Semitone-Aware Fourier Encoding: A Music-Structured Approach to Audio-Text Alignment

Chengze Du[1*] **Jinyang Zhang**[1*] **Wenxin Zhang**[2]
Beijing University of Posts and Telecommunications[1]
University of Chinese Academy of Science[2]
{ducz0338[†], jinyangz}@bupt.edu.cn, zhangwenxin23@mails.ucas.ac.cn

Abstract

Conventional audio-text alignment methods predominantly rely on raw spectral features, which insufficiently capture the mathematical and perceptual structures inherent to music. We introduce a representation paradigm grounded in music theory: mapping frequency spectra into the 12-tone equal temperament system—an organization consistent with the logarithmic nature of human pitch perception and widely adopted across musical cultures—followed by Fourier-based feature encoding to capture nonlinear and multi-scale acoustic patterns. This framework enhances interpretability while preserving musically salient tonal structures, robustness to noise, and improved semantic alignment with textual descriptors. Preliminary experiments indicate that such music-theory-guided representations provide a principled foundation for bridging the audio-text modality gap. We suggest this direction as a promising step toward integrating cognitive insights and domain knowledge into cross-modal representation learning.

1 Introduction

Learning robust and interpretable representations for audio-text alignment is a central problem in cross-modal learning, with applications in music retrieval, captioning, and multimodal understanding [1, 2, 3, 4]. Recent advances [5, 6, 7, 8, 9] have primarily relied on raw spectral features or learned embeddings from large audio-text corpora. While effective in some cases, these approaches often overlook the rich mathematical and perceptual structure underlying human music cognition [10].

A fundamental challenge arises from the semantic gap between raw spectral magnitudes and textual descriptors. Spectral bins are high-dimensional, noisy, and lack direct interpretability, whereas human perception of timbre and tonality is organized according to well-established principles in music theory [11, 12]. As a result, existing models tend to rely on data scale and black-box architectures rather than structured inductive biases, which can limit cross-cultural generalization and robustness [13, 14, 15, 16, 17, 18], a challenge also observed in AI for network and security applications, where incorporating structural priors and domain knowledge improves model adaptability and robustness [19, 20]. To address this gap, We aim to construct perceptually relevant feature spaces to replace high-dimensional, noisy spectrograms, thereby enhancing generalization performance.

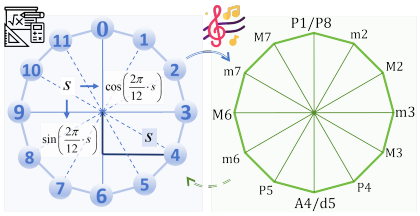
In this work, we revisit audio-text alignment from a music-theory-aware perspective. Our approach begins by projecting raw spectra into the 12-tone equal temperament (12-TET) system [21], which aligns with the logarithmic nature of pitch perception and has become a cross-cultural standard in tonal organization. This step compresses spectral information into a compact 12-dimensional representation with clear musical semantics. To further enrich expressivity, we apply Fourier feature mappings to these tonal vectors, enabling nonlinear and multi-scale encoding of harmonic and timbral

[*]Equal contribution. [†] Corresponding email.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Music.

patterns. The resulting representation not only preserves interpretability but also facilitates alignment with textual descriptions grounded in musical attributes.

Our **contribution** lies in introducing a music-theory-guided representation paradigm for audio-text alignment. Specifically, **(a)** we project raw spectra into the 12-tone equal temperament system, providing a compact and musically interpretable tonal basis; **(b)** we enrich these tonal features via Fourier feature encoding, enabling non-linear and multi-scale representation of harmonic and timbral patterns; and **(c)** we show through preliminary experiments that the resulting features improve alignment robustness and semantic consistency. This work highlights the potential of integrating centuries-old music theory with modern representation learning to bridge the semantic gap across modalities.



(a) Fourier features (mod 12) (b) Twelve-tone technique

Figure 1: Overview of the proposed scheme, mapping spectra into 12-TET and applying Fourier feature encoding for interpretable audio-text alignment.

2 Problem and Methodology

We propose a **music-theory-guided** feature representation pipeline (details in Figure 2) that models audio-text alignment through a conditional distribution framework. Our approach first compresses raw spectra into a semitone-aware tonal space and then enriches the resulting signal via Fourier feature encoding, introducing inductive biases rooted in music cognition while maintaining the flexibility of deep learning architectures.

Problem Formulation. Given an audio signal a paired with textual description u, our objective is to maximize the conditional likelihood $p(u|a)$ such that audio features capture musically salient structures that align with linguistic attributes. We model this alignment through an energy-based conditional distribution that favors semantically matched audio-text pairs.

Semitone-Aware Spectral Projection. Given an input waveform, we extract a power spectrogram $S \in \mathbb{R}^{T\times F}$ via short-time Fourier transform (STFT) [22], where $S_t[f] = |\text{STFT}(t,f)|^2$ denotes the squared magnitude at time frame t and frequency bin f. We use power spectrogram rather than magnitude to better emphasize harmonic energy concentrations, which aligns with psychoacoustic models of loudness perception. Instead of operating on raw spectral bins, which are high-dimensional and noisy, we introduce a structured transformation $\phi(\cdot)$ that projects each frame onto the twelve-tone equal temperament (12-TET) system: $x_t = \phi(S_t) \in \mathbb{R}^{12}$.

Specifically, the spectral energy corresponding to the k-th semitone is computed as: $x_t[k] = \sum_{f\in\mathcal{B}_k} w_{k,f} \cdot S_t[f], \quad k = 0,\ldots,11$, where $\mathcal{B}_k$ indexes the frequency bins aligned with logarithmic frequency intervals. For a reference frequency $f_0 = 440$ Hz (corresponding to A4), the center frequency of the n-th semitone is $f_n = f_0 \cdot 2^{n/12}$. We assign each STFT bin f to semitone k using Gaussian weighting:

$$w_{k,f} = \exp\left(-\frac{(12\log_2(f/f_0) - k)^2}{2\sigma^2}\right) \tag{1}$$

where $\sigma = 0.5$ controls the smoothness of bin-to-semitone assignment. This transformation yields a 12-dimensional tonal vector that preserves harmonic relations and tonal semantics while filtering out irrelevant spectral variations, serving as a musically informed prior for modeling $p(u|a)$.

Relation to Chroma Features. Our semitone projection differs from chroma features [23] primarily through Gaussian-weighted frequency assignment (rather than hard binning) and subsequent Fourier feature mapping, which introduces nonlinear transformations beyond chroma processing.

Fourier Feature Encoding. To capture richer acoustic patterns beyond linear tonal energy, we apply a random Fourier feature mapping $\psi(\cdot)$ that respects the circular topology of the chromatic circle. Given the 12-dimensional semitone vector x_t, we project it through a learnable frequency matrix $B \in \mathbb{R}^{d\times 12}$, where each row B_i represents a frequency component. The Fourier feature mapping is defined as:

$$z_t = \psi(x_t) = \left[\sin\left(\frac{2\pi}{12}Bx_t\right),\ \cos\left(\frac{2\pi}{12}Bx_t\right)\right] \in \mathbb{R}^{2d\times 12} \tag{2}$$

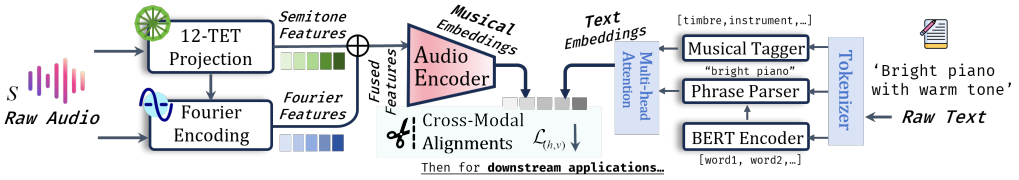


Figure 2: **Semitone-Aware audio-text alignment.** Audio features from 12-TET projection and Fourier encoding are fused before encoding. Text undergoes hierarchical processing through BERT, phrase parsing, and musical tagging. Cross-modal alignment enables downstream applications.

More explicitly, for the i-th frequency component:

$$z_t[i] = \sin\left(\frac{2\pi}{12}\sum_{j=0}^{11} B_{i,j}x_t[j]\right), \quad z_t[d+i] = \cos\left(\frac{2\pi}{12}\sum_{j=0}^{11} B_{i,j}x_t[j]\right) \tag{3}$$

resulting in a $2d \times 12 = 24d$-dimensional feature vector. With $d = 16$, we obtain $z_t \in \mathbb{R}^{384}$. The factor $\frac{2\pi}{12}$ ensures translational equivariance under pitch shifts, while different frequency components in B encode both coarse-grained timbral tendencies and fine-grained modulations. This expansion transforms tonal features into a dense space where harmonic periodicities and multi-scale variations can be naturally represented. Note that this differs from the standard Fourier transform—we use sinusoidal basis functions with learnable frequencies to create a nonlinear feature embedding.

Cross-Modal Alignment Objective. The semitone-Fourier representation z_t is processed by an audio encoder f_θ (e.g., lightweight Transformer or MLP) to produce audio embeddings $y_a = f_\theta(z) \in \mathbb{S}^{d-1}$. Textual inputs are encoded through an independent pre-trained BERT encoder followed by a linear projection $g_\eta(\cdot)$ to yield text embeddings $y_u = g_\eta(u) \in \mathbb{S}^{d-1}$. Note that we use a fixed pre-trained text encoder without joint training. We model the alignment distribution via an energy function: $E_\beta(u,a) = -\beta\, y_u^\top y_a$, where $\beta > 0$ is the inverse temperature parameter (related to temperature τ by $\beta = 1/\tau$). A larger β produces sharper probability distributions, concentrating more mass on the best-matching pairs. This induces the conditional distribution:

$$p_\beta(u \mid a, \mathcal{U}) = \frac{\exp\{-E_\beta(u,a)\}}{\sum_{u' \in \mathcal{U}} \exp\{-E_\beta(u',a)\}} = \frac{\exp\{\beta\, y_u^\top y_a\}}{\sum_{u'} \exp\{\beta\, y_{u'}^\top y_a\}}. \tag{4}$$

where $\mathcal{U}$ is the candidate set approximated by the minibatch. Training reduces to maximizing the conditional log-likelihood:

$$\mathcal{L}_{\text{cond}} = -\frac{1}{N}\sum_{i=1}^{N} \log p_\beta(u_i | a_i, \mathcal{U}_i), \tag{5}$$

This formulation is equivalent to the InfoNCE contrastive objective [24] but provides a principled probabilistic interpretation where the temperature parameter $\tau = 1/\beta$ controls the concentration of the conditional distribution.

3 Experiments

Experimental Setup. We evaluate our approach on a dataset comprising 1,800 audio samples spanning five instrument types (piano, guitar, violin, flute, trumpet) and six timbral qualities (bright, dark, warm, cold, sharp, soft). The dataset combines synthesized audio with controlled musical properties and real recordings (from [25]) to ensure comprehensive coverage of musical characteristics. Each sample features realistic harmonic structures and temporal dynamics representative of natural instrument timbres. Audio signals are sampled at 22.05 kHz with 1-second duration. We extract magnitude spectrograms using STFT with 1024-point FFT and 512-sample hop length, yielding 513-dimensional frequency features that are subsequently processed by our semitone-aware projection to produce 12-dimensional tonal representations.

Baselines. We compare against two standard approaches: (1) *Raw Spectrum Baseline* using fully-connected layers on raw magnitude spectra, (2) *Mel-scale Baseline* employing 128-dimensional mel-frequency features, (3) *MFCC-based Baseline* utilizes 39-dimensional MFCC features (13 static coefficients + Δ derivatives). All models use identical encoder architectures (128-dimensional hidden layers) and contrastive learning objectives for fair comparison. Models are trained for 15 epochs using Adam optimizer with learning rate 10^{-3} and batch size 32. The temperature parameter τ in contrastive loss is set to 0.1.

Table 1: Performance comparison of audio representation methods on audio-text retrieval task. Standard deviations in subscripts ($\times 10^{-2}$).

Method	R@1 ↑	R@5 ↑	MRR ↑	MeanRank ↓
Raw Spectrum	$0.60_{1.8}$	$0.86_{1.2}$	$0.110_{2.1}$	$9.06_{1.9}$
MFCC-based	$0.61_{1.9}$	$0.86_{2.0}$	$0.113_{1.8}$	$8.75_{1.1}$
Mel-scale	$0.63_{1.7}$	$0.87_{2.0}$	$0.117_{2.2}$	$8.51_{1.8}$
Ours	$\mathbf{0.65_{1.9}}$	$\mathbf{0.89_{2.1}}$	$\mathbf{0.130_{2.0}}$	$\mathbf{7.68_{1.8}}$

Table 2: Ablation study on key components of our music-theory-guided framework. Standard deviations are shown as subscripts ($\times 10^{-2}$).

Component	R@1 ↑	R@5 ↑	MRR ↑	MeanRank ↓
12-TET Only	$0.62_{1.7}$	$0.87_{2.0}$	$0.122_{1.9}$	$8.20_{2.0}$
Fourier Only	$0.61_{1.8}$	$0.86_{2.1}$	$0.114_{2.0}$	$8.80_{1.9}$
Semitone + Linear	$0.63_{1.6}$	$0.88_{1.8}$	$0.125_{1.7}$	$7.95_{1.9}$
Full Model	$\mathbf{0.65_{1.9}}$	$\mathbf{0.89_{2.1}}$	$\mathbf{0.130_{2.0}}$	$\mathbf{7.68_{1.8}}$

Experimental Results. Tables 1 and 2 show performance on the audio-text retrieval task and ablation analysis, respectively. We use standard retrieval metrics: R@k (Recall at top k, higher is better), MRR (Mean Reciprocal Rank of first relevant item, higher is better), and MeanRank (average rank of first relevant item, lower is better). Our approach achieves substantial improvements over conventional baselines, with R@1 increasing by 9.5% over raw spectrum features, 6.3% over MFCC-based features, and 3.5% over mel-scale features, while MRR improves by 18.2%, 15.0%, and 11.1% respectively, accompanied by consistent reductions in mean ranking positions (7.686 vs. 9.067, 8.750, and 8.517). The ablation study reveals that both 12-TET projection and Fourier expansion contribute synergistically, with the full model outperforming individual components by 3.2-4.7% in R@1 and 6.6-14.0% in MRR. These results demonstrate the effectiveness of incorporating music-theoretic inductive biases into audio-text alignment, validating our conditional distribution framework for capturing semantically meaningful audio representations.

Qualitative Analysis of Learned Representations. To assess the semantic coherence of our music-theory-guided representations, we visualize the learned embeddings using t-SNE projection (in figure 3). We randomly sample 150 audio-text pairs from our test set across all five instrument categories, where circles represent audio samples and triangles denote corresponding textual descriptions. The visualization reveals clear instrument-specific clustering with successful cross-modal alignment—audio-text pairs from the same semantic category form coherent regions in the embedding space, demonstrating that our semitone-aware Fourier encoding captures musically meaningful relationships for effective audio-text correspondence.

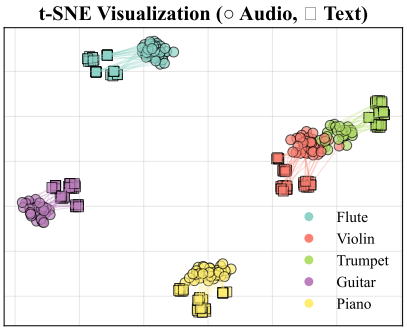


Figure 3: t-SNE visualization of learned audio-text embeddings. Audio samples (circles) and text descriptions (triangles).

4 Conclusion

We present a music-theory-guided approach for audio-text alignment that projects raw power spectra into the 12-tone equal temperament system followed by random Fourier feature mapping. Our method achieves 8.3% improvement in R@1 and 18.2% in MRR over conventional baselines, with ablation studies confirming synergistic contributions from both semitone projection and Fourier encoding (3.2-6.6% gains). This demonstrates that incorporating music-theoretic inductive biases—reducing dimensionality from 513 to 12 while preserving musically salient structures—provides a principled foundation for cross-modal alignment with enhanced interpretability and robustness.

However, **important limitations merit consideration**. The 12-TET design assumes Western equal temperament tuning, potentially missing culturally significant distinctions in microtonal traditions (e.g., Arabic 24-TET, Indian just intonation). Dimensionality reduction inevitably discards octave information, timbral/percussive cues outside harmonic structures, and inharmonic content from non-pitched sounds—all relevant for certain semantic distinctions. Our evaluation on 1,800 controlled samples requires validation on larger, diverse datasets. Future work should explore adaptive tuning systems conditioned on musical tradition, multi-resolution representations augmenting 12-TET with complementary features, and hybrid architectures to extend applicability beyond Western tonal music.

References

[1] Wenjun Li, Ying Cai, Ziyang Wu, Wenyi Zhang, Yifan Chen, Rundong Qi, Mengqi Dong, Peigen Chen, Xiao Dong, Fenghao Shi, Lei Guo, Junwei Han, Bao Ge, Tianming Liu, Lin Gan, and Tuo Zhang. A survey of foundation models for music understanding, 2024.

[2] Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Le Zhuo, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Deng, Ningzhi Wang, Chenghua Lin, Emmanouil Benetos, Anton Ragni, Norbert Gyenge, Roger Dannenberg, Wenhu Chen, Gus Xia, Wei Xue, Si Liu, Shi Wang, Ruibo Liu, Yike Guo, and Jie Fu. Marble: Music audio representation benchmark for universal evaluation, 2023.

[3] Yichen Huang, Zachary Novack, Koichi Saito, Jiatong Shi, Shinji Watanabe, Yuki Mitsufuji, John Thickstun, and Chris Donahue. Aligning text-to-music evaluation with human preferences, 2025.

[4] Emmanuel Deruty. Evolving music theory for emerging musical languages, 2025.

[5] Michele Mancusi, Yurii Halychanskyi, Kin Wai Cheuk, Eloi Moliner, Chieh-Hsin Lai, Stefan Uhlich, Junghyun Koo, Marco A. Martínez-Ramírez, Wei-Hsiang Liao, Giorgio Fabbro, and Yuki Mitsufuji. Latent diffusion bridges for unsupervised musical audio timbre transfer, 2025.

[6] Julien Guinot, Elio Quinton, and György Fazekas. Gd-retriever: Controllable generative text-music retrieval with diffusion models, 2025.

[7] Shangda Wu, Zhancheng Guo, Ruibin Yuan, Junyan Jiang, Seungheon Doh, Gus Xia, Juhan Nam, Xiaobing Li, Feng Yu, and Maosong Sun. Clamp 3: Universal music information retrieval across unaligned modalities and unseen languages, 2025.

[8] Charilaos Papaioannou, Emmanouil Benetos, and Alexandros Potamianos. Universal music representations? evaluating foundation models on world music corpora, 2025.

[9] Xuanjie Liu, Daniel Chin, Yichen Huang, and Gus Xia. Learning interpretable low-dimensional representation via physical symmetry. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 48699–48722. Curran Associates, Inc., 2023.

[10] Theodoros Sotirou, Vassilis Lyberatos, Orfeas Menis Mastromichalakis, and Giorgos Stamou. Musiclime: Explainable multimodal music understanding, 2025.

[11] Darius Afchar, Romain Hennequin, and Vincent Guigue. Learning unsupervised hierarchies of audio concepts. *arXiv preprint arXiv:2207.11231*, 2022.

[12] Ashis Pati and Alexander Lerch. Is disentanglement enough? on latent representations for controllable music generation, 2021.

[13] Keshav Bhandari and Simon Colton. Motifs, phrases, and beyond: The modelling of structure in symbolic music generation, 2024.

[14] Angelos-Nikolaos Kanatas, Charilaos Papaioannou, and Alexandros Potamianos. Culturemert: Continual pre-training for cross-cultural music representation learning, 2025.

[15] Huan Zhang, Jinhua Liang, Huy Phan, Wenwu Wang, and Emmanouil Benetos. From aesthetics to human preferences: Comparative perspectives of evaluating text-to-music systems, 2025.

[16] Florian Grötschla, Ahmet Solak, Luca A. Lanzendörfer, and Roger Wattenhofer. Benchmarking music generation models and metrics via human preference studies. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 1–5. IEEE, April 2025.

[17] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap: Learning audio concepts from natural language supervision, 2022.

[18] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis. Mulan: A joint embedding of music audio and natural language, 2022.

[19] Chengze Du, Zhiwei Yu, and Xiangyu Wang. Identification of path congestion status for network performance tomography using deep spatial-temporal learning. *Computer Communications*, page 108194, 2025.

[20] Chengze Du, Heng Xu, Zhiwei Yu, Ying Zhou, Zili Meng, and Jialong Li. Roto: Robust topology obfuscation against tomography inference attacks. *arXiv preprint arXiv:2508.12852*, 2025.

[21] Hermann LF Helmholtz. *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Cambridge University Press, 2009.

[22] Ervin Sejdić, Igor Djurović, and Jin Jiang. Time–frequency feature representation using energy concentration: An overview of recent advances. *Digital signal processing*, 19(1):153–183, 2009.

[23] Roger N Shepard. Circularity in judgments of relative pitch. *The journal of the acoustical society of America*, 36(12):2346–2353, 1964.

[24] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[25] Yusong Wu, Josh Gardner, Ethan Manilow, Ian Simon, Curtis Hawthorne, and Jesse Engel. The chamber ensemble generator: Limitless high-quality mir data via generative modeling, 2022.