Int. J. Advance Soft Compu. Appl, Vol. 18, No. 2, July 2026
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Self-Supervised Temporal Transformer Learning for Behavioral Segmentation in Digital Banking

Bassam Mohammad Maali[1], Faisal Aburub[2] and Mohyi Aldin Abu Al Houl[3]

[1]Department of International Accounting, German Jordanian University, Amman, Jordan
e-mail:Bassam.Maali@gju.edu.jo
[2]Business Intelligence & Data Analytics, University of Petra, Amman, Jordan
e-mail:faburub@uop.edu.jo
[3] Faculty of Business, Ajloun National University, Ajloun, Jordan
e-mail:mohialden@anu.edu.jo

Abstract

Digital banking systems continuously record detailed traces of how customers interact with financial services, offering rich opportunities for behavioral analysis. However, using these data for customer segmentation remains difficult due to the heterogeneity of banking events, strong temporal dependencies, and the lack of labeled segment definitions. This paper introduces a sequence-based segmentation approach that models customer behavior as a stream of multi-modal banking events and learns representations directly from raw interaction logs. The proposed framework relies on a temporal Transformer encoder trained with self-supervised objectives, allowing categorical attributes, numerical values, and inter-event time information to be jointly encoded. Masked event modeling and next-event prediction are used to guide representation learning, while a deep embedded clustering objective integrates segmentation directly into the training process. Experiments on a large-scale real-world digital banking dataset show that the proposed method produces more compact and better separated behavioral segments than traditional aggregation-based techniques and neural baselines. In addition, the resulting segments remain stable across monthly evaluation windows and reflect persistent behavioral characteristics related to activity regularity, spending variability, and engagement patterns. In the future, we also plan to extend this work in several directions, such as incorporating the segmentation objectives with downstream outcomes, performing online segmentation, and conducting a deeper causal analysis of segment jumps. Plus, it would be interesting to inject domain knowledge or expert intervention into the segmentation process.

Keywords: behavioral segmentation, digital banking, self-supervised learning, Transformer models, deep clustering, temporal modeling

1 Introduction

Nowadays, digital banking channels produce continuous customer behavioral data via transactions, service requests, and other channel activities [1]. Contrary to conventional

Received 16 February 2026; Accepted 21 April 2026

banking data, these are a collection of event sequences capturing customers' interactions with services over time [2–4]. These records are therefore able to capture behavioral trajectories, as opposed to isolated profiles of accounts, ushering new horizons for customer analytics [5].

Despite its importance, behavioral segmentation is useful to further digital banking applications such as personalized service delivery, risk monitoring, product recommendation, and customer re- lationship management [6,7]. Segmentation helps in distinguishing customers who might be seem similar in terms of overall statistics but differ in temporal regularity, transaction dispersion, and engagement dynamics [8, 9]. But, it remains challenging to extract stable and interpretable behavior segments directly from event data, owing to the high-dimensional, heterogeneously featured, and strongly correlated temporal data [10].

Current segmentation techniques used in banks mostly consider the recency, frequency, and monetary value of aggregate data or other demographic features that have a relatively weak association with digital behaviors [11]. Although they are easy to use, they miss any sequential pattern and are affected by short-term behavioral changes. Thus, the derived segments are not stable over time and miss repeated patterns such as salary-related activity, periodical bill payment, or occasional usage [12].

A recent work utilizes representation learning methods such as autoencoders and se- quence models to learn customer embeddings directly from interaction data [13]. Sequence-aware models have been applied with success in capturing the evolution of user behavior in transaction or interaction logs [14]. But, in many cases, learning representations and clustering are performed separately, where embeddings are learned with unrelated objectives and clustering is performed as post-processing. This results in a mismatch between learned representations and the segmentation task, yielding suboptimal cluster structure and unstable assignments [15].

The key challenge addressed in this work is to learn customer representations from multi-modal banking event sequences that can directly help temporally coherent and behaviorally meaningful segmentation without manual labels or predefined segment definitions.

The following research questions guide this study:

- Item How can we jointly embed heterogeneous banking events, containing categorical, numeral, and temporal information, into a unified behavioral embedding space?
- Item Can self-supervised sequence modeling be leveraged to learn customer representations that capture the inherent behavioral traits rather than variations in recent activities?
- Item Does learning representations in conjunction with clustering yield more stable and interpretable behavioral segments than applying clustering after the fact?

To answer these questions, we propose a multi-modal temporal segmentation model based on a Transformer encoder with deep embedded clustering. We treat customer behavior as a sequence of events, and each event is represented by a combination of categorical features, numerical features, and inter-event times. A temporal Transformer is utilized to learn the sequences of events to capture the long-term dependencies and repeated patterns of customer behavior.

We train the model using masked event modeling and next-event prediction in a self-supervised way, which makes it possible to fully use unlabeled interaction logs to improve the model performance. Finally, we use a clustering objective in the training process to directly improve the learned representation towards the segmentation task. Jointly optimizing representation learning and clustering encourages behavioral segments to be compact, separated, and temporally smooth.

The experimental evaluation is conducted on a large-scale real-world dataset of digital bank customers, focusing on the quality, stability, and interpretability of the clusters.

The key contributions of the thesis are as follows:

- We propose a unified representation learning framework for behavioral segmentation in digital banking by jointly modeling categorical, numerical, and temporal event information.
- We use deep clustering as an integral component of our self-supervised sequence model, rather than applying clustering as a post hoc step to learned embeddings.
- We conduct an extensive empirical study that reveals better separated and more stable clusters over time compared to classical and neural baselines.
- We show that the segments produced can be mapped onto distinct and interpretable behavioural profiles that reflect actual usage patterns.

2 Related work

Behavioral segmentation has also been used extensively by customer relationship management. During the short-term customer relationship management, customers are commonly identified and segmented by aggregated features such as recency, frequency, and monetary values (RFM). RFM-based behavioral segmentation methods are still widely used in practice due to their simplicity and intuitive meanings. Several studies have used clustering algorithms and rule extraction methods to improve the usefulness and ease of understanding of RFM-based segmentation. For instance, Cheng and Chen proposed a combination of RFM features, clustering methods, and rough set theories to analyze customer segments and extract the rules describing each segment. They demonstrated that even simple statistics can be useful in understanding customer behaviors when combined with rule extraction methods [16]. More recent studies have used time-dependent methods to address the issues of losing temporal information caused by aggregation. Song et al. proposed a statistic-based approach to segment customers' RFM values over time on large-scale datasets. They directly considered temporal information in quantitative customer value measurement and segmentation [17]. These studies highlight the importance of temporal information, but their feature design heavily relies on predefined aggregates of interactions.

Another line of work tries to learn customer representations from high-dimensional raw behavioral records. In the banking domain, Mancisidor et al. proposed learning latent representations of bank customers with a variational autoencoder and demonstrated that the latent space has meaningful clustering structures consistent with clusterings relevant to creditworthiness [18]. This representation learning perspective circumvents the need for heuristic aggregation steps. That said, in most processing pipelines, clustering is performed on the learned embeddings, which may lead to incompatible learning of representations with segmen- tation in mind.

Sequence modeling with attention mechanisms provides a natural framework for representing longitudinal behavior. The Transformer architecture introduced by Vaswani et al. replaces recurrence with self-attention and enables efficient learning of long-range dependencies [19]. In recommendation and user modeling, self-attention has been adapted to encode behavioral sequences; SASRec demonstrated that self-attentive sequence encoders can capture both short- and long-term dependencies in user action histories [20]. Bidirectional masked modeling has also been used to learn sequence representations: BERT4Rec adopted a Cloze-style objective to predict masked items in a user sequence, improving representation quality by leveraging both left and right context [21]. Building on these ideas, S^3-Rec proposed self-supervised pretraining objectives for sequential recommendation based on mutual information maximization, showing that auxiliary self-supervision can improve sequence representations under sparse supervision [22]. Although these works are typically evaluated for prediction and ranking, their modeling choices (masked modeling, next-step prediction, and attention pooling) are directly relevant to unsupervised behavioral representation learning.

Finally, deep clustering methods integrate representation learning and clustering into a single objective, reducing reliance on post hoc clustering of fixed embeddings. Xie et al. proposed Deep Embedded Clustering (DEC), which jointly optimizes an embedding space and cluster assignments via a KL-divergence-based objective and a sharpened target distribution [23]. In the context of behavioral segmentation, the DEC principle is attractive because it encourages embeddings to become cluster-friendly while training, rather than treating clustering as an external step [24].

The present work connects these threads by learning multi-modal event embeddings from digital banking logs using a temporal Transformer trained with self-supervised objectives inspired by masked modeling and next-event prediction, while integrating a DEC-style clustering loss to align representation learning with segmentation. [25].

3 Methodology

3.1 Dataset and preprocessing

Experiments are conducted on the Multimodal Banking Dataset (MBD), a large-scale anonymized dataset released for research on digital banking behavior [26]. The dataset consists of longitudinal event logs collected from real-world digital banking systems, comprising millions of interaction records associated with individual customers. Each record corresponds to a discrete banking event described by a timestamp, categorical attributes, and numerical values.

For each customer u, the dataset provides a time-ordered sequence of events

$$S_u = \{e_{u,1}, e_{u,2}, \ldots, e_{u,T_u}\},$$

where each event is represented as

$$e_{u,t} = \left(x_{u,t}^{\text{cat}},\ x_{u,t}^{\text{num}},\ \tau_{u,t}\right).$$

Here, $x_{u,t}^{\text{cat}}$ denotes categorical descriptors such as event type, category, channel, and merchant or service identifier (when available), while $x_{u,t}^{\text{num}}$ includes numerical attributes such as transaction amount and balance-related indicators. The timestamp $\tau_{u,t}$ enables the computation of inter-event time gaps $\Delta t_t = \tau_t - \tau_{t-1}$.

Customers with extremely short activity histories are excluded to ensure stable representation learning. Numerical attributes are log-transformed when appropriate and standardized, while categorical variables are encoded via learnable embeddings.

Table 1 illustrates a simplified sample of the event-level data structure.

Table 1: Illustrative sample from the Multimodal Banking Dataset

Customer	Time	Event type	Category	Channel	Amount	Δt
u_1	t_1	Payment	Utilities	Mobile	75.3	–
u_1	t_2	Transfer	P2P	Mobile	120.0	2.1d
u_1	t_3	Card use	Retail	POS	34.8	0.3d
u_2	t_1	Salary	Income	Bank	1800.0	–

3.2 Model architecture

The proposed approach learns customer representations from raw event sequences using a multi-modal temporal Transformer combined with joint deep clustering. Figure 1 presents an overview of the architecture.

Each event is represented by a fixed-dimensional embedding that integrates categorical, numerical, and temporal information. At time step t, the event embedding is defined as

$$\mathbf{h}_t = \sum_{v \in \{\text{type,cat,chan,mch}\}} \mathbf{E}_v(x_t^v) + \mathbf{P}(x_t^{\text{num}}) + \mathbf{T}(\Delta t_t), \tag{1}$$

where $\mathbf{E}_v(\cdot)$ denote trainable embedding matrices associated with categorical attributes, $\mathbf{P}(\cdot)$ is a multi-layer perceptron that projects numerical features into the embedding space, and $\mathbf{T}(\cdot)$ encodes discretized logarithmic inter-event time gaps. This formulation allows heterogeneous event attributes to be mapped into a unified latent representation while preserving temporal information.

The resulting sequence of event embeddings is processed by a temporal Transformer encoder composed of stacked self-attention layers, which capture contextual dependencies across the entire customer history. The encoder produces a sequence of contextualized representations

$$\mathbf{H}_u = \{\mathbf{z}_{u,1}, \ldots, \mathbf{z}_{u,T_u}\}. \tag{2}$$

To obtain a fixed-length customer representation, an attention-based pooling mechanism is applied over the contextualized event embeddings:

$$\mathbf{z}_u = \sum_{t=1}^{T_u} \alpha_t \mathbf{z}_{u,t}, \qquad \alpha_t = \frac{\exp(\mathbf{w}^\top \tanh(\mathbf{W}\mathbf{z}_{u,t}))}{\sum_j \exp(\mathbf{w}^\top \tanh(\mathbf{W}\mathbf{z}_{u,j}))}, \tag{3}$$

where the attention weights α_t reflect the relative contribution of each event to the customer's behavioral representation.

3.3 Training objective and clustering

Model training is performed in a self-supervised manner and does not rely on predefined customer labels. Two complementary objectives are employed.

First, masked event modeling randomly masks components of selected events (event type, category, or discretized transaction amount) and trains the model to reconstruct them using a cross-entropy loss:

$$\mathcal{L}_{\text{mem}} = \sum_{t \in \mathcal{M}} \text{CE}(\hat{x}_t, x_t).$$

Second, next-event prediction encourages temporal coherence by predicting both the type of the next event and the time until its occurrence:

$$\mathcal{L}_{\text{next}} = \text{CE}(\hat{x}_{t+1}^{\text{type}}, x_{t+1}^{\text{type}}) + \text{Huber}(\log(\widehat{1+\Delta t}_{t+1}), \log(1+\Delta t_{t+1})).$$

The self-supervised objective is given by

$$\mathcal{L}_{\text{ssl}} = \lambda_1 \mathcal{L}_{\text{mem}} + \lambda_2 \mathcal{L}_{\text{next}}.$$

To obtain behaviorally coherent segments, clustering is integrated directly into representation learning. Let $\{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K\}$ denote trainable cluster centroids. Soft cluster assignments are computed as

$$q_{uk} = \frac{(1 + \|\mathbf{z}_u - \boldsymbol{\mu}_k\|^2)^{-1}}{\sum_j (1 + \|\mathbf{z}_u - \boldsymbol{\mu}_j\|^2)^{-1}}.$$

A sharpened target distribution p_{uk} is derived and used to define the clustering loss

$$\mathcal{L}_{\text{clust}} = \sum_u \sum_k p_{uk} \log \frac{p_{uk}}{q_{uk}}.$$

The final training objective is

$$\mathcal{L} = \mathcal{L}_{\text{ssl}} + \beta \mathcal{L}_{\text{clust}}.$$

3.4 Learning procedure

Algorithm 1 summarizes the training procedure.

Algorithm 1 Training procedure of the proposed model

Initialize Transformer parameters and cluster centroids $\{\boldsymbol{\mu}_k\}_{k=1}^K$
for each training epoch **do**
  for each mini-batch of customer sequences **do**
    Embed events and compute contextual representations via Transformer
    Apply masked event modeling and next-event prediction
    Compute customer embeddings $\mathbf{z}_u$
    Update soft cluster assignments and target distribution
    Minimize $\mathcal{L}_{\text{ssl}} + \beta \mathcal{L}_{\text{clust}}$
  end for
end for
Output customer embeddings and final segment assignments

3.5 Evaluation protocol

Segmentation quality is evaluated using internal clustering metrics, including the Silhouette score and Davies–Bouldin index, as well as segment size distributions. Temporal stability is assessed by tracking segment transitions across consecutive monthly windows.

Baseline comparisons include RFM-based clustering, PCA followed by K-means, autoencoder embeddings with K-means, and Transformer-based embeddings clustered without joint optimization.

4 Results

4.1 Clustering quality and separation

Clustering performance is first evaluated using standard internal validity metrics that assess intra-cluster compactness and inter-cluster separation. Table 2 reports the Silhouette score and Davies–Bouldin index for all compared methods.

Table 2: Clustering performance comparison

Method	Silhouette ↑	Davies–Bouldin ↓
RFM + K-means	0.21	1.84
PCA + K-means	0.26	1.52
Autoencoder + K-means	0.31	1.27
Transformer + K-means	0.36	1.05
Proposed model	**0.43**	**0.81**

The proposed approach achieves the highest Silhouette score and the lowest Davies–Bouldin index among all methods. This indicates that jointly learning representations and cluster assignments leads to clusters that are both more compact and better separated than those obtained by post hoc clustering of fixed embeddings.

To complement numerical metrics, Figure 2 visualizes customer embeddings projected into two dimensions using UMAP. While baseline methods exhibit substantial overlap between clusters, the proposed model produces well-defined regions with clearer boundaries.

4.2 Temporal stability of behavioral segments

Beyond static separation, effective behavioral segmentation in digital banking requires temporal consistency. Segment stability is therefore evaluated by tracking customer assignments across consecutive monthly windows and computing the proportion of customers remaining in the same segment.

Table 3 reports the average monthly retention rate for representative methods.

Table 3: Average monthly segment retention

Method	Retention rate
RFM + K-means	0.62
Transformer + K-means	0.71
Proposed model	**0.83**

The proposed model shows substantially higher retention, indicating that the learned segments correspond to persistent behavioral patterns rather than short-term fluctuations. Static aggregation-based methods are more sensitive to temporary changes in activity intensity, leading to frequent segment switching.

Table 4 further illustrates this effect through a monthly segment transition matrix. Strong diagonal dominance confirms that most customers remain within the same behavioral segment over time, while off-diagonal values reflect limited but interpretable transitions between segments.

Table 4: Monthly segment transition matrix for the proposed model.

	Ending segment				
Starting segment	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	**0.86**	0.02	0.03	0.01	0.08
Cluster 2	0.02	**0.81**	0.04	0.01	0.11
Cluster 3	0.02	0.03	**0.86**	0.06	0.03
Cluster 4	0.01	0.01	0.02	**0.82**	0.15
Cluster 5	0.05	0.09	0.04	0.08	**0.82**

The transition matrix in Table 4 shows strong diagonal dominance across all segments, with self-transition probabilities ranging from 0.81 to 0.86. This indicates that customer behavior is largely stable at the monthly scale, and that the identified segments capture persistent usage patterns rather than short-term fluctuations.

Off-diagonal transitions occur at comparatively low rates and are unevenly distributed across segments, suggesting structured behavioral shifts rather than random noise. In particular, transitions involving Clusters 4 and 5 exhibit slightly higher outward probabilities, which may reflect customers with more variable or evolving interaction patterns. Overall, the transition structure supports the temporal coherence of the proposed segmentation.

4.3 Behavioral interpretation of segments

Analysis of cluster-level statistics reveals distinct and interpretable behavioral profiles. The identified segments include customers with highly regular salary-driven activity, frequent low-value digital spenders, sporadic high-value transactors, and low-activity or dormant users.

Figure 3 summarizes normalized behavioral indicators for each segment using radar plots. Differences between segments are driven by combinations of transaction frequency, monetary dispersion, temporal regularity, and diversity of event types, rather than by any single aggregate feature.

Inspection of attention weights from the pooling mechanism indicates that segment assignment is influenced by recurring temporal structures, such as regular salary deposits or periodic bill payments, as well as by deviations from these patterns. This supports the use of attention-based aggregation for summarizing long behavioral histories.

4.4 Ablation study

To assess the contribution of individual components, an ablation analysis is conducted. Removing the time-gap encoding results in a decrease of approximately 0.06 in Silhouette score and a noticeable reduction in segment stability, particularly for customers with irregular activity cycles.

Similarly, replacing joint deep clustering with post hoc K-means on fixed embeddings leads to poorer separation and increased segment volatility. These results confirm that both temporal modeling and integrated clustering play a central role in the observed performance gains.

5 Conclusion

This research focused on sequence-based behavioral segmentation in digital banking to understand the value of historical customer interactions in extracting meaningful and stable segments. The proposed method does not employ aggregated features or predefined customer characteristics; rather, it treats banking behavior as heterogeneous event sequences and extracts meaningful representations from these sequences.

To this end, we proposed a multi-modal temporal transformer that encodes categorical, numerical, and temporal information of the event entry. Self-supervised objectives guide the representation learning, along with a deep embedded clustering method, so that customer embeddings are learned for the clustering purpose and not as an aftereffect. This is in contrast to the existing pipeline where the representation learning and clustering tasks are disjoint.

Experiments on a real-world large-scale bank dataset show that the proposed model achieves more compact and better separated segments than classical aggregation-based models and neural embedding-based baseline models. The resulting segments are consistent over monthly evaluation windows, indicating the model captures customers' enduring characteristics rather than short-term behavioral trends. We also examine the characteristics captured by each segment profile and show they relate to interesting aspects of bank behaviors such as the consistency of activity, volatility of spending, and engagement patterns.

Despite its enormous potential, the proposed methodology still suffers from some limitations that need to be addressed. In particular, the current version of the paper does not connect unsupervised segmentation to subsequent downstream outcomes related to churn, profitability, or credit risk. Also, we have tested the method on only one large-scale dataset. Experiments on other datasets from different banks would show the effectiveness and flexibility of the framework more thoroughly.

Future research could also extend the framework to Islamic banks. Due to Shariah principles such as profit-and-loss sharing and the prohibition of interest, customer behavior and risk patterns may differ in ways that affect the stability and interpretation of the identified segments [27] [28]. Examining this context would provide a useful test of the model's generalization.

In the future, we also plan to extend this work in several directions, such as incorporating the segmentation objectives with downstream outcomes, performing online segmentation, and conducting a deeper causal analysis of segment jumps. Plus, it would be interesting to inject domain knowledge or expert intervention into the segmentation process.

Availability of data and materials Data will be made available on request.
Competing interests The author declares that there are no conflicts of interest related

to this study.
Funding The author has not received funding for conducting this research.
Authors' contributions The author contributed to this manuscript and reviewed the paper prior to submission.
Disclosures and Declarations The author declares that this research received no external funding, has no financial or non-financial competing interests, did not involve human participants or animals, and therefore ethical approval and informed consent were not required

References

[1] M. Loukili, F. Messaoudi, and R. El Youbi, "Enhancing financial transaction security: A deep learning approach for e-payment fraud detection," in *Internet of Things and Big Data Analytics for a Green Environment*. Boca Raton, FL, USA: Chapman and Hall/CRC, 2024, pp. 238–252.

[2] P. C. Verhoef, P. K. Kannan, and J. J. Inman, "From multi-channel retailing to omni-channel retailing: Introduction to the special issue on multi-channel retailing," *Journal of Retailing*, vol. 91, no. 2, pp. 174–181, 2015.

[3] M.-H. Huang, R. T. Rust, and M. Z. Hu, "Managing the service profit chain in the era of digital marketing," *Journal of the Academy of Marketing Science*, vol. 45, no. 1, pp. 44–65, 2017.

[4] M. Loukili, F. Messaoudi, and M. E. Ghazi, "Defending against digital thievery: A machine learning approach to predict e-payment fraud," *Int. J. Management Practice*, vol. 17, no. 5, pp. 522–538, 2024.

[5] M. Loukili, F. Messaoudi, and H. Azirar, "E-payment fraud detection in e-commerce using supervised learning algorithms," in *Advances in Emerging Financial Technology and Digital Money*. Boca Raton, FL, USA: CRC Press, 2024, pp. 27–35.

[6] E. W. T. Ngai, L. Xiu, and D. C. K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2592–2602, 2009.

[7] M. Haenlein and A. Kaplan, "A brief history of artificial intelligence: On the past, present, and future of artificial intelligence," *California Management Review*, vol. 61, no. 4, pp. 5–14, 2019.

[8] R. El Youbi, F. Messaoudi, M. Loukili, and R. Loukili, "From click to checkout: Deep learning for real-time fraud detection in e-payment systems," *Statistics, Optimization & Information Computing*, vol. 14, no. 6, pp. 3398–3408, 2025.

[9] F. Messaoudi, M. Loukili, and R. El Youbi, "A text-based approach for product clustering and recommendation in e-commerce using machine learning," in *Internet of Things and Big Data Analytics for a Green Environment*. Boca Raton, FL, USA: Chapman and Hall/CRC, 2024, pp. 124–137.

[10] E. Ascarza, "Retention futility: Targeting high-risk customers might be ineffective," *Journal of Marketing Research*, vol. 55, no. 1, pp. 80–98, 2018.

[11] J. Z. Kolter and M. A. Maloof, "Dynamic weighted majority: An ensemble method for drifting concepts," *Journal of Machine Learning Research*, vol. 8, pp. 2755–2790, 2007.

[12] W. Buckinx and D. Van den Poel, "Customer base analysis: Partial defection of behaviorally loyal clients in a non-contractual FMCG retail setting," *European Journal of Operational Research*, vol. 164, no. 1, pp. 252–268, 2007.

[13] I. Bose and X. Chen, "Hybrid models using unsupervised learning for retail customer segmentation," *Decision Support Systems*, vol. 111, pp. 11–22, 2018.

[14] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," in *Proc. Int. Conf. Learning Representations (ICLR)*, San Juan, Puerto Rico, 2016.

[15] X. Guo, L. Gao, X. Liu, and J. Yin, "Improved deep embedded clustering with local structure preservation," in *Proc. Int. Joint Conf. Artificial Intelligence (IJCAI)*, Melbourne, Australia, 2017, pp. 1753–1759.

[16] C.-H. Cheng and Y.-S. Chen, "Classifying the segmentation of customer value via RFM model and RS theory," *Expert Systems with Applications*, vol. 36, no. 3, pp. 4176–4184, 2009.

[17] M. Song, X. Zhao, H. E, and Z. Ou, "Statistics-based CRM approach via time series segmenting RFM on large-scale data," *Knowledge-Based Systems*, vol. 132, pp. 21–29, 2017.

[18] R. A. Mancisidor, M. Kampffmeyer, K. Aas, and R. Jenssen, "Learning latent representations of bank customers with the variational autoencoder," *Expert Systems with Applications*, vol. 164, Art. no. 114020, 2021.

[19] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[20] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Singapore, 2018, pp. 197–206.

[21] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer," in *Proc. 28th ACM Int. Conf. Information and Knowledge Management (CIKM)*, Beijing, China, 2019, pp. 1441–1450.

[22] K. Zhou, H. Wang, W. X. Zhao, Y. Zhu, S. Wang, F. Zhang, Z. Wang, and J.-R. Wen, "S^3-Rec: Self-supervised learning for sequential recommendation with mutual information maximization," in *Proc. 29th ACM Int. Conf. Information and Knowledge Management (CIKM)*, Galway, Ireland, 2020, pp. 1893–1902.

[23] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. 33rd Int. Conf. Machine Learning (ICML)*, New York, NY, USA, 2016, pp. 478–487.

[24] A. R. Alshehadeh, M. Al-Zaqeba, A. Qtaishat, H. Al-khawaja, and E. Al-Wreikat, "Digitalization and Sustainable Development Goals: Enhancing Electronic Financial Reports Quality in Banking," Data and Metadata, vol. 4, pp. 734–734, 2025.

[25] A. R. Alshehadeh, M. A. A. Al-Zaqeba, G. A. Elrefae, H. A. Al-Khawaja, and N. M. Aljawarneh, "The effect of digital zakat and accounting on corporate sustainability through financial transparency," Asian Economic and Financial Review, vol. 14, no. 3, pp. 228–249, 2024.

[26] D. Dzhambo, A. Voronkov, and A. Filippova, "Multimodal Banking Dataset: A Large-Scale Dataset for Research on Digital Banking Behavior," 2024. [Online]. Available: https://huggingface.co/datasets/ai-lab/MBD

[27] Maali, B. M., Fendi, U. A., & Atmeh, M. A. (2021). The economic reality of Islamic banks' transactions: A qualitative inquiry. *International Journal of Islamic and Middle Eastern Finance and Management*, 14(2), 286–300.

[28] Maali, B. M., Hassan, M. K., & Rashid, M. (2024). Creating a religious identity and impression management by Islamic banks. *International Journal of Islamic and Middle Eastern Finance and Management*, 17(4), 693–710.

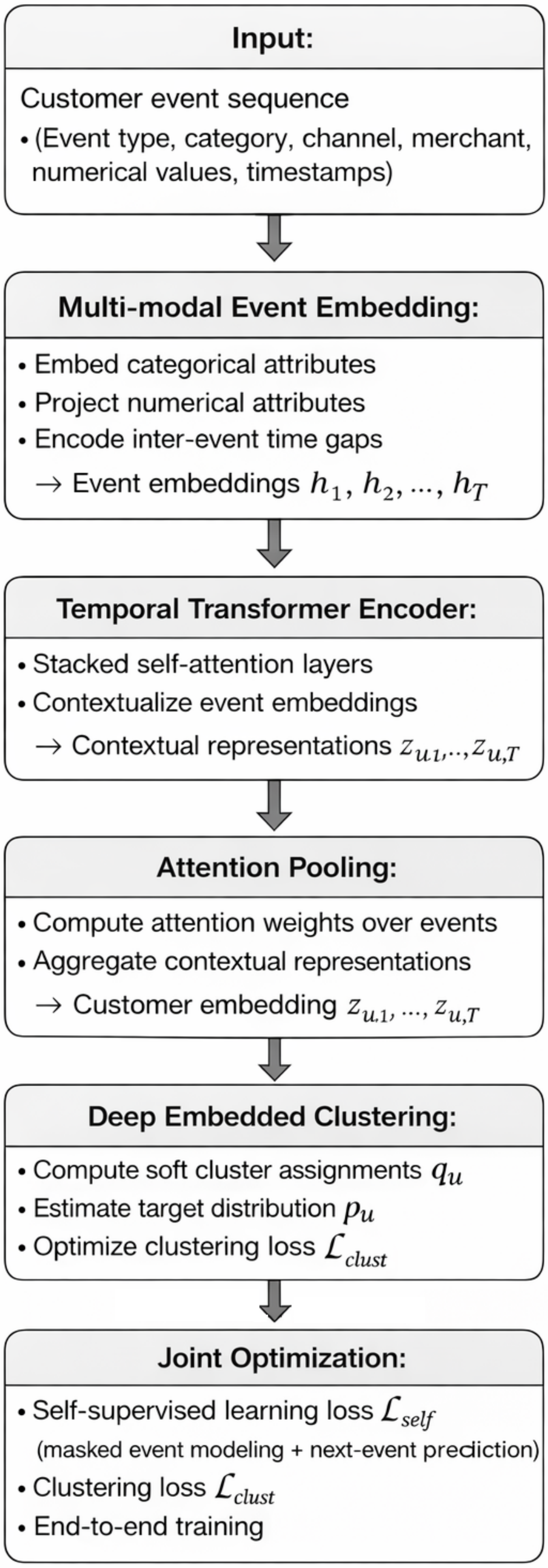

Figure 1: The proposed multi-modal temporal Transformer with deep embedded clustering architecture.

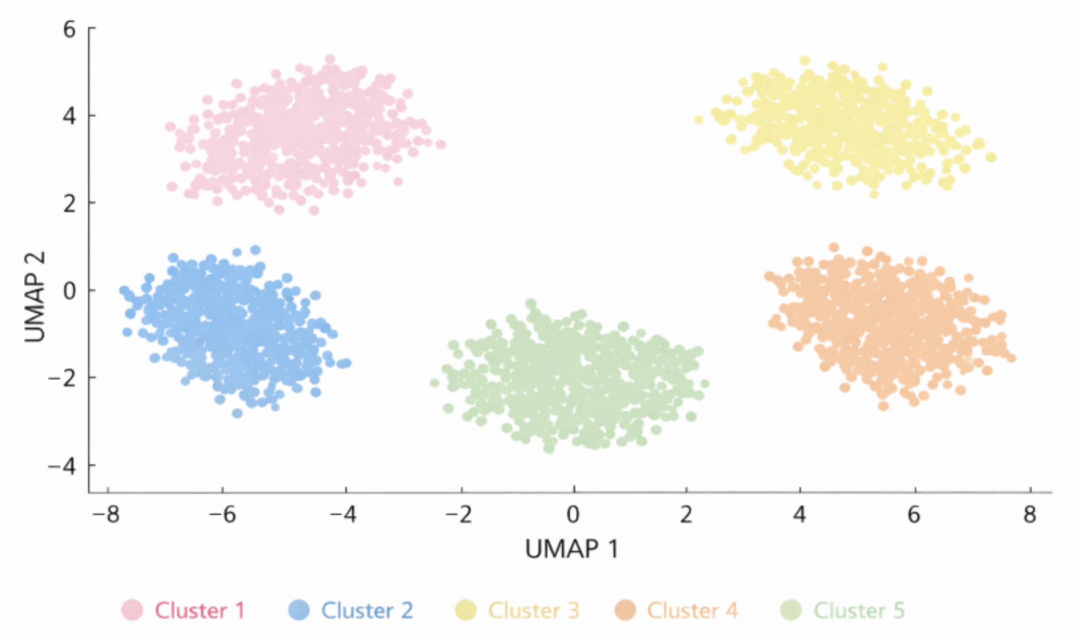

Figure 2: UMAP projection of customer embeddings.

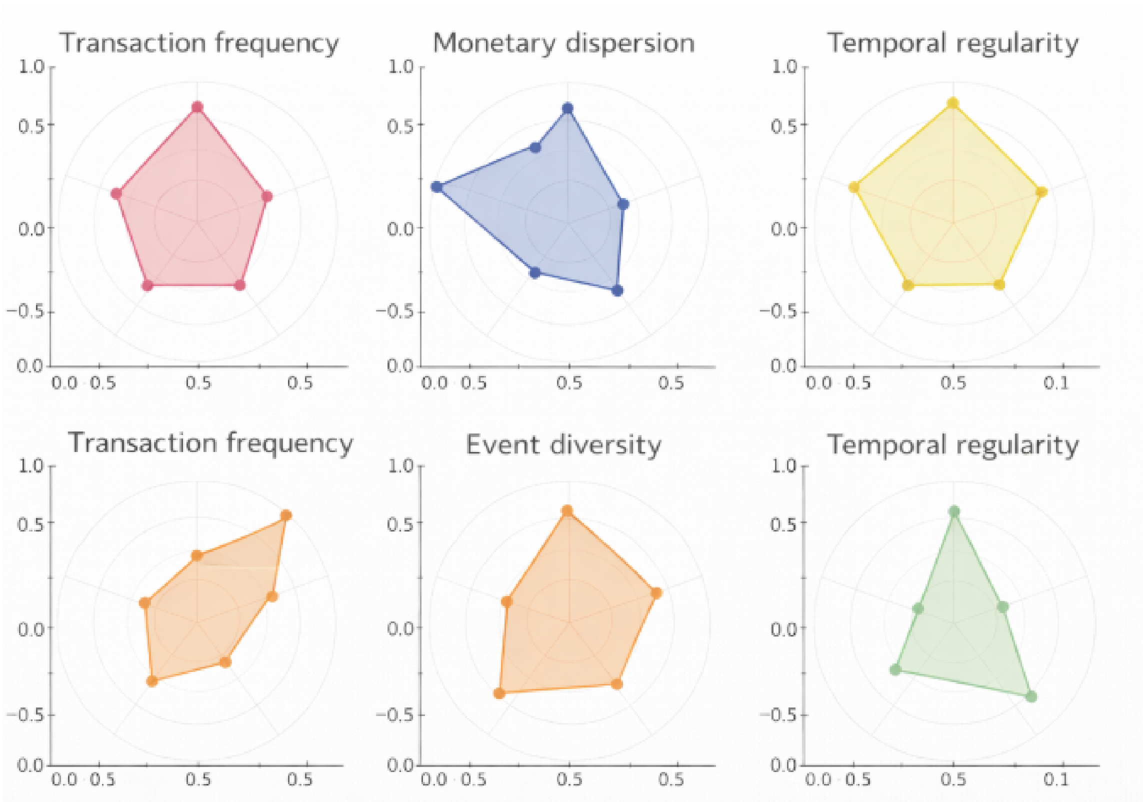

Figure 3: Radar plots of normalized behavioral indicators for the discovered segments.