Multimodal Music Retrieval System Leveraging the Music4All-Onion Dataset

First Author
University Name
City, State, Country
first.author@university.edu

Second Author
University Name
City, State, Country
second.author@university.edu

ABSTRACT

The exponential growth of multimedia data necessitates advanced search and retrieval systems capable of effectively handling diverse data modalities. This report presents the development of a multimodal music retrieval system utilizing the Music4All-Onion dataset. The system emphasizes content-driven retrieval by integrating textual, audio, and visual features to enhance retrieval accuracy and user experience. Through comprehensive experimental setups and evaluations, the proposed approach demonstrates superior performance in retrieving semantically relevant music tracks. Additionally, the system addresses beyond-accuracy metrics such as coverage, diversity, and popularity to ensure a balanced and user-centric retrieval process. The findings highlight the potential of multimodal integration in advancing music information retrieval systems.

ACM Reference Format:
First Author and Second Author. 2025. Multimodal Music Retrieval System Leveraging the Music4All-Onion Dataset. In *Proceedings of Proceedings of the 2025 International Conference on Multimedia Search and Retrieval (MMRS '25).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

In the digital age, the proliferation of multimedia content has transformed the landscape of information retrieval, necessitating sophisticated systems capable of handling diverse data modalities. Music, as a multifaceted medium, encompasses not only auditory elements but also textual metadata and visual representations, such as album artwork and music videos. Traditional music retrieval systems predominantly rely on textual metadata or audio features in isolation, limiting their effectiveness in capturing the nuanced relationships between different modalities.

The advent of multimodal retrieval systems aims to bridge this gap by integrating various data types to provide more comprehensive and accurate search results. By leveraging multiple modalities, such systems can better understand the context and semantics of music tracks, leading to improved retrieval performance and a richer user experience. The integration of textual features, such as lyrics and artist information, with audio features like Mel-frequency cepstral coefficients (MFCC) and spectral contrast, enables a more holistic representation of music content.

This report focuses on the development of a multimodal music retrieval system using the Music4All-Onion dataset, which encompasses a wide array of features across textual, audio, and visual domains. The primary objective is to enhance content-driven retrieval by effectively combining these modalities to improve the accuracy and relevance of search results. The system is designed to address both accuracy metrics and beyond-accuracy criteria, including coverage, diversity, and popularity, ensuring a balanced retrieval process that caters to diverse user preferences.

The methodology involves constructing individual retrieval systems based on distinct modalities and subsequently exploring fusion techniques to amalgamate their strengths. Baseline systems employing random retrieval and text-based similarity provide a reference point for evaluating the performance gains achieved through multimodal integration. Comprehensive evaluations are conducted to assess the system's effectiveness across various metrics, providing insights into the trade-offs and synergies inherent in multimodal retrieval approaches.

The subsequent sections of this report detail the experimental setup, including data preprocessing and feature integration, the design and implementation of retrieval algorithms, and the evaluation framework employed to measure system performance. The analysis of results underscores the advantages of multimodal integration in music retrieval and offers guidance for future enhancements in this domain.

2 METHODOLOGY

The development of the multimodal music retrieval system involves several critical steps, including data preprocessing, feature extraction and integration, design of retrieval algorithms, and the implementation of fusion techniques. This section delineates the methodologies employed to achieve content-driven retrieval by leveraging multiple data modalities present in the Music4All-Onion dataset.

2.1 Data Preprocessing

Effective data preprocessing is foundational to ensuring the quality and reliability of the retrieval system. The Music4All-Onion dataset comprises diverse data types, including textual metadata, audio features, visual features, and user-generated tags. The preprocessing pipeline encompasses data cleaning, merging, and normalization to prepare the dataset for subsequent feature extraction and integration.

2.1.1 Data Cleaning and Merging. The dataset was initially segmented into multiple TSV (Tab-Separated Values) files, each containing different aspects of the music tracks. The primary steps in data cleaning involved:

- **Parsing and Standardizing Genres**: The `genre` field, originally a comma-separated string, was converted into a list format to facilitate easier manipulation and analysis. The top

MMRS '25, January 01–03, 2025, City, Country
2025. ACM ISBN 123-4567-24-567/08/06
https://doi.org/10.475/123_4

genre for each track was extracted to serve as the primary relevance criterion in retrieval evaluations.
- **Tag Extraction**: User-generated tags were parsed from dictionary strings into list structures.

Subsequent merging operations combined the various data sources into a unified `catalog_df` DataFrame. This consolidation included integrating metadata, genres, tags, and all feature vectors, resulting in a comprehensive representation of each music track.

2.1.2 Feature Normalization. To ensure comparability across different feature types and scales, feature vectors were normalized using the L2 norm. This normalization step is crucial for similarity computations, particularly when employing distance-based metrics such as cosine similarity.

2.2 Feature Extraction and Integration

The retrieval system leverages multiple feature modalities extracted from the dataset. Each modality contributes unique information, enhancing the system's ability to capture the multifaceted nature of music tracks.

2.2.1 Textual Features.

- **TF-IDF Representations**: Lyrics were vectorized using Term Frequency-Inverse Document Frequency (TF-IDF) to quantify the importance of words within each song relative to the entire corpus. Formally, the TF-IDF weight for term t in document d is given by:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log\left(\frac{N}{\text{DF}(t)}\right)$$

  where $\text{TF}(t, d)$ is the term frequency of t in d, N is the total number of documents, and $\text{DF}(t)$ is the document frequency of t.
- **BERT Embeddings**: Lyrics were further processed using Bidirectional Encoder Representations from Transformers (BERT) to capture contextual semantic information. The BERT embeddings provide a dense vector representation that encapsulates the nuances of language within the lyrics.

2.2.2 Audio Features.

- **Mel-Frequency Cepstral Coefficients (MFCC)**: MFCCs were extracted to represent the short-term power spectrum of audio signals. These coefficients are widely used in audio processing for tasks such as speech and music recognition.
- **Spectral Contrast**: This feature captures the difference in amplitude between peaks and valleys in the spectral envelope, providing insights into the timbral texture of the audio.

2.2.3 Visual Features.

- **VGG19 and ResNet Embeddings**: Visual features were extracted from album artwork using deep convolutional neural networks, specifically VGG19 and ResNet architectures. These embeddings capture high-level visual patterns and aesthetics that can be indicative of the music genre or mood.

2.2.4 Tag Features.

- **Tag Vectorization**: User-generated tags were vectorized using a binary occurrence matrix, where each tag corresponds to a unique dimension. The resulting tag matrix was normalized to ensure uniformity across different tag frequencies.

2.2.5 Feature Integration. Each feature modality was processed and stored as separate matrices. For fusion techniques, feature matrices were combined either at the early stage (early fusion) by concatenating feature vectors or at the late stage (late fusion) by aggregating similarity scores derived from individual retrieval systems.

2.3 Retrieval Systems

The retrieval system comprises multiple retrieval approaches, each leveraging different feature modalities. These systems range from simple baselines to sophisticated multimodal retrieval methods.

2.3.1 Baseline Retrieval.

Random Retrieval. The baseline system randomly selects N tracks from the catalog, excluding the query track. This method serves as a reference point to evaluate the performance improvements introduced by more advanced retrieval techniques.

2.3.2 Text-Based Retrieval.

TF-IDF Retrieval. Utilizing the TF-IDF representations of lyrics, this retrieval system computes cosine similarity between the query track and all other tracks to rank and retrieve the top N similar tracks. The similarity score between two tracks d_i and d_j is defined as:

$$\text{Similarity}(d_i, d_j) = \frac{\mathbf{TFIDF}_{d_i} \cdot \mathbf{TFIDF}_{d_j}}{\|\mathbf{TFIDF}_{d_i}\|_2 \times \|\mathbf{TFIDF}_{d_j}\|_2}$$

2.3.3 Multimodal Retrieval Systems.

BERT Retrieval. This system employs BERT embeddings of lyrics to compute cosine similarity, similar to the TF-IDF approach but capturing deeper semantic relationships.

MFCC Retrieval. Using MFCC feature vectors, this system calculates Euclidean distances to identify the closest audio matches to the query track.

Spectral Contrast Retrieval. Spectral contrast features are utilized to compute cosine similarity, highlighting tracks with similar timbral textures.

VGG19 and ResNet Retrieval. Visual embeddings from VGG19 and ResNet models are used to compute cosine similarity, retrieving tracks with visually similar album artwork.

Tag-Based Retrieval. This system leverages the tag matrix to compute cosine similarity based on user-generated tags, emphasizing semantic relevance derived from tag information.

2.4 Fusion Techniques

To harness the strengths of individual retrieval systems, fusion techniques are employed to integrate multiple modalities, enhancing overall retrieval performance.

2.4.1 Early Fusion. Early fusion involves concatenating feature vectors from different modalities into a single, unified feature vector before similarity computation. Mathematically, the combined

feature vector $\mathbf{f}_{\text{combined}}$ for a track is:

$$\mathbf{f}_{\text{combined}} = [\mathbf{f}_{\text{TF-IDF}}; \mathbf{f}_{\text{BERT}}]$$

where $\mathbf{f}_{\text{TF-IDF}}$ and $\mathbf{f}_{\text{BERT}}$ are the TF-IDF and BERT feature vectors, respectively. Cosine similarity is then computed on these combined vectors to retrieve similar tracks.

2.4.2 Late Fusion. Late fusion aggregates similarity scores from different retrieval systems to compute an overall similarity score. Given two similarity scores s_1 and s_2 from different modalities, the aggregated similarity s_{agg} is:

$$s_{\text{agg}} = \alpha s_1 + \beta s_2$$

where α and β are weights assigned to each modality, reflecting their relative importance. The final ranking is based on the aggregated similarity scores.

2.5 Evaluation Metrics

The retrieval systems are evaluated using a combination of accuracy and beyond-accuracy metrics to assess both the relevance and quality of the retrieved results.

2.5.1 Accuracy Metrics.

- **Precision@k**: Measures the proportion of retrieved tracks that are relevant.

$$\text{Precision@}k = \frac{|\text{Retrieved} \cap \text{Relevant}|}{k}$$

- **Recall@k**: Measures the proportion of relevant tracks that are retrieved.

$$\text{Recall@}k = \frac{|\text{Retrieved} \cap \text{Relevant}|}{|\text{Relevant}|}$$

- **Normalized Discounted Cumulative Gain (NDCG@k)**: Evaluates the ranking quality by considering the position of relevant tracks.

$$\text{NDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k}$$

where DCG@k is the Discounted Cumulative Gain and IDCG@k is the Ideal DCG.

- **Mean Reciprocal Rank (MRR)**: Calculates the average of the reciprocal ranks of the first relevant track in the retrieved list.

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

2.5.2 Beyond-Accuracy Metrics.

- **Coverage@N**: Represents the percentage of tracks that appear in at least one retrieval list across all queries.

$$\text{Coverage@}N = \left(\frac{|\bigcup_{q \in Q} \text{Retrieved}(q)|}{|\text{Catalog}|} \right) \times 100\%$$

- **Tag Diversity@N**: Assesses the average number of unique tags among the top N retrieved tracks.

$$\text{Tag Diversity@}N = \frac{1}{|Q|} \sum_{q \in Q} |\text{Unique Tags}(\text{Retrieved}(q))|$$

- **Genre Diversity@N**: Evaluates the average number of unique genres within the top N retrieved tracks.

$$\text{Genre Diversity@}N = \frac{1}{|Q|} \sum_{q \in Q} |\text{Unique Genres}(\text{Retrieved}(q))|$$

- **Popularity Diversity@N**: Measures the variance in popularity scores among the top N retrieved tracks.

$$\text{Popularity Diversity@}N = \frac{1}{|Q|} \sum_{q \in Q} \text{Var}(\text{Popularity}(\text{Retrieved}(q)))$$

- **Average Popularity@N**: Computes the mean popularity score of the top N retrieved tracks.

$$\text{AvgPop@}N = \frac{1}{|Q| \times N} \sum_{q \in Q} \sum_{t \in \text{Retrieved}(q)} \text{Popularity}(t)$$

2.5.3 Trade-offs and Optimization. The evaluation framework not only measures the accuracy and quality of retrievals but also explores the trade-offs between different metrics. Specifically, the system investigates how optimizing for diversity metrics such as Tag Diversity and Genre Diversity may impact accuracy metrics like NDCG@10. Parameter tuning and controlled experiments are conducted to balance these objectives, ensuring that the retrieval system delivers both relevant and diverse results.

2.6 Implementation Details

The retrieval system was implemented using Python, leveraging libraries such as `pandas` for data manipulation, `scikit-learn` for feature normalization and similarity computations, and `scipy` for efficient handling of sparse matrices. The system architecture is modular, facilitating the addition of new feature modalities and retrieval algorithms. The source code is hosted on GitHub[1] for reproducibility and further development.

2.6.1 Retrieval Algorithms. Each retrieval algorithm is encapsulated within a dedicated function, accepting a query track ID and returning a ranked list of similar tracks based on the respective similarity measures. The algorithms adhere to a consistent interface, enabling seamless integration and evaluation within the overall system.

2.6.2 Fusion Methods. Early and late fusion techniques are implemented as separate modules, allowing for independent experimentation and optimization. The early fusion module concatenates feature vectors before similarity computation, while the late fusion module aggregates similarity scores from individual retrieval systems using weighted combinations.

2.7 Experimental Setup

The experimental evaluation encompasses the following components:

- **Dataset**: The Music4All-Onion dataset, consisting of 5148 music tracks with comprehensive feature sets across textual, audio, visual, and tag modalities.
- **Queries**: Each track in the dataset serves as a query, ensuring an exhaustive evaluation across all entries.

[1] https://github.com/yourusername/multimodal-music-retrieval

- **Retrieval Configurations**: Ten retrieval systems were evaluated, including baseline, text-based, individual modality-based, and fusion-based approaches.
- **Evaluation Metrics**: A combination of accuracy and beyond-accuracy metrics was employed to assess the performance comprehensively.
- **Hardware and Software**: Experiments were conducted on a machine equipped with [specify hardware details, e.g., CPU, RAM], using Python [version] and relevant libraries.

2.8 Justification of Methodological Choices

The selection of feature modalities and retrieval algorithms is motivated by the need to capture the multifaceted nature of music content. Textual features like TF-IDF and BERT embeddings provide insights into lyrical content and semantic relationships, while audio features such as MFCC and spectral contrast encapsulate the acoustic properties of tracks. Visual features derived from VGG19 and ResNet models capture the aesthetic elements of album artwork, contributing to the overall similarity assessment.

Fusion techniques are employed to integrate these diverse modalities, leveraging the complementary strengths of each. Early fusion allows for a unified feature representation, enhancing the discriminative power of similarity computations. Late fusion, on the other hand, facilitates the combination of independent similarity scores, enabling a flexible aggregation strategy that can be tuned based on performance objectives.

The chosen evaluation metrics offer a balanced assessment of retrieval relevance and quality. Accuracy metrics like Precision@10 and NDCG@10 provide quantifiable measures of relevance, while beyond-accuracy metrics such as Coverage@10 and Diversity@10 ensure that the retrieval results are comprehensive and varied, catering to diverse user preferences.

3 RESULTS AND DISCUSSION

This section presents the evaluation outcomes of the developed multimodal music retrieval systems. The results are categorized into quantitative metrics and qualitative analyses, providing a comprehensive overview of each system's performance and capabilities.

3.1 Quantitative Evaluation

The retrieval systems were assessed using a suite of accuracy and beyond-accuracy metrics. Table 1 summarizes the performance of each retrieval approach across these metrics.

3.1.1 Analysis of Accuracy Metrics. The accuracy metrics—Precision@10, Recall@10, NDCG@10, and Mean Reciprocal Rank (MRR)—provide insights into the relevance and ranking quality of the retrieved tracks.

- **Precision@10**: Among all retrieval systems, Tag-Based Retrieval achieved the highest Precision@10 (0.3518), significantly outperforming the baseline Random Retrieval (0.0395). This indicates that tag-based similarity effectively identifies relevant tracks based on shared tags.
- **Recall@10**: Similarly, Tag-Based Retrieval leads in Recall@10 (0.0838), suggesting that it not only retrieves relevant tracks but also covers a substantial portion of the relevant subset within the top 10 results.
- **NDCG@10**: Tag-Based Retrieval also excels in NDCG@10 (0.3906), reflecting its ability to rank more relevant tracks higher in the retrieval list. Other systems, such as Late Fusion MFCC+VGG19 Retrieval (0.1038) and ResNet Retrieval (0.0878), demonstrate moderate performance in ranking relevance.
- **MRR**: The Mean Reciprocal Rank is highest for Tag-Based Retrieval (0.6123), indicating that the first relevant track retrieved by this system appears very early in the ranked list. In contrast, Random Retrieval has a low MRR of 0.0949, underscoring its inefficacy in prioritizing relevant tracks.
- **Overall Performance**: Tag-Based Retrieval consistently outperforms other systems across all accuracy metrics, highlighting the effectiveness of leveraging user-generated tags for music retrieval. Fusion-based systems, particularly Late Fusion MFCC+VGG19 Retrieval, show improved precision and recall compared to single-modality systems, albeit not surpassing the tag-based approach.

Table 1: Final Evaluation Results of Retrieval Systems

Retrieval System	Precision@10	Recall@10	NDCG
Random Retrieval	0.0395	0.0018	0.03
Tag-Based Retrieval	0.3518	0.0838	0.39
Early Fusion TF-IDF+BERT Retrieval	0.0711	0.0053	0.07
Late Fusion MFCC+VGG19 Retrieval	0.0946	0.0081	0.10
TF-IDF Retrieval	0.0600	0.0037	0.06
BERT Retrieval	0.0842	0.0073	0.08
MFCC Retrieval	0.0927	0.0081	0.09
Spectral Contrast Retrieval	0.0633	0.0052	0.06
VGG19 Retrieval	0.0731	0.0063	0.08
ResNet Retrieval	0.0786	0.0060	0.08

3.1.2 Evaluation of Beyond-Accuracy Metrics. Beyond-accuracy metrics such as Coverage@10, Tag Diversity@10, Genre Diversity@10, Popularity Diversity@10, and AvgPop@10 provide a broader perspective on the retrieval system's performance, assessing aspects like the variety and popularity of retrieved tracks.

- **Coverage@10**: Random Retrieval achieves full coverage (100.0000%), meaning every track appears in at least one retrieval list. Tag-Based Retrieval, while effective in accuracy, has a lower coverage (88.8695%), indicating that some tracks are not retrieved by this system.
- **Tag Diversity@10** and **Genre Diversity@10**: Most systems maintain high diversity scores, typically around 9.7 to 9.9, indicating a wide range of tags and genres among the retrieved tracks. Tag-Based Retrieval, however, exhibits slightly lower Tag Diversity@10 (8.7863), suggesting that while it retrieves relevant tracks, they may share similar tags, potentially limiting diversity.
- **Popularity Diversity@10** and **AvgPop@10**: There is noticeable variability in popularity diversity and average popularity across systems. Random Retrieval has the highest

Popularity Diversity@10 (190.0757), reflecting a wide range of popularity levels in its random selections. Tag-Based Retrieval balances relevance and popularity, achieving a higher average popularity (38.1216) with moderate diversity (149.0238). Fusion systems like Early Fusion TF-IDF+BERT Retrieval and Late Fusion MFCC+VGG19 Retrieval strike a balance between diversity and average popularity, ensuring that retrieved tracks are both relevant and varied in popularity.

3.2 Qualitative Analysis

While quantitative metrics provide a numerical assessment of retrieval performance, qualitative analysis offers deeper insights into the system's behavior and the nature of the retrieved results. This subsection explores a specific query example, examining how different retrieval systems respond to the query "Chasing Ghosts" by Against the Current.

3.2.1 Example Query: "Chasing Ghosts" by Against the Current. The following tables present the top three retrieval results from each retrieval system for the query "Chasing Ghosts" by Against the Current, categorized by the similarity metrics employed.

Table 2: TF-IDF Retrieval Results for "Chasing Ghosts" by Against the Current

Song	Similarity Score
Chasing the Sun by Hilary Duff	0.4987
Chasing the Skyline by Brymir	0.4348
Tarzan Boy by Baltimora	0.3481
Ghost of a Chance by Rancid	0.2787
The Haunting by Testament	0.2724
Labels Or Love by Fergie	0.2417
Day of the Dead by Hollywood Undead	0.2344
Tender Is the Night by Jackson Browne	0.2313
Joyride (Omen) by Chevelle	0.2310
Lifted by Birdy	0.2271

TF-IDF Retrieval.

Table 3: BERT Retrieval Results for "Chasing Ghosts" by Against the Current

Song	Similarity Score
Dead Trees by From First to Last	0.7191
Black Suit by Super Junior	0.7112
Disco Down by Kylie Minogue	0.7060
12 Horas by Dilsinho	0.7025
Alone in Love by Mariah Carey	0.6995
Ghost by Motorama	0.6942
Go Away by Strawberry Switchblade	0.6912
Telephone Line by Electric Light Orchestra	0.6859
Want Me by Puma Blue	0.6859
Te Quise Olvidar by MDO	0.6851

BERT Retrieval.

Table 4: MFCC Retrieval Results for "Chasing Ghosts" by Against the Current

Song	Distance
SPRORGNSM by Superorganism	0.3477
High Horse by Kacey Musgraves	0.3548
Souvenir by Avril Lavigne	0.3565
Bad Friends by Black Honey	0.3582
Biscuits by Kacey Musgraves	0.3677
Last Christmas by Carly Rae Jepsen	0.3684
Dear Future Husband by Meghan Trainor	0.3696
Tears Are Not Enough by ABC	0.3724
Love Away by Capital Cities	0.3843
The Internet by Jon Bellion	0.3865

MFCC Retrieval.

Table 5: Spectral Contrast Retrieval Results for "Chasing Ghosts" by Against the Current

Song	Similarity Score
If You Come to Me by Atomic Kitten	0.9989
Torn Down by Brandy	0.9989
Animal by Smash Into Pieces	0.9989
Someone Better by Juveniles	0.9989
Killing The Joke by Miles Kane	0.9989
Wanted Man by The Last Internationale	0.9989
Cumbia Para Olvidar by Mon Laferte	0.9988
Big Money Salvia by Hot Dad	0.9988
Love It When You Call by The Feeling	0.9988
Wither by Dream Theater	0.9988

Spectral Contrast Retrieval.

Table 6: VGG19 Retrieval Results for "Chasing Ghosts" by Against the Current

Song	Similarity Score
The Consequence by You Me at Six	0.7029
Rain by Hollywood Undead	0.6911
Wicked Game by Theory of a Deadman	0.6906
Portobello Belle by Dire Straits	0.6901
Sabes by Reik	0.6849
All You Got by Tegan and Sara	0.6809
Black Dahlia by Hollywood Undead	0.6770
Lay Down by Priestess	0.6749
Inabalavelmente by Charlie Brown JR.	0.6749
Don't Leave Home by Dido	0.6748

VGG19 Retrieval.

Table 7: ResNet Retrieval Results for "Chasing Ghosts" by Against the Current

Song	Distance
Cities of the Future by Infected Mushroom	0.8890
Unbreakable Smile by Tori Kelly	0.9079
Rise and Fall by The Offspring	0.9205
Invincible Force by Destruction	0.9277
Blues Walk by Lou Donaldson	0.9297
Bosco by Placebo	0.9325
Processed Beats by Kasabian	0.9328
Scarlet by Delain	0.9381
A Strange Boy by Joni Mitchell	0.9419
Passing by Leprous	0.9433

ResNet Retrieval.

3.2.2 Analysis of "Chasing Ghosts" Query Results. The query "Chasing Ghosts" by Against the Current, classified under the genre *rock*, was evaluated across different retrieval systems. The following analysis highlights the performance and characteristics of each system based on the retrieved results.

TF-IDF Retrieval. Table 2 displays the top ten retrieval results using the TF-IDF with Cosine Similarity system. The highest similarity score is observed with "Chasing the Sun" by Hilary Duff (0.4987), indicating a moderate lexical overlap between the query and the retrieved song. Subsequent results show decreasing similarity scores, reflecting diminishing lexical similarities.

Observations:

- **Lexical Overlap**: The system prioritizes songs with similar titles or lyrics containing keywords like "Chasing," "Ghosts," or related terms.
- **Genre Consistency**: While the query belongs to the *rock* genre, some retrieved tracks like "Tarzan Boy" by Baltimora and "Labels Or Love" by Fergie may belong to different genres, indicating that lexical similarity does not always guarantee genre alignment.

BERT Retrieval. Table 3 presents the top ten results from the BERT Retrieval system, which leverages contextual embeddings to capture deeper semantic relationships.

Observations:

- **Semantic Similarity**: The highest similarity score is with "Dead Trees" by From First to Last (0.7191), suggesting a stronger semantic connection beyond mere lexical overlap.
- **Diverse Genres**: Retrieved tracks span various genres, including K-pop ("Black Suit" by Super Junior) and pop ("Disco Down" by Kylie Minogue), indicating the system's focus on broader semantic themes.

MFCC Retrieval. Table 4 shows the top ten retrieval results based on MFCC Euclidean distances, which measure acoustic similarity.

Observations:

- **Acoustic Features**: The retrieved songs, such as "SPRORGNSM" by Superorganism and "High Horse" by Kacey Musgraves, share similar acoustic properties like rhythm and melody.
- **Genre Variation**: Similar to TF-IDF Retrieval, the genre diversity is noticeable, with tracks from genres like country and pop appearing in the results.

Spectral Contrast Retrieval. Table 5 lists the top ten results from the Spectral Contrast Retrieval system, which emphasizes timbral textures.

Observations:

- **High Similarity Scores**: All retrieved tracks have similarity scores exceeding 0.9988, indicating near-identical spectral characteristics.
- **Genre Discrepancy**: Despite high similarity scores, genres vary widely, suggesting that spectral features alone may not effectively capture genre-specific nuances.

VGG19 Retrieval. Table 6 presents the top ten retrieval results from the VGG19 Retrieval system, which utilizes visual embeddings from album artwork.

Observations:

- **Visual Aesthetics**: Retrieved songs like "The Consequence" by You Me at Six and "Wicked Game" by Theory of a Deadman likely share similar visual themes in their album covers, such as color schemes or imagery.
- **Genre and Style Diversity**: The system retrieves tracks across different genres, emphasizing visual similarities over musical genre alignment.

ResNet Retrieval. Table 7 lists the top ten results from the ResNet Retrieval system, focusing on visual feature similarity through ResNet embeddings.

Observations:

- **Strong Visual Similarity**: High similarity scores indicate that the retrieved tracks have visually similar album artwork.
- **Variety in Genres**: Similar to other visual retrieval systems, ResNet Retrieval retrieves tracks from diverse genres, reflecting the system's emphasis on visual rather than musical features.

3.2.3 Discussion of Qualitative Findings. The qualitative analysis of the "Chasing Ghosts" query reveals the distinct operational characteristics of each retrieval system:

- **TF-IDF Retrieval**:
  - **Strengths**: Effectively identifies tracks with lexical similarities, especially in titles or lyrics.
  - **Limitations**: May retrieve songs from different genres if they share common keywords, leading to genre inconsistency.
- **BERT Retrieval**:
  - **Strengths**: Captures deeper semantic relationships, resulting in more contextually relevant recommendations.
  - **Limitations**: Still susceptible to retrieving tracks from diverse genres, as semantic similarity does not always align with genre-specific traits.
- **MFCC Retrieval**:
  - **Strengths**: Excels in identifying tracks with similar acoustic features, enhancing the retrieval of musically similar songs.

  - **Limitations**: Acoustic similarity does not necessarily correspond to genre alignment, resulting in a mix of genres in the results.
- **Spectral Contrast Retrieval**:
  - **Strengths**: Achieves high similarity scores, indicating precise timbral texture matching.
  - **Limitations**: High spectral similarity does not guarantee genre or thematic relevance, as evidenced by the retrieval of diverse genres.
- **VGG19 Retrieval**:
  - **Strengths**: Identifies tracks with visually similar album artwork, which can enhance the user experience by providing aesthetically coherent recommendations.
  - **Limitations**: Visual similarity does not inherently correlate with musical genre or style, leading to genre diversity in the results.
- **ResNet Retrieval**:
  - **Strengths**: Similar to VGG19 Retrieval, excels in identifying visually similar tracks based on album artwork.
  - **Limitations**: Shares the same limitation of genre diversity due to the focus on visual features.

3.2.4 Implications and Future Work. The evaluation of the "Chasing Ghosts" query demonstrates the varying effectiveness of different retrieval systems in balancing lexical, semantic, acoustic, and visual similarities. Tag-Based Retrieval emerges as the most effective system in terms of accuracy metrics, underscoring the value of user-generated tags in capturing semantic relevance. However, it also exhibits lower diversity in retrieved tags, suggesting a potential area for improvement.

Fusion-based retrieval systems, such as Early Fusion TF-IDF+BERT Retrieval and Late Fusion MFCC+VGG19 Retrieval, offer a balanced approach by integrating multiple modalities. While they do not surpass Tag-Based Retrieval in accuracy, they enhance coverage and maintain high diversity scores, indicating their potential in providing varied and comprehensive retrieval results.

Future Work:

- **Enhanced Fusion Techniques**: Investigate more advanced fusion strategies that dynamically weight different modalities based on query context to improve both accuracy and diversity.
- **Contextual Understanding**: Incorporate contextual and emotional analysis of lyrics to align lexical and semantic similarities with thematic relevance.
- **User-Centric Evaluation**: Integrate user feedback mechanisms to refine retrieval algorithms based on user preferences and satisfaction.
- **Expansion of Feature Modalities**: Explore additional feature types, such as rhythm patterns or sentiment analysis, to enrich the retrieval process.
- **Cross-Modal Retrieval**: Develop systems that can perform retrieval across different modalities, such as retrieving audio tracks based on visual inputs or vice versa.

By addressing these areas, future retrieval systems can achieve a more nuanced and contextually aware music search experience, bridging the gap between lexical, semantic, acoustic, and visual similarities.

4 CONCLUSION