Digital Human Interactive Recommendation Decision-Making Based on Reinforcement Learning

Xiong Junwu[1], Xiaoyun Feng[1] , YunZhou Shi[2] , James Zhang[1] , Zhongzhou Zhao[2] , Wei Zhou[2]

1：AntGroup 2：Damo Academy, Alibaba Group

1 Background

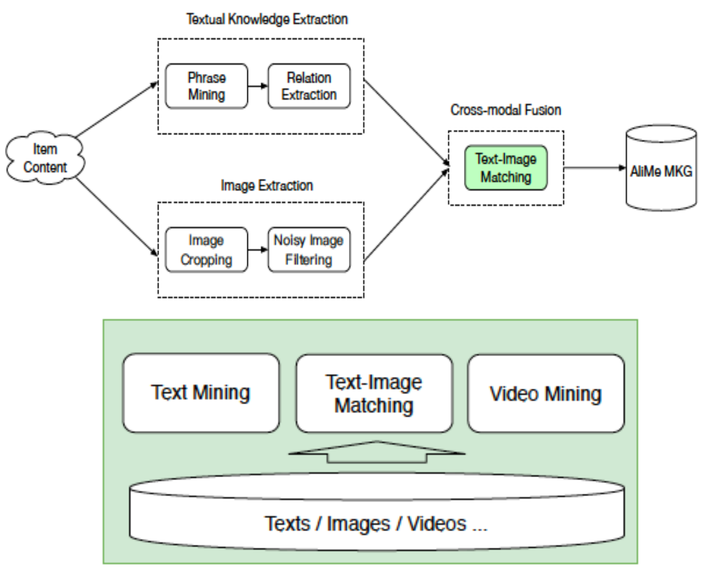
Digital human recommendation system: It helps customers find their favorite products and is playing an active role in various recommendation contexts. However, how to timely catch and learn the dynamics of the preferences of the customers, while meeting their exact requirements?

- **Conventional Recommendation System**: Adapt to passive display-based recommendation contexts, the customer can only passively consume the prepared items and often with a single chance of action (e.g. watching or clicking only once among the recommended items).
- **Reinforcement Learning based (RL) Recommendation System**: Mainly applied to passive display-based recommendation contexts so far
- **Digital Human Recommendation System**: Encounter the same problems as traditional recommendation systems

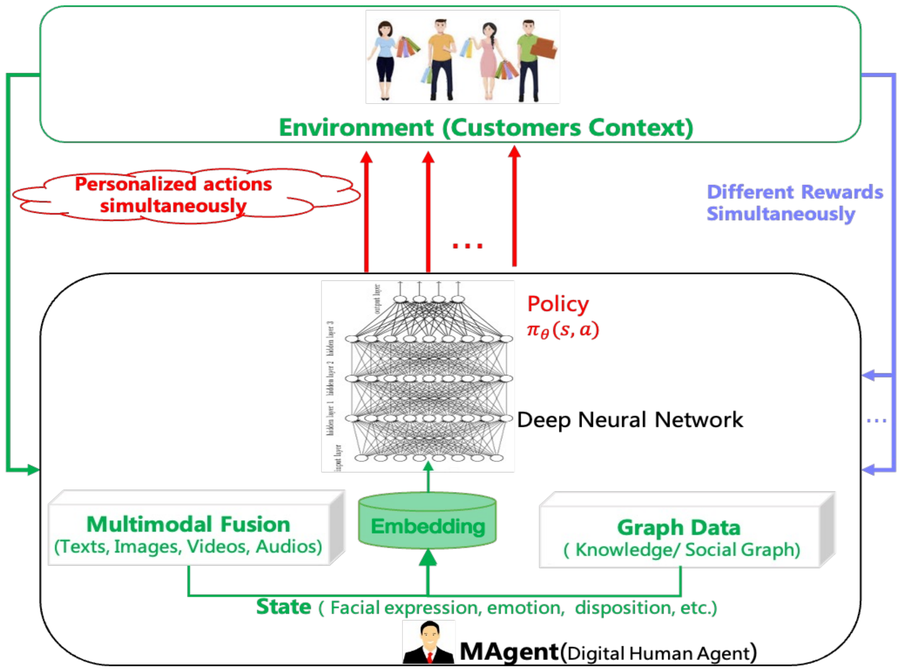
2 Our Proposed Framework

We design a novel and practical digital human recommendation agent framework based on RL to improve the efficiency of decision-making by leveraging both the digital human features and the superior flexibility of RL.

(a) Multi-modal and graph embedding[1,2]



(b) Our proposed MAgent framework

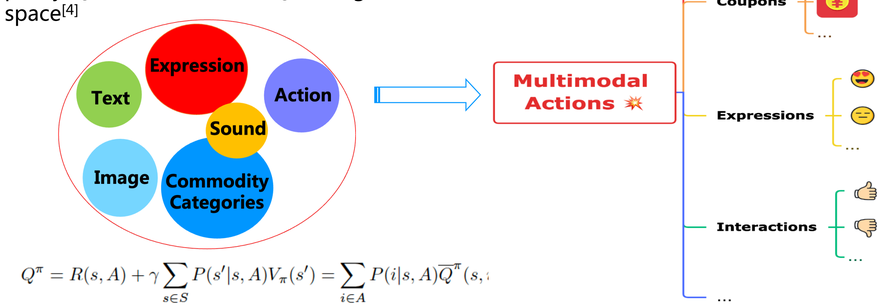


(c) The action explosion problem and tuning

(i) Digital human policy learning with SAC[3]

$$\pi^* = \arg\max_\pi \mathbb{E}_{(s_t,a_t)} [\sum_t \underbrace{R(s_t,a_t)}_{reward} + \alpha \underbrace{H(\pi(\cdot|s_t))}_{entropy}]$$

(ii) Multimodal behaviors explosion (users, items and their interactions): Learning through on-policy Q funtion with SlateQ for large action space[4]



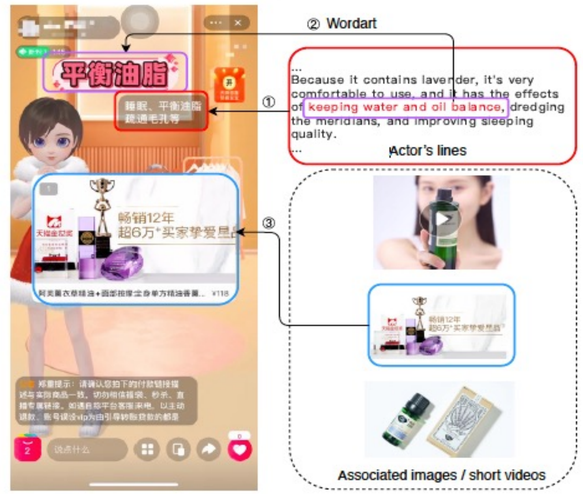
$$Q^\pi = R(s,A) + \gamma \sum_{s\in S} P(s'|s,A)V_\pi(s') = \sum_{i\in A} P(i|s,A)\overline{Q}^\pi(s,$$

Our proposed framework learns through real-time interactions between the digital human and customers dynamically through the state-of-the-art RL algorithms[3,4], combined with multi-modal embedding and graph embedding, to improve the accuracy of personalization and thus enable the digital human agent to timely catch the attention of the customer.

3 Virtual Live Broadcast Example

Our proposed framework can be easily adapted to fully dynamic contexts appropriately, especially in interactive recommendation decision-making contexts such as in a virtual live broadcast room.

A demo of Alime Avatar prodcut recommendation[2]



4 Performance Evaluation

Evaluate the performance of **MAgent** under the context of live-streaming broadcast with real-world business data and compare the corresponding conversion rate of transactions on regular days, as well as on marketing campaign days.

Digital human recommendation performance based on RL framework (DFM[5]: Deep factorization model, SAC[3]: Soft actor-critic, **MRR**[6]: Mean reciprocal rank)

	MRR	Hit1@Recall
DFM	0.255	3.9%
SAC	0.308	4.2%

References

[1] Guohai Xu, Hehong Chen, Feng-Lin Li, Fu Sun, Yunzhou Shi, Zhixiong Zeng, Wei Zhou, Zhongzhou Zhao, and Ji Zhang. Alime mkg: A multi-modal knowledge graph for live-streaming e-commerce. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pages 4808–4812, 2021.

[2] Feng-Lin Li, Zhongzhou Zhao, Qin Lu, Xuming Lin, Hehong Chen, Bo Chen, Liming Pu, Jiashuo Zhang, Fu Sun, Xikai Liu, et al. Alime avatar: Multi-modal content production and presentation for live-streaming e-commerce. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2635–2636, 2021

[3] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. arXiv preprint arXiv:1812.05905, 2018.

[4] Eugene Ie, Vihan Jain, Jing Wang, Sanmit Narvekar, Ritesh Agarwal, Rui Wu, Heng-Tze Cheng, Morgane Lustman, Vince Gatto, Paul Covington, et al. Reinforcement learning for slate-based recommender systems: A tractable decomposition and practical methodology. arXiv preprint arXiv:1905.12767, 2019.

[5]Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. Deepfm: a factorization-machine based neural network for ctr prediction. arXiv preprint arXiv:1703.04247,2017.

[6] Quoc V Le, Alex Smola, Olivier Chapelle, and Choon Hui Teo. Optimization of ranking measures. Journal of Machine Learning Research, 1:1–48, 2010.