Multimodal Recommender Systems: A Comprehensive Study

Sri Spoorthi Vattem - SE22UARI165
Artificial Intelligence
Mahindra University
Hyderabad, India
Team-ID S165

***Abstract*— Multimodal machine learning, integrating vision, text, audio, and beyond has driven breakthroughs in tasks such as Visual Question Answering, Visual Common-sense Reasoning, and Phrase Grounding. Yet inconsistent or outdated definitions of "modality" hinder systematic progress. In this work, we discuss a task-relative definition: two inputs constitute distinct modalities if, after preprocessing, they are represented differently and their atomic information units lack a bijective mapping. We illustrate this definition through cases ranging from image–text to infrared–visible data. In the domain of recommender systems, we identify a harmful "siloization" effect, where models built for one modality are compared only within that modality. By recognising that all modalities ultimately yield vectors or matrices, we repurpose five representative recommender models, text-based (CDL, CDR), image-based (VMF, VBPR), and graph-based (MCF) to operate on text, image, and graph inputs interchangeably. Across six Amazon review datasets, we find that (1) any auxiliary modality improves performance over preference-only baselines; (2) graph data yields the largest gains, followed by text, then images; and (3) many models achieve their best results with non-native modalities. These findings advocate for a unified approach to multimodal modelling and cross-modality evaluation in recommender systems and beyond. Moreover, explored the three dedicated multimodal recommender architectures, a deep-learning model combining review texts and product images, a multimodal movie recommender leveraging poster imagery and metadata, and a Disentangled Multimodal Representation Learning framework to validate our insights across diverse real-world settings. These findings advocate for a unified approach to multimodal modelling and cross-modality evaluation in recommender systems and beyond.**

Keywords—multimodal machine learning; cross-modality utilisation; recommender systems; deep learning; disentangled representation learning.

I. Introduction

The rapid expansion of multimodal machine learning, the fusion of heterogeneous data sources such as images, text, and graph structures, has yielded remarkable advances in areas from Visual Question Answering to Visual Dialogue. By leveraging complementary sensory channels, these systems achieve richer representations and deeper understanding than unimodal approaches allow. However, the field remains hampered by a lack of consensus on what constitutes a "modality." Traditional conceptions fall into two camps:

1. Human-Centered Definitions
Modalities equate to human senses (vision, hearing, touch). While intuitive, this view cannot account for inputs beyond human perception (e.g., ultraviolet imagery, gene expression data) or the subtleties of modern ML preprocessing.

2. Machine-Centered Definitions
Modalities correspond to data encodings (e.g., raw images vs. tokenized text). This perspective breaks down when a single architecture (such as a Transformer) processes multiple data types identically, or when trivial file-format changes (JPEG ↔ PNG) are conflated with new modalities.

To overcome these limitations, we propose a task-relative definition: a machine-learning task is multimodal if its inputs (or outputs)

- are represented differently after preprocessing, and
- contain distinct atomic units (pixels, tokens, graph edges) lacking a bijective mapping between them.

This definition accommodates non-human-sensory data, clarifies when two inputs truly offer different information, and grounds multimodality in the context of the learning task itself.

In parallel, multimodal recommender systems have suffered from "siloization," where image-based methods are compared only to other image models, text-based methods only to text, and so on. Yet regardless of origin, all modalities ultimately convert to vector or matrix inputs for downstream learning. We therefore conduct a cross-modality evaluation of five representative models; CDL, CDR (text), VMF, VBPR (image), and MCF (graph) - across six Amazon review datasets. Each model, originally tied to one modality, is repurposed to ingest text, image, or graph data via simple substitution of its input representation.

The study reveals three key insights:

1. Auxiliary modalities always boost performance over preference-only baselines.
2. Graph data delivers the greatest improvements, followed by text, then images; though this ranking varies by domain.
3. Best model-modality pairings often occur with non-native modalities (e.g., text models excel on graph inputs).

These findings call for a unified approach to multimodal modelling, where definitions, architectures, and evaluations span modalities rather than remain confined to narrow silos.

II. What is Multimodality ?

A. In recent years, multimodal machine learning, the integration of vision, text, audio, and beyond has gained tremendous traction. Tasks such as Visual Question Answering, Visual Common-sense Reasoning, Visual Dialogue, and Phrase Grounding showcase the power of combining modalities to achieve richer understanding and interaction. Yet progress remains hindered by outdated or inconsistent definitions of "modality."

Most prior work adopts either a human-centered view (e.g. modalities = human senses) or a machine-centered view (e.g. modalities = data encodings), but each alone proves inadequate:

- Human-centered: Fails to account for non-human-perceivable signals. Example: Human senses are limited (e.g., no UV perception). Machines can process non-human-perceivable signals.

• Machine-centered: Breaks down when a single architecture (e.g. a Transformer) encodes both images and text identically.

A task-relative definition, where modalities are distinguished by their atomic information units and by whether a bijective mapping exists, better accommodates the diversity of inputs and the needs of modern ML. This redefinition grounds language and other data in task-specific representations, paving the way to true Natural Language Understanding (NLU) and beyond.

B. Defining "Modality" for Machine Learning

Shortcomings of Existing Views

• No Definition/Etymological: Multimodal means simply multiple modalities. Leaves "modality" itself undefined.

• Human-Centered: We perceive the world through vision, hearing, touch, etc.
-Cannot classify non-human-sensory inputs.
-Contradicts community practice of treating text and images as distinct.

• Machine-Centered: A modality is the data encoding used before ML processing.
-Breaks when a single encoder (e.g. Transformer) handles multiple inputs identically.
-Equates any format change (PNG↔JPEG, adjacency matrix ↔ graph) with a new modality.

C. A Task-Relative Definition

Definition: A machine-learning task is multimodal if its inputs (or outputs)

Are represented differently after preprocessing, *and* Contain different *atomic units* of information i.e. there is *no* one-to-one (bijective) mapping between their fundamental elements.

• Atomic units: pixels, phonemes, tokens, graph edges, etc.

• Bijective mapping: each atomic unit in modality A corresponds exactly to one in modality B.

D. Illustrative Cases

Case	Same Modality?	Justification
Image vs Text	**No**	Pixels ≠ tokens; no one-to-one mapping.
Infrared vs Visible Light Images	**Yes**	Both are grids of photon-intensity values; bijective mapping.
Hateful Meme Challenge (pixels + OCR text)	**Yes (ideally)**	If a model processes raw pixels (including embedded text) uniformly, it's unimodal—mirroring human vision.
CLIP (image + text pairs)	**No**	Separate image and text inputs; no bijection between pixel regions and word tokens.

III. From Siloization to Cross-Modality Utilisation

A. Siloization by Modality and Its Pitfalls

One of the key issues in multimodality recommender systems is that models designed for a specific modality are tested and compared only within that same modality. For example, an image-based model is evaluated against other image-based models, while text-based models are benchmarked only against their textual counterparts. This artificial segregation assumes that modalities are fundamentally incompatible, even though they often share representational structures at a deeper level. This assumption restricts the reuse of valuable modelling innovations across different modalities and limits the exploration of generalisable models.

B. Cross-Modality Utilisation: A Broader Perspective

One of the central insights in this research is that many multimodal recommender algorithms are essentially machine learning models that treat auxiliary data as features. Regardless of the data's origin - text, image, or graph, the final representations used for learning (e.g., vectors or embeddings) often share structural similarities. For instance, textual descriptions may be represented using TF-IDF vectors, while graph-based information may be encoded via adjacency matrices or node embeddings. After preprocessing, these seemingly different modalities converge into comparable input formats. Recognising this commonality, the researchers propose a systematic investigation into whether models built for one modality can generalise to others. Specifically, they examine three commonly used auxiliary data types — text, image, and graph and explore.

C. Experimental Setup

Datasets: Six Amazon-review datasets, each containing:

• User–item preferences (ratings)

• Textual descriptions of items

• Product images

• Co-purchase graphs ("Also-Viewed" links)

Dataset	#Users	#Items	#Ratings	#Text	#Images	#Graph Edges
Cellphones	3,383	2,170	9,214	2,170	2,170	2,012
Clothing	5,377	3,393	13,689	3,393	3,393	9,198
Electronics	55,930	30,074	2,12,863	30,074	30,074	63,242
Movies & TV	28,566	10,116	1,96,277	10,116	10,116	19,763
Office	24,232	13,520	99,255	13,520	13,520	2,06,719
Tools	19,902	12,522	58,419	12,522	12,522	44,509

Preprocessing:

• Retain items with all three modalities.

• Keep users with ≥ 3 ratings; items with ≥ 2.

• 80% of ratings → training; 20% → testing.

D. Models

Five representative models, each with a Base variant using only preference data: All models implemented in the Cornac framework for consistent evaluation.

Modality	Model	Base Version	Description
Text	CDL	MF	Matrix-factorisation + SDAE on text (Gaussian loss)
	CDR	BPR-MF	Same as CDL but pairwise ranking loss
Image	VMF	MF	MF + visual-feature vectors from CNN
	VBPR	BPR-MF	BPR + visual features
Graph	MCF	MF	Joint factorization of user–item & item–item matrices

E. Cross-Modality Repurposing:

Each model is tested on all three modalities by substituting its original input with:

- CDL/CDR (Text models) can process image or graph data by replacing the document-word matrix with a visual feature matrix or an adjacency matrix.
- VBPR/VMF (Image models) can handle text or graph data by similarly substituting the input representations.
- MCF (Graph model) can use text or image data by converting them into similarity graphs (e.g., cosine similarity for text vectors or visual embeddings).

This flexibility enables robust experimentation across modality boundaries.

F. Evaluation Metrics

- Recall@N and NDCG@N (higher is better) for top-N recommendation.
- RMSE and MAE for rating-prediction tasks where applicable.

G. Key Findings & Insights

Analysis shows that all models perform better with any auxiliary data than with preference-only "Base" versions. The improvement is modality-agnostic, even models used with mismatched modalities (e.g., CDL with images) outperform their Base counterparts. This clearly affirms the value of leveraging auxiliary modalities.

Which Modality to rely on?

- Graph data generally yields the largest gains (captures co-consumption patterns).
- Textual data often trumps images, even in visually-driven categories.
- Images add the least but still offer improvements.

Can Models Perform Better with Other Modalities?

Surprisingly, most models perform best with a modality other than the one they were designed for.

Examples: CDL (text model) works best with graph on multiple datasets. VMF (image model) also excels with graph or text inputs. VBPR (image model) performs best with text on *MoviesTV* and *Office* datasets.

Should We Use Models Designed for Other Modalities?

Surprisingly, most models perform best with a modality other than the one they were designed for.

Examples: CDL (text model) works best with graph on multiple datasets. VMF (image model) also excels with graph or text inputs. VBPR (image model) performs best with text on *MoviesTV* and *Office* datasets.

IV. Recommendation Fundamentals

A. Classic approaches

- Collaborative Filtering (CF):

User-based CF: find similar users by rating patterns.

Item-based CF: find similar items by co-ratings.

Suffers from data sparsity & cold-start.

- Content-Based Filtering (CBF):

Matches item attributes to user profiles.

Lacks novelty; ignores community preferences.

- Hybrid Systems:

Combine CF + CBF to mitigate individual weaknesses.

B. Types of Data in Recommendation systems:

- Behavioural data: Ratings, purchases, clicks, history
- Textual data: item descriptions, user profiles, tags

Persistent Challenges: Sparsity, cold-start, limited diversity, information decay.

Broadly, these methods of integrating multimodal information can be categorised into two types (Multimodal Integration Strategies) :

- Fusion-based models that merge features from various modalities either linearly or non-linearly to form joint representations.
- Regularisation-based models that enforce constraints on representations from different modalities to encourage consistency in the learned latent space.

Recent research shows multimodal fusion (text+image+audio) leads to Higher-quality recommendations, improved user satisfaction and better handling of unstructured data.

Different ways of combining modalities to improve the recommendation systems -

- A multimodal recommender system combining review texts and product images using deep learning techniques.
- Multimodal Movie Recommendation System Using Deep Learning
- Disentangled Multimodal Representation Learning for Recommendation.

V. Representative Multimodal Models

A. CAMRec: Co-Attention Multimodal Recommender

A multimodal recommender system combining review texts and product images using deep learning techniques. This method improves recommendations by capturing user preferences more accurately through sentiment-rich reviews and visual features from product images. The goal of this method is to build a model (CAMRec) that learns from both visual and textual data to generate better recommendations. CAMRec uses the RoBERTa and VGG-16 models to extract textual and visual information from online reviews and applies a co-attention mechanism to capture the complementarity between the two modalities. This model shows the impact of visual representations on recommender performance.

Objective: Leverage review texts & product images via a co-attention mechanism.

1.1 Model Overview

- RoBERTa is used to extract textual information and VGG-16 model to extract visual information, then the co-attention mechanism to obtain the joint representations of textual and visual information through dependencies between them.
- A co-attention-based fusion strategy is implemented that models interactive dependencies between the two modalities.
- Finally integrates the overall features to predict user preference for a specific item; for the evaluation of the model 2 datasets have been used from Amazon.

1.2 Textual Review Modelling

- Reviews are tokenised and passed through an embedding layer.
- Followed by a Bidirectional LSTM, which captures context from both directions of the text.
- Mathematical representation:

$h_t = \text{BiLSTM}(e_t, h_{t-1})$

where e_t is the embedding at time step and h_t is the hidden state.

- The final representation is a context vector H, aggregating all hidden states.

1.3 Visual Feature Extraction

- Product images are input to a pre-trained VGG-16 CNN model.
- Extracted from the second fully connected layer (fc2), yielding a 4096-dimensional vector V

1.4 Multimodal Fusion Layer

Simple fusion (e.g., addition, concatenation) lacks capacity to model inter-modal dependencies. Co-attention mechanisms dynamically learns what part of one modality attends to the other.

- The fusion step combines:

$$F = \text{ReLU}\,(W_t H + W_v + b)$$

where W_t, W_v = weight matrices for text and visual features
b = bias, ReLU = activation function

This fused representation F is passed to a regression layer to predict the rating:

$$\hat{r}_{ui} = \sigma\,(W_f F + b_f)$$

where $\hat{r}ui$ is the predicted rating for user u and item i

Methodology used:

The CAMRec model includes:

User–Item Interaction Network

Users/items are one-hot encoded, then embedded into dense vectors, then further are concatenated. The vectors are passed through MLP with leaky ReLU activations, the output is a latent user-item interaction.

Feature Extraction Network

- RoBERTa encodes reviews and Uses [CLS] token as text representation.
- Images passed through VGG-16 (4096-dim output per image), for multiple images multihued Co-attention mechanism in implemented

Preference Prediction Network

Concatenate fused feature F and user–item interaction vector E_M.

VI. Multimodal Movie Recommendation (Deep-CNN Approach)

Objective: Fuse user metadata (ID, gender, age, profession) and movie metadata (ID, genre, title, poster) via a deep CNN.

Personalised multimodal movie recommendation system that combines multimodal data like text (title), metadata (genre, type), visual data (movie posters), etc. Applies deep learning to extract hidden features and model complex patterns in the data.

This model is trained on a Real-world MovieLens datasets, 100K and 1M. The method used, first the hidden features of users and movies are mined using deep learning then these features are used to train a network model to predict movie scores. The evaluation metric is RMSE. This method outperforms CF and SVD by mitigating sparse data issues while offering better personalisation. Deep learning helps extracting the deep features/hidden dependencies using multi-layered neural networks to model non-linear, high-level feature interactions and provide good results even if the data is incomplete or high dimensional in data. DL can generate better representations of users/items than traditional methods. Common DL architecture for RS consists of encoding layers to extract agent features from noisy output and decoding layers for predicting the ratings. The inputs are movie content and user ratings

1.1 Model Overview

Based on deep learning technology and multimodal data analysis a system architecture is proposed for the movie recommendation.

- Leverages user features (gender, age, profession, ID) and movie features (title, genre, poster, ID).
- Fuses these features via a deep convolutional neural network (CNN) architecture.
- Finally trained and evaluated using MovieLens datasets (100K, 1M).

1.2 Framework of the proposed model:

- The dataset contains multimodal information:
  For users: ID, gender, age, profession.
  For movies: ID, genre/type, title, poster (image).
- Transformation: Inputs are converted into single-value matrices that contain non-zero singular values. This refers to applying dimensionality reduction or embedding techniques to compress input data while preserving key information.
- CNN Training: A CNN with multiple convolution layers is trained to learn high-level feature representations. These layers perform feature extraction and dimensionality compression.
- Recommendation generation: Once trained the model uses similarity-based matching to identify relationships between users and movies in the latent feature space. The final recommendation is done using top-N filtering; the redundant data is removed, scores are normalised and filtered, finally top-N most relevant movies are recommended to the user.
- Feature extraction: CNN is used to extract the hidden features of users and movies in the MovieLens dataset. CNNs are best here as they are good at hierarchical feature learning, they work well with unstructured data especially images and unlike traditional models, CNNs do not require hand crafted features or explicit mathematical models.
- Output layer: Produces top 10 recommended movies, uses fully connected layer for final prediction.
- Evaluation metrics: The train-test split is 80:20. The metric used to test is RMSE (Root Mean Square Error). Lower RMSE → better prediction accuracy. As the number of training batches increases, RMSE continues to decline and stabilises, indicating effective learning.
- Conclusion: Model with posters (multimodal) slightly outperforms the non-multimodal version. Multimodal input (poster images) improves accuracy marginally but consistently. Suggests that adding more modalities (e.g., audio, trailer features) may lead to further improvements.

VII. DMRL: Disentangled Multimodal Representation Learning

Objective: Disentangle modality-specific factors and learn user-specific modality preferences per factor.

While multimodal recommender systems have advanced by incorporating data such as reviews and images to better capture user preferences, they still suffer from a critical limitation: they fail to disentangle the distinct contributions of each modality with respect to specific item factors. User preferences are inherently diverse, a user might prioritise brand and quality for one product but focus on aesthetics for another. Reviews often reflect abstract aspects like comfort

or usability, whereas images capture tangible visual elements like style or colour. However, most existing models treat multimodal inputs as unified blocks, ignoring the unique, factor-specific strengths of each modality. Even when attention mechanisms are used, they typically lack the granularity needed to differentiate which modality informs which factor, leading to entangled and less accurate user/item representations.

To mitigate these issues a Disentangled Multimodal Representation Learning (DMRL) is proposed, which aims to explicitly model user attention to each modality on a per-factor basis. The core idea is to disentangle representations so that features corresponding to different item factors are independent within each modality. Then, a multimodal attention mechanism is applied to assign personalised weights to these factors across modalities. The final recommendation is generated by combining factor-wise preference scores based on these modality-specific weights.

Disentangled Representation learning seeks to uncover and separate the underlying explanatory factors in the data, making representations more interpretable and robust to complex variations. Recent works focus on using disentangled representations to capture the diverse intents behind user preferences. Although several models use disentangled learning, they typically apply it to interaction data alone, ignoring valuable multimodal side information. This leaves out rich semantic cues available in images and reviews.

Model: Disentangled Multimodal Representation Learning for Recommendation

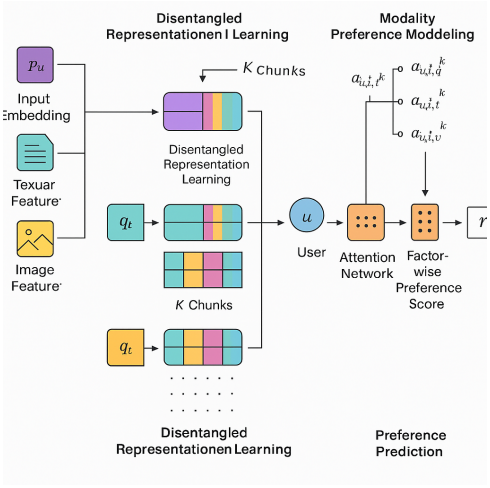
DMRL is composed of three core components:

I. Disentangled Representation Learning

II. Modality Preference Modeling

III. Preference Prediction

Each is elaborated below.



I. Disentangled Representation Learning

To model the fact that users may care about different factors (e.g., appearance, quality, usability) of items, DMRL splits user and item embeddings into K equal-sized chunks. Each chunk is intended to capture a distinct latent factor. However, simply splitting isn't enough, these chunks might still capture overlapping information, i.e., they're entangled.

To fix this, the model uses distance correlation, which is a statistical method that measures whether two chunks (like chunk 1 and chunk 2) are independent. If they're not, it adds a penalty during training.

Regularisation Loss: To enforce independence across K chunks, the total disentanglement loss for a vector y is:

$L_y = \sum_{k=1}^{K} \sum_{k'=k+1}^{K} dCor(y^k, y^{k'})$

Here, dCor measures dependence between chunk k and chunk k'.

Regularisation for All Modalities:
Separate regularisation terms (L_p, L_q, L_t, L_v) are defined for user ID, item ID, textual, and visual embeddings. The final disentanglement loss is:

$L_d = L_p + L_q + L_t + L_v$

In general weights for these can be tuned, but are kept equal in this model.

II. Modality Preference Modeling

Motivation for Modality-Specific Attention:
Different users may value different modalities for the same factor (e.g., some prefer visual cues for appearance, others prefer reviews). Thus, the model introduces an attention mechanism to learn user-specific modality weights per factor.

Multimodal Attention Mechanism: The attention network is weight-shared across factors to reduce complexity.

Attention Scores:

$$\hat{a}_{u,i}^k = W_v \tanh(W.[p_u^k; q_i^k; q_t^k; q_v^k] + b)$$

This outputs raw attention scores for each modality. These are then passed through softmax to get normalised attention weights for each modality.

III. Preference Prediction

- Factor-Wise Preference Estimation:

For each factor k and modality x, the user's preference score is calculated as:

$$r_{u,i,x}^k = a_{u,i,x}^k \cdot \sigma(p_u^k \cdot q_x^k)$$

Here, $a_{u,i,x}^k$ is the attention weight and

$p_u^k \cdot q_x^k$ is the dot product between user and item features for modality x and factor k.

- Aggregating Factor Scores

Scores for a factor are summed across all three modalities:

$$r_{u,i}^k = r_{u,i,q}^k + r_{u,i,t}^k + r_{u,i,v}^k$$

Final Preference Score: $r_{u,i} = \sum_{k=1}^{K} r_{u,i}^k$

- Model Learning

Objective Function is the Pairwise Ranking Loss : The model uses Bayesian Personalised Ranking (BPR) for optimisation. The goal is to ensure that observed (positive) items are ranked higher than unobserved (negative) ones.

BPR Loss: $L_{BPR} = \sum_{(u,i^+,i^-) \in O} -\ln \phi(r_{u,i^+} - r_{u,i^-})$

Here, i^+ is a positive item, i^- is a sampled negative item, and ϕ is the sigmoid function.

Hard Negative Sampling: Among several randomly sampled negatives (n=4), the model chooses the one most similar to the user (based on dot product), ensuring more informative training.

- Optimisation:

The final loss includes:
BPR Loss for ranking, L2 Regularisation,
Disentanglement Regularisation L_d

Total Loss: loss = $L_{BPR} + \lambda_\theta L_\theta + \lambda_d L_d$

λ_θ and λ_d balance the contributions of the regularisation terms.

- Training Details

The model uses Stochastic Gradient Descent (SGD) with Adam optimiser for efficiency and performance. Learning rates and hyperparameters are tuned empirically.

Used Amazon Review Dataset for evaluation;
train-test split:

- 80% of each user's interactions are used for training.

- 20% for testing.
- From the training set, 10% is used as a validation set for hyperparameter tuning.

Why DMRL Performs Best:

- Effective use of multimodal side information.
- Modelling modality contributions at the factor level.
- Multimodal attention to capture user-specific preferences.
- Disentangled learning to reduce noise and redundancy.

The final insight here is that Modelling modality preference and semantic disentanglement jointly leads to more interpretable, flexible, and powerful recommender systems.

Conclusions

In this work, we have addressed two fundamental challenges that have long impeded progress in multimodal machine learning and recommender systems. First, by introducing a task-relative definition of modality, one based on preprocessing representations and the uniqueness of atomic information units. We provide a clear, operational criterion for when two inputs truly constitute distinct modalities. This unified definition dissolves the ambiguities inherent in purely human or machine-centered views and lays a rigorous foundation for all future multimodal research.

Second, through an extensive cross-modality evaluation of five canonical recommender models (CDL, CDR, VMF, VBPR, MCF) across six real-world Amazon datasets, we demonstrate that auxiliary modalities consistently improve recommendation performance even when those modalities differ from a model's original design. This finding overturns the conventional wisdom of modality silos and underscores the untapped potential in repurposing existing architectures beyond their native inputs.

The experiments also reveal a clear modality hierarchy, graph data yields the largest average gains, followed by text, then images while reminding us that domain-specific factors can shift this ranking. Beyond simple fusion, we show that advanced techniques such as co-attention (CAMRec) and disentangled multimodal representations (DMRL) markedly outperform baseline fusion and regularisation schemes, capturing fine-grained, factor-level interactions between modalities.

By breaking down modality silos, unifying notation, and rigorously evaluating models under cross-modality conditions, we chart a path toward more flexible, powerful, and interpretable multimodal recommender systems. We believe that this integrated approach not only elevates recommendation quality but also moves us closer to true multimodal understanding in machine learning where any model can dynamically leverage the richest possible combination of textual, visual, and relational cues.

Acknowledgements

Acknowledging the insights and foundational contributions of the following works, which have informed and inspired this study

- https://www.mdpi.com/2076-3417/14/20/9206
- https://arxiv.org/abs/2203.05406
- https://ieeexplore.ieee.org/document/9354572
- https://arxiv.org/abs/2103.06304
- https://www.mdpi.com/2227-7390/11/4/895