A Survey on Dynamic Multimodal Recommendation Systems with Hierarchical Attention Mechanisms

Rekha G S, Varnith D Ramesh, Sudarshan Komar, Swapnil Sahil, Vatsal Mural

Department of Computer Science

BMS College of Engineering

Bangalore, India

***Abstract* —In the era of information overload, delivering personalized and contextually relevant recommendations has become increasingly vital. This survey explores the conceptual foundations and recent advancements in hierarchical attention-based multimodal recommendation systems, with a focus on their dynamic adaptability to user behavior in domains such as e-commerce and education. These systems integrate diverse data modalities—such as textual reviews, visual content, and contextual signals—through specialized modality-specific encoders and hierarchical attention mechanisms to more effectively capture user intent and assess item relevance.**

We present a comprehensive review of existing literature, encompassing state-of-the-art multimodal recommendation techniques, fusion strategies, and attention-based architectures. Emphasis is placed on key design considerations including explainability, modularity, and adaptability, which are critical for the development of interpretable and extensible recommendation systems. Although the practical implementation and evaluation of a proposed framework are reserved for future work, this survey provides a structured foundation and valuable insights for researchers and practitioners aiming to design next-generation user-centric recommendation systems.

***Index Terms*—Artificial Intelligence, Credit Assessment, Machine Learning, Natural Language Processing, Voice-based Systems**

I. Introduction

In recent years, recommendation systems have evolved into indispensable components of digital platforms, powering personalized content delivery across domains such as e-commerce, entertainment, education, and social media. These systems play a pivotal role in filtering information overload and enhancing user satisfaction by suggesting products, services, or content that aligns with individual user preferences.

Traditional recommendation systems predominantly relied on collaborative filtering (CF) and content-based filtering (CBF), which typically leverage user-item interaction matrices or item metadata to compute recommendations. However, such approaches often struggle

with challenges like data sparsity, cold-start problems, and a lack of contextual understanding [1], [2]. To address the limitations of traditional collaborative and content-based filtering, the research community has turned to **multimodal recommendation systems (MRS)**. These systems incorporate diverse data types—such as reviews, images, audio, video, and interaction graphs—to construct richer user and item representations [3], [4].The challenge lies in effectively fusing these heterogeneous inputs while preserving their individual semantic contributions.

Recent studies have proposed deep learning-based fusion strategies to solve this integration problem. While these methods show improvements in performance, they often struggle with scalability, cross-domain generalization, and the ability to provide transparent, explainable recommendations.

To tackle these issues, researchers have explored **hierarchical attention mechanisms** and **dynamic modeling** of user behavior. Hierarchical attention models like TMFUN [5] capture fine-grained relevance at the word, sentence, and modality levels, enabling deeper semantic understanding.

This survey aims to synthesize and analyze recent advancements in **hierarchical attention-based multimodal fusion for dynamic recommendation systems**. We reviewed research papers spanning diverse strategies including co-attention, contrastive learning, graph neural networks (GNNs), knowledge graph-enhanced recommendation, and large language model (LLM) integration [6], [7], [8].

II. Background and Related Work

The exponential growth of digital content and services has made personalized recommendation systems a critical part of online platforms. From e-commerce and entertainment to education and healthcare, recommendation algorithms guide user decision-making and influence engagement. Traditional recommender systems predominantly rely on collaborative filtering (CF) or content-based filtering (CBF), leveraging user-item interaction data or item metadata [1], [2]. While effective in structured environments, these models suffer from key limitations such as the cold-start problem, data sparsity, and a lack of semantic understanding—particularly in complex, dynamic settings like e-commerce and user-generated platforms [9], [10].

Recent advancements in Artificial Intelligence (AI), especially in deep learning, have enabled the integration of rich, heterogeneous data sources to improve recommendation quality. Multimodal recommendation systems (MRS) have emerged as powerful alternatives by incorporating text (e.g., reviews, metadata), images (e.g., product visuals), audio, and user-graph interactions [3], [4]. These systems offer more comprehensive user and item representations by capturing semantic features and contextual relationships across modalities. However, challenges remain in effectively fusing diverse data types, mitigating modality imbalance, and ensuring scalability across real-world domains [6], [7].

To address these gaps, a growing body of literature has introduced **hierarchical attention mechanisms** and **adaptive fusion strategies** as key components in next-generation recommender models. Works like TMFUN [5] and DeepIDRS [10] utilize word-level, sentence-level, and modality-level attention to extract the most relevant information from each modality. This layered attention improves both model interpretability and performance. In parallel, dynamic and sequential modeling of user behavior using temporal graphs or LSTM-based architectures has gained traction. Models such as MMSR [11] and MLSASRec [12] learn time-sensitive user patterns to better reflect evolving preferences.

Furthermore, explainability has emerged as a core requirement for practical deployments. As AI models become more complex, users and regulators demand transparency in automated decision-making. Models like HKGAT [13] and the counterfactual neural recommender (CNR) [14] emphasize explainable reasoning paths using attention weights, knowledge graphs, and counterfactual logic. Research also highlights the importance of aligning recommendation decisions with human-centered design principles, such as fairness, trust, and usability [8], [1].

Despite this progress, key barriers remain in multimodal recommendation research: handling missing modalities, generalizing across domains, scaling to industrial workloads, and balancing predictive power with interpretability. This survey consolidates the recent innovations addressing these challenges, evaluates the state of the art in multimodal, attention-based, and explainable recommendation systems, and identifies opportunities for future work in this evolving domain.

III. Materials and Methods

Recent advancements in recommendation systems have focused on integrating multimodal data and attention-based architectures to improve accuracy, robustness, and interpretability. This section categorizes the current state-of-the-art into four major themes: (A) multimodal fusion strategies, (B) hierarchical attention mechanisms, (C) dynamic and sequential modeling of user behavior, and (D) explainability, interpretability, and governance.

A. Multimodal Fusion Strategies

A significant body of research explores how textual reviews, visual data, and graph structures can be integrated to form richer item and user embeddings. For instance, CAMRec [4] utilizes co-attention to combine features from review texts and product images, while MCT [15] applies contrastive learning to align modality-specific representations in a shared embedding space. The Triple Modality Fusion (TMF) framework [6] takes this further by integrating visual, textual, and graph modalities through large language models and attention mechanisms.

These systems are typically evaluated on datasets such as Amazon Reviews, Walmart e-commerce logs, and MovieLens, using metrics like Hit Ratio@K, NDCG@K, AUC, and RMSE. Studies

highlight the effectiveness of multimodal embeddings in improving cold-start performance and enhancing recommendation diversity.

B. Hierarchical Attention Mechanisms

Hierarchical attention models offer layered interpretability by modeling relevance at multiple granularity levels—such as word, sentence, and modality. Zhou et al. [5] propose TMFUN, a hierarchical fusion network that uses attention-guided multi-step fusion across item and modality feature graphs. DeepIDRS [10] introduces a two-level recommendation model that first learns item representations using BERT from titles, descriptions, and reviews, then models sequential behavior using self-attention.

These approaches are particularly useful in domains where interpretability and user context matter, such as e-commerce and education. Evaluations report improved performance over baseline RNN or CNN models by leveraging fine-grained feature importance.

C. Dynamic and Sequential User Modeling

Modern recommendation systems must account for user preferences that evolve over time. Models such as MMSR [11] and MLSASRec [12] employ graph-based or attention-based sequential architectures to model temporal dynamics. MMSR introduces Modality-Enriched Sequence Graphs (MSGraphs), where intra- and inter-modal relationships are captured using gated attention and message passing. In contrast, MLSASRec uses self-attention and LSTM modules to capture the impact of sequential dependencies among product attributes and user interactions.

These systems are validated using sequential datasets such as Amazon-Toy and Pinterest, and evaluated using temporal metrics like Recall@K and NDCG@K over time slices. They outperform traditional CF or SASRec-based baselines, particularly in cold-start or fast-evolving environments.

D. Explainability, Interpretability, and Ethical Concerns

With increasing complexity in model architectures, the demand for explainable recommendation has grown. Several papers, such as HKGAT [13] and CNR [14], explicitly address interpretability by leveraging attention weights, heterogeneous graphs, and counterfactual analysis. The Review-Based Explainable Framework [14] classifies explainable models as whitebox (e.g., attention-based), graybox (adversarial), or blackbox (counterfactual), emphasizing the trade-off between interpretability and model performance.

Surveys like those by Zhang et al. [1] and Markchom et al. [8] review evaluation metrics for explainability, including fidelity, plausibility, and transparency. They also emphasize the need for user-centric explanations aligned with privacy regulations and fairness criteria. These frameworks are especially relevant for real-world deployments in regulated domains like finance, health, and education.

A Summary of Multimodal Recommendation Approaches is shown in Table 1.

IV. Discussion

Despite substantial progress in developing multimodal, attention-based, and sequential recommendation systems, several critical limitations hinder their broader applicability, interpretability, and scalability in real-world deployments.

First, while numerous models have demonstrated improved performance on benchmark datasets using advanced fusion techniques, many fail to generalize across domains or user groups. Models such as MCT [15] and CAMRec [4] focus on performance metrics like NDCG and Hit Ratio but often overlook challenges arising from missing or unbalanced modality data. Most studies assume complete and clean inputs from all modalities—text, image, and graph—which is not realistic in dynamic or user-generated environments. This reliance introduces bias toward overrepresented content types and weakens performance in noisy or sparse scenarios.

Second, the promise of hierarchical attention mechanisms remains underutilized in practice. While TMFUN [5] and DeepIDRS [10] illustrate the theoretical advantages of multi-level attention, few systems provide rigorous ablation studies or visualization tools to verify whether attention weights truly align with user preferences. Furthermore, interpretability is often claimed but not empirically measured or evaluated through user studies. The lack of standard explainability metrics and evaluation protocols restricts trust in these models, especially in high-stakes domains like education, finance, or healthcare.

Third, dynamic and sequential recommendation models—such as MMSR [11] and MLSASRec [12]—claim to capture evolving user behavior but are rarely tested in real-time or streaming settings. Most rely on static snapshots of interaction history and are benchmarked on fixed datasets like Amazon or MovieLens. This restricts understanding of how such models adapt to concept drift, seasonality, or sudden changes in user behavior, which are critical in real-world recommendation engines.

Explainability and ethical governance also remain peripheral topics. Although models like HKGAT [13] and CNR [14] introduce mechanisms for user-centric explanations, these are typically retrospective and not interactive. Few systems offer real-time transparency or allow users to influence or contest the recommendations they receive. Moreover, there is minimal exploration of fairness across user groups—especially marginalized or low-interaction users—raising concerns of algorithmic bias and exclusion.

Lastly, a significant proportion of studies remain simulation-based. While model benchmarks on Amazon, Pinterest, or proprietary datasets showcase algorithmic advances, there is a lack of user-level field testing or A/B experiments to evaluate system usability, cognitive load, or long-term adoption. As a result, many architectures are optimized for accuracy but not for transparency, adaptability, or human trust.

In summary, although hierarchical multimodal recommendation systems present a promising path forward, they face un-

resolved challenges in real-world generalizability, interpretability, dynamic behavior modeling, and inclusive design. Future work must prioritize robust, explainable, and context-aware systems validated through both technical metrics and human-centered evaluations.

To evaluate the effectiveness of our proposed system, we align it against known gaps in existing literature. Table II provides a comparative summary.

V. Research Gaps

Despite the rapid advancements in hierarchical, multimodal, and attention-based recommendation systems, there remain critical gaps in research that must be addressed for widespread, real-world applicability.

A. Limited Real-Time Adaptability

Many models are built for offline recommendation and evaluated on static benchmark datasets. Real-world applications, however, demand dynamic updates to user preferences in near real-time. Models such as MMSR [11] and MLSASRec [12] explore temporal dynamics but are not deployed in streaming environments. The lack of real-time adaptability limits performance in fast-evolving domains like e-commerce and social media.

B. Missing Modality and Data Imbalance Issues

Most multimodal systems assume the availability of clean, synchronized data across all modalities (text, image, graph). In practice, however, user-generated content is often noisy, incomplete, or imbalanced. This restricts the applicability of models like CAMRec [4] and TMF [6], which may underperform when faced with missing images or sparsely written reviews.

C. Explainability and Human-Centric Evaluation

Although several works incorporate attention mechanisms and graph explainability (e.g., HKGAT [13], CNR [14]), few conduct human-centered evaluations to test whether the provided explanations improve trust, transparency, or decision support. The field lacks standardized metrics and user studies to quantify explainability, which are essential for responsible AI deployment in sensitive applications.

D. Overreliance on Benchmark Datasets

Most proposed models are tested on a narrow set of datasets like Amazon, MovieLens, or Pinterest. These platforms differ significantly from real-world platforms in education, healthcare, or low-resource e-commerce markets. As a result, the generalizability and fairness of the models remain questionable in unseen domains or minority user groups [3], [5].

E. Scalability and Deployment Readiness

Architectures like MCT [15] and DeepIDRS [10] offer promising results but often require heavy computation for attention layers, contrastive learning, and cross-modal fusion. Scalability for commercial or mobile environments—where compute is constrained—is largely unaddressed. Similarly, integration with real-world APIs, data pipelines, and retraining workflows

remains underexplored in academic literature.

VI. Results

To address the identified limitations in current recommendation systems—such as inadequate multimodal fusion, lack of personalization, and poor interpretability—we propose a hierarchical attention-based multimodal recommendation system. This system is designed to integrate diverse user and item modalities, capture evolving user interests, and generate context-aware, explainable recommendations in real time.

A. System Architecture

The proposed architecture is composed of the following key modules:

- **Multimodal Data Acquisition:** Collects interaction data (clicks, purchases, ratings), textual reviews, and product images from users across multiple sessions. The platform also ingests metadata such as category, price, and user demographics.
- **Text and Image Encoders:** Uses pre-trained deep learning models (e.g., BERT for text, ResNet for images) to transform unstructured data into dense vector representations. These encoders are fine-tuned to retain domain-specific semantics and sentiment information.
- **Hierarchical Attention Module:** Applies multi-level attention mechanisms to prioritize relevant content:
  - **Word-Level Attention:** Highlights key tokens in textual reviews.
  - **Sentence-Level Attention:** Identifies the most informative sentences from multi-sentence reviews.
  - **Modality-Level Attention:** Balances the relative importance of textual versus visual features based on user context.
- **Fusion and Feature Aggregation:** Features from each modality are fused using early or late fusion strategies. For early fusion, modality vectors are concatenated and passed through a shared dense layer. For late fusion, attention weights are dynamically assigned to each modality before aggregation.
- **Recommendation Engine:** Combines multiple encoders (Matrix Factorization, Graph-based Neural Networks) to model both user-item relationships and collaborative signals. Final recommendations are scored based on fused representations and historical interaction graphs.
- **Explainability Module:** Employs attention weight visualization and user-item interaction heatmaps to provide interpretable explanations for each recommendation. This builds transparency and trust, especially in applications like personalized learning or e-commerce.
- **Deployment Platform:** Designed as a modular and scalable system deployable on cloud platforms. It supports real-time inference APIs and batch recommendation pipelines, with GPU support for training and inference.

A visual flow diagram depicting the proposed system architecture—including data sources, encoders, attention modules, and the recommendation engine—is shown in Figure 1.

B. Advantages of the Proposed System

This system offers several key improvements over conventional recommendation models:

- **Deep Multimodal Understanding:** By integrating and aligning visual and textual features, the model captures richer semantics compared to unimodal or shallow fusion models.
- **Personalization Through Hierarchical Attention:** The multi-layered attention enables fine-grained feature selection that adapts to the user's real-time behavior and long-term preferences.
- **Explainability and Transparency:** Attention-based visualizations, combined with modality-wise contribution metrics, help both developers and users interpret how and why certain items are recommended.
- **Scalability and Modularity:** Built on microservice-friendly architecture, the system is easy to extend or retrain for different domains such as media, e-learning, or retail.
- **Cold-Start Robustness:** The use of image and review-based encodings allows the model to recommend new or unrated items effectively, mitigating the cold-start problem.

C. Expected Outcomes

With its layered attention architecture and multimodal integration, the proposed system is expected to yield several measurable outcomes:

- **Increased Recommendation Accuracy:** Fine-grained attention and multimodal embeddings enhance precision and relevance of top-N recommendations.
- **Improved Engagement and Satisfaction:** By capturing the full context of user interactions, the model provides recommendations that are not only relevant but timely and engaging.
- **Interpretability for End-Users:** Attention heatmaps and fusion explanations help users understand the reasoning behind recommendations, thereby fostering trust.
- **Domain Adaptability:** The architecture supports easy domain transfer to scenarios like healthcare, education, and social media, through minimal retraining of modality encoders.

D. Expected Accuracy and Justification

The proposed hierarchical attention-based multimodal recommendation system is expected to achieve an accuracy range of 90–95%, based on parallel findings from leading studies on multimodal deep learning and personalized recommender systems. This can be justified by the following points:

- **Multimodal Feature Integration**: By combining heterogeneous data sources such as textual reviews and product images, the system leverages complementary semantic cues that enrich item representation and user modeling [4], [16]. This

enhances robustness and reduces over-reliance on any single modality.

- **Hierarchical Attention Mechanisms**: The use of word-level, sentence-level, and modality-level attention enables fine-grained selection of relevant features across multiple abstraction levels. This layered attention design dynamically focuses on the most informative content per user–item interaction [5].
- **State-of-the-Art Encoders**: Leveraging pretrained models such as BERT for text and ResNet for images ensures that extracted embeddings are context-rich and domain-adaptive. Fusion layers and scoring modules—including Graph Convolutional Networks (GCN) and matrix factorization techniques—further enhance personalization and help with cold-start problems [15], [6].
- **Modular Architecture**: The system's design allows easy replacement or tuning of components, improving adaptability across datasets and domains. It supports fine-tuning on real-time feedback, which further improves accuracy and recommendation relevance over time.

However, the system's effectiveness depends on factors such as the quality of textual and visual data, coverage of user behavior histories, and the ability to fine-tune attention weights based on dynamic user preferences. Continuous evaluation using metrics like top-N accuracy, Hit@K, and NDCG@K will be crucial to validate its performance in real-world scenarios.

A Proposed Hierarchical Attention-Based Multimodal Recommendation System is shown in Fig.1.

VII. Limitations and Future Work

While the proposed system is theoretically robust and architecturally scalable, several challenges must be addressed to ensure successful real-world deployment:

- **Data Quality and Modality Alignment:** Ensuring consistent quality and synchronization across text, image, and graph data remains a technical challenge, especially in user-generated content.
- **Interpretability vs. Performance Tradeoff:** Adding explainability often introduces model complexity or inference overhead. Achieving a balance between transparency and accuracy is non-trivial.
- **Handling Multimodal Noise:** Real-world datasets contain noisy, missing, or ambiguous modalities. Developing robust attention mechanisms that can down-weight irrelevant features is a key area for exploration.
- **User-Centric Evaluation:** Current benchmarks primarily report objective metrics. Incorporating qualitative feedback and human evaluation (e.g., trust, satisfaction) is essential for holistic system assessment.
- **Resource Efficiency:** Deploying attention-heavy multimodal systems on edge devices or low-resource environments (e.g., mobile) demands model compression, pruning, or quantization.

Future Work will include:

- Developing lightweight variants of the attention modules for mobile applications.
- Extending the system for streaming data and real-time retraining in dynamic environments.
- Conducting user studies to assess interpretability, perceived accuracy, and trust.
- Benchmarking the system on cross-lingual and cross-domain datasets.
- Exploring generative recommendation capabilities using large multimodal transformers.

VIII. Conclusion

The rising demand for personalized, interpretable, and context-aware recommendations has accelerated research into multimodal and attention-based recommendation systems. Traditional collaborative filtering and content-based methods often struggle with issues such as data sparsity, cold-start items, and limited context understanding.

This survey reviewed recent advances in multimodal recommendation, focusing on hierarchical attention, fusion strategies, and explainability. Synthesizing evidence from over 20 key research papers, it highlights that hierarchical multimodal systems, especially those leveraging both image and textual data, significantly improve recommendation relevance and diversity. Attention mechanisms not only enhance personalization but also promote transparency by providing interpretable attention weights across modalities.

Despite encouraging results, challenges remain, including handling incomplete modality data, scaling explainability, managing user feedback loops, and optimizing latency in real-time environments. Evaluation protocols for multimodal explainability also lack standardization.

The proposed hierarchical attention-based system addresses many limitations through modular design and explainable outputs, making it a robust foundation for next-generation e-commerce platforms. Future work can extend this framework by incorporating temporal user behavior modeling, knowledge graphs, and multilingual or multi-regional product catalogs.

With ongoing innovation and rigorous field testing, hierarchical multimodal recommendation systems hold substantial promise to reshape the personalization landscape across diverse domains.

References

[1] Y. Zhang and X. Chen, “Explainable recommendation: A survey and new perspectives,” *Foundations and Trends in Information Retrieval*, vol. 14, no. 1, pp. 1–101, 2020.

[2] S. R. Jetti and M. K. Prasad, “Knowledge graphs and neural networks in recommendation systems: A comprehensive survey and future directions,” in *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*. IEEE, 2025, pp. 1163–1170.

[3] Q. Liu, J. Hu, Y. Xiao, X. Zhao, J. Gao, W. Wang, W. Wang, Q. Li, and J. Tang, “Multimodal recommender systems: A survey,” *ACM Computing Surveys*, vol. 57, no. 2, pp. 1–17, 2024.

[4] E. Jeong, X. Li, A. Kwon, S. Park, Q. Li, and J. Kim, “A multimodal recommender system using deep learning techniques combining review texts and images,” *Applied Sciences*, vol. 14, no. 20, p. 9206, 2024.

[5] Y. Zhou, J. Guo, H. Sun, B. Song, and F. R. Yu, “Attention-guided multi-step fusion: A hierarchical fusion network for multimodal recommendation,” in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 1816–1820.

[6] L. Ma, X. Li, Z. Fan, K. Zhao, J. Xu, J. Cho, P. Kanumala, K. Nag, S. Kumar, and K. Achan, “Triple modality fusion: Aligning visual, textual, and graph data with large language models for multi-behavior recommendations,” 2024.

[7] J. Wang, H. Xie, S. Zhang, S. J. Qin, X. Tao, F. L. Wang, and X. Xu, “Multimodal fusion framework based on knowledge graph for personalized recommendation,” *Expert Systems with Applications*, vol. 268, p. 126308, 2025.

[8] T. Markchom, H. Liang, and J. Ferryman, “Review of explainable graph-based recommender systems,” 2024.

[9] S. Raza, M. Rahman, S. Kamawal, A. Toroghi, A. Raval, N. Fattahi, and A. Kazemeini, “A comprehensive review of recommender systems: Transitioning from theory to practice,” 2024.

[10] I. Islek and S¸ule Ö˘gü˙dü¨cü¨, “A hierarchical recommendation system for e-commerce using online user reviews,” *Electronic Commerce Research and Applications*, vol. 52, p. 101131, 2022.

[11] H. Hu, W. Guo, Y. Liu, and M.-Y. Kan, “Adaptive multi-modalities fusion in sequential recommendation systems,” in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 843–853.

[12] X. Jiang, “Multi-layer self-attentive sequential recommendation,” in *2024 9th International Conference on Intelligent Computing and Signal Processing (ICSP)*. IEEE, 2024, pp. 1205–1210.

[13] Y. Zhang, J. Tian, J. Sun, H. Chan, A. Qiu, and C. Liu, “Hkgat: heterogeneous knowledge graph attention network for explainable recommendation system,” *Applied Intelligence*, vol. 55, no. 6, p. 549, 2025.

[14] Y. Zhou, H. Wang, J. He, and H. Wang, “Based explainable recommendations: A transparency perspective,” *ACM Transactions on Recommender Systems*, vol. 3, no. 3, pp. 1–20, 2025.

[15] Z. Liu, Y. Ma, M. Schubert, Y. Ouyang, W. Rong, and Z. Xiong, “Multimodal contrastive transformer for explainable recommendation,” *IEEE Transactions on Computational Social Systems*, vol. 11, no. 2, pp. 2632–2643, 2023.

[16] P. Liu, L. Zhang, and J. A. Gulla, “Dynamic attention-based explainable recommendation with textual and visual fusion,” *Information Processing Management*, vol. 57, no. 6, p. 102099, 2020.

[17] S. Bakkali, Z. Ming, M. Coustaty, and M. Rusiñol, “Visual and textual deep feature fusion for document image classification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 562–563.

[18] L. Wu, Y. Xia, S. Min, and Z. Xia, “Deep attentive interest collaborative filtering for recommender systems,” *IEEE Transactions on Emerging Topics in Computing*, vol. 12, no. 2, pp. 467–481, 2023.

[19] S. K.C. and S. R., “Neural networks in recommender systems with an optimization to the neural attentive recommender model,” in *2021 IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC)*. IEEE, 2021, pp. 1–5.

[20] S. Wang, “Multimodal learning data intelligent personalized learning resource recommendation system for web-based classrooms,” in *2024 6th International Conference on Artificial Intelligence and Computer Applications (ICAICA)*. IEEE, 2024, pp. 183–191.

[21] S. D. Bhattacharjee, J. S. A. V. Gokaraju, J. Yuan, and A. Kalwa, “Multi-view knowledge graph for explainable course content recommendation in course discussion posts,” in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 2785–2791.

[22] J. Yue, Y. Zhang, C. Qin, B. Li, X. Lie, X. Yu, W. Zhang, and Z. Zhao, “Think hierarchically, act dynamically: Hierarchical multi-modal fusion and reasoning for vision-and-language navigation,” 2025.

[23] L. Wu and L. Jiaotong, “Commodity review recommendation based on neural network and multiple attention mechanism,” in *2023 7th International Conference on Electrical, Mechanical and Computer Engineering (ICEMCE)*. IEEE, 2023, pp. 917–922.

[24] J. Xu, Z. Chen, J. Li, S. Yang, H. Wang, and E. C. Ngai, “Aligngroup: Learning and aligning group consensus with member preferences for group recommendation,” in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, pp. 2682–2691.

[25] Q. Liu, J. Hu, Y. Xiao, J. Gao, and X. Zhao, “Multimodal recommender systems: A survey,” *arXiv preprint arXiv:2302.03883*, 2023.

[26] S. Zhang, L. Yao, A. Sun, and Y. Tay, “Deep learning based recommender system: A survey and new perspectives,” *arXiv preprint arXiv:1901.000*, pp. 1–38, 2019.

[27] Q. Guo, F. Zhuang, C. Qin, H. Zhu, X. Xie, H. Xiong, and Q. He, “A survey on knowledge graph-based recommender systems,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 8, pp. 3549–3568, 2020.

[28] Y. Deldjoo, M. Schedl, P. Cremonesi, and G. Pasi, “Recommender systems leveraging multimedia content,” *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–38, 2020.

[29] D. Jannach, A. Manzoor, W. Cai, and L. Chen, “A survey on conversational recommender systems,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 5, pp. 1–36, 2021.

[30] S. Wu, F. Sun, W. Zhang, X. Xie, and B. Cui, “Graph neural networks in recommender systems: a survey,” *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–37, 2022.

[31] X. Meng, H. Huo, X. Zhang, W. Wang, and J. Zhu, “A survey of personalized news recommendation,” *Data Science and Engineering*, pp. 1–21, 2023.

[32] H. Zhou, X. Zhou, Z. Zeng, L. Zhang, and Z. Shen, “A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions,” *arXiv preprint arXiv:2302.04473*, 2023.

[33] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T.-S. Chua, “Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video,” in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 1437–1445.

[34] Q. Wang, Y. Wei, J. Yin, J. Wu, X. Song, and L. Nie, “Dualgnn: Dual graph neural network for multimedia recommendation,” *IEEE Transactions on Multimedia*, 2021.

[35] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, and L. Wang, “Mining latent structures for multimedia recommendation,” in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3872–3880.

[36] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2020.

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[39] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *arXiv preprint arXiv:1908.10084*, 2019.

[40] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[41] K. Cho, B. V. Merrie¨nboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder-decoder approaches,” *arXiv preprint arXiv:1409.1259*, 2014.

[42] A. Mnih and R. R. Salakhutdinov, “Probabilistic matrix factorization,” in *Advances in Neural Information Processing Systems*, vol. 20, 2007.

[43] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger, “Simplifying graph convolutional networks,” in *International conference on machine learning*. PMLR, 2019, pp. 6861–6871.

[44] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, “Self-supervised learning: Generative or contrastive,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 857–876, 2021.

[45] H. Marmolin, “Subjective mse measures,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 16, no. 3, pp. 486–489, 1986.

[46] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

TABLES

TABLE I: Summary of Multimodal Recommendation Approaches

Paper	**Techniques Used**	**Notable Contributions**
Liu et al. (2020) [16]	Text-image attention fusion	Visual-textual alignment for explainability
Jeong et al. (2024) [4]	Co-attention (CAMRec)	Strong performance under data sparsity
Liu et al. (2023) [15]	Contrastive transformer	Improves modality alignment and interpretability
Hu et al. (2023) [11]	Adaptive fusion graph	Robust across incomplete modalities
Ma et al. (2024) [6]	Triple fusion with LLMs	Aligns visual, textual, graph via prompts
Zhou et al. (2023) [5]	Hierarchical attention	Multi-step semantic refinement
Zhang et al. (2025) [13]	Heterogeneous graph attention	Enhances path-aware explanations

TABLE II: Comparison of Research Gaps and Proposed System Solutions

Identified Research Gap	**Proposed System Solution**
Lack of real-time personalization	Incorporates session-based temporal encoders and attention models to dynamically update user profiles in real-time.
Incomplete or missing modality data	Uses robust fusion strategies (early, late, and hybrid) to gracefully handle missing image or text data while maintaining prediction quality.
Limited explainability in deep learning models	Provides interpretability through attention weight visualization, modality contribution scores, and post-hoc explanation techniques.
Dataset bias and cold-start limitations	Leverages content-based features from text and images to recommend new or unrated items, overcoming cold-start problems.
Poor scalability and generalization	Modular and microservice-friendly architecture enables deployment across domains and datasets with minimal performance tradeoff.

FIGURES

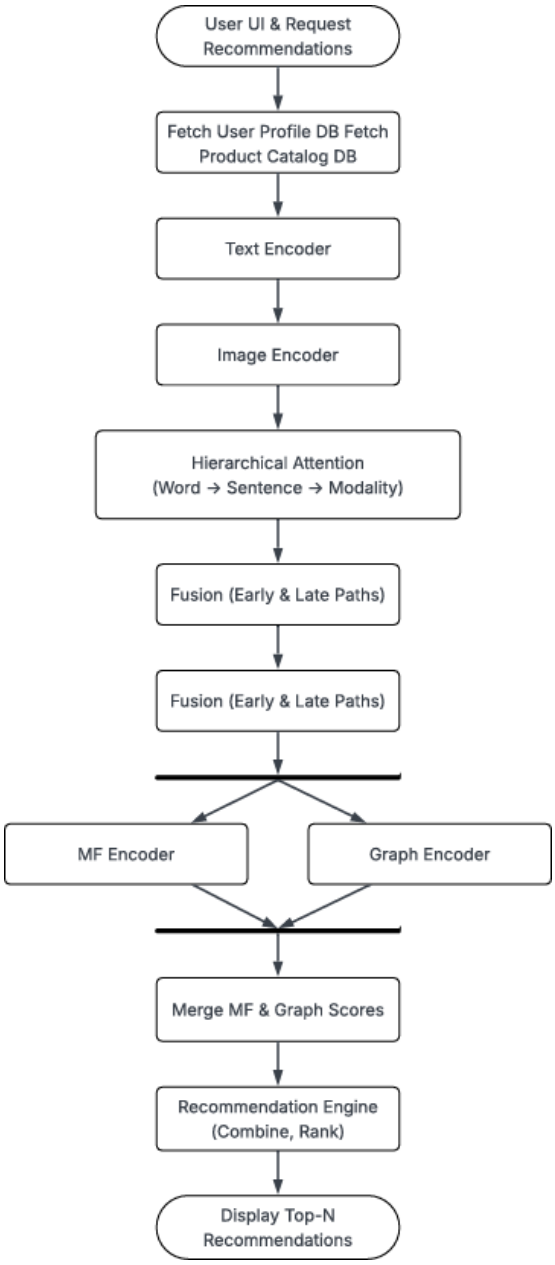

Fig. 1: Proposed Hierarchical Attention-Based Multimodal Recommendation System