A Large Language Model Enhanced Sequential Recommender for Joint Video and Comment Recommendation

Bowen Zheng
Renmin University of China
Beijing, China
bwzheng0324@ruc.edu.cn

Zihan Lin
Kuaishou Inc.
Beijing, China
linzihan03@kuaishou.com

Enze Liu
Beijing Institute of Technology
Beijing, China
enzeeleo@gmail.com

Chen Yang
Renmin University of China
Beijing, China
flust@ruc.edu.cn

Enyang Bai
Kuaishou Inc.
Beijing, China
baienyang@kuaishou.com

Cheng Ling
Kuaishou Inc.
Beijing, China
lingcheng@kuaishou.com

Wayne Xin Zhao ✉
Renmin University of China
Beijing, China
batmanfly@gmail.com

Ji-Rong Wen
Renmin University of China
Beijing, China
jrwen@ruc.edu.cn

ABSTRACT

In online video platforms, reading or writing comments on interesting videos has become an essential part of the video watching experience. However, existing video recommender systems mainly model users' interaction behaviors with videos, lacking consideration of comments in user behavior modeling.

In this paper, we propose a novel recommendation approach called **LSVCR** by leveraging user interaction histories with both *videos* and *comments*, so as to jointly conduct personalized video and comment recommendation. Specifically, our approach consists of two key components, namely sequential recommendation (SR) model and supplemental large language model (LLM) recommender. The SR model serves as the primary recommendation backbone (*retained in deployment*) of our approach, allowing for efficient user preference modeling. Meanwhile, we leverage the LLM recommender as a supplemental component (*discarded in deployment*) to better capture underlying user preferences from heterogeneous interaction behaviors. In order to integrate the merits of the SR model and the supplemental LLM recommender, we design a two-stage training paradigm. The first stage is personalized preference alignment, which aims to align the preference representations from both components, thereby enhancing the semantics of the SR model. The second stage is recommendation-oriented fine-tuning, in which the alignment-enhanced SR model is fine-tuned according to specific objectives. Extensive experiments in both video and comment recommendation tasks demonstrate the effectiveness of LSVCR. Additionally, online A/B testing on the *KuaiShou* platform verifies the actual benefits brought by our approach. In particular, we achieve a significant overall gain of 4.13% in comment watch time.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Language models**.

KEYWORDS

Large Language Model, Comment, Sequential Recommendation

ACM Reference Format:
Bowen Zheng, Zihan Lin, Enze Liu, Chen Yang, Enyang Bai, Cheng Ling, Wayne Xin Zhao ✉, and Ji-Rong Wen. 2018. A Large Language Model Enhanced Sequential Recommender for Joint Video and Comment Recommendation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Nowadays, recommender systems play an essential role in alleviating information overload by providing personalized recommendation services with high-quality content resources. As a typical application scenario, online video platforms (*e.g., YouTube, TikTok, KuaiShou*[1]) employ recommender systems to deliver customized videos to users, in which they primarily concentrate on capturing users' video watching interests based on user-video interaction logs [18, 24, 28, 30].

With the growth of online video communities, the textual comments posted to the videos become increasingly crucial for enhancing the overall watching experience of users, providing complement or divergence information compared to the original video content. Our statistics show that more than 60% of users on the Kuaishou platform regularly view corresponding comments when watching the videos and express their interests through comments. In light

✉ Corresponding author.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

[1]https://www.kuaishou.com/

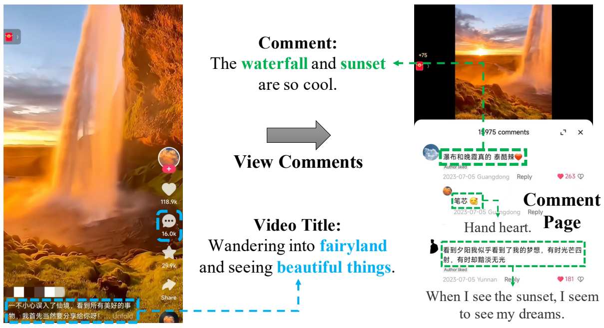

Figure 1: A snapshot of one micro-video and its corresponding comment area on KuaiShou.

of this, this paper aims to leverage both video and comment data to improve the recommendation quality and user engagement.

To be more precise, we propose to study the joint tasks of video and comment recommendation in the specific scenario of KuaiShou, which is a popular Chinese micro-video platform (similar to TikTok). Figure 1 presents a snapshot of a micro-video and its associated comments on KuaiShou. When a user is watching a video and clicks the "*comment*" button, the comment page below the video will unfold, displaying a list of comments to the user. To develop a comprehensive solution, we integrate the modeling of videos and comments into a unified framework, considering the following three aspects:

(1) While comments provide supplemental textual information for a video, their semantic modeling should be grounded in the video content. For example, in Figure 1, the keywords "*sunset*" and "*waterfall*" in the comment complement the video features and should be understood within the context of "*fairyland*" and "*beautiful things*".
(2) User's interaction and feedback behaviors on comments indicate potential interest in watching the corresponding video. When a user actively interacts with the comments of a particular video, he/she is more likely to enjoy videos with similar content or comments.
(3) In KuaiShou, a user could make multiple interactions with videos (*e.g., like,collect, share*) and corresponding comments (*like, reply*). Both interaction histories can capture users' personalized video and comment preferences, which might suggest delivering high-quality video and customized comments instead of monotonous hot comments.

Recently, large language models (LLMs) have demonstrated great potential in recommender systems [4, 8, 36, 53], due to their excellent semantic understanding and knowledge reasoning capabilities. In our task setting, we have abundant textual input such as *video title* and *comment content*, which naturally motivates us to leverage the impressive semantic modeling capability of LLM for developing our approach. Another potential advantage of using LLM in our setting is that the interaction histories of videos and comments essentially provide heterogeneous expressions of user behaviors, and we can leverage LLM to reason about the underlying user preferences from the two kinds of signals. However, the unaffordable computation costs of LLM restrict its efficient deployment as an online recommender in large-scale industrial systems [20, 44, 49]. Considering this issue, we only adopt LLM in training as a supplement to enhance the semantics of our recommendation backbone.

To this end, in this paper, we propose a novel framework that consolidates the benefits of **L**LM with the **S**equential recommendation model for joint **V**ideo and **C**omment **R**ecommendation, namely **LSVCR**. Our framework consists of two key components: conventional sequential recommendation (SR) model and supplemental LLM recommender. To begin with, we develop our recommendation backbone (*retained in deployment*) based on the SR model to achieve efficient user preference modeling. Furthermore, we introduce a supplemental LLM recommender (*discarded in deployment*) to enhance the preference semantics of the SR model. By verbalizing both types of user interaction histories in the format of textual instructions, we aim to leverage LLM's semantic understanding and knowledge reasoning capabilities for enhanced preference modeling. In order to integrate the merits of the above two components, we design a training paradigm consisting of two stages: personalized preference alignment and recommendation-oriented fine-tuning. For the alignment stage, we jointly learn the video and comment recommendation tasks while aligning the preference representation of the SR model with that of the supplemental LLM recommender. For the fine-tuning stage, we discard the supplemental LLM recommender and further adapt the alignment-enhanced SR model to specific tasks for better recommendation performance.

To evaluate our approach, we conduct extensive experiments on a large real-world industrial dataset. The experimental results demonstrate the significant effectiveness of LSVCR compared to competitive baselines for both video and comment recommendation tasks. Additionally, online A/B testing verifies the benefits of LSVCR in the industrial recommendation system of KuaiShou. Especially in the comment end, we achieve a 4.13% increase in watch time and a 1.36% gain in interaction number.

2 PROBLEM FORMULATION

In our task, we consider the online micro-video scenario, where users can freely watch interesting videos and further view comments below specific videos. Different from previous video recommendation studies [18, 24, 28, 30], we consider both *video recommendation* and *comment recommendation*, in which comment recommendation is helpful to enhance the user engagement and improve the user experience. Formally, we denote the sets of users and videos as $\mathcal{U}$ and $\mathcal{V}$ respectively. Each video $v \in \mathcal{V}$ is associated with a textual title and a list of corresponding comments, denoted as C_v. For a user $u \in \mathcal{U}$, there exist two chronological interaction sequences: (1) *Video interactions* $\mathcal{S}_v = [v_1, v_2, \dots, v_n]$, where v_i is the i-th watched video by the user. (2) *Comment interactions* $\mathcal{S}_c = [r_1, r_2, \dots, r_m]$, where r_j is the j-th comment interaction record of the user, and each record contains a set of comments that the user has interacted with below video $\tilde{v}_j$, denoted as $(c_j^1, \dots, c_j^k)$ In addition, we denote the original videos of these comments as $\mathcal{S}_{\tilde{v}} = [\tilde{v}_1, \tilde{v}_2, \dots, \tilde{v}_m]$ for further elaboration. In KuaiShou, a user can only engage in the comment thread without interacting with the corresponding video. Thus, the videos in $\mathcal{S}_v$ and $\mathcal{S}_{\tilde{v}}$ may not be the same, but likely share overlapped ones.

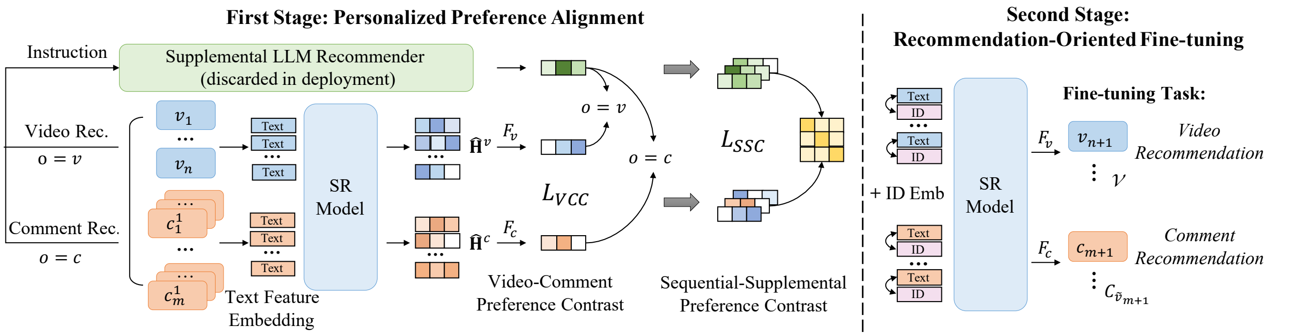

Figure 2: The overall framework of our LSVCR. v_i **and** c_j^k **denote the video and comment.** $o \in \{v, c\}$ **represents the instruction/task objective (video or comment recommendation) of each data instance. "SR model" is the short of Sequential Recommendation Model.** F_v **and** F_c **denote the preference extraction function corresponding to each task in Eq.** (8) **and Eq.** (9)**.**

Based on the above notations, we focus on the tasks of sequential video and comment recommendation. Regarding video recommendation, given the video interaction history $\mathcal{S}_v$ and the auxiliary comment interaction history $\mathcal{S}_c$, the task aims to predict the next video that the user may be interested in. We attempt to enhance user preference modeling with feedback on video comments. For comment recommendation, the primary focus is on $\mathcal{S}_c$, with $\mathcal{S}_v$ serving as a supplement. When the user is watching video $\tilde{v}_{m+1}$ and clicks the "comment"' button, the task objective is to predict the comments that the user may interact with for the current video and display personalized comment content to the user.

3 METHODOLOGY

In this section, we present **LSVCR** for joint video and comment recommendation. Its overall framework is depicted in Figure 2. We first introduce the SR model for sequential user preference modeling in Section 3.1. Then we incorporate the supplemental LLM recommender in Section 3.2 to enhance preference modeling. After that, a personalized preference alignment approach is proposed in Section 3.3 to align the SR model with the supplemental LLM recommender. Finally, we fine-tune the alignment-enhanced SR model for video or comment recommendation tasks in Section 3.4.

3.1 Sequential User Preference Modeling

As we mentioned in Section 1, it is resource-intensive to directly employ LLM as a recommender for online service [44]. Thus, we consider a more economical way to utilize LLM as an offline text feature encoder and utilize a sequential recommendation model (called SR model) as the backbone for user behavior modeling [19, 47].

3.1.1 Text Feature Embedding with LLM. Following our problem formulation in Section 2, we use the user's historical interaction sequences, *i.e.,* $\mathcal{S}_v$ and $\mathcal{S}_c$, to construct inputs. First, we employ LLM to encode the titles and comments involved in the inputs. Then, we combine the output embeddings as the representation of each interaction, which can be formally written as:

$$\mathbf{z}_i^v = [\mathrm{LLM}(t_i) \,||\, \mathrm{LLM}(c_i)]\mathbf{W}_1, \tag{1}$$

$$\mathbf{z}_j^c = [\mathrm{LLM}(t_j) \,||\, \mathrm{MEAN}\left(\mathrm{LLM}(c_j^1), \ldots, \mathrm{LLM}(c_j^k)\right)]\mathbf{W}_1, \tag{2}$$

where t_i and c_i denote the title and a popular comment of video v_i in $\mathcal{S}_v$ respectively. t_j denotes the title of video $\tilde{v}_j$ corresponding to the interaction record r_j in $\mathcal{S}_c$. "$||$" denotes concatenation operation, $\mathbf{W}_1 \in \mathbb{R}^{2d \times d}$ is a trainable parameter for linear projection, and $\mathrm{MEAN}(\cdot)$ denotes mean pooling. $\mathbf{z}_i^v$ and $\mathbf{z}_j^c$ represent the text feature embeddings for one video/comment interaction, and the text feature embeddings of interaction sequences are denoted as $\mathbf{Z}^v \in \mathbb{R}^{n \times d}$ and $\mathbf{Z}^c \in \mathbb{R}^{m \times d}$.

3.1.2 Randomized Positional Encoding for Length Extension. In order to model the sequential pattern of user interaction histories, we incorporate two learnable positional encoding matrices, denoted as $\mathbf{P}^v \in \mathbb{R}^{L_v \times d}$ and $\mathbf{P}^c \in \mathbb{R}^{L_c \times d}$, respectively. Here L_v and L_c represent the maximum lengths of the historical interaction sequences. In conventional sequential recommendation, the length can generally be set longer [18, 24, 33]. However, in the personalized preference alignment training stage, it is set shorter which is limited by LLM input length and computational consumption [13, 49, 53]. Inspired by the length extension technique in the NLP field [32], we randomly select an ordered subset of positional encodings during the alignment stage, while applying the entire positional encodings in subsequent fine-tuning and inference processes. Formally, the positional encodings of $\mathcal{S}_v^u$ are sampled as follows:

$$\widetilde{\mathbf{P}}^v = [\mathbf{p}_{i_1}, \mathbf{p}_{i_2}, \ldots, \mathbf{p}_{i_n}], \tag{3}$$

where $\mathbf{p}_{i_k}$ denotes the i_k-th positional encoding in $\mathbf{P}^v$. Each index i_k is sampled from $\{1, \ldots, L_v\}$ without replacement, and all sampled indices are arranged in ascending order. Additionally, the positional encodings $\widetilde{\mathbf{P}}^c$ for $\mathcal{S}_c^u$ are sampled using a similar process.

3.1.3 Sequence Representation Learning. Given the user interaction histories and sampled position encodings, we use the Transformer backbone [35] to model video and comment interaction sequences. The above sequence embeddings (*i.e.,* $\mathbf{E}^v$ and $\mathbf{E}^c$) and positional encodings (*i.e.,* $\widetilde{\mathbf{P}}^v$ and $\widetilde{\mathbf{P}}^c$) are summed up as inputs of Transformer layers. The calculations can be formally written as:

$$\mathbf{H}^v = \mathrm{Transformer}^v(\mathbf{E}^v + \widetilde{\mathbf{P}}^v), \tag{4}$$

$$\mathbf{H}^c = \mathrm{Transformer}^c(\mathbf{E}^c + \widetilde{\mathbf{P}}^c), \tag{5}$$

where $\mathbf{H}^v \in \mathbb{R}^{n\times d}$ and $\mathbf{H}^c \in \mathbb{R}^{m\times d}$ are sequence representations of user video and comment interaction histories.

Next, a pair of fusion encoders based on multi-head attention (denoted by $\mathrm{MHA}(\mathbf{Q},\mathbf{K},\mathbf{V})$) are used to achieve the cross-fusion between two user historical sequences:

$$\widetilde{\mathbf{H}}^v = \mathrm{MHA}^v(\mathbf{H}^\mathbf{v},\mathbf{H}^\mathbf{c},\mathbf{H}^\mathbf{c}), \quad \widehat{\mathbf{H}}^v = [\mathbf{H}^v||\widetilde{\mathbf{H}}^v]\mathbf{W}_2, \tag{6}$$

$$\widetilde{\mathbf{H}}^c = \mathrm{MHA}^c(\mathbf{H}^\mathbf{c},\mathbf{H}^\mathbf{v},\mathbf{H}^\mathbf{v}), \quad \widehat{\mathbf{H}}^c = [\mathbf{H}^c||\widetilde{\mathbf{H}}^c]\mathbf{W}_3, \tag{7}$$

where $\mathbf{W}_2 \in \mathbb{R}^{2d\times d}$ and $\mathbf{W}_3 \in \mathbb{R}^{2d\times d}$ are learnable parameters for representation fusion, and $\widehat{\mathbf{H}}^v \in \mathbb{R}^{n\times d}$ and $\widehat{\mathbf{H}}^c \in \mathbb{R}^{m\times d}$ represent the fused video and comment sequence representations.

3.1.4 Preference Extraction from Sequence Representations. For the tasks of video and comment recommendation, we employ different methods to extract user interests from sequence representations. Specifically, we exploit additive attention [3] in video recommendation to manage the importance of different representations:

$$\mathbf{s}^v = F_v(\widehat{\mathbf{H}}^v) = \sum_{i=1}^{n} \alpha_i \mathbf{h}_i^v, \quad \alpha_i = \frac{\exp(f(\mathbf{h}_i^v))}{\sum_{k=1}^{n}\exp(f(\mathbf{h}_k^v))}, \tag{8}$$

where $\mathbf{h}_i^v$ is the i-th vector of the matrix $\widehat{\mathbf{H}}^v$, f is an attention weight function implemented as a multi-layer perceptron (MLP), and $\mathbf{s}^v$ denotes the representation of user video preference.

As for the comment recommendation task, which involves the video that the user is watching, we regard the title embedding (denoted as $\mathbf{e}_{m+1}^t = \mathrm{LLM}(t_{m+1})$) of the current video as a query for representation aggregation:

$$\mathbf{s}^c = F_c(\widehat{\mathbf{H}}^c) = \sum_{j=1}^{m} \beta_j \mathbf{h}_j^c, \quad \beta_j = \frac{\exp(g(\mathbf{e}_{m+1}^t,\mathbf{h}_j^c))}{\sum_{k=1}^{m}\exp(g(\mathbf{e}_{m+1}^t,\mathbf{h}_k^c))}. \tag{9}$$

where $\mathbf{h}_j^c$ is the j-th vector of $\widehat{\mathbf{H}}^c$, $g(\cdot)$ is an attention weight function (similar to multi-head attention) with $\mathbf{e}_{m+1}^t$ as the query, $\mathbf{s}^c$ denotes the representation of user comment preference on video $\tilde{v}_{m+1}$.

3.2 Enhanced Preference Modeling via LLM

In recommendation scenarios, LLM generally plays two roles: as a text feature encoder (Section 3.1) or a recommender [4, 49]. The former offers greater flexibility and efficiency, while the latter is beneficial for the more effective utilization of LLM's excellent semantic understanding and knowledge reasoning capabilities. Therefore, in this section, we further perform enhanced user preference modeling through a LLM recommender to improve the semantics of the SR model. It is worth noting that this component is only included *in the alignment training stage* and does not exist in the subsequent fine-tuning and practical deployment.

Following previous works [9, 20, 49], we format the user interaction sequences in the form of textual instructions and utilize LLM for generative recommendation. Specifically, the instruction templates for two tasks are as follows:

Video Recommendation: *A user has interacted with the following videos:* [*VideoInterList*]*, and his/her comment interaction history is as follows:* [*CommentInterList*]*. Please recommend the next video for him/her based on these interaction histories:*
Comment Recommendation: *A user's comment interaction history is as follows:* [*CommentInterList*]*, and he/she has interacted with the following videos:* [*VideoInterList*]*. Now he/she is watching a video:* [*VideoTitle*]*. Please generate a potentially interactive comment based on the above historical interactions and the current video content:*

In the instruction templates, "[*VideoInterList*]" represents a list of video interactions, where each record contains a video title and a popular comment, *e.g.,* "Title: *Review of the 2022 Beijing Winter Olympics*; Popular comment: *Figure skating is so beautiful!*". As can be seen from this example, the associated comment provides additional illustration for this video, with a special focus on figure skating. "[*CommentInterList*]" represents a list of comment interactions, where each record contains the comments that the user has interacted with and the corresponding video title. For the video recommendation task, LLM is instructed to generate the title of the next video. As for the comment recommendation task, it is based on the video that the user is watching (*i.e.,* "[*VideoTitle*]"). LLM is instructed to generate the comment that the user may interact with. Finally, we denote the instruction as $I(\mathcal{S}_v,\mathcal{S}_c)$ and feed it into LLM:

$$\tilde{\mathbf{s}} = \mathrm{LLM}(I(\mathcal{S}_v,\mathcal{S}_c))[-1], \tag{10}$$

where $[-1]$ means taking the hidden state corresponding to the final token of $I(\mathcal{S}_v,\mathcal{S}_c)$, and $\tilde{\mathbf{s}}$ is the enhanced preference representation from the LLM recommender.

3.3 Personalized Preference Alignment

In our approach, the SR model and the supplemental LLM recommender learn user preference representations from different perceptions. To enhance the SR model by incorporating more informative preference knowledge from the LLM recommender, we introduce personalized preference alignment as the first stage of training. Specifically, we propose two alignment perspectives: sequential-supplemental preference contrast and video-comment preference contrast, to align preference semantics from both components.

3.3.1 Sequential-Supplemental Preference Contrast. Contrastive learning is typically used to align instance representations across different latent spaces [6, 12]. In our LSVCR, the primary objective is to achieve alignment of personalized preference representations between the SR model and the supplemental LLM recommender. Specifically, we first perform data augmentation via comment diversification on inputs and then employ InfoNCE [15] with in-batch negatives to align the output representations of two models.

Input augmentation via comment diversification. In our scenario, there is abundant and diverse comment data, which can be leveraged for data augmentation. Specifically, we apply comment diversification with a certain probability to augment user video and comment interaction histories: (1) For each video interaction, we select two distinct popular comments to pair with the same video as inputs for the SR model and the LLM recommender. (2) For each comment interaction, we sample a couple of virtual interaction comments and mix them with the real interaction comments. Then, the mixed interaction comments are randomly allocated to the inputs of two components.

Preference representation contrastive learning. In the personalized preference alignment stage, we jointly learn video and comment recommendation tasks. For each data instance, the sequential preference representation of the SR model is calculated by Eq. (8) or Eq. (9). For the sake of brevity, here we no longer distinguish between the two tasks, and the representation is denoted as $\mathbf{s}$. Additionally, by the approach mentioned in Section 3.2, we obtain the enhanced preference representation of the LLM recommender, denoted as $\tilde{\mathbf{s}}$. The sequential-supplemental preference contrastive loss can be formulated as:

$$\mathcal{L}_{SSC} = \frac{1}{2}\left(\text{InfoNCE}(\mathbf{s}, \tilde{\mathbf{s}}, \mathcal{R}_{\tilde{\mathbf{s}}}) + \text{InfoNCE}(\tilde{\mathbf{s}}, \mathbf{s}, \mathcal{R}_{\mathbf{s}})\right), \quad (11)$$

where $\mathcal{R}_{\mathbf{s}}$ and $\mathcal{R}_{\tilde{\mathbf{s}}}$ denote batch preference representations that are generated by the SR model and the LLM recommender, respectively. InfoNCE$(\cdot,\cdot,\cdot)$ represents InfoNCE loss, which can be written as:

$$\text{InfoNCE}(\mathbf{x}, \mathbf{y}^+, O_{\mathbf{y}}) = -\log \frac{\exp(\cos(\mathbf{x}, \mathbf{y}^+)/\tau)}{\sum_{\mathbf{y}\in O_{\mathbf{y}}} \exp(\cos(\mathbf{x}, \mathbf{y})/\tau)}, \quad (12)$$

where $\mathbf{x}$ and $\mathbf{y}^+$ denote a pair of positive instances for contrastive learning. $O_{\mathbf{y}}$ denotes a set of contrastive instances, consisting of both positive and negative samples. τ is a temperature coefficient, and $\cos(\cdot,\cdot)$ represents the cosine similarity function.

3.3.2 Video-Comment Preference Contrast. Furthermore, we distinguish between user preferences that are extracted from video and comment sequence representations, based on the enhanced preference representation provided by the LLM recommender. The core idea is that when LLM is instructed to make video recommendations, the preference representation (*i.e.,* $\tilde{\mathbf{s}}$) should be aligned with that extracted from the video sequence (*i.e.,* $\widehat{\mathbf{H}}^v$), rather than from the comment sequence (*i.e.,* $\widehat{\mathbf{H}}^c$), and vice versa. Formally, we denote the instruction/task objective (video or comment recommendation) of the instance as $o \in \{v, c\}$. The video-comment preference contrastive loss is formulated as follows:

$$\mathcal{L}_{VCC} = -\log \frac{\exp(\cos(\tilde{\mathbf{s}}, F_o(\widehat{\mathbf{H}}^o))/\tau)}{\sum_{\tilde{o}\in\{v,c\}} \exp(\cos(\tilde{\mathbf{s}}, F_o(\widehat{\mathbf{H}}^{\tilde{o}}))/\tau)}, \quad (13)$$

where F_o denotes the preference extraction function corresponding to each instance in Eq. (8) (*i.e.,* F_v) or Eq. (9) (*i.e.,* F_c).

3.3.3 Alignment Training Objective. In our approach, we consider the next video title as the target text for video recommendation and the comment that the user may interact with as the target text for comment recommendation. LLM is trained by following a typical procedure of instruction tuning [37]:

$$\mathcal{L}_{LM} = -\sum_{i=1}^{|T|} \log P(T_i | I, T_{<i}), \quad (14)$$

where I and T represent the instruction and target text, T_i is the i-th token of T, and $T_{<i}$ denotes the tokens before T_i. In terms of the SR model, we harness the following loss for optimization:

$$\mathcal{L}_{SR} = \text{InfoNCE}(\mathbf{s}, \mathbf{e}, \mathcal{T}_{\mathbf{e}}) \quad (15)$$

where $e = \text{LLM}(T)$, and $\mathcal{T}_{\mathbf{e}}$ denotes a batch of target embeddings.

In the end, the overall loss of personalized preference alignment, which combines the above objectives, is as follows:

$$\mathcal{L}_{ALI} = \mathcal{L}_{LM} + \lambda\mathcal{L}_{SR} + \mu(\mathcal{L}_{SSC} + \mathcal{L}_{VCC}), \quad (16)$$

where λ and μ are hyperparameters for the trade-off between various objectives.

3.4 Recommendation-Oriented Fine-tuning

After the personalized preference alignment stage, we further fine-tune the alignment-enhanced SR model for video and comment recommendation tasks. The supplemental LLM recommender will be discarded during the fine-tuning stage and practical deployment.

3.4.1 Fine-tuning. Since the first stage only utilizes textual features, we incorporate video ID embedding into our framework during the fine-tuning stage to enhance the overall performance. The particular method is to add the corresponding ID embedding to the inputs of the SR model in Eq. (1) and Eq. (2). Moreover, to ensure representation compatibility between ID and text embeddings, we introduce the following ID-text regularization loss:

$$\mathcal{L}_{REG} = \text{InfoNCE}(\mathbf{e}_i^{id}, \mathbf{z}_i^v, \mathcal{Z}_i^v) + \text{InfoNCE}(\mathbf{e}_j^{id}, \mathbf{z}_j^c, \mathcal{Z}_j^c), \quad (17)$$

where $\mathbf{e}_i^{id}$ and $\mathbf{e}_j^{id}$ denote the ID embeddings of video v_i in $\mathcal{S}_v$ and $\tilde{v}_j$ in $\mathcal{S}_{\tilde{v}}$, respectively. The sets $\mathcal{Z}_i^v$ and $\mathcal{Z}_j^c$ consist of both the positive and negative text embeddings sampled from the batch data.

For the video recommendation task, our objective is similar to target text contrastive learning, but we use all videos as candidates instead of in-batch negatives. The loss can be written as:

$$\mathcal{L}_{VRT} = \text{InfoNCE}(\mathbf{s}^v, (\mathbf{e}_{n+1}^{id} + \mathbf{e}_{n+1}^t), \mathcal{V}), \quad (18)$$

where $\mathcal{V}$ is the entire set of candidate videos. $\mathbf{e}_{n+1}^t = \text{LLM}(t_{n+1})$ represents the title embedding of the next video v_{n+1}.

Finally, our overall optimization loss function in the fine-tuning stage can be denoted as follows:

$$\mathcal{L}_{FT} = \mathcal{L}_{VRT} + \eta\mathcal{L}_{REG}, \quad (19)$$

where η is the hyperparameter for the regularization loss. Regarding the comment recommendation task, we use comment preference representation $\mathbf{s}^c$ for prediction, and all candidates (*i.e.,* $C_{\tilde{v}_{m+1}}$) are selected from comments of video $\tilde{v}_{m+1}$ that the user is watching.

3.4.2 Time Complexity Analysis. In practical deployment, we retain the fine-tuned SR model for recommendation, and discard the supplemental LLM recommender. Besides, the textual information associated with the videos and comments is encoded offline to be directly accessed during online inference. Therefore, we just discuss the time complexity of the subsequent sequential user preference modeling process. Formally, the time consumption of Transformer layers in Eq. (4) and Eq. (5) is $O(n^2d + m^2d)$, where d denotes the model dimension, n and m denote the lengths of user video and comment interaction histories, respectively. Furthermore, $O(nd^2 + md^2)$ is the complexity of intermediate linear operations. And the time used by fusion encoders in Eq. (6) and Eq. (7) is $O(mnd)$. In addition, some low-order time complexities are ignored, such as the $O(nd)$ or $O(md)$ in preference extraction (Section 3.1.4). Overall, the time complexity of sequential user preference modeling is $O(n^2d + m^2d + nd^2 + md^2 + mnd)$. In the end, the time consumed for calculating scores with candidates is $O(Nd)$, where N denotes the number of candidates. In general, when it comes to online deployment, our LSVCR model has comparable time complexity to mainstream sequential recommendation models (*e.g.,* SASRec [24], BERT4Rec [33]).

Table 1: Statistics of the preprocessed dataset. "V" and "C" denote video and comment interactions, respectively.

#Users	#Videos	#Comments	#Interactions-V	#Interactions-C
19,691	100,772	14,163,262	789,986	100,593

4 EXPERIMENT

In this section, we first set up the experiments and then present the results and analyses of our proposed LSVCR.

4.1 Experiment Setup

4.1.1 Dataset. We evaluate the proposed framework on the large-scale industrial dataset constructed from a popular Chinese micro-video platform KuaiShou for video and comment recommendation tasks. To be specific, the dataset is built from the interaction logs of 19,691 users from October 24 to October 31, 2023. For data preprocessing, we split the user interactions into training, validation, and testing sets according to the timestamp. The data from the last two days is used for validation and testing respectively, while the remaining data is utilized to train our framework. The detailed statistics of the dataset are summarized in Table 1.

4.1.2 Baseline Models. We compare LSVCR with the following various competitive baselines:

Video recommendation. The baselines for video recommendation can be divided into three categories: (1) Traditional sequential recommendation models: **Caser** [34] employs CNN to model user preference through horizontal and vertical convolutional filters. **GRU4Rec** [18] is a RNN-based model that utilizes GRU to encode historical behavior sequences. **SASRec** [24] uses a unidirectional Transformer for sequence modeling and next item prediction. **BERT4Rec** [33] adopts a bidirectional Transformer with the mask prediction objective to model the item sequence. **NARM** [25] combines the GRU with an attention mechanism to capture users' interests. (2) Multi-behavior models: **DMT** [14] exploits multiple Transformers with MoEs to model diverse user behavior sequences. **MBHT** [45] empowers the Transformer with a multi-behavior hypergraph to capture diverse multi-behavior dependencies. (3) Text-enhanced models: **FDSA** [50] applies separate self-attention networks to model item-level and feature-level sequences. **S^3-Rec** [54] proposes pre-training with mutual information maximization to learn the correlation between items and attributes. **UniSRec** [19] incorporates textual item embeddings with a MoE-enhanced adapter to learn universal sequence representations.

Comment recommendation. The baseline methods for comment recommendation include: (1) Traditional recommendation model: **DSSM** [22] uses dual encoders to separately encode user and candidate comments. (2) Sequential modeling methods: **GRU** [7] intuitively utilizes GRU to derive short- and long-term user representations [2, 18, 25]. **ATT** [35] employs an additive attention (ATT) mechanism to model the user comment interaction sequence [38]. **MHA** [35] adopts multi-head attention (MHA) and additive attention to obtain the representation of the user comment interaction history [38, 39]. (3) Query-based models: **ZAM** [1] introduces a zero vector into the attention mechanism to dynamically control the influence of user interaction history on product search. **TEM** [5] uses Transformer to encode the query and user interaction sequence for more flexible personalization adaptation. For all comment recommendation baselines, we adopt trainable BERT [10] as their text encoder. Furthermore, we consider the title embedding of the video being watched by the user as the query for query-based models.

Table 2: Performance comparison of different methods for video recommendation. The best and second-best results are highlighted in bold and underlined font, respectively.

Methods	Recall@5	Recall@10	NDCG@5	NDCG@10	MRR
Caser	0.2230	0.2886	0.1601	0.1813	0.1481
GRU4Rec	0.2249	0.2896	0.1647	0.1856	0.1535
BERT4Rec	0.2193	0.2815	0.1612	0.1814	0.1504
SASRec	0.2311	0.2953	0.1706	0.1913	0.1592
NARM	0.2372	0.2989	0.1742	0.1942	0.1617
DMT	0.2236	0.2903	0.1633	0.1848	0.1522
MBHT	0.2421	0.3038	0.1780	0.1980	0.1651
FDSA	0.2277	0.2922	0.1667	0.1875	0.1551
S^3-Rec	0.2226	0.2899	0.1619	0.1837	0.1509
UniSRec	<u>0.2570</u>	<u>0.3159</u>	<u>0.1875</u>	<u>0.2135</u>	<u>0.1737</u>
LSVCR	**0.2719**	**0.3322**	**0.2037**	**0.2233**	**0.1893**

Table 3: Performance comparison of various models on the comment recommendation task.

Methods	Recall@5	Recall@10	NDCG@5	NDCG@10	MRR
DSSM	0.2092	0.3386	0.1447	0.1884	0.1573
GRU	0.2386	0.3708	0.1644	0.2094	0.1750
ATT	0.2284	0.3653	0.1537	0.2004	0.1634
MHA	0.2339	<u>0.3768</u>	0.1646	0.2131	0.1775
ZAM	0.2516	0.3699	0.1879	0.2283	0.2031
TEM	<u>0.2533</u>	0.3685	<u>0.1898</u>	<u>0.2302</u>	<u>0.2057</u>
LSVCR	**0.2821**	**0.3901**	**0.2175**	**0.2541**	**0.2303**

4.1.3 Evaluation Settings. To evaluate the performance of video and comment recommendation tasks, we adopt three widely used metrics Recall@K, NDCG@K, and MRR, where K is set to 5 and 10. We perform negative sampling evaluations for both tasks. For video recommendation, we pair the ground-truth item with 99 randomly sampled items. Regarding comment recommendation, the candidate number is limited to 100. If the number of comments below the current video is less than 100, all comments will be used. Conversely, negative sampling will be conducted until there are 100 candidates.

4.1.4 Implementation Details. For the personalized preference alignment stage, we utilize ChatGLM3 [11, 48] as our LLM backbone and perform low-rank adaptation based on LoRA [21]. We employ the AdamW optimizer with a learning rate of 3e-4 for optimization. With the application of data parallelism and gradient accumulation, the total batch size reaches 128. The loss coefficients, λ and μ, are set to 1 and 0.5 respectively. The temperature coefficient τ is assigned a value of 0.07. The maximum lengths of user video and comment interaction histories are set to 20 for the alignment stage and 50 for

Table 4: Ablation study of LSVCR on the tasks of video and comment recommendation.

Methods	Video Rec.		Comment Rec.	
	Recall@10	NDCG@10	Recall@10	NDCG@10
LSVCR	**0.3322**	**0.2233**	**0.3901**	**0.2541**
w/o $\mathcal{L}_{SSC}$	0.3211	0.2120	0.3822	0.2487
w/o $\mathcal{L}_{VCC}$	0.3301	0.2211	0.3892	0.2528
w/o ComDiv	0.3280	0.2189	0.3853	0.2525
w/o RandPosi	0.3267	0.2178	0.3881	0.2533
w/o Alignment	0.3071	0.1998	0.3505	0.2176
w/o $\mathcal{L}_{REG}$	0.3258	0.2171	0.3733	0.2428

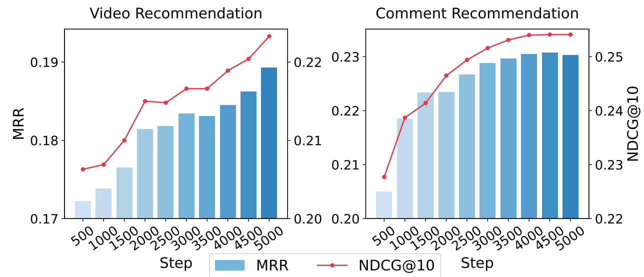

Figure 3: Performance over different alignment steps.

the fine-tuning stage. Finally, we conduct a total of 5000 alignment steps. For the recommendation-oriented fine-tuning stage, the embedding dimension is set to 64, and the learning rate is tuned in {0.005, 0.003, 0.001, 0.0005, 0.0001}, which are consistent with all baseline models. The hyperparameter η is tuned in {0.1, 0.5 }. More details can be found in our implementation code[2].

4.2 Overall Results

Video recommendation performance. The overall results of the video recommendation task are shown in Table 2. Based on the results, we have the following observations: Intuitively incorporating heterogeneous comment interaction behaviors (*i.e.,* DMT) is not inherently better than traditional sequential recommendation models. In contrast, the hypergraph-enhanced Transformer (*i.e.,* MBHT) excels in modeling multi-behavior dependencies and outperforms all traditional sequential recommendation baselines. In terms of text-enhanced methods, UniSRec achieves the best performance, thanks to its contrastive learning pre-training, which leads to high-quality universal sequence representation. As for FDSA and S^3-Rec, they do not benefit from the extra text embeddings. One possible reason is the poor quality of text present on our platform. Consequently, addressing this issue requires the adoption of a robust text semantic modeling approach. Finally, LSVCR consistently maintains the best results compared to the baseline methods.

Comment recommendation rerformance. The results of different methods on the comment recommendation task are shown in Table 3. From these results, we can find: DSSM, which does not involve user comment interaction history, exhibits inferior performance in comparison to other methods due to the lack of personalized user information. The personalized search methods that apply the title of the current video as a query (*i.e.,* ZAM, TEM) demonstrate better performance than the simple sequential modeling methods (*i.e.,* GRU, ATT, MHA). Based on the findings, it is evident that the personalized comment recommendation task requires user interests closely aligned with the specific video content, rather than overall generalized user comment preference. By comparing our approach with all baselines, it is clear that LSVCR achieves the most effective results.

All in all, our proposed approach shows significant improvements compared to existing baselines on both video and comment recommendation tasks. This phenomenon indicates that joint modeling of videos and comments can simultaneously enhance user video and comment preference learning. Based on personalized preference alignment, we effectively integrate the powerful semantic understanding and knowledge reasoning capabilities of the supplemental LLM recommender into the SR model. Furthermore, these results also show that the two-stage training paradigm can coordinate well for improving overall recommendation.

4.3 Ablation Study

In order to investigate how the proposed techniques affect the final performance, we conducted an ablation study on personalized video and comment recommendation tasks. Specifically, we consider the following six variants of LSVCR: (1) w/o $\mathcal{L}_{SSC}$ without the sequential-supplemental preference contrastive loss (Eq. (11)). (2) w/o $\mathcal{L}_{VCC}$ without the video-comment preference contrastive loss (Eq. (13)). (3) w/o ComDiv without the comment diversification introduced in Section 3.3.1 and maintaining inputs consistency between the two components. (4) w/o RandPosi without the randomized positional encoding introduced in Section 3.1.2. It expands the positional encoding during fine-tuning and randomly initializes the expanded part. (5) w/o Alignment without the personalized preference alignment and learning the SR model directly. (6) w/o $\mathcal{L}_{REG}$ without the ID-text regularization loss (Eq. (17)).

The results presented in Table 4 clearly demonstrate that removing any of the above techniques would lead to a decline in the overall effect. Concretely, the absence of personalized preference alignment (*i.e.,* w/o Alignment) results in significantly weaker performance. Moreover, both contrastive losses (*i.e.,* $\mathcal{L}_{SSC}$ and $\mathcal{L}_{VCC}$) contribute to the downstream tasks. These observations highlight the importance of our proposed various techniques in integrating the merits of the SR model and the supplemental LLM recommender.

4.4 Further Analysis

4.4.1 Performance Comparison w.r.t. Alignment Steps. In this section, we investigate how the number of personalized preference alignment steps affects performance. In the experiment, we utilize alignment-enhanced SR models with different numbers of alignment steps for fine-tuning. From the results in Figure 3, we can see that LSVCR primarily benefits from about the first 3000 alignment steps. For the video recommendation task, there is still considerable improvement in performance with continued training, which may

[2]https://github.com/RUCAIBox/LSVCR/

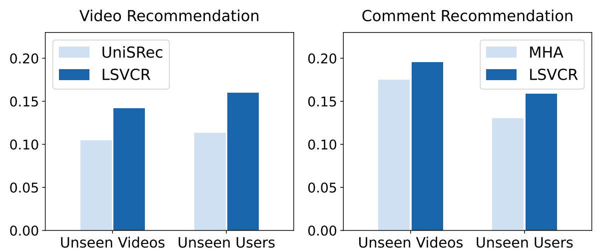

Figure 4: Performance comparison on unseen videos and users. We present the results of NDCG@10.

be attributed to the availability of more diverse video recommendation data. For the comment recommendation task, a duration of 3000 steps for alignment is deemed sufficient, which adequately showcases the superiority of LLM in swiftly adapting to new tasks.

4.4.2 Performance Comparison w.r.t. Unseen Videos/Users. In this part, we attempt to apply our approach to videos or users that were not previously seen during training. Specifically, we first discard the ID embedding involved in our approach and baselines (*i.e.,* UniSRec, MHA), allowing adaptation in the presence of unseen videos. Then, we introduce two special datasets: (1) Unseen Videos: Select instances from the original test set where the target video never appeared during training. (2) Unseen Users: Select additional users who have never been seen and organize their interaction logs into a new dataset. As shown in Figure 4, we can observe that LSVCR achieves more outstanding performance compared to the baselines on more challenging datasets in both video and comment recommendation tasks. These findings indicate that with the assistance of personalized preference alignment between the SR model and the supplemental LLM recommender, LSVCR demonstrates robust capability in user preference modeling.

4.5 Online A/B Testing

In order to further verify the effectiveness of LSVCR, we deploy it to the Kuaishou platform for two weeks of online A/B testing. Specifically, we incorporate our method into the current recommendation workflow for comparison, and the experimental results are shown in Table 5. Due to the limited computational resources, we select 20K users for testing and randomly split them into two groups. We consider two metrics to measure user engagement: (1) *Watch Time*: Average video/comment watch time. (2) *Interaction Num.*: Average number of video/comment interactions. Video interactions include "like", "collect" and "share". Comment interactions include "like" and "reply". The results indicate that LSVCR achieves significant improvements on both video and comment recommendation tasks. It is worth noting that the promotion of comment recommendation is much more effective compared to video recommendation which may be attributed to the integration of LLM's powerful understanding capability on natural language in the proposed LSVCR. Overall, our framework reveals the great potential of LLM for joint video and comment recommendation on online video platforms.

Table 5: Results of online A/B testing on KuaiShou.

Online Metrics	Video	Comment
Watch Time	+0.3649%	+4.1264%
Interaction Num.	+0.7821%	+1.3557%

5 RELATED WORK

Sequential recommendation Sequential recommendation aims to mine users' personalized preferences and recommend the next item for each user. This field has attracted a lot of studies due to its advantages in capturing the dynamic characteristics of user behaviors [18, 24, 33]. Early methods follow the Markov Chain assumption and apply matrix factorization to model item-item transfer relationships [17, 31]. Recently, typical works are mostly based on various deep neural networks to better user behavior modeling, including RNN [18, 25], CNN [34, 46], GNN [41, 43], and Transformer [24, 33]. Furthermore, several studies focus on utilizing additional contextual information (*e.g.,* title, description, category) to enhance item sequence modeling [42, 50, 54]. Unlike these works, this paper specifically concentrates on video comments, incorporating abundant textual information and user feedback from comment pages into sequential recommendation.

Text-enhanced recommendation. On various application platforms, items are associated with abundant textual features, which motivates researchers to engage in text-enhanced recommendation models. Recently, pre-trained large language models (LLMs) have gained significant popularity and shown great potential in wide-ranging fields [23, 52]. In the realm of text-enhanced recommendation, LLM generally plays distinct roles: (1) LLM serves as a text feature encoder, allowing the text features to be encoded asynchronously. Subsequently, a conventional recommendation model utilizes knowledge-aware text embedding for recommendation [16, 19, 29, 40, 47]. Although effective, this approach often requires special mechanisms to enhance text embedding for better performance, such as contrastive learning [19, 29], and further embedding optimization [16, 47]. (2) LLM serves as a recommender. This approach typically constructs the user's interaction history and other contexts into textual instructions, and then LLM is instructed to directly generate recommendation targets [4, 20, 44, 49**?**]. More recently, several works leverage various methods to enhance the adaptation of LLM for recommendation scenarios, including incorporating collaborative embedding [26, 27, 51], semantic alignment [53, 55], and so on. These methods fully utilize the excellent capabilities of LLM and achieve astonishing results, but they still cannot avoid the high computational cost and slow inference time caused by LLM. As a comparison, we aim to consolidate the merits of these two approaches through personalized preference alignment in order to achieve effective textual semantics understanding and heterogeneous interaction behavior modeling.

6 CONCLUSION

In this paper, we propose a novel framework called LSVCR, which leverages user interaction histories with both videos and comments for joint video and comment recommendation. There are two key components, namely conventional sequential recommendation (SR) model and supplemental LLM recommender. The former serves as the recommendation backbone to ensure efficiency, and the latter is used to enhance the semantics of the SR model. Furthermore, a two-stage training paradigm is introduced to integrate the merits of both components. For the personalized preference alignment stage, we jointly learn the video and comment recommendation tasks while performing preference alignment between the two components. For the recommendation-oriented fine-tuning stage, we discard the supplemental LLM recommender and further fine-tune the alignment-enhanced SR model to improve performance on video and comment recommendation tasks. Offline evaluation of large-scale industrial data and online A/B testing on the KuaiShou platform demonstrated the effectiveness of LSVCR. In particular, we achieve an overall watch time gain of 4.13% in the comment end. For future work, we will explore more efficient ways of utilizing LLM in joint video and comment recommendation. Additionally, we also attempt to incorporate a wider range of video and comment interaction behaviors into our framework.

REFERENCES

[1] Qingyao Ai, Daniel N. Hill, S. V. N. Vishwanathan, and W. Bruce Croft. 2019. A Zero Attention Model for Personalized Product Search. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019.* 379–388.

[2] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural News Recommendation with Long- and Short-term User Representations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers.* 336–345.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.*

[4] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023.* 1007–1014.

[5] Keping Bi, Qingyao Ai, and W. Bruce Croft. 2020. A Transformer-based Embedding Model for Personalized Product Search. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020.* 1521–1524.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119).* 1597–1607.

[7] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL.* 1724–1734.

[8] Zeyu Cui, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. M6-Rec: Generative Pretrained Language Models are Open-Ended Recommender Systems. *CoRR* abs/2205.08084 (2022). arXiv:2205.08084

[9] Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering ChatGPT's Capabilities in Recommender Systems. In *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023.* 1126–1132.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers).* 4171–4186.

[11] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022.* 320–335.

[12] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021.* 6894–6910.

[13] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5). In *RecSys '22: Sixteenth ACM Conference on Recommender Systems, Seattle, WA, USA, September 18 - 23, 2022.* 299–315.

[14] Yulong Gu, Zhuoye Ding, Shuaiqiang Wang, Lixin Zou, Yiding Liu, and Dawei Yin. 2020. Deep Multifaceted Transformers for Multi-objective Ranking in Large-Scale E-commerce Recommender Systems. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020.* 2493–2500.

[15] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010 (JMLR Proceedings, Vol. 9).* 297–304.

[16] Jesse Harte, Wouter Zorgdrager, Panos Louridas, Asterios Katsifodimos, Dietmar Jannach, and Marios Fragkoulis. 2023. Leveraging Large Language Models for Sequential Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023.* 1096–1102.

[17] Ruining He and Julian J. McAuley. 2016. Fusing Similarity Models with Markov Chains for Sparse Sequential Recommendation. In *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain.* 191–200.

[18] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based Recommendations with Recurrent Neural Networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings.*

[19] Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2022. Towards Universal Sequence Representation Learning for Recommender Systems. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022.* 585–593.

[20] Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian J. McAuley, and Wayne Xin Zhao. 2023. Large Language Models are Zero-Shot Rankers for Recommender Systems. *CoRR* abs/2305.08845 (2023). arXiv:2305.08845

[21] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022.*

[22] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry P. Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013.* 2333–2338.

[23] Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xin Zhao, and Ji-Rong Wen. 2023. StructGPT: A General Framework for Large Language Model to Reason over Structured Data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023.* 9237–9251.

[24] Wang-Cheng Kang and Julian J. McAuley. 2018. Self-Attentive Sequential Recommendation. In *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018.* 197–206.

[25] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017. Neural Attentive Session-based Recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017.* 1419–1428.

[26] Xinhang Li, Chong Chen, Xiangyu Zhao, Yong Zhang, and Chunxiao Xing. 2023. E4SRec: An Elegant Effective Efficient Extensible Solution of Large Language Models for Sequential Recommendation. *CoRR* abs/2312.02443 (2023). arXiv:2312.02443

[27] Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, and Xiang Wang. 2023. LLaRA: Aligning Large Language Models with Sequential Recommenders. *CoRR* abs/2312.02445 (2023). arXiv:2312.02445

[28] Zihan Lin, Hui Wang, Jingshu Mao, Wayne Xin Zhao, Cheng Wang, Peng Jiang, and Ji-Rong Wen. 2022. Feature-aware Diversified Re-ranking with Disentangled Representations for Relevant Recommendation. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022.* 3327–3335.

[29] Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2023. Representation Learning with Large Language Models for Recommendation. *CoRR* abs/2310.15950 (2023). arXiv:2310.15950

[30] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI 2009,*

Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009. 452–461.

[31] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized Markov chains for next-basket recommendation. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010.* 811–820.

[32] Anian Ruoss, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg, and Joel Veness. 2023. Randomized Positional Encodings Boost Length Generalization of Transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023.* 1889–1903.

[33] 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, CIKM 2019, Beijing, China, November 3-7, 2019.* 1441–1450.

[34] 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, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018.* 565–573.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA.* 5998–6008.

[36] Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023. Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023.* 10052–10065.

[37] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models are Zero-Shot Learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022.*

[38] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Attentive Multi-View Learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019.* 3863–3869.

[39] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Multi-Head Self-Attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019.* 6388–6393.

[40] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Empowering News Recommendation with Pre-trained Language Models. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021.* 1652–1656.

[41] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-Based Recommendation with Graph Neural Networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.* 346–353.

[42] Yueqi Xie, Peilin Zhou, and Sunghun Kim. 2022. Decoupled Side Information Fusion for Sequential Recommendation. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022.* 1611–1621.

[43] 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 *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019.* 3940–3946.

[44] Lanling Xu, Junjie Zhang, Bingqian Li, Jinpeng Wang, Mingchen Cai, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Prompting Large Language Models for Recommender Systems: A Comprehensive Framework and Empirical Analysis. *CoRR* abs/2401.04997 (2024). arXiv:2401.04997

[45] Yuhao Yang, Chao Huang, Lianghao Xia, Yuxuan Liang, Yanwei Yu, and Chenliang Li. 2022. Multi-Behavior Hypergraph-Enhanced Transformer for Sequential Recommendation. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022.* 2263–2274.

[46] Fajie Yuan, Alexandros Karatzoglou, Ioannis Arapakis, Joemon M. Jose, and Xiangnan He. 2019. A Simple Convolutional Generative Network for Next Item Recommendation. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019.* 582–590.

[47] Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. 2023. Where to Go Next for Recommender Systems? ID- vs. Modality-based Recommender Models Revisited. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023.* 2639–2649.

[48] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: An Open Bilingual Pre-trained Model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023.*

[49] Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2023. Recommendation as Instruction Following: A Large Language Model Empowered Recommendation Approach. *CoRR* abs/2305.07001 (2023). arXiv:2305.07001

[50] Tingting Zhang, Pengpeng Zhao, Yanchi Liu, Victor S. Sheng, Jiajie Xu, Deqing Wang, Guanfeng Liu, and Xiaofang Zhou. 2019. Feature-level Deeper Self-Attention Network for Sequential Recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019.* 4320–4326.

[51] Yang Zhang, Fuli Feng, Jizhi Zhang, Keqin Bao, Qifan Wang, and Xiangnan He. 2023. CoLLM: Integrating Collaborative Embeddings into Large Language Models for Recommendation. *CoRR* abs/2310.19488 (2023). arXiv:2310.19488

[52] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. *CoRR* abs/2303.18223 (2023). arXiv:2303.18223

[53] Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. 2023. Adapting Large Language Models by Integrating Collaborative Semantics for Recommendation. *CoRR* abs/2311.09049 (2023). arXiv:2311.09049

[54] 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 *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020.* 1893–1902.

[55] Yaochen Zhu, Liang Wu, Qi Guo, Liangjie Hong, and Jundong Li. 2023. Collaborative Large Language Model for Recommender Systems. *CoRR* abs/2311.01343 (2023). arXiv:2311.01343