Large-scale Generative and Multimodal Recommendation Systems: An Overview

Junwei Pan

Tencent 腾讯

CIKM 2025 SEOUL
November 10 - 14

Outline

Generative Recommendation

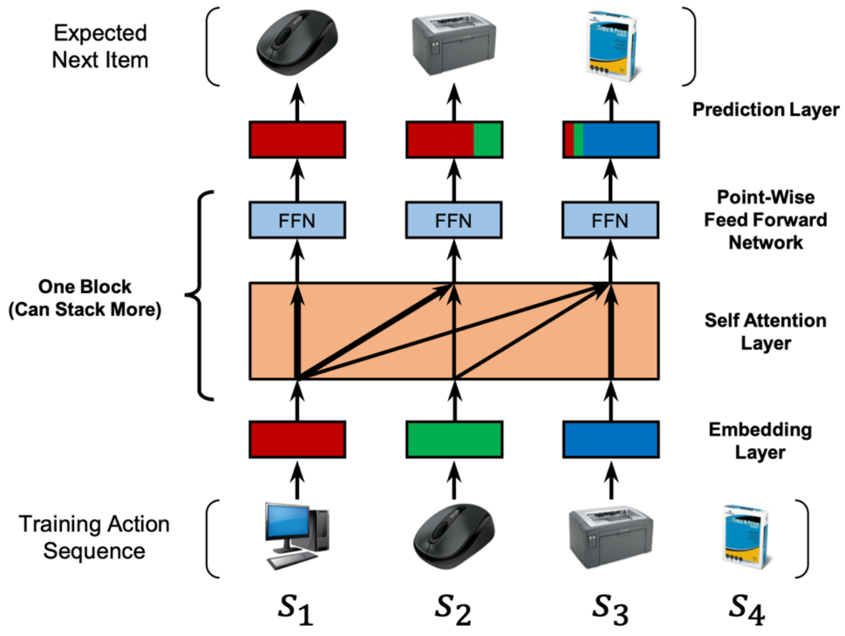
- Token Organization
- Action Handling

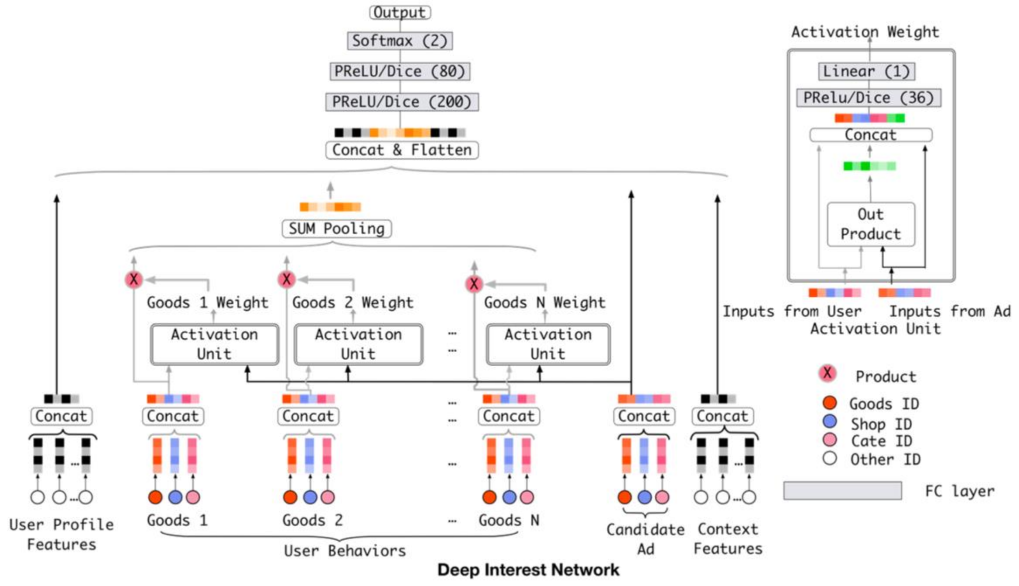
Multimodal Recommendation

- Alignment
- Distance Transfer
- Semantic IDs

Item-Oriented: SASRec, DIN

- Next-item prediction paradigm
- Fuse all side infos

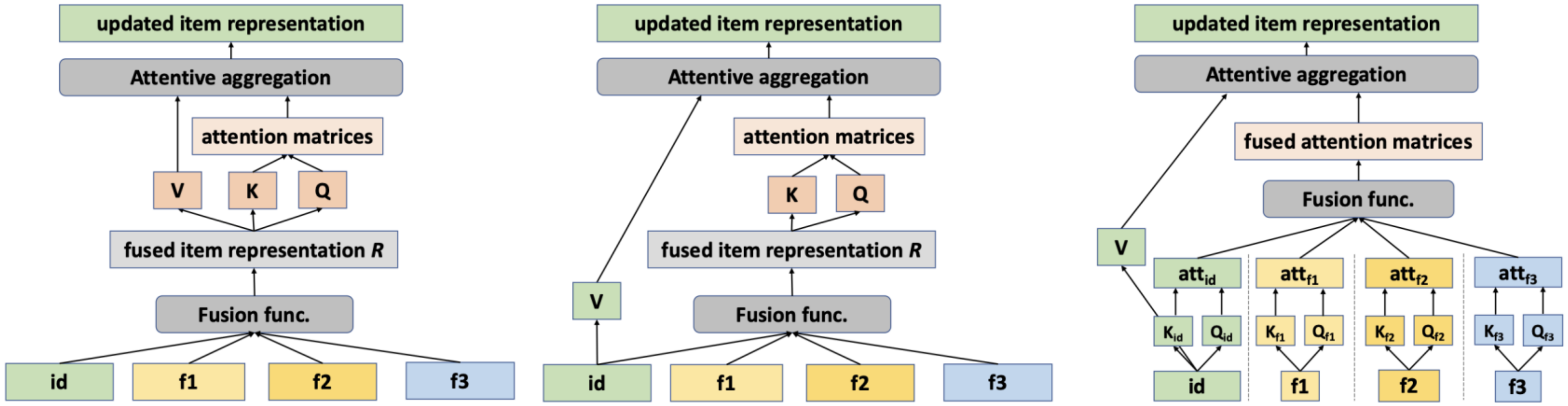




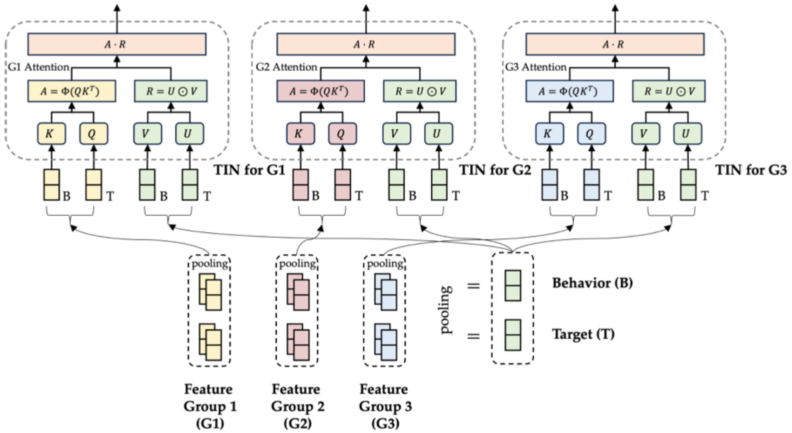
Actions Speak Louder than Words- Trillion-Parameter Sequential Transducers for Generative Recommendations. ICML 2024.

Side Infos

- SASRec fuse all side infos
- NOVA employs only the item ID in the **Value**
- DIF-SR further decouple each feature in the **Attention (Q, K)**
- DSI-TIN decouples feature groups in the **Attention**



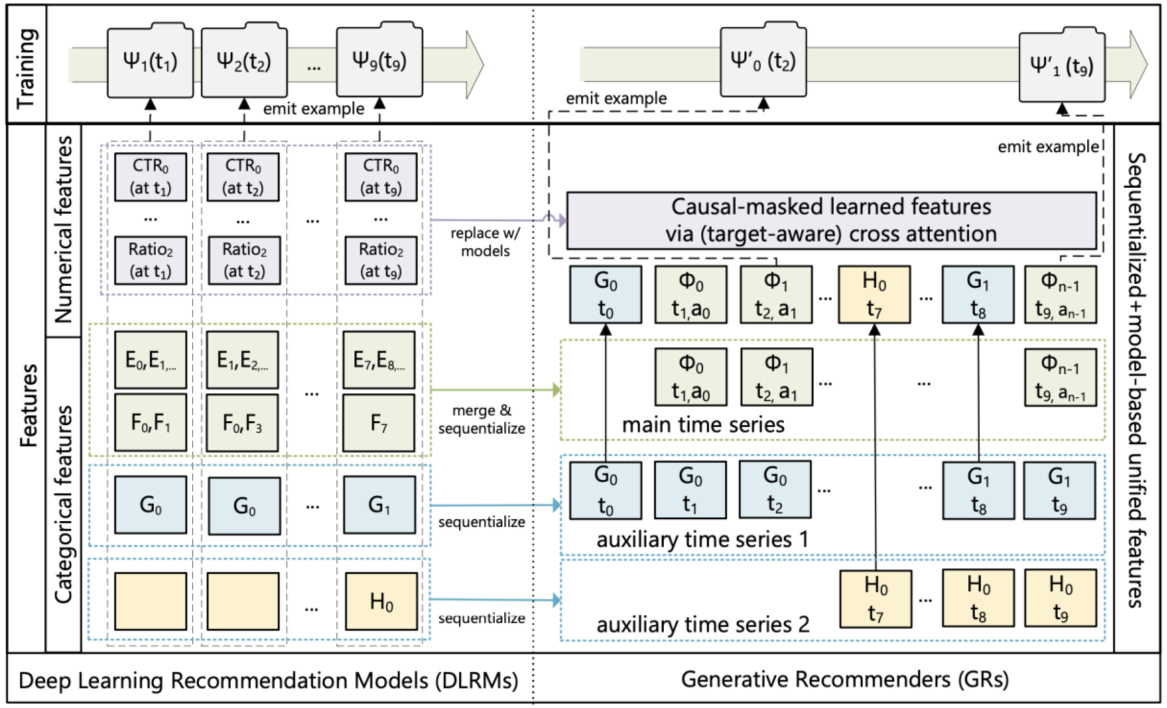
(a) SASRec$_F$. (b) NOVA-SR. (c) DIF-SR.



Decoupled Side Information Fusion for Sequential Recommendation. SIGIR 2022.
Non-invasive Self-attention for Side Information Fusion in Sequential Recommendation. AAAI 2021.
Practice on Long Behavior Sequence Modeling in Tencent Advertising. Arxiv 2025.

HSTU

- HSTU first selects the longest time series, typically by merging the features that represent items user engaged with, i.e., item ID, as the **main time series**
- It then *compress* the remaining time series by keeping the earliest entry per consecutive segment and then merge the results into the main time series
- **Remove numerical features**, relying on **the target attention** mechanism to handle them



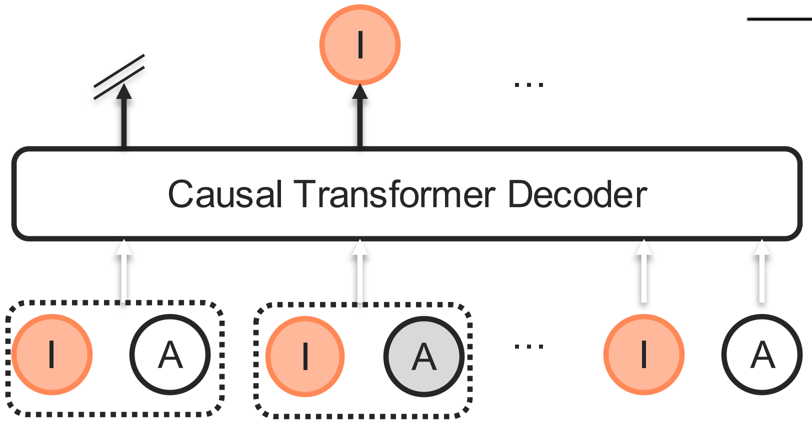
Task		Specification (Inputs / Outputs)
Ranking	x_is	$\Phi_0, a_0, \Phi_1, a_1, \ldots, \Phi_{n_c-1}, a_{n_c-1}$
	y_is	$a_0, \varnothing, a_1, \varnothing, \ldots, a_{n_c-1}, \varnothing$
Retrieval	x_is	$(\Phi_0, a_0), (\Phi_1, a_1), \ldots, (\Phi_{n_c-1}, a_{n_c-1})$
	y_is	$\Phi'_1, \Phi'_2, \ldots, \Phi'_{n_c-1}, \varnothing$ ($\Phi'_i = \Phi_i$ if a_i is positive, otherwise $\varnothing$)

Actions Speak Louder than Words- Trillion-Parameter Sequential Transducers for Generative Recommendations. ICML 2024.

HSTU

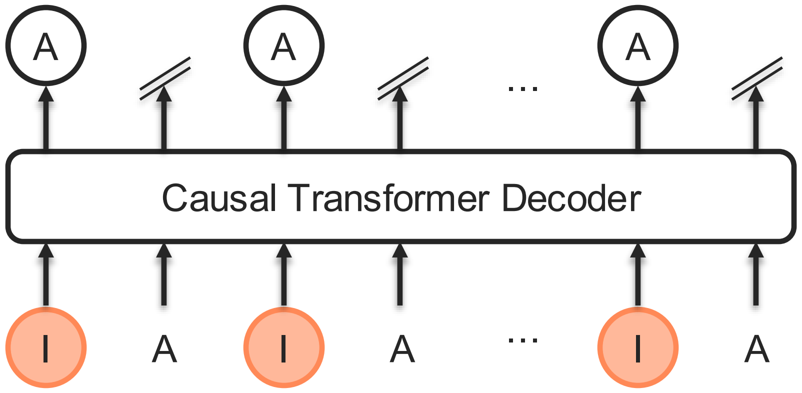
Task		Specification (Inputs / Outputs)
Ranking	x_is	$\Phi_0, a_0, \Phi_1, a_1, \ldots, \Phi_{n_c-1}, a_{n_c-1}$
	y_is	$a_0, \varnothing, a_1, \varnothing, \ldots, a_{n_c-1}, \varnothing$
Retrieval	x_is	$(\Phi_0, a_0), (\Phi_1, a_1), \ldots, (\Phi_{n_c-1}, a_{n_c-1})$
	y_is	$\Phi'_1, \Phi'_2, \ldots, \Phi'_{n_c-1}, \varnothing$ $(\Phi'_i = \Phi_i$ if a_i is positive, otherwise $\varnothing)$

Retrieval



- If there is **positive** feedback on the next item, then predict this item; otherwise, predict **empty**.

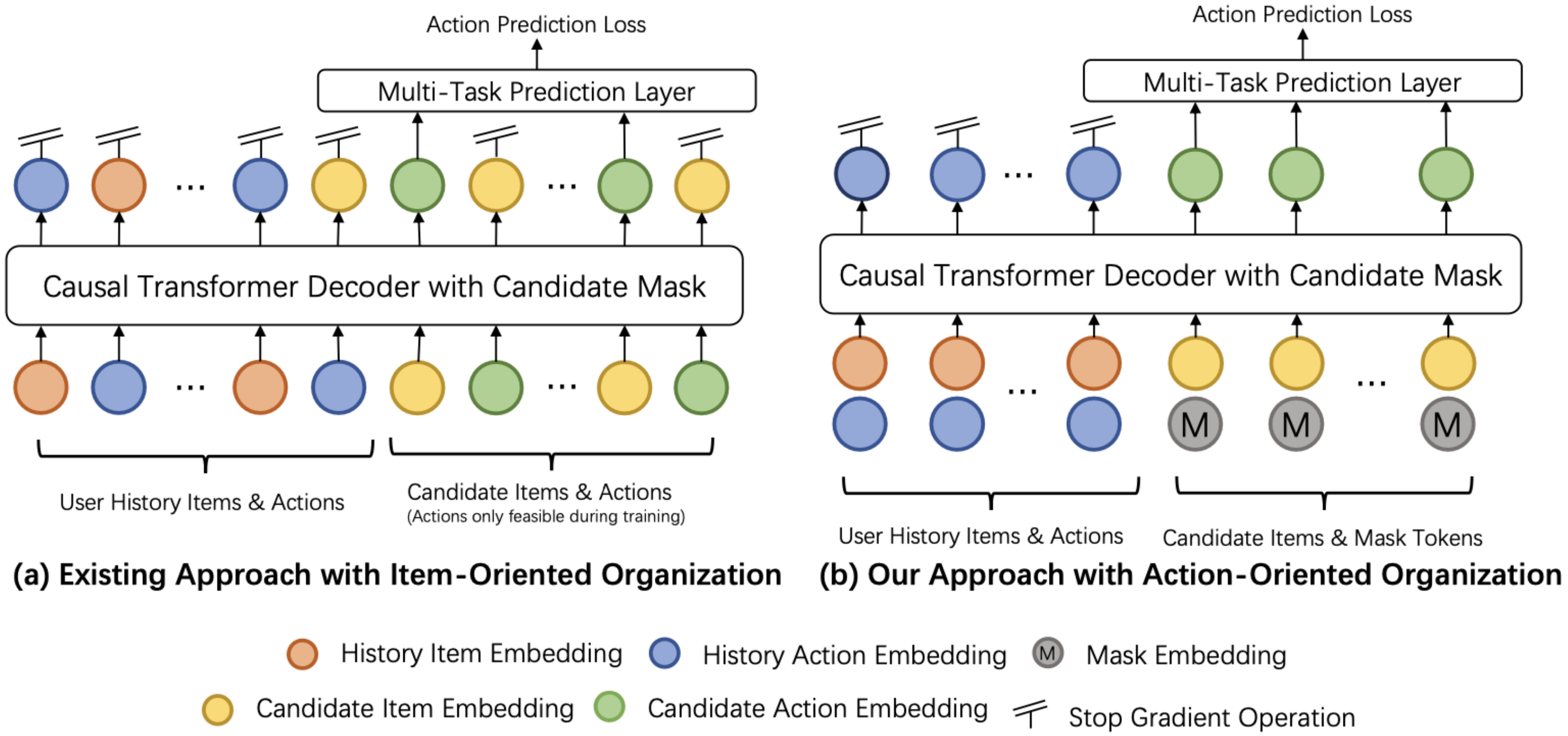
Ranking



- **Interleave** items and actions in the sequence. Predict the action for each item, and predict **empty** for the action.

GenRank

- **DON’T interleave** items and actions in the sequence, but treat actions as side info, as done in traditional methods like SASRec or DIN for the history behaviors
- For the target, predict the action of each item
- **Left side (history): SASRec; right side (target): HSTU**



Towards Large-scale Generative Ranking. 2025.

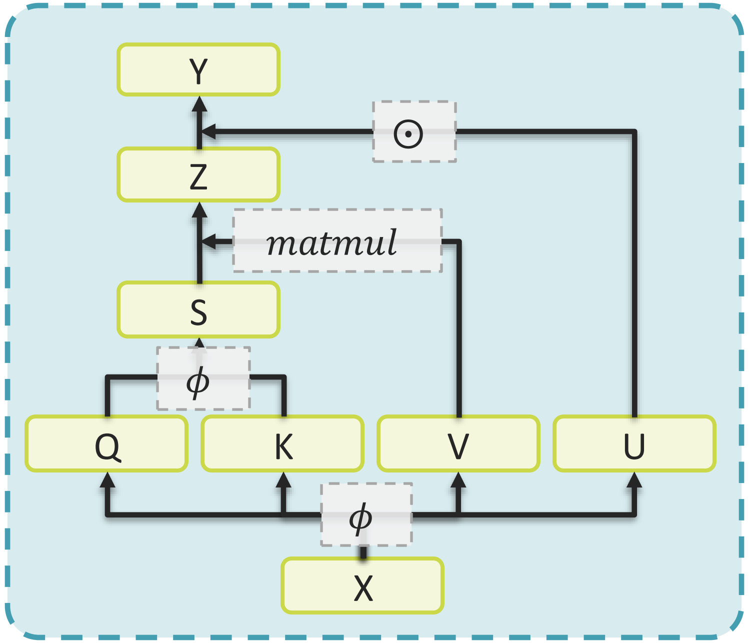
Overview

Generative Recommendation
- Token Organization
- Action Handling

Multimodal Recommendation
- Alignment
- Distance Transfer
- Semantic IDs

HSTU

- DNN-based methods **employs MLPs** $f_{MLP}(b_i, t)$to learn the interaction between behaviors and the target
- Employs a SiLU or SwiGLU activation function, due to "**the difficulty of approximating dot products with learned MLPs**"



Task		Specification (Inputs / Outputs)
Ranking	x_is	$\Phi_0, a_0, \Phi_1, a_1, \ldots, \Phi_{n_c-1}, a_{n_c-1}$
	y_is	$a_0, \varnothing, a_1, \varnothing, \ldots, a_{n_c-1}, \varnothing$
Retrieval	x_is	$(\Phi_0, a_0), (\Phi_1, a_1), \ldots, (\Phi_{n_c-1}, a_{n_c-1})$
	y_is	$\Phi'_1, \Phi'_2, \ldots, \Phi'_{n_c-1}, \varnothing$ ($\Phi'_i = \Phi_i$ if a_i is positive, otherwise $\varnothing$)

$$U(X), V(X), Q(X), K(X) = \text{Split}(\phi_1(f_1(X)))$$

$$A(X)V(X) = \phi_2\left(Q(X)K(X)^T + \text{rab}^{p,t}\right) V(X)$$

$$Y(X) = f_2\left(\text{Norm}\left(A(X)V(X)\right) \odot U(X)\right)$$

Actions Speak Louder than Words- Trillion-Parameter Sequential Transducers for Generative Recommendations. ICML 2024.

Temporal Interest Network (TIN)

- Target-aware Temporal Encoding
- Target-aware Attention
- Target-aware Representation

$$u_{TIN} = \alpha(Q,K) \odot (U \odot V)$$

$$= \alpha(\tilde{\boldsymbol{v}}_t W_Q, \tilde{\boldsymbol{e}}_i Q_K) \cdot (\tilde{\boldsymbol{v}}_t W_U \odot \tilde{\boldsymbol{e}}_i W_V)$$

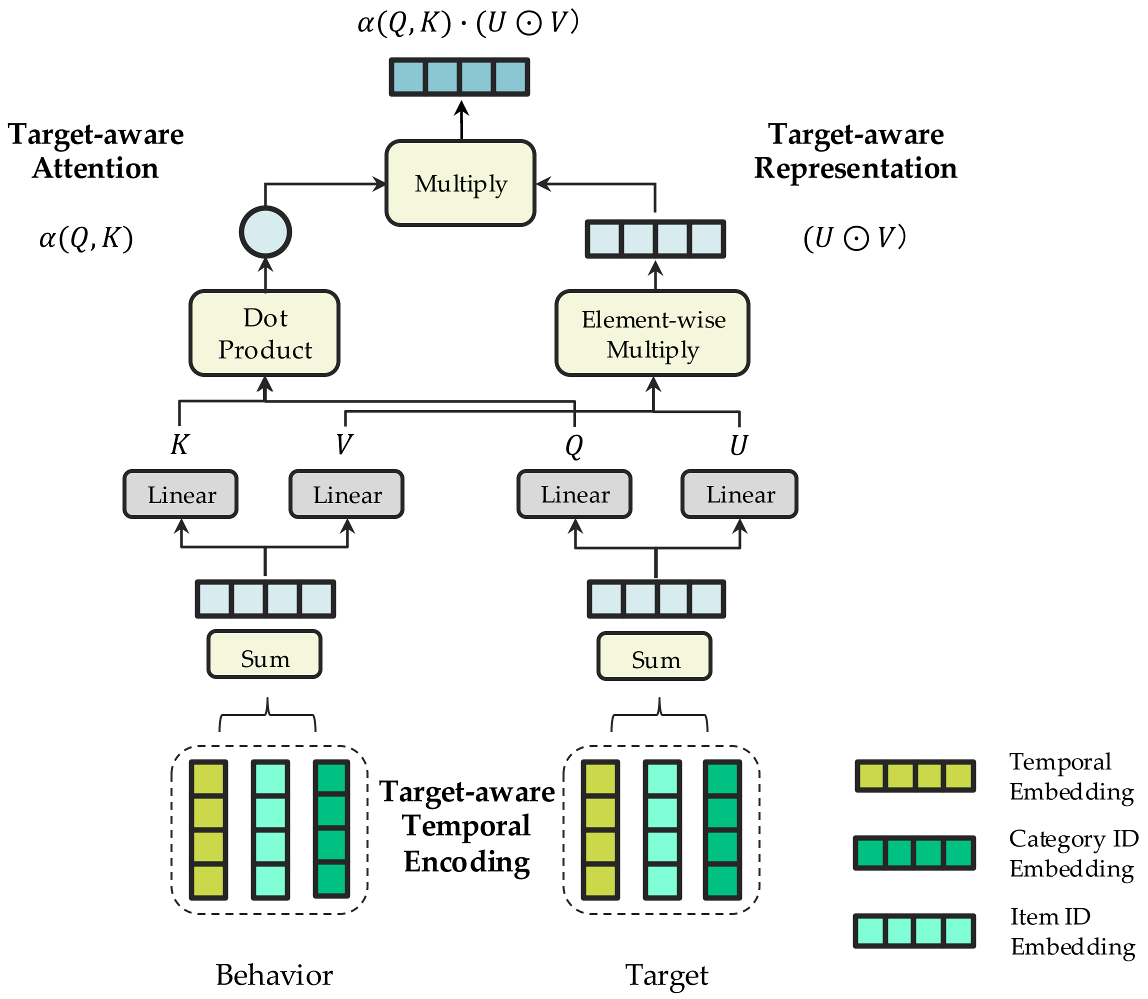
Formulation of TIN

$$U(X), V(X), Q(X), K(X) = \text{Split}(\phi_1(f_1(X)))$$

$$A(X)V(X) = \boxed{\phi_2\left(Q(X)K(X)^T + \text{rab}^{p,t}\right)}\boxed{V(X)}$$

$$Y(X) = f_2\left(\text{Norm}\left(A(X)V(X)\right)\boxed{\odot\, U(X)}\right)$$

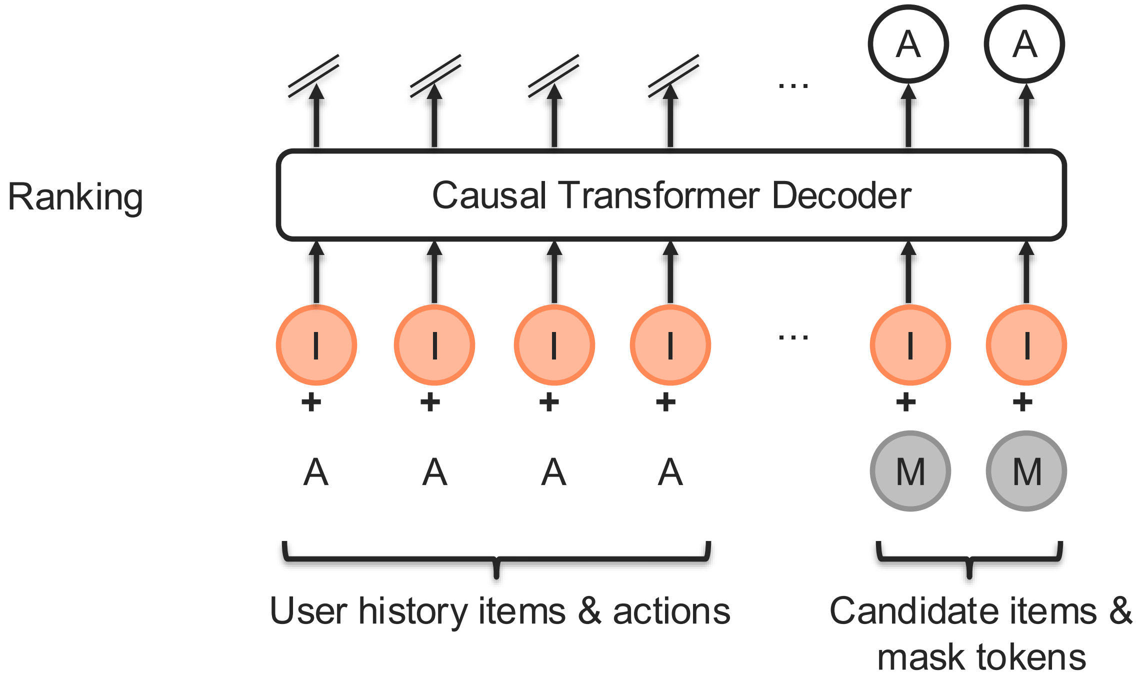
Formulation of HSTU

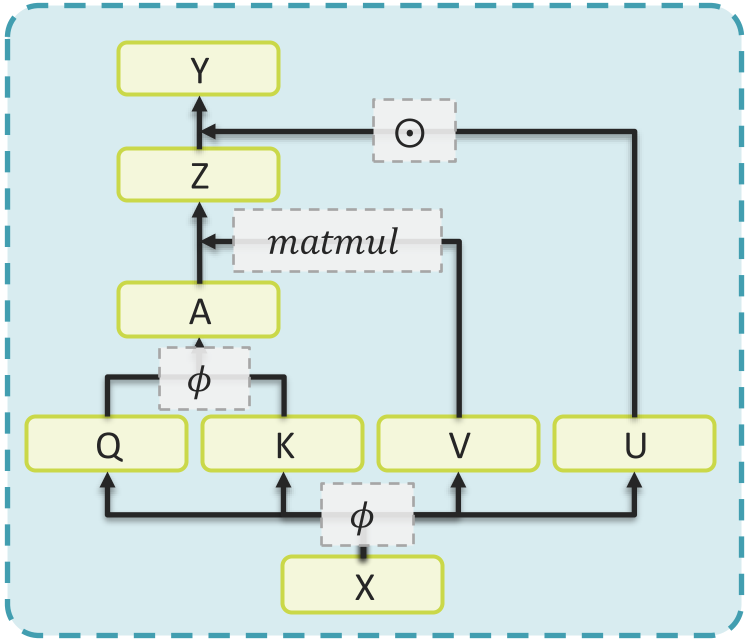


Temporal Interest Network for User Response Prediction, camera-ready. WWW 2024.

GenRank

- Similar architecture with HSTU: SwiGLU activation function





$X = I + A$ or $X = I + M$

$$Q = \phi(W^Q X), K = \phi(W^K X), V = \phi(W^V X), U = \phi(W^U X)$$

$$Z = \phi(QK^T + rab)V$$

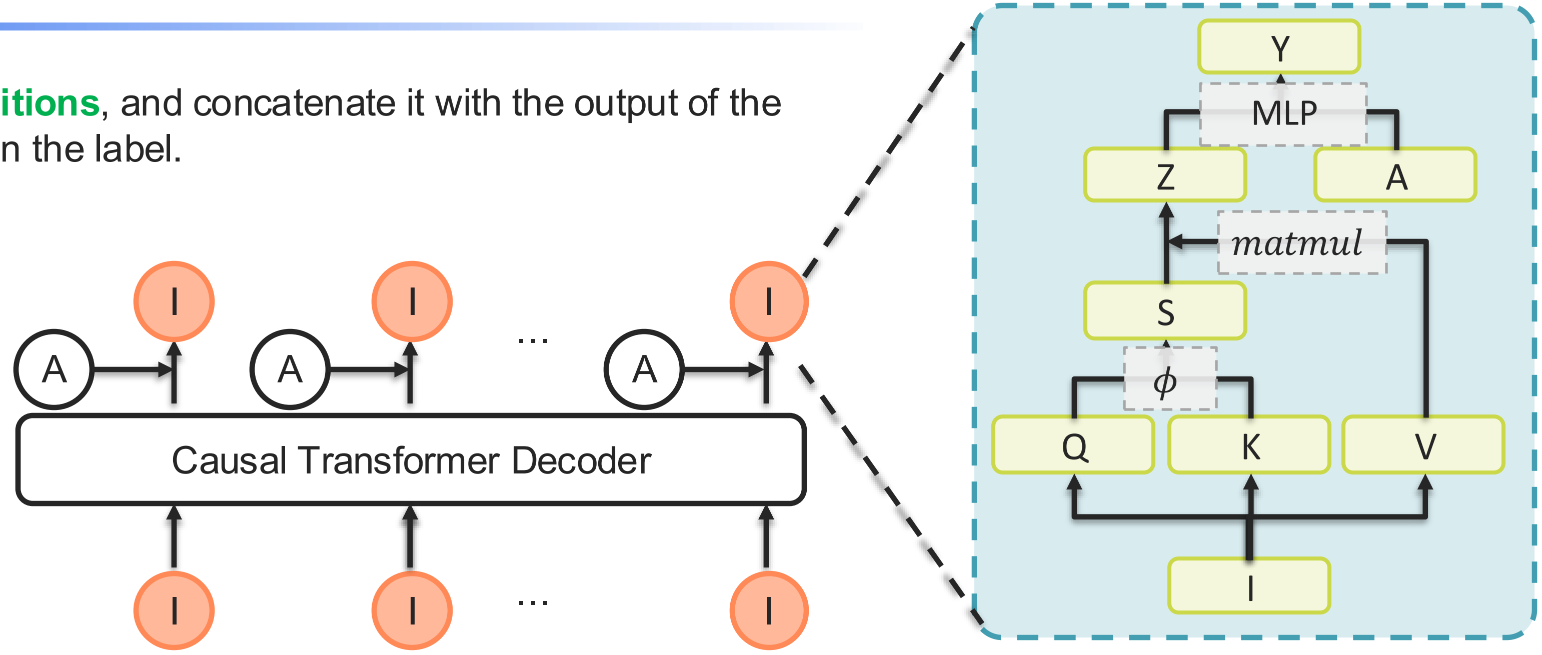
$$Y = W^Y(\mathrm{Norm}(Z) \odot U)$$

Towards Large-scale Generative Ranking. 2025.

PinRec

- **Treat actions as conditions**, and concatenate it with the output of the transformer to prediction the label.

Retrieval



$$Q = \phi(W^Q I), K = \phi(W^K I), V = \phi(W^V I)$$

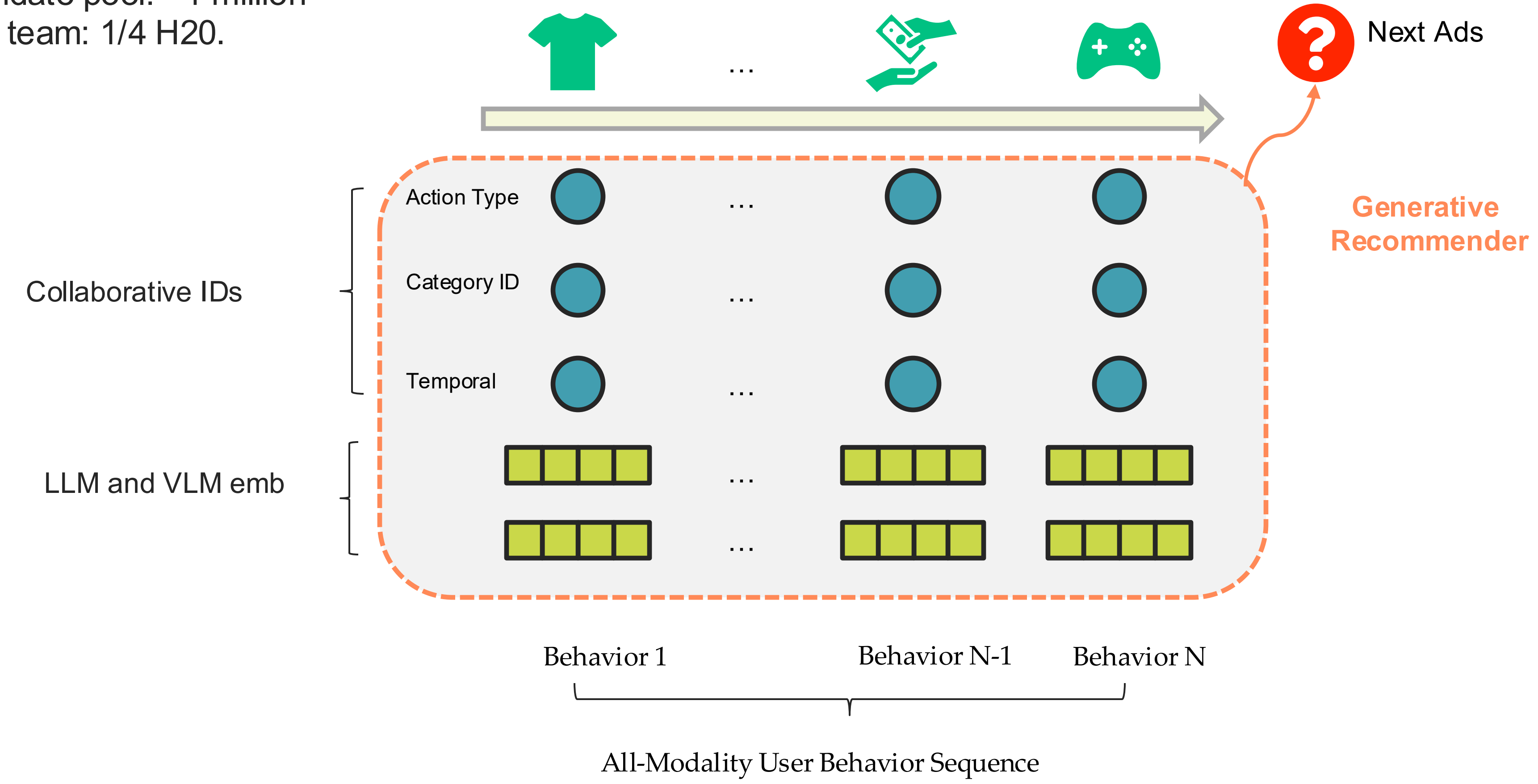
$$Z = \phi(QK^T)V \qquad Y = \phi(W^Y([Z; A]))$$

CIKM 2025 SEOUL
November 10 - 14

Tencent Advertising Algorithm Competition

All-Modality Generative Recommendation

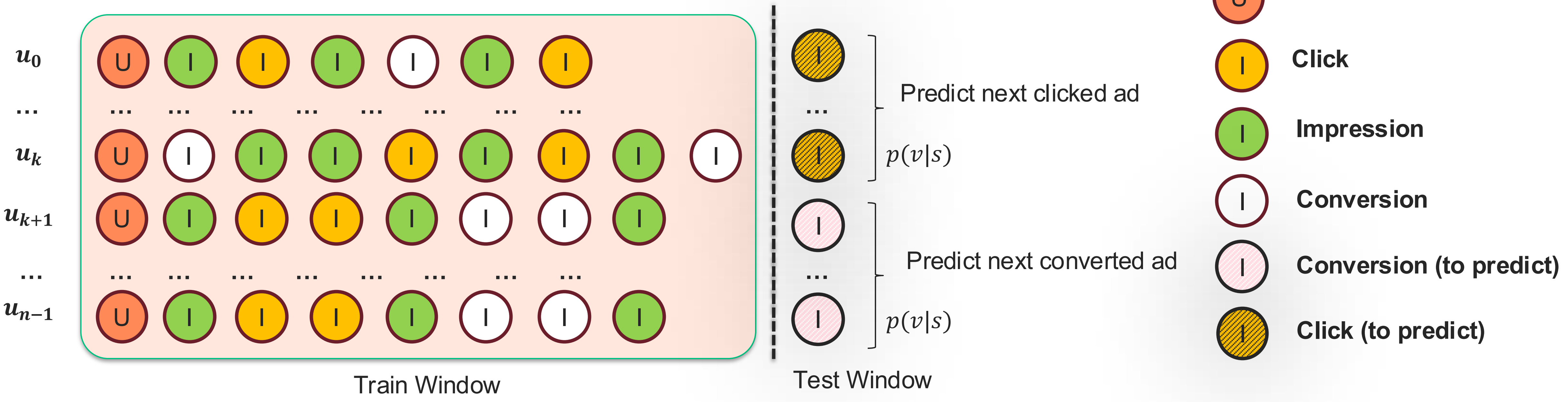
Setting
- 1 millions of user sequences
- Candidate pool: ~1 million
- Each team: 1/4 H20.



Tencent Advertising Algorithm Competition

All-Modality Generative Recommendation

Setting
- **10 millions** of user sequences
- Include **conversions** in both features and target
- Each team: **7 H20 GPU**
- In evaluation, conversions have larger scores (2.5x) than the clicks



TAAC

AWARD

Total Prize Pool of 3.6 Million RMB Ready!

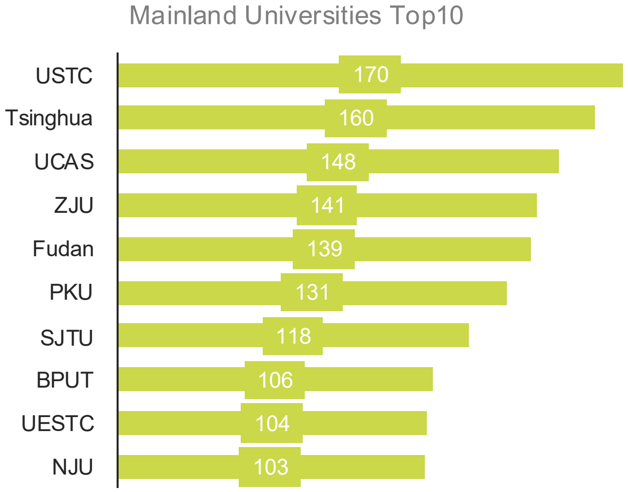
Note: Based on comprehensive evaluation by the competition,the award may go unassigned. Disbursement of the prize is subject to the rules established post-award.

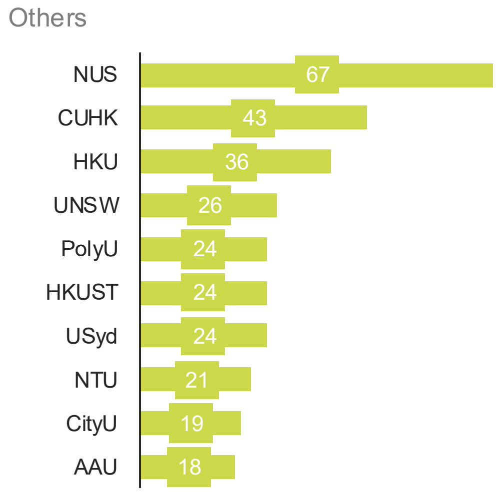
1	2	3	4-10
Champion	**Runner-Up**	**Second Runner-Up**	**4th-10th Places**
1 team	1 team	1 team	per team
¥2,000,000 RMB	¥600,000 RMB	¥300,000 RMB	¥100,000 RMB

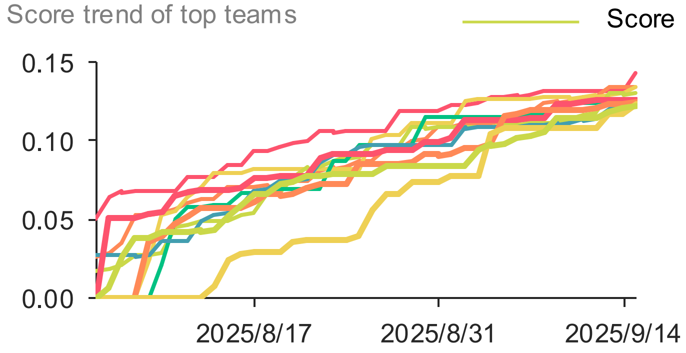
腾讯广告算法大赛
Tencent Advertising Algorithm Competition

TAAC

8400+	4600+	2800+
Registration	Enroll	#Teams







CIKM 2025 SEOUL
November 10 - 14

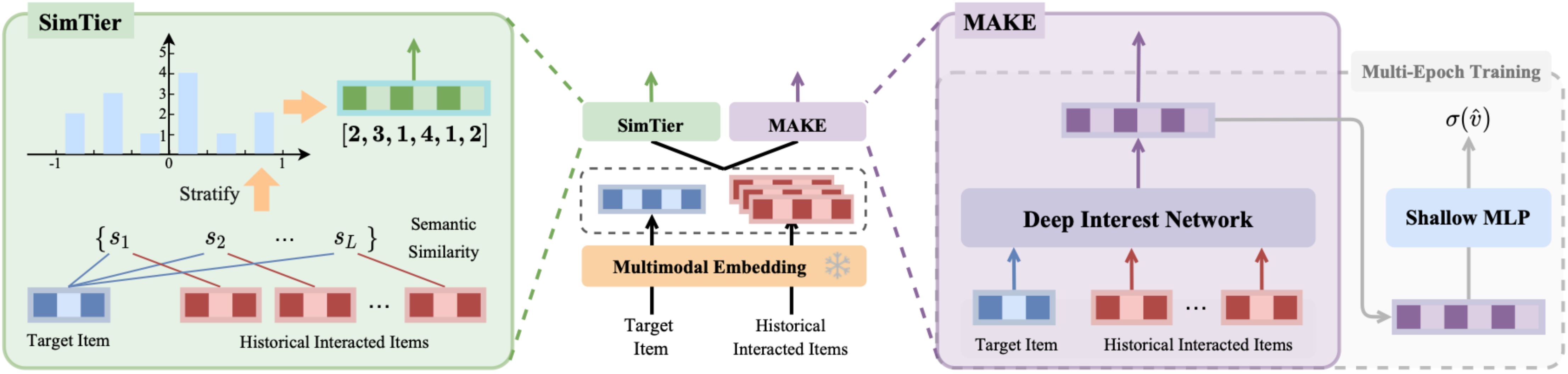
Overview

Generative Recommendation
- Token Organization
- Action Handling

Multimodal Recommendation
- Distance Transfer
- Alignment
- Semantic IDs

SimTier

- Calculate the **similarity score** between multi-modal representation of behaviors and the target
- **Histogram** of the scores in a N pre-defined buckets
- Use the N-dimension vector as a new representation



(a). SimTier (b). MAKE

MNSE

- Calculate the **distance** between two features based on the source embedding
- Encoding the distance with **n-ary**
- Train the encoded embeddings in the target task, together with other embeddings in the target space

$$f_{\mathrm{MNS}}(v) = \left[\sum_{k=1}^{K_2} \mathbf{X}^{(2)}_{2k+\mathbb{B}_k}, \sum_{k=1}^{K_3} \mathbf{X}^{(3)}_{3k+\mathbb{C}_k}, \ldots, \sum_{k=1}^{K_n} \mathbf{X}^{(n)}_{nk+\mathbb{N}_k}\right]$$

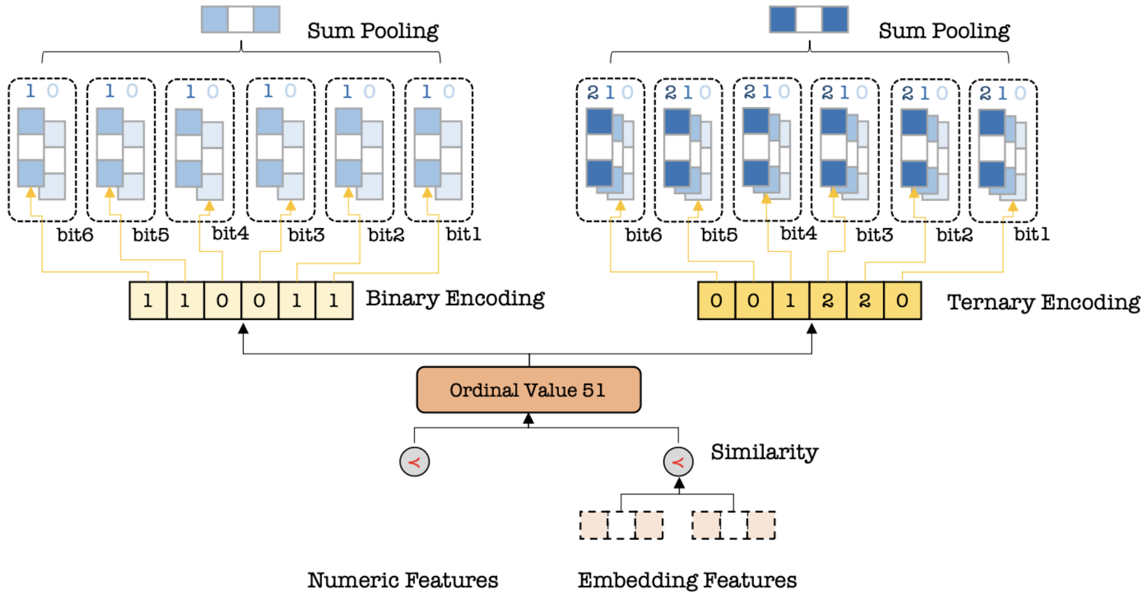
$$\mathbb{B} = \mathrm{func\_binary}(v),\ \mathbb{C} = \mathrm{func\_ternary}(v), \ldots$$

Numerical Feature (Decimal)	Binary	
45	0000101101	Continuity
46	0000101110	Discriminability
957	1110111101	

Property

Issue

Numerical Feature (Decimal)	Binary		Ternary	
63	0000111111	Carry	000002100	Continuity
64	0001000000	Modulo	000002101	Discriminability
575	1000111111		000210022	



Key properties of n-ary: continuity, discriminability

CIKM 2025 SEOUL November 10 - 14

Ads Recommendation in a Collapsed and Entangled World. KDD 2024.
Numerical Feature Representation with Hybrid N-ary Encoding. CIKM 2022.

Overview

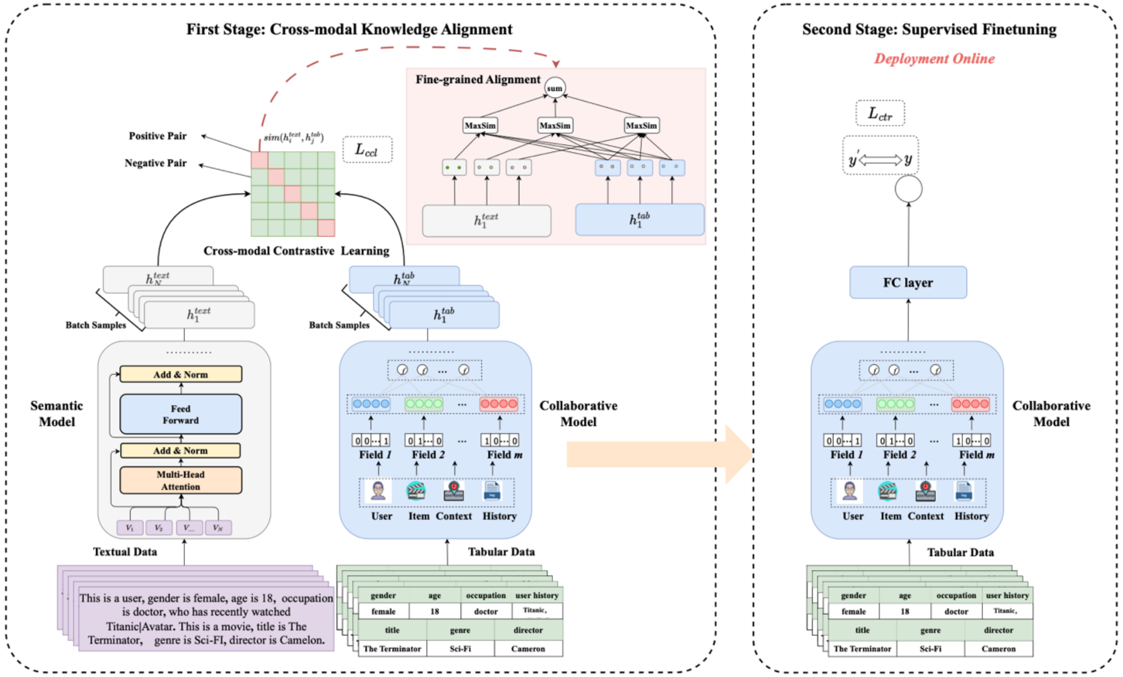
Generative Recommendation
- Token Organization
- Action Handling

Multimodal Recommendation
- Distance Transfer
- Alignment
- Semantic IDs

CTRL

- Clip-like Textual-to-Tabular **contrastive loss**

$$\mathcal{L}^{textual2tabular} = -\frac{1}{N}\sum_{k=1}^{N} log \frac{exp(sim(\mathbf{h}_k^{text}, \mathbf{h}_k^{tab})/\tau)}{\sum_{j=1}^{N} exp(sim(\mathbf{h}_k^{text}, \mathbf{h}_j^{tab})/\tau)},$$



CTRL: Connect Collaborative and Language Model for CTR Prediction.

PAD (Pre-train, Align and Disentangle)

- Adopt **MK-MMD** (multi-kernel maximum mean discrepancy) as the alignment loss to capture **all information about the distribution**
- Combine the alignment with the BCE loss to avoid **catastrophic forgetting**

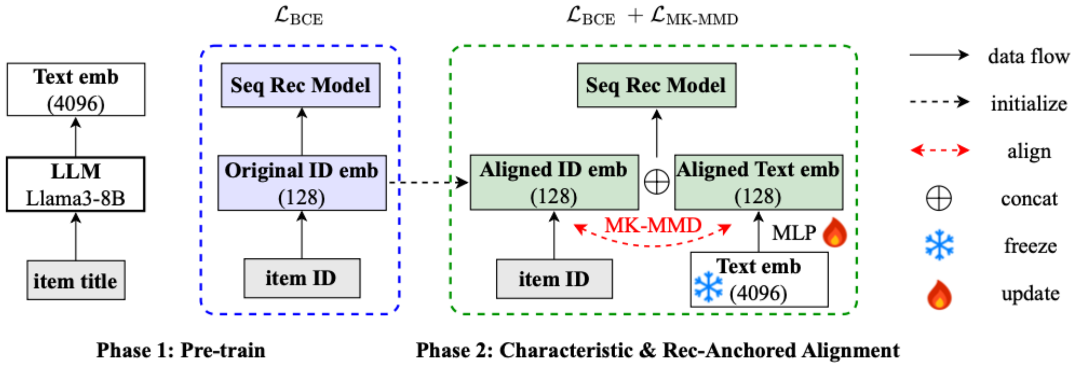
Avoid **catastrophic forgetting** **Space Alignment**

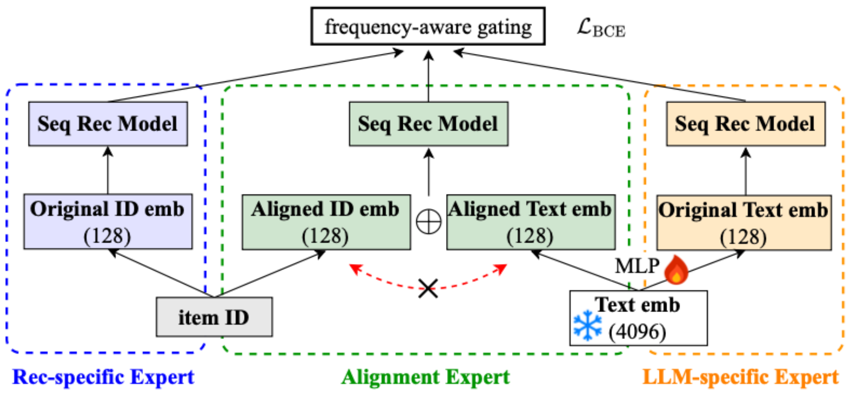
$$\mathcal{L} = \boxed{\mathcal{L}_{\text{REC}}} + \gamma \cdot \boxed{\mathcal{L}_{\text{MK-MMD}}}$$

$$\mathcal{L}_{\text{MK-MMD}} = D_k^2\left(f_{\text{MLP}}(\mathsf{SG}(\mathcal{D}_{\text{text}}), \boldsymbol{w}), \mathcal{D}_{\text{rec}}\right)$$

$$\mathcal{L}_{\text{REC}} = \frac{1}{n}\sum_{i=1}^{n} \text{BCE}\left(f_\theta\left(\{\mathbf{h}_i^s\}, \{\mathbf{h}_i^c\}, \mathbf{x}_i^s, \mathbf{x}_i^c\right), y_i\right)$$

$$\text{MK-MMD}^2(X_s, X_t) = \left\| \frac{1}{n}\sum_{i=1}^{n} \phi_k(x_s^i) - \frac{1}{m}\sum_{j=1}^{m} \phi_k(x_t^j) \right\|_{\mathcal{H}_k}^2$$





Pre-train, Align and Disentangle: Empowering Sequential Recommendation with Large Language Models. SIGIR 2025.

Overview

Generative Recommendation

- Token Organization
- Action Handling

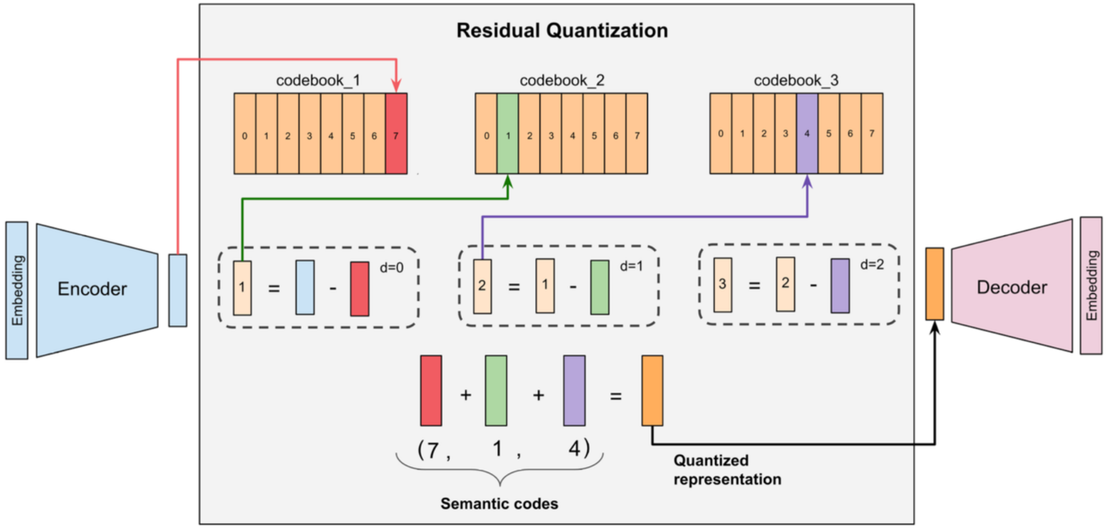
Multimodal Recommendation

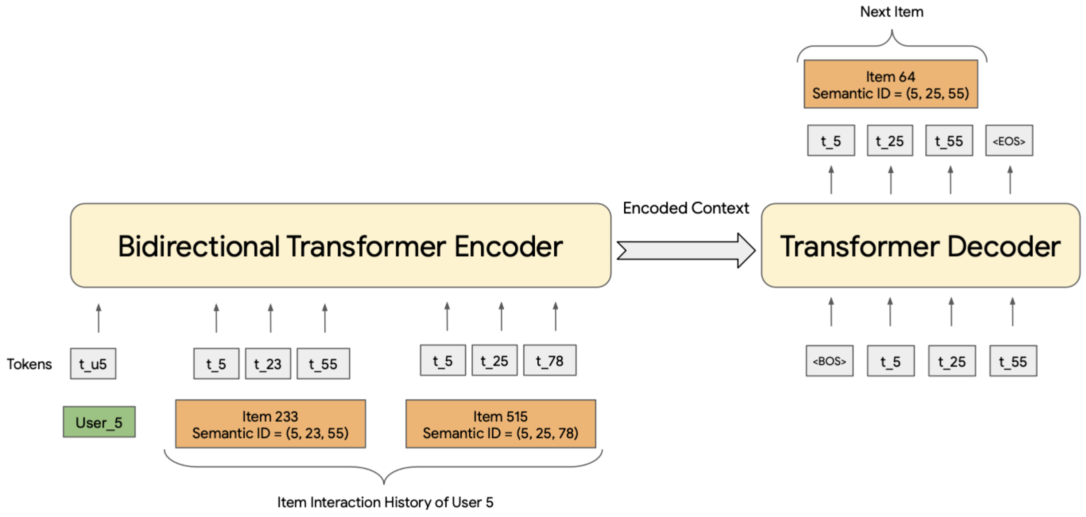
- Distance Transfer
- Alignment
- Semantic IDs

CIKM 2025 SEOUL
November 10 - 14

Tiger

- Use **RQ-VAE** on the LLM representation to get **semantic IDs**
- Use semantic IDs in the downstream recommendation models





Why it works?

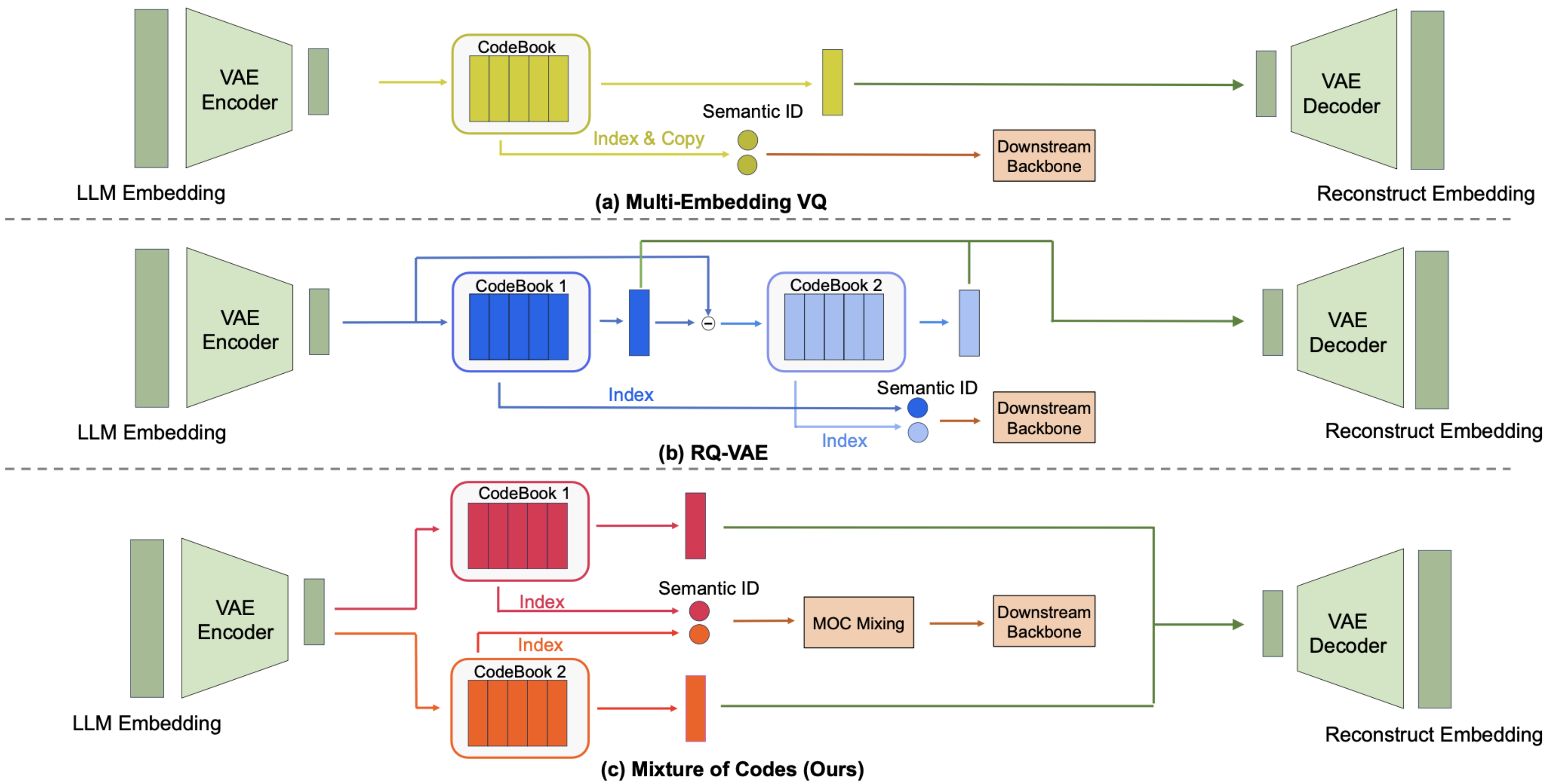
- **RQ-VAE** and **Semantic IDs** to capture the **source** space structures.
- **Semantic ID Embeddings** to align with the **target** space.

Recommender Systems with Generative Retrieval, NIPS 2023.

Reference

Parallel Semantic IDs - MoC

- RQ-VAE scales semantic IDs in a cascading way
- MoC scales semantic IDs in **a parallel way**

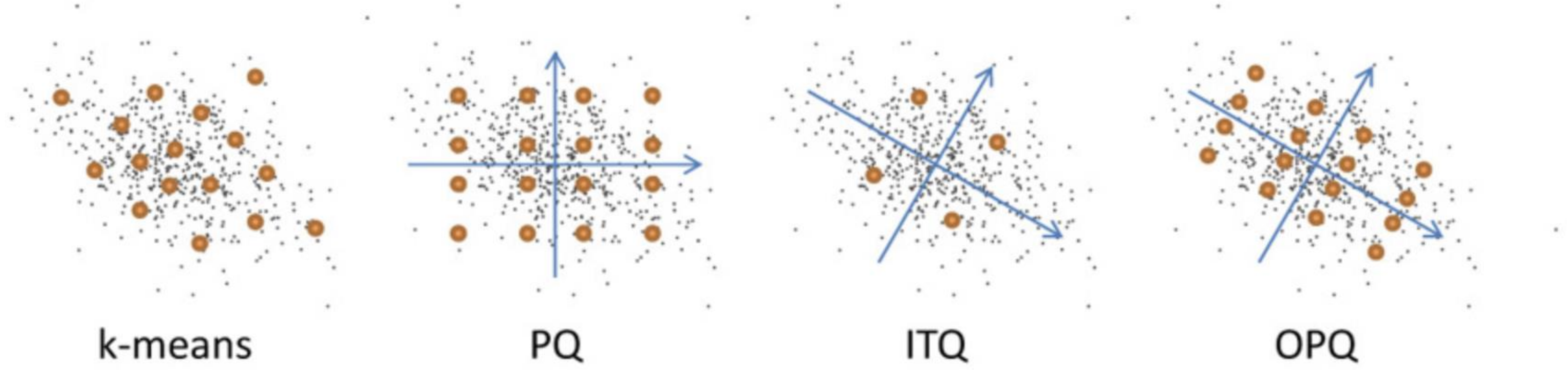


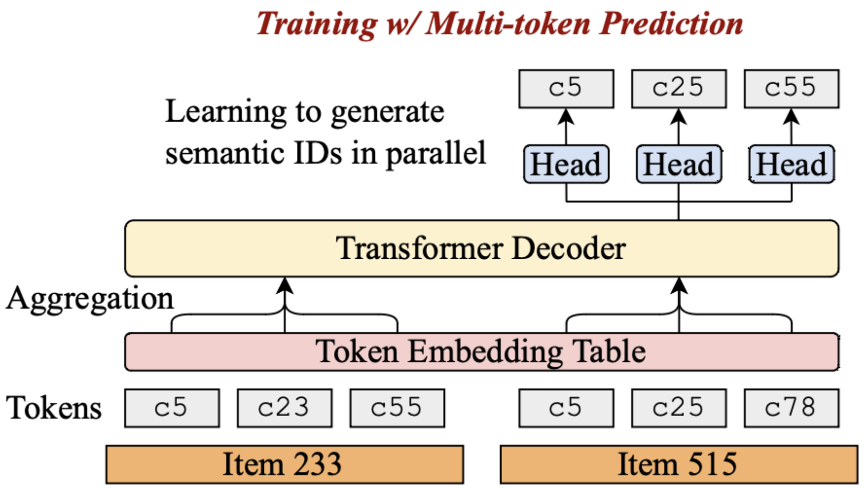
Towards Scalable Semantic Representation for Recommendation. Arxiv 2024.

Parallel Semantic IDs

- Generates parallel VQs with a **multi-token prediction loss**
- Employ **OPQ** to split sub-space

$$\mathcal{L} = -\sum_{j=1}^{m} \log \mathbb{P}^{(j)}(c_{t,j}|s) = -\sum_{j=1}^{m} \log \frac{\exp(\boldsymbol{e}_{c_{t,j}}^{\top} \cdot \mathrm{g}_j(\boldsymbol{s})/\tau)}{\sum_{c \in C^{(j)}} \exp(\boldsymbol{e}_c^{\top} \cdot \mathrm{g}_j(\boldsymbol{s})/\tau)},$$





Generating Long Semantic IDs in Parallel for Recommendation. KDD 2025.

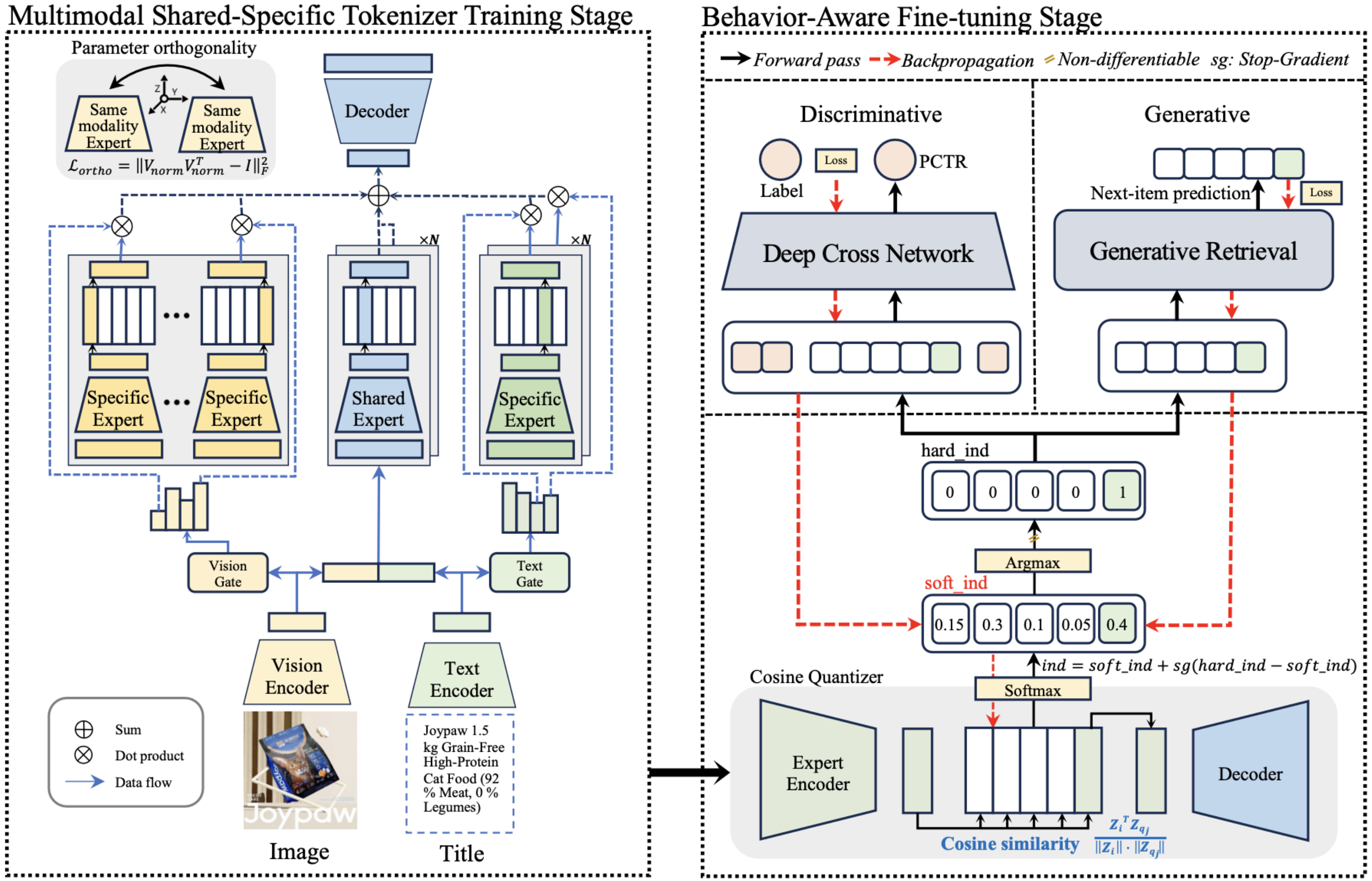
Parallel Semantic IDs - MMQ

- Modality-shared and –specific experts
- **Orthogonality constraints** on experts

$$\mathcal{L}_{\text{ortho\_shared}} = \left\| \mathbf{V}_{\text{norm}} \mathbf{V}_{\text{norm}}^{\top} - \mathbf{I} \right\|_F^2 ,$$

$$\mathcal{L}_{\text{ortho\_specific}} = \left\| \mathbf{V'}_{\text{norm}} \mathbf{V'}_{\text{norm}}^{\top} - \mathbf{I} \right\|_F^2 ,$$

- Similar to Redundancy Reduction or Expert De-correlation Principle

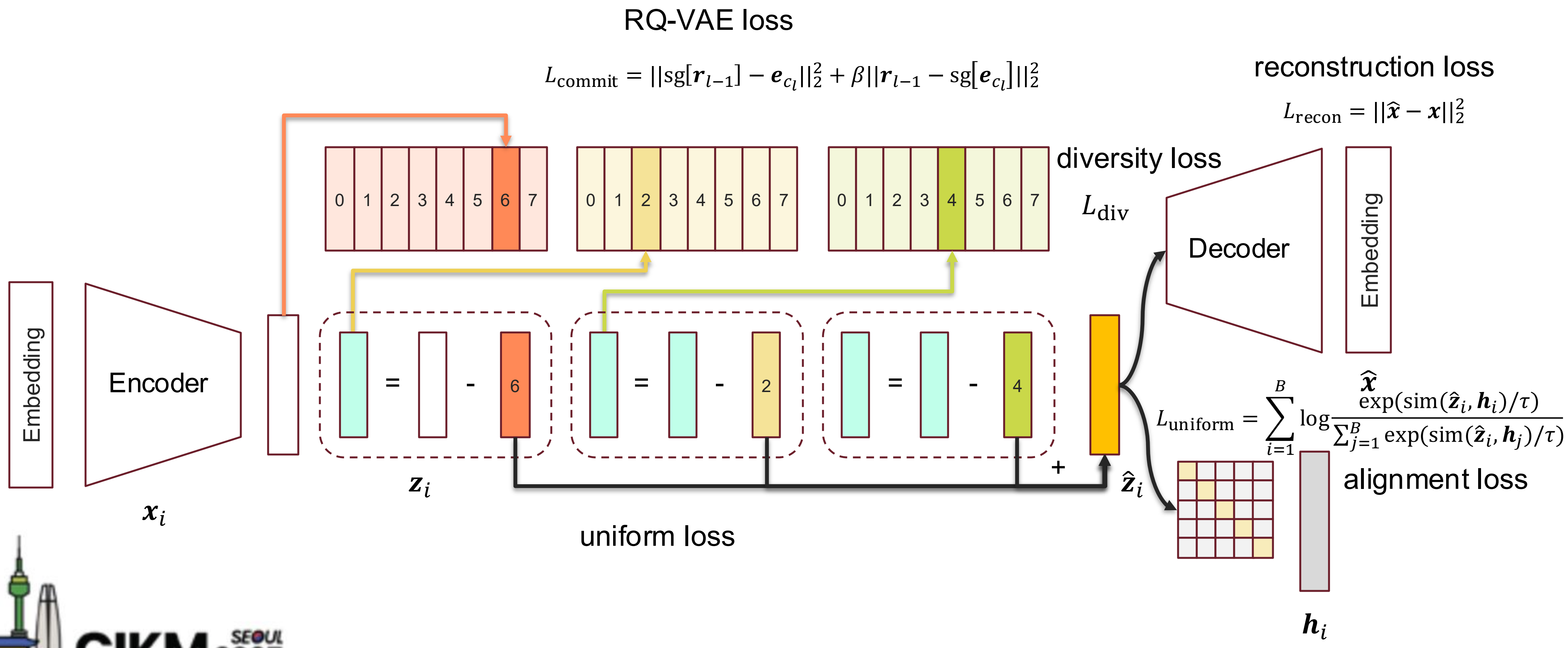


MMQ- Multimodal Mixture-of-Quantization Tokenization for Semantic ID Generation and User Behavioral Adaptation. 2025.

Collaborative-aware - LETTER

- LETTER **aligns** the semantic IDs with the **downstream recommendation models**
- **Diversity loss** to promote the uniform distribution of the code embeddings

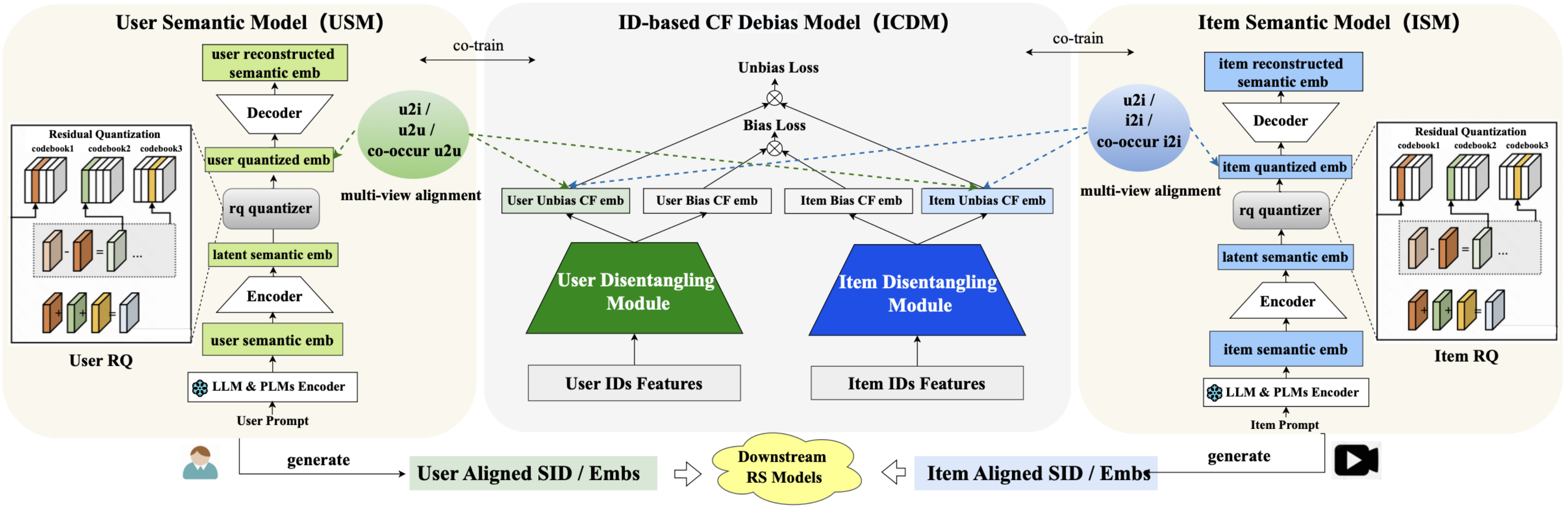
$$\mathcal{L}_{\text{CF}} = -\frac{1}{B}\sum_{i=1}^{B} \frac{\exp(<\hat{\boldsymbol{z}}_i, \boldsymbol{h}_i>)}{\sum_{j=1}^{B}\exp(<\hat{\boldsymbol{z}}_i, \boldsymbol{h}_j>)},$$



Learnable Item Tokenization for Generative Recommendation. CIKM 2024.

Collaborative-aware - DAS

- **Various contrastive/alignment losses** between users and items across the semantic and collaborative embeddings



Dual-Aligned Semantic IDs Empowered Industrial Recommender System. 2025.

Q & A

jonaspan@tencent.com