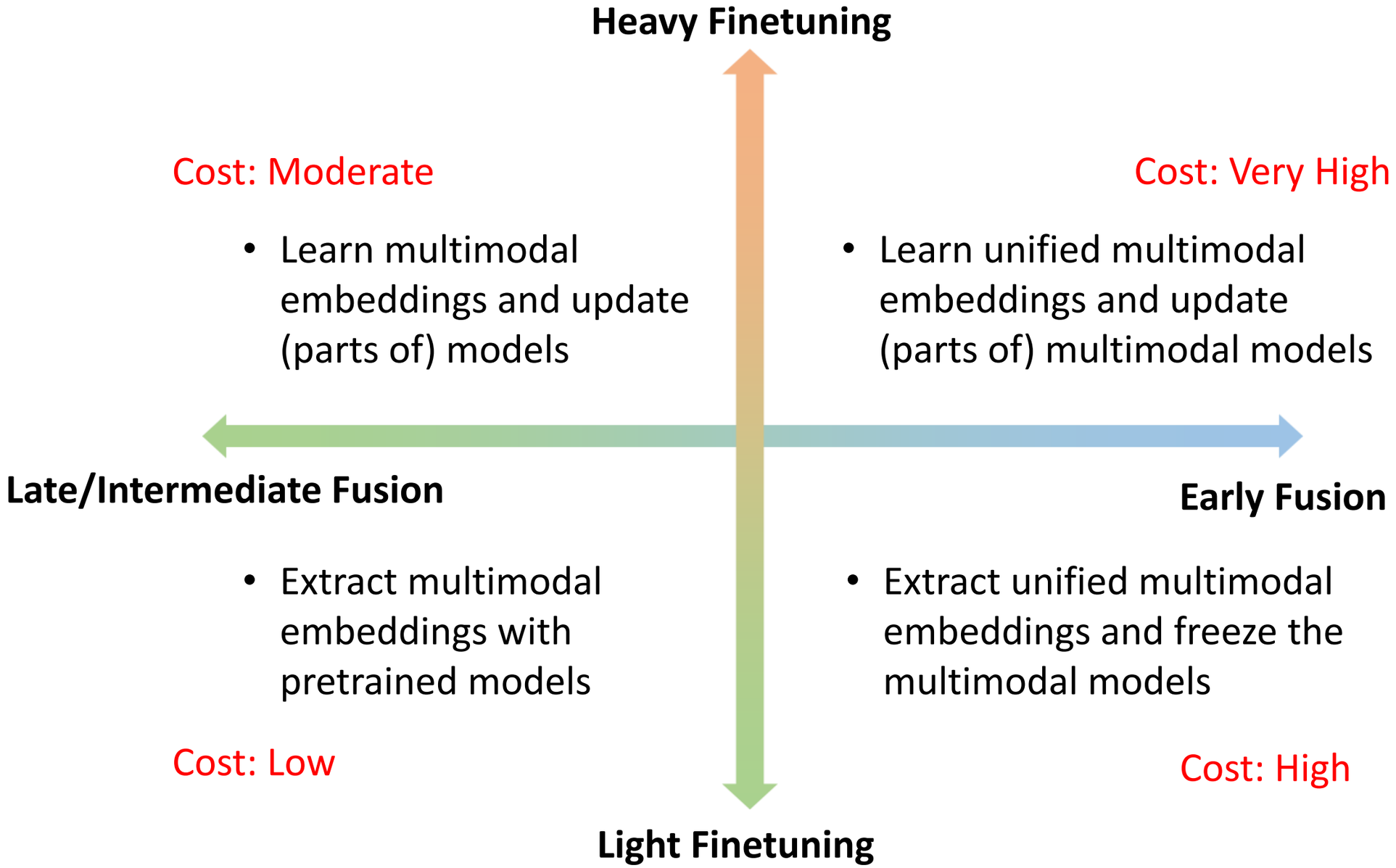
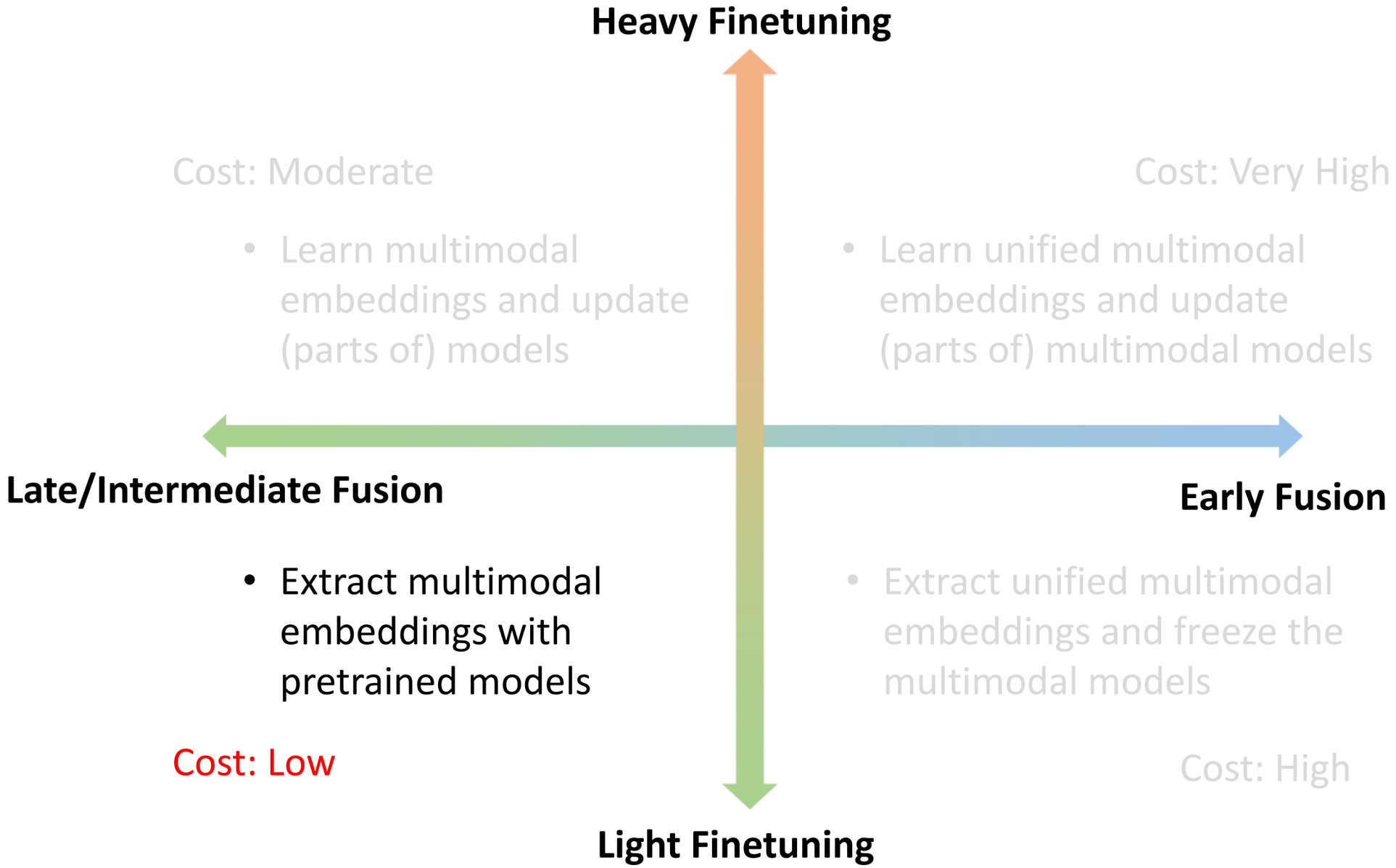
Industrial Applications and Open Challenges in Multimodal Recommendation

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

Multimodal Recommendation: Fusion and Finetuning Matters

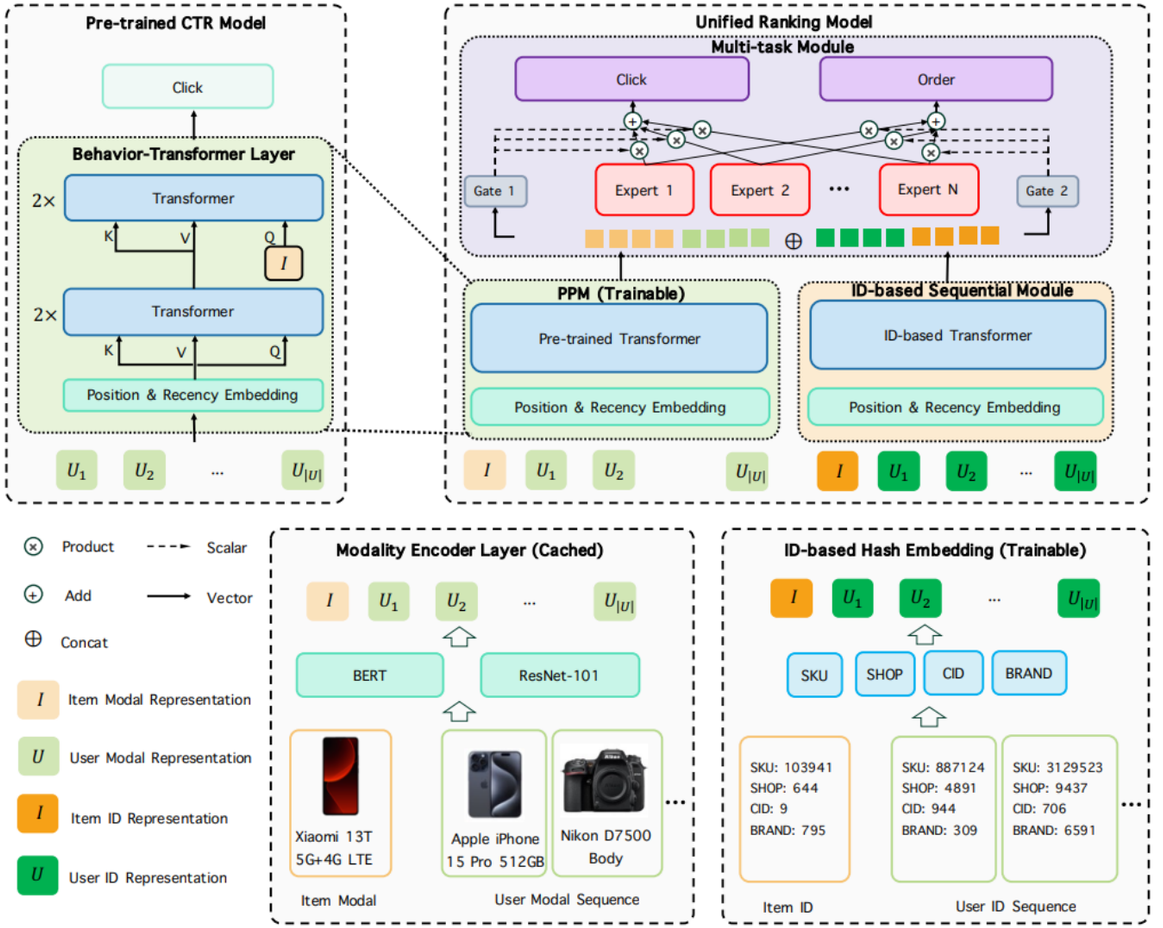


Multimodal Recommendation: Fusion and Finetuning Matters



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

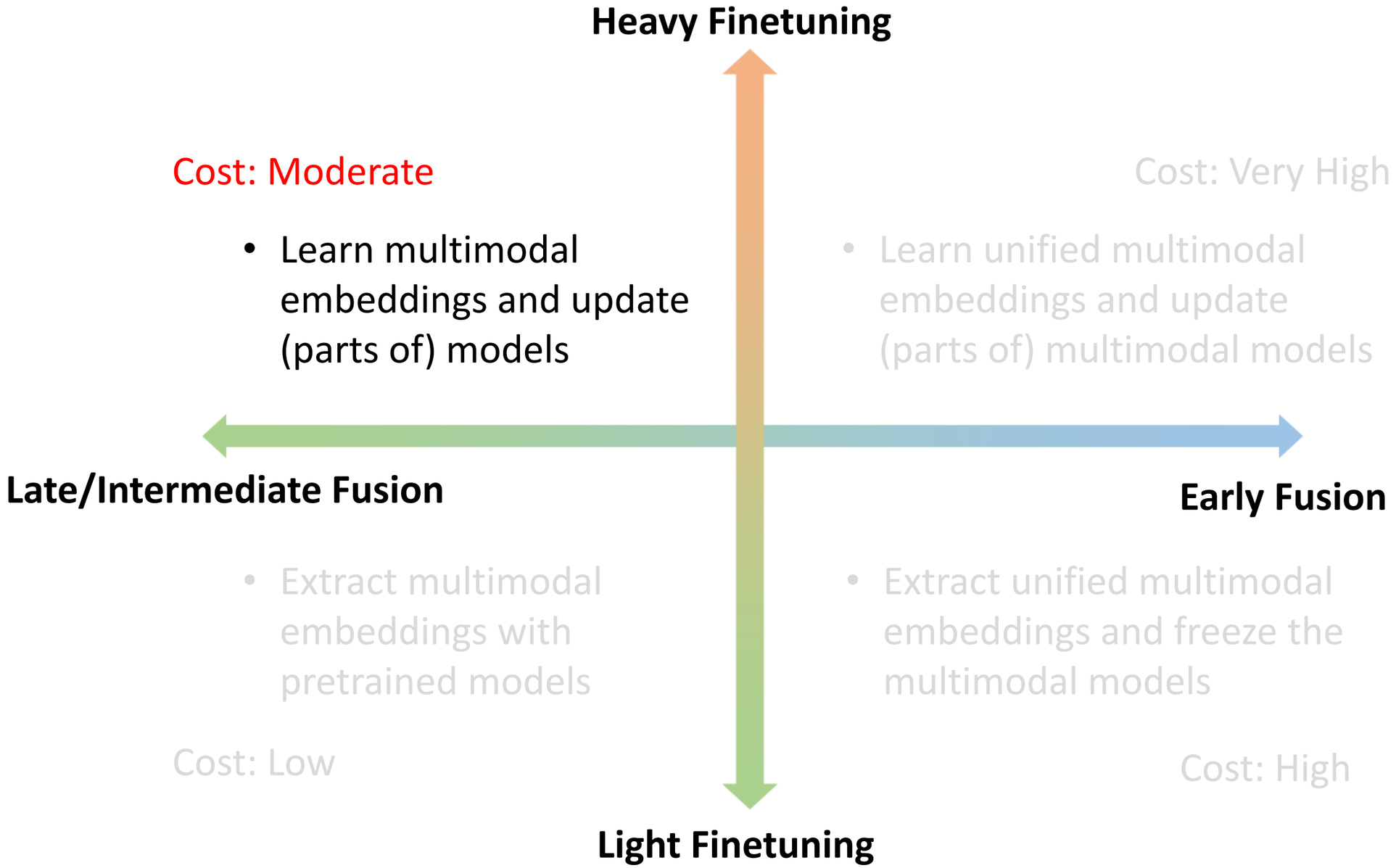
Dataset	Model	wo.PPM						w.PPM					
		Click		Order		Average		Click		Order		Average	
		AUC	P@2	AUC	P@2	$\overline{AUC}$	$\overline{P@2}$	AUC	P@2	AUC	P@2	$\overline{AUC}$(Improv)	$\overline{P@2}$(Improv)
Small	Wide&Deep	0.6094	0.1886	0.6030	0.1655	0.6062	0.1770	0.6797	0.2029	0.6948	0.1685	**0.6873**(0.0811)	**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)

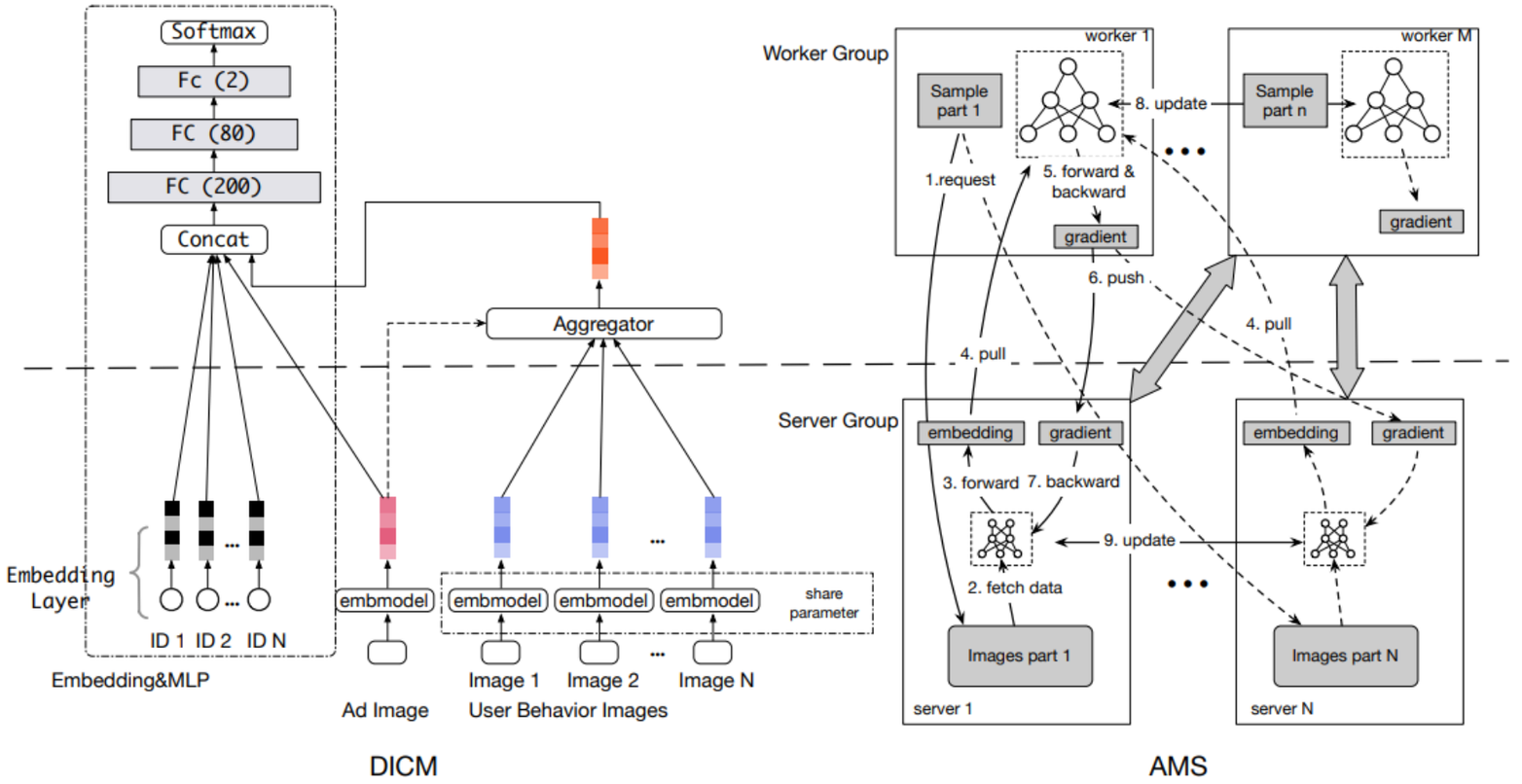
Table 3: The performance of contrast models in terms of AUC and P@2 for click and order tasks.

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

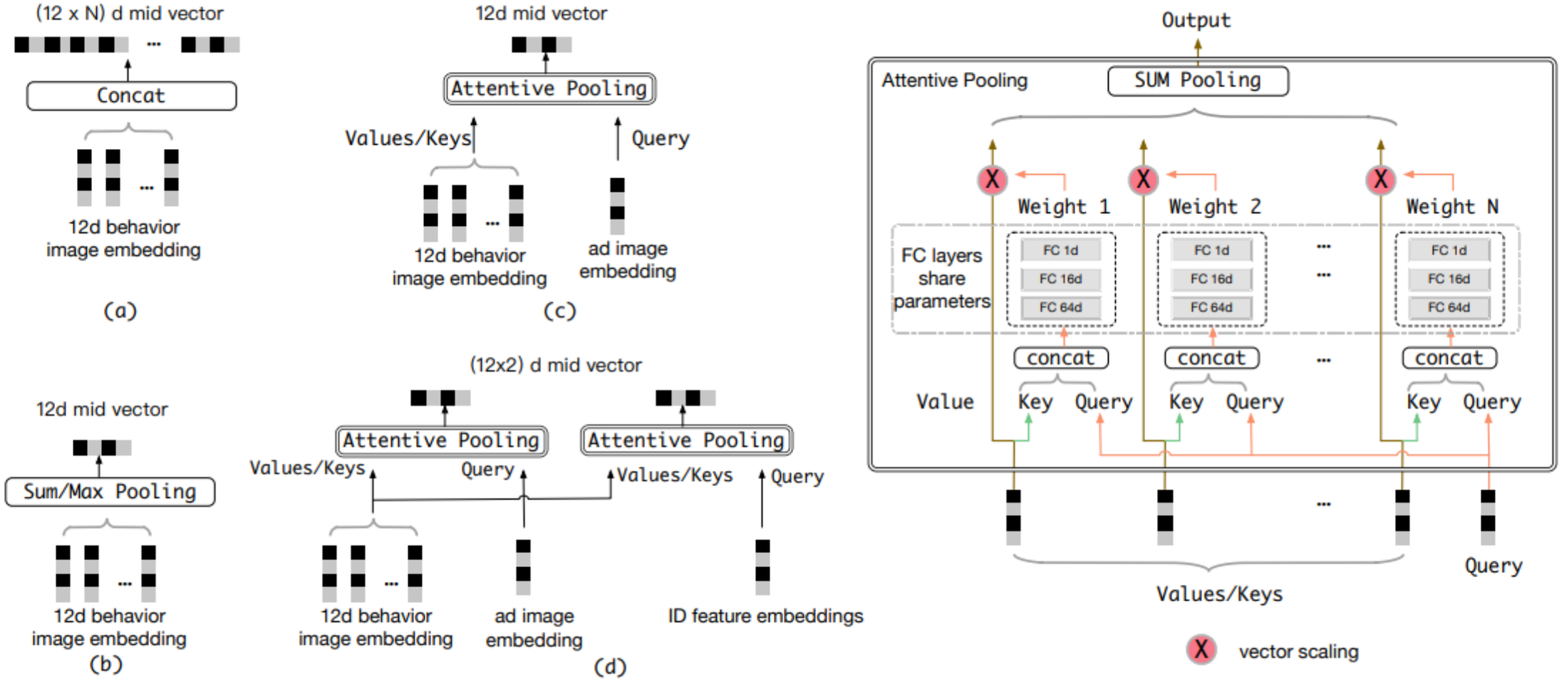


Figure 3: Aggregator architectures. (a) Concatenate (b) Sum/Max Pooling (c) Attentive Pooling (d) MultiQuery-AttentivePooling

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**

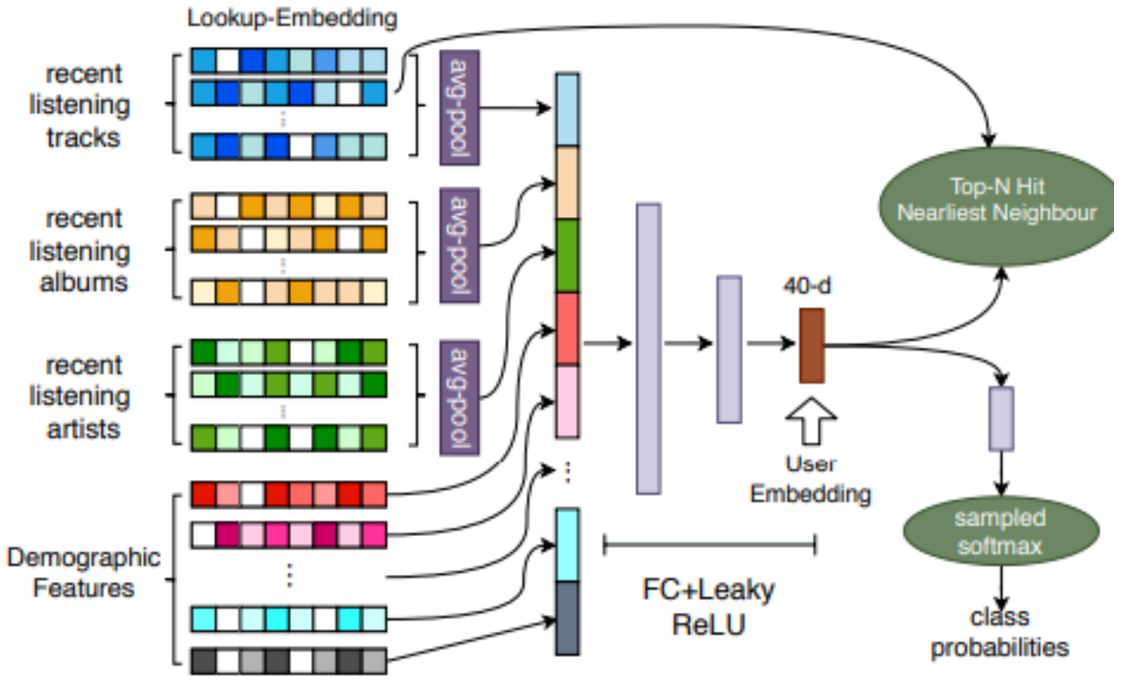
Table 4: Result of different aggregator. Aggregators are investigated jointly with ad image.

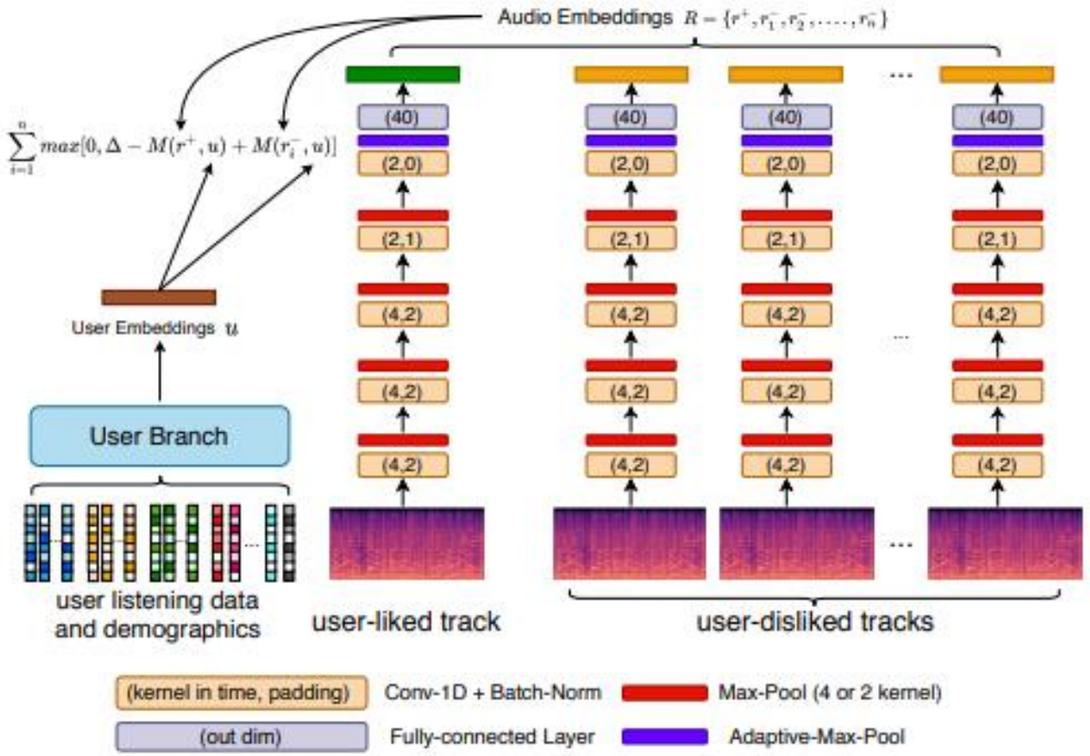
Strategy	Storage		Communication	
	Worker	Server	All	Image
store-in-worker	5.1G(332T)	0	128M	0
store-in-server	134M(8.8T)	30.3M(2T)	5.1G	5.0G
AMS	134M(8.8T)	30.3M(2T)	158M	30M

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.

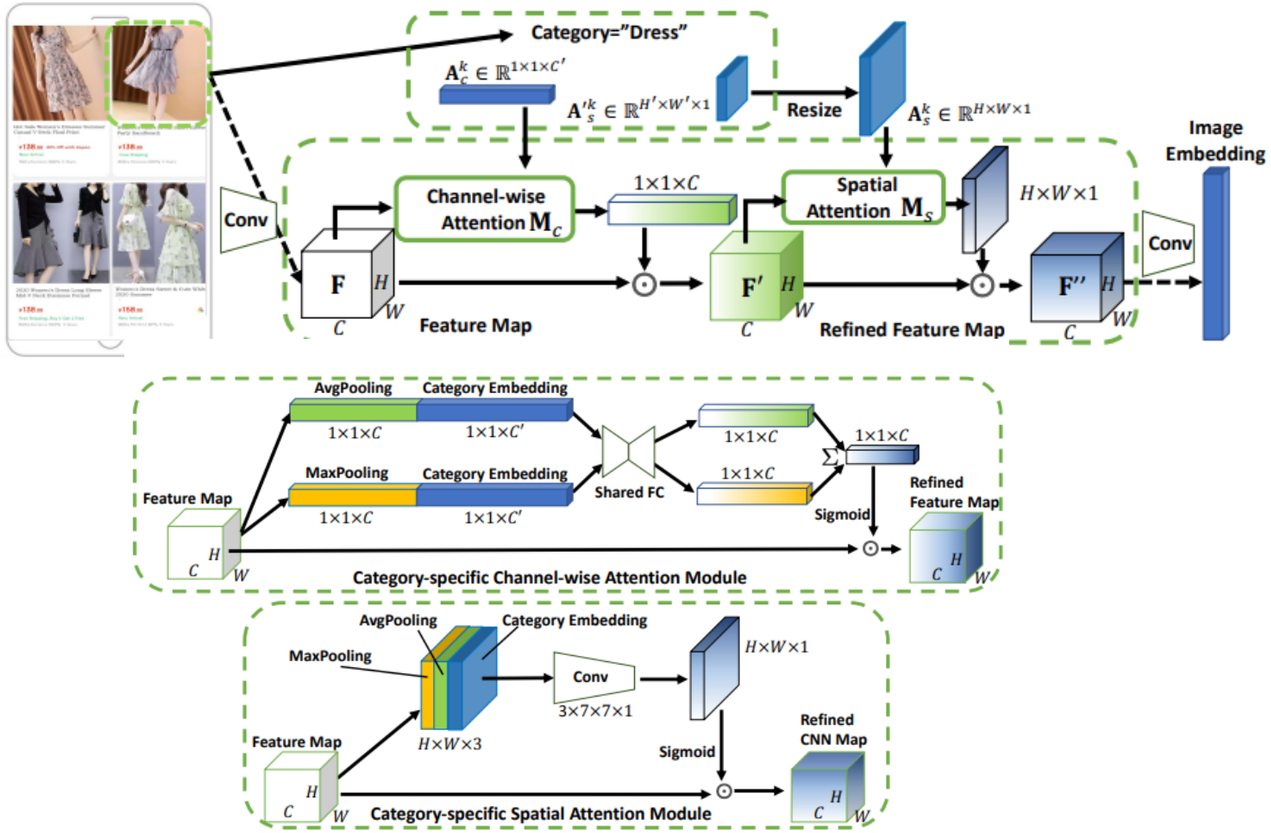
Model	Precision	AUC
Basic-Binary	0.696	0.762
DCUE-1vs1	0.644	0.697
Metric-1vs1	0.717	0.788
Metric-1vs4	0.701	0.792

(b) context duration equals 10 seconds.

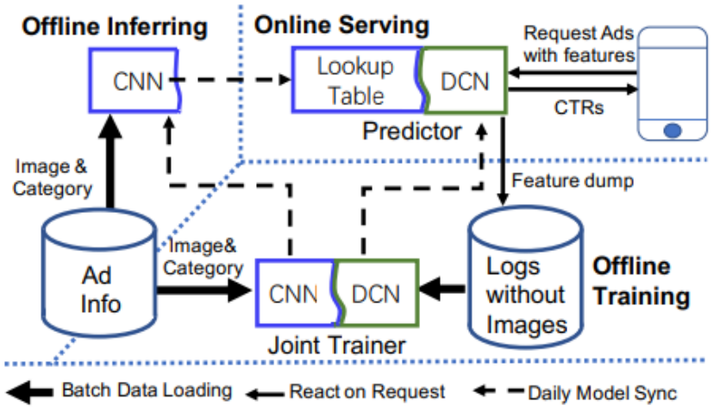
Learning Audio Embeddings with User Listening Data for Content-Based Music Recommendation

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

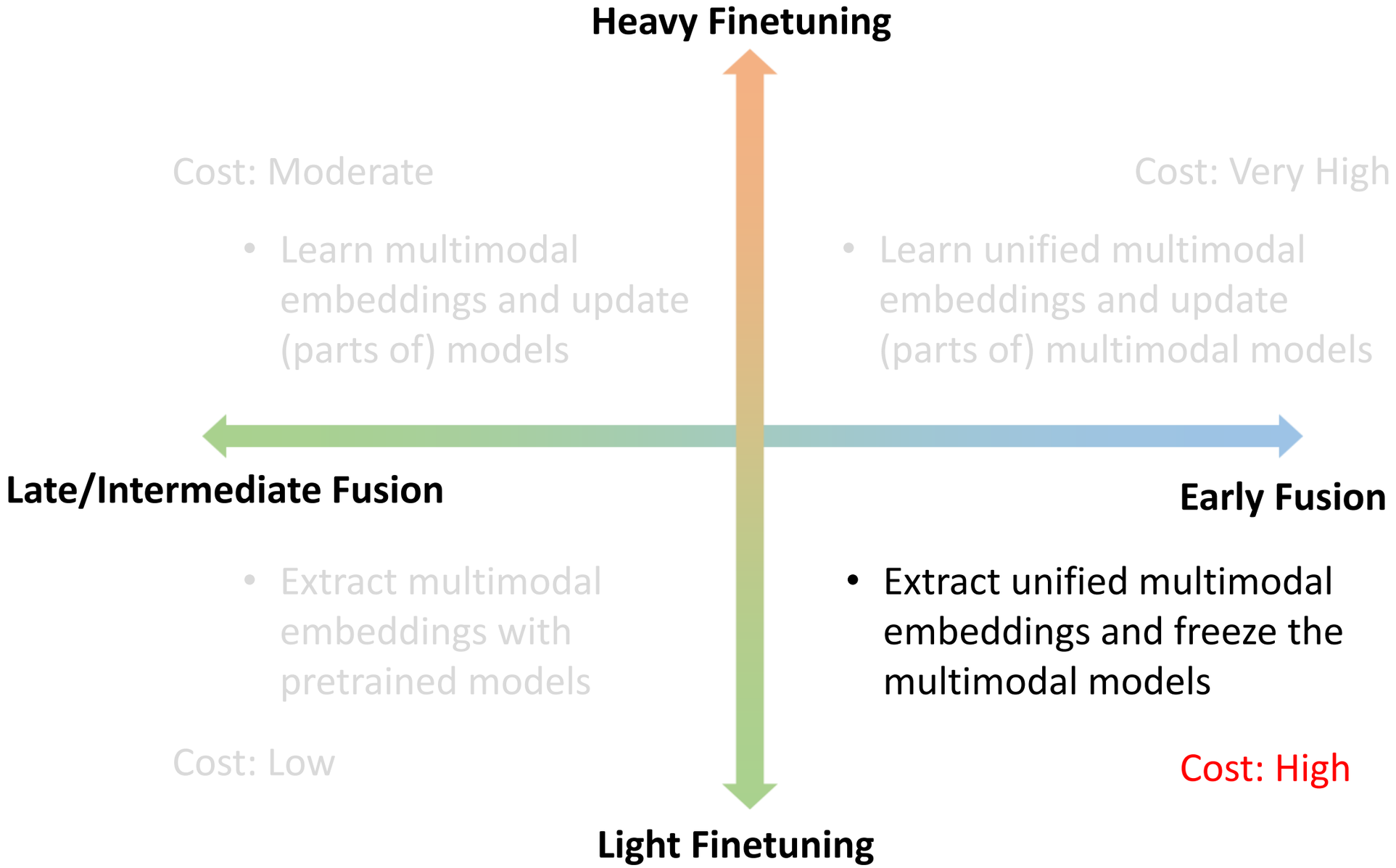
CSCNN (JD.com, 2020)

- 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**

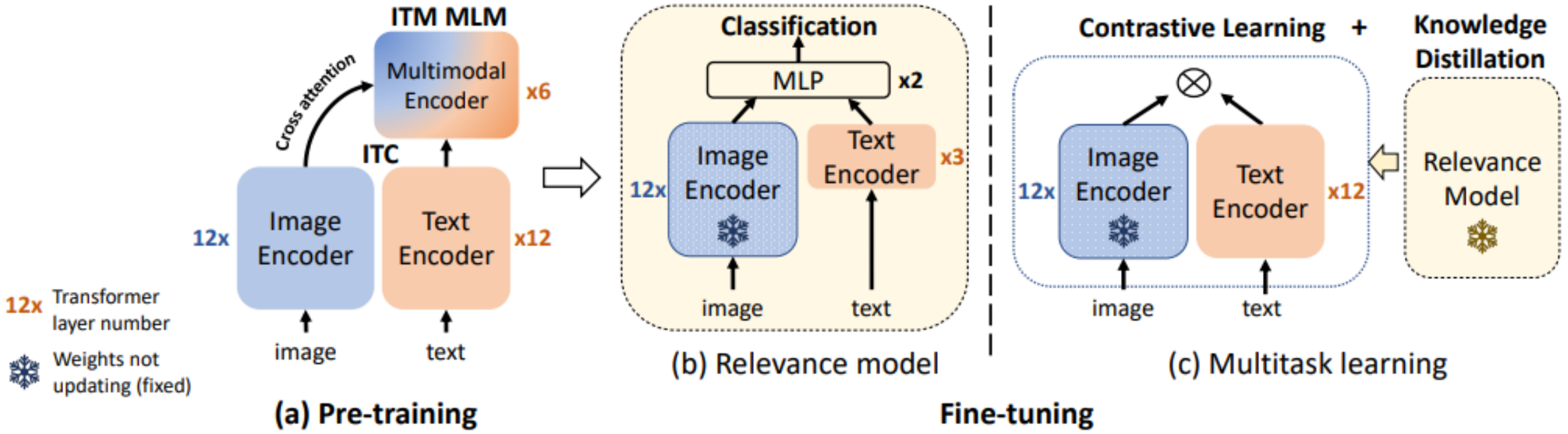
	Original		+CSCNN	
	All	Cold	All	Cold
No Attention	0.7491	0.6985	–	–
SE	0.7500	0.6989	**0.7673**	**0.7153**
CBAM-Channel	0.7506	0.7002	**0.7683**	**0.7184**
CBAM-All	0.7556	0.7075	**0.7749**	**0.7315**

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

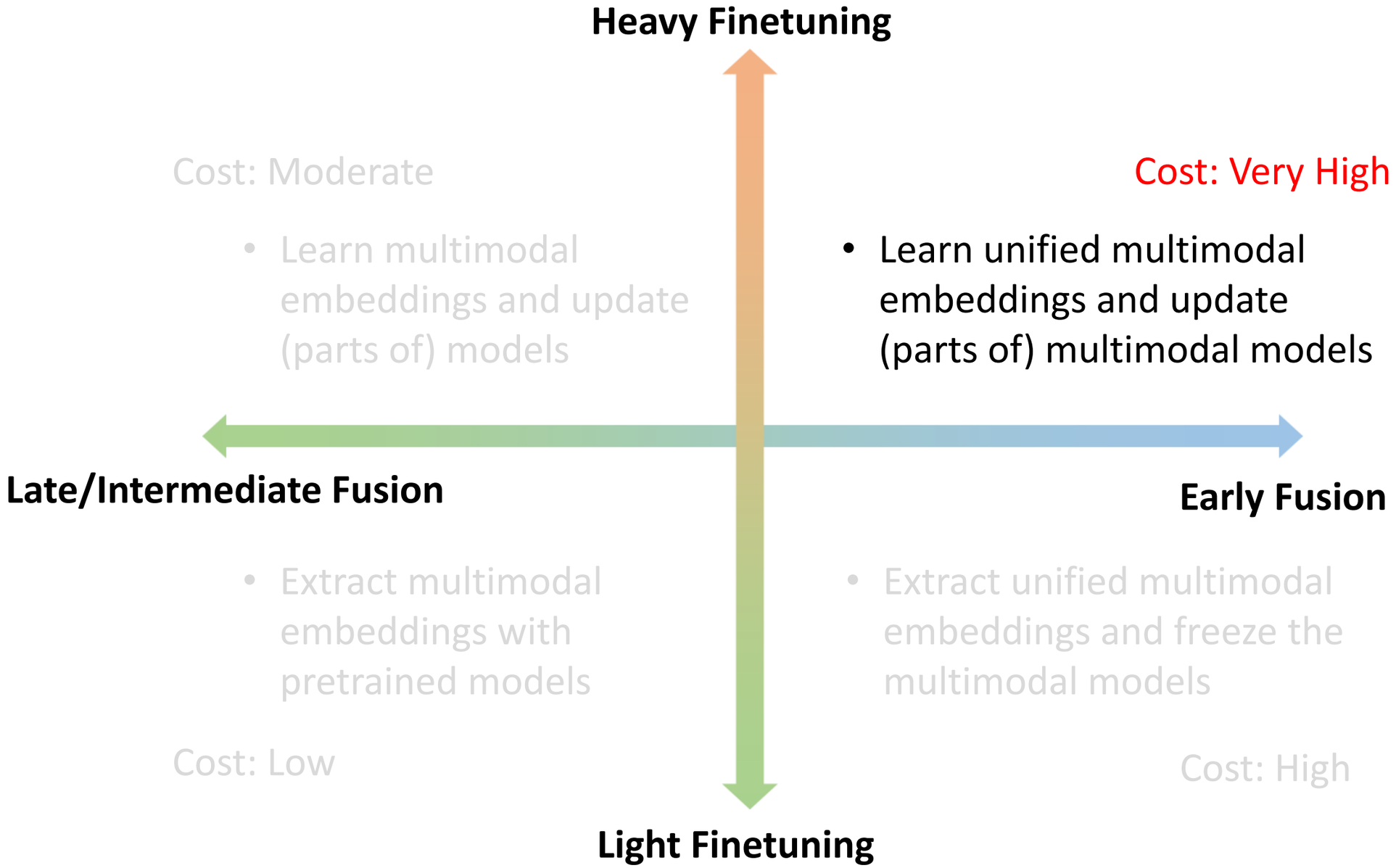
Method	Wukong			MSCOCO-CN			Flickr30-CN		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CN-CLIP	45.6	72.4	79.8	**62.2**	**86.6**	**94.9**	**62.7**	**86.9**	**92.8**
ours$_{base}$	**56.3**	**82.9**	**88.0**	51.0	80.8	91.2	45.8	75.2	84.2
Method	search			advertising			e-commerce		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	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**

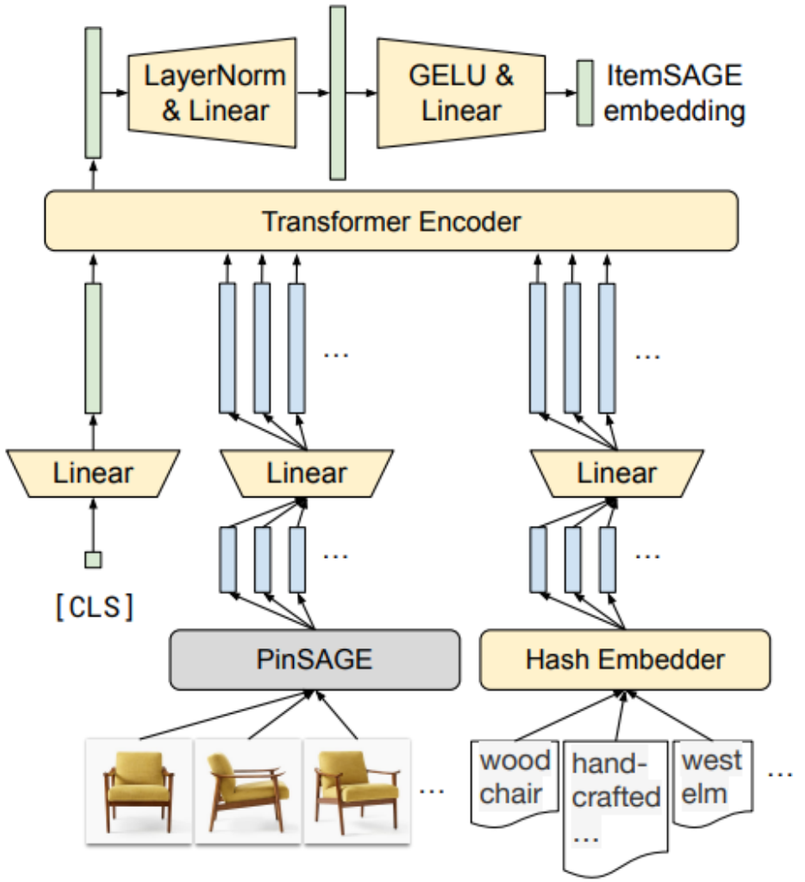
Enhancing Dynamic Image Advertising with Vision-Language Pre-training

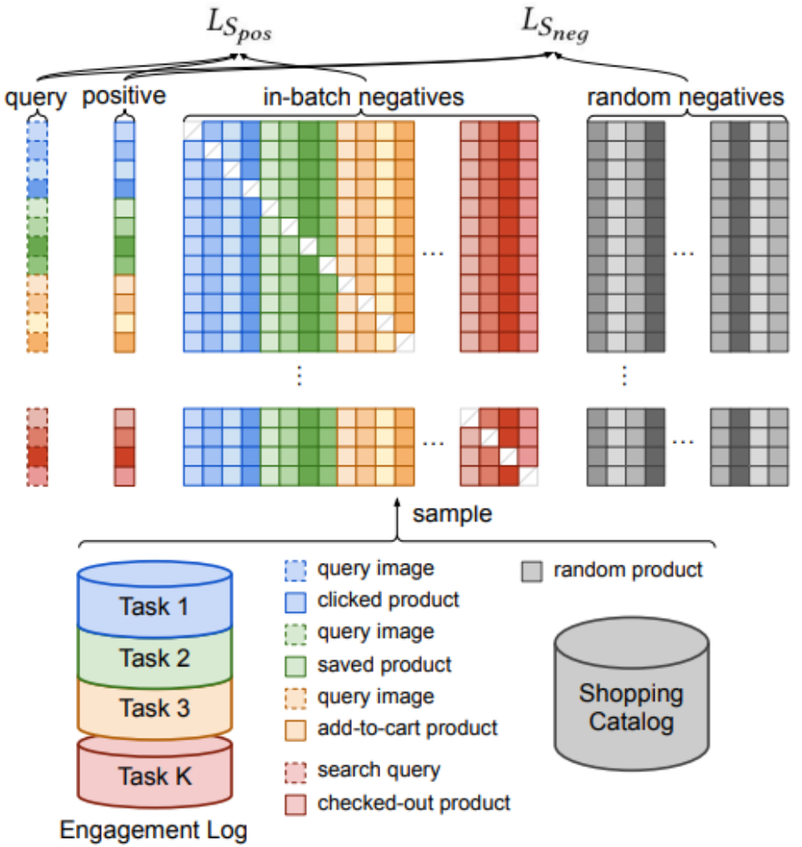
Multimodal Recommendation: Fusion and Finetuning Matters



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

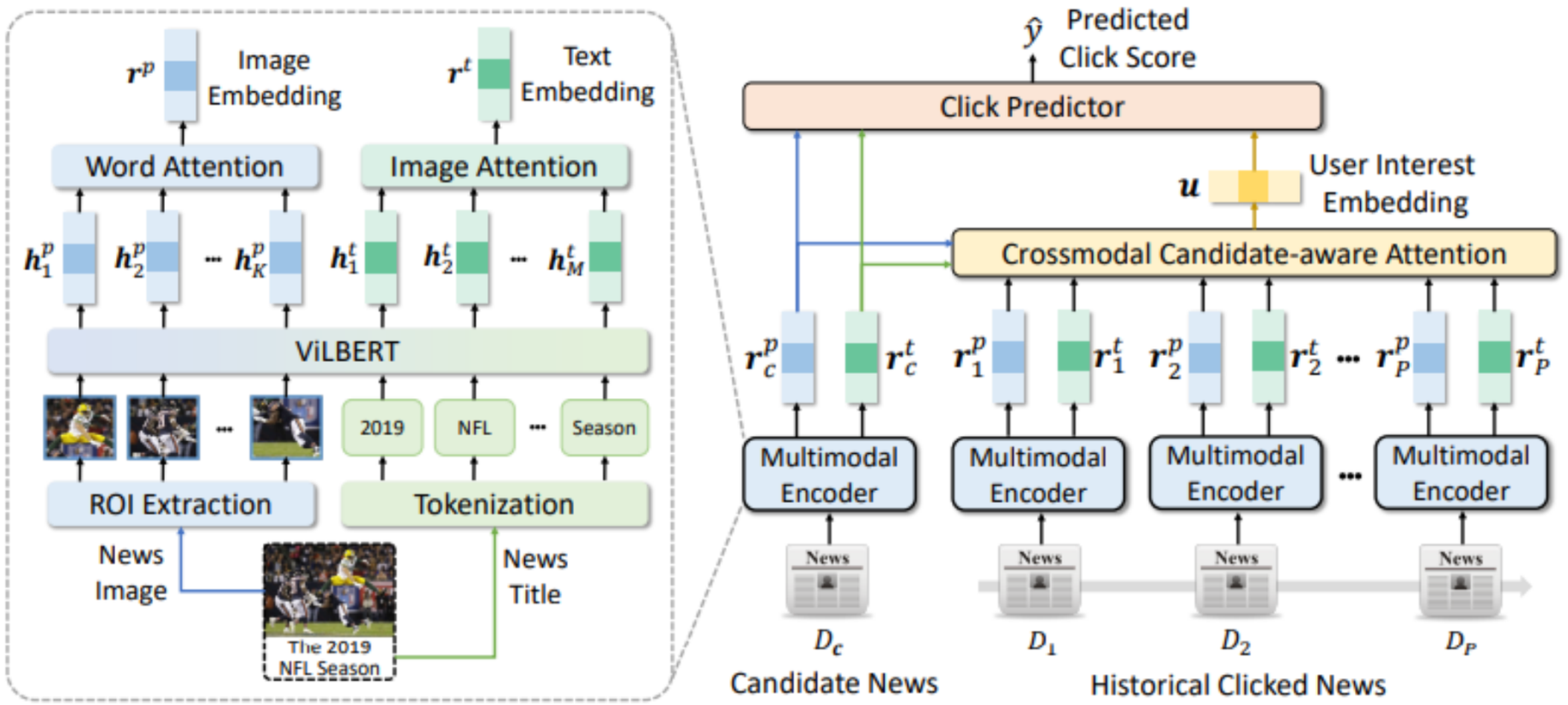
	Number of Parameters	Closeup Clicks	Closeup Saves	Closeup Add-to-Cart	Closeup Checkouts	Search Clicks	Search Saves	Search Add-to-Cart	Search Checkouts
Sum	-	0.663	0.647	0.669	0.699	-	-	-	-
Sum-MLP	-	-	-	-	-	0.577	0.533	0.561	0.629
MLP-Concat-MLP	30.8M	0.805	0.794	0.896	**0.916**	0.723	0.736	0.834	0.861
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

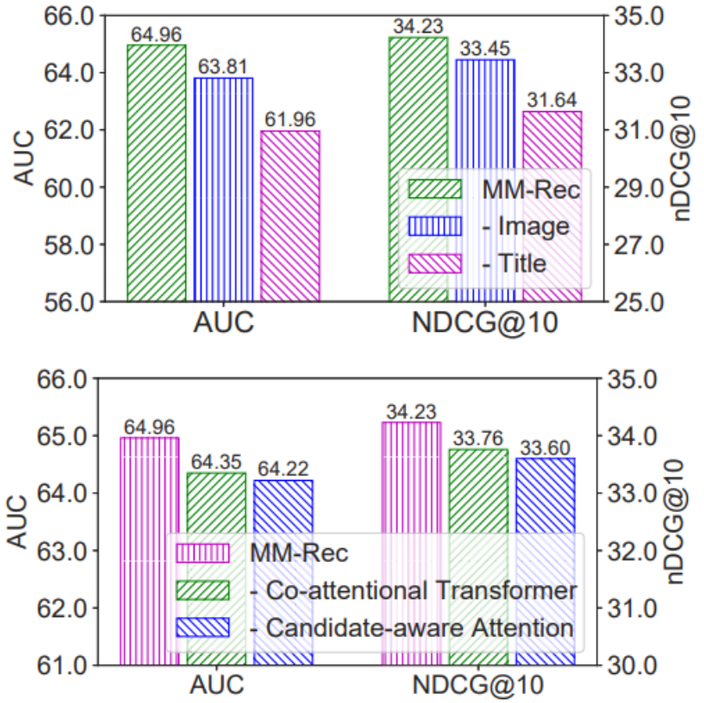


MM-Rec: Visiolinguistic Model Empowered Multimodal News Recommendation

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).