Sequential Recommendation Based on Multi-Modal and Multi-Lingual

Yongqing Wu *, Yu Xing

***Abstract*—Sequential recommendation aims to predict and recommend the next item of interest based on a user's historical behavior. Despite advancements in this field, current methods often focus on limited item co-occurrence patterns within a session and overlook the potential of rich multi-modal information on a page. Moreover, they fail to consider the correlation between user behaviors and preferences across different languages. Existing methods treat user interests as static, inadequately capturing their dynamic nature. To address these challenges, we propose a novel sequential recommendation method that leverages both multi-modal and multi-lingual information (SRMML). Our approach employs a multi-modal fusion mechanism to enhance the representation of diverse information, embedding items as stochastic distributions that incorporate both mean and covariance embeddings. Additionally, we introduce a multi-lingual Gate Neural Unit to capture personalized user preferences across different languages. We also utilize a probabilistic model to describe the positional relationships of items, accurately simulating dynamic changes in user preferences and collaborative transitivity. Experiments on six real world datasets validate the effectiveness of our algorithm, demonstrating significant improvements over existing methods.**

***Index Terms*—Multi-modal, Multi-lingual, Sequential recommendation, Self-attention**

I. Introduction

RECOMMENDER systems [1], [2], [3], [4], [5] filter information by leveraging user behavior and preference data to deliver personalized recommendations. Sequential Recommendation (SR) [6] plays a huge role in academia and industry due to its ability to capture dynamic interests. Unlike traditional recommendation systems, SR considers the sequence of user behaviors, effectively capturing the evolution and trends of user interests to provide more accurate recommendations. The primary objective of SR consists of modeling users' dynamic preferences and estimating their subsequent selections. Early SR methods employ Recurrent Neural Network (RNN) [7] to capture the temporal features of user behavior. With the advent of the Transformer model, the self-attention mechanism has become a cornerstone of sequential recommendation. The self-attention-based Transformer[8], [9] has recently achieved remarkable advances in SR.SASRec[10], a pioneering work that introduced Transformers to sequential recommendation, uses scaled dot-product self-attention to learn item-item correlation weights. BERT4Rec[11] employs sequence-based bidirectional modeling, whereas TiSASRec[12] and SSE-PT[13] enhance SASRec by incorporating time interval details and user regularization. However, most existing methods predominantly rely on short-term item ID patterns, which restricts their accuracy. Fortunately, incorporating multi-modal information about items presents a promising approach to improving SR accuracy.

Multi-modal recommendation systems have gained substantial research interest, as humans naturally interpret and synthesize information through multiple modalities concurrently[14], [15]. Multi-modal information about items presents a compelling approach to enhance recommendation system performance. Descriptive information, such as images and text on pages, vividly conveys an item's style and attributes. When deciding whether to click on an item, users typically assess all available information, proceeding only if each aspect aligns with their preferences. Therefore, a user's choice is jointly determined by multi-modal information. Although some models attempt to incorporate additional information, such as item categories [16] and description texts [17], accurately depicting user intent remains challenging. The main challenges include:

(1) Descriptive information fusion: In SR scenarios involving multimodal information processing, images and text each emphasize different properties.These modalities are often weakly correlated, and certain modalities may lack relevant or useful information. Therefore, the first challenge is to effectively assess the similarity of features across different modalities and evaluate their influence on the decision-making process in SR.

(2) Multi-lingual model: Recommender systems must account for the correlation between user behaviors and preferences across different linguistic environments to accurately predict users' needs and interests. Accordingly, to provide personalized recommendations, another challenge is how to address the diversity of user preferences and the complexity of language translation.

(3) Dynamic interest model: Current SR approaches treat user interests as deterministic. For example, SASRec and BERT4Rec encodes items through static vector embeddings, neglecting the inherent uncertainty in sequential interactions. Thus, the last challenge is how to model uncertainty in user interests.

To tackle aforementioned issues, we introduce an innovative model, **S**equential **R**ecommendation Based on **M**ulti-modal and **M**ulti-**l**ingual (SRMML), which fuses multi-modal information to generate comprehensive item representations. This model incorporates a Gate Neural Unit for processing of multi-lingual text and uses a probabilistic model for predicting SR outcomes. Specifically, to fuse descriptive information, we utilize a CLIP-based learning method to capture semantic information and directly measure the alignment between textual and visual modalities,

Manuscript received July 18, 2024; revised December 24, 2024.

This work was supported by the National Natural Science Foundation of China (Grant No. 52174184).

Y. Q. Wu is an associate professor of School of Software, Liaoning Technical University, Huludao, 125105, China. (corresponding author to provide phone: +86 135-919-93792; e-mail: yqwuyywu@163.com).

Y. Xing is a postgraduate student of School of Software, Liaoning Technical University, Huludao, 125105, China. (e-mail: xy000710@163.com).

integrating this correlation as a weighting factor. Dynamic tuning of single-peak features and fusion features is achieved by introducing an attention layer which dynamically generates channel weights to quantify the contributions of each modality.

In multi-lingual model, we introduce a gating network to process a priori information, effectively capturing users' diverse preference patterns in response to linguistic variations. We add item embeddings into the bottom layer to generate personalized embedded gates. These gates personalize a selection of raw embeddings from multiple languages to obtain a score of personalized gates. Different personalization semantics process more a priori information and inject it into the model.

Moreover, to dynamically model users' interests, items are embedded as stochastic distributions, incorporating mean embedding (representing base interest) and covariance embedding (representing interest variability). Additionally, item embeddings are represented as Gaussian distributions, with distance metrics from metric learning [18] employed to measure item transitions. The Wasserstein distance [19] in the self-attention mechanism delineates inter-item positional relationships within sequences, integrating uncertainty into model training. Our primary contributions are summarized as follows:

(1) We propose SRMML, an advanced SR method that leverages multi-modal information to more accurately reflect the user decision-making process, surpassing the limitations of traditional co-occurrence based approaches.

(2) We introduce a Gate Neural Unit to process multilingual textual information, enabling the model to capture users' preference characteristics for different language items.

(3) We employ a probabilistic model with innovative techniques to comprehensively uncover user intent. By employing the Wasserstein distance, which satisfies the triangular inequality, our approach effectively captures collaborative transitivity within sequence modeling, enhancing collaborative proximity and significantly improving performance in the item cold-start problem.

II. METHODOLOGY

We design a model SRMML, which is illustrated in Fig. 1. We generate multi-modal features by fusing image and text information, and then introduce a stochastic embedding method to address dynamic uncertainty. We build a multilingual Gate Neural Unit to capture the user's personalized preference for information in different languages and design a probabilistic model to capture the co-passing signals. We employ Feed-forward Networks utilizing the Exponential Linear Unit (ELU) [20] activation function, ensuring the covariance matrices maintain their positive definiteness.

A. Multi-modal Feature Generation

The pipeline consists of two stages: CLIP-guided Feature Generation and Feature Aggregation.

1) CLIP-guided Feature Generation: Since CLIP is pretrained on numerous diverse datasets, it can embed text and image into a unified mathematical space, it is conducive to the computation of cross-modal correlation [21]. We employ CLIP to encode text and image for features f_T and f_I. To enhance the representation in single-peak, we extract features from several encoders for each branch.

In the CLIP-guided feature generation phase, it is challenging to mine the obtained text feature f_T and image feature f_I for their intrinsic semantic relevance due to their notable cross modal semantic gaps. These two features are combined as follows:

$$f_M = concat(f_T, f_I) \tag{1}$$

As a complement to uni-modal feature, multi-modal feature can enhance semantic representation. To eliminate sentiment, noise, and other irrelevant features, we first use pretrained CLIP models for the uni-modal task. CLIP learns to extract semantics from large-scale image-text pairs. To remove redundant information separately, different projection heads are used for resizing. The three projection heads, P_I, P_T, P_M, have identical structures but different weights. Simply combining feature combiners based on CLIP into multi-modal features does not yield accurate semantic information, so we design a fusion adjustment module to eliminate blurring. This module measures cosine similarity across text and image features derived from CLIP and modifies the strength of the combined features accordingly. The cosine similarity is normalized to [0, 1] and computed as follows:

$$similarity = \frac{f_T \cdot (f_I)^T}{\|f_T\| \|f_I\|} \tag{2}$$

As a result, three channels are generated: image feature m_T, text feature m_I, and fusion feature m_M. This process is illustrated as follows:

$$\begin{cases} m_T = P_T(f_T) \\ m_I = P_I(f_I) \\ m_M = M(similarity) P_M(f_M) \end{cases} \tag{3}$$

where M(·) represents a linear map function.

2) Feature Aggregation: Prior to feature aggregation, we reweight the m_T, m_I, and m_M channels using a modal cross attention mechanism. Motivated by SE-Net framework [22], we propose the feature aggregation module shown in Fig. 1 to recalculate the channel weights of m_T, m_I, and m_M. Specifically, we perform mean pooling and max pooling operations on the three connected features to obtain a 1 × 3 vector. This vector is then normalized using GELU [23] and Sigmoid in the 3 × 3 fully connected layer to obtain the attention weights att_T, att_I, and att_M for each feature. Finally, the aggregated feature is obtained as follows:

$$m_{Agg} = att_T \cdot m_T + att_I \cdot m_I + att_M \cdot m_M \tag{4}$$

B. Stochastic Embedding

In our model, items are embedded as stochastic distributions, including mean and covariance embeddings. This stochastic representation spans a wider space, allowing for the inclusion of denser collaborative neighborhoods. Each item is parameterized by a mean and covariance embedding, introducing an element of uncertainty and reflecting the variability of user-item interactions. The framework enriches item embeddings with the potential to capture more nuanced user preferences. A mean embedding table $M^{\mu} \in \mathbb{R}^{|v| \times d}$ and a covariance embedding table $M^{\Sigma} \in \mathbb{R}^{|v| \times d}$ are defined to represent all entities, where d is the number of latent dimensions. To account for the different information represented

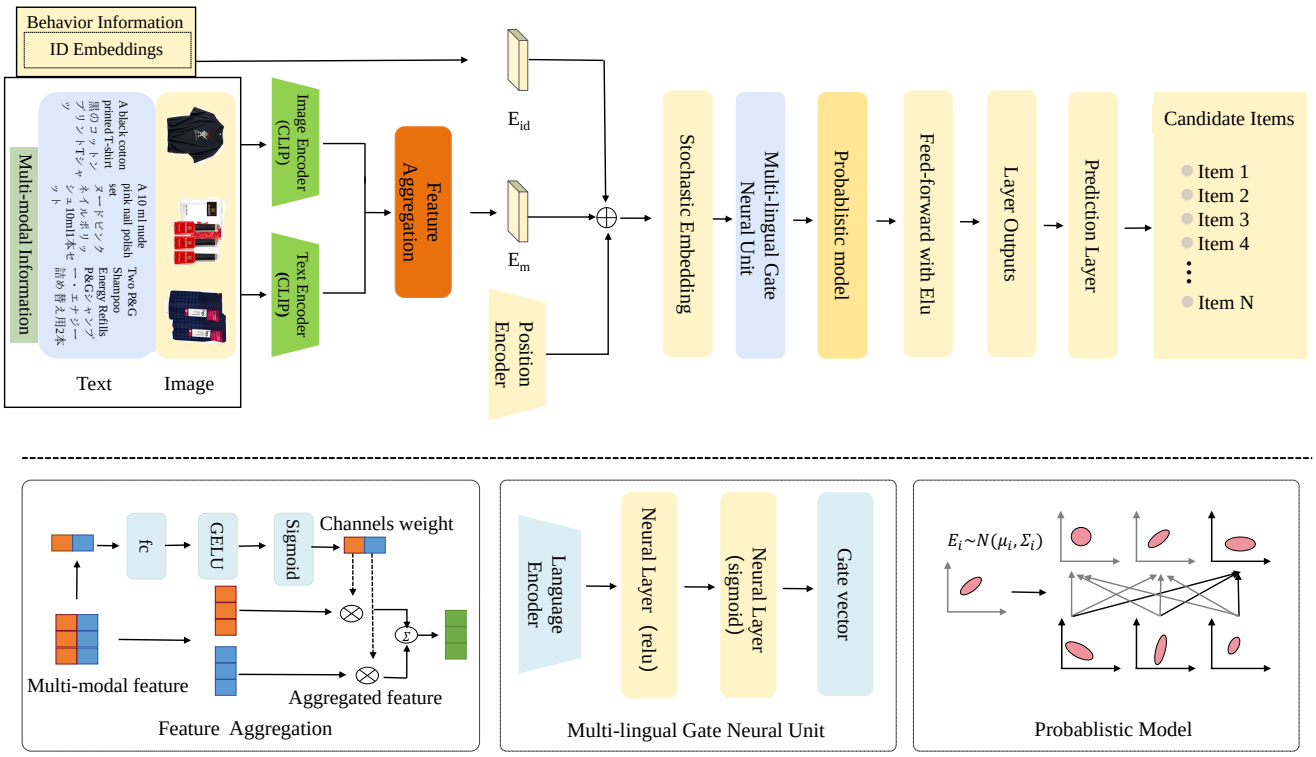


Fig. 1. Schematic diagram illustrating the architecture of the SRMML network

by the mean and covariance, we employ additional positional embeddings $P^{\mu} \in \mathbb{R}^{n \times d}$ and $P^{\Sigma} \in \mathbb{R}^{n \times d}$ for them. Thus, we obtain embeddings for the mean and covariance sequences of users:

$$\begin{cases} E^{\Sigma}_{S_u} = [E^{\Sigma}_{S_1}, \ldots, E^{\Sigma}_{S_n}] \\ \quad\quad = [M^{\Sigma}_{s_1} + P^{\Sigma}_{s_1}, \ldots, M^{\Sigma}_{s_n} + P^{\Sigma}_{s_n}] \\ E^{\mu}_{S_u} = [E^{\mu}_{S_1}, \ldots, E^{\mu}_{S_n}] \\ \quad\quad = [M^{\mu}_{s_1} + P^{\mu}_{s_1}, \ldots, M^{\mu}_{s_n} + P^{\mu}_{s_n}] \end{cases} \tag{5}$$

C. Multi-lingual Gate Neural Unit

In recommender systems, we need to deal with a priori information with different personalized semantics to obtain an exhaustive understanding of user tendencies. Specifically, item IDs and description languages are key to matching users and items. According to several research [24], users exhibit unique individual preference patterns for diverse items. Inspired by this, we design a Gate Neural Unit comprising two neural network layers, which are then injected into the model. We represent the input to Gate Neural Unit as y and describe the first layer's formulation in the following way:

$$y' = ReLU(yw + b) \tag{6}$$

where w and b represent the trainable weights and biases. ReLU serves as the nonlinear activation function. The first layer integrates multiple features with prior knowledge. Next, we modify the gate scores according to the second layer as below:

$$G = \lambda * Sigmoid(y'w' + b'), G \in [0, \lambda] \tag{7}$$

where w' and b' represent the learnable weights and bias, and y' is derived from the first layer's output. The $Sigmoid$ function is applied to produce gate vectors G, limiting the output to the range [0, λ], where λ is a scaling factor set to 2.

Gate Neural Unit creates personalized gates G by using a priori knowledge y. It adaptively controls the importance of a priori information and uses hyper-parameters to compress and amplify the effective signal. We take the item's language information $E(f_L)$, including the language ID and language-specific personalized statistical features, as input to Gate Neural Unit. The Ω_{ep} represents the Gate Neural Unit of our model in the embedding layer. Then we obtain a personalized gating vector through a gating network:

$$G_L = \Omega_{ep}(E(f_L)) \tag{8}$$

We personalize the multi-modal embeddings using gating vectors to transform them in a way that preserves the original embedding layer. This method adaptively controls the importance of plurality and adjusts the degree of user preference for various linguistic messages. The transformed embeddings are calculated by element-wise product as follows:

$$\begin{cases} \widetilde{E}^{\Sigma}_{S_i} = G_L \otimes E^{\Sigma}_{S_i} \\ \widetilde{E}^{\mu}_{S_i} = G_L \otimes E^{\mu}_{S_i} \end{cases} \tag{9}$$

D. Probabilistic Model

The self-attention mechanism is commonly used to model embedded behavioral sequences. To address the challenges of shifting and distributing items in the model dynamics and aggregating these sequential signals to acquire a representation of the user, we employ the Wasserstein distance. This distance assesses the pairwise relationships between the sequence items. We use the linear combination characteristic of the Gaussian distribution to combine the historical items in order to generate the sequential representation.

1) Gaussian Distribution Embedding: We capture the probabilistic nature of item interactions by representing items as multidimensional elliptical Gaussian distributions instead of static vectors. For instance, the stochastic embedding of the i-th item in the sequence can be depicted with a d-dimensional Gaussian distribution: $N(\mu_{s_i}, \Sigma_{s_i}) \in \mathbb{R}^{d\times d}$, where $\mu_{s_i} = \widetilde{E}^{\mu}_{S_i}$ and $\Sigma_{s_i} = diag(\widetilde{E}^{\Sigma}_{S_i})$.

2) Wasserstein Attention: We design a Wasserstein self-attention mechanism, where $A \in \mathbb{R}^{n\times n}$ denotes the self-attention values. A_{kt} indicates the attention value between item s_k at k-th position and item s_t at t-th position in the sequence, where $k \leqslant t$. The attention weights in the conventional self-attention mechanism are determined as follows:

$$A_{kt} = Q_k K_t^T / \sqrt{d} \tag{10}$$

where $Q_k = \widetilde{\Sigma}_{S_k} W^Q$, $K_t = \widetilde{\Sigma}_{S_t} W^K$. W^Q, $W^K \in \mathbb{R}^{d\times d_k}$ are trainable query and key weight matrices respectively. However, traditional self-attention computes the similarity between dot product vector embeddings, which is not suitable for our setup. Consequently ,we estimate the distance between stochastic embeddings using the Wasserstein distance[25]. Formally, the corresponding stochastic embeddings for items s_k and s_t are $N(\mu_{s_k}, \Sigma_{s_k})$ and $N(\mu_{s_t}, \Sigma_{s_t})$, where $\mu_{s_k} = \widetilde{E}^{\mu}_{S_k} W^{\mu}_K$, $\Sigma_{s_k} = ELU(diag(\widetilde{E}^{\Sigma}_{S_k} W^{\Sigma}_K)) + 1$, $\mu_{s_t} = \widetilde{E}^{\mu}_{S_t} W^{\mu}_Q$, $\Sigma_{s_t} = ELU(diag(\widetilde{E}^{\Sigma}_{S_t} W^{\Sigma}_Q)) + 1$. W^{μ}_K, W^{Σ}_K represents the key weight matrix for mean embedding and covariance embedding respectively, and W^{μ}_Q, W^{Σ}_Q represents the query weight matrix. ELU is used to ensure the positive characterization of the covariance, which maps the inputs into $[-1, +\infty)$. The attention weight is defined as the negative 2-Wasserstein distance $W_2(\cdot,\cdot)$, and it is quantified in the following way:

$$\begin{aligned} A_{kt} &= -(W_2(s_k, s_t)) \\ &= -(\|\mu_{s_k} - \mu_{s_t}\|_2^2 \\ &\quad + trace(\Sigma_{s_k} + \Sigma_{s_t} - 2(\Sigma_{s_k}^{\frac{1}{2}} \Sigma_{s_t}^{\frac{1}{2}})^{\frac{1}{2}})) \end{aligned} \tag{11}$$

3) Wasserstein Attentive Aggregation: The output embedding is the sum of the weights of the embeddings from the previous step for each item at each location in the sequence. The weights are the normalized attention values $\widetilde{A}$ as follows:

$$\widetilde{A}_{kt} = \frac{A_{kt}}{\sum_{j=1}^{t} A_{jt}} \tag{12}$$

Each item consists of a linear combination of stochastic embeddings of the mean and covariance, which are represented as follows:

$$\begin{cases} z^{\mu}_{s_t} = \sum\limits_{k=1}^{t} \widetilde{A}_{k_t} V^{\mu}_k \\ z^{\Sigma}_{s_t} = \sum\limits_{k=0}^{t} \widetilde{A}^2_{k_t} V^{\Sigma}_k \end{cases} \tag{13}$$

where $V^{\mu}_{s_k} = \widetilde{E}^{\mu}_{s_t} W^{\mu}_V, V^{\Sigma}_{s_k} = diag(\widetilde{E}^{\Sigma}_{S_k}) W^{\Sigma}_V$, and $k \leq t$ for causality. W^{μ}_V, W^{Σ}_V represents the value weight matrix for mean embedding and covariance embedding respectively. The outputs $Z^{\mu} = (z^{\mu}_{s_1}, z^{\mu}_{s_2}, ..., z^{\mu}_{s_n})$ and $Z^{\Sigma} = (z^{\Sigma}_{s_1}, z^{\Sigma}_{s_2}, ..., z^{\Sigma}_{s_n})$ together form the newly generated sequence's stochastic embeddings, which combine previous sequential signals while taking uncertainty into account.

E. Feed-Forward Network and Layer Outputs

The nonlinear activation function is capable of capturing intricate relations. To introduce nonlinearity in the learning of stochastic embeddings, we employ two fully connected layers with ELU activations applied point-wise:

$$\begin{cases} FFN^{\mu}(z^{\mu}_{s_t}) = ELU(z^{\mu}_{s_t} W^{\mu}_1 + b^{\mu}_1) W^{\mu}_2 + b^{\mu}_2) \\ FFN^{\Sigma}(z^{\Sigma}_{s_t}) = ELU(z^{\Sigma}_{s_t} W^{\Sigma}_1 + b^{\Sigma}_1) W^{\Sigma}_2 + b^{\Sigma}_2) \end{cases} \tag{14}$$

where W^{μ}_1, $W^{\Sigma}_1 \in \mathbb{R}^{d\times d}$, W^{μ}_2, $W^{\Sigma}_2 \in \mathbb{R}^{d\times d}$, b^{μ}_1, $b^{\Sigma}_1 \in \mathbb{R}^d$, and b^{μ}_2, $b^{\Sigma}_2 \in \mathbb{R}^d$ are learnable parameters. Because of its better numerical stability, we choose ELU instead of ReLU. Additionally, we incorporate elements like residual connection, layer normalization, and dropout layers. The outputs of each layer are formulated in the following way:

$$\begin{cases} Z^{\mu}_{s_t} = z^{\mu}_{s_t} + Dropout(FFN^{\mu}(LayerNorm(z^{\mu}_{s_t}))) \\ Z^{\Sigma}_{s_t} = ELU(z^{\Sigma}_{s_t} + Dropout(FFN^{\Sigma}(LayerNorm \\ \qquad (z^{\Sigma}_{s_t})))) + 1 \end{cases} \tag{15}$$

ELU activation along with additive covariance embeddings is applied to ensure covariances remain positively definite. Layer superscripts are omitted for simplicity, particularly when layers are stacked, with Z^{μ} and Z^{Σ} serving as inputs for the subsequent Wasserstein self-attention layer.

F. Prediction Layer

The probability of a potential item j following item s_t at position t-th is derived by computing the 2-Wasserstein distance between their Gaussian distributions, represented as $N(\mu_{s_t}, \Sigma_{s_t})$ and $N(\mu_j, \Sigma_j)$:

$$\hat{y_j} = Softmax(W_2(s_t, j)) \tag{16}$$

where $\mu_{s_t} = Z^{\mu}_{s_t}$ and $\Sigma_{s_t} = Z^{\Sigma}_{s_t}$ are inferred representations; The embeddings for μ_j and Σ_j and are derived from the input stochastic embeddings. To boost the effectiveness of recommendations, we apply the cross-entropy method:

$$\mathcal{L} = -\sum_{j=1}^{n} y_j \log(\hat{y_j}) + (1 - y_j) \log(1 - \hat{y_j}) \tag{17}$$

where y_j corresponds to the true click status of item j and $\hat{y_j}$ signifies the estimated probability that item j will be clicked.

III. EXPERIMENTS

This section addresses the following five research questions:

- **RQ1:** Does SRMML generate recommendations that outperform baselines?
- **RQ2:** What impact do the various innovative techniques introduced in SRMML have on performance?
- **RQ3:** How does SRMML mitigate the item cold-start problem?
- **RQ4:** How does SRMML perform with different sequence lengths?
- **RQ5:** What effects do different hyper-parameter settings have on SRMML?

A. Experimental Settings

1) Dataset: The dataset used in this experiment is derived from the Multi-lingual Shopping Sessions Dataset [26] provided by Amazon. This dataset includes anonymous customer session data across six different regions: English, German, Japanese, French, Italian, and Spanish. The dataset contains two primary components: product characteristics and user behavior. User behavior is represented as a chronological list of product interactions, and product attributes include details such as title, price (in local currency), brand, color, and description. To convert the dataset into an implicit one, each rating or review is treated as a user-item interaction. We then grouped the interactions by region and excluded sequences that were either too short (length 1) or too infrequent (occurring fewer than 5 times). Additionally, items with missing or invalid images or texts were removed. Table I provides the statistical overview of the dataset.

TABLE I
DATASET STATISTICS AFTER PRE-PROCESSING

Dateset	#user	#Item	Density
German(DE)	1111416	513811	0.00063%
Japanese(JP)	979119	389888	0.00094%
English(UK)	1182181	494409	0.00061%
Spanish(ES)	89047	41341	0.09796%
French(FR)	117516	43033	0.07131%
Italia(IT)	126925	48788	0.05824%

2) Compared Methods: We evaluate the proposed SRMML model against two baseline categories: General Models and Cross-domain Models. These categories are defined as follows:

a) General Models:

- **SASRec** [10] captures sequences of user-item interactions.
- **BERT4Rec** [27] captures bidirectional dependencies for the SR.
- **STOSA** [19] utilizes self-attention mechanisms to capture and represent spatiotemporal patterns in sequential data.

b) Cross-domain Models:

- **NATR** [28] recommends items by learning shared features between domains.
- **PiNet** [29]enhances recommendation accuracy by using a personalized interest network.
- **MiFN** [30] is a session-based recommendation model that utilizes a mixed information flow network.

3) Evaluation Protocols: We follow a full ranking protocol to assess the top-K recommendation performance: Recall@N , NDCG@N and MRR@N. The average of these metrics is reported across all users, with N set to 10 and 20.

4) Implementation Details: The SRMML model is implemented using PyTorch on an Nvidia 3090 64GB GPU. We conduct a grid search to tune the parameters and evaluate performance using the top validation outcomes. For all baseline models, embedding dimensions are chosen from $\{16, 32, 64, 128, 256\}$. Since the models incorporate both mean and covariance embeddings, we examine SRMML embedding dimensions in $\{8, 16, 32, 64, 128\}$ to ensure a fair comparison. The learning rate is adjusted within the range $\{10^{-3}, 10^{-4}\}$ and the dropout rate is varied between $\{0.3, 0.5, 0.7\}$. For the sequential approach, we search the depth of the layers in $\{1, 2, 3, 4\}$ and the number of heads in $\{1, 2, 4, 8, 16\}$. An early stopping strategy is employed, halting optimization if the validation MRR does not improve within 50 epochs.

B. Overall Performance(RQ1)

Table II provides a detailed comparison of the performance of all models. It shows the significant effectiveness and advantages of SRMML. Instead of simply combining image and text representations, we introduce a novel deblurring tuning mechanism. This mechanism employs the cosine similarity calculated from textual and visual representations obtained through the CLIP model, refining the combined weights to enhance semantic precision. Additionally, we introduce a multi-lingual gate mechanism to processes and integrate a priori information with different semantics. This mechanism effectively captures personalized user preferences across different languages. Moreover, we incorporate a probabilistic model and a Wasserstein self-attention mechanism to describe the positional relationships between items in a sequence and to model dynamic changes in user preferences. This approach enables the model to provide insights into user preferences from multivariate features and accurately capture preferences for items in various languages and changing user interests. Experiments conducted on six different datasets demonstrate that our SRMML model outperforms current methods, especially on the IT dataset, where it shows a significant improvement over the best baseline.

Cross-domain recommendation methods utilize data and resources from multiple domains to improve system performance and coverage. TiSASRec utilizes temporal interval information between items in user behavior sequences. CoNet learns relevant information across domains synergistically. MiFN focuses on leveraging multi-modal information for recommendation. Nevertheless, they only perform basic modal fusion and do not take into account the impact of language diversity on user preferences, which restricts their ability to comprehensively model user preferences. In contrast, our approach addresses these limitations and enhances recommendation accuracy.

C. Ablation Study(RQ2)

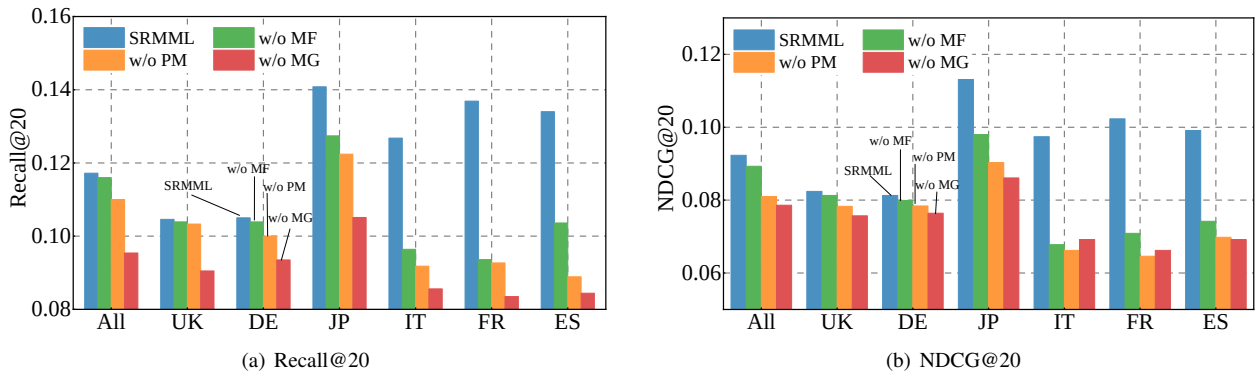
We conduct an ablation study focused on assessing the importance of various modules in SRMML.

1) Effect of Modules: We create the following model variants to evaluate the impact of key components in SRMML.

- **w/o MF**: The multi-modal fusion module is removed. Instead, image and text features are simply combined and directly input into the multi-modal information encoder.
- **w/o MG**: The multi-lingual Gate Neural Unit is removed, which means the model does not personalize multi-lingual text.
- **w/o PM**: The probabilistic model module is removed, and item embeddings are treated as fixed vectors.

TABLE II
THE PERFORMANCE COMPARISON OF DIFFERENT RECOMMENDATION MODELS

Dataset	Metric	General Model			Cross-domain Model			Ours
		SASRec	BERT4Rec	STOSA	NATR	PiNet	MiFN	
UK	Recall@20	0.0445	0.0541	0.0570	0.0913	0.0935	0.1011	0.1046
	NDCG@20	0.0370	0.0391	0.0435	0.0744	0.0774	0.0778	0.0824
	MRR	0.0330	0.0380	0.0415	0.6927	0.6990	0.0708	0.0756
DE	Recall@20	0.0446	0.0564	0.0601	0.0897	0.0935	0.1028	0.1050
	NDCG@20	0.0367	0.0399	0.0444	0.0733	0.0773	0.0776	0.0813
	MRR	0.0342	0.0376	0.0428	0.6847	0.0698	0.0701	0.0741
JP	Recall@20	0.0511	0.0648	0.0661	0.0961	0.1002	0.1195	0.1408
	NDCG@20	0.0421	0.0459	0.0495	0.0793	0.0809	0.0890	0.1131
	MRR	0.0382	0.0447	0.0456	0.0743	0.0788	0.0808	0.1047
IT	Recall@20	0.0365	0.0513	0.0521	0.0848	0.0863	0.0895	0.1268
	NDCG@20	0.0303	0.0342	0.0346	0.0642	0.0692	0.0704	0.0974
	MRR	0.0287	0.0298	0.0305	0.0581	0.0589	0.0604	0.0886
FR	Recall@20	0.0344	0.0500	0.0495	0.0854	0.0872	0.0909	0.1369
	NDCG@20	0.0288	0.0331	0.0360	0.0648	0.0660	0.0705	0.1023
	MRR	0.0283	0.0293	0.0337	0.0587	0.0574	0.0566	0.0918
ES	Recall@20	0.0484	0.0495	0.0557	0.0878	0.0934	0.0889	0.1340
	NDCG@20	0.0335	0.0357	0.0379	0.0728	0.0781	0.0797	0.0991
	MRR	0.0320	0.0335	0.0344	0.0683	0.0675	0.0640	0.0886
ALL	Recall@20	0.0485	0.0568	0.0597	0.0923	0.0968	0.1024	0.1172
	NDCG@20	0.0380	0.0447	0.0465	0.0789	0.0804	0.0852	0.0923
	MRR	0.0327	0.0433	0.0446	0.0705	0.0721	0.0768	0.0847



(a) Recall@20 (b) NDCG@20

Fig. 2. The performance comparison of SRMML with different multi-modal, multi-lingual and probabilistic model setting

Fig. 2 shows Recall@20 and NDCG@20 for these variants on six datasets. The model without the multi-modal fusion (MF) performs significantly worse than SRMML, indicating that simple combination of modal features is insufficient. The multi-modal fusion module effectively retains more information and strengthens the associations between different features, thereby improving recommendation performance.

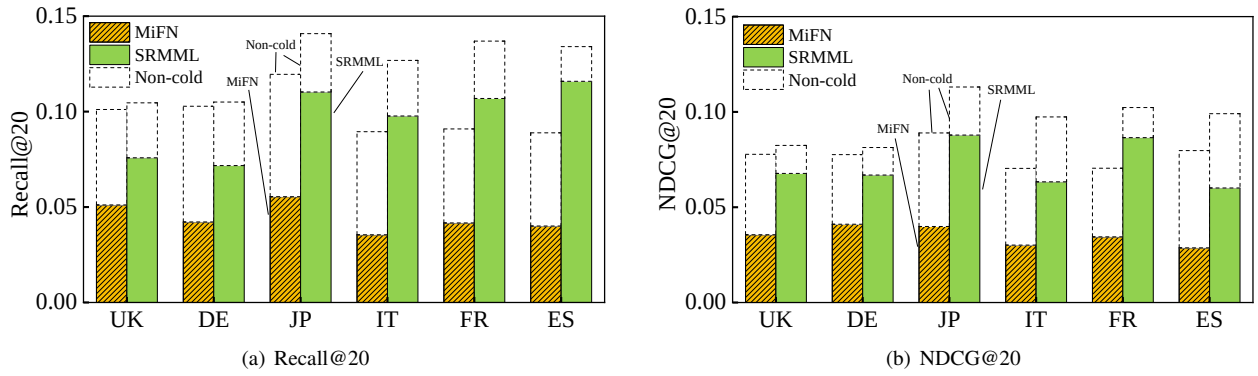
When the multi-lingual Gated Neural Unit (MG) is removed, performance for items in multiple languages deteriorates significantly. This underscores the importance of personalizing multi-lingual information, as it helps capture language-specific user preferences and semantic patterns.

The removal of the probabilistic model (PM) also results in a noticeable decline in recommendation effectiveness. This indicates the essential role of distributional representations and uncertainty modeling in sequence-based recommendation tasks. Moreover, the self-attention mechanism proves to be essential for sequential recommender systems.

2) Effect of Modalities: As shown in Table III,we conduct a series of experiments with varying input conditions to assess how different modalities affect SRMML's efficiency . The text modality includes textual data, the image modality incorporates image data, and the multi-modal condition combines both types of information. Experimental results indicate that both text and image features contribute positively to recommendation performance, with text features having a more pronounced effect. This observation can be attributed to the multi-lingual Gate Neural Units, which effectively capture user preferences and behavioral patterns across different languages. Additionally, image data often contains misleading information, which may overwhelm the model's ability to extract useful patterns from user preferences.

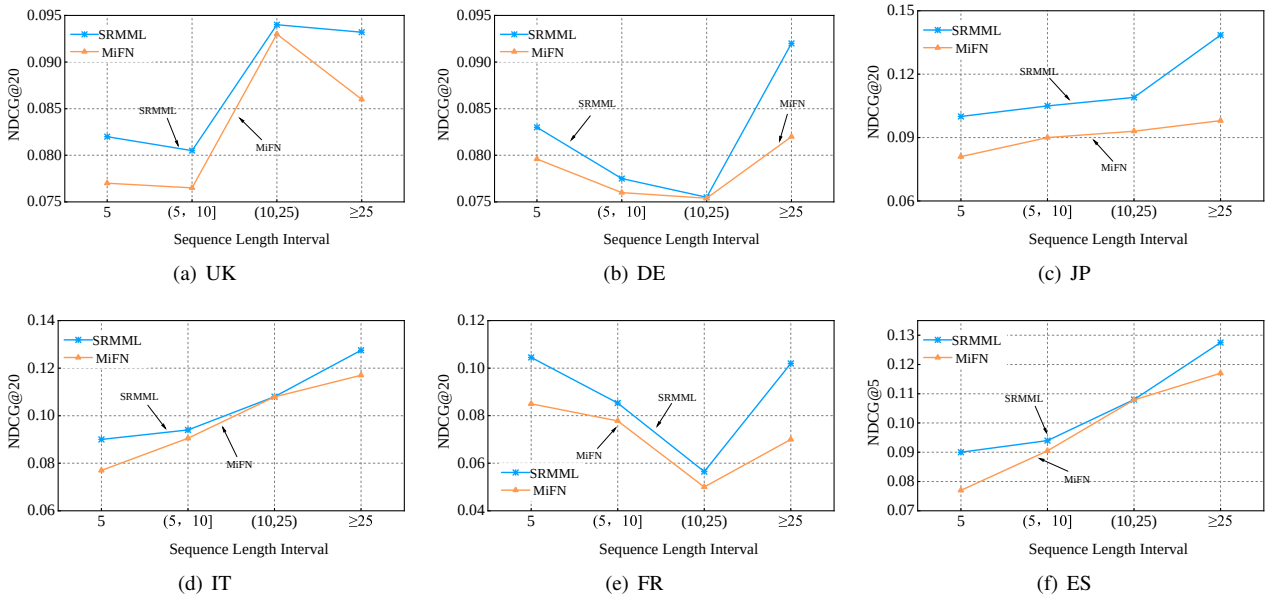
D. Performance in Cold-start Scenario(RQ3)

Recommender systems face a persistent challenge known as the cold-start problem, where new items that have not been



(a) Recall@20 (b) NDCG@20

Fig. 3. The performance in cold-start scenario



(a) UK (b) DE (c) JP

(d) IT (e) FR (f) ES

Fig. 4. The performance on different sequence lengths

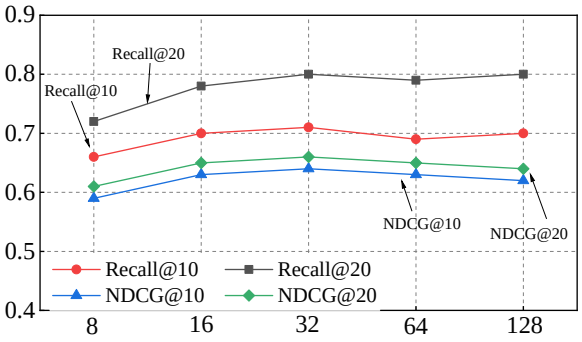


Fig. 5. The influence of embedding dimension on ALL dataset

seen in the training set need to be recommended. To assess SRMML's performance in such a scenario, we test the model with new items that were not part of the training data. The statistics are shown in Table IV with insights drawn from Fig. 3:

Both models exhibit reduced performance when dealing with new items, highlighting the significant challenges of the cold-start problem in SR. However, the concept of collaborative transitivity proves to be an effective solution. By identifying similarities between items within the same item-project transition pair, this technique can inductively introduce collaborative similarities beyond the limited item-project pairs in the dataset. As a result, collaborative transitivity helps mitigate the cold-start problem by incorporating a broader range of collaboratively similar items. The key difference between the two models lies in SRMML's ability to capture collaborative transitivity through the Wasserstein self-attention mechanism within its probabilistic model. This feature, which is absent in MiFN, enables SRMML to generalize collaborative signals to cold items more effectively.

E. Performance with different sequence length(RQ4)

We analyze the performance of SRMML across varying sequence lengths. We initially categorize users based on the number of interactions they engage in during the training phase. The category with the shortest sequences contains the largest number of users, and as the sequence length increases, the number of users in each category correspondingly decreases. In Fig. 4, the proposed SRMML achieves a greater improvement over the baseline on the maximum sequence length interval compared to short sequences. Obviously, it is difficult for traditional methods to accurately predict user behavior in long sequences as users with frequent interactions tend to exhibit a greater diversity of interests. In contrast,

TABLE III
THE PERFORMANCE COMPARISON UNDER DIFFERENT MODALITIES

Datasets	Modality	Recall@20	NDCG@20	MRR
UK	Text	0.0954	0.0768	0.0711
	Image	0.0906	0.0758	0.0713
	Multi-modal	0.1046	0.0824	0.0756
DE	Text	0.0955	0.0762	0.0704
	Image	0.0935	0.0765	0.0712
	Multi-modal	0.1050	0.0813	0.0741
JP	Text	0.1020	0.0830	0.0772
	Image	0.1051	0.0861	0.0804
	Multi-modal	0.1408	0.1131	0.1047
IT	Text	0.0903	0.0687	0.0621
	Image	0.0856	0.0693	0.0645
	Multi-modal	0.1268	0.0974	0.0886
FR	Text	0.0891	0.0696	0.0637
	Image	0.0836	0.0662	0.0612
	Multi-modal	0.1368	0.1023	0.0918
ES	Text	0.0957	0.0775	0.0719
	Image	0.0845	0.0693	0.0647
	Multi-modal	0.1340	0.0990	0.0886
ALL	Text	0.0971	0.0781	0.0723
	Image	0.0955	0.0786	0.0735
	Multi-modal	0.1172	0.0923	0.0847

TABLE IV
DATASET STATISTICS WITH COLD-START ITEMS

Dateset	#user	#Item	Density
German(DE)	1224161(+112745)	562723(+48912)	0.00058%
Japanese(JP)	1019446(+40327)	418610(+28722)	0.00093%
English(UK)	1306660(+124479)	528888(+34479)	0.00058%
Spanish(ES)	98868(+9821)	44620(+3279)	0.09035%
French(FR)	175185(+57669)	46247(+3214)	0.04920%
Italia(IT)	161546(+34621)	52560(+3772)	0.04694%

SRMML proves the efficiency of stochastic embeddings in modeling uncertainty within user behavior. Moreover, SRMML achieves relatively good results in all cases, which again proves its effectiveness in sequential recommendation.

F. Parameter analysis(RQ5)

This section evaluates how hyper parameters influence model performance and recommendation accuracy. Specifically, we focus on the embedding dimension in multi-modal fusion and the number of layers and heads in the Wasserstein Attention module. First, we varies the embedding dimension in the set {8, 16, 32, 64, 128} to assess its influence on SRMML. The results, shown in Fig. 5, reveal that performance improves progressively as the embedding dimension increases. However, the gains plateau after a certain threshold, indicating that higher-dimensional representations enhance the model's capacity to capture richer and more accurate information. Next, we performed experiments on the ALL dataset by adjusting the number of layers and heads in Wasserstein Attention. The results are presented in Fig. 6 and Fig. 7. These experiments demonstrate that as the network structure evolves, recommendation accuracy initially improves but eventually stabilizes. A shallow network fails to adequately capture the complex dependencies between multi-behavioral sequences. In contrast, increasing the network

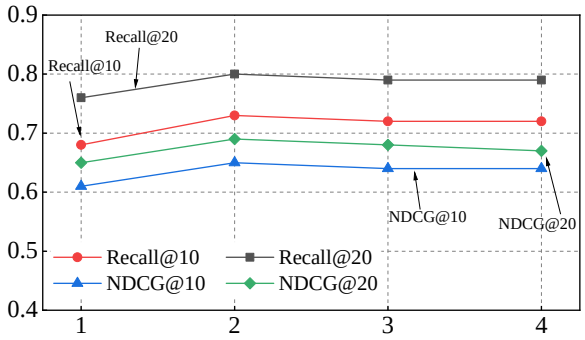


Fig. 6. The influence of the number of layers in Wasserstein Attention module on ALL dataset

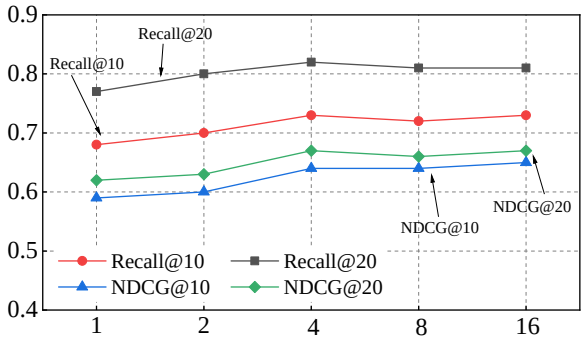


Fig. 7. The influence of the number of heads in Wasserstein Attention module on ALL dataset

depth yields only marginal improvements beyond a certain point. Increasing the number of attention heads enhances the model's ability to focus on both local and global sequence patterns, which in turn improves performance. However, excessive increases in the number of attention heads can lead to redundancy, diminishing the model's overall accuracy.

IV. CONCLUSION

In this paper, we propose a novel sequential recommendation method SRMML improves recommendation performance by integrating multi-modal information and multi-lingual text. Experimental results reveal that SRMML offers significant improvements in both recommendation accuracy and personalization. Our multi-modal fusion mechanism is designed to maximize the mutual information between fused multi-modal and behavioral features, effectively capturing both complementary and supplementary preference information. Additionally, the multi-lingual Gate Neural Unit captures correlations between user behaviors and preferences across varied linguistic environments, thus improving the semantic accuracy of recommendations. Furthermore, the probabilistic model adeptly simulates the dynamic evolution of user interests, enabling a more flexible and responsive recommendation process. Overall, SRMML demonstrates considerable promise in advancing recommender systems by providing highly accurate and personalized recommendations. This work contributes novel methodologies and insights to the field, underscoring its potential to significantly enhance user experience within complex and multi-lingual environments.

REFERENCES

[1] Linyuan Lü, Matúš Medo, Chi Ho Yeung, Yicheng Zhang, Zike Zhang, and Tao Zhou, "Recommender systems," Physics Reports, vol. 519, no. 1, pp1-49, 2012

[2] Robin Burke, Alexander Felfernig, and Mehmet H Göker, "Recommender systems: An overview," Ai Magazine, vol. 32, no. 3, pp13-18, 2011

[3] Ziwei Fan, Zhiwei Liu, Jiawei Zhang, Yun Xiong, Lei Zheng, and Philip S Yu, "Continuous-time sequential recommendation with temporal graph collaborative transformer," in The 30th ACM International Conference on Information & Knowledge Management 2021, pp433-442

[4] Qiushi Wang, and Wenyu Zhang, "Session-based Recommendation Algorithm Based on Heterogeneous Graph Transformer," IAENG International Journal of Computer Science, vol. 50, no. 4, pp1347-1353, 2023

[5] Yue Teng, and Kai Yang, "Research on Enhanced Multi-head Self-Attention Social Recommendation Algorithm Based on Graph Neural Network," IAENG International Journal of Computer Science, vol. 51, no. 7, pp754-764, 2024

[6] Xu Chen, Hongteng Xu, Yongfeng Zhang, Jiaxi Tang, Yixin Cao, Zheng Qin, and Hongyuan Zha, "Sequential recommendation with user memory networks," in The Eleventh ACM International Conference on Web Search and Data Mining 2018, pp108-116

[7] S Hochreiter, "Long Short-term Memory," Neural Computation MIT-Press, vol. 385, pp37-45, 1997

[8] Zhiwei Liu, Ziwei Fan, Yu Wang, and Philip S Yu, "Augmenting sequential recommendation with pseudo-prior items via reversely pre-training transformer," in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval 2021, pp1608-1612

[9] A Vaswani, "Attention is all you need," Advances in Neural Information Processing Systems, vol. 30, 2017

[10] Wang-Cheng Kang, and Julian McAuley, "Self-attentive sequential recommendation," in IEEE International Conference on Data Mining (ICDM) 2018, pp197-206

[11] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang, "BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer," in Proceedings of the 28th ACM International Conference on Information and Knowledge Management 2019, pp1441-1450

[12] Jiacheng Li, Yujie Wang, and Julian McAuley, "Time interval aware self-attention for sequential recommendation," in Proceedings of the 13th International Conference on Web Search and Data Mining 2020, pp322-330

[13] Liwei Wu, Shuqing Li, Chojui Hsieh, and James Sharpnack, "SSE-PT: Sequential recommendation via personalized transformer," in Proceedings of the 14th ACM Conference on Recommender Systems 2020, pp328-337

[14] Venkataravana Nayak K, Sharathkumar S K, Arunalatha J S, and Venugopal K R, "Single and Cross Domain Image Retrieval using Multi-Modal Feature Fusion," IAENG International Journal of Computer Science, vol. 50, no. 2, pp793-802, 2023

[15] Dongping Li, Yingchun Yang, Shikai Shen, Jun He, Haoru Shen, Qiang Yue, Sunyan Hong, and Fei Deng, "Research on Deep Learning Model of Multimodal Heterogeneous Data Based on LSTM," IAENG International Journal of Computer Science, vol. 49, no. 4, pp1016-1022, 2022

[16] Siqi Lai, Erli Meng, Fan Zhang, Chenliang Li, Bin Wang, and Aixin Sun, "An attribute-driven mirror graph network for session-based recommendation," in The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval 2022, pp1674-1683

[17] Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Jirong Wen, "Towards universal sequence representation learning for recommender systems," in The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2022, pp585-593

[18] Chengkang Hsieh, Longqi Yang, Yin Cui, Tsungyi Lin, Serge Belongie, and Deborah Estrin, "Collaborative metric learning," in The 26th International Conference on World Wide Web 2017, pp193-201

[19] Ziwei Fan, Zhiwei Liu, Yu Wang, Alice Wang, Zahra Nazari, Lei Zheng, Hao Peng, and Philip S Yu, "Sequential recommendation via stochastic self-attention," in The ACM Web Conference 2022, pp2036-2047

[20] Djork Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter, "Fast and accurate deep network learning by exponential linear units," arXiv preprint arXiv:1511.07289, 2015

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark, "Learning transferable visual models from natural language supervision," in International Conference on Machine Learning 2021, pp8748-8763

[22] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in The IEEE Conference on Computer Vision and Pattern Recognition 2018, pp7132-7141

[23] Dan Hendrycks, and Kevin Gimpel, "Gaussian error linear units," arXiv preprint arXiv:1606.08415, 2016

[24] Qi Pi, Guorui Zhou, Yujing Zhang, Zhe Wang, Lejian Ren, Ying Fan, Xiaoqiang Zhu, and Kun Gai, "Search-based user interest modeling with lifelong sequential behavior data for click-through rate prediction," in The 29th ACM International Conference on Information & Knowledge Management 2020, pp2685-2692

[25] Ludger Rüschendorf, "The Wasserstein distance and approximation theorems," Probability Theory and Related Fields, vol. 70, no. 1, pp117-129, 1985

[26] Wei Jin, Haitao Mao, Zheng Li, Haoming Jiang, Chen Luo, Hongzhi Wen, Haoyu Han, Hanqing Lu, Zhengyang Wang, and Ruirui Li, "Amazon-M2: A multilingual multi-locale shopping session dataset for recommendation and text generation," in Advances in Neural Information Processing Systems 2023, pp8006-8026

[27] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang, "BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer," in The 28th ACM International Conference on Information and Knowledge Management 2019, pp1441-1450

[28] Chen Gao, Xiangning Chen, Fuli Feng, Kai Zhao, Xiangnan He, Yong Li, and Depeng Jin, "Cross-domain recommendation without sharing user-relevant data," in The World Wide Web Conference 2019, pp491-502

[29] Muyang Ma, Pengjie Ren, Yujie Lin, Zhumin Chen, Jun Ma, and Maarten de Rijke, "π-Net: A parallel information-sharing network for shared-account cross-domain sequential recommendations," in The 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval 2019, pp685-694

[30] Muyang Ma, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Lifan Zhao, Peiyu Liu, Jun Ma, and Maarten de Rijke, "Mixed information flow for cross-domain sequential recommendations," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 16, no. 4, pp1-32, 2022

Yongqing Wu received the PhD degree in computational mathematics from Lanzhou University, China, in 2011. He worked as a visiting scholar with the University of Texas at Arlington, Texas, USA, from 2019 to 2020. Currently, he is an Associate Professor in the the School of Software, Liaoning Technical University, Huludao, China. His current research interests include recommender systems, machine learning algorithms, complex systems and complex networks.

Yu Xing M.S. candidate. Her research interests include graph neural network, recommender systems.