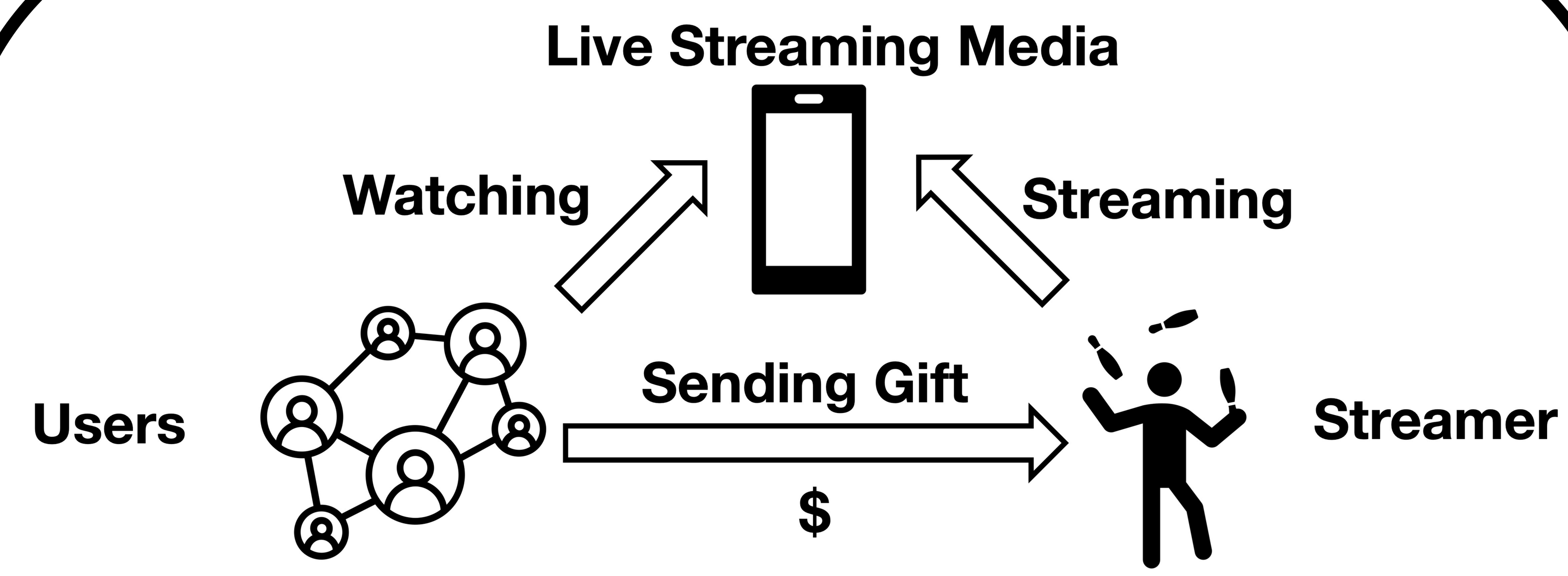
MMBee: Live Streaming Gift-Sending Recommendations via Multi-Modal Fusion and Behaviour Expansion

Jiaxin Deng[1,2], Shiyao Wang[3], Yuchen Wang[3], Jiansong Qi[3], Liqin Zhao[3], Guorui Zhou[3] and Gaofeng Meng[1]
[1] MAIS, Institute of Automation, Chinese Academy of Science
[2] University of Chinese Academy of Science, [3] Kuaishou Inc.

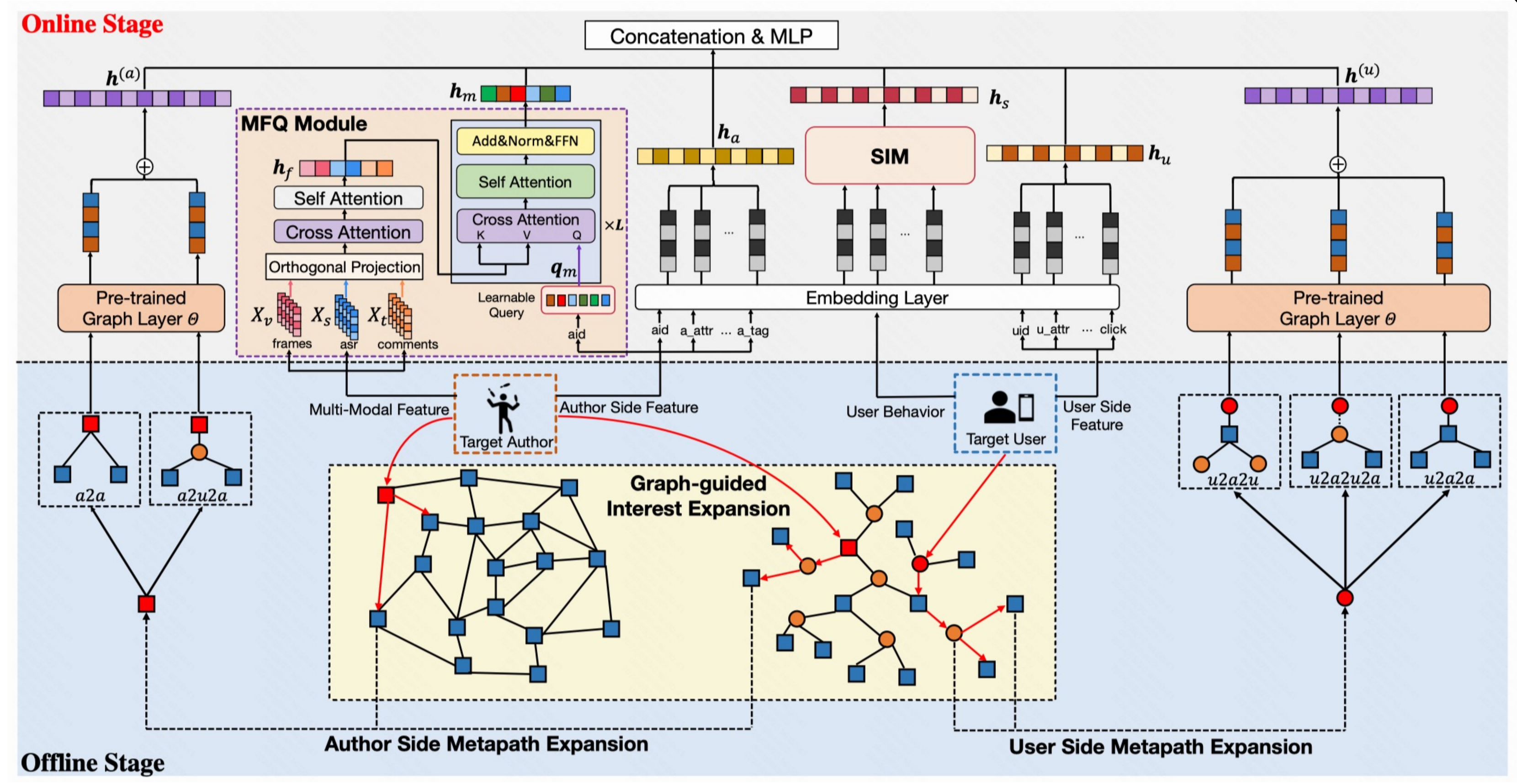
Introduction



The Live Streaming Gifting Scenario

- Motivation
  - Different from conventional recommendation problem, it is challenging to precisely describe the ***real-time content changes*** in live streaming recommendation.
  - Due to the ***sparsity of gifting behaviors***, capturing the preferences and intentions of users is quite difficult.
- Contributions
  - The proposed Multi-modal Fusion with Learnable Query (MFQ) module leverages the dynamic multimodal content of live streaming and captures the distinct characteristics among streamers.
  - Graph-guided Interest Expansion (GIE) module largely enriches the observed history behaviors of users and streamers with both self-supervised graph representation learning and metapath-based behavior expansion to alleviate the sparsity problem.
  - Online A/B tests further show that MMBee brings significant online benefits and we build efficient industrial infrastructure to deploy MMBee on the real-world online live streaming recommendation.

Method



Multi-modal Fusion with Learnable Query

The proposed Multi-modal Fusion Module with Learnable Query (MFQ) module helps the model to perceive the real-time content changes in live streaming through processing the complex visual frames, comments and audio in each streaming segment.

$$\boldsymbol{h}_v = \mathrm{CrossAttention}(X_v \boldsymbol{W}_v^Q, Y_v \boldsymbol{W}_v^K, Y_v \boldsymbol{W}_v^V), Y_v = OP(X_v, X_s, X_t)$$

$$\boldsymbol{h}_s = \mathrm{CrossAttention}(X_s \boldsymbol{W}_s^Q, Y_s \boldsymbol{W}_s^K, Y_s \boldsymbol{W}_s^V), Y_s = OP(X_s, X_t, X_v)$$

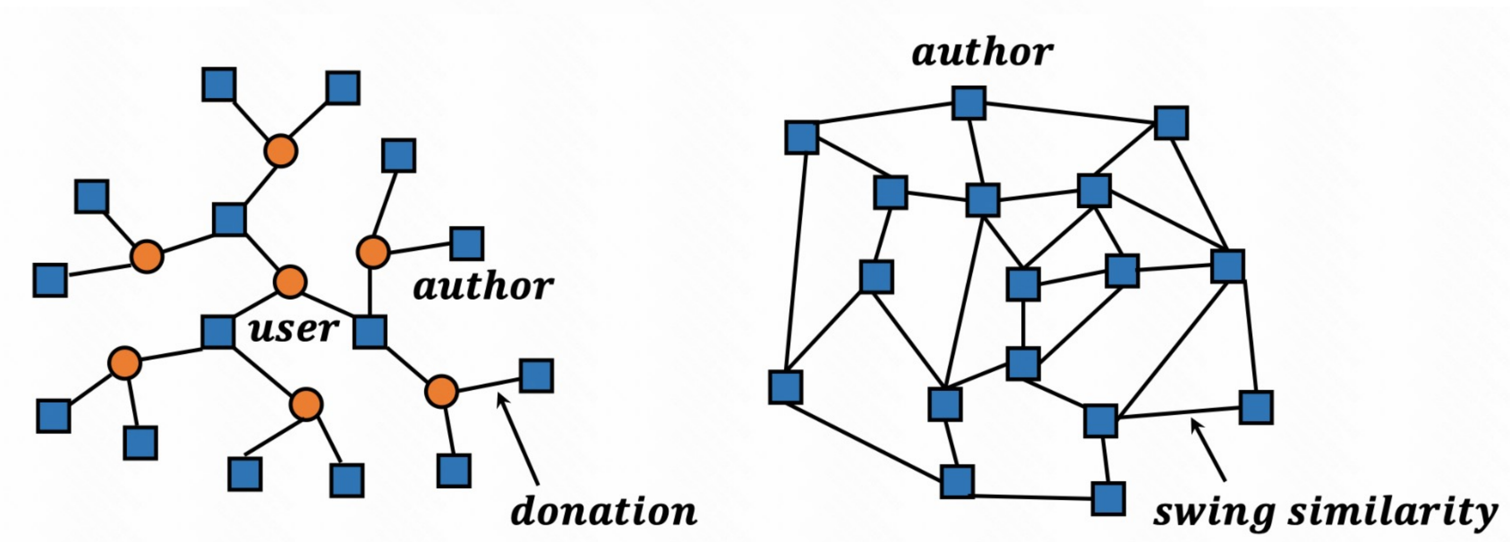
$$\boldsymbol{h}_t = \mathrm{CrossAttention}(X_t \boldsymbol{W}_t^Q, Y_t \boldsymbol{W}_t^K, Y_t \boldsymbol{W}_t^V), Y_t = OP(X_t, X_s, X_v)$$

$$\boldsymbol{h}_m' = \mathrm{CrossAttention}(\boldsymbol{q}_m \boldsymbol{W}_c^Q, \boldsymbol{h}_f \boldsymbol{W}_c^K, \boldsymbol{h}_f \boldsymbol{W}_c^V)$$

Graph-guided Interest Expansion

$$\mathbb{E}^{(u)} = \{\Theta(v_i) | v_i \in \mathcal{N}^{(2)}_{\rho_{u2a2u}}(u_t) \cup \mathcal{N}^{(3)}_{\rho_{u2a2u2a}}(u_t) \cup \mathcal{N}^{(2)}_{\rho_{u2a2a}}(u_t)\}$$

$$\mathbb{E}^{(a)} = \{\Theta(v_i) | v_i \in \mathcal{N}^{(1)}_{\rho_{a2a}}(a_t) \cup \mathcal{N}^{(2)}_{\rho_{a2u2a}}(a_t)\}$$



(a) User-to-Author **(b) Author-to-Author**

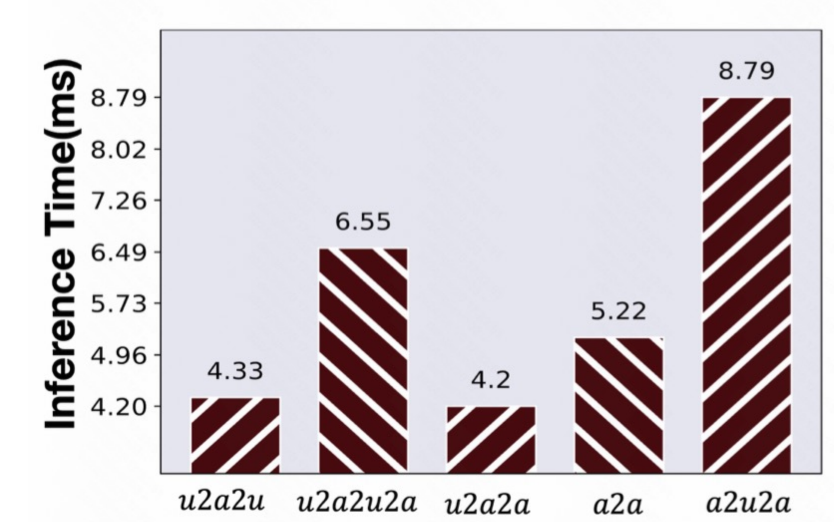
Experiments

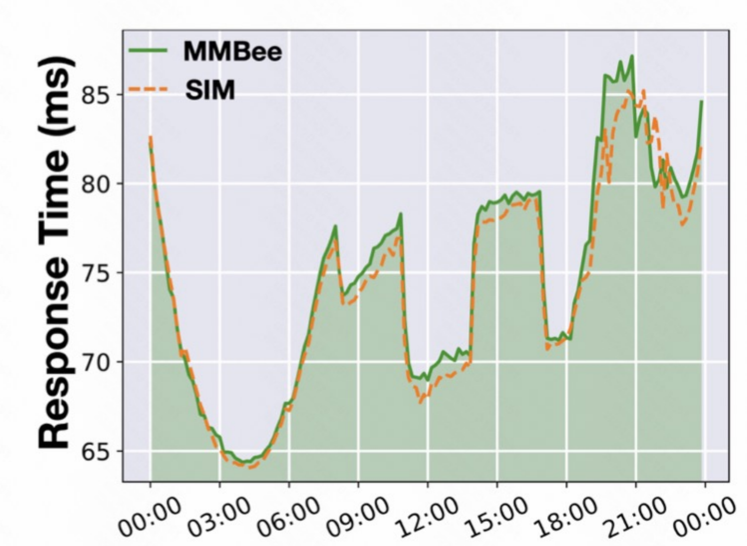
Methods	GTR					
	AUC	Impr.*	UAUC	Impr.*	GAUC	Impr.*
MMoE [16]	0.956230	-	0.730186	-	0.746711	-
MMoE+BDR [39]	0.956908	+0.0678 %	0.730625	+0.0439 %	0.747136	+0.0425 %
MMoE+MTA [32]	0.957095	+0.0865 %	0.731450	+0.1264 %	0.747327	+0.0616 %
MMoE+EgoFusion [4]	0.956952	+0.0722 %	0.731418	+0.1232 %	0.747275	+0.0564 %
MMoE+MFQ	0.956902	+0.0672 %	0.731975	+0.1789 %	0.747275	+0.1764 %
MMoE+GIE	0.957064	+0.0834 %	0.733853	+0.3667 %	0.751239	+0.4528 %
MMoE+Ours(MFQ+GIE)	**0.95723**	**+0.1001 %**	**0.735776**	**+0.5590 %**	**0.753017**	**+0.6306 %**
SIM [20]	0.958656	-	0.732239	-	0.748383	-
SIM+BDR [39]	0.958419	-0.0237 %	0.734757	+0.2518 %	0.750684	+0.2301 %
SIM+MTA [32]	0.958867	+0.0211 %	0.734921	+0.2682 %	0.750802	+0.2419 %
SIM+EgoFusion [4]	0.959387	+0.0085 %	0.735608	+0.3369 %	0.751669	+0.3286 %
SIM+MFQ	0.959202	+0.0546 %	0.735717	+0.3478 %	0.751780	+0.3397 %
SIM+GIE	0.959802	+0.1146 %	0.738309	+0.6070 %	0.755154	+0.6771 %
SIM+Ours(MFQ+GIE)	**0.960302**	**+0.1646 %**	**0.743678**	**+1.1439 %**	**0.76044**	**+1.2057 %**
p-value	$1.02e^{-3}$		$2.01e^{-3}$		$5.12e^{-3}$	

Performance on Kuaishou Dataset.

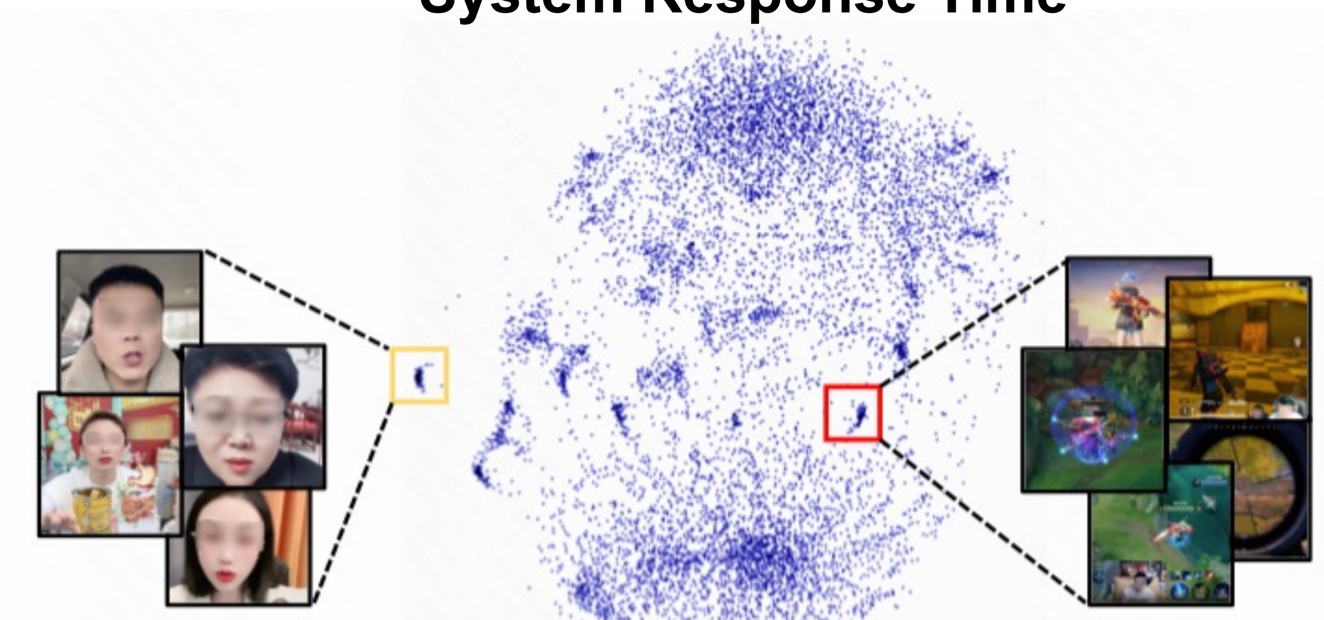
Methods	TikTok			Movielens		
	Recall@10	Precision@10	NDCG@10	Recall@10	Precision@10	NDCG@10
NGCF [28]	0.0292	0.0045	0.0156	0.1198	0.0289	0.0750
LightGCN [8]	0.0448	0.0082	0.0261	0.1992	0.0479	0.1324
MMGCN [30]	0.0544	0.0089	0.0297	0.2028	0.0506	0.1361
GRCN [29]	0.0392	0.0065	0.0221	0.1402	0.0338	0.0882
EgoGCN [4]	<u>0.0569</u>	<u>0.0093</u>	<u>0.0330</u>	0.2155	<u>0.0524</u>	<u>0.1444</u>
DIN [42]	0.0403	0.0074	0.0235	0.1372	0.0330	0.0912
SASRec [9]	0.0435	0.0043	0.0215	0.1914	0.0191	0.1006
SIM [20]	0.0413	0.0079	0.0245	0.1470	0.0429	0.1011
MMMLP [15]	0.0509	0.0081	0.0297	0.1842	0.0484	0.1328
MMSSL [20]	0.0553	0.0055	0.0299	**0.2482**	0.0170	0.1113
Ours	**0.0605**	**0.0097**	**0.0347**	<u>0.2317</u>	**0.0566**	**0.1573**
p-value	$1.29e^{-5}$	$6.23e^{-6}$	$7.29e^{-5}$	$2.75e^{-5}$	$2.81e^{-3}$	$1.61e^{-2}$

Performance on TikTok and ML Dataset.





System Response Time



Visualization Study