804	0.2914	0.2817	0.2901	0.2905	0.2921	0.2956	0.2970	0.2916	0.3091 (4.1)	0.3076 (3.6)	**0.3122** (5.1)	0.3098 (4.3)
	CC@5	0.3156	0.3137	0.3143	0.3102	0.3127	0.3200	0.3110	0.3115	0.3116	0.3180 (-0.6)	**0.3249** (1.5)	0.3241 (1.3)	0.3245 (1.4)
	CC@20	0.5604	0.5564	0.5613	0.5539	0.5603	0.5615	0.5577	0.5585	0.5437	0.5620 (0.1)	0.5688 (1.3)	**0.5694** (1.4)	0.5690 (1.3)
	ILD@5	0.7840	0.7735	0.7692	0.7638	0.7781	0.7834	0.7806	0.7702	0.7629	0.7891 (0.7)	**0.7944** (1.3)	0.7852 (0.2)	0.7935 (1.2)
	ILD@20	0.8079	0.8076	0.8159	0.8142	0.7869	0.8120	0.8028	0.8071	0.8004	0.8124 (-0.4)	0.8246 (1.1)	0.8147 (-0.1)	**0.8265** (1.3)
	F1@5	0.3158	0.3395	0.3298	0.3365	0.3383	0.3408	0.3416	0.3402	0.3381	0.3552 (4.0)	0.3575 (4.7)	**0.3629** (6.2)	0.3615 (5.8)
	F1@20	0.5320	0.5395	0.5328	0.5361	0.5328	0.5381	0.5401	0.5441	0.5328	0.5560 (2.2)	0.5587 (2.7)	0.5612 (3.1)	**0.5618** (3.2)
ML-1M	Re@5	0.0759	0.0791	0.0726	0.0780	0.0775	0.0795	0.0812	0.0804	0.0797	0.0850 (4.7)	0.0864 (6.4)	**0.0867** (6.8)	0.0860 (5.9)
	Re@20	0.1870	0.2015	0.1864	0.2018	0.1923	0.2010	0.2053	0.2010	0.2029	0.2181 (6.2)	0.2155 (5.0)	**0.2186** (6.5)	0.2180 (6.2)
	ND@5	0.0490	0.0505	0.0480	0.0493	0.0491	0.0502	0.0518	0.0509	0.0523	0.0542 (3.6)	0.0551 (5.4)	0.0559 (6.9)	**0.0560** (7.1)
	ND@20	0.0826	0.0849	0.0796	0.0836	0.0845	0.0860	0.0859	0.0877	0.0864	0.0896 (2.2)	0.0891 (1.6)	0.0902 (2.9)	**0.0904** (3.1)
	CC@5	0.3057	0.3042	0.2967	0.3013	0.3014	0.3032	0.2974	0.3011	0.2980	0.3082 (0.8)	**0.3095** (1.2)	0.3046 (-0.4)	0.3089 (1.0)
	CC@20	0.5613	0.5542	0.5651	0.5612	0.5594	0.5617	0.5558	0.5590	0.5610	0.5693 (0.7)	**0.5699** (0.8)	0.5647 (-0.1)	0.5690 (0.7)
	ILD@5	0.7987	0.7830	0.7926	0.7782	0.7911	0.7860	0.7816	0.7914	0.7895	0.8030 (0.5)	**0.8043** (0.7)	0.7976 (-0.1)	0.8034 (0.6)
	ILD@20	0.8250	0.8175	0.8218	0.8158	0.8116	0.8233	0.8241	0.8232	0.8200	0.8276 (0.3)	0.8291 (0.5)	0.8285 (0.4)	**0.8294** (0.5)
	F1@5	0.1122	0.1158	0.1086	0.1139	0.1135	0.1159	0.1184	0.1172	0.1177	0.1237 (4.5)	0.1255 (6.0)	**0.1263** (6.6)	0.1259 (6.4)
	F1@20	0.2257	0.2369	0.2232	0.2364	0.2303	0.2377	0.2405	0.2388	0.2392	0.2522 (4.9)	0.2501 (4.0)	**0.2528** (5.1)	0.2527 (5.1)

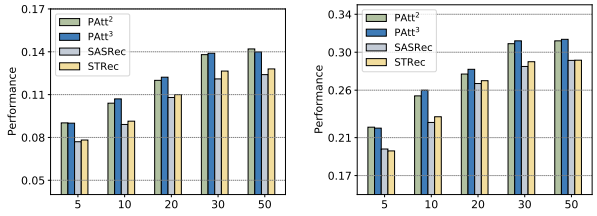


(a) Recall on CDs (b) NDCG on CDs (c) Recall on Anime (d) NDCG on Anime

Figure 3: NDCG and Recall performance comparison among different PAtt mechanisms *w.r.t.* Top-N.

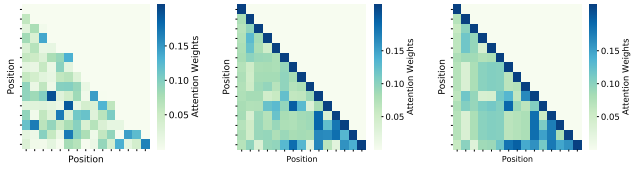
Figure 3, we can observe that: (i) the "+K" variant performs the poorest (but still better than SASRec), possibly because introducing superfluous Key could interfere with the learning of the DPP kernel, thereby weakening its ability to distribute probabilities for subsets in the global context; (ii) The "+VK" variant, despite introducing more weight parameters, does not perform well, indicating that PAtt does not require redundant parameters to excessively match the predictive task of SR.

In Figure 4, we investigate the influence of the maximum length T on the recommendation accuracy (NDCG@20) of attention-based



(a) Beauty (b) Anime

Figure 4: NDCG@20 performance *w.r.t.* sequence length.



(a) SASRec (b) PAtt² (c) PAtt³

Figure 5: Attention Weights Visualizations on ML-1M.

methods. When the value of T is set to be small, items in the previous positions of the long sequence are truncated, which reduces the number of feedbacks in datasets, resulting in further sparse datasets. Analyzing the performance variations across different methods in response to various T values provides a comprehensive answer to **RQ4**. We select two representative self-attention-based models, the seminal work SASRec and the recent STRec, for comparison with PAtt^2 and PAtt^3. Our findings are as follows: (i) All methods experience a certain enhancement with the increase of T, indicating the impact of sparsity (when T is small) on SR. Our methods demonstrate a very significant advantage compared to the two baselines, especially at $T = 5$ and $T = 10$. This suggests that our probabilistic attention is more skilled at handling data sparsity issues compared to other self-attention models; (ii) At $T = 30$, our methods outperform other baselines, even those at $T = 50$. When set to $T = 20$, they achieve results comparable to other baselines at $T = 30$. This suggests that our methods pave the way for achieving commendable recommendation results, even with limited previous information, which highlights a significant advantage of PAtt.

In addition, we conduct a comparative analysis of the attention weights heatmap learned by SASRec, PAtt^2, and PAtt^3, for the same user of ML-1M in Figure 5. We observe that, in contrast to the relatively scattered attention demonstrated by SASRec, both PAtt^2 and PAtt^3 are more capable of capturing attention with a discernible pattern. This indicates that leveraging probabilistic attention within a global context enables the capture of complex dependency interactions within the sequence, consequently enhancing performance. Moreover, PAtt^3 that introduces triple dependency demonstrates a capability to capture a broader range of behavior patterns compared to PAtt^2 (only considers dual dependency), as evidenced by similar attention weights encompassing more items. This underscores a further validation of our proposed definition of dependency interactions, and also elucidates why PAtt^3 outperforms PAtt^2, as it embraces a wider and richer set of dependencies.

The training efficiency of different models when handling various maximum lengths T on two datasets (Anime and ML-1M) is also compared. We discover that when data is sparse ($T = 5$), the PAtt-based models (PAtt^2 and PAtt^3) require a similar amount of time per epoch for training as SASRec. Although training efficiency tends to decrease as T increases, due to the increase in the number of subsets related to probabilistic attention, it is important to acknowledge that, in real-world recommendation applications, the number of available previous sequences for users is inherently limited ($T = 5$ is usually set to 30 or 50 in existing studies [10, 29, 31, 53]). Consequently, this does not impede the utilization of PAtt. Even when the number of items in the sequence is substantial, the training time used by our PAtt^2 and PAtt^3 methods is on the same order of magnitude as traditional self-attention models, while the relative improvement in performance is indeed significant. Moreover, our model, at $T = 30$, consumes close time to SASRec at $T = 50$, while significantly outperforming it in terms of accuracy. This balance between efficiency and effectiveness highlights the robustness and practicality of PAtt models, making them a preferable choice for handling complex user behavioral sequences in dynamic recommendation environments.

5 Conclusion

In this work, we delve into the intrinsic nature of sequential recommendation tasks, proposing a conceptualization of the potent Transformer's self-attention as mutual dependencies among items within a sequence. To formulate these dependency interactions, we transmute the well-formulated repulsion interactions from Determinantal Point Processes (DPP) and, adopting a novel perspective, envision the distribution of attention as probability allocations within global DPP probabilistic models, thereby deriving an innovative probabilistic attention model, PAtt. Remarkably, our model steers clear of the traditional Query-Key-Value setup, achieving commendable results with a mere Sampler. Furthermore, the concept of dependency interactions, as defined at first, incorporates associations among a greater number of items into the formulation of attention weight, enhancing the model's capacity to represent sequences. We also design a novel DPP kernel, capable of integrating DPP kernels with varied inclinations and balancing different tendencies (similarity and diversity), which is a departure from the traditional quality *vs.* diversity decomposition of DPP kernel [28]. The introduction of this kernel affirms the extensibility of PAtt. By providing precise definitions and formulations for new attention models and architectures, PAtt stands to expand researchers' understanding and development of the Transformer. The insights gained from this work could serve as a foundation for developing more sophisticated models that further intertwine the notions of attention and probabilistic modeling, paving the way for breakthroughs in creating models that can understand and predict sequences with higher accuracy and diversity. Expanding PAtt to cover additional sequence analysis tasks is a key future research direction.

6 Acknowledgments

We thank the ANU College of Engineering, Computing & Cybernetics for providing the experimental environment support for this work. We thank Alex Kulesza and Ben Taskar for authoring the tutorial on DPPs [28], and for kindly providing us the code for reference. This work is supported by high performance computing center of Qinghai University.

References

[1] Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. arXiv preprint arXiv:2005.00928 (2020).

[2] Jinze Bai, Chang Zhou, Junshuai Song, Xiaoru Qu, Weiting An, Zhao Li, and Jun Gao. 2019. Personalized bundle list recommendation. In The World Wide Web Conference. 60–71.

[3] Alex Beutel, Paul Covington, Sagar Jain, Can Xu, Jia Li, Vince Gatto, and Ed H Chi. 2018. Latent cross: Making use of context in recurrent recommender systems. In Proceedings of the eleventh ACM international conference on web search and data mining. 46–54.

[4] Renqin Cai, Jibang Wu, Aidan San, Chong Wang, and Hongning Wang. 2021. Category-aware collaborative sequential recommendation. In Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. 388–397.

[5] Laming Chen, Guoxin Zhang, and Eric Zhou. 2018. Fast greedy map inference for determinantal point process to improve recommendation diversity. Advances in Neural Information Processing Systems 31 (2018).

[6] Qi Chen, Guohui Li, Quan Zhou, Si Shi, and Deqing Zou. 2022. Double attention convolutional neural network for sequential recommendation. ACM Transactions on the Web 16, 4 (2022), 1–23.

[7] Tianwen Chen and Raymond Chi-Wing Wong. 2020. Handling information loss of graph neural networks for session-based recommendation. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 1172–1180.

[8] Chen Cheng, Haiqin Yang, Michael R Lyu, and Irwin King. 2013. Where you like to go next: Successive point-of-interest recommendation. In Twenty-Third international joint conference on Artificial Intelligence.

[9] Peizhe Cheng, Shuaiqiang Wang, Jun Ma, Jiankai Sun, and Hui Xiong. 2017. Learning to recommend accurate and diverse items. In Proceedings of the 26th international conference on World Wide Web. 183–192.

[10] Xinyu Du, Huanhuan Yuan, Pengpeng Zhao, Jianfeng Qu, Fuzhen Zhuang, Guanfeng Liu, Yanchi Liu, and Victor S Sheng. 2023. Frequency enhanced hybrid attention network for sequential recommendation. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 78–88.

[11] Ziwei Fan, Zhiwei Liu, Yu Wang, Alice Wang, Zahra Nazari, Lei Zheng, Hao Peng, and Philip S Yu. 2022. Sequential recommendation via stochastic self-attention. In Proceedings of the ACM Web Conference 2022. 2036–2047.

[12] Lu Gan, Diana Nurbakova, Léa Laporte, and Sylvie Calabretto. 2020. Enhancing recommendation diversity using determinantal point processes on knowledge graphs. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2001–2004.

[13] Mike Gartrell, Ulrich Paquet, and Noam Koenigstein. 2017. Low-rank factorization of determinantal point processes. In Thirty-First AAAI Conference on Artificial Intelligence.

[14] Izrail Moiseevich Gelfand. 1989. Lectures on linear algebra. Courier Corporation.

[15] Jennifer A Gillenwater, Alex Kulesza, Emily Fox, and Ben Taskar. 2014. Expectation-maximization for learning determinantal point processes. Advances in Neural Information Processing Systems 27 (2014).

[16] Qipeng Guo, Xipeng Qiu, Pengfei Liu, Xiangyang Xue, and Zheng Zhang. 2020. Multi-scale self-attention for text classification. In Proceedings of the AAAI conference on artificial intelligence, Vol. 34. 7847–7854.

[17] Insu Han, Prabhanjan Kambadur, Kyoungsoo Park, and Jinwoo Shin. 2017. Faster greedy MAP inference for determinantal point processes. In International Conference on Machine Learning. PMLR, 1384–1393.

[18] Yongjing Hao, Tingting Zhang, Pengpeng Zhao, Yanchi Liu, Victor S Sheng, Jiajie Xu, Guanfeng Liu, and Xiaofang Zhou. 2023. Feature-Level Deeper Self-Attention Network With Contrastive Learning for Sequential Recommendation. IEEE Transactions on Knowledge and Data Engineering (2023).

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

[20] Ruining He and Julian McAuley. 2016. Fusing similarity models with markov chains for sparse sequential recommendation. In 2016 IEEE 16th international conference on data mining (ICDM). IEEE, 191–200.

[21] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In proceedings of the 25th international conference on world wide web. 507–517.

[22] Cheng Hsu and Cheng-Te Li. 2021. Retagnn: Relational temporal attentive graph neural networks for holistic sequential recommendation. In Proceedings of the web conference 2021. 2968–2979.

[23] Xiaowen Huang, Shengsheng Qian, Quan Fang, Jitao Sang, and Changsheng Xu. 2018. Csan: Contextual self-attention network for user sequential recommendation. In Proceedings of the 26th ACM international conference on Multimedia. 447–455.

[24] Mingi Ji, Weonyoung Joo, Kyungwoo Song, Yoon-Yeong Kim, and Il-Chul Moon. 2020. Sequential recommendation with relation-aware kernelized self-attention. In Proceedings of the AAAI conference on artificial intelligence, Vol. 34. 4304–4311.

[25] Juyong Jiang, Jae Boum Kim, Yingtao Luo, Kai Zhang, and Sunghun Kim. 2022. AdaMCT: adaptive mixture of CNN-transformer for sequential recommendation. arXiv preprint arXiv:2205.08776 (2022).

[26] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In 2018 IEEE international conference on data mining (ICDM). IEEE, 197–206.

[27] Alex Kulesza and Ben Taskar. 2011. k-DPPs: Fixed-size determinantal point processes. In ICML.

[28] Alex Kulesza, Ben Taskar, et al. 2012. Determinantal point processes for machine learning. Foundations and Trends® in Machine Learning 5, 2–3 (2012), 123–286.

[29] Chengxi Li, Yejing Wang, Qidong Liu, Xiangyu Zhao, Wanyu Wang, Yiqi Wang, Lixin Zou, Wenqi Fan, and Qing Li. 2023. STRec: Sparse Transformer for Sequential Recommendations. In Proceedings of the 17th ACM Conference on Recommender Systems. 101–111.

[30] Jiacheng Li, Yujie Wang, and Julian McAuley. 2020. Time interval aware self-attention for sequential recommendation. In Proceedings of the 13th international conference on web search and data mining. 322–330.

[31] Jiacheng Li, Tong Zhao, Jin Li, Jim Chan, Christos Faloutsos, George Karypis, Soo-Min Pantel, and Julian McAuley. 2022. Coarse-to-fine sparse sequential recommendation. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2082–2086.

[32] Yang Li, Tong Chen, Peng-Fei Zhang, and Hongzhi Yin. 2021. Lightweight self-attentive sequential recommendation. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 967–977.

[33] Yunyi Li, Yongjing Hao, Pengpeng Zhao, Guanfeng Liu, Yanchi Liu, Victor S Sheng, and Xiaofang Zhou. 2023. Edge-enhanced global disentangled graph neural network for sequential recommendation. ACM Transactions on Knowledge Discovery from Data 17, 6 (2023), 1–22.

[34] Yile Liang and Tieyun Qian. 2021. Recommending accurate and diverse items using bilateral branch network. arXiv preprint arXiv:2101.00781 (2021).

[35] Yong Liu, Zhiqi Shen, Yinan Zhang, and Lizhen Cui. 2021. Diversity-promoting deep reinforcement learning for interactive recommendation. In 5th International Conference on Crowd Science and Engineering. 132–139.

[36] Yuli Liu, Christian Walder, and Lexing Xie. 2022. Determinantal Point Process Likelihoods for Sequential Recommendation. arXiv preprint arXiv:2204.11562 (2022).

[37] Shalini Pandey and Jaideep Srivastava. 2020. RKT: relation-aware self-attention for knowledge tracing. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 1205–1214.

[38] Shameem A Puthiya Parambath, Nicolas Usunier, and Yves Grandvalet. 2016. A coverage-based approach to recommendation diversity on similarity graph. In Proceedings of the 10th ACM Conference on Recommender Systems. 15–22.

[39] Ruihong Qiu, Zi Huang, Hongzhi Yin, and Zijian Wang. 2022. Contrastive learning for representation degeneration problem in sequential recommendation. In Proceedings of the fifteenth ACM international conference on web search and data mining. 813–823.

[40] Lakshmanan Rakkappan and Vaibhav Rajan. 2019. Context-aware sequential recommendations withstacked recurrent neural networks. In The World Wide Web Conference. 3172–3178.

[41] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. arXiv preprint arXiv:1803.02155 (2018).

[42] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Sen Wang, and Chengqi Zhang. 2018. Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling. arXiv preprint arXiv:1801.10296 (2018).

[43] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In Proceedings of the 28th ACM international conference on information and knowledge management. 1441–1450.

[44] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In Proceedings of the eleventh ACM international conference on web search and data mining. 565–573.

[45] Viet-Anh Tran, Guillaume Salha-Galvan, Bruno Sguerra, and Romain Hennequin. 2023. Attention Mixtures for Time-Aware Sequential Recommendation. arXiv preprint arXiv:2304.08158 (2023).

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems 30 (2017).

[47] Romain Warlop, Jérémie Mary, and Mike Gartrell. 2019. Tensorized determinantal point processes for recommendation. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 1605–1615.

[48] Jibang Wu, Renqin Cai, and Hongning Wang. 2020. Déjà vu: A contextualized temporal attention mechanism for sequential recommendation. In Proceedings of The Web Conference 2020. 2199–2209.

[49] Qiong Wu, Yong Liu, Chunyan Miao, Binqiang Zhao, Yin Zhao, and Lu Guan. 2019. PD-GAN: Adversarial Learning for Personalized Diversity-Promoting Recommendation.. In IJCAI, Vol. 19. 3870–3876.

[50] Yongji Wu, Defu Lian, Neil Zhenqiang Gong, Lu Yin, Mingyang Yin, Jingren Zhou, and Hongxia Yang. 2021. Linear-time self attention with codeword histogram for efficient recommendation. In Proceedings of the Web Conference 2021. 1262–1273.
[51] Bin Xia, Yun Li, Qianmu Li, and Tao Li. 2017. Attention-based recurrent neural network for location recommendation. In 2017 12th international conference on intelligent systems and knowledge engineering (ISKE). IEEE, 1–6.
[52] Tong Xiao, Yinqiao Li, Jingbo Zhu, Zhengtao Yu, and Tongran Liu. 2019. Sharing attention weights for fast transformer. arXiv preprint arXiv:1906.11024 (2019).
[53] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive learning for sequential recommendation. In 2022 IEEE 38th international conference on data engineering (ICDE). IEEE, 1259–1273.
[54] Chengfeng Xu, Jian Feng, Pengpeng Zhao, Fuzhen Zhuang, Deqing Wang, Yanchi Liu, and Victor S Sheng. 2021. Long-and short-term self-attention network for sequential recommendation. Neurocomputing 423 (2021), 580–589.
[55] Chengfeng Xu, Pengpeng Zhao, Yanchi Liu, Victor S Sheng, Jiajie Xu, Fuzhen Zhuang, Junhua Fang, and Xiaofang Zhou. 2019. Graph contextualized self-attention network for session-based recommendation.. In IJCAI, Vol. 19. 3940–3946.
[56] Chengfeng Xu, Pengpeng Zhao, Yanchi Liu, Jiajie Xu, Victor S Sheng S. Sheng, Zhiming Cui, Xiaofang Zhou, and Hui Xiong. 2019. Recurrent convolutional neural network for sequential recommendation. In The world wide web conference. 3398–3404.
[57] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning. PMLR, 2048–2057.
[58] An Yan, Shuo Cheng, Wang-Cheng Kang, Mengting Wan, and Julian McAuley. 2019. CosRec: 2D convolutional neural networks for sequential recommendation. In Proceedings of the 28th ACM international conference on information and knowledge management. 2173–2176.
[59] Baosong Yang, Longyue Wang, Derek F Wong, Shuming Shi, and Zhaopeng Tu. 2021. Context-aware self-attention networks for natural language processing. Neurocomputing 458 (2021), 157–169.
[60] Yuhao Yang, Chao Huang, Lianghao Xia, Chunzhen Huang, Da Luo, and Kangyi Lin. 2023. Debiased Contrastive Learning for Sequential Recommendation. In Proceedings of the ACM Web Conference 2023. 1063–1073.
[61] Yuhao Yang, Chao Huang, Lianghao Xia, Yuxuan Liang, Yanwei Yu, and Chenliang Li. 2022. Multi-behavior hypergraph-enhanced transformer for sequential recommendation. In Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining. 2263–2274.
[62] Lu Yu, Chuxu Zhang, Shangsong Liang, and Xiangliang Zhang. 2019. Multi-order attentive ranking model for sequential recommendation. In Proceedings of the AAAI conference on artificial intelligence, Vol. 33. 5709–5716.
[63] Mingliang Zhai, Xuezhi Xiang, Rongfang Zhang, Ning Lv, and Abdulmotaleb El Saddik. 2019. Optical flow estimation using dual self-attention pyramid networks. IEEE Transactions on Circuits and Systems for Video Technology 30, 10 (2019), 3663–3674.
[64] Mi Zhang and Neil Hurley. 2008. Avoiding monotony: improving the diversity of recommendation lists. In Proceedings of the 2008 ACM conference on Recommender systems. 123–130.
[65] Shuai Zhang, Yi Tay, Lina Yao, and Aixin Sun. 2018. Next item recommendation with self-attention. arXiv preprint arXiv:1808.06414 (2018).
[66] Tingting Zhang, Pengpeng Zhao, Yanchi Liu, Victor S Sheng, Jiajie Xu, Deqing Wang, Guanfeng Liu, Xiaofang Zhou, et al. 2019. Feature-level Deeper Self-Attention Network for Sequential Recommendation.. In IJCAI. 4320–4326.
[67] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In Proceedings of the 29th ACM international conference on information & knowledge management. 1893–1902.

A Appendix

A.1 Diversity-aware Attention

Consider two independent positive semi-definite kernel matrices, $\mathbf{L}$ and $\mathbf{C}$, an integrated kernel $\mathbf{T}$, and a subset Y. For clarity, $\mathbf{L}$, $\mathbf{C}$, and $\mathbf{T}$ respectively represent $\mathbf{L}^{(S)}$, $\mathbf{C}_{\mathcal{S}_u}$, and $\mathbf{T}^{(S)}$. We examine:

$$\begin{aligned}\mathcal{P}(Y) &= \frac{\det(\mathbf{T}_Y)}{\det(\mathbf{T}+\mathbf{I})} \\ &= \frac{\det(\mathbf{L}_Y)}{\det(\mathbf{C}_Y)} / \det\left(\mathbf{C}^{-1}(\mathbf{L}+\mathbf{C})\right) \\ &= \frac{\det(\mathbf{L}_Y)}{\det(\mathbf{C}_Y)} / \frac{\det(\mathbf{L}+\mathbf{C})}{\det(\mathbf{C})}.\end{aligned} \tag{13}$$

The objective is to analyze the manner in which $\mathcal{P}(Y)$ evolves as $\det(\mathbf{C}_Y)$ augments, with $\mathbf{L}$ and $\mathbf{C}$ held as independent kernels.

Theorem A.1. *Given independent kernels* $\mathbf{L}$ *and* $\mathbf{C}$, $\mathcal{P}(Y)$ *diminishes as the category scope of* Y *expands.*

Proof. Consider the expression for $\mathcal{P}(Y)$:

$$\mathcal{P}(Y) = \frac{\det(\mathbf{L}_Y)}{\det(\mathbf{C}_Y)} \times \frac{\det(\mathbf{C})}{\det(\mathbf{L}+\mathbf{C})} \tag{14}$$

Firstly, observe that $\det(\mathbf{L}_Y)$ remains invariant for changes in $\det(\mathbf{C}_Y)$ due to the presumed independence of matrices $\mathbf{L}$ and $\mathbf{C}$. Thus, an increment in $\det(\mathbf{C}_Y)$, while keeping $\det(\mathbf{L}_Y)$ constant, will precipitate a reduction in the fraction $\frac{\det(\mathbf{L}_Y)}{\det(\mathbf{C}_Y)}$.

Secondly, the fraction $\frac{\det(\mathbf{C})}{\det(\mathbf{L}+\mathbf{C})}$ remains unaffected by alterations in $\det(\mathbf{C}_Y)$, thus holding constant with respect to changes in Y.

Given these observations, it is evident that while the second component of the expression retains its magnitude, the increase in $\det(\mathbf{C}_Y)$ conduces a decrease in $\mathcal{P}(Y)$, as the first component diminishes while the latter remains constant. In alignment with the objective function described in Equation 10, when a subset Y is diverse, indicating a broad category scope, the corresponding $\det(\mathbf{C}_Y)$ is consequently learned to be a larger value. Hence, it is derived that an increase in category scope of the subset Y reduces the probability of $\mathcal{P}(Y)$. Consequently, the corresponding attention is augmented, as it takes the negation of the probability. □

A.2 Baselines and Implementations

We carry out the implementation of PAtt using PyTorch on a NVIDIA Quadro P2000 GPU. A thorough grid search is undertaken to examine all hyper-parameters across the compared methods, with test performance reported based on peak validation results. For all methods in consideration, the embedding dimension is explored within the set $\{32, 64, 128\}$. The maximum sequence length of our models is examined from 5 to 50 for various analyses and is set to 30 for all experiments, while the learning rate is adjusted within $\{10^{-3}, 10^{-4}\}$. In addition, we investigated the dropout rate within $\{0.3, 0.5, 0.7\}$ and λ within $\{0.5, 1.0, 4.0, 8.0, 16.0\}$ specifically for PAtt. For methods grounded on self-attention, the number of layers was explored within $\{1, 2, 3\}$, the number of heads within $\{1, 2, 4\}$, the maximum sequence length within $\{30, 50, 100\}$ for baselines. An early stopping strategy is applied, model optimization will be ceased if NDCG@20 of validation does not exhibit improvement for 10 consecutive epochs. Following is a detailed introduction and hyper-parameter search ranges of the baselines:

- **Caser**[4] is proposed in [2], which aims to capture high-order patterns by applying horizontal and vertical convolutional operations for sequential recommendation. We search the length L from $\{5, 10\}$, and T from $\{1, 3, 5\}$.

[4] https://github.com/graytowne/caser_pytorch

- **GC-SAN**[5] [55] integrates GNN with a self-attention mechanism to identify both local and long-range transitions of neighboring items hidden within each interaction session. We tune weight factor ω from 0.4 to 0.8.
- **CL4SRec**[6] [53] is the first work that incorporates CL into sequential recommendation based on the basic self-attention model. We test the crop/mask/reorder proportion of items from 0.2 to 0.7 according to the experimental results of [53].
- **DuoRec**[7] [39] provides a model-level augmentation based on Dropout to enable better semantic preserving and address the representation degeneration problem. The scale weight λ is chose from $\{0.1, 0.2, 0.3, 0.4, 0.5\}$.
- **SASRec**[8] [26] is a seminal sequential recommendation method that depends on the self-attention mechanism. We search the dropout rate from $\{0.3, 0.5, 0.7\}$.
- **BERT4Rec**[9] [43] employs a masked item training scheme, analogous to the masked language model sequential in NLP, using the bi-directional self-attention mechanism as its backbone. We tune the mask proportion ρ in $\{0.2, 0.4, 0.6\}$.
- **STOSA**[10] [11] is recently proposed, utilizing the difference between distributions as a substitute for the traditional attention weight calculation method. We search the L2 regularization weight from $\{10^{-1}, 10^{-2}, 10^{-3}\}$.
- **MOJITO**[11] [45] employs Gaussian mixtures of attention-based temporal context and item embedding representations for sequential modeling. We test weight λ for balancing long- and short-term representation in $\{0.1, 0.2, 0.5, 0.8, 1.0\}\}$.
- **STRec**[12] [29] is a recently proposed self-attention based sequential recommendation model that identifies the sparse attention phenomenon by replacing self-attention with cross-attention. We search the dropout rates (attention and hidden state dropout) from $\{0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$ to obtain the best combination.

A.3 Experiments

Recommendation models leveraging DPP to enhance diversity are often designed for traditional recommendation scenarios, as indicated by [5, 12, 13, 49]. A direct comparison with these models is neither sufficient nor fair, given our method's specific focus on sequential recommendation. To address this, we not only consider existing DPP-based models like **PD-GAN** [49] and **CDSL** [36] but also introduce a newly designed model, **AMAP**. The implementation details are:

- **PD-GAN** utilizes DPP MAP to generate items for the discriminator in traditional recommendation systems. We initialize the PD-GAN generator parameters using BPR matrix factorization, adhering to the approach in the original work.
- **AMAP**, our proposed DPP-refined baseline based on SASRec, employs the MAP generation approach from DPP [5] to select recommendations, which is a departure from the conventional method used in SASRec where recommendations are chosen based on ranking items by their predicted relevance scores.
- **CDSL** [36] is a state-of-the-art DPP-based method for sequential recommendation. It uses DPP likelihood to guide the learning process for SR in balancing recommendations, and we have applied this approach on the SASRec model.

[5]https://github.com/johnny12150/GC-SAN
[6]https://github.com/RuihongQiu/DuoRec
[7]https://github.com/RuihongQiu/DuoRec
[8]https://github.com/zfan20/STOSA
[9]https://github.com/jaywonchung/BERT4Rec-VAE-Pytorch
[10]https://github.com/zfan20/STOSA
[11]https://github.com/deezer/sigir23-mojito
[12]https://github.com/ChengxiLi5/STRec

Table 3: Performance comparison between DPP-based SR models and PAtt.

Dataset	Metric	PD-GAN	AMAP	CDSL	PAtt^2 (%)	DPAtt^2 (%)
Beauty	ND@5	0.0556	0.0790	0.0829	**0.0960** (15.80)	0.0954 (14.11)
	ND@20	0.0941	0.1197	0.1225	**0.1386** (10.44)	0.1380 (9.96)
	CC@5	**0.0560**	0.0536	0.0547	0.0538 (-3.92)	0.0550 (-1.78)
	CC@20	0.1195	0.1184	0.1181	0.1180 (-1.25)	**0.1209** (1.17)
	F1@5	0.0558	0.0639	0.0659	0.0690 (4.70)	**0.0698** (5.92)
	F1@20	0.1053	0.1191	0.1202	0.1275 (6.07)	**0.1289** (7.24)
Anime	ND@5	0.1213	0.2018	0.2032	0.2134 (5.02)	**0.2146** (5.61)
	ND@20	0.1670	0.2912	0.2922	**0.3091** (5.82)	0.3076 (5.31)
	CC@5	0.3168	0.3196	0.3165	0.3180 (-0.50)	**0.3249** (1.66)
	CC@20	0.5653	0.5624	0.5598	0.5620 (-0.58)	**0.5688** (0.62)
	F1@5	0.1754	0.2473	0.2471	0.2554 (3.36)	**0.2585** (4.44)
	F1@20	0.2578	0.3837	0.3839	0.3988 (3.88)	**0.3993** (4.01)

In Table 3, F1 measurement is calculated based on NDCG and CC. For these three baselines, the DPP kernel construction is based on the standard quality (predicted relevance score) *vs.* diversity (diverse kernel **C**) decomposition.

The above Table 3 offers observations:

- While PD-GAN shows some advantages in diversity in certain scenarios, its NDCG scores are significantly lower compared others. This illustrates that traditional recommendation models based on DPP are not adequately equipped to handle tasks anticipating dynamic preferences. It also underscores the rationale behind our initiative to design new DPP-refined model AMAP, tailored to meet these evolving requirements.
- When comparing AMAP and CDSL against their base model SASRec (as shown in Table 2), it can be observed that the extent of improvement in results is not significant. This suggests that the performance of AMAP and CDSL is still largely constrained by the underlying model, as these methods serve to adjust or guide the base model to balance accuracy and diversity of suggestions, rather than designing a new model.

Furthermore, we find that the three baselines, which utilize the standard quality *vs.* diversity kernel, often experience a trade-off where an increase in diversity comes at the cost of accuracy. However, the diversity-aware attention DPAtt excels in balancing accuracy and diversity aspects and achieves the best F1 performance. This suggests that the existing quality *vs.* diversity DPP kernel is less effective compared to our newly proposed composite kernel in Equation 11 (detailed and mathematically analyzed in Appendix A.1), which adopts a novel construction form. The composite kernel $\mathbf{T}^{(S)}$ is directly applied in the probabilistic attention model with the aim of balancing the dependency and category-aware diversity in attention distribution, rather than trading off accuracy for diversity in MAP generation.