BEYOND GENRE: DIAGNOSING BIAS IN MUSIC EMBEDDINGS USING CONCEPT ACTIVATION VECTORS

Anonymous Authors
Anonymous Affiliations
anonymous@ismir.net

ABSTRACT

Music representation models are widely used for tasks such as tagging, retrieval, and music understanding. Yet, their potential to encode cultural bias remains underexplored. In this paper, we apply Concept Activation Vectors (CAVs) to investigate whether non-musical singer attributes—such as gender and language—influence genre representations in unintended ways. We analyze four state-of-the-art models (MERT, Whisper, MuQ, MuQ-MuLan) using the STraDa dataset, carefully balancing training sets to control for genre confounds. Our results reveal significant model-specific biases, aligning with disparities reported in MIR and music sociology. Furthermore, we propose a post-hoc debiasing strategy using concept vector manipulation, demonstrating its effectiveness in mitigating these biases. These findings highlight the need for bias-aware model design and show that conceptualized interpretability methods offer practical tools for diagnosing and mitigating representational bias in MIR.

1. INTRODUCTION

Model bias is a well-known challenge across machine learning domains. While extensively studied in NLP and computer vision [1, 2], it has recently also gained growing attention in MIR [3–5]. Beyond degrading performance, model bias also poses serious challenges for ML fairness [3]. In MIR, models may learn to reflect or amplify societal imbalances present in the music industry, reinforcing stereotypes in classification, recommendation, and musical understanding. This phenomenon has been empirically demonstrated in prior work on artist gender bias in music recommendation systems [4]. Recent research in explainable AI for MIR has shown that multimodal large language models (LLMs) trained for music understanding often struggle to meaningfully utilize audio content, at times ignoring it entirely in favor of accompanying text data [6]. This overreliance on the textual modality can lead to auditory hallucinations, where models generate plausible-sounding but inaccurate descriptions. We hypothesize that such failures may also partially stem from biased or entangled audio representations—where stereotypical musical patterns or demographic correlations overly influence the internal encoding. Already flawed audio representations will naturally hinder the model's ability to generalize and reason faithfully about music. We therefore shift our focus to analyzing the audio representations themselves—where such biases may originate but remain largely underexplored. To address this gap, we employ Concept Activation Vectors (CAVs) [7] to systematically probe for unwanted concept entanglement. We investigate whether non-musical factors like singer gender and language influence genre representations. While these attributes should not affect genre representation, we hypothesize that skews in training data lead models to associate them with specific genres. Prior work suggests such imbalances are genre-dependent [8–11]: *Metal* and *Hip-Hop* are heavily male-dominated, while genres like *Pop*, *Electronic*, and *R&B* are more balanced [10]. A model might thus associate male vocals with *Metal* and female vocals with *Pop*, even though vocal gender is not a genre-defining trait. Similarly, while language may serve as a genre cue in specific cases (e.g., Portuguese in *Brazilian music*), overreliance on dominant languages risks marginalizing others [12]. To investigate these biases, we quantify how strongly non-musical attributes are reflected in the model's latent space. By adapting Testing with CAVs (TCAV) [7] for frozen audio encoders, we estimate how consistently genre-specific audio embeddings align with a given concept direction. Secondly, we explore the possibilities of applying CAVs for concept removal or addition, representing a simple post-hoc de-biasing strategy. We publish our code on Github. [1]

This work aims to provide insights into how state-of-the-art music representation models encode and propagate bias, advocating for more fairness-aware design in MIR. While we focus on audio encoders, our approach generalizes to other music-related models. It offers a lightweight, interpretable framework to surface and mitigate biases using small, targeted datasets. The remainder of this paper is structured as follows: Section 2 reviews related work in concept-based explainability and bias analysis in MIR. Sections 3 and 4 introduce the models and the dataset used in our experiments. Section 5 outlines our methodology, including how we construct and evaluate Concept Activation Vectors, before in Section 6, we present our results. We discuss Limitations and Future Work in Section 7 and

[1] upon camera-ready submission

close with our Conclusion and Ethics statement in Sections 8 and 9.

2. RELATED WORK

2.1 Concept based Explanations

Concept-based explanations, notably Concept Activation Vectors (CAVs), have emerged in machine learning as a way to make models more interpretable by aligning their latent representations with human reasoning and intuitive concepts, rather than solely relying on low-level input features such as individual pixels or raw data points [13]. Originally introduced by Kim et al. [7], CAVs represent high-level, user-defined concepts as linear directions within a neural network's latent space. They facilitate intuitive explanations and have seen successful adoption across domains including audio, medical imaging, and generative image modeling [14–16]. The related Testing with CAV (TCAV) method further quantifies interpretability by measuring how sensitive a model's predictions are to specific human-defined concepts through directional derivatives along the CAVs [7].

While alternative interpretability approaches, including saliency maps on image data - or spectrograms analogously in the audio domain - provide local explanations highlighting which individual input elements affect model decisions [17], they may lack the broader conceptual insights that methods such as CAVs offer by directly associating predictions with human-defined concepts [7]. Recent research applies CAV-based methods to verify whether models focus on desired semantic concepts, like object shapes, or undesired spurious signals [18]. Therefore, CAV-based methods have had successful applications in explainability, understanding model representations, concept entanglement detection, spatial dependency evaluations [19] and bias evaluation [20]. Hence, CAVs have proven particularly effective due to their explicit alignment with human-understandable and domain-relevant concepts.

In the domain of MIR, interpretability has often been pursued through disentanglement methods, which attempt to align latent representations explicitly with interpretable musical attributes during model training [21]. Such methods have supported MIR tasks like song generation, cover song identification, and enhanced music search capabilities [22–24]. However, these approaches operate under the assumption that a disentangled and complete set of underlying explanatory factors exist, which is rarely the case in practice, and that a corresponding labeled dataset is available for training.

More recent MIR-specific research incorporates concept-based methods, such as CAVs, for structuring complex genre and mood categories into hierarchical representations [25], and generating explanations tailored explicitly to musicologists by visualizing interpretable musical concepts [14]. Building upon these developments, our work applies CAV-based approaches similarly motivated by aligning model bias explicitly with human-defined concepts, in line with methodologies introduced in bias exploration and debiasing using CAVs.

2.2 Bias Exploration and Mitigation

Existing studies interpret bias as individual neurons, higher-dimensional subspaces, or linear directions in latent space [26–28]. We employ the linear-directions approach, defining concepts as linear directions learned through supervised training on model activations.

Methods concerned with bias in the audio domain can be grouped into techniques applied before or during model training, and post-hoc approaches aimed at identifying and mitigating biases after training, primarily through debiasing latent representations. Pre-or-during strategies include mitigation of data and annotation biases including careful dataset selection and improved transparency through detailed documentation, counterfactual attention learning, and token masking to prevent overfitting to the dataset during learning. [29–31]. Post-hoc methods in MIR that employ bias exploration utilize statistical significance tests to compare performance distributions across affected and advantaged groups, and analyze performance disparities by evaluating differences between universal and culturally adapted models [32, 33].

Our approach in this paper is most similar to that of Wang et al. [5], who apply a dimensionality-reduction method, Linear Discriminant Analysis (LDA), to identify the bias direction in pre-trained audio embeddings by training LDA to separate datasets. In contrast, our CAV-based method defines concepts as undesirable biases which allows us to analyze high-level biases related to how the artist representations manifest themselves in the embeddings, while the method proposed in the work [5] addresses the domain sensitivity bias that is influenced by both the training approach of the embeddings and the alignment of class vocabularies between audio datasets. Therefore, our work expands upon existing post-hoc bias exploration methods by utilizing CAV-based interpretability techniques to address the influence of demographic and sociocultural attributes in music representation models.

3. MUSIC REPRESENTATION MODELS

We evaluate four state-of-the-art music representation models: **MERT**, **Whisper-large-v2**, **MuQ**, and **MuQ-MuLan**. All four can be used to generate audio embeddings for tasks such as *music tagging*, *zero-shot and downstream classification*, *music retrieval*, and as *audio encoders* in music understanding LLMs.

MERT [34] is a self-supervised Transformer model trained on large-scale music datasets using masked modeling and pseudo-labels from acoustic and musical teacher models. It captures musical structure and semantics, making it particularly effective for tasks like music-text retrieval and genre classification. It is the most common audio encoder used in state-of-the-art open-source Music LLMs [6].

Whisper-large-v2 [35] is a speech recognition model trained on 680,000 hours of multilingual and multitask audio data. While originally designed for *automatic speech recognition (ASR)*, it has shown some capability in processing music audio, particularly for transcription tasks. Like MERT, Whisper is often used as an encoder in music LLMs pipelines [6], contributing to music retrieval and captioning tasks. Unlike the other models, Whisper was not trained with music; instead, it is optimized to capture linguistic structure, phonetic features, and prosody, which presumably could lead to less entanglement between language and musical genre as the model's main focus should be the singers' voice instead of the musical content.

MuQ [36] is a self-supervised model that learns discrete music representations via Mel Residual Vector Quantization (Mel-RVQ). It is trained *solely on audio data*, without manual labels, and excels at zero-shot tagging and instrument classification.

MuQ-MuLan [36] extends MuQ by incorporating a *joint training objective with a text encoder (MuLan)* on a large-scale corpus of paired music–text data. This allows MuQ-MuLan to align musical and textual features in a shared embedding space, enabling music–text retrieval and captioning.

In our study, we include both MuQ and MuQ-MuLan to assess how text supervision affects concept encoding and entanglement. While MuQ captures structure-driven representations grounded in audio alone, MuQ-MuLan—through its alignment with text—may encode stronger correlations between musical and non-musical attributes, potentially leading to increased cultural or linguistic bias.

4. DATASET

We use STraDa (Singer Traits Dataset) [37], a large-scale dataset designed for analyzing singer-related attributes in music. Specifically, we leverage the automatic-strada subset, which includes metadata for over 25,000 tracks. This metadata—covering lead singer gender, language, and year of birth—is cross-validated across multiple sources to ensure reliability. To obtain audio, we use the Deezer API to retrieve 30-second audio previews. Despite some files being unavailable, we successfully collect 22,168 tracks. To address underrepresentation of certain genre–gender combinations in STraDa, we supplement the dataset with 251 additional tracks from Deezer playlists specifically curated around the underrepresented concepts [2]. These playlists provide an external, non-manual source of curation. For quality assurance, we manually annotate each track's genre, and the singer's assumed gender and language, discarding those that do not match the intended playlist theme. All associated playlist IDs, track IDs, and metadata are made publicly available in the additional material.

From this corpus, we construct balanced training datasets for nine binary classification tasks: gender (male, female) and the seven most common languages in STraDa (en, fr, it, pt, ja, es, de). We initially explored age (binned by estimated age at release) as a candidate concept but discarded it due to poor CAV projection performance. Following [7], we consider poorly projecting CAVs as incapable of reliably representing a concept and thus unsuitable for bias analysis.

To construct the CAV training and test sets, we split the data at the level of language–genre–gender combinations. When sufficient data is available, we reserve a fixed number of 50 samples per subgroup for training and assign the remainder to the test set. When data is limited, we reserve a smaller portion proportionally. This ensures broad subgroup diversity in the test set while avoiding overlap with the training data. Within both sets, we enforce an exact balance of genre distributions across concept positive samples and randomly selected non-positive samples. This prevents genre from acting as a confounding variable. Additionally, we limit the number of samples across the joint distribution of language, genre, and gender attributes in the training set to avoid subgroup overrepresentation, thereby reducing cross-concept entanglement. This careful balancing is especially crucial for our method to ensure that any observed bias reflects the model's internal representations, not imbalances in the data. While our strategy reduces confounding effects based on available metadata, we state that we cannot guarantee to eliminate biases stemming from unobserved or latent factors. In other words, we assume that the presence of a particular concept (e.g., gender) within a given genre does not fundamentally alter further acoustic features of the genre. For instance, we expect female-led and male-led English-language jazz to be musically comparable, despite differences in vocal timbre. Consequently, we design our CAVs to point exclusively toward the target concept, acknowledging the possibility of residual entanglement from unknown variables.

5. METHOD

CAVs assume that the target concept is *linearly separable* in the model's latent space—that is, there exists a hyperplane that distinguishes between samples with and without the concept. We construct a CAV by training a linear classifier [3] on the latent embeddings, yielding the decision function:

$$\hat{y} = \mathbf{w}^\top \cdot \mathbf{x} + b \tag{1}$$

Here, $\mathbf{x}$ denotes a sample's embedding, $\mathbf{w}$ is the learned weight vector, and b is the bias term. The hyperplane defined by this function—where $\hat{y} = 0$—forms the decision boundary, and the CAV corresponds to the normal vector $\mathbf{w}$, which is orthogonal to that boundary and points in the direction most aligned with the concept. To validate its reliability, we can evaluate the classifier's accuracy. High performance indicates that the concept is linearly encoded; otherwise, the CAV is considered unreliable. Once a CAV is learned, it can be used to rank audio samples by

[2] Affected genres are marked with * in Figure 1

[3] We use a neural network without hidden layers, though any linear classifier is applicable.

how strongly their latent embeddings geometrically align with the concept. This offers an intuitive and interpretable way to explore the model's internal representation—for example, revealing whether certain genres tend to exhibit stronger alignment with a given concept. While useful for qualitative inspection, this approach also enables a more systematic analysis through Testing with Concept Activation Vectors (TCAV), which quantifies concept alignment across categories.

5.1 Measuring Concept Alignment Using TCAV

TCAV [7] provides a statistical framework to quantify the extent to which a model's internal representations rely on a given concept, enabling structured and scalable analysis of concept influence.

In contrast to the original TCAV formulation—which measures the sensitivity of class logits to a concept direction via directional derivatives—we adapt the method for frozen audio encoders, where no downstream classifier is available. Since no gradient information can be extracted, we evaluate concept alignment directly using the full decision function of the trained CAV, including its bias.

For a given genre, we extract the latent embeddings from the test dataset and compute the CAV projection for each sample x as:

$$p_{\text{CAV}}(\mathbf{x}) = \mathbf{CAV}^{\top} \cdot \mathbf{x} + b \tag{2}$$

This projection reflects the sample's alignment with the learned concept, based on the decision boundary. Including the bias b is essential here—unlike in the original TCAV formulation—since we do not compute directional derivatives, where the bias would vanish. Instead, we interpret the CAV as a complete decision function. The TCAV score is then defined as the fraction of samples with a positive projection, indicating how often the model's representations align with the concept:

$$\text{TCAV} = \frac{1}{N} \sum_{i=1}^{N} I\left(p_{\text{CAV}}(\mathbf{x}_i) > 0\right) \tag{3}$$

By comparing TCAV scores across different genres, we assess the extent to which the model's representation of genre relates to the given concept. A higher TCAV score implies stronger alignment—and potential bias. To ensure that these findings are statistically sound, we follow the original TCAV protocol and train 500 CAVs per concept on independently sampled, balanced training subsets with 25% of the data. This yields a distribution of TCAV scores for each concept-genre pair. We conduct a two-sided t-test to determine whether the mean TCAV score significantly deviates from 0.5—the expected value under the null hypothesis of no alignment—and apply a Bonferroni correction to account for multiple comparisons.

5.2 Adjusting Bias via Concept Vector Manipulation

To further explore and mitigate bias in genre representations, we adjust genre-specific CAVs by incorporating bias-related signals through vector operations. As a case study, we select *Hip-Hop* due to its expected negative bias toward the *Female vocal* concept. We train a *Hip-Hop* genre classifier without balancing constraints and apply its CAV to a gender-balanced *Hip-Hop* test set. If there is cultural bias, tracks with male vocals are expected to rank higher. By adding the *Female vocal* CAV—or subtracting the *Male vocal* CAV—we examine whether rankings shift to promote female-associated tracks. The adjusted CAV is defined as:

$$\mathbf{CAV}_{\text{hiphop}}^{\text{adj}} = (1-\lambda) \cdot \mathbf{CAV}_{\text{hiphop}} + \lambda \cdot \mathbf{CAV}_{\text{female}}, \tag{4}$$

or equivalently,

$$\mathbf{CAV}_{\text{hiphop}}^{\text{adj}} = (1-\lambda) \cdot \mathbf{CAV}_{\text{hiphop}} - \lambda \cdot \mathbf{CAV}_{\text{male}}, \tag{5}$$

where $\lambda \in [0, 1]$ controls the adjustment strength. This technique reveals how bias-related concepts influence ranking and highlights potential entanglement in genre representations.

6. RESULTS

To investigate the influence of non-musical attributes on genre representation, we analyze the model responses to two representative concepts in depth: **Female vocals** and **Portuguese language**. The two discussed concepts were selected for their illustrative power—*Female vocals* as a proxy for the singer's gender (noting that the *Male vocals* concept yields largely inverse results), and *Portuguese* as a representative example of linguistic variation in music. *Portuguese* was specifically chosen due to its strong association with genres such as *Latin American Music* and *Brazilian Music*, where meaningful entanglement might be expected, in contrast to other genres where such associations should be less likely. Additional results for all concepts are provided in the supplementary material and largely mirror the trends described in the following sections.

6.1 Evaluation of the Female Concept

We first assess the TCAV scores across genres for the concept of *Female vocals* across the four models, displayed in Figure 1. The average classification accuracy of the trained CAVs exceeds 80% for all models except MuQ-MuLan, indicating that the concept is linearly encoded in their latent spaces and thus reliably captured by the CAVs. Most TCAV scores deviate significantly from the chance level of 0.5, as indicated by their non-overlapping 95% confidence intervals, suggesting the presence of bias in the internal genre representations of all models. The fact that these scores often diverge in direction between models—despite being trained on the same balanced data—strongly suggests that the observed biases reflect genuine differences in how each model encodes the concept.

MERT displays significant negative biases for genres such as *metal*, *rock*, and *Hip-Hop*, with TCAV scores substantially below 0.5. In contrast, genres like *Electronic*,

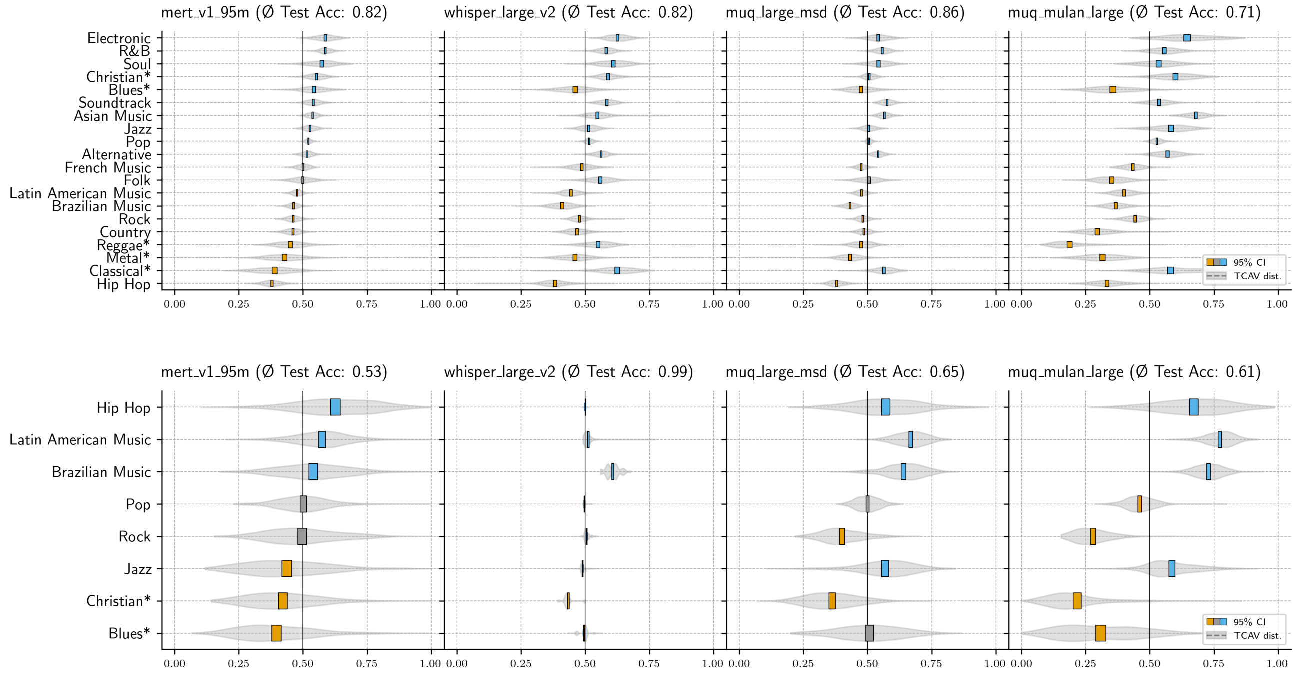

Figure 1. TCAV-based bias evaluation across genres and four music representation models for two non-musical concepts: **Gender-Female** (upper) and **Language-Portuguese** (lower).

R&B, and *Soul* show clear positive associations with the *Female vocals* concept. These patterns align with common vocal stereotypes associated with these genres. Interestingly, *Classical* music reveals the second-strongest negative bias in MERT, despite showing a significantly positive association with female vocals in all other models.

Whisper demonstrates a notably different pattern. While it agrees with expectations for a few genres (e.g., *Metal*, *Hip-Hop*), its TCAV scores diverge—sometimes substantially—in other genres. These inconsistencies suggest that Whisper's internal representations differ from those of MERT. One plausible explanation is Whisper's origin as an automatic speech recognition (ASR) model, which may render it more sensitive to vocal characteristics such as pitch, timbre, or even lyrics, leading to model-specific associations between vocal traits and genre.

MuQ's distribution patterns interestingly resemble those of Whisper, with both models showing relatively subtle genre-specific deviations. Neither model is exposed to music-specific textual labels during training. This absence of genre- or culturally-aligned text input may encourage more structurally grounded representations, resulting in similar concept entanglement with non-musical attributes.

MuQ-MuLan, while aligned in general bias direction with its sibling MuQ, reveals much stronger and more polarized TCAV scores. The model shows significantly negative scores for many genres, and amplified positive scores in others. This suggests that MuQ-MuLan, trained with text supervision, encodes stronger cultural associations between gender and genre. Notably, MuQ-MuLan exhibits strong and statistically significant TCAV scores despite lower CAV test accuracy. This suggests that the model's representations are aligned with the concept in a more entangled or diffuse manner—capturing meaningful bias even when the concept is not cleanly linearly separable. In summary, all models exhibit genre-specific biases related to female vocals, with significant variation in both magnitude and direction. These findings suggest that the training objective, modality, and supervision signal have a substantial influence on how gendered information becomes encoded, and underscore the need for awareness of such effects in downstream MIR applications.

6.2 Evaluation of the Portuguese Language Concept

We now turn to the TCAV evaluation of the *Portuguese language* concept across genres (Figure 1). While Whisper's accuracy is close to perfect, the average CAV classification accuracy of the other models is lower than for the gender-based concepts (ranging between 0.54 and 0.65), it is still above chance and should be interpreted with caution.

MERT shows strong positive TCAV scores for *Latin American Music* and *Brazilian Music*, which aligns with expectations given the natural linguistic-cultural overlap. Surprisingly, MERT shows the strongest bias for *Hip-Hop*. In contrast, genres such as *Rock*, *Christian*, and *Blues* exhibit significant negative biases, suggesting a possible entanglement of Portuguese with stylistic or rhythmic features more prevalent in other genres.

Whisper, by contrast, yields TCAV scores that remain close to the null hypothesis of 0.5 for most genres, indicating relatively little concept alignment. This supports the

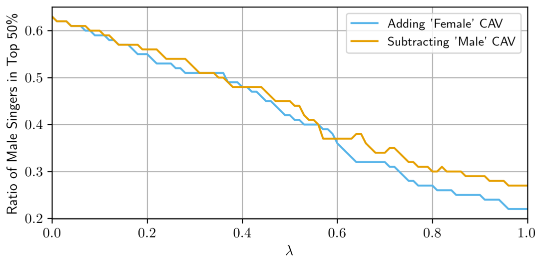

Figure 2. Effect of vector-based concept debiasing on the *Hip-Hop* CAV. The plot shows the ratio of male singers among the top 50% ranked *Hip-Hop* tracks when gradually adding the *Female vocals* CAV (blue) or subtracting the *Male vocals* CAV (orange), as a function of the debiasing weight λ.

interpretation that Whisper—trained primarily for multilingual speech recognition—encodes language in a more disentangled fashion, possibly ignoring most musical information. A notable exception is a small but significant positive bias for *Brazilian Music*, hypothetically reflecting Whisper's heightened sensitivity to vocal or phonetic traits present in Brazilian Portuguese singing.

MuQ captures the expected positive associations between Portuguese and the *Latin American* and *Brazilian Music* genres, while showing negative or neutral biases for other genres. These biases are more strongly amplified in **MuQ-MuLan**, where all genre-specific TCAV scores exhibit clearer polarity. This again may highlight the effect of multimodal training: while MuQ learns from acoustic features alone, MuQ-MuLan's text alignment appears to reinforce cultural and linguistic correlations, magnifying concept entanglement.

6.3 Concept Debiasing

Figure 2 visualizes the effect of our vector-based debiasing strategy applied to the *Hip-Hop* CAV for **MuQ-MuLan**. Here, $\lambda = 0$ represents the original CAV-based sorting, where a strong male bias is observed, as expected from the earlier TCAV analysis. As λ increases, we either add the *Female vocals* CAV or subtract the *Male vocals* CAV, and monitor the proportion of male singers among the top 50% ranked tracks in a gender-balanced Hip-Hop test set.

Notably, both operations result in a nearly linear reduction of male dominance as we interpolate with $\lambda \in [0, 1]$. This suggests that we are traversing a meaningful semantic direction in the latent space and that our *Male* and *Female vocals* CAVs likely encode well-isolated representations of vocal gender. The consistency between the two approaches further supports the intuitive symmetry of these concept directions, which is expected given their mutually exclusive and balanced construction in our training setup. A qualitative review of male-labeled tracks that remained highly ranked after debiasing revealed that many were in fact wrongly labeled, and instead feature female vocals, further reinforcing the reliability of the learned CAVs in capturing vocal gender. Our findings of this experiment highlight that our learned CAVs capture meaningful and robust concept directions, and that concept vector manipulation offers a simple yet effective post-hoc strategy for debiasing model behavior.

7. LIMITATIONS AND FUTURE WORK

Concept disentanglement is essential for interpretable representations. In our work, we mitigate spurious associations between non-musical concepts (e.g., gender, language) and genre by carefully balancing datasets to avoid subgroup overrepresentation. However, even with this balancing, learned CAVs may still be entangled with latent or unobserved factors, especially when the target concept is not cleanly separable in the embedding space. Noisy or ambiguous concept labels may further degrade the clarity of the resulting CAVs. As a result, TCAV scores may reflect not only the intended concept but also correlated dimensions, limiting interpretability. Future work could explore orthogonal CAV training strategies (e.g., [16]) to better isolate individual concepts by explicitly reducing overlap between concept vectors in the latent space. Additionally, our linear analysis assumes a roughly interpretable embedding geometry, which may not capture complex concept interactions. Simple, acoustically salient concepts (like gendered voice timbre) may yield clearer, more interpretable CAVs than abstract, culturally embedded ones (like the singers' age or regional associations). This could unintentionally bias the analysis toward more acoustically grounded attributes.

8. CONCLUSION

We systematically investigated non-musical bias in state-of-the-art music embedding models using Concept Activation Vectors (CAVs) and an adapted TCAV pipeline. Our results reveal significant and meaningful entanglements between genre representations and attributes such as singer gender and language, with variation across models. These patterns reflect known disparities in the music industry and highlight the need for bias-aware model development in MIR. Beyond diagnostic insights, we demonstrate that CAVs can serve as an intuitive and lightweight tool for post-hoc debiasing through concept vector manipulation. Our approach generalizes beyond music representation models: it can be readily applied to any MIR system that produces latent embeddings, including genre classifiers, taggers, or retrieval models. Crucially, it requires only a small set of curated concept examples, making it practical and accessible for real-world deployment.

With this work, we aim to encourage broader research into how MIR models understand and represent music—and the social and cultural implications that follow. While we use concept-based analysis, regardless of method, our primary goal is to foster critical reflection on the biases and assumptions embedded in music technologies.

9. ETHICS STATEMENT

This work investigates representational biases in music embedding models, focusing on demographic and linguistic attributes. Our goal is to expose how models may encode and propagate social and cultural imbalances, aiming to promote fairer and more inclusive MIR systems.

We acknowledge that concepts like gender and language are complex, fluid, and socially constructed. Our binary treatment of gender (male/female) reflects limitations in available metadata and is not an endorsement of reductive framings. We recognize the broader spectrum of gender identities and emphasize the need for more inclusive data collection practices in future research. The absence of non-binary classes in our study is due to insufficient annotated data, and we encourage the community to expand upon these axes with more representative datasets. In our data augmentation process, we were not able to formally verify genre identity, and relied on vocal characteristics to infer gender, introducing a potential source of labeling uncertainty.

All datasets used in this study were sourced from publicly available resources and supplemented with carefully annotated samples to improve representation across groups. We are committed to transparency and reproducibility in our research practices and publish the supplemented metadata alongside this work.

While our focus is on diagnosing and mitigating biases, we also acknowledge the broader ethical implications of our work. This includes the potential misuse of debiasing techniques and the unintended consequences of highlighting biases. Engaging with communities affected by these biases is crucial for ensuring that our research is grounded in real-world experiences and needs.

Our findings are intended to foster critical reflection on the biases and assumptions embedded in music technologies. We hope this work encourages broader research into how MIR models understand and represent music, and the social and cultural implications that follow.

10. REFERENCES

[1] I. Garrido-Muñoz, A. Montejo-Ráez, F. Martínez-Santiago, and L. A. Ureña-López, "A survey on bias in deep nlp," *Applied Sciences*, vol. 11, no. 7.

[2] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis *et al.*, "Bias in data-driven artificial intelligence systems—an introductory survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3.

[3] A. Holzapfel, B. L. Sturm, and M. Coeckelbergh, "Ethical dimensions of music information retrieval technology," *Transactions of the International Society for Music Information Retrieval*, Sep 2018.

[4] D. Shakespeare, L. Porcaro, E. Gómez, and C. Castillo, "Exploring artist gender bias in music recommendation," *arXiv preprint arXiv:2009.01715*, 2020.

[5] C. Wang, G. Richard, and B. McFee, "Transfer learning and bias correction with pre-trained audio embeddings," in *Proc. of the 24th International Society for Music Information Retrieval Conference (ISMIR)*, 2023.

[6] B. Weck, I. Manco, E. Benetos, E. Quinton, G. Fazekas, and D. Bogdanov, "Muchomusic: Evaluating music understanding in multimodal audio-language models," in *Proc. of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[7] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond feature attribution: Testing with concept activation vectors (tcav)," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[8] B. C. Richardson, R. Yoder, and T. F. P. II, "Gender and perception of music genre in college students," *Modern Psychological Studies*, vol. 28, no. 1, 2022.

[9] C. Tabak, "Gender and music: Gender roles and the music industry," *THE JOURNAL OF WORLD WOMEN STUDIES*, vol. 8, no. 1, 2023.

[10] A. Epps-Darling, H. Cramer, and R. T. Bouyer, "Artist gender representation in music streaming," in *Proc. of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[11] A. Ferraro, X. Serra, and C. Bauer, "Break the loop: Gender imbalance in music recommenders." New York, NY, USA: Association for Computing Machinery, 2021.

[12] S. Howard, C. N. Silla, and C. G. Johnson, "Automatic lyrics-based music genre classification in a multilingual setting," 2011.

[13] C.-K. Yeh, B. Kim, and P. Ravikumar, "Human-centered concept explanations for neural networks," arXiv preprint arXiv:2202.12451, 2022.

[14] F. Foscarin, K. Hoedt, V. Praher, A. Flexer, and G. Widmer, "Concept-based techniques for 'musicologist-friendly' explanations in a deep music classifier," in *Proc. of the Int. Society for Music Information Retrieval Conf.*, 2022.

[15] K. A. Thakoor, S. C. Koorathota, D. C. Hood, and P. Sajda, "Robust and interpretable convolutional neural networks to detect glaucoma in optical coherence tomography images," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 8, 2021.

[16] E. Erogullari, S. Lapuschkin, W. Samek, and F. Pahde, "Post-hoc concept disentanglement: From correlated to isolated concept representations," 2025.

[17] T. Brima and M. Atemkeng, "Saliency-driven explainable deep learning in medical imaging," *BioData Mining*, 2024.

[18] C. J. Anders, L. Weber, D. Neumann, W. Samek, K. R. Müller, and S. Lapuschkin, "Finding and removing clever hans: Using explanation methods to debug and improve deep models," *Information Fusion*, vol. 77, 2022.

[19] A. Ghorbani, J. Wexler, and B. Kim, "Explaining explainability: Recommendations for effective use of concept activation vectors (cavs)," *arXiv preprint arXiv:2404.03713*, 2024.

[20] Z. Wei, A. Caines, P. Buttery, and M. Gales, "Analysing bias in spoken language assessment using concept activation vectors," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021.

[21] Y. Zhang, D. S. Carvalho, and A. Freitas, "Learning disentangled semantic spaces of explanations via invertible neural networks," in *Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[22] A. Pati, S. Gururani, and A. Lerch, "dmelodies: A music dataset for disentanglement learning," in *Proc. of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, Montréal, Canada, 2020.

[23] J. Lee, N. J. Bryan, J. Salamon, Z. Jin, and J. Nam, "Metric learning vs classification for disentangled music representation learning," in *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, Montréal, Canada, 2020.

[24] J. Xun, S. Zhang, J. Z. Y. Yang, L. Deng, Z. Zhao, Z. Dong, L. Z. R. Li, , and F. Wu, "Discover: Disentangled music representation learning for cover song identification," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

[25] D. Afchar, R. Hennequin, and V. Guigue, "Learning unsupervised hierarchies of audio concepts," *arXiv preprint arXiv:2207.11231*, 2022.

[26] Z. Yu and S. Ananiadou, "Understanding and mitigating gender bias in llms via interpretable neuron editing," *arXiv preprint*, vol. arXiv:2501.14457, 2025.

[27] A. Krishnan, B. M. Abdullah, and D. Klakow, "On the encoding of gender in transformer-based asr representations," in *Proc. of Interspeech 2024*. ISCA, September 2024.

[28] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *arXiv preprint*, vol. arXiv:1607.06520, 2016.

[29] L. S. Maia, M. Rocamora, L. W. P. Biscainho, and M. Fuentes, "Selective annotation of few data for beat tracking of latin american music using rhythmic features," *Transactions of the International Society for Music Information Retrieval*, May 2024.

[30] Y.-X. Lin, J.-C. Lin, W.-L. Wei, and J.-C. Wang, "Learnable counterfactual attention for music classification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 570–585, 2025.

[31] Z. Zhao, "Let network decide what to learn: Symbolic music understanding model based on large-scale adversarial pre-training," 2025.

[32] A. Holzapfel, F. Krebs, and A. Srinivasamurthy, "Tracking the "odd": Meter inference in a culturally diverse music corpus," in *Proc. of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[33] F. Yesiler, M. Miron, J. Serrà, and E. Gómez, "Assessing algorithmic biases for musical version identification," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, ser. WSDM '22, 2022.

[34] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Lin, A. Ragni, E. Benetos, N. Gyenge, R. Dannenberg, R. Liu, W. Chen, G. Xia, Y. Shi, W. Huang, Y. Guo, and J. Fu, "Mert: Acoustic music understanding model with large-scale self-supervised training," 2023.

[35] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[36] H. Zhu, Y. Zhou, H. Chen, J. Yu, Z. Ma, R. Gu, Y. Luo, W. Tan, and X. Chen, "Muq: Self-supervised music representation learning with mel residual vector quantization," *arXiv preprint arXiv:2501.01108*, 2025.

[37] Y. Kong, V.-A. Tran, and R. Hennequin, "Strada: A singer traits dataset," in *Interspeech 2024*, 2024.