Exploring the Internal Mechanisms of Music LLMs: A Study of Root and Quality via Probing and Intervention Techniques

Wenye Ma [1] **Gus Xia** [1 2]

Abstract

How do music large language models (LLMs) interpret musical concepts? This study investigates the representational abilities of MusicGen, a transformer-based music LLM, using single chords to assess how these models process structured musical entities. We develop a novel probe-via-intervene approach to enhance our understanding of the model's internal interpretability. Our findings indicate that although the model faces challenges in forming linearly separable representations for certain musical concepts such as chord *quality*, the integration of directional vectors from other musical concepts into the transformer's residual stream substantially improves the probing results. Notably, significant enhancements are achieved by intervening in just one head across all layers. These insights underscore the differences between human and machine perception of music and suggest important considerations for future design of music LLMs.

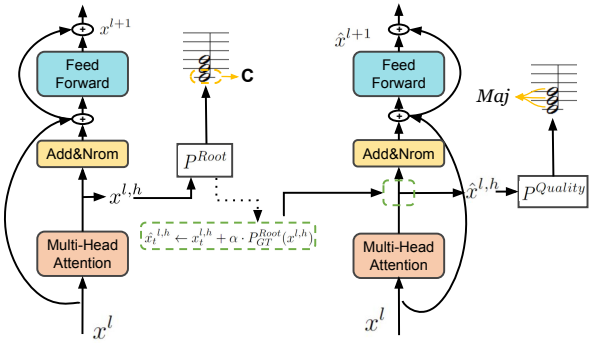

Figure 1. Overall pipeline including probing and augment-intervention. The left section illustrates the probing of various musical concepts (only the chord *root* probe is shown for simplicity). The right section demonstrates the augment-intervention process, where chord *root* directions obtained from the probing phase are added to the hidden states, enhancing the model's representation of chord *quality*. $P_{GT}^{ROOT}(x^{l,h})$ indicates the direction added is the ground truth root direction of $x^{l,h}$

1. Introduction

In the field of computer music, particularly in music generation and understanding, music large language models have demonstrated remarkable development. Among them, MERT (Li et al., 2023c) is a masked language model (MLM) tailored for music information retrieval (MIR) tasks. while Jukebox (Dhariwal et al., 2020), MusicLM (Agostinelli et al., 2023), and MusicGen (Copet et al., 2024) are notable auto-regressive models designed primarily for music generation. The success of these models underscores the potential of AI to not only replicate human-like musical abilities but also to provide tools that can enhance the creative process for musicians and producers.

Despite the strong capabilities of music large language models, one persistent issue remains unresolved: can models that have not been explicitly taught human-centric summaries and understandings of music still extract and utilize concepts that parallel human comprehension? This study explores how AI models, through their interaction with music data, naturally form relationships and between different musical concepts. By doing so, these models provide us with unique insights into which concepts are more fundamental or intrinsic to music as an art form.

Interpretability is often closely linked to linear representations (Mikolov et al., 2013; Nanda et al., 2023). However, confirming a model's mastery of a concept is challenging if the concept representation is not linearly separable. (Li et al., 2023a) investigates the nonlinear representation of board states and has achieved the anticipated outcomes by intervening using gradient ascent. Yet, for more complex music LLMs compared to Othello-GPT, this method may be ineffective. Consequently, we develop a novel approach, probe-via-intervene, to unlock the potential of certain concept representations that a linear classifier cannot probe, in

[1]MBZUAI, Machine Learning Department, Abu Dhabi, United Arab Emirates [2]NYU Shanghai, Computer Science Department, Shanghai, China. Correspondence to: Wenye Ma <wenye.ma@mbzuai.ac.ae>, Gus Xia <gus.xia@mbzuai.ac.ae>.

a manner that is interpretable.

We focus on two primary tasks: detecting chord *roots* and *qualities* using probe-via-intervene. Besides the conventional probing for each concept, to demonstrate the interconnections among different musical concepts within the model, we probe the *quality* concept from the hidden states of MusicGen after intervention, specifically by introducing *root* directions. The result shows that after the intervention, the linear probing accuracy improved. This suggests that MusicGen had actually learned something about *quality* although it can't be well probed by linear classifier at first. Remarkably, even if we only intervene one head from all transformer layers, the accuracy can improve over 0.2.

Our principal contributions are twofold:

- We introduce a novel method, probe-via-intervene, to enhance our understanding of whether a model has learned a specific concept.
- We gain valuable insights into the model's internal mechanisms for processing music through our experimental observations.

2. Preliminaries

In this section, we briefly describe about music terminology chord, and MusicGen-small, the model we probe.

2.1. Chord

Chords, made of multiple simultaneous notes, are crucial in songs for supporting and guiding the melody. When saying a chord like *C major*, we can see two concept that we concern in the following paper. The chord *root* **C** is the fundamental note on which a chord is built. It is the base note from which the other notes of the chord are derived. Chord *quality* **major** refers to the specific characteristics of a chord that define its sound. This is determined by the intervals between the three notes in the chord. Different chord *quality* have different color, even people without music training can tell major and minor chords for the former is bright and the latter of sorrow.

2.2. MusicGen

MusicGen (Copet et al., 2024) is a transformer-based generative music model capable of generating up to 30-second continuations from an audio or text prompt. In this work, we only use audio prompt and set the text prompt empty. We use the small size version of MusicGen, which has 24 transformer decoder layers, where each layer consists of 16 heads, yields a 1024-dimensional hidden state per frame at a resolution of 75 frames per second.

3. Methodology

3.1. Process of MusicGen Transformer

In the model, the forward pass begins by embedding the input tokens from the prompt audio into a high-dimensional space, denoted as $x^0_{1:T}$, which serves as the initial state of the residual stream. Subsequently, within the transformer's structure, each sublayer $l \in [1, \ldots, L]$ calculates H self-attention heads as:

$$x^{l,h}_{1:T} = \text{Att}^h_l(x^l_{1:T}) \tag{1}$$

Then, the projection matrix $W_{l,h}$ and MLP block fuse the information in this layer to the residual stream.

3.2. Audio Level Feature Extraction

Before projection, we extract features $x^{l,h}_{1:T} \in \mathbb{R}^{T\times D}$ (D is the hidden state dimension of each head). By ensuring that the tonality remains consistent across all data in the dataset, we can employ mean pooling to define the representation of the entire audio track at layer l and head h as follows:

$$x^{l,h} = \frac{\sum_{t=1}^{T} x^{l,h}_t}{T} \tag{2}$$

where $x^{l,h} \in \mathbb{R}^D$.

3.3. Probes

We train different probes of music concept c at each layer l and head h. For music concept c that requires n_c-category classification, we designate i to represent the i-th category, where i ranges from 1 to n_c, inclusively, denoted as $i \in [1, \ldots, n_c]$. Our linear probe can be written as $P^c_i(x) = \text{Softmax}(Wx)$. When x is a feature that is extracted from a single head, $W \in \mathbb{R}^{D\times n_c}$. When we probe the whole layer using all heads in that layer, $W \in \mathbb{R}^{DH\times n_c}$, where $x \in \mathbb{R}^{DH}$, obtaining by simply concatenating all the H heads' feature in that layer.

Non-Linear probes are 2-layer MLP models: $P^c_i(x) = \text{Softmax}(W_1 \text{ ReLU } (W_2 x))$. Where $W_1 \in \mathbb{R}^{512\times n_c}$, and $W_2 \in \mathbb{R}^{DH\times 512}$ when probing the whole layer and $W_2 \in \mathbb{R}^{D\times 512}$ when just probing a single head.

3.4. Intervene

Inspired by previous works (Li et al., 2023b; Nanda et al., 2023; Koo et al., 2024), our intervention is done by adding a vector to the heads before multiplying by the projection matrix:

$$\hat{x}^{l,h}_t \leftarrow x^{l,h}_t + \alpha \cdot P^{c,l,h}_i(x^{l,h}),\ \forall t \in [1, \ldots, T] \tag{3}$$

The modified head feature $\hat{x}^{l,h}$ is then passed forward to the residual stream by projection matrix and feed-forward blocks. In this function, $P^{c,l,h}_i(x^{l,h})$ can be considered as

the direction of category i for the concept c with the given x, while scale α denote an intervention strength.

Initially, we want to see whether there is some concept A that can help the model learn a better concept B representation. To evaluate this, we use method denote as **augment-intervention**. The direction i in this case is the **ground truth** direction of the concept A. To examine the impact of augment-intervention, we extract $\hat{x}^{l,h}$ features from the layers being intervened, and do the probing tasks on the concept B (probe-via-intervene). We contrast the probing result of concept B before and after augment-intervention. Additionally, using the direction of each category derived from the probes, we can potentially alter music from one category to another, a process referred to as **shift-intervention**, as discussed in Appendix C.3.

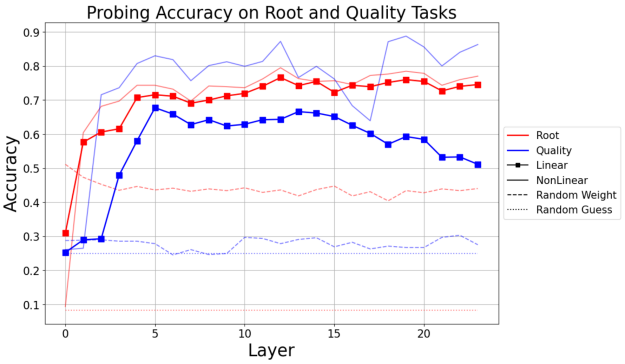

Figure 2. Probing accuracy across layers on two tasks

4. Experiment

4.1. Dataset

We construct a synthetic dataset comprising 38,160 audio clips, each representing a distinct chord with root notes spanning C2 to C7. It is important to note that for classification purposes, root notes in different octaves are grouped into the same category (e.g., D3 and D5 are both classified under the root 'D'). The chords are rendered in root position, first inversion, or second inversion, and they encompass four triad types: Major, Minor, Augmented, and Diminished. Each clip played using one of 53 different instruments, lasts approximately one second. For more details, please see Appendix A. The dataset facilitates two main tasks: identifying the chord *root* from 12 possible categories and classifying the chord *quality* into one of the 4 triad types.

4.2. Probing Analysis

We employ the MusicGen-small model for our experiments, which consists of 24 layers with 16 attention heads each, summing up to a total of 384 heads. We initially probe all layers using both linear and nonlinear methods to assess their performance. In addition, we set up two comparative baselines: one using features derived from a randomly initialized MusicGen model, and another based on a strategy of random guessing. The outcomes of these analyses are illustrated in Figure 2.

The figure demonstrates that the overall linear probing accuracy for chord *root* is higher than that for chord *quality*, despite there being more categories for *root* than for *quality*. Additionally, the analysis indicates a consistent, albeit slight, increase in representational strength in deeper layers for *root* detection; however, accuracy for *quality* detection tends to decrease in these layers. Moreover, while the probing results for *root* detection show similar outcomes with both linear and nonlinear probes, a significant difference is evident in the *quality* detection results between these probing methods. This finding contrasts with common perceptions, as *quality*, compared to *root*, is generally considered a simpler concept for humans to discern.

We further conduct a detailed examination of all heads across each layer. The results show considerable variation among the heads within each layer for the chord *root* concept, with some heads effectively capturing *root* information and others not. Conversely, the variance among heads for *quality* detection is notably lower. The visualization of this result can be seen in Figure 6. This suggests that we can inject concept c into only a portion of heads that have a good understanding of that concept in the following experiments.

4.3. Intervention

4.3.1. Root Information Impact on Quality Detection

During the augment-intervention phase, we utilize a $Top-K$ strategy, intervening only in the K heads with the highest accuracy for predicting musical concepts. A grid search is performed to optimize K, scale α, and the intervention layers L. For the parameter details, see Appendix C.1

To make the result clear, we only present probe-via-intervene results of *quality* at layer 23 with one group of settings (Figure 3).

Initially, the features extracted from layer 23 achieved an accuracy of 0.51 in the *quality* detection task. After the augment-intervention, the model demonstrates improved linear representation for *quality* detection. This information indicates that, although Music LLMs struggle to form linearly separable representations of *quality*, the concept of *quality* is still embedded within its representations. It can be activated by *root* information.

Remarkably, intervening on top-1 head across all layers results in the best outcome. This suggests that the model is

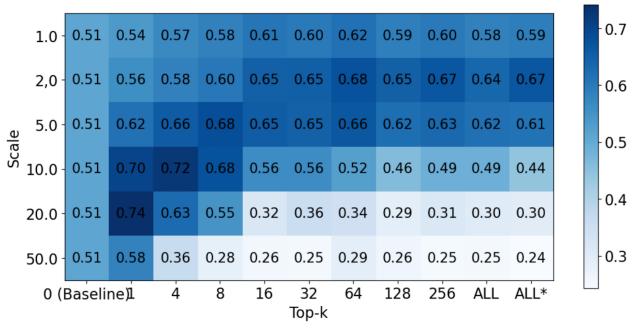

Figure 3. Linear probing accuracy at layer 23 after augment-intervention with settings (Start-Layer = 4, Intervention-Mode = 2).

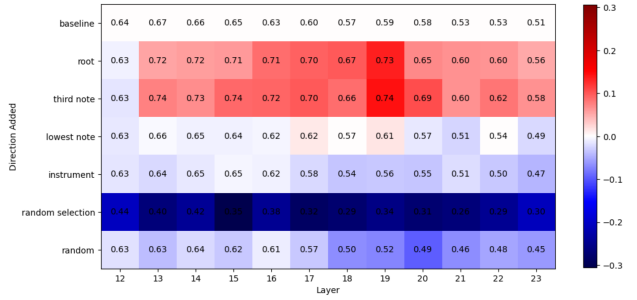

Figure 4. Linear probing accuracy after augment-intervention with the setting (Start-Layer = 12, Intervention-Mode = 1, Top-K = 16, Scale = 1.0). The grid score represents accuracy, with the color indicating accuracy relative to the baseline. Red signifies improved accuracy compared to the baseline, while blue signifies a decrease, with deeper hues indicating greater deviation from the baseline.

sensitive to *root* information, and demonstrates the potential for achieving better representations through relatively lightweight interventions.

4.3.2. Influence of Other Musical Information on Quality Detection

We conduct several experiments to explore whether *root* information uniquely contributes to the model's ability to learn the *quality* concept. We investigate the effects of directions related to *third note*, *lowest note* and the *instrument* on the model's performance, the result is shown in Figure 4. (For detailed illustration about terminology *third note* and *lowest note*, please see Appendix A.) The probing results before augment-intervention of all related tasks are shown in Table B.1.

The *third note* in a chord, which is a third interval above the root, plays a crucial role in determining whether the chord is major or minor given a root. Note that we does not use root directions along with *third note* direction, thus not offering the direct intervals among the triad to provide explicit hints. Our findings indicate that the direction corresponding to the *third note* also facilitates the model's learning of the *quality* concept, with improvements comparable to those observed with *root* information.

The *lowest note* in a chord varies depending on the chord's inversion. Our experiments demonstrate that incorporating directions for the *lowest note* into the hidden states does not enhance the model's ability to discern *quality*. Moreover, *instrument* direction also doesn't help the model to learn a better *quality* features, even if the linear probing accuracy of *instrument* classification is over 90 %.

Additionally, we conduct two baseline experiments for comparative analysis. The first baseline involves adding directions representing alternative *roots* that do not correspond to the actual chord (random selection), which adversely affects the model's *quality* representation. The second baseline introduces a random vector (random) matching the distribution of the directions, which slightly deteriorates the model's *quality* detection capabilities.

These findings reveal the intriguing logic models use to detect musical concepts. The success in using *roots* and *third notes* as augmentation directions indicates that chord components information can enhance each other. Conversely, the failure in using the *lowest note* as an augmentation direction suggests that the model assesses chord *quality* beyond superficial audio information. Instead, it likely integrates musical knowledge acquired from various pretrained data sources.

5. Conclusion

In this study, we employ a probe-via-intervene method to closely examine the representations from music LLM. We demonstrate that a better chord *quality* representation can emerge with the help of other music concepts like chord *root*, even by only intervening 1 head in the transformer. We also show the relationship of music concepts from the perspective of model.

In a higher level, the results and the comparison of success and failure case from augment-intervention experiments give insights that can assist researchers in aligning model perception with human understanding and in designing next generation music models with enhanced performance.

For future research, we intend to conduct a more in-depth investigation into the underlying mechanisms driving this phenomenon and identify additional related concept pairs. Moreover, it is important to further explore the interpretability of music LLMs in both the audio and symbolic domains.

References

Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., and Frank, C. Musiclm: Generating music from text, 2023.

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., and Défossez, A. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., and Sutskever, I. Jukebox: A generative model for music, 2020.

Koo, J., Wichern, G., Germain, F. G., Khurana, S., and Roux, J. L. Smitin: Self-monitored inference-time intervention for generative music transformers, 2024.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Emergent world representations: Exploring a sequence model trained on a synthetic task, 2023a.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model, 2023b.

Li, Y., Yuan, R., Zhang, G., Ma, Y., Chen, X., Yin, H., Lin, C., Ragni, A., Benetos, E., Gyenge, N., et al. Mert: Acoustic music understanding model with large-scale self-supervised training. *arXiv preprint arXiv:2306.00107*, 2023c.

Mikolov, T., Yih, W.-t., and Zweig, G. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746–751, 2013.

Nanda, N., Lee, A., and Wattenberg, M. Emergent linear representations in world models of self-supervised sequence models, 2023.

A. Music Theory of Chord

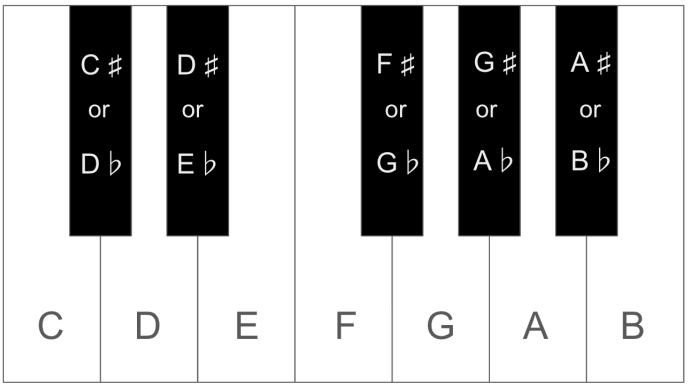

Figure 5. Positions of each note on the piano keys, where each pair of adjacent keys, whether black or white, differs by a semitone

There are four types of triads. In this work, they correspond to the four chord *qualities*:

A **major chord**, characterized by its bright and happy sound, consists of three notes: the root, a major third (four semitones above the root), and a perfect fifth (seven semitones above the root).
A **minor chord**, which sounds more somber, also comprises three notes: the root, a minor third (three semitones above the root), and a perfect fifth.
An **augmented chord**, with its mysterious and unsettled tone, consists of the root, a major third, and an augmented fifth (eight semitones above the root).
A **diminished chord**, known for its tense and dissonant quality, includes the root, a minor third, and a diminished fifth (six semitones above the root).

Inversions of these chords provide different voicings by rearranging the order of the notes. The **root position** has the root note as the lowest note, the **first inversion** places the third as the lowest note, and the **second inversion** positions the fifth as the lowest note. These inversions offer varying harmonic textures while maintaining the chord's fundamental structure including roots and qualities. The table below uses chords based on the root note C to demonstrate its four qualities and the arrangement of notes after each inversion. The *third note* mentioned in 4.3.2 is marked in blue, and the *lowest note* is underlined. As shown, for a chord with a fixed root and quality, the third note remains unchanged regardless of the inversion, but the lowest note changes with each inversion.

Inversion / Quality	Root Position	First Inversion	Second Inversion
Major Chord	C - E - G	E - G - C	G - C - E
Minor Chord	C - E♭ - G	E♭ - G - C	G - C - E♭
Augmented Chord	C - E - G♯	E - G♯ - C	G♯ - C - E
Diminished Chord	C - E♭ - G♭	E♭ - G♭ - C	G♭ - C - E♭

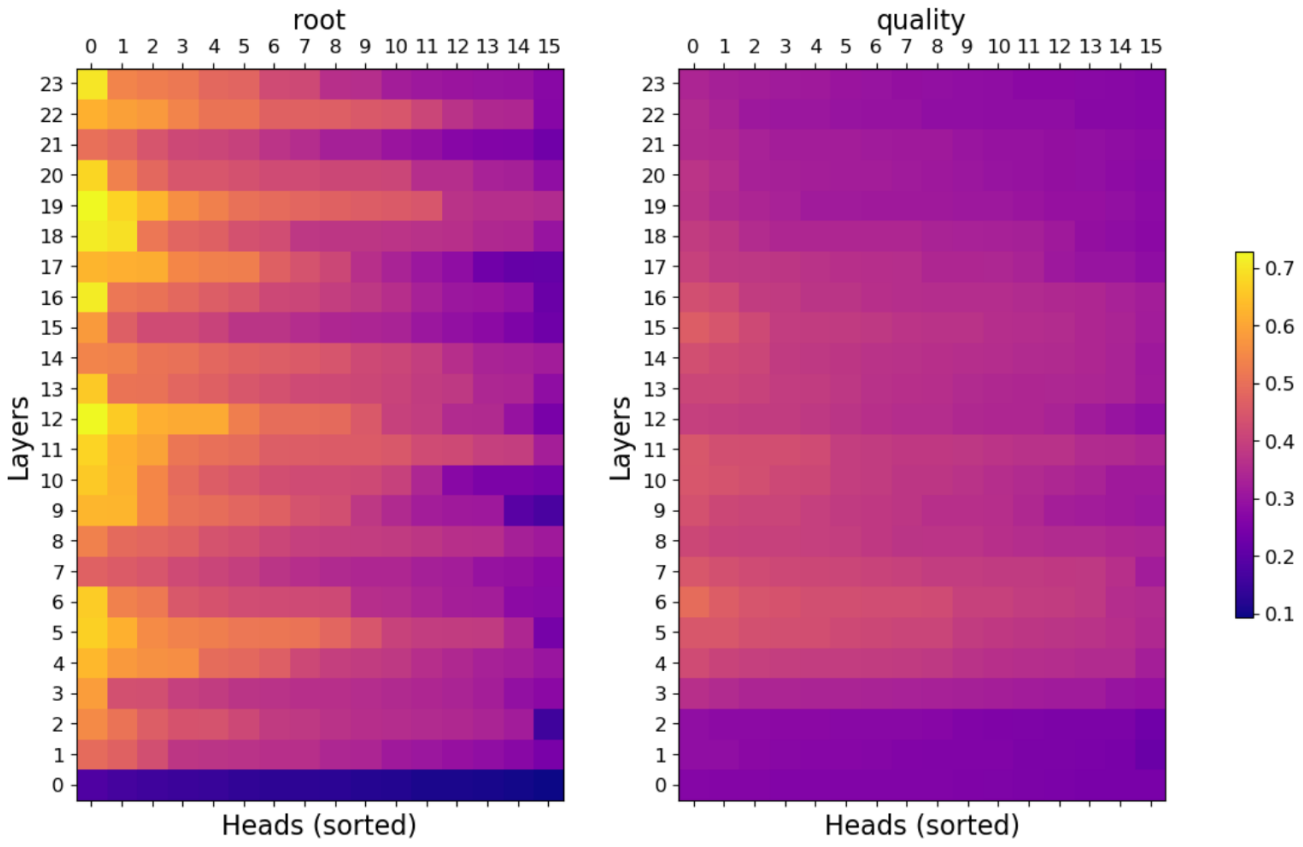

Figure 6. Linear probing result of individual attention head on root (left) and quality (right)

B. Results: Linear Probing on Different Heads of Chord Root and Quality

B.1. Probing Results before Augment-intervention

Task	Layer 0	Layer 3	Layer 6	Layer 9	Layer 12	Layer 15	Layer 18	Layer 21
Root	0.309	0.615	0.711	0.712	0.766	0.723	0.752	0.727
Quality	0.253	0.479	0.659	0.624	0.643	0.652	0.570	0.532
Instrument	0.536	0.932	0.941	0.945	0.935	0.942	0.934	0.938
Lowest Note	0.243	0.415	0.554	0.521	0.626	0.602	0.648	0.565
Third Note	0.292	0.640	0.708	0.723	0.748	0.742	0.753	0.743

C. Intervention Details

C.1. Grid Search Parameters

The grid search parameters of intervention is as follows:

- Intervention-Mode: {0, 1, 2}
- Start-Layer: {4, 8, 12, 16, 20}
- Top-K: {1, 4, 8, 16, 32, 64, 128, 256, All, All*}
- Scale: {1.0, 2.0, 5.0, 10.0, 20.0, 50.0}

C.2. Details of Different Intervention Modes

In *Intervention-Mode 0*, the intervention is applied to just one layer, leveraging the fact that instructions are propagated through all timesteps in the residual stream, thereby affecting all subsequent $x_t^{l,h}$ in all following layers. Alternatively, in *Intervention-Mode 1* , we do the intervention from the *Start-Layer* to the final layer of transformer layer by layer. In this two mode, the range of *Top-K* is restricted to $\{1, 4, 8, 16, \text{All}^*\}$, where *All** denotes intervening the whole layer's feature x^l, while *All* pertains to intervening in every head-specific $x^{l,h}$. *Intervention-Mode 2*, derived from prior research (Li et al., 2023b; Koo et al., 2024), involves intervening in the *top-K* best-performing heads across all layers.

C.3. Case Study of Shift-Intervention

We utilize *quality* directions after *root* augment-interventions to intervene in the model's hidden states. In this part, our goal is shift-intervention rather than augment-intervention, specifically altering a major chord to a minor one by adding a minor direction, and vice versa. (We do not attempt augmented or diminished triads due to the lack of support from our rule-based chord recognition tool for result evaluation). Here are two cases that successfully changed the root *quality*.

In Audio 1, a G minor chord is changed to C major, while in Audio 2, an A major chord is changed to F minor. (The results are verified both by ear and the recognition tool).

Steering on the directions easily makes the texture of the audio become noisy and disrupts the accuracy of the recognition tool. We hypothesize that adding directions cannot ensure that the features remain within the music subspace, which is a potential area for future improvement in this work.

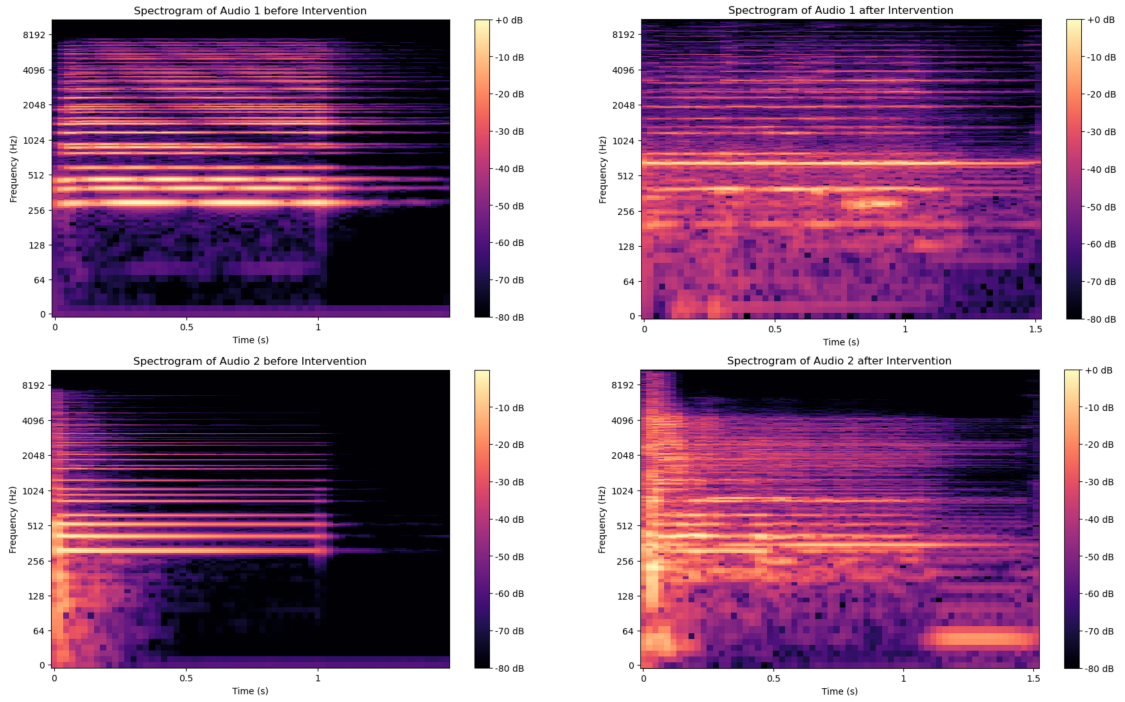

Figure 7. Spectorgrams of two audios before and after intervention