# Aligned Side Information Fusion Method for Sequential Recommendation


Shuhan Wang
hanshu.wsh@antgroup.com
Ant Group
Hangzhou, China

Bin Shen
ringo.sb@antgroup.com
Ant Group
Hangzhou, China

Xu Min
minxu.mx@antgroup.com
Ant Group
Hangzhou, China

Yong He
heyong.h@antgroup.com
Ant Group
Hangzhou, China

Xiaolu Zhang
yueyin.zxl@antgroup.com
Ant Group
Hangzhou, China

Liang Zhang
zhuyue.zl@antgroup.com
Ant Group
Hangzhou, China

Jun Zhou
jun.zhoujun@antgroup.com
Ant Group
Hangzhou, China

Linjian Mo
linyi01@antgroup.com
Ant Group
Hangzhou, China



## ABSTRACT

Combining contextual information (i.e., side information) of items beyond IDs has become an important way to improve the performance in recommender systems. Existing self-attention-based side information fusion methods can be categorized into early, late, and hybrid fusion. In practice, naive early fusion may interfere with the representation of IDs, resulting in negative effects, while late fusion misses effective interactions between IDs and side information. Some hybrid methods have been proposed to address these issues, but they only utilize side information in calculating attention scores, which may lead to information loss. To harness the full potential of side information without noisy interference, we propose an **A**ligned **S**ide **I**nformation **F**usion (**ASIF**) method for sequential recommendation, consisting of two parts: Fused Attention with Untied Positions and Representation Alignment. Specifically, we first decouple the positions to exclude the noisy interference in the attention scores. Secondly, we adopt the contrastive objective to maintain the semantic consistency between IDs and side information and then employ orthogonal decomposition to extract the homogeneous parts. By aligning the representations and fusing them together, ASIF makes full use of the side information without interfering with IDs. Offline experimental results on four datasets demonstrate the superiority of ASIF. Additionally, we successfully deployed the model in Alipay's advertising system and achieved 1.09% and 1.86% improvements on clicks and Cost Per Mille (CPM).



## CCS CONCEPTS

• **Information systems** → **Online advertising**; • **Computing methodologies** → *Artificial intelligence.*




*WWW '24 Companion, May 13–17, 2024, Singapore, Singapore*



ACM ISBN 979-8-4007-0172-6/24/05
https://doi.org/10.1145/3589335.3648308

## KEYWORDS

Sequential Recommendation, Side Information Fusion, Orthogonal Decomposition

**ACM Reference Format:**
Shuhan Wang, Bin Shen, Xu Min, Yong He, Xiaolu Zhang, Liang Zhang, Jun Zhou, and Linjian Mo. 2024. Aligned Side Information Fusion Method for Sequential Recommendation. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3589335.3648308


## 1 INTRODUCTION

Sequential recommendation plays an important role in industrial scenarios such as e-commerce, advertising, and search systems, and its main goal is to model the user's historical behavior to predict the next item that may be of interest to the user. Among various solutions [6, 16], the attention-based models [10, 15, 20, 21] are gradually becoming the mainstream due to excellent performance.

Early self-attention-based models like BERT4Rec [15] and SASRec [7] only consider item IDs, lacking the ability to capture item attributes beyond IDs. This limitation becomes apparent when IDs change frequently. For example, in a typical recommendation scenario on the Alipay membership page, users are shown items that can be redeemed using points and money. The product pool is frequently updated with advertising programs, causing rapid changes in item IDs. Attributes such as categories and brands offer a more stable representation of a user's long-term preferences. Thus, we aim to incorporate side information into the recommendation model to boost performance.

Based on the varying fusion locations, the existing self-attention-based side information fusion methods can be categorized into three types: early, late, and hybrid fusion. The early fusion fuses IDs with side information together before feeding them into the attention block. In contrast, the late fusion applies separated self-attention blocks on item-level and feature-level sequences and fuses them until the final stage. It has been pointed that the early fusion may not always improve performance but instead impair the representation

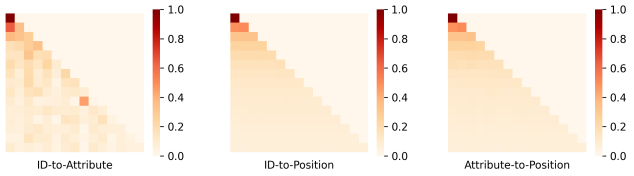
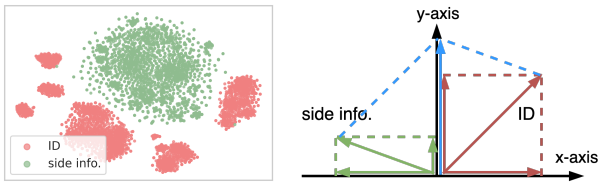



**Figure 1: Visualization of attention scores in SASRec$_F$ on Yelp dataset.**




(a) Macroscopic perspective (b) Microscopic perspective

**Figure 2: Visualization of information invasion.**

of the IDs, causing a phenomenon known as information invasion [11]. The late fusion, on the other hand, lacks the interaction between IDs and side information and losses some prior information. Consequently, some hybrid-fusion methods have emerged recently. They avoid information invasion by incorporating side information in attention score calculation and explore interesting structures for attention correlations.

Despite the remarkable improvements, these approaches still suffer from two limitations: (1) Correlations between IDs and attributes can vary, with some being strong and others being weak, making it difficult to eliminate interference and learn meaningful correlations effectively. (2) Methods that completely exclude side information from the final representation to prevent information invasion may inadvertently discard crucial information within the side information itself.

In this work, we try to enhance the utilization of side information by mitigating noise interference. Inspired by [8], we expand the fusion form of attention scores of early-fusion method SASRec$_F$. As shown in Fig. 1, IDs have a strong relationship with attributes, while the correlations between position encoding and others are relatively weak. This indicates the common way to fuse position as ordinary side information may potentially introduce noise into the attention scores. We also examine the representation spaces of IDs and side information in SASRec$_F$ on the Yelp dataset to provide an explanation for information invasion. From a macroscopic perspective, we can observe a significant dissimilarity between the two distributions (see Fig. 2(a)), indicating that the representation space after fusion will deviate considerably from the original ID space. From a microscopic perspective, by projecting both the ID and side information embeddings onto a coordinate system, we uncover that if the directions of the two are opposite on certain axes, these segments of vectors may cancel each other out, leading to a loss of information (see Fig. 2(b)).

To address the above issues, we propose a novel method called Aligned Side Information Fusion (ASIF). First, we introduce Fused Attention with Untied Positions, which separates the ID-attributes from the position encoding during attention score calculation, eliminating noise interference and preserving the strong correlation. Second, we propose Representation Alignment, consisting of two steps: Representation Space Alignment (RSA) and Homogeneous Information Extraction (HIE). The RSA approach employs a contrastive objective for paired ID and attribute at the interaction granularity within the sequence to ensure their semantic consistency. Although this operation brings the two distributions closer together, it still can not avoid the existence of heterogeneous parts. Therefore, HIE performs orthogonal decomposition on IDs and side information to extract the homogeneous parts, thus avoiding information invasion. In summary, our main contributions can be summarized as follows:

- We meticulously design the ASIF framework, based on Fused Attention With Untied Position and Representation Alignment, to enhance the recommendation performance by leveraging side information.
- In terms of Representation Alignment, we propose RSA and HIE. By employing contrastive loss and orthogonal decomposition, we align the representation space of IDs and side information in both macroscopic and microscopic aspects, effectively preventing the problem of information invasion.
- Offline and online experiments demonstrate the effectiveness of our proposed method.

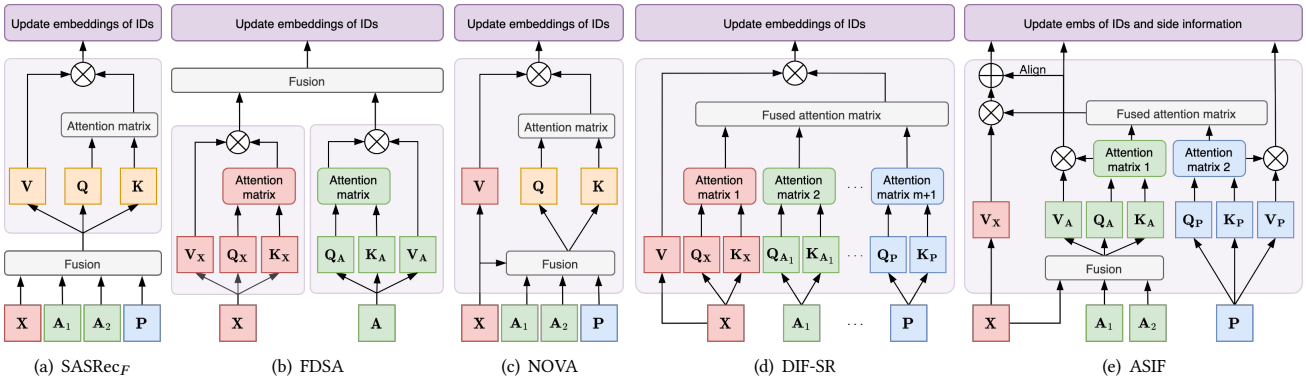
## 2 RELATED WORKS

### 2.1 Sequential Recommendation

Sequential recommendation aims to predict the next item that is most likely to be interacted with based on the user's historical behaviors. With the development of deep learning techniques recent years, many neural network based methods such as Convolutional Neural Networks (CNNs) [16, 19], Recurrent Neural Networks (RNNs) [13], Graph Neural Networks (GNNs) [3] and attention-based models start to emerge. Among them, the self-attention-based methods have made significant progress. SASRec [7] introduces self-attention into the SR model to capture long-range dependencies . BERT4Rec [15] adopts the Cloze objective and improves the performance by bidirectional self-attention mechanism. Recent SR methods also use contrastive learning to augment the data, including CL4SRec [17] and DuoRec [12]. These works utilize only item IDs, ignoring other attributes associated with the item, which may potentially help to extract comprehensive sequence patterns.

### 2.2 Side Information Fusion for Sequential Recommendation

Instead of using item IDs only as the above solutions, the side information, such as other item attributes and ratings, is taken into consideration to capture meaningful supervision signals. $S^3$Rec [24] notices the important information contained in the attributes and devises four auxiliary self-supervised tasks to learn the intrinsic relationship. Besides utilizing side information in auxiliary tasks, the end-to-end fusion approaches are beginning to be explored.




(a) SASRec$_F$ (b) FDSA (c) NOVA (d) DIF-SR (e) ASIF

**Figure 3: Single layer structure comparison of existing self-attention-based side information fusion approaches: SASRec$_F$ is early fusion, FDSA is late fusion, while NOVA, DIF-SR and ASIF is hybrid fusion.**

Following the classification system of multi-modal fusion [1, 2], we categorize the self-attention-based side information fusion methods into three types: early, late, and hybrid fusion. In early fusion, IDs and side information are combined at the shallow layers of the model, which are then fed into the network and generate outputs. For example, SASRec$_F$ [24] combines ID and attributes and feeds them into the self-attention block as input (see Fig. 3(a)). In late fusion, the networks of ID and side information are independent, and fusion takes place just before the predict layer. FDSA [22] is the late-fusion method, which applies separated self-attention blocks on item-level and feature-level sequences and concatenates their hidden states until the final stage (see Fig. 3(b)).

Both early and late fusion have their own limitations. The former cannot exclude noisy interference and may result in information invasion, while the latter lacks effective interaction between IDs and attributes. Hybrid fusion lies between them, allowing IDs and side information to interact in the middle layer. NOVA [11] first defines the information invasion problem caused by naive early fusion and proposes only to incorporate attributes in the calculation of attention scores to mitigate it (see Fig. 3(c)). However, it regards the position as an ordinary attribute, introducing noise into mixed attention. Furthermore, DIF-SR [18] decouples the attention scores for IDs and side information, allowing higher-rank attention matrices and flexible gradients (see Fig. 3(d)). Unfortunately, it abandons the implicit cross-relationships between IDs and attributes. Both methods utilize side information only in the attention scores, completely discarding it in the value matrices, which may result in a loss of information. Our work aims to fill these gaps, reducing noisy interference while enhancing the utilization of side information.

## 3 METHODOLOGY

The overall framework of ASIF is shown in Fig. 4, and the details will be introduced next.

### 3.1 Problem Formulation

In sequential recommendation with side information, let $\mathcal{U}$, $\mathcal{V}$, $\mathcal{X}$ and $\mathcal{A}_j$ denote the sets of users, items, item IDs and the $j$-th type of attributes, respectively. Let $\mathcal{S}_u = [\mathbf{v}_u^{(1)}, \mathbf{v}_u^{(2)}, \cdots, \mathbf{v}_u^{(n)}]$ denotes the historical sequence of interactions in chronological order for user $u \in \mathcal{U}$, where $\mathbf{v}_u^{(t)} \in \mathcal{V}$ is the $t$-th item in the user interaction sequence and $n$ is the maximum length of the sequence. Suppose we have $m$ types of side information, then $\mathbf{v}_u^{(t)} = \{\mathbf{x}_u^{(t)}, \mathbf{a}_{1,u}^{(t)}, \mathbf{a}_{2,u}^{(t)}, \cdots, \mathbf{a}_{m,u}^{(t)}\}$, where $\mathbf{x}_u^{(t)} \in \mathcal{X}$ is the item ID of the $t$-th interaction, and $\mathbf{a}_{j,u}^{(t)} \in \mathcal{A}_j$ represents the $j$-th type of the attributes of the $t$-th interaction. Given the interaction history $\mathcal{S}_u$, the goal of sequential recommendation is to predict the next item that the user $u$ may be interested in. It can be formalized as modeling the probability over all candidate items for user $u$: $P(\mathbf{v}_u^{(n+1)} = \mathbf{v} \mid \mathcal{S}_u)$.

### 3.2 Fused Attention with Untied Positions

For attention-based models, the naive way to incorporate side information is to fuse it with IDs and input into the attention block (see Fig. 3(a)). NOVA follows this structure but excludes side information from the value matrix (see Fig. 3(c)), while DIF-SR advises applying decoupled attention calculation of various side information and IDs representations (see Fig. 3(d)), ensuring flexible gradients. However, according to Fig.1, IDs have a strong relationship with attributes, but position encoding has a weak relationship with IDs and attributes, which may not be suitable to be fused with others. Therefore, we propose the Fused Attention (FA) with Untied Positions (UP) (see Fig. 3(e)).

Let $\mathbf{X}$ and $\mathbf{A}$ represent the embedding matrices of item IDs and attributes, we first fuse them and compute the correlation matrix as

$$\mathbf{C}_{\mathbf{XA}} = \mathcal{F}(\mathbf{X}, \mathbf{A})\mathbf{W}_{q,1}\mathbf{W}_{k,1}^T\mathcal{F}(\mathbf{X}, \mathbf{A})^T, \tag{1}$$

where $\mathbf{W}_{q,1} \in \mathbb{R}^{d \times d_h}$, $\mathbf{W}_{k,1} \in \mathbb{R}^{d \times d_h}$. $\mathcal{F}$ denotes the fusion function, e.g., $\mathcal{F}_{sum}(\mathbf{X}, \mathbf{A}) = \mathbf{X} + \sum_{j=1}^{m} \mathbf{A}_j$. Next, we compute the correlation matrix of the position encoding as

$$\mathbf{C}_{\mathbf{P}} = \mathbf{P}\mathbf{W}_{q,2}\mathbf{W}_{k,2}^T\mathbf{P}^T, \tag{2}$$

where $\mathbf{P}$ denotes the absolute position embedding matrix, $\mathbf{W}_{q,2} \in \mathbb{R}^{d \times d_h}$, and $\mathbf{W}_{k,2} \in \mathbb{R}^{d \times d_h}$.

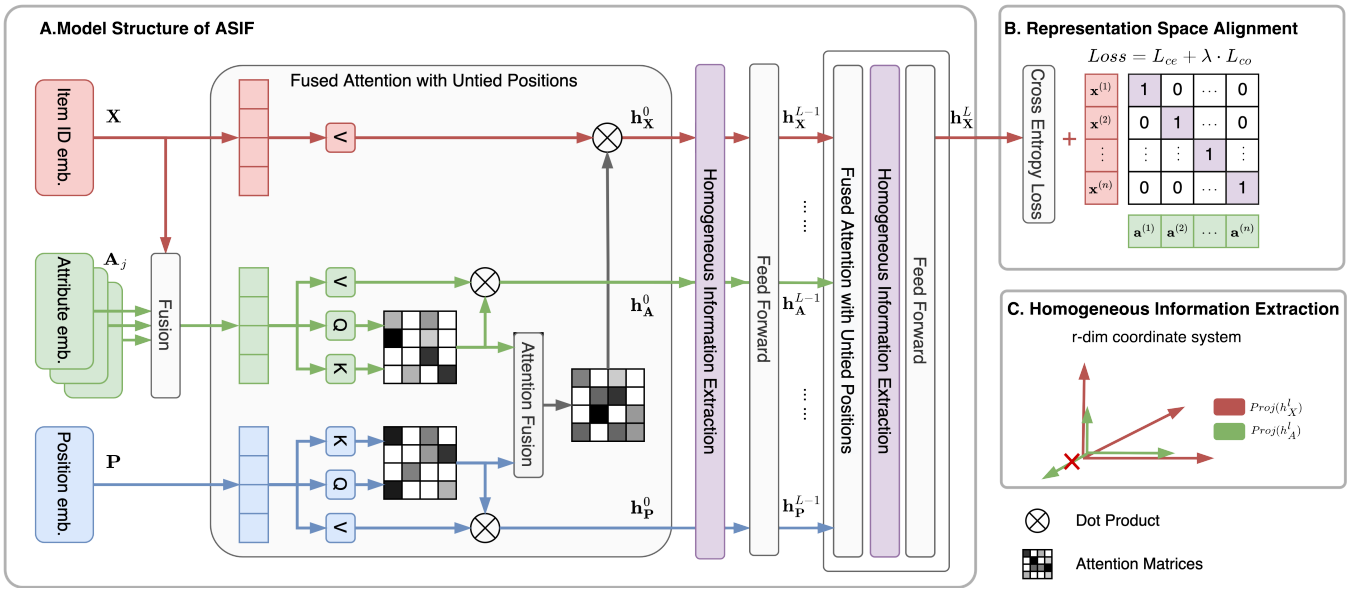



**Figure 4: An overview of ASIF. The model is stacked with the Fused Attention with Untied Positions block, which decouples the computation of position and focuses on effective interaction between IDs and attributes. Through the Representation Space Alignment (RSA) and Homogeneous Information Extraction (HIE), item IDs and attributes' representations are aligned and the homogeneous parts are accurately captured.**

Then, we fuse the two correlation matrices and obtain the final attention formula as follows:

$$\begin{aligned}\mathbf{h}_{\mathbf{X}} &= \text{FusedAttention}(\mathbf{X}, \mathbf{A}_1, \cdots, \mathbf{A}_m, \mathbf{P}) \\ &= \text{Softmax}\left(\frac{\mathbf{C}_{\mathbf{XA}} + \mathbf{C}_{\mathbf{P}}}{\sqrt{d_h}}\right)\mathbf{X}\mathbf{W}_{v,1}, \end{aligned} \tag{3}$$

where $\mathbf{W}_{v,1} \in \mathbb{R}^{d\times d}$. $\mathbf{h}_{\mathbf{X}}$ denotes the hidden state of the item IDs. Finally, we consider the side information important enough to be fully learned, so we also pass and update them between different Transformer layers as follows:

$$\begin{aligned}\mathbf{h}_{\mathbf{A}} &= \text{FusedAttention}(\mathbf{X}, \mathbf{A}_1, \cdots, \mathbf{A}_m) \\ &= \text{Softmax}\left(\frac{\mathbf{C}_{\mathbf{XA}}}{\sqrt{d_h}}\right)\mathcal{F}(\mathbf{X}, \mathbf{A})\mathbf{W}_{v,2}, \end{aligned} \tag{4}$$

$$\mathbf{h}_{\mathbf{P}} = \text{FusedAttention}(\mathbf{P}) = \text{Softmax}\left(\frac{\mathbf{C}_{\mathbf{P}}}{\sqrt{d_h}}\right)\mathbf{P}\mathbf{W}_{v,3}, \tag{5}$$

where $\mathbf{W}_{v,2} \in \mathbb{R}^{d\times d}, \mathbf{W}_{v,3} \in \mathbb{R}^{d\times d}$, and $\mathbf{h}_{\mathbf{A}}$ and $\mathbf{h}_{\mathbf{P}}$ denote the hidden states of the attributes and the positions, respectively.

## 3.3 Representation Alignment

From macroscopic and microscopic perspectives, the occurrence of invasion phenomenon may be due to excessive distribution deviation and vector offset, respectively. To solve this, we propose Representation Space Alignment (RSA) and Homogeneous Information Extraction (HIE) to align the representations of IDs and attributes. The goal of the former is to narrow the representation space of both item IDs and attributes to improve the semantic consistency at the interaction granularity (see Fig. 5(a)). The latter extracts the information in the attributes that is homogeneous with the item IDs and fuses it into the IDs representation (see Fig. 5(b)).

*3.3.1 Representation Space Alignment (RSA).* Taking inspiration from CLIP's alignment operation [14], we leverage a contrastive loss to align the embedding spaces of item IDs and attributes, intending to bring the two distributions closer (see Fig. 5(a)). However, unlike CLIP, our alignment occurs at the interaction granularity within a sequence rather than at the sample granularity. Specifically, $\mathbf{X}$ and $\mathbf{A}$ represent the embedding matrices of item IDs and attributes:

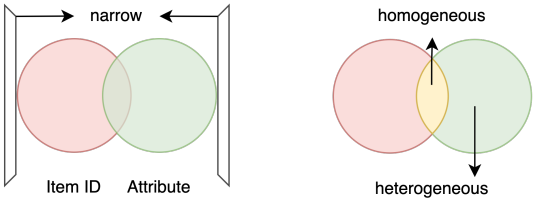
$$\mathbf{X} = \begin{bmatrix}\mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ \vdots \\ \mathbf{x}^{(n)}\end{bmatrix}, \quad \mathbf{A} = \sum_{j=1}^{m}\mathbf{A}_j = \begin{bmatrix}\mathbf{a}^{(1)} \\ \mathbf{a}^{(2)} \\ \vdots \\ \mathbf{a}^{(n)}\end{bmatrix}, \tag{6}$$

where $\mathbf{x}^{(t)}, \mathbf{a}^{(t)} \in \mathbb{R}^{1\times d}$ denote the embeddings of the item ID and the attribute of the $t$-th interaction in the sequence, and $\mathbf{X}, \mathbf{A} \in \mathbb{R}^{n\times d}$. Next, we calculate the cosine similarity between the two embeddings to get the final matching scores as follows:

$$\widetilde{\mathbf{X}} = \begin{bmatrix}\mathbf{x}^{(1)}/\|\mathbf{x}^{(1)}\| \\ \mathbf{x}^{(2)}/\|\mathbf{x}^{(2)}\| \\ \vdots \\ \mathbf{x}^{(n)}/\|\mathbf{x}^{(n)}\|\end{bmatrix}, \quad \widetilde{\mathbf{A}} = \begin{bmatrix}\mathbf{a}^{(1)}/\|\mathbf{a}^{(1)}\| \\ \mathbf{a}^{(2)}/\|\mathbf{a}^{(2)}\| \\ \vdots \\ \mathbf{a}^{(n)}/\|\mathbf{a}^{(n)}\|\end{bmatrix}, \tag{7}$$

$$\widehat{\mathbf{Y}}_{\mathbf{X}} = \text{Softmax}\left(\widetilde{\mathbf{X}}\widetilde{\mathbf{A}}^T/\tau\right), \quad \widehat{\mathbf{Y}}_{\mathbf{A}} = \text{Softmax}\left(\widetilde{\mathbf{A}}\widetilde{\mathbf{X}}^T/\tau\right), \tag{8}$$

where Softmax$(\cdot)$ is executed for each row of the similarity matrix and $\tau$ denotes the learnable temperature coefficient. Finally, we




(a) RSA (b) HIE

**Figure 5: Two steps of Representation Alignment.**

calculate the contrastive loss in the following form:

$$L_{co} = -\frac{1}{2N}\sum_{i=1}^{N}\sum\left(\mathbf{Y}^i \odot \log \widehat{\mathbf{Y}}^i_{\mathbf{X}} + \mathbf{Y}^i \odot \log \widehat{\mathbf{Y}}^i_{\mathbf{A}}\right), \tag{9}$$

where $\odot$ is the element-wise product, $N$ is the sample size, and $\mathbf{Y}^i$ is the ground truth of the $i$-th sample, which is an identity matrix $\mathbf{Y}^i = \mathbf{I}_n = \text{diag}(1, 1, \cdots, 1)$, meaning that only the paired item IDs and attributes are positive examples.

*3.3.2 Homogeneous Information Extraction (HIE).* The space alignment brings the two distributions closer, but still cannot avoid the existence of the heterogeneous part. Therefore, we propose performing orthogonal decomposition on each layer's hidden states to extract the homogeneous parts (see Fig. 5(b)).

Intuitively, if an attribute's representation is in the same direction as ID's, it should be maximally preserved. Otherwise, there may be a conflict, and it should be discarded. Thus we need an $r$-dim orthogonal coordinate system as the comparison granularity, which needs to fully accommodate all the IDs' representations in a user's interaction sequence. Specifically, we first perform a QR decomposition of the IDs' hidden state: ${\mathbf{h}_{\mathbf{X}}}^T = \mathbf{QR}$, where $\mathbf{Q} \in \mathbb{R}^{d\times n}$ is an orthogonal matrix, and $\mathbf{R} \in \mathbb{R}^{n\times n}$ is an upper triangular matrix. Then, we map both hidden states into $\mathbf{Q}$ to get the coordinate matrices as follows:

$$\text{Proj}(\mathbf{h}_{\mathbf{X}}) = \mathbf{h}_{\mathbf{X}}\mathbf{Q}, \quad \text{Proj}(\mathbf{h}_{\mathbf{A}}) = \mathbf{h}_{\mathbf{A}}\mathbf{Q}, \tag{10}$$

where $\text{Proj}(\mathbf{h}_{\mathbf{X}}), \text{Proj}(\mathbf{h}_{\mathbf{A}}) \in \mathbb{R}^{n\times n}$. Thus we can obtain the homogeneous part $\mathbf{h}^*_{\mathbf{A}} \in \mathbb{R}^{n\times d}$ as follows:

$$\widetilde{\text{Proj}}(\mathbf{h}_{\mathbf{A}}) = \phi\left(\text{Proj}(\mathbf{h}_{\mathbf{X}}) \odot \text{Proj}(\mathbf{h}_{\mathbf{A}})\right) \odot \text{Proj}(\mathbf{h}_{\mathbf{A}}), \tag{11}$$

$$\mathbf{h}^*_{\mathbf{A}} = \widetilde{\text{Proj}}(\mathbf{h}_{\mathbf{A}})\mathbf{Q}^T, \tag{12}$$

where $\odot$ is the element-wise product, $\phi(\cdot)$ is the indicator function, which outputs 1 if the value is greater than 0, and 0 for the rest. Since $\mathbf{h}^*_{\mathbf{A}}$ is homogeneous with $\mathbf{h}_{\mathbf{X}}$, we can directly fusing it into the item representation, Eq.3 can be updated:

$$\begin{aligned}\mathbf{h}_{\mathbf{X}} &= \text{FusedAttention}(\mathbf{X}, \mathbf{A}_1, \cdots, \mathbf{A}_m, \mathbf{P}) + {\mathbf{h}_{\mathbf{A}}}^* \\ &= \text{Softmax}\left(\frac{\mathbf{C}_{\mathbf{XA}} + \mathbf{C}_{\mathbf{P}}}{\sqrt{d_h}}\right)\mathbf{X}\mathbf{W}_{v,1} + {\mathbf{h}_{\mathbf{A}}}^*.\end{aligned} \tag{13}$$

Since the average sequence length of users is often lower than $n$, we can reduce the dimension of ${\mathbf{h}_{\mathbf{X}}}^T$ as ${\mathbf{h}_{\mathbf{X}}}^T\mathbf{W}_r$, where $\mathbf{W}_r \in \mathbb{R}^{n\times r}$, to reduce the computational complexity as well as the redundancy of parameters before doing QR decomposition.

**Table 1: Statistics of datasets.**

| Dataset | # Users | # Items | # Actions | # Avg. len |
|---|---|---|---|---|
| Yelp | 30450 | 20039 | 316541 | 10.4 |
| AliEC | 34148 | 18654 | 290490 | 8.5 |
| Beauty | 22364 | 12102 | 198502 | 8.9 |
| Industrial | 33061 | 19873 | 290000 | 8.8 |

## 3.4 Model Prediction and Learning

After $L$ layers of Transformer structure, we get the final hidden state $\mathbf{h}^L_{\mathbf{X}}$ of the item ID, and calculate the prediction score as:

$$\widehat{y} = \text{Softmax}(\mathbf{h}^L_{\mathbf{X}} \cdot \mathbf{V}), \tag{14}$$

where $\mathbf{V} \in \mathbb{R}^{|\mathbf{V}|\times d}$ is the candidate item matrix. For the sequential recommendation task, we adopt the cross-entropy loss function as

$$L_{ce} = -\frac{1}{N}\sum_{i=1}^{N} y^i \log \widehat{y}^i, \tag{15}$$

where $y^i$ and $\widehat{y}^i$ denote the ground truth and the predictive probability of the $i$-th sample. Finally, combining the contrastive loss in RSA, we define the loss function with the balance coefficient $\lambda$:

$$\begin{aligned}L &= L_{ce} + \lambda \cdot L_{co} \\ &= -\frac{1}{N}\sum_{i=1}^{N}\left(y^i \log \widehat{y}^i + \frac{\lambda}{2}\sum\left(\mathbf{Y}^i \odot \log \widehat{\mathbf{Y}}^i_{\mathbf{X}} + \mathbf{Y}^i \odot \log \widehat{\mathbf{Y}}^i_{\mathbf{A}}\right)\right).\end{aligned} \tag{16}$$

# 4 OFFLINE EXPERIMENTS

In this section, offline experiments are designed to evaluate the performance and effectiveness of ASIF.

## 4.1 Datasets and Settings

*4.1.1 Dataset.* We conduct experiments on three publicly available datasets and an industrial dataset:

- **Yelp**[1] dataset is a well-known business recommendation dataset. Category of business and position are regarded as side information.
- **Amazon Beauty**[2] dataset is collected from Amazon review datasets. Category of the goods and position information are supplementary attributes.
- **AliEC**[3] is a Taobao display advertising dataset provided by Alibaba. We utilize category and position as side information.
- **Industrial** dataset is collected from a scenario in the commercial advertising system in Alipay, which is desensitized and encrypted, and does not contain any Personal Identifiable Information (PII). The position and item's entity such as category and brand, are utilized as side information.

[1] https://www.yelp.com/dataset
[2] http://jmcauley.ucsd.edu/data/amazon/
[3] https://tianchi.aliyun.com/dataset/56

**Table 2: Overall Performance (HR and NDCG) on public datasets. The best results are boldfaced, while the second-best results are underlined. We pick the best model with the highest NDCG@20 on the validation set. Impr. (%) is the performance gain of ASIF against the best baseline method.**

| Model | Yelp | | | | AliEC | | | | Beauty | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | H@10 | H@20 | N@10 | N@20 | H@10 | H@20 | N@10 | N@20 | H@10 | H@20 | N@10 | N@20 |
| Bert4Rec | 0.0354 | 0.0580 | 0.0189 | 0.0246 | 0.0503 | 0.0756 | 0.0263 | 0.0327 | 0.0542 | 0.0793 | 0.0315 | 0.0378 |
| Caser | 0.0357 | 0.0573 | 0.0177 | 0.0231 | 0.0336 | 0.0522 | 0.0171 | 0.0218 | 0.0416 | 0.0672 | 0.0211 | 0.0275 |
| GRU4Rec | 0.0350 | 0.0579 | 0.0175 | 0.0232 | 0.0361 | 0.0567 | 0.0182 | 0.0234 | 0.0510 | 0.0766 | 0.0268 | 0.0333 |
| SASRec | 0.0647 | 0.0936 | 0.0398 | 0.0471 | 0.0903 | 0.1300 | 0.0449 | 0.0549 | 0.0861 | 0.1225 | 0.0424 | 0.0516 |
| LightSANs | 0.0658 | 0.0970 | 0.0402 | 0.0480 | 0.0942 | 0.1354 | 0.0470 | 0.0574 | 0.0871 | 0.1242 | 0.0441 | 0.0535 |
| FMLP | 0.0657 | 0.0935 | 0.0400 | 0.0470 | 0.0936 | 0.1346 | 0.0463 | 0.0566 | 0.0855 | 0.1190 | 0.0450 | 0.0534 |
| GRU4Rec$_F$ | 0.0362 | 0.0605 | 0.0182 | 0.0243 | 0.0471 | 0.0743 | 0.0237 | 0.0305 | 0.0532 | 0.0820 | 0.0274 | 0.0347 |
| SASRec$_F$ | 0.0467 | 0.0749 | 0.0249 | 0.0319 | 0.0719 | 0.1081 | 0.0383 | 0.0474 | 0.0776 | 0.1082 | 0.0447 | 0.0540 |
| LightSANs$_F$ | 0.0641 | 0.0925 | 0.0390 | 0.0461 | 0.0944 | 0.1382 | 0.0469 | 0.0579 | 0.0880 | 0.1244 | 0.0448 | 0.0540 |
| FMLP$_F$ | 0.0629 | 0.0884 | 0.0385 | 0.0448 | 0.0997 | 0.1431 | 0.0495 | 0.0604 | 0.0871 | 0.1220 | 0.0452 | 0.0540 |
| CL4SRec | 0.0666 | 0.0965 | 0.0390 | 0.0465 | 0.0922 | 0.1287 | 0.0464 | 0.0556 | 0.0825 | 0.1180 | 0.0437 | 0.0526 |
| DuoRec | 0.0667 | 0.0962 | 0.0407 | 0.0481 | 0.0863 | 0.1272 | 0.0432 | 0.0535 | 0.0878 | 0.1244 | 0.0451 | 0.0543 |
| FDSA | 0.0668 | 0.0966 | 0.0403 | 0.0478 | 0.0900 | 0.1327 | 0.0456 | 0.0563 | 0.0839 | 0.1209 | 0.0439 | 0.0532 |
| NOVA | 0.0670 | 0.0952 | 0.0407 | 0.0478 | 0.0951 | 0.1382 | 0.0467 | 0.0575 | 0.0866 | 0.1240 | 0.0441 | 0.0535 |
| DIF-SR | 0.0673 | 0.0988 | 0.0412 | 0.0491 | 0.0983 | 0.1419 | 0.0482 | 0.0592 | 0.0871 | 0.1234 | 0.0434 | 0.0526 |
| **ASIF** | **0.0768** | **0.1131** | **0.0452** | **0.0543** | **0.1131** | **0.1631** | **0.0574** | **0.0700** | **0.0922** | **0.1322** | **0.0453** | **0.0554** |
| Impr. | 14.12% | 14.47% | 9.71% | 10.59% | 13.44% | 13.98% | 15.96% | 15.89% | 4.77% | 6.27% | 0.22% | 2.03% |

**Table 3: Performance on the industrial dataset.**

| Model | Industrial | | | |
|---|---|---|---|---|
| | H@10 | H@20 | N@10 | N@20 |
| Bert4Rec | 0.0706 | 0.1187 | 0.0355 | 0.0476 |
| Caser | 0.0808 | 0.1315 | 0.0417 | 0.0544 |
| GRU4Rec | 0.0322 | 0.0575 | 0.0190 | 0.0250 |
| SASRec | 0.0942 | 0.1518 | 0.0480 | 0.0625 |
| LightSANs | 0.0935 | 0.1556 | 0.0466 | 0.0622 |
| FMLP | 0.0939 | 0.1553 | 0.0454 | 0.0608 |
| GRU4Rec$_F$ | 0.0830 | 0.1364 | 0.0433 | 0.0567 |
| SASRec$_F$ | 0.0877 | 0.1385 | 0.0463 | 0.0591 |
| LightSANs$_F$ | 0.0889 | 0.1457 | 0.0454 | 0.0596 |
| FMLP$_F$ | 0.0863 | 0.1438 | 0.0420 | 0.0564 |
| CL4SRec | 0.0683 | 0.1134 | 0.0342 | 0.0455 |
| DuoRec | 0.0917 | 0.1475 | 0.0475 | 0.0615 |
| FDSA | 0.0913 | 0.1496 | 0.0479 | 0.0626 |
| NOVA | 0.0933 | 0.1517 | 0.0456 | 0.0602 |
| DIF-SR | 0.0951 | 0.1559 | 0.0459 | 0.0612 |
| **ASIF** | **0.0996** | **0.1653** | **0.0495** | **0.0660** |
| Impr. | 4.73% | 6.03% | 3.13% | 5.43% |

Following the same data pre-processing ways in [7, 18, 24], we remove all items and users that occur less than five times in public datasets. For the industrial dataset, we retain all users and items that have appeared due to the frequent updating of item IDs. The statistics of all processed datasets are summarized in Tab. 1.

*4.1.2 Baseline Methods.* We compare our model with the following state-of-the-art sequential recommendation methods.

- **Methods without side information.** We take GRU-based model GRU4Rec [6], self-attention-base model BERT4Rec [15], SASRec [7], LightSANs [5], CNN-based model Caser [16] and MLP-based model FMLP [25] as basic baselines.
- **Naive early-fusion methods.** GRU4Rec$_F$, SASRec$_F$, LightSANs$_F$, FMLP$_F$ is the naive early-fusion variants of GRU4Rec, SASRec, LightSANs and FMLP, which fuse IDs and side information together before feeding them into the networks.
- **Advanced self-attention-based side information fusion methods.** We include the late-fusion method FDSA [22], the hybrid-fusion methods NOVA [11] and DIF-SR [18], which are highly related to our work. For a fair comparison, we implement NOVA based on SASRec as in [18].
- **Other relevant methods.** CL4SRec [17] and DuoRec [12] are SR models using contrastive learning objectives.

*4.1.3 Evaluation Metrics.* Following the previous works [7, 18], the leave-one-out strategy is used for evaluation. For each user sequence, we use the last item for testing, the second last item for validation, and the rest items for training. Models are evaluated in a full ranking manner as in [5, 11, 18] rather than negative sampling, which is often criticized for bias [4, 9]. Two widely used metrics are employed: top-K Hit Rate (HR@K) and top-K Normalized Discounted Cumulative Gain (NDCG@K) with K={10, 20}.

*4.1.4 Implementation Details.* We run all the models on the open-source recommendation framework Recbole [23] and evaluate them with the same setting. We set the maximum sequence length to 50 and the embedding size to 256 for all datasets. All the networks

are 3 layers and 4 heads, and the Adam optimizer is adopted for 200 epochs with batch size 2048 and learning rate 1e-4. Fusion functions for side information fusion methods are searched among sum, concat and gating. For other hyperparameters, we follow the best setting mentioned in previous papers.

## 4.2 Performance Comparison

*4.2.1 Overall Performance.* Tab. 2 and Tab. 3 report the overall performance of three public datasets and an industrial dataset. We can make the following observations from four aspects: (1) In line with intuition, some fusion methods perform better than those use only IDs, revealing that side information can improve model's performance by capturing better sequence patterns. And this emphasizes the importance of side information fusion works. (2) On the contrary, under the vanilla self-attention framework, SASRec$_F$ considers more kinds of side information but brings a significant decrease compared with SASRec on all datasets, indicating that the information invasion does exist with self-attention-based naive early-fusion methods. (3) NOVA and DIF-SR are carefully designed to alleviate the invasion phenomenon, thus achieving better results than SASRec. At the same time, we note that, due to its lack of interaction caused by separating ID and feature into two channels, the effect of FDSA is not significantly better than that of NOVA and DIF-SR. (4) It is clear to see that ASIF achieves significantly better results than other SOTA baseline methods on all datasets. These results demonstrate the efficiency and validity of ASIF for eliminating noisy interference and solving the information invasion problem in side information fusion.

*4.2.2 Ablation Study.* We analyze the effectiveness of each component of ASIF via an ablation study. Tab. 4 shows the performance of ASIF and its ablation versions on three public datasets.

- **w/o Representation Space Alignment (RSA)**. We disable the contrastive loss to verify the effectiveness of RSA. The significant decline implies that bringing the two spaces closer appropriately can help alleviate the invasion phenomenon and thus improve performance.
- **w/o Homogeneous Information Extraction (HIE)**. Without the HIE component, attributes and position information can only participate in the calculation of attention scores, instead of directly being integrated into the hidden state of item representation. In this case, metrics drop on all datasets.
- **w/o Untied Positions (UP)**. This version removes the independent position channel and treats position as a common attribute. It can be observed that the interactions between the position encoding and other terms increase the noise and lead to a decrease in performance.
- **w/o Fused Attention (FA)**. We decouple the correlation calculation of IDs and attributes, i.e., each learns its own correlation matrix. The results show a decrease in most metrics. It means that it is necessary to retain the intersectionality between IDs and attributes.

All four ablated versions of ASIF are significantly better than SASRec$_F$ in Tab. 4. RSA and HIE are the most effective components in ASIF, proving there is indeed valid information in side information that should be carefully incorporated into item representation.

**Table 4: Ablation results (HR@20 and NDCG@20) on three public datasets. Each row removes a single component from the model except the last row.**

| Model | Yelp | | AliEC | | Beauty | |
|---|---|---|---|---|---|---|
| | H@20 | N@20 | H@20 | N@20 | H@20 | N@20 |
| w/o RSA | 0.1075 | 0.0524 | 0.1558 | 0.0668 | 0.1292 | 0.0540 |
| w/o HIE | 0.0996 | 0.0493 | 0.1439 | 0.0603 | 0.1255 | 0.0543 |
| w/o UP | 0.1077 | 0.0522 | 0.1572 | 0.0673 | 0.1298 | 0.0550 |
| w/o FA | 0.1108 | 0.0534 | 0.1601 | 0.0683 | 0.1317 | 0.0544 |
| ASIF | **0.1131** | **0.0543** | **0.1631** | **0.0700** | **0.1322** | **0.0554** |

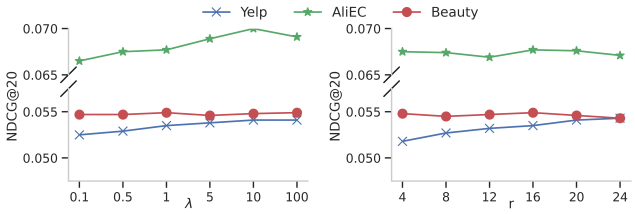



**Figure 6: Influence of balance parameter $\lambda$ and number of orthogonal bases $r$.**

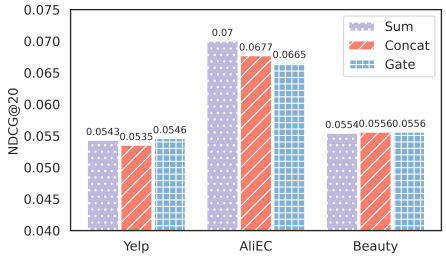



**Figure 7: Impact of fusion func $\mathcal{F}$.**

*4.2.3 Hyper-parameters Study.* **Influence of loss balance parameter $\lambda$.** We investigate the influence of hyper-parameter $\lambda$ which controls the balance of prediction loss and the contrastive loss. NDCG@20 is reported with $\lambda \in \{0.1, 0.5, 1, 5, 10, 100\}$ in Fig. 6 on three datasets. For Beauty dataset, the performance is robust with little variance across different $\lambda$, while on both Yelp and AliEC datasets, our ASIF achieves the best performance when $\lambda = 10$.

**Influence of number of orthogonal bases $r$.** ASIF's performance with a varying number of bases $r \in \{4, 8, 12, 16, 20, 24\}$ on three public datasets is reported in Fig. 6, respectively. A bigger number of bases usually means a finer granularity of decomposition. However, finer granularity does not always mean better. As we can see, the optimal number of orthogonal bases for three public datasets is around 16 to 24.

**Impact of fusion function $\mathcal{F}$.** We compare the performance of three different fusion functions: Sum, Concat, and Gate. Fig. 7 illustrates the results, showing that ASIF with all three fusion functions outperforms the state-of-the-art baselines mentioned in the paper. This highlights the robustness and superiority of ASIF.

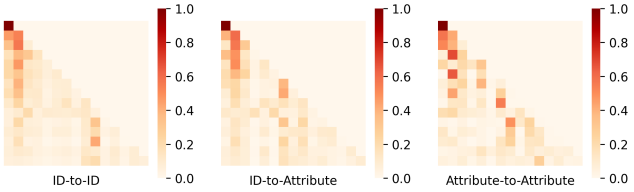
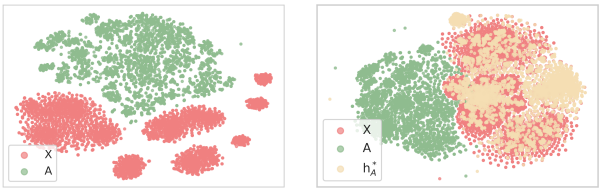



**Figure 8: Visualization of attention correlations in ASIF.**




(a) DIF-SR (b) ASIF

**Figure 9: Visualization of clustered embeddings on Yelp.**

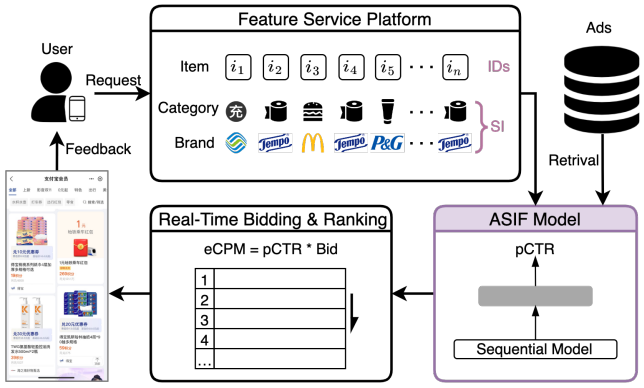



**Figure 10: Online deployment of ASIF in Alipay.**

*4.2.4 Case Study.* To provide more discussions on ASIF's interpretability, we visualize the correlations of ASIF using the same sample as SASRec$_F$ in Fig. 1. As shown in Fig. 8, the ID-to-ID, ID-to-Attribute, and Attribute-to-Attribute correlations show a strong pattern, indicating that ASIF has a better ability to capture associations among data after excluding noisy interference of position encoding. Moreover, we visualize ASIF's clustered embeddings on the Yelp dataset in Fig. 9(b). Compared with DIF-SR's in Fig. 9(a)) and SASRec$_F$'s in Fig. 2(a), the representation spaces of IDs and side information are closer after the alignment, and then the homogeneous part is basically aligned with ID representation after Homogeneous Information Extraction.

## 5 ONLINE DEPLOYMENT

In the online advertising system, the Click-Through Rate (CTR) prediction task is an important part, responsible for predicting the probability of users clicking on candidate items. Xlight is a traffic platform in the online app Alipay which provides advertisement services for small program merchants and so on. To further verify the effectiveness of the proposed model ASIF, we deploy it into the advertising system shown in Fig. 10. In Alipay's membership scenario, most of the ads are real goods, which are sold to users in the form of points with money. In order for ASIF to fully utilize its superiority in side information fusion, we select the category and brand of goods as side information for the items recommended in the this scenario. For offline training, ASIF collects the recent click samples in the past 7 days as the training dataset. For online service, when a user visits the membership page, the system will initiate a request for the user's historical behavior from an online feature service platform, which are truncated to a length of 50. ASIF will estimate the pCTR for some of the ads retrieved from the ads pool. In Xlight's Real-Time Bidding and Ranking system, each advertisement will be ranked based on its Effective Cost Per Mille (eCPM), which is estimated based on the pCTR and bid. Therefore, accurate estimation of CTR is pivotal for the Xlight platform.

**Table 5: Online performance on membership scene in Alipay.**

| exp_name | #clk | CPM | AUC | Time Cost (p99) |
|---|---|---|---|---|
| ASIF | +1.09% | +1.86% | +0.97% | +2ms |

Due to industrial constraints, it was not feasible to compare all baseline models in the online system. Therefore, we selected SASRec as the baseline model for comparison. After conducting two-week online A/B test, our model improved clicks by 1.09% and delivered a significant 1.86% increase in Cost Per Mille (CPM). Meanwhile, it enhanced multi-day online AUC by 0.97% with additional negligible computational cost (p99 latency 2ms). In conclusion, combined with offline evaluation, ASIF demonstrates strong performance in real-world industrial scenarios.

## 6 CONCLUSION

In this paper, we present a novel method ASIF for side information fusion in Sequential Recommendation. Our method addresses the challenges of noisy interference and information invasion in the mixed embedding space. Specifically, we first introduce Fused Attention with Untied Positions, which calculates position correlations individually to avoid noisy interference in the mixed attention scores. Secondly, we propose Representation Alignment, consisting of RSA and HIE, to solve the information invasion problem. RSA aligns the embedding spaces of IDs and attributes using the contrastive objective to improve their semantic consistency at the interaction level. HIE employs orthogonal decomposition to extract the homogeneous part in attributes and then integrate it into item representation, further enhancing the utilization of side information. Through extensive experiments, we have demonstrated that our proposed method surpasses previous approaches in side information fusion, and the visualization and ablation experiments demonstrate its rationality. The online A/B test on Alipay's advertising system showed that ASIF obtains a 1.09% improvement on clicks and 1.86% on CPM. In future research, we aim to further improve the denoising techniques and explore automatic methods to enhance the utilization for side information fusion.

# REFERENCES


[1] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems* 16 (2010), 345–379.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[3] Jianxin Chang, Chen Gao, Yu Zheng, Yiqun Hui, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. 2021. Sequential recommendation with graph neural networks. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 378–387.

[4] Alexander Dallmann, Daniel Zoller, and Andreas Hotho. 2021. A Case Study on Sampling Strategies for Evaluating Neural Sequential Item Recommendation Models. In *Fifteenth ACM Conference on Recommender Systems*. 505–514.

[5] Xinyan Fan, Zheng Liu, Jianxun Lian, Wayne Xin Zhao, Xing Xie, and Ji-Rong Wen. 2021. Lighter and better: low-rank decomposed self-attention networks for next-item recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 1733–1737.

[6] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).

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

[8] Guolin Ke, Di He, and Tie-Yan Liu. 2020. Rethinking Positional Encoding in Language Pre-training. In *International Conference on Learning Representations*.

[9] Walid Krichene and Steffen Rendle. 2020. On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 1748–1757.

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

[11] Chang Liu, Xiaoguang Li, Guohao Cai, Zhenhua Dong, Hong Zhu, and Lifeng Shang. 2021. Noninvasive self-attention for side information fusion in sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 4249–4256.

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

[13] Massimo Quadrana, Alexandros Karatzoglou, Balázs Hidasi, and Paolo Cremonesi. 2017. Personalizing session-based recommendations with hierarchical recurrent neural networks. In *proceedings of the Eleventh ACM Conference on Recommender Systems*. 130–137.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

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

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

[17] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive learning for sequential recommendation. In *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE, 1259–1273.

[18] Yueqi Xie, Peilin Zhou, and Sunghun Kim. 2022. Decoupled side information fusion for sequential recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1611–1621.

[19] 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*. 582–590.

[20] Xu Yuan, Dongsheng Duan, Lingling Tong, Lei Shi, and Cheng Zhang. 2021. ICAI-SR: Item Categorical Attribute Integrated Sequential Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1687–1691.

[21] Shuai Zhang, Yi Tay, Lina Yao, Aixin Sun, and Jake An. 2019. Next item recommendation with self-attentive metric learning. In *Thirty-Third AAAI Conference on Artificial Intelligence*, Vol. 9.

[22] 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 *IJCAI*. 4320–4326.

[23] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, et al. 2021. Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 4653–4664.

[24] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1893–1902.

[25] Kun Zhou, Hui Yu, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Filter-enhanced MLP is all you need for sequential recommendation. In *Proceedings of the ACM Web Conference 2022*. 2388–2399.