Query-Aware Multi-Modal Recommendation with Graph Neural Networks

Joseph Li, Christopher Stanulet, Bryan Pineda
Department of Computer Science, Stanford

1 Motivation

Traditional recommender systems in the consumer domain often rely on static user profiles, explicit ratings, or simple collaborative filtering, which limits their ability to adapt to dynamic, context-rich user interactions. As users increasingly express preferences through queries and conversations—often containing multi-modal information such as text, dialogue, and images—existing models struggle to capture the full intent and context behind these requests, resulting in less relevant recommendations.

We propose a query-aware, multi-modal GNN-based recommender system that dynamically adapts to the user's current query or conversational context with no prior user history. By integrating text (queries, dialogue), images (movie posters), and metadata within a unified graph framework, our approach enables richer user and item representations and models complex relationships among users, items, and queries. This method promises to deliver more accurate and contextually relevant recommendations, addressing key challenges in cold-start and ambiguous scenarios, and is highly applicable to next-generation streaming platforms, conversational agents, and e-commerce systems.

2 Datasets

We utilize three complementary datasets in our study: MovieLens 32M[2], ReDial[3], and TMDB[5]. The MovieLens 32M dataset is a large-scale, publicly available collection containing 32 million explicit user ratings for movies, along with rich metadata such as titles, genres, and user-generated tags. This dataset provides a robust foundation for modeling user preferences and item characteristics in traditional recommendation scenarios. The ReDial dataset, also public, offers a unique conversational recommendation setting, capturing real user dialogues, social context, and movie mentions. This enables us to model dynamic, query-driven interactions that are increasingly relevant in modern recommender systems.

To incorporate multi-modal information, we leverage the TMDB (The Movie Database) API, which provides access to visual data such as movie posters. By extracting image features from these posters, we can enrich our item representations and explore the impact of visual signals on recommendation quality. The combination of these datasets allows us to comprehensively evaluate our query-aware, multi-modal GNN approach across explicit ratings, conversational context, and visual modalities, ensuring our model is both robust and applicable to real-world, interactive recommendation scenarios.

3 Graph

Our proposed graph (Figure 2) consists of heterogeneous nodes representing users, movies, dialogues (queries), tags, and genres, with edges encoding key relationships such as user-movie interactions (e.g., ratings or mentions), dialogue-movie associations (capturing conversational context), and movie-tag/genre connections (providing semantic metadata). Additionally, movies are linked to their visual representations (posters) to enable multi-modal feature integration. This unified graph structure allows our GNN-based model to effectively capture complex, context-aware, and multi-modal relationships essential for dynamic and accurate movie recommendations.

4 Model

Our proposed model is a query-aware, multi-modal Graph Attention Network (GAT) architecture, as illustrated in Figure 1.The model is designed to address the dynamic and context-rich nature of movie recommendation tasks by leveraging both graph structure and query information. The architecture consists of three main components: a GATEncoder, QueryAwarePooling, and a simplified Fusion and Scorer module.

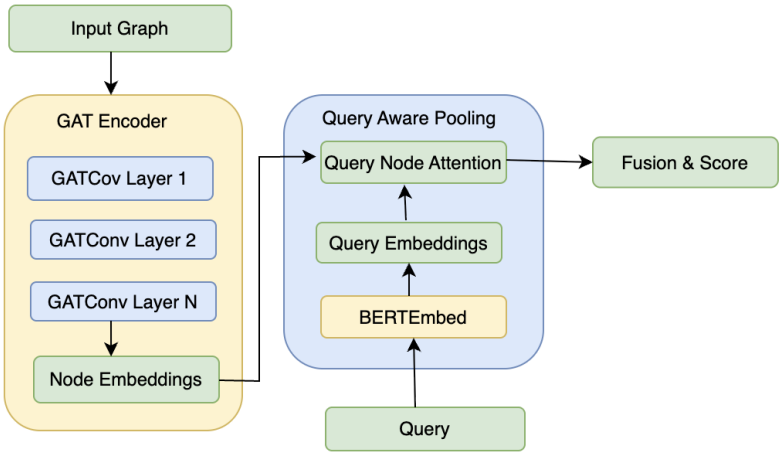

Figure 1: Proposed Model Architecture

4.1 GAT Encoder[4][6]

$$h_i' = \sigma\left(\sum_{j \in \mathcal{N}(i) \cup \{i\}} \alpha_{ij} \mathbf{W} h_j\right) \tag{1}$$

where

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top [\mathbf{W}h_i \,\|\, \mathbf{W}h_j]\right)\right)}{\sum_{k \in \mathcal{N}(i) \cup \{i\}} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top [\mathbf{W}h_i \,\|\, \mathbf{W}h_k]\right)\right)} \tag{2}$$

4.2 Query Aware Pooling[1]

$$g = \sum_{i=1}^{N} \frac{\exp\left(\text{MLP}(h_i, q)\right)}{\sum_{j=1}^{N} \exp\left(\text{MLP}(h_j, q)\right)} h_i \tag{3}$$

4.3 Model Justification

We employ Graph Attention Networks (GATs) for their ability to model heterogeneous graphs and selectively focus on the most relevant nodes and edges via attention mechanisms. This is well-suited to our datasets, which include diverse node types (movies, queries, tags, genres) and require context-aware reasoning. Attention-based pooling enables the model to dynamically adapt recommendations to each query, overcoming the limitations of static, profile-based methods.

4.4 Task and Evaluation Metrics

The primary task is to predict a relevance score for each (query, movie) pair, enabling effective ranking of candidate movies for a given query. We will evaluate our model using Mean Average Precision (MAP), as well as additional metrics such as Recall@K, Precision@K, and Normalized Discounted Cumulative Gain (NDCG), to comprehensively assess the quality of top-ranked recommendations.

References

[1] Vibhor Agrawal, Fay Wang, and Rishi Puri. *Query-Aware Graph Neural Networks for Enhanced Retrieval-Augmented Generation*. arXiv:2508.05647 [cs]. July 2025. DOI: 10.48550/arXiv.2508.05647. URL: http://arxiv.org/abs/2508.05647 (visited on 10/19/2025).

[2] F. Maxwell Harper and Joseph A. Konstan. "The MovieLens Datasets: History and Context". en. In: *ACM Transactions on Interactive Intelligent Systems* 5.4 (Jan. 2016), pp. 1–19. ISSN: 2160-6455, 2160-6463. DOI: 10.1145/2827872. URL: https://dl.acm.org/doi/10.1145/2827872 (visited on 10/19/2025).

[3] Raymond Li et al. "Towards Deep Conversational Recommendations". In: *Advances in Neural Information Processing Systems 31 (NIPS 2018)*. 2018.

[4] Christopher Morris et al. *Weisfeiler and Leman Go Neural: Higher-order Graph Neural Networks*. arXiv:1810.02244 [cs]. Nov. 2021. DOI: 10.48550/arXiv.1810.02244. URL: http://arxiv.org/abs/1810.02244 (visited on 10/19/2025).

[5] The Movie Database (TMDb). *The Movie Database (TMDb)*. https://www.themoviedb.org/. 2025. (Visited on 10/19/2025).

[6] Petar Veličković et al. *Graph Attention Networks*. arXiv:1710.10903 [stat]. Feb. 2018. DOI: 10.48550/arXiv.1710.10903. URL: http://arxiv.org/abs/1710.10903 (visited on 10/19/2025).

A Appendix: Graph

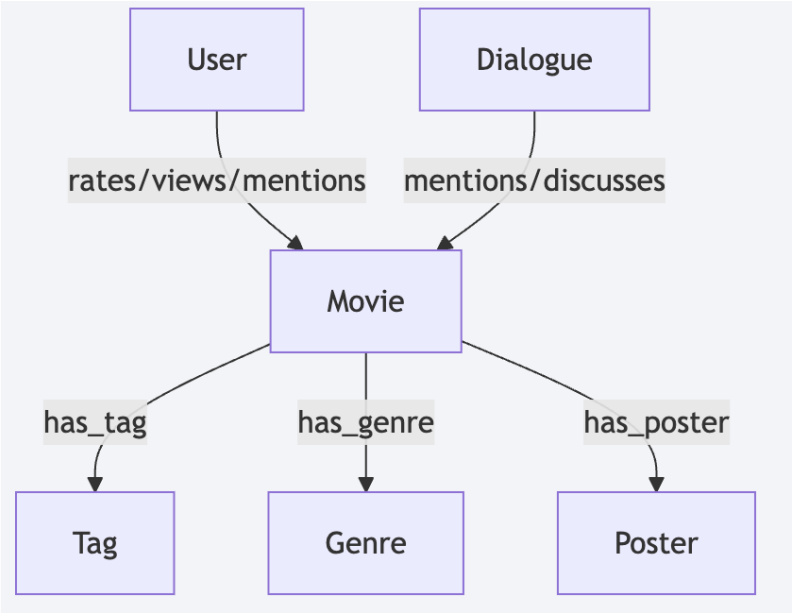

Figure 2: Movie Graph