Emotion control in MusicGen via Activation Steering

Lucia Fornetti and Tommaso Federici

{fornetti.2214370, federici.2214368}@studenti.uniroma1.it

Abstract

We propose a lightweight framework for controlling high-level emotional attributes in MusicGen via Activation Steering, eliminating the need for resource-intensive fine-tuning. By analyzing the model's internal representations, we identify linearly separable directions corresponding to "Happy" and "Sad" concepts and introduce a Multi-Block injection strategy to modulate both rhythmic and timbral features. Our experiments demonstrate that this method induces consistent emotional shifts, validated by both objective semantic metrics (CLAP) and human evaluation, offering a precise and efficient alternative for affect-driven music generation.

The source code and audio samples are available at: https://github.com/TommasoFederici/Emotion-Control-in-MusicGen-via-Steering.

1 Introduction

Generative models like MusicGen [1] have achieved impressive high-fidelity audio generation but often struggle with the precise control of abstract nuances, particularly **emotional expression**. Relying solely on text prompts is frequently insufficient, as models may conflate complex emotions with simple acoustic features (e.g., interpreting "sad" merely as slow tempo), missing crucial timbral qualities, or vice versa.

In this report, we investigate emotional control in MusicGen without fine-tuning. Inspired by activation steering in LLMs [6] and recent audio domain studies [3, 5], we developed a framework to extract internal representations and inject steering vectors during inference. While our framework is designed to be general-purpose, supporting arbitrary semantic directions, this study focuses on the **'Happy' vs. 'Sad'** dichotomy. By identifying layers with maximum cluster separability, we demonstrate that injecting computed vectors effectively steers the musical mood, offering a lightweight alternative to model retraining.

2 Related Work

Our research sits at the intersection of controllable music generation, interpretability in deep learning, and multi-modal semantic evaluation.

Music Generation Transformers. Our work relies on **MusicGen** (Copet et al., 2023) [1], an autoregressive Transformer operating on discrete **EnCodec** [2] tokens. Specifically, we utilize the `musicgen-melody` checkpoint. Although this model is trained to accept melodic audio conditioning, we employ it in a **text-only mode** for our steering pipeline. This choice is deliberate: as detailed in Appendix B.1, our experiments showed that explicit melodic conditioning rigidly constrains the generation, hindering the effectiveness of high-level emotional steering.

Activation Steering. Originally popularized in LLMs by **Turner et al.** (2023) [6], activation steering involves injecting vectors into hidden states to alter behavior without fine-tuning. Adapted to music, **Facchiano et al.** (2025) [3] analyzed layer-specific information encoding via activation patching, while **Panda et al.** (2025) [5] demonstrated fine-grained control over musical attributes. Building on these works, we developed a framework to specifically target the separability and steerability of high-level emotional concepts (Happy vs. Sad).

Semantic Evaluation via CLAP. To address the subjectivity of audio evaluation, we employ **CLAP** (Wu et al., 2023) [7], which maps audio and text into a shared latent space. We utilize this embedding space to compute a quantitative **"Happiness Score"**, defined as the cosine similarity between the generated audio and the target emotion labels, to objectively measure steering success.

3 Proposed Method Explained

We propose a three-step framework to steer the emotional output of MusicGen. Our method first analyzes the model's latent space to identify the layers where emotional concepts are linearly separable. Based on this analysis, we extract a steering vector representing the direction of the target emotion and inject it during inference.

3.1 Internal Representation Analysis

To effectively steer the model, we must first determine *where* emotional information is encoded. Hypothesizing that high-level concepts like "Happiness" and "Sadness" cluster in specific layers rather than being uniformly distributed, we extract hidden states $h^{(l)} \in \mathbb{R}^d$ using a dataset of paired contrasting prompts (same context, opposite sentiment). We identify optimal intervention points via a two-step analysis:

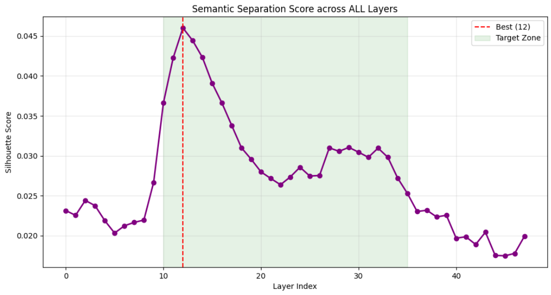

Figure 1: **Layer Separability Analysis (Happy vs Sad).** The Silhouette Score across layers shows peaks at Layer 12 (rhythm) and Layer 27 (timbre), selected as optimal injection points.

1. **Silhouette Score (Quantitative):** We rank layers by cluster separability using the Silhouette Score $s(i)$:
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{1}$$
where $a(i)$ is the mean intra-cluster distance and $b(i)$ the mean nearest-cluster distance. As shown in **Figure 1**, the score reveals distinct peaks (e.g., Layers 12 and 27), pinpointing where emotional distinction is maximized.

2. **PCA Visualization (Qualitative):** To validate these findings, we project the hidden states of top-performing layers onto a 2D plane using Principal Component Analysis (PCA). This visual inspection confirms that "Happy" and "Sad" representations form distinct, non-overlapping clusters in the selected layers (see Appendix C.1 for the layer 12 analysis).

3.2 Steering Vector Extraction

Once the optimal layers are identified, we proceed to compute the steering vector. We employ a **Contrastive Pair** approach to ensure that the extracted direction captures purely emotional features rather than other acoustic variations.

We construct a dataset of N pairs of prompts. Each pair i consists of a positive prompt $p_{pos}^{(i)}$ and a negative prompt $p_{neg}^{(i)}$ that share the same musical context (e.g., genre, instruments) but differ in emotional sentiment. For every pair, we extract the hidden states $h_{pos,i}^{(l)}$ and $h_{neg,i}^{(l)}$ at layer l and compute the difference vector $\delta_i^{(l)}$:

$$\delta_i^{(l)} = h_{pos,i}^{(l)} - h_{neg,i}^{(l)} \tag{2}$$

By computing this difference for each pair individually, we effectively cancel out the shared stylistic information, isolating the specific "emotional direction" for that sample. To obtain a robust global steering vector, we average these differences across the entire dataset. Finally, to ensure consistent application across layers, we apply L_2 normalization:

$$v^{(l)} = \frac{\frac{1}{N}\sum_{i=1}^{N}\delta_i^{(l)}}{\left\|\frac{1}{N}\sum_{i=1}^{N}\delta_i^{(l)}\right\|_2} \tag{3}$$

This results in a unit-norm vector $v^{(l)}$ that represents the pure direction of the semantic shift required to transform the internal representation from the source emotion to the target emotion.

3.3 Inference Injection

The final step is the runtime intervention. During the autoregressive generation loop, we intercept the forward pass of the Transformer. For a selected layer l, we modify the residual stream $x^{(l)}$ by adding the pre-computed steering vector.

However, a naive application of a constant vector often leads to signal degradation (e.g., noise or silence) as the sequence length increases. To address this, we implement a **Dynamic Decay** mechanism and a **Multi-Block Strategy**.

3.3.1 Dynamic Steering with Decay

To prevent the accumulation of artifacts and maintain audio fidelity, the steering strength is not constant but decays over time. The activation at time step t for layer l is modified as follows:

$$\hat{x}_t^{(l)} = x_t^{(l)} + \alpha(t) \cdot v^{(source)} \tag{4}$$

where $\alpha(t)$ is defined by an exponential decay function:

$$\alpha(t) = \alpha_0 \cdot \gamma^t \tag{5}$$

Here, α_0 is the initial injection strength and $\gamma \in (0, 1]$ is the decay factor. This ensures that the steering is strong at the beginning of the generation to set the emotional tone, but gradually fades to allow the model to maintain structural coherence without saturation (see Appendix B.4 for a comparison of decay strategies).

3.3.2 Multi-Block One-to-Many Strategy

Rather than extracting and injecting a unique vector for every single layer (One-to-One), we adopt a **One-to-Many** approach. We extract a robust steering vector from a highly separable "source" layer and broadcast it to a block of adjacent target layers.

Our selection of target blocks is not arbitrary but grounded in recent interpretability findings. In our final configuration, we target two discontinuous blocks to control different musical attributes simultaneously:

- **Mid-Block (Layers 11-14):** Injected with the vector extracted from Layer 12. This choice aligns with **Facchiano et al.** [3], who identified the middle section of the Transformer (approximately layers 10-18) as the primary locus for low-level features such as **tempo and brightness**.

- **Deep-Block (Layers 27-29):** Injected with the vector extracted from Layer 27. Following the insights of **Panda et al.** [5], who observed optimal results **for style and timbre transfer** in the deeper layers (around layer 27), we target this region to influence the texture of the generation without disrupting the rhythmic structure established in the earlier layers (limitations of single-block approaches are discussed in Appendix B.3).

Crucially, our framework allows setting distinct α_0 values for each block, enabling fine-grained balancing between rhythmic and timbral steering.

4 Dataset And Evaluation

In this section, we present the experimental framework used to validate our proposed emotion steering method. We first detail the curation of a custom dataset designed to enable both the extraction of steering vectors and the assessment of their generalization capabilities. Subsequently, we outline our multi-modal evaluation strategy, which combines objective semantic and acoustic metrics with subjective human assessment to ensure a robust analysis of the steering effects.

4.1 Dataset Construction

To address the lack of standard benchmarks for emotion steering, we curated a custom dataset where all textual prompts were generated by Gemini [4] and served as inputs for MusicGen to synthesize the audio tracks. The data is organized into two distinct subsets:

- **The Extraction Dataset**, used to compute the steering vectors, consists of 50 pairs of contrastive prompts (p_{pos}, p_{neg}). To ensure the vectors capture the pure emotional direction rather than instrument-specific features, we restricted this dataset to four distinct instruments: **Piano, Guitar, Violin, and Flute** (we discuss the limitations of single-instrument extraction in Appendix B.2). For each, we generated matched pairs representing opposite emotional poles within the same context (e.g., "A happy and cheerful classical violin solo" vs. "A sad and depressive classical violin solo").
- **The Inference Dataset** comprises 20 neutral prompts designed to evaluate steering on unbiased inputs. It is structured to test both specific and general capabilities:
  - The first 16 prompts focus on the **same four single-instrument categories** used in extraction (4 prompts each for Piano, Guitar, Violin, and Flute) to validate in-domain performance.
  - The remaining 4 prompts describe **complex, articulated musical scenarios** (e.g., orchestral or multi-instrument pieces) to assess the method's generalization capabilities on unseen contexts.

4.2 Evaluation Approach

To ensure a comprehensive assessment of our steering framework, we adopt a multi-modal evaluation strategy that combines quantitative analysis with qualitative human perception. We first employ objective metrics to measure semantic alignment and acoustic structural changes, followed by a subjective listening test to validate the perceptual effectiveness of the induced emotional shifts.

4.2.1 Objective Evaluation: CLAP Score

To quantify emotional intensity, we adopted a zero-shot classification approach using the **CLAP** model (checkpoint `laion/clap-htsat-unfused`). We defined a *Valence Score* based on the model's output probabilities. For a given audio x, the pipeline compares the audio embedding against two opposing text anchors (L_{pos} and L_{neg}), returning a probability score for each. The final metric is the difference between these scores:

$$\text{Score}(x) = P(L_{pos}|x) - P(L_{neg}|x) \tag{6}$$

The labels ("Happiness vibes" vs. "Sadness vibes") were selected empirically to maximize separation on the Extraction Dataset. The resulting score ranges from -1 (Sad) to $+1$ (Happy).

Finally, we computed the *Steering Delta* (ΔS) as the shift relative to the original generation:

$$\Delta S = \text{Score}(x_{steered}) - \text{Score}(x_{orig}) \tag{7}$$

A positive ΔS indicates a shift towards the positive pole, while a negative value indicates a shift towards the negative pole.

4.2.2 Subjective Evaluation: Blind Listening Test

Given the subjective nature of music perception, we complemented the objective metrics with a human evaluation campaign, collecting approximately 40 responses via a custom web-based interface.

We selected the top 5 generated examples, producing their respective positive and negative variations for a total of **15 audio tracks**. To prevent comparative bias, tracks were not presented in groups; instead, users listened to the 15 samples in a completely mixed order, unaware that the list contained related variations of the same musical ideas. This ensured that each track was evaluated independently based solely on its acoustic properties.

For each sample, participants were asked to classify the predominant emotion into one of three categories:

- **Happy:** The music feels joyful, bright, or energetic.
- **Sad:** The music feels melancholic, gloomy, or depressive.
- **Neutral:** The music lacks a strong emotional direction.

This rigorous approach provides insight into whether the steering effect is perceptibly significant to human listeners without the influence of side-by-side comparisons.

5 Experimental Results

This section details our experimental findings, starting with an objective analysis of the semantic and structural shifts on the inference dataset, followed by a human evaluation to validate perceptual effectiveness.

5.1 Objective Evaluation Results: CLAP Score

The injection of steering vectors resulted in a consistent and directionally accurate shift in emotional valence. As shown in Table 1, the positive steering induced a positive delta (Δ+0.23), effectively raising the valence scores to align the audio with the ”Happiness” concept. Conversely, the negative steering reduced the valence with a negative delta ($\Delta - 0.13$), driving the model's output towards ”Sadness”. Notably, the absolute CLAP scores observed during inference (Pos: -0.09, Neg: -0.45) are remarkably aligned with the baseline values recorded on the extraction dataset (Pos: -0.11, Neg: -0.48; detailed score distributions are provided in Appendix C.2). This convergence confirms that the steering vectors capture a generalized emotional representation that transfers robustly to unseen prompts.

Table 1: Objective Semantic Evaluation (CLAP Valence). Comparison of average scores between the Extraction Dataset and Inference Dataset (Test) across steering conditions.

	Extraction Set		**Inference Set**		
Metric (CLAP)	**Pos**	**Neg**	**Pos**	**Orig**	**Neg**
Valence Score	-0.11	-0.48	-0.09	-0.32	-0.45
Delta (Δ vs Orig)	-	-	**+0.23**	-	**-0.13**

5.2 Subjective Evaluation: Blind Listening Test

The human evaluation, summarized in Table 2, reveals a much stronger perceptual shift than the objective metrics alone suggested. The positive steering achieved a substantial subjective valence increase (Δ+0.377), while the negative steering proved even more effective ($\Delta - 0.507$). These results demonstrate that listeners, on average, correctly classified the emotional valence of the tracks, consistently aligning their perception with the intended steering direction.

Table 2: Blind Listening Test Results. Average Subjective Valence Scores and shift (Δ) from original.

Condition	**Avg. Score**	**Δ from Orig.**
Original	0.085	-
Positive Steering	0.470	**+0.385**
Negative Steering	-0.415	**-0.500**

6 Conclusions and Future Work

In this work, we investigated activation steering to control high-level emotional attributes in MusicGen without fine-tuning. Unlike prior studies focused on objective properties like instrumentation or tempo [cite references], we targeted the subjective domain of emotion, which emerges from a complex interplay of acoustic features rather than a single attribute. Our findings suggest that MusicGen encodes emotion as a distributed, multi-level representation—dependent on the synergy of tempo, brightness, melody, and timbre—rather than isolating it within a single dedicated mechanism. However, the contrast between clear human perception and noisier objective metrics highlights the complexity of steering abstract concepts compared to simple structural elements.

Future work should extend beyond the ”Happy vs. Sad” dichotomy to analyze other emotion pairs, investigating whether different emotional dimensions are localized in the same model layers. Additionally, exploring the continuous Valence-Arousal plane and developing robust metrics that better correlate with human perception remain critical challenges for controllable music generation.

References

[1] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[2] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

[3] S. Facchiano et al. Activation patching for interpretable steering in music generation, 2025. Unpublished manuscript/In press.

[4] Google Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[5] Dipanshu Panda et al. Fine-grained control over music generation with activation steering, 2025. Unpublished manuscript/In press.

[6] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv e-prints*, pages arXiv–2308, 2023.

[7] Yusong Wu, Ke Chen, Tianyu Zhang, Yutong Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

A Dataset Samples

To ensure reproducibility, we provide real examples of the prompts used in our pipeline, extracted directly from our custom dataset.

A.1 Extraction Dataset Samples

The Extraction Dataset consists of pairs of prompts designed to be semantically identical except for the emotional valence. Table 3 shows examples for the Piano and Violin categories.

Category	Prompt Pair (Positive vs. Negative)
Piano	***Pos:*** A solo piano melody, exhibiting a happy, cheerful and bright mood. ***Neg:*** A solo piano melody, exhibiting a sad, depressive and dark mood.
Violin	***Pos:*** A happy and cheerful classical violin solo. ***Neg:*** A sad and depressive classical violin solo.

Table 3: **Extraction Dataset.** Sample contrastive pairs used to compute the steering vectors. The context remains fixed while the sentiment is inverted.

A.2 Inference Dataset Samples

The Inference Dataset contains neutral prompts to test the steering effect on unbiased inputs. Table 4 presents a simple, single-instrument prompt (Guitar) used for in-domain testing, and a complex, multi-instrument prompt (Orchestral) used to test generalization.

Complexity	Neutral Prompt
Simple	A solo acoustic guitar strumming simple chords.
Complex	A full orchestral intro, cinematic and dense texture.

Table 4: **Inference Dataset.** Examples of neutral prompts. The "Complex" category challenges the model with dense textures unseen during the vector extraction phase.

B Ablation Studies and Design Choices

This section details alternative strategies explored during development that yielded suboptimal results, justifying our final architectural choices.

B.1 Explicit Melodic Conditioning

While our final method utilizes the `musicgen-melody` checkpoint in a text-only mode, we initially experimented with providing explicit **neutral melodic inputs** alongside the text prompts.

However, we observed that conditioning the generation on a pre-existing audio melody significantly reduced the steering effectiveness. As shown in Figure 2, the Silhouette Score analysis revealed a much flatter profile compared to the text-only approach: the peak in the mid-layers (critical for rhythmic and brightness control) was approximately more than 3x lower. We hypothesize that the explicit melodic input "locks" the structural features of the generation, leaving less freedom for the steering vector to modulate the emotional attributes. Consequently, we opted to use the model without audio conditioning to maximize steerability.

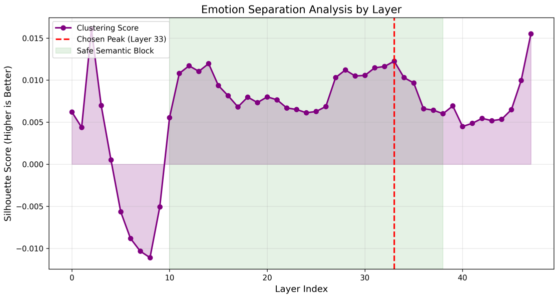

Figure 2: **Layer Analysis (With Melodic Conditioning).** When the model is conditioned on an input melody, the cluster separability in the mid-layers (10-18) collapses. This constraint limits the effectiveness of the steering compared to the text-only generation.

B.2 Piano-Only Extraction Dataset

We hypothesized that extracting vectors solely from Piano tracks (considered the most expressively versatile instrument) would produce cleaner emotional directions. Contrary to expectations, this approach resulted in poor generalization. When applied to other instruments (e.g., violin, flute), the steering vector tended to transfer the *timbre* of the piano rather than the abstract emotional valence, suggesting that the vector encoded "piano-ness" rather than "happiness".

B.3 Single-Block vs. Multi-Block Injection

Interventions on single layers or single blocks proved insufficient.

- **Single-Block:** Injecting only in the Mid-Block affected tempo but lacked textural depth, while the Deep-Block alone changed timbre without altering the energy.
- **Balanced Multi-Block:** Our final strategy uses a **Multi-Block** approach with differential gains. We assign a higher α to the Mid-Block (due to its higher Silhouette Score and robustness) and a lower α to the Deep-Block, as the latter proved more prone to introducing audio artifacts when heavily perturbed.

B.4 Alpha Decay Calibration

Finding the correct decay rate was critical. We initially tested an "aggressive" decay (e.g., $\gamma = 0.99$) to minimize artifacts. However, as visualized in Figure 3, this caused the steering value to drop to near-zero too rapidly. This sudden loss of conditioning signal created a "contextual shock" for the model: the drastic shift from a steered state to a neutral state within few seconds introduced severe noise and incoherence. A slow decay ($\gamma = 0.998$) was necessary to maintain a smooth, consistent emotional trajectory.

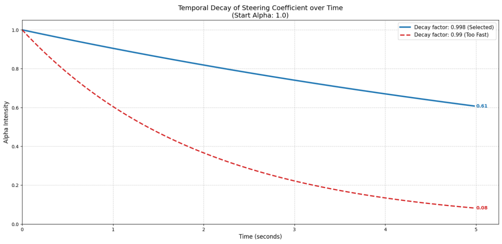

Figure 3: **Decay Strategy Comparison.** Aggressive decay (red) vanishes too quickly, causing discontinuity. The selected slow decay (blue) maintains the emotional context throughout the generation.

B.5 Acoustic Feature Metrics

We attempted to quantify changes in tempo and brightness using standard signal processing metrics (BPM and Spectral Centroid). However, these measurements proved unreliable. The trade-off required to achieve strong emotional steering inevitably introduces some audio artifacts and background noise; this signal degradation interfered with the feature extraction algorithms, yielding inconsistent results compared to the semantic (CLAP) and perceptual (Human) evaluations.

C Additional Visualizations

This section provides a deeper visual insight into the model's internal representations and the distribution of evaluation scores.

C.1 Layer 12 PCA Analysis

Figure 4 visualizes the latent space of the selected intervention point. We chose Layer 12 because it represents the peak of emotional separability within the model, despite the geometric complexity inherent to the mid-layers (associated with tempo and brightness [3]).

C.2 Detailed Score Distribution

While Section 5 reported the average metrics, this section analyzes the distribution of individual samples to assess the consistency of the steering mechanism.

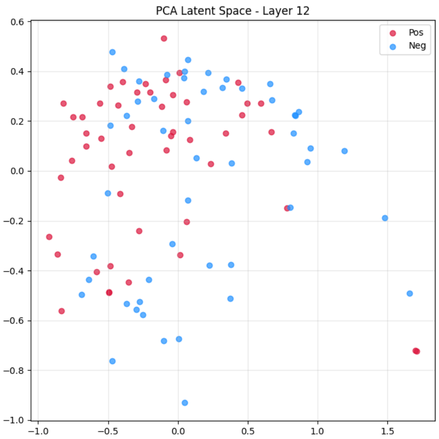

Figure 4: **PCA Projection of Layer 12.** Although this layer offers the highest relative separability, the absolute Silhouette Score remains low (≈ 0.045). The visual overlap suggests a complex, non-linear boundary, indicating that features like rhythm and brightness are difficult to disentangle in the context of emotional representation.

C.2.1 Extraction Dataset Validation

To validate the source data for vector computation, we analyzed the score distributions of the Extraction Dataset. As shown in Figures 5 and 6, while the metric acts more as a comparative tool than an absolute classifier, the systematic separation between the two distributions confirms that the dataset successfully captures the necessary emotional opposition.

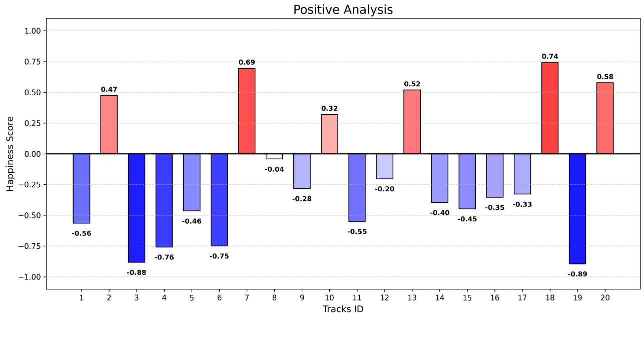

Figure 5: **CLAP Distribution (Positive Extraction).** While absolute scores remain low (mean ≈ -0.11), the distribution is significantly shifted up compared to the negative set, validating the semantic distinction of the "Happy" prompts.

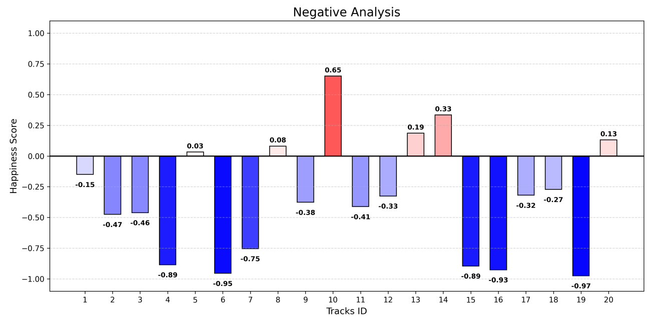

Figure 6: **CLAP Distribution (Negative Extraction).** The distribution mass is consistently lower (mean ≈ -0.48), confirming that the "Sad" prompts occupy a distinct, more negative region of the embedding space.

C.2.2 Inference Analysis

We extended the analysis to unseen prompts (Inference Dataset). As seen in Figure 7, the neutral baseline is unstable and instrument-dependent. However, the steering mechanism proves robust against this variance. Figures 8 and 9 show the raw score distributions, while Figures 10 and 11 isolate the net effect of the intervention (ΔS). The consistent shift across both metrics confirms that the steering vector effectively overrides the initial bias of the prompt.

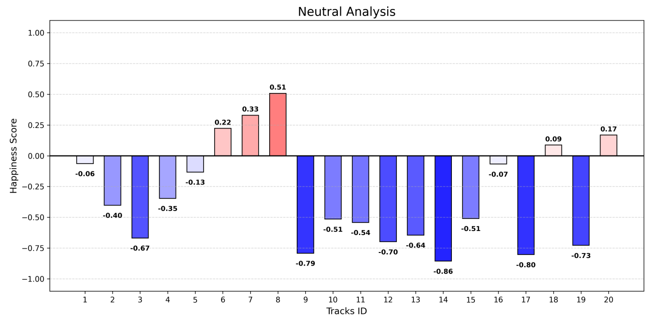

Figure 7: **CLAP Neutral Baseline Distribution.** The lack of centering around 0.0 highlights the intrinsic bias of different instruments (e.g., guitar vs. piano) and the inherent challenge of generating truly neutral audio.

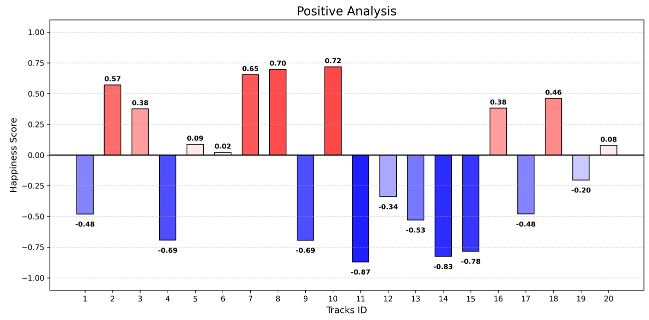

Figure 8: **CLAP Positive Steering (Inference).** Regardless of the neutral baseline, the injection shifts the probability mass towards higher values, aligning the generation with the positive emotional pole.

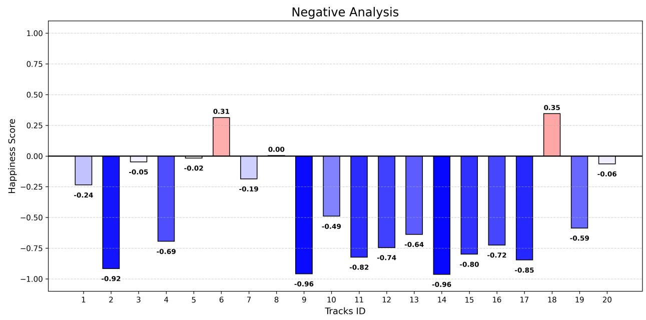

Figure 9: **CLAP Negative Steering (Inference).** Symmetrically, the negative steering consistently pushes the scores towards the lower end of the spectrum.

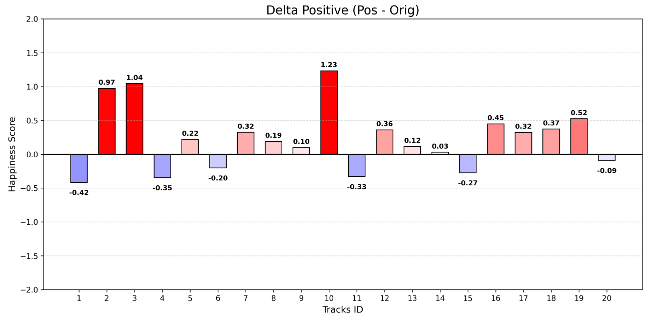

Figure 10: **Positive Delta Distribution (ΔS).** This plot isolates the steering effect by subtracting the original score. The positive mean ($\approx +0.23$) confirms a consistent emotional uplift.

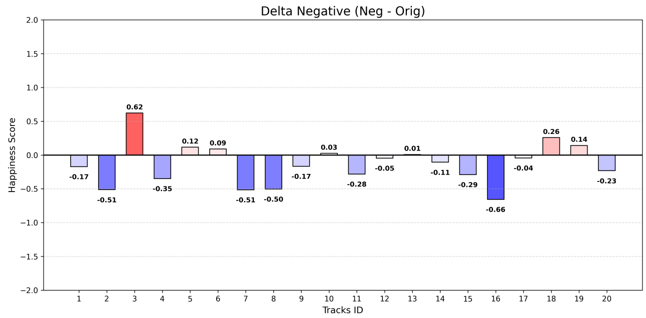

Figure 11: **Negative Delta Distribution (ΔS).** The negative mean (≈ -0.13) demonstrates that the method successfully dampens the emotional tone, shifting it towards sadness.

C.2.3 Human Perception Analysis

Finally, the Human Evaluation serves as the perceptual ground truth. While the **neutral baseline** (Figure 12) displays high variance due to subjective interpretation, the steered conditions (Figures 13 and 14) exhibit a **sharp polarization** of listener responses.

The injection of the vectors effectively overrides the inherent ambiguity of the generation, aligning the consensus decisively with the target emotion. This confirms that the mathematical shift observed in the embedding space translates into a tangible acoustic transformation clearly perceptible by human listeners.

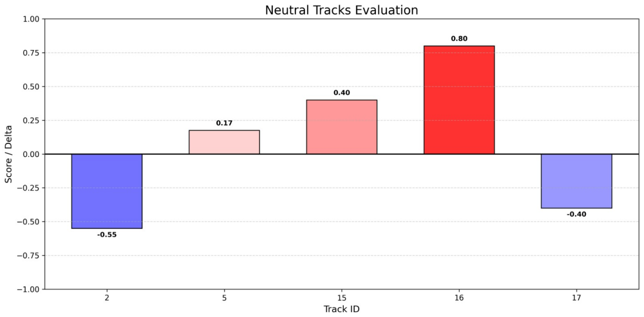

Figure 12: **Human Evaluation (Neutral Baseline).** The distribution of user responses for unsteered tracks shows no dominant consensus. This highlights the subjective nature of ”neutral” music, where listeners project different emotions based on subtle acoustic cues.

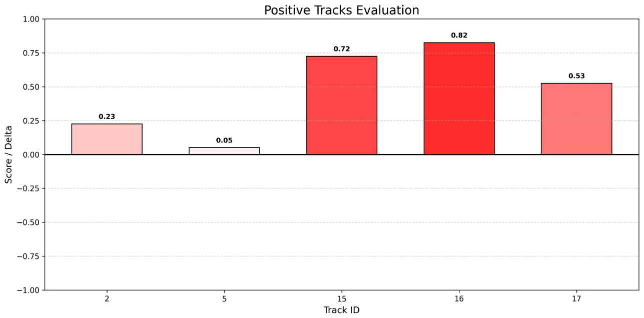

Figure 13: **Human Evaluation (Positive Steering).** The injection of the ”Happy” vector creates a clear consensus. The majority of listeners correctly classified these tracks as ”Happy”, demonstrating the perceptual effectiveness of the steering.

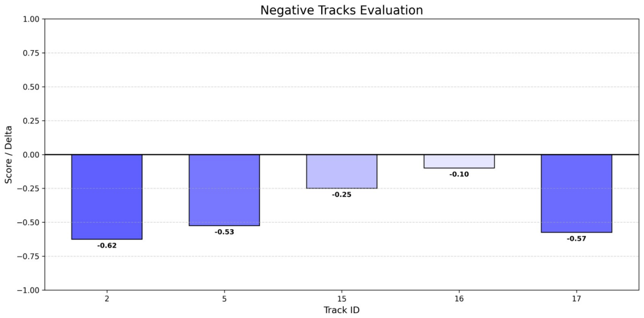

Figure 14: **Human Evaluation (Negative Steering).** Similarly, the ”Sad” vector induces a strong perceptual shift. The overwhelming majority of responses align with the target emotion, validating the symmetry of our method.