Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

An Empirical Study of Attention-Based Cross-Modal Retrieval for Movies

Mohamed Elrfaey[†,*], T. Aaron Gulliver[†]
[†] Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada

Abstract

We present an empirical study of attention-based cross-modal retrieval for movies. Our approach combines text overviews, poster images, and trailer thumbnails using a cross-attention fusion module to learn unified item representations. To support this study, we augment MovieLens 1M with metadata from The Movie Database (TMDB), including overview text, poster images, and static trailer thumbnails. We evaluate text-only, image-only, and fused representations on top-K retrieval metrics, and compare them with interaction-only baselines based on Bayesian Personalized Ranking (BPR) and LightGCN. The results show that image-only retrieval achieves the strongest Recall@K and NDCG@K performance, while the fused model produces qualitatively more semantically balanced recommendations but does not outperform the strongest unimodal baseline. These findings suggest that attention-based multimodal fusion can improve recommendation coherence and interpretability, while also highlighting the challenge of translating cross-modal signals into stronger ranking performance.

Keywords: multimodal recommendation, cross-modal retrieval, movie recommendation, attention mechanisms, content-based recommendation, collaborative filtering

1. Introduction

Movie recommendations increasingly depend on signals drawn from multiple content modalities. In addition to structured interaction data, movies are naturally described through textual overviews, poster artwork, and trailer visuals, each of which captures a different aspect of an item's identity. Text can reflect narrative and theme, while visual assets often convey style, tone, and genre cues. This motivates multi-modal retrieval methods that aim to combine complementary signals into a unified representation for recommendation.

Despite this promise, effective fusion remains difficult. Many multi-modal recommendation systems rely on simple concatenation or shallow projection, which may not adequately capture relationships across modalities [1]. Other recent approaches enrich recommendations with retrieval-augmented or multimodal components [2], but often study only a subset of modalities or emphasize pipeline extensions beyond the retrieval setting itself. In practice, it remains unclear when attention-based fusion improves retrieval quality relative to strong unimodal representations.

In this paper, we present an empirical study of attention-based cross-modal retrieval for movies. We construct a multimodal benchmark by augmenting MovieLens 1M [3] with metadata from The Movie Database (TMDB) [4], including text overviews, poster images, and static trailer thumbnails. We then study whether cross-attention can combine these signals into more useful item embeddings for retrieval. Our focus is on item representation learning and top-K retrieval behavior rather than on a full end-to-end ranking pipeline.

The main contributions of this work are as follows.

- **A multimodal movie benchmark.** We enrich MovieLens 1M with aligned TMDB metadata, including overview text, poster images, and trailer-thumbnail proxies, to support multimodal retrieval experiments in the movie domain.

* elrefaey@uvic.ca

- **An attention-based fusion model for item retrieval.** We study a cross-modal fusion approach that combines text and visual signals into unified item embeddings for retrieval.
- **An empirical comparison across unimodal, fused, and interaction-only baselines.** We compare text-only, image-only, and fused representations against Bayesian Personalized Ranking (BPR) [5] and LightGCN [6]. Our results show that image-only retrieval achieves the strongest Recall@K and NDCG@K, while the fused model yields qualitatively more semantically balanced recommendations without outperforming the strongest unimodal baseline.

These findings highlight both the value and the limitations of attention-based multimodal fusion for movie retrieval. While fusion can improve recommendation coherence and interpretability, translating cross-modal signals into stronger ranking performance remains challenging.

2. Related Work

Multi-modal recommendation and retrieval aim to combine complementary signals from different content modalities, such as text and images, within a shared representation space. Prior work has explored joint embedding methods, contrastive learning objectives, and lightweight cross-modal representation learning to improve alignment across modalities [1, 7]. Attention-based mechanisms have also been used to model dependencies across heterogeneous inputs, showing that cross-modal interactions can be more expressive than shallow fusion strategies such as concatenation or projection [8].

Recent recommendation research has also considered richer retrieval settings, including retrieval-augmented and multimodal approaches that combine textual and visual information [2, 9]. However, many existing studies either focus on a subset of modalities, emphasize broader retrieval-ranking pipelines, or rely on fusion strategies that are not the primary object of analysis. As a result, the empirical benefit of attention-based fusion itself remains unclear, especially when compared against strong unimodal representations.

Our work focuses on this narrower question in the movie domain. We construct a multimodal benchmark by enriching MovieLens 1M with TMDB metadata, including text overviews, poster images, and static trailer thumbnails, and study whether cross-attention improves item representations for retrieval. Unlike a full ranking-system paper, our emphasis is on empirical comparison across text-only, image-only, fused, and interaction-only baselines.

3. Methodology

We study attention-based multimodal item retrieval in the movie domain using aligned textual and visual metadata.

3.1. Dataset and Representations

We augment MovieLens 1M with metadata from The Movie Database (TMDB). For each movie, we collect three aligned modalities: overview text, poster image, and a static trailer thumbnail. Movies missing any of these modalities are excluded so that all evaluated items have consistent multimodal representations.

Overview text is cleaned and lowercased, then encoded with a pre-trained Sentence-BERT model. Poster images and trailer thumbnails are resized and normalized, and visual features are extracted using ResNet-50. We treat trailer thumbnails as a static visual proxy rather than as full video input.

3.2. **Fusion Model and Retrieval**

ARMoR learns item representations by combining modality-specific embeddings with a cross-modal attention module. Given text, poster, and trailer-thumbnail embeddings, the fusion block produces a unified item embedding $\mathbf{z}_{\text{item}}$ in a shared latent space.

To train the fused representation, we use triplet margin loss, which encourages related items to be closer than unrelated items in the learned embedding space. User representations are formed by averaging the embeddings of movies that the user rated highly (≥ 4.0). At inference time, candidate items are retrieved by cosine similarity between the user embedding and item embeddings.

3.3. **Experimental Setup**

We use an 80/10/10 train/validation/test split and evaluate retrieval with Recall@K and NDCG@K, averaged over test users with at least five historical interactions. We compare four retrieval settings: text-only, image-only, fused multimodal retrieval, and interaction-only baselines based on Bayesian Personalized Ranking (BPR) [5] and LightGCN. The collaborative-filtering baselines are trained on binarized implicit feedback derived from ratings ≥ 4.0 and are included as interaction-only reference baselines.

4. **ARMoR Architecture**

Figure 1 illustrates the ARMoR fusion architecture used in our experiments. The model combines three aligned modalities for each movie: overview text, poster image, and a static trailer thumbnail. Modality-specific encoders first produce fixed-dimensional embeddings, which are then fused through a cross-modal attention block to obtain a unified item representation for retrieval.

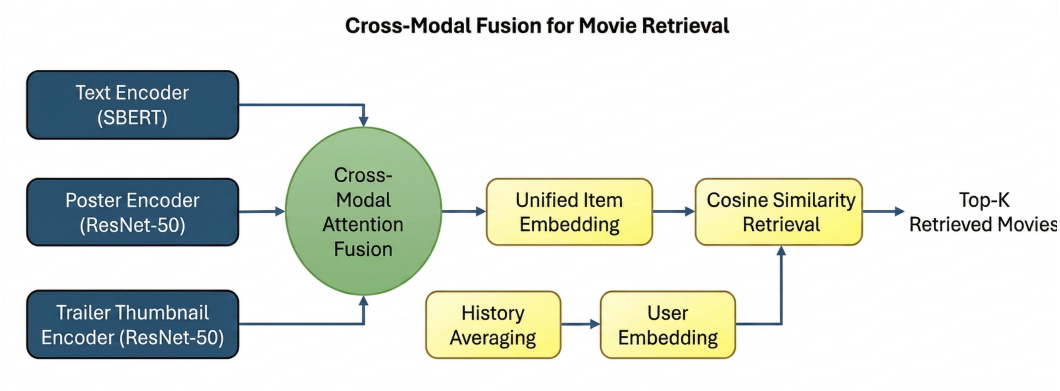


Figure 1. ARMoR architecture used in our experiments: modality-specific encoders for overview text, poster images, and static trailer thumbnails are fused through cross-modal attention to produce shared item embeddings for cosine-similarity retrieval.

4.1. **Cross-Modal Fusion**

Let

$$\mathbf{e}_{\text{text}}, \mathbf{e}_{\text{poster}}, \mathbf{e}_{\text{thumb}} \in \mathbb{R}^d, \tag{4.1}$$

denote the text, poster, and trailer-thumbnail embeddings, respectively. In our implementation, text embeddings are obtained from Sentence-BERT, while poster and trailer-thumbnail embeddings are extracted with ResNet-50.

To model interactions across modalities, we apply cross-modal attention with the text embedding as the query and the visual embeddings as keys and values:

$$\mathbf{Q} = W^Q \mathbf{e}_{\text{text}}, \tag{4.2}$$

$$\mathbf{K} = \begin{bmatrix} W^K \mathbf{e}_{\text{poster}} \\ W^K \mathbf{e}_{\text{thumb}} \end{bmatrix}, \tag{4.3}$$

$$\mathbf{V} = \begin{bmatrix} W^V \mathbf{e}_{\text{poster}} \\ W^V \mathbf{e}_{\text{thumb}} \end{bmatrix}. \tag{4.4}$$

The fused representation is computed as

$$\mathbf{e}_{\text{fused}} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}. \tag{4.5}$$

We then project the fused representation into the shared retrieval space:

$$\mathbf{z}_{\text{item}} = W_p \,\text{ReLU}(\mathbf{e}_{\text{fused}}) + b_p. \tag{4.6}$$

4.2. **Training Objective and Retrieval**

The fused item embeddings are trained using triplet margin loss:

$$\mathcal{L}_{\text{triplet}} = \max(0,\, \text{sim}(\mathbf{z}_a, \mathbf{z}_n) - \text{sim}(\mathbf{z}_a, \mathbf{z}_p) + \alpha)\,. \tag{4.7}$$

This objective encourages related items to lie closer in the learned embedding space than unrelated items.

For retrieval, a user embedding $\mathbf{z}_u$ is computed by averaging the embeddings of movies the user rated highly (≥ 4.0). Candidate items are then ranked by cosine similarity:

$$s(\mathbf{z}_u, \mathbf{z}_j) = \cos(\mathbf{z}_u, \mathbf{z}_j). \tag{4.8}$$

5. **Results and Discussion**

We evaluate retrieval quality using Recall@K and NDCG@K for $K \in \{5, 10, 20\}$ under a leave-one-out protocol. For each test user, the user representation is computed as the average of the embeddings of highly rated movies (≥ 4.0), excluding the held-out test item. We compare three content-based settings—text-only, image-only, and fused multimodal retrieval—as well as two interaction-only reference baselines, BPR and LightGCN.

Table 1 summarizes the main quantitative results. Among the content-based methods, image-only retrieval achieves the strongest performance across all metrics. The fused model underperforms both text-only and image-only retrieval on Recall@K and NDCG@K. Among the interaction-only baselines, BPR outperforms LightGCN, and both exceed the fused model on the reported ranking metrics.

Table 1. Recall@K and NDCG@K results. Best values are shown in **bold**.

Model	R@5	R@10	R@20	N@5	N@10	N@20
Text-only	0.018	0.043	0.107	0.0070	0.0160	0.0320
Image-only	**0.052**	**0.122**	**0.196**	**0.0260**	**0.0480**	**0.0650**
Fused	0.004	0.012	0.036	0.0025	0.0050	0.0105
BPR	0.017	0.035	0.073	0.0100	0.0156	0.0250
LightGCN	0.010	0.030	0.065	0.0055	0.0110	0.0201

These results suggest that poster-based visual features are particularly strong for this benchmark. At the same time, the fused model remains useful qualitatively. Figure 2 shows the top-5 retrieved movies for *The Matrix (1999)* using attention-based fusion. In manual inspection using *The Matrix (1999)* as a query, text-only retrieval emphasizes narrative

and thematic similarity, image-only retrieval emphasizes visual style, and fused retrieval produces more semantically balanced results by combining complementary cues from text, posters, and static trailer thumbnails.

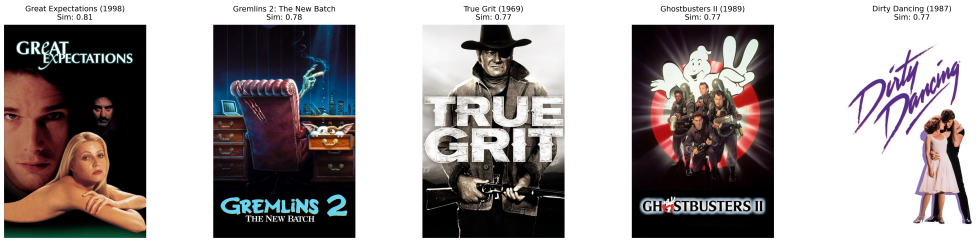


Figure 2. Top-5 retrieved movies for *The Matrix (1999)* using attention-based fusion of overview text, poster images, and static trailer thumbnails.

This contrast between quantitative and qualitative behavior is important. Although attention-based fusion did not improve top-K ranking metrics in the current setup, the qualitative example in Figure 2 suggests that it can produce recommendations that are more interpretable and contextually balanced than those from single-modality retrieval alone. We therefore view the main empirical results of this study as twofold: strong unimodal visual representations remain difficult to beat on this benchmark, while multimodal fusion offers potential benefits in coherence and interpretability.

A key limitation is that the third modality is represented by static trailer thumbnails rather than full video, so it serves as a visual proxy rather than a true spatio-temporal signal. In addition, BPR and LightGCN are interaction-only baselines and therefore provide contextual reference rather than a fully matched comparison to content-based retrieval. Future work should study richer video representations and stronger multimodal training strategies, and should evaluate against more directly comparable multimodal baselines.

6. Conclusion and Future Work

We presented ARMoR, an attention-based approach for cross-modal movie retrieval that combines overview text, poster images, and static trailer thumbnails in a shared embedding space. To study this setting, we enriched MovieLens 1M with aligned metadata from TMDB and compared text-only, image-only, fused, and interaction-only baselines.

Our results show that image-only retrieval achieves the strongest Recall@K and NDCG@K performance on this benchmark, while the fused model does not outperform the strongest unimodal or interaction-only baselines quantitatively. At the same time, qualitative inspection suggests that attention-based fusion can produce more semantically balanced and interpretable recommendations by combining narrative and visual cues. These results highlight both the promise and the difficulty of multimodal fusion for retrieval. Strong visual representations remain difficult to beat, but cross-modal fusion may still offer value in coherence, interpretability, and recommendation diversity.

Future Work. Future work will study richer multimodal training strategies, more directly comparable multimodal baselines, and stronger visual representations beyond static trailer thumbnails. We also plan to investigate contrastive objectives and hybrid approaches that combine interaction signals with multimodal item representations.

References

[1] H. Won, B. Oh, H. Yang, and K.-H. Lee. "Cross-Modal Contrastive Learning for Aspect-Based Recommendation". In: *Information Fusion* 99 (2023), p. 101858. DOI: 10.1016/j.inffus.2023.101858.

[2] A. Tourani, F. Nazary, and Y. Deldjoo. "RAG-VisualRec: An Open Resource for Vision- and Text-Enhanced Retrieval-Augmented Generation in Recommendation". In: *arXiv preprint arXiv:2506.20817* (2025). DOI: 10.48550/arXiv.2506.20817. URL: https://arxiv.org/abs/2506.20817.

[3] F. M. Harper and J. A. Konstan. "The MovieLens Datasets: History and Context". In: *ACM Transactions on Interactive Intelligent Systems* 5.4 (2015), pp. 1–19. DOI: 10.1145/2827872.

[4] The Movie Database (TMDB). *TMDB API*. https://developer.themoviedb.org/docs/getting-started. Accessed: 2026-04-11. 2026.

[5] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. "BPR: Bayesian personalized ranking from implicit feedback". In: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI 2009)*. 2009, pp. 452–461. URL: https://dl.acm.org/doi/10.5555/1795114.1795167.

[6] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang. "LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020, pp. 639–648. DOI: 10.1145/3397271.3401063.

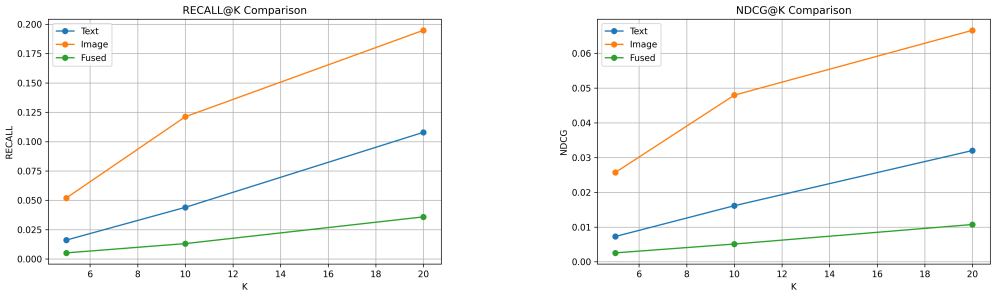
[7] B. Faye, H. Azzag, M. Lebbah, and D. Bouchaffra. "Lightweight Cross-Modal Representation Learning". In: *Proceedings of the 32nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2024)*. 2024. DOI: 10.14428/esann/2024.ES2024-96.

[8] X. Song, H. Chao, X. Xu, H. Guo, S. Xu, B. Türkbey, B. J. Wood, T. Sanford, G. Wang, and P. Yan. "Cross-modal attention for multi-modal image registration". In: *Medical Image Analysis* 82 (2022), p. 102612. DOI: 10.1016/j.media.2022.102612.

[9] J. Deng, S. Wang, K. Cai, L. Ren, Q. Hu, W. Ding, Q. Luo, and G. Zhou. "OneRec: Unifying Retrieve and Rank with Generative Recommender and Iterative Preference Alignment". In: *arXiv preprint arXiv:2502.18965* (2025). URL: https://arxiv.org/abs/2502.18965.

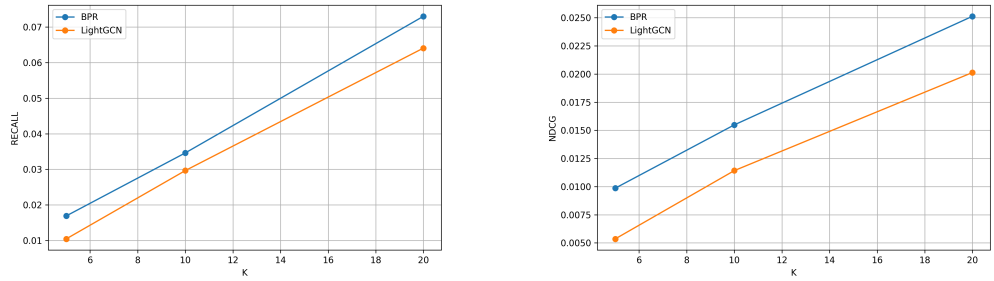
Appendix A. Supplementary Quantitative Results

For completeness, we include additional Recall@K and NDCG@K plots that complement Table 1 in the paper.



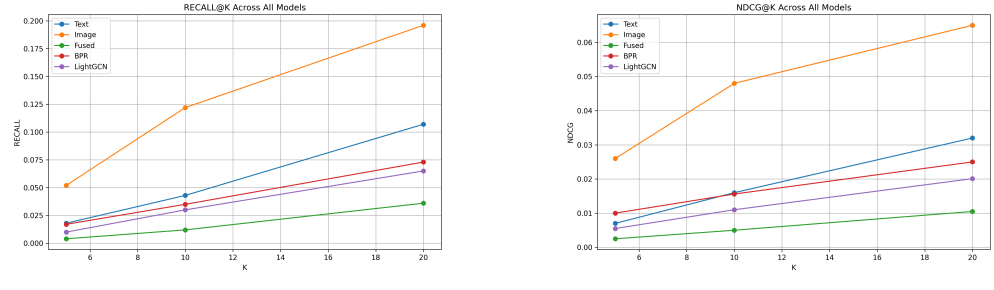
(a) Recall@K *(b)* NDCG@K

Figure 3. Text-only, image-only, and fused retrieval results.



(a) Recall@K *(b)* NDCG@K

Figure 4. Interaction-only baseline results for BPR and LightGCN.



(a) Recall@K *(b)* NDCG@K

Figure 5. Results for the text-only, image-only, fused, BPR, and LightGCN models.

Appendix B. Supplementary Qualitative Results

We provide additional qualitative retrieval examples for *The Matrix (1999)* using text-only and image-only retrieval.

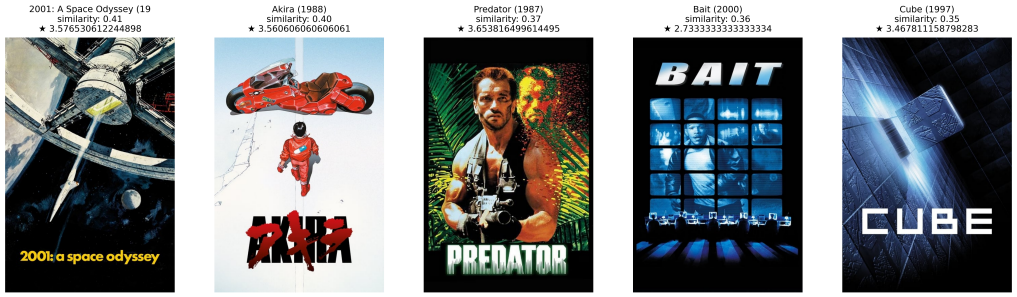


Figure 6. The top 5 retrieved movies for *The Matrix (1999)* using text-only retrieval.

Figure 7. The top 5 retrieved movies for *The Matrix (1999)* using image-only retrieval.

Appendix C. Text Embedding Visualization

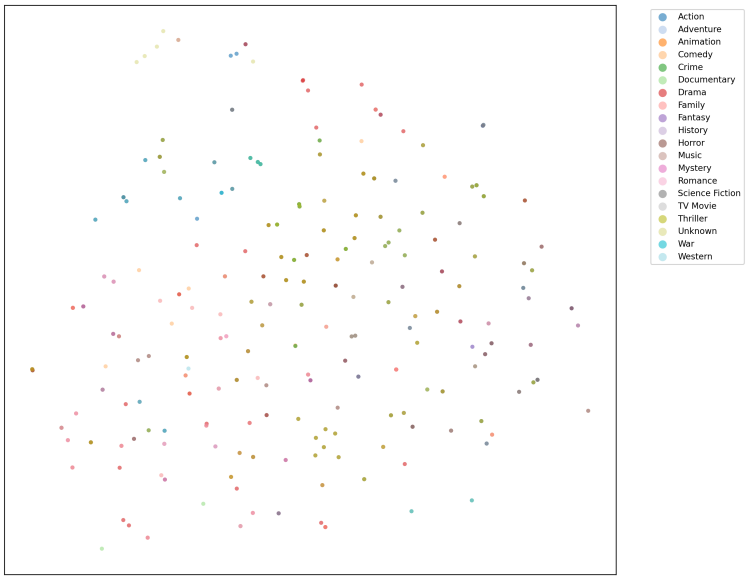


Figure 8. t-SNE visualization of SBERT text overview embeddings, colored by primary genre.