

Multimodal Pre-training for Recommendation

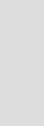
Jieming Zhu

Huawei Noah's Ark Lab

<https://jiemingzhu.github.io>

WWW 2024 @ Singapore





Multimodal models are reshaping the world
with 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.

Fig. from: 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.

Fig. from: OpenAI et al., GPT-4 Technical Report, 2023

Cross-Modal Retrieval

Text-to-image retrieval

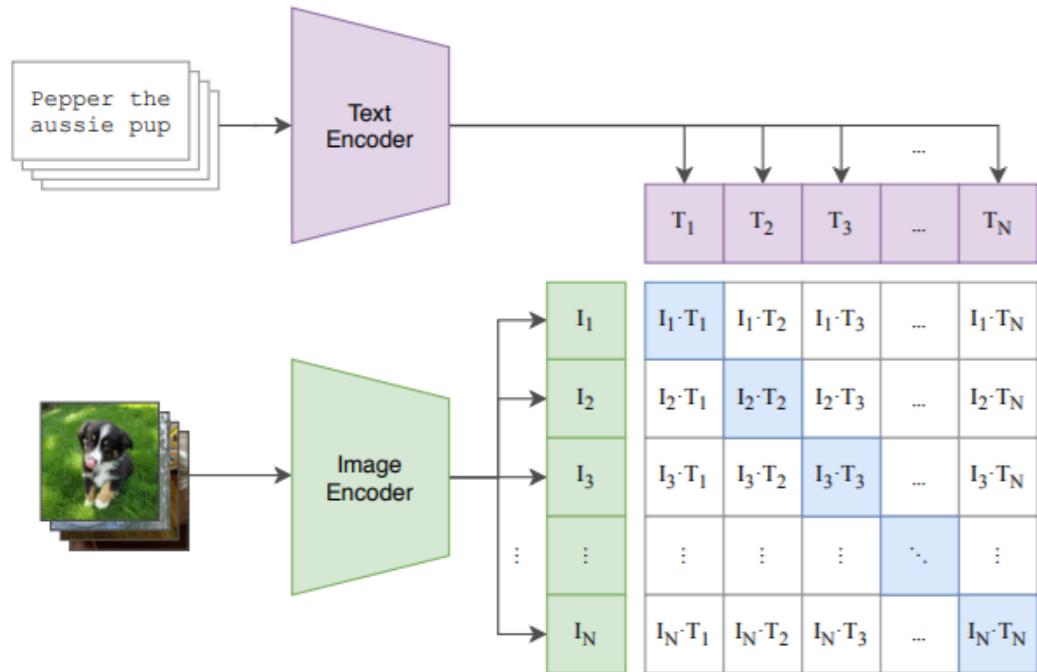


Image-to-audio retrieval

↓ Select an image



↓ Explore audio retrievals

▶ 0:19 / 0:26 🔍

Distance: 1.138

▶ 0:00 / 0:16 🔍

Distance: 1.155

Fig. from: Radford et al., Learning Transferable Visual Models From Natural Language Supervision, 2021

Fig. from: Girdhar et al., ImageBind: One Embedding Space To Bind Them All, 2023

Multimodal Generation

Text-to-image generation

Text-to-Image



Fig. from: Stable Diffusion 2. <https://github.com/Stability-AI/stablediffusion>

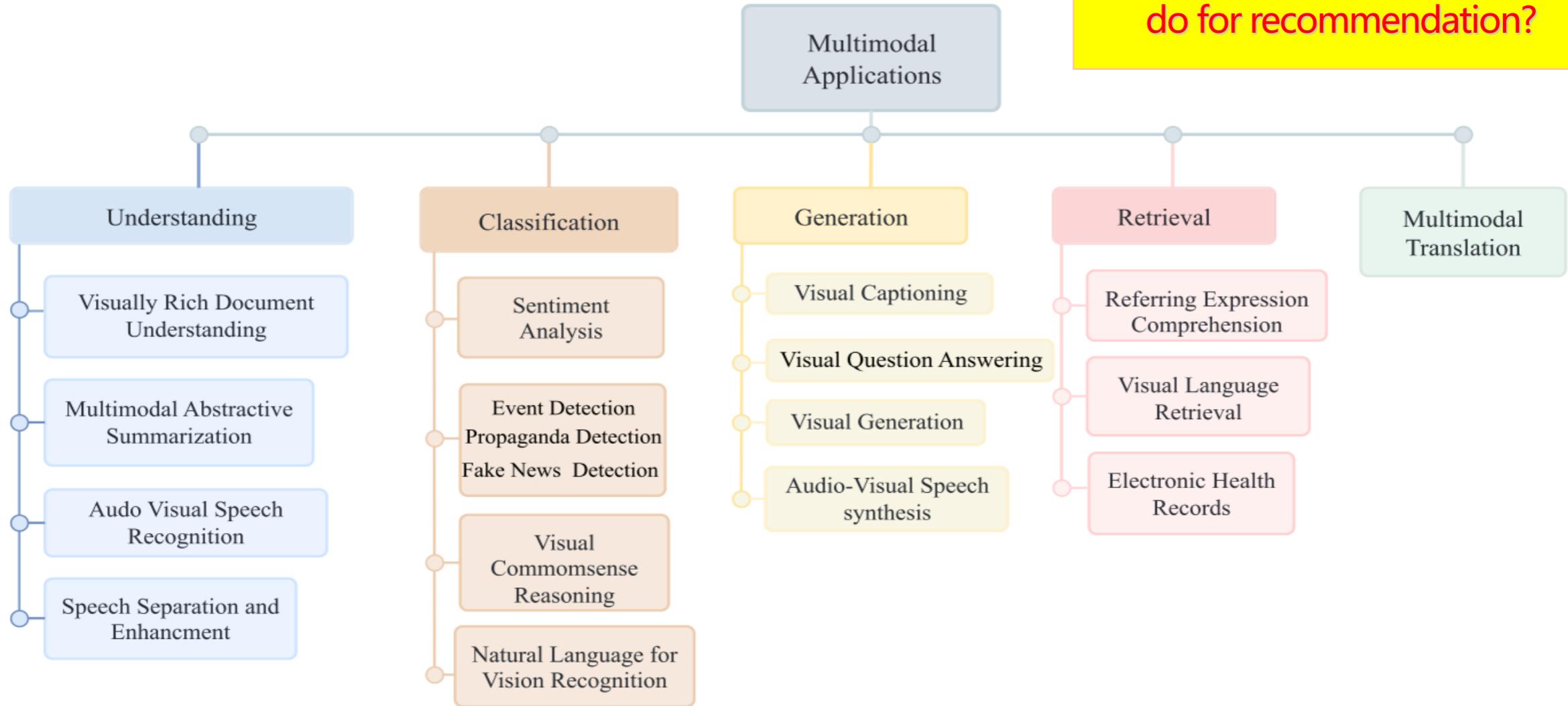
Text-to-video generation

Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress,...



Fig. from: <https://openai.com/index/video-generation-models-as-world-simulators/>

More and More Applications...



What can multimodal models do for recommendation?



THE WEB CONFERENCE 2024 IN SINGAPORE

MAY 13 - 17, 2024

MULTIMODAL PRETRAINING
FOR RECOMMENDATION

Outline

- **General-Domain Multimodal Pretraining**
 - Pretraining tasks
 - Pretraining modalities
 - Pretrained models
- **Multimodal Pretraining for Recommendation**
 - Domain data
 - In-domain pretraining tasks
 - Downstream recommendation tasks
- **Open Challenges and Opportunities**
 - Pretraining
 - Adaptation



THE WEB CONFERENCE 2024 IN SINGAPORE

MAY 13 - 17, 2024

MULTIMODAL PRETRAINING
FOR RECOMMENDATION

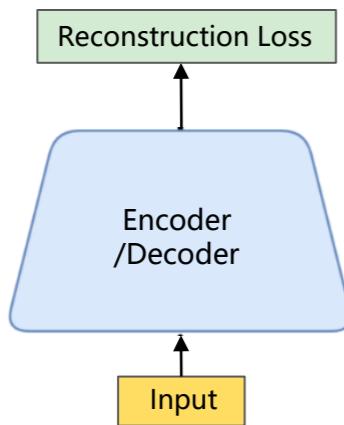
General-Domain Multimodal Pretraining

- **Pretraining tasks**
 - Reconstructive
 - Contrastive
 - Generative
- **Pretraining modalities**
 - Text
 - Image
 - Audio
 - More modalities
- **Pretrained models**
 - CLIP
 - BLIP-2
 - ImageBind
 - UnifiedIO-2

Self-supervised Pretraining

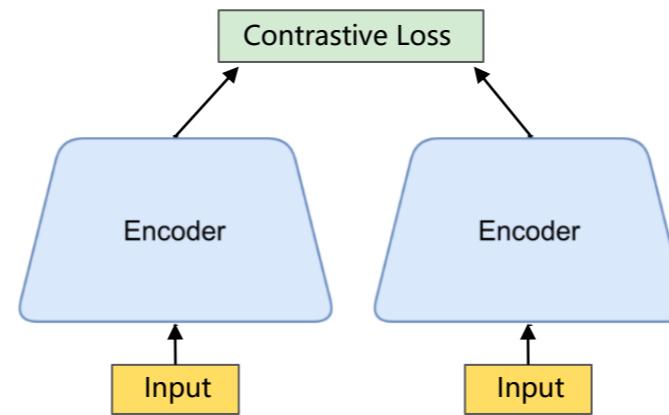
Typical pretraining paradigms: reconstructive, contrastive, and generative

BERT
StableDiffusion



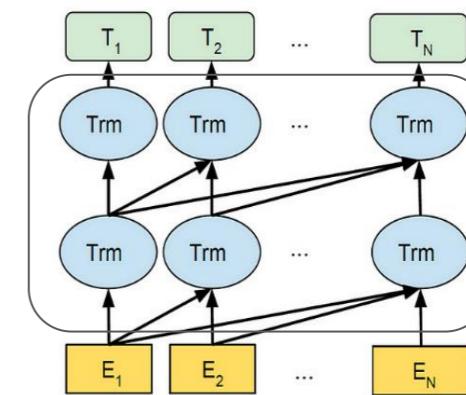
Reconstructive

CLIP
ImageBind



Contrastive

GPT
DALL-E



Generative

Pretrained Multimodal Models (2019~2022)

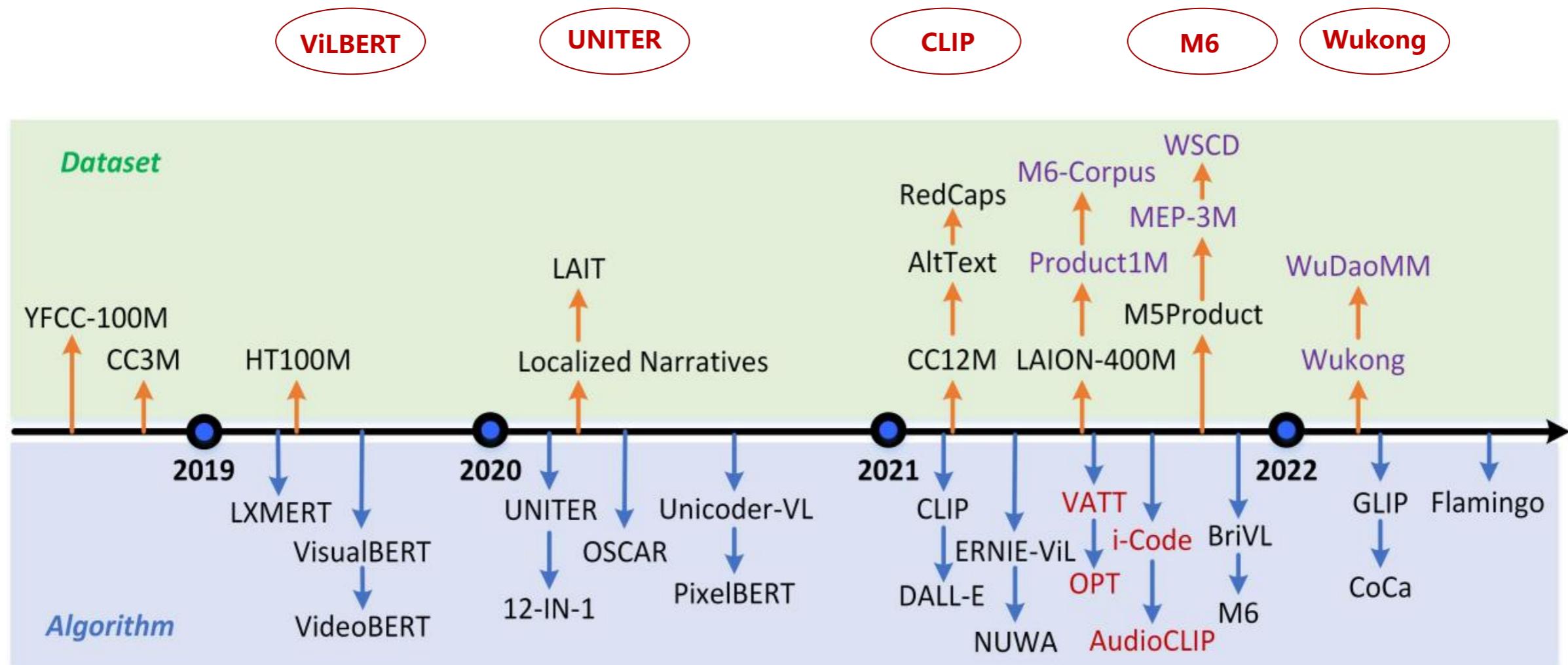


Fig. from: Wang et al., Large-scale Multi-Modal Pre-trained Models: A Comprehensive Survey, 2023

Learning Paradigm: Pretraining-Finetuning

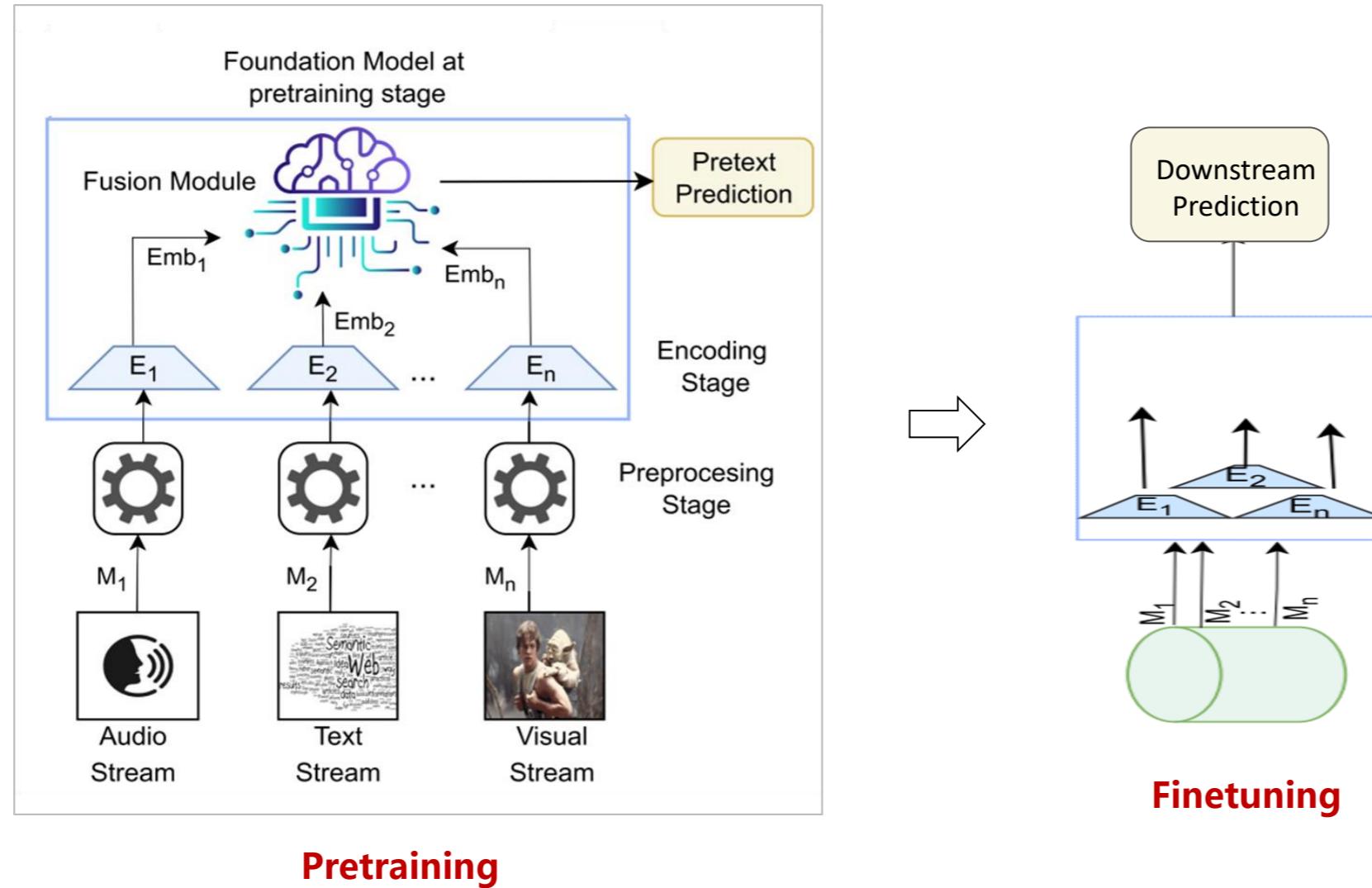
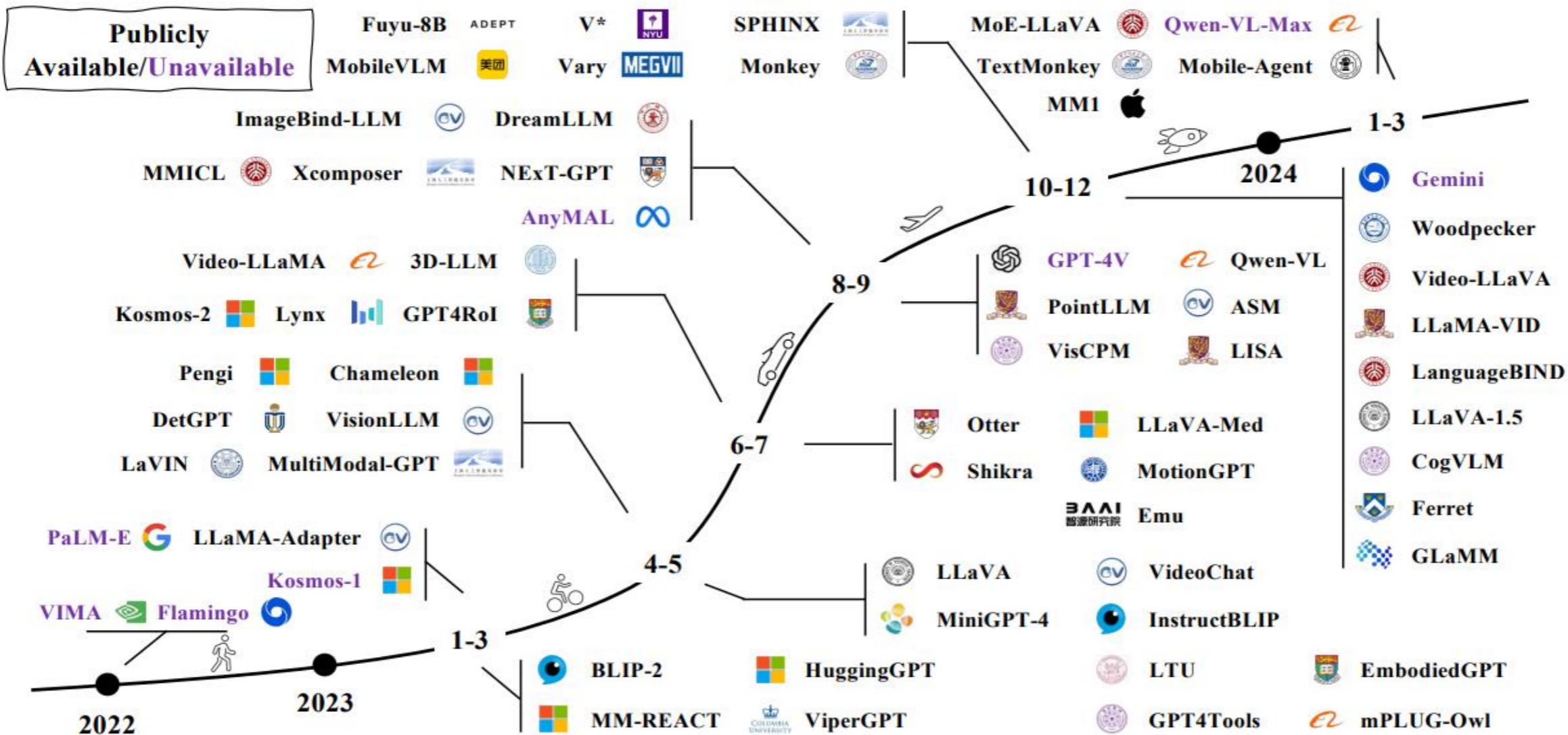


Fig. from: MANZOOR et al., Multimodality Representation Learning: A Survey on Evolution, Pretraining and Its Applications, 2023

Multimodal Large Language Models (2023~2024)



Learning Paradigm: Prompting + In-Context Learning

- **In-context learning** involves the injection of few-shot samples into the prompts, allowing the model to learn from in-context examples and output similar logits.
- **Chain-of-thought prompting** centers on the user's approach to crafting interconnected prompts to guide the model through a conversation or task.

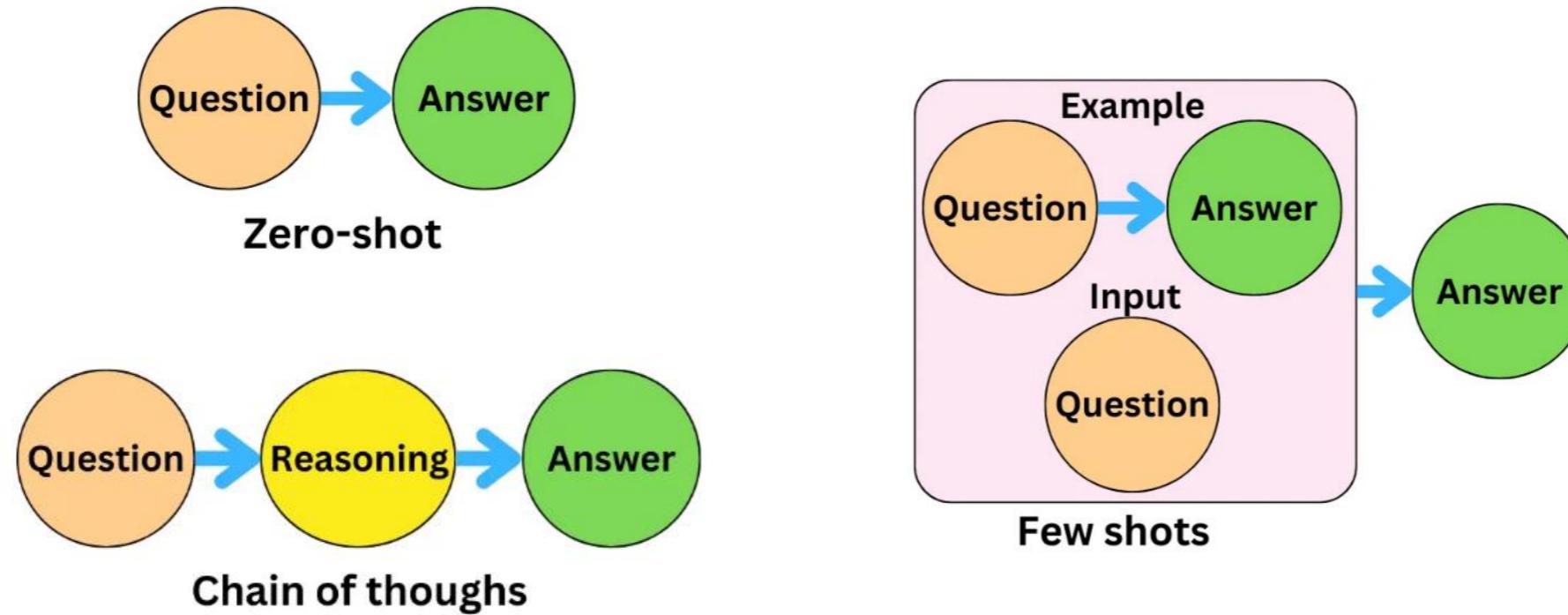
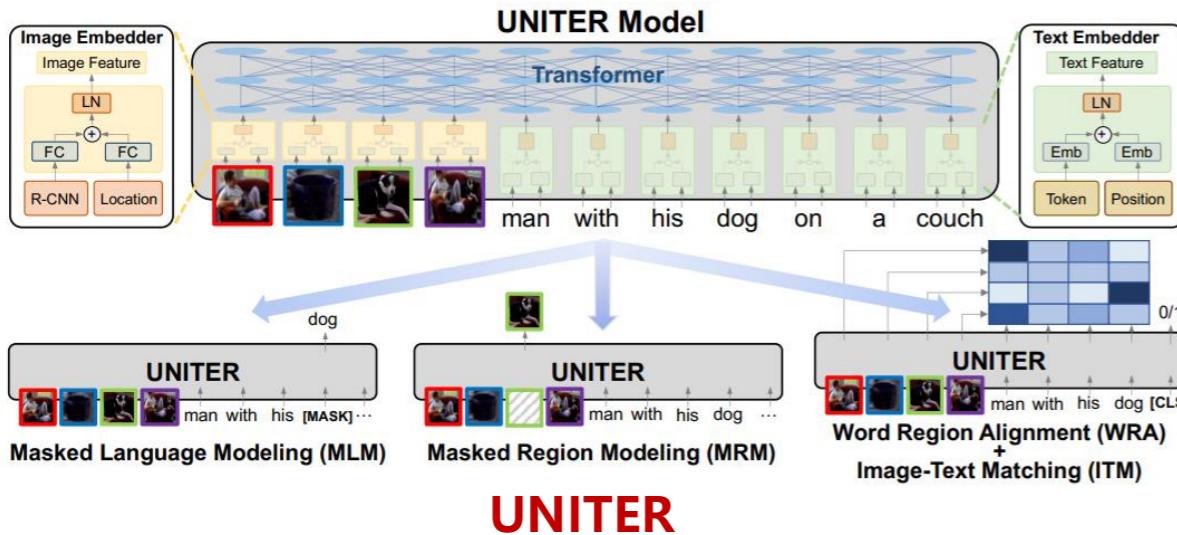
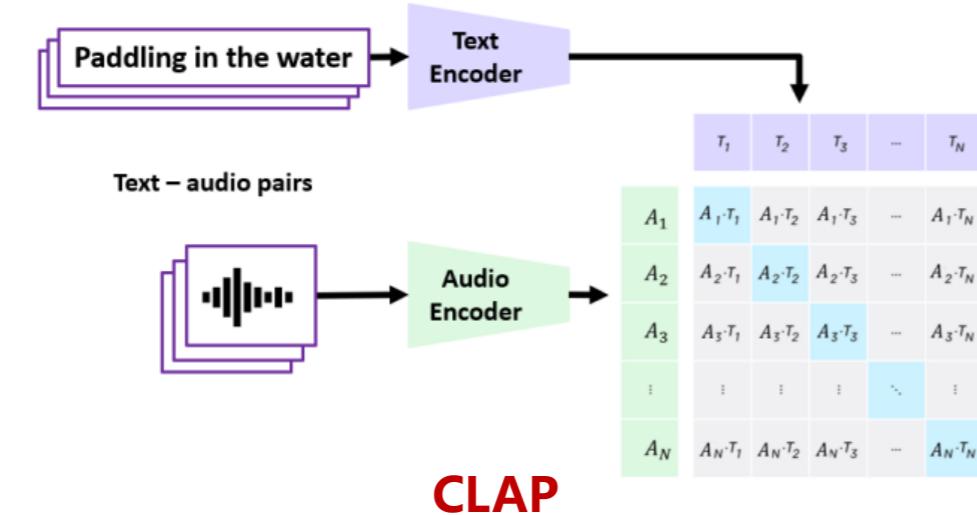
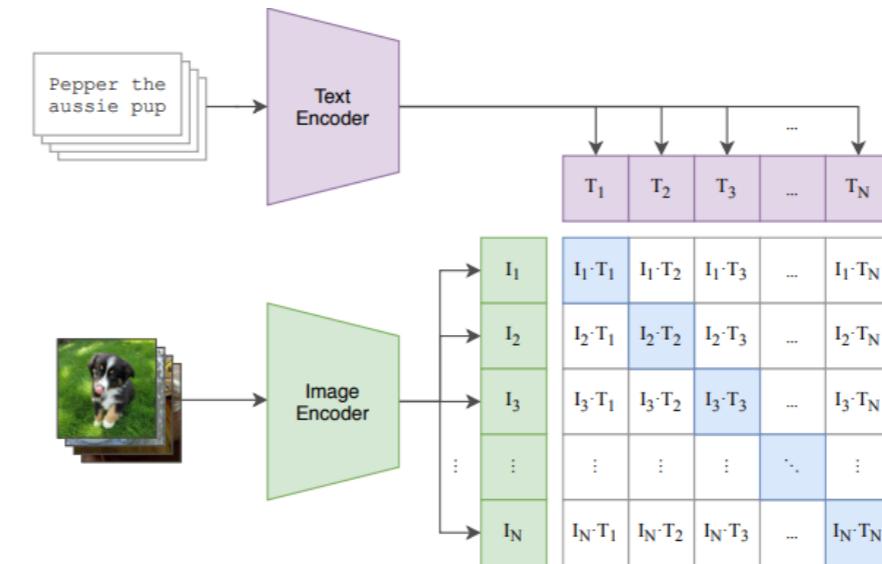


Fig. from <https://newsletter.theaiedge.io/p/prompt-engineering-and-llmops-building>

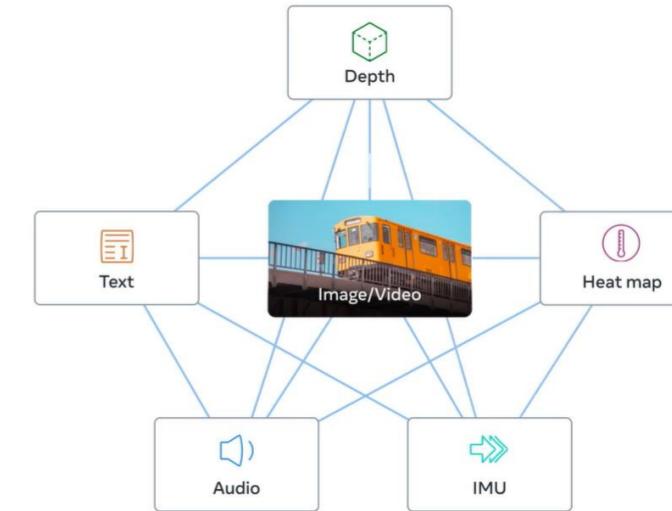
Multimodal Large Models (1)



CLIP



ImageBind



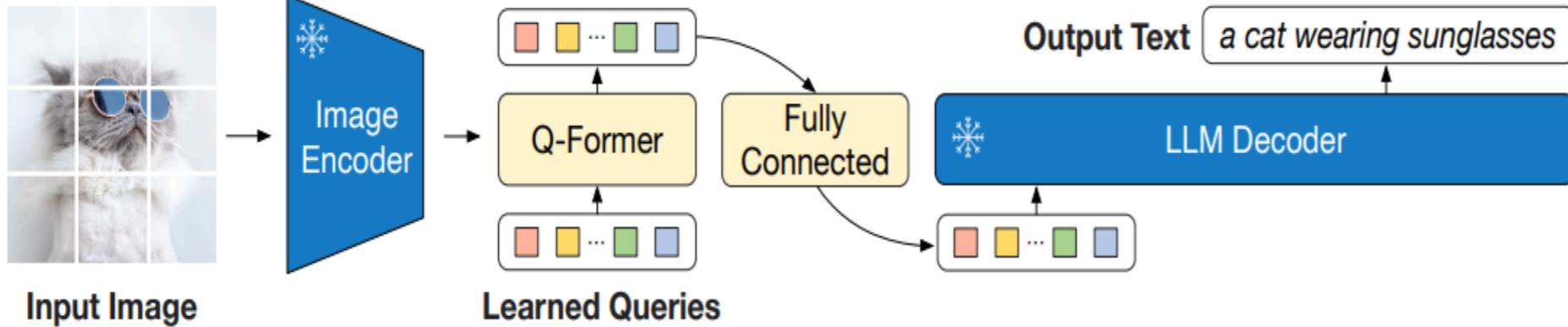
Chen et al., UNITER: UNiversal Image-TExt Representation Learning, 2019

Lu et al., Learning Transferable Visual Models From Natural Language Supervision, 2021

Elizalde et al., CLAP: Learning Audio Concepts From Natural Language Supervision, 2022

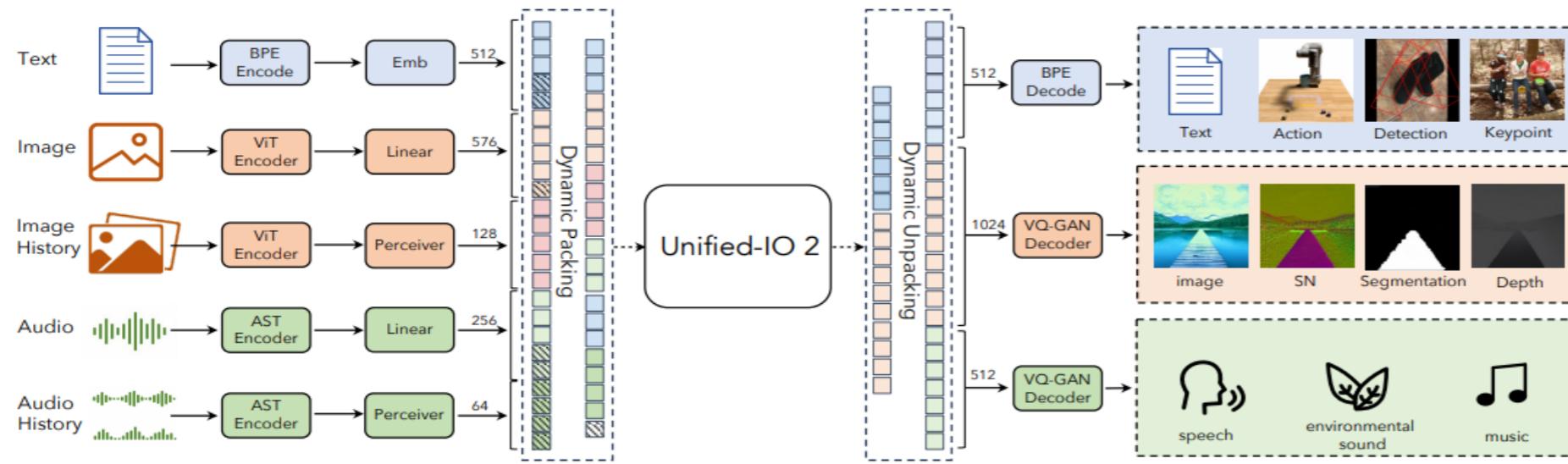
Girdhar et al., ImageBind: One Embedding Space To Bind Them All, 2023

Multimodal Large Models (2)



BLIP-2

- **Image-to-text:** Visual understanding and conversation



Unified-IO-2

- **Any-to-any:** Taking multimodal inputs and generating multimodal outputs

Li et al., BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, 2023

Lu et al., Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action, 2023



THE WEB CONFERENCE 2024 IN SINGAPORE

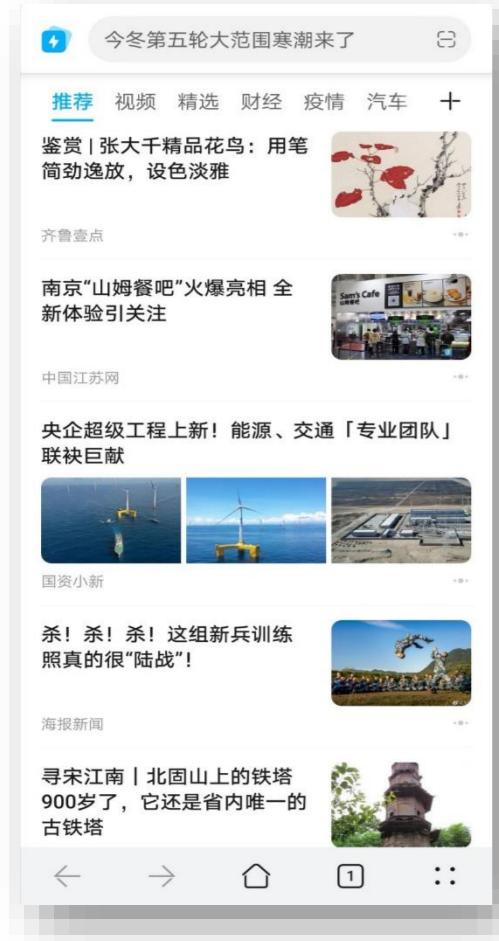
MAY 13 - 17, 2024

MULTIMODAL PRETRAINING
FOR RECOMMENDATION

Multimodal Pretraining for Recommendation

- **Domain data**
 - Item contents
 - User behavior sequences
 - User/item graphs
- **In-domain pretraining tasks**
 - Representation learning
 - Masked item prediction
 - Next item prediction
- **Downstream recommendation tasks**
 - Tagging
 - Matching: Sequential recommendation
 - Ranking: CTR prediction
 - Reranking: diversity awareness

Multimodal Recommendation Scenarios



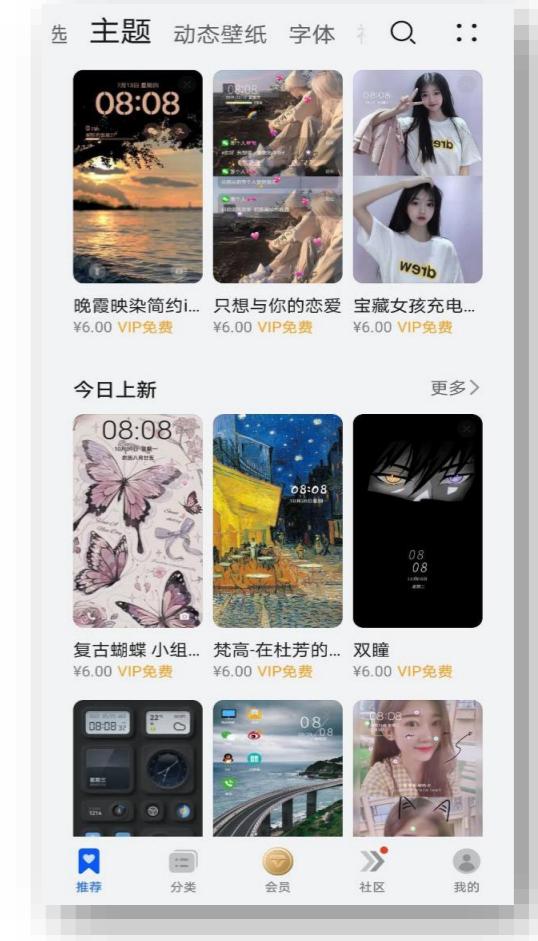
News Feed



Micro-Video



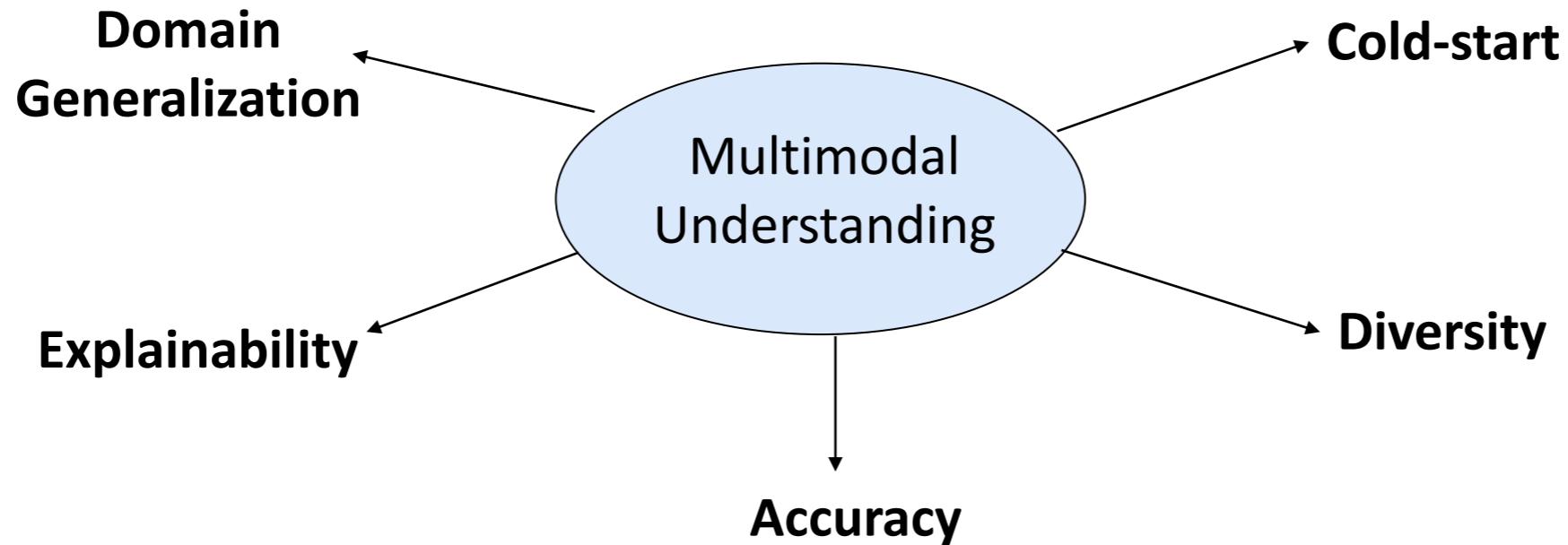
Music Streaming



Mobile Theme

And many more...

Why Multimodal Recommendation?



Why Multimodal Pretraining for Recommendation?

Multimodal pretraining for recommendation

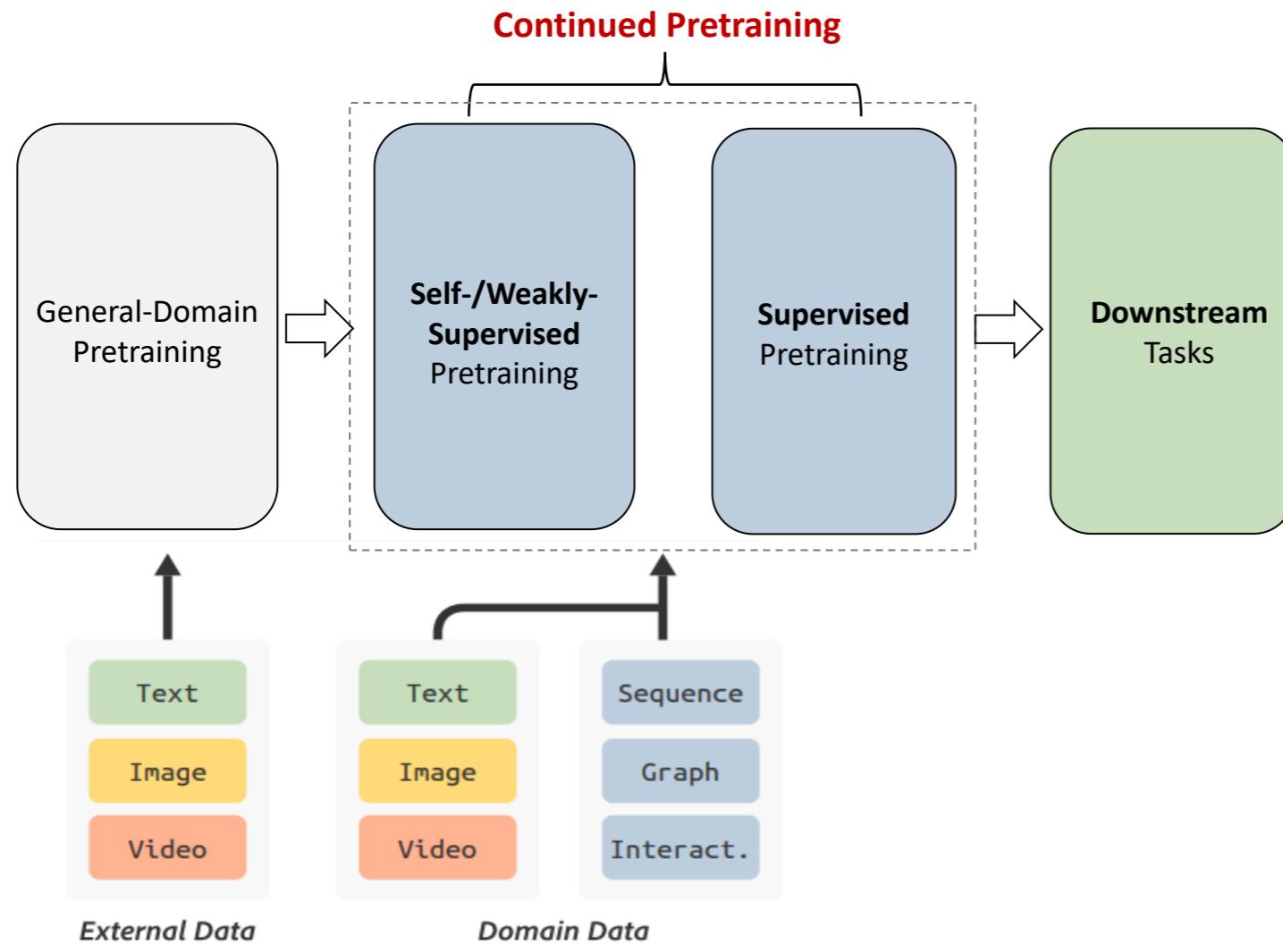
How to learn a good representation
for recommendation?



Traditional multimodal recommendation

How to learn a good recommender
given pretrained representations?

Overall Framework



Domain Data:

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

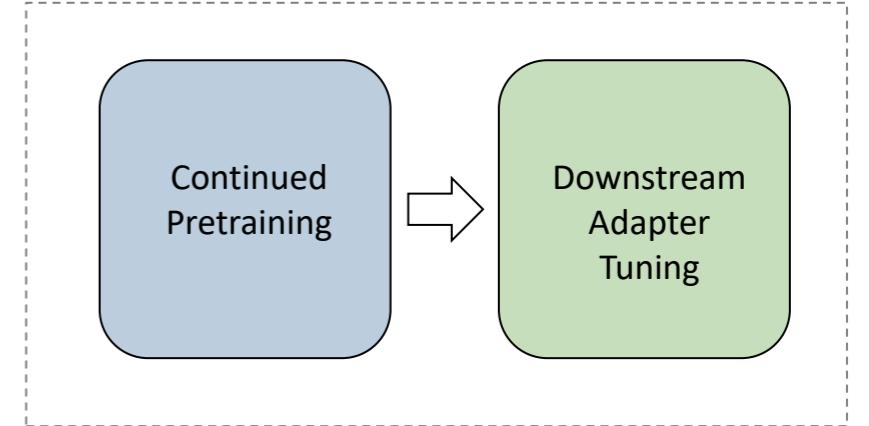
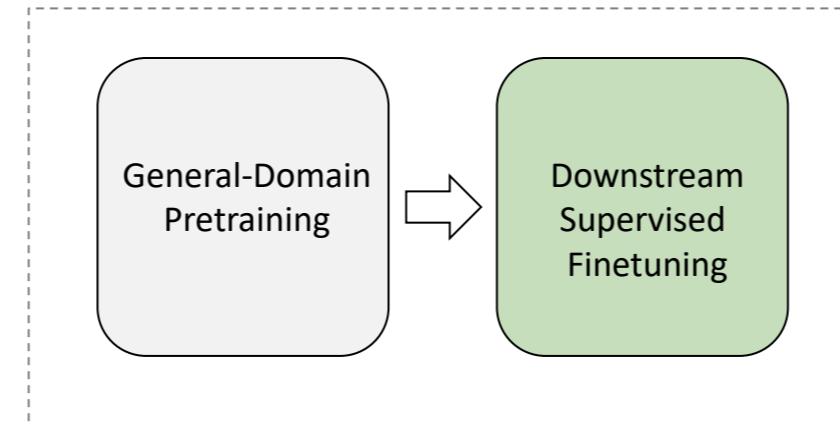
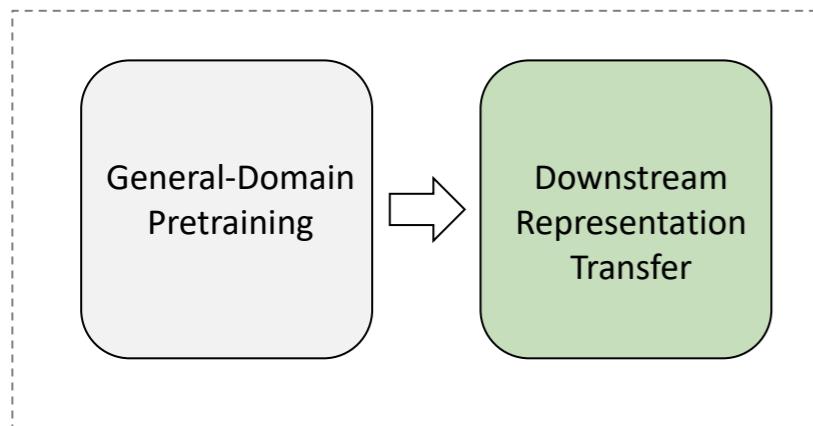
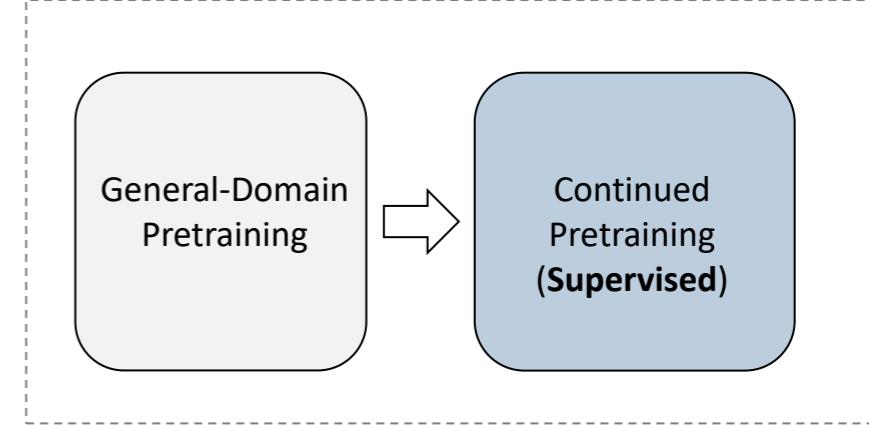
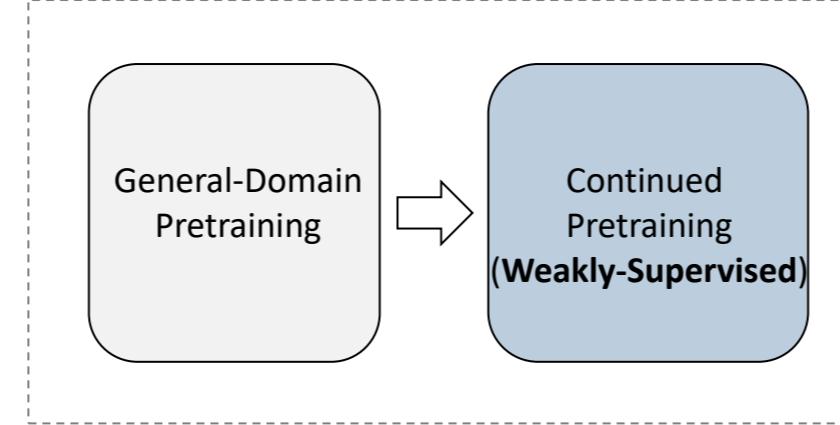
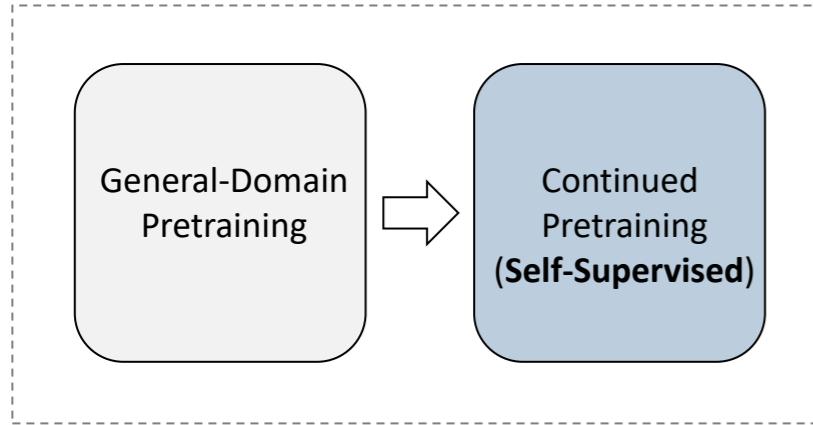
Continued Pretraining:

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

Downstream Tasks

- Representation transfer
- Supervised finetuning
- Adapter tuning

Categorization of Existing Work



C1

C2

C3

C4

C5

C6

C1: Continued Pretraining w/ Self-Supervised Tasks

Reconstructive:

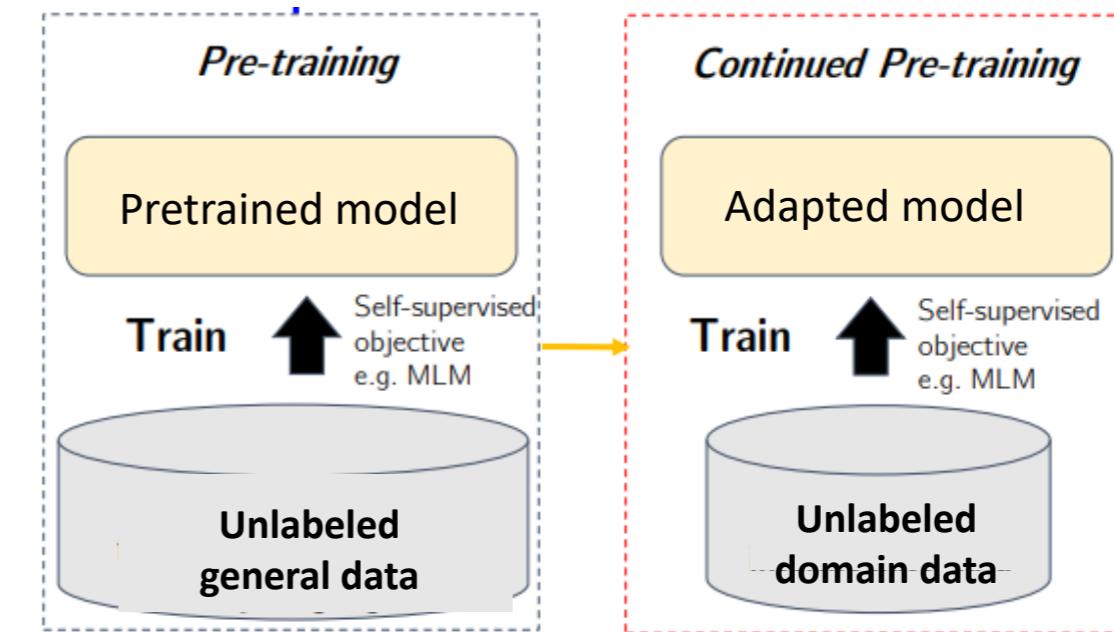
- Masked token prediction (PREC、RecoBERT)
- Masked attribute prediction (PREC、S3-Rec)
- Masked item prediction (PREC、S3-Rec)

Contrastive:

- Contrastive learning (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

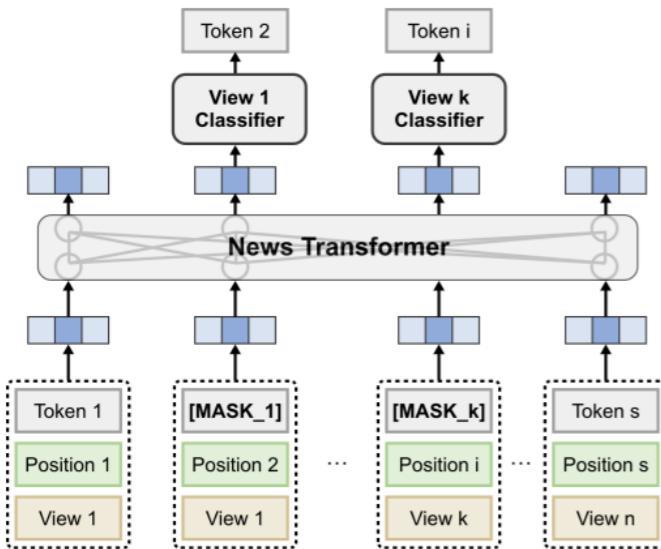
Malkiel et al., RecoBERT: A Catalog Language Model for Text-Based Recommendations, 2020

Zhou et al., S³-Rec: Self-Supervised Learning for Sequential Recommendation with Mutual Information Maximization, 2020

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

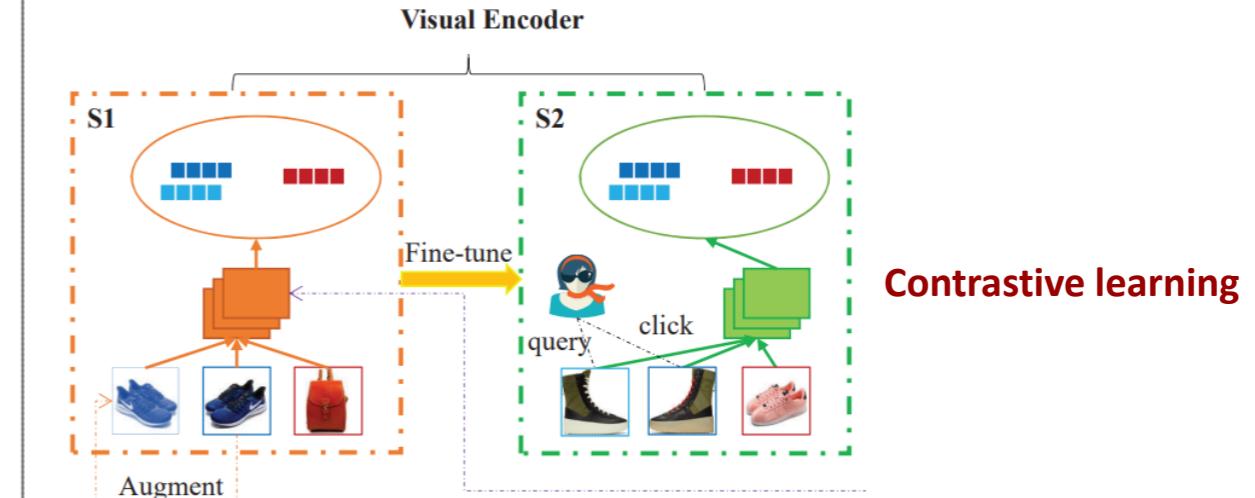
Chen et al., Visual Encoding and Debiasing for CTR Prediction, 2022

C1: Continued Pretraining w/ Self-Supervised Tasks

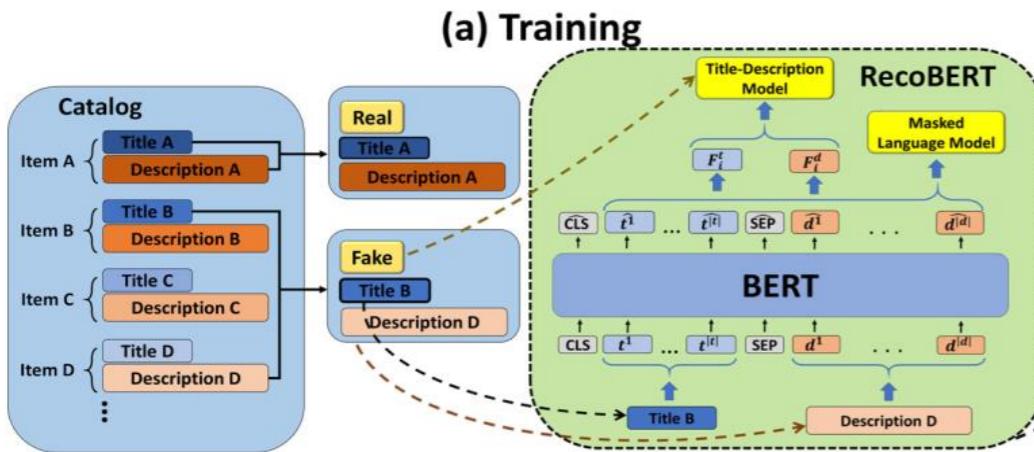


(a) Masked news token prediction.
PREC [Huawei] Liu et al., 2022

Masked token/category prediction

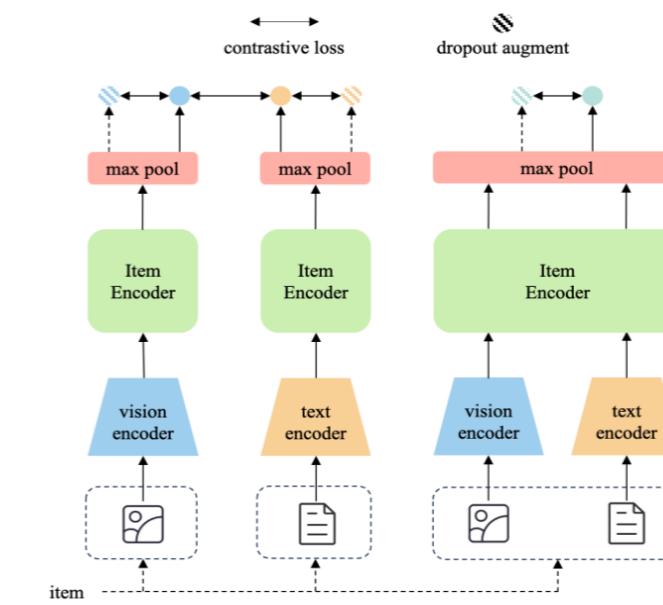


VisualEncoder Chen et al., 2022



RecoBERT Malkiel et al., 2020

Masked language model / alignment between title and description



MMSRec Song et al., 2023

Contrastive learning

Contrastive learning within modality & across modalities

C2: Continued Pretraining w/ Weakly-Supervised Tasks

Weakly-Supervised Signals:

- Category/tags/topics
- Correlated item pairs
- Knowledge graphs
- Query-document pairs
-

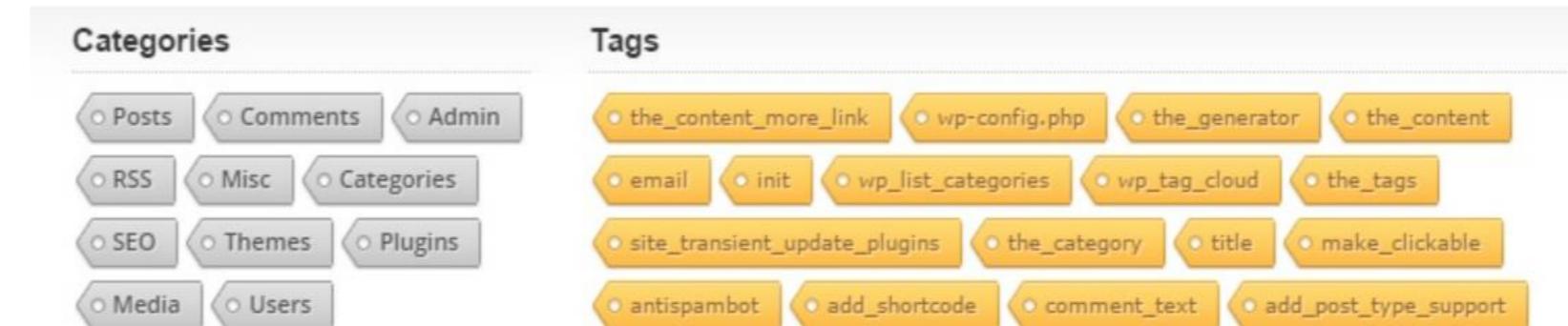


Fig. from: <https://www.wpxplorer.com/plugins-add-categories-tags/>

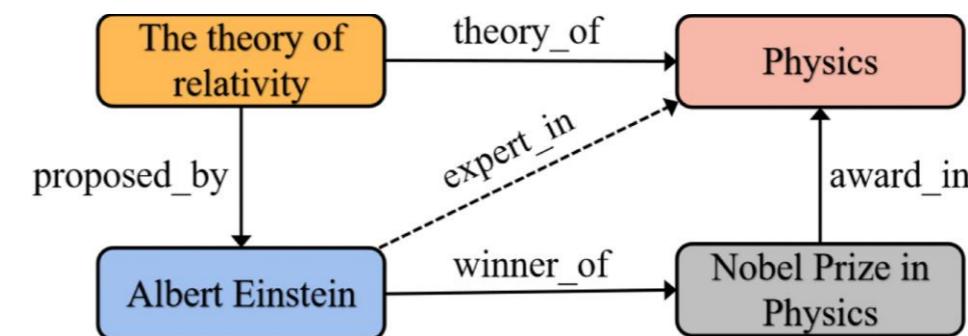
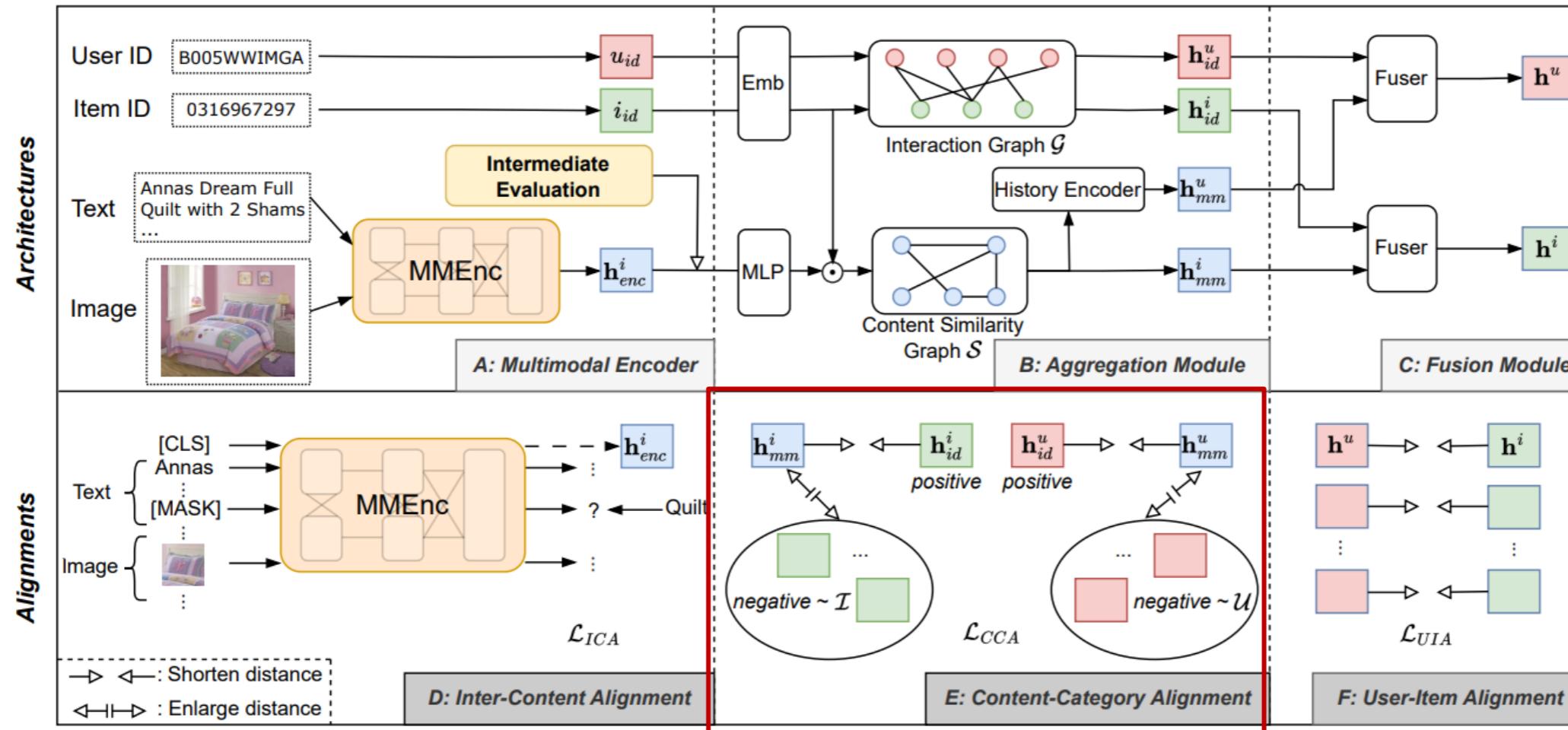


Fig. from: TDN: Triplet Distributor Network for Knowledge Graph Completion, 2023

C2: Continued Pretraining w/ Weakly-Supervised Tasks

AlignRec

- **Content-category alignment:** aligning the item representations of the same category



C2: Continued Pretraining w/ Weakly-Supervised Tasks

NewsEmbed

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

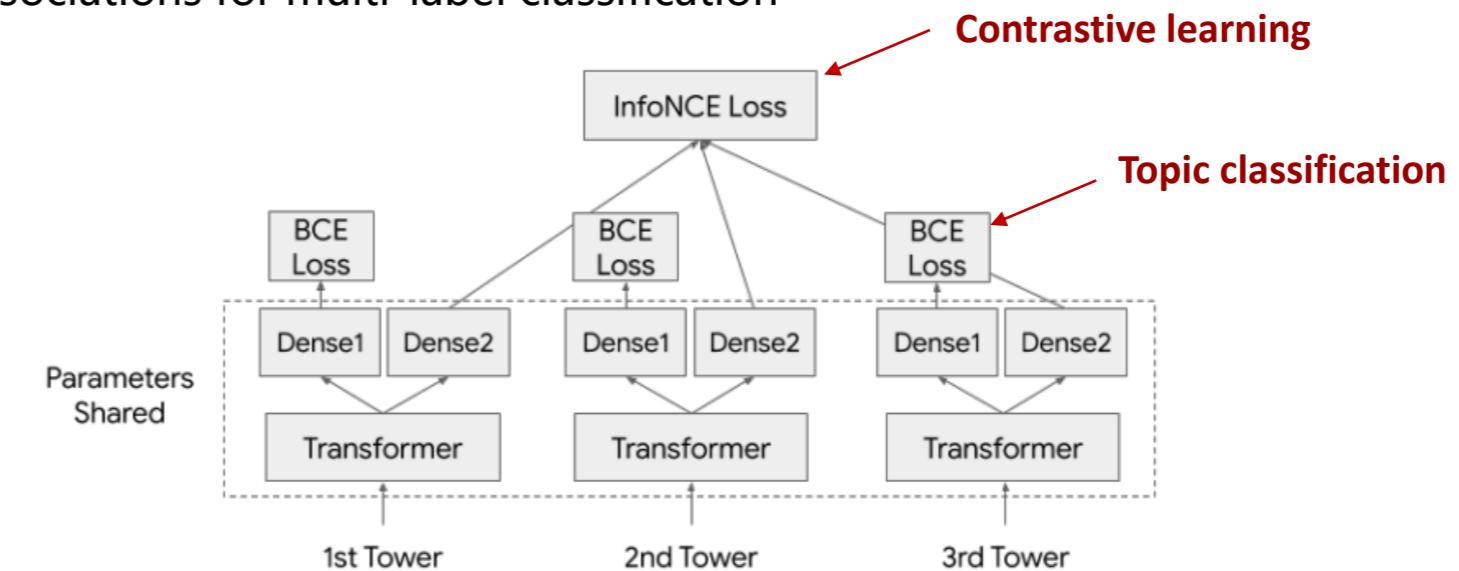
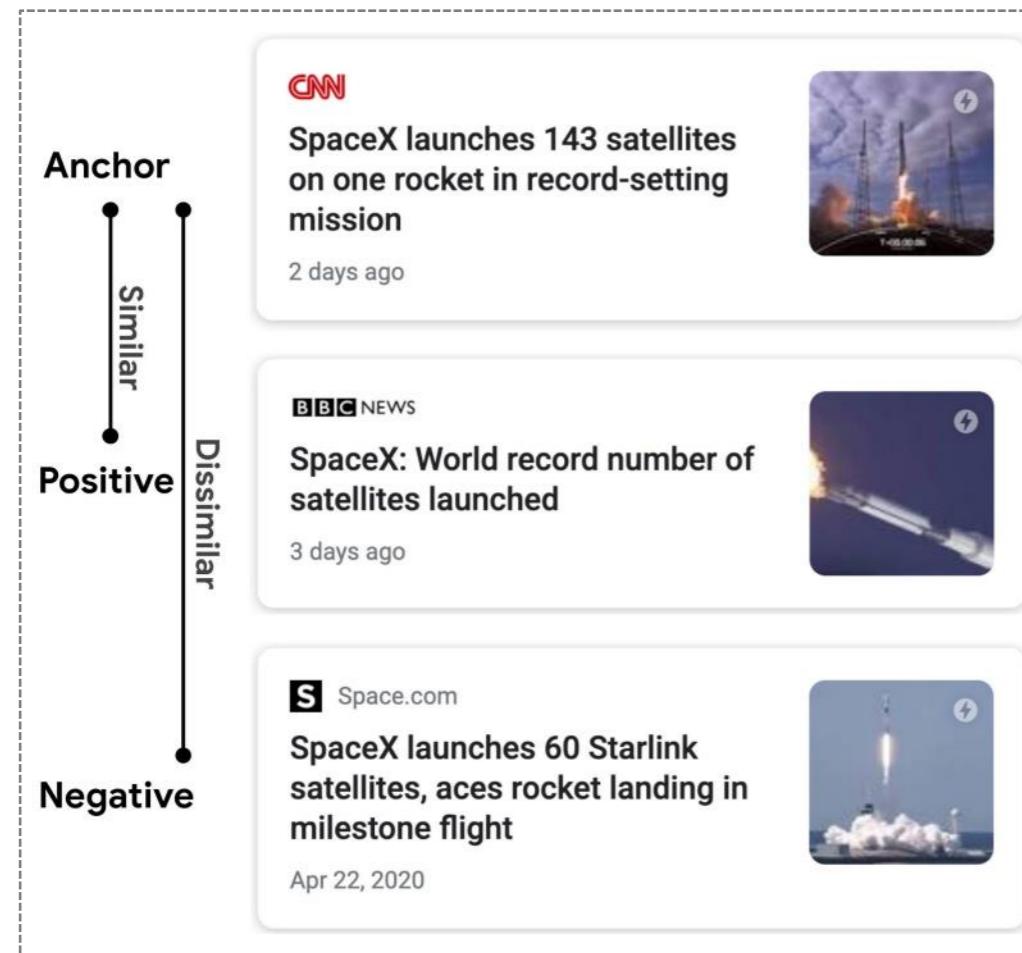


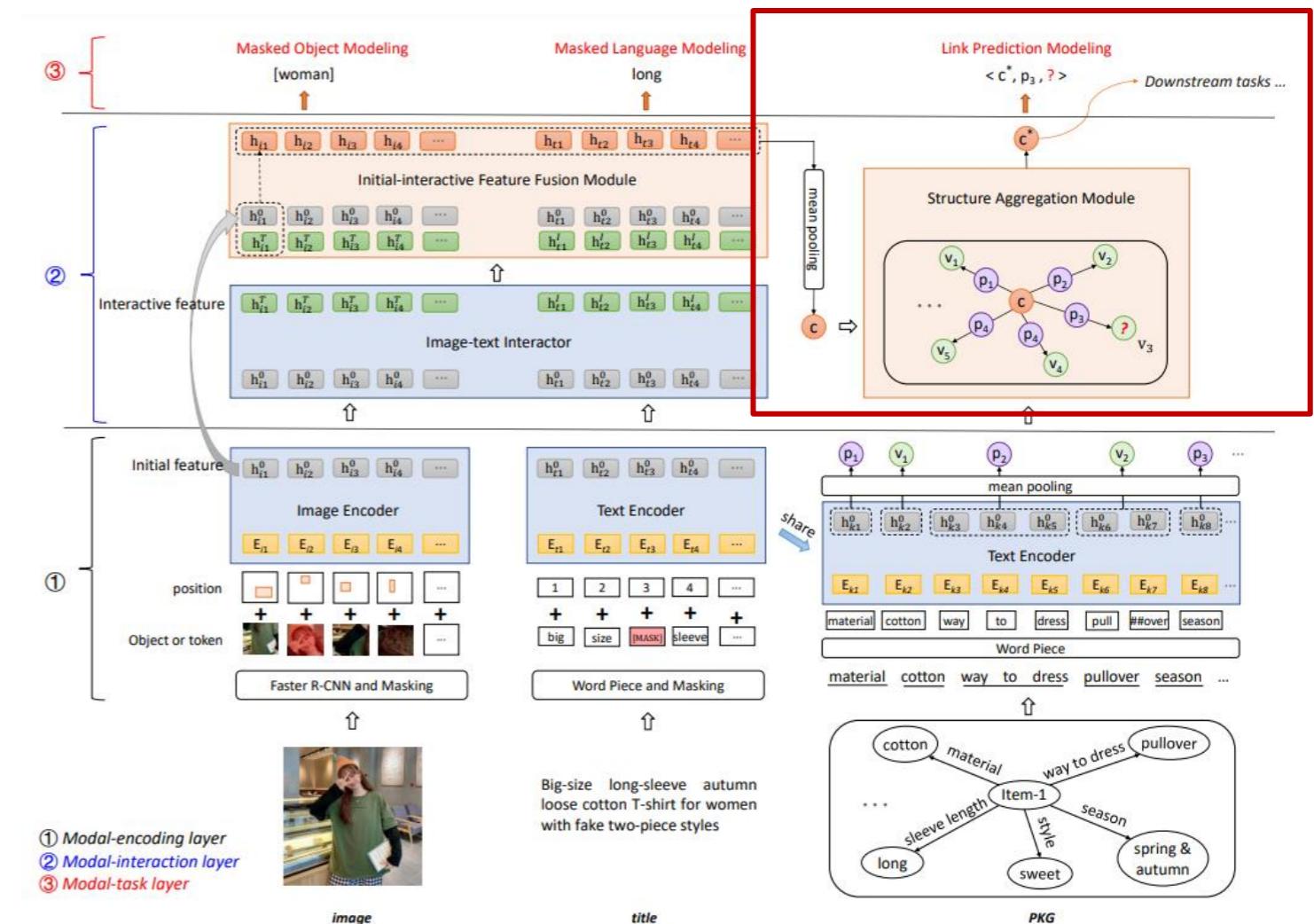
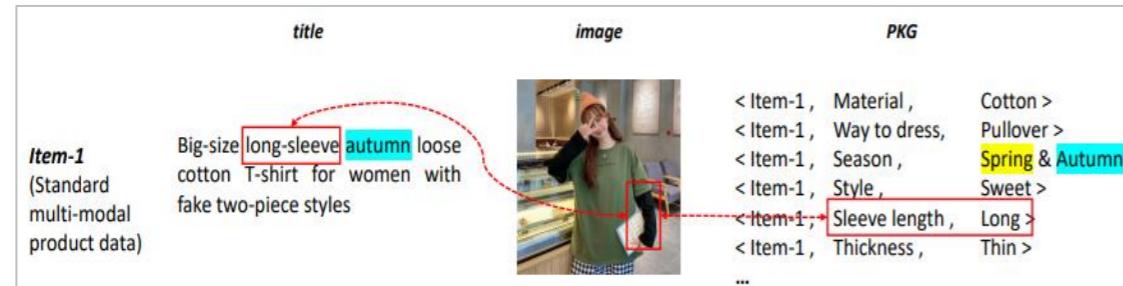
Figure 4: NewsEmbed model structure.

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

C2: Continued Pretraining w/ Weakly-Supervised Tasks

K3M

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



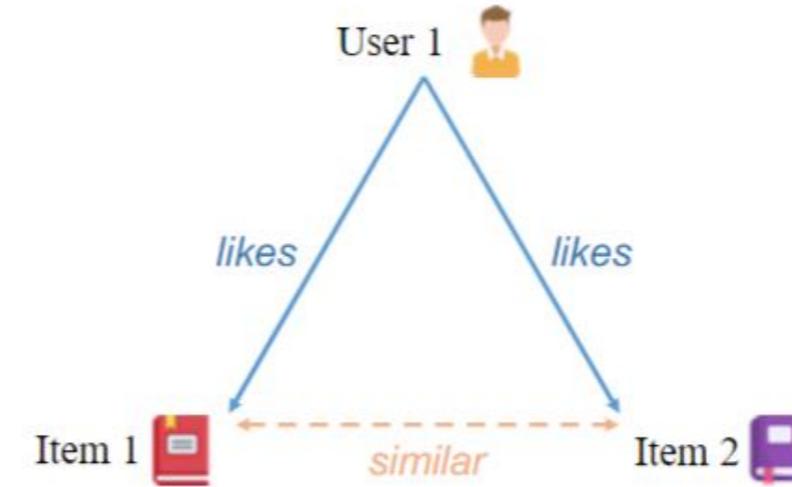
- Masked Object Modeling (MOM)
- Masked Language Modeling (MLM)
- Link Prediction Modeling (LPM)

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

C3: Continued Pretraining w/ Supervised Tasks

Supervised Signals:

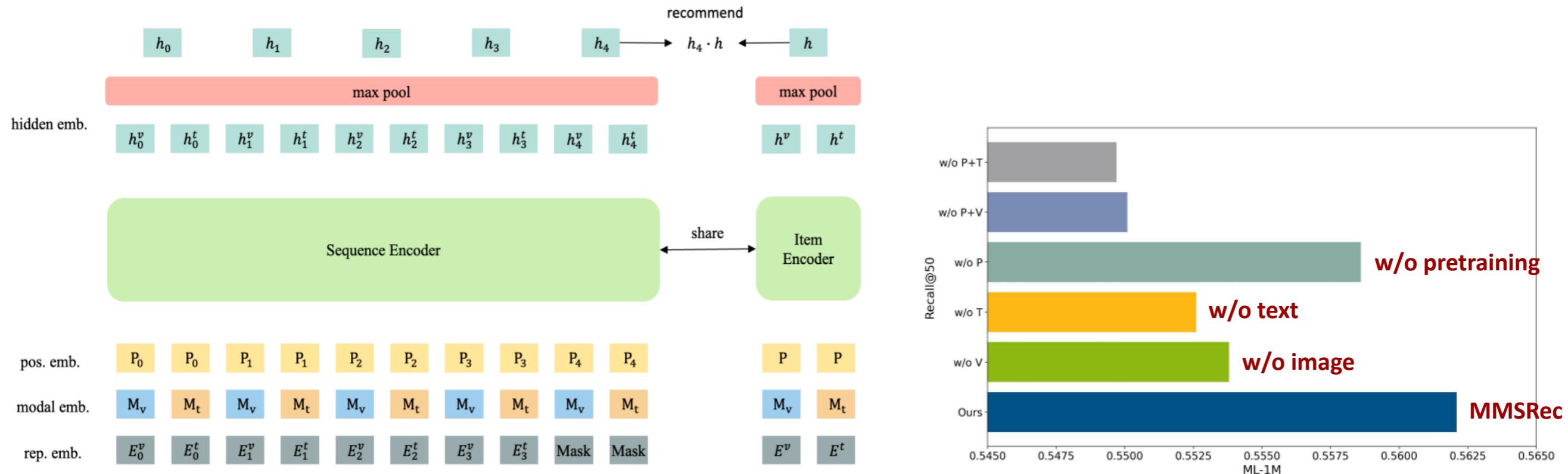
- User-to-item matching (MMSRec, MISSRec)
- Item-to-item matching (CB2CF, ItemSage)
- ID-to-modality alignment (CLCRec)
- User-to-item cross-encoder



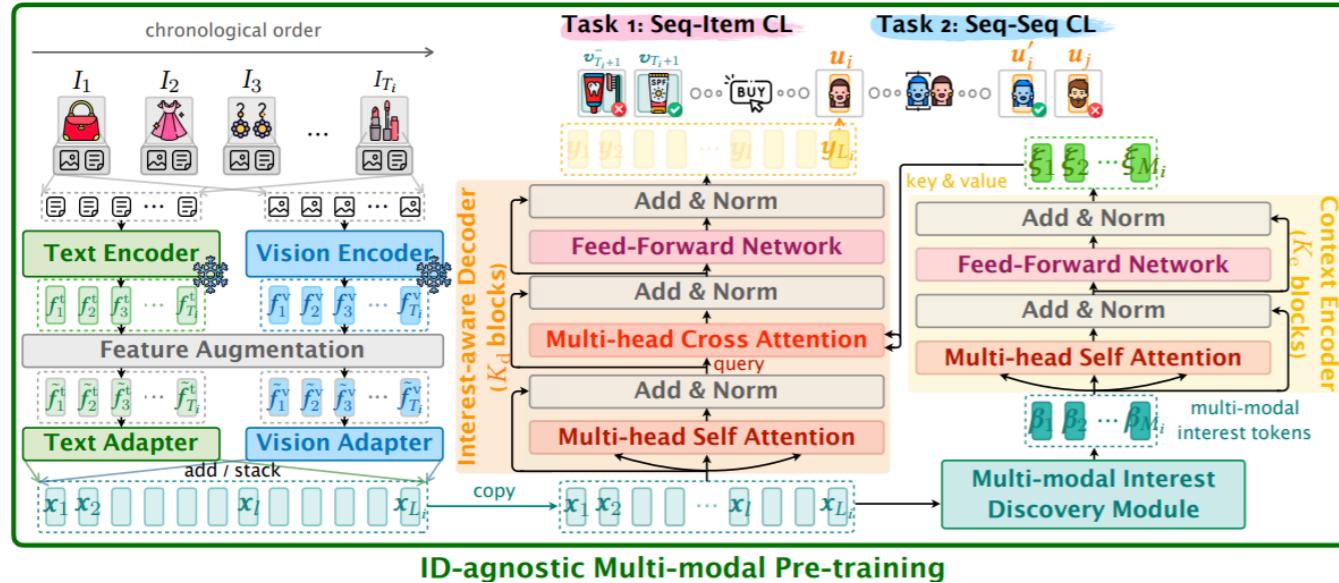
C3: Continued Pretraining w/ User-to-Item Matching

MMSRec

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

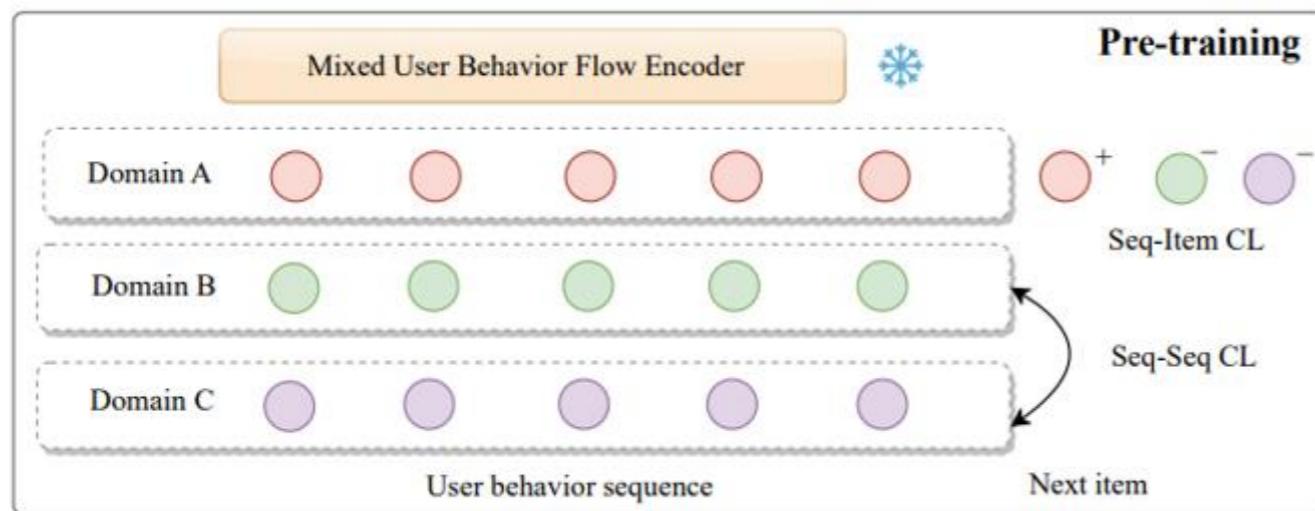


C3: Continued Pretraining w/ User-to-Item Matching



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



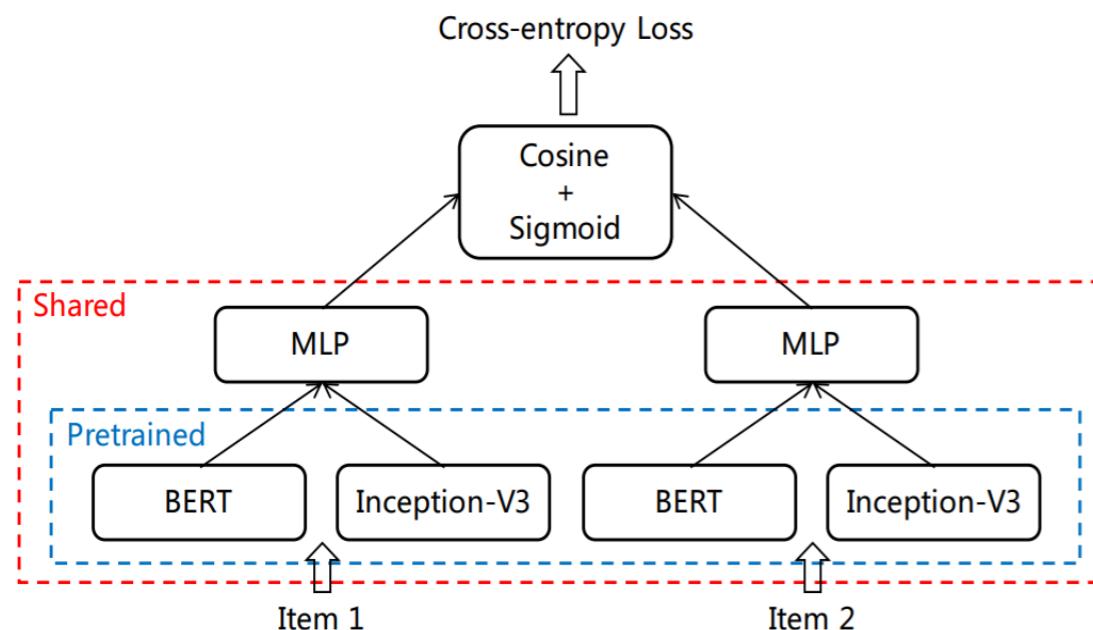
UniM2Rec

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

C3: Continued Pretraining w/ Item-to-Item Matching

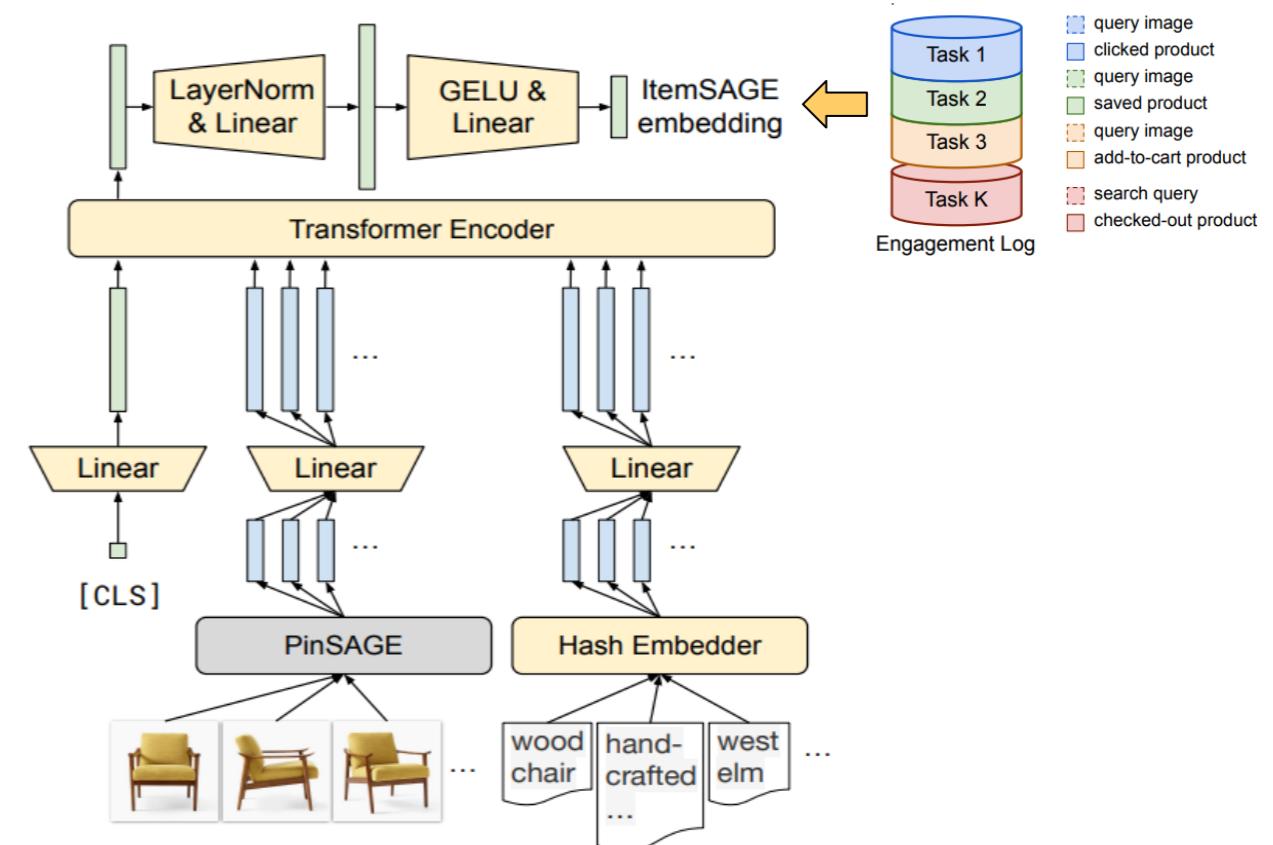
SSD

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



ItemSage

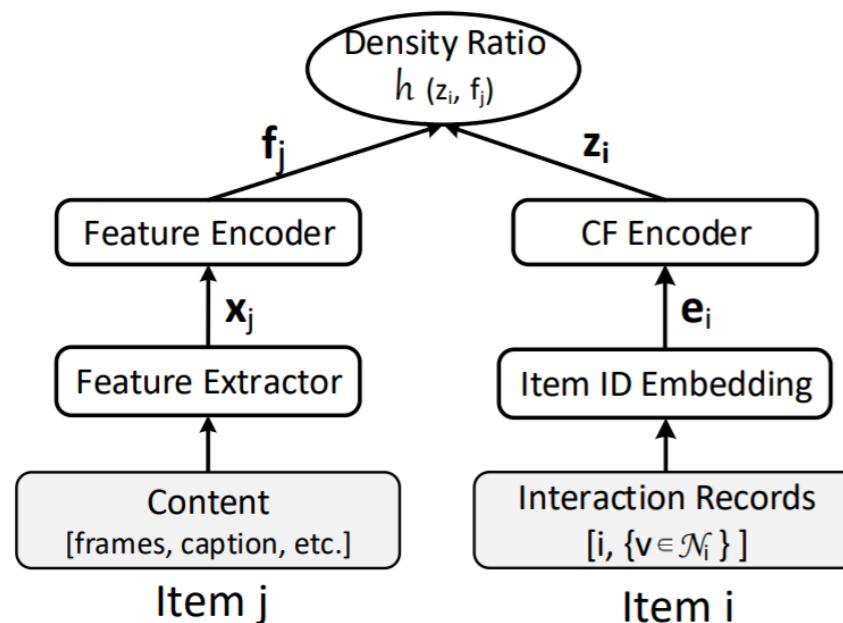
- Applying all engagement logs with multi-tasks to train multimodal item embeddings



C3: Continued Pretraining w/ ID-to-Modality Alignment

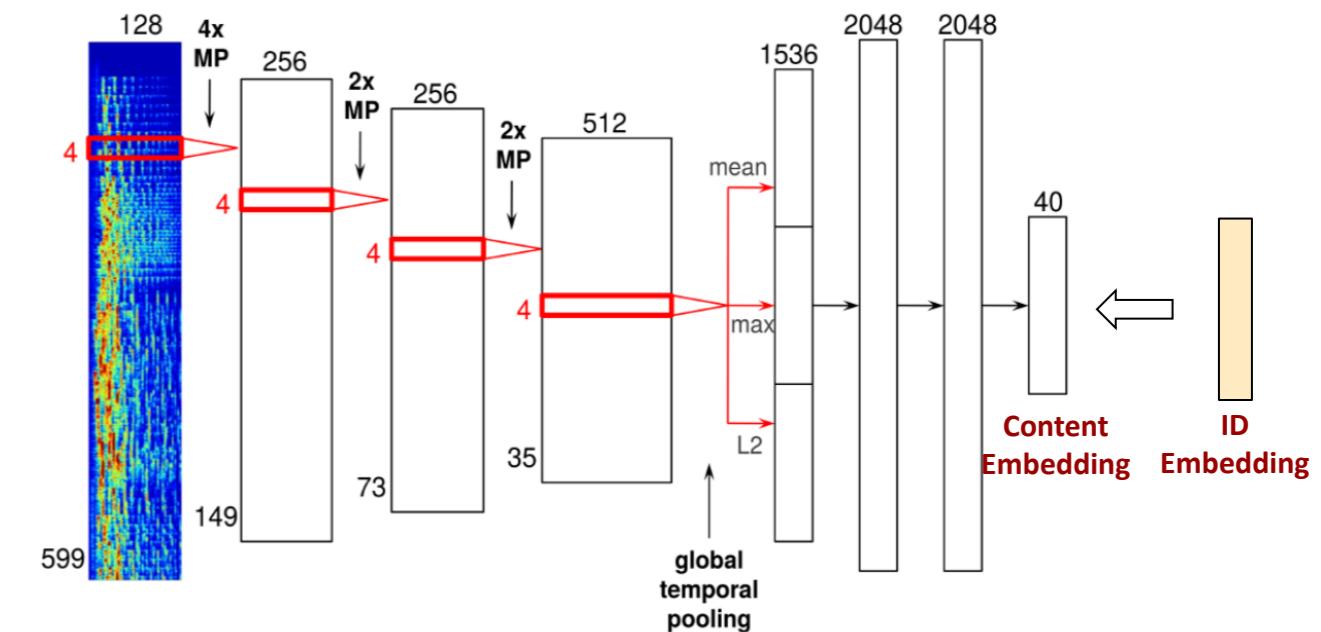
CLCRec

- Contrastive learning between CF encoder and content encoder



DCMR

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



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

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

Summarization

Continued Pretraining	Quality	Cost
Self-Supervised	Representation quality [★] Representation robustness [★★★]	Training cost [\$\$] Large unlabeled corpus
Weakly-Supervised	Representation quality [★★] Representation robustness [★★]	Training cost [\$] Small labelled data
Supervised	Representation quality [★★★] Representation robustness [★]	Training cost [\$\$\$] Huge engagement logs

Transfer to Downstream Tasks

Representation transfer:

- Multimodal features as side information
- User representation learning from multimodal sequences
- Multimodal sequence denoising and aggregation

Joint Finetuning:

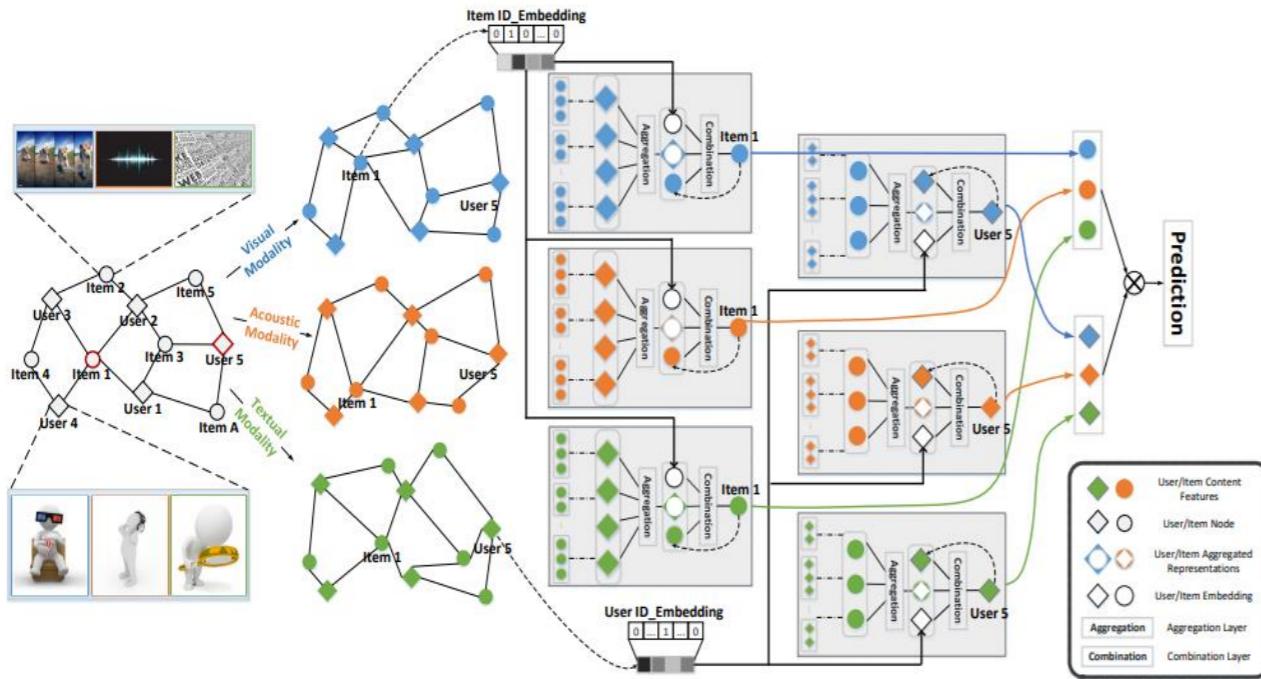
- Only finetuning item-side encoder
- Finetuning both item-side and user-side encoders
- Finetuning user-item cross encoder

Adapter Tuning:

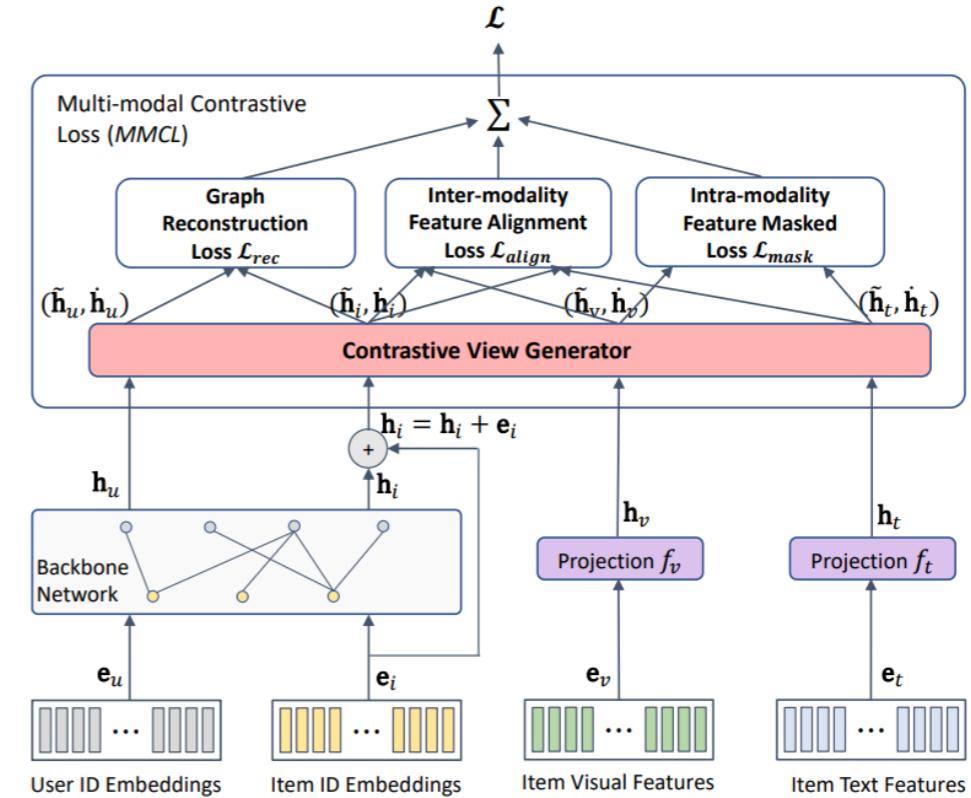
- Multimodal fusion adapter
- LLM adapter for multimodal recommendation

C4: Representation Transfer

- Multimodal features as side information



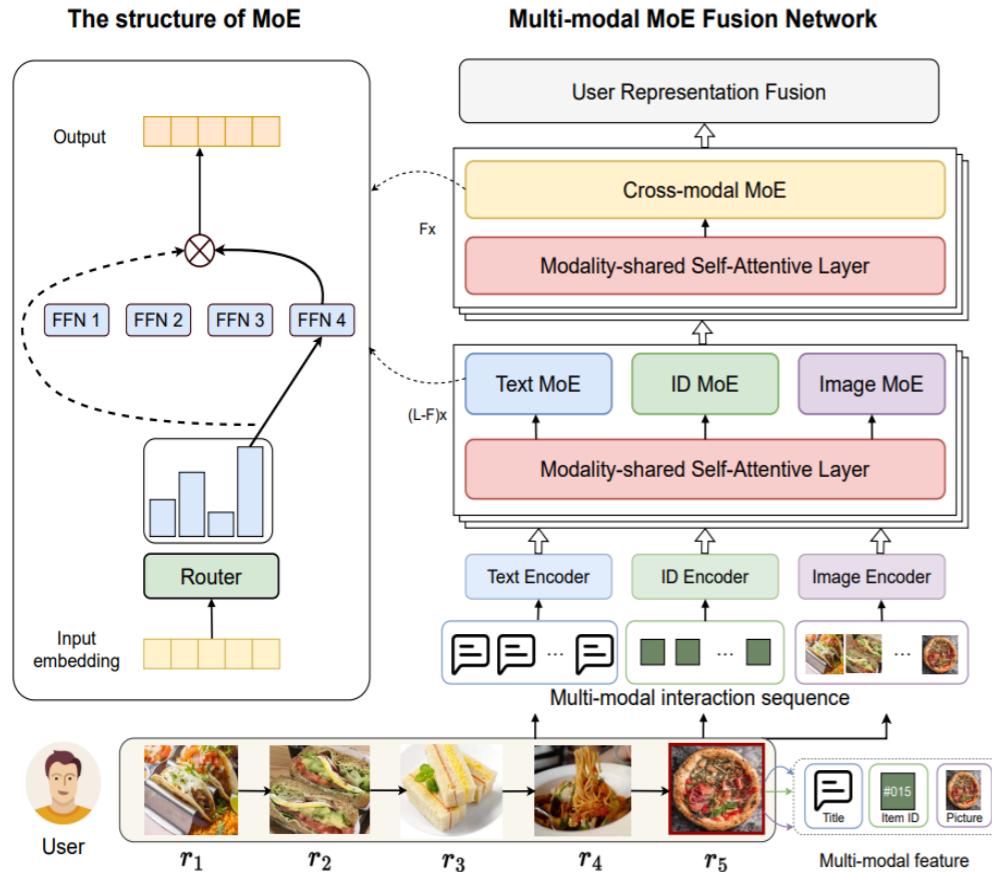
MMGCR: leveraging multimodal graph networks



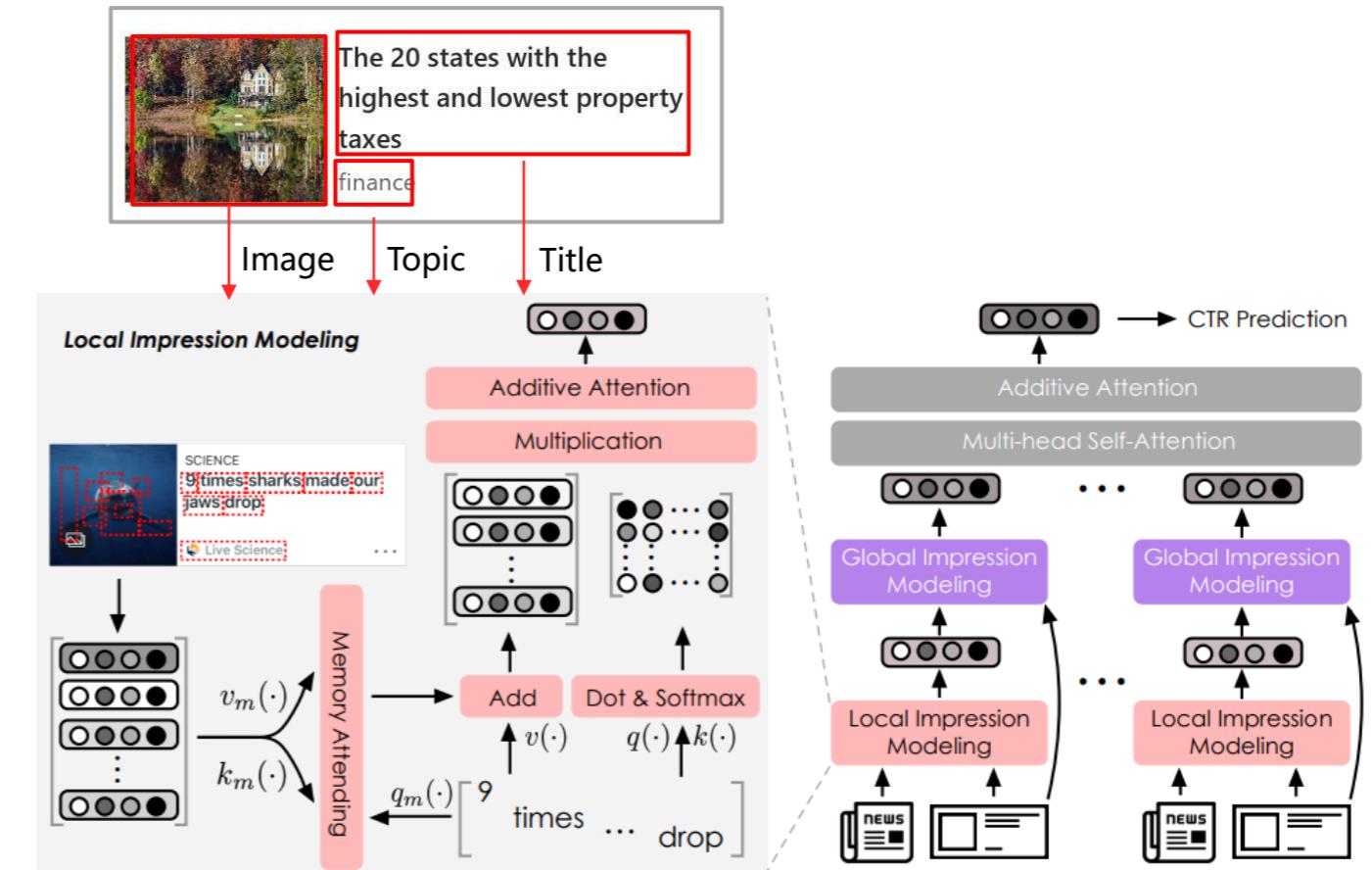
BM3: leveraging ID-modality alignment

C4: Representation Transfer

- User representation learning from multimodal sequences



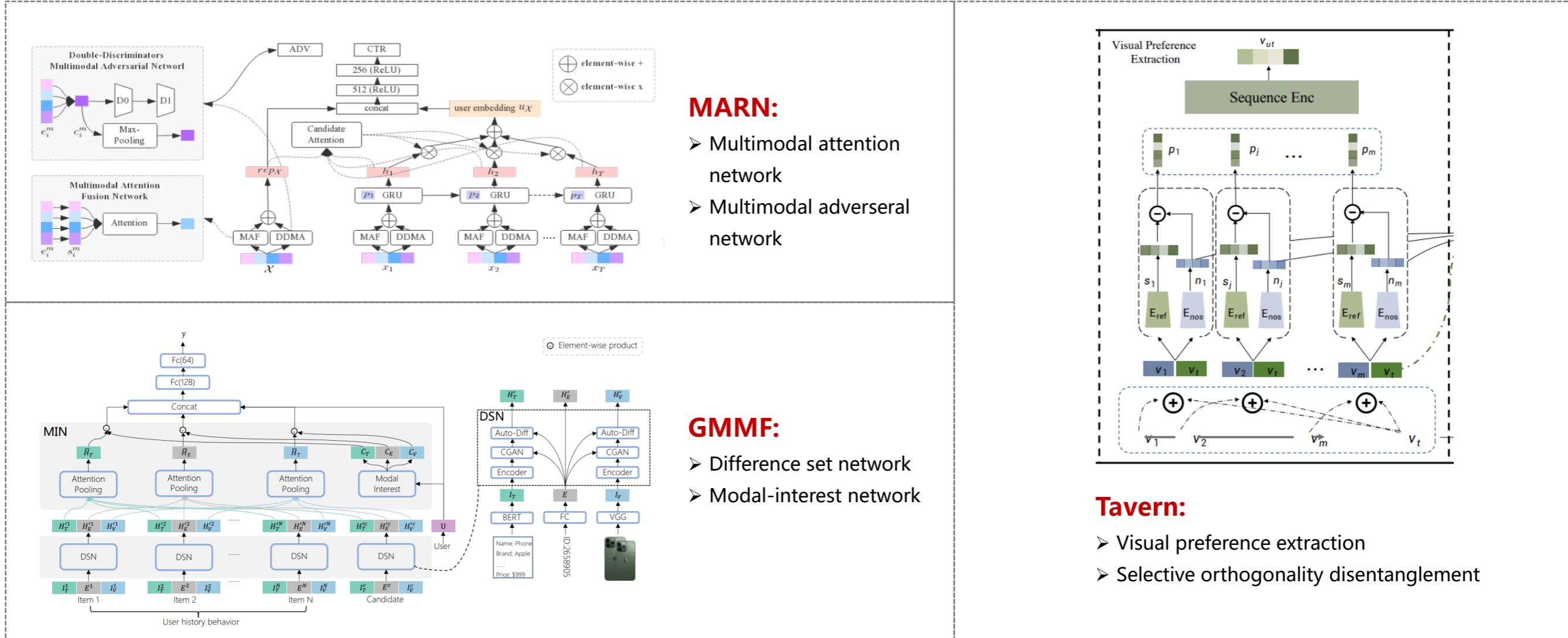
M3SRec: MOE-based multimodal fusion



IMRec [Huawei]: local and global fusion

C4: Representation Transfer

- Multimodal sequence denoising and aggregation



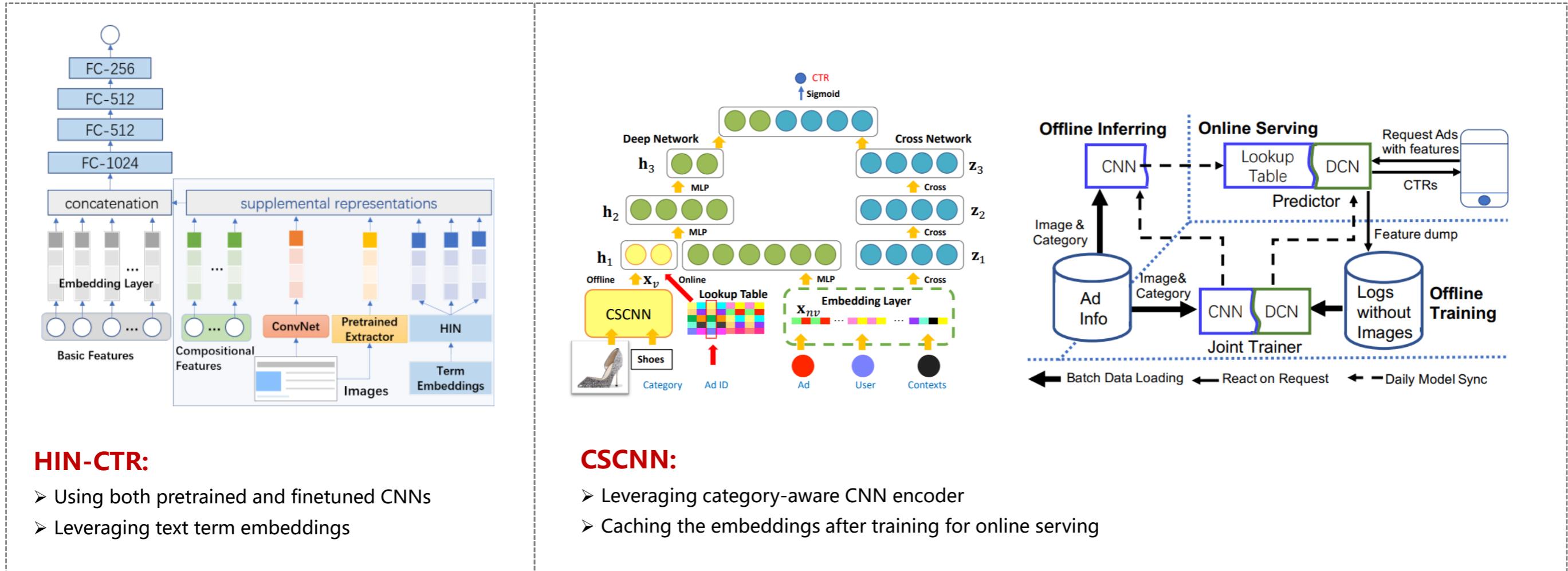
Li et al., Adversarial Multimodal Representation Learning for Click-Through Rate Prediction, 2020

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

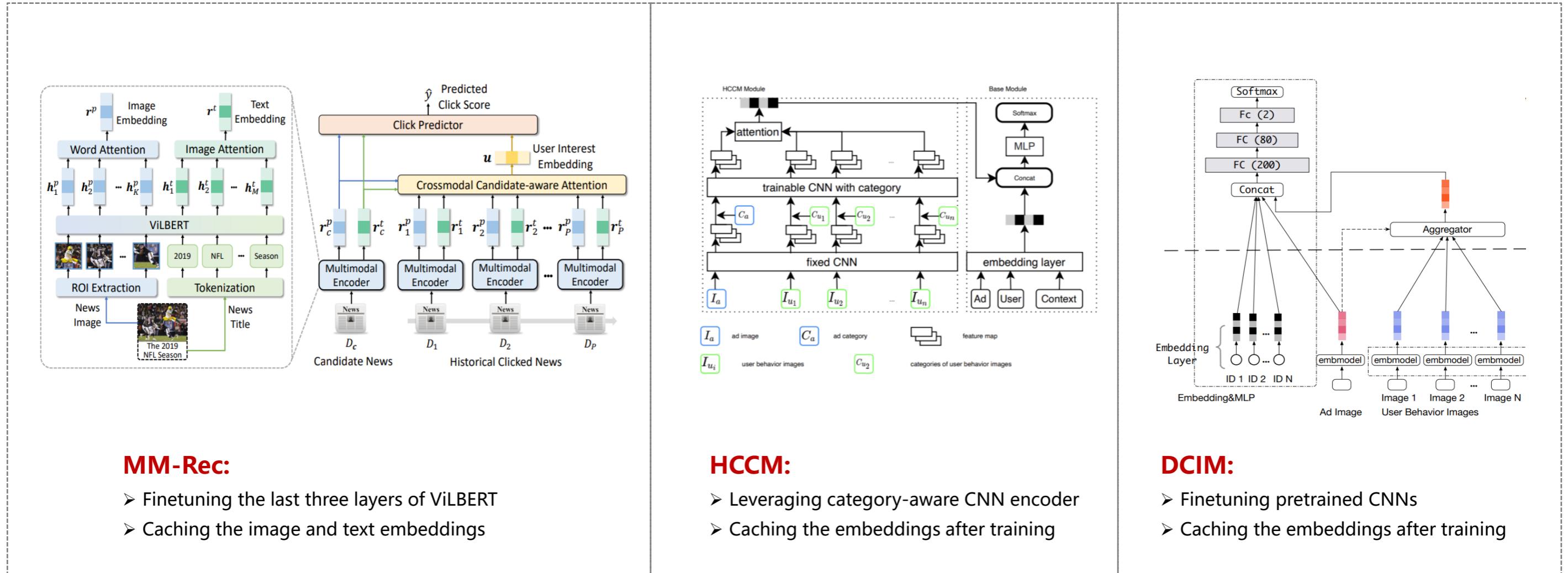
C5: Joint Finetuning

- Only finetuning item-side encoder



C5: Joint Finetuning

- Finetuning both item-side and user-side encoders



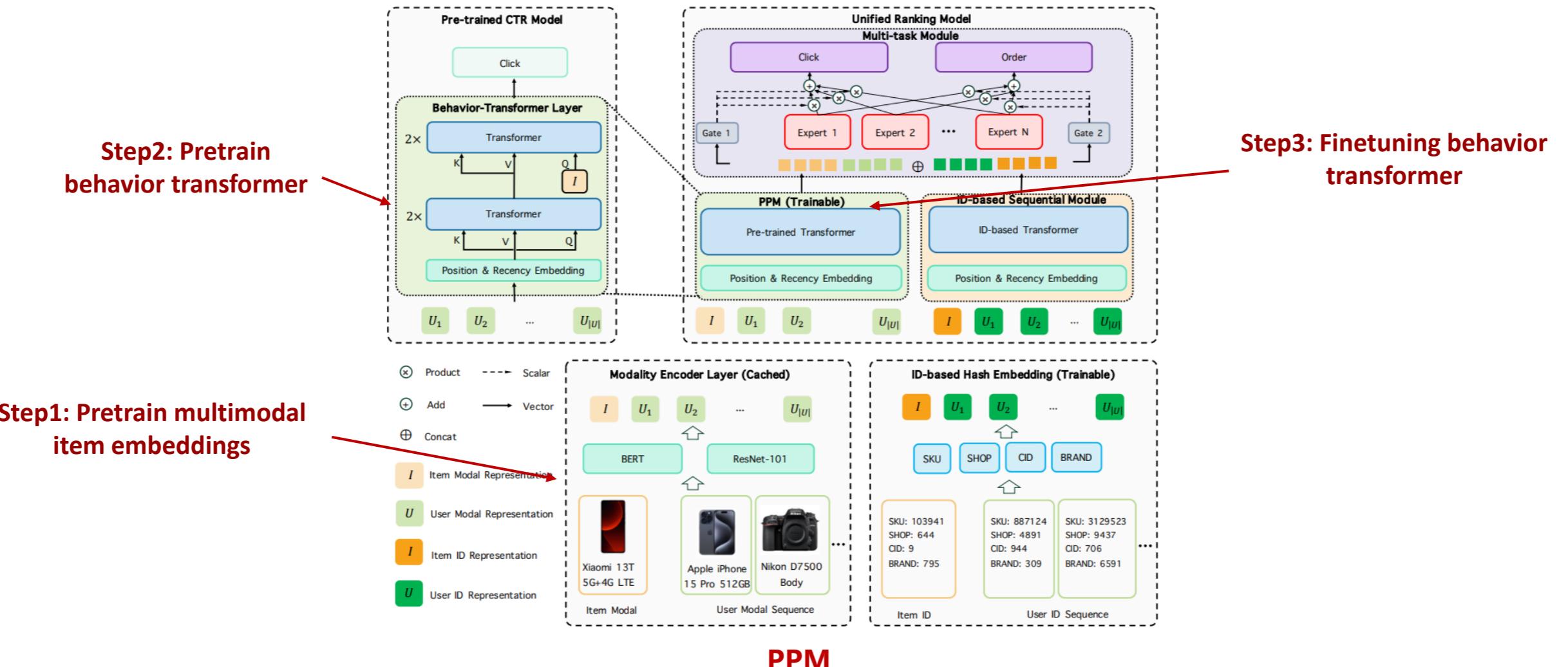
Wu et al., MM-Rec: Multimodal News Recommendation, 2022

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

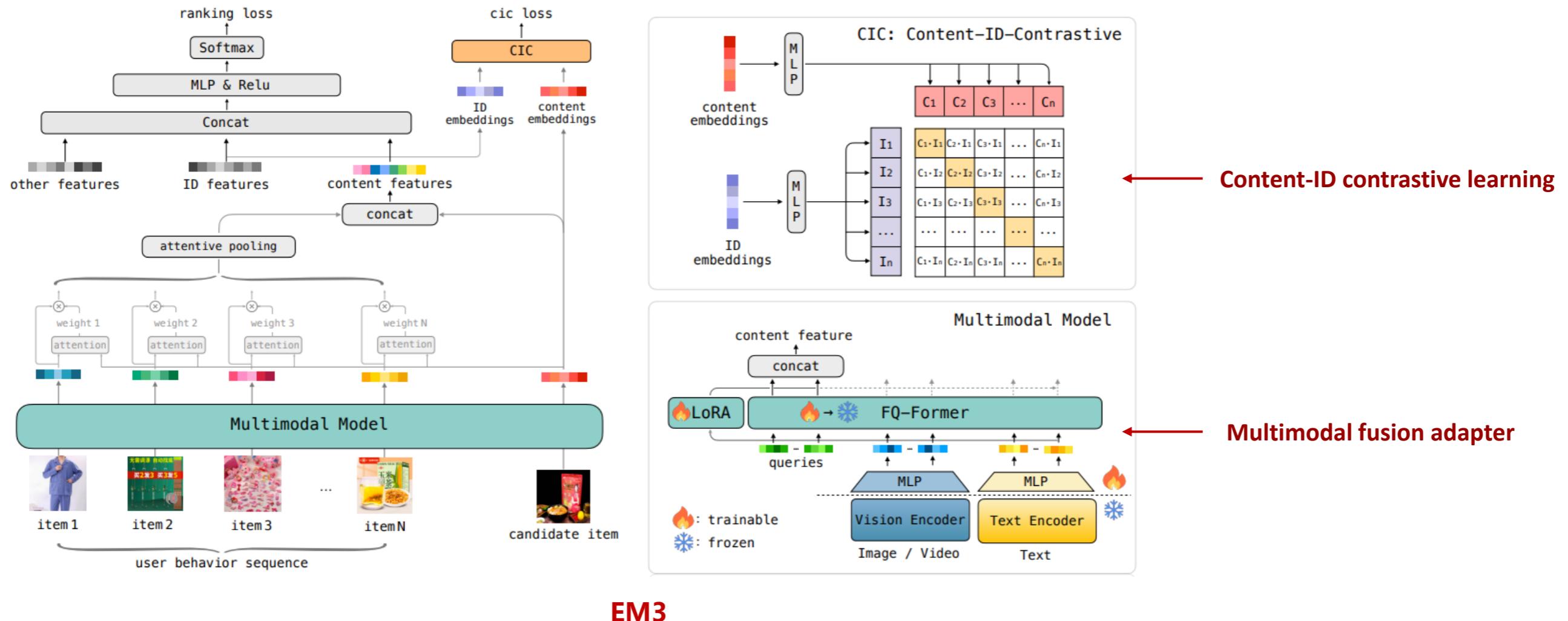
C5: Joint Finetuning

- Finetuning user-item cross encoder



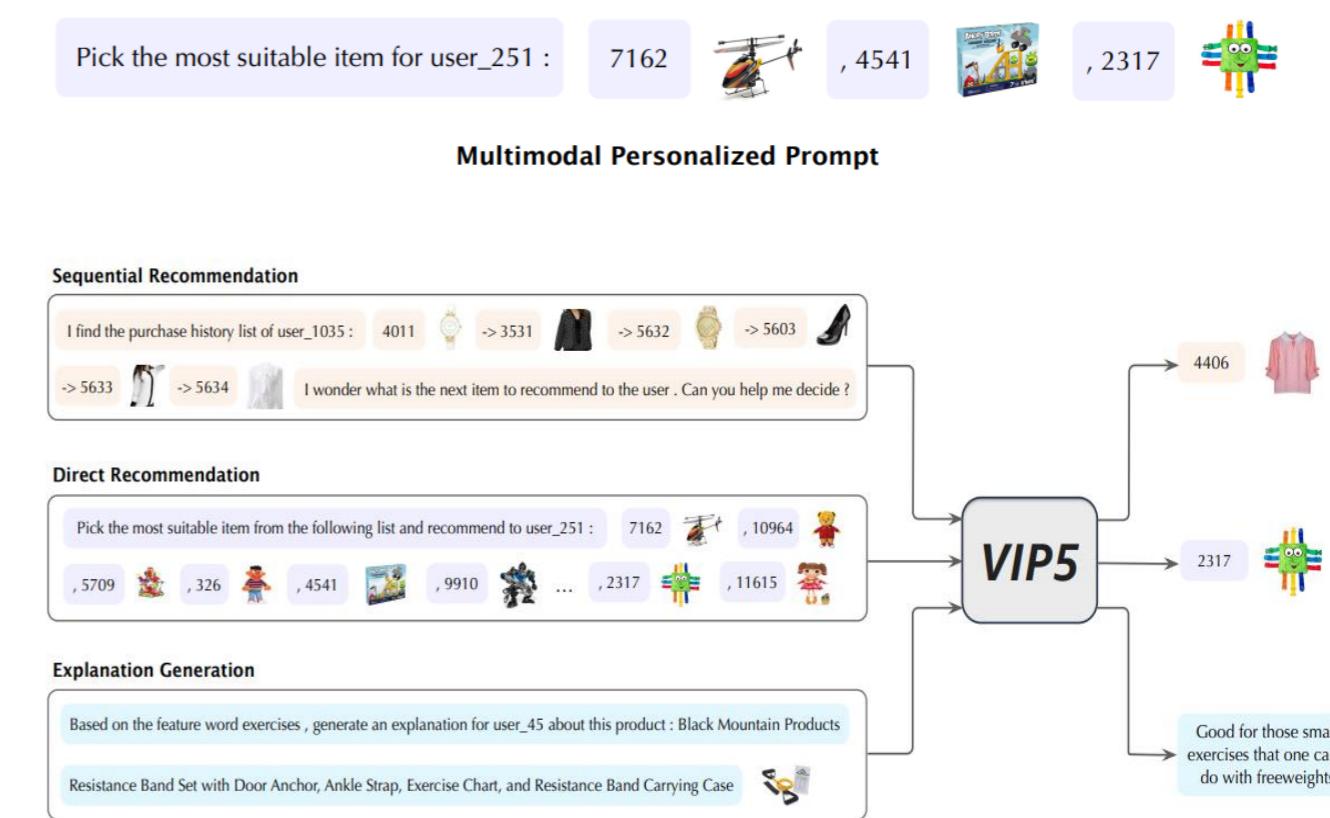
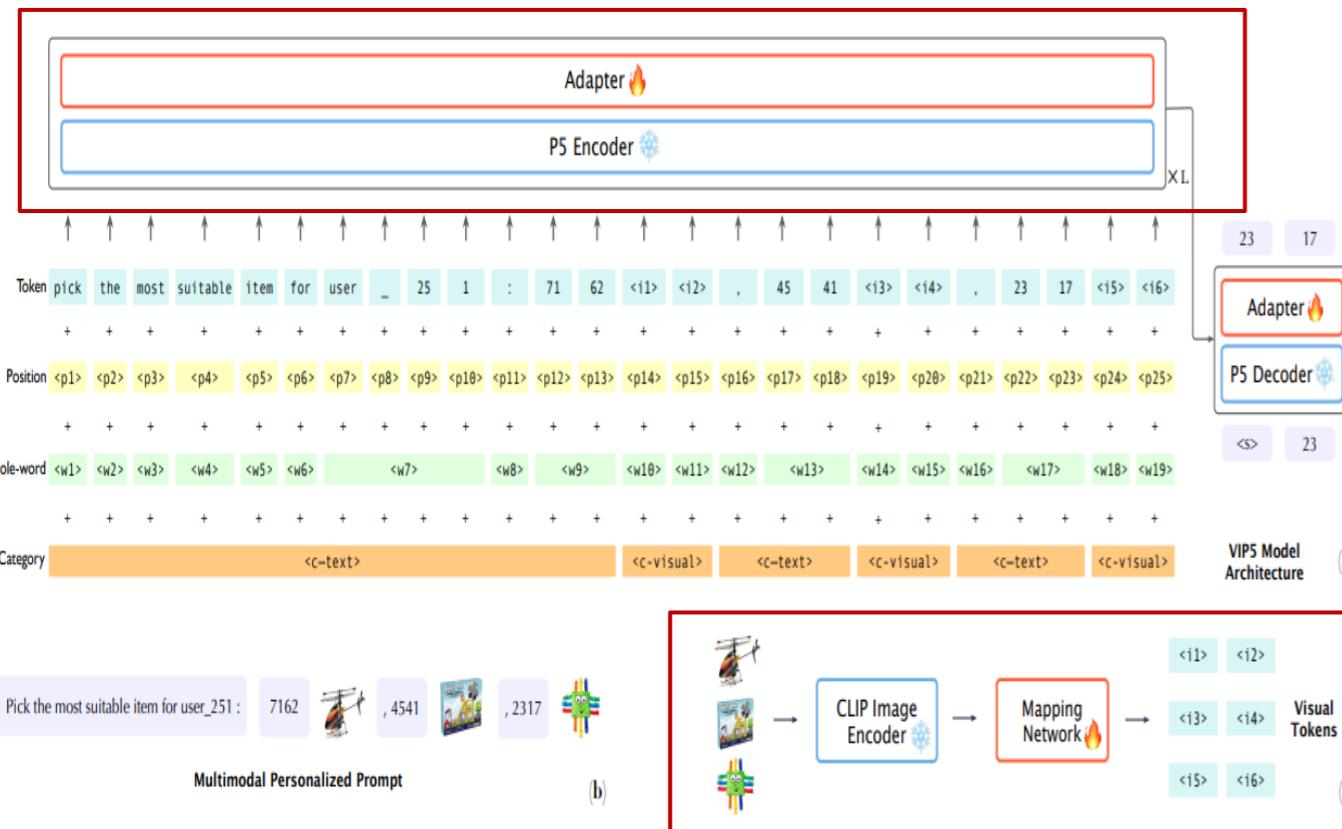
C6: Adapter Tuning

➤ Multimodal fusion adapter



C6: Adapter Tuning

- LLM adapter for multimodal recommendation



Summarization

Downstream Adaptation	Quality	Cost
Representation Transfer	Alignment quality [★]	Training cost [\$]
Joint Finetuning	Alignment quality [★★★]	Training cost [\$\$\$]
Adapter Tuning	Alignment quality [★★]	Training cost [\$\$]



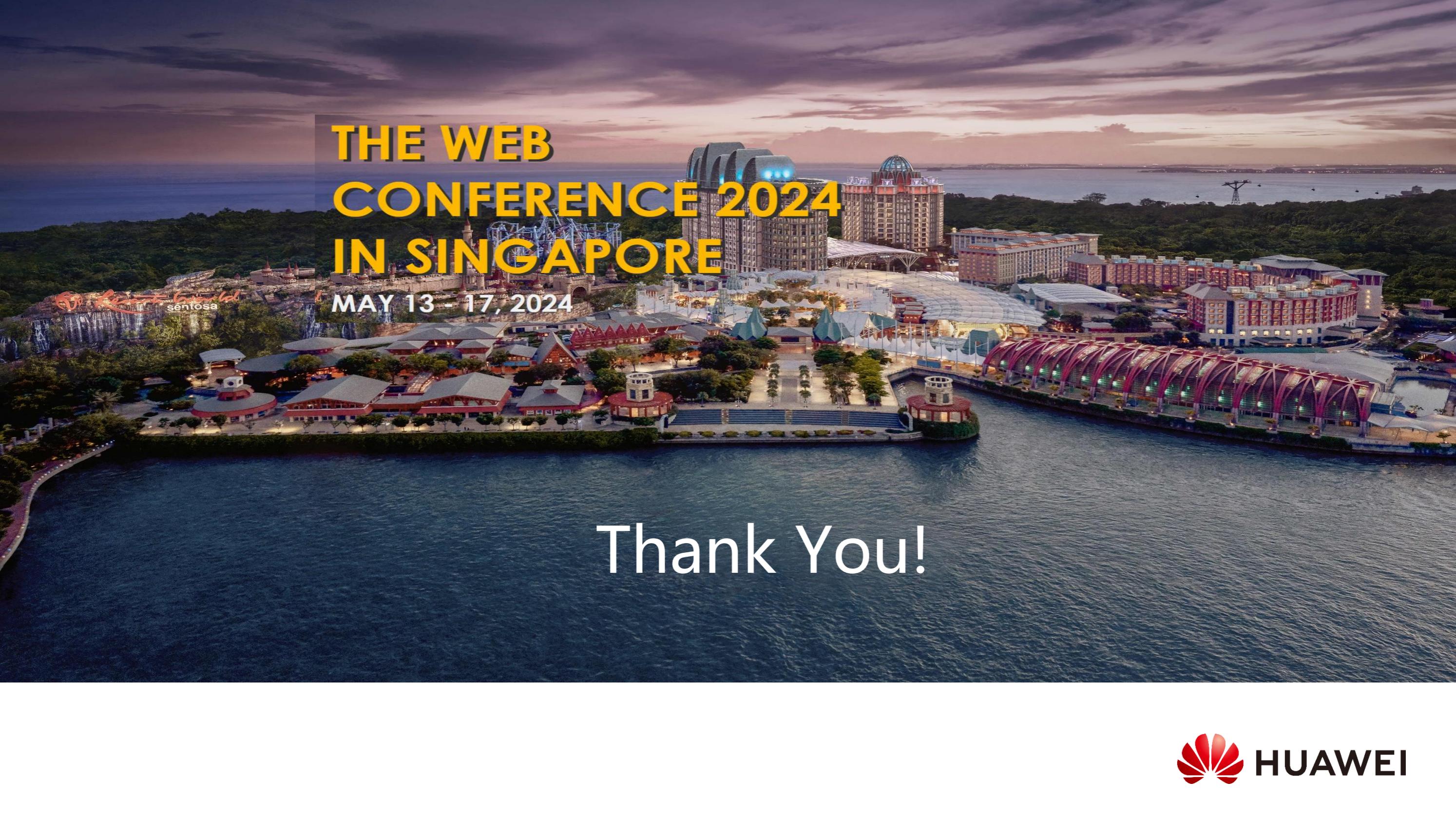
THE WEB CONFERENCE 2024 IN SINGAPORE

MAY 13 - 17, 2024

MULTIMODAL PRETRAINING
FOR RECOMMENDATION

Potential Future Directions

- **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/...
 - Amazon/MIND/PixelRec/MicroLens



THE WEB CONFERENCE 2024 IN SINGAPORE

MAY 13 - 17, 2024

Thank You!

