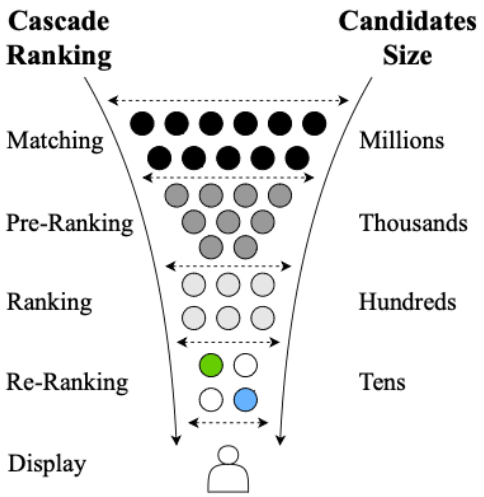
SIGIR 2025
Padova
ITALY

蚂蚁集团
ANT GROUP

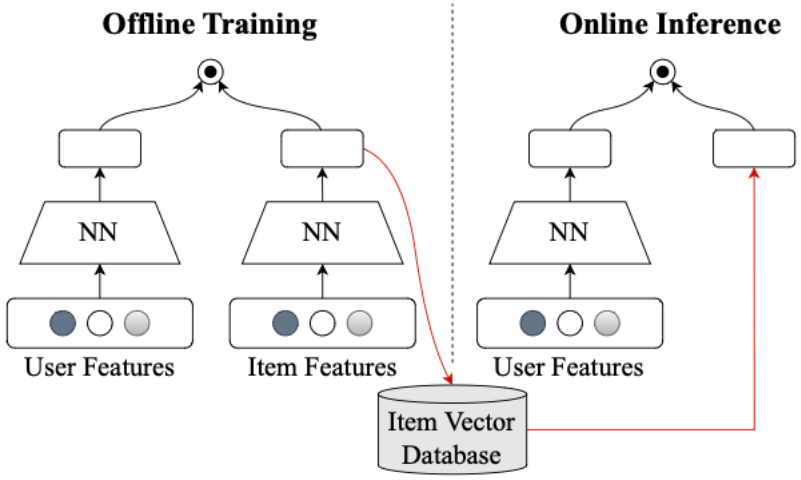
A Learnable Fully Interacted Two-Tower Model for Pre-Ranking System

Chao Xiong, Xianwen Yu, Wei Xu, Lei Cheng, Chuan Yuan, Linjian Mo

A Learnable Fully Interacted Two-Tower Model for Pre-Ranking System - *Background*



(a) Cascade Ranking System



(b) Two-Tower Model

The two-tower model has become the dominant architecture for the pre-ranking considering both effectiveness and efficiency.

Two-Tower Model Advantages:

Excellent Efficiency:

- The item representations can be pre-calculated offline.
- Only the user representations need to be computed in real time.
- Minimal calculation in output layer. (ex dot product)

Two-Tower Model Limitations:

Less interaction:

- No early interaction: User-item signals isolated until output layer.
- Simple late interaction: dot product or cos similarity considering efficiency.

A Learnable Fully Interacted Two-Tower Model for Pre-Ranking System
- *Current Advancements vs. Unsolved Challenges*

SIGIR 2025 Padova ITALY

蚂蚁集团 ANT GROUP

Early-fusion

DAT relies on augmented vectors to implicitly learn information from the opposite tower.

Unresolved Issue:
- Implicitly learned representations have limited expressiveness

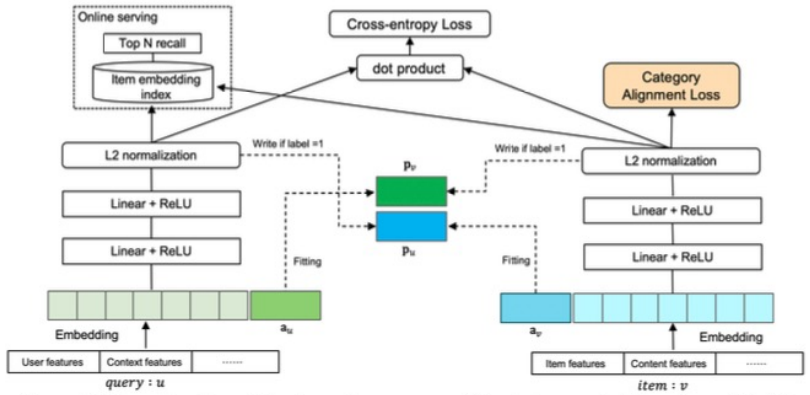
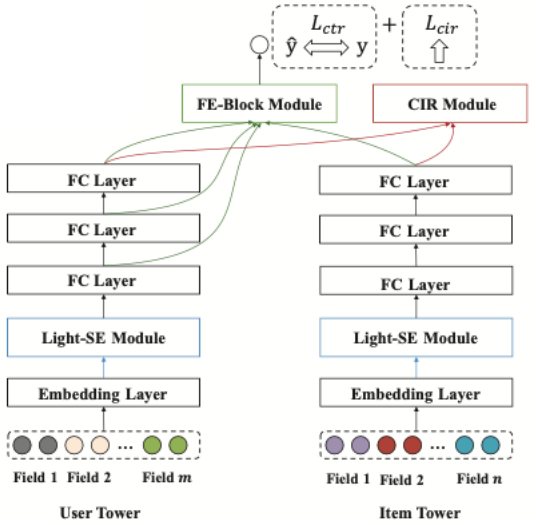


Figure 1: The network architecture of our proposed Dual Augmented Two-tower Model

Late-fusion

IntTower replaces dot product with sum-max scoring between user and item representations.

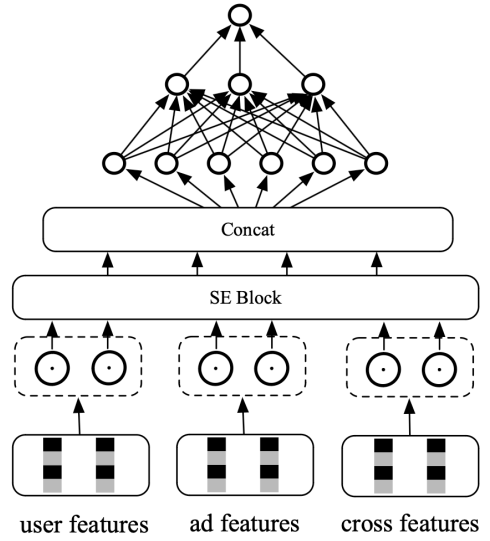
Unresolved Issue:
- Cannot use explicit interactions between user and item features.
- Hand-crafted reduction cannot capture arbitrary user-item interaction.



Single tower

COLD uses a single-tower model to achieve fully feature interactions with several optimization tricks for inference acceleration.

Unresolved Issue:
- Single-tower always sacrifices performance.



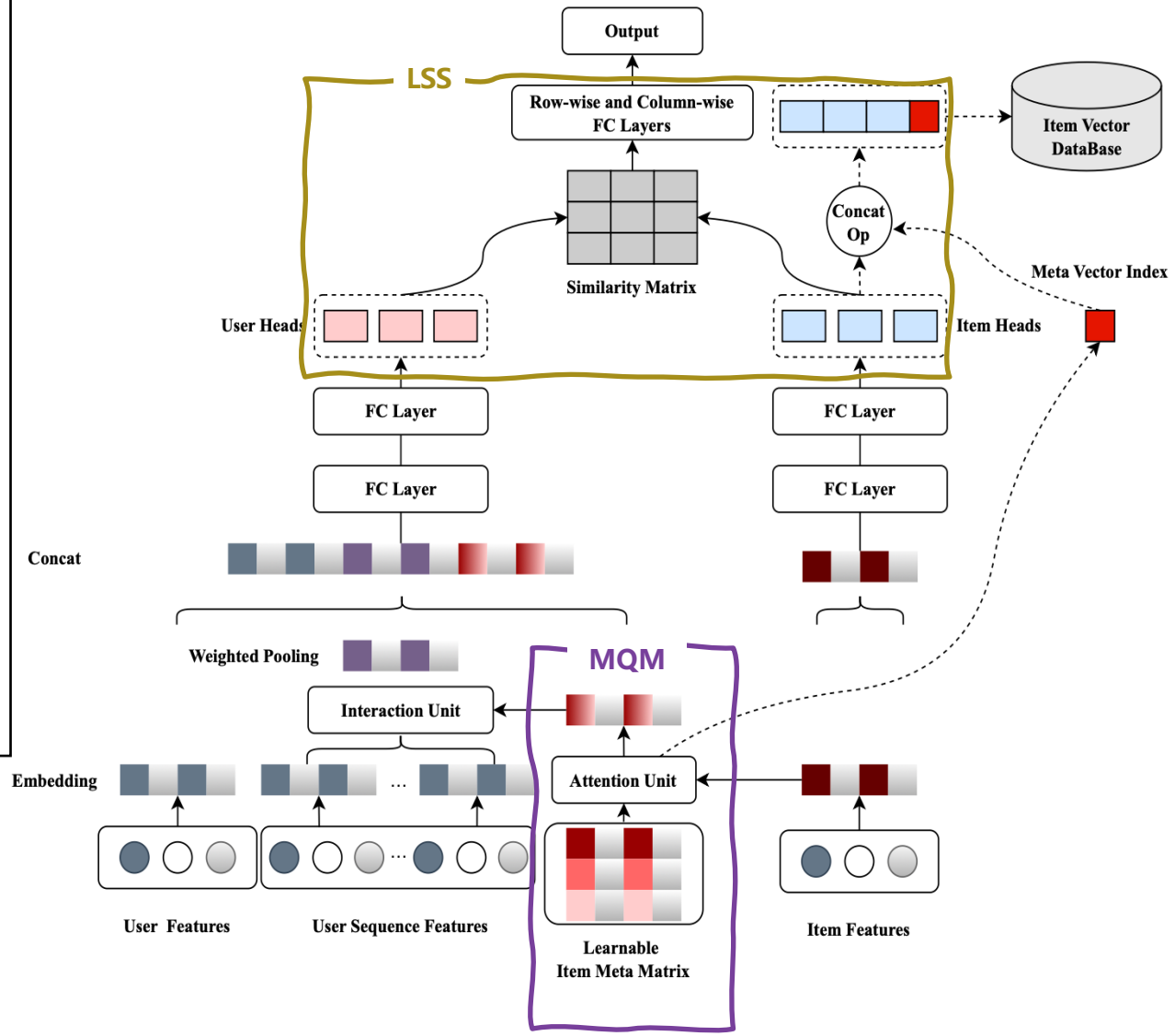
A Learnable Fully Interacted Two-Tower Model for Pre-Ranking System
-Our proposed model

SIGIR 2025 Padova ITALY

蚂蚁集团 ANT GROUP

Main contributions:

- **Meta Query Module (MQM)**: Makes **arbitrary interaction** methods available while maintaining user-item decoupling architecture.
- **Lightweight Similarity Scorer** (LSS): Approachs the **theoretical bound** using only a small number of parameters.
- **SOTA performance** on 4 public datasets
- The deployment in **industrial online** advertising systems has yielded measurable revenue increases。 (not include in Paper)



A Learnable Fully Interacted Two-Tower Model for Pre-Ranking System
- *Meta Query Module* (MQM)

SIGIR 2025 Padova ITALY

蚂蚁集团 ANT GROUP

How it works?

1. Meta matrix : A trainable item matrix each row represents a clustered item embedding
2. The attention unit： Selects the most similar meta vector(query) to represent the original item by computing similarity scores between the meta matrix and item features.
3. Interaction unit： Performs **explicit feature-level interaction** by processing the selected query with user features through DIN (Deep Interest Network) to achieve **early-fusion**.

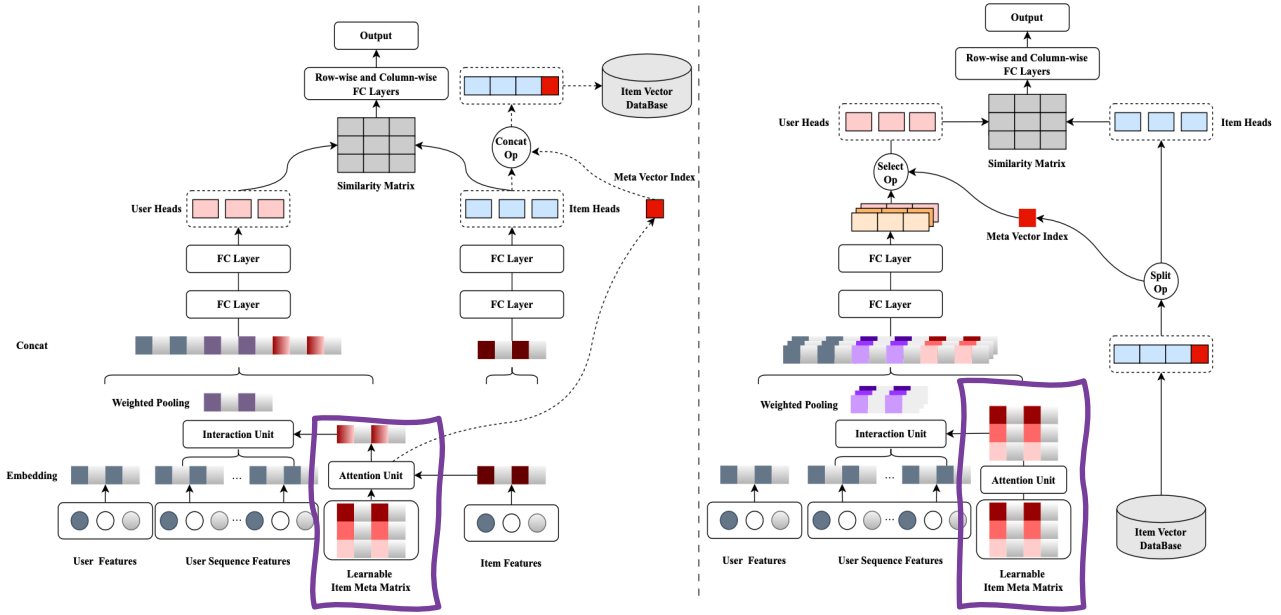


Figure 2: The overall architecture of our proposed FIT. The left part illustrates the training phase and the right part shows the inference phase.

Differences between training and infer phase

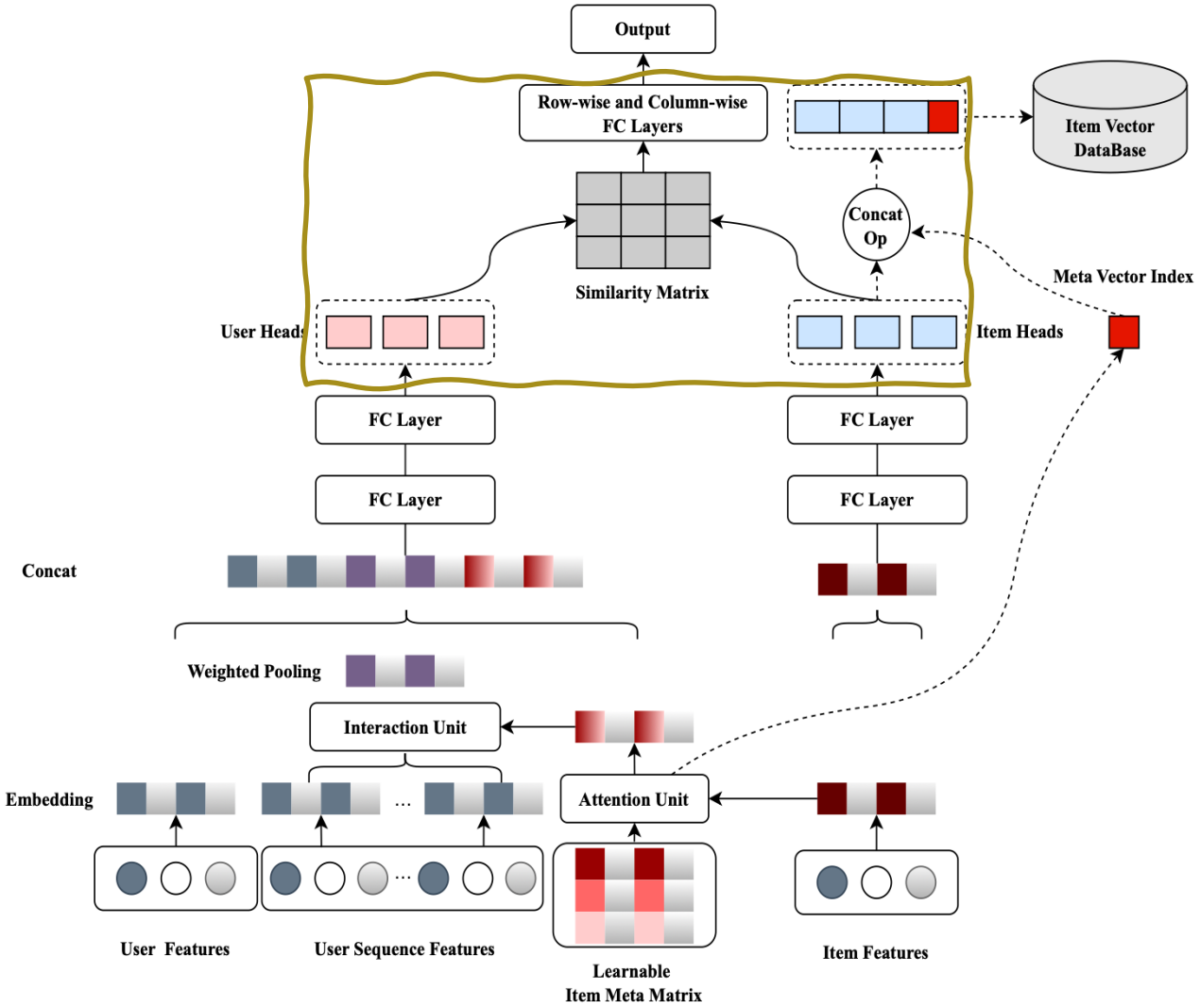
Training phase:

- Retains **only the most similar** meta vector.

Inference phase：

- **All meta vectors** in the meta matrix perform explicit cross-attention with user features. This generates **multiple** intermediate user representations.
- The output layer selects the final representation solely through the **index** of the item's most similar meta vector.

A Learnable Fully Interacted Two-Tower Model for Pre-Ranking System
-Lightweight Similarity Scorer (LSS)



How it works?

1. Maps user/item representations into muti-head sub-spaces to extract diverse information.
2. Builds similarity matrix by dot products of all user-item head pairs.
3. Sequentially perform the row-wise and column-wise computation to the similarity matrix and obtain the final scalar score.

Why better?

1. Row-wise and column-wise FC layers have fewer parameters, which can reduce the computational cost of online inference, especially when the head size Hu and Hv are large
2. As proved in [12], this scorer universally approximates continuous functions in ℓ^2 space, even with constrained heads.

A Learnable Fully Interacted Two-Tower Model for Pre-Ranking System
-*Experiments*

SIGIR 2025 Padova ITALY

蚂蚁集团 ANT GROUP

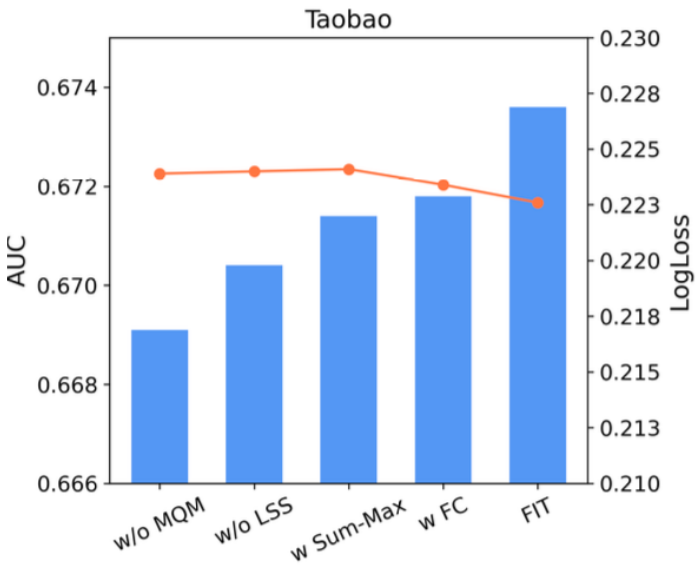
Comparison of Model Effectiveness

Model	ML-1M			Amazon Electro			Amazon Books			Taobao		
	AUC	Logloss	RelaImpr	AUC	Logloss	RelaImpr	AUC	Logloss	RelaImpr	AUC	Logloss	RelaImpr
Two-Tower	0.8695	0.4508	0.00%	0.8441	0.4870	0.00%	0.8742	0.4484	0.00%	0.6577	0.2257	0.00%
DAT	0.8784	0.4363	2.41%	0.8465	0.4858	0.70%	0.8802	0.4413	1.61%	0.6605	0.2241	1.74%
COLD	0.9009	0.4316	8.51%	0.8504	0.4833	1.82%	0.8996	0.4072	6.80%	0.6645	0.2236	4.28%
IntTower	0.9036	0.4142	9.24%	0.8490	0.4807	1.42%	0.8972	0.4444	6.15%	0.6638	0.2245	3.81%
RankTower	0.9095	0.3986	10.82%	0.8520	0.4797	2.31%	0.9031	0.4149	7.75%	0.6630	0.2250	3.34%
FIT	**0.9225***	**0.3766***	**14.34%**	**0.8598***	**0.4707***	**4.58%**	**0.9171***	**0.3731***	**11.47%**	**0.6736***	**0.2226***	**10.03%**

Comparison of Model Efficiency

Model	Training Time	Latency	Storage
Two-Tower	228 s	25.4 ms	1×
DAT	238 s	25.6 ms	0.25×
COLD	284 s	31.4 ms	0×
IntTower	280 s	27.4 ms	8×
RankTower	231 s	33.3 ms	1×
FIT	260 s	26.9 ms	1×

Component Effectiveness Analysis



- **w/o MQM** : remove MQM .
- **w/o LSS** : remove LSS and the output layer is the same as the Two-Tower.
- **w SumMax** : replace LSS with the sum-max similarity score.
- **w FC** : replace LSS with a FC layer.

Online AB test（Alipay ads）

- FIT has been deployed for **handling major online traffic** in Alipay’s advertising platform.
- GMV： **+1.41%**
- AUC： **+1.2%**
- Latency： +0.8ms

SIGIR 2025
Padova
ITALY

蚂蚁集团
ANT GROUP

Thanks