From Representation to Generative Learning: A Deep Dive into Multimodal IR Innovation at Huawei

Jieming Zhu
Huawei Noah's Ark Lab
https://jiemingzhu.github.io
WWW 2025 @ Sydney

About Me

- Researcher, Huawei Noah's Ark Lab
- Research areas:
  - Recommender systems
  - Personalized LLMs
  - RAG/Agents
- Find me at https://jiemingzhu.github.io

Huawei Noah's Ark Lab for AI Research

2012 Laboratory


Network Intelligence


Enterprise Intelligence


Terminal Intelligence


Vehicles Intelligence

Business Success

Noah's Ark Laboratory

AI System Engineering

AI Algorithm Application

Computer Vision

Natural Language Processing

Search & Recommendation

Decision Making & Reasoning

AI Theory

Advanced Technology

AI Research Collaboration

open source

START UP

Professional Advisory Committee

10+ Country, 25~University, 50~ projects, 1,000+ Researchers

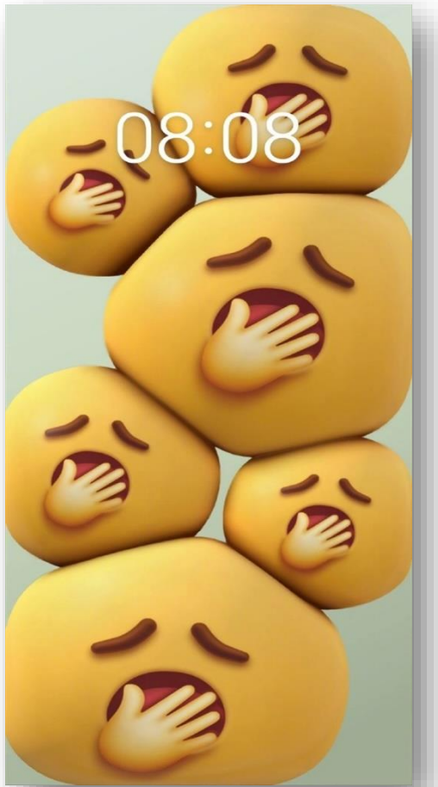
Healthy Eco-system

Multimodal IR Scenarios at Huawei

News Feeds

Micro-Videos

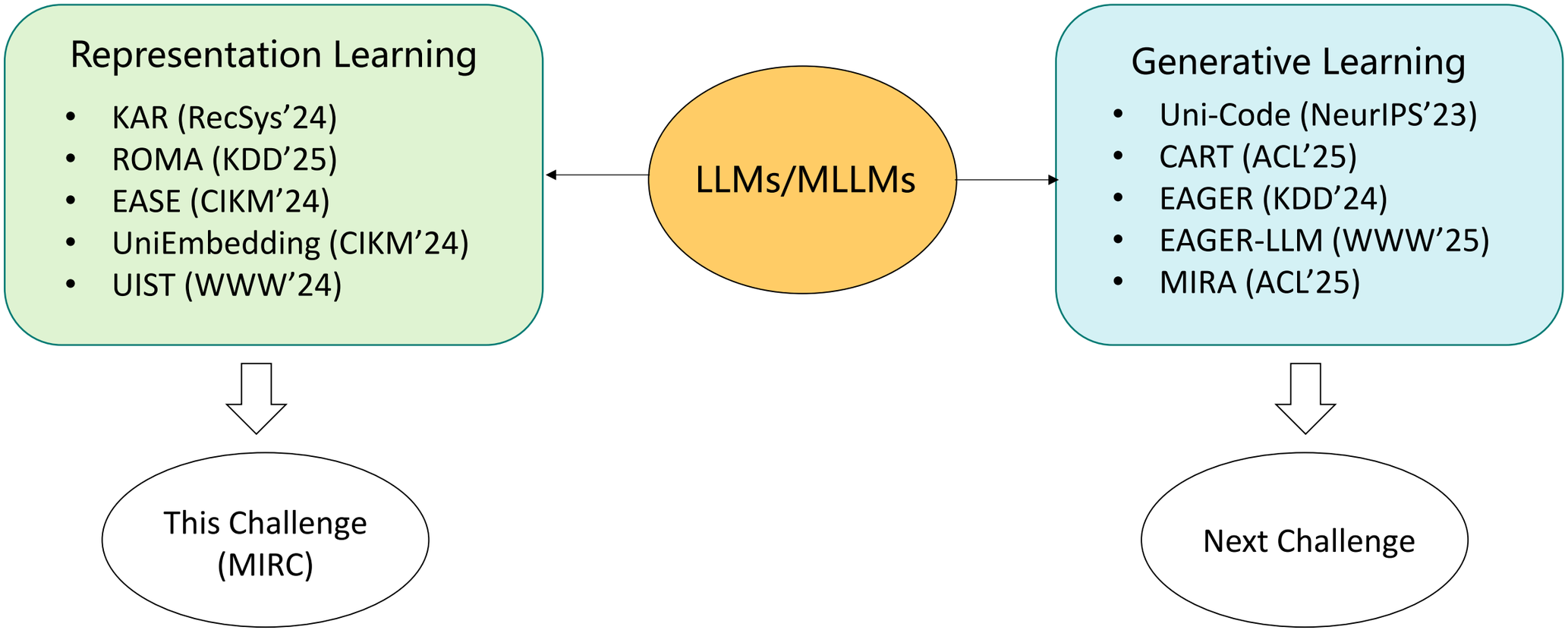
Music Streaming



Mobile Themes

And many more…

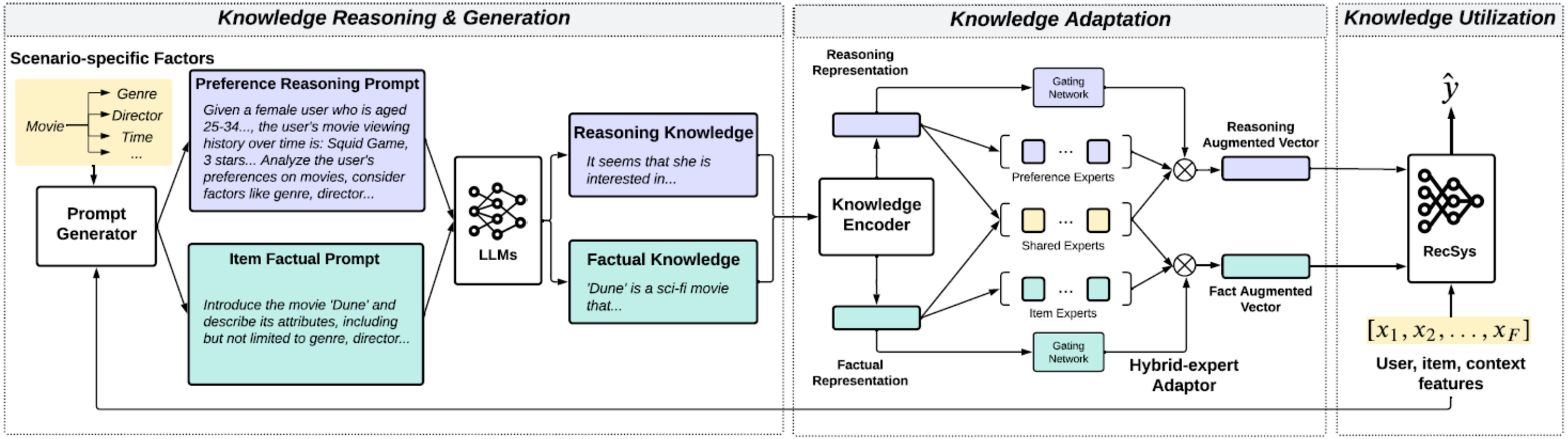
This Talk


Representation Learning
KAR (RecSys'24)
ROMA (KDD'25)
EASE (CIKM'24)
UniEmbedding (CIKM'24)
UIST (WWW'24)
LLMs/MLLMs
Generative Learning
Uni-Code (NeurIPS'23)
CART (ACL'25)
EAGER (KDD'24)
EAGER-LLM (WWW'25)
MIRA (ACL'25)
This Challenge (MIRC)
Next Challenge

Outline

- **Representational IR**
  - LLMs as Extractors
  - LLMs as Encoders
  - LLMs as Decoders
  - Embedding across Domains

KAR: (M)LLMs as Extractors



Knowledge Extraction

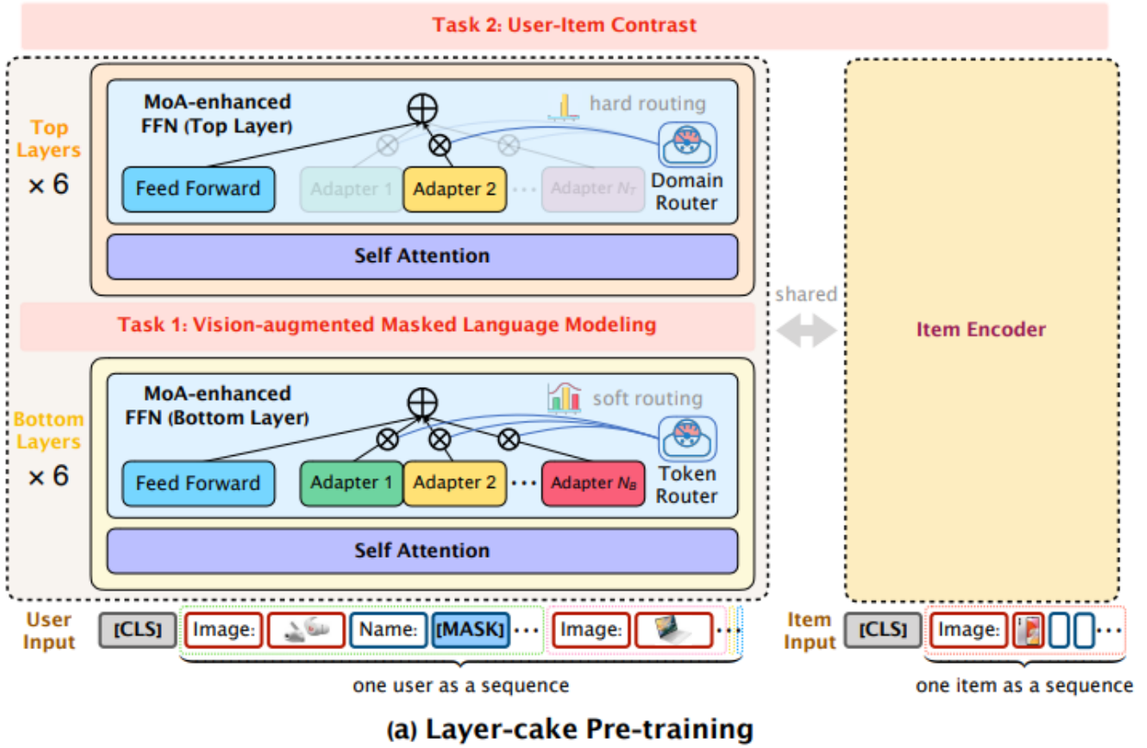
- **Prompting:** Leveraging prompt templates to define preference reasoning prompt and item factual prompt
- **Extraction:** Summarizing and generating descriptive text for items and users

Knowledge Adaptation

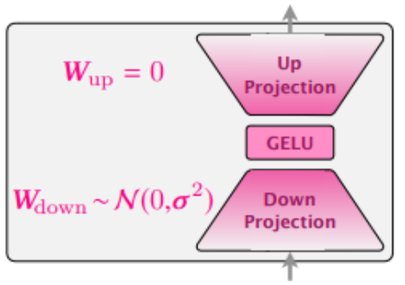
- **Knowledge Encoder:** Transforming extracted texts into hidden representations
- **Knowledge Adaptor:** Leveraging MOE adaptors to fuse knowledge as informative embedding features

Xi et al. Towards Open-World Recommendation with Knowledge Augmentation from Large Language Models, RecSys 2024

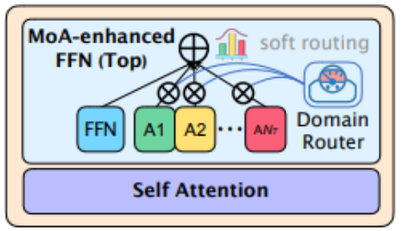
ROMA: (M)LLMs as Encoders



(a) Layer-cake Pre-training



(b) Adapter Parameterization



(c) Top Transformer Layer in Downstream Fine-tuning

Multi-modal			
MISSRec	MMSASRec	ROMA	(Imprv.)
0.0793	0.0977	**0.1139**	10.91%
0.1407	0.1373	**0.1619**	11.81%
0.0675	0.0869	**0.1058**	11.25%
0.0531	0.0732	**0.0891**	21.72%
0.1142	0.1143	**0.1426**	24.76%
0.0449	0.0681	**0.0823**	20.85%
0.0765	0.0842	**0.0959**	13.90%
0.1324	0.1126	0.1295	-
0.0668	0.0809	**0.0919**	10.46%
0.0852	0.1161	**0.1343**	10.08%
0.1506	0.1649	**0.1735**	5.22%
0.0719	0.1155	**0.1283**	11.08%
0.0890	0.1135	**0.1288**	12.88%
0.1384	0.1428	**0.1596**	11.76%
0.0783	0.1105	**0.1237**	11.95%
0.0841	0.0944	**0.1049**	7.26%
0.1249	0.1188	**0.1307**	4.64%
0.0772	0.0888	**0.1013**	8.23%

Multi-modal Pretraining

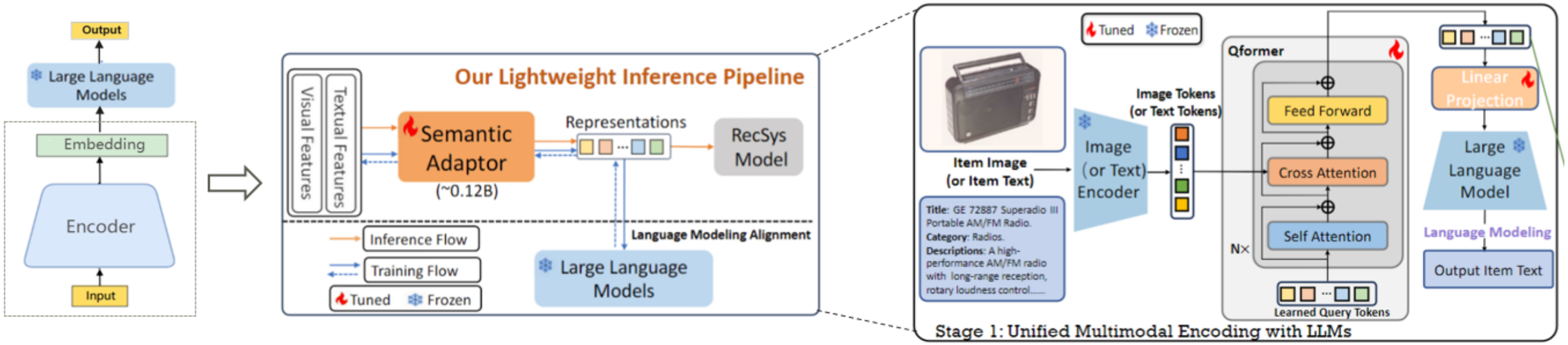
- **Vision-augmented Masked Language Modeling:** aligning vision tokens and language tokens to unify their modeling
- **User-Item Contrastive Learning :** matching between user sequence and target item

Parameter-Efficient Finetuning

- **Modality Adaptors:** Learning LoRA-based adapters
- **MoA-enhanced FFN:** Leveraging MOE-like mixture of adapters with domain-aware routers

Lu et al., ROMA: Recommendation-Oriented Language Model Adaptation Using Multi-Modal Multi-Domain Item Sequences, KDD 2025

EASE: (M)LLMs as Decoders



EASE:

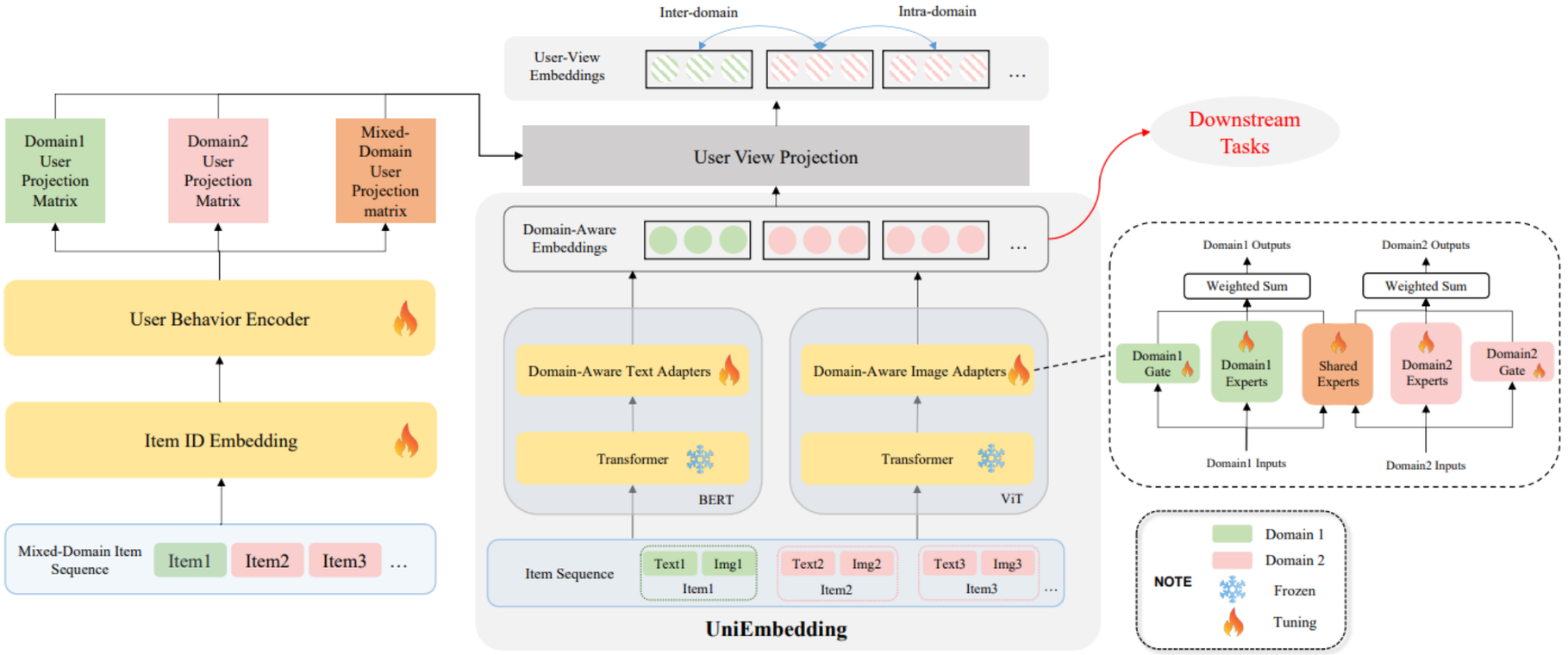
- **LLM Autoencoder**: Proposing an Autoencoder Architecture with LLM as the Decoder.
- **Distillation-like Training**: Extracting Semantic Information from Large Models as Supervisory Signals to Enhance Multi-Modal Implicit Representations.
- **Efficient and Universal:** Inference Does Not Depend on LLM, Supports Any Encoder Architecture.

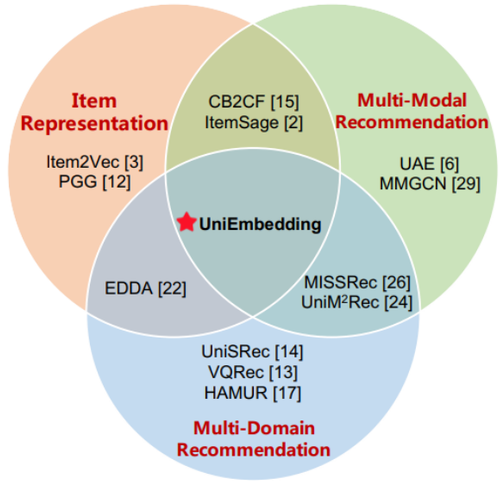
Please describe what is in the image. Pay attention to the text information shown in the image and provide a detailed response:

On the left side of the image is a male model wearing a red shirt and jeans, sitting on a white stool in a relaxed and confident posture. On the right side is the promotional advertisement text, which reads "FARFETCH 发发奇 尚品集结 限时75折". "FARFETCH" refers to Farfetch.com, an e-commerce platform specializing in international fashion brands. Below the text is the phrase "11.11 发发星愿", indicating that this is a promotional offer for the Double 11 shopping event.

Qiu et al., EASE: Learning Lightweight Semantic Feature Adapters from Large Language Models for CTR Prediction, CIKM 2024

UniEmbedding: Universal Embedding across Domains





UniEmbedding:

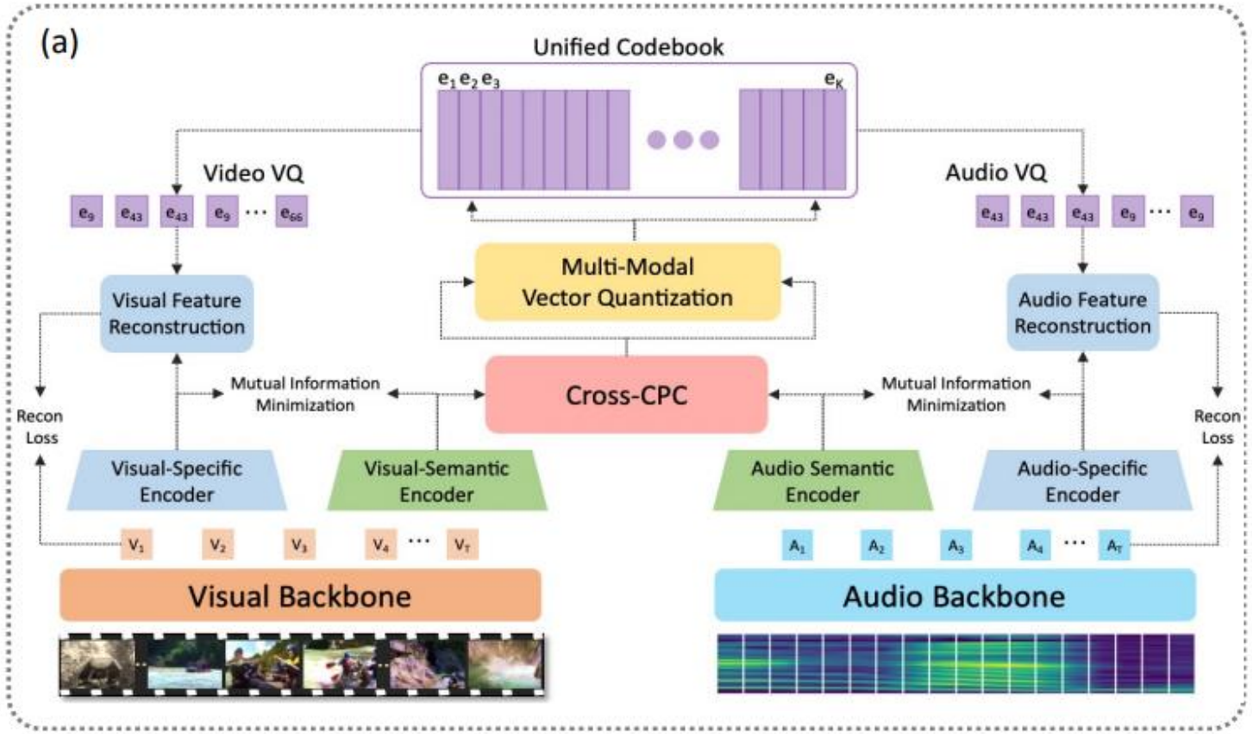
- **Multi-modal & Multi-domain**: Proposing an pretraining method for embedding learning with multimodal and multi-domain data
- **User-Specific Contrastive Learning**: Learning a user-specific projection to obtain different user views
- **Contrastive Learning across Domains:** Aligning item within domain and across domains

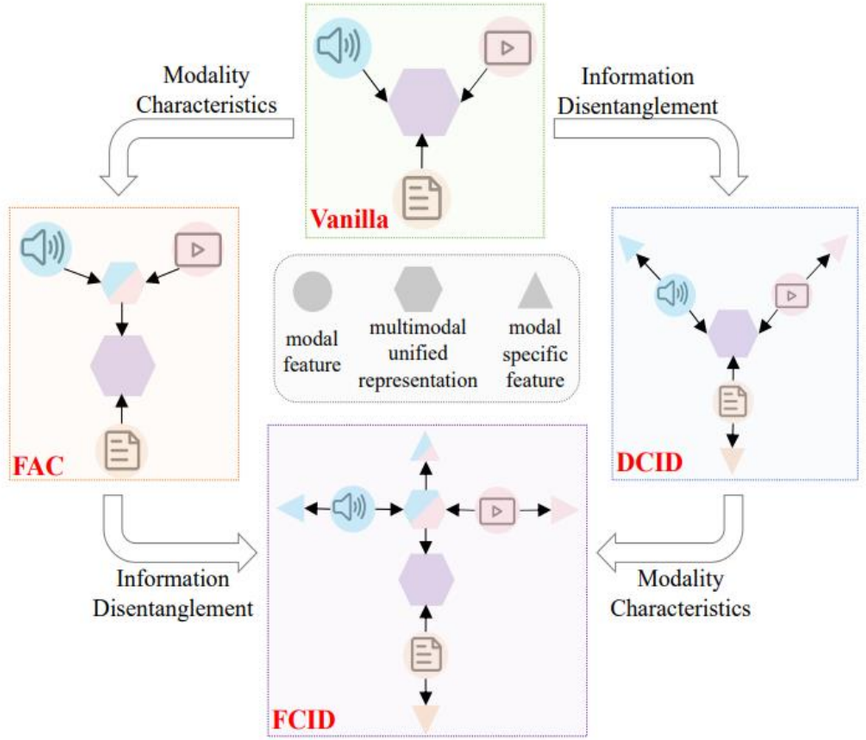
Dai et al., UniEmbedding: Learning Universal Multi-Modal Multi-Domain Item Embeddings via User-View Contrastive Learning, CIKM 2024

Outline

- **Generative IR**
  - Unified Multimodal Tokenizer
  - Generative Retrieval
  - Generative Recommendation

Uni-Code: Unified Multimodal Tokenizer



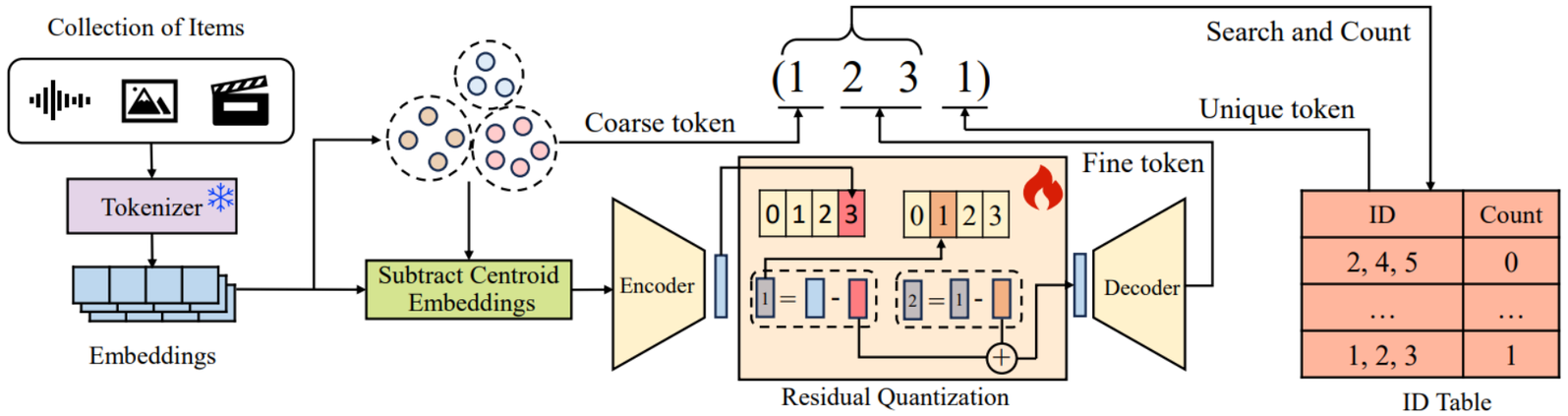


Multimodal Unified Representation:

- **Unified Codebook Learning:** Aligning all modalities into one codebook space and leverage it for quantization
- **Intra-modality Disentanglement and Inter-modality Alignment:** Leveraging Cross-CPC to align pairwise modalities and utilizing mutual information minimization for disentangling.

Xia et al., Achieving Cross Modal Generalization with Multimodal Unified Representation, NeurIPS 2023
Huang et al., Enhancing Multimodal Unified Representations for Cross Modal Generalization, ACL 2025

CART: Generative Cross-Modal Retrieval



Task	Method	R@1
Text-to-Image Retrieval	CLIP	55.9
	OpenCLIP	63.9
	MobileCLIP	74.9
	ONE-PEACE(Pretrained)	73.4
	ImageBind(Huge)	74.9
	LanguageBind(Image)	69.5
	CART(ours)	**81.8**

Task	Method	R@1
Text-to-Audio Retrieval	CLAP-HTSAT	16.1
	HTSAT-BERT	19.7
	LanguageBind(Audio)	16.3
	VALOR-B	17.5
	ONE-PEACE(Pretrained)	22.4
	CART(ours)	**46.4**

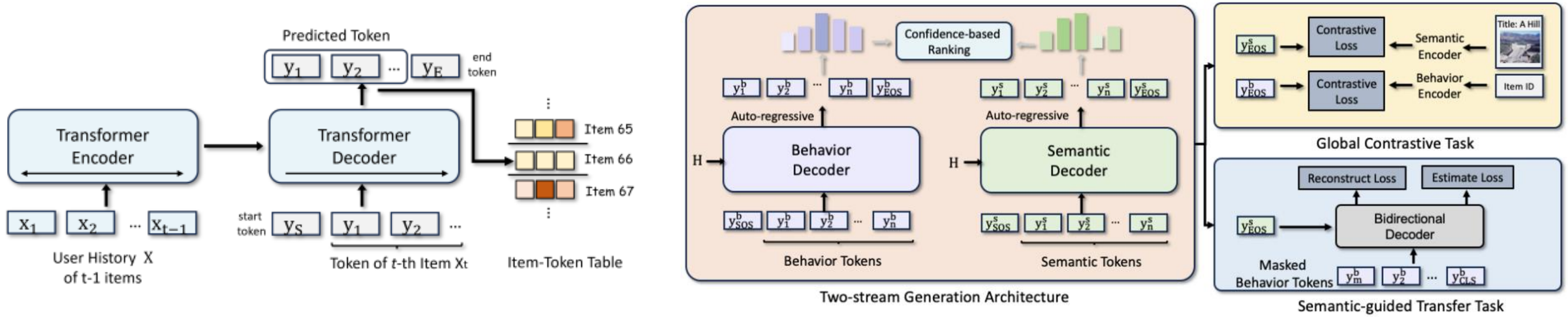
Task	Method	R@1
Text-to-Video Retrieval	CLIP2Video	45.6
	CLIP4Clip	44.5
	UMT-B	46.3
	Cap4Video	49.3
	ImageBind(Huge)	36.8
	LanguageBind(Video)	42.7
	CART(ours)	**52.6**

Generative Cross-Modal Retrieval:

- **Unified Discrete Encoding Module:** Encodes objects from different modalities into discrete token sequences using a shared codebook space.
- **Autoregressive Generation Module:** Takes the input text query tokens as input, encodes them, and decodes into a target discrete token sequence, which is then mapped to the corresponding retrieval object. The autoregressive model can flexibly adopt either a Decoder-Only or Encoder-Decoder architecture.
- **Training Module:** Trains both the unified discrete encoding module and the autoregressive generation module.

Fang et al., CART: A Generative Cross-Modal Retrieval Framework with Coarse-To-Fine Semantic Modeling, ACL 2025

EAGER: Encoder-Decoder Generative Recommendation

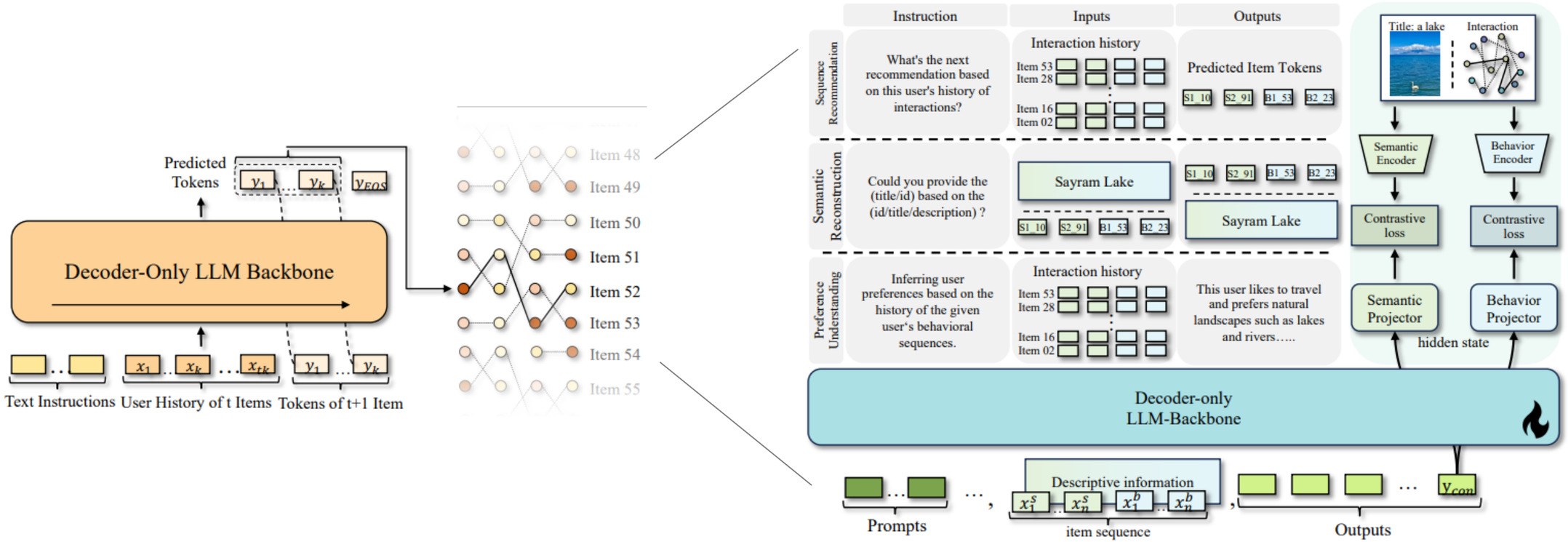


Dual-stream generative recommendation:

- **Item Code Encoding Module:** This module separately encodes item content and user behavior to generate two independent discrete code sequences — the item behavior code sequence and the item semantic code sequence.
- **Dual-Decoder Module:** A dual-decoder architecture that utilizes both the behavior code sequence and the semantic code sequence. It decodes based on behavioral information to generate the corresponding target item code sequence.
- **Auxiliary Training Tasks:** A variety of auxiliary training tasks are proposed, including global contrastive learning and content transfer training, to improve the overall model training performance.

Wang et al., EAGER: Two-Stream Generative Recommender with Behavior-Semantic Collaboration, KDD 2024

EAGER-LLM: Decoder-only Generative Recommendation

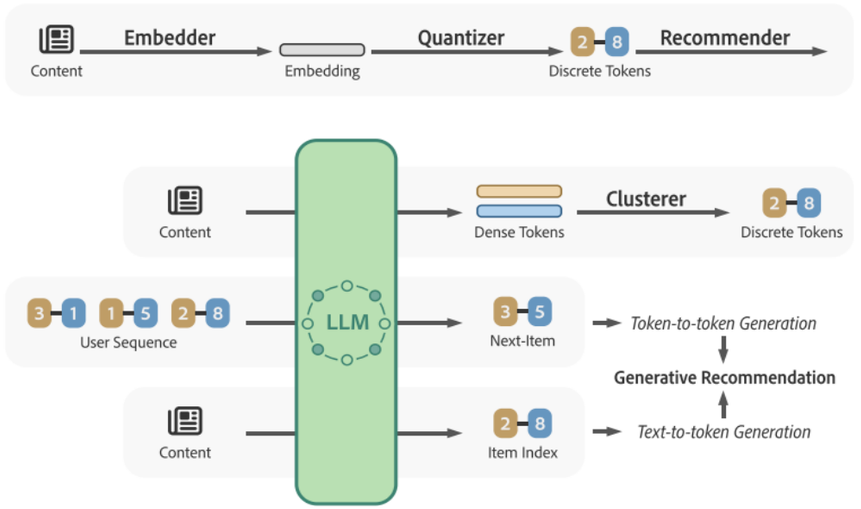


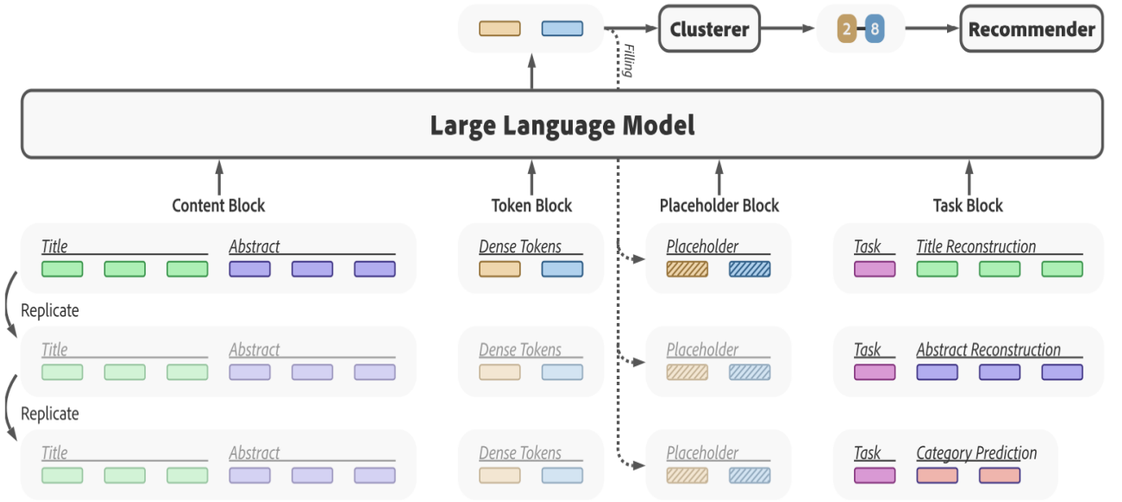
LLM-based decoder-only generative recommendation:

- **Preference Understanding:** Inferring user preferences based on the history of the given user's behavior sequences.
- **Semantic Reconstruction:** Providing the title or ID based on the ID/title/description.
- **Sequence Recommendation:** Predicting the next recommendation based on the user's history of interactions

Hong et al., EAGER-LLM: Enhancing Large Language Models as Recommenders through Exogenous Behavior-Semantic Integration, WWW 2025

STORE: Streamlined Tokenization and Generation





A unified language model-based generative framework that uses the same language model to handle both tokenization and generative recommendation tasks:

- **Text-to-Text:** Inferring the title of the next item based on the description of the given user's behavior sequences.
- **Text-to-Token:** Providing the tokens based on the item description.
- **Token-to-Token:** Predicting the next item based on the user's history of interactions

Liu et al., STORE: Streamlining Semantic Tokenization and Generative Recommendation with A Single LLM, Arxiv 2024

Outline

- **MIRC Challenge**
  - Text2Doc Retrieval
  - Multimodal CTR Prediction

Track 1: Multimodal Document Retrieval

Multimodal Document Retrieval Challenge ERel@MIR

Multimodal Document Retrieval Challenge Track at the 1st ERel@MIR Workshop at WWW

Overview | Data | Code | Models | Discussion | Leaderboard | Rules

Overview

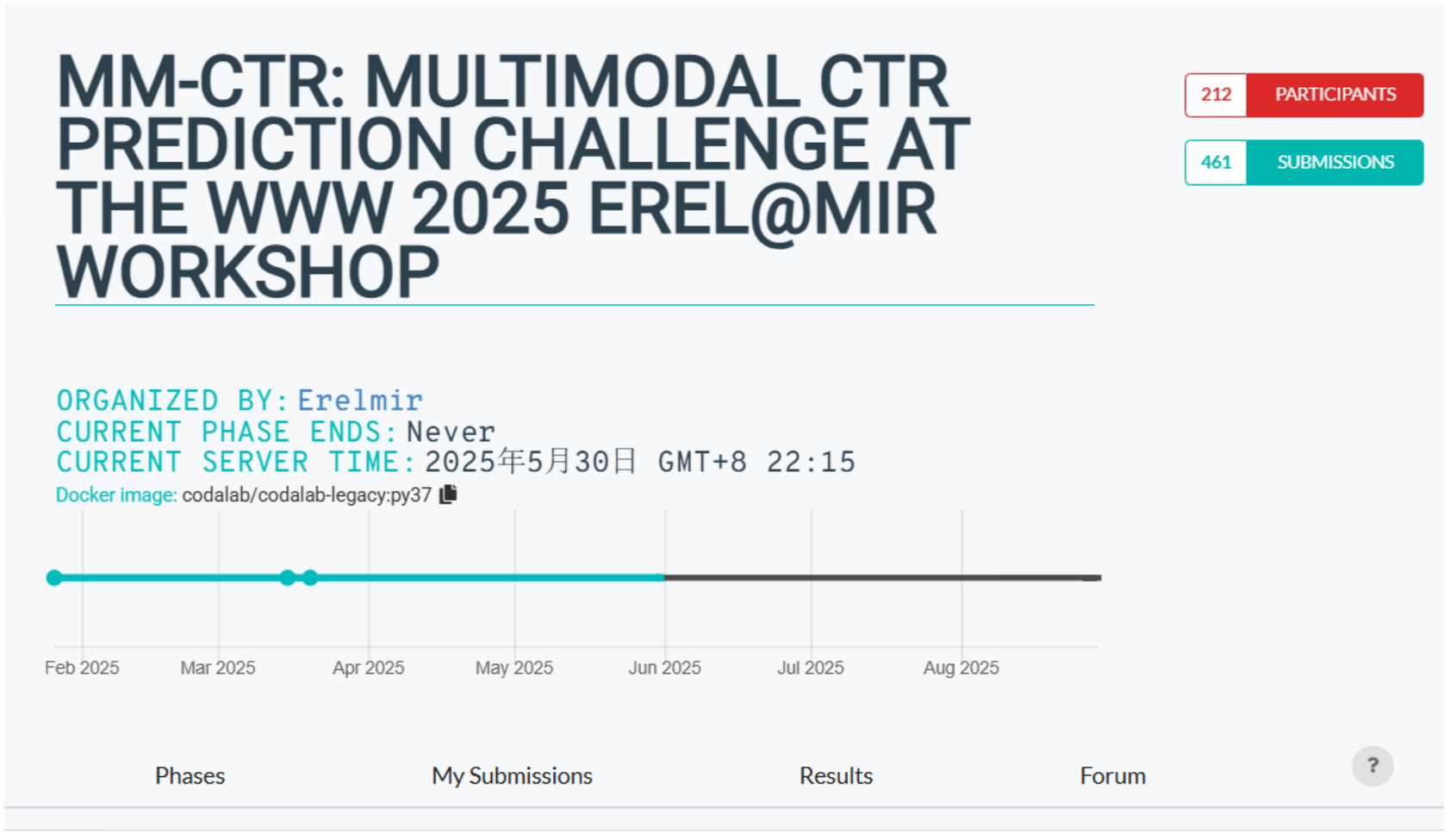
Currently, the majority of retrieval-augmented generation (RAG) models are designed to retrieve relevant text documents based on user queries. However, in real-world scenarios, users often need to retrieve multimodal documents, including text, images, tables, and charts. This requires a more sophisticated retrieval system that can handle multimodal information and provide relevant documents or passages based on user queries. Retrieving multimodal documents will help AI chatbots, search engines, and other applications provide more accurate and relevant information to users.

The **Multimodal Document Retrieval Task** focuses on modeling passages from multimodal documents or web pages, leveraging textual and multimodal information for embedding modeling. The ultimate goal is to retrieve the relevant multimodal document or passage based on the user's text or multimodal query.

The challenge website can be accessed from: https://erel-mir.github.io/challenge/mdr-track1/.

https://www.kaggle.com/competitions/multimodal-document-retrieval-challenge

Track 2: Multimodal CTR Prediction



MM-CTR: Multimodal CTR Prediction Challenge

https://www.codabench.org/competitions/5372/

Summary

Track 1 MMDocIR	Track 2 MM-CTR
41 Participants	176 Participants
586 Submissions	428 Submissions

Track 1 Winners-Multimodal Document Retrieval Challenge Track

Rank/Award	Team Name	Code Link
1st Place	iLearn	https://github.com/hbhalpha/MDR
2nd Place	LLMHunter	https://github.com/i2vec/MMDocRetrievalChallenge
3rd Place	GPU is all you need	https://github.com/bargav25/MultiModal_InformationRetrieval

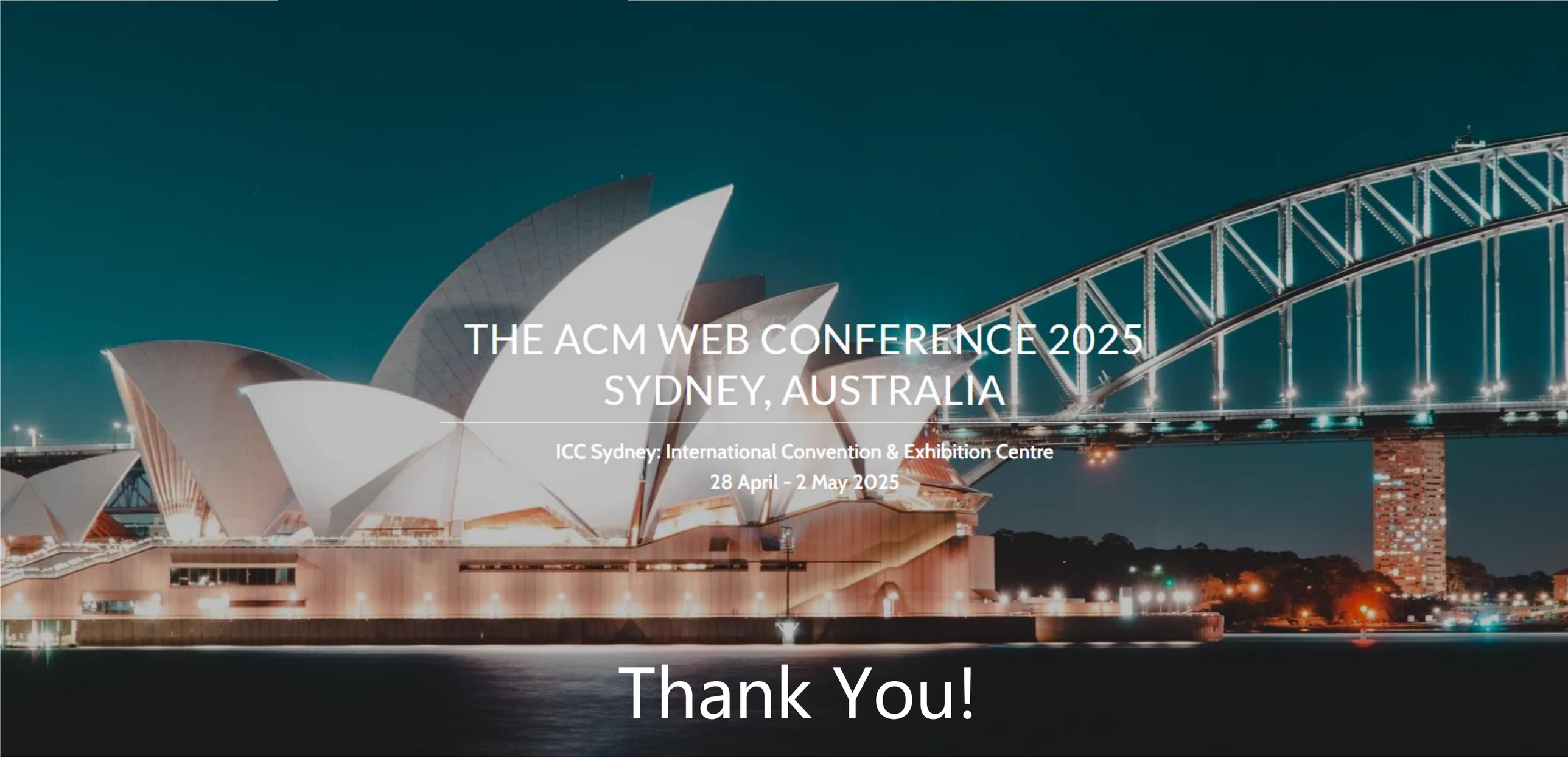
Track 2 Winners-Multimodal CTR Prediction Challenge Track

Rank/Award	Team Name	Code Link
1st Place	momo	https://github.com/pinskyrobin/WWW2025_MMCTR
2nd Place	DISCO.AHU	https://github.com/salmon1802/QIN
3rd Place	jzzx.NTU	https://github.com/ZiaoGao/MMCTR
Task 1 Winner	delorean	https://github.com/AdamLTy/mmctr-solusion-by-delorean
Task 2 Winner	zhou123	https://github.com/Lattice-zjj/MMCTR_Code

We are hiring!

- **Full-time jobs**
  - Shenzhen, Beijing, Shanghai, Singapore…
- **Research interns**
  - Recommendation & search
  - Multimodal models
  - RAG/Agents

Welcome to reach out via jiemingzhu@ieee.org



Thank You!

HUAWEI