Multimodal Pre-training and Generation for Recommendation

Jieming Zhu
Huawei Noah's Ark Lab
jiemingzhu@ieee.org

Chuhan Wu
Huawei Noah's Ark Lab
wuchuhan1@huawei.com

Rui Zhang
www.ruizhang.info
rayteam@yeah.net

Zhenhua Dong
Huawei Noah's Ark Lab
dongzhenhua@huawei.com

Outline of Tutorial

- **Introduction**
- Multimodal Pre-training for Recommendation
- Multimodal generation for personalization
- Practices and Open Challenges in Multimodal Recommendation

Motivation of tutorial

- Classic recommendation models are trained by ID feature, categorical features, which are good at modeling the collaborative signals, but they always overlook the raw contents across multiple modalities such as text, image, audio and video.
- The recent advancement in pre-trained and generation multimodal models, like LLM, CLIP, ChatGPT, DALL.E, offer new opportunities in understanding item and user, and developing better recommender systems.
- We would like to share our research, survey and industrial practices in multimodal recommendations and propose challenging questions:
  - *How to enhance recommendation with multimodal pre-training technologies?*
  - *How to align and fuse the user preference modality to other content modality?*
  - *How to generate the personalized content for each individual user?*
  - *How to apply the multimodal technologies in recommender system?*

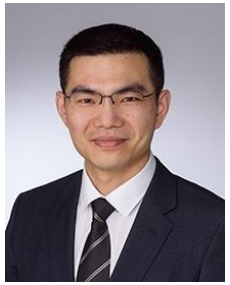
Main content and speakers

- Introduction and challenges in industrial recommender systems

By Dr. Zhenhua Dong (30 mins)

- Multimodal Pre-training and its applications in recommendation by Dr. Jieming Zhu (45mins)

- Multimodal generation for personalization by Prof. Rui Zhang (45mins)



- Practices about multimodal recommendation in products by Dr. Chuhan Wu (45mins)

Targeted audiences

- Researchers and practitioners in multimodal learning: the tutorial offers the insights into how to integrate multimodal technologies into recommender system, like practices and challenges.
- Researchers and practitioners in recommender system: the tutorial introduces the knowledge about the recent and prospective progress in multimodality and generation technologies, and how to apply them to enhance the recommendation.

About us

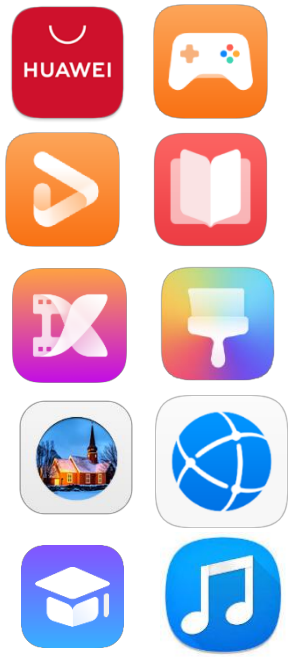
- Huawei's vision: *bring digital to every person, home and organization for a fully connected, intelligent world.*
- Huawei Noah's ark lab: *Building an intelligent world*
  - 7 labs: Computer vision, decision making & reasoning, AI theory, speech and language processing, recommendation & search, AI system, AI application
  - World wide labs: China, Singapore, U.K., France, Canada, Russia
- Recommendation & search research lab: *gets the right information to the right people*
  - Academic research and industrial practice are two wheels of horse drawn carriage[1]
  - Collaboration with product team: advanced AI for products, practical scenario for RQs
  - Collaboration with academia: learn from the best

[1] A Brief History of Recommender Systems, Zhenhua Dong et. al.

1. Industry practices

- Various recommendation scenarios, serving hundreds of millions of users in each day

Product	Scenario
App gallery	App and game recommendation & search
Instant service	Service recommendation
Ads. Platform	CTR/CVR prediction
Browser	News recommendation, search ads.
Music	Songs recommendation & search
Education	Lessons and learning method recommendation
Theme	Theme recommendation and search
GTS	Cases recommendation and search
Internal IT/HR system	Document and staff recommendation and search



2. Impactful research (10000+ citations):

- Recommendation model structure: *DeepFM, PNN, AutoML4RecSys*
- Causal Recommendation: *Counterfactual/intervention recommendations, de-bias*
- LLM4Rec: *NOVA-BERT, LLM4CTR, Survey*
- Benchmark: *FuxiCTR, BARS, SimpleX, REASONER*

Based on the great **missions**, **opportunities** for practices and our research **experiences**, we summarize 10 challenges of industrial recommender system.

Problems
1. Missing information
2. Individual treatment effect
3. Biases

Methods
4. Models reusing
5. Large language model enhancement
6. Multiple modalities
7. Simulations

Goals
8. Lifetime value modeling
9. Trustworthy
10. Win-win ecosystem

1. Missing information

- Research question (RQ): How to handle the missing information in recommender system?
  - Missing features (column data)
    - RecSys may miss some information such as item's popularity, user's thought about the item.
    - RecSys may don't know the causal features: user watched a movie for her friend's suggestion.
  - Missing samples (raw data)
    - RecSys exposes only a few items of all items in one interaction.
    - RecSys can't collect a user's behaviors on the items in other systems.
- Solutions:
  - Counterfactual learning [1].
  - Predict the missing features.
  - What else?

[1] Counterfactual learning for recommender system, RecSys20, Zhenhua Dong et. al.

2. Individual treatment effect

- RQ: How to find the causal attributes of a user's decision (e.g. click, rate) or preference?
- Causal attributes:
  - Example: both user A and B rate one movie C 5 star, A likes C for director, B likes C for cast. The same attribute or treatment (director/cast) may have different effects on user's decision, so it is individual treatment effect (ITE).
  - Accurate causal attribute can help in user profiling, explanation, accuracy.
- Solution:
  - Conditional counterfactual causal effect[1].
  - More novel methods for computing ITE under more generalized assumptions.

[1] Conditional counterfactual causal effect for individual attribution, UAI23, Ruiqi Zhao et. al.

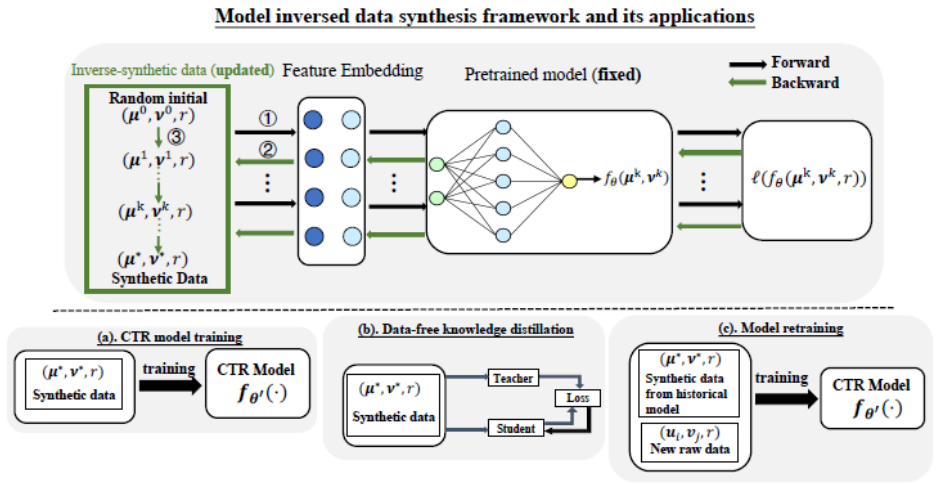
3. Biases

- RQ: How to mitigate the biases in recommender systems?
- There are so many biases [1] in recommender systems due to missing information, confounders and closed feedback loop.
- Solution:
  - Causal analysis: potential outcome model [2], structural causal model
  - Inverse propensity score, direct methods, doubly robust methods
- More research opportunities:
  - How to handle new biases such as duration bias, trust bias, confounder bias?
  - How to collect unbiased data?
  - How to train an unbiased model?
  - How to evaluate the biases?

[1] Workshop keynote -- How to De-Bias for Industrial Recommender System? A causal Perspective, SIGIR21, Zhenhua Dong
[2] On the Opportunity of Causal Learning in Recommendation Systems: Foundation, Estimation, Prediction and Challenges, IJCAI22, Peng Wu et. al.

4. Model reusing

- RQ: How to reuse the historical models efficiently?
- The industrial recommendation models should be updated with recent data for better performance, but the historical models are always underutilized.
- Solutions:
  - Online learning, ensemble learning.
  - Model inversed data synthesis framework [1].
- More research opportunities:
  - Machine unlearning, learnware, etc.
  - Can we train large recommendation model like large language model?

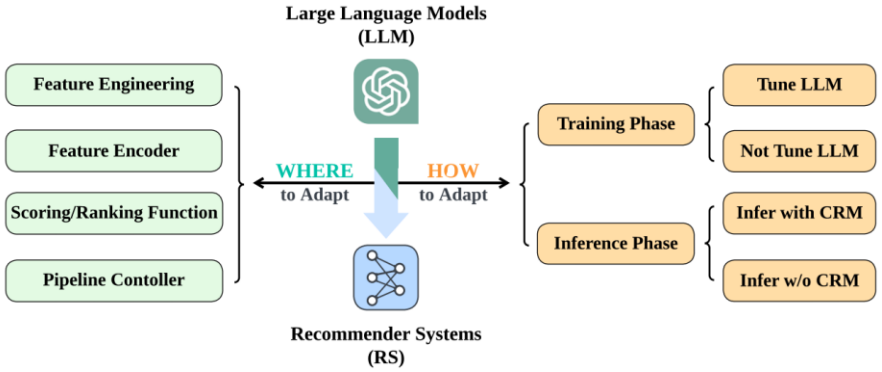


[1] Data-free Knowledge Distillation for Reusing Recommendation Models , RecSys23, Cheng Wang et. al.

5. Large language model enhanced recommendation

- RQ: ChatGPT has demonstrated great capabilities of LLMs, how to improve recommendation with large language models?
- Solutions[1]:
  - Where: feature, embedding, prediction, controller
  - How: fine tune, collaboration with classic models
- Let us embrace LLM:

"It does not matter if you love it or not
It is standing right there
With no emotion
Not going to change" By Sangs-Rgyas Rgya Mtsho



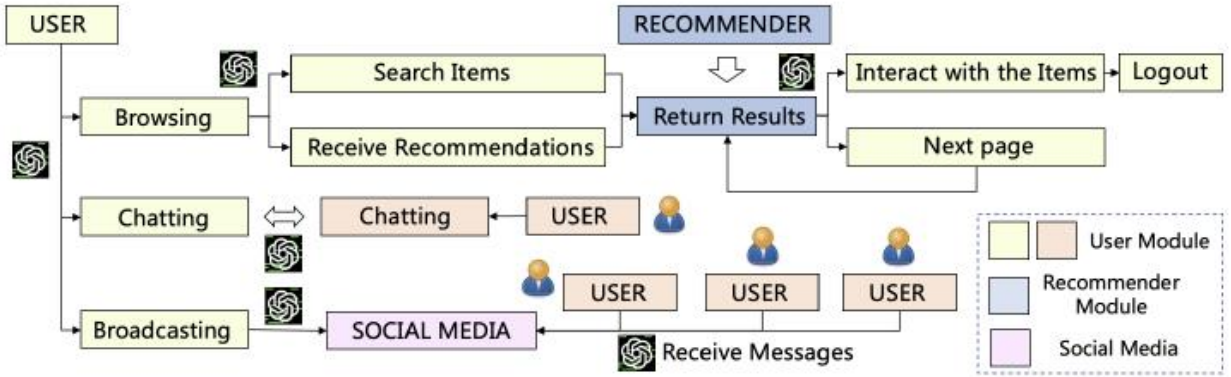
[1] How Can Recommender Systems Benefit from Large Language Models: A Survey, Jianghao Lin et. al.

6. Multiple modalities

- RQ: How to align the user preference modality to content modality?
  - Classic recommendation train models with preference modality data like user behavior, which is good at modeling the collaborative signals.
  - Content modalities like text, image and video are good at content semantic understanding.
  - There is huge gap between preference modality and content modalities leads to minor improvement in recommendation.
- Solution:
  - Deeply understanding about the different modalities and their relation to user’s preferences and decisions.
  - Recommendation focused multi-modal pre-trained model that bridges the gap.

7. Simulation

- RQ: How to model user preference with simulation?
- User modeling is a classical research topic
  - ACM UMAP is 31 years old, many research topics have been studied.
  - User modeling is still hard since people and context are complex, e.g. in some ads. scenarios, the CTR is less than 1%.
- Solution: RecAgent [1]: digital twin of recommender system, simulate user behaviors, and align with real human understanding.



- More opportunities:
  - How to simulate more users and more behaviors efficient and accurately?
  - How to evaluate the simulation ?

[1] RecAgent: A Novel Simulation Paradigm for Recommender Systems, Lei Wang et. al.

8. Lifetime value modeling

- RQ: How to predict users' long term satisfaction?
- Most recommendation studies focus on optimizing short term objectives like click, rating, dwell time, which can not align the goal to improve user's long term satisfaction like deep conversion task.
- Solution: in the tutorial[1], we introduce the definitions and scenarios of LTV, some typical LTV prediction technique, and products practices.
- There are still many hard problems:
  - Delayed and sparse feedback.
  - Cold start, offline evaluation, multi-task optimizations.

[1] Tutorial -- Customer Lifetime Value Prediction: Towards the Paradigm Shift of Recommender System Objectives, RecSys23, Chuhan Wu et. al.

9. Trustworthy

- RQ: How to build trustworthy recommender systems?
- We consider 8 perspectives such as accountability, security, fairness [1], privacy, robustness, transparency, assisting or serving people, and long term enhancement of the happiness of human, society and environment.
- Most current recommender system focus on the accuracy such as AUC, Logloss, CTR, which is not enough to be a trustworthy recommender system.
- We hope more scholars can help industry to define and build trustworthy and social good recommender systems from wider perspectives such as society, economy, user-centric, ecological systems and natural environments.

[1] Workshop keynote --Two perspectives about biases in recommender system: OoD and unfairness, ICDM2023, Zhenhua Dong

10. Win-win ecosystem

- RQ: How to satisfy multi-stakeholders in dialog based IR or RecSys?
- Mainly 4 kinds stakeholders: user, content provider(CP), information system and advertisers.
- Dialog based IR or RecSys can directly providing answer or satisfying users' request. However, there are key challenges for each stakeholder:
  - Content provider: How to protect their intellectual property and benefits?
  - Advertisers: How to appropriately expose the Ads. during the dialog?
  - Users: How to ensure the generated information is objective and accountable?
  - Information system: How to design win-win interactions and mechanisms for long-term benefits?

Based on the great **missions**, **opportunities** for practices and our research **experiences**, we summarize 10 challenges of industrial recommender system.

Problems
1. Missing information
2. Individual treatment effect
3. Biases (SIGIR21)

Methods
4. Models reusing
5. Large language model enhancement
6. Multiple modalities (Web2024)
7. Simulations

Goals
8. Lifetime value modeling (RecSys23)
9. Trustworthy (ICDM23)
10. Win-win ecosystem

We desire to collaborate with the passionate, talented people and change world together!

noahlab1@huawei.com
dongzhenhua@huawei.com

WeChat

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

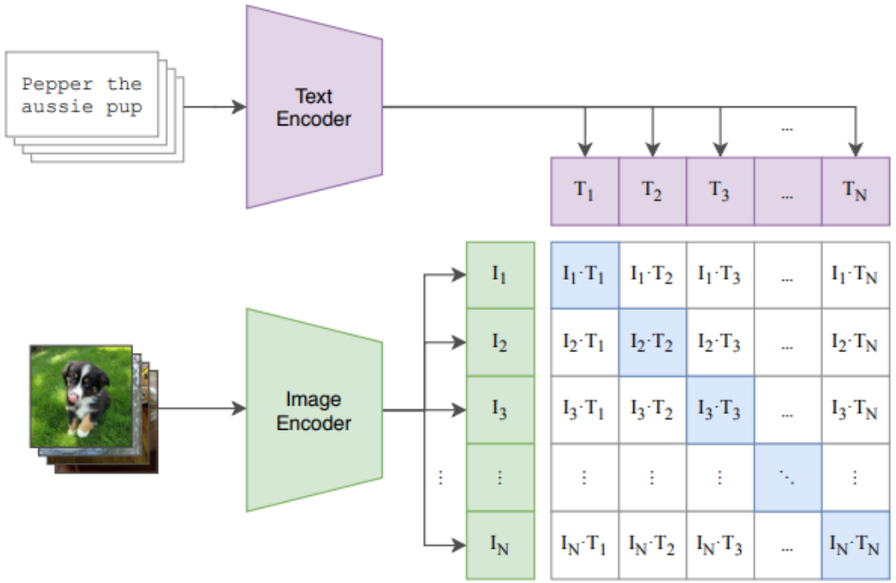


Fig. from: Radford et al., Learning Transferable Visual Models From Natural Language Supervision, 2021

Image-to-audio retrieval

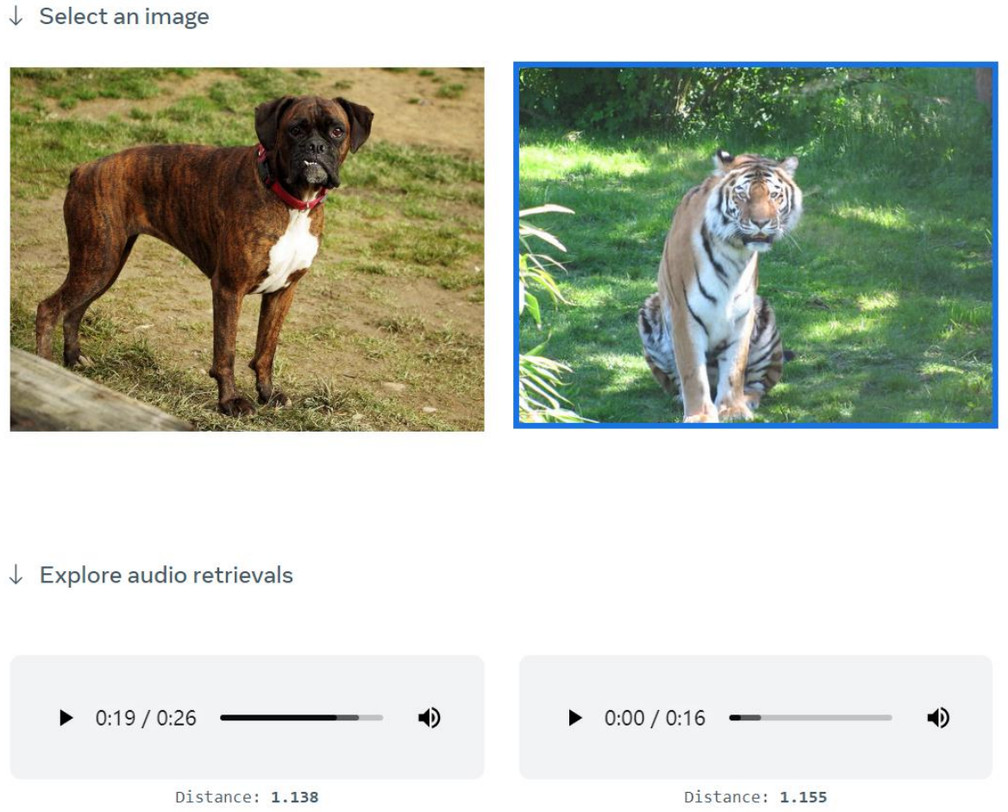


Fig. from: Girdhar et al., ImageBind: One Embedding Space To Bind Them All, 2023

Multimodal Generation

Text-to-image generation

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 help for recommendation?

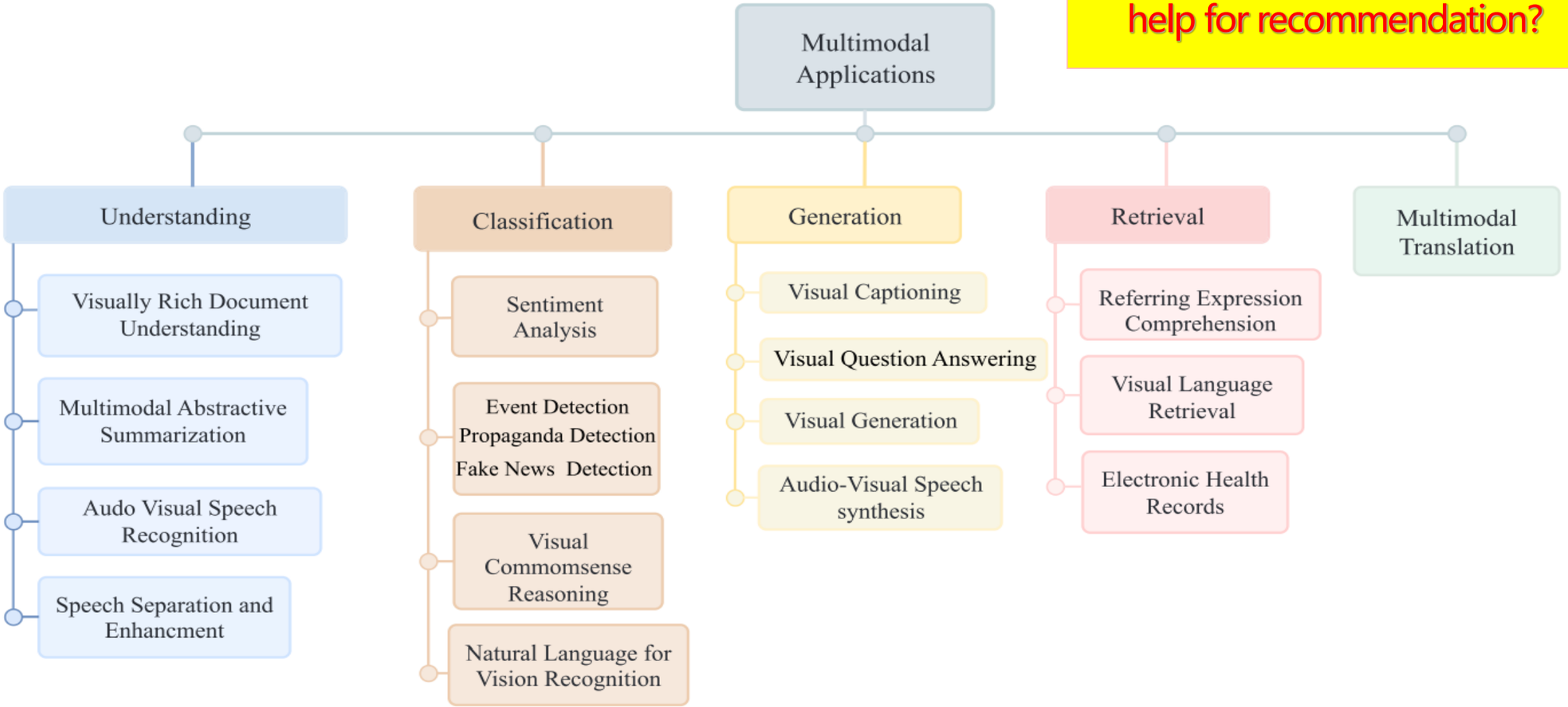


Fig. from: Manzoor et al., Multimodality Representation Learning: A Survey on Evolution, Pretraining and Its Applications, 2023

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

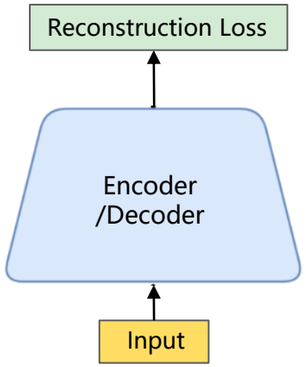
General-Domain Multimodal Pretraining

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

Self-supervised Pretraining

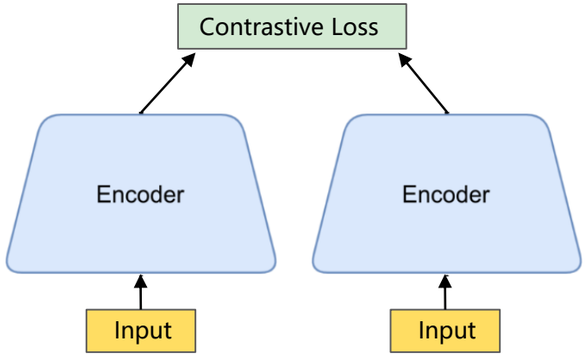
Typical pretraining paradigms: reconstructive, contrastive, and generative

BERT
StableDiffusion



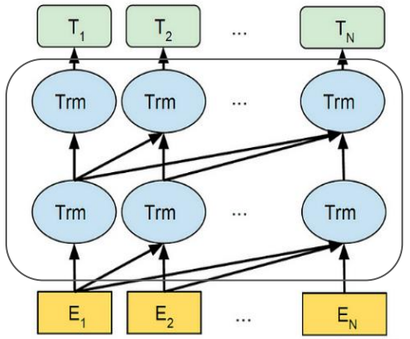
Reconstructive

CLIP
ImageBind



Contrastive

GPT
DALL-E



Generative

Pretrained Multimodal Models (2019~2022)

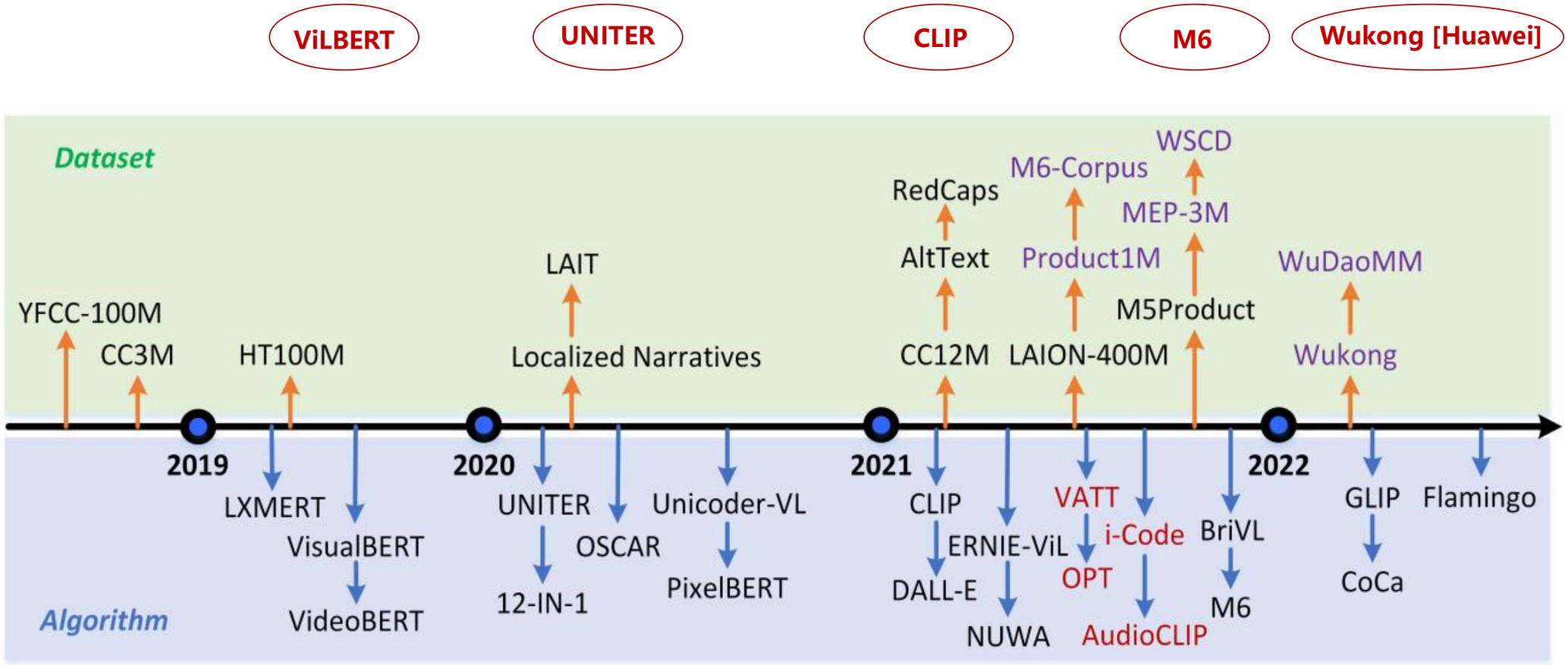
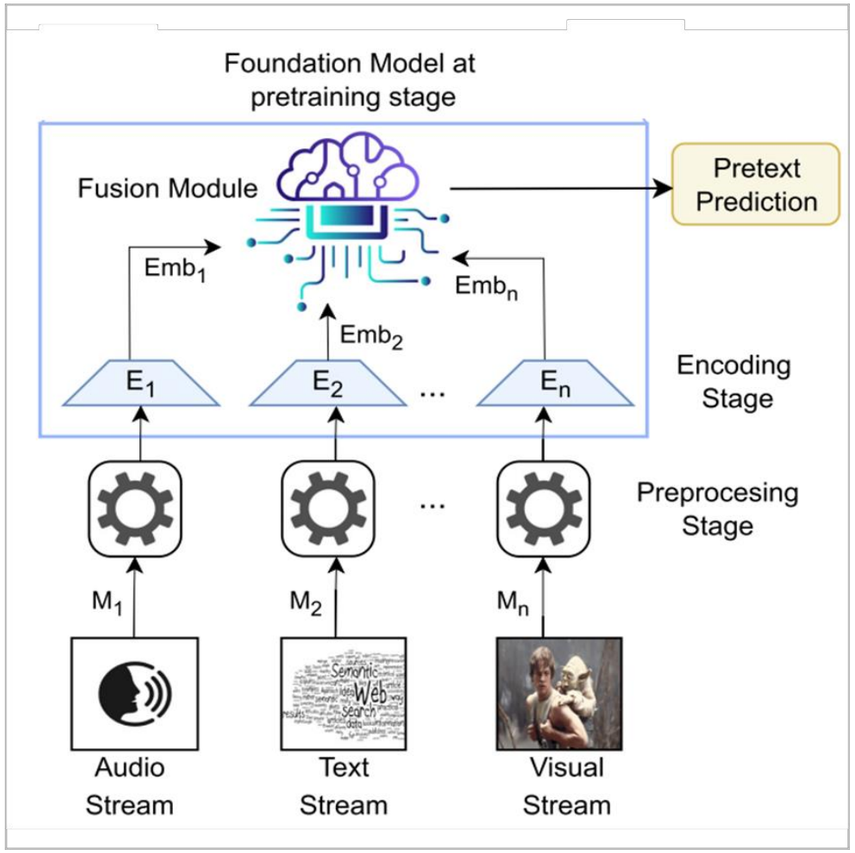
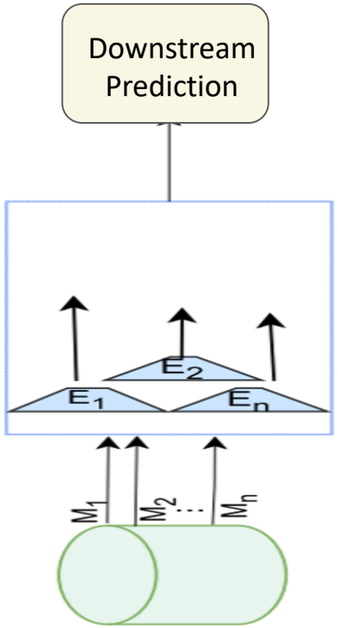


Fig. from: Wang et al., Large-scale Multi-Modal Pre-trained Models: A Comprehensive Survey, 2023

Learning Paradigm: Pretraining-Finetuning



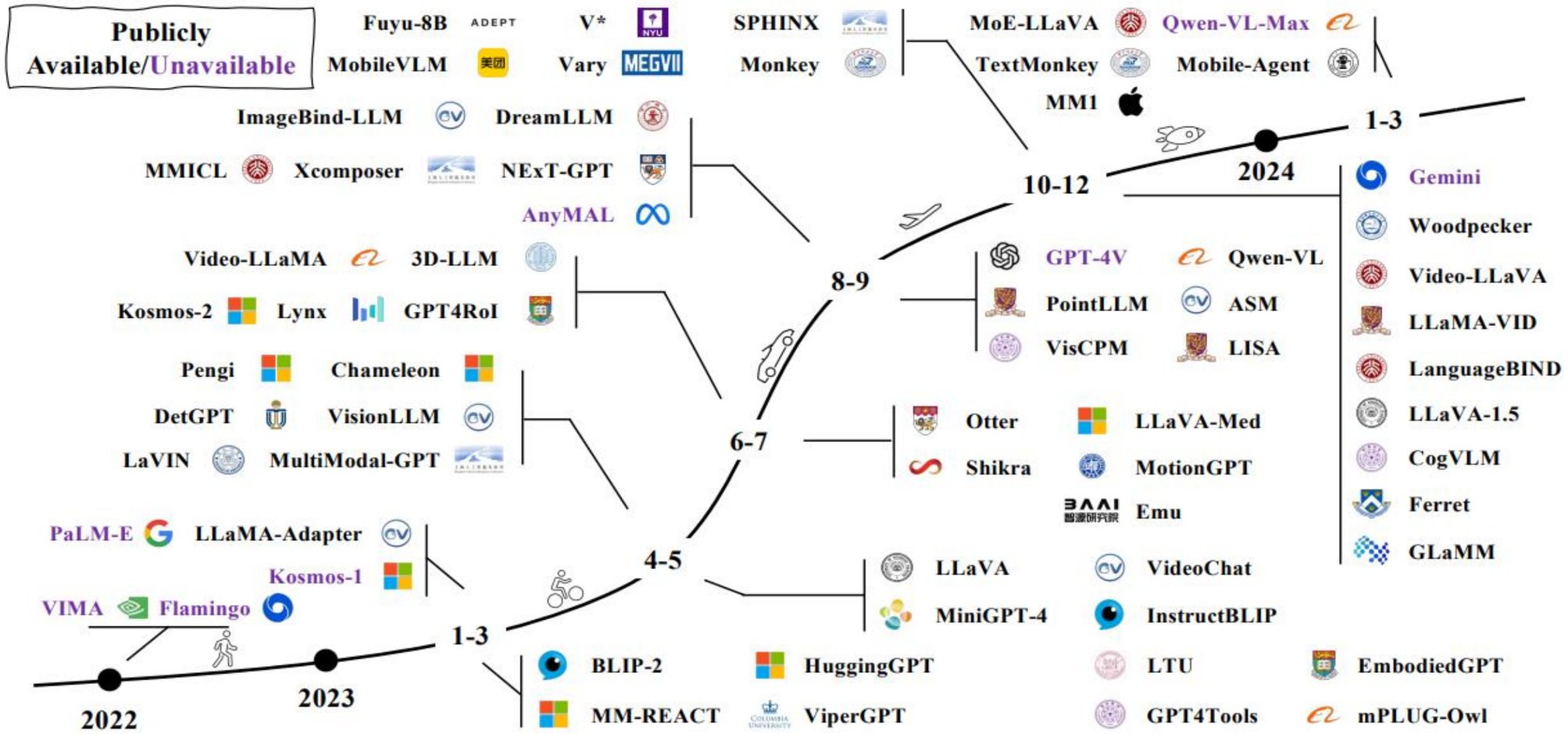
Pretraining



Finetuning

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

Multimodal Large Language Models (2023~2024)



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

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.

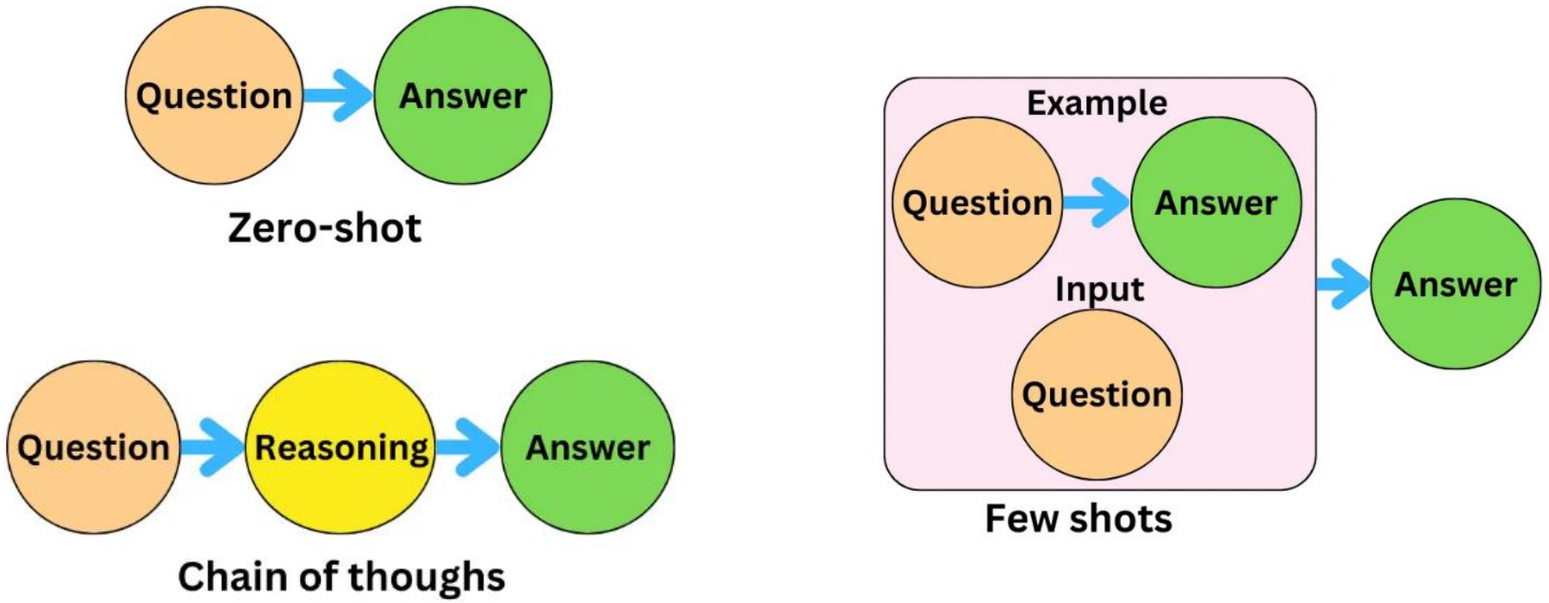
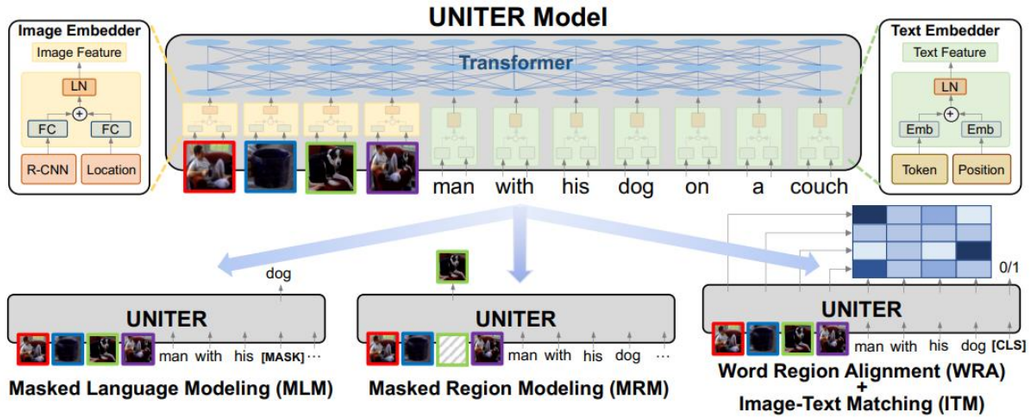
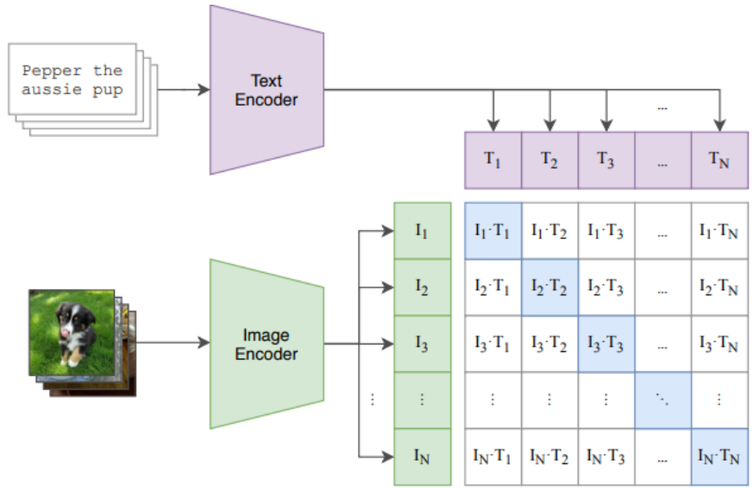


Fig. from https://newsletter.theaiedge.io/p/prompt-engineering-and-llmops-building

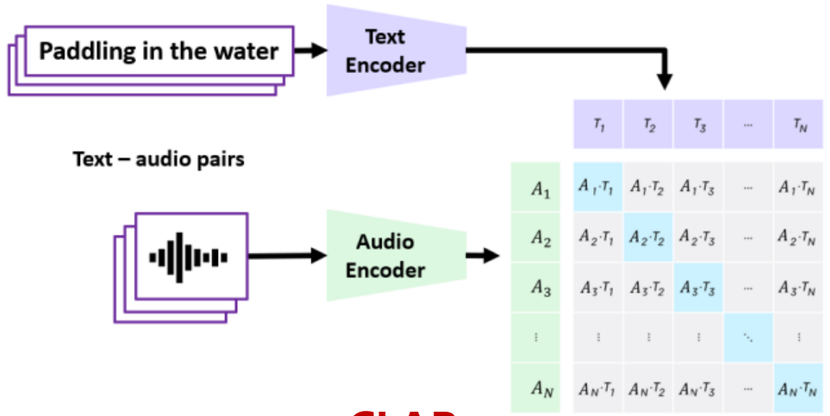
Multimodal Large Models (1)



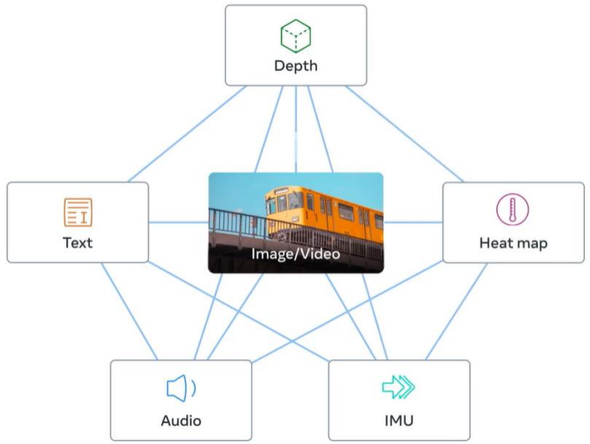
UNITER



CLIP



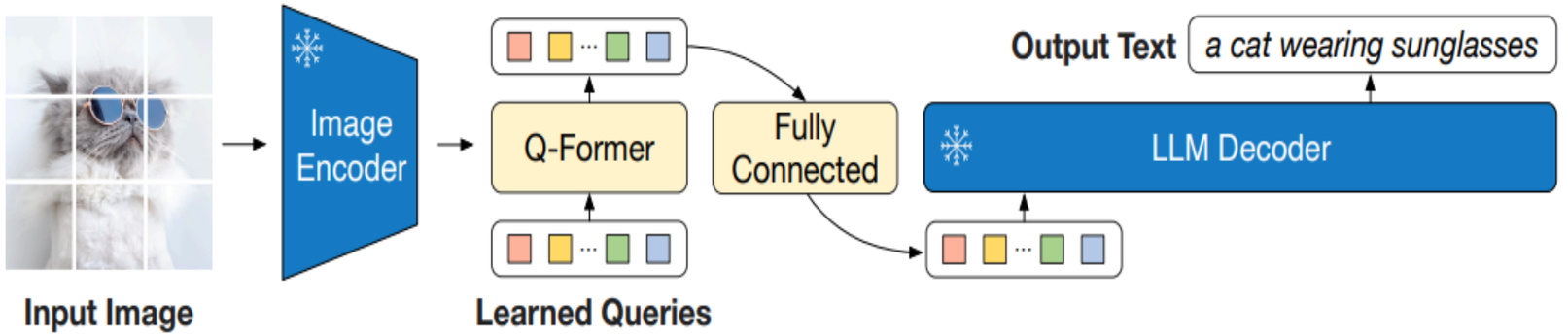
CLAP



ImageBind

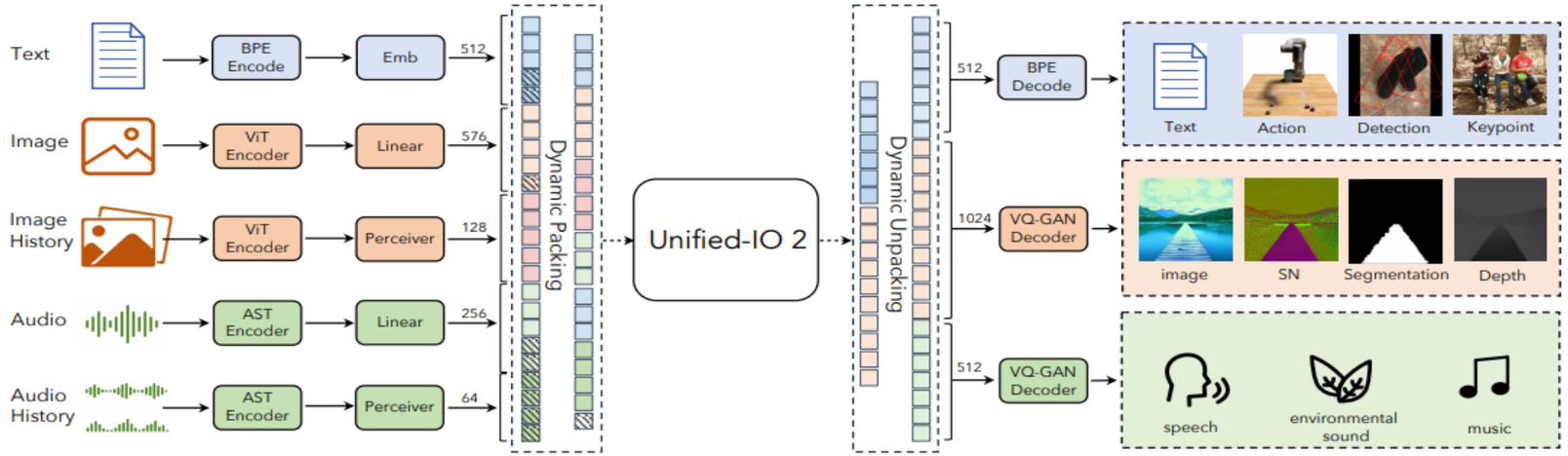
Chen et al., UNITER: UNiversal Image-TExt Representation Learning, 2019
Lu et al., Learning Transferable Visual Models From Natural Language Supervision, 2021
Elizalde et al., CLAP: Learning Audio Concepts From Natural Language Supervision, 2022
Girdhar et al., ImageBind: One Embedding Space To Bind Them All, 2023

Multimodal Large Models (2)



BLIP-2

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



UnifiedIO-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

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**
  - User/Item Tagging
  - Matching: Sequential recommendation
  - Ranking: CTR prediction
  - Reranking: diversity awareness

Multimodal Recommendation Scenarios

News Feed

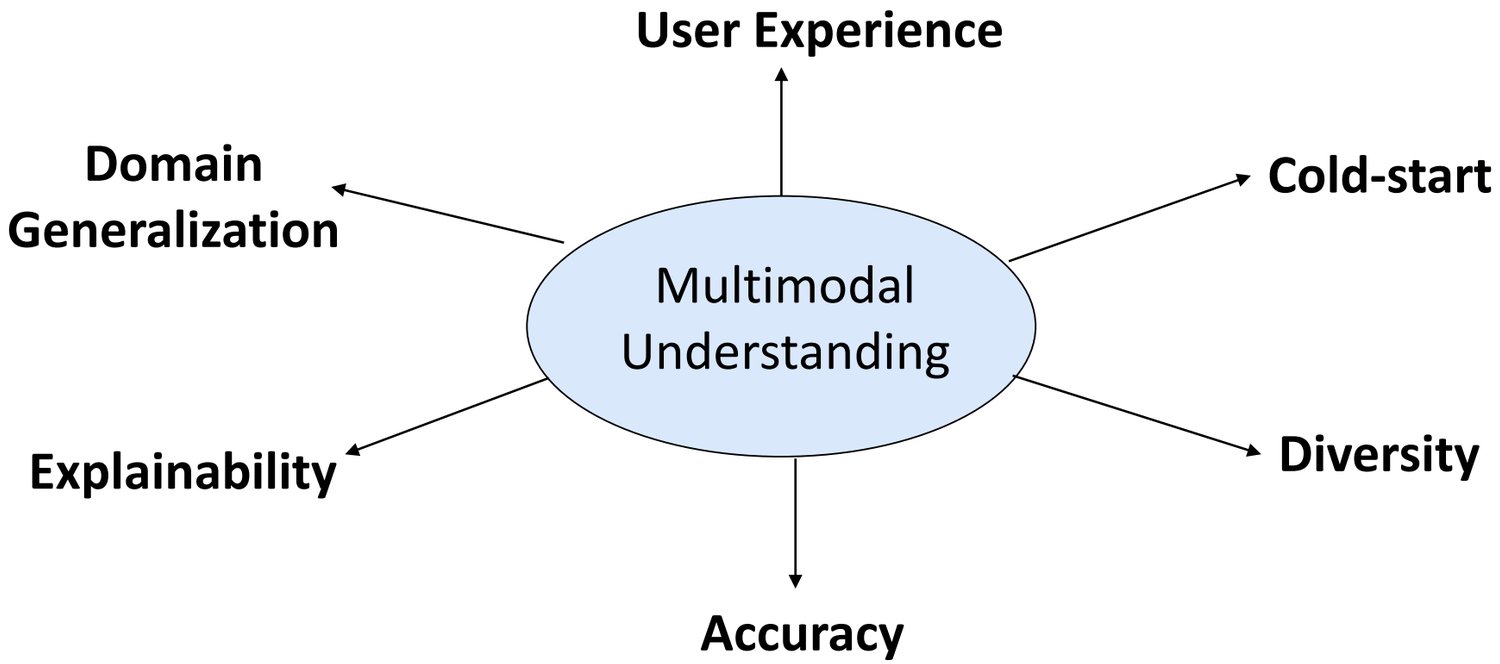
Micro-Video

Music Streaming

Mobile Theme

And many more...

What Can Multimodal Models Help?



Why Is Multimodal Pretraining Necessary?

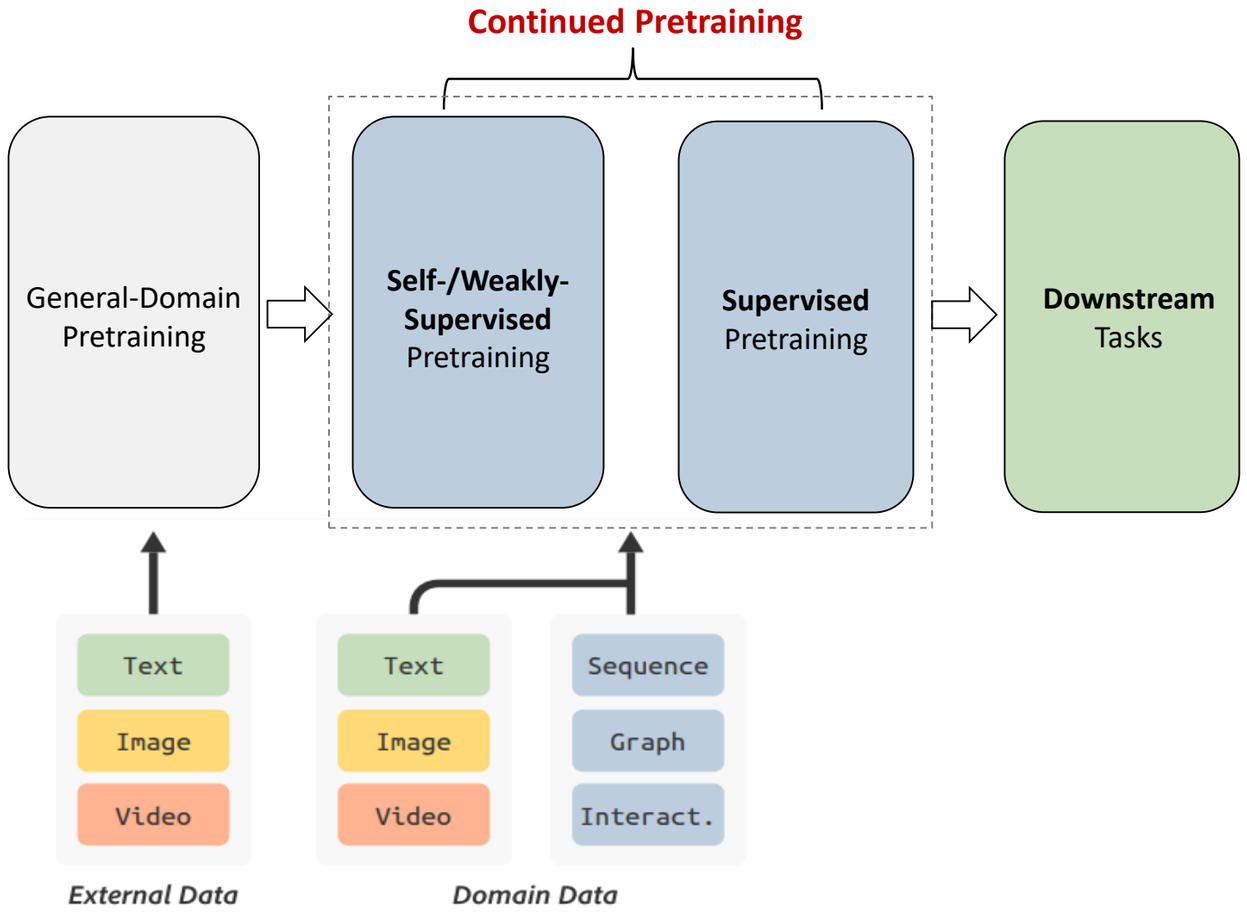
Multimodal pretraining for recommendation

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…

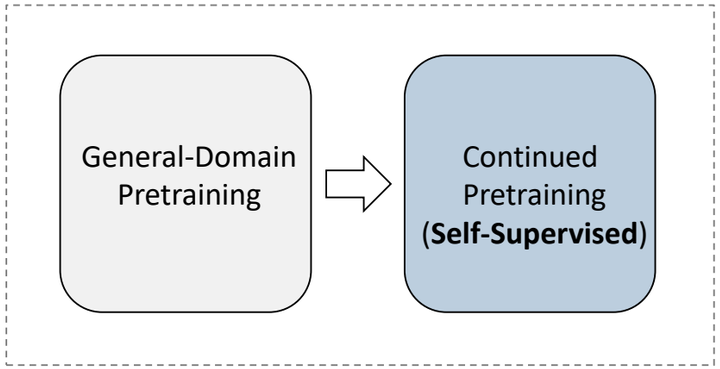
Continued Pretraining:

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

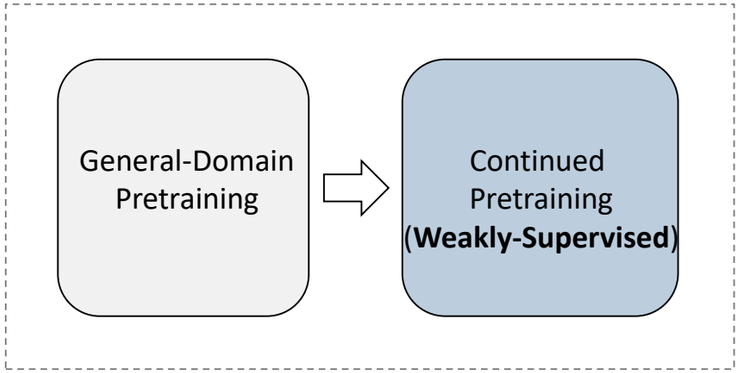
Downstream Tasks

- Representation transfer
- Supervised finetuning
- Adapter tuning

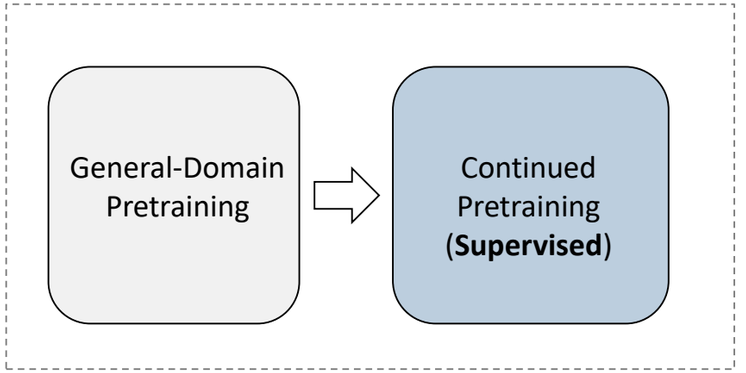
Categorization of Existing Work



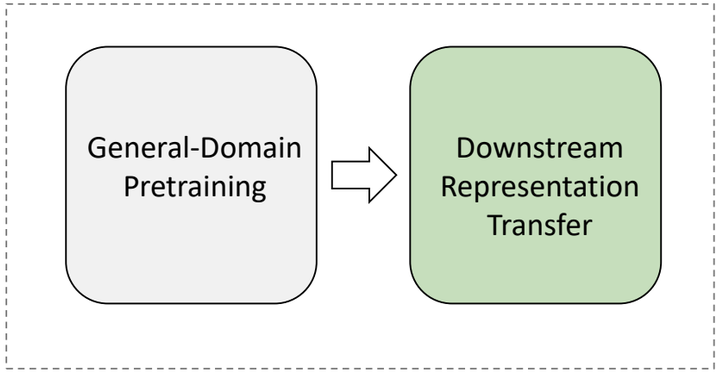
C1



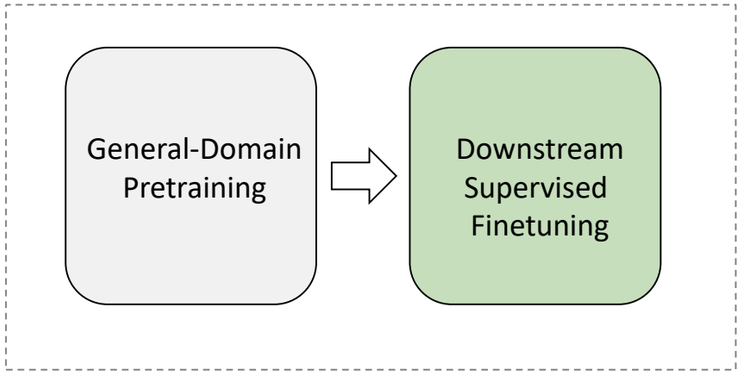
C2



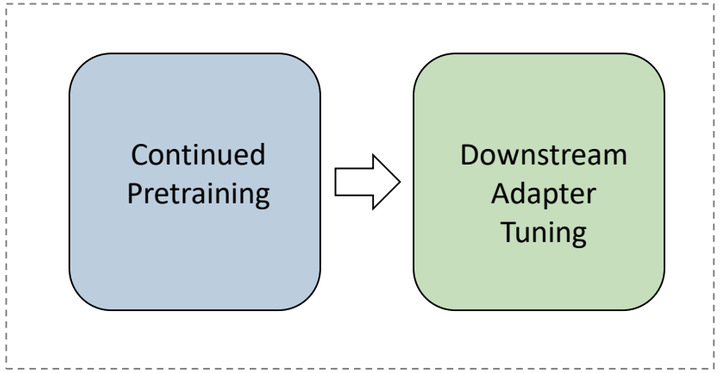
C3



C4



C5



C6

C1: Continued Pretraining w/ Self-Supervised Tasks

Reconstructive:

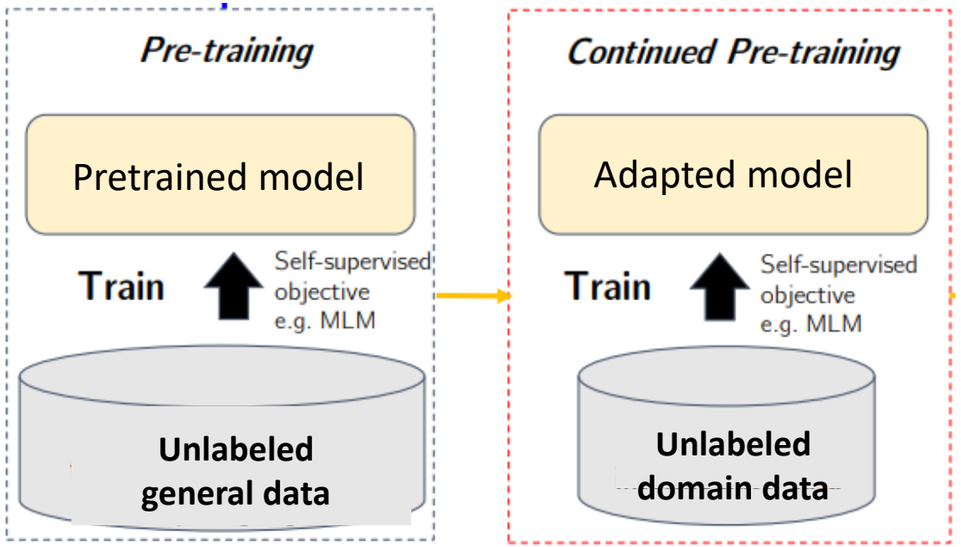
- Masked token prediction (PREC、RecoBERT)
- Masked attribute prediction (PREC、S3-Rec)
- Masked item prediction (PREC、S3-Rec)

Contrastive:

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

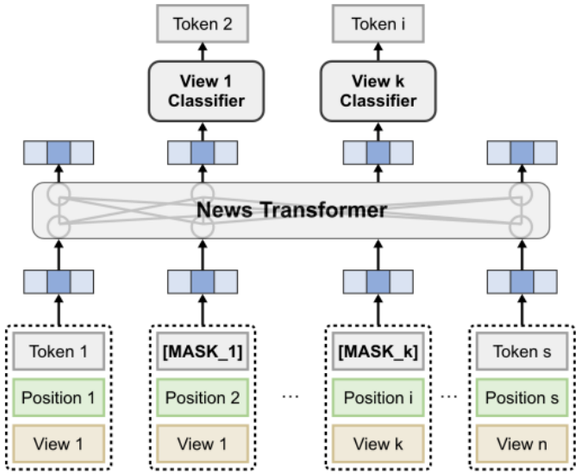
Generative:

- ……



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., 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

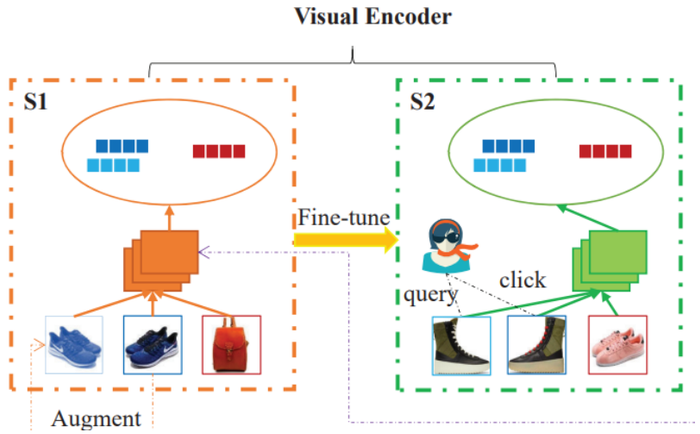
C1: Continued Pretraining w/ Self-Supervised Tasks



(a) Masked news token prediction.

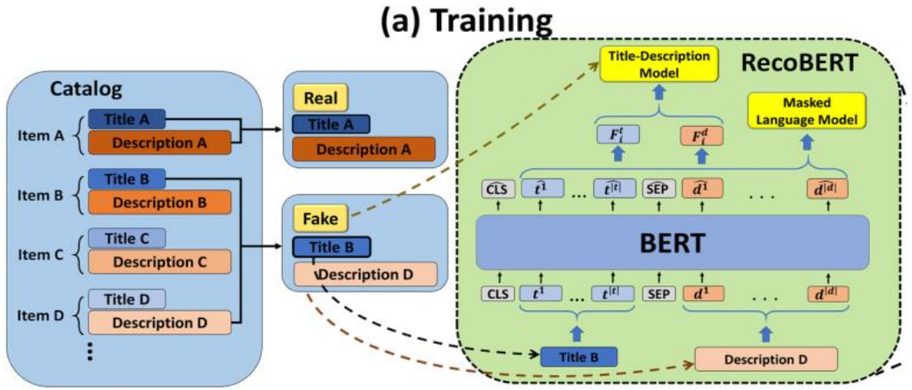
PREC [Huawei] *Liu et al., 2022*

Masked token/category prediction



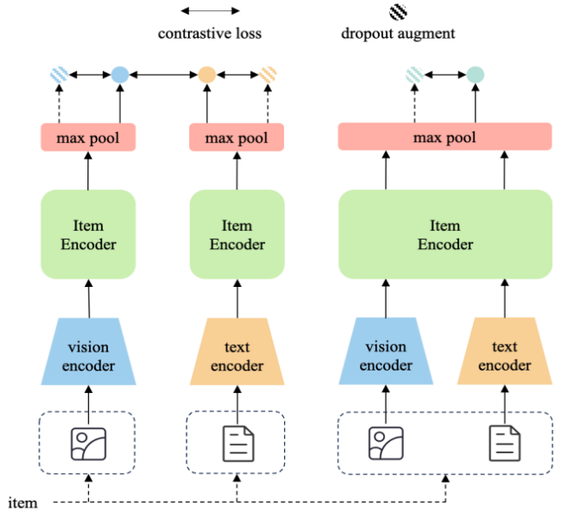
VisualEncoder *Chen et al., 2022*

Contrastive learning



RecoBERT Malkiel *et al., 2020*

Masked language model / alignment between title and description



MMSRec *Song et al., 2023*

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
- ……

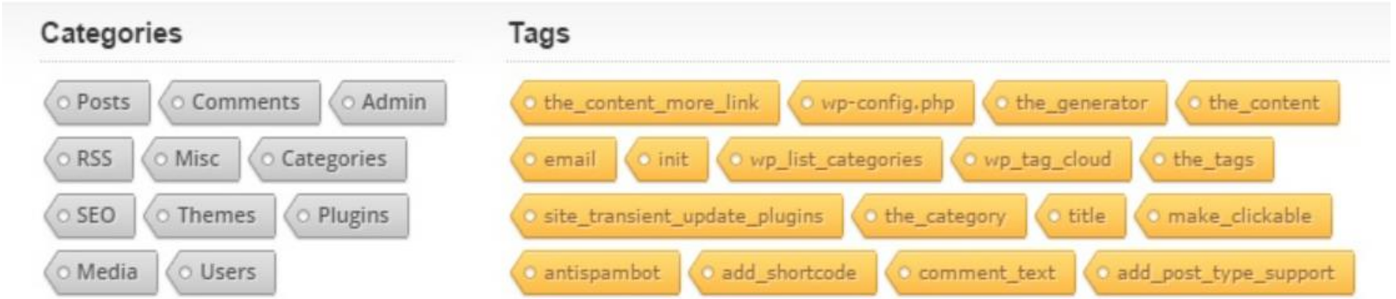


Fig. from: https://www.wpexplorer.com/plugins-add-categories-tags/

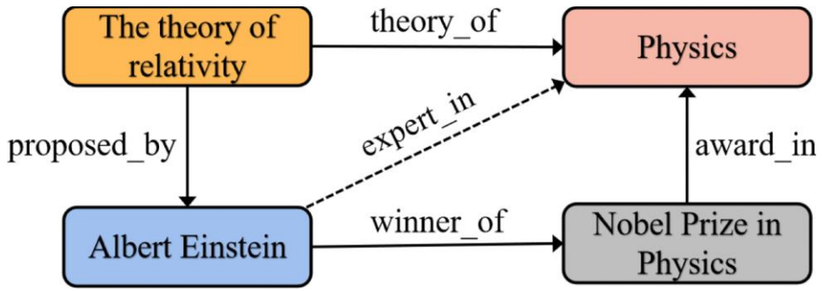
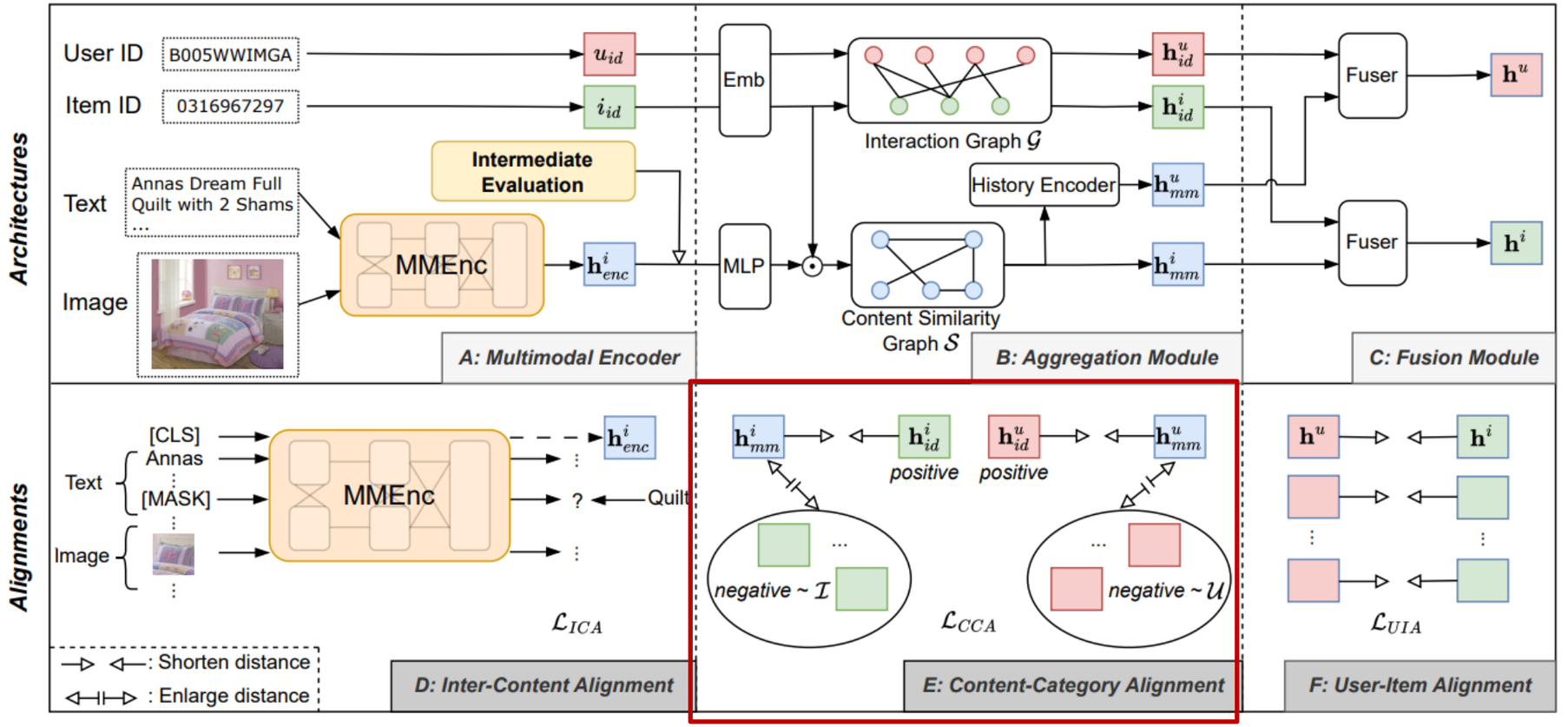


Fig. from: TDN: Triplet Distributor Network for Knowledge Graph Completion, 2023

C2: Continued Pretraining w/ Weakly-Supervised Tasks

AlignRec

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

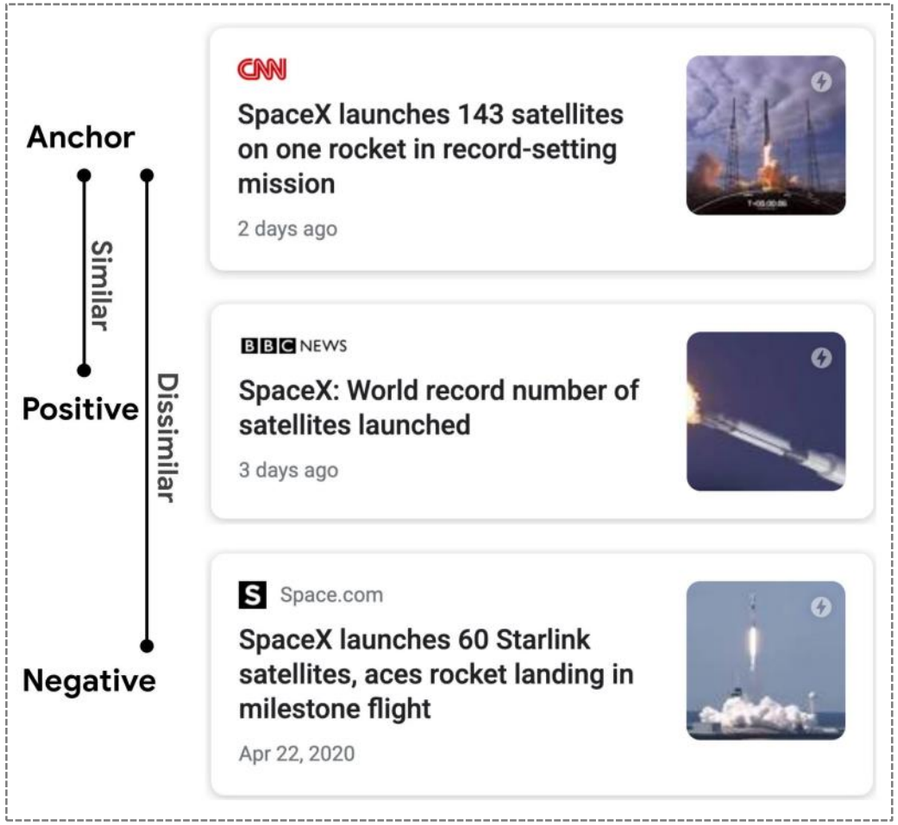


Liu et al., An Aligning and Training Framework for Multimodal Recommendations, 2024

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



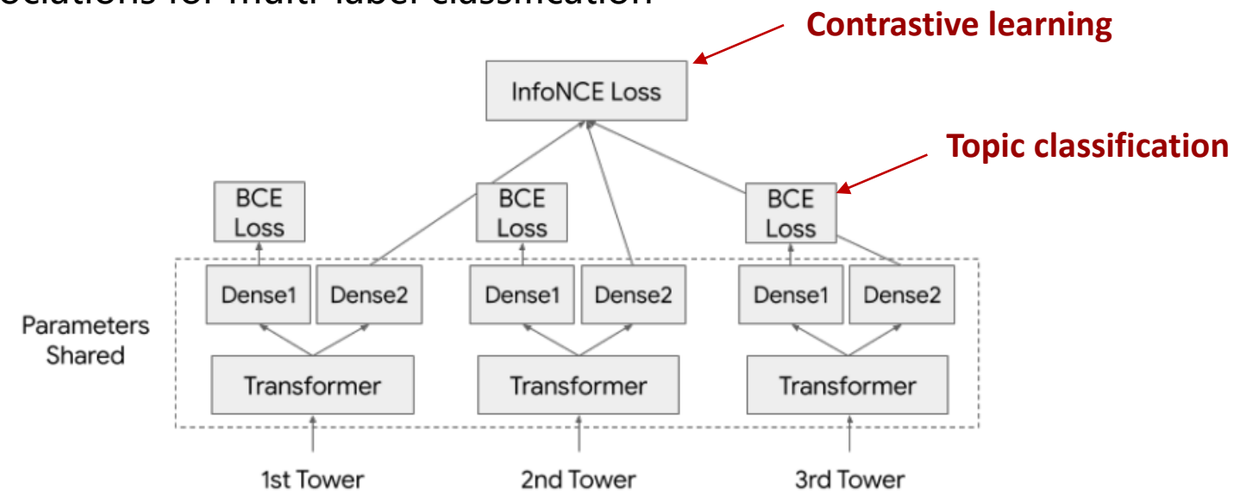


Figure 4: NewsEmbed model structure.

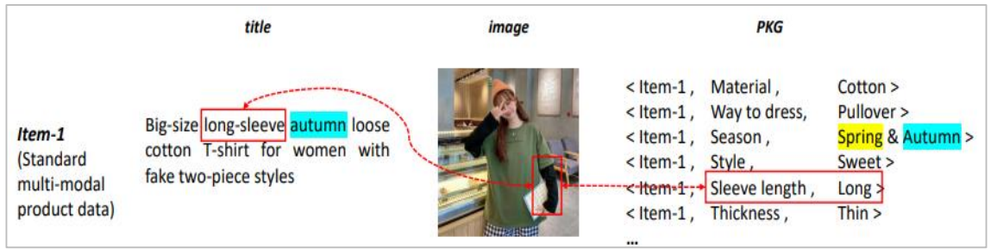
Model	BBC	20 Newsgroup	AG News	MIND	Avg.
SBERT	96.0	64.5	86.5	75.2	80.5
USE	97.3	66.4	85.8	75.0	81.1
mUSE	97.3	71.4	86.8	75.8	82.8
LaBSE	96.6	70.3	86.3	73.9	81.8
Laser	92.1	63.5	81.4	63.1	75.0
NewsEmbed	**97.5**	**73.9**	**89.1**	**77.0**	**84.2**

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

C2: Continued Pretraining w/ Weakly-Supervised Tasks

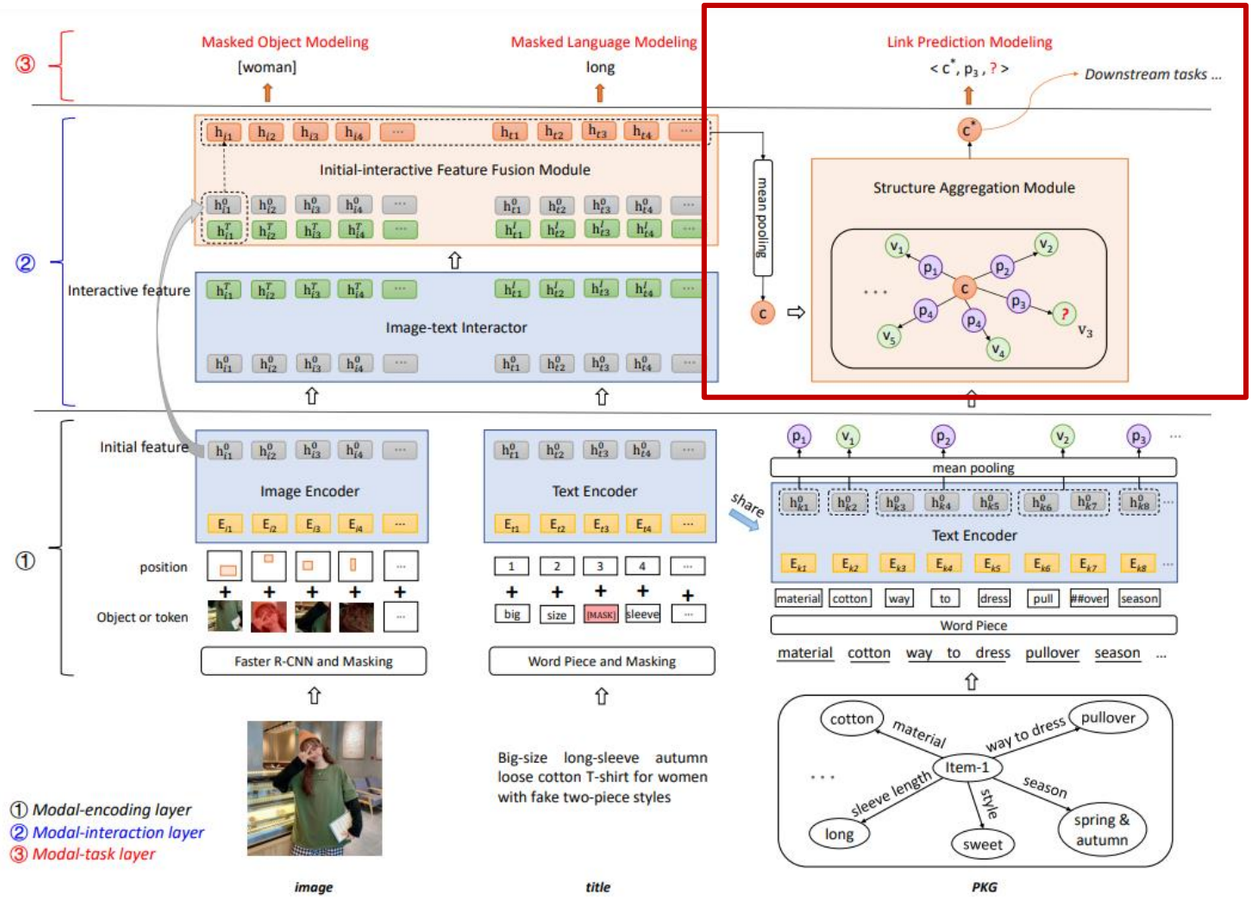
K3M

- **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)

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

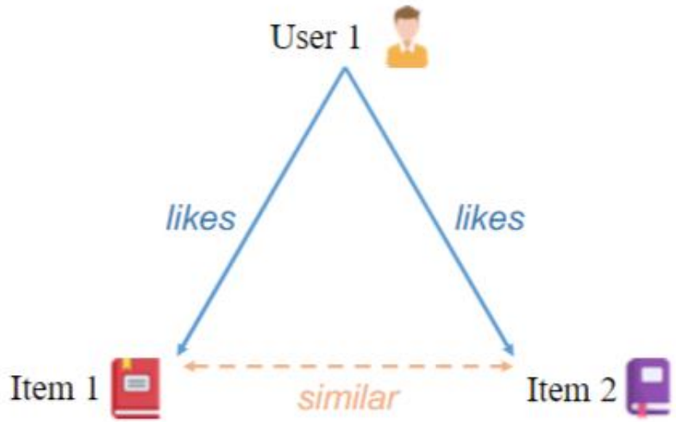


Zhu et al., Knowledge Perceived Multi-modal Pretraining in E-commerce, 2021

C3: Continued Pretraining w/ Supervised Tasks

Supervised Signals:

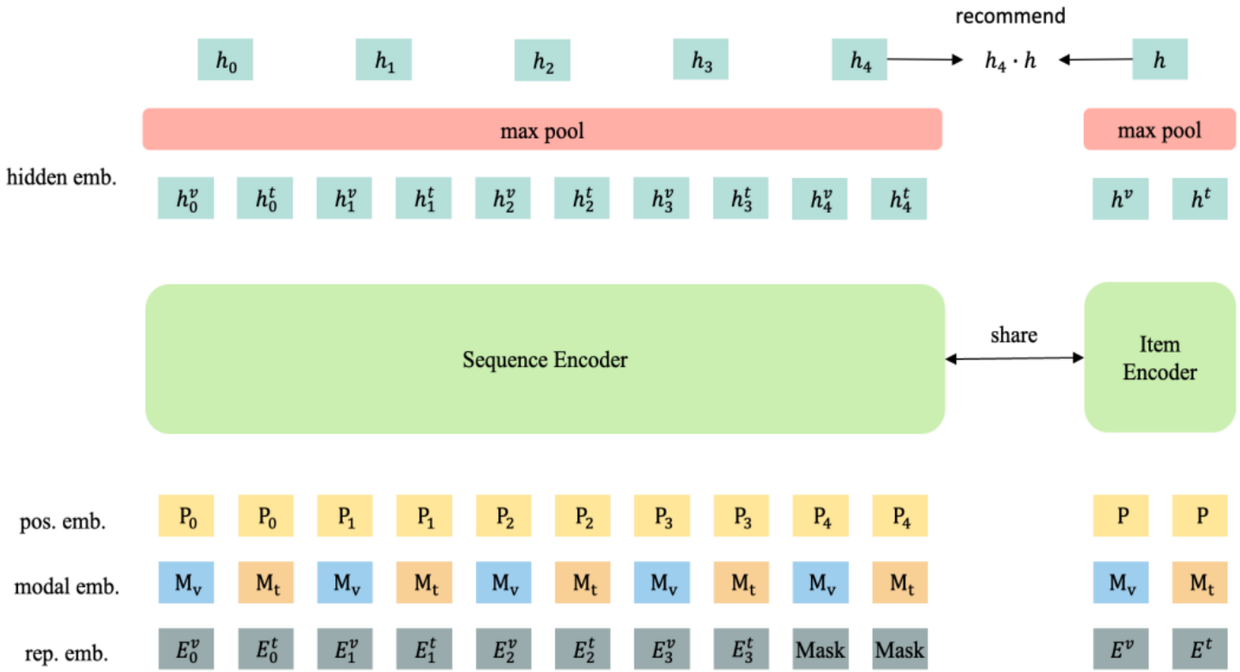
- User-to-item matching (MMSRec、MISSRec)
- Item-to-item matching (CB2CF、ItemSage)
- ID-to-modality alignment (CLCRec)
- User-to-item cross-encoder

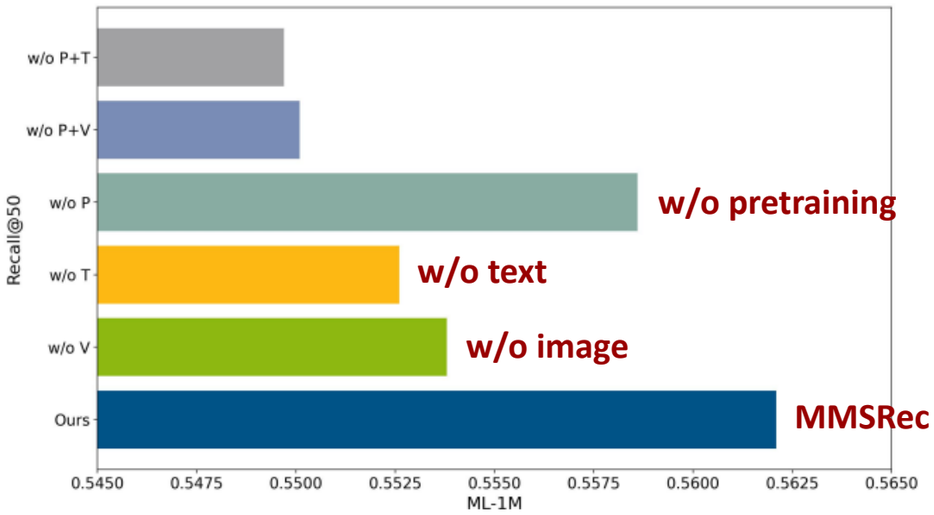


C3: Continued Pretraining w/ User-to-Item Matching

MMSRec

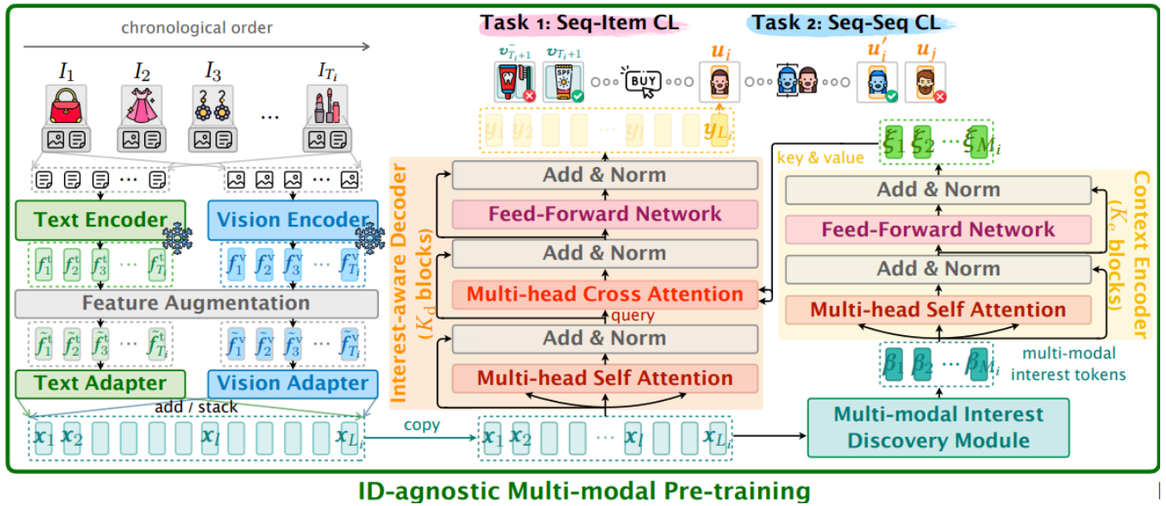
- **Masked Item Prediction**: mask the last position of the behavior sequence for next item prediction





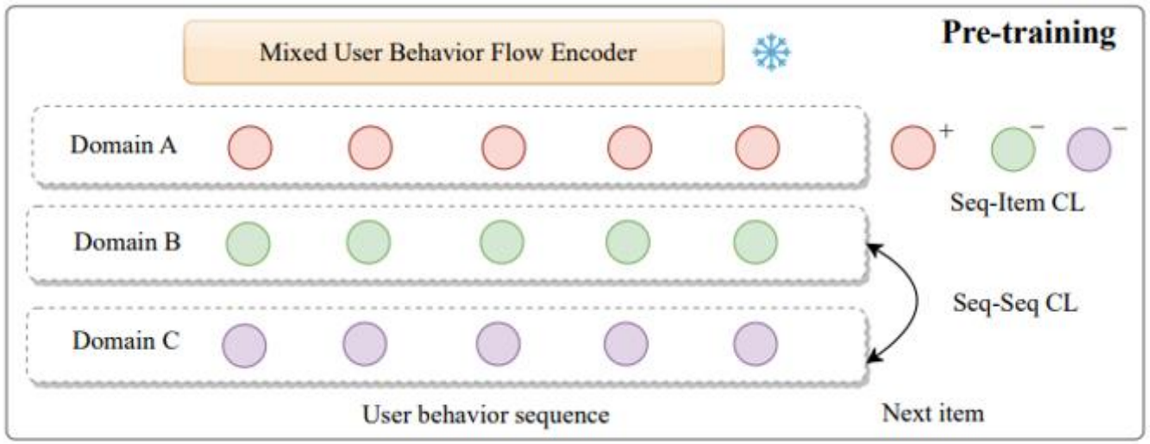
Song et al., Self-Supervised Multi-Modal Sequential Recommendation, 2023

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

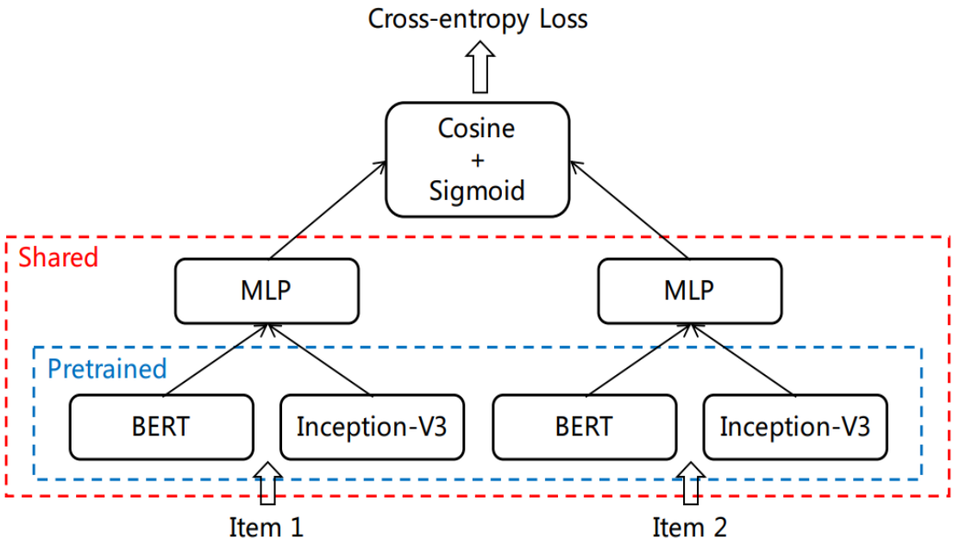
- **Seq-item contrastive learning**: matching between user sequence and masked item
- **Multi-domain matching**: matching across domains

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

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

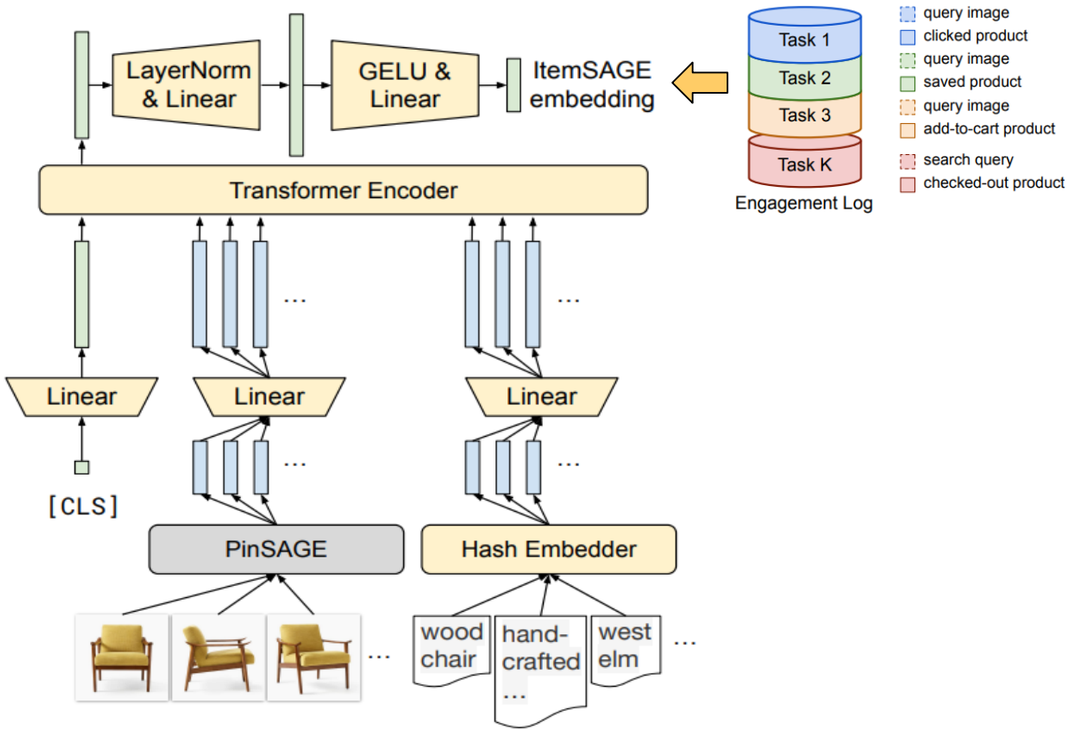
SSD

- **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

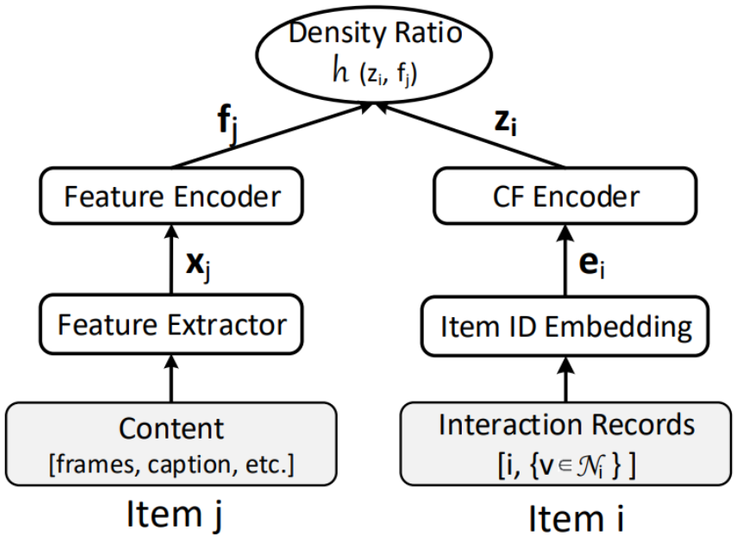


Huang et al., Sliding Spectrum Decomposition for Diversified Recommendation, 2021.
Baltescu et al., ItemSage: Learning Product Embeddings for Shopping Recommendations at Pinterest, 2022

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

CLCRec

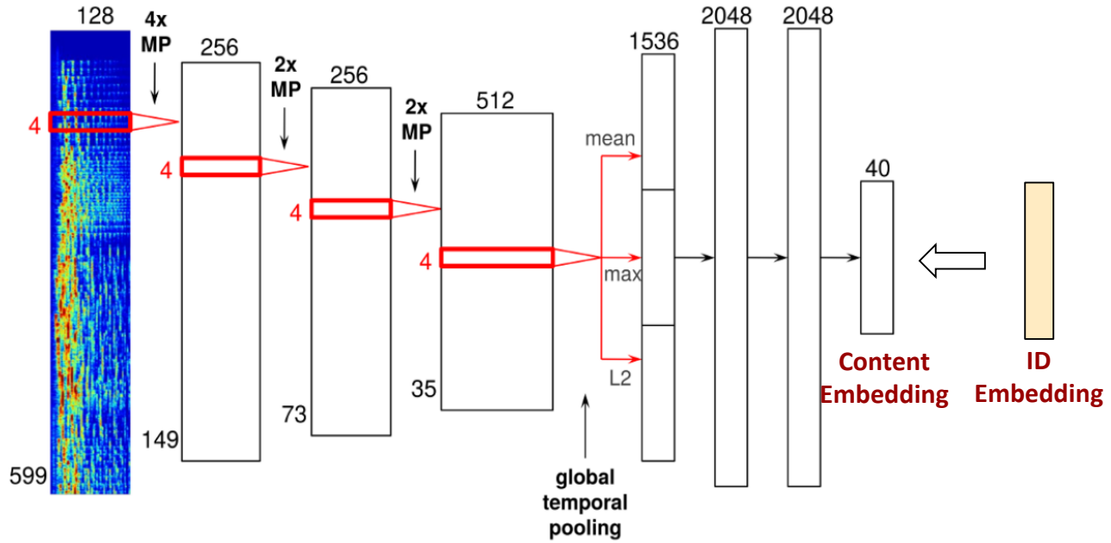
- Contrastive learning between CF encoder and content encoder



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

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

Downstream Tasks

User/Item Tagging:

- Item tagging

Matching: Sequential recommendation:

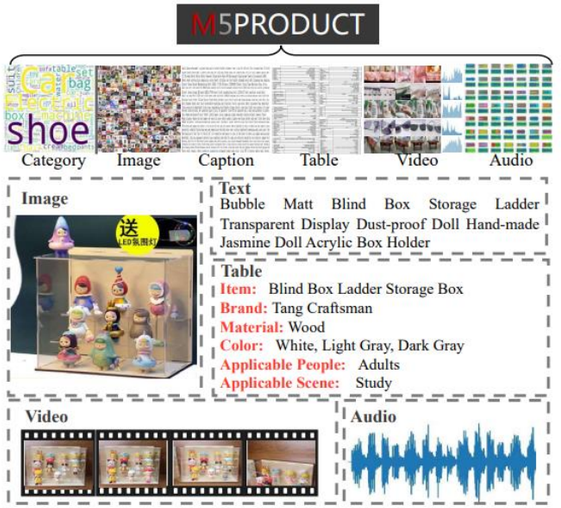
- MM-Rec [Microsoft]
- IMRec [Huawei]

Ranking: CTR prediction:

- PPM [JD.com]
- EM3 [KuaiShou]

Reranking: diversity awareness

- SSD [Xiaohongshu]



Item tagging

Diversity awareness

Image Credits: [Left] Dong et al., M5Product: Self-harmonized Contrastive Learning for E-commercial Multi-modal Pretraining, 2022
[Right] Huang et al., Sliding Spectrum Decomposition for Diversified Recommendation, 2021

How to Adapt to Downstream Tasks

Representation transfer:

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

Joint Finetuning:

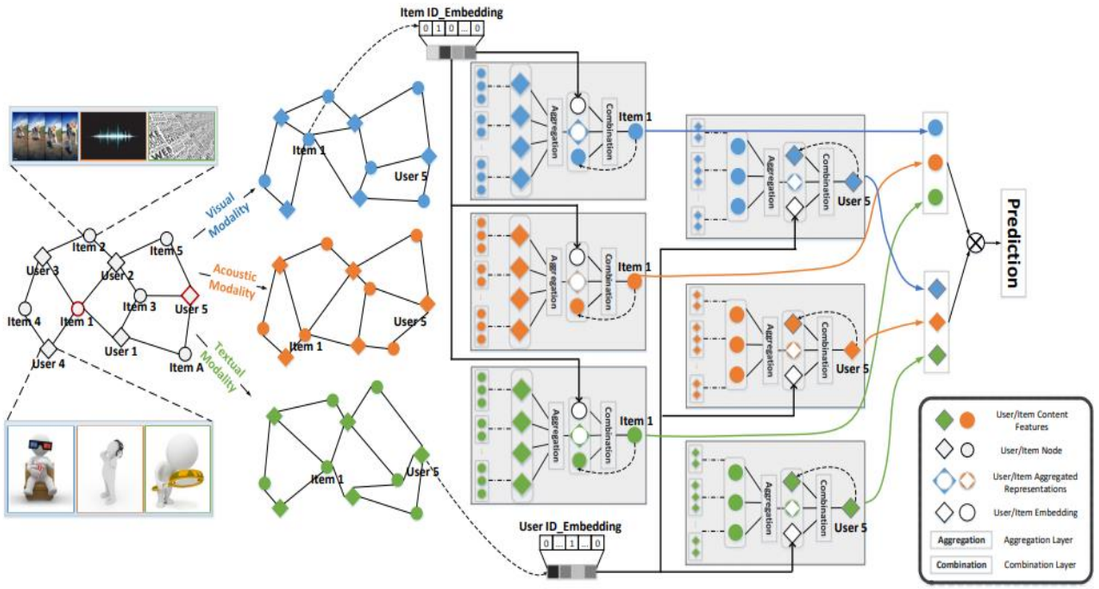
- Only finetuning item-side encoder
- Finetuning both item-side and user-side encoders
- Finetuning user-item cross encoder

Adapter Tuning:

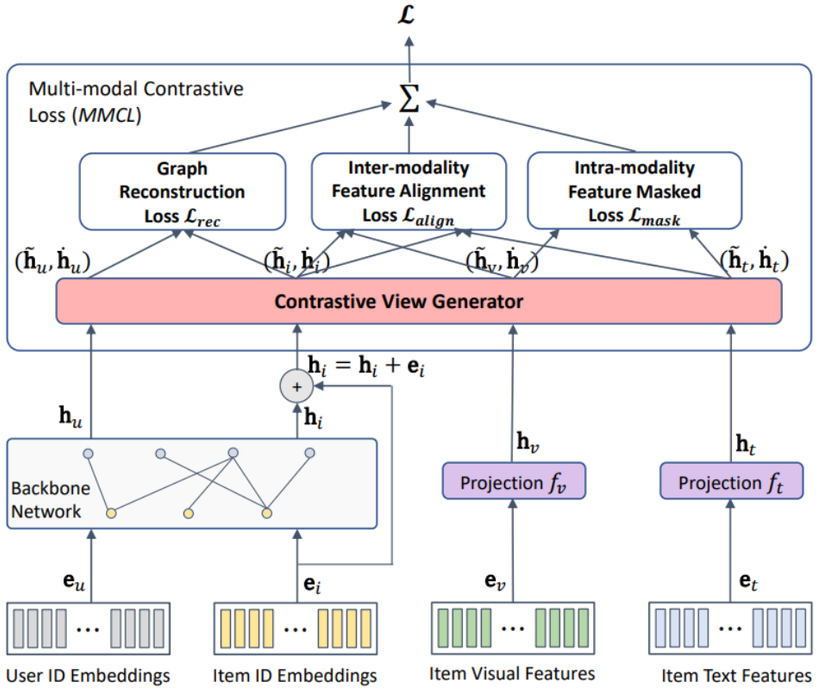
- Multimodal fusion adapter
- LLM adapter for multimodal recommendation

C4: Representation Transfer

- Multimodal features as side information



MMGCN: leveraging multimodal graph networks

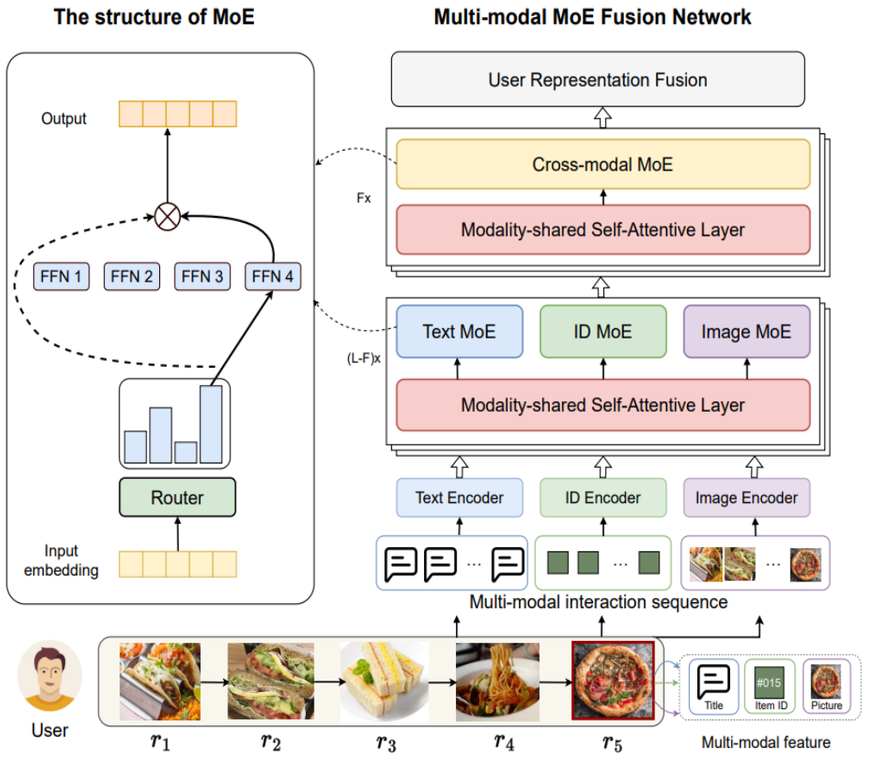


BM3: leveraging ID-modality alignment

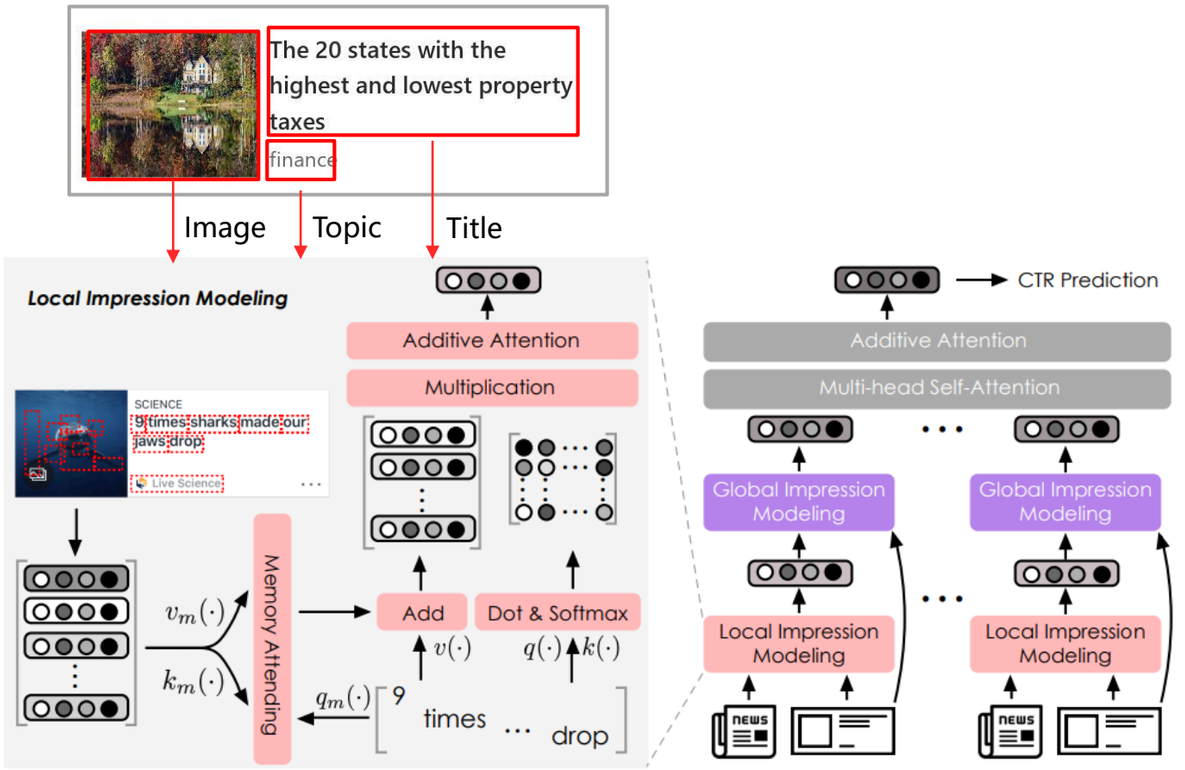
Wei et al., MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video, 2019
Zhou et al., Bootstrap Latent Representations for Multi-modal Recommendation, 2023

C4: Representation Transfer

- User representation learning from multimodal sequences



M3SRec: MOE-based multimodal fusion

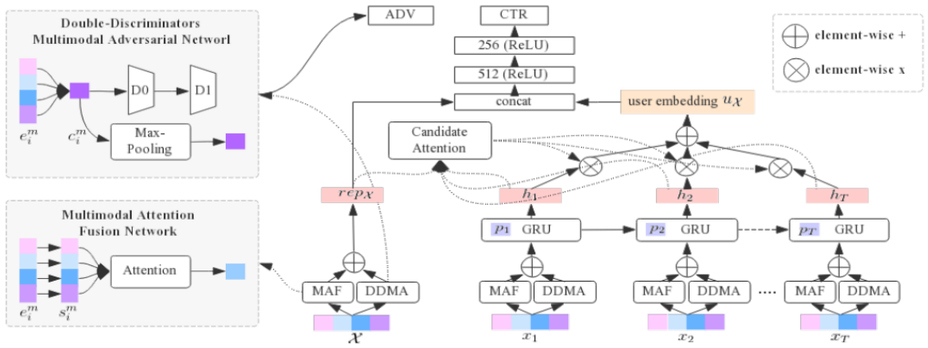


IMRec [Huawei]: local and global fusion

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

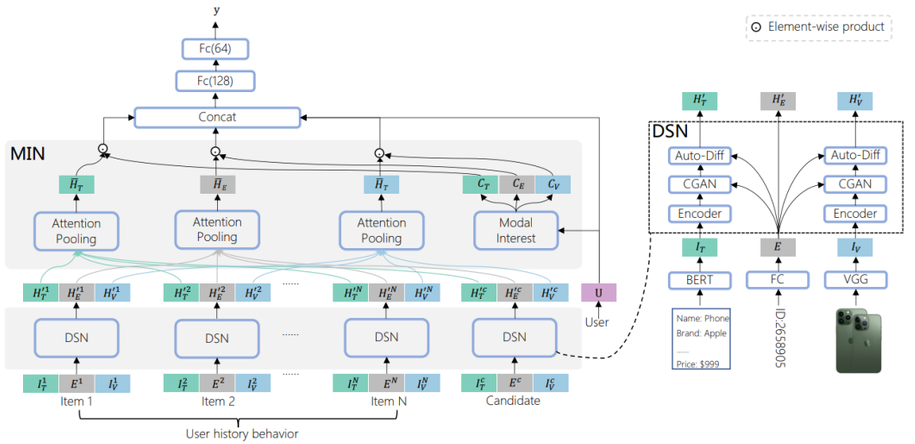
C4: Representation Transfer

- Multimodal sequence denoising and aggregation



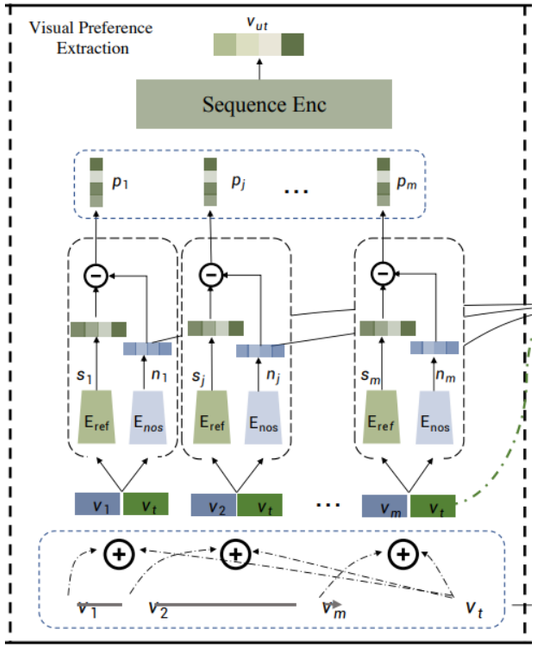
MARN:

- Multimodal attention network
- Multimodal adverseral network



GMMF:

- Difference set network
- Modal-interest network



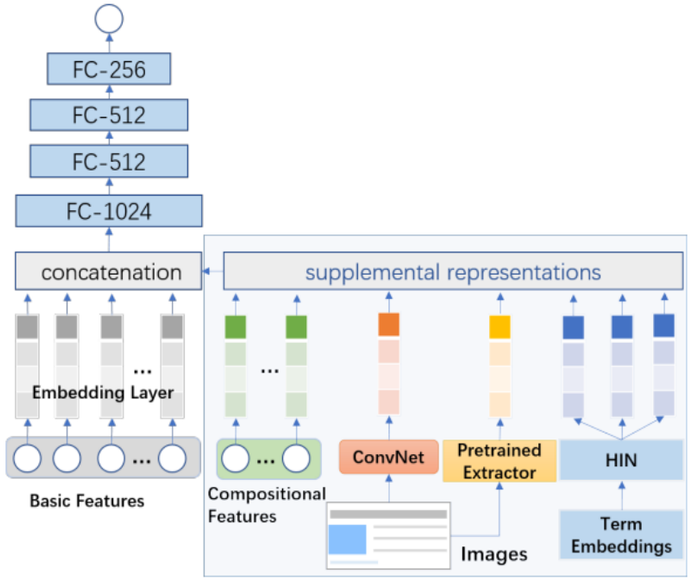
Tavern:

- 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

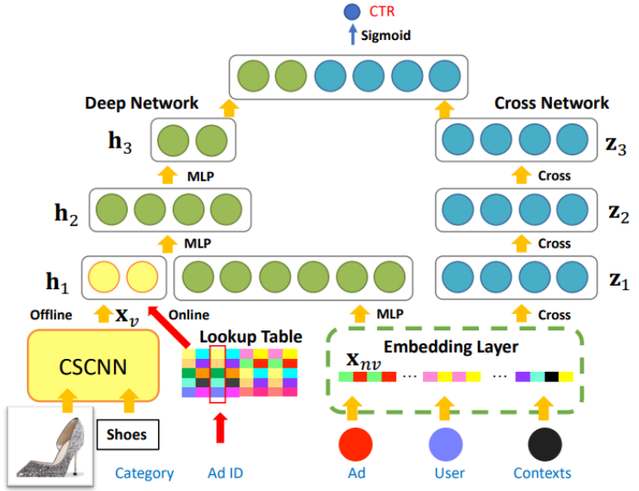
C5: Joint Finetuning

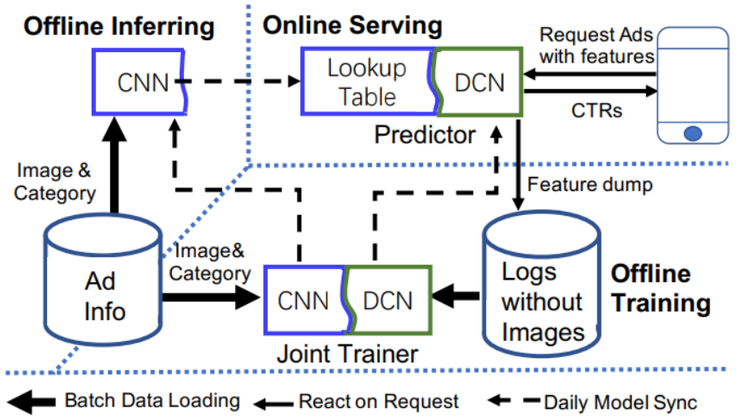
- Only finetuning item-side encoder



HIN-CTR:

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





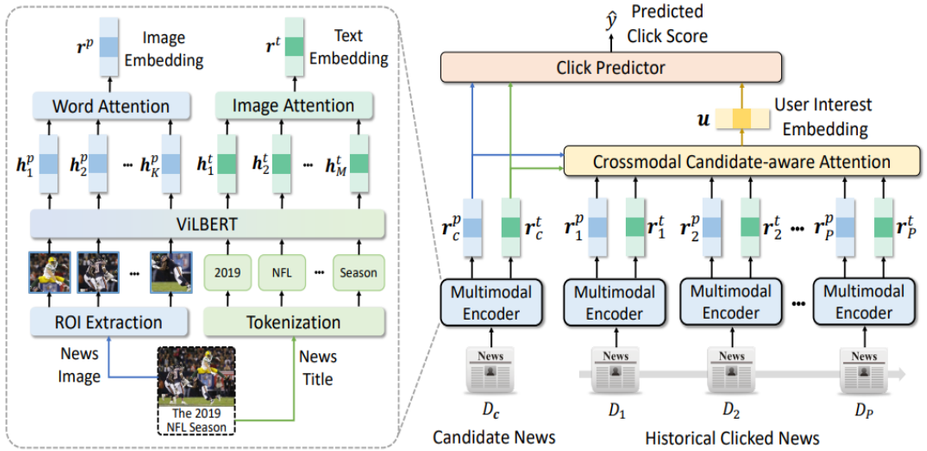
CSCNN:

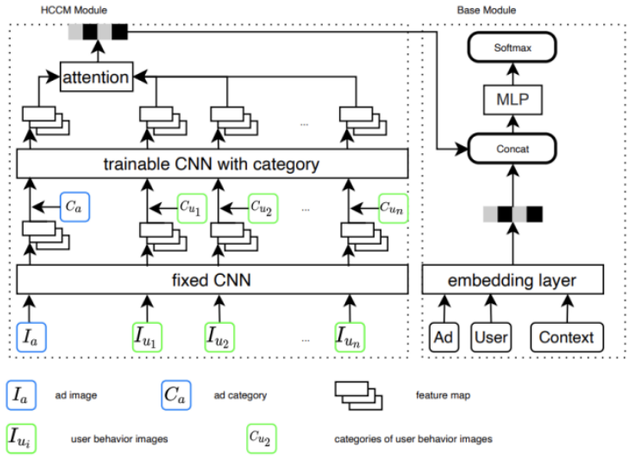
- Leveraging category-aware CNN encoder
- Caching the embeddings after training for online serving

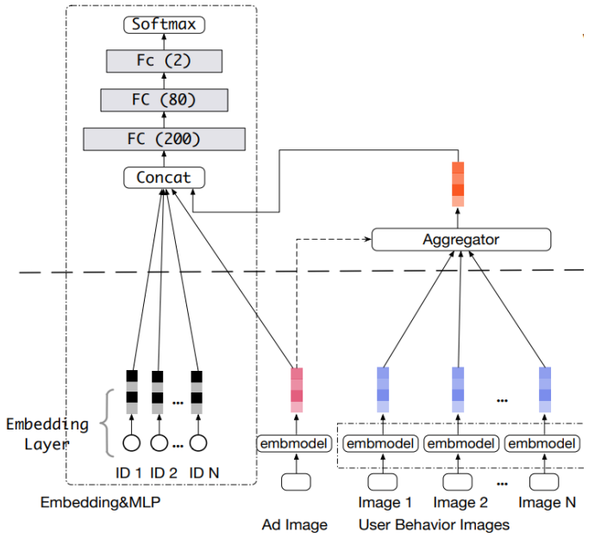
Yang et al., Learning Compositional, Visual and Relational Representations for CTR Prediction in Sponsored Search, 2019
Liu et al., Category-Specific CNN for Visual-aware CTR Prediction at JD.com, 2020

C5: Joint Finetuning

- 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

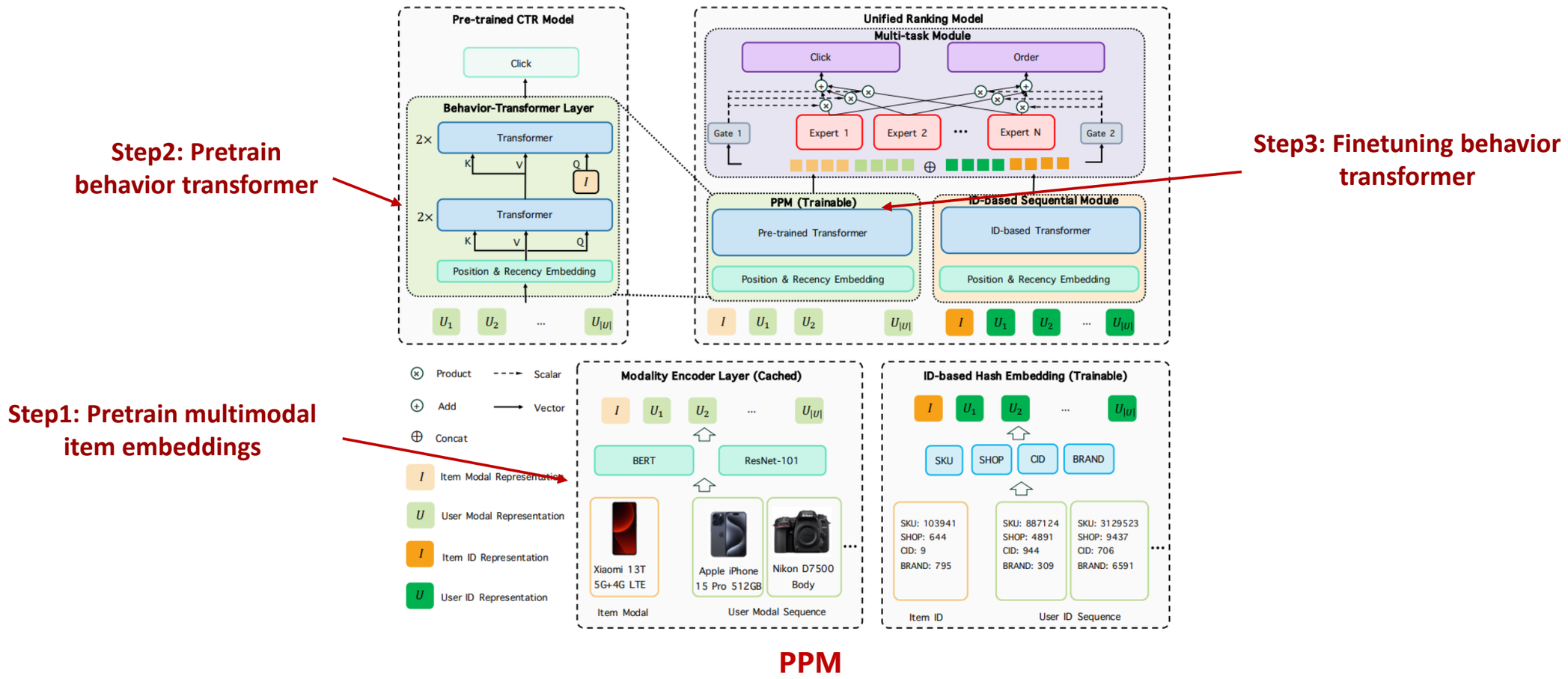
DCIM:

- 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

C5: Joint Finetuning

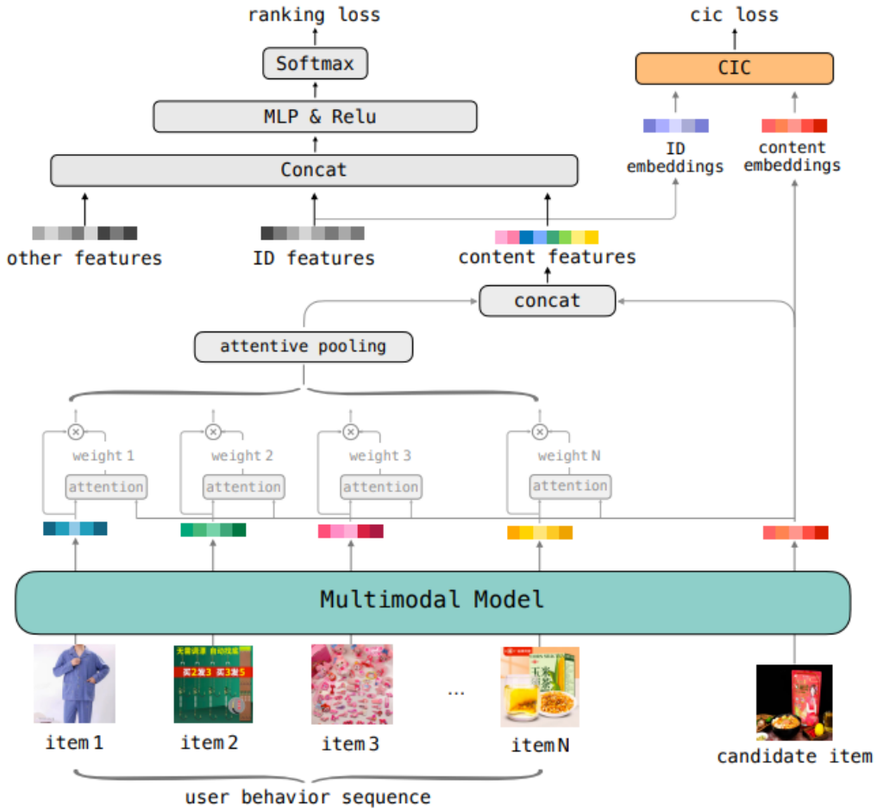
- Finetuning user-item cross encoder

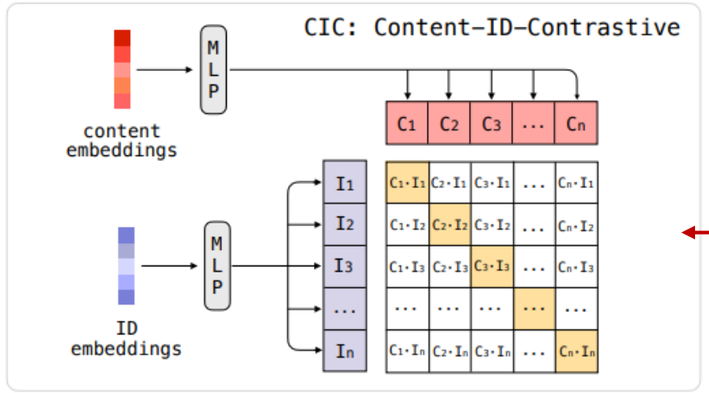


Gao et al, PPM : A Pre-trained Plug-in Model for Click-through Rate Prediction, 2024

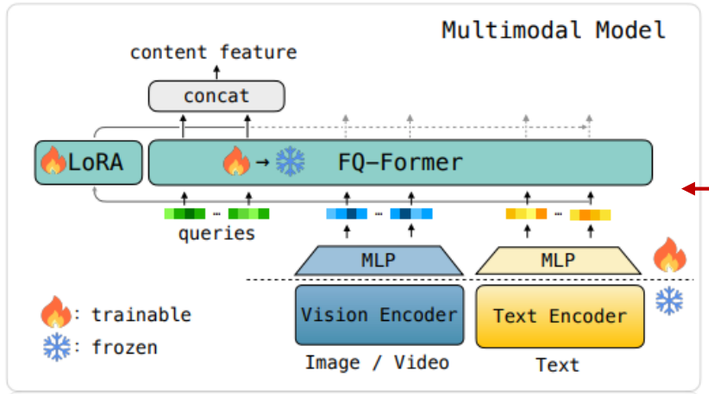
C6: Adapter Tuning

- Multimodal fusion adapter





Content-ID contrastive learning



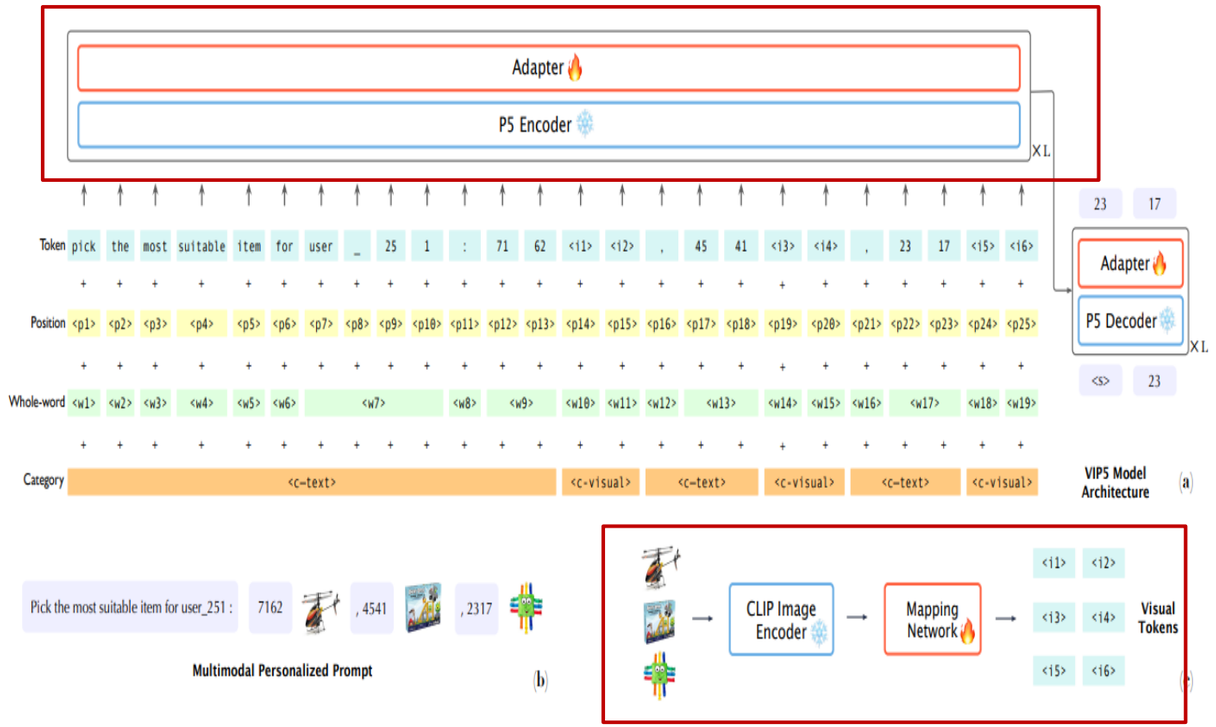
Multimodal fusion adapter

EM3

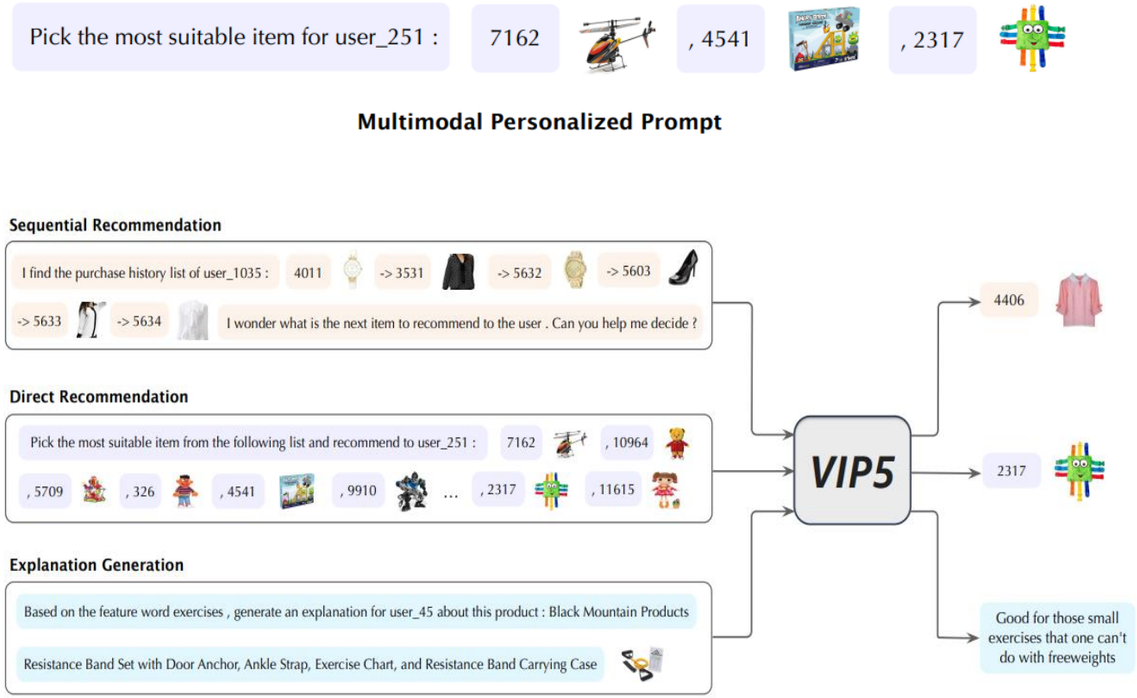
Deng et al., End-to-end Training of Multimodal Model and Ranking Model, 2024

C6: Adapter Tuning

- LLM adapter for multimodal recommendation



VIP5: P5 + Adapter



Geng et al., VIP5: Towards Multimodal Foundation Models for Recommendation, 2024

Summarization

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

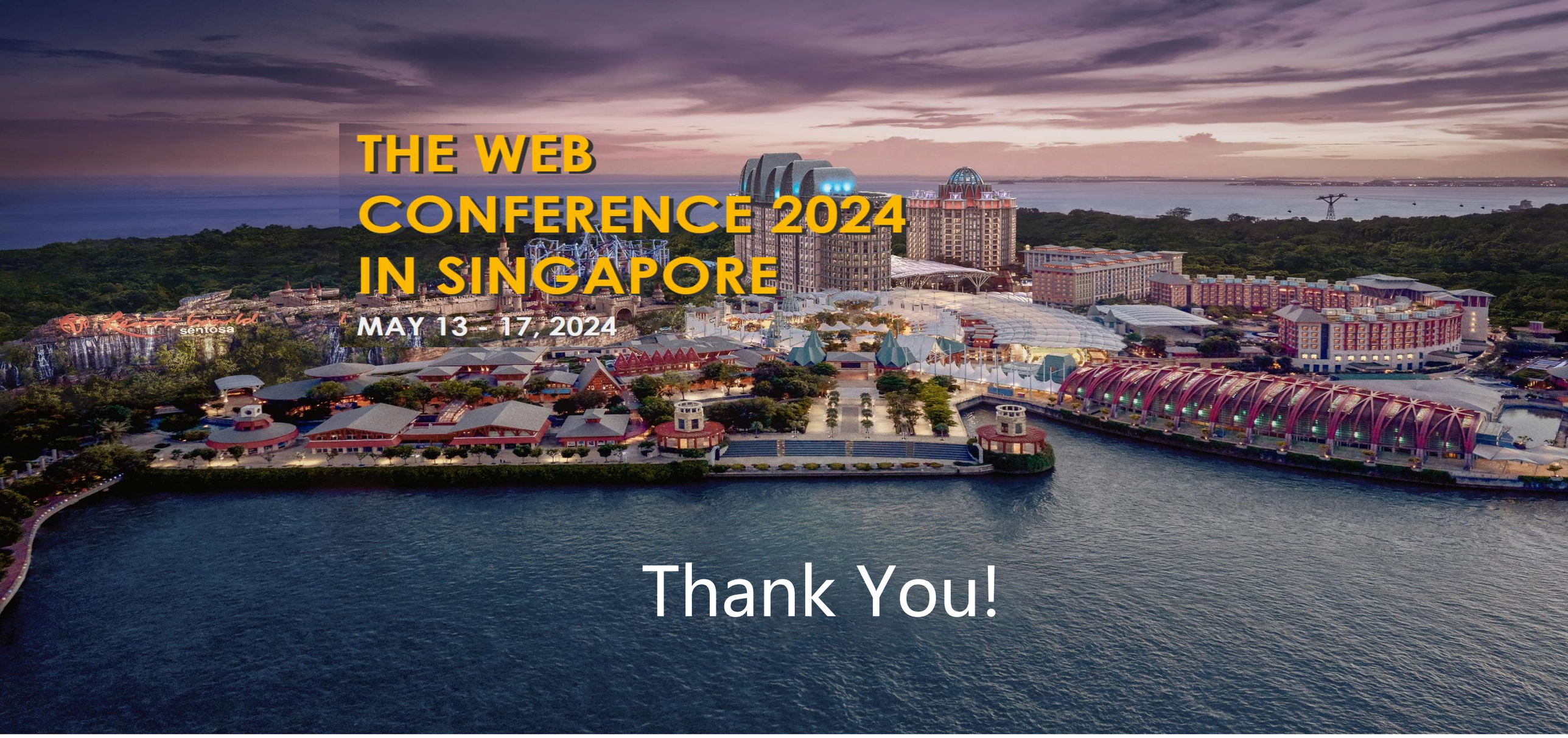
What's Next?

- **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

We are hiring!

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

Please send your CV to jiemingzhu@ieee.org



Thank You!

HUAWEI

Multimodal Generation for Recommendation

Rui Zhang

www.ruizhang.info

Multimodal Generation

midjourney

https://www.midjourney.com/

Sora

https://openai.com/index/sora/

Can we make them personal?

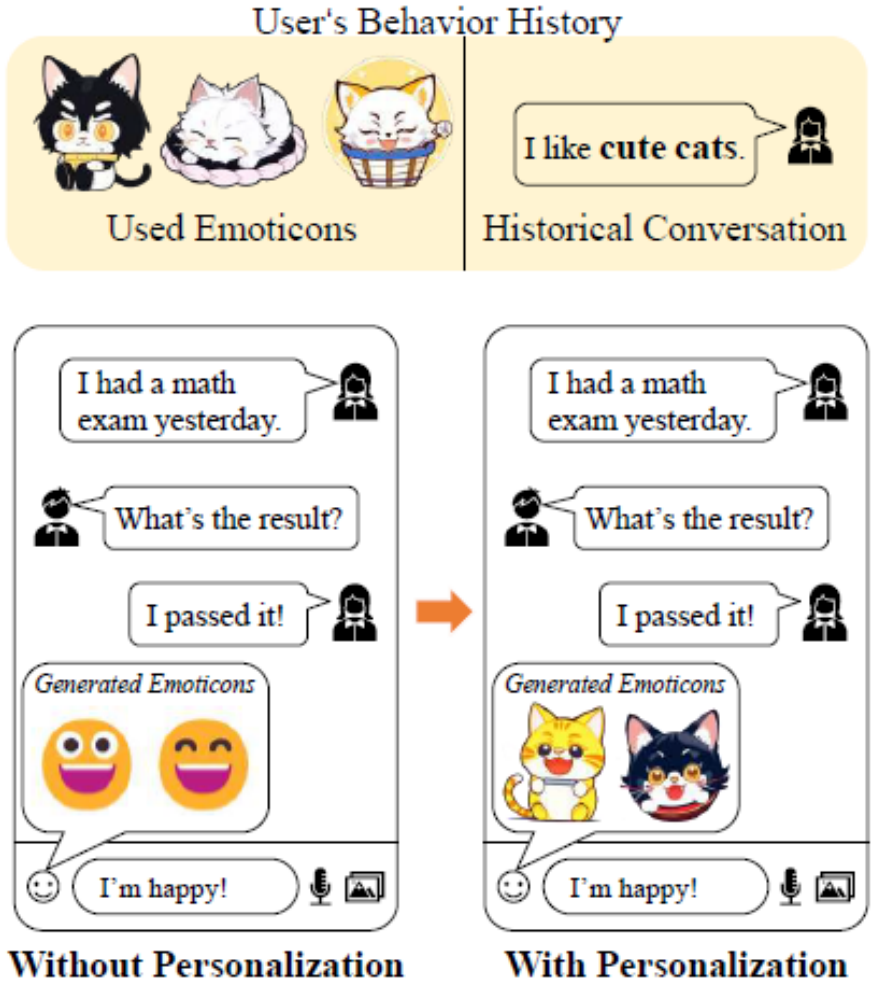
Table of Contents

Term: LLM – language models with capabilities similar to chatgpt, such as llama, claude, gemini, etc

Multimodal Pretraining and Generation for Recommendation: A Tutorial, Web Conference 2024
Multimodal Pretraining, Adaptation, and Generation for Recommendation: A Survey, arXiv:2404.00621

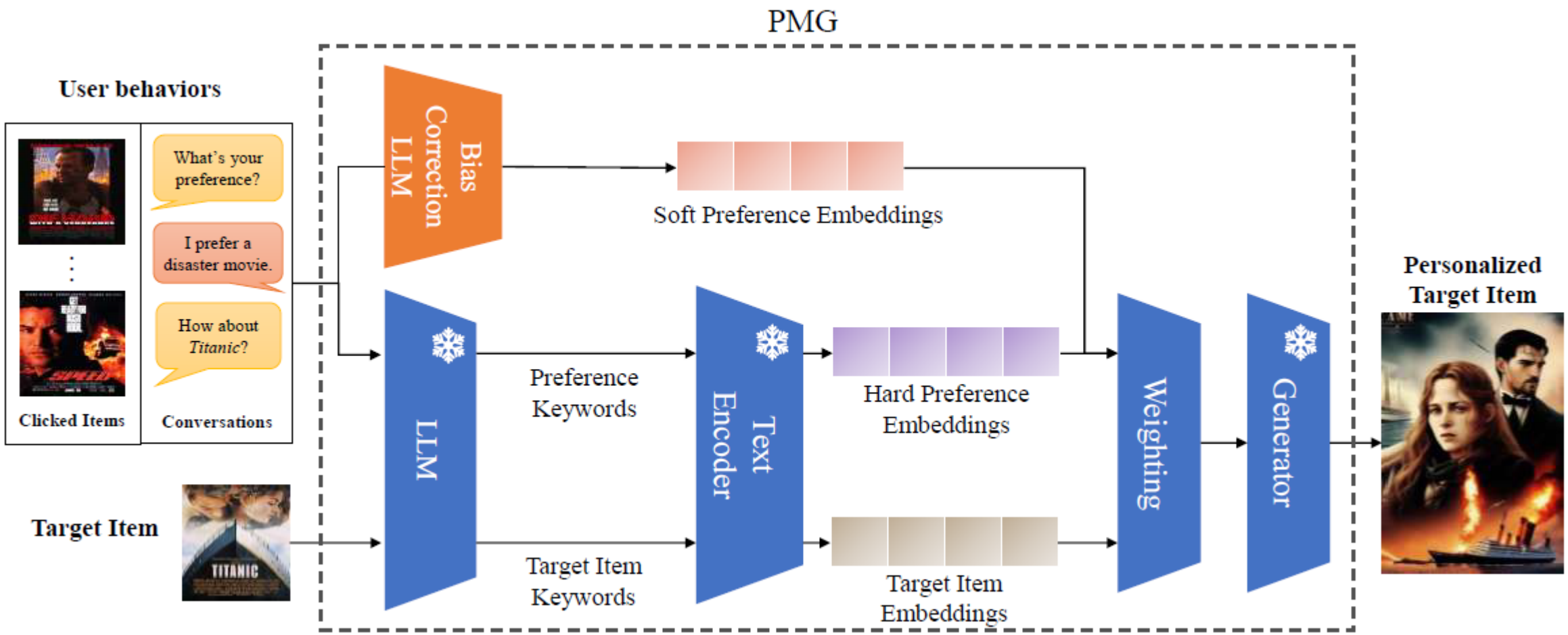
PMG for Recommendation: multimodal → image w/ LLM

- **PMG: Personalized Multimodal Generation with LLM**
  - Converts user behaviors (conversations, clicks, etc) into natural language
  - Extract user preference descriptions, both hard and soft preference embeddings
  - Preference conditioned multimodal generation
  - Improves 8% in terms of personalization measure

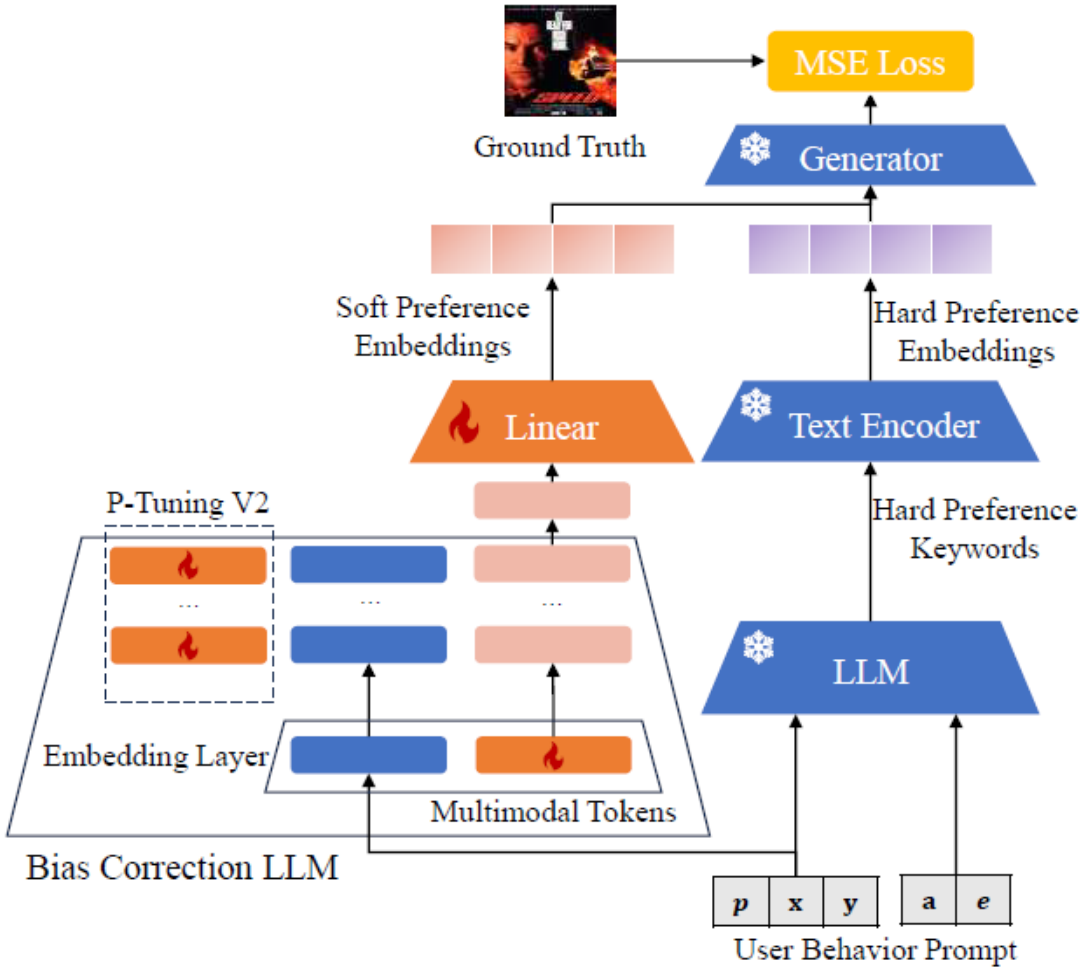


PMG : Personalized Multimodal Generation with Large Language Models, The Web Conference 2024
Friday 17 May 2024: 2:30 - 4pm Poster Session

PMG for Recommendation: multimodal → image w/ LLM



PMG for Recommendation: multimodal → image w/ LLM



$$\mathbf{E}^p = concatenate(\mathbf{E}_m, \mathbf{E}_k)$$

$$M_n = M_s + \epsilon,$$

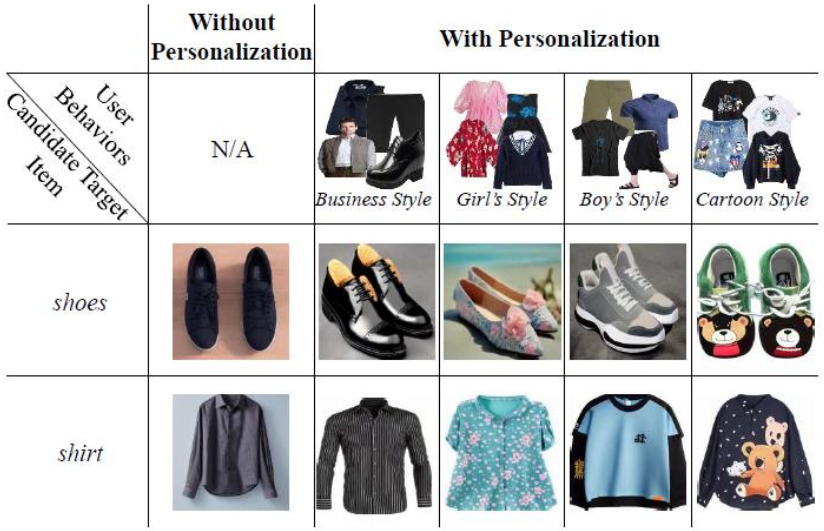
$$M_d = Unet(\mathbf{E}^p, M_n).$$

The loss is calculated as MSE loss of M_s and M_d:

$$loss = MSE(M_s, M_d).$$

Figure 3: Model designed to train soft preference embeddings.

PMG for Recommendation: multimodal → image w/ LLM



$$d_p = \frac{e_M \cdot e_p}{\|e_M\|_2 \|e_p\|_2},$$

$$d_t = \frac{e_M \cdot e_t}{\|e_M\|_2 \|e_t\|_2}.$$

Finally, our objective is to optimize the weighted sum of d_p and d_t.

$$z = \alpha \cdot \log d_p + (1 - \alpha) \cdot \log d_t.$$

(a) $w_p : w_t = 0 : 4$

(b) $w_p : w_t = 1 : 3$

(c) $w_p : w_t = 2 : 2$

(d) $w_p : w_t = 3 : 1$

(e) $w_p : w_t = 4 : 0$

Figure 7: Generated poster of movie *Titanic* with different weights of conditions. w_p is the weight of preference conditions, which prefer disaster movie. w_t is the weight of target item conditions, which consider it as a romantic movie. When $w_p : w_t = 1 : 3$ it achieves the highest z score and the generated poster is a combination of romance and disaster.

PMG for Recommendation: multimodal → image w/ LLM

Data

1) Generating personalized images of products whose original images are missing according to the historically clicked products of the user. POG dataset, a multimodal dataset of fashion clothes. We selected 2,000 users and 16,100 items for experiments.
2) Generating personalized posters of movies according to historical watched movies of user. MovieLens Latest Datasets, 9,000 movies, 600 users, and 100,000 rating interactions.
3) Generating emoticons in instant messaging according to current conversation and historically used emoticons of the user. We do not train soft preference embeddings and only use keywords to generate images.

	Movie Posters Scenario	Clothes Scenario
PMG	2.587	2.001
Textual Inversion	1.952	1.725
No personalization	1.462	1.495

Human evaluation score, range (1, 2, 3)

PMG for Preference Questions: multimodal → multimodal w/ V-LM

- **Multi-task Multimodal generation, answering different types of questions**

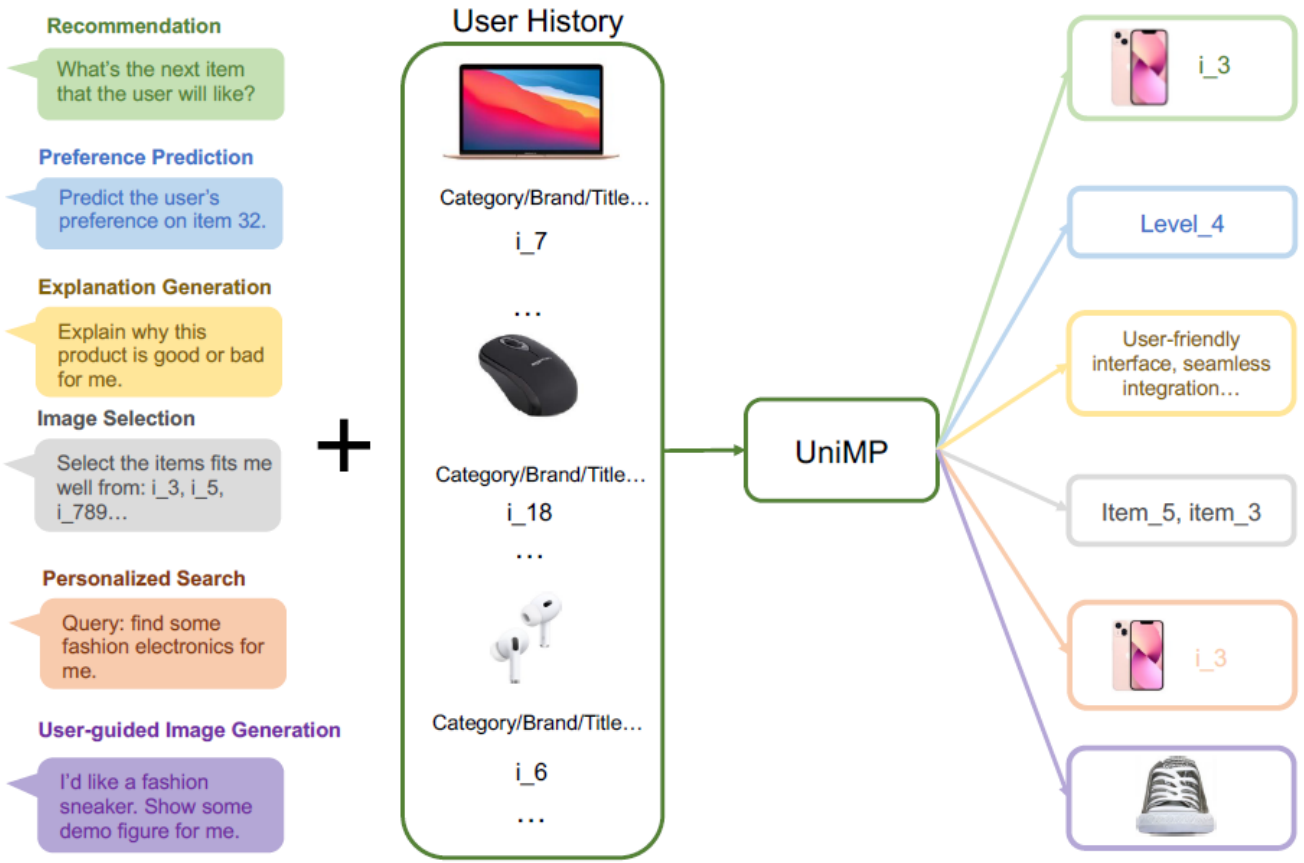


Figure 2: Through multi-task, multi-modal instruction tuning, the model can adapt to a range of user requirements. By altering the instructions, it can generate diverse responses to suit user needs. For

Towards Unified Multi-Modal Personalization: Large Vision-Language Models for Generative Recommendation and Beyond, ICLR 2024

PMG for Preference Questions: multimodal → multimodal w/ V-LM

- **Item contextual data is serialized and processed through fine-grained cross-modal fusion**

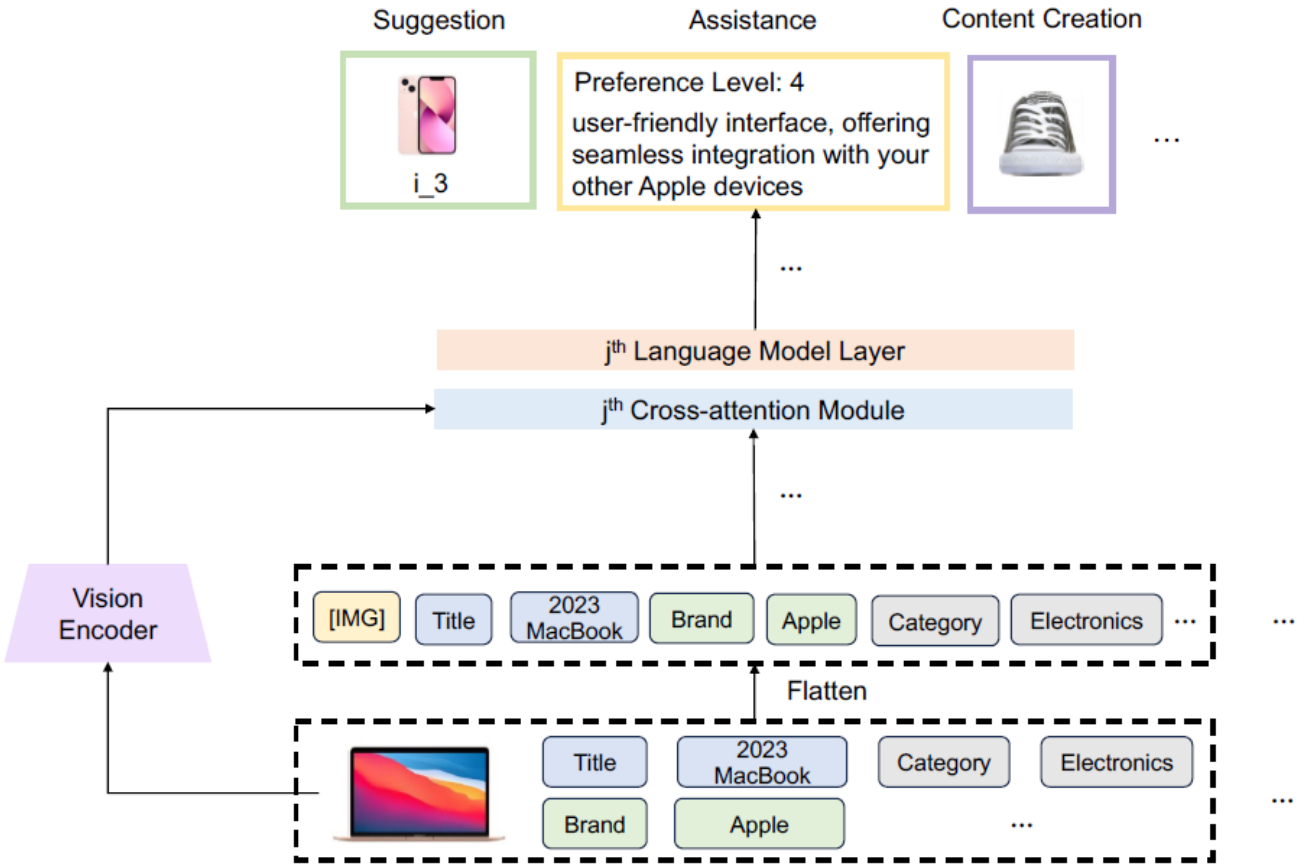


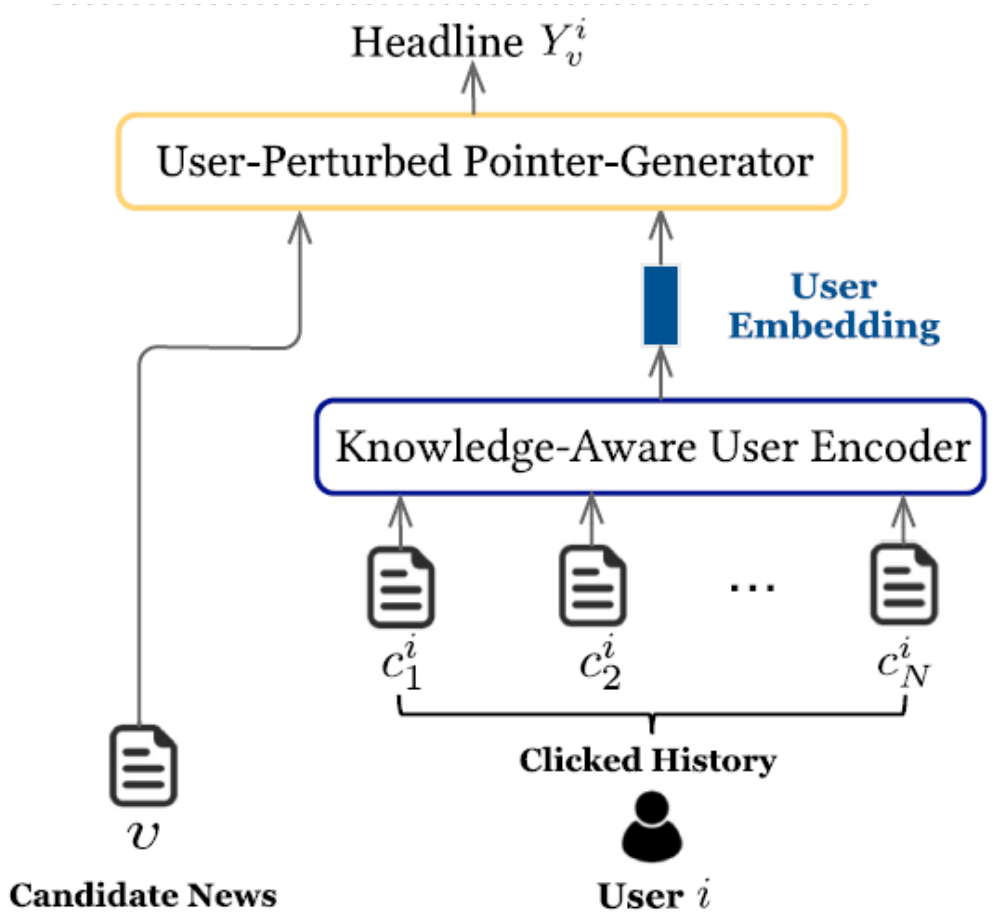
Figure 1: Our proposed UniMP framework operates as follows: Item contextual data is streamlined into a user sequence, which is then processed through fine-grained cross-modal fusion. Depending on the instructions, the output is tailored to produce diverse response types.

Personalized Generation: text → text w/o LLM

News Headline Generation

Put Your Voice on Stage: Personalized Headline Generation for News Articles, TKDD 2023

- Framework
- Evaluation
  - Automtaic
    - Informativeness: F1 ROUGE
    - Fluency: longest common subsequence (ROUGE-L)
  - Human evaluation



Framework

Put Your Voice on Stage: Personalized Headline Generation for News Articles, TKDD 2023

Personalized Generation: item → text w/o LLM

■ Personalized Answer Generation in E-commerce

Product	Question	Does this lens work with Nikon camera?
 *Tokina Lens*	**Relevant Review Snippets**	1. …Using it on Nikon D7100. It's **cheaper** than others! … 2. … I did great testing with D5000, **D5100**, and D7200. … 3. … **Light weight glass**, at first seemed elegant. … …

Personalized Answers

- This lens will work and **pay a little less** probably. — *User1*
- Yes, it will work. It works beautifully with **D5100**. — *User2*
- Yes the lens works great, and it is a **light weight glass**. — *User3*

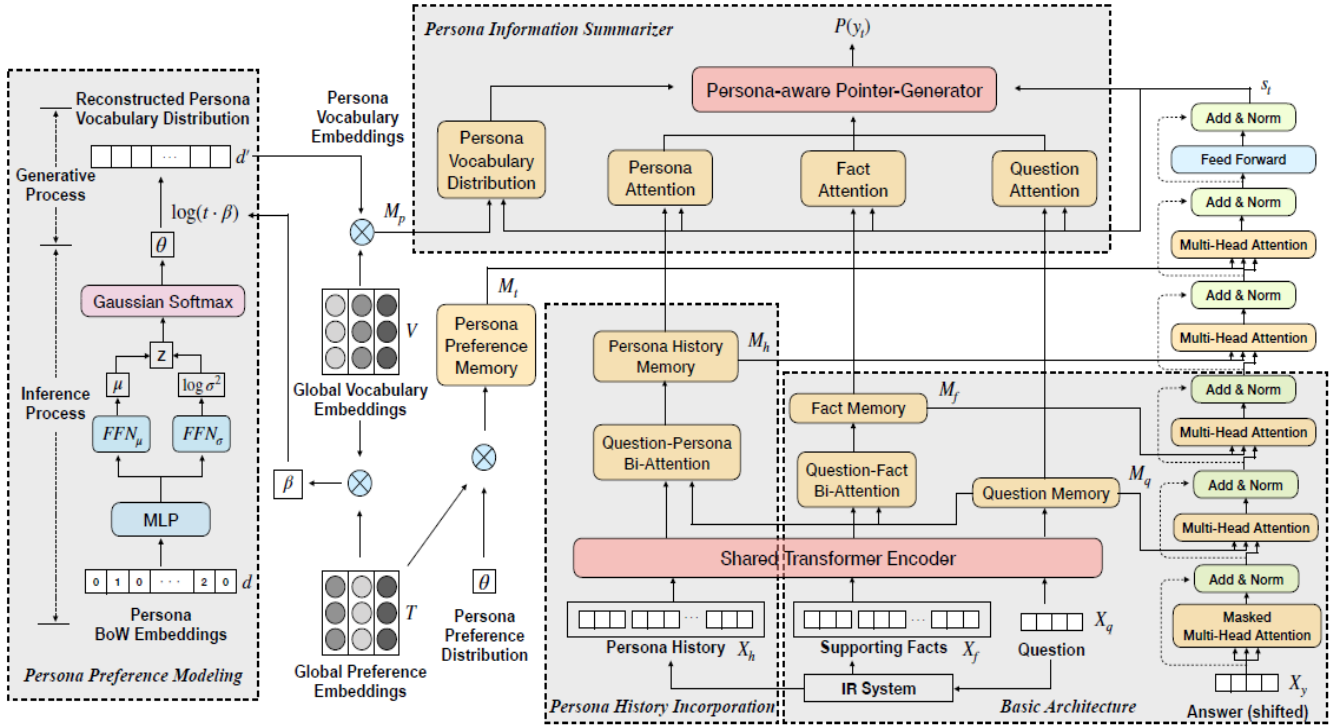


Fig. 3. Overview of the proposed method PAGE, including four components: (1) Basic Encoder-decoder Architecture, (2) Persona History Incorporation, (3) Persona Preference Modeling, and (4) Persona Information Summarizer.

Towards Personalized Answer Generation in E-Commerce via Multi-Perspective Preference Modeling, TOIS 2022

Personalized Generation: text → text w/ LLM

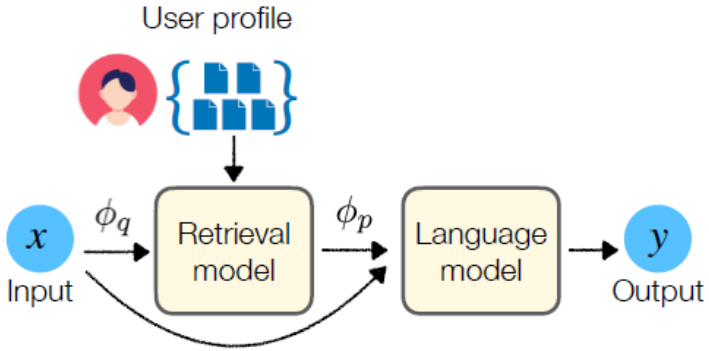
■ Benchmark, RAG (Retrieval Augmented Generation) paradigm

LaMP: When Large Language Models Meet Personalization, arXiv:2304.11406

- 7 Tasks

  - **Personalized Text Classification**
    - (1) Personalized Citation Identification
    - (2) Personalized Movie Tagging
    - (3) Personalized Product Rating
  - **Personalized Text Generation**
    - (4) Personalized News Headline Generation
    - (5) Personalized Scholarly Title Generation
    - (6) Personalized Email Subject Generation
    - (7) Personalized Tweet Paraphrasing

- Using RAG paradigm



Personalized Generation: text → text w/ LLM & Human

- **LLM-assisted news headline generation**
  - Human-AI Text Co-Creation

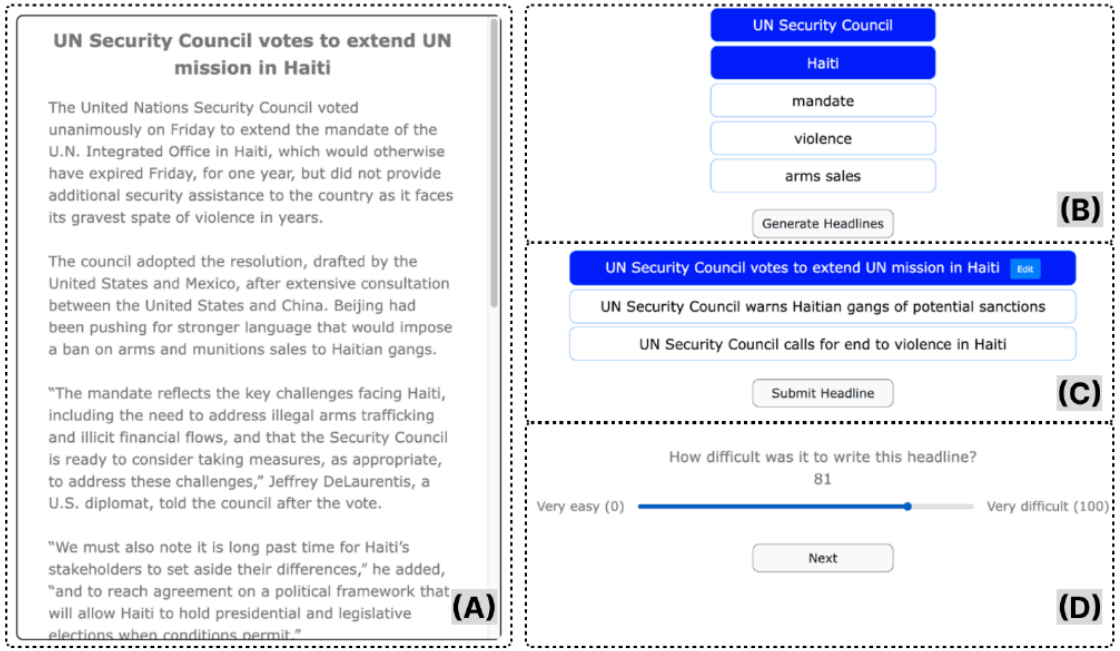


Figure 2: Interface for human-AI news headline co-creation for *guidance* + *selection* + *post-editing* condition: (A) news reading panel, (B) perspectives (keywords) selection panel (multiple keywords can be selected), (C) headline selection panel with post-editing capability, and (D) difficulty rating slider. Note: (B), (C) and (D) are hidden from the user until the requisite step is finished (e.g., the user does not see the difficulty

Harnessing the Power of LLMs: Evaluating Human-AI Text Co-Creation through the Lens of News Headline Generation, EMNLP 2023

(non-Personalized) Multimodal Generation: multimodal → multimodal

■ Multi-modal News Headline Generation

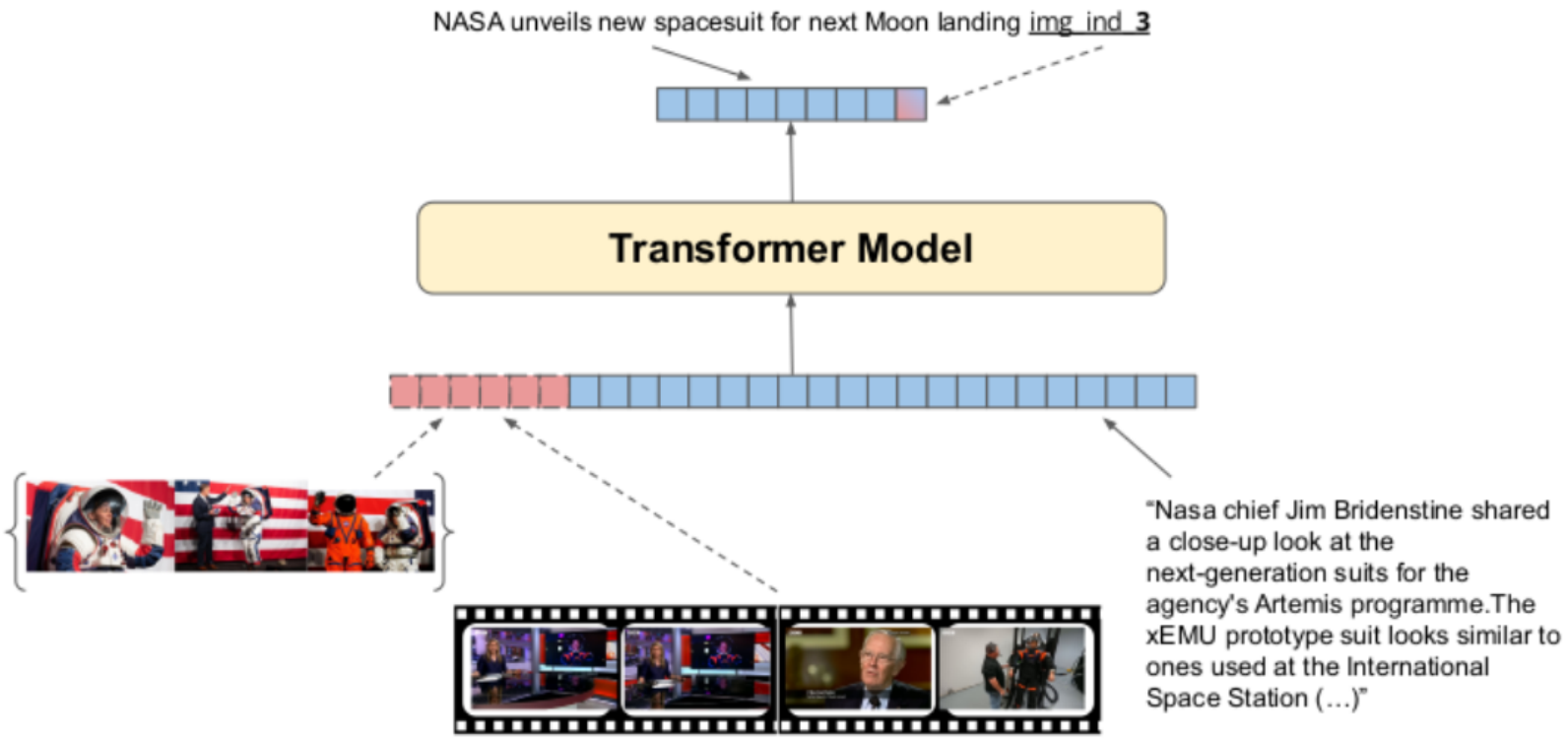


Figure 1: Overview of the proposed unified approach to MSMO. The visual tokens are appended to the text representation. The generated output includes the textual summary and the *index token* that indicates which input image (first, second, third, etc.) is picked as the pictorial summary. During training, a mixture of video-based, image-based, and text-only data is used.

Towards Unified Uni- and Multi-modal News Headline Generation, EACL 2024

Other Tasks of Multimodal Generation for Recommendation

Marketing Copy Generation

- Generate the promotional copy

GCOF: Self-iterative Text Generation for Copywriting Using Large Language Model, arXiv:2402.13667

Explanation Generation

- Generate reasons why an item is recommended

Personalized Reason Generation for Explainable Song Recommendation. TIST 2019

Dialogue Generation

- Generate questions for clarification during conversational search

Zero-shot Clarifying Question Generation for Conversational Search, Web Conference 2023

What's Next

- **Multimodal → multimodal for Recommendation**
- **Improve the control of correctness (text, image, video, etc)**
- **Include more modalities, such as audio, video**
- **Interactive multimodal generation**

Thanks and Questions?

Hiring junior academics, postdocs, PhD students

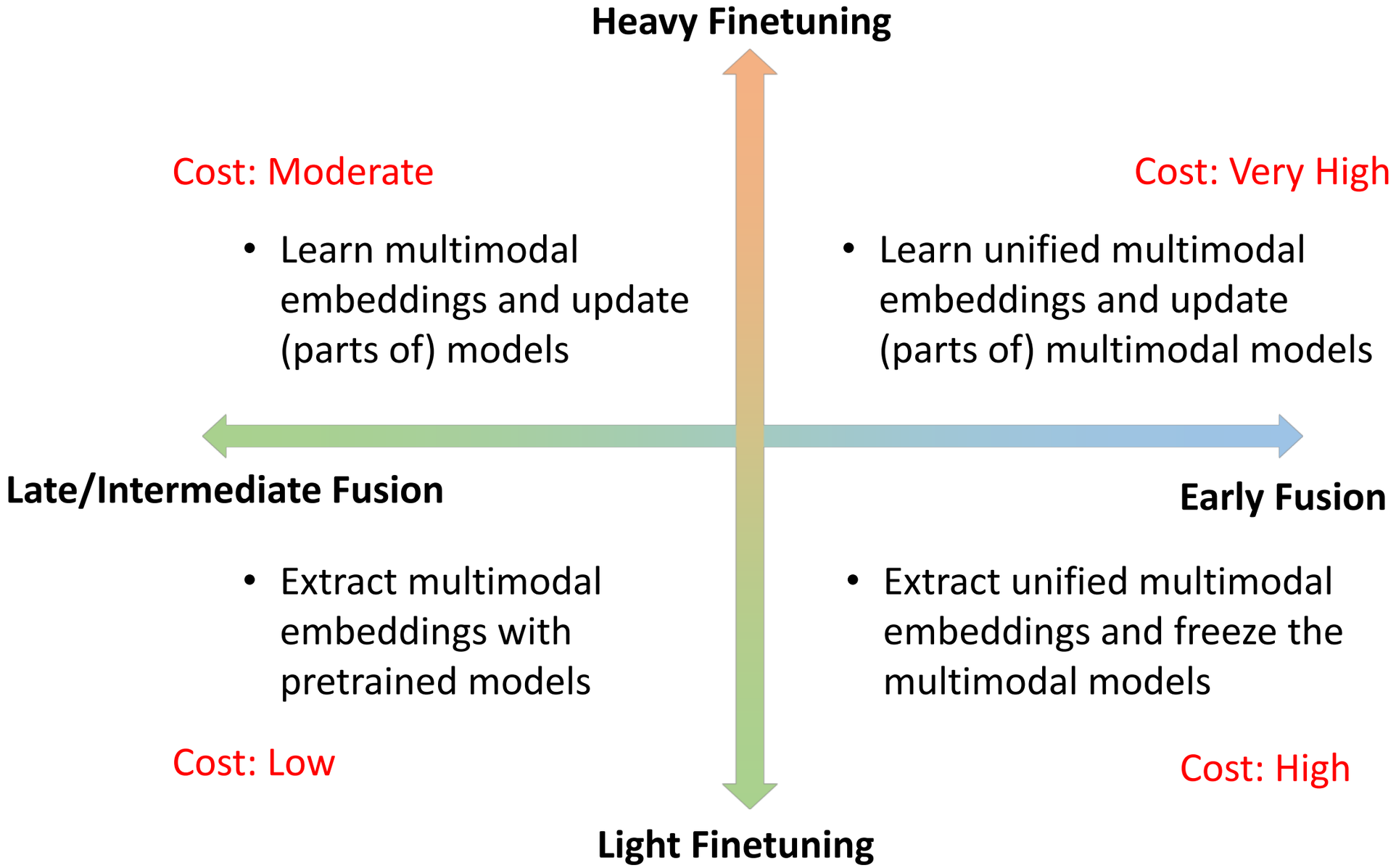
Contact email:

rayteam@yeah.net

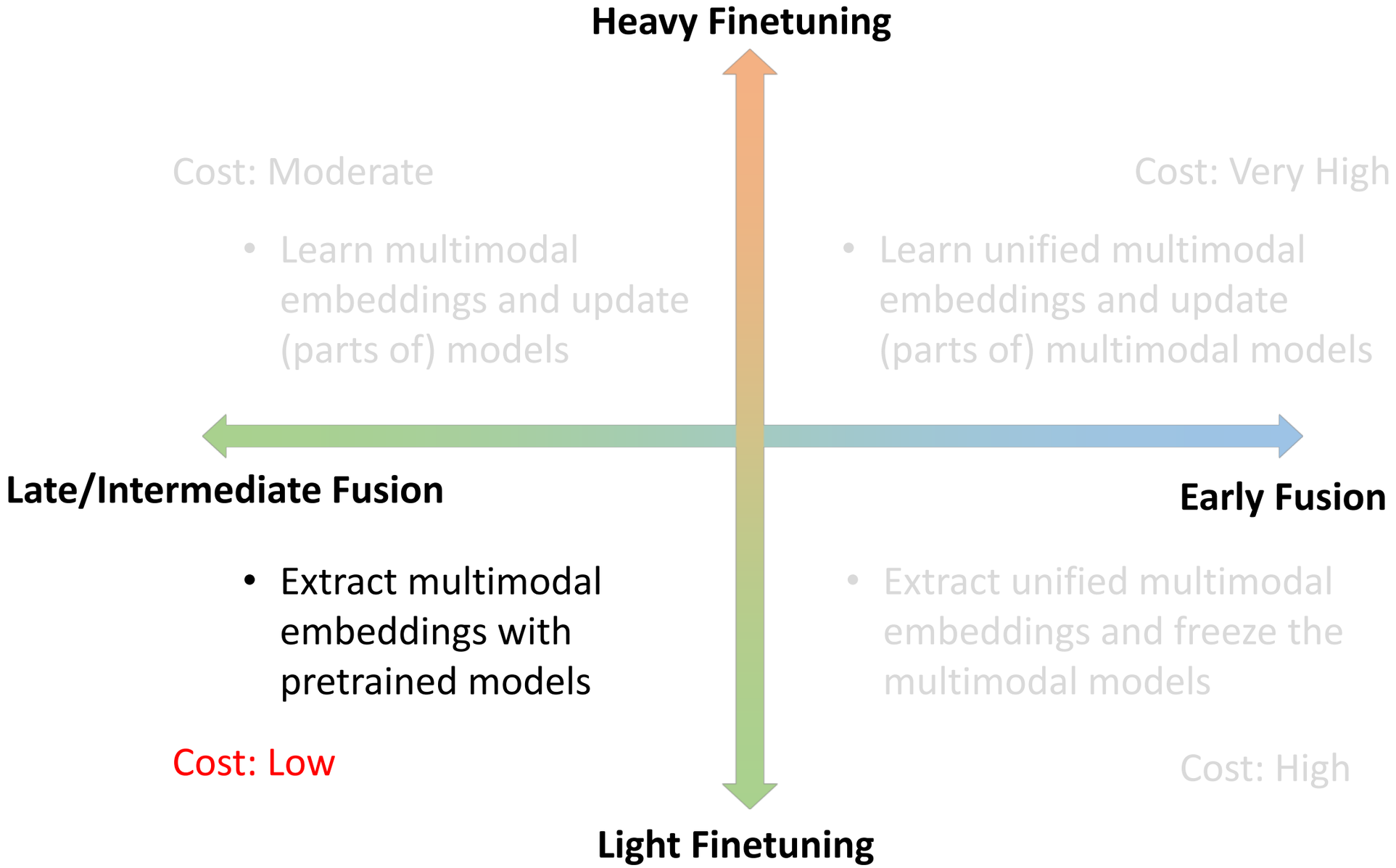
Industrial Applications and Open Challenges in Multimodal Recommendation

Chuhan Wu
Huawei Noah's Ark Lab
https://wuch15.github.io
WWW 2024 @ Singapore

Multimodal Recommendation: Fusion and Finetuning Matters

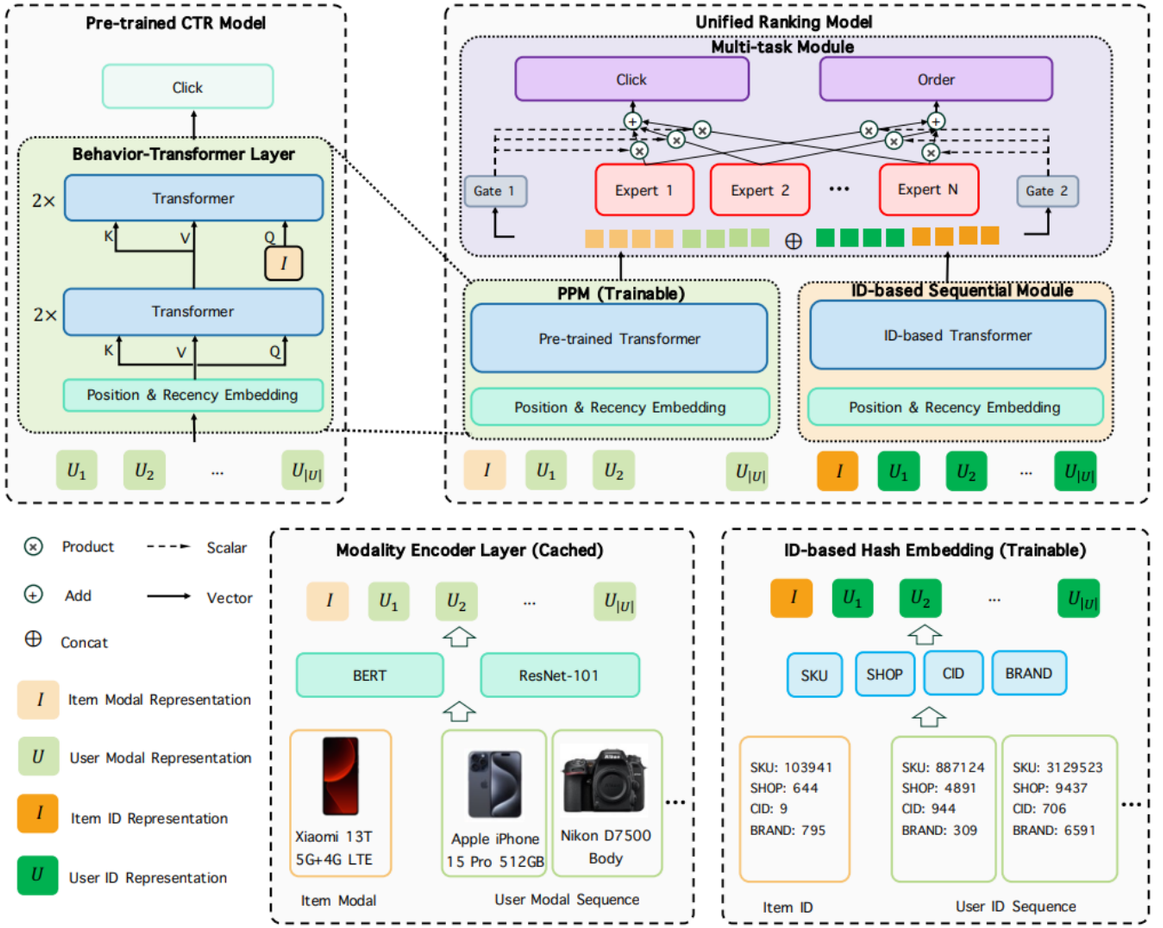


Multimodal Recommendation: Fusion and Finetuning Matters



PPM (JD.com, 2024)

- Extract multimodal item embeddings with adapted BERT and ResNet



- BERT: adapted on Query Matching Task to learn relevance signal
- ResNet: adapted on Entity Prediction Task to capture key entities
- Cache the multimodal item embeddings during recommendation model training

PPM : A Pre-trained Plug-in Model for Click-through Rate Prediction

PPM (JD.com, 2024)

- Multimodal features benefit product recommendation
- Adapted BERT/ResNet are better than the original ones

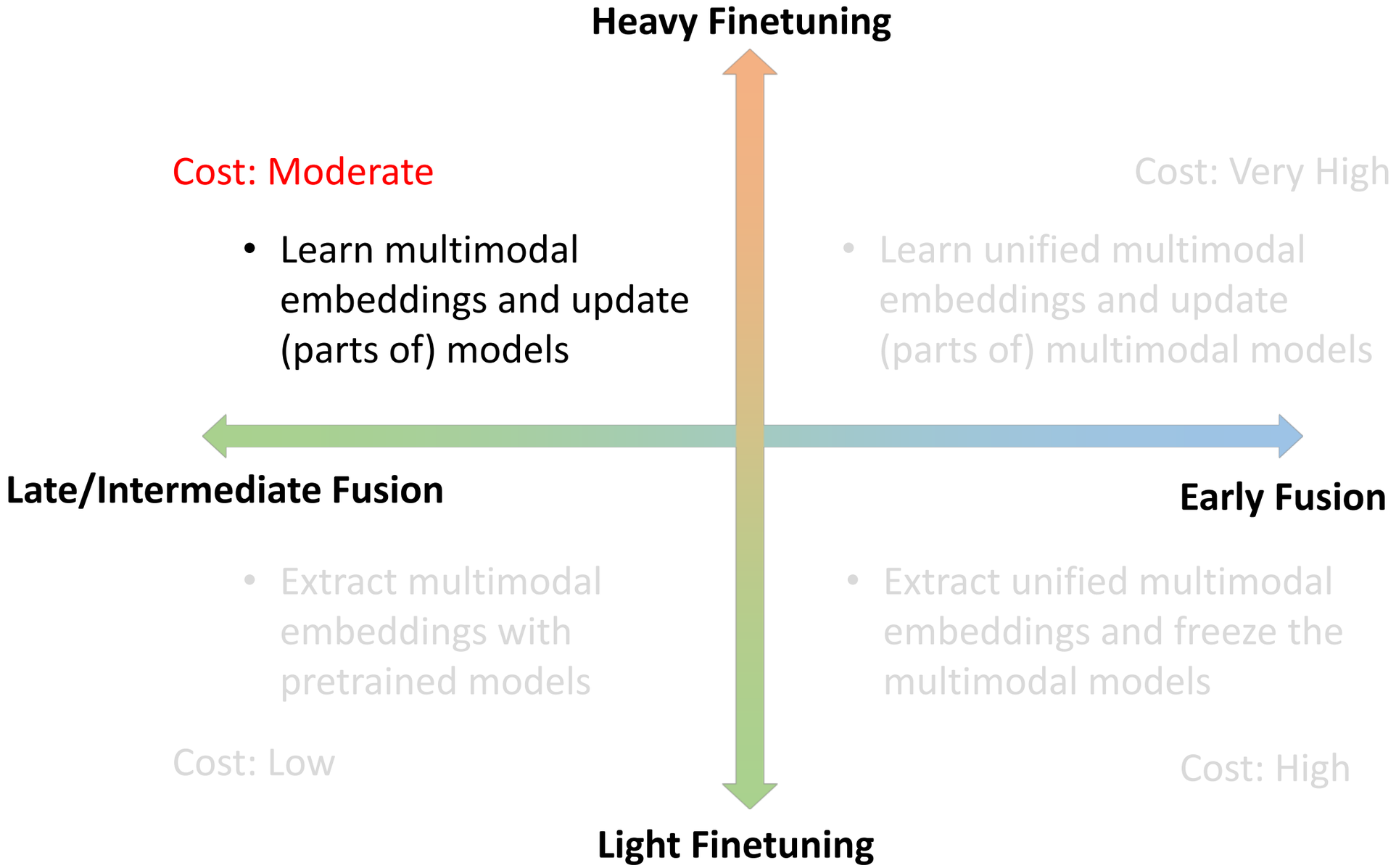
Dataset	Model	wo.PPM						w.PPM					
		Click		Order		Average		Click		Order		Average	
		AUC	P@2	AUC	P@2	$\overline{AUC}$	$\overline{P@2}$	AUC	P@2	AUC	P@2	$\overline{AUC}$(Improv)	$\overline{P@2}$(Improv)
Small	Wide&Deep	0.6094	0.1886	0.6030	0.1655	0.6062	0.1770	0.6797	0.2029	0.6948	0.1685	**0.6873**(0.0811)	**0.1857**(0.0087)
	DeepFM	0.6051	0.1882	0.6138	0.1688	0.6095	0.1785	0.6784	0.2028	0.6920	0.1686	**0.6852**(0.0758)	**0.1857**(0.0072)
	DIN	0.6946	0.2168	0.7231	0.2013	0.7089	0.2091	0.7093	0.2242	0.7451	0.2099	**0.7272**(0.0183)	**0.2171**(0.0080)
	DSIN	0.7000	0.2189	0.7332	0.2055	0.7166	0.2122	0.7107	0.2241	0.7422	0.2084	**0.7265**(0.0098)	**0.2163**(0.0041)
	MAIN	0.6859	0.2170	0.7174	0.1973	0.7016	0.2072	0.6926	0.2225	0.7279	0.2030	**0.7102**(0.0086)	**0.2127**(0.0056)
	URM	0.7228	0.2283	0.7648	0.2301	0.7438	0.2292	0.7279	0.2324	0.7676	0.2325	**0.7477**(0.0040)	**0.2324**(0.0023)
Large	Wide&Deep	0.6011	0.1883	0.6129	0.1668	0.6070	0.1775	0.6848	0.2065	0.7021	0.1734	**0.6935**(0.0865)	**0.1899**(0.0124)
	DeepFM	0.6046	0.1875	0.6233	0.1749	0.6140	0.1812	0.6853	0.2069	0.7024	0.1737	**0.6939**(0.0799)	**0.1903**(0.0091)
	DIN	0.7038	0.2208	0.7373	0.2063	0.7205	0.2135	0.7130	0.2261	0.7482	0.2106	**0.7306**(0.0100)	**0.2184**(0.0048)
	DSIN	0.7069	0.2232	0.7416	0.2087	0.7243	0.2159	0.7125	0.2254	0.7464	0.2107	**0.7295**(0.0052)	**0.2181**(0.0021)
	MAIN	0.6945	0.2215	0.7266	0.2014	0.7106	0.2114	0.6996	0.2251	0.7287	0.2038	**0.7141**(0.0036)	**0.2145**(0.0030)
	URM	0.7279	0.2326	0.7685	0.2323	0.7482	0.2324	0.7343	0.2363	0.7722	0.2313	**0.7532**(0.0050)	**0.2338**(0.0014)

Table 2: The overall performance comparison with other baseline methods

Model	Click		Order		Average	
	AUC	P@2	AUC	P@2	$\overline{AUC}$ (Improv)	$\overline{P@2}$ (Improv)
Base	0.7252	0.2310	0.7612	0.2305	0.7432 (-)	0.2308 (-)
Base+QM&EP	0.7279	0.2326	0.7685	0.2323	0.7482 (0.0050)	0.2324 (0.0017)
Base+QM&EP+PPM (random initialized)	0.7277	0.2321	**0.7735**	**0.2351**	0.7506 (0.0074)	0.2336 (0.0028)
Base+QM&EP+PPM (frozen)	0.7318	0.2347	0.7710	0.2317	0.7514 (0.0082)	0.2332 (0.0024)
Base+QM&EP+PPM (finetune)	**0.7343**	**0.2363**	0.7722	0.2313	**0.7532** (0.0100)	**0.2338** (0.0031)

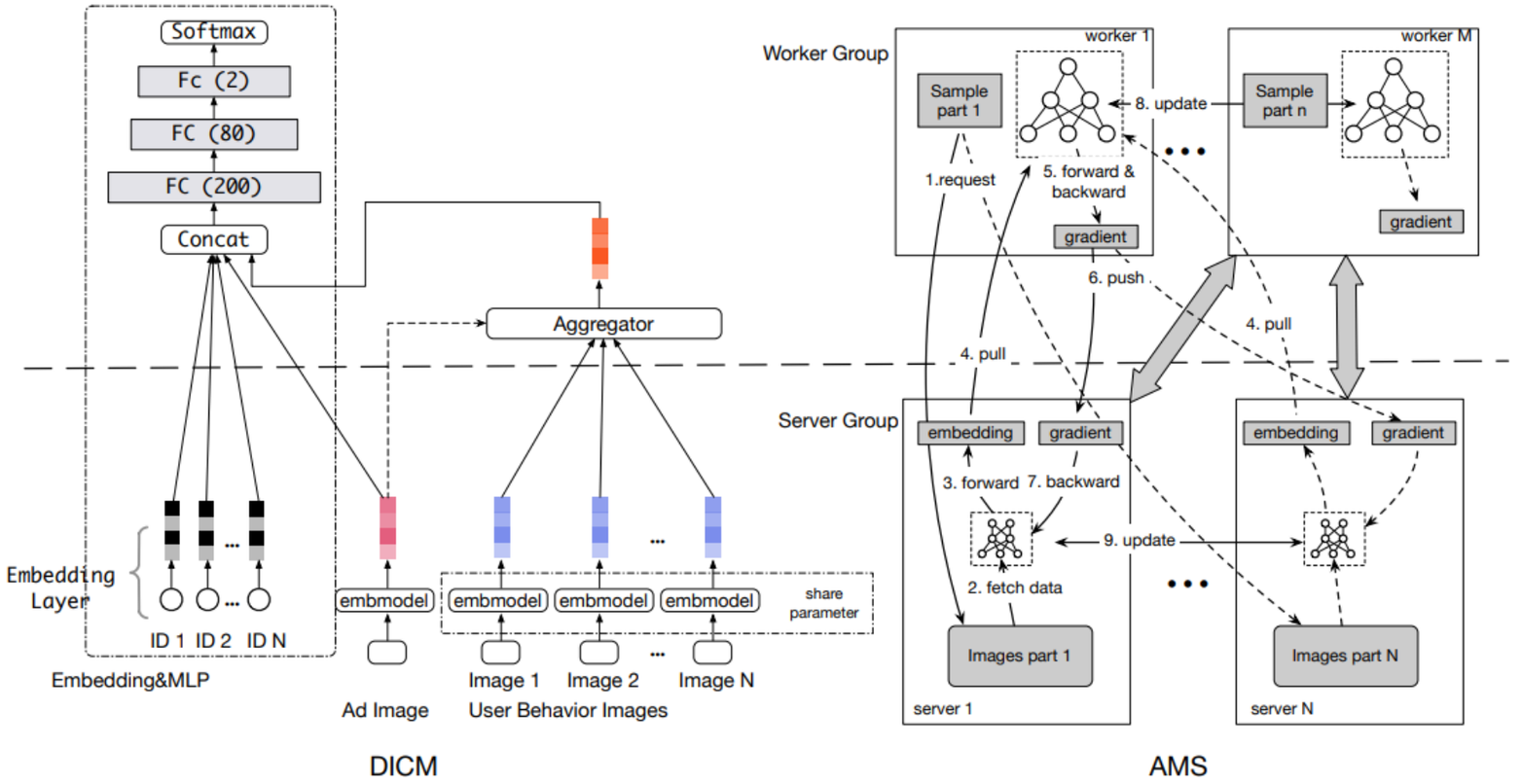
Table 3: The performance of contrast models in terms of AUC and P@2 for click and order tasks.

Multimodal Recommendation: Fusion and Finetuning Matters



DICM (Alibaba, 2018)

- Learn the image embedding model (part of the VGG16 network)



- Using a model server to learn the embedding model and other parts separately (split learning)
- Balance storage and communication costs

Image Matters: Visually modeling user behaviors using Advanced Model Server

DICM (Alibaba, 2018)

- Multiquery attentive pooling to aggregate embeddings

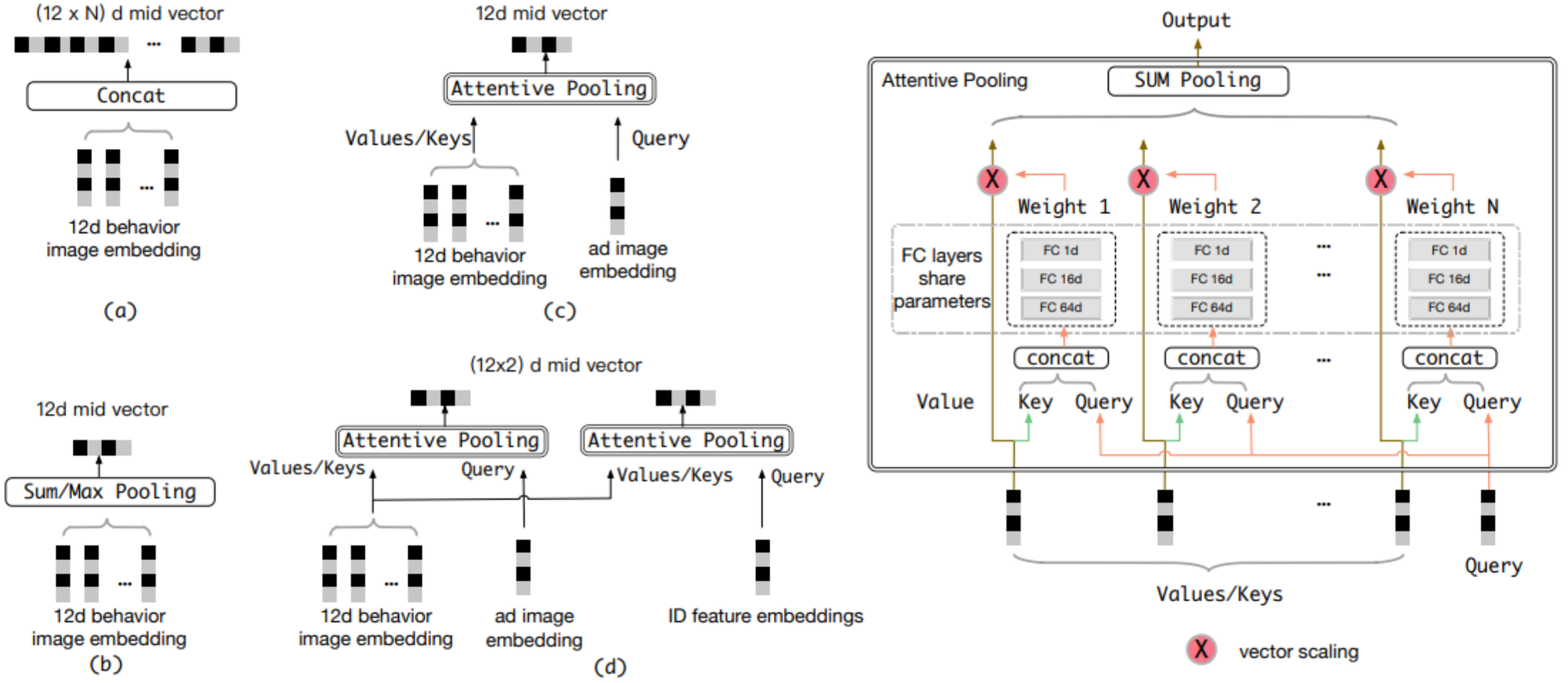


Figure 3: Aggregator architectures. (a) Concatenate (b) Sum/Max Pooling (c) Attentive Pooling (d) MultiQuery-AttentivePooling

Image Matters: Visually modeling user behaviors using Advanced Model Server

DICM (Alibaba, 2018)

- Image signals may be very strong
- Model server balances storage and communication
- Attentive pooling is helpful

Method	GAUC	GAUC gain	AUC	AUC gain
baseline	0.6205	-	0.6758	-
ad image	0.6235	0.0030	0.6772	0.0014
behavior images	0.6219	0.0014	0.6768	0.0010
joint	**0.6260**	**0.0055**	**0.6795**	**0.0037**

Table 3: Comparison of behavior images and ad image, and their combination in DICM.

Aggregator	GAUC
baseline	0.6205
Only ad images	0.6235
Concatenation	0.6232
MaxPooling	0.6236
SumPooling	0.6248
AttentivePooling	0.6257
MultiQueryAttentivePooling	**0.6260**

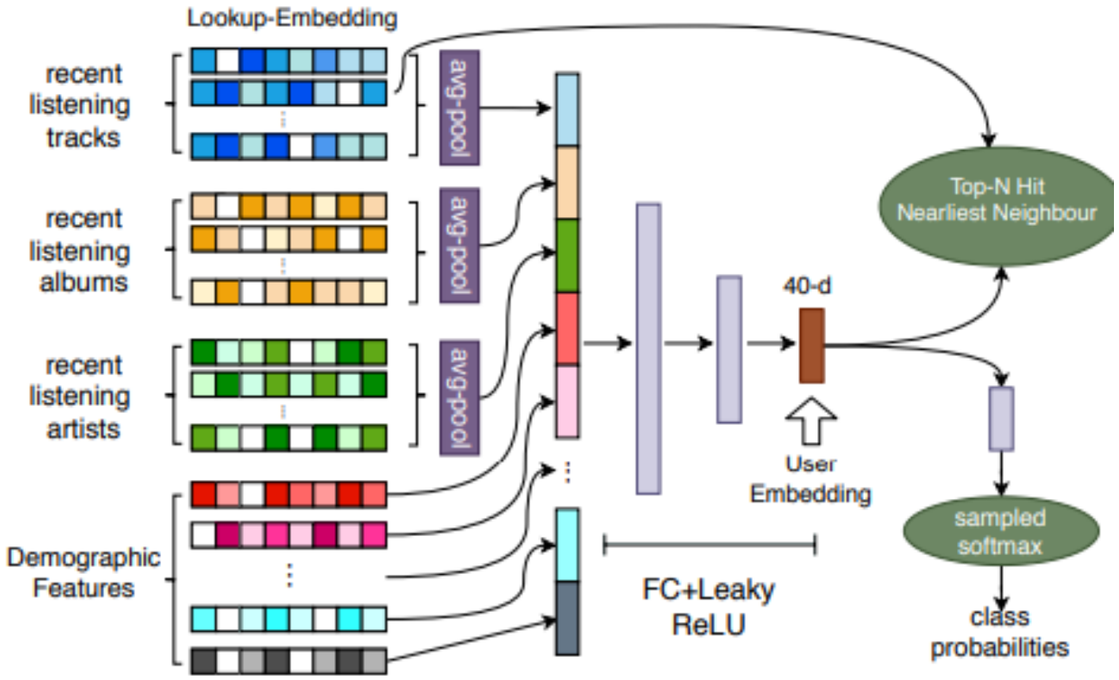
Table 4: Result of different aggregator. Aggregators are investigated jointly with ad image.

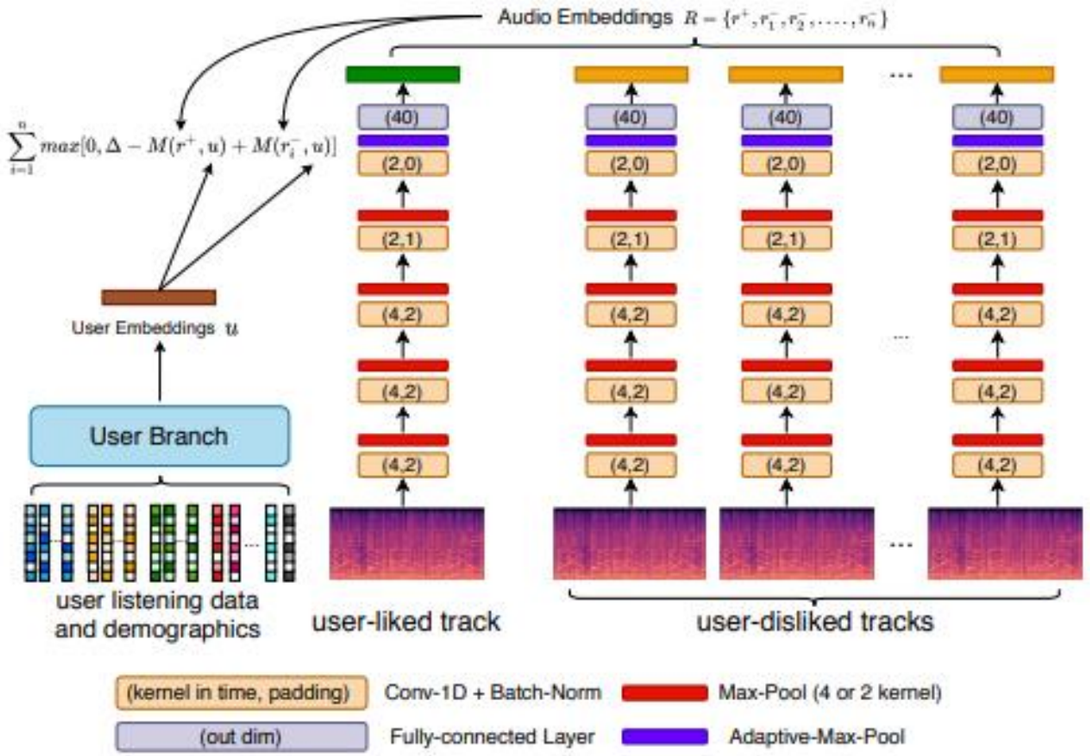
Strategy	Storage		Communication	
	Worker	Server	All	Image
store-in-worker	5.1G(332T)	0	128M	0
store-in-server	134M(8.8T)	30.3M(2T)	5.1G	5.0G
AMS	134M(8.8T)	30.3M(2T)	158M	30M

Date	CTR	eCPM	GPM
Day1	+10.0%	+5.5%	+3.3%
Day2	+10.0%	+6.8%	+8.0%
Day3	+9.1%	+6.6%	+1.8%
Day4	+9.9%	+4.8%	+7.9%
Day5	+8.2%	+5.0%	+2.7%
Day6	+8.2%	+5.4%	+9.9%
Day7	+9.0%	+5.7%	+8.0%
Average	9.2(±0.7)%	5.7(±0.7)%	5.9(±4.0)%

Siamese networks (Tencent, 2021)

- Modeling music information from music audio
  - Using DSSM-like architecture to learn user/audio embeddings





Learning Audio Embeddings with User Listening Data for Content-Based Music Recommendation

Siamese networks (Tencent, 2021)

- Longer audio context yields better performance
- More negative samples yield better AUC but lower precision

Model	Precision	AUC
Basic-Binary	0.677	0.747
DCUE-1vs1	0.623	0.675
Multi-1vs1	0.745	0.752
Multi-1vs4	0.687	0.749
Metric-1vs1	0.691	0.765
Metric-1vs4	0.681	0.778

(a) context duration equals 3 seconds.

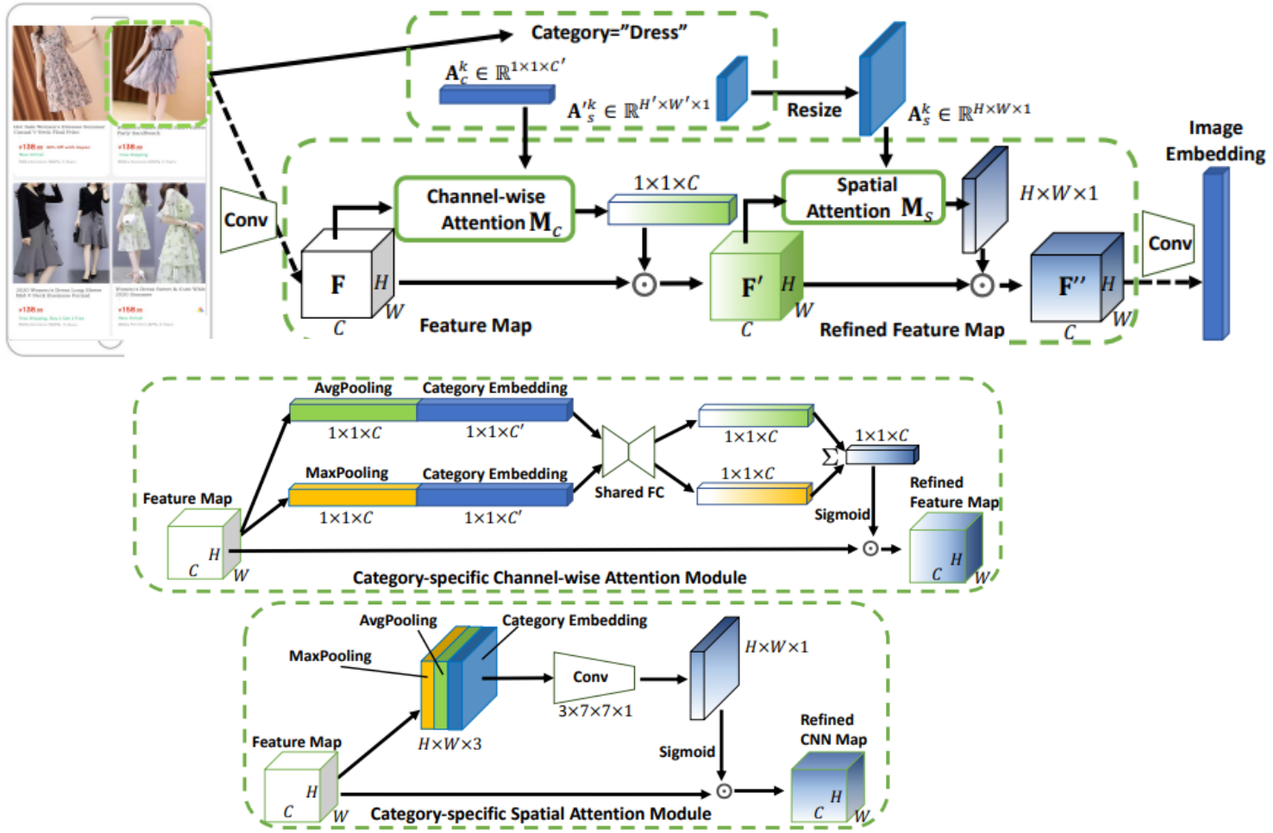
Model	Precision	AUC
Basic-Binary	0.696	0.762
DCUE-1vs1	0.644	0.697
Metric-1vs1	0.717	0.788
Metric-1vs4	0.701	0.792

(b) context duration equals 10 seconds.

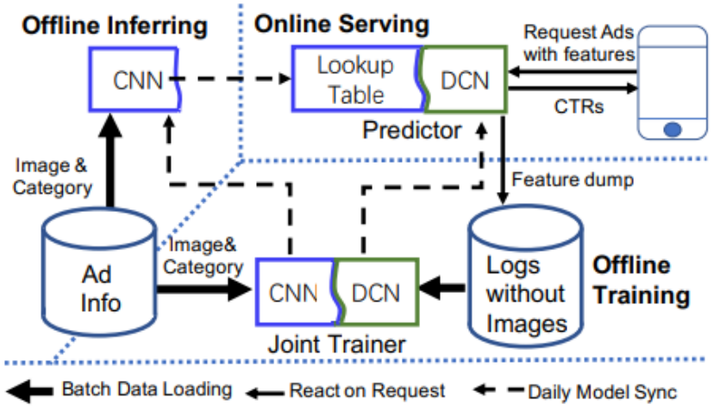
Learning Audio Embeddings with User Listening Data for Content-Based Music Recommendation

CSCNN (JD.com, 2020)

- Using category embedding to customize image embedding
  - Channel-wise attention
  - Spatial attention



- Combine both image and other features in model training
- Cache CNN features in online serving



Category-Specific CNN for Visual-aware CTR Prediction at JD.com

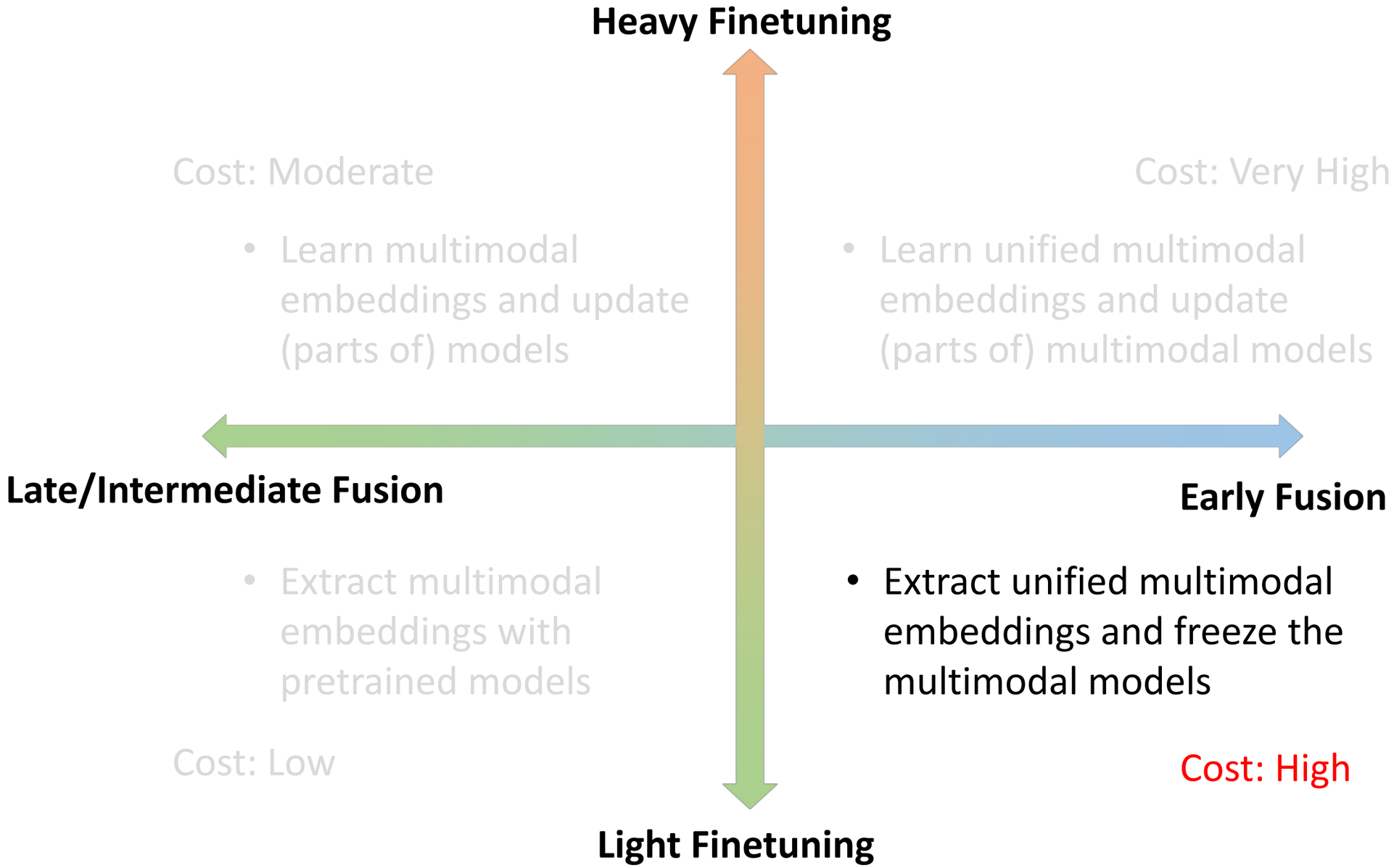
CSCNN (JD.com, 2020)

- Category information helps learn better image representations
- CSCNN can be used to enhance other vision backbones

Datasets		No Image	With Image		With Image + Category				
		BPR-MF	VBPR	DVBPR	DVBPR-C	Sherlock	DeepStyle	DVBPR-SCA	Ours
Fashion	All	0.6147	0.7557	0.8011	0.8022	0.7640	0.7530	0.8032	**0.8156**
	Cold	0.5334	0.7476	0.7712	0.7703	0.7427	0.7465	0.7694	**0.7882**
Women	All	0.6506	0.7238	0.7624	0.7645	0.7265	0.7232	0.7772	**0.7931**
	Cold	0.5198	0.7086	0.7078	0.7099	0.6945	0.7120	0.7273	**0.7523**
Men	All	0.6321	0.7079	0.7491	0.7549	0.7239	0.7279	0.7547	**0.7749**
	Cold	0.5331	0.6880	0.6985	0.7018	0.6910	0.7210	0.7048	**0.7315**

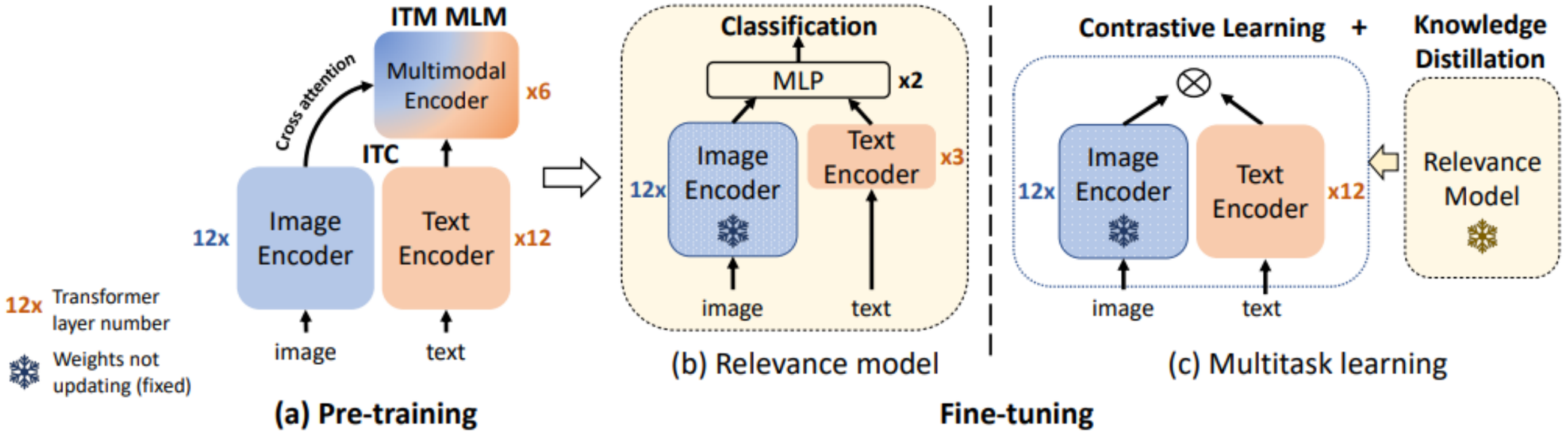
	Original		+CSCNN	
	All	Cold	All	Cold
No Attention	0.7491	0.6985	–	–
SE	0.7500	0.6989	**0.7673**	**0.7153**
CBAM-Channel	0.7506	0.7002	**0.7683**	**0.7184**
CBAM-All	0.7556	0.7075	**0.7749**	**0.7315**

Multimodal Recommendation: Fusion and Finetuning Matters



Adapted CLIP (Baidu, 2023)

- A VLM-based framework for Ad recall
  - Pretraining CLIP on the vision MLM task
  - Finetune the relevance model on high-quality Ad domain data
  - Distill knowledge from the relevance model on the full data



Enhancing Dynamic Image Advertising with Vision-Language Pre-training

Adapted CLIP (Baidu, 2023)

- The adapted CLIP model outperforms the base model on both general and industrial datasets
- Knowledge distillation helps model learning on large-scale noisy data

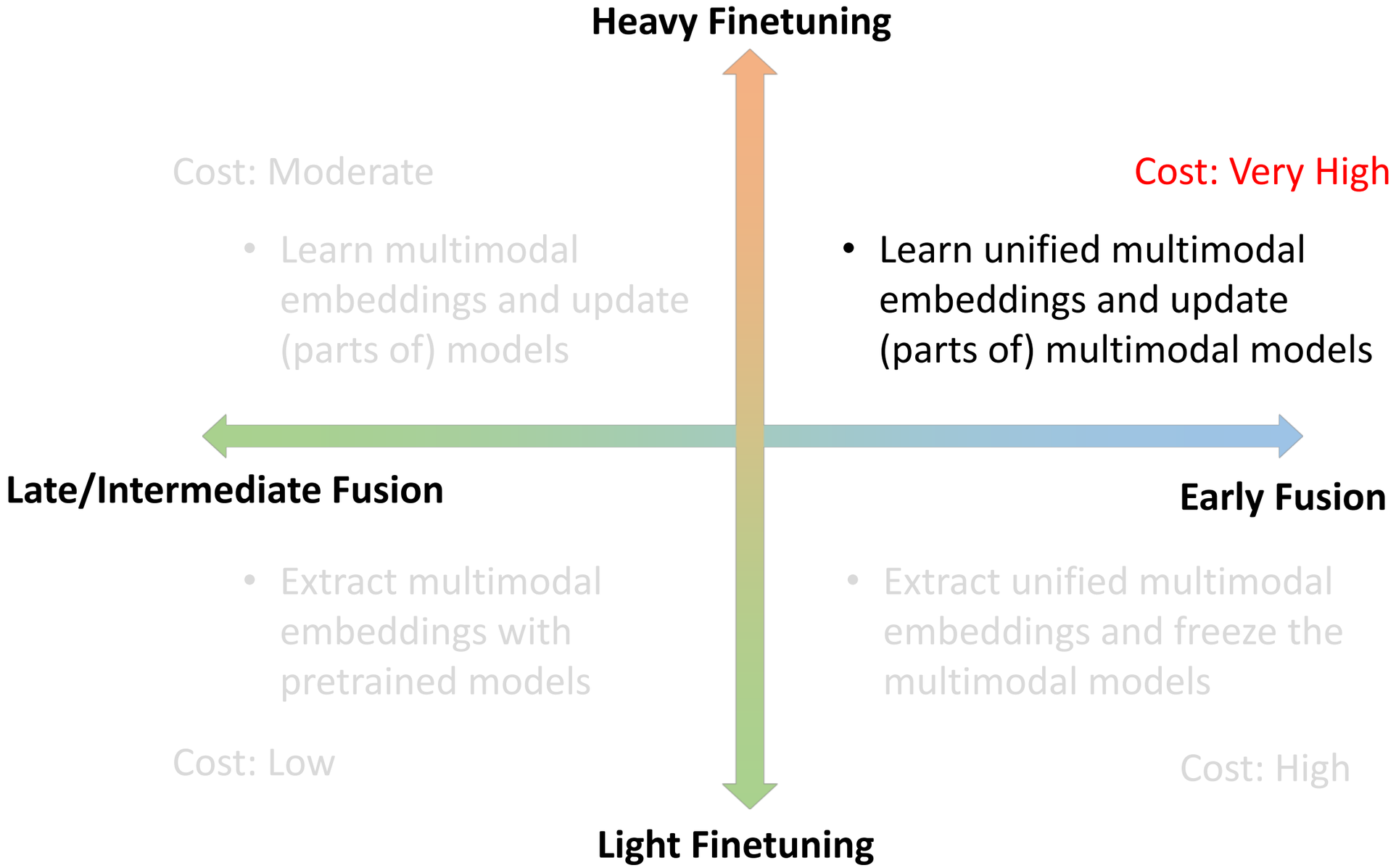
Method	Wukong			MSCOCO-CN			Flickr30-CN		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CN-CLIP	45.6	72.4	79.8	**62.2**	**86.6**	**94.9**	**62.7**	**86.9**	**92.8**
ours$_{base}$	**56.3**	**82.9**	**88.0**	51.0	80.8	91.2	45.8	75.2	84.2
Method	search			advertising			e-commerce		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CN-CLIP	23.6	48.9	59.0	8.0	22.1	30.6	**58.9**	79.7	84.8
ours$_{base}$	**36.1**	**67.4**	**76.7**	**9.8**	**27.0**	**37.4**	58.0	**80.0**	**85.5**

Method	Diversity Ratio	Irrelevant Ratio
previous$_{retrieval}$	6.11	32.67
ours$_{base}$	**9.40**	**21.78**

Method	Recall@10	Relscore@10
base	75.3	78.5
retrieval w/o KD	92.8	77.7
retrieval w/ KD	**94.1**	**79.5**

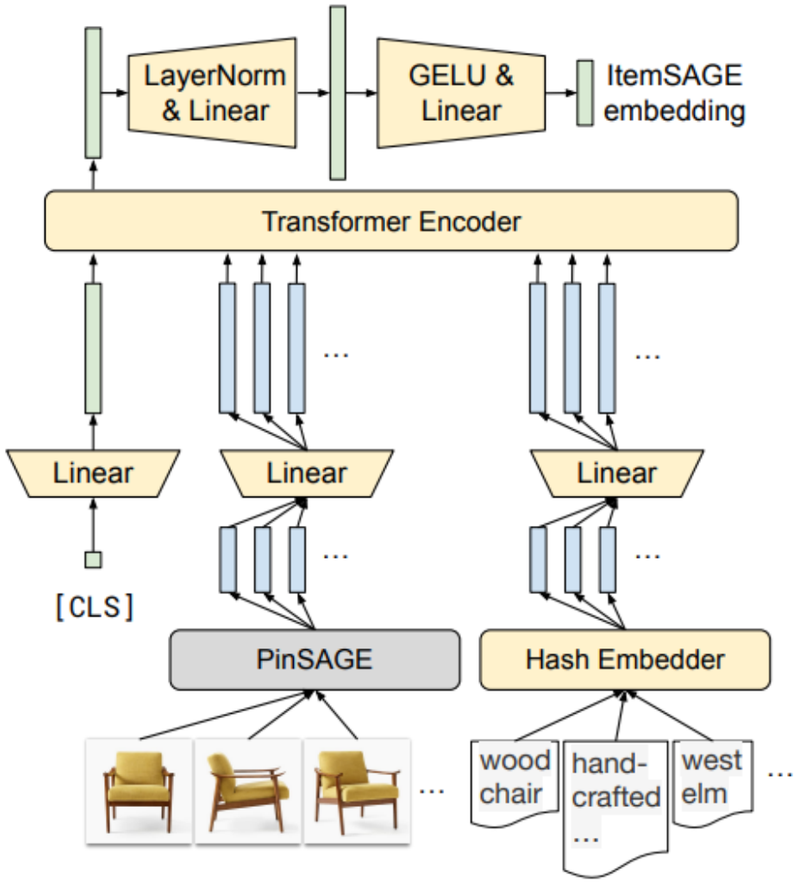
Enhancing Dynamic Image Advertising with Vision-Language Pre-training

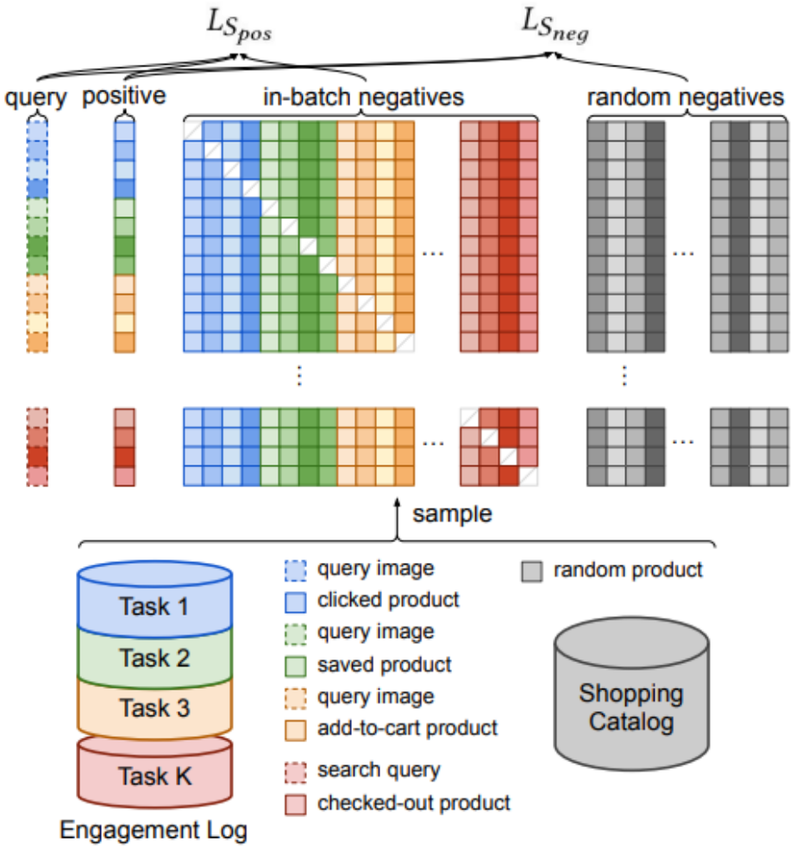
Multimodal Recommendation: Fusion and Finetuning Matters



ItemSage (Pinterest, 2022)

- Text-to-image retrieval in multiple tasks





ItemSage: Learning Product Embeddings for Shopping Recommendations at Pinterest.

ItemSage (Pinterest, 2022)

- ItemSage achieves better results in various tasks
- Deep model may not help
- Features and task signals are critical

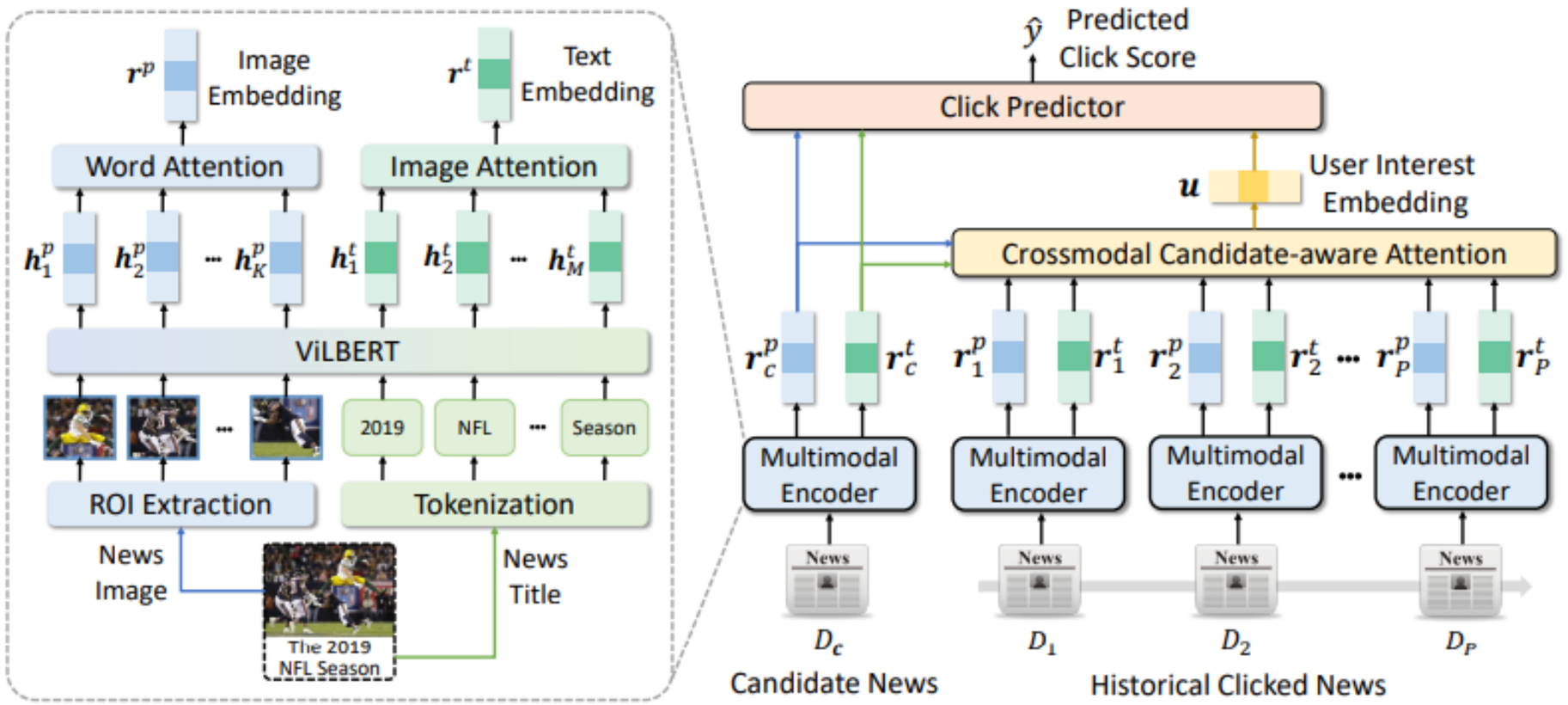
	Number of Parameters	Closeup Clicks	Closeup Saves	Closeup Add-to-Cart	Closeup Checkouts	Search Clicks	Search Saves	Search Add-to-Cart	Search Checkouts
Sum	-	0.663	0.647	0.669	0.699	-	-	-	-
Sum-MLP	-	-	-	-	-	0.577	0.533	0.561	0.629
MLP-Concat-MLP	30.8M	0.805	0.794	0.896	**0.916**	0.723	0.736	0.834	0.861
ItemSage	33.1M	**0.816**	0.812	**0.897**	**0.916**	0.749	0.762	**0.842**	**0.869**
2-Layer Transformer	36.3M	0.815	0.809	0.895	0.913	0.745	0.759	0.837	0.867
3-Layer Transformer	39.4M	0.815	0.810	0.896	0.915	0.747	0.758	0.841	**0.869**
4-Layer Transformer	42.6M	**0.816**	**0.813**	**0.897**	0.915	**0.750**	**0.764**	0.840	**0.869**

		Closeup Clicks	Closeup Saves	Closeup Add Cart	Closeup Checkouts	Search Clicks	Search Saves	Search Add Cart	Search Checkouts
	ItemSage	0.816	0.812	0.897	0.916	0.749	0.762	0.842	0.869
Feature	Image Only	0.795 (-2.6%)	0.787 (-3.1%)	0.882 (-1.7%)	0.908 (-0.9%)	0.670 (-10.5%)	0.698 (-8.4%)	0.798 (-5.2%)	0.830 (-4.5%)
	Text Only	0.683 (-16.3%)	0.658 (-19.0%)	0.832 (-7.2%)	0.859 (-6.2%)	0.669 (-10.7%)	0.665 (-12.7%)	0.790 (-6.2%)	0.820 (-5.6%)
	Image + Text + Graph	0.814 (-0.2%)	0.812 (0.0%)	0.893 (-0.4%)	0.905 (-1.2%)	0.743 (-0.8%)	0.767 (0.7%)	0.842 (0.0%)	0.860 (-1.0%)
Negative Sampling	$L_{S_{pos}}$ Only	0.597 (−26.8%)	0.602 (-25.9%)	0.717 (-20.1%)	0.772 (-15.7%)	0.553 (-26.2%)	0.544 (-28.6%)	0.662 (-21.4%)	0.724 (-16.7%)
	$L_{S_{neg}}$ Only	0.774 (-5.1%)	0.768 (-5.2%)	0.868 (-3.2%)	0.897 (-2.1%)	0.655 (-12.6%)	0.670 (-12.1%)	0.804 (-4.5%)	0.840 (-3.3%)
	$L_{S_{mixed}}$	0.781 (-4.3%)	0.774 (-4.7%)	0.860 (-4.1%)	0.884 (-3.5%)	0.687 (-8.3%)	0.706 (-7.3%)	0.809 (-3.9%)	0.838 (-3.6%)
Surface	Closeup	0.815 (-0.1%)	0.811 (-0.1%)	0.891 (-0.7%)	0.909 (-0.8%)	- -	- -	- -	- -
	Search	- -	- -	- -	- -	0.760 (1.5%)	0.766 (0.5%)	0.830 (-1.4%)	0.861 (-0.9%)
Engagement Type	Clicks + Saves	0.819 (0.4%)	0.812 (0.0%)	0.869 (-3.1%)	0.894 (-2.4%)	0.755 (0.8%)	0.768 (0.8%)	0.689 (-18.2%)	0.765 (-12.0%)
	Add Cart + Checkouts	0.503 (-38.4%)	0.503 (-38.1%)	0.850 (-5.2%)	0.882 (-3.7%)	0.382 (-49.0%)	0.392 (-48.6%)	0.768 (-8.8%)	0.793 (-8.7%)

ItemSage: Learning Product Embeddings for Shopping Recommendations at Pinterest.

MM-Rec (Microsoft, 2022)

- Finetuning VLM during model training

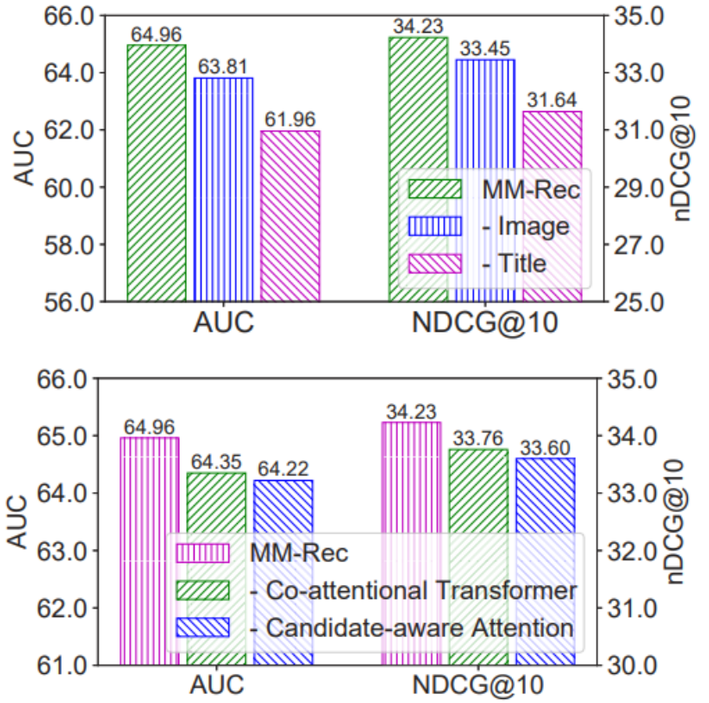


MM-Rec: Visiolinguistic Model Empowered Multimodal News Recommendation

MM-Rec (Microsoft, 2022)

- Cross-modal matching is useful for recommendation

Methods	AUC	MRR	NDCG@5	NDCG@10
EBNR	60.34±0.29	20.79±0.25	22.43±0.26	30.76±0.23
DKN	60.18±0.24	20.56±0.22	22.24±0.20	30.53±0.18
DAN	61.03±0.22	21.69±0.19	23.12±0.23	31.48±0.20
NAML	61.55±0.18	22.13±0.16	23.57±0.17	31.92±0.17
NRMS	62.01±0.13	22.68±0.15	24.08±0.15	32.38±0.15
GERL	62.21±0.17	22.82±0.16	24.36±0.18	32.55±0.19
FIM	62.18±0.15	22.79±0.14	24.35±0.13	32.52±0.16
PLM-NR	63.67±0.10	24.17±0.09	25.42±0.11	33.31±0.12
MM-Rec	**64.96**±0.12	**25.22**±0.11	**26.67**±0.12	**34.23**±0.10



MM-Rec: Visiolinguistic Model Empowered Multimodal News Recommendation

Open Challenges: 5A

- **A**lignment
- **A**ggregation
- **A**daptation
- **A**cceleration
- **A**tmosphere

Alignment: More Modalities, More Information

- Recommender systems need to process more and more modalities
  - Text, audio, image, video, signal, tabular data…
  - How to align so many modalities?

- How to align so many modalities?
  - Which one should be the center?

- How to align new modalities to existing ones?
  - Motivated by GPT-4, GPT-4V, DALLE, and Sora

Aggregation: Multimodal Information Fusion

- Recommender systems need to fuse representations of different modalities
  - Different modalities have commonality and diversity

- Early fusion is difficult and expensive
  - Needs low-level understanding of different modalities

- Late fusion may be ineffective
  - Many useful signals are lost during representation learning

Adaptation: Foundation Model to Recommenders

- Multimodal foundation models are not born as multimodal recommender systems
  - Need to adapt pretrained models in recommendation tasks

- Pretrained models are often general domain oriented
  - May be suboptimal in specific domains

- How to develop good tasks and data to adapt foundation models to novel tasks and domains?

Acceleration: Bigger and Faster

- Multimodal models usually have high computational & memory costs
  - Large foundation models are much more slower than traditional recommendation models

- How to accelerate multimodal models to meet latency requirements?
  - Distillation/Quantization/Compression
  - Cache
  - Speculative Decoding
  - MoE
  - …

Atmosphere: How to Embrace AIGC?

- The influence of AIGC on content delivery platforms
  - Users can use AI tools to create better content in a shorter time
  - May pollute the content ecosystem and affect the quality of UGC
- The influence of AIGC on recommendation algorithms
  - Difficult to balance the exposure chance of AIGC and UGC
  - Some models may prefer AIGC than UGC[1]
- The influence of AIGC on users
  - May distort users' perception and views (this can be intentional)

[1]Llms may dominate information access: Neural retrievers are biased towards llm-generated texts." arXiv preprint arXiv:2310.20501 (2023).