MMCRec: Towards Multi-Modal Generative AI in Conversational Recommendation

Tendai Mukande[1,3][0000−0002−0654−7141], Esraa Ali[1,4][0000−0003−1600−3161], Annalina Caputo[1,4][0000−0002−7144−8545], Ruihai Dong[2,5][0000−0002−2509−1370], and Noel E. O'Connor[1,3,5][0000−0002−4033−9135]

[1] Dublin City University, Dublin 9, Ireland https://www.dcu.ie/
[2] University College Dublin, Dublin 4, Ireland https://www.ucd.ie/
[3] SFI ML-LABS, Dublin, Ireland https://www.ml-labs.ie/
[4] ADAPT Centre, Dublin, Ireland https://www.adaptcentre.ie/
[5] Insight SFI Research Centre for Data Analytics, Dublin, Ireland https://www.insight-centre.org/
tendai.mukande2@mail.dcu.ie, esraa.ali@adaptcentre.ie, ruihai.dong@ucd.ie, {annalina.caputo,noel.oconnor}@dcu.ie

Abstract. Personalized recommendation systems have become integral in this digital age by facilitating content discovery to users and products tailored to their preferences. Since the Generative Artificial Intelligence (GAI) boom, research into GAI-enhanced Conversational Recommender Systems (CRSs) has sparked great interest. Most existing methods, however, mainly rely on one mode of input such as text, thereby limiting their ability to capture content diversity. This is also inconsistent with real-world scenarios, which involve multi-modal input data and output data. To address these limitations, we propose the Multi-Modal Conversational Recommender System (MMCRec) model which harnesses multiple modalities, including text, images, voice and video to enhance the recommendation performance and experience. Our model is capable of not only accepting multi-mode input, but also generating multi-modal output in conversational recommendation. Experimental evaluations demonstrate the effectiveness of our model in real-world conversational recommendation scenarios.

Keywords: Generative AI, Large Language Model, Conversational Recommendation, Graph Neural Network, Diffusion Model

1 Introduction

The data generated in the increasingly digital world is a valuable source of information to create personalized recommendations. This data comes in multiple forms such as text, audio, videos, and images. Recommender Systems (RSs) have become essential components of e-commerce platforms, content streaming services, and social media networks [22]. Most traditional RSs rely on historical offline user-item interaction data, and this constrains their ability to fully capture dynamic user preferences [47]. To resolve this issue, CRSs have been proposed

to enable proactive dialogue with users as well as generate more accurate and explainable recommendations [18]. Most state-of-the-art CRSs however, rely on unimodal data such as text, which limits their ability to capture the complexity of real-world user intent [5]. This limitation has spurred research into CRSs that are multi-modal and interactive [6,17]. The motivation behind multi-modal CRSs is to harness the power of multiple modalities [10] in order to enrich user experiences and offer explainable recommendations that align more closely with user preferences [23]. In this way, users would be able to express their preferences through several ways which include textual queries, images and voice commands. Moreover, CRSs incorporating diverse modalities would provide richer and more nuanced representations of user intent, thereby comprehending and responding to user needs more effectively [31]. Several methods have been proposed for multi-modal CRSs but they do not sufficiently address the current CRS challenges [28] such as incorporating multi-modal input/output combinations of the data forms [17,21,43] as well as limited explainability [7], thereby limiting the transparency of the recommendations [14].

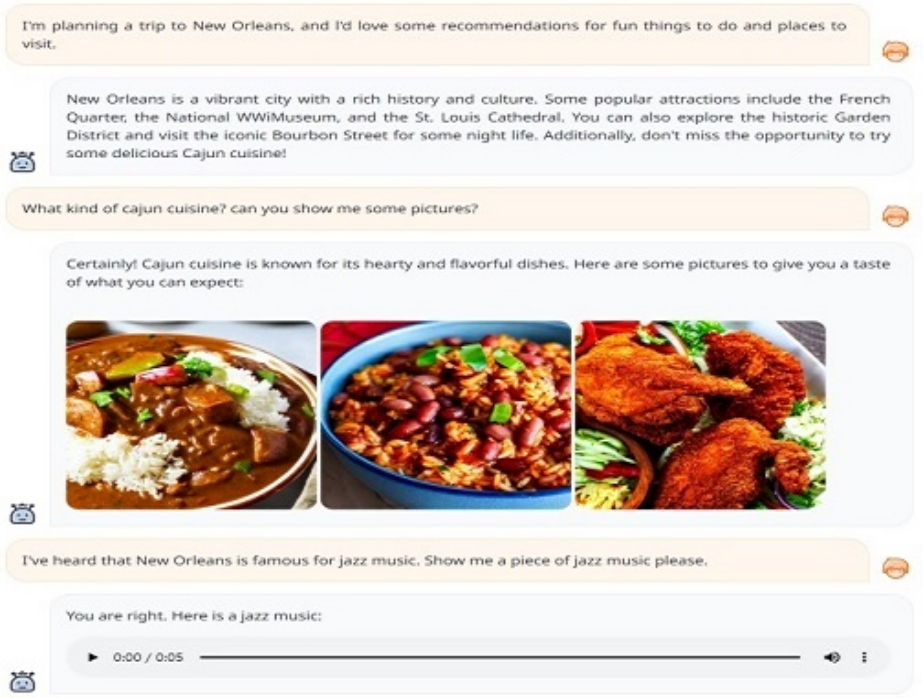

Fig. 1. An example of multi-modal Conversational Recommendation. Source:[1]

To address the highlighted challenges, we propose the Multi-Modal Conversational Recommender System (MMCRec) model capable of handling input and output combinations of text, images, videos and audio data. Using a Large Language Model (LLM) as the backbone of the model, multi-modal input encoders

and output diffusion decoder modules enable a context-aware recommendation experience. Users can interact with the model through text, share images, audio and videos of their desired items. This multi-modal approach does not only enrich user engagement but also facilitates better understanding of user intent and context, thereby improving the quality of the recommendations [15]. Our **contributions** are as follows:

- We extend the NExT-GPT [1] model to the conversational recommendation scenario. Leveraging on the multi-modal input encoding as well as output image, audio and video diffusion modules, the resultant framework is a *multi-modal input-multi-modal output* LLM-based CRS which has understanding, reasoning and explanation abilities.
- We apply a Graph-of-Thought (GoT) [48] prompting model to enhance the LLM reasoning process which leverages on cross-attention among the text, video, image and video modalities [40]. We hypothesize that this approach adds additional context and semantic understanding of the queries by the model [41].

In our experiments, we apply the proposed model to the M5Product dataset for real world cross-modal recommendation scenarios.

2 Related Work

GAI Models such as ChatGPT, LLaMA and BARD have been shown to demonstrate remarkable abilities in natural language understanding, multi-step reasoning and undertaking dialogue tasks [12,29]. In zero-shot or few-shot learning settings, they have been shown to adapt to handle to novel tasks [2,3,20]. Methods such as Reinforcement Learning from Human Feedback (RLHF) [11,16], Chain-of-Thought Prompting [40], self-consistency and boosting techniques have been explored by prior research to enhance contextual understanding and reasoning in GAIs [35]. To improve perception in real-world scenarios, the integration of multi-modal capabilities into GAI models has also been studied, such as understanding of other modalities which include image video, audio, etc [9,13]. A notable approach involves the use of adapters that align pre-trained encoders in other modalities to textual LLM [44]. In this line of research, multi-modal Language-Vision models such as PandaGPT [27] have been proposed. With these models, multi-modal input with text output is achieved for general-purpose conversational tasks [33]. Other works have also explored the use of generative AI models in conversational recommendation tasks [20,24,42]. The main limitation with these models is that they do not generate multi-modal recommendations, which is inconsistent with practical scenarios. Wu et al. proposed the Next-GPT LLM in which they added multi-modal output generation with multi-modal input content [1]. To address the limitations of state-of-the-art methods, we build upon this approach in conversational recommendation settings, which is consistent with real-world scenarios.

3 Methodology

In this section, we introduce our MM-MMCRec model. The novelty of our approach lies in the multi-modal graph-based derivation of potential user-item interactions. We hypothesize that the generated multi-modal recommendations enhance the user experience. Input instructions to the LLM module are passed in the form of audio, video, text or image. We leverage on the multi-modal joint embedding capabilities of ImageBIND [25] to enable multi-modal input queries to the CRS. We adopt a GNN cross-attention mechanism [8] to capture semantic relationships among the text, image, video and audio modalities. In this way multi-modal context is added to the queries, thereby complementing the reasoning abilities of the model. We also adopt Multi-modal Alignment Learning and Modality-switching Instruction Tuning as proposed by Wu et al. [1] to understand user input queries and generate the requested multi-modal recommendation outputs.

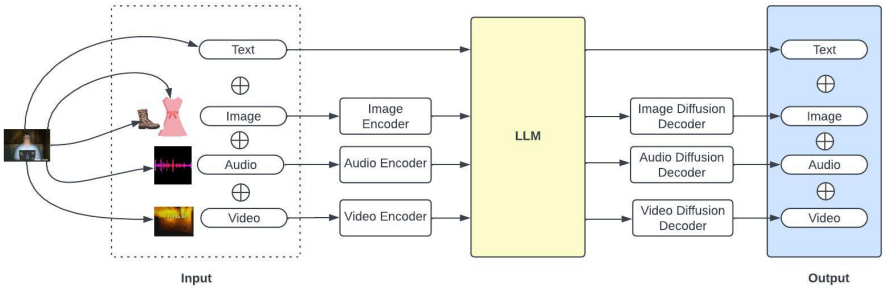

Fig. 2. MMCRec model overview. Input instructions to the LLM module can be passed in the form of audio, video, text or image. Graph-engineered prompts enable cross-attention between the multi-modal input data. Audio, Image and Video diffusion models decode and synthesize the required output as one or a combination of the required modalities.

We adopt the Vicuna LLM [26] which has understanding abilities to multi-modal input representations [27]. Following the NextGPT model [1], we adopt the transformer-based multi-modal output generation whereby multi-modal signal token representations from the LLM are mapped to Image Diffusion, Audio Diffusion and Video Diffusion decoders. We also adopt the following latent diffusion models to generate the multi-modal outputs: Stable Diffusion [37] for image synthesis, Zeroscope [6] for video synthesis, and AudioLDM5 [38] for audio synthesis. For our model instruction tuning, we leverage on the Low-Rank Adaptation (LoRA) approach to minimize the number of trainable parameters for the conversational recommendation task [39].

[6] https://huggingface.co/cerspense.

In adopting the GoT prompting model, multi-modal inputs to the LLM are represented as nodes and edges to capture the non-sequential complex scenarios in real-world scenarios. In this way, the intermediate LLM reasoning model is enhanced to mimic the human-like thought process in generating the recommendations. Following the GoT framework [48], the graph input (representation of the user-item interactions) to the LLM models the reasoning stage to improve the model's contextual understanding of complex real-world heterogeneous recommendation scenarios. This reasoning process is modelled as a heterogeneous graph G = (V, E) where V is a set of nodes representing the users/items and E is a set of edges representing the node relationships. The recommendation task, involves the graph-enhanced reasoning multi-modal model which maps node relationships based on the respective user queries.

4 Experiments

We consider the multi-modal product recommendation task for our experiments. The task aims to find the most relevant target products using one or a combinations of or more modalities, enhanced by the LLM and the multi-modal interaction with the CRS. User input query is in the form of a combination of text and other modalities to provide better context whereas the recommendation output can be enhanced in an interactive way. In conducting our experiments, we aim to answer the **research questions** outlined below:

- **RQ1:** Does the integration of multi-modal data in the CRS model contribute to the improvement of the recommendation performance?
- **RQ2:** To what extend does the multi-modal input modules in the proposed MMCRec model impact the recomendation performance?

Dataset: We use the publicly available M5Product real world e-commerce dataset [36] which contains 6,313,067 samples of multi-modal information (images, text, table, video, and audio) consisting of 6,232 product categories and 5,679 attributes such as **appearance, usage, specification, selling point, production, material** and **category descriptions** for products which include clothes, cosmetics and instruments.

Training: The model is trained using a combination of positive and negative samples. The training objective is to minimize the ranking loss function, which measures the difference between predicted scores and true interactions.

Metrics: We use the Recall (R) and Normalized Discounted Cumulative Gain (NDCG) metrics to evaluate the recommendation performance. A scoring function is applied to the recommendations for each item. The model returns Top-N items with the highest scores which are recommended to the user. We adopt the GoT ranking model to evaluate the recommendation scores [48].

Baselines: We compare our approach to the following methods: **P5**: A unified "Pretrain, Personalized Prompt & Predict Paradigm" which learns related recommendation tasks through a unified sequence-to-sequence framework [43] as well as **Macaw-LLM**: A multi-modal model which integrates image, video, audio, and text information features into the LLM input sequence [45]. We also include **CHATRec**: A ChatGPT-augmented model for conversational recommendation which uses prompts based on historical user profiles and interactions [4]. In addition, **CLIP**: A Contrastive Language-Image Pre-Training model which jointly learns image-text pairs and can be instructed in natural language to predict the most relevant text snippets, given an image [46].

5 Results and Discussion

Performance Evaluation (RQ1): From the recommendation performance comparison results in Table 1, multi-modal recommendation models Macaw-LLM and MMCRec achieve better performance compared to the uni-modal approaches P5 and CHATRec, which shows the effectiveness of incorporating multi-modal inference into the CRS. The CLIP model, which combines text and image input, performs better than the unimodal approaches, but worse than MMCRec and Macaw-LLM approaches which combine more modalities. Our model, which incorporates graph-based reasoning into the LLM, outperforms the other approaches. This is mainly attributed to the additional context into the recommendation scenarios.

Model	Modality	R@1	R@5	R@10	NDCG@1	NDCG@5	NDCG@10
P5	T	0.398	0.286	0.384	0.312	0.249	0.298
CHATRec	T	0.387	0.415	0.482	0.423	0.344	0.465
CLIP	T+I	0.486	0.547	0.576	0.408	0.442	0.451
Macaw-LLM	V+A+I+T	0.545	0.597	0.628	0.496	0.537	0.522
MMCRec	T+V+A+I	**0.625***	**0.688***	**0.734***	**0.584***	**0.607***	**0.692***

Table 1. Comparing the recommendation performance against baselines. Bold indicates the performance improvement against the second-best baseline at 0.05 significance.

Effect of the input module components (RQ2): We study the effect of each multi-modal input modules to our model performance. Experimental results conducted by deactivating the model components indicated in Table 2 show degraded performance, which is consistent with the earlier observation that the incorporation of the multi-modal components provides additional context to the model, resulting in better recommendation performance. In these experiments, the MMCRec model outperforms the variants in which the audio, video and image input modules are removed, which indicates their positive influence on the overall recommendation performance.

Model	R@1	R@5	NDCG@1	NDCG@5
MMCRec	**0.625**	**0.688**	**0.584**	**0.607**
w/o Audio	0.581	0.640	0.552	0.56
w/o Video	0.38	0.471	0.340	0.459
w/o Image	0.32	0.356	0.307	0.338

Table 2. Comparing the influence of module components. Removal of the audio, image and video input modules from the MMCRec model shows degraded performance.

6 Ethical Considerations

As the development of multi-modal CRSs involves the use of data forms such as images, audio and videos, several ethical issues need to be addressed. Paramount among these is user data privacy and security to ensure that personal information is not misused or exposed without consent. To achieve this, robust security measures should be implemented to protect user data from unauthorized access or breaches. Informed user consent should be emphasized before data collection or usage and users should be aware of how their data is being used and have the option to opt in or out. CRSs should be inclusive, for instance, provide alternative interaction modalities to accommodate different users with diverse abilities and needs. Content moderation mechanisms should be implemented to filter out inappropriate or harmful content. Accountability mechanisms are essential to ensure compliance with relevant data privacy and protection regulations such as the General Data Protection Regulation (GDPR) [7] and other applicable laws. In conclusion, collaboration with stakeholders which include users, researchers, and policymakers is necessary to address any ethical lapses or unintended consequences.

7 Conclusion

In this work, we propose a multi-modal GAI-enhanced Conversational Recommender System MMCRec. By integration of an LLM module with multi-modal encoders and diffusion decoders, MMCRec is capable of accepting GNN enhanced multi-modal input and generating multi-modal output in a combination of text, audio, video and images based on the input request. This approach offers users a more dynamic and engaging way to explore content and products tailored to their preferences and contexts. In future we plan to extend our model to include other modalities such as tabular and web data as well as take into account user satisfaction metrics.

8 Acknowledgements

This publication has emanated from research supported by Science Foundation Ireland under Grant number 18/CRT/6183.

[7] https://gdpr-info.eu/.

References

1. Wu, S., Fei, H., Qu, L., Ji, W. & Chua, T. NExT-GPT: Any-to-Any Multimodal LLM. *ArXiv Preprint ArXiv:2309.05519*. (2023)
2. Cui, Z., Ma, J., Zhou, C., Zhou, J. & Yang, H. M6-rec: Generative pretrained language models are open-ended recommender systems. *ArXiv Preprint ArXiv:2205.08084*. (2022)
3. Hou, Y., Zhang, J., Lin, Z., Lu, H., Xie, R., McAuley, J. & Zhao, W. Large Language Models are Zero-Shot Rankers for Recommender Systems. *ArXiv Preprint ArXiv:2305.08845*. (2023)
4. Gao, Y., Sheng, T., Xiang, Y., Xiong, Y., Wang, H. & Zhang, J. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *ArXiv Preprint ArXiv:2303.14524*. (2023)
5. Salah, A., Truong, Q. & Lauw, H. Cornac: A comparative framework for multimodal recommender systems. *The Journal Of Machine Learning Research*. **21**, 3803-3807 (2020)
6. Liu, Q., Hu, J., Xiao, Y., Gao, J. & Zhao, X. Multimodal Recommender Systems: A Survey. *ArXiv Preprint ArXiv:2302.03883*. (2023)
7. Chen, X., Chen, H., Xu, H., Zhang, Y., Cao, Y., Qin, Z. & Zha, H. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. *Proceedings Of The 42nd International ACM SIGIR Conference On Research And Development In Information Retrieval*. pp. 765-774 (2019)
8. Gu, R., Wang, X. & Yang, Q. Multimodal Cross-Attention Graph Network for Desire Detection. *International Conference On Artificial Neural Networks*. pp. 512-523 (2023)
9. Yao, Y., Liu, Z., Lin, Y. & Sun, M. Cross-Modal Representation Learning. *Representation Learning For Natural Language Processing*. pp. 211-240 (2023)
10. Zhu, L., Wang, T., Li, F., Li, J., Zhang, Z. & Shen, H. Cross-Modal Retrieval: A Systematic Review of Methods and Future Directions. *ArXiv Preprint ArXiv:2308.14263*. (2023)
11. Tao, S., Qiu, R., Ping, Y. & Ma, H. Multi-modal knowledge-aware reinforcement learning network for explainable recommendation. *Knowledge-Based Systems*. **227** pp. 107217 (2021)
12. Huang, H., Zheng, O., Wang, D., Yin, J., Wang, Z., Ding, S., Yin, H., Xu, C., Yang, R., Zheng, Q. & Others ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. *International Journal Of Oral Science*. **15**, 29 (2023)
13. Hu, Z., Cai, S., Wang, J. & Zhou, T. Collaborative recommendation model based on multi-modal multi-view attention network: Movie and literature cases. *Applied Soft Computing*. pp. 110518 (2023)
14. Yan, A., He, Z., Li, J., Zhang, T. & McAuley, J. Personalized Showcases: Generating multi-modal explanations for recommendations. *Proceedings Of The 46th International ACM SIGIR Conference On Research And Development In Information Retrieval*. pp. 2251-2255 (2023)
15. Wu, Y., Macdonald, C. & Ounis, I. Goal-Oriented Multi-Modal Interactive Recommendation with Verbal and Non-Verbal Relevance Feedback. *Proceedings Of The 17th ACM Conference On Recommender Systems*. pp. 362-373 (2023)

16. Xin, X., Pimentel, T., Karatzoglou, A., Ren, P., Christakopoulou, K. & Ren, Z. Rethinking reinforcement learning for recommendation: A prompt perspective. *Proceedings Of The 45th International ACM SIGIR Conference On Research And Development In Information Retrieval*. pp. 1347-1357 (2022)
17. Chen, X., Lu, Y., Wang, Y. & Yang, J. CMBF: Cross-modal-based fusion recommendation algorithm. *Sensors*. **21**, 5275 (2021)
18. Friedman, L., Ahuja, S., Allen, D., Tan, T., Sidahmed, H., Long, C., Xie, J., Schubiner, G., Patel, A., Lara, H. & Others Leveraging Large Language Models in Conversational Recommender Systems. *ArXiv Preprint ArXiv:2305.07961*. (2023)
19. Hou, Y., Zhang, J., Lin, Z., Lu, H., Xie, R., McAuley, J. & Zhao, W. Large language models are zero-shot rankers for recommender systems. *ArXiv Preprint ArXiv:2305.08845*. (2023)
20. Dai, S., Shao, N., Zhao, H., Yu, W., Si, Z., Xu, C., Sun, Z., Zhang, X. & Xu, J. Uncovering ChatGPT's Capabilities in Recommender Systems. *ArXiv Preprint ArXiv:2305.02182*. (2023)
21. Bao, K., Zhang, J., Zhang, Y., Wang, W., Feng, F. & He, X. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. *ArXiv Preprint ArXiv:2305.00447*. (2023)
22. Wang, W., Lin, X., Feng, F., He, X. & Chua, T. Generative recommendation: Towards next-generation recommender paradigm. *ArXiv Preprint ArXiv:2304.03516*. (2023)
23. Li, J., Zhang, W., Wang, T., Xiong, G., Lu, A. & Medioni, G. GPT4Rec: A generative framework for personalized recommendation and user interests interpretation. *ArXiv Preprint ArXiv:2304.03879*. (2023)
24. Wang, X., Tang, X., Zhao, W., Wang, J. & Wen, J. Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models. *ArXiv Preprint ArXiv:2305.13112*. (2023)
25. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K., Joulin, A. & Misra, I. Imagebind: One embedding space to bind them all. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 15180-15190 (2023)
26. Chiang, W., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. & Others Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See Https://vicuna. Lmsys. Org (accessed 14 April 2023)*. (2023)
27. Su, Y., Lan, T., Li, H., Xu, J., Wang, Y. & Cai, D. Pandagpt: One model to instruction-follow them all. *ArXiv Preprint ArXiv:2305.16355*. (2023)
28. Wang, X. & Qin, J. Multimodal recommendation algorithm based on Dempster-Shafer evidence theory. *Multimedia Tools And Applications*. pp. 1-16 (2023)
29. Luo, L., Ju, J., Xiong, B., Li, Y., Haffari, G. & Pan, S. ChatRule: Mining Logical Rules with Large Language Models for Knowledge Graph Reasoning. *ArXiv Preprint ArXiv:2309.01538*. (2023)
30. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X. & Others MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *ArXiv Preprint ArXiv:2306.13394*. (2023)
31. Wu, Y., Liao, L., Zhang, G., Lei, W., Zhao, G., Qian, X. & Chua, T. State graph reasoning for multimodal conversational recommendation. *IEEE Transactions On Multimedia*. (2022)
32. Jannach, D., Manzoor, A., Cai, W. & Chen, L. A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)*. **54**, 1-36 (2021)

33. Liao, L., Long, L., Zhang, Z., Huang, M. & Chua, T. MMConv: an environment for multimodal conversational search across multiple domains. *Proceedings Of The 44th International ACM SIGIR Conference On Research And Development In Information Retrieval*. pp. 675-684 (2021)
34. Viswanathan, S., Guillot, F., Chang, M., Grasso, A. & Renders, J. Addressing Hiccups in Conversations with Recommender Systems. *Designing Interactive Systems Conference*. pp. 1243-1259 (2022)
35. Viswanathan, S., Guillot, F. & Grasso, A. What is Natural? Challenges and Opportunities for Conversational Recommender Systems. *Proceedings Of The 2nd Conference On Conversational User Interfaces*. pp. 1-4 (2020)
36. Dong, X., Zhan, X., Wu, Y., Wei, Y., Kampffmeyer, M., Wei, X., Lu, M., Wang, Y. & Liang, X. M5product: Self-harmonized contrastive learning for e-commercial multi-modal pretraining. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 21252-21262 (2022)
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 10684-10695 (2022)
38. Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W. & Plumbley, M. Audioldm: Text-to-audio generation with latent diffusion models. *ArXiv Preprint ArXiv:2301.12503*. (2023)
39. Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. & Chen, W. Lora: Low-rank adaptation of large language models. *ArXiv Preprint ArXiv:2106.09685*. (2021)
40. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G. & Smola, A. Multimodal chain-of-thought reasoning in language models. *ArXiv Preprint ArXiv:2302.00923*. (2023)
41. Liu, Z., Yu, X., Fang, Y. & Zhang, X. Graphprompt: Unifying pre-training and downstream tasks for graph neural networks. *Proceedings Of The ACM Web Conference 2023*. pp. 417-428 (2023)
42. Wu, L., Zheng, Z., Qiu, Z., Wang, H., Gu, H., Shen, T., Qin, C., Zhu, C., Zhu, H., Liu, Q. & Others A Survey on Large Language Models for Recommendation. *ArXiv Preprint ArXiv:2305.19860*. (2023)
43. Geng, S., Liu, S., Fu, Z., Ge, Y. & Zhang, Y. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). *Proceedings Of The 16th ACM Conference On Recommender Systems*. pp. 299-315 (2022)
44. Lin, J., Dai, X., Xi, Y., Liu, W., Chen, B., Li, X., Zhu, C., Guo, H., Yu, Y., Tang, R. & Others How Can Recommender Systems Benefit from Large Language Models: A Survey. *ArXiv Preprint ArXiv:2306.05817*. (2023)
45. Lyu, C., Wu, M., Wang, L., Huang, X., Liu, B., Du, Z., Shi, S. & Tu, Z. Macaw-LLM: Multi-Modal Language Modeling with Image, Audio, Video, and Text Integration. *ArXiv Preprint ArXiv:2306.09093*. (2023)
46. Radford, A., Kim, J., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. & Others Learning transferable visual models from natural language supervision. *International Conference On Machine Learning*. pp. 8748-8763 (2021)
47. He, M., Wang, J., Ding, T. & Shen, T. Conversation and recommendation: knowledge-enhanced personalized dialog system. *Knowledge And Information Systems*. **65**, 261-279 (2023)

48. Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Gianinazzi, L., Gajda, J., Lehmann, T., Podstawski, M., Niewiadomski, H., Nyczyk, P. & Others Graph of thoughts: Solving elaborate problems with large language models. *ArXiv Preprint ArXiv:2308.09687*. (2023)