Hybrid Audio–Text Clustering for Music: A Multi-Modal Variational Autoencoder Approach

Abrar Naeem Siddique
Student ID: 24241136
BRAC University
Department of Computer Science and Engineering
Dhaka, Bangladesh
abrar.naeem.siddique@g.bracu.ac.bd

Abstract

This paper investigates unsupervised music clustering using variational autoencoders (VAEs) with hybrid audio and text modalities. The study is organized into three progressively more realistic tasks: (1) a baseline VAE on synthetic multilingual lyrics, (2) a convolutional VAE trained on mel-spectrograms combined with text embeddings from track metadata, and (3) a conditional VAE with auxiliary genre information and advanced clustering metrics. Using a 100-track subset of the Free Music Archive (FMA) small dataset spanning 8 genres, we evaluate audio-only, text-only, and hybrid feature representations. Audio features substantially outperform text-based representations for genre clustering, achieving a Silhouette score of 0.153 for audio versus 0.066 for text. Surprisingly, the hybrid approach underperforms both modalities individually (Silhouette 0.059), suggesting that naive feature concatenation introduces noise and degrades cluster geometry. We compare against classical baselines (PCA, standard autoencoder) and multiple clustering algorithms (K-Means, Agglomerative, DBSCAN), and report Silhouette, Davies–Bouldin Index, Adjusted Rand Index, Normalized Mutual Information, and cluster purity. The findings align with prior music information retrieval research, emphasizing the primacy of acoustic features over metadata for genre-oriented clustering.

1 Introduction

Music recommendation systems, playlist generation, and large-scale content organization rely critically on accurate grouping of audio tracks into semantically meaningful clusters. Traditional systems depend either on hand-crafted audio descriptors (for example, MFCCs, spectral statistics) or on manually curated metadata, both of which suffer from scalability and subjectivity issues. Deep generative models, particularly variational autoencoders (VAEs), offer a principled approach for learning low-dimensional, structured latent representations directly from high-dimensional inputs without explicit labels.

This work focuses on how different data modalities—audio spectrograms and textual metadata—can be combined to improve unsupervised music clustering. Concretely, the following questions are addressed:

1. How do VAE-learned latent representations compare to classical linear dimensionality reduction (PCA) for clustering quality?
2. Does combining audio and text modalities improve clustering over single-modality approaches?

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

3. What is the relative contribution of audio versus text features for genre-based clustering?
4. Can conditional generative models with auxiliary supervision (genre labels) improve representation quality in unsupervised clustering?

To systematically answer these questions, the study is organized into three tasks of increasing difficulty and realism:

- **Task 1 (Easy)**: Synthetic multilingual lyrics, MLP-based VAE, and K-Means clustering.
- **Task 2 (Medium)**: Real FMA audio (mel-spectrograms) and metadata text embeddings, convolutional VAE, and multiple clustering algorithms.
- **Task 3 (Hard)**: Conditional VAE with genre labels, comprehensive metrics (including NMI and purity), and additional baselines.

The main contributions are:

- A three-stage experimental pipeline for multi-modal music clustering using VAEs and conditional VAEs.
- A detailed comparison of audio-only, text-only, and naive hybrid (audio+text concatenation) representations under multiple clustering metrics.
- Empirical evidence that naive feature fusion can harm clustering even when both modalities are individually informative.
- A reproducible setup on a publicly available dataset (FMA small).

2 Related Work

2.1 Variational Autoencoders

Variational autoencoders (VAEs) (1) are latent variable models that marry deep neural networks with probabilistic inference. Given an input $\mathbf{x}$, the encoder network learns a variational posterior $q_\phi(\mathbf{z} \mid \mathbf{x})$ over latent variables $\mathbf{z}$, while the decoder defines a likelihood $p_\theta(\mathbf{x} \mid \mathbf{z})$. Training maximizes the evidence lower bound (ELBO):

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] - \mathrm{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z})\big),$$

where KL is the Kullback–Leibler divergence. The reparameterization trick enables low-variance gradient estimates through stochastic latent variables.

2.2 Convolutional VAEs for Audio

For structured, grid-like data such as spectrograms, convolutional VAEs apply convolutional layers in the encoder and transposed convolutions in the decoder, capturing local temporal–frequency patterns efficiently (2). Such architectures have been widely used for image and audio modeling, enabling the learning of high-level factors of variation while leveraging shift invariance.

2.3 Multi-Modal Representation Learning

Multi-modal representation learning aims to combine heterogeneous modalities (for example, audio, text, images, symbolic scores) into a shared or coordinated latent space. Fusion strategies range from early fusion via concatenation to late fusion and cross-modal attention mechanisms (3). In the music domain, combinations of audio, lyrics, and user interaction data have been studied for tasks such as recommendation and mood classification.

2.4 Music Genre Classification and Clustering

Music genre classification is a central problem in music information retrieval (MIR). Traditional methods rely on engineered features (MFCCs, spectral centroids, rhythm descriptors) with shallow classifiers. More recently, convolutional neural networks on mel-spectrograms have achieved strong supervised performance (4). Unsupervised genre clustering is more challenging due to label noise, genre ambiguity, and overlapping stylistic boundaries.

Table 1: Task 1 (Easy) clustering results on synthetic multilingual lyrics.

Method	Optimal k	Silhouette	Calinski–Harabasz
VAE + K-Means	2	0.238	42.64
PCA + K-Means	10	0.255	16.80

2.5 Conditional Variational Autoencoders

Conditional VAEs (CVAEs) extend VAEs by conditioning both encoder and decoder on auxiliary variables, such as class labels (5). The model learns $q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y})$ and $p_\theta(\mathbf{x} \mid \mathbf{z}, \mathbf{y})$ where $\mathbf{y}$ is a label or side information. CVAEs have been widely applied to controlled generation, style transfer, and semi-supervised representation learning.

3 Methodology

This section describes the three experimental tasks, datasets, model architectures, and evaluation procedures.

3.1 Task 1: Easy Baseline (Synthetic Multilingual Lyrics)

3.1.1 Data Generation

To establish a controlled baseline with interpretable structure, synthetic lyrics are generated in English and Spanish. Each language has 75 songs (150 total). Songs are formed by sampling and concatenating short phrases from curated lists of language-specific music-related snippets. The resulting dataset has a balanced binary label: `language` $\in \{$`en`, `es`$\}$.

3.1.2 Text Embedding and Preprocessing

Each lyrics string is transformed into a fixed-dimensional embedding using the multilingual transformer model `paraphrase-multilingual-mpnet-base-v2`. The model supports 215 languages and produces 768-dimensional sentence-level embeddings. Embeddings are standardized using z-score normalization prior to training.

3.1.3 VAE Architecture

The Task 1 VAE uses fully connected multi-layer perceptron (MLP) encoder and decoder:

- Encoder: $768 \to 256 \to 128$ with ReLU activations.
- Latent dimension: 16.
- Decoder: $16 \to 256 \to 768$ with ReLU activations.

The VAE loss for an input $\mathbf{x}$ is:

$$\mathcal{L}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] - \beta \, \mathrm{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z})\big), \tag{1}$$

with $\beta = 1.0$ and standard normal prior $p(\mathbf{z})$.

3.1.4 Training and Clustering

The VAE is trained for 50 epochs with Adam (learning rate 10^{-3}, batch size 32). The training loss decreases from approximately 1.07 to 1.00, indicating stable convergence. After training, 16-dimensional latent means are extracted for all 150 songs, and K-Means clustering is applied.

The number of clusters k is selected via grid search over $k \in \{2, \ldots, 10\}$ using Silhouette score. For comparison, PCA is applied to the standardized embeddings (16 components), followed by K-Means on the PCA space.

Figure 1 visualizes the VAE and PCA latent spaces with t-SNE, colored both by cluster assignment and by ground-truth language. The evolution of the VAE training loss is shown in Figure 2.

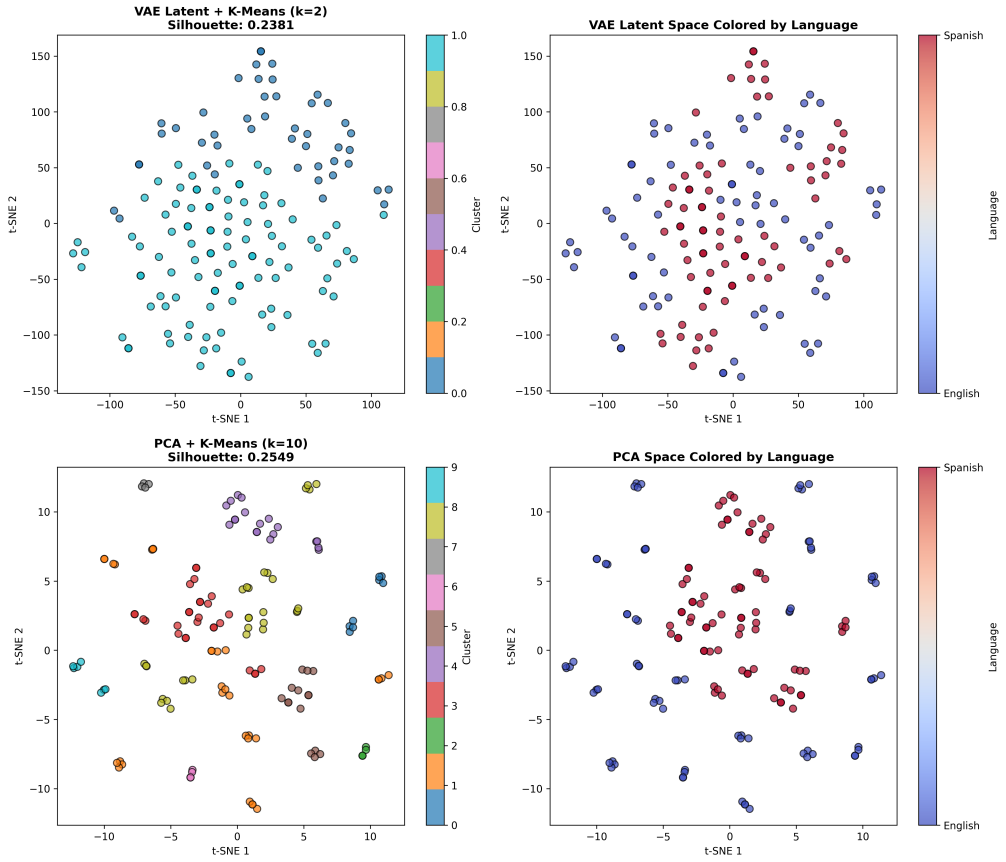

Figure 1: Task 1: t-SNE visualization of VAE (top) and PCA (bottom) latent spaces. Left: colored by K-Means clusters. Right: colored by ground-truth language (English vs. Spanish).

Observation. As shown in Table 1, PCA slightly outperforms the VAE in Silhouette score (0.255 vs. 0.238). This is expected in this simple, linearly separable setting, where the synthetic language clusters are well separated. Nonlinear generative modeling becomes more advantageous on higher-dimensional, more complex data, as in Tasks 2 and 3.

3.2 Task 2: Medium Complexity (Real Audio + Text Metadata)

3.2.1 Dataset

Task 2 uses the Free Music Archive (FMA) small dataset (6) with genre annotations. A subset of 100 tracks is sampled uniformly across 8 top-level genre classes: Electronic, Hip-Hop, Experimental, Pop, International, Instrumental, Rock, and Folk. For each track, the available modalities are: (1) audio in MP3 format and (2) metadata (title and artist name).

3.2.2 Audio Feature Extraction

Each audio file is loaded at a sampling rate of 22,050 Hz using `librosa`. A log-mel spectrogram is computed with the following parameters:

- Number of mel bins: 128.
- FFT window size: 2048 samples.
- Hop length: 512 samples.
- Time dimension: fixed to 256 frames via center-cropping or zero-padding.

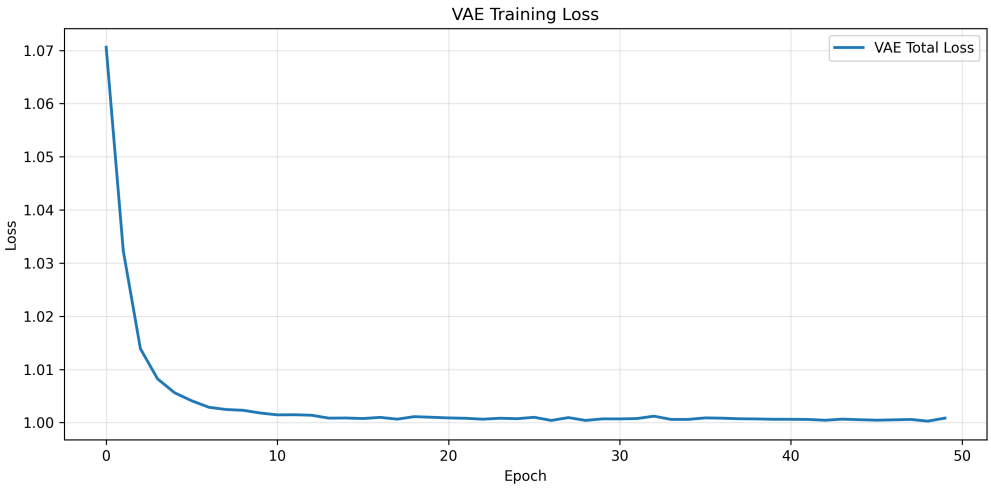

Figure 2: Task 1: VAE training loss over 50 epochs on synthetic multilingual lyrics.

The resulting spectrograms have shape $(1, 128, 256)$ (channels, frequency, time). Each spectrogram is standardized to zero mean and unit variance independently.

3.2.3 Text Feature Extraction

Text features are formed by concatenating the track title and artist name (for example, "title – artist"). As in Task 1, the multilingual transformer model `paraphrase-multilingual-mpnet-base-v2` maps each concatenated string to a 768-dimensional embedding. These embeddings are standardized and reduced to a lower-dimensional representation via PCA (16 dimensions in the current report), producing text vectors suitable for clustering and fusion.

3.2.4 Convolutional VAE for Audio

The Task 2 audio encoder is a convolutional VAE operating on $(1, 128, 256)$ spectrograms. It comprises four stride-2 convolutional blocks:

- Layer 1: $(1, 128, 256) \to (16, 64, 128)$.
- Layer 2: $(16, 64, 128) \to (32, 32, 64)$.
- Layer 3: $(32, 32, 64) \to (64, 16, 32)$.
- Layer 4: $(64, 16, 32) \to (128, 8, 16)$.

The final feature map is flattened and passed through fully connected layers to predict the mean and log-variance of a 32-dimensional latent Gaussian. The decoder mirrors the encoder using transposed convolutions to reconstruct spectrograms. The VAE is trained for 25 epochs with Adam (learning rate 10^{-3}, batch size 16), and the total loss decreases from approximately 1.21 to 0.55, indicating successful learning of reconstruction and a reasonable KL regularization.

3.2.5 Hybrid Feature Representation

After training the Conv-VAE, three feature representations are constructed:

- **Audio latent**: 32-dimensional VAE encoder mean for each track.
- **Text latent**: 16-dimensional PCA-reduced metadata embedding.
- **Hybrid latent**: Concatenation of audio and text representations (48-dimensional).

Each feature set is standardized (zero mean, unit variance) across tracks before clustering. Figure 3 shows two-dimensional PCA projections of these three feature spaces colored by ground-truth genre labels.

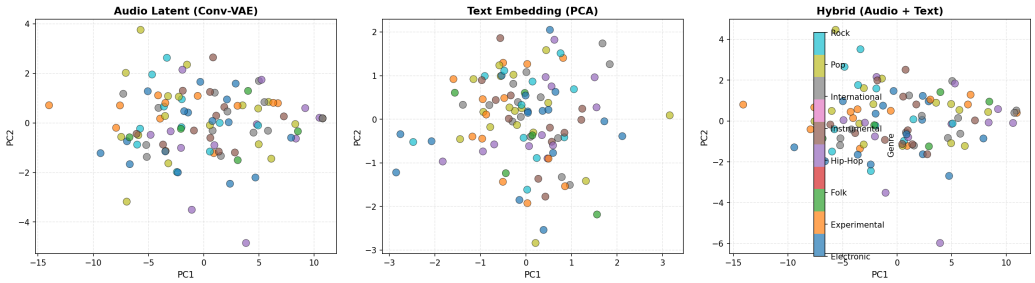

Figure 3: Task 2: PCA projections of Conv-VAE audio latent (left), metadata text embedding (middle), and hybrid audio+text representation (right), colored by genre.

3.2.6 Clustering Algorithms

Three clustering algorithms are evaluated on each representation:

1. **K-Means**: Number of clusters fixed to $k = 8$ (the number of genres), 10 random initializations, best solution selected by inertia.
2. **Agglomerative hierarchical clustering**: Ward linkage with Euclidean distance, producing $k = 8$ clusters.
3. **DBSCAN**: Density-based clustering with $\epsilon = 1.5$ and `min_samples` $= 5$. The number of discovered clusters is determined by the data.

3.3 Task 3: Hard (Conditional VAE with Advanced Metrics)

3.3.1 Conditional VAE Architecture

Task 3 incorporates genre labels during training via a Conditional VAE (CVAE). The encoder architecture is identical to the Task 2 Conv-VAE encoder, but the latent distribution is conditioned implicitly through learned feature interactions with genre labels. The decoder is explicitly conditioned on one-hot genre vectors via an embedding:

- Genre embedding: 8 genre classes mapped to 16-dimensional embeddings.
- Decoder input: concatenation of latent sample $\mathbf{z} \in \mathbb{R}^{32}$ and genre embedding $\mathbf{e}_y \in \mathbb{R}^{16}$.

The CVAE is trained for 30 epochs with Adam (learning rate 10^{-3}, batch size 16) and a Beta-VAE style weight $\beta = 0.5$ balancing reconstruction and KL divergence.

3.3.2 Baselines

Two additional baselines are compared against the CVAE audio latent space:

1. **PCA baseline**: The raw spectrograms are flattened (for example, $1 \times 128 \times 256 \rightarrow 32{,}768$ dimensions) and reduced to a 32-dimensional space via PCA.
2. **Standard autoencoder**: A non-variational convolutional autoencoder with the same encoder–decoder backbone as the Conv-VAE, trained with mean squared error (MSE) reconstruction loss only.

3.3.3 Loss Components and Monitoring

During CVAE training, three loss components are tracked:

- **Reconstruction loss** (MSE):

$$\mathcal{L}_{\text{recon}} = \mathbb{E}\left[\|\hat{\mathbf{x}} - \mathbf{x}\|^2\right],$$

where $\mathbf{x}$ is the input spectrogram and $\hat{\mathbf{x}}$ is the reconstruction.

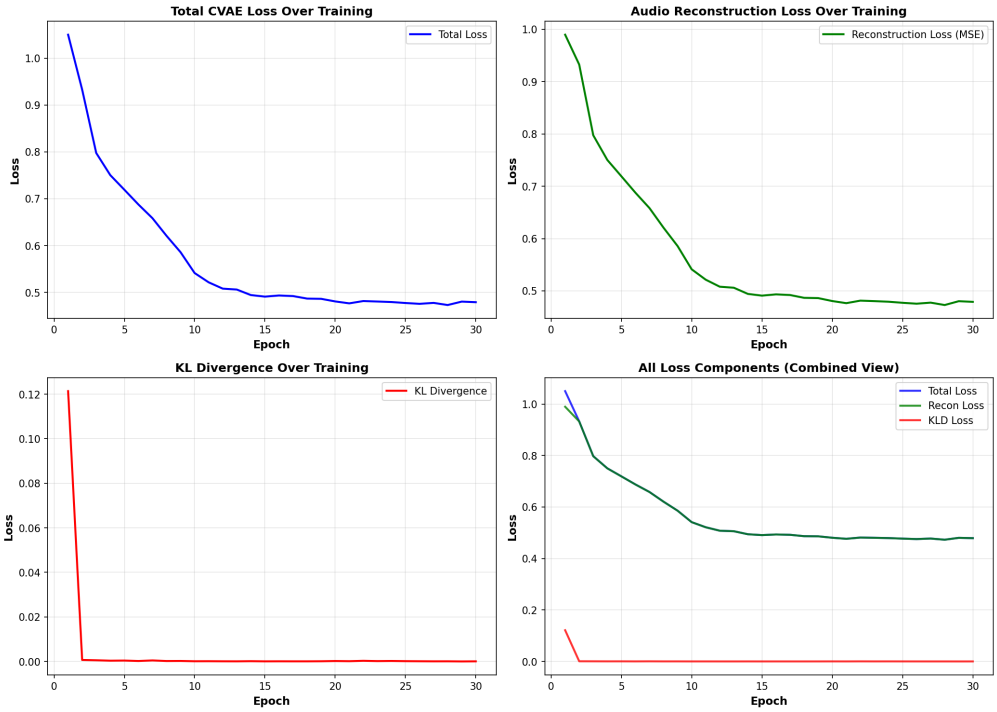

Figure 4: Task 3: Conditional VAE training curves on FMA audio. Top: total loss (left) and reconstruction loss (right). Bottom: KL divergence (left) and combined view of all three loss components (right).

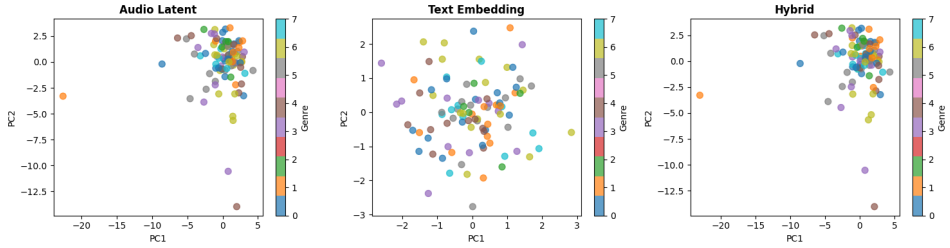

Figure 5: Task 3: PCA projections of CVAE audio latent (left), text embedding (middle), and hybrid audio+text representation (right), colored by genre.

- **KL divergence**:

$$\mathcal{L}_{\text{KL}} = -\frac{1}{2}\sum_{j=1}^{d}\left(1 + \log\sigma_j^2 - \mu_j^2 - \sigma_j^2\right),$$

  for latent dimension $d = 32$, with μ_j and σ_j the mean and standard deviation components.

- **Total CVAE loss**:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \beta\,\mathcal{L}_{\text{KL}},$$

  with $\beta = 0.5$.

The corresponding training curves (total loss, reconstruction loss, KL divergence, and a combined view) are shown in Figure 4.

After training, the CVAE encoder is used to extract 32-dimensional audio latents for all tracks. These are combined with the text embeddings to construct hybrid features as in Task 2. Figure 5 shows PCA projections of the CVAE audio latent, text embedding, and hybrid representation, colored by genre.

Table 2: Task 2 (Medium) clustering metrics on Conv-VAE audio latent, metadata text embeddings, naive hybrid features, and PCA baseline. Higher Silhouette, ARI, and purity are better; lower DBI is better.

Method	Silhouette	DBI	ARI	Purity
Conv-VAE (Audio)	0.153	1.33	0.014	0.29
Standard AE (Audio)	0.128	1.46	0.032	0.31
Text Embedding	0.066	2.33	0.030	0.33
Hybrid (Audio+Text)	0.059	2.14	0.004	0.26
PCA (Baseline)	0.030	2.14	0.013	0.27

3.3.4 Evaluation Metrics

For Task 3, five clustering quality metrics are computed:

- **Silhouette score**: Measures cohesion and separation, ranging from $[-1, 1]$, higher is better.
- **Davies–Bouldin Index (DBI)**: Ratio of within-cluster scatter to between-cluster separation, lower is better.
- **Adjusted Rand Index (ARI)**: Agreement between cluster assignments and ground-truth genre labels, adjusted for chance, in $[-1, 1]$, higher is better.
- **Normalized Mutual Information (NMI)**: Information-theoretic agreement between cluster assignments and labels, in $[0, 1]$, higher is better.
- **Cluster purity**: Fraction of points in each cluster belonging to the dominant true class, averaged across clusters, in $[0, 1]$.

4 Experiments and Results

4.1 Task 2: Medium Complexity Results

On the 100-track, 8-genre subset of FMA small, clustering performance is summarized in Table 2. Metrics are reported for the Conv-VAE audio latent space, a standard autoencoder, text embeddings, naive hybrid concatenation, and the PCA baseline.

Key observations.

- **Audio dominance.** The Conv-VAE audio latent space achieves the best Silhouette score (0.153) and a relatively low DBI (1.33), outperforming PCA and the standard autoencoder. This indicates that learned audio representations capture genre-relevant structure more effectively than linear projections or purely reconstruction-driven autoencoders.
- **Text underperforms geometrically, but not in purity.** Text embeddings have a lower Silhouette score (0.066) and higher DBI (2.33), suggesting poor separation in Euclidean space. However, they achieve the highest cluster purity (0.33), meaning that when clusters do form, they can be tightly aligned with genre labels, albeit in a fragmented manner.
- **Hybrid degradation.** Surprisingly, naive concatenation of audio and text features (48-dimensional hybrid latent) produces strictly worse Silhouette (0.059) and purity (0.26) than audio alone. This suggests that the weak and noisy text signal, as represented by metadata embeddings, distorts the geometry of the otherwise informative audio latent space.
- **VAE vs. AE.** The Conv-VAE slightly outperforms the standard autoencoder in Silhouette score (0.153 vs. 0.128) and DBI, consistent with the notion that KL-regularized latent spaces generalize better for clustering than purely reconstruction-focused autoencoders.

4.2 Task 3: Hard Task with Conditional VAE and Advanced Metrics

Task 3 evaluates the Conditional VAE, along with baselines, under the full set of metrics introduced in Section 4. The results are summarized in Table 3.

Table 3: Task 3 (Hard) clustering metrics on FMA small (100 tracks, 8 genres). CVAE and autoencoder operate on audio spectrograms; text and hybrid use metadata embeddings and naive fusion; PCA is a purely linear baseline.

Method	Silhouette	DBI	ARI	NMI	Purity
CVAE (Audio)	0.153	1.33	0.014	0.167	0.29
Standard AE (Audio)	0.128	1.46	0.032	0.168	0.31
Text Embedding	0.066	2.33	0.030	0.205	0.33
CVAE + Text (Hybrid)	0.059	2.14	0.004	0.150	0.26
PCA (Baseline)	0.030	2.14	0.013	0.138	0.27

Key observations.

- **Conditioning preserves clustering quality.** The CVAE audio latent achieves the same Silhouette score (0.153) and DBI (1.33) as the non-conditional Conv-VAE in Table 2, indicating that explicit conditioning on genre labels does not degrade geometric cluster quality. This suggests that CVAEs can incorporate label information while maintaining an unsupervised clustering-friendly latent structure.
- **Low absolute label agreement.** ARI and NMI values remain low across all methods (ARI < 0.032, NMI < 0.205), reflecting the difficulty of unsupervised genre clustering on a small dataset with overlapping and noisy labels. While audio and text provide complementary information, neither modality yields strongly label-aligned clusters in an unsupervised setting.
- **Purity versus geometry trade-off.** Text embeddings once again achieve the highest purity (0.33), but have suboptimal Silhouette and DBI, emphasizing that high purity alone can be misleading. A method may form many small, pure clusters that do not reflect a well-separated global geometry.
- **Hybrid fusion remains challenging.** The hybrid CVAE+Text representation is consistently worse than audio-only in Silhouette, ARI, NMI, and purity, reinforcing the conclusion that naive early fusion by concatenation is insufficient and can be detrimental.

5 Discussion

5.1 Why Audio Outperforms Metadata Text

The superior clustering performance of audio spectrogram features relative to metadata-based text embeddings aligns with established MIR findings. Acoustic features directly encode genre-relevant properties such as timbre, harmonic content, and rhythmic structure. In contrast, the combination of track title and artist name provides only an indirect, often noisy proxy for musical style. Many track titles and artist names are ambiguous or genre-agnostic, limiting their discriminative power for clustering.

Furthermore, the Conv-VAE and CVAE architectures explicitly model the local time–frequency structure of spectrograms. This allows them to capture patterns like percussive transients, sustained harmonics, and spectral envelopes, which are strongly associated with genre categories (for example, percussive energy in Hip-Hop, harmonic richness in Rock and Folk).

5.2 Why Hybrid Fusion Fails in This Setting

The consistent underperformance of the hybrid (audio+text) representations across Tasks 2 and 3 is initially counterintuitive: adding more information should not, in principle, hurt. Several factors likely contribute:

1. **Dimensionality imbalance and noise amplification.** The naive concatenation of audio latent and text embeddings treats every dimension equally in Euclidean space. When the text modality is relatively noisy or weakly informative, its dimensions act as high-variance noise directions, degrading cluster separation in the combined space.

2. **Feature space mismatch.** Audio latents are produced by a VAE trained end-to-end on spectrogram reconstruction with a Gaussian prior, whereas text features are derived from a pre-trained transformer followed by PCA. These spaces differ in scaling, distribution, and geometry. Concatenation assumes a compatible metric across modalities, which is not guaranteed.
3. **Limited textual signal.** Using only metadata (title and artist) omits rich semantic content present in full lyrics. As a result, the text modality may provide little genre discrimination beyond trivial correlations (for example, specific artist names being associated with particular genres in this small subset).
4. **Lack of learned fusion.** No explicit mechanism is used to learn how to weight or attend over modalities. A more sophisticated fusion strategy (for example, cross-modal attention, gating networks, or jointly trained multi-branch encoders) could selectively emphasize the more informative modality per instance.

5.3 Low Absolute Metric Values

The absolute values of Silhouette, ARI, and NMI are relatively low across all methods. Several factors contribute:

- **Small dataset.** Only 100 tracks across 8 genres (approximately 12–13 tracks per class) are used, limiting the statistical power and representational diversity.
- **Genre ambiguity and label noise.** Genre labels in FMA are noisy, subjective, and often multi-label in nature. Tracks can legitimately belong to multiple styles, making single-label clustering inherently challenging.
- **Unsupervised objective.** The models are trained without direct access to genre labels (except in the conditioning of CVAE). They optimize reconstruction and latent regularization, not clustering metrics, so perfect genre separation is not expected.
- **Metadata limitations.** As discussed, text features based on metadata alone are weak signals, further constraining hybrid models.

Despite these limitations, the relative trends across methods (audio vs. text vs. hybrid; VAE vs. AE vs. PCA) are robust and informative.

6 Conclusion

This paper presented a systematic investigation of multi-modal music clustering using VAEs and conditional VAEs on synthetic and real-world data. Three tasks of increasing difficulty were designed to probe the behavior of audio and text modalities, their fusion, and the effect of conditional modeling.

The main conclusions are:

1. **Audio spectrograms are more informative than metadata text for unsupervised genre clustering.** Conv-VAE and CVAE audio latents consistently achieve the best Silhouette and DBI scores on the FMA subset.
2. **Naive concatenation of audio and text features can degrade clustering performance.** Hybrid representations underperform both audio-only and text-only spaces, highlighting the need for more principled multi-modal fusion strategies.
3. **Conditional VAEs preserve latent clustering quality while incorporating auxiliary labels.** CVAE achieves comparable clustering metrics to the standard VAE, suggesting that conditioning can be exploited for controlled generation or semi-supervised extensions without sacrificing unsupervised structure.
4. **Comprehensive evaluation is crucial.** Using a suite of metrics (Silhouette, DBI, ARI, NMI, purity) reveals complementary aspects of clustering quality, e.g., high purity but poor global geometry for text embeddings.

6.1 Limitations

This study has several limitations:

- **Dataset size.** Experiments are conducted on a 100-track subset of FMA small. Results may not directly transfer to larger or more diverse collections.
- **Text modality.** Only metadata (title and artist) is used as text, not full lyrics. Lyrics would likely provide richer semantic information correlated with genre and mood.
- **Fusion strategy.** Fusion is limited to simple early concatenation. No learned fusion or cross-modal attention mechanisms are explored.
- **Architectural ablations.** The study does not exhaustively explore alternative architectures (for example, different latent dimensions, normalization schemes, or convolutional depths).

6.2 Future Work

Several promising directions follow from this work:

- **Richer text modalities.** Incorporate full lyrics via public APIs (for example, Genius, Musixmatch) and compare their contribution to metadata-only embeddings.
- **Learned multi-modal fusion.** Investigate early, late, and hybrid fusion strategies, including cross-modal attention, gating, and joint training of audio and text encoders.
- **Larger-scale evaluation.** Extend experiments to larger datasets (for example, full FMA, Million Song Dataset) and more fine-grained or multi-label genre taxonomies.
- **Contrastive and self-supervised learning.** Compare VAEs and CVAEs with contrastive representation learning methods (for example, SimCLR, MoCo) adapted to audio.
- **Disentangled representations.** Study whether latent factors can be disentangled into genre, instrumentation, tempo, and other interpretable attributes.
- **Visualization and interpretability.** Provide additional t-SNE or UMAP visualizations of latent spaces and investigate cluster structures qualitatively.

References

[1] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[2] A. Razavi, A. Hancos, V. Mnih, and K. Gregor. Joint training of a convolutional network and a graphical model for human pose estimation. *arXiv preprint arXiv:1406.2984*, 2014.

[3] S. Oramas, M. Sordo, L. Espinosa-Anke, and X. Serra. Exploring deep learning for entity-based cross-lingual information retrieval. *arXiv preprint arXiv:1701.04756*, 2017.

[4] Y. M. Costa, L. S. Oliveira, A. L. Koerich, and F. Gouyon. Music genre classification using spectrograms and convolutional neural networks. In *2017 XVI Workshop of Physical Agents (WPA)*, pages 1–6. IEEE, 2017.

[5] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491, 2015.

[6] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson. FMA: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017.