Mixture of Semantic Modalities for Recommendation

Kangrui Yu
ky2390@nyu.edu

Ruokai Gu
rg4014@nyu.edu

Mingke Zhou
mz2995@nyu.edu

Siqi Xu
sx2492@nyu.edu

Abstract

Content metadata is important in movie recommender systems as it serves as valuable side information about various modalities of the movie. Recently, Large Language Models (LLMs) are widely used to generate such metadata, and further aggregate such metadata according to the user preference description. However, existing methods typically merge multiple metadata modalities into a single natural language description, overlooking the explicit modeling of how different users attend to different modalities. In this work, we propose a Mixture of Semantic Modalities for Recommendation (MSMRec) model, to acquire accurate capture of user-specific preferences across different modalities. We utilize retrieval-augmented generation to enhance the reliability of LLM-generated movie metadata, mitigating hallucination issues. The metadata generated is segmented into the natural language description of clearly defined modalities, including plot summaries, cast details, poster visuals, release information, audio tone descriptions… Each modality is encoded into embedding using a pretrained text embedding model. We employ the Mixture of Modality Experts (MoE) model to capture user preference on modalities. Furthermore, we provide a comprehensive analysis of how user preferences across these modalities evolve. Our code is publicly available at github.com/ukangrui/NLU-Final-Project.

1 Introduction

Movie Recommender systems play a crucial role in navigating vast item sets. Traditionally, movie recommendations have relied on structured metadata like genre and cast. However, the emergence of Large Language Models (LLMs) has enabled the generation of detailed natural language descriptions that contain multiple aspects of movies, from plot summaries and cast details to visual and audio elements(Zhou et al., 2023)(Zhao et al., 2024).

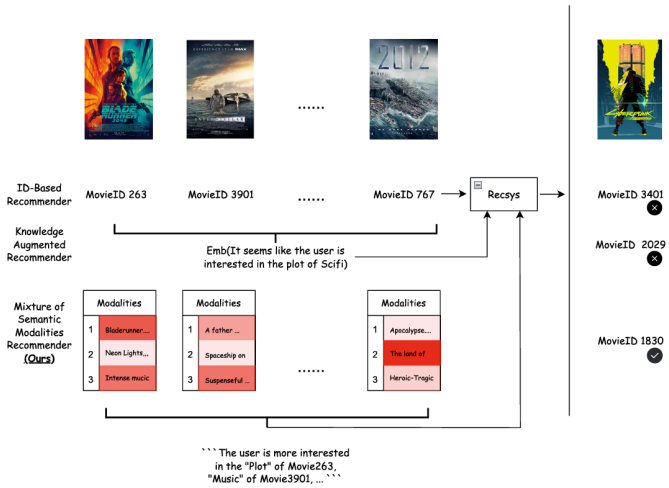

Figure 1: High-level MSMRec pipeline: RAG-enhanced metadata generation, per-modality embeddings, MoE routing, and temporal modality analysis.

Many current approaches(Xi et al., 2023), However, merging these different modalities into one narrative overlooks the fact that users often prioritize certain aspects over others.

For example, plot and character development may be the main reason one user watches a movie, while other users may lean towards visual or auditory features. Additionally, many current approaches overlook the hallucination issues inherent in LLMs. These hallucinations can introduce inaccuracies or misleading information into the generated metadata, adversely influencing the recommendation task.

The main contributions of this work addressing this gap can be summarized as:

- We employ retrieval-augmented generation techniques to produce reliable and objective movie metadata, thereby mitigating hallucination issues commonly associated with LLM

outputs.

- By segmenting the generated metadata into clearly defined modalities (e.g., plot, cast, visuals, release details, audio tone...), with each modality encoded using pretrained text embedding models, we allow for explicit modeling user preference on each modality.
- We introduce a novel Mixture of Modality Experts framework that models the interactions between user preferences and movie modality to enhance recommendation accuracy.
- We perform comprehensive analyses and visualizations of modality utilization, offering insights into the impact of different encoding methods and revealing how shifts in user interests correlate with changes in modality utilization.

2 Related Works

2.1 LLM Metadata Integration

Recent studies have shown that Large Language Models can generate rich textual descriptions that capture diverse aspects of movies, including plot summaries, cast details, and even visual or audio features(Acharya et al., 2023). In addition, research on multi-modal recommender systems highlights the benefits of integrating various metadata modalities to better reflect user preferences(Liu et al., 2024). However, many existing methods fuse these modalities into a single narrative, which limits the ability to model the differential attention that users give to specific aspects of movie content.

2.2 Retrieval-Augmented Generation & Hallucination Mitigation

Retrieval-augmented generation techniques have been developed to improve the reliability of LLM outputs by grounding them in verifiable external data(Gao et al., 2024). This approach addresses the hallucination issues by ensuring that the generated metadata accurately reflects factual movie information(Yehudai et al., 2024). Despite these improvements, the application of such techniques to enhance recommendation systems by preserving modality-specific information remains underexplored.

2.3 MoE for User Preferences

The Mixture-of-Experts (MoE) framework is crucial for capturing complex data subdistributions by allowing specialized models to focus on separate data splits.(Bian et al., 2023). In the context of movie recommendation, MoE-based approaches can assign distinct weights to various modalities—such as plot, visuals, and audio—thereby providing a more nuanced and interpretable modeling of user interests(Qin et al., 2020). This approach contrasts with traditional methods that rely on uniform fusion of all modalities, and it lays the foundation for improved recommendation accuracy by aligning recommendations with the specific aspects valued by individual users.

3 Methodology

In this section, we detail the pipeline of the proposed Mixture of Semantic Modalities for Recommendation (MSMRec) model. The methodology comprises four major components: (i) retrieval-augmented metadata generation and modality segmentation, (ii) modality-specific semantic embedding, (iii) a Mixture-of-Experts (MoE) framework for user preference modeling, and (iv) temporal dynamics for capturing evolving user interests.

3.1 Retrieval-Augmented Metadata Generation and Modality Segmentation

For each movie i, let x_i denote its title. We first retrieve relevant context $R(x_i)$ from an external knowledge base. This context is then input to a Large Language Model (LLM) to generate detailed movie metadata:

$$M_i = \text{LLM}\big(R(x_i); \theta_{\text{LLM}}\big), \tag{1}$$

where θ_{LLM} represents the model parameters. The generated metadata M_i is segmented into K distinct modalities:

$$M_i = \{m_{i,1}, m_{i,2}, \ldots, m_{i,K}\}, \tag{2}$$

with each $m_{i,k}$ corresponding to a modality (e.g., plot, cast, visuals, release details, audio tone).

3.2 Modality-Specific Semantic Embedding

Each segmented modality $m_{i,k}$ is encoded using a pretrained text embedding model $E(\cdot)$ to obtain a fixed-length semantic vector:

$$\mathbf{e}_{i,k} = E\big(m_{i,k}\big) \in R^d, \tag{3}$$

The set $\{\mathbf{e}_{i,1}, \mathbf{e}_{i,2}, \ldots, \mathbf{e}_{i,K}\}$ represents the semantic embeddings of movie i across different modalities.

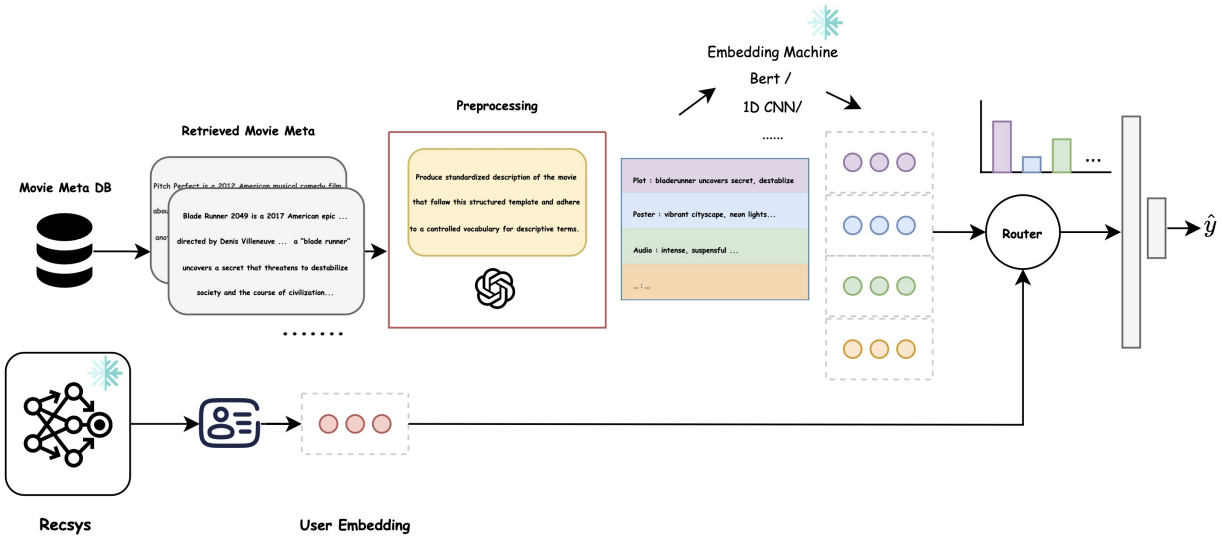

Figure 2: Visualization of MSMRec Model

3.3 Mixture-of-Experts Framework for User Preference Modeling

Let $\mathbf{u_t} \in R^d$ denote the latent representation of a user at timestep t

$$u_t = AVGPOOL(e_1, e_2, \dots e_t)$$

For each modality k of item i, we compute a modality-specific compatibility score:

$$s_{i,k} = \langle \mathbf{u_t}, \mathbf{e}_{i,k} \rangle, \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes a learnable compatibility function. We use *softargmax* activation to assign a normalized weight $\alpha_{i,k}$ to each modality:

$$\alpha_{i,k} = \frac{\exp\left(\langle \mathbf{u}, \mathbf{e}_{i,k} \rangle\right)}{\sum_{j=1}^{K} \exp\left(\langle \mathbf{u}, \mathbf{e}_{i,j} \rangle\right)} \tag{5}$$

The item aggregated modality representation is then added to its original id-based representation as the final representation of item i for user t.

$$\hat{e}_i = e_i + \sum_{k=1}^{K} \alpha_{i,k}\, e_{i,k} \tag{6}$$

The item aggregated modality representation $\hat{e}_i$ is fed forward into a recommendation netwoek (SASRec) for pointwise CTR prediction.

4 Experiment

1. **RQ1:** How effective is the retrieval-augmented generation in mitigating LLM hallucinations and enhancing metadata quality for recommendations?
2. **RQ2:** Does the proposed MSMRec model achieve better sequential recommendation results compared to baseline methods?
3. **RQ3:** What is the impact of different embedding models and the use of RAG, does it affect the performance of MSMRec?
4. **RQ4:** How do different groups of users utilize each modality, and how does the user's utilization of modality evolve?

4.1 Setup

We conduct experiments on two widely-used MovieLens datasets: MovieLens 100K and MovieLens 1M.(Harper and Konstan, 2015)

Dataset	#Users	#Items	#Interactions	#Avg.Seq	Sparsity
MovieLens 100K	943	1682	100,000	106.04	93.70%
MovieLens 1M	6040	3407	1,000,000	165.56	95.14%

Table 1: Dataset statistics for MovieLens 100K and MovieLens 1M.

4.2 Baseline

For our experiments, we compare our proposed MSMRec model, which integrates modality segmentation and a Mixture-of-Experts (MoE) framework into the SASRec architecture, against the following baselines:

- **SASRec** A sequential recommendation model based on self-attention, serving as a standard benchmark.(Kang and McAuley, 2018)

- **SASRec+KAR** A variant that incorporates Knowledge Augmented Recommendation (KAR) into SASRec to enhance metadata quality.(Xi et al., 2023)

4.3 Evaluation Metrics and Training Details

We evaluate the models using standard metrics for sequential recommendation tasks such as Hit Rate (HR@K) and Normalized Discounted Cumulative Gain (NDCG@K) for various values of K. We also assess the modality generated by LLMs via human evaluation.

5 Experiments

5.1 RQ1

How does RAG enhance semantic quality

We prompted GPT-3.5-Turbo, via in-context learning, to generate descriptive text for each movie modality. Table 2 presents an example of a GPT-generated modality description. We then compared the generated descriptions with and without employing retrieval-augmented generation (RAG) with the IMDB Dataset. Our findings indicate that incorporating RAG results in fewer hallucinations and greater accuracy. Additionally, we observed that GPT-3.5-Turbo produces higher-quality descriptions for newer movies (post-1970s) and for films with distinctive titles that serve as clear identifiers (e.g., "Escape to Witch Mountains" vs. "Kim"). We also observed that in the T-SNE visualizations of modality and aspect, some modalities are more clustered among genres, such as "Narrative", "Direction" and "Acting", as we assume in later experiments, these modalities will serve as valuable side information for the backbone recommender system. On the other hand, although the visualization of some modality, such as "Dialogue" seems comparatively noisy, this could also be attributed to the ambiguity of the segmentation of genres, as a movie with the genre "Action" is highly probable to also be related with genre "Adventure".

5.2 RQ2

How does MSMRec improve recommendation performance

Modality	Description
Narrative	Heartfelt toy adventure centered on friendship and growth.
Dialogue	Clever, playful dialogue that resonates with both kids and adults.
Visual	Vibrant, imaginative animation with expressive, dynamic camera work.
Set	Richly detailed toy-world environments with creative set design.
Audio	Upbeat musical score and whimsical sound effects heighten emotions.
Pace	Smooth transitions and energetic pacing drive the narrative.
Direction	Pixar's innovative storytelling direction with heart.
Acting	Expressive voice acting that brings charm to every character.
Poster	Iconic posters and trailers reflect the film's playful essence.

Table 2: Semantic Modalities of Toy Story (1995)

Movie / Modality	gpt-3.5	gpt-3.5-RAG
Heat (1995) / Narrative	Intense crime drama ... between ... and a dedicated detective.	Intense cat-and-mouse ... between ... and a relentless detective
Escape to Witch Mountain (1975) / Dialogue	Engaging dialogue that blends intrigue and heartwarming moments.	Exchanges that convey the twins' unique abilities and the challenges they face.
Kim (1950) / Poster	Vintage posters that evoke a sense of nostalgia.	Evocative posters that hint at the adventure elements of the story.

Table 3: Comparative Analysis of Modalities Generated by GPT-3.5 with and without RAG

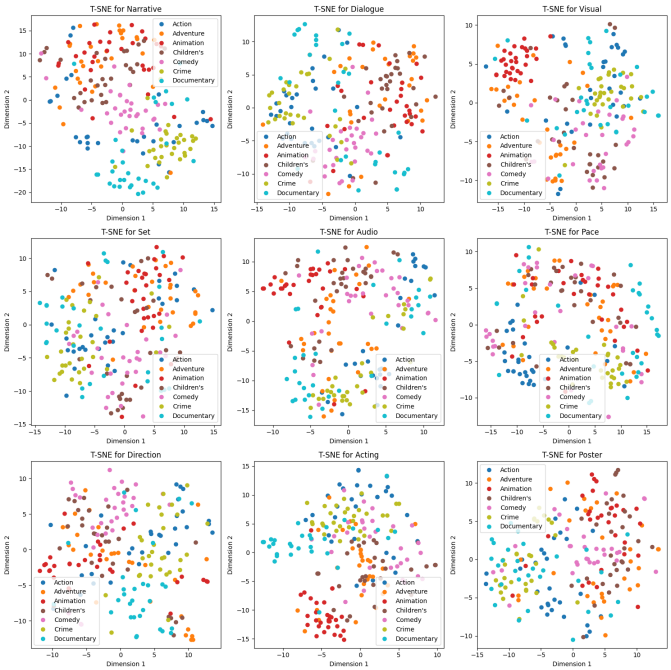

Figure 3: T-SNE Visualization of Modality and Aspect

Model	Dataset	HR@100	HR@10	NDCG@10
SASRec	ML-1M	0.6263	0.2351	0.1211
	ML-100K	0.6341	0.1463	0.0670
SASRec + KAR	ML-1M	0.6311	0.2389	0.1227
	ML-100K	**0.6421**	**0.1497**	0.0691
MSMRec (**Ours**)	ML-1M	**0.6365**	**0.2444**	**0.1282**
	ML-100K	0.6377	0.1479	**0.0701**

Table 4: Comparison of recommendation metrics on MovieLens 1M and MovieLens 100K.

Parameter	Value
Modality	
num_modality	9
llm_model	"gpt-4o-mini"
embedding_model	"text-embedding-ada-002"
use_RAG	True
SASRec	
sequence len	200
num_attention_blocks	2
hidden_units	50
lr	1e-3
weight_decay	0
optimizer	Adam
num_epochs	1000

Table 5: Hyperparameters for Modality and SASRec

The experiment results, as shown in Table 4, indicate that MSMRec outperforms KAR on larger datasets such as ML-1M, and is comparable on smaller datasets such as ML-100K. MSMRec generates a substantial 1.6% increase in Hit Rate @ 100 on ML-1M.

5.3 RQ3

How does embedding method impact recommendation performance

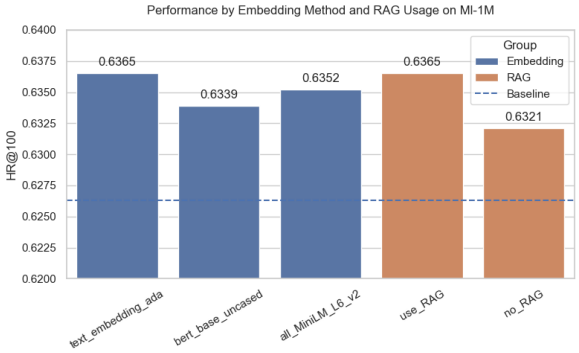

Figure 4: Performance comparison among different embedding models and if use RAG

To analyze the impact of different embedding methods and RAG, we run the experiment again varying those arguments. The experiment result is shown in Figure 4. We observe that RAG substantially improves the performance on ML-1M, as the model tends to follow a more structured template and output less hallucination. In addition, we observe that encoding text with text-embedding-ada-002 yields better results over bert and sentence transformer models. This could be attributed to the better representation power from the higher embedding dimension from text-embedding-ada-002 (1536).

5.4 RQ4

How does different user subgroup utilize modalities

We split users into different subgroups based on their tendency to consume popular items. We first generate the set of popular items:

$$\{I_{pop}\} = \{i \in I | \frac{COUNT(interaction(\cdot, i))}{COUNT(interaction(\cdot, \cdot))} > \lambda\}$$

Then, we calculate the proportion of popular items in the user's interaction history, sort in descending orders, and cut into equal-width bins of Blockbuster, Diverse, and Niche subgroups. We then analyze the pattern of the utilization of modalities within and across subgroups. The analysis results, as shown in Figure 5 and Figure 6, exhibit

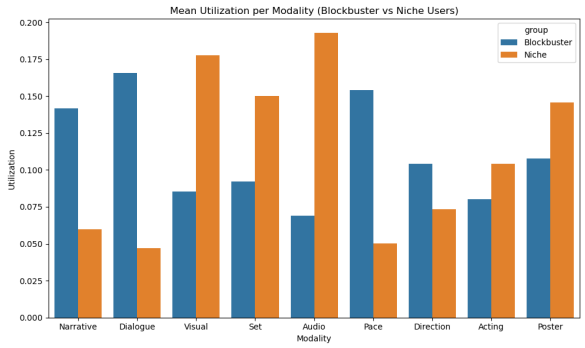

Figure 5: Normalized modality weights for Blockbuster and Niche user groups

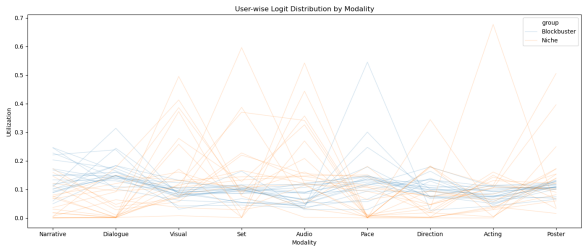

Figure 6: Individual user modality-weight curves within each subgroup

that across Blockbuster and Niche user groups, the modalities utilization are dramatically different. We observe that Blockbuster Users (i.e. Users that tend to consume more popular items(movies)) attend to mostly the "Visual", "Poster", "Audio", and "Poster" aspect of a movie, whereas Niche users mostly attend to "Narrative", "Dialogue", and "Pace" of a movie. This corresponds with the intuitive interpretation that popular movies more often come with a more eye-catching visual and poster design, whereas niche movies focus more on storytelling. In addition, as shown in Figure 6, we also analyze the modality utilization for each user (as shown as a line in the plot). We observe that on one hand, the use of modality is diverse across different users. Some users have a more uniform utilization of modalities while some have a more skewed utilization of modalities. Nonetheless, we observe that trend shown in Figure 5 is preserved when we analyze each user's behavior, Blockbuster users still attend mostly to "Visual", "Poster", "Audio", and "Poster" and Niche users to "Narrative", "Dialogue", and "Pace". We provide an example of User

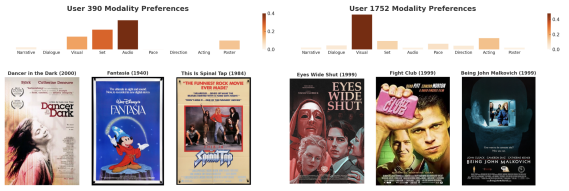

Figure 7: Modality-weight profiles for User 390 (audio-centric) versus User 1752 (visual-centric) on their top films.

390 and User 1752 leveraging different modalities as shown in Figure 7. User390 who mostly utilized the "Audio" modality watched "Dancer in the Dark (2000)", "Fantasia(1940)" and "This is Spinal Tap (1984)", which these movies are indeed strongly rooted in their audio design and unique set. In comparison, User1752 who mostly utilized the "Visual" modality watched "Eyes Wide Shut(1999)", "Fight Club(1999)", and "Being John Malkovich(1999)", where these movies, in the genre of horror, are indeed focused on the visual punch. We, therefore, conclude that our model, MSMRec, successfully captured different users' different interests in different modalities of a given movie.

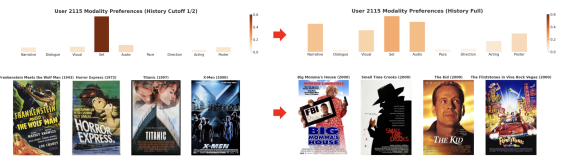

Figure 8: Modality weights for User 2115 in the first vs. second half of their history, illustrating a shift from predominantly "Set" to a mix of "Set," "Narrative," and "Audio."

How does the user's utilization of modality evolve?

We conducted a qualitative analysis to examine the effect of our model capturing the user's shift of utilization pattern. We calculate the user's utilization mode with the history cut off at half, and calculate again with full history. We find that our model well captured the user's interest shift. As shown in Figure 8, User 2115 was primarily utilizing the "Set" modality in the first half of history, which is reflected in its movie history consisting of drama, and horror films. We notice that user 2115 then shifted its interest to family and comedy shows in the second half of the watch history, utilizing a combination of "Set", "Narrative" and "Audio".

6 Limitations

6.1 Scalability

Prompting large language models to generate detailed, modality-specific text for every item in a large catalog introduces significant computational and monetary overhead. Each RAG call requires retrieving external knowledge and executing one or more LLM queries per modality, and then embedding the resulting text with a high-dimensional encoder (e.g., text-embedding-ada-002). In a real-world deployment with millions of items, this pipeline can become prohibitively slow and expensive, particularly if embeddings must be refreshed periodically to account for new content or changing external knowledge. Moreover, the storage and indexing of multi-modality embeddings at scale may pose additional engineering challenges in terms of data throughput and memory footprint.

6.2 Stability Across Datasets

While MSMRec consistently outperforms SASRec+KAR on the relatively large MovieLens-1M dataset, its gains diminish or even reverse on smaller, sparser collections such as MovieLens-

100K. This variability suggests that the effectiveness of explicit modality segmentation and the Mixture-of-Experts routing depends on having sufficient interaction data to learn reliable modality weights. In low-data regimes, the model may overfit to noisy or under-represented modalities, and the relative benefit of RAG-enhanced metadata can be overshadowed by sparse user histories. Consequently, MSMRec may require careful tuning of regularization, expert capacity, or modality grouping strategies when applied to datasets with limited samples.

7 Future Work

7.1 TopK Gating for Mixture of Modalities

Our experiments indicate that the top four modalities contribute over 84% of the total modality weight in aggregate. In future work, we plan to replace the current softargmax over all modalities with a TopK gating mechanism that selects and renormalizes only the highest-scoring subsets of modalities. This modification is expected to reduce computational overhead and limit noise from less relevant modalities, leading to more efficient training and inference. We will compare static versus dynamic TopK selection strategies and measure their impact on convergence speed, memory usage, and recommendation accuracy.

7.2 Fusion of Modality and ID beyond Inner Product

At present, we compute user–modality compatibility via a simple inner product between the user representation and each modality embedding. Going forward, we intend to explore more expressive fusion techniques, such as multi-headed attention layers that learn cross-modal interactions or small feed-forward networks that take the concatenated user, modality, and item ID embeddings as input. We are also interested in lightweight LLM-based scoring approaches that could leverage reasoning chains to capture intricate user preferences. These alternatives will be evaluated in terms of predictive performance, interpretability, and computational cost.

7.3 Robustness of Generated Semantic Text

While our retrieval-augmented generation pipeline reduces hallucinations relative to vanilla LLM outputs, the consistency and representativeness of the generated modality descriptions have not yet been rigorously quantified. In future studies, we will conduct large-scale human assessments and automated evaluations (for factual fidelity, diversity, and coverage) across a diverse set of movies. We will also perform stress tests by varying retrieval sources, prompt templates, and LLM versions to uncover potential failure modes. Finally, we will investigate fine-tuning and reinforcement learning approaches to further align generated text with authoritative metadata sources, ensuring reliable downstream embeddings for recommendation.

8 Conclusion

In this work, we introduced MSMRec, a novel recommendation framework that explicitly models user preferences over multiple semantic modalities. By leveraging retrieval-augmented generation to produce reliable, modality-specific text descriptions and encoding them with pretrained embedding models, we obtain rich semantic vectors for the plot, cast, visuals, release details, audio tone, and more. A Mixture-of-Experts routing mechanism then learns personalized weights over these modalities, allowing us to capture the diverse ways in which users value different aspects of movie content. Extensive experiments on MovieLens-1M and MovieLens-100K demonstrate the effectiveness of our approach. MSMRec achieves a 1.6 % absolute improvement in Hit Rate @ 100 on ML-1M compared to SASRec+KAR and remains competitive on ML-100K, while our ablation studies confirm the benefits of RAG and high-dimensional embeddings. Moreover, our temporal and subgroup analyses provide interpretable insights into how modality utilization varies across user segments and evolves over time.

We also highlighted key limitations—namely scalability concerns when generating and embedding metadata at large scale, and reduced stability on smaller, sparser datasets—and outlined concrete future directions, including TopK gating for efficient modality selection, richer fusion methods beyond inner products, and rigorous robustness evaluation of generated semantic text. We believe that modality-aware recommendation, when combined with efficient architectures and robust metadata generation, holds great promise for more personalized and interpretable recommender systems.

9 Author Contribution

- Kangrui Yu: Conceived the project, implemented the MSMRec model, and drafted the final report.
- Ruokai Gu: Designed and ran all experiments on the MovieLens-1M and MovieLens-100K datasets.
- Mingke Zhou: Performed the modality usage analyses and generated the associated visualizations.
- Siqi Xu: Implemented and validated the SASRec+KAR baseline integration.

References

Arkadeep Acharya, Brijraj Singh, and Naoyuki Onoe. 2023. Llm based generation of item-description for recommendation system. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 1204–1207, New York, NY, USA. Association for Computing Machinery.

Shuqing Bian, Xingyu Pan, Wayne Xin Zhao, Jinpeng Wang, Chuyuan Wang, and Ji-Rong Wen. 2023. Multi-modal mixture of experts represetation learning for sequential recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 110–119, New York, NY, USA. Association for Computing Machinery.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.

F. Maxwell Harper and Joseph A. Konstan. 2015. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4).

Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation.

Qidong Liu, Jiaxi Hu, Yutian Xiao, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Qing Li, and Jiliang Tang. 2024. Multimodal recommender systems: A survey.

Zhen Qin, Yicheng Cheng, Zhe Zhao, Zhe Chen, Donald Metzler, and Jingzheng Qin. 2020. Multitask mixture of sequential experts for user activity streams. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3083–3091, New York, NY, USA. Association for Computing Machinery.

Yunjia Xi, Weiwen Liu, Jianghao Lin, Xiaoling Cai, Hong Zhu, Jieming Zhu, Bo Chen, Ruiming Tang, Weinan Zhang, Rui Zhang, and Yong Yu. 2023. Towards open-world recommendation with knowledge augmentation from large language models.

Asaf Yehudai, Boaz Carmeli, Yosi Mass, Ofir Arviv, Nathaniel Mills, Assaf Toledo, Eyal Shnarch, and Leshem Choshen. 2024. Genie: Achieving human parity in content-grounded datasets generation.

Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li. 2024. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6889–6907.

Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen. 2023. A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions.

Appendix

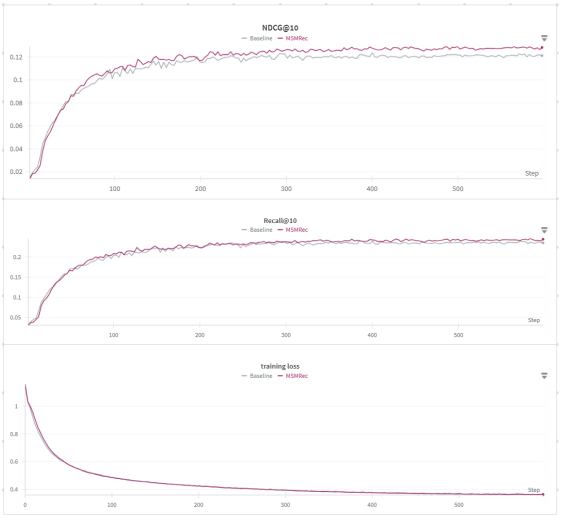

Figure 9: Training Dynamics on ML-1M