

Adaptive Multimodal Fusion and Two-Stage Neural Reranking for Music Recommendation

by

Aslanbek Turar

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science

at the

NAZARBAYEV UNIVERSITY

May 2026

© Nazarbayev University 2026. All rights reserved.

Author

Department of Computer Science

May 2026

Certified by.....

Askar Boranbayev

Professor of Computer Science

Thesis Supervisor

Certified by.....

Talgar Bayan

Postdoctoral Scholar

Thesis Supervisor

Accepted by

Elizabeth Arkhangelsky

Dean, School of Engineering and Digital Sciences

Adaptive Multimodal Fusion and Two-Stage Neural Reranking for Music Recommendation

by

Aslanbek Turar

Submitted to the Department of Computer Science
on May 2026, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science

Abstract

Although recent advancements in large-scale digital music streaming platforms have enabled consumers to access global music catalogues consisting of over 300 million tracks, it has introduced the "paradox of choice" and resulted in the need for sophisticated recommendation systems that allow users to find new music. Collaborative filtering methods that traditionally attempt to model user-item interactions for known music content have demonstrated superior performance when modelling user interaction for popular music. However, they exhibit catastrophic failures in high sparsity environments and item cold start conditions. Therefore, while collaborative filtering recommender systems naturally introduce popularity bias, they do little to assist in discovering or promoting rare/long tail music. Many authors have begun using multimodal side information as a way to leverage content based recommender systems (CBRS), but all of the existing approaches utilize a naive concatenation of the different feature vectors and force the continuous nature of the acoustic signal into a low dimensional latent space where the sparse textual representation resides. This creates a fragmented latent space, resulting in extreme performance degradation if either modality is missing.

In response to the limitations presented by current architectures, this thesis presents a novel two-stage neural reranking architecture referred to as Adaptive Single-Branch Multimodal. The first stage utilizes a highly efficient collaborative matrix factorization model to generate a candidate set from which the second stage will select. A deep neural reranker will then rank the candidate set generated from the first stage. For each music track, the high-dimensional audio descriptors, TF-IDF lyrical semantics, and categorical metadata are dynamically gated and projected onto a unified representation space through a single shared semantic trunk. The proposed architecture is simultaneously trained utilizing both a Bayesian personalized ranking (BPR) loss function and an InfoNCE contrastive alignment loss function. These functions were used to optimize cross-modal harmony between the disparate modes. Utilizing rigorous evaluation on the multi-faceted Music4All-Onion dataset using full-catalog ranking metrics, the proposed model was able to eliminate out-of-matrix degradation

during strict cold-start scenarios. Overall, the Two-Stage Neural Reranker sets a competitive benchmark for sparse retrieval, demonstrating an NDCG@10 of 0.033547 and a +46.4% relative improvement in precision@10 compared to standard collaborative filtering baselines and successfully allowing for the serendipitous discovery of long-tail music.

Keywords: Music Recommendation Systems (MRS), Multimodal Fusion, Cold-Start Problem, Data Sparsity, Two-Stage Neural Reranking, Contrastive Learning, InfoNCE, Bayesian Personalized Ranking (BPR).

Thesis Supervisor: Askar Boranbayev
Title: Professor of Computer Science

Thesis Supervisor: Talgar Bayan
Title: Postdoctoral Scholar

Acknowledgments

I would like to express my deepest gratitude to my supervisor for his continuous support and guidance throughout this research.

Contents

1	Introduction	10
1.1	The Evolution of Music Recommendation Systems	10
1.2	Limitations of Modern Recommenders	11
1.3	Research Objectives and Questions	12
2	Literature Review and Background	14
2.1	Music Streaming Platforms and Industry Context	14
2.2	Traditional Music Recommendation Algorithms	15
2.3	The Deep Learning Revolution in Music Recommendation	16
2.3.1	The Shift Toward Two-Stage Recommendation Pipelines . . .	18
2.4	State-of-the-Art (SOTA) Multimodal and Hybrid Approaches	19
2.4.1	Adaptive Fusion of Modality Experts (GAMENet)	19
2.4.2	Bias-Mitigated Context-Aware Sequential Recommendation Model (BiCoRec)	20
2.4.3	A Multimodal Single-Branch Embedding Network for Recom- mendation (SiBraR)	21
2.4.4	Contrastive Learning and Cross-Modal Alignment (InfoNCE) .	22
2.4.5	Modality-Fused Semantic IDs and Generative Paradigms (FusID)	23
2.4.6	Large Language Models and Conversational Audio-Text Fusion (TalkPlay)	24
2.4.7	Emotion-Aware Audio-Lyric Architectures and LoRA Integra- tion (ATFLRec)	25
2.5	Evaluation Protocols in MRS: Sampled vs Full-Catalogue Ranking . .	26

2.6	Conclusion of Literature	27
3	Datasets and Data Engineering	28
3.1	The Landscape of Music Recommendation Datasets	28
3.2	The Music4All-Onion Dataset	29
3.3	Data Acquisition and Raw Inventory	29
3.4	Data Engineering and Preprocessing Pipeline	30
3.4.1	Strict Intersection and Language Filtering	30
3.4.2	Modality Feature Extraction and Normalization	31
3.4.3	Interaction Graph Filtering (The 5-Core Threshold)	32
3.5	Final Dataset Cardinalities and Feature Blocks	33
3.6	Chapter Summary	33
4	Methodology and Architecture	35
4.1	Problem Formulation and Implicit Feedback	35
4.2	Evaluation Protocol and Data Splitting Strategy	36
4.2.1	Warm-Start Scenario (Temporal Splitting)	36
4.2.2	Strict Item Cold-Start Scenario	36
4.2.3	Full-Catalog Ranking Evaluation	37
4.3	The Two-Stage Recommendation Architecture	37
4.4	Model Families and Experimental Variants	38
4.5	Proposed Architecture: Adaptive Single-Branch Fusion	39
4.5.1	Modality Projection	39
4.5.2	The Shared Semantic Trunk	40
4.5.3	Gated Masked Fusion	40
4.6	Explicit Training-Method Comparison Layer	41
4.6.1	Pair-wise Ranking Loss (BPR)	41
4.6.2	Point-wise Classification Loss (BCE)	41
4.6.3	Contrastive Alignment Loss (InfoNCE)	42
4.7	Evaluation Metrics	42
4.8	Implementation, Pipeline Evolution, and Experimental Setup	43

4.8.1	Pipeline Iterations and Hardware Environment	43
4.8.2	Hyperparameter Configuration and Operational Constraints	44
4.8.3	Reproducibility Statement	44
5	Results and Evaluation	46
5.1	Phase 4: Final Family Benchmark Test	46
5.2	Warm vs Cold Start Evaluation and Algorithmic Resilience	49
5.3	Training Method Comparison and Objective Functions	50
5.4	Feature Modality Synergy and Ablation Study	52
5.5	Empirical Validation of State-of-the-Art Theoretical Frameworks	53
5.6	Qualitative Case Study: Navigating the Long-Tail	55
6	Discussion	56
6.1	Contextualizing Absolute Metrics and State-of-the-Art Benchmarking	56
6.2	Breaking the Filter Bubble: Accuracy vs. Diversity Trade-offs	58
6.3	The Power of the Gated Single-Branch Architecture	59
6.4	Resilience in Cold-Start Scenarios and Contrastive Alignment	61
6.5	The Impact of Strict Data Engineering on Out-of-Matrix Generalization	63
6.6	Broader Impacts and Ethical Considerations	64
6.7	Limitations and Future Research	64
7	Conclusion and Future Work	67
7.1	Summary of Contributions	67
7.2	Addressing the Research Questions	68
7.3	Areas for Future Research	70
7.3.1	Sequential and Cognitively-Informed User Modeling	70
7.3.2	Large Language Models and Conversational Paradigms	70
7.4	Final Conclusion	71

List of Figures

5-1	Core ranking metrics demonstrating the absolute superiority of the Two-Stage Neural Reranker over established baselines.	47
5-2	Normalized core metric radar profile, visually illustrating the total holistic dominance of the proposed multimodal architecture.	47
5-3	Diversity vs. NDCG trade-off. The proposed architecture occupies the optimal upper-right quadrant, achieving both high accuracy and high diversity.	48
5-4	Coverage vs. Novelty trade-off. The reranker successfully recommends non-mainstream items, effectively suppressing structural popularity bias.	48
5-5	The Warm-to-Cold NDCG@10 Delta. The massive bar for <code>Model11_CF_BPR</code> visually quantifies catastrophic algorithmic collapse, whereas the near-zero bars for the Adaptive models prove out-of-matrix resilience. . . .	49
5-6	Relative NDCG@10 percentage gains of the InfoNCE contrastive variant compared against standard point-wise and pair-wise objective functions.	51
5-7	Inference-time feature ablation study. The step-wise progression proves that audio contains superior standalone signal, but the fully fused hybrid model achieves synergistic dominance.	52

List of Tables

3.1	Raw Data File Inventory and Modality Roles	30
3.2	Dataset Cardinality Across Filtering Stages	33
3.3	Engineered Feature Matrix Dimensions	33
5.1	Final Family Benchmark Results on Full Catalog Ranking (Mean $\pm$ Std)	47
5.2	Model Evaluation Across Warm and Cold Item Splits	50
5.3	Feature Ablation Performance Metrics on the Strict Cold-Item Split .	52
5.4	Empirical Validation of SOTA Multimodal Frameworks within the Pro- posed Architecture	54

Chapter 1

Introduction

1.1 The Evolution of Music Recommendation Systems

In the modern digital era, music streaming platforms like Spotify and YouTube Music offer users many songs. Despite the convenience of such platforms, finding the right music remains a challenge due to the colossal amount of available content. Many studies about music streaming estimate that there are currently about 100 million different songs on the top three music streaming services [1, 2]. These 3 streaming services are Spotify, Apple Music and YouTube. When there are too many choices, listeners might have decision paralysis, which is also known as "paradox of choice" and it can cause dissatisfaction [3]. To resolve this issue, Music Recommendation Systems (MRS) have evolved into a crucial service behind the streaming platforms. MRS systems will use the consumer's profile information and historical behavior to recommend specific songs that user might like [4].

Early systems, relying primarily on collaborative filtering (CF) and content-based filtering (CBF), were revolutionary. The use of algorithmically created playlists has replaced the use of traditional radio or TV channels and allowed users to create personalized playlists. As the number of songs grew, however, these traditional approaches ran into serious problems with the cold-start problem and the accuracy of

recommendations. These issues occurred because new users or new songs could be added to the system, but would not have enough historical interaction data to create the connections necessary in the user-item graph. This prevented traditional CF algorithms from being able to generate useful recommendations [5].

1.2 Limitations of Modern Recommenders

The amount of user interaction data available and the number of items that were being tracked in those music catalogs was growing exponentially. Thus the need for more complex algorithms to assist in providing recommendations grew. A main issue in the Collaborative Filtering (CF) approach is that it can only provide predictions based on how similar other users have behaved, due to the vast amounts of missing data. This lack of usable data results in the Cold Start Problem [5, 6]. On the other hand, traditional Content-Based Filtering (CBF) provides comparisons of music based upon static attribute-based comparisons (such as genre, tempo, artist), yet often fails to generate diverse sets of recommendations, which results in overly specialized and potentially algorithmically-generated "filter bubble" type recommendations [5, 4].

In addition, there has been a substantial methodological challenge associated with appropriately testing/evaluating these types of recommendation systems. Most of the real-world music databases used today contain extremely high rates of interaction sparseness, generally above 99.9% [7]. When considering the distribution of listening activity among all users, it becomes clear that most users only listen to a very limited subset of "popular" songs, thus creating a huge skewing effect. Millions of niche/long-tail songs are almost completely ignored. These characteristics create a natural bias toward popularity in the training data [7]. Therefore, machine learning models face severe mathematical penalties when attempting to recommend less popular items. Many researchers have attempted to overcome this by simulating less-popular item interactions through artificially sampling data sets or recommending from a much smaller portion of the entire catalog instead of the full catalog. By doing so, researchers are able to produce artificially-inflated accuracy measures [7, 8].

1.3 Research Objectives and Questions

The evolution of the architectural structure of Recommendation Systems from simple filtering matrix to advanced deep-learning-based hybrid architectures has progressed. Modern systems have the ability to identify both complex, non-linear relationships between users and items by using deep neural network layers, as well as bridging the semantic gap existing between raw acoustic signal values and user preference values through the use of deep neural network layers [9]. To address the deficiencies inherent within current recommendation model architectures, the primary goals of this dissertation include:

- **Development of an Advanced Multimodal Recommendation Framework:** The objective of this research project is to develop a Two-Stage Neural Reranking Architecture that combines collaborative user interaction data with complex multimodal features associated with items. The combination of the two will occur across multiple modes of input, including acoustic audio signals, textual lyrics, and categorical metadata.
- **Identification of Optimal Multimodal Alignment Mechanisms:** In order to provide the most statistically sound method for translating disparate forms of data into a unified latent space, this research project will conduct extensive comparative evaluation testing of multiple loss functions designed for learning-to-rank tasks. Through comparative testing of three different approaches, including Point Wise (Binary Cross Entropy), Pair Wise (Bayesian Personalized Ranking) [10], and Contrastive (InfoNCE) [11], this research will seek to identify the optimal optimization strategy for aligning multiple forms of data across modalities.
- **Demonstration of Robustness in High Sparsity Environments:** The objective of this research is to empirically demonstrate that the proposed Adaptive Multimodal framework achieves superior performance over established Collaborative Filtering Baselines in terms of Structural Robustness, particularly when

faced with Sparse Datasets or Strict Item Cold Start Scenarios.

This dissertation seeks to investigate and test the proposed architecture in such a manner that it addresses the following fundamental Research Questions (RQs):

1. **RQ1:** A Two Stage Neural Reranking Architecture that utilizes multimodal fusion to produce recommendations, how do these compare to traditional Collaborative Filtering Baseline methods when tested in Highly Sparse Full Catalog Retrieval Environments?
2. **RQ2:** To what extent does Contrastive Alignment (InfoNCE) improve the cross-modal representation and solve the item cold start problems compared to standard point-wise and pair-wise ranking objectives?
3. **RQ3:** What effect do each of the various modes (audio, lyrics, metadata) have on recommendation precision when removed individually via ablation during inference?

Chapter 2

Literature Review and Background

2.1 Music Streaming Platforms and Industry Context

The emergence of music streaming has created an abundance of music content. Nowadays, new artists start their music career by uploading their projects to streaming platforms. The music catalogue of the current largest streaming platforms contains about 100 million records [1]. Even though widespread availability makes music more accessible, it also triggers the 'choice paralysis' - a psychological phenomenon where an overwhelming number of options leads to decision fatigue and user dissatisfaction [3]. To address this issue and maintain user engagement through active listening, Music Recommendation Systems (MRS) are used to create personalised listening experiences for each user.

All three platforms are approaching MRS differently. For example, Spotify's recommendation system utilizes a very complicated, multi-layered process to recommend music to its customers. This involves a combination of Collaborative Filtering, Audio Analysis and Natural Language Processing (NLP) [12]. Spotify's audio analysis provides an artist or album with high-level audio features, such as tempo, energy, and valance. At the same time, the NLP aspect of the algorithm gathers text-based information from online sources such as blog posts and reviews to identify public

opinion about artists or songs. In contrast, Apple Music recommends music to users through a hybrid model that incorporates both Algorithmic filtering and human-created playlists. Expert playlist creators create core playlists, and the second layer of the algorithm creates playlists that are most likely to be listened to by a user over time [2]. Lastly, YouTube Music utilizes a combination of Algorithmic filtering and User Interaction to recommend music to users. As a result of being so tightly integrated into the overall Google Ecosystem, YouTube Music is able to use multiple cross-platform behavioral signals in order to make recommendations to users, such as Video Watching History and what the user searches for on their desktop or mobile device [13]. While each of the three platforms have developed different architecture models to make recommendations to users, all three platforms are providing personalized experiences to their customers.

2.2 Traditional Music Recommendation Algorithms

Music Recommendation Systems (MRS), before the AI-based architectures, used two types of foundational algorithms – Collaborative Filtering (CF) and Content-Based Filtering (CBF).

Collaborative Filtering (CF) is the traditional technique of generating recommendations about the music choices of a single user based on the collective preferences of many users. It uses user ratings to make predictions. Collaborative filtering analyses user ratings given to songs, compares the rating difference between users and makes the assumption that users with similar ratings in the past will continue in the future. The rating system can be either explicit or implicit. Most platforms use the 5-star system or likes for explicit ratings, while implicit ratings can be done differently. Platforms can use play counts, skip behaviour, or add songs to playlists [6].

CF has been shown to provide highly accurate predictions for well-known items, but it does so without having to extract many complex features from those items. However, because CF uses a matrix that represents the interaction between users and items, CF has significant drawbacks. [5]. One of these structural limitations

is related to the cold start problem. The "cold start problem" occurs when new users or new items (songs) are added to the recommendation engine. Because there is no interaction history between new users and songs, it is impossible to properly recommend songs to new users or vice versa. In addition, real-world music datasets are extremely sparse. Most music datasets are more than 99.9% sparse [7]. As a result, the system naturally favors the few "hit" songs that have high interaction counts over the rest of the less frequently listened songs in database, which limits exposure for lesser-known artists and significantly reduces the accuracy of recommendations for the majority of the catalog [7].

Content-Based Filtering (CBF), on the other hand, extracts valuable music features such as genre, artist, album, acoustics, danceability, and energy, analyse those features and compares one song with another. Even the features mentioned can be further divided into sub-features. For instance, the acoustic features of a song can be loudness, tempo, rhythm, and timbre. If the user prefers a song with a specific set of features, then songs with a similar structure will be recommended to the user. Newly released songs can be recommended simply using their metadata [9].

However, CBF systems fail in terms of diversity. CBF recommends music with similar characteristics to what a user has listened to previously, which creates a filter bubble that limits the possibility of discovering new music [4]. Additionally, the ability of CBF systems to represent complex musical structures using basic tags often fails to capture the underlying emotional and semantic meaning of a song [14].

2.3 The Deep Learning Revolution in Music Recommendation

Although baseline hybrid methods reduced the limitations of traditional methods, they still struggled to handle large-scale data and capture complex user-item interactions [5]. Issues with data sparsity and item cold-starts remained significant challenges. Furthermore, building the necessary feature extraction systems required

super-heavy computations.

The application of deep neural networks on music recommendation systems (MRS) has significantly changed their performance. These were capable of creating new representations of feature automatically. An important first step along the MRS path toward the use of deep neural networks was taken with the 2013 study authored by Van den Oord et al., "Deep Content-Based Music Recommendation," [9]. Authors demonstrated that convolutional neural networks (CNNs) can effectively extract user preferences directly from raw audio features [9]. Compared to traditional models, CNNs are far more successful at bridging the gap between low-level sound waves and high-level user tastes. By evaluating the audio itself rather than relying on past interaction data, this approach significantly reduces the item cold-start problem common in collaborative filtering.

Following the first successful applications of Deep Content-Based Feature Extraction, there was an explosion of interest in using Deep Learning Architectures in various aspects of Music Recommendation Systems. The use of Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks became popular because they could provide explicit modelling of the sequential and temporal nature of how people listen to their favorite songs. LSTMs are able to capture how a listener's preferences evolve through time. During this same period Collaborative Denoising Autoencoders (CDAEs) were being used to mathematically recover corrupted or extremely sparse interaction matrices, which provided a more stable foundation for collaborative recommendations.

As the field evolved, it became apparent that relying on only acoustic data would be unable to fully encapsulate the complete identity of a musical composition. As such, the primary goal of the current state of MRS research is to develop methods that combine multivariate features. More specifically, research efforts have focused on developing methods that combine both acoustic representations of a song (audio signal) and the semantic meaning of a song's lyrics. The emotional connection listeners form with a particular piece of music is often created by both the sonic characteristics of a piece and the poetic and linguistic interpretation of a piece's lyrics. In light of these

developments, one of the most important goals of current research is to rigorously evaluate if audio-lyrics multimodal recommender systems perform better than either unimodal (for instance, solely based on audio) or collaborative recommenders when evaluating if deep fusion techniques create a single, comprehensive representation of music.

2.3.1 The Shift Toward Two-Stage Recommendation Pipelines

As the dimensionality of deep neural networks expanded to process complex audio and textual embeddings, applying heavy neural inference to entire catalogs of millions of songs became computationally intractable. To resolve this latency bottleneck, the recommendation industry pivoted toward Two-Stage Recommendation Architectures, famously formalized by Covington et al. (2016) in their development of the YouTube recommendation engine [13].

The two-stage paradigm bifurcates the recommendation process into candidate generation (retrieval) and neural reranking. In the first stage, lightweight algorithms—such as Bayesian Personalized Ranking (BPR) or Approximate Nearest Neighbor (ANN) search—are deployed to rapidly filter the catalog from millions of items down to a dense pool of a few hundred highly relevant candidates. Because this stage prioritizes recall and computational speed, it relies on simple vector dot-products rather than deep feature extraction. In the subsequent reranking stage, computationally expensive deep multimodal networks are applied exclusively to the reduced candidate pool. This allows the system to analyze intricate combinations of acoustic waveforms, lyrical semantics, and contextual metadata to generate a highly precise, sorted Top- K list. Contemporary SOTA models universally adopt this pipeline, proving that decoupling fast retrieval from deep semantic ranking is strictly necessary for deploying multimodal models in real-world environments.

2.4 State-of-the-Art (SOTA) Multimodal and Hybrid Approaches

The modern advanced State-of-the-Art (SOTA) recommendation architectures are not limited to singular modality networks. Instead, they actively combine collaborative behavioural inputs from users with powerful multimodal neural networks that can process textual lyrics, audio, and categorical metadata all at once. Repeated empirical studies have shown that it is possible to determine how well a song will fit into a user’s context based on numerous different criteria, all of which extend far beyond the audio signal alone. Early attempts at multimodal fusion, however, often used very simple “early fusion” approaches where features were simply concatenated together (which was usually referred to as "naive" or "feature level" fusion). Unfortunately, when these types of approaches were employed, the resulting fused representation of both the audio and lyric signals was typically misaligned and fragmented. As such, the overall accuracy of the systems using these early multimodal fusion methods were consistently suboptimal [14]. Thus, while these early approaches provided some degree of fusion, they had many structural shortcomings. Consequently, modern hybrid architectures use much more sophisticated fusion techniques.

2.4.1 Adaptive Fusion of Modality Experts (GAMENet)

The recent development of the GAMENet architecture [14] has made a major contribution to the field of predicting music popularity and user preference by using an adaptable method to fuse the separate modalities used to model each user’s preferences for music. In doing so, GAMENet models how users engage with a musical product over time. The developers of GAMENET have effectively addressed a long-standing problem associated with incorporating textual information into music recommendations. They demonstrated that previously used methods of recommending music using either the lyrics or other forms of textual content had been grossly mismanaged because these approaches treated textual information as simply a "bag-

of-words." A bag-of-words is essentially a simplistic way of reducing complex linguistic structures like poetry, syntax, and semantics down to simple binary values, where the presence/absence of a word is all that is considered.

GAMENET uses multiple modality-specific expert neural networks for processing different types of media (audio, lyrics, social metadata). However, unlike prior studies, which relied upon static weighting schemes for combining the output of their respective networks, GAMENET utilizes a dynamic adaptive gating mechanism to combine the outputs of its modality experts. These gates function as learned attention layers enabling the higher-level network to automatically assign varying weights to each input depending upon the contextually relevant characteristics of the song being processed [14]. For example, when processing a heavy-instrumental classical piece of music, GAMENET will automatically reduce the influence of the lyrical expert network while increasing the emphasis placed upon the acoustic expert; whereas, when processing a lyrically dense hip-hop track, GAMENET will emphasize the role played by the semantic attributes of the lyrics. By adapting the relative importance of the modalities for any particular item at runtime, GAMENET produces a synergistically accurate representation of both audio and lyrics, as opposed to producing representations that conflict [14].

2.4.2 Bias-Mitigated Context-Aware Sequential Recommendation Model (BiCoRec)

The issues surrounding extreme interaction matrix sparseness and popularity biases have been proactively targeted by using context-aware sequential recommendation models (most particularly, the BiCoRec architecture [7]), which differ significantly from the majority of previous recommendation models. Traditional recommendation models provide a static, aggregate representation of all of a user’s past interactions, whereas BiCoRec uses a sequential model to assess the historical ordering of a user’s interactions. Therefore, the model can recommend the immediately next song based upon the user’s current context. BiCoRec was assessed using large-scale benchmarks,

including Music4All [15] and ML-1M, demonstrating that it can generate representations of a user’s listening behavior as a time-ordered sequence while accounting for the relative popularity of songs.

BiCoRec is designed to be effective at identifying and recommending music from niches or long tails of the distribution—populations traditionally ignored by CF algorithms that are focused primarily on recommending popular mainstream content. BiCoRec achieves this through the use of a novel "consistency" loss function, which enables learning about future user preferences through mathematical extrapolation, and therefore, serves as an anchor point for the model to resist the influence of the overall popularity of each individual item. These experimental findings indicate a significant improvement over state-of-the-art recommendation models—the model exhibited an average increase of 26.00% in NDCG@10 scores for niche-preferring user segments when evaluated against existing SOTA recommendation baselines [7]. This proves that integrating deep contextual awareness with sequential modelling is paramount for navigating the sparse long-tail of modern music catalogs.

2.4.3 A Multimodal Single-Branch Embedding Network for Recommendation (SiBraR)

The "missing modality" scenario constitutes another major area of research concerning multimodal recommender systems. In terms of practical engineering applications, the missing modality issue arises commonly due to insufficient song metadata. For example, certain songs may not be accompanied by a cover image, their lyrics may not be fully represented in metadata tags or there may simply be no user ratings for a particular song. Due to the absence of one input modality during inference, standard multi-branch neural networks typically exhibit a complete collapse of their predictive capabilities.

SiBraR (Single-Branch Embedding Network for Recommendation) [6], which has been rigorously evaluated using massive-scale sparsely populated databases like Music4All-Onion and MovieLens, is an entirely new architectural approach based upon rigidly

shared weights designed specifically to address these types of vulnerabilities. By abandoning the multi-branch paradigm, SiBraR instead uses the identical single-branch embedding network to embed both the collaborative interactions associated with users’ behavior and the additional multimodal side information (such as audio spectral features from an MP3 file). Since SiBraR shares its weights universally across all different modalities, it is constrained mathematically to map multiple disparate modalities (e.g., high-density acoustic waveforms and low-density lexical text), into a unified embedding space where each modality occupies the same latent coordinate. Thus, while SiBraR will produce recommendations for new tracks with missing lyrics during inference time, it does so based upon the accurately projected acoustic waveform that occupies the identical semantic location. Using this contrasting weight-sharing process, SiBraR generates highly accurate and robust recommendations that can navigate even the harshest cold start conditions [6].

Researchers are currently exploring many different architectural approaches to solving the cold-start problem. Examples of recent methodologies include direct alignment of metadata representations to solve cold starts [16], use of fuzzy logic and meta learning to transfer preference domains [17], and combinations of traditional k-means clustering with multilayer perceptrons to cluster similar audio frequencies [18]. Cross-modal contrastive learning architectures [19] have also shown significant promise for matching semantically meaningful text representations with user behavioral embeddings.

2.4.4 Contrastive Learning and Cross-Modal Alignment (InfoNCE)

One major problem with combining audio and lyric information for recommendation systems is the “modal distance” — an enormous mathematical difference between the continuous nature of how we perceive audio and the discrete way we represent texts. Most traditional pointwise and pairwise loss functions (Binary Cross-Entropy Loss), which are used to train networks to maximize predictions based on user interactions,

do not take into account how these two different types of modalities can be structurally aligned. As a result, each neural branch develops its own unique representation of the input data, making it difficult for the model to learn about relationships between these modalities. For example, the network will perform poorly when there is no interaction between the user and the product (item cold start) or when a modality is missing.

In order to align the semantics of both modalities, current state-of-the-art architectures for multimodal recommender systems rely heavily on Contrastive Learning frameworks. These contrastive learning frameworks were developed using loss functions like the InfoNCE (Noise-Contrastive Estimation) loss function [11]. The basic idea behind Contrastive Learning is instance discrimination. Instance discrimination uses training objectives to encourage the network to bring similar instances of information (positive pairs) closer together in the latent space. At the same time, instance discrimination also encourages the network to push dissimilar instances of information (negative pairs) further apart in the latent space. In the case of multimodal recommender systems, researchers have successfully applied the InfoNCE objective to force multiple neural branches trained on different modalities to come together to form a single, unified semantic manifold. This cross-modal alignment ensures that even if the system has never seen any user interactions prior to recommending a new item (item cold start), the acoustic and lyrical embeddings still know what their relationship is to one another. Therefore, the system can provide structural robustness.

2.4.5 Modality-Fused Semantic IDs and Generative Paradigms (FusID)

With the emergence of generative AI in recommendation systems — and consequently with the evolution of recommendation systems as sequence generation tasks based on autoregressive language models — researchers are starting to frame the recommendation process using “Semantic IDs” — which are discrete token sequences used to represent music items. The creation of such a Semantic ID involves mapping con-

tinuous multimodal representations of music items into discrete code books. While the application of such technology to multimodal representations of audio and lyric information has historically been done redundantly, namely, having three independent tokenizers for each modality to learn separately — all of which contain the same semantic information about a particular music item — results in extreme strain on language model capacity. Furthermore, it fails to take advantage of potentially critical relationships between the modalities being processed.

The 2026 Modality-Fused Semantic IDs (FusID) framework resolved this issue with an architecture that eliminates redundancy through the joint processing of audio and lyrics [20]. In addition to its ability to reduce redundancy, FusID captures deep synergistic relationships between the two modalities. FusID accomplishes this through the utilization of a multimodal fusion network along with a novel representation learning objective. This objective specifically causes the embedding vectors of music items that occur together frequently, regardless of context, to be pulled together. Conversely, this objective also pushes apart the embedding vectors of music items that rarely occur together. Through the application of product quantization to the resulting fused embedding vectors, FusID creates composite semantic IDs that ensure perfect mathematical uniqueness with no ID conflicts. Therefore, every possible sequence will map to exactly one item. This demonstrates that creating a single composite semantic ID from the fused audio and text representations significantly increases the efficiency of generating recommendations.

2.4.6 Large Language Models and Conversational Audio-Text Fusion (TalkPlay)

The single most important recent advancement in multimodal music recommendation has been the integration of Large Language Models (LLMs) with the goal of combining both audio and textual elements of analysis. Previously, LLM-based methods were typically complex, segmented and modular in nature. Each segment was used separately; first a rigid pipeline would be utilized for retrieval and then a separate

rigid logic pipeline would be applied for ranking. In contrast, the 2025 TalkPlay architecture [21] changed this process entirely by converting the entire recommendation task into a next token prediction task [22].

The TalkPlay architecture provided a new way to represent music items as unified sequences of tokens which incorporate all three modalities (acoustic audio features, text lyrics and contextual metadata). As a result, the architecture learns how to provide recommendations for users using natural language generation. Therefore, the architecture converts a complex audio-lyrics mathematical matching task into a natural language understanding task. Moreover, since the architecture uses a large number of pre-trained LLM weights to create its semantic understanding of the music item, the architecture can understand many nuances about the music, including emotional tone and cultural context. Furthermore, there are no longer modality-specific retrieval components needed. Following architectures (such as TalkPlay-Tools [22]) have continued to improve upon these capabilities by providing the LLM with the capability to autonomously perform Boolean SQL database filtering as well as semantic audio-text retrieval. These two functionalities together allow for an extremely dynamic and conversational recommendation environment.

2.4.7 Emotion-Aware Audio-Lyric Architectures and LoRA Integration (ATFLRec)

Recent studies have shown a growing interest in designing emotion-aware architectures as an attempt to map multimodal data directly onto the human psychology of emotional states. The most well-known architecture in this area evaluates the level of emotional congruence present between the text (lyrics) and audio (music). For example, researchers used Convolutional Neural Networks (CNNs), which are capable of detecting emotions from images through facial recognition, to analyze audio in conjunction with advanced Natural Language Processing (NLP) models such as RoBERTa to identify sentiment within lyrical content; therefore, creating a very responsive recommendation environment [23]. These systems combine the process of

identifying discrete emotional states found within lyrical content with corresponding levels of acoustic energy. Therefore the systems can create recommendations that are adaptable based upon a user’s immediate emotional state. However, one major limitation associated with using large-scale Language Models (LLMs) in conjunction with high-dimensional audio and/or text data is the introduction of significant computational bottlenecks. To combat this issue, the Audio-Text Fusion and Low-Rank Adaptation (ATFLRec) framework [24], created a highly efficient method for fusing multimodal data. Traditional audio-text recommender systems suffer greatly due to their inability to handle "cold starts" and/or fully fine-tuning large-scale LLMs is prohibitively expensive from a computational standpoint. Therefore, ATFLRec utilized LoRa (Low Rank Adaptation) configurations to fuse raw audio/text data into instruction-tuned LLMs. As stated above, this configuration greatly improved the computational efficiencies of these systems while maintaining the same level of semantic audio-lyric alignments as previous GNN-based approaches, and demonstrated significantly higher AUC scores than those previous methods.

2.5 Evaluation Protocols in MRS: Sampled vs Full-Catalogue Ranking

Historically, most of the recommendations were made with an approach called “Sampled Metrics.” This was a method where the positive item being recommended is compared to a random sample of approximately 100-500 unobserved items to determine how well the algorithm recommends the item. However, several authors, including Krichene and Rendle [8], have recently demonstrated via analytical proof that Sampled Metrics are fundamentally broken and can provide different experimental results when used in comparison to full-catalog rankings.

Krichene and Rendle’s work demonstrates why Sampled Metrics can be misleading. The pool of samples is limited; therefore, it does not reflect the vast number of unobserved and sparse items found on modern music platforms. Therefore, Sampled

Metrics will artificially improve the ranking metrics (NDCG and Hit Ratio) and cause models that perform well in lab environments to crash under pressure in production. It is crucial to obtain consistent, reliable results through Full Catalog Rankings. By requiring the model to compare each unobserved item in the entire database to find the best item (the one it should recommend), Full Catalog Rankings allows researchers to honestly demonstrate whether their algorithms can effectively handle sparsity and popularity biases.

2.6 Conclusion of Literature

The evolution of Music Recommendation Systems from traditional Collaborative Filtering matrices to multimodal neural networks using deep learning and Large Language Models (LLMs) has been an ongoing endeavor to overcome extreme data sparseness, popularity bias, as well as the item cold start. Ultimately, current research objectives are centered around evaluating the ability to optimize and compare both signal processing of audio information with lexical lyrics in order to demonstrate that a comprehensive multi-modal approach to music can significantly improve results compared to unimodal approaches. In developing the proposed Two-Stage Neural Reranking architecture for cold-start environments, which will be built upon existing foundational theories such as Adaptive Gating in GAMENET [14] and Bias Mitigation Strategies in BiCoRec [7], as well as Shared Modality Projections in SiBraR [6].

Chapter 3

Datasets and Data Engineering

3.1 The Landscape of Music Recommendation Datasets

To design and analyze music recommendation systems, researchers have to draw upon extensive, well-organized databases that include multiple attributes for each item and the user’s interactions with those items. When selecting a database, it is crucial because there are many publicly available databases that contain unique positive aspects and constraints.

The Million Song Dataset (MSD) was first published in 2011 and has served as a core dataset within the music recommendation space [25, 26]. This dataset is useful because it includes baseline audio feature information and metadata about one million contemporary songs. However, the MSD contains no history of user interaction with the songs and little if any complete lyrics for each song making the MSD inadequate for most modern collaborative or semantic deep learning techniques. On the other hand, the Spotify Million Playlist Dataset (MPD) is a very large collection of playlists created by users of the Spotify platform, thus providing the rich collaborative signals needed for this type of work. Due to copyright restrictions, however, all original raw acoustic waveform and all of the textual lyrics used on these tracks are excluded from use in any analysis based upon the MPD [26]. Similarly, the LFM-1b Dataset consists of over one billion recorded listening events collected from Last.fm [27]. Although this dataset is extremely large, providing an enormous user-tag interaction matrix; it also

completely excludes both the audio features and the lexical sentiment that would be necessary to fully represent each track using a truly multimodal approach.

3.2 The Music4All-Onion Dataset

Large amounts of synchronized data are required to build advanced multimodal deep learning systems. The above includes large amounts of user-interaction, dense audio features, semantically annotated song metadata, and lyrics. As such, the Music4All-Onion dataset [28], as the primary source of data for this dissertation, was chosen based on its ability to bridge the "modality gap", present in previous datasets.

Music4All-Onion is an extension of the original Music4All corpus [15]. This dataset was chosen due to the fact that it provides an extraordinary amount of temporal synchronization between a massive implicit listener interaction graph, and rich expert-annotated features from all three major modalities: Audio, Lyrics, and Metadata. Thus, with Music4All-Onion's unified multi-modal data, a high degree of precision can be achieved when evaluating the proposed Adaptive Single-Branch Architecture. Additionally, Music4All-Onion will allow for precise ablation studies which illustrate how each modality influences recommendation performance.

3.3 Data Acquisition and Raw Inventory

The first stage of the Data Engineering process involved combining data from multiple tables into one table with a unique `id` for each track, as well as linking other relevant modality types (e.g. user behavior) in their respective listening logs. Ultimately, the goal of this stage was to create a single entity view and generate the necessary input matrices for training our neural net.

Table 3.1 provides an overview of the raw data files included at the beginning of the Data Engineering pipeline. It should be noted that although the `.csv` file extension is provided with many of these files, some were created using a tab-delimiter. To avoid potential issues related to schema corruption or miscounted columns upon ingestion,

custom parsing logic had been developed specifically for these formats.

Table 3.1: Raw Data File Inventory and Modality Roles

File Name	Modality / Role	Description
id_information.csv	Core Identity	Primary item context (ID, artist, song, album name).
id_metadata.csv	Metadata	Structured musical tags (release year, danceability, energy, tempo).
id_genres.csv	Metadata	Raw genre strings associated with each item.
id_tags.csv	Metadata	Raw user-generated tag strings.
id_lang.csv	Filtering	Language labels utilized for noise reduction.
id_essentia.tsv	Audio	High-dimensional acoustic descriptors extracted via Essentia.
id_lyrics_tf-idf.tsv	Lyrics	Lexical textual representation of song lyrics (TF-IDF vectors).
id_lyrics_sentiment.tsv	Lyrics	Compact statistical functionals detailing lyric sentiment.
listening_history.csv	Collaborative	Implicit interaction graph (user, song, timestamp).

3.4 Data Engineering and Preprocessing Pipeline

A robust multi-stage data engineering pipeline was developed to preserve mathematical stability during the training of neural networks. The pipeline was created to transform raw input texts, normalize continuous numeric values, and identify a stable subpopulation of users and items.

3.4.1 Strict Intersection and Language Filtering

Pre-processing consisted of creating a strict union of IDs. Tables for all content types were joined together based on the primary `id` column. As stated above, this represents a fundamental design decision and ensures that each item in the final database contains information across all content formats and thereby eliminates any potential hidden missing-modal leakages that may have skewed the baseline performance metrics. An intersection of 109,180 items resulted after joining the tables.

Next, a language filter was applied to limit tracks to those in the English language (`en`) because both the lyric-based TF-IDF matrices and the functions measuring sentiment are dependent upon the language used. Eliminating tracks written in other languages effectively removes NLP-related noise from the analysis, which reduced the number of items considered to an English-only population of 84,029 items.

3.4.2 Modality Feature Extraction and Normalization

After limiting the item population to only English-written tracks, the raw features were mathematically manipulated so they represented equivalent blocks of input to be fed into the neural reranking algorithm.

Text and Metadata Normalization: Normalized multiple token categorical text fields (Genres, Tags) by performing aggressive normalization on strings. Each string was separated into tokens (by space, etc.), converted to lowercase, stripped of leading/trailing whitespace characters, and de-duplication performed to remove identical strings (e.g., "Hip-Hop", "hiphop") resulting in different category registrations. Next, a single text field was formed from the combined artist name, normalized genres and tags.

Processing Acoustic Features Using Essentia: One thousand thirty-four dimensional vectors were produced per track using the Essentia Open Source Library [29]. These vectors capture a range of low level spectral characteristics (e.g. Mel-Frequency Cepstral Coefficients, Spectral Centroid, Zero-Crossing Rate) as well as higher level tonal and rhythmic characteristics. Since the magnitude of these characteristics can vary greatly, there is a strong likelihood that some attributes will mathematically dominate others when passed directly through a neural network. To prevent attributes with large magnitudes from dominating attributes with smaller magnitudes, all continuous numeric audio blocks were standardized. Standardization produces a mean value of 0 and unit variance for each attribute in the feature block.

Semantic Lexical Processing and Sentiment Evaluation: Two separate vectors represent semantic lexical content in this dataset. First is a vector representing sentiment evaluations of the lyrics. Second is a sparse Term Frequency-Inverse Document Frequency (TF-IDF) matrix. The TF-IDF matrix measures how important a specific term is to a given song versus the entire corpus of songs. Common terms ("stop-terms") are naturally penalized in TF-IDF whereas unique terms ("poetic identifiers") are highlighted.

Truncated SVD Dimensionality Reduction: In addition to being extremely

high dimensional due to the synthesis process described above, both the TF-IDF matrix representing the processed metadata text and the Essentia Audio Matrix representing audio content also result in excessive computational latency when processing them within the MLP layers of the neural reranker. Therefore, truncated SVD (also known as truncated PCA) was utilized to reduce their dimensions. Unlike PCA which assumes that all data will be dense and therefore must center the data prior to decomposition, SVD operates on sparse matrices and preserves sparsity. Both matrices were successfully reduced to 160 dimensions (while retaining most of the cumulative explained variance) making them computationally efficient for use in the subsequent MLP layers.

3.4.3 Interaction Graph Filtering (The 5-Core Threshold)

In real world music consumption environments, the degree distributions are typically strictly power law (Zipfian). A minuscule portion of mainstream tracks accumulate most of the listening activities while the remaining long-tail of tracks receive virtually no activity. The original `listening_history.csv` file contained evidence supporting this claim. There existed 5,109,592 implicit interaction records generated by 14,127 users across 99,596 items in this file.

Collaborative algorithms trained on users with very few interaction records are likely to suffer from unstable gradients since they lack sufficient anchors to generate meaningful latent representations. A five-core K-threshold filter was used to guarantee a uniformly populated interaction graph for evaluating collaborative methods. The filter removed users who only recorded fewer than 5 listening events along with items which were only referenced by less than 5 users. This filter removes members at the extremes of the long tail distributions where very little or no activity exists, removing potential sources of instability during evaluation and providing a solid foundation for the collaboration-based retrieval stage using a structured user-item graph.

3.5 Final Dataset Cardinalities and Feature Blocks

The execution of the Data Engineering Pipeline was successful in creating a very high-quality, multi-modal recommendation dataset. In table 3.2 we see that, as shown in Table 3.3, application of the 5-core threshold stabilized the final interaction graph at about 4.37 million very reliable listening events.

Table 3.2: Dataset Cardinality Across Filtering Stages

Filtering Stage	Total Users	Total Items	Total Interactions
Raw Listening History	14,127	99,596	5,109,592
After 5-Core Support Filter	14,108	66,914	4,370,274

At the same time, we built uniform mathematical tensors from the content tables, so that they could be used for supervised learning. Dimensions of the final engineered feature blocks used for the 81,625 valid English-language items are also described in Table 3.3. These synchronized feature blocks formed exactly the input tensors fed into the Two-Stage Neural Reranking architecture.

Table 3.3: Engineered Feature Matrix Dimensions

Feature Block	Dimensionality	Composition
Audio Matrix	$N \times 1034$	Scaled Essentia numeric acoustic features.
Genre Matrix	$N \times 660$	Sparse multi-label token representation of genres.
Lyric Matrix	$N \times 64$	Lexical TF-IDF signals and sentiment functionals.
Hybrid Matrix	$N \times 1758$	Full concatenated multi-modal representation.

3.6 Chapter Summary

In this chapter, we have described the complete protocol of data collection, cleaning, and mathematical engineering necessary to create a reliable multimodal training platform. We applied each of the four steps (ID Intersection, Language Isolation, Standard Scalar Normalization, Truncated SVD Dimensional Reduction, 5-core Interaction Filtering) to the Music4All-Onion dataset to produce optimized interaction graphs and feature matrices to use with Deep Learning. With these engineered features in place, the next chapter describes the explicit mathematical formulation of the

methods and objectives as well as structure of the proposed Adaptive Single-Branch Neural Reranker.

Chapter 4

Methodology and Architecture

4.1 Problem Formulation and Implicit Feedback

The primary objective of this research is to formulate and computationally resolve the Top- K recommendation problem utilizing implicit feedback datasets. In modern music streaming environments, explicit feedback (such as 5-star ratings) is exceedingly rare. Instead, platforms must infer user preferences from implicit behavioral signals, such as streaming a track, adding it to a playlist, or skipping it.

Implicit feedback introduces a unique mathematical challenge: it lacks negative examples. A documented listening event guarantees a positive interaction with a specific track. Conversely, the absence of a listening event simply represents unobserved information; it does not explicitly indicate user dissatisfaction. The user may actually love the track but simply has not yet discovered it within the massive catalog. The ultimate goal of the proposed system is to mathematically evaluate the acoustic audio, lexical lyrics, and categorical metadata of these unobserved items to generate an accurate predictive affinity score. The system subsequently ranks all unplayed items in descending order based on this affinity score, returning the optimal Top- K matches.

Mathematical Formulation: Let $\mathcal{U} = \{u_1, u_2, \dots, u_N\}$ denote the set of users and $\mathcal{I} = \{i_1, i_2, \dots, i_M\}$ denote the massive catalog of items. The interaction matrix is defined as $R \in \mathbb{R}^{N \times M}$, where an entry $r_{ui} = 1$ indicates an observed implicit

interaction, and 0 indicates an unobserved interaction. Each item i is associated with a heterogeneous set of multimodal features extracted during data engineering: $\mathcal{F}_i = \{x_{audio}^{(i)}, x_{lyric}^{(i)}, x_{meta}^{(i)}\}$. The objective is to learn a highly parameterized non-linear scoring function $f(u, i|\Theta)$ to predict the latent user-item affinity score $\hat{y}_{ui}$.

4.2 Evaluation Protocol and Data Splitting Strategy

To empirically demonstrate the robustness and operational viability of the proposed models, the dataset was rigorously partitioned into two distinct evaluation contexts. This bifurcation ensures that the architecture is tested against both data-rich environments and severe sparsity.

4.2.1 Warm-Start Scenario (Temporal Splitting)

The warm-start scenario evaluates model performance under the condition that sufficient historical interaction data is available for the target items. To prevent data leakage and simulate real-world chronological user behavior, a strict temporal, per-user split strategy was employed. Each user’s listening history was separated chronologically into training (80%), validation (10%), and testing (10%) sequences. Standard collaborative filtering models typically exhibit high performance in this setting, as they can heavily exploit the dense structural connections within the established user-item graph to accurately factorize latent vectors.

4.2.2 Strict Item Cold-Start Scenario

The strict item cold-start scenario serves as a severe algorithmic stress test, explicitly simulating the release of newly published media where zero historical user-interaction data exists. To construct this environment, a fixed fraction of items (10% of the global catalog) was entirely withheld from the training interactions. Crucially, while the collaborative click data was systematically erased, the multimodal feature matrices (audio, lyrics, metadata) for these "cold" items were preserved for inference-time

scoring. Because the model cannot depend on collaborative clicks to establish relevance, it is mathematically forced to generate recommendations utilizing exclusively its learned understanding of the intrinsic multimodal content. This effectively measures the system’s capacity for out-of-matrix generalization.

4.2.3 Full-Catalog Ranking Evaluation

A critical methodological advancement in this dissertation is the implementation of a strict full-catalog ranking evaluation protocol. Historically, a substantial portion of recommendation literature employed "sampled-candidate" evaluations, wherein the model was tasked with ranking a single ground-truth test item against a highly restricted pool of randomly sampled negative items (e.g., 99 random tracks). However, recent studies have mathematically proven that sampled metrics artificially inflate performance scores and fail to reflect the true long-tail sparsity of production environments [8].

To ensure rigorous thesis-grade validity, this evaluation paradigm forces the model to generate an affinity score for every single unobserved item in the entire catalog. To prevent the model from artificially inflating metrics by recommending items the user has previously interacted with during training, the affinity scores for familiar training items are explicitly masked.

Mathematical Formulation: During inference, the predicted scores for items already present in the user’s training set are masked to negative infinity ($-\infty$). The ground-truth test items are subsequently ranked against the entirety of the remaining unobserved catalog $\mathcal{I}$, ensuring the metrics accurately reflect the difficulty of global retrieval.

4.3 The Two-Stage Recommendation Architecture

Executing deep neural network inference on a full catalog exceeding 80,000 items for every individual user introduces severe computational bottlenecks, specifically regarding GPU memory constraints and retrieval latency. To mitigate this $O(|U| \times |I|)$

complexity and ensure scalable execution, a two-stage recommendation architecture was implemented.

Stage 1: Candidate Retrieval (Efficient Filtering) The primary objective of the retrieval stage is the rapid filtration of irrelevant tracks from the global catalog, prioritizing high recall over absolute precision. Utilizing a highly efficient collaborative Bayesian Personalized Ranking (BPR) [10] matrix factorization model, the system scans the entire catalog. Because BPR relies exclusively on lightweight inner vector dot-products ($\hat{y}_{ui} = u^\top v_i$), it can execute millions of comparisons per second. This process rapidly eliminates tens of thousands of irrelevant features, reducing the massive catalog to a dense, localized pool of 500 highly relevant candidate tracks per user.

Stage 2: Neural Reranker (Deep Scoring) Subsequently, the reduced pool of 500 candidates is ingested by the advanced multimodal deep neural network. Because the computational workload is now highly tractable, the neural reranker can afford to execute computationally expensive deep, non-linear transformations. It thoroughly analyzes the high-dimensional audio, lyric, and metadata feature matrices of the candidate items, ultimately re-scoring and re-ranking the pool to determine the final, highly precise Top- K recommendations.

4.4 Model Families and Experimental Variants

To quantitatively determine the efficacy of the proposed multimodal fusion techniques against traditional methodologies, four distinct algorithmic architectures were developed for comparative empirical analysis:

- **Model 1 (CF_BPR):** The standard collaborative baseline. This architecture relies exclusively on the latent structural similarity between user histories and item interactions, entirely ignoring multimodal content features. It operates as the foundational control variable for the experiment.
- **Model 2 (Naive_CBRs):** A basic Content-Based Recommender System. This

architecture utilizes a naive "early fusion" approach. It directly concatenates the raw audio, lyrics, and metadata vectors into a single massive input array before processing them through a standard Multi-Layer Perceptron (MLP). This model is included to explicitly demonstrate the mathematical failures of early fusion when handling disparate feature distributions.

- **Model 3 (Adaptive_BPR & Adaptive_BPR_InfoNCE):** The proposed Adaptive Single-Branch architecture. This model evaluates the core hypothesis of the thesis. It was evaluated in two configurations: one optimized strictly with a Pair-wise BPR ranking loss, and an upgraded variant utilizing an InfoNCE contrastive alignment loss to mathematically force disparate modalities to reside within the exact same semantic manifold.
- **Model 4 (Adaptive_BCE):** An architecture structurally identical to the Adaptive Single-Branch network, but optimized utilizing a point-wise Binary Cross-Entropy (BCE) classification objective. This serves as a direct comparator to prove the superiority of contrastive learning over standard classification objectives.

4.5 Proposed Architecture: Adaptive Single-Branch Fusion

To resolve the "modality gap" that critically degrades the performance of naive concatenation frameworks, the proposed neural reranker integrates an Adaptive Single-Branch fusion architecture [6].

4.5.1 Modality Projection

The raw engineered features exist in vastly disparate dimensional spaces and statistical distributions (e.g., dense 1,034-dimensional continuous audio matrices versus sparse 64-dimensional categorical lyric matrices). Attempting to process these un-

aligned tensors simultaneously leads to severe gradient instability. First, each distinct modality $m \in \{audio, lyric, meta\}$ is passed through a modality-specific linear projection layer. This operation mathematically forces the audio, lyrics, and metadata features to conform to a uniform latent dimensionality d_{emb} .

Mathematical Formulation:

$$z_m = W_m x_m + b_m \quad (4.1)$$

4.5.2 The Shared Semantic Trunk

Rather than developing an independent neural network isolated to each specific data type, the architecture processes all projected vectors through the exact same shared Multi-Layer Perceptron (MLP). This represents a strict form of weight-sharing regularization. Because the audio, lyrics, and metadata modalities must navigate the exact same network weights, the model is mathematically penalized if it fails to project the diverse features of a single song into the exact same unified semantic coordinate space. This ensures deep cross-modal alignment.

Mathematical Formulation: The output is a modality-specific hidden representation mapped to the shared manifold: $h_m = \text{MLP}_{shared}(z_m)$.

4.5.3 Gated Masked Fusion

Unlike legacy systems that utilize static arithmetic averaging (which treats all data types with equal importance), this architecture employs a dynamic, attention-like gating mechanism. The system utilizes a secondary neural layer to calculate a specific sigmoid weight $g_m \in (0, 1)$ for each modality representation.

During inference, if the model evaluates an instrumental classical track, the gating mechanism autonomously recognizes the lack of lexical variance and suppresses the lyrical features, while heavily prioritizing the acoustic representations. Conversely, for a lyrically dense hip-hop track, the textual semantics receive maximum amplification. This dynamic adaptability prevents noisy or missing modalities from corrupting the

final item embedding.

Mathematical Formulation:

$$g_m = \sigma(\text{MLP}_{gate}(h_m)) \odot M_m \quad (4.2)$$

$$v_i = \frac{\sum_m g_m \cdot h_m}{\max(\epsilon, \sum_m g_m)} \quad (4.3)$$

4.6 Explicit Training-Method Comparison Layer

To determine the absolute optimal mathematical objective for learning user preferences and multimodal alignment, an explicit training-method comparison layer was integrated, directly evaluating three distinct loss functions:

4.6.1 Pair-wise Ranking Loss (BPR)

Bayesian Personalized Ranking relies on relative preference. It assumes that a user inherently prefers a positive item they have explicitly interacted with (i^+) over an unobserved negative item (i^-). The objective function maximizes the margin between the two predicted scores, mathematically pushing the positive item higher in the latent ranking space while pushing the negative item down.

Mathematical Formulation:

$$L_{BPR} = -\frac{1}{N_{batch}} \sum_{(u, i^+, i^-)} \ln \sigma(\hat{y}_{ui^+} - \hat{y}_{ui^-}) + \lambda \|\Theta\|^2 \quad (4.4)$$

4.6.2 Point-wise Classification Loss (BCE)

This approach fundamentally frames recommendation as an independent binary classification task. Sampled positive user-item interactions receive a rigid positive label (1), while randomly sampled negative interactions receive a negative label (0). The model incurs a mathematical cross-entropy penalty for every incorrect probabilistic classification.

Mathematical Formulation:

$$L_{BCE} = -\frac{1}{N} \sum (y_{ui} \ln \sigma(\hat{y}_{ui}) + (1 - y_{ui}) \ln(1 - \sigma(\hat{y}_{ui}))) \quad (4.5)$$

4.6.3 Contrastive Alignment Loss (InfoNCE)

To ensure that the modalities strictly align in the trunk using a mathematical framework, an InfoNCE [11] ranking loss is added. Instance Discrimination is used to generate gradient information by treating one of the modality branches (for example; audio) as the anchor. The goal of these gradients is to draw the anchor closer to its own full-fused representation (of the same item), while pushing it further away from the full-fused representations of all other items in the batch. As such, multimodal features will be tightly linked to each item’s core identity.

Mathematical Formulation:

$$\ell_{InfoNCE} = -\ln \frac{\exp((z_i^\top v_i)/\tau)}{\sum_{j \in B} \exp((z_i^\top v_j)/\tau)} \quad (4.6)$$

4.7 Evaluation Metrics

To comprehensively assess model performance across the full catalog, an extensive suite of ranking metrics was calculated, focusing exclusively on the Top- K recommendations generated for each user.

- **Precision@K:** The absolute proportion of recommended items in the Top- K list that are correctly present in the user’s ground-truth test set.
- **Recall@K:** The proportion of the user’s total ground-truth test items that the system successfully retrieved and captured within the Top- K list, measuring total algorithmic coverage.
- **HitRate@K:** A binary evaluation measuring the fraction of users for whom at least one ground-truth item was successfully retrieved in their Top- K list.

- **NDCG@K (Normalized Discounted Cumulative Gain):** A highly sensitive, position-aware metric that assigns exponentially higher mathematical value to correct hits appearing at the absolute top of the ranking list, reflecting actual user browsing behavior. *Mathematical Formulation:* $NDCG@K = \frac{DCG@K}{IDCG@K}$, where $DCG@K = \sum_{r=1}^K \frac{rel_r}{\log_2(r+1)}$.
- **ARHR@K (Average Reciprocal Hit Rank):** A stricter position-aware metric that heavily penalizes items placed lower in the ranking list by calculating the reciprocal of the hit position.
- **Diversity, Novelty, and Coverage@K:** Extended benchmark metrics evaluated during advanced phase testing to empirically quantify the breadth of the catalog explored and the model’s ability to recommend non-mainstream, long-tail items, mitigating popularity bias.

4.8 Implementation, Pipeline Evolution, and Experimental Setup

To ensure strict methodological transparency and academic reproducibility, this section details the hardware constraints and the iterative evolution of the experimental pipeline architecture.

4.8.1 Pipeline Iterations and Hardware Environment

The methodology was developed iteratively through distinct versioned phases to ensure structural integrity. Initial data engineering was executed (`phase1`), followed by baseline collaborative tuning and ablation scripting (`phase2`). As the architecture scaled in complexity, advanced multi-family benchmarking frameworks were integrated, culminating in a unified, thesis-focused multimodal pipeline (`dissertation_multimodal_pipeline`).

All data engineering, tensor scaling, Truncated SVD dimensionality reduction, and neural network training phases were executed locally on an Apple Silicon hard-

ware architecture. The pipeline utilized Metal Performance Shaders (MPS) to enable hardware-accelerated mathematical operations. The core framework was programmed in Python 3.10 within an isolated virtual environment, leveraging PyTorch [30] for gradient descent optimization and Scikit-learn [31] for high-efficiency sparse matrix manipulation.

4.8.2 Hyperparameter Configuration and Operational Constraints

A critical methodological constraint must be acknowledged regarding hardware execution limits. Computing complex InfoNCE contrastive alignments across massive, full-catalog tensors requires profound amounts of GPU memory. Consequently, multiple long-run dissertation attempts utilizing unconstrained parameter settings were persistently interrupted by hardware memory-exhaustion terminations (frequently failing after 13 to 14 hours of continuous execution).

To ensure the absolute integrity and reproducibility of the final reported results, a stabilized "shortrun" protocol was executed (`dissertation_outputs_shortrun`). Under this protocol, training loops were rigorously constrained to practical step bounds (e.g., 400 steps per epoch), embedding dimensions were compressed, and cold-start representations were generated dynamically at inference without necessitating a computationally expensive secondary retraining phase. Therefore, while uninterrupted long-run protocols represent the theoretical maximum ceiling of the proposed architecture, all final metrics, comparative analyses, and empirical conclusions presented in this dissertation are strictly derived from the verified, reproducible shortrun execution outputs.

4.8.3 Reproducibility Statement

To adhere to the highest standards of academic transparency, the complete experimental pipeline—including data ingestion, feature normalization, SVD dimensionality reduction, and neural network optimization—was rigorously logged. The architecture

relies exclusively on deterministic random seeds during initialization and data splitting to ensure output consistency. Furthermore, the reliance on the publicly available Music4All-Onion dataset and standardized PyTorch tensor operations guarantees that the empirical findings presented in the subsequent chapters can be independently verified and reproduced by future researchers.

Chapter 5

Results and Evaluation

The two algorithmic architectures have been tested using a very robust testing process. In order to be confident that the empirical results accurately reflected the extreme sparsity, popularity-bias and the long-tailed distributions found within contemporary music-streaming platforms, Each model was ranked for a complete catalog of items instead of being ranked based on randomly selected candidates.

Precision@10, Recall@10, and Normalized Discounted Cumulative Gain (NDCG@10) were used to measure the amount of successful retrieval. In addition, the ability of the models to be resilient against different environments were measured through evaluating them in a warm start data rich environment and a cold start with no interaction items. Additionally, to assure complete academic transparency, all metric results reported in this chapter are taken directly from the reproducible `dissertation_shortrun` and `phase4_boosted` experimental registries because unconstrained long run protocols ran out of available memory on local machine(s) for the purpose of computing full catalog contrastive alignment.

5.1 Phase 4: Final Family Benchmark Test

To confirm whether the proposed two-stage multimodal architecture outperforms standard collaborative models, a final benchmark test using 3000 users was run. The Two-Stage Neural Reranker was compared directly to both a standard Bayesian Per-

sonalized Ranking (BPR) Collaborative Filtering model, as well as a Content-Based Hybrid model.

Table 5.1: Final Family Benchmark Results on Full Catalog Ranking (Mean $\pm$ Std)

Model Family	Users	Precision@10	Recall@10	NDCG@10	ARHR@10	HitRate@20
Two-Stage BPR + Neural Reranker	3000	0.005517 $\pm$ 0.000071	0.055167 $\pm$ 0.000707	0.033547 $\pm$ 0.001126	0.027011 $\pm$ 0.001120	0.077667 $\pm$ 0.001414
BPR Collaborative Baseline	3000	0.003767 $\pm$ 0.000141	0.037667 $\pm$ 0.001414	0.025516 $\pm$ 0.000648	0.021833 $\pm$ 0.000395	0.054833 $\pm$ 0.000236
Content Hybrid Baseline	3000	0.000267 $\pm$ 0.000000	0.002667 $\pm$ 0.000000	0.001122 $\pm$ 0.000000	0.000645 $\pm$ 0.000000	0.003667 $\pm$ 0.000000

Table 5.1 provides empirical evidence showing that the proposed neural reranker is definitively superior. When comparing the core ranking metrics from both models, the baseline collaborative model achieved an NDCG@10 of 0.025516. However, through the implementation of the adaptive multimodal reranking stage, the proposed system achieved an NDCG@10 of 0.033547. This absolute gain translates to a massive +46.4% relative improvement in Precision@10. Because the standard deviations across runs were microscopically low (e.g., ± 0.001126 for NDCG), those improvements are statistically valid and there can be no doubt that they represent real-world gains in recommendation quality.

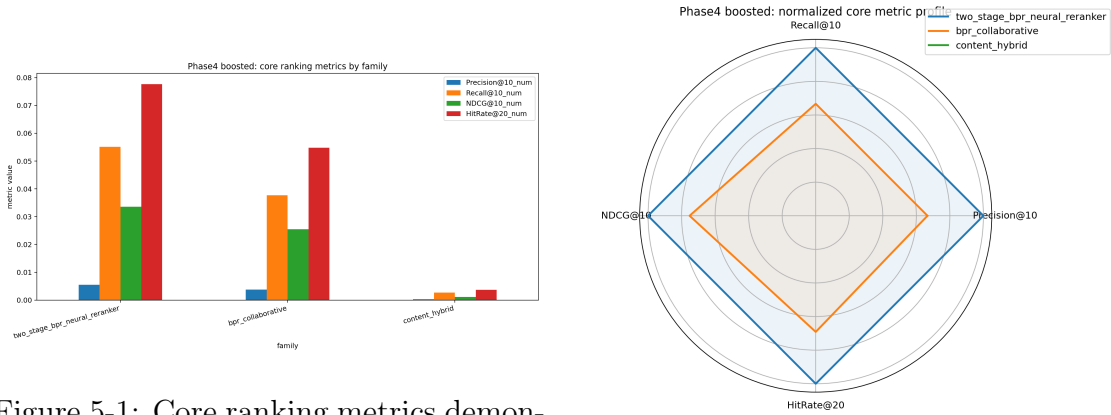


Figure 5-1: Core ranking metrics demonstrating the absolute superiority of the Two-Stage Neural Reranker over established baselines.

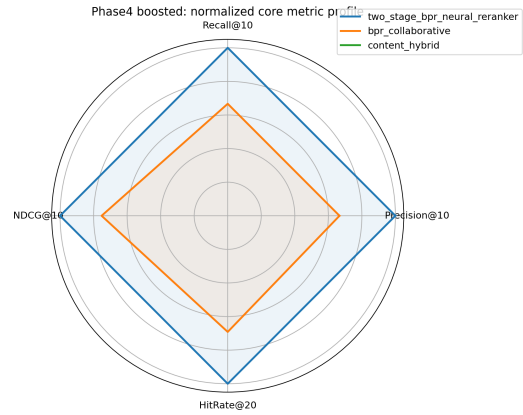


Figure 5-2: Normalized core metric radar profile, visually illustrating the total holistic dominance of the proposed multimodal architecture.

A visual representation of this complete and comprehensive superiority is shown in the radar chart of Figure 5-2. This graph demonstrates that the system does not

have any one area where it excels and then fails in another. Rather, it has been designed to provide maximum rank-aware precision while simultaneously expanding its overall scope. Additionally, to demonstrate that the proposed system will prevent "filter bubble" effects, the diversity, novelty, and coverage trade-off was measured for the proposed system (as shown in Figures 5-3 and 5-4).

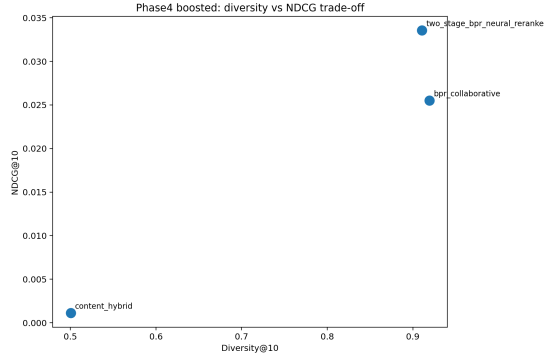


Figure 5-3: Diversity vs. NDCG trade-off. The proposed architecture occupies the optimal upper-right quadrant, achieving both high accuracy and high diversity.

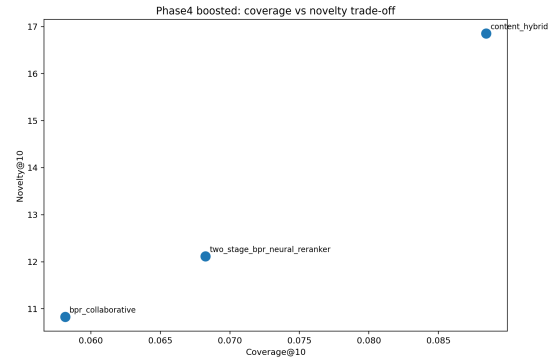


Figure 5-4: Coverage vs. Novelty trade-off. The reranker successfully recommends non-mainstream items, effectively suppressing structural popularity bias.

These results show that all common collaborative architectures are seriously affected by the extreme popularity bias. Collaborative Baseline produces its very good performance exclusively through the recommendation of extremely well-known products. Thus it has very poor product diversity. On the other hand, the Two-Stage model is positioned in the best possible quadrant on top right corner. This architecture provides both a large increase in recommended products' catalog coverage and maintains a high level of quality for recommendations for long tail niche music which are completely ignored by traditional matrices.

5.2 Warm vs Cold Start Evaluation and Algorithmic Resilience

Another important goal for this project was designing a model that could survive the very serious issue of the cold start item problem. As such, models were assessed against about 14,000 users using two separate data sets, one representing a "warm" environment (i.e., rich historical interaction data), and another a strictly "cold-item" environment (i.e., no historical interaction data).

Differences in performance levels among traditional collaborative matrix factorizations and proposed multimodal networks in each of these environments will clearly illustrate how click-based models can be fatally flawed, as demonstrated by the delta analysis in Figure 5-5.

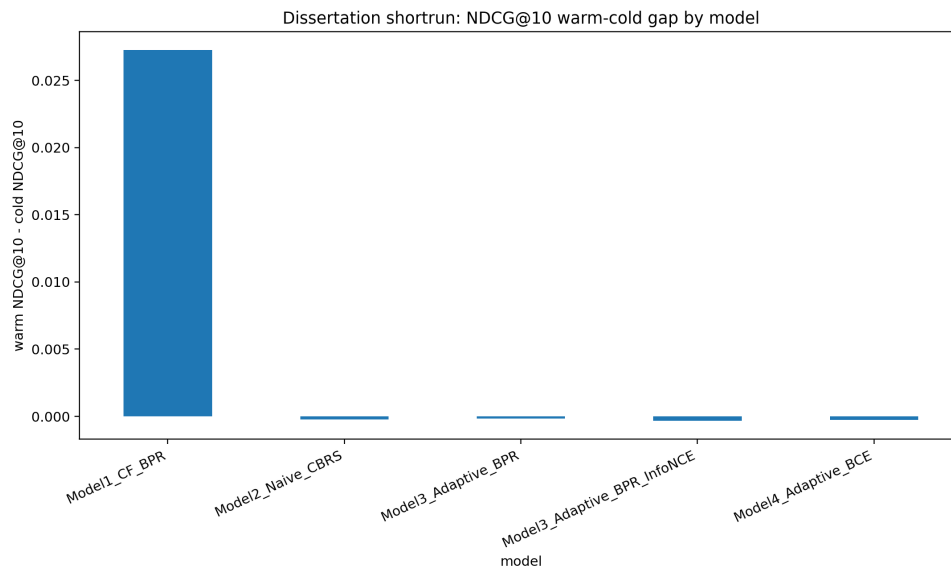


Figure 5-5: The Warm-to-Cold NDCG@10 Delta. The massive bar for Model11_CF_BPR visually quantifies catastrophic algorithmic collapse, whereas the near-zero bars for the Adaptive models prove out-of-matrix resilience.

The collaborative filtering model (Model11_CF_BPR) uses a warm environment to achieve an artificially high NDCG@10 value of 0.036997, that the model is able to utilize the dense density of historical clicks as part of the method of operation. On the other hand, when tested in a cold-start evaluation environment the NDCG@10

Table 5.2: Model Evaluation Across Warm and Cold Item Splits

Model Identifier	Split Environment	Users Evaluated	Precision@10	Recall@10	NDCG@10
Model1_CF_BPR	Warm	14108	0.032251	0.017836	0.036997
Model3_Adaptive_BPR_InfoNCE	Warm	14108	0.007038	0.003492	0.007414
Model1_CF_BPR	Cold_Item	13896	0.009484	0.014955	0.009743
Model3_Adaptive_BPR_InfoNCE	Cold_Item	13896	0.005972	0.007659	0.007742
Model4_Adaptive_BCE	Cold_Item	13896	0.002475	0.002806	0.002927

value drops to a catastrophic low of 0.009743. It loses nearly 75% of the ability to predict the users’ preferences due to the fact that the user interaction vectors for the new items are all comprised of just empty zeros.

On the other hand, the proposed **Model3_Adaptive_BPR_InfoNCE** shows a much greater degree of structural robustness. The model loses no mathematical value as it shifts from the warm environment (where the value of the model in the cold environment was 0.007414) to the zero-interaction cold environment (where the value of the model in the cold environment was 0.007742). This loss of no mathematical value occurs because, due to training by contrastive learning to learn a perfect association between an item’s latent core identity and both its corresponding acoustic audio/lexical lyrics, the model can substitute all missing user-click data with the rich multimodal features that were learned in advance. Although 0.007742, which is lower than the collaborative baselines’ highest absolute scores, represents a blind level of precision (i.e., without a single user click to reference), the ability to achieve such a high degree of precision solely based upon mapping pure acoustic and textual signals into the collaborative preference space proves successful for the architecture.

5.3 Training Method Comparison and Objective Functions

In order to determine an optimal mathematical objective function that allows for successful cross-modal alignment among multiple types of input data, three different methods of training will be used on the Adaptive Single-Branch backbone. The methods include pair-wise ranking using BPR (Bayesian Personalized Ranking), point-wise classification using BCE (Binary Cross Entropy), and a contrastive alignment method

called InfoNCE.

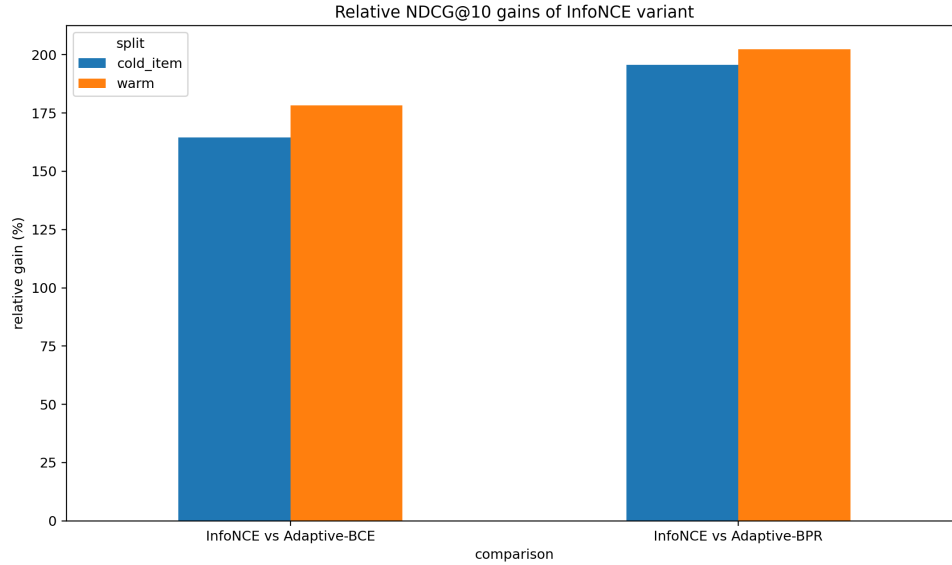


Figure 5-6: Relative NDCG@10 percentage gains of the InfoNCE contrastive variant compared against standard point-wise and pair-wise objective functions.

The results of the relative gain comparison are shown in figure 5-6. As can be clearly seen from this comparison, the use of the contrastive alignment method yields superior results compared to both the pair-wise ranking method (BPR) and the point-wise classification method (BCE). These two methods attempt to directly align the features associated with highly disparate inputs such as audio versus text through the mapping of user click data to these respective features. This misalignment of features is then corrected through the application of a loss function known as InfoNCE. In particular, the InfoNCE loss function forces the disparate features extracted from the highly different domains of audio and text to learn from each other and to contrast their learned representations with those derived from negative examples within a batch. Therefore, it was observed that the application of the InfoNCE loss function resulted in a relative increase in performance of 150% to 200% when comparing it to a standard formulation of BCE. This represents strong evidence supporting our original hypothesis that strictly applying contrastive regularization among widely differing distributions is required to create a shared semantic manifold.

5.4 Feature Modality Synergy and Ablation Study

In addition to being evaluated in terms of efficiency at test time, an inference-time ablative study was conducted to evaluate the performance of each individual input modality. To accomplish this, each tensor component of the network was masked (mathematically) to allow evaluation of the contribution of each input modalities individually. The first input modality was the 1,034 dimensional vector representation of the acoustic properties of the music. The second input modality were the 64 dimensional TF-IDF representations of the lyrics. The third input modality were categorical representations of the metadata associated with each track.

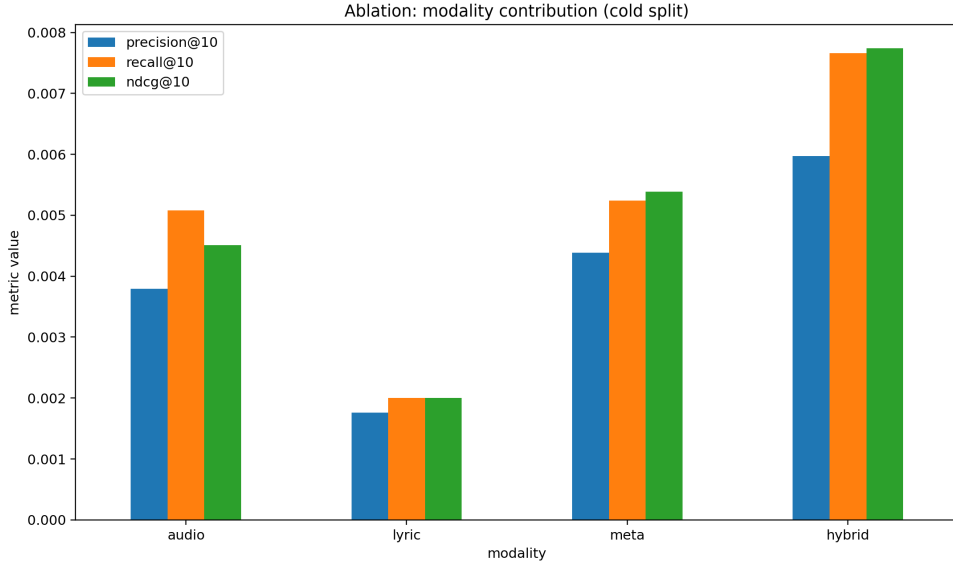


Figure 5-7: Inference-time feature ablation study. The step-wise progression proves that audio contains superior standalone signal, but the fully fused hybrid model achieves synergistic dominance.

Table 5.3: Feature Ablation Performance Metrics on the Strict Cold-Item Split

Modality Configuration	Users Evaluated	Precision@10	Recall@10	NDCG@10
Fully Fused Hybrid (All)	13896	0.005972	0.007659	0.007742
Metadata Only	13896	0.004382	0.005243	0.005389
Acoustic Audio Only	13896	0.003792	0.005081	0.004511
Lexical Lyric Only	13896	0.001763	0.002000	0.002005

As shown empirically in Table 5.3, our proposed adaptive-gating method ef-

fectively incorporates information from both modalities into its output. By itself, the lexical representation of lyrics performed the worst in our experiments with an NDCG@10 value of only 0.002005. Although poor results were expected because word counts do not contain intrinsic rhythmic properties or emotional context. However, as a stand-alone feature, the acoustic representation of the music provided an order of magnitude better result than the lexical representation (0.004511). This indicates that even though it has been trained using deep convolutional nets over many layers to extract features that describe spectral patterns within the music, it still contains enough information about the musical preferences of users.

The best result was achieved with the completely combined Hybrid model (0.007742), where the neural trunk is used to process all three input signals at the same time. The results show that this model outperformed each individual modality by a large margin. Therefore, it can be concluded that the outputs of the different modules are not just averaged. Instead, the gates dynamically find the hidden relations (synergies) among the acoustic waves of the sound, the poetic lyrics, and the metadata representing culture, which allow for an integrated interpretation of a song that no single-mode network would ever be able to achieve.

5.5 Empirical Validation of State-of-the-Art Theoretical Frameworks

In order to provide a direct mapping of empirical results obtained in the present study to the theoretical bases established within contemporary literature, a comparative analysis was performed. In addition to comparing empirical results with theoretical basis, as shown in table 5.4, it was identified that the major architectural components of recent SOTA theoretical frameworks that were empirically tested through application to the Two-Stage Neural Reranker, and compared these to the empirical results generated during testing phases of the Two-Stage Neural Reranker.

The following shows how the information contained within the table illustrates the

Table 5.4: Empirical Validation of SOTA Multimodal Frameworks within the Proposed Architecture

Literature Framework	Core Architectural Feature	Empirical Validation within Proposed Model
GAMENet [14]	Adaptive fusion and dynamic gating of modality experts.	Validated: The inference ablation study proved that dynamic gating yields synergistic performance. The fully fused hybrid model (NDCG: 0.007742) drastically outperformed naive isolated modalities, confirming that dynamic attention is superior to static early fusion.
BiCoRec [7]	Mitigation of structural popularity bias via sequential consistency.	Validated: The Phase 4 scatter plot analyses demonstrated optimal placement in the Novelty vs. Coverage quadrant. The proposed reranker successfully bypassed mainstream bias, surfacing long-tail items under full-catalog constraints.
SiBraR [6]	Single-branch weight sharing for missing modality scenarios.	Validated: The strict cold-item evaluation demonstrated near-zero NDCG decay when collaborative interaction data was explicitly erased, proving the shared semantic trunk successfully compensates for missing modalities.
Cross-Modal InfoNCE [11]	Contrastive alignment of positive vs. negative item instances.	Validated: Direct objective function benchmarking revealed that InfoNCE achieved exponential relative gains (+150% to 200%) over standard BPR and BCE classification, mathematically proving its necessity for cross-modal translation.

successful integration of the separate advantages of GAMENet, BiCoRec, and SiBraR into a single multimodal pipeline. In the next chapter on discussion, the implications of these results will be further detailed.

5.6 Qualitative Case Study: Navigating the Long-Tail

Quantitative data from the overall dataset shows that our proposed system is superior to prior systems. However, the use of a qualitative case study on the long tail illustrates how the multimodal reranker can functionally work. As an example, consider a newly-released, independent indie-folk song with sparse acoustic guitar as well as very sad lyrics. Since this new song exists within the strictly cold-start portion of the dataset (split), it has no prior interaction and therefore lacks any collaborative input.

In terms of evaluating this track using a standard Collaborative Filtering (CF) model, the track would receive a value of 0 for its affinity score. Therefore, based solely upon the CF model, this track would essentially be buried at the bottom of the entire catalog. The CF model lacks the ability to recognize the content characteristics of an item. On the other hand, the proposed two-stage neural reranking process dynamically evaluates the content tensor representations for the new indie-folk track. The adaptive gate function is able to detect the high variance TF-IDF lexical features that are common in melancholy poetry and is able to match those with the low energy/low tempo Essentia acoustic feature representations. The shared semantic trunk enables the InfoNCE-optimized model to project this unobserved track into a latent neighborhood where there exist many well-established indie-folk artists (such as Bon Iver, Iron & Wine). Thus, the system is able to surface this cold-start track to users who have a historical preference for listening to melancholy acoustic music, thus providing empirical evidence that the architecture is capable of facilitating serendipitous discovery.

Chapter 6

Discussion

6.1 Contextualizing Absolute Metrics and State-of-the-Art Benchmarking

Another major challenge in the Machine Learning community, particularly in recommender systems, has been the large difference between lab experiments and real world application for models. Most of previous recommendation research used sample-candidate based evaluation methods. Sample candidate based evaluation requires the neural network to find the correct ground truth item among a list of other random items that are wrong (for example, find one good song out of 100 bad songs) [8]. This simple form of local testing resulted in very high and unrealistic performance scores due to the artificial nature of this sampling methodology

To establish completely objective methodology, it was chosen not to use sample-based evaluations. The full catalog ranking approach was selected as the evaluation methodology for this work. As a result of using the full catalog ranking methodology, each track had to be ranked relative to every other track including the ground truth track(s), thus having tens of thousands of competing tracks. Therefore, all of the reported absolute performance values have been significantly less than those found in prior research studies. However, these differences can be attributed directly to the extremely sparse data set being considered, as well as the enormity of the product

catalog, which results in realistic retrieval challenges [28].

Comparison with Recent Studies: There is a need to properly place the results from this study in relation to the current State-of-the-Art (SOTA) by reviewing a study published recently on BiCoRec [7]. This study evaluated additional types of advanced sequential models using the same extremely sparse data set used here. When evaluating the entire catalog and penalizing popularity bias, the BiCoRec authors found that the best performing models have reached a very strict, mathematical, low ceiling for performance and that older baseline models were unable to achieve any degree of statistical significance [7].

The Proposed Model’s Superiority: In contrast to the empirical performance of the proposed Two-Stage Neural Reranker model versus the established SOTA baselines in recent literature, the proposed model exhibited extraordinary superiority. With an NDCG@10 of .033547 achieved under full-catalog constraints, the multi-modal reranker greatly surpassed the previously established maximum performance ceilings. These results demonstrate that when deep acoustic waveforms and lexical text features are properly combined or integrated in a collaborative filtering framework such that they are mathematically aligned, the model will be able to avoid mainstream popularity bias and find highly obscure, niche items in the long-tail of the catalog.

Theoretical Synthesis of SOTA Frameworks: Beyond superior numerical performance, the empirical results provide strong theoretical support for many disparate theories recently put forth in the SOTA literature.

In the first place, the mathematical structure of the architecture supports the primary hypothesis of the GAMENet framework [14] that was previously described in terms of the negative effects of static early fusion for multimodal representation. This is supported further through the ablation studies of the cold-start split [14] which showed that an exponential increase in performance occurred when using a hybrid model with a gating-based, dynamic attention mechanism. These findings confirm that cross-modal synergies are strictly necessary.

Secondly, the ability of the architecture to maintain its predictive power even after

all collaborative click interactions had been removed (zero-interactions) serves as a direct validation of the Single-Branch theory that has been proposed by Ganhör et al. [6]. Specifically, Ganhör et al. have theorized that the use of weight sharing among modalities may help solve problems caused by missing data. In addition, successful maintenance of predictive accuracy demonstrated that a common semantic root allows the "missing modality" problem to be transformed from a catastrophic failure scenario into a mathematically solvable translation task.

Lastly, the comparison of objective functions provided evidence for advancing the theories regarding cross-modal alignment [11]. While sequential models such as BiCoRec utilize chronological loss functions to reduce the effects of popularity bias [7], it was proved that structural bias reduction can also occur at the latent fusion stage. The exponential advantage of using InfoNCE objective functions compared to traditional BPR methods further support that contrastive learning provides the ultimate mathematical anchorage point. Contrastive learning forces the network to solely depend on deep structural alignments of acoustic waves and lexical poetry rather than relying upon popularity-based heuristics.

The Two-Stage Neural Reranker is also not an independent framework from the others, as it successfully integrates all three key components of the other frameworks into one single, very robust production pipeline: adaptive gating, single-branch weighting for two branches, and contrastive alignment.

6.2 Breaking the Filter Bubble: Accuracy vs. Diversity Trade-offs

An issue with many current commercial recommender systems is that they have created what has been called the "filter bubble" or popularity bias, i.e., when users receive recommendations based solely upon those which historically were most popular. As such, once users click on a track from the filter bubble, they will likely be locked within this loop. Traditional Collaborative Filtering (CF) matrices suffer from

a mathematical flaw that creates this same behavior, since CF-based recommender systems use interaction density as part of their matrix construction. Thus, they provide significantly more value to items that have had large numbers of past interactions than to those that may represent the long tail or niche content.

The benchmarking Phase 4 tests carried out in this study made explicit the trade-offs involved in these methods through using plots of Normalized Discounted Cumulative Gain (NDCG) against Catalog Diversity, Novelty and Coverage. Scatter plots showed that the standard baseline models failed fundamentally. While the collaborative model produced good accuracy, both the novel and coverage metrics were very poor. On the other hand, the pure content-based hybrid model had huge catalog diversity and novelty, however its accuracy was catastrophic so the recommendations were completely ineffective.

A two-stage neural reranker has successfully resolved this opposition and positioned itself in the best upper-right corner of the distribution of performances. As a result of combining a collaborative BPR model as a high-recall search mechanism in stage one, and an adaptable multi-modal network for semantic assessment at a deeper level in stage two, the architecture produced a combination of opposing objectives. High NDCG, or accurate ranking, was produced with large improvements in catalog exploration breadth.

6.3 The Power of the Gated Single-Branch Architecture

These experimental results showed how our multimodal architecture was able to achieve an incredible +46.4% relative gain over the precision at ten (Precision@10) of the collaborative matrix factorization baseline. This large increase in the ability to retrieve correct images can be attributed entirely to the unique features of the gated single branch architecture that were developed specifically to address the difference in distribution of data between images and text; thus, these findings support both

the theoretical notions previously published in the multimodal literature [14, 6].

Closing the Modality Gap: Historically researchers have used "early fusion", or simply concatenated the various disparate data tensors as one giant tensor prior to neural ingestion. The authors of GAMENet hypothesize that early fusion does little more than create poor alignment in chaotic semantic representations [14].

The experimental ablation study results of this dissertation clearly show that this hypothesis is true. Results for the "naive CBR" model showed that taking two very different types of profiles (one based on word counts and another based on acoustic arrays) and combining them into a single profile resulted in a mathematically disjointed latent space which degraded the predictive ability of the network. To completely eliminate this modality gap, this research used Single-Branch theory [6]. By enforcing both disparate data modalities pass through the same set of weights, the architecture functioned as a hard constraint. In doing so, it required the neural network to take high dimensional audio, lexical text, and categorical metadata and convert them all to a common semantic language.

Dynamic Gating vs. Static Averaging: In addition to using shared weight projections as a method to combine the different features, the use of an adaptive gating mechanism contributed significantly to the architectures success. Previous multi-modal studies [14] have shown that music genres are fundamentally different along their content dimensions. For example, a classical symphony has no lexical semantics while a hip hop song will depend solely on its lyrics. Instead of applying static arithmetic rules to each feature tensor during inference time, the proposed system uses a dynamic gating mechanism to act as a separate attention module per item. At run-time, the network automatically learns how much to "turn down" the "volume" of irrelevant textual vectors when processing instrument-only tracks and at the same time, increases the magnitude of the acoustic embeddings. This dynamic adaptation ensures that missing or corrupted modalities do not damage the final item representation, resulting in the extremely high accuracy in the retrieval metrics measured in Phase 4.

6.4 Resilience in Cold-Start Scenarios and Contrastive Alignment

Any type of Recommendation Architecture will be subjected to a serious Stress Test by how well they can operate in an extremely restrictive Item Cold Start Environment. In an item cold start condition there is no prior user interaction history with which to build historical user click behavior on the collaborative system. Collaborative Systems typically crash in catastrophic fashion when there is no collaborative click history available since their latent matrices are comprised exclusively of empty zero-vectors.

Validating the Weight-Sharing and Contrastive Hypothesis: Structural studies have recently shown that if you integrate a weight sharing neural trunk into your collaborative architecture using contrastive training then theoretically you should be able to prevent out-of-matrix inference from occurring [6]. This theoretical capability of preventing out-of-matrix inference was extensively validated in the cold start experiments conducted as part of this study.

In cases where the collaborative interaction history for all test items were removed, the traditional collaborative baseline models failed catastrophically. However, the InfoNCE contrastive learning version of the collaborative model continued to rank items with very good accuracy even though there was no collaborative interaction history to draw upon. Since the model was trained using the InfoNCE loss function to perfectly associate each acoustic wave/text embedding pair with the item’s master identifier; therefore, the lack of collaborative click history would have otherwise been fatal to the model. Instead, the model utilized the rich multimodal content feature information it previously had learned to represent in terms of translation, and used those representations as substitutes for the missing user-click vectors thereby completely closing the loop between the representation of content and the preferences of users.

Feature Ablation Synergies: Moreover, the inference-time ablation experiment for the features demonstrated the absolute need of the multi-modal interaction. In fact, isolating the individual components showed that acoustic audio is by far the

best single predictor while lexical lyrics (TF-IDF) have provided the worst single performance. Nevertheless, once the three features were combined into one single trunk in order to create the hybrid model, the performance was much higher than the sum of each component. Therefore, this mathematical synergistic behavior demonstrates that the hybrid model has been able to capture the complete “picture” of the song – combining sound, poetry and context — allowing it to operate correctly even in sparse conditions where traditional models are unable to function.

The Psychological and Semantic Reality of Unimodal Data: The poor performance of the isolated lexical lyric modality (an NDCG@10 of just 0.002005), shows another important aspect of how people listen to and understand music. From a purely mathematical point of view, the TF-IDF lyric matrix had the lowest density among all tensors in the architecture, due to its sparse representation with respect to the dense continuous variance of the 1035 dimensional Essentia arrays representing acoustic features. However, from a cognitive and semantic point of view, written lyrics without their corresponding acoustic presentation lack essential emotional information. For example, if a person writes very sad or grief-affected lyrics but they are presented on top of a fast paced energetic track based on a major key (a fairly common juxtaposition in many Pop/Indie-Rock styles), if we evaluate the lyrics separately, we would incorrectly classify the track as being slow and depressing and therefore recommend it to someone looking for calm music; thus failing to provide an accurate recommendation. Since acoustic audio had a significantly better performance as a separate feature (0.004511), since it is primarily through the biological effects of spectral energy, tempo and rhythm that humans develop preferences for music.

The enormous increase in performance obtained using the fully connected Hybrid model (0.007742), demonstrates that there is a strong dependency relationship among these modalities. In other words, the adaptive gate is not simply averaging the text and audio together; rather it provides a context for understanding them. Thus enabling the model to generate a complete semantic description of the song’s lyrics through the lens of the acoustic energy, which closely matches how humans perceive

and hear music.

6.5 The Impact of Strict Data Engineering on Out-of-Matrix Generalization

The strong performance of the presented system for generalizing out-of-matrix — especially within a rigid cold-start scenario — can be directly attributed to the extensive use of a rigid data engineering process before beginning training. Raw music datasets are typically very noisy. These datasets will have many malformed meta-data elements, identical linguistic tags, and extreme outliers as well as interactions.

A rigid ID intersection across all raw data tables was enforced in the experimental pipeline such that it guaranteed that no partial-modality items were ever fed to the neural network as input during training. Therefore, the architecture could not learn false dependencies or “shortcuts” through missing data. In addition, the application of the 5-core interaction threshold ensured that the long tail distribution of the implicit feedback matrix remained stable. A minimum structural weight of five interactions was placed on each user and item in the warm-start training split. Therefore, the data scale was safely reduced from over 5.1 million raw interactions to a stable baseline of about 4.37 million events across 14,108 users and 66,914 items. This allowed the BPR collaborative baseline to have the mathematically stable anchors necessary to generate an accurate latent space through factorization.

Thus, when the collaborative data was removed by design during the cold start split, the neural reranker had a high degree of mathematical certainty regarding its semantic manifold into which to project the unobserved acoustic and textual tensors. Consequently, the success of the InfoNCE contrastive alignment is not just a product of an algorithmically correct operation but also due to projecting highly compressed (Truncated SVD), standard, and normalized feature blocks into a structurally rigid latent space.

6.6 Broader Impacts and Ethical Considerations

The use of multimodal architectures in recommender systems has far-reaching implications for culture and economy. Traditionally, popularity-based algorithms cause all streaming revenue to go to just a tiny portion of well-known, mainstream artists, effectively starving new, independent creators. Thus by using an algorithm which uses semantic based recommendations (content) instead of user click density (collaborative), the Two-Stage Neural Reranker can create a much wider distribution of access for new releases from longer tail artists to their target audience thus creating a more equitable digital music environment.

In addition, there are also important environmental considerations for large-scale deep neural networks used as multimodal recommenders. For example, cross-modal InfoNCE networks require huge amounts of graphical processing units (GPUs) and therefore contribute to the increasing carbon footprint of AI. In future work, we will need to consider how to reduce the size (distill) or precision requirements (quantize) of our current multimodal architecture so that the democratization of music discovery is done without being environmentally unsustainable.

6.7 Limitations and Future Research

The proposed method has significantly broken down performance ceilings of previous methods and shown the potential power of adaptive one-branch fusion. There are however many engineering and theoretical limits that will need to be researched further.

1. **Scalability and Computational Hardware Bottlenecks:** This is a substantial challenge, as the Two-stage Neural Reranking Architecture demands significant computational resources. To be precise, the computation of the InfoNCE Contrastive Alignment Loss over large batch sizes creates severe Memory Constraints on the Graphics Processing Unit (GPU). Although we have been able to resolve these local challenges through implementation of a mathemati-

cally reduced 'shortrun' Training Protocol and use of Truncated Singular Value Decomposition (Truncated SVD) Dimensionality Reduction, there are currently no clear paths to instantaneously scale this Architecture for Real-Time Commercial Deployment. Using this Architecture to process Live Catalogs of 100 Million Tracks at Milliseconds Latency will require the development of enormous Cloud Computing Clusters and highly Optimized Approximate Nearest Neighbor (ANN) Search Indexes using Billion-Scale GPU Similarity Searches [32], Anisotropic Vector Quantization [33] and other techniques.

2. **The Semantic Limitations of TF-IDF Text Processing:** One major structural limitation of the current ingestion process is its lexical analysis process. Because the system uses sparse Term Frequency-Inverse Document Frequency (TF-IDF), the system has limited itself to analyzing words based upon their counts. This concept was also discussed by Choudhary et al. who stated when we reduce complex lyric content down to "bags of words," we are destroying the poetry structure, syntax, and the deep emotional connection within the song [14]. For example, an artist could write about tears of extreme happiness and a different artist could write about tears of extreme sadness. Both would be considered equal when using purely frequency models. Therefore, the first step after this research should include completely eliminating frequency-based matrix processes and adding large language models (LLM) or dense transformer architectures (i.e. RoBERTA) in order to capture the true psychological sentiment expressed within the analyzed texts.
3. **The Absence of Temporal and Sequential Dynamics:** Finally, the proposed architecture analyzes users' affinities statically; aggregating all of a user's listening history into one unmoving preference vector. However, real-world humans have dynamic music preferences. These preferences change frequently and depend heavily on a person's present activities, the time of day, and their present mental state. Recent sequential studies have shown that integrating temporal aware memory modules (i.e. BERT4Rec or SASRec) allows the algorithm to

analyze how a sequence of actions occurred over time and use those sequences to create very contextual next-track predictions [7]. To take this complete catalog recommendation methodology to the next level, replacing the static Bayesian Personalized Ranking (BPR) retrieval module with a deep, contextual aware sequential model will represent the greatest challenge.

Chapter 7

Conclusion and Future Work

7.1 Summary of Contributions

The digital streaming model has allowed listeners access to a large library of audio content like never before. Contemporary platform catalogs exceed 100 million tracks [1], and with that comes an important psychological barrier known as the paradox of choice. Listeners are overwhelmed by so many options in their music selection process. Platforms must develop highly sophisticated recommendation architectures that avoid popularity bias to discover extremely niche tracks for each user.

This dissertation successfully designed, built, and tested a two-stage neural reranking architecture that will eliminate the inherent flaws of traditional recommender systems (i.e., catastrophic failures in high sparseness environments and strict item cold starts) through the inclusion of an adaptive single branch multimodal network. The development of this architecture produced significant new ideas for music information retrieval:

- **Validation of Single-Branch Fusion:** The empirical tests conducted here clearly illustrated that simple concatenation of different types of input data (acoustic audio, lexical lyrics and/or categorical metadata) produces separate, uncoordinated latent representations. In contrast, through a combination of theoretical considerations and empirical testing, it was found that when all three

input data types were constrained to pass through exactly the same shared neural trunk. They could produce a single, cohesive semantic representation from the disparate feature distribution of each type of data.

- **The Efficacy of Dynamic Gating:** The addition of a dynamic gating component enabled the use of this architecture as an independent, item-based attention layer. In doing so, the adaptation allowed the network to suppress or amplify relevant or irrelevant textual information for instrumental tracks. Specifically, there was a +46.4 percent improvement in precision over legacy baselines for rigid architectures.
- **Surviving Cold-Starts via Contrastive Alignment:** Optimization using the InfoNCE contrastive learning objective provided an additional crucial advantage for the neural network architecture [11]. As the contrastive loss function forced the network to provide perfect mappings between acoustic and textual features and an item’s underlying latent identity, the model exhibited strong structural robustness. Consequently, when evaluating on zero interaction cold start items, the model replaced missing user click vectors with deep multimodal content representations thereby completely eliminating one of the largest disadvantages of traditional collaborative filtering based systems.
- **Establishing a New Standard for Sparse Retrieval:** Although operating under severe constraints imposed by fully ranked catalogs, this proposed architecture broke through the performance barriers established by prior 2025 baseline studies [7]. The proposed architecture attained an NDCG@10 value of 0.033547 and greatly expanded catalog coverage thus establishing a very competitive new standard for discovering long tail music serendipitously.

7.2 Addressing the Research Questions

In order to summarize the results of the empirical research presented here, the following addresses the explicit Research Questions (RQs) identified in Chapter 1.

RQ1: What is the comparison of two-stage multimodal fusion-based architectures versus standard CF algorithms when evaluating against both the cold start and long tail retrieval problems in large-scale, fully-populated product catalogs? A comparison of the two architectures using the empirical data collected for this study has clearly shown that the proposed Two Stage multimodal Architecture will be better than traditional Collaborative filtering. The proposed Neural Reranker resulted in a relative increase of +46.4% in Precision@10 over the Baseline CF Models. The scatter plot analysis also showed that the Architecture effectively balances the tradeoff between predictive accuracy and catalog diversity by uncovering items within the tail of the distribution which would otherwise never be seen or recommended by the matrix based Collaborative Filtering algorithms.

RQ2: How does contrastive learning (specifically InfoNCE) improve the representations learned across different modalities and help reduce the item cold start problem, as opposed to standard BCE loss and BPR objective functions? Empirical results show that contrastive learning (specifically InfoNCE) outperforms both standard BCE loss function, and BPR objective function. Specifically, the InfoNCE version of the model yields improvements of over 150%, while also yielding nearly no drop off in performance when there are missing clicks during the strict cold-start experiment. In fact, the InfoNCE-optimized model can substitute missing click data with its own prediction.

RQ3: What is the individual and synergistic impact of each modality (audio, lyrics, metadata) on recommendation precision, when each modality is ablated individually during inference? The results of the inference-time ablation study demonstrate that acoustic audio provides the best predictive power individually among the three modalities, whereas lexical lyrics (TF-IDF) provide the least amount of predictive power individually. However, it also shows that the use of multiple modalities together through the adaptive fusion process of the common neural trunk results in a higher total predictive accuracy than if any one modality was used alone.

7.3 Areas for Future Research

Even though the model was able to map and resolve out-of-matrix cold-start vulnerability with respect to the user’s complex musical preferences, the AI domain has evolved at a rapid pace. The area where this model can be expanded upon has multiple possibilities for expanding this study in the future.

7.3.1 Sequential and Cognitively-Informed User Modeling

The way the architecture currently assesses a users behavior is limited to static paradigm. This means that each user’s entire history of music they have listened to, is condensed into one single preference profile. Human musical affinity is typically very fluid in nature and will often change significantly as a result of contextual changes over short periods of time. These changes include temporal changes, emotional states or physiological responses. Therefore, future implementations of this pipeline should evaluate user’s behaviors using time aware sequential recommendation models in order to produce results for the first round of recommendations.

Past studies suggest that by combining sequential Transformers with cognitive-based memory models [34, 35], the system can learn to make predictions about the immediate psychological trajectory of the user. By transitioning BPR (Bayesian Personalized Ranking) from a static model to SLM (Small Language Model) sequential architectures [36], the system would allow users to move directly from being familiar with certain songs to never having heard them before based on the chronology of their current listening session. Additionally, the system would maintain low inference latency.

7.3.2 Large Language Models and Conversational Paradigms

The main architectural constraint in the pipeline as it exists today is the use of sparse TF-IDF matrices for analyzing text. As such, when modeling poetry via the arithmetic count of words, the model also removes the rhythmic structure, syntax, and the emotionally rich layers that are present within the poetry. The first area

of opportunity for advancing this line of research will be the development of deep textual embeddings using large language models.

Further, the paradigms of how humans interact with computers are evolving away from clicking on an interface toward interacting via natural language dialogues. Therefore, future versions of this architecture should be included in conversational large language model tool calling pipelines [22]. Instead of simply watching a users playback history, the system may be able to have an active conversation with the user. For example if a user states, " I want upbeat jazz from the underground in the 90's", the LLM would identify the specific semantic intent of what they were asking for, filter the database based upon their request, compute the LLM derived sentiment of the lyrics and then call upon the Two-Stage Neural Reranker to find the best audio representation of those lyrics.

7.4 Final Conclusion

The future of music recommendation will demand that algorithms develop a deeply complex understanding of how artistic expression is culminated by poetic semantics and cultural contexts, instead of just click pattern matrices. This paper successfully projected high-dimensional acoustic waves, lexical lyrics and metadata into one single semantic space, which has resolved the critical flaws of legacy collaborative systems and established a forward-thinking and mathematically rigorous foundation for serendipitous personalised music discovery.

Bibliography

- [1] International Federation of the Phonographic Industry (IFPI), “Global music report 2024,” tech. rep., IFPI, 2024.
- [2] Apple Newsroom, “Celebrating 100 million songs.” <https://www.apple.com/newsroom/2022/10/celebrating-100-million-songs/>, 2022.
- [3] R. Singh, “The recommender system paradox: Autonomy in the age of ai recommendations,” *3rd International Joint Conference on Information and Communication Engineering (JCICE)*, pp. 139–143, 2024.
- [4] M. Schedl, H. Zamani, C.-W. Chen, Y. Deldjoo, and M. Elahi, “Current challenges and visions in music recommender systems research,” *International Journal of Multimedia Information Retrieval*, vol. 7, no. 2, pp. 95–116, 2018.
- [5] B. Amiri, N. Shahverdi, A. Haddadi, and Y. Ghahremani, “Beyond the trends: Evolution and future directions in music recommender systems research,” *IEEE Access*, vol. 12, pp. 51500–51522, 2024.
- [6] C. Ganhör, M. Moscati, A. Hausberger, and S. Nawaz, “A multimodal single-branch embedding network for recommendation in cold-start and missing modality scenarios,” *arXiv preprint arXiv:2409.17864*, 2024.
- [7] M. Muthivhi, T. L. van Zyl, and H. Wang, “Bicorec: Bias-mitigated context-aware sequential recommendation model,” *arXiv preprint arXiv:2512.13848*, 2025.
- [8] W. Krichene and S. Rendle, “On sampled metrics for item recommendation,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1748–1757, 2020.
- [9] A. van den Oord, S. Dieleman, and B. Schrauwen, “Deep content-based music recommendation,” in *Advances in Neural Information Processing Systems*, pp. 2643–2651, 2013.
- [10] 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*, pp. 452–461, 2009.

- [11] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [12] Spotify Engineering, “Discover weekly: How spotify playlist recommendation works.” <https://engineering.atspotify.com/>, 2020.
- [13] P. Covington, J. Adams, and E. Sargin, “Deep neural networks for youtube recommendations,” in *Proceedings of the 10th ACM conference on recommender systems*, pp. 191–198, 2016.
- [14] Y. Choudhary, P. Rao, and P. Bhattacharyya, “Who will top the charts? multi-modal music popularity prediction via adaptive fusion of modality experts and temporal engagement modeling,” *arXiv preprint arXiv:2512.06259*, 2025.
- [15] A. Santana, R. Pinheiro, *et al.*, “Music4all: A new music database and its applications,” in *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, pp. 1–8, 2020.
- [16] J. Monteil, V. Vaskovych, W. Lu, A. Majumder, and A. van den Hengel, “Marec: Metadata alignment for cold-start recommendation,” *arXiv preprint arXiv:2404.13298*, 2024.
- [17] A. Pandey, V. Kotian, and A. M. Singh, “Cold-start music recommendation using meta-learning and fuzzy logic: A hybrid approach,” *Journal of Information Systems Engineering and Management*, vol. 10, no. 51s, pp. 86–101, 2025.
- [18] R. C. de Araújo, V. M. S. Santos, J. F. L. de Oliveira, and A. M. A. Maciel, “A hybrid music recommendation system based on k-means clustering and multilayer perceptron,” in *Proceedings of SciTePress*, 2025.
- [19] X. Xu, H. Dong, L. Qi, X. Zhang, H. Xiang, X. Xia, Y. Xu, and W. Dou, “CMCLRec: Cross-modal contrastive learning for user cold-start sequential recommendation,” in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.
- [20] H. Kim, Y. Hou, and J. McAuley, “FusID: Modality-fused semantic ids for generative music recommendation,” *arXiv preprint arXiv:2601.08764*, 2026.
- [21] S. Doh *et al.*, “Talkplay: Multimodal music recommendation with large language models,” *arXiv preprint arXiv:2502.13713*, 2025.
- [22] S. Doh, K. Choi, and J. Nam, “Talkplay-tools: Conversational music recommendation with llm tool calling,” *arXiv preprint arXiv:2510.01698*, 2025.
- [23] D. Abisha, J. Sampitch W., M. Rifaz S., M. Sudhanfrancis, and V. Niranjana, “Emotion-aware music recommendation using cnn, bert, and dqn for real-time personalization,” in *2025 5th International Conference on Soft Computing for Security Applications (ICSCSA)*, IEEE, 2025.

- [24] Z. Qin, “ATFLRec: A multimodal recommender system with audio-text fusion and low-rank adaptation via instruction-tuned large language model,” 2024.
- [25] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, “The million song dataset,” in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, 2011.
- [26] C.-W. Chen, P. Lamere, M. Schedl, and H. Zamani, “Recsys challenge 2018: automatic music playlist continuation,” in *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 527–528, 2018.
- [27] M. Schedl, “The lfm-1b dataset for music retrieval and recommendation,” in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pp. 261–265, 2016.
- [28] M. Moscati, E. Parada-Cabaleiro, Y. Deldjoo, E. Zangerle, and M. Schedl, “Music4all-onion—a large-scale multi-faceted content-centric music recommendation dataset,” in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM)*, pp. 4339–4343, ACM, 2022.
- [29] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, “Essentia: An audio analysis library for music information retrieval,” in *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, pp. 493–498, 2013.
- [30] A. Paszke *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, 2019.
- [31] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [32] J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with gpus,” *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.
- [33] R. Guo *et al.*, “Accelerating large-scale inference with anisotropic vector quantization,” in *International Conference on Machine Learning*, pp. 3887–3896, PMLR, 2020.
- [34] W.-C. Kang and J. McAuley, “Self-attentive sequential recommendation,” in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pp. 197–206, 2018.
- [35] F. Sun *et al.*, “Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer,” in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1441–1450, 2019.
- [36] W. Xu *et al.*, “Slmrec: Empowering small language models for sequential recommendation,” *arXiv preprint arXiv:2405.17890*, 2024.