A APPENDIX

A.1 Summary of Notations

We summarize the important notations used in Section 3 in the following table:

Table 6: Important Notations Used in Section 3.

f	predictor	σ	sigmoid	$\mathbf{x}$	feature vector
$\mathcal{D}$	dateset	$\mathbb{R}$	real number set	$\hat{y}$	predicted CTR
$\mathcal{L}$	loss	d_0	feature dimension	y	ground truth label
S	historical behaviors	$\mathbf{K}$	cluster embeddings of C		
C	clustered behaviors	P	playing completion ratio of S		
s_j	j-th behavior in S	p_j	playing completion ratio of s_j		
c_i	i-th cluster in C	M	group number		
T	length of S	L_m	m-th item group		
$\hat{T}$	length of C	γ	maximum cluster size		
$\mathbf{K}_s$	item embeddings of S	δ	adaptive cluster number		
$\mathbf{x}_v$	feature vector of item v	$\mathbf{c}_i$	feature vector of cluster c_i		
$\mathbf{x}_{1:N_1}^{(v)}$	categorical part of $\mathbf{x}_v$	$\mathbf{c}_{1:N_1}^{(i)}$	categorical part of $\mathbf{c}_i$		
$\mathbf{x}_{1:N_2}^{(v)}$	numerical part of $\mathbf{x}_v$	$\mathbf{c}_{1:N_2}^{(i)}$	numerical part of $\mathbf{c}_i$		
$\mathbf{k}_i$	embedding vector of c_i	d	embedding dimension		
$\mathbf{q}$	target item's inherent embedding	$\boldsymbol{\beta}$	learnable cross feature weight		
$\mathbf{K}_h$	inherent feature part of $\mathbf{K}$	$\mathbf{K}_c$	cross feature part of $\mathbf{K}$		
H	inherent feature dimension	C	cross feature dimension		
$\boldsymbol{\alpha}$	attention weight	$\boldsymbol{\alpha}'$	adjusted attention weight		
$\mathbf{n}$	cluster size of C	$\mathbb{N}$	natural number set		
$\mathbf{W}^q, \mathbf{W}^h, \mathbf{W}^c, \mathbf{W}^v, \mathbf{W}^o$	linear projection parameters				

A.2 Behavior Feature Splits and Linear Projection

Following TWIN [2], we define the feature representations of a length $\hat{T}$ clustered behavior sequence $[c_1, c_2, ..., c_{\hat{T}}]$ as matrix $\mathbf{K}$, where each row denotes the features of one behavior. In practice, the linear projection of $\mathbf{K}$ in the attention score computation of MHTA is the key computational bottleneck that hinders the application of multi-head target attention (MHTA) on ultra-long user behavior sequences. We thus propose the following to reduce its complexity.

We first split the behavior features matrix $\mathbf{K}$ into two parts,

$$\mathbf{K} = [\mathbf{K}_h, \mathbf{K}_c] \in \mathbb{R}^{\hat{T}\times(H+C)}, \tag{12}$$

We define $\mathbf{K}_h \in \mathbb{R}^{\hat{T}\times H}$ as the *inherent* features of behavior items (e.g. video id, author, topic, duration) which are independent of the specific user/behavior sequence, and $\mathbf{K}_c \in \mathbb{R}^{\hat{T}\times C}$ as the user-item cross features (e.g. user click timestamp, user play time, clicked page position, user-video interactions). This split allows highly efficient computation of the following linear projection $\mathbf{K}_h\mathbf{W}^h$ and $\mathbf{K}_c\mathbf{W}^c$.

For the inherent features $\mathbf{K}_h$, although the dimension $\mathbf{H}$ is large (64 for each id feature), the linear projection is not costly. The inherent features of a specific item are shared across users/behavior sequences. With essential caching strategies, $\mathbf{K}_h\mathbf{W}^h$ could be efficiently "calculated" by a look-up and gathering procedure.

For the user-item cross features $\mathbf{K}_c$, caching strategies are not applicable because: 1). Cross features describe the interaction details between a user and a video, thus not shared across users' behavior sequences. 2). Each user watches a video at most once. Namely, there is no duplicated computation in projecting cross features. We thus reduce the computational cost by simplifying the linear projection weight.

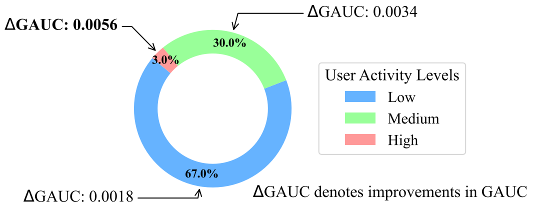

Figure 5: GAUC improvement among users of three activity levels with different proportions. The improvement in GAUC is calculated by subtracting the GAUC of TWIN from the GAUC of TWIN-V2. TWIN-V2 demonstrates more improvements for users with a longer history.

Given J cross features, each with embedding dimension 8 (since not id features with huge vocabulary size). We have $C = 8J$. We simplify the linear projection as follows,

$$\mathbf{K}_c\mathbf{W}^c \triangleq \Big[\ \mathbf{K}_{c,1}\mathbf{w}_1^c, \quad \dots \quad , \mathbf{K}_{c,J}\mathbf{w}_J^c\ \Big], \tag{13}$$

where $\mathbf{K}_{c,j} \in \mathbb{R}^{\hat{T}\times 8}$ is a column-wise slice of $\mathbf{K}_c$ for the j-th cross feature, and $\mathbf{w}_j^c \in \mathbb{R}^8$ is its linear projection weight. Using this simplified projection, we compress each cross feature into one dimension, i.e., $\mathbf{K}_c\mathbf{W}^c \in \mathbb{R}^{\hat{T}\times J}$. Note that this simplified projection is equivalent to restricting $\mathbf{W}^c$ to a diagonal block matrix.

A.3 Analysis of User Activity Levels

We postulate that enhancing recommendation model performance can be achieved by extending the length of user history input. Given that users with different activity levels exhibit varied lengths of historical behavior, the effect of extending long-term interest modeling to the life-cycle level is likely to differ among them. Consequently, we grouped users by different history lengths and reported the performance improvement across these groups.

We categorized users in the dataset into three groups based on the number of their historical behaviors: Low, Medium, and High. Additionally, we calculated the GAUC for TWIN and TWIN-V2 models across these groups. The improvements in GAUC for TWIN-V2 in different groups and their respective proportions of the total user count are shown in Figure 5. It is observable that TWIN-V2 achieves performance improvements across all user groups, validating the effectiveness of our approach. It is also evident that the absolute increase in GAUC is greater in user groups with a higher number of historical behaviors. This occurs as users with a greater number of historical actions possess a broader spectrum of interests within their life-cycle behaviors, thereby presenting more significant opportunities for enhancement. Additionally, TWIN-V2 also achieves performance improvements in groups of users with shorter histories. This is due to the incorporation of clustered behavior features in the cluster-aware target attention mechanism. These features represent the aggregated characteristics of all items within the cluster. Consequently, the data fed into GSU and ESU reflects a more comprehensive scope of behaviors compared to TWIN, leading to improved performance.