Multimodal Pre-training, Adaptation, and Generation for Recommendation

Qijiong Liu, Jieming Zhu, Yanting Yang, Quanyu Dai, Zhaocheng Du, Xiao-Ming Wu, Zhou Zhao, Rui Zhang, and Zhenhua Dong

KDD 2024 @ Barcelona

THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學
HUAWEI
ZHEJIANG UNIVERSITY
HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Tutorial Speakers

Dr. Jieming ZHU

Huawei Noah's Ark Lab

Mr. Qijiong LIU

The HK PolyU

Prof. Xiao-Ming WU

The HK PolyU

Prof. Rui ZHANG

HUST (www.ruizhang.info)

Tutorial Schedule

Time	Event	Speaker
10:00 AM - 10:40 AM	Multimodal Representation Pretraining and Adaptation for Recommendation [slides]	Jieming Zhu
10:40 AM - 11:00 AM	Coffee Break	
11:00 AM - 11:40 AM	Multimodal Generation for Recommendation [slides]	Rui Zhang
11:40 AM - 12:20 PM	Enhancing Multimodal Retrieval and Generation with Unified Vision-Language Models	Xiao-Ming Wu
12:20 PM - 13:00 PM	Benchmarking Recommendation Ability of Foundation Models: Legommenders and RecBench	Qijiong Liu

Tutorial Materials

Tutorial: https://mmrec.github.io/tutorial/kdd2024/
Survey: https://arxiv.org/abs/2404.00621

Multimodal Representation Pre-training and Adaptation for Recommendation

Jieming Zhu
Huawei Noah's Ark Lab
https://jiemingzhu.github.io

KDD 2024 @ Barcelona

Multimodal models are reshaping the world
with new understanding, creation, and interaction capabilities

Multimodal Understanding

Image captioning

Input Image

Human Describe this image as detailed as possible.

BLIP-2 A city street with many people walking and riding motorcycles.

MiniGPT-4 The image shows a busy city street with a clock tower in the background. The street is lined with shops and restaurants on both sides. There are several motorcycles parked on the side of the road, and people are walking down the street. The clock tower has a large clock face with Roman numerals and a small spire on top. The buildings on either side of the street have ornate facades and balconies. The street is paved with cobblestones and there are streetlights on either side of the road. The sky is clear and blue, and there are a few clouds in the distance.

Image credit: Zhu et al., MiniGPT-4…, 2023

Visual question answering

GPT-4 visual input example, Extreme Ironing:

User What is unusual about this image?

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

Image credit: OpenAI et al., GPT-4 Technical Report, 2023

Cross-Modal Retrieval

Text-to-image retrieval

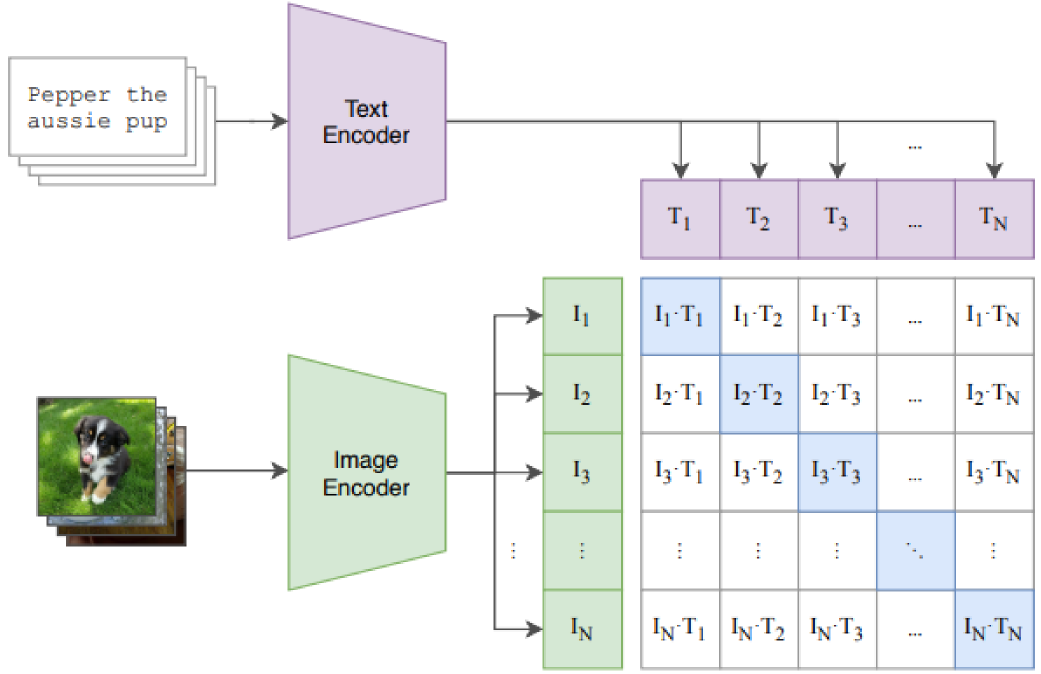


Image credit:: Radford et al., Learning Transferable Visual Models From Natural Language Supervision, 2021

Image-to-audio retrieval

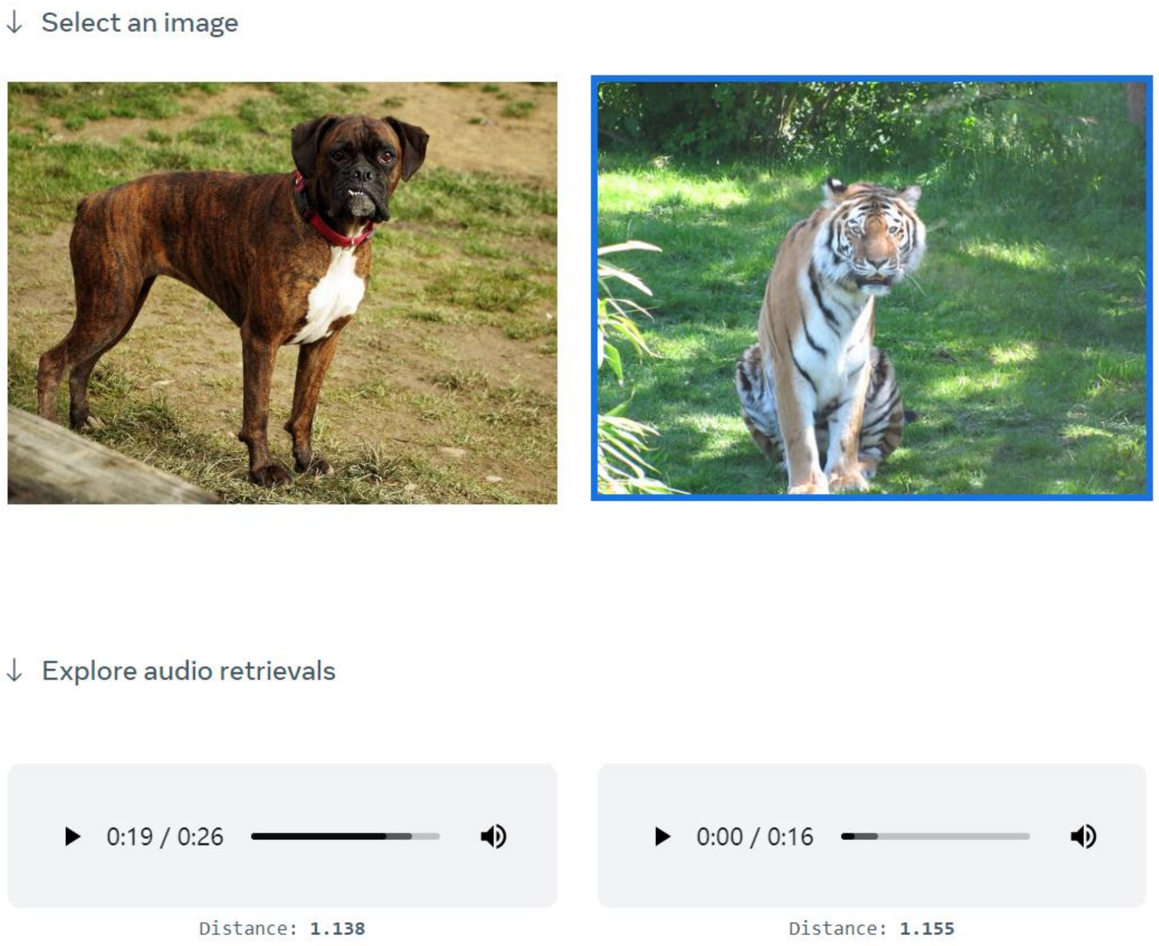


Image credit: Girdhar et al., ImageBind: One Embedding Space To Bind Them All, 2023

Multimodal Generation

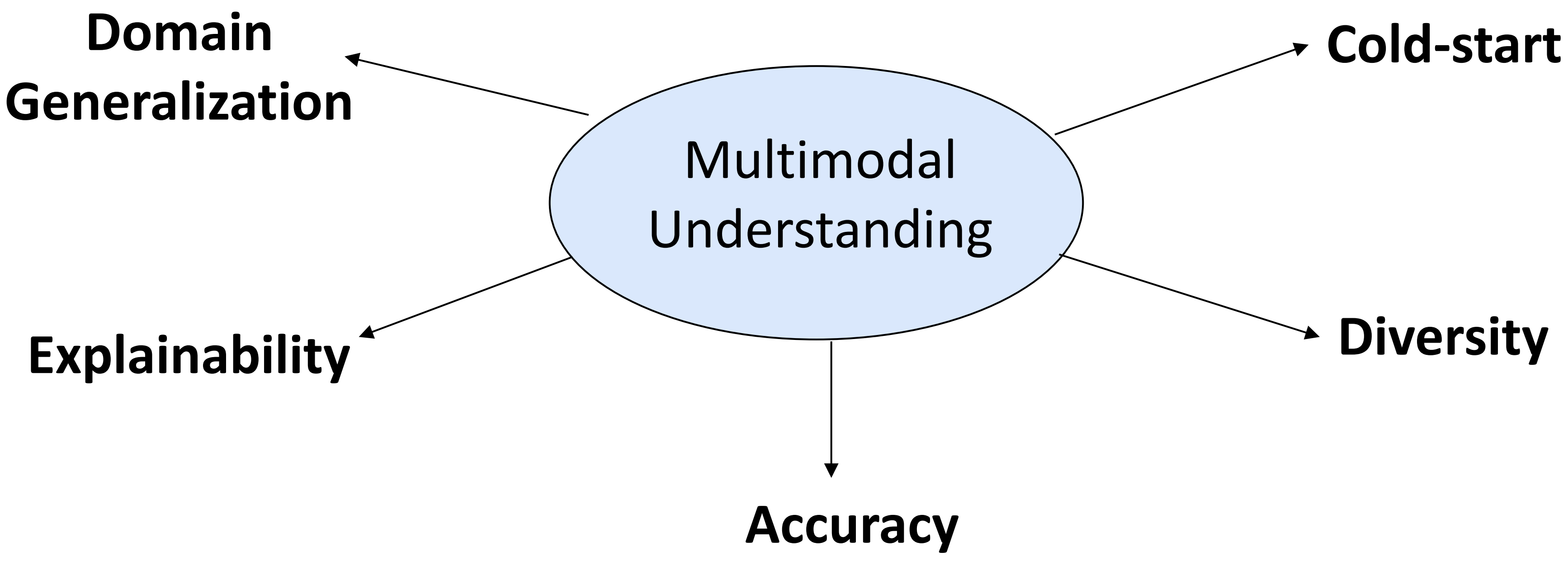
Text-to-image generation

Image credit: Stable Diffusion 2. https://github.com/Stability-AI/stablediffusion

Text-to-video generation

Image credit: https://openai.com/index/video-generation-models-as-world-simulators

What Can Multimodal Models Offer for Recommendation?



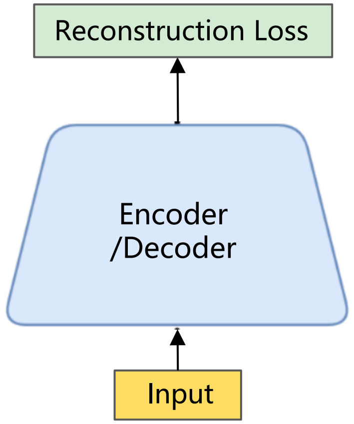
Outline

- **Multimodal Pretraining Techniques for Recommendation**
  - Pretraining paradigms
  - User-to-item matching
  - Item-to-item matching
  - ID-to-modality alignment
- **Multimodal Adaptation Techniques for Recommendation**
  - Representation transfer
  - Joint Finetuning
  - Adapter/Prompt tuning
- **Open Challenges**

Self-supervised Pretraining

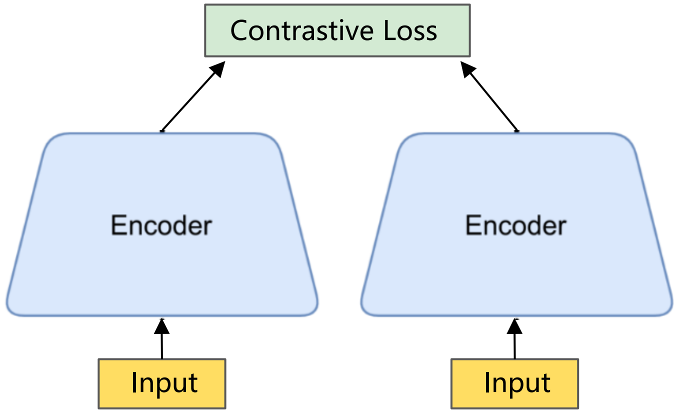
Representative pretraining tasks: reconstructive, contrastive, and generative

BERT
StableDiffusion



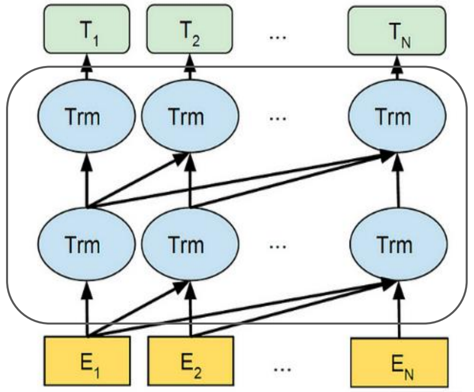
Reconstructive

CLIP
ImageBind



Contrastive

GPT
DALL-E



Generative

Pretrained Multimodal Models (2019~2022)

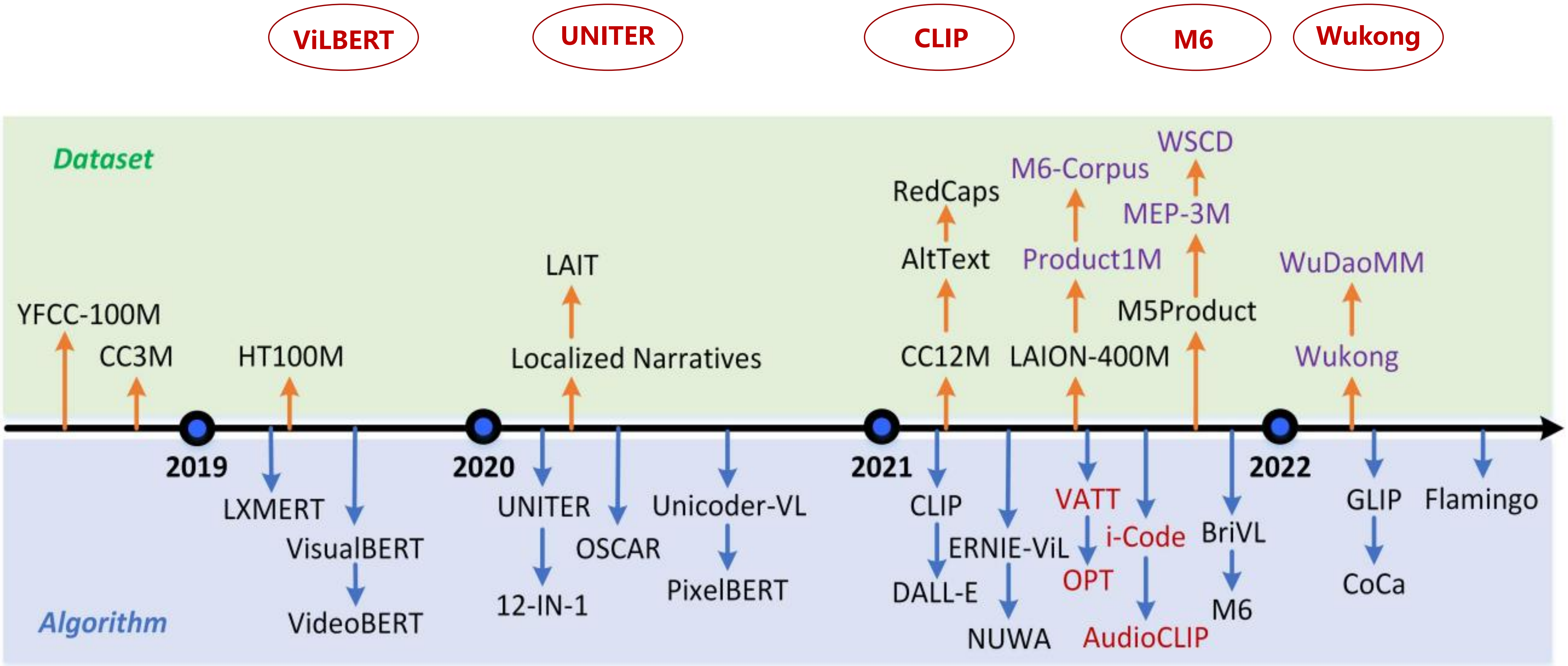


Image credit: Wang et al., Large-scale Multi-Modal Pre-trained Models: A Comprehensive Survey, 2023

Multimodal Large Language Models (2023~2024)

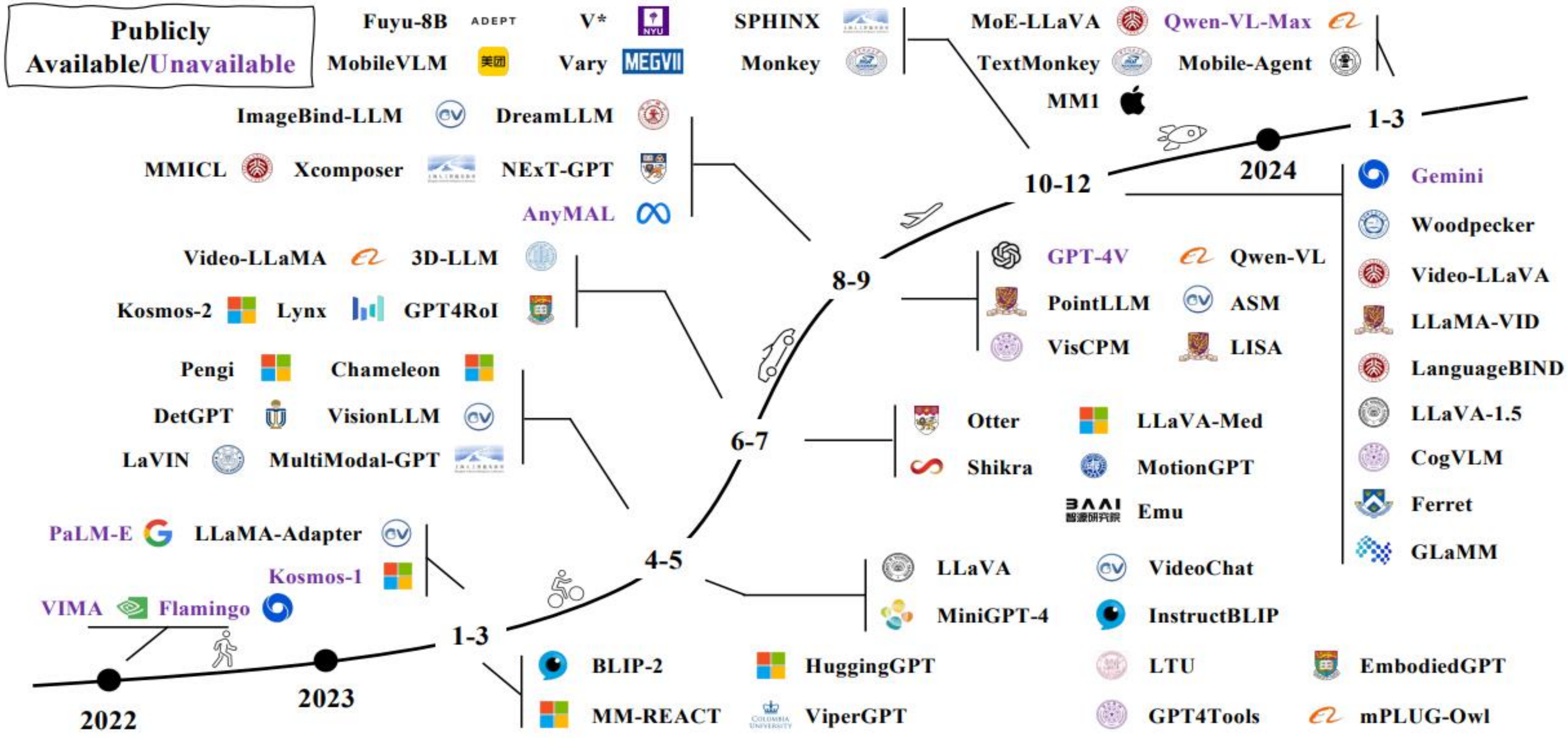
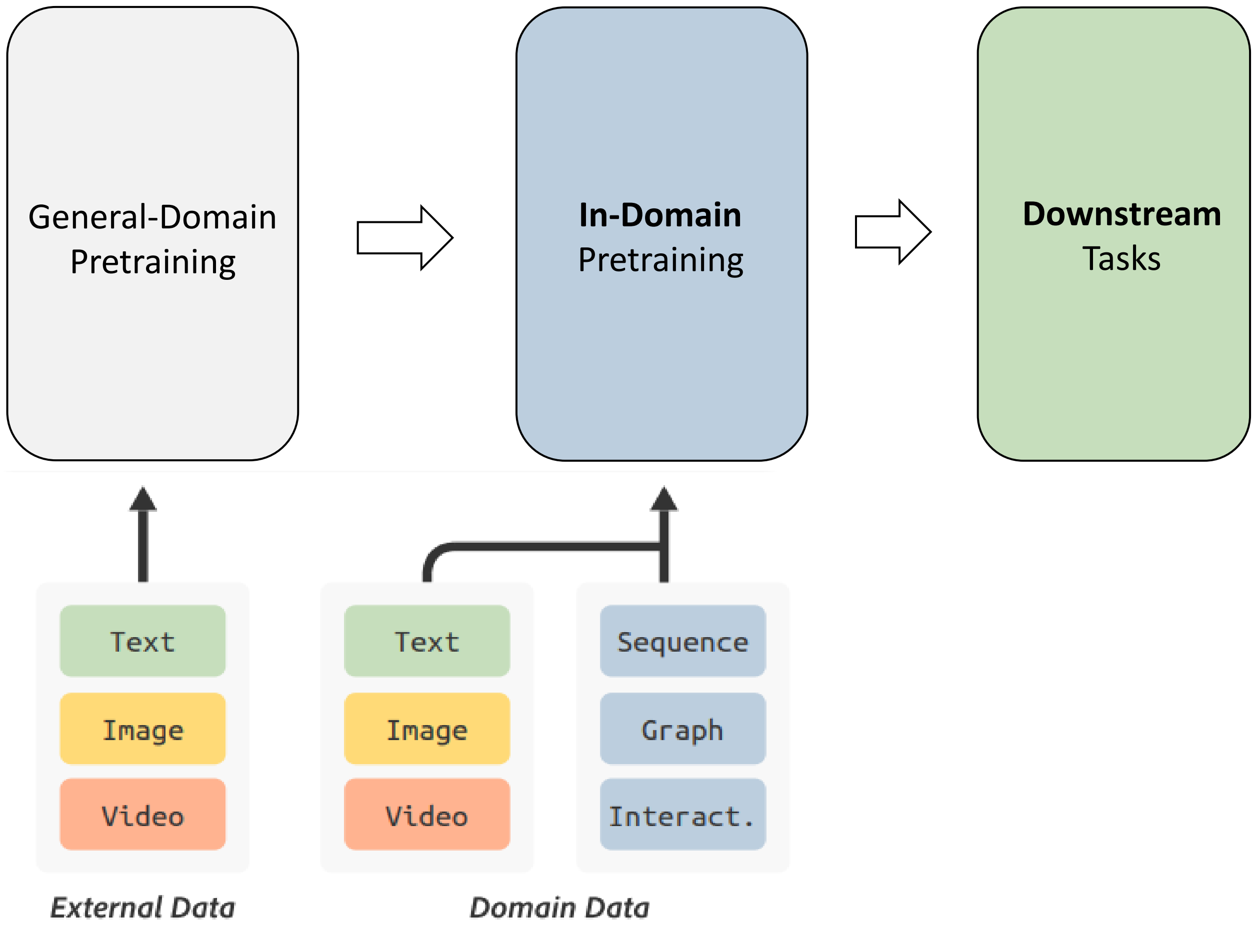


Image credit: Yin et al., A Survey on Multimodal Large Language Models, 2023

Overall Framework



Domain-Specific Data:

- **Content data**: text、image、video、category、tags…
- **Behavior data**: user-item pairs, sequences、graphs…

In-Domain Pretraining:

- A way to incorporate domain knowledge into a pretrained model without having to retrain it from scratch

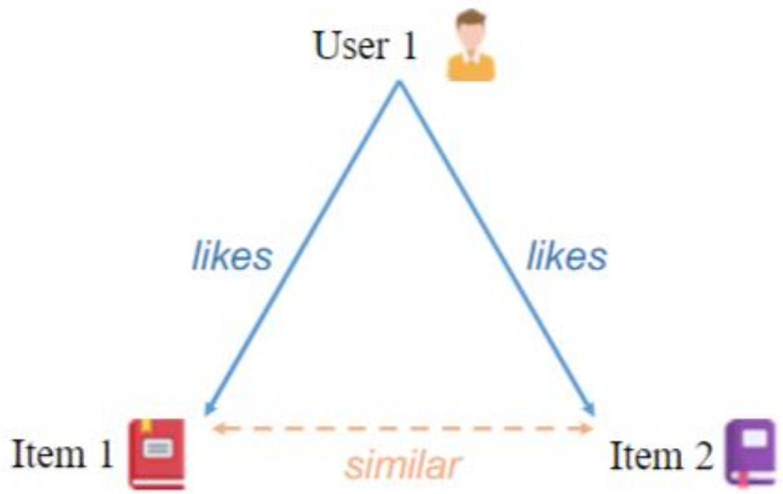
Downstream Tasks

- Representation transfer
- Supervised finetuning
- Adapter/Prompt tuning

In-domain Pretraining w/ Collaborative Signals

Collaborative Signals:

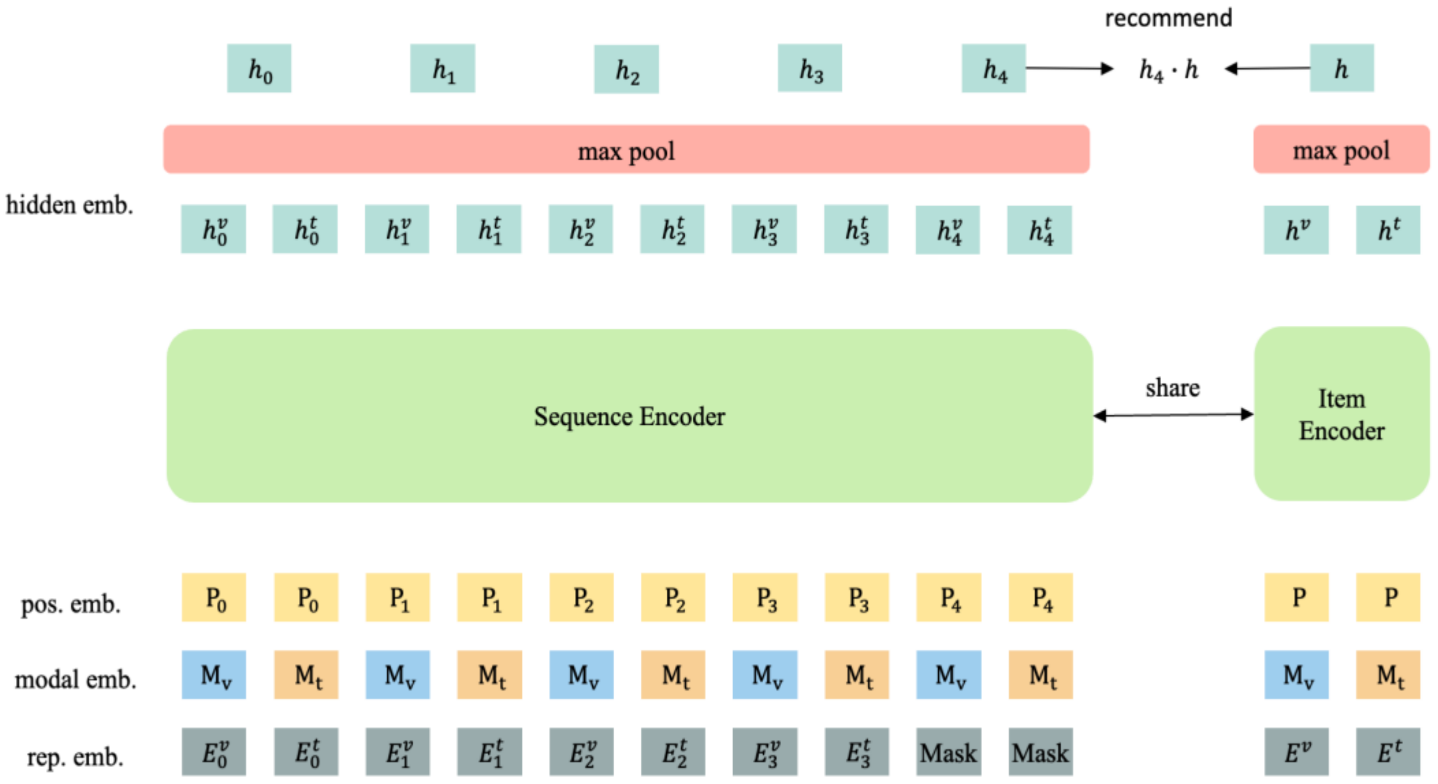
- User-to-item matching (e.g., MMSRec、MISSRec)
- Item-to-item matching (e.g., CB2CF、ItemSage)
- ID-to-modality alignment (e.g., CLCRec)
- ……

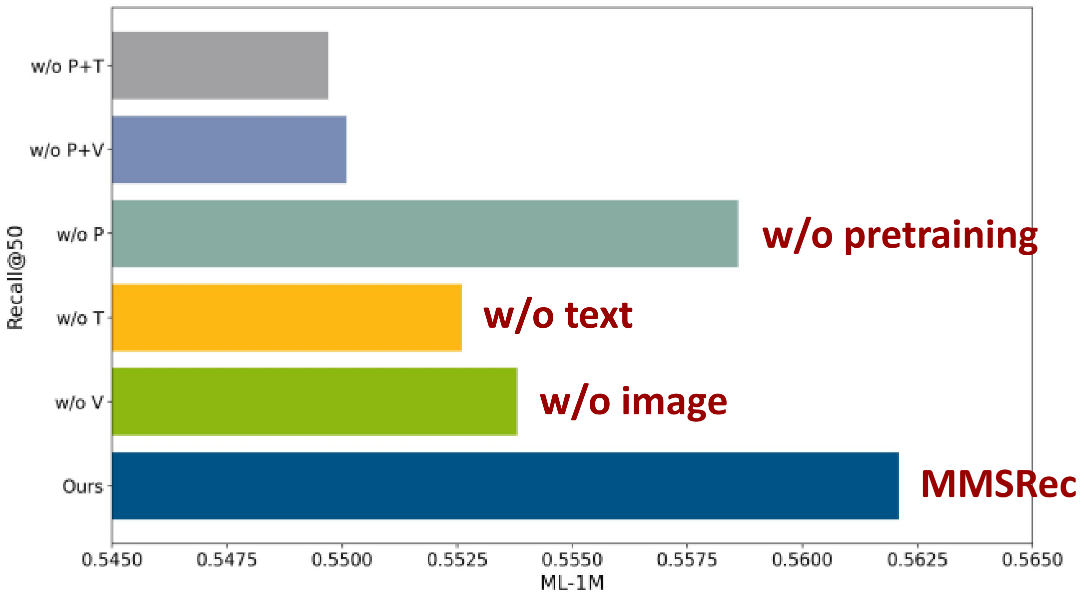


In-domain Pretraining w/ User-to-Item Matching

MMSRec

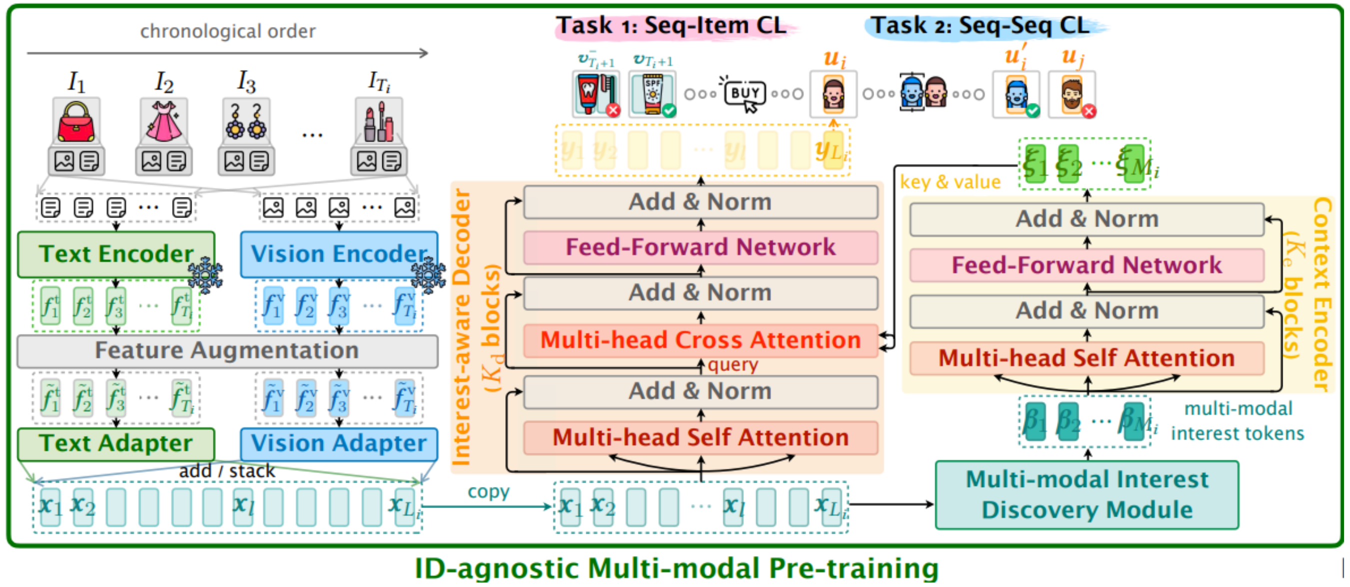
➢ **Masked Item Prediction**: mask the last position of the behavior sequence for next item prediction





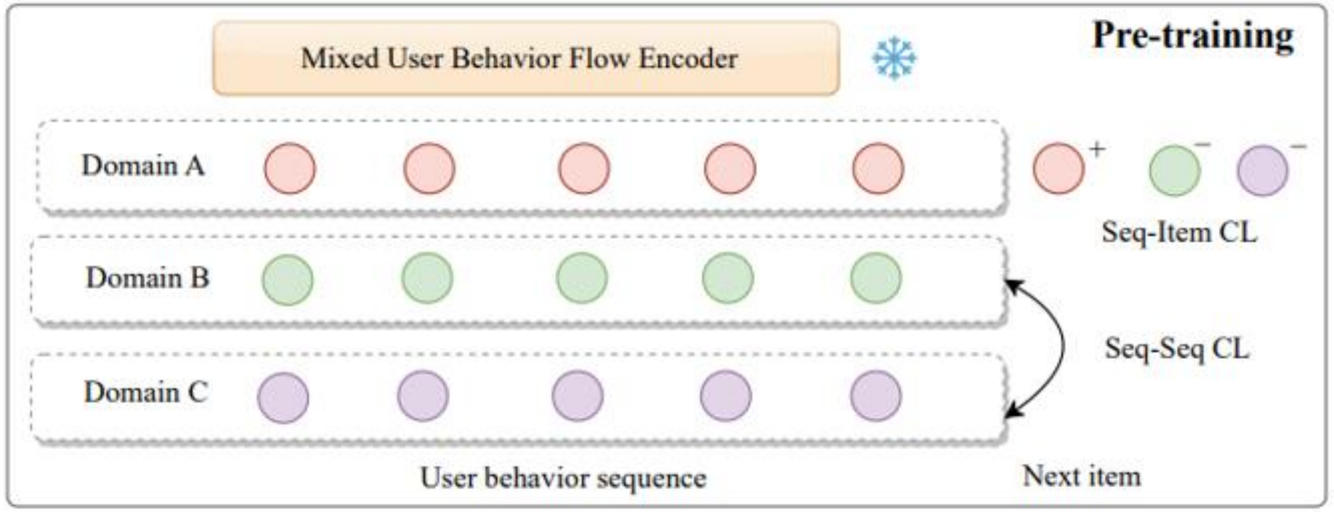
Song et al., Self-Supervised Multi-Modal Sequential Recommendation, 2023

In-domain Pretraining w/ User-to-Item Matching (continued)



MISSRec [Huawei]

- **Seq-item contrastive learning**: matching between user sequence and masked item
- **Multi-interest matching**: grouping interest prototypes and performing interest-aware sequence modeling



UniM2Rec [Tencent]

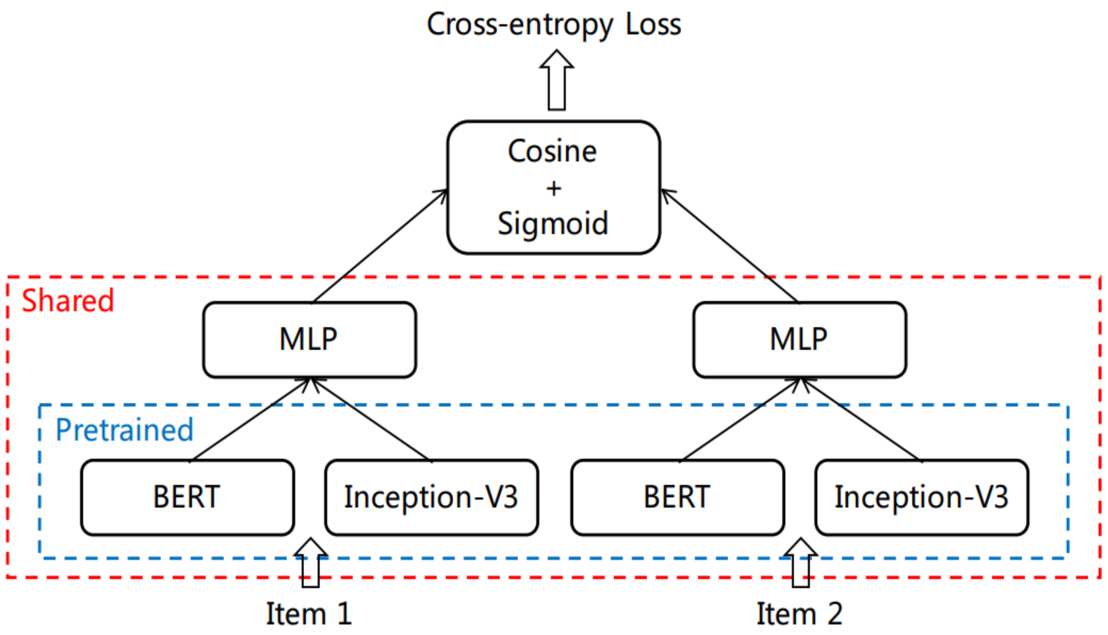
- **Seq-item contrastive learning**: matching between user sequence and masked item
- **Multi-domain matching**: matching across domains

Wang et al., MISSRec: Pre-training and Transferring Multi-modal Interest-aware Sequence Representation for Recommendation, 2023
Sun et al., Universal Multi-modal Multi-domain Pre-trained Recommendation, 2023

In-domain Pretraining w/ Item-to-Item Matching

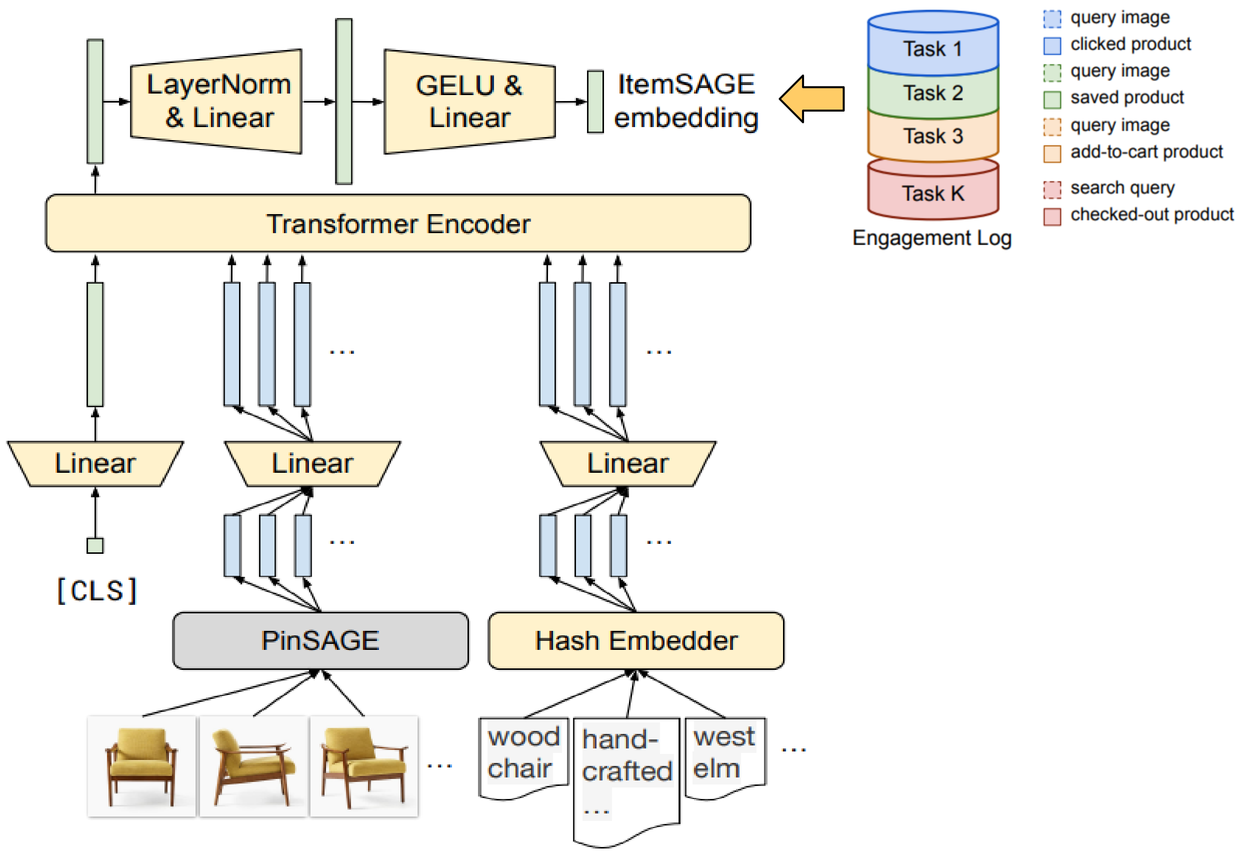
SSD [Xiaohongshu]

- **CB2CF**: leveraging item-to-item towers to pretrain multimodal item embeddings for recommendation diversity



ItemSage [Pinterest]

- Applying all engagement logs with multi-tasks to train multimodal item embeddings

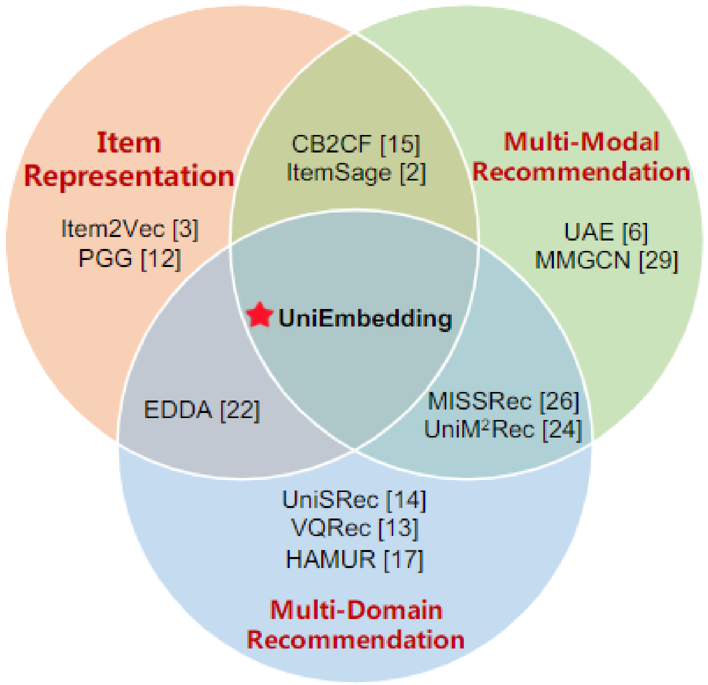


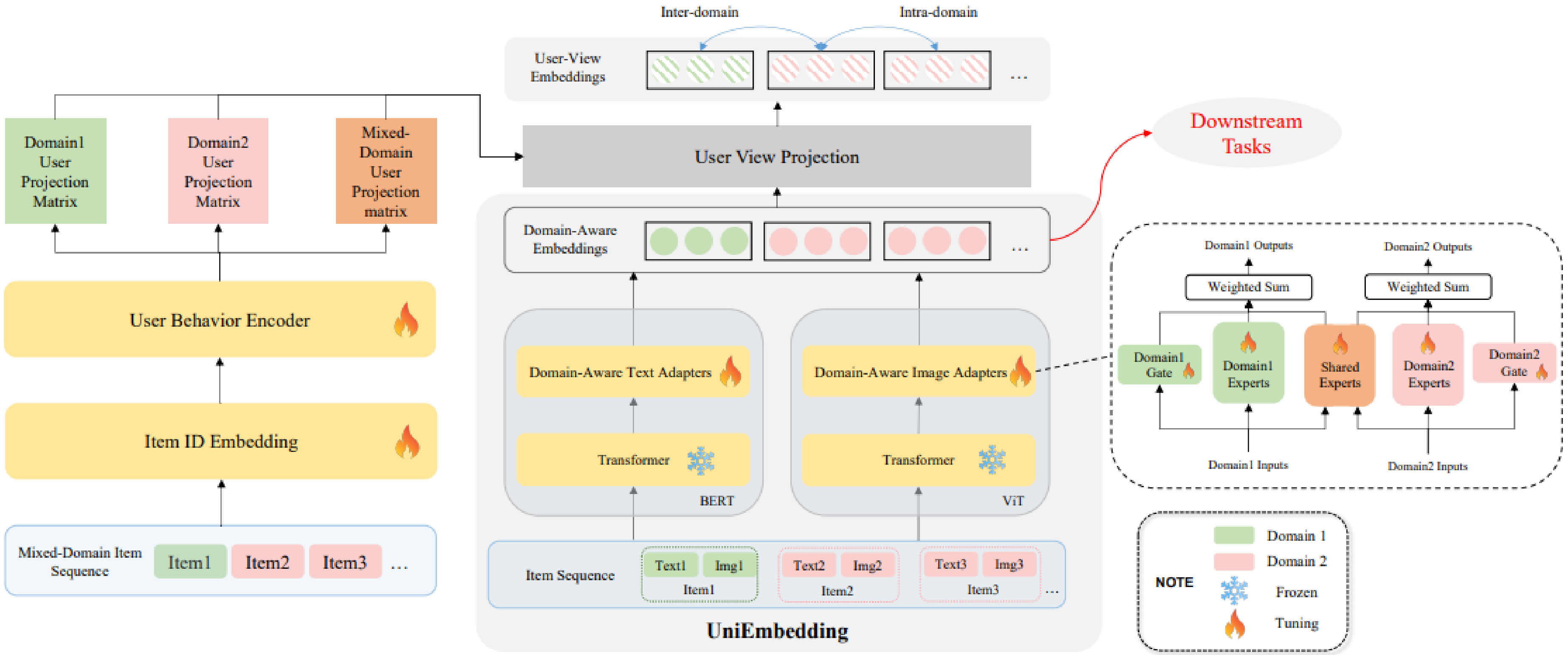
Huang et al., Sliding Spectrum Decomposition for Diversified Recommendation, 2021.
Baltescu et al., ItemSage: Learning Product Embeddings for Shopping Recommendations at Pinterest, 2022

In-domain Pretraining w/ Item-to-Item Matching

UniEmbedding [Huawei]

- Universal multi-modal multi-domain item embedding
- Intra-domain/cross-domain contrastive learning
- Domain-aware multimodal adapter



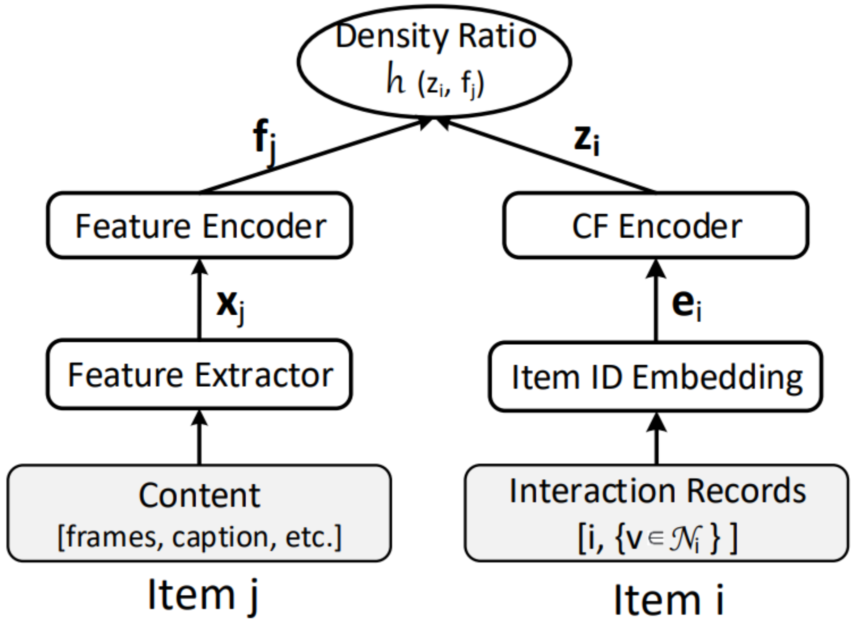


Dai et al., UniEmbedding: Learning Universal Multi-Modal Multi-Domain Item Embeddings via User-View Contrastive Learning, CIKM 2024

In-domain Pretraining w/ ID-to-Modality Alignment

CLCRec

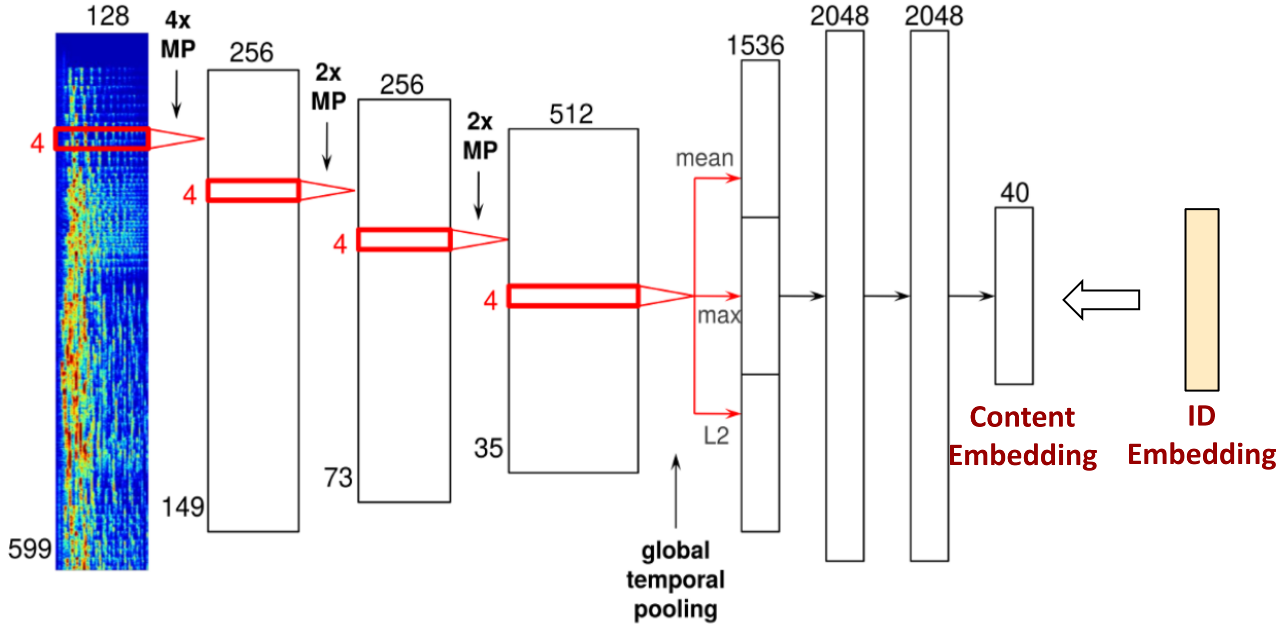
- Contrastive learning between CF encoder and content encoder



Wei et al., Contrastive Learning for Cold-Start Recommendation, 2021

DCMR

- Leveraging pretrained item ID embeddings as signals to train content encoders



Oord et al., Deep Content-based Music Recommendation, 2013
https://sander.ai/2014/08/05/spotify-cnns.html

In-domain Pretraining w/ Self-supervised Tasks

Reconstructive:

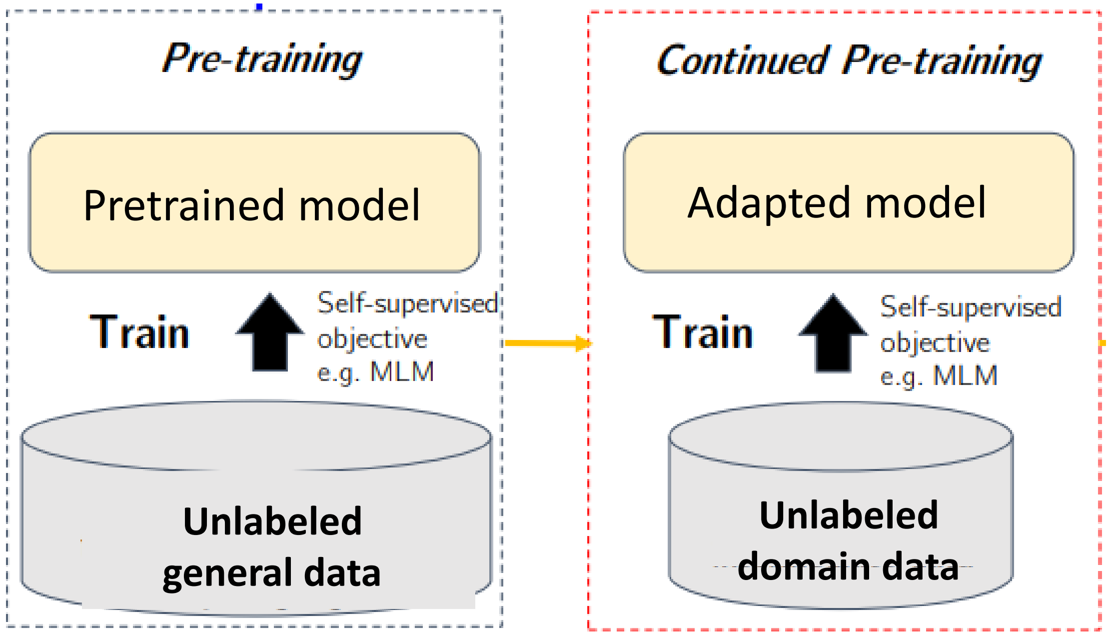
- Masked token prediction (PREC、RecoBERT)
- Masked attribute prediction (PREC、S3-Rec)
- Masked item prediction (PREC、S3-Rec)

Contrastive:

- Contrastive view alignment (MMSRec、VisualEncoder)
- Title-body alignment (RecoBERT)
- Cross-modal alignment

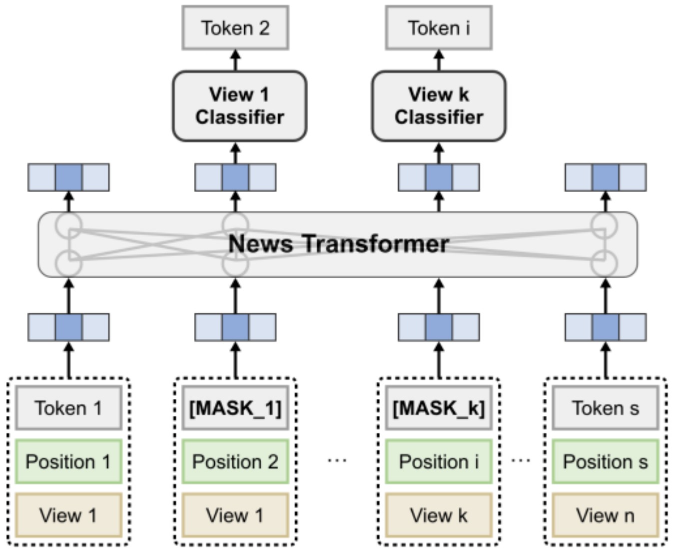
Generative:

- ……



Liu et al., Boosting Deep CTR Prediction with a Plug-and-Play Pre-trainer for News Recommendation, 2022
Malkiel et al., RecoBERT: A Catalog Language Model for Text-Based Recommendations, 2020
Zhou et al., S^3-Rec: Self-Supervised Learning for Sequential Recommendation with Mutual Information Maximization, 2020
Song et al., Self-Supervised Multi-Modal Sequential Recommendation, 2023
Chen et al., Visual Encoding and Debiasing for CTR Prediction, 2022

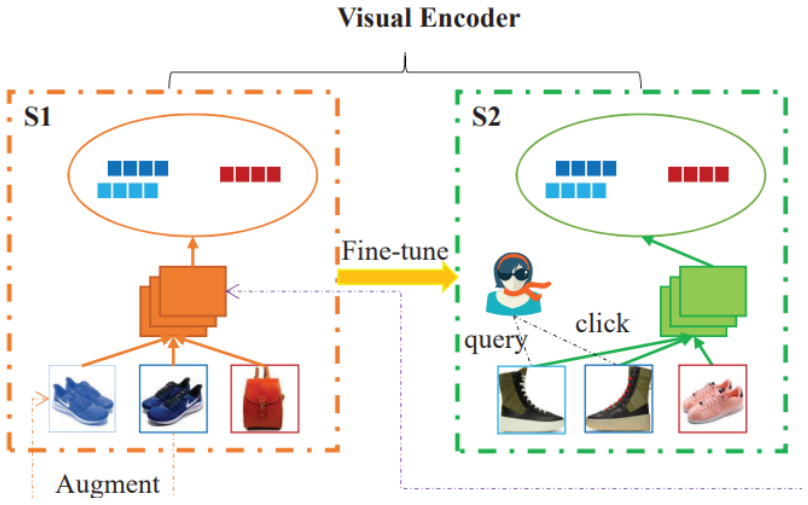
In-domain Pretraining w/ Self-supervised Tasks (continued)



(a) Masked news token prediction.

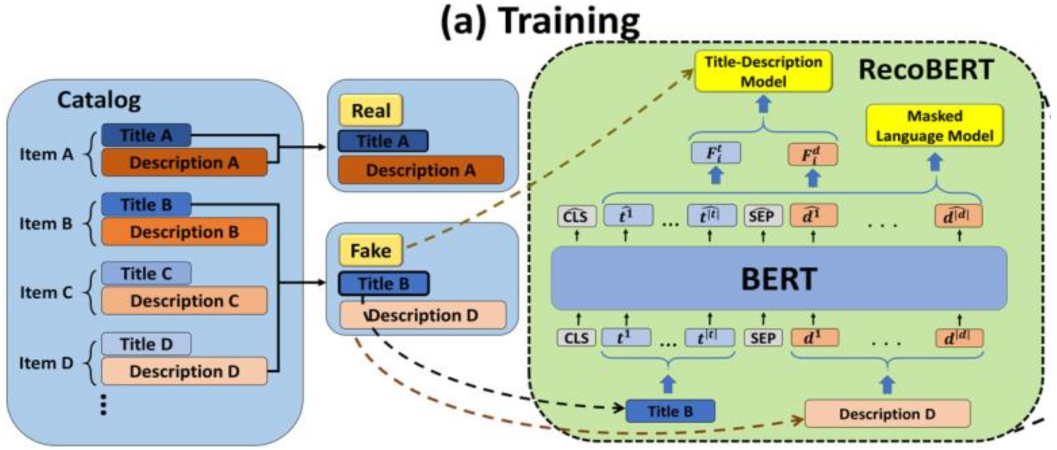
Masked token prediction

PREC [Huawei] *Liu et al., 2022*



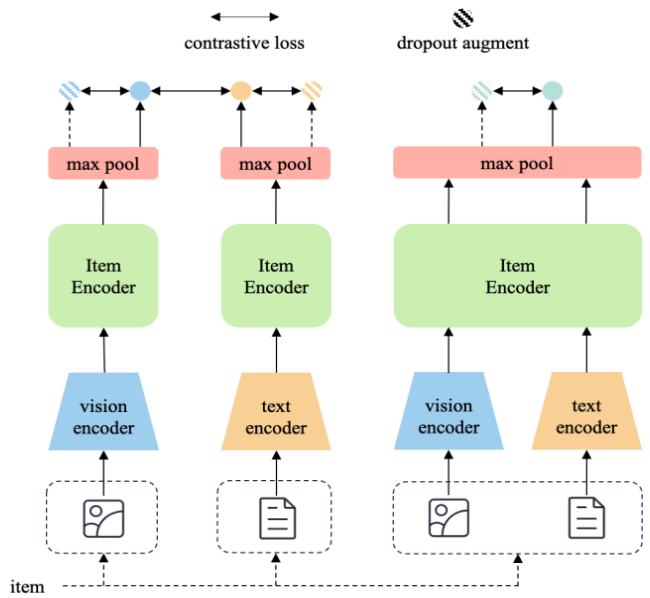
Contrastive learning

VisualEncoder *Chen et al., 2022*



Masked language model / alignment between title and description

RecoBERT Malkiel *et al., 2020*



Contrastive learning within modality & across modalities

MMSRec *Song et al., 2023*

In-domain Pretraining w/ Augmented Knowledge

Knowledge Sources:

- Category/tags/topics
- Knowledge graphs
- Query-document pairs
- ……

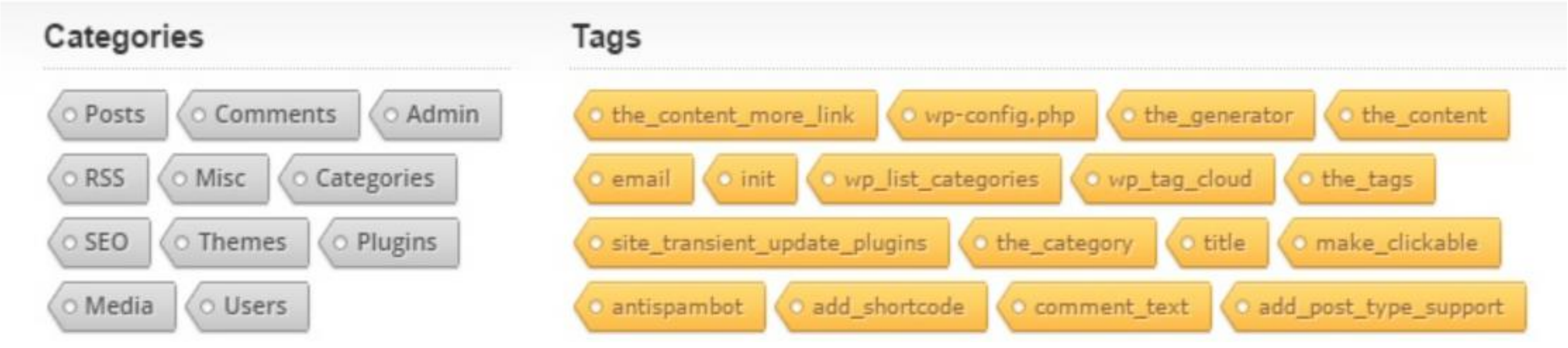


Image credit: https://www.wpexplorer.com/plugins-add-categories-tags/

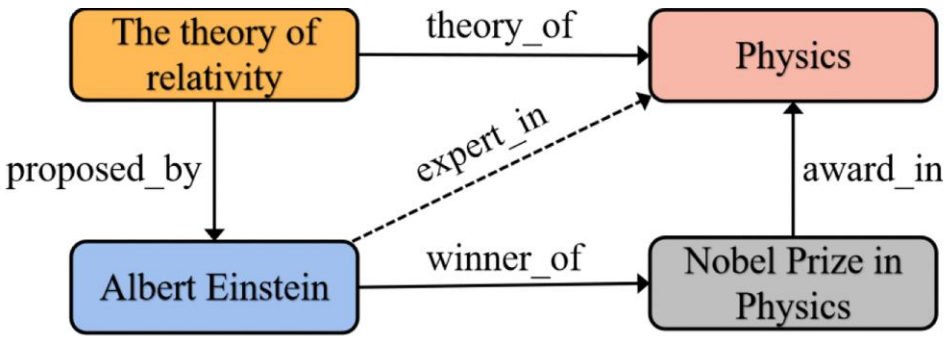
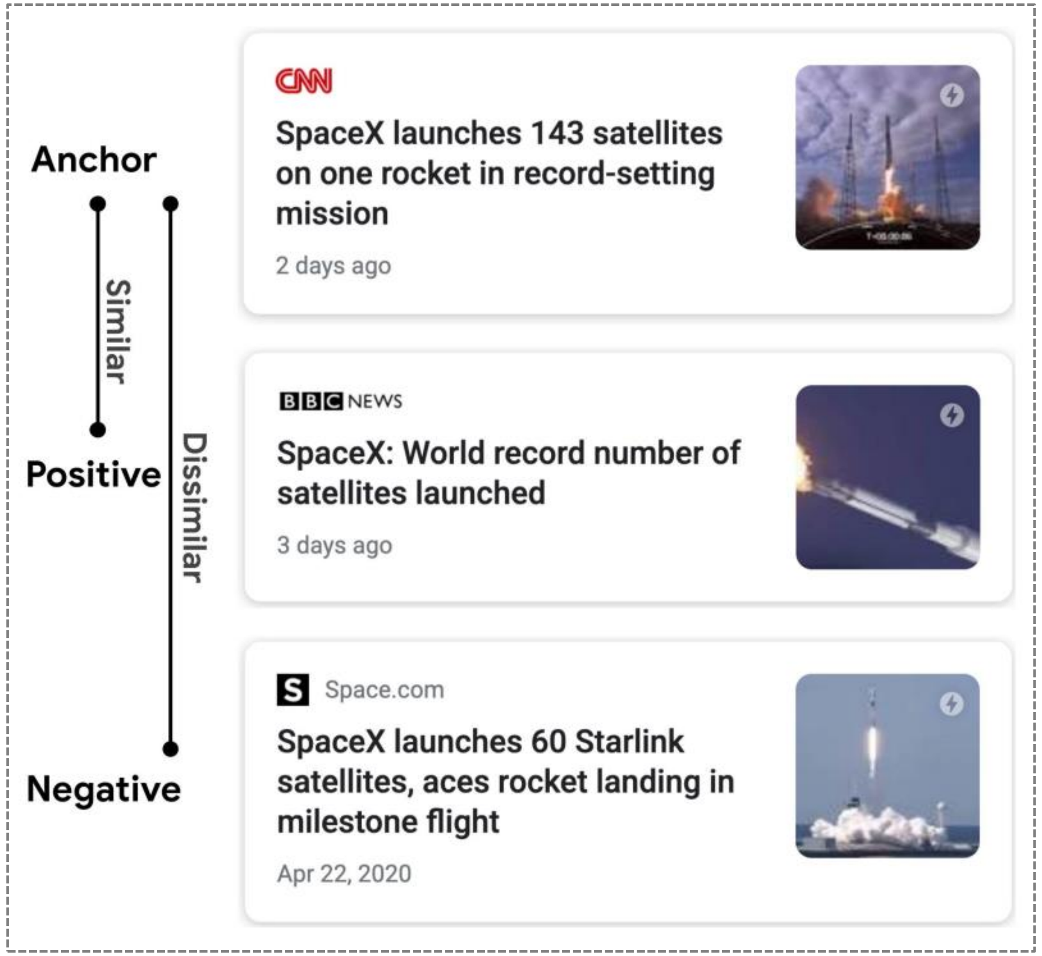


Image credit: TDN: Triplet Distributor Network for Knowledge Graph Completion, 2023

In-domain Pretraining w/ Augmented Knowledge (continued)

NewsEmbed

- **Contrastive learning**: aligning the crawled news document triplets
- **Topic classification**: leveraging document-topic associations for multi-label classification



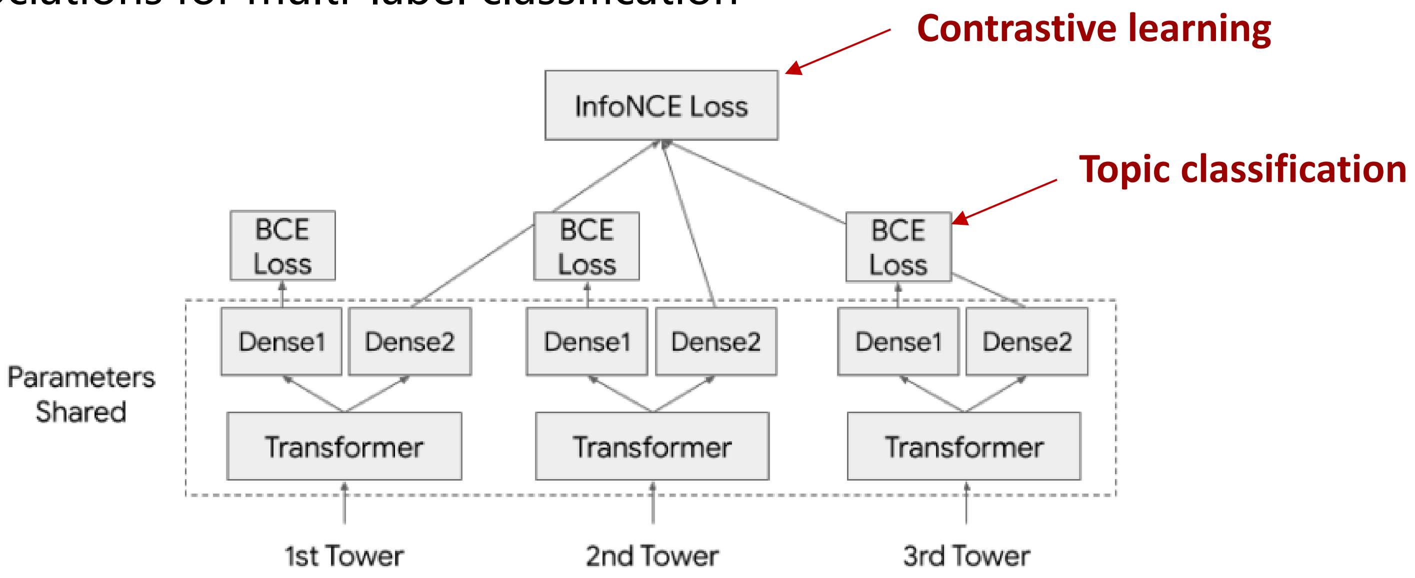


Figure 4: NewsEmbed model structure.

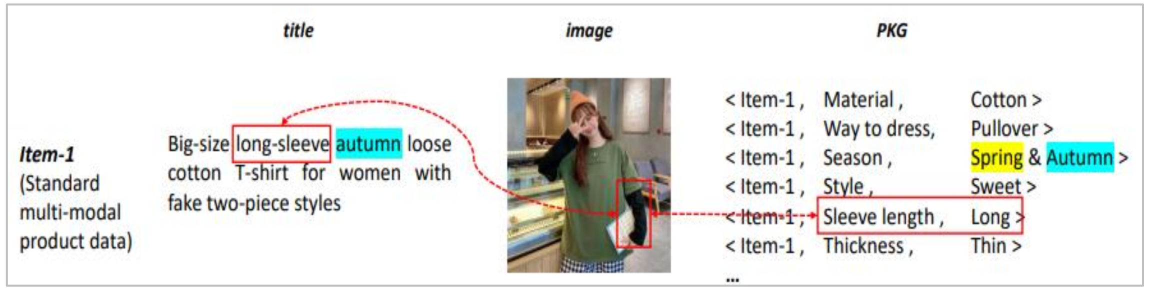
Model	BBC	20 Newsgroup	AG News	MIND	Avg.
SBERT	96.0	64.5	86.5	75.2	80.5
USE	97.3	66.4	85.8	75.0	81.1
mUSE	97.3	71.4	86.8	75.8	82.8
LaBSE	96.6	70.3	86.3	73.9	81.8
Laser	92.1	63.5	81.4	63.1	75.0
NewsEmbed	**97.5**	**73.9**	**89.1**	**77.0**	**84.2**

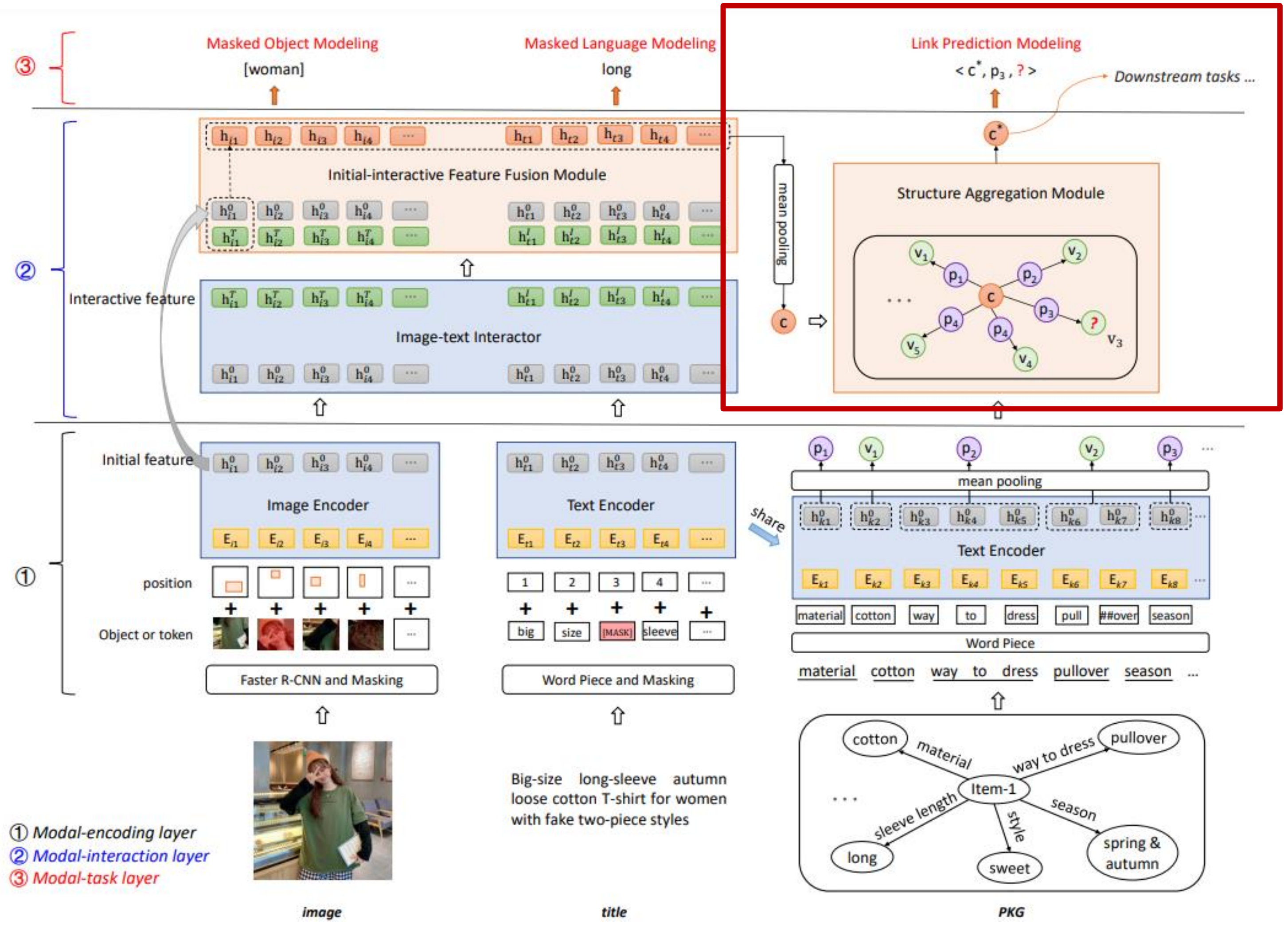
Liu et al., NewsEmbed: Modeling News through Pre-trained Document Representations, 2021

In-domain Pretraining w/ Augmented Knowledge (continued)

K3M

- **Link Prediction Modeling (LPM)**: learning modality encoders to perform link prediction in KGs





- Masked Object Modeling (MOM)
- Masked Language Modeling (MLM)
- Link Prediction Modeling (LPM)

Pretraining loss: $L_{pre} = l_{MLM} + l_{MOM} + l_{LPM}.$

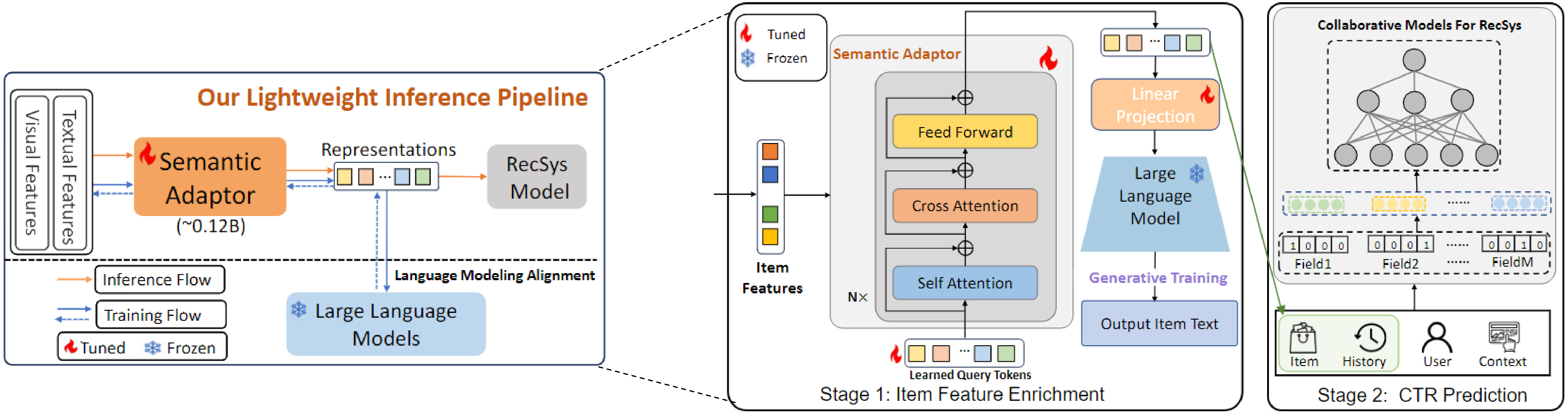
Zhu et al., Knowledge Perceived Multi-modal Pretraining in E-commerce, 2021

In-domain Pretraining w/ Augmented Knowledge (continued)

EASE

- **Enrich with LLMs**: LLMs serve as a world knowledge base
- Semantic adapter bridge item features and LLMs
- Reconstruction text from LLMs to transfer knowledge

$$\log p(t|x) = \sum_{l=1}^{L} \log p(t_l|\tilde{Z}_x, t_1, t_2, \cdots, t_{l-1})$$
$$= \sum_{l=1}^{L} \log p(t_l|i_1, i_2, \cdots, i_{n_q}, t_1, t_2, \cdots, t_{l-1}),$$



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

Summarization

- **Multimodal Pretraining Techniques for Recommendation**
  - With **collaborative signals**
    - User-to-item matching
    - Item-to-item matching
    - ID-to-modality alignment
  - With **self-supervised tasks**
    - Mask prediction
    - Contrastive alignment
  - With **augmented knowledge**
    - Categories/Tags/Topics
    - Knowledge graphs
    - LLMs

Multimodal Pretraining vs. Adaptation

Multimodal pretraining for recommendation

How to learn a good representation for recommendation?

Pretrained large models

VS

Recommendation models

Multimodal adaptation for recommendation

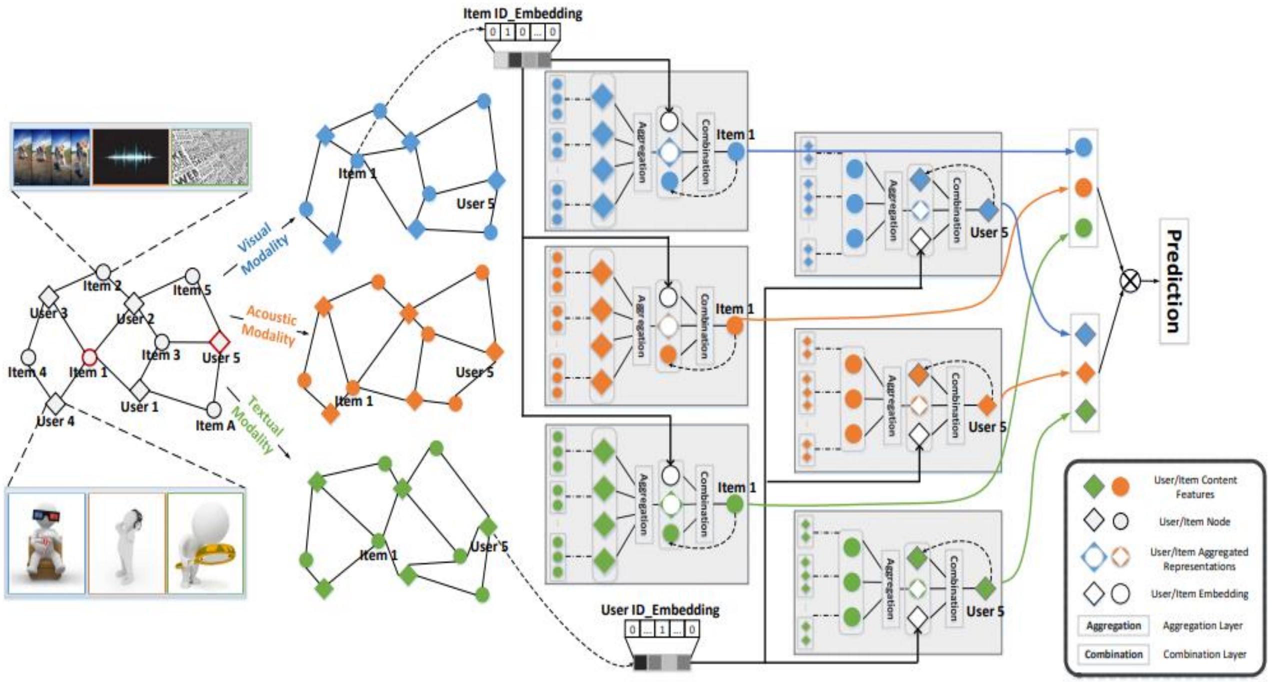
How to learn a good recommender given pretrained representations?

Multimodal Adaptation for Recommendation

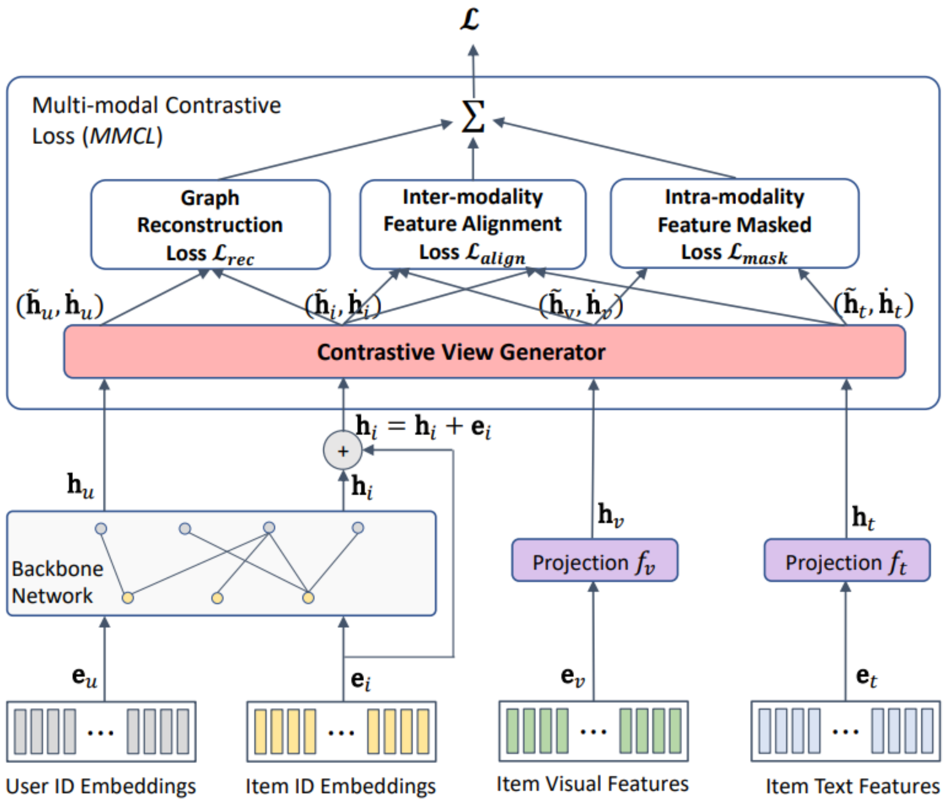
- **Representation transfer**
  - Multimodal features as side information
  - Multimodal sequence learning for user representation
  - Multimodal sequence denoising and aggregation
- **Joint Finetuning**
  - Only finetuning item-side encoder
  - Finetuning both item-side and user-side encoders
  - Finetuning user-item cross encoder
- **Adapter/Prompt Tuning**
  - Multimodal fusion adapter
  - LLM adapter for multimodal recommendation

Representation Transfer (1)

- Multimodal features as side information



MMGCN: leveraging multimodal graph networks

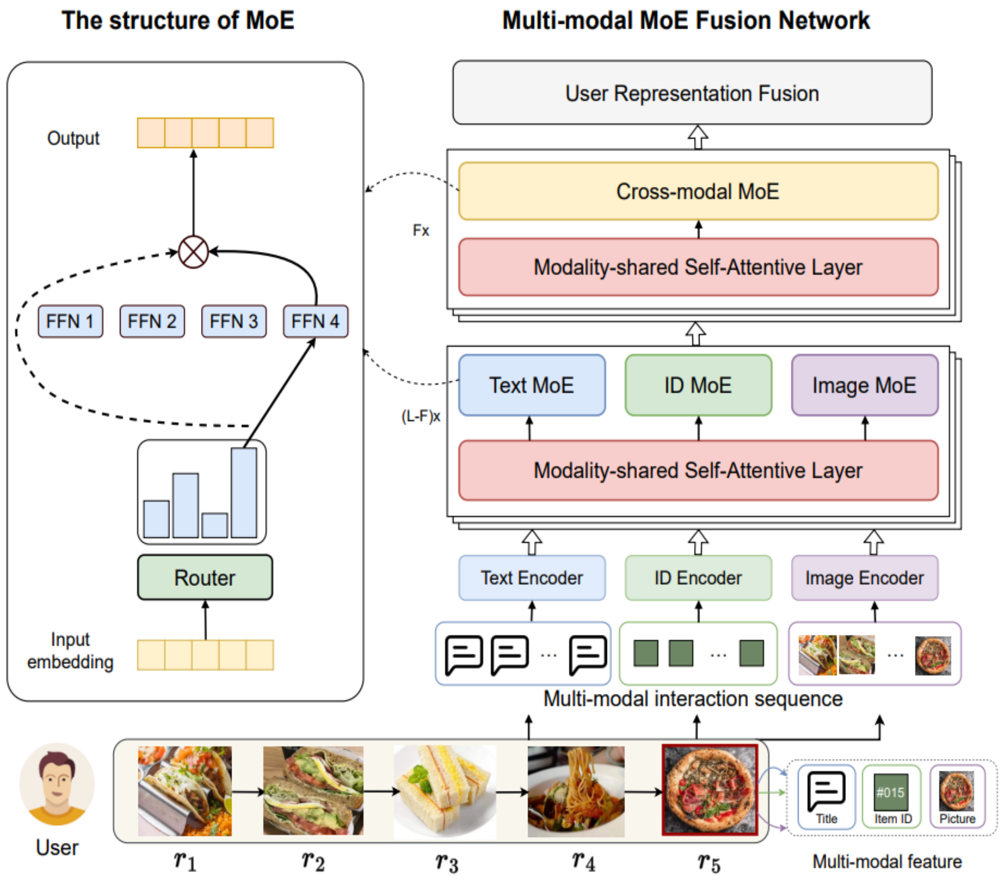


BM3: leveraging ID-modality alignment

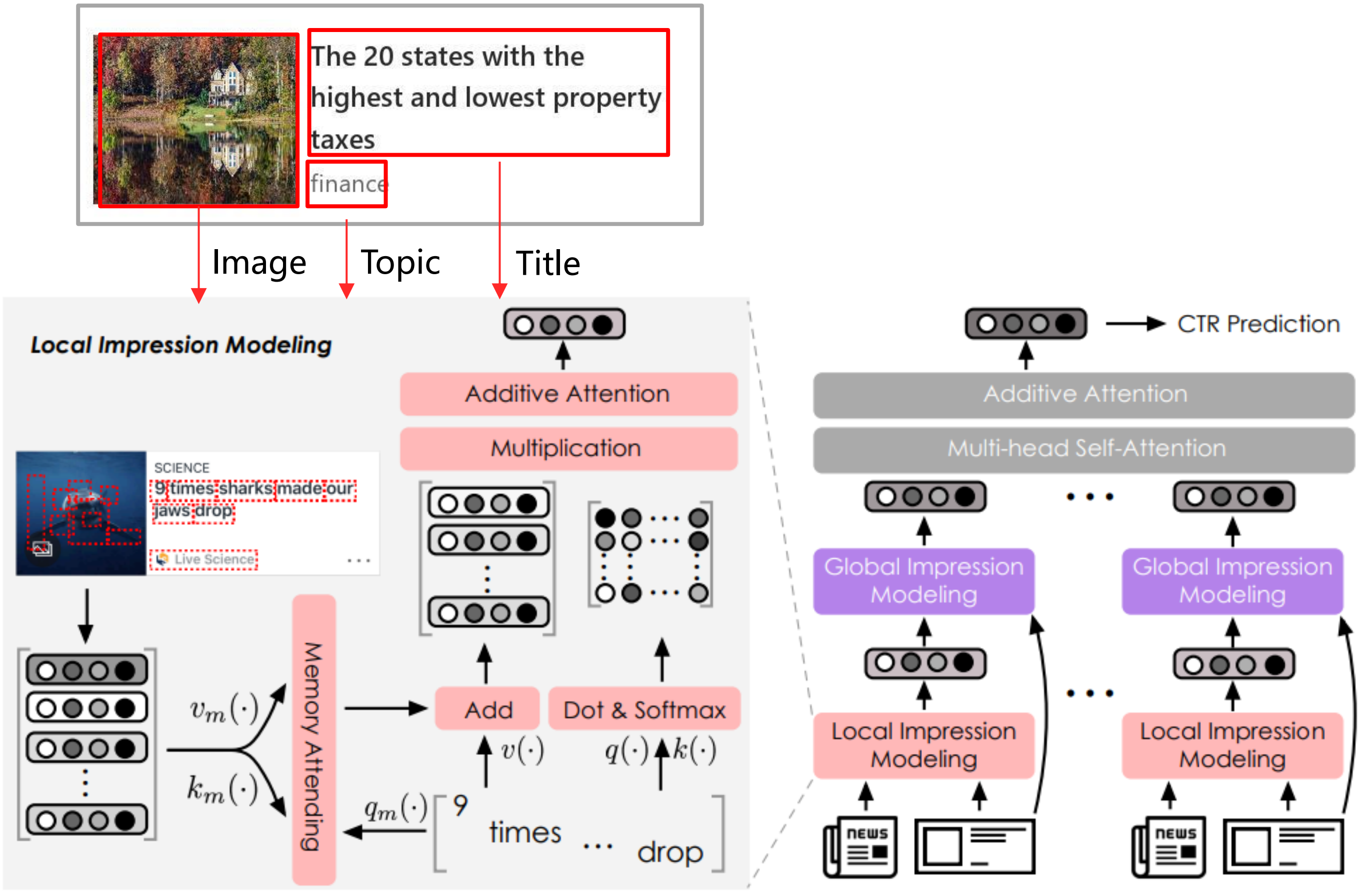
Wei et al., MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video, 2019
Zhou et al., Bootstrap Latent Representations for Multi-modal Recommendation, 2023

Representation Transfer (2)

- Multimodal sequence learning for user representation



M3SRec: MOE-based multimodal fusion

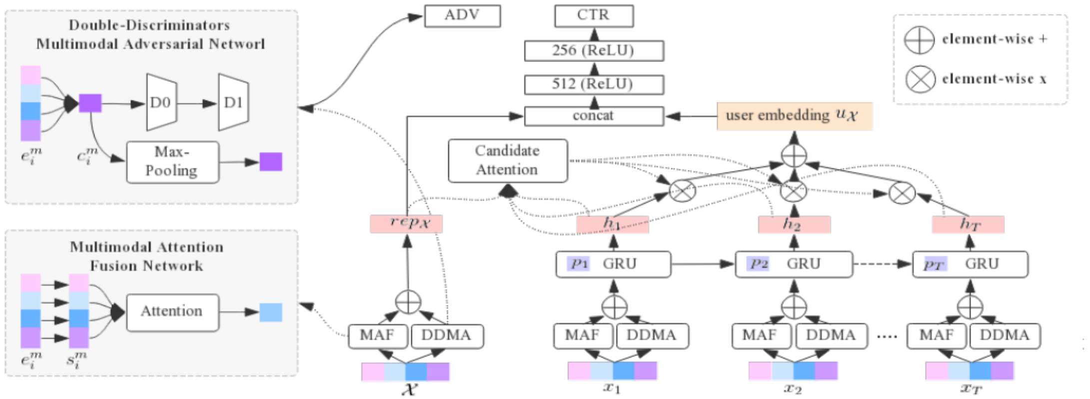


IMRec [Huawei]: local and global fusion

Bian et al., Multi-modal Mixture of Experts Represetation Learning for Sequential Recommendation, 2023
Xun et al., Why Do We Click: Visual Impression-aware News Recommendation, 2021

Representation Transfer (3)

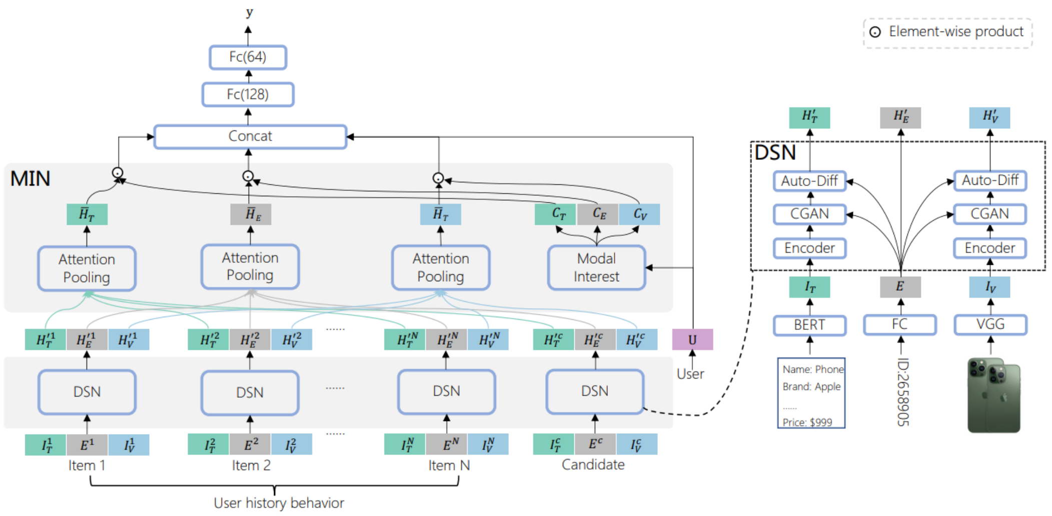
➢ Multimodal sequence denoising and aggregation



MARN:

➢ Multimodal attention network

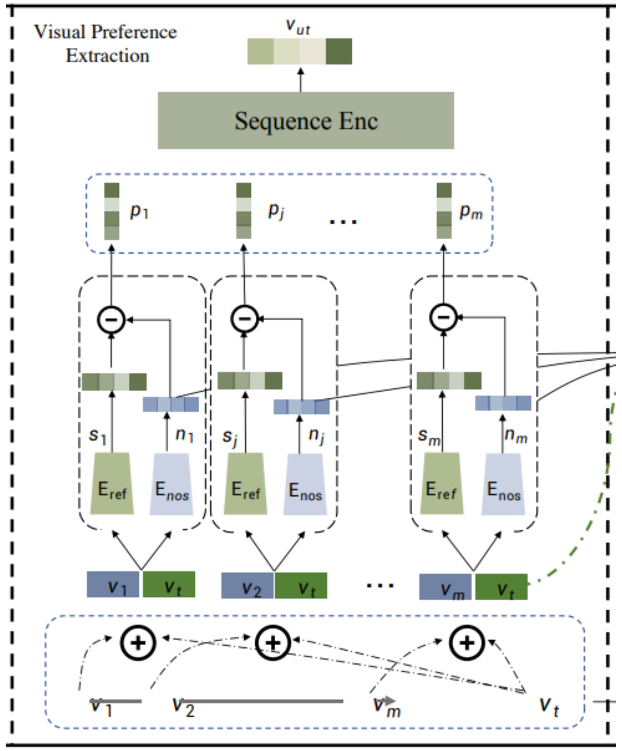
➢ Multimodal adverseral network



GMMF:

➢ Difference set network

➢ Modal-interest network



Tavern:

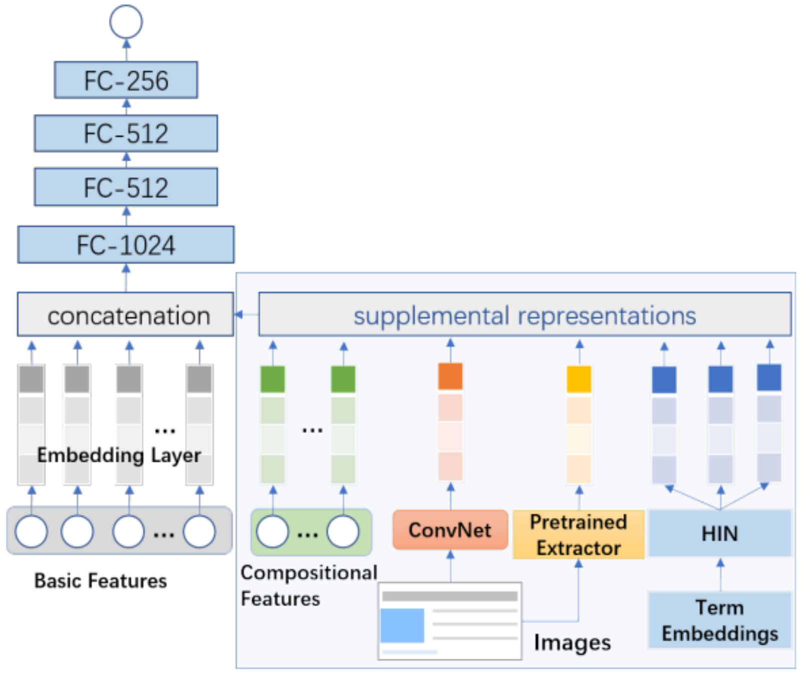
➢ Visual preference extraction

➢ Selective orthogonality disentanglement

Li et al., Adversarial Multimodal Representation Learning for Click-Through Rate Prediction, 2020
Xiao et al., From Abstract to Details: A Generative Multimodal Fusion Framework for Recommendation, 2022
Wen et al., Unified Visual Preference Learning for User Intent Understanding, 2024

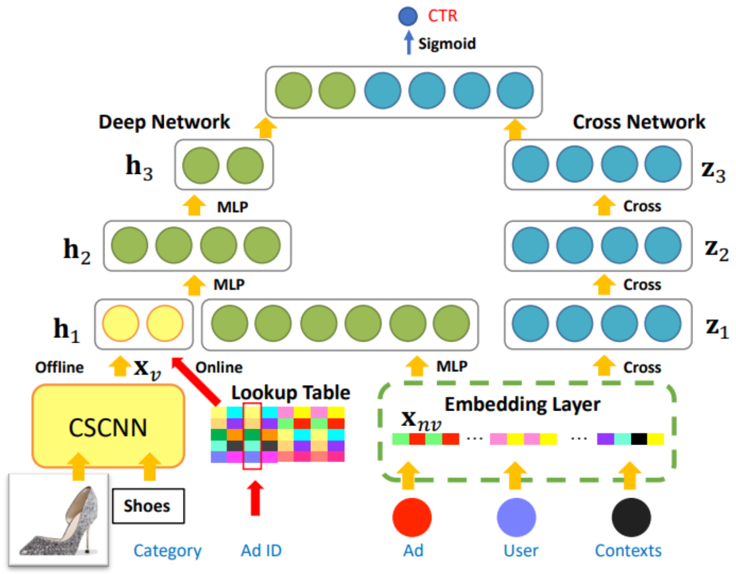
Joint Finetuning (1)

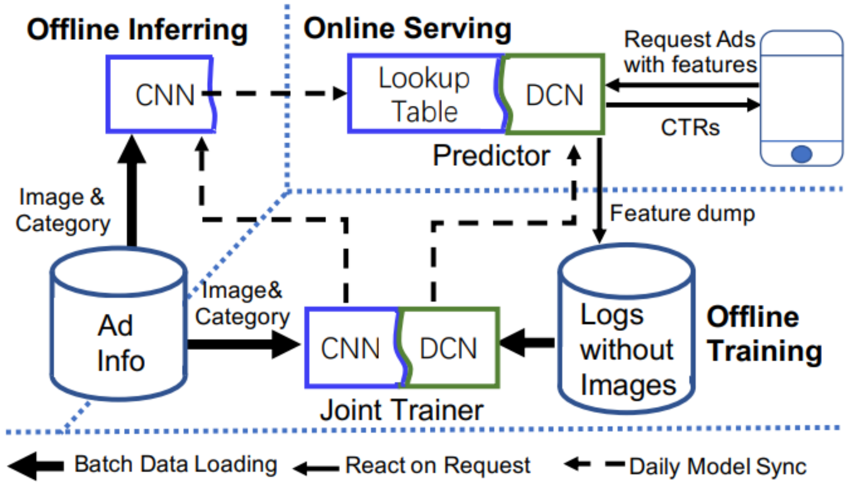
- Only finetuning item-side encoder



HIN-CTR:

- Using both pretrained and finetuned CNNs
- Leveraging text term embeddings





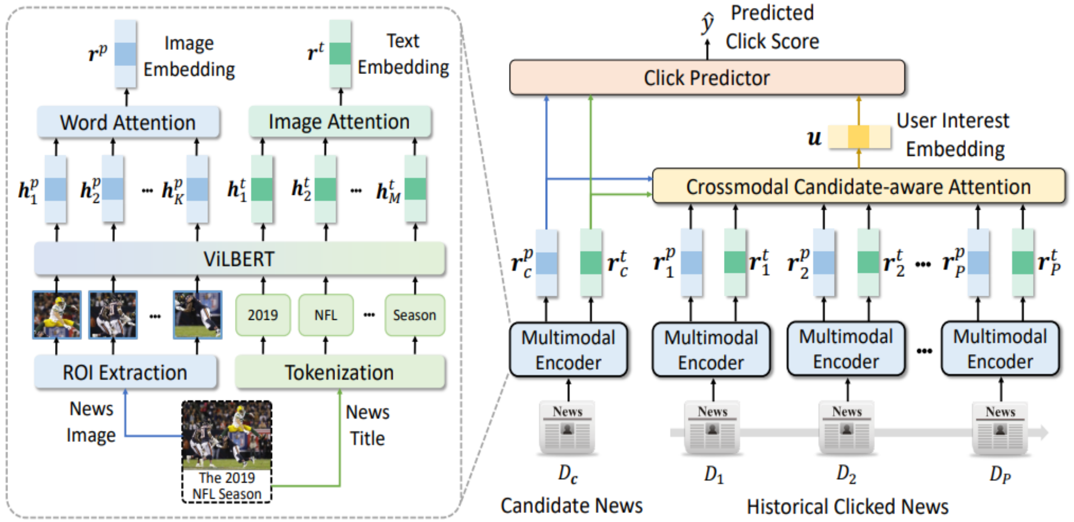
CSCNN:

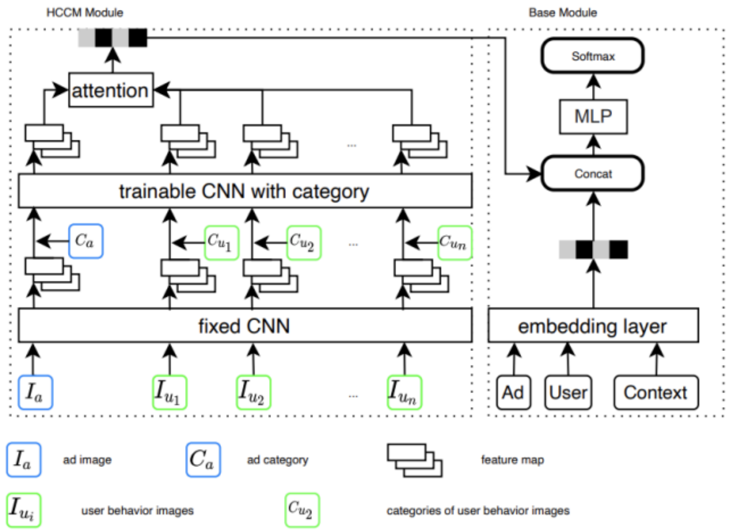
- Leveraging category-aware CNN encoder
- Caching the embeddings after training for online serving

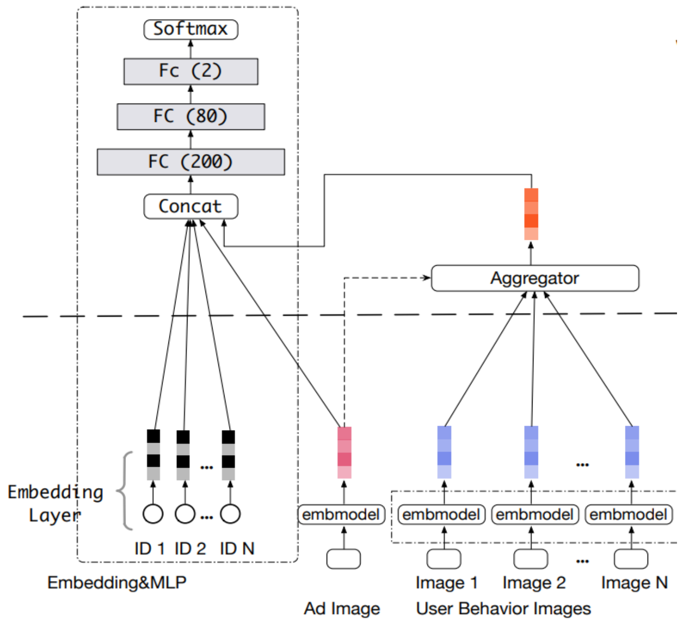
Yang et al., Learning Compositional, Visual and Relational Representations for CTR Prediction in Sponsored Search, 2019
Liu et al., Category-Specific CNN for Visual-aware CTR Prediction at JD.com, 2020

Joint Finetuning (2)

➢ Finetuning both **item-side** and **user-side** encoders







MM-Rec:

➢ Finetuning the last three layers of ViLBERT

➢ Caching the image and text embeddings

HCCM:

➢ Leveraging category-aware CNN encoder

➢ Caching the embeddings after training

DCIM:

➢ Finetuning pretrained CNNs

➢ Caching the embeddings after training

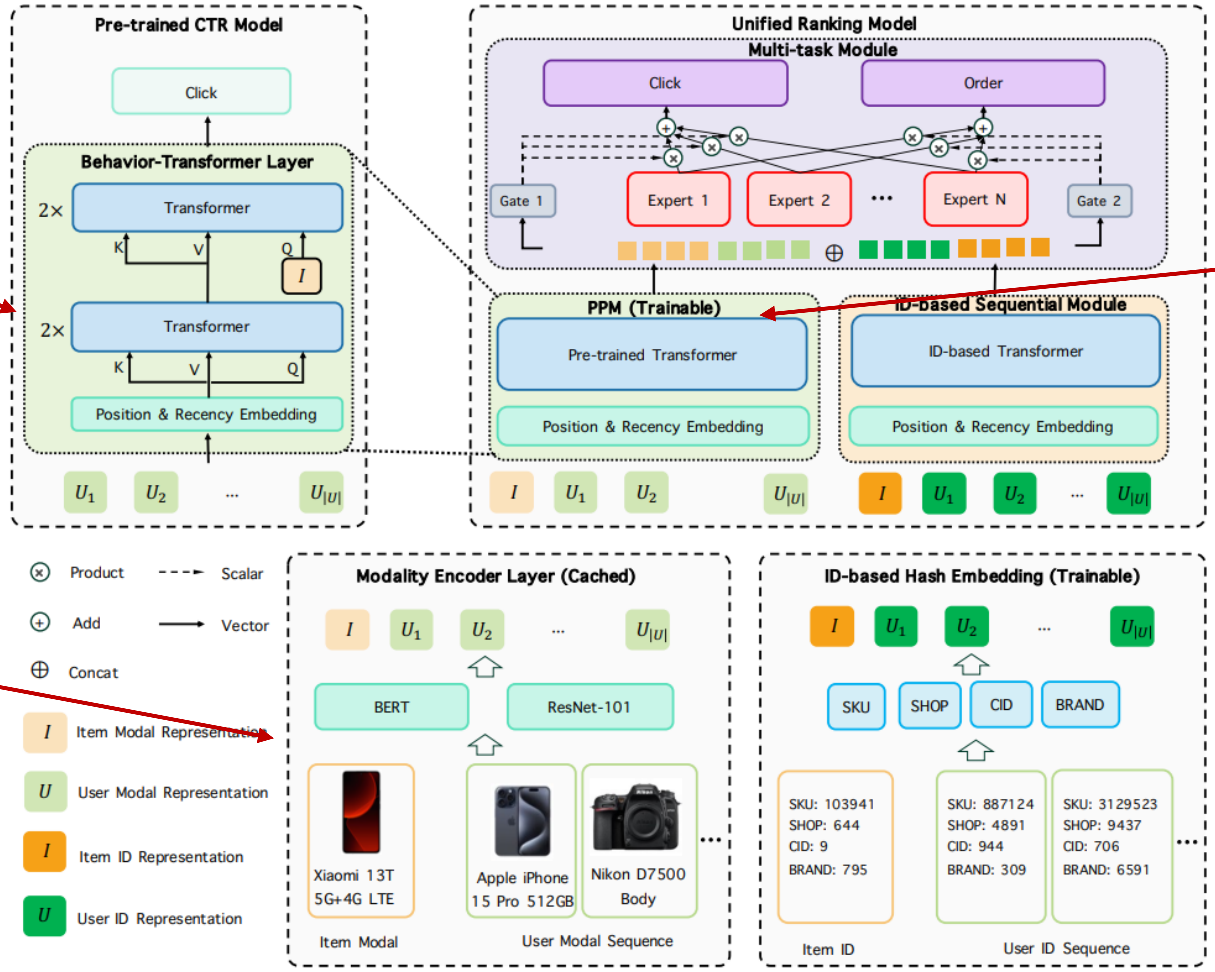
Wu et al., MM-Rec: Multimodal News Recommendation, 2022
Chen et al., Hybrid CNN Based Attention with Category Prior for User Image Behavior Modeling, 2022
Ge et al., Image Matters: Visually modeling user behaviors using Advanced Model Server, 2018

Joint Finetuning (3)

➢ Finetuning **user-item cross encoder**

Step2: Pretrain behavior transformer

Step1: Pretrain multimodal item embeddings



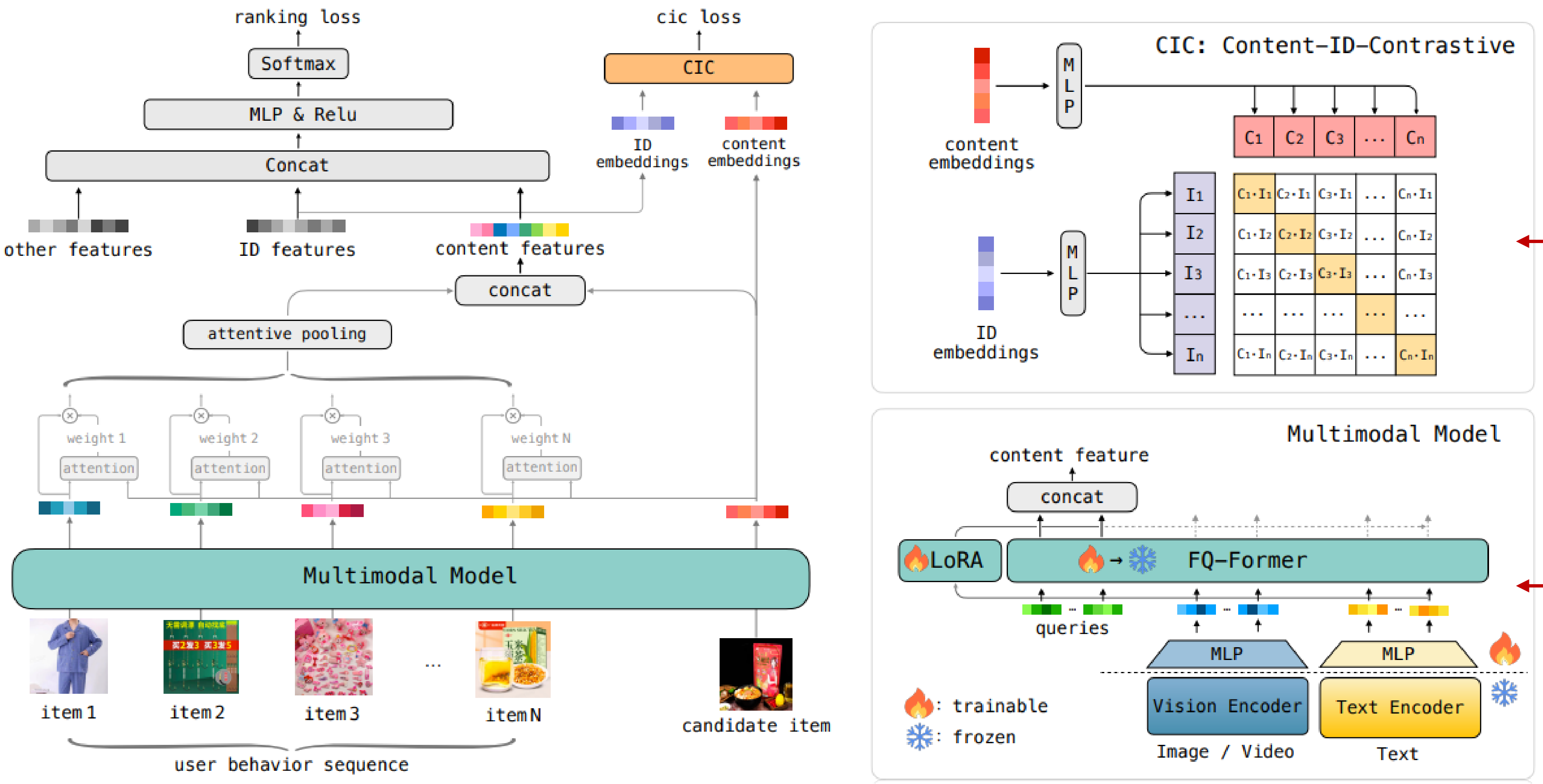
Step3: Finetuning behavior transformer

PPM

Gao et al, PPM : A Pre-trained Plug-in Model for Click-through Rate Prediction, 2024

Adapter Tuning (1)

- Multimodal fusion adapter



Content-ID contrastive learning

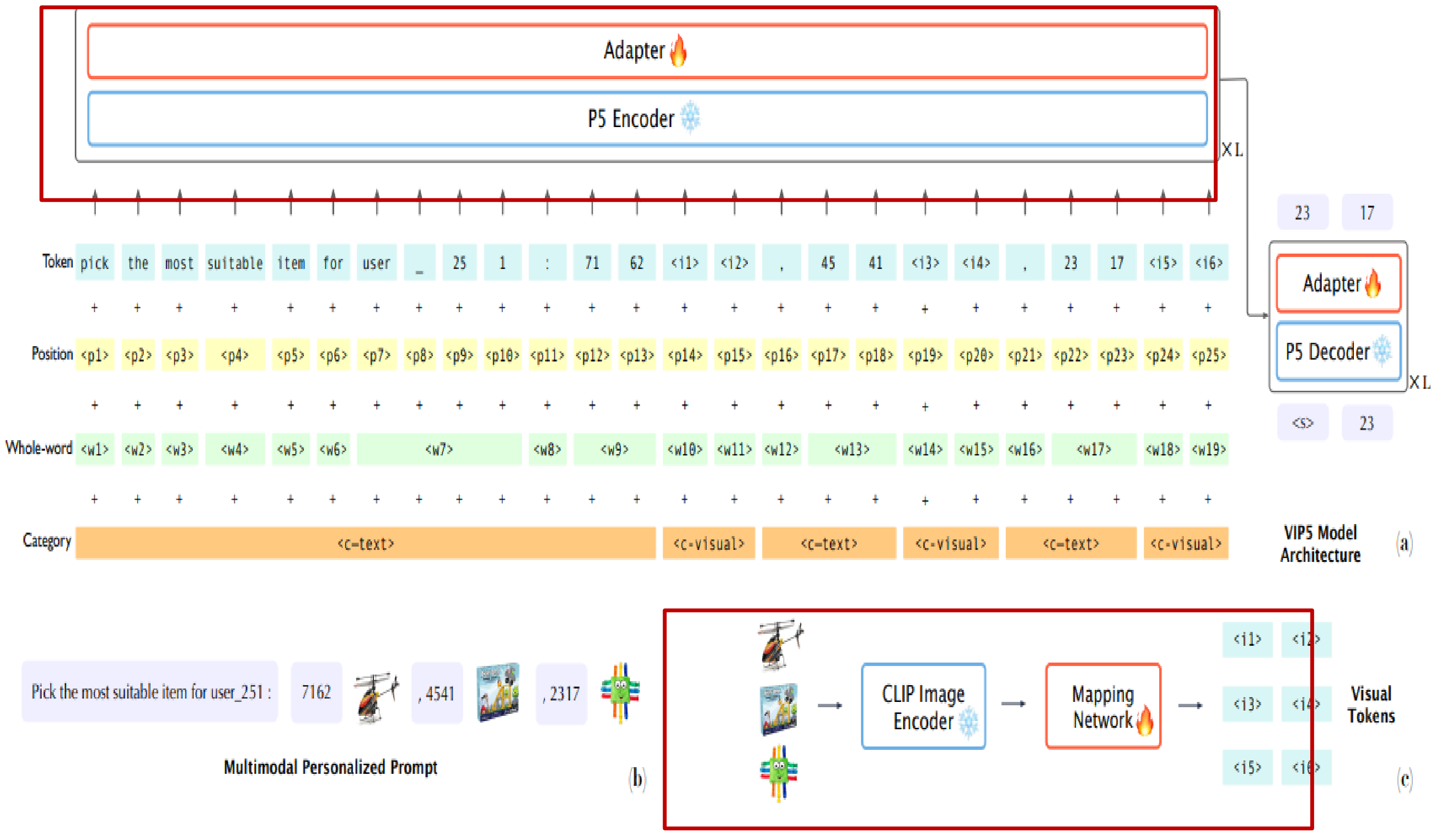
Multimodal fusion adapter

EM3

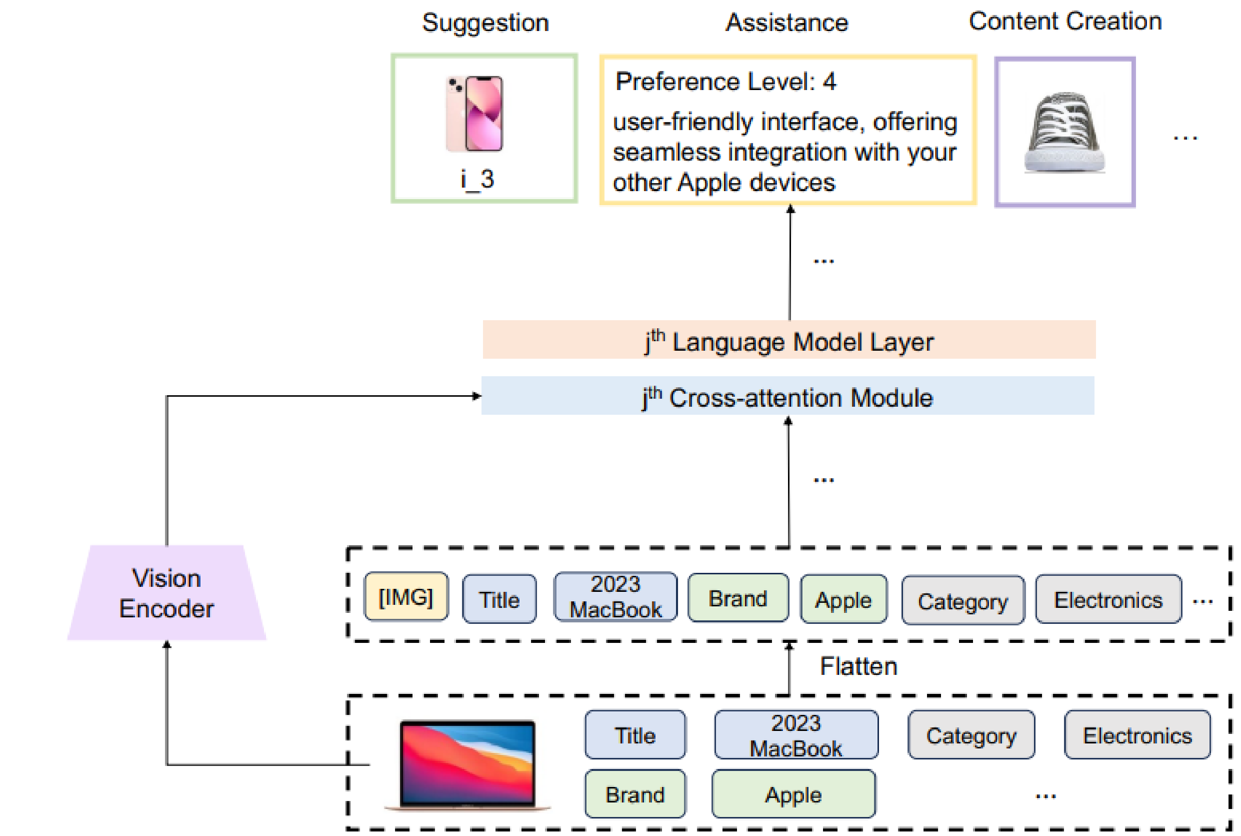
Deng et al., End-to-end Training of Multimodal Model and Ranking Model, 2024

Adapter Tuning (2)

➢ LLM adapter for multimodal fusion



VIP5: P5 + Adapter

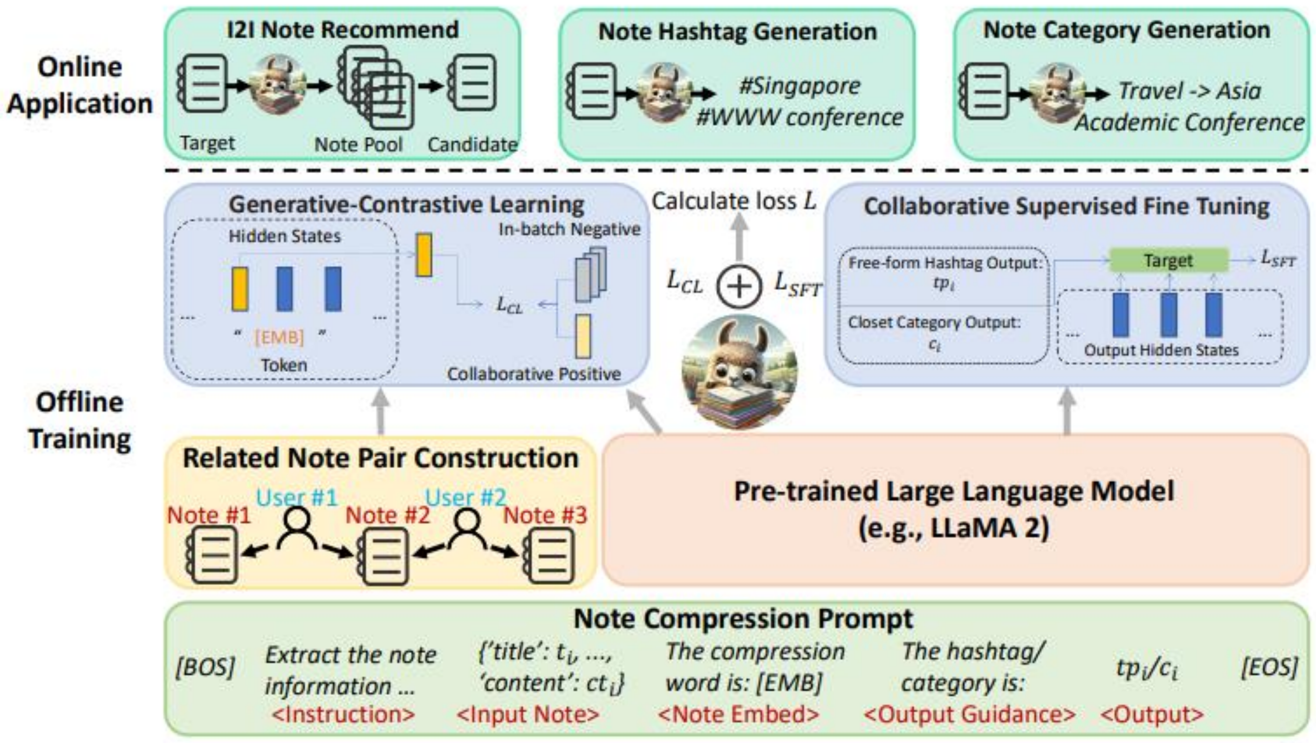


UniMP: 3B LLM + Adapter

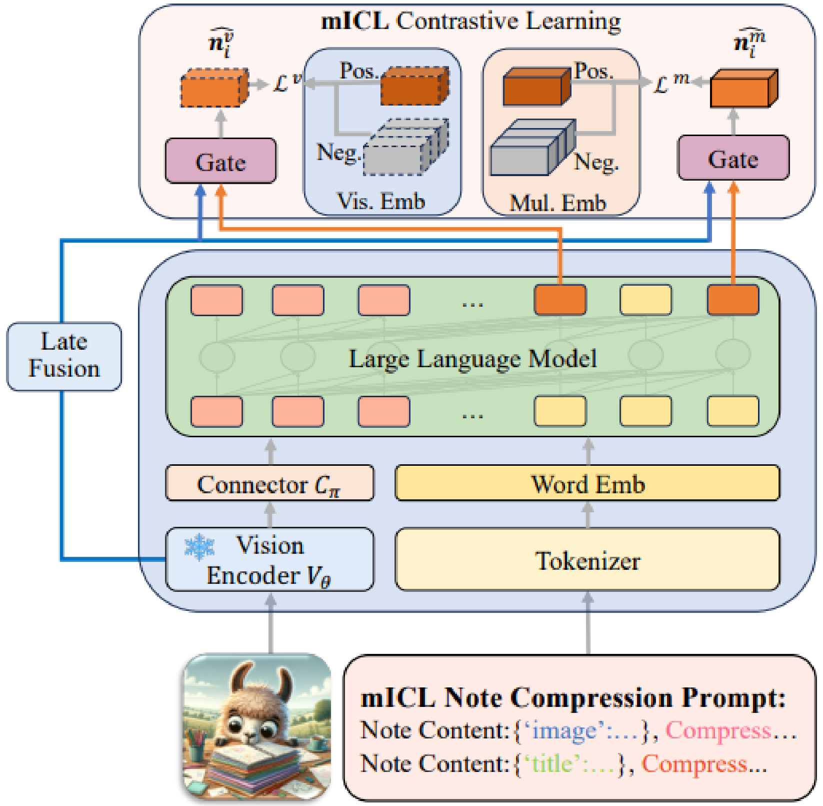
Geng et al., VIP5: Towards Multimodal Foundation Models for Recommendation, 2024
Wei et al., Towards Unified Multi-Modal Personalization: Large Vision-Language Models for Generative Recommendation and Beyond, 2024

Prompt Tuning

- Learnable prompt token as embedding



Prompt: [BOS]<Instruction> <Input Note> The compression word is: "$\boxed{\text{[EMB]}}$". <Output Guidance> <Output>[EOS]



mICL Note Compression Prompt.
Note content: {'image': <IMG>}, Compress this note into one word: $\boxed{\text{"<IMG\_EMB>"}}$. Note content: {'title': t_i , 'topic': tp_i , 'content': ct_i}. Compress this note into one word:"

Zhang et al., NoteLLM: A Retrievable Large Language Model for Note Recommendation, 2024
Zhang et al., NoteLLM-2: Multimodal Large Representation Models for Recommendation, 2024

Summarization

- **Multimodal Adaptation Techniques for Recommendation**
  - **Representation transfer**
    - Multimodal features as side information
    - Multimodal sequence learning for user representation
    - Multimodal sequence denoising and aggregation
  - **Joint Finetuning**
    - Only finetuning item-side encoder
    - Finetuning both item-side and user-side encoders
    - Finetuning user-item cross encoder
  - **Adapter Tuning**
    - Multimodal fusion adapter
    - LLM adapter for multimodal recommendation
  - **Prompt Tuning**
    - Learnable prompt token

Open Challenges

- **Vertical-domain foundational model**
  - Multimodal inputs
  - Multi-domain data
  - One model for multi-tasks
  - Unified modeling
- **Interplay with MLLMs**
  - Adapting LLMs/MLLMs to recommendation
  - From representation to generative modeling
  - Model scaling law
  - Model efficiency
- **Benchmarks and evaluation**
  - BARS/GreenRec/RecBench…
  - Amazon/MIND/PixelRec/MicroLens

Thank You!