FACTORIZED COUPLED SEQUENCE MODELING FOR ROBUST SEQUENTIAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequential prediction in real-world systems often involves coupled sequence pairs, *i.e.* a primary action sequence and a co-evolving environment-level context sequence (*e.g.*, app sub-interfaces and publisher sources, as opposed to item-intrinsic attributes like category or brand). Unlike item-level context, such environment-level context is susceptible to alignment noise: element-wise correspondence between the two sequences is inherently corrupted by delayed feedback and context switching. Existing approaches either ignore the context sequence or encode it jointly with actions at the element level. The former loses useful contextual signals, while the latter can introduce erroneous pairwise information and perturb the learned representations. We factorize the joint sequence model into two interacting sequence models that capture their own temporal dynamics and exchange information only at the representation level. We instantiate this design as *Factorized Coupled Sequence Networks (FCSNet)*, with parallel sequence encoders, cross-sequence representation fusion, and dual prediction heads. Under mean-field posterior and factorized-prior assumptions, its training objective can be interpreted as an approximation to a factored evidence lower bound (ELBO) of the joint likelihood. We further introduce *Conditional Contrastive Regularization (CCR)*, which uses similarity-adaptive weights to emphasize informative contrastive pairs under heavy-tailed sequential data. We validate FCSNet on two domains where coupled sequences with alignment noise naturally arise: e-commerce purchasing and online news browsing. Experiments on one public news benchmark and two industrial e-commerce datasets examine predictive performance and robustness to controlled alignment noise. Production deployment in an e-commerce search system further confirms real-world impact, achieving a 0.04 absolute-point CTR increase, 0.78% deal growth, and 0.64% GMV rise. Code is available in the anonymous repository at https://anonymous.4open.science/r/FCSNet-ForPublish/.

1 INTRODUCTION

Sequential prediction is fundamental to modern information retrieval systems (Zhang et al., 2022; Li et al., 2024c; Dai et al., 2025). An important characteristic of real-world sequential data, often overlooked in existing literature, is its *coupled* nature: user behaviors simultaneously generate a primary *action sequence* (which items were interacted with) and a co-evolving *context sequence* (under what conditions the interactions occurred). We focus on *environment-level* context, which means features that describe the interaction environment rather than the item itself, such as app sub-interfaces (channels) in e-commerce and publisher sources in online news. Unlike item-level context (*e.g.*, category, brand) that is intrinsically tied to each item, environment-level context forms a sequence whose element-wise alignment with the action sequence is inherently noisy — a phenomenon we term *alignment noise*, arising from delayed feedback, context switching, and cross-context browsing. How to leverage such context-sequence dependencies while remaining robust to alignment noise is a fundamental challenge for sequential prediction.

This coupled structure arises in multiple applications. In e-commerce, user purchase interactions span multiple *channels*, *i.e.* the sub-interfaces such as "search", "recommendation", and "live shopping", with each exhibiting distinct item distributions and user engagement patterns. The item sequence and context sequence form natural coupled sequences, yet their element-wise alignment is

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

frequently disrupted. For instance, a user's purchase intention for a product may originate in the "recommendation" channel but materialize days later in the "favorite" channel due to a delayed promotion, producing a channel-item misalignment at that position. In **online news**, user browsing sequences are coupled with document-source sequences. Alignment noise arises here as well: cross-source recommendation platforms display recommendations across sources, so a user browsing on Source A may click a recommended article hosted by Source B, and the recorded interaction is $(doc_B, source_B)$, but the behavioral context that triggered the click was $source_A$, producing a source-document misalignment. These examples illustrate that alignment noise between coupled sequences is a general phenomenon across domains.

Existing sequential methods face a fundamental tension when handling coupled sequences with alignment noise. On one end, many methods (Sun et al., 2019; Fu et al., 2023; Wang et al., 2023a; Ma et al., 2024; Chai et al., 2025; Zhang et al., 2026) omit the environment-level context sequence, losing contextual signals that influence behavior occurrence. On the other end, methods that incorporate context (Tian et al., 2023; Papso, 2023; Wan et al., 2024; Li et al., 2024a; Huang et al., 2026) typically bind it to actions at the element level (*e.g.*, concatenating their embeddings at each position), so a misaligned pair introduces erroneous pairwise information that can perturb the learned representation. *Neither strategy addresses the core challenge: how to leverage cross-sequence dependencies while remaining robust to element-wise alignment noise.*

In this work, we propose *sequence-level factorized modeling*, which captures each sequence's temporal dynamics independently and fuses cross-sequence information only at the representation level. Because the action encoder operates on $\mathcal{V}$ alone, the pre-fusion action representation is structurally invariant to context perturbations; alignment noise can only reach predictions through the context encoder and the subsequent fusion module. We instantiate this design as **Factorized Coupled Sequence Networks (FCSNet)**, with parallel action and context encoders, cross-sequence representation fusion, and dual prediction heads. Under explicit mean-field posterior and factorized-prior assumptions, the FCSNet objective can be interpreted as an approximation to a *factored evidence lower bound (ELBO)* of the joint likelihood, providing a probabilistic motivation. To further improve robustness against the heavy-tailed distributions prevalent in real-world user behaviors, we introduce **Conditional Contrastive Regularization (CCR)**, a distribution-aware contrastive loss that learns similarity-adaptive weights that emphasize harder contrastive pairs. Online A/B tests in a production e-commerce search system confirm that FCSNet achieves a 0.04 absolute-point CTR improvement, 0.78% deal growth, and 0.64% GMV rise. The contributions are summarized as:

- We formalize the problem of sequential prediction from coupled sequences under noisy element-wise alignment, and show that the factorized architecture structurally isolates the action representation from context perturbations, confining alignment-noise sensitivity to a single context-to-prediction pathway.
- We propose FCSNet, which instantiates factorized coupled sequence modeling through dual encoders with cross-sequence representation fusion and admits an approximate factored-ELBO interpretation under explicit assumptions. FCSNet further incorporates Conditional Contrastive Regularization (CCR) with similarity-adaptive weights to encourage consistent representations of augmented sequences, enhancing the model's robustness in learning sequence dynamics.
- We validate FCSNet across two domains where alignment noise naturally arises, online news browsing and e-commerce purchasing, and evaluate predictive performance and robustness to controlled alignment noise. Production deployment serving hundreds of millions of users further confirms its real-world impact.

2 Related Work

Sequential Encoding Architectures: Early works employ Markov chains (Ching & Ng, 2006; Rendle et al., 2010) for sequential modeling, but struggle with higher-order dependencies. Later, RNNs (Medsker et al., 2001; Hidasi et al., 2016; Hidasi & Karatzoglou, 2018), CNNs (Tang & Wang, 2018), LSTMs (Graves & Graves, 2012; Duan et al., 2023) and self-attention models (Vaswani et al., 2017; Kang & McAuley, 2018; Sun et al., 2019) significantly advanced sequential behavior modeling. SASRec (Kang & McAuley, 2018) and BERT4Rec (Sun et al., 2019) established self-attention as the dominant paradigm. Recent works explore LLMs (Hu et al., 2024; Li et al., 2024b; Zheng et al., 2024; Liao et al., 2024) and diffusion models (Ma et al., 2024; Yang et al., 2024; Wang

et al., 2024) for sequential recommendation. These methods model a single action sequence, without explicitly addressing the coupled context sequence or alignment noise.

Training Objectives and Self-Supervised Learning: Beyond next-item prediction, SUMN (Gu et al., 2021) and MSDP (Fu et al., 2023) predict behavioral distributions over time periods. Self-supervised objectives have gained considerable attention (Yao et al., 2021; Zhou et al., 2020; Qiu et al., 2021; Wang et al., 2023b). S3-Rec (Zhou et al., 2020) designs four self-supervised tasks. CL4SRec (Qiu et al., 2021) and ContraRec (Wang et al., 2023b) leverage contrastive learning on augmented sequences. Recent works continue to advance this direction, for instance, (He et al., 2025) disentangles long and short-term interests via self-supervised graph neural networks, (Wang et al., 2025) proposes triplet contrastive learning with learnable augmentation strategies, and (Cui et al., 2025) augments contrastive learning with semantic retrieval. However, these contrastive losses assign uniform weights to samples, ignoring the varying importance of positives and negatives. Our CCR loss addresses this with similarity-adaptive weights that emphasize harder contrastive pairs.

Contextual Information in Sequential Modeling: Various contextual factors have been studied, including behavior types (Meng et al., 2020; Ni et al., 2018; Su et al., 2023; Xia et al., 2023; Xuan et al., 2023; Chen et al., 2023a), product categories (Cai et al., 2021), and item attributes (Papso, 2023). DUPN (Ni et al., 2018) incorporates behavior types via multi-task learning. CARCA (Papso, 2023) concatenates item attributes and contextual data as input to attention blocks. (Peng et al., 2025) incorporate temporal-aware graph contrastive learning to augment sequential features. (Zhang et al., 2025) design a hierarchical time-aware mixture of experts for multi-modal sequential recommendation. Some works use different context definitions: (Wang et al., 2021) defined contexts as predefined item category collections, and (Li et al., 2024a) defined contexts as 200 predefined topics. These works treat context as element-level item attributes, which is reasonable because behavior type, item category, and item topic are intrinsically tied to each item and thus naturally aligned. In contrast, the context features we study, such as app sub-interfaces (channels) in e-commerce and publisher sources in online news, are *environment-level* rather than item-level. They describe where or how the interaction occurred, not properties of the item itself. This distinction is critical because environment-level context is susceptible to action-context alignment noise. For instance, in e-commerce a user may form purchase intent in the "recommendation" channel but complete the transaction in the "favorite" channel days later. In web browsing, cross-publisher recommendation widgets cause the recorded source to differ from the discovery context. Element-level encoding of such misaligned action–context pairs introduces erroneous pairwise information that can perturb the input representations.

3 Factorized Coupled Sequence Networks

3.1 Overview

In sequential user behavior prediction, we aim to predict a user's future behaviors based on historical behaviors. Given a historical behavior sequence $\mathcal{T}$ from user $\boldsymbol{u}$, it is defined as:

$$\mathcal{T} = \{(\boldsymbol{v}_1, \boldsymbol{c}_1), (\boldsymbol{v}_2, \boldsymbol{c}_2), ..., (\boldsymbol{v}_j, \boldsymbol{c}_j), ..., (\boldsymbol{v}_{|\mathcal{T}|}, \boldsymbol{c}_{|\mathcal{T}|})\}, \tag{1}$$

where $\boldsymbol{v}_j \in \boldsymbol{V} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{N_v}\}$ and $\boldsymbol{c}_j \in \boldsymbol{C} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{N_c}\}$ denote an action and its recorded environment-level context (*e.g.*, an item/channel pair in e-commerce or a document/source pair in online news), respectively. $\boldsymbol{V}$ and $\boldsymbol{C}$ denote the action and context vocabularies with sizes N_v and N_c, respectively. $|\mathcal{T}|$ is the number of historical interactions. The history contains two *coupled sequences*: the action sequence $\mathcal{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{|\mathcal{T}|}\}$ and the context sequence $\mathcal{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_{|\mathcal{T}|}\}$. We aim to predict future behaviors based on the historical behaviors. **To better study the issue, we only consider one-type behavior in sequence learning.**

FCSNet consists of two components: **dual sequence learning** and **conditional contrastive regularization (CCR)**. The first component encodes the coupled histories separately and learns their interaction through representation-level fusion, structurally isolating the action branch from context perturbations. The second one learns representation invariance with adaptive weights, improving robustness against random and heavy-tailed user behaviors. The model architecture is in Figure 1.

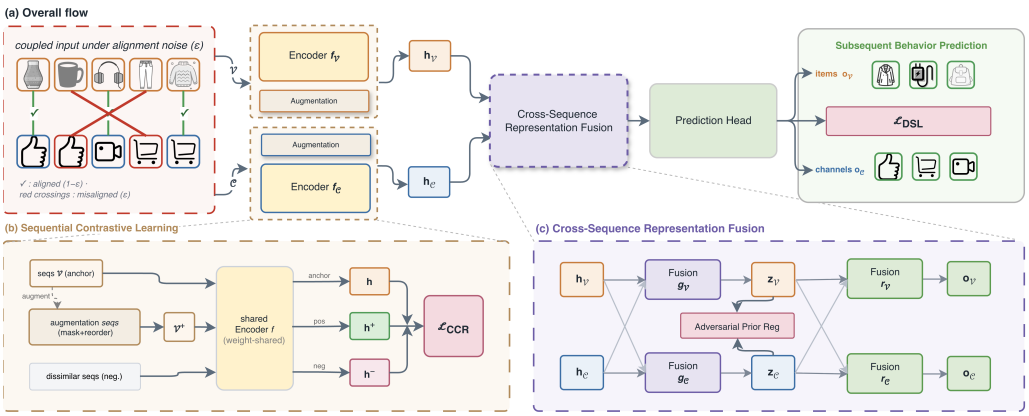


Figure 1: The architecture of FCSNet. Dual sequence learning independently encodes the coupled action and context sequences and captures their interaction through representation-level fusion. CCR uses similarity-adaptive weights to learn representation invariance. $\oplus$ denotes concatenation.

3.2 Dual Sequence Learning

Dual Sequence Dynamics: Online user behaviors occur under different contexts chronologically, reflecting users' dynamic interests over time. Context can influence action occurrence, while past actions can inform subsequent contexts. We therefore encode the historical action sequence $\mathcal{V}$ and context sequence $\mathcal{C}$ independently before representation-level fusion. Let $f_{\mathcal{V}}$ and $f_{\mathcal{C}}$ denote their sequence encoders:

$$\boldsymbol{h}_{\mathcal{C}} = f_{\mathcal{C}}(\mathcal{C}),\ \boldsymbol{h}_{\mathcal{V}} = f_{\mathcal{V}}(\mathcal{V}), \tag{2}$$

where $\boldsymbol{h}_{\mathcal{C}}$ and $\boldsymbol{h}_{\mathcal{V}}$ mean the encoded latent representations of $\mathcal{C}$ and $\mathcal{V}$, respectively. The choice of sequential model is flexible (such as RNN, LSTM and transformer) and we employ the powerful transformer model following recent works (Sun et al., 2019; Fu et al., 2023; Wang et al., 2023a).

Cross-Sequence Representation Fusion: As discussed, both the action and context sequences mutually influence subsequent behavior occurrence. When predicting future actions, context dynamics provide complementary signals, and action history can likewise inform context prediction. In the factorized design, this cross-sequence interaction is captured at the representation level through a fusion module. Let $\boldsymbol{z}_{\mathcal{C}}$ and $\boldsymbol{z}_{\mathcal{V}}$ denote the fused context and action representations, respectively:

$$\boldsymbol{z}_{\mathcal{C}} = g_{\mathcal{C}}(\boldsymbol{h}_{\mathcal{C}} \oplus \boldsymbol{h}_{\mathcal{V}}),\ \ \boldsymbol{z}_{\mathcal{V}} = g_{\mathcal{V}}(\boldsymbol{h}_{\mathcal{C}} \oplus \boldsymbol{h}_{\mathcal{V}}), \tag{3}$$

where $g_{\mathcal{C}}(\cdot)$ and $g_{\mathcal{V}}(\cdot)$ denote the fusion MLP layers. Both $\boldsymbol{z}_{\mathcal{C}}$ and $\boldsymbol{z}_{\mathcal{V}}$ can be considered as the user representation $\boldsymbol{z}_{\boldsymbol{u}}$ that encodes mutual effects and dynamics from historical behaviors. *We maintain these two enhanced representations to provide diverse aspects of user interests and better facilitate subsequent action and context prediction tasks.*

Subsequent Behavior Prediction: When conducting subsequent behavior prediction, we incorporate both enhanced representations $\boldsymbol{z}_{\mathcal{C}}$ and $\boldsymbol{z}_{\mathcal{V}}$ for behavior prediction:

$$\boldsymbol{o}_{\mathcal{C}} = r_{\mathcal{C}}(\boldsymbol{z}_{\mathcal{C}} \oplus \boldsymbol{z}_{\mathcal{V}}),\ \ \boldsymbol{o}_{\mathcal{V}} = r_{\mathcal{V}}(\boldsymbol{z}_{\mathcal{C}} \oplus \boldsymbol{z}_{\mathcal{V}}), \tag{4}$$

where $\boldsymbol{o}_{\mathcal{C}}$ and $\boldsymbol{o}_{\mathcal{V}}$ are outputs of the feature selection functions. Additionally, since the next user behavior may be stochastic and noisy, we choose to predict the subsequent behaviors over a period of time by following (Fu et al., 2023). Let $\boldsymbol{C}^{\text{cand}} \subseteq \boldsymbol{C}$ and $\boldsymbol{V}^{\text{cand}} \subseteq \boldsymbol{V}$ denote the candidate context and action sets. Their multi-hot target vectors are $\boldsymbol{y}^{c} \in \{0,1\}^{K^{c}}$ and $\boldsymbol{y}^{v} \in \{0,1\}^{K^{v}}$, respectively. The prediction losses for future contexts and actions are:

$$\mathcal{L}_{\text{DSL}}^{\mathcal{C}/\mathcal{V}} = -\frac{1}{K}\sum_{k=1}^{K}[\boldsymbol{y}_{k}^{\boldsymbol{c}/\boldsymbol{v}}\log(\hat{\boldsymbol{y}}_{k}^{c/v}) + (1-\boldsymbol{y}_{k}^{c/v})\log(1-\hat{\boldsymbol{y}}_{k}^{c/v})], \tag{5}$$

where $\hat{\boldsymbol{y}}_{k}^{\boldsymbol{c}} = \sigma(\boldsymbol{o}_{\mathcal{C}} \cdot \boldsymbol{e}_{k}^{\boldsymbol{c}})$ and $\hat{\boldsymbol{y}}_{k}^{\boldsymbol{v}} = \sigma(\boldsymbol{o}_{\mathcal{V}} \cdot \boldsymbol{e}_{k}^{\boldsymbol{v}})$ denote the prediction probabilities of the k-th context and action candidates. $\sigma(\cdot)$ is the sigmoid function and $\cdot$ means inner-product operation.

Adversarial Prior Regularization: Since user behaviors usually face severe data sparsity problem and user representations may overfit to samples, we impose adversarial prior regularization on learned user representations (Makhzani et al., 2016a; Chen et al., 2023b). This approach is more advantageous than Kullback-Leibler (KL) divergence regularization because of its flexibility and easy implementation.

Let $D_{\mathcal{C}}$ and $D_{\mathcal{V}}$ be the discriminator of $\boldsymbol{z}_{\mathcal{C}}$ and $\boldsymbol{z}_{\mathcal{V}}$ respectively, the adversarial learning based prior regularization is written as:

$$\begin{aligned}\min_{g_{\mathcal{C}},g_{\mathcal{V}},f_{\mathcal{C}},f_{\mathcal{V}}} \max_{D_{\mathcal{C}},D_{\mathcal{V}}} \mathcal{L}_{\text{APR}} =& \mathbb{E}_{\boldsymbol{z}_{\mathcal{C}}\sim p(\boldsymbol{z}_{\mathcal{C}})}[\log D_{\mathcal{C}}(\boldsymbol{z}_{\mathcal{C}})] + \mathbb{E}_{\boldsymbol{z}_{\mathcal{C}}\sim g_{\mathcal{C}}(\cdot)}[\log(1 - D_{\mathcal{C}}(\boldsymbol{z}_{\mathcal{C}}))] \\ &+ \mathbb{E}_{\boldsymbol{z}_{\mathcal{V}}\sim p(\boldsymbol{z}_{\mathcal{V}})}[\log D_{\mathcal{V}}(\boldsymbol{z}_{\mathcal{V}})] + \mathbb{E}_{\boldsymbol{z}_{\mathcal{V}}\sim g_{\mathcal{V}}(\cdot)}[\log(1 - D_{\mathcal{V}}(\boldsymbol{z}_{\mathcal{V}}))], \end{aligned} \tag{6}$$

where $p(\boldsymbol{z}_{\mathcal{C}})$ and $p(\boldsymbol{z}_{\mathcal{V}})$ are prior distributions. FCSNet has a behavior prediction task that matches complex behavior distributions, which can largely prevent mode collapse in adversarial learning.

We provide a probabilistic interpretation of dual sequence learning as follows:

Proposition 1 *Assume (i) a mean-field posterior $q_\phi(\boldsymbol{z}_{\mathcal{V}}, \boldsymbol{z}_{\mathcal{C}}|\mathcal{V}, \mathcal{C}) = q_{\phi_1}(\boldsymbol{z}_{\mathcal{V}}|\mathcal{V}, \mathcal{C})q_{\phi_2}(\boldsymbol{z}_{\mathcal{C}}|\mathcal{V}, \mathcal{C})$, (ii) a factorized prior $p(\boldsymbol{z}_{\mathcal{V}}, \boldsymbol{z}_{\mathcal{C}}) = p(\boldsymbol{z}_{\mathcal{V}})p(\boldsymbol{z}_{\mathcal{C}})$, and (iii) conditionally independent action and context targets given the latent representations. Then the latent-variable objective underlying FCSNet yields the following factored evidence lower bound:*

$$\begin{aligned}\max_{\theta_1,\theta_2,\phi_1,\phi_2} \mathcal{L}_{ELBO} =& \mathbb{E}_{q_{\phi_1}(\boldsymbol{z}_{\mathcal{V}}|\mathcal{V},\mathcal{C})q_{\phi_2}(\boldsymbol{z}_{\mathcal{C}}|\mathcal{V},\mathcal{C})}[\log p_{\theta_1}(\boldsymbol{v}|\boldsymbol{z}_{\mathcal{V}}, \boldsymbol{z}_{\mathcal{C}})p_{\theta_2}(\boldsymbol{c}|\boldsymbol{z}_{\mathcal{V}}, \boldsymbol{z}_{\mathcal{C}})] \\ &- D_{KL}[q_{\phi_1}(\boldsymbol{z}_{\mathcal{V}}|\mathcal{V}, \mathcal{C})||p(\boldsymbol{z}_{\mathcal{V}})] - D_{KL}[q_{\phi_2}(\boldsymbol{z}_{\mathcal{C}}|\mathcal{V}, \mathcal{C})||p(\boldsymbol{z}_{\mathcal{C}})]. \end{aligned} \tag{7}$$

Here, $\boldsymbol{v}$ and $\boldsymbol{c}$ are the future action and context targets, and $\boldsymbol{z}_{\mathcal{V}}$ and $\boldsymbol{z}_{\mathcal{C}}$ are latent representations of the historical action and context sequences. The detailed derivation is provided in Appendix A. The two likelihood terms motivate dual prediction, while adversarial prior regularization approximates matching each aggregated posterior to its prior. Thus, Eq. 7 provides a probabilistic interpretation to the implemented deterministic objective.

Remark: Modeling the joint likelihood captures interactions between the action and context sequences. In contrast, element-level feature fusion entangles this interaction with potentially noisy pairwise correspondence, while our factorized design postpones cross-sequence interaction until after independent encoding. Moreover, because $f_{\mathcal{V}}$ takes only $\mathcal{V}$ as input, $\boldsymbol{h}_{\mathcal{V}}$ is structurally invariant to any context perturbation. In a joint encoder, this isolation is lost as context perturbations alter shared attention weights at action positions, propagating noise into the action representation.

3.3 Sequential Contrastive Learning

Unlike human language, sequential user behaviors are often random and noisy, with heavy-tailed frequency distributions. To better capture robust dynamics, we employ sequential contrastive learning with two parts: sequence augmentation and conditional contrastive regularization (CCR).

Sequence Augmentation: The sequence augmentation must not alter the user's intention in input sequence. Drawing insights from recent studies (Sun et al., 2019; Wang et al., 2023a), we utilize masking and reordering to perform sequence augmentation. The masking augmentation involves randomly masking a percentage of elements from the input sequence. Reorder augmentation consists of two steps: first, we randomly select a size that ranges from 2 up to the length of the sequence. Then, we uniformly choose a continuous subsequence of this size and shuffle its elements, while the elements outside of this subsequence retain their original order. Let $\mathcal{A}(\cdot)$ represent a function that applies augmentation to the original sequence. The augmented context and action sequences are $\mathcal{C}^+ = \mathcal{A}(\mathcal{C})$ and $\mathcal{V}^+ = \mathcal{A}(\mathcal{V})$, respectively. These augmented samples offer valuable signals for learning representation invariance.

Conditional Contrastive Regularization: To learn sequence invariance, CCR increases agreement between original and augmented sequences while separating sampled negative sequences. It uses two similarity-adaptive weights to account for the different hardness of positive and negative pairs.

Table 1: The statistics of datasets.

Dataset	#sequences	#items	#channels	#avg. length	#density
Outbrain	46,676	238,653	3,508	2.36	9.89e-4%
AllChannelPay-1m	1,000,000	7,871,700	330	25.35	3.22e-4%
AllChannelPay-10m	10,000,000	32,766,762	801	25.33	7.70e-5%

Let $\boldsymbol{h}_{\mathcal{V}}^{+}$ be the representation of an augmented action sequence $\mathcal{V}^{+}$ and $\boldsymbol{h}_{\mathcal{V}}^{-}$ the representation of a sampled negative action sequence; our contrastive loss with two similarity-adaptive weights is:

$$\mathcal{L}_{\text{CCR}}^{\mathcal{V}} = -\mathbb{E}_{\boldsymbol{h}_{\mathcal{V}}}[\sum_{i=1}^{N_+} \underbrace{\frac{e^{-s(\boldsymbol{h}_{\mathcal{V}}, \boldsymbol{h}_{\mathcal{V},i}^{+})}}{\sum_{i}^{N_+} e^{-s(\boldsymbol{h}_{\mathcal{V}}, \boldsymbol{h}_{\mathcal{V},i}^{+})}}}_{\text{adaptive weight: } \boldsymbol{w}_{\mathcal{V},i}^{+}} s(\boldsymbol{h}_{\mathcal{V}}, \boldsymbol{h}_{\mathcal{V},i}^{+})] + \mathbb{E}_{\boldsymbol{h}_{\mathcal{V}}}[\sum_{j=1}^{N_-} \underbrace{\frac{e^{s(\boldsymbol{h}_{\mathcal{V}}, \boldsymbol{h}_{\mathcal{V},j}^{-})}}{\sum_{j}^{N_-} e^{s(\boldsymbol{h}_{\mathcal{V}}, \boldsymbol{h}_{\mathcal{V},j}^{-})}}}_{\text{adaptive weight: } \boldsymbol{w}_{\mathcal{V},j}^{-}} s(\boldsymbol{h}_{\mathcal{V}}, \boldsymbol{h}_{\mathcal{V},j}^{-})], \tag{8}$$

where $s(\boldsymbol{h}_{\mathcal{V}}, \boldsymbol{h}_{\mathcal{V}}^{+}) = \boldsymbol{h}_{\mathcal{V}}^{T}\boldsymbol{h}_{\mathcal{V}}^{+}/\tau^{+}$ and $s(\boldsymbol{h}_{\mathcal{V}}, \boldsymbol{h}_{\mathcal{V}}^{-}) = \boldsymbol{h}_{\mathcal{V}}^{T}\boldsymbol{h}_{\mathcal{V}}^{-}/\tau^{-}$ calculate the similarity between two vectors. τ^{+} and τ^{-} are two temperature hyper-parameters. N_{+} and N_{-} indicate the number of augmented sequences (*i.e.* positive samples) and dissimilar sequences (*i.e.* negative samples). $\boldsymbol{w}_{\mathcal{V},i}^{+}$ and $\boldsymbol{w}_{\mathcal{V},j}^{-}$ are similarity-adaptive weights that emphasize harder pairs during optimization. The conditional contrastive loss of context sequence $\mathcal{L}_{\text{CCR}}^{\mathcal{C}}$ can be written in similar formula with $(\boldsymbol{h}_{\mathcal{C}}, \boldsymbol{h}_{\mathcal{C}}^{+}, \boldsymbol{h}_{\mathcal{C}}^{-})$ as input.

Remark: The weights $\boldsymbol{w}^{+}$ and $\boldsymbol{w}^{-}$ provide similarity-adaptive, hardness-aware reweighting. Specifically, $\boldsymbol{w}^{+}$ emphasizes augmented sequences farther from the anchor, whereas $\boldsymbol{w}^{-}$ emphasizes negative sequences that remain close in representation space. This directs optimization toward informative contrastive pairs and is particularly useful for sparse, short-history sequences.

3.4 Overall Objective Function

To sum up, we can write the whole training objective function as:

$$\mathcal{L}_{\text{FCSNet}} = \mathcal{L}_{\text{DSL}}^{\mathcal{V}} + \lambda * \mathcal{L}_{\text{DSL}}^{\mathcal{C}} + \alpha * \mathcal{L}_{\text{APR}} + \beta * \mathcal{L}_{\text{CCR}}, \tag{9}$$

where $\mathcal{L}_{\text{CCR}} = \mathcal{L}_{\text{CCR}}^{\mathcal{C}} + \mathcal{L}_{\text{CCR}}^{\mathcal{V}}$. The λ, α and β are hyper-parameters to weight the importance of loss terms. *Because future-action prediction is the primary task, $\mathcal{L}_{DSL}^{\mathcal{V}}$ is unscaled and λ weights the auxiliary context-prediction loss.* To sum up, FCSNet offers an efficient and principled approach for factorized modeling of coupled action-context sequences. Dual sequence learning admits the factored-ELBO interpretation under the stated assumptions and preserves the pre-fusion action representation under context permutations; CCR enhances representation learning through similarity-adaptive weighting. **The model complexity analysis is given in Appendix B.**

4 Experiments and Analysis

4.1 Experiment Setup

Datasets: We evaluate on Outbrain and two industrial e-commerce datasets; their statistics are summarized in Table 1. In Outbrain[1], viewed documents and publisher sources form the action and context sequences, respectively. Records before timeoffset "181468800" (2016-06-14) are used as history and the remainder as targets; users must have at least one interaction in each period, and the data are split 8:1:1 for training, validation, and testing. AllChannelPay-1m and AllChannelPay-10m are random samples of one and ten million purchase sequences collected from 2024-07-01 to 2024-08-07. July records form the history and records from 2024-08-01 to 2024-08-07 form the targets. We retain users with at least three interactions in both periods and use an 80/10/10 split.

Evaluation Protocol: Next-item prediction targets the first future item. Period-item prediction uses the first 100 future interactions (or all available interactions if fewer), retaining repeated items as separate positive entries. Candidate lists contain 1,000 entries on Outbrain and 10,000 on AllChannelPay-1m and AllChannelPay-10m, with the remaining entries taken from the supplied frozen negative samples. We report Recall@k and NDCG@k for $k \in \{5, 10\}$, averaged over users.

[1] https://www.kaggle.com/competitions/outbrain-click-prediction/overview

Table 2: Performance comparison on next item prediction. R@k and N@k are Recall@k and NDCG@k. We use "w/o" to denote FCSNet without a particular part. The best results are bolded and *the most competitive public baselines* are underlined. Historical rows are retained; their checkpoint and test-split provenance remains under verification. Dashes denote results not reported here.

Dataset	Outbrain				AllChannelPay-1m				AllChannelPay-10m			
Method	R@5	N@5	R@10	N@10	R@5	N@5	R@10	N@10	R@5	N@5	R@10	N@10
BERT4Rec	0.0943	0.0676	0.1384	0.0819	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
MSDP	0.2703	0.2181	0.2994	0.2275	0.0006	0.0004	0.0010	0.0005	0.0005	0.0003	0.0011	0.0005
ContraRec	0.3619	0.2468	0.4701	0.2820	0.0753	0.0533	0.1010	0.0616	0.1414	0.1026	0.1925	0.1191
DIN	0.4177	0.3133	0.5296	0.3495	0.0591	0.0400	0.0869	0.0490	0.1301	0.1005	0.1645	0.1116
DIEN	0.4188	0.3138	0.5216	0.3471	0.0689	0.0470	0.1000	0.0570	0.1400	0.1018	0.1827	0.1155
GRU4Rec	0.4394	0.3309	0.5394	0.3635	0.0721	0.0513	0.0990	0.0599	0.1424	0.1051	0.1855	0.1190
ContextCTC	0.4811	0.4068	0.5232	0.4205	0.0735	0.0517	0.1023	0.0610	0.1459	0.1027	0.1974	0.1193
ContextContraRec	0.4979	0.4027	0.5448	0.4182	0.0762	0.0544	0.1045	0.0635	0.1455	0.1028	0.1983	0.1199
CARCA	0.5126	0.4373	0.5430	0.4472	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
FCSNet–	0.5324	0.4612	0.5604	0.4703	0.0742	0.0527	0.1047	0.0625	0.1443	0.1028	0.1987	0.1201
FCSNet(w/o $\mathcal{L}_{\text{APR}}$, $\mathcal{L}_{\text{CCR}}$)	0.6115	0.5292	0.6625	0.5459	0.0843	0.0617	0.1123	0.0707	0.1680	0.1241	0.2206	0.1411
FCSNet(w/o $\mathcal{L}_{\text{CCR}}$)	0.6109	0.5327	0.6674	0.5511	0.0845	0.0616	0.1121	0.0704	0.1710	0.1266	0.2239	**0.1437**
FCSNet(w/o $\mathcal{L}_{\text{APR}}$)	0.6198	**0.5388**	0.6684	**0.5545**	0.0870	0.0630	**0.1158**	0.0723	0.1711	0.1266	0.2229	0.1433
FCSNet	**0.6248** (+11.22%)	0.5368 (+9.95%)	**0.6717** (+12.69%)	0.5520 (+10.48%)	**0.0870** (+1.07%)	**0.0632** (+0.88%)	0.1155 (+1.10%)	**0.0725** (+0.90%)	**0.1712**	**0.1267**	**0.2240**	0.1436

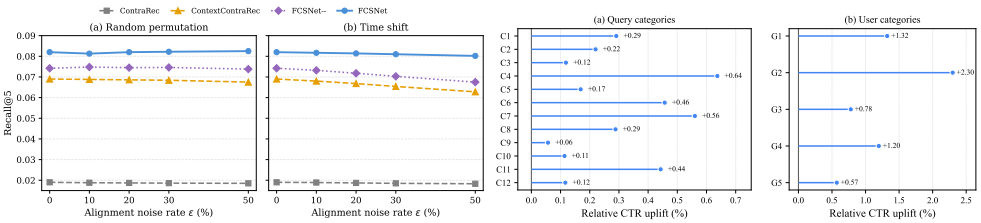
Table 3: Performance comparison on period item prediction. R@k and N@k are Recall@k and NDCG@k. We use "w/o" to denote FCSNet without a particular part. Highlighting is retained for the historical rows. Historical checkpoint and test-split provenance remains under verification. Dashes denote results not reported here.

Dataset	Outbrain				AllChannelPay-1m				AllChannelPay-10m			
Method	R@5	N@5	R@10	N@10	R@5	N@5	R@10	N@10	R@5	N@5	R@10	N@10
BERT4Rec	0.0946	0.0694	0.1388	0.084	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
MSDP	0.2671	0.2191	0.2967	0.229	0.0007	0.0006	0.0011	0.0008	0.0008	0.0007	0.0012	0.0009
ContraRec	0.3573	0.2481	0.4654	0.2844	0.0759	0.0790	0.0888	0.0839	0.1534	0.1610	0.1796	0.1707
DIN	0.4150	0.3162	0.5253	0.3529	0.0661	0.0677	0.0822	0.0747	–	–	–	–
DIEN	0.4159	0.3168	0.5184	0.3509	0.0773	0.0791	0.0950	0.0869	–	–	–	–
GRU4Rec	0.4361	0.3337	0.5370	0.3674	0.0819	0.0852	0.0958	0.0905	–	–	–	–
ContextCTC	0.4754	0.4080	0.5184	0.4225	0.0758	0.0785	0.0913	0.0848	0.1538	0.1585	0.1813	0.1694
ContextContraRec	0.4902	0.4031	0.5388	0.4197	0.0790	0.0821	0.0933	0.0876	0.1532	0.1587	0.1820	0.1702
CARCA	0.5092	0.4393	0.5402	0.4499	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
FCSNet–	0.5281	0.4634	0.5570	0.4372	0.0867	0.0905	0.1029	0.0968	0.1636	0.1703	0.1922	0.1812
FCSNet(w/o $\mathcal{L}_{\text{APR}}$, $\mathcal{L}_{\text{CCR}}$)	0.6062	0.5307	0.6583	0.5482	0.0997	0.1053	0.1121	0.1090	0.1895	0.1998	0.2149	0.2083
FCSNet(w/o $\mathcal{L}_{\text{CCR}}$)	0.6069	0.5347	0.6635	0.5537	0.0995	0.1049	0.1120	0.1088	0.1930	0.2033	0.2186	**0.2118**
FCSNet(w/o $\mathcal{L}_{\text{APR}}$)	0.6151	**0.5396**	0.6651	**0.5562**	0.1007	0.1060	0.1148	0.1107	**0.1930**	**0.2037**	0.2178	0.2117
FCSNet	**0.6198** (+11.06%)	0.5388 (+9.95%)	**0.6682** (+12.80%)	0.5549 (+10.50%)	**0.1015** (+2.25%)	**0.1071** (+2.50%)	**0.1149** (+2.16%)	**0.1114** (+2.38%)	0.1926 (+3.88%)	0.2028 (+4.18%)	**0.2187** (+3.67%)	0.2115 (+4.08%)

Baselines: We compare with BERT4Rec (Sun et al., 2019), MSDP (Fu et al., 2023), ContraRec (Wang et al., 2023a), and CARCA (Papso, 2023), covering sequential encoding, distribution prediction, contrastive learning, and contextual modeling. We also evaluate item-only adaptations of DIN, DIEN, and GRU4Rec for next-item prediction on all three datasets and for period-item prediction on Outbrain and AllChannelPay-1m. DIN (Zhou et al., 2018) aggregates historical behaviors through a candidate-conditioned local activation unit. DIEN (Zhou et al., 2019) extracts sequential interests with a GRU and auxiliary next-behavior supervision, then models candidate-specific interest evolution with a GRU with an attentional update gate (AUGRU). GRU4Rec (Hidasi et al., 2016; Hidasi & Karatzoglou, 2018) encodes item sequences with recurrent units; our adaptation uses the BPR-max ranking loss. ContextCTC uses context-target contrastive learning with channel embeddings, whereas ContextContraRec augments ContraRec with channel embeddings. FCSNet– replaces our dual encoders with element-level joint encoding. Results combining FCSNet with LLM-ESR (Liu et al., 2024) are reported in Appendix C.

Parameter Settings: We train for at most 100 epochs with validation-based early stopping and a batch size of 32 per GPU, using one GPU for Outbrain and eight for the industrial datasets. Baseline hyperparameters follow their original papers or are tuned on the validation set. The AllChannelPay-10m GRU4Rec result is evaluated from the epoch-32 checkpoint. FCSNet uses a two-layer BERT4Rec Transformer and two-layer fusion MLPs. The item/context embedding dimensions are 256/256 on Outbrain and 16/4 on the industrial datasets; Outbrain uses a shared linear prediction head, $r_{\mathcal{C}} = r_{\mathcal{V}}$. CCR uses two positives, in-batch negatives, $\tau^+ = 1.0$, and $\tau^- = 0.07$. We use a standard Gaussian prior and set (λ, α, β) to $(1.0, 2 \times 10^{-7}, 5 \times 10^{-6})$ on Outbrain and $(0.2, 10^{-9}, 10^{-7})$ on the industrial datasets[2]..

[2]The value of $\mathcal{L}_{DSL}$ is quite small due to large number of negative samples. We set α and β to a small scale to ensure they do not dominate $\mathcal{L}_{DSL}$.



(a) Controlled alignment noise (b) Online A/B test

Figure 2: (a) Recall@5 under random-permutation and time-shift alignment noise at varying noise rates on AllChannelPay-1m. (b) Relative CTR uplift across query and user categories.

Table 4: Study of cross-sequence representation fusion.

Dataset	Outbrain				AllChannelPay-1m				AllChannelPay-10m			
Model	R@5	N@5	R@10	N@10	R@5	N@5	R@10	N@10	R@5	N@5	R@10	N@10
w/o fusion	0.3661	0.2997	0.4028	0.3118	0.0701	0.0488	0.0994	0.0582	0.1412	0.1000	0.1953	0.1175
FCSNet(MLP_layers=1)	0.5633	0.4795	0.6211	0.4984	0.0868	0.0634	0.1138	0.0720	0.1676	0.1241	0.2194	0.1409
FCSNet(MLP_layers=3)	0.6175	0.5360	0.6644	0.5513	0.0857	0.0623	**0.1165**	0.0723	0.1621	0.1207	0.2111	0.1365
FCSNet(MLP_layers=2)	**0.6248**	**0.5368**	**0.6717**	**0.5520**	**0.0870**	**0.0632**	0.1155	**0.0725**	**0.1712**	**0.1267**	**0.2240**	**0.1436**

4.2 Performance Comparison

Overall Comparison: Tables 2 and 3 report next-item and period-item prediction results, respectively, with the latter reducing sensitivity to stochastic next actions. FCSNet exceeds all listed public baselines in the reported next-item metrics on all three datasets. On AllChannelPay-10m, GRU4Rec is the strongest of the three added item-only baselines, while FCSNet obtains higher Recall and NDCG at both cutoffs. FCSNet's advantage over FCSNet– supports sequence-level factorization, while the ablations show that both $\mathcal{L}_{\text{APR}}$ and $\mathcal{L}_{\text{CCR}}$ generally contribute to performance. CARCA and BERT4Rec run out of memory on the industrial datasets.

Robustness under Controlled Alignment Noise: Since true action–context correspondence noise is not directly observable, we use controlled perturbations to isolate its effect. On AllChannelPay-1m, we corrupt a fraction of the action–context correspondences under two noise types: *random permutation* (uniformly reassign context positions) and *time shift* (shift context assignments by k positions to simulate delayed feedback). Figure 2(a) reports absolute Recall@5. FCSNet achieves the best performance and remains the most stable under both noise types. Its gap over joint-encoding baselines widens as noise increases, especially under time-shift noise, confirming that the factorized design effectively isolates the action representation from context perturbations.

Online A/B Test and Deployment: We deployed FCSNet as a pretrained user-representation model in a production e-commerce search system. Over a 7-day A/B test, it achieved a 0.04-point absolute CTR increase, 0.78% deal growth, and 0.64% GMV growth. Figure 2(b) shows relative CTR uplift across query and user categories, which is positive for all displayed categories. FCSNet now serves hundreds of millions of users.

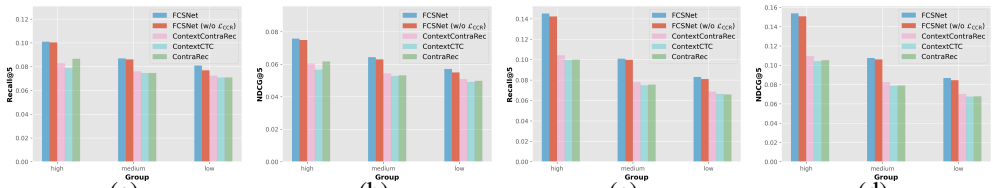
Study of Cross-Sequence Representation Fusion: In our dual sequence learning, cross-sequence representation fusion is a key module for capturing the interplay between the two historical sequences. To study its effect, we remove this component (denoted "w/o fusion"). As shown in Table 4, this leads to a clear performance drop, confirming the value of cross-sequence fusion.

4.3 Further Experimental Analysis

Prediction Performance on Channel: Table 5 evaluates next-channel prediction. FCSNet consistently outperforms all baselines on both datasets, showing that factorized modeling benefits both action and context prediction.

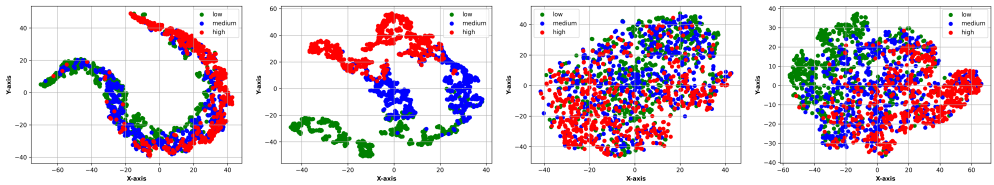
Table 5: Performance comparison on next-channel prediction.

Dataset	Outbrain				AllChannelPay-1m			
Method	R@5	N@5	R@10	N@10	R@5	N@5	R@10	N@10
BERT4Rec	0.3043	0.2763	0.3345	0.2861	0.8253	0.6365	0.9205	0.6483
MSDP	0.2638	0.1798	0.3789	0.2167	0.8464	0.6489	0.9371	0.6788
ContraRec	0.4071	0.3468	0.4662	0.3661	0.8710	0.6635	0.9508	0.6903
ContextCTC	0.6246	0.5492	0.6937	0.5715	0.8872	0.6692	0.9547	0.6918
ContextContraRec	0.6175	0.5386	0.6858	0.5606	0.8692	0.6629	0.9528	0.6911
FCSNet	**0.6567**	**0.5770**	**0.7200**	**0.5975**	**0.8944**	**0.6816**	**0.9629**	**0.7045**



(a) (b) (c) (d)

Figure 3: Performance across sequence-length groups. (a)–(b) show next-item prediction results, and (c)–(d) show period-item prediction results.



(a) ContraRec (b) ContextContraRec (c) FCSNet(w/o $\mathcal{L}_{CCR}$) (d) FCSNet

Figure 4: The t-SNE visualization of learned user representations on AllChannelPay-1m dataset.

Generalization on Sequence Distributions: Sequence lengths exhibit a long-tailed distribution, so we partition test sequences into long-, medium-, and short-history groups. Figure 3 shows that FCSNet consistently outperforms the baselines across all groups on both prediction tasks, demonstrating robust generalization across the sequence-length distribution. Its larger gain over FCSNet (w/o $\mathcal{L}_{CCR}$) in the short-history group further indicates CCR's benefit for tail sequences.

Effect of Different Priors: FCSNet uses adversarial prior regularization to embed prior knowledge into learned representations. We test several priors: Uniform $\mathcal{U}(0,1)$, Laplace Laplace$(0,1)$, standard Gaussian $\mathcal{N}(0,1)$, Lognormal Lognormal$(0,1)$, and a two-component Gaussian mixture $0.5\mathcal{N}(0,1) + 0.5\mathcal{N}(3,1)$. As shown in Figure 5, the standard Gaussian (darkest row) consistently achieves the best or near-best performance across all metrics on both datasets, with Lognormal the closest competitor on Outbrain and Laplace lagging most. The standard Gaussian is thus a simple and effective default.

	Outbrain				AllChannelPay-1m			
	R@5	N@5	R@10	N@10	R@5	N@5	R@10	N@10
Uniform	0.6025	0.5190	0.6537	0.5355	0.0851	0.0621	0.1129	0.0711
Laplace	0.5867	0.5031	0.6400	0.5206	0.0839	0.0610	0.1127	0.0703
Gauss. Mix.	0.5982	0.5181	0.6485	0.5343	0.0842	0.0618	0.1114	0.0705
Lognormal	0.6177	**0.5373**	0.6648	**0.5525**	0.0843	0.0615	0.1131	0.0708
Std. Gaussian	**0.6248**	0.5368	**0.6717**	0.5520	**0.0870**	**0.0632**	**0.1155**	**0.0725**

Figure 5: Effect of different prior distributions. Darker cells indicate higher values; bold marks each column's best.

Visualization of User Representations: We visualize 500 randomly sampled test-user representations from each of the "low", "medium", and "high" groups using t-SNE in Figure 4. FCSNet exhibits a more structured representation geometry, while CCR produces a more compact distribution for the "low" group.

5 Conclusion and Future Work

In this work, we formalize sequential prediction from coupled sequences under noisy element-wise alignment and propose FCSNet, which uses parallel encoders and cross-sequence representation fusion to reduce reliance on noisy pairwise correspondence. Under explicit assumptions, its objective admits an approximate factored-ELBO interpretation. We further propose CCR, a distribution-aware contrastive regularizer with similarity-adaptive weights for heavy-tailed sequential data. Extensive experiments show state-of-the-art results on three datasets, and production deployment confirms significant online gains for hundreds of millions of users. Controlled noise experiments further validate the robustness insight independently of the application domain. In future work, we will extend the framework to more than two coupled sequences and validate it on non-recommendation domains such as multi-modal sequential prediction.

AI USE STATEMENT

We used AI-based tools solely for proofreading and language polishing. All research contributions including the ideas, methods, experiments, results, and claims are our own, and we take full responsibility for the final content of this paper.

ETHICS STATEMENT

This work takes industrial user e-commerce app logs as industrial datasets to validate model performance. These industrial datasets are used under the privacy rules of our company, without leakage to public.

REPRODUCIBILITY STATEMENT

We provide a clear description of the model architecture, training procedure, hyperparameter settings, and evaluation protocol to facilitate reproducibility. The codes and used public datasets are also provided in an anonymous Github repository to enable reproduction of the reported results.

REFERENCES

Renqin Cai, Jibang Wu, Aidan San, Chong Wang, and Hongning Wang. Category-aware collaborative sequential recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pp. 388–397, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462832. URL https://doi.org/10.1145/3404835.3462832.

Zheng Chai, Qin Ren, Xijun Xiao, Huizhi Yang, Bo Han, Sijun Zhang, Di Chen, Hui Lu, Wenlin Zhao, Lele Yu, et al. Longer: Scaling up long sequence modeling in industrial recommenders. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pp. 247–256, 2025.

Xiaoqing Chen, Zhitao Li, Weike Pan, and Zhong Ming. A survey on multi-behavior sequential recommendation. *arXiv preprint arXiv:2308.15701*, 2023a.

Xu Chen, Ya Zhang, Ivor W. Tsang, Yuangang Pan, and Jingchao Su. Toward equivalent transformation of user preferences in cross domain recommendation. *ACM Trans. Inf. Syst.*, 41(1), jan 2023b. ISSN 1046-8188. doi: 10.1145/3522762. URL https://doi.org/10.1145/3522762.

Wai-Ki Ching and Michael K Ng. Markov chains. *Models, algorithms and applications*, 2006.

Ziqiang Cui, Yunpeng Weng, Xing Tang, Xiaokun Zhang, Shiwei Li, Peiyang Liu, Bowei He, Dugang Liu, Weihong Luo, Xiuqiang He, and Chen Ma. Semantic retrieval augmented contrastive learning for sequential recommendation, 2025. URL https://arxiv.org/abs/2503.04162.

Tianjie Dai, Xu Chen, Yunmeng Shu, Jinsong Lan, Xiaoyong Zhu, Jiangchao Yao, and Bo Zheng. Enhancing sequential recommendation with world knowledge from large language models, 2025. URL https://arxiv.org/abs/2511.20177.

Jiasheng Duan, Peng-Fei Zhang, Ruihong Qiu, and Zi Huang. Long short-term enhanced memory for sequential recommendation. *World Wide Web*, 26(2):561–583, 2023.

Chilin Fu, Weichang Wu, Xiaolu Zhang, Jun Hu, Jing Wang, and Jun Zhou. Robust user behavioral sequence representation via multi-scale stochastic distribution prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pp. 4567–4573, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701245. doi: 10.1145/3583780.3614714. URL https://doi.org/10.1145/3583780.3614714.

Alex Graves and Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

Jie Gu, Feng Wang, Qinghui Sun, Zhiquan Ye, Xiaoxiao Xu, Jingmin Chen, and Jun Zhang. Exploiting behavioral consistence for universal user representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4063–4071, 2021.

Liang He, Wujie Yan, Tingzhou Yi, and Heli Sun. Self-supervised graph neural sequential recommendation with disentangling long and short-term interest. *ACM Trans. Recomm. Syst.*, 4(1), July 2025. doi: 10.1145/3723173. URL https://doi.org/10.1145/3723173.

Balázs Hidasi and Alexandros Karatzoglou. Recurrent neural networks with top-k gains for session-based recommendations. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pp. 843–852, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. doi: 10.1145/3269206.3271761. URL https://doi.org/10.1145/3269206.3271761.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks, 2016. URL https://arxiv.org/abs/1511.06939.

Jun Hu, Wenwen Xia, Xiaolu Zhang, Chilin Fu, Weichang Wu, Zhaoxin Huan, Ang Li, Zuoli Tang, and Jun Zhou. Enhancing sequential recommendation via llm-based semantic embedding learning. In *Companion Proceedings of the ACM on Web Conference 2024*, pp. 103–111, 2024.

Xu Huang, Hao Zhang, Zhifang Fan, Yunwen Huang, Zhuoxing Wei, Zheng Chai, Jinan Ni, Yuchao Zheng, and Qiwei Chen. Mixformer: Co-scaling up dense and sequence in industrial recommenders. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 7422–7431, 2026.

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

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Wenhao Li, Jie Zhou, Chuan Luo, Chao Tang, Kun Zhang, and Shixiong Zhao. Scene-wise adaptive network for dynamic cold-start scenes optimization in ctr prediction. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, pp. 370–379, New York, NY, USA, 2024a. Association for Computing Machinery. ISBN 9798400705052. doi: 10.1145/3640457.3688115. URL https://doi.org/10.1145/3640457.3688115.

Yaoyiran Li, Xiang Zhai, Moustafa Alzantot, Keyi Yu, Ivan Vulić, Anna Korhonen, and Mohamed Hammad. Calrec: Contrastive alignment of generative llms for sequential recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pp. 422–432, 2024b.

Zijian Li, Ruichu Cai, Fengzhu Wu, Sili Zhang, Hao Gu, Yuexing Hao, and Yuguang Yan. Tea: A sequential recommendation framework via temporally evolving aggregations. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):2628–2639, 2024c. doi: 10.1109/TNNLS.2022.3190534.

Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. Llara: Large language-recommendation assistant. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1785–1795, 2024.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yejing Wang, Zijian Zhang, Feng Tian, and Yefeng Zheng. Large language models enhanced sequential recommendation for long-tail user and item. *arXiv preprint arXiv:2405.20646*, 2024.

Haokai Ma, Ruobing Xie, Lei Meng, Xin Chen, Xu Zhang, Leyu Lin, and Zhanhui Kang. Plug-in diffusion model for sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8886–8894, 2024.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian Goodfellow. Adversarial autoencoders. In *International Conference on Learning Representations*, 2016a. URL http://arxiv.org/abs/1511.05644.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian Goodfellow. Adversarial autoencoders. In *International conference on learning representation*, 2016b. URL http://arxiv.org/abs/1511.05644.

Larry R Medsker, Lakhmi Jain, et al. Recurrent neural networks. *Design and Applications*, 5(64-67): 2, 2001.

Wenjing Meng, Deqing Yang, and Yanghua Xiao. Incorporating user micro-behaviors and item knowledge into multi-task learning for session-based recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pp. 1091–1100, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. doi: 10.1145/3397271.3401098. URL https://doi.org/10.1145/3397271.3401098.

Yabo Ni, Dan Ou, Shichen Liu, Xiang Li, Wenwu Ou, Anxiang Zeng, and Luo Si. Perceive your users in depth: Learning universal user representations from multiple e-commerce tasks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pp. 596–605, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219828. URL https://doi.org/10.1145/3219819.3219828.

Rastislav Papso. Complementary product recommendation for long-tail products. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, pp. 1305–1311, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702419. doi: 10.1145/3604915.3608864. URL https://doi.org/10.1145/3604915.3608864.

Tianhao Peng, Haitao Yuan, Yongqi Zhang, Yuchen Li, Peihong Dai, Qunbo Wang, Senzhang Wang, and Wenjun Wu. Tagrec: Temporal-aware graph contrastive learning with theoretical augmentation for sequential recommendation. *IEEE Trans. on Knowl. and Data Eng.*, 37(5):3015–3029, May 2025. ISSN 1041-4347. doi: 10.1109/TKDE.2025.3538706. URL https://doi.org/10.1109/TKDE.2025.3538706.

Ruihong Qiu, Zi Huang, Hongzhi Yin, and Zijian Wang. Contrastive learning for representation degeneration problem in sequential recommendation. *CoRR*, abs/2110.05730, 2021.

Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 811–820, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605587998. doi: 10.1145/1772690.1772773. URL https://doi.org/10.1145/1772690.1772773.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

Jiajie Su, Chaochao Chen, Zibin Lin, Xi Li, Weiming Liu, and Xiaolin Zheng. Personalized behavior-aware transformer for multi-behavior sequential recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 6321–6331, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3611723. URL https://doi.org/10.1145/3581783.3611723.

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*, CIKM '19, pp. 1441–1450, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369763. doi: 10.1145/3357384.3357895. URL https://doi.org/10.1145/3357384.3357895.

Jiaxi Tang and Ke Wang. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pp. 565–573, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355810. doi: 10.1145/3159652.3159656. URL https://doi.org/10.1145/3159652.3159656.

Yu Tian, Bofang Li, Si Chen, Xubin Li, Hongbo Deng, Jian Xu, Bo Zheng, Qian Wang, and Chenliang Li. Multi-scenario ranking with adaptive feature learning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pp. 517–526, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3591736. URL https://doi.org/10.1145/3539618.3591736.

Jakub M Tomczak and Max Welling. Vae with a vampprior. *arXiv preprint arXiv:1705.07120*, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Zhizhong Wan, Bin Yin, Junjie Xie, Fei Jiang, Xiang Li, and Wei Lin. Larr: Large language model aided real-time scene recommendation with semantic understanding. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, pp. 23–32, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705052. doi: 10.1145/3640457.3688135. URL https://doi.org/10.1145/3640457.3688135.

Chenyang Wang, Weizhi Ma, Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. Sequential recommendation with multiple contrast signals. *ACM Trans. Inf. Syst.*, 41(1), jan 2023a. ISSN 1046-8188. doi: 10.1145/3522673. URL https://doi.org/10.1145/3522673.

Chenyang Wang, Weizhi Ma, Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. Sequential recommendation with multiple contrast signals. *ACM Transactions on Information Systems*, 41 (1):1–27, 2023b.

Gang Wang, Ziyi Guo, Xiang Li, Dawei Yin, and Shuai Ma. Scenerec: Scene-based graph neural networks for recommender systems. *ArXiv*, abs/2102.06401, 2021. URL https://api.semanticscholar.org/CorpusID:231918600.

Wei Wang, Yujie Lin, Moyan Zhang, Hongyu Lu, Jianli Zhao, Jie Sun, Xianye Ben, Pengjie Ren, and Yujun Li. Triplet contrastive learning with learnable sequence augmentation for sequential recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, pp. 1519–1529, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715921. doi: 10.1145/3726302.3730101. URL https://doi.org/10.1145/3726302.3730101.

Yu Wang, Zhiwei Liu, Liangwei Yang, and Philip S Yu. Conditional denoising diffusion for sequential recommendation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 156–169. Springer, 2024.

Lianghao Xia, Chao Huang, Yong Xu, and Jian Pei. Multi-behavior sequential recommendation with temporal graph transformer. *IEEE Trans. on Knowl. and Data Eng.*, 35(6):6099–6112, June 2023. ISSN 1041-4347. doi: 10.1109/TKDE.2022.3175094. URL https://doi.org/10.1109/TKDE.2022.3175094.

Hongrui Xuan, Yi Liu, Bohan Li, and Hongzhi Yin. Knowledge enhancement for contrastive multi-behavior recommendation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, pp. 195–203, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394079. doi: 10.1145/3539597.3570386. URL https://doi.org/10.1145/3539597.3570386.

Zhengyi Yang, Jiancan Wu, Zhicai Wang, Xiang Wang, Yancheng Yuan, and Xiangnan He. Generate what you prefer: Reshaping sequential recommendation via guided diffusion. *Advances in Neural Information Processing Systems*, 36, 2024.

Tiansheng Yao, Xinyang Yi, Derek Zhiyuan Cheng, Felix Yu, Ting Chen, Aditya Menon, Lichan Hong, Ed H Chi, Steve Tjoa, Jieqi Kang, et al. Self-supervised learning for large-scale item recommendations. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pp. 4321–4330, 2021.

Mingzhang Yin and Mingyuan Zhou. Semi-implicit variational inference. *International conference on machine learning*, 2018.

Qi Zhang, Longbing Cao, Chongyang Shi, and Zhendong Niu. Neural time-aware sequential recommendation by jointly modeling preference dynamics and explicit feature couplings. *IEEE Transactions on Neural Networks and Learning Systems*, 33(10):5125–5137, 2022. doi: 10.1109/TNNLS.2021.3069058.

Shengzhe Zhang, Liyi Chen, Dazhong Shen, Chao Wang, and Hui Xiong. Hierarchical time-aware mixture of experts for multi-modal sequential recommendation. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, pp. 3672–3682, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746. doi: 10.1145/3696410.3714676. URL https://doi.org/10.1145/3696410.3714676.

Zhaoqi Zhang, Haolei Pei, Jun Guo, Tianyu Wang, Yufei Feng, Hui Sun, Shaowei Liu, and Aixin Sun. Onetrans: Unified feature interaction and sequence modeling with one transformer in industrial recommender. In *Proceedings of the ACM Web Conference 2026*, pp. 8162–8170, 2026.

Zhi Zheng, Wenshuo Chao, Zhaopeng Qiu, Hengshu Zhu, and Hui Xiong. Harnessing large language models for text-rich sequential recommendation. In *Proceedings of the ACM on Web Conference 2024*, pp. 3207–3216, 2024.

Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1059–1068. ACM, 2018. doi: 10.1145/3219819.3219823. URL https://doi.org/10.1145/3219819.3219823.

Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. Deep interest evolution network for click-through rate prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):5941–5948, 2019. doi: 10.1609/aaai.v33i01.33015941. URL https://doi.org/10.1609/aaai.v33i01.33015941.

Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pp. 1893–1902, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3411954. URL https://doi.org/10.1145/3340531.3411954.

A Derivation of the Factored-ELBO Interpretation

Let $\boldsymbol{v}$ and $\boldsymbol{c}$ denote the future action and context targets for user $\boldsymbol{u}$, respectively. The joint log-likelihood sums $\log p_\theta(\boldsymbol{v}, \boldsymbol{c})$ over users, where $p_\theta(\boldsymbol{v}, \boldsymbol{c})$ is the corresponding probability mass function. Given the future action–context pair $(\boldsymbol{v}, \boldsymbol{c})$, we denote the preceding action and context histories by $\mathcal{V}$ and $\mathcal{C}$, respectively. The corresponding encoded latent representations of $\mathcal{V}$ and $\mathcal{C}$ are denoted as $\boldsymbol{z}_\mathcal{V}$ and $\boldsymbol{z}_\mathcal{C}$, respectively. Then, drawing from the idea of maximizing the marginal log-likelihood in Variational Autoencoders (VAEs) (Kingma & Welling, 2013), $\log p_\theta(\boldsymbol{v}, \boldsymbol{c})$ is written as:

$$\log p_\theta(\boldsymbol{v}, \boldsymbol{c}) = D_{KL}[q_\phi(\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C}|\mathcal{V}, \mathcal{C})||p(\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C}|\mathcal{V}, \mathcal{C})] + \mathcal{L}_{\text{ELBO}}, \tag{10}$$

where the first term denotes the non-negative KL divergence, so the second term is the *evidence lower bound (ELBO)* on $\log p_\theta(\boldsymbol{v}, \boldsymbol{c})$.

Following the derivation in VAE (Kingma & Welling, 2013), when maximizing the above joint log-likelihood, we can maximize the following *ELBO*:

$$\begin{aligned}\max_{\theta,\phi} \mathcal{L}_{\text{ELBO}} = &\underbrace{\mathbb{E}_{q_\phi(\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C}|\mathcal{V}, \mathcal{C})}[\log p_\theta(\boldsymbol{v}, \boldsymbol{c}|\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C})]}_{\text{encoder-decoder}} \\ &\underbrace{-D_{KL}[q_\phi(\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C}|\mathcal{V}, \mathcal{C})||p(\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C})]}_{\text{joint prior regularization}},\end{aligned} \tag{11}$$

where $p_\theta(\boldsymbol{v}, \boldsymbol{c}|\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C})$ is the conditional distribution parameterized by θ. The first term actually shows an encoder-decoder architecture, while the second term indicates a joint prior regularization.

The Encoder-Decoder: Under a mean-field posterior assumption, the encoded latent representations $\boldsymbol{z}_\mathcal{V}$ and $\boldsymbol{z}_\mathcal{C}$ are conditionally independent given $\mathcal{V}$ and $\mathcal{C}$, and the posterior can be written as:

$$\begin{aligned}q_\phi(\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C}|\mathcal{V}, \mathcal{C}) &= q_{\phi_1}(\boldsymbol{z}_\mathcal{V}|\boldsymbol{z}_\mathcal{C}, \mathcal{V}, \mathcal{C}) q_{\phi_2}(\boldsymbol{z}_\mathcal{C}|\mathcal{V}, \mathcal{C}) \\ &= q_{\phi_1}(\boldsymbol{z}_\mathcal{V}|\mathcal{V}, \mathcal{C}) q_{\phi_2}(\boldsymbol{z}_\mathcal{C}|\mathcal{V}, \mathcal{C}),\end{aligned} \tag{12}$$

This factorization is applied after representation-level fusion: both latent representations may depend on the action and context histories, while the approximate posterior factorizes across the two latent variables.

Similarly, we assume that $\boldsymbol{v}$ and $\boldsymbol{c}$ are conditionally independent given $\boldsymbol{z}_\mathcal{V}$ and $\boldsymbol{z}_\mathcal{C}$, the conditional distribution is written as:

$$\begin{aligned}p_\theta(\boldsymbol{v}, \boldsymbol{c}|\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C}) &= p_{\theta_1}(\boldsymbol{v}|\boldsymbol{c}, \boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C}) p_{\theta_2}(\boldsymbol{c}|\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C}) \\ &= p_{\theta_1}(\boldsymbol{v}|\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C}) p_{\theta_2}(\boldsymbol{c}|\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C}),\end{aligned} \tag{13}$$

which indicates we employ both the information from $\boldsymbol{z}_\mathcal{V}$ and $\boldsymbol{z}_\mathcal{C}$ to make the individual prediction of $\boldsymbol{v}$ and $\boldsymbol{c}$.

The Joint Prior Regularization: The second term in Eq. 11 represents a joint prior on the posterior $q_\phi(\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C}|\mathcal{V}, \mathcal{C})$ for $\boldsymbol{z}_\mathcal{V}$ and $\boldsymbol{z}_\mathcal{C}$. Given the complexity of the joint prior $p(\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C})$, we simplify its implementation by assuming $p(\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C}) = p(\boldsymbol{z}_\mathcal{V})p(\boldsymbol{z}_\mathcal{C})$. This choice aligns with recent works (Chen et al., 2023b; Tomczak & Welling, 2017), allowing for a more straightforward and efficient implementation. By integrating it with Eq. 12, the joint prior regularization (Eq. 11) can be formulated as:

$$\begin{aligned}&D_{KL}[q_\phi(\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C}|\mathcal{V}, \mathcal{C})||p(\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C})] \\ &= D_{KL}[q_\phi(\boldsymbol{z}_\mathcal{V}|\mathcal{V}, \mathcal{C}) q_\phi(\boldsymbol{z}_\mathcal{C}|\mathcal{V}, \mathcal{C})||p(\boldsymbol{z}_\mathcal{V})p(\boldsymbol{z}_\mathcal{C})] \\ &= D_{KL}[q_\phi(\boldsymbol{z}_\mathcal{V}|\mathcal{V}, \mathcal{C})||p(\boldsymbol{z}_\mathcal{V})] + D_{KL}[q_\phi(\boldsymbol{z}_\mathcal{C}|\mathcal{V}, \mathcal{C})||p(\boldsymbol{z}_\mathcal{C})].\end{aligned} \tag{14}$$

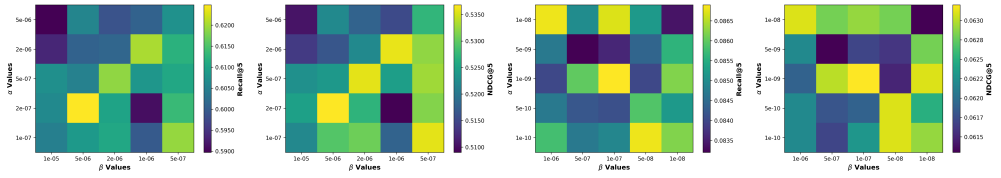
The factorized prior is a simplifying assumption rather than a claim that the latent variables are intrinsically independent. More expressive joint priors could be considered in future work (Tomczak & Welling, 2017; Rezende & Mohamed, 2015; Yin & Zhou, 2018).

Rewrite the $\mathcal{L}_{\text{ELBO}}$: By integrating Eq. 12, Eq. 13 and Eq. 14 together, we can rewrite the *ELBO* in Eq. 11 as follows:

$$\begin{aligned}&\max_{\theta_1,\theta_2,\phi_1,\phi_2} \mathcal{L}_{\text{ELBO}} = \\ &\mathbb{E}_{q_{\phi_1}(\boldsymbol{z}_\mathcal{V}|\mathcal{V}, \mathcal{C}) q_{\phi_2}(\boldsymbol{z}_\mathcal{C}|\mathcal{V}, \mathcal{C})}[\log p_{\theta_1}(\boldsymbol{v}|\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C}) p_{\theta_2}(\boldsymbol{c}|\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C})] \\ &- D_{KL}[q_{\phi_1}(\boldsymbol{z}_\mathcal{V}|\mathcal{V}, \mathcal{C})||p(\boldsymbol{z}_\mathcal{V})] - D_{KL}[q_{\phi_2}(\boldsymbol{z}_\mathcal{C}|\mathcal{V}, \mathcal{C})||p(\boldsymbol{z}_\mathcal{C})].\end{aligned} \tag{15}$$

Table 6: Time complexity of different models on behavior prediction part.

Method	Behavior Prediction	Remark
Bert4Rec	$O(B * \rho * \|\mathcal{T}\| * K^v)$	ρ: the ratio of sequence tokens for Cloze task
CARCA	$O(B * K^v + B * H' * (K^v * N') * d)$	H': the number of heads in cross attention N': the number of user-side features
ContextCTC	$O(B * K^v)$	no contrastive loss
MSDP	$O(B * K^v + B^2 * d)$	conventional contrastive loss of input sequence
ContraRec	$O(B * K^v + B^2 * d)$	conventional contrastive loss of input sequence
ContextContraRec	$O(B * K^v + B^2 * d)$	conventional contrastive loss of input sequence
FCSNet	$O(B * K^v + 2 * B^2 * d)$	CCR loss of two input sequences



(a) Recall@5 (b) NDCG@5 (c) Recall@5 (d) NDCG@5

Figure 6: Sensitivity to α (vertical axis) and β (horizontal axis). (a) and (b)are on Oubrain, (c) and (d) are on AllChannlelPay-1m.

The Objective Function: Maximizing the above ELBO is equivalent to minimizing its negative version, which is:

$$\begin{aligned}\min_{\theta_1,\theta_2,\phi_1,\phi_2} \mathcal{L} = &- \mathbb{E}_{q_{\phi_1}(\boldsymbol{z}_\mathcal{V}|\mathcal{V},\mathcal{C})}[\log p_{\theta_1}(\boldsymbol{v}|\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C})] \\ &- \mathbb{E}_{q_{\phi_2}(\boldsymbol{z}_\mathcal{C}|\mathcal{V},\mathcal{C})}[\log p_{\theta_2}(\boldsymbol{c}|\boldsymbol{z}_\mathcal{V}, \boldsymbol{z}_\mathcal{C})] \\ &+ D_{KL}[q_{\phi_1}(\boldsymbol{z}_\mathcal{V}|\mathcal{V},\mathcal{C})||p(\boldsymbol{z}_\mathcal{V})] \\ &+ D_{KL}[q_{\phi_2}(\boldsymbol{z}_\mathcal{C}|\mathcal{V},\mathcal{C})||p(\boldsymbol{z}_\mathcal{C})]. \end{aligned} \tag{16}$$

The first two terms correspond to predicting the future action $\boldsymbol{v}$ and context $\boldsymbol{c}$ from the historical sequences. The last two terms regularize the latent distributions; in FCSNet, this distribution matching is approximated through adversarial learning (Makhzani et al., 2016b).

B Model Complexity Comparison

We compare model complexity in terms of two components common to sequential behavior prediction: feature encoding and behavior prediction. Let B be the batch size, L be the number of transformer layers, $|\mathcal{T}|$ be the sequence length of samples, H be the head number and d be the dimension of each head. All methods use a transformer encoder with time complexity $O(B * L * H * |T|^2 * d)$, which is similar across models. The main differences lie in behavior prediction, where K^v (the number of candidate items) can reach millions in industrial scenarios, and $|\mathcal{T}|$ is typically set to 100 in our applications. As shown in Table 6, FCSNet has much lower complexity than CARCA and Bert4Rec, and is comparable to methods like ContextContraRec. This makes FCSNet practical for large-scale industrial use, and we have successfully deployed it on 16 A100 GPUs for daily user representation update.

C More Experiments

Parameter Sensitivity: Figure 6 reports Recall@5 and NDCG@5 over the loss weights α and β on Outbrain and AllChannelPay-1m. The heatmaps show the local sensitivity around the selected configurations.

Transferability of CCR to ContextContraRec: To evaluate whether CCR transfers beyond FCSNet, we replace the conventional contrastive loss in ContextContraRec with CCR. As shown in Table 7, CCR improves six of eight metrics across Outbrain and AllChannelPay-1m, demonstrating its compatibility with other sequential models.

FCSNet with LLM-based Sequential Recommendation: FCSNet is compatible with recent LLM-based sequential recommendation models. We combine it with popular LLM-ESR (Liu et al., 2024)

Table 7: Results of integrating CCR into ContextContraRec.

Dataset	Outbrain				AllChannelPay-1m			
Method	R@5	N@5	R@10	N@10	R@5	N@5	R@10	N@10
ContextContraRec	**0.4979**	0.4027	0.5448	0.4182	0.0762	0.0544	0.1045	0.0635
ContextContraRec+CCR	0.4966	**0.4088**	**0.5587**	**0.4250**	**0.0787**	**0.0571**	**0.1033**	**0.0651**

Table 8: Effect of $\mathcal{L}_{\text{APR}}$ under different training ratios.

Dataset	AllChannelPay-1m			
Train Ratio	40%	60%	80%	100%
FCSNet(w/o $\mathcal{L}_{\text{APR}}$)	0.0299	0.0611	0.0973	0.1158
FCSNet	**0.0475**	**0.0685**	**0.1002**	**0.1155**

on AllChannelPay-10m. The Recall@5 of FCSNet, LLM-ESR and **FCSNet+LLM-ESR** is 0.1712, 0.2338, **0.2667** respectively.

Study of CCR Variants: In CCR, similarity-adaptive weights emphasize harder negative pairs, while a reverse-CCR variant emphasizes more distinct ones. On AllChannelPay-1m, Recall@5 for uniform weights, reverse-CCR, and CCR is 0.0851, 0.0790, and 0.0870, respectively, showing that the reweighting strategy affects performance.

Effect of $\mathcal{L}_{\text{APR}}$ Under Different Train Data: Table 8 shows that prior regularization provides larger Recall@10 gains in sparse-data regimes, while its effect becomes negligible at the full training ratio.