Beyond Frequency: Dissonance Spectrum for Perceptually Motivated Music Understanding

Tianle Wang[1,2], Xinyi Tong[1,2], Liangke Zhao[1,2], Jishang CHEN[1,2], Sirui Zhang[1,2], Haoxin Zhang[1,2], Xin Jin[1], Duo XU[1], Xiaobing Li[2], Song-Chun Zhu[1,3]

[1]Beijing Institute for General Artificial Intelligence, Beijing, China
[2]Central Conservatory of Music, Beijing, China
[3]Peking University, Beijing, China

Abstract

Conventional music representations describe acoustic energy over time and frequency but do not explicitly expose relations among simultaneous frequency components. We introduce the *Dissonance Spectrum* (DS), a nonnegative time–frequency representation that applies a tolerance-based rational pitch-relation kernel with logarithmic harmonic distance to a constant-Q spectrum and attributes aggregate pairwise interactions back to individual frequency bins. Controlled music-theory tests show strong ordinal agreement for intervals, harmonic-function connections, and church modes, and moderate positive agreement across diverse chord voicings. DS is then encoded by a lightweight parallel branch whose zero-initialized residual projection preserves the baseline function at initialization. Across six paired training seeds in open-ended music question answering and categorical and dimensional music emotion recognition, DS obtains the highest mean on every reported endpoint relative to the unchanged baseline, a parameter-matched Gaussian-input branch, and an architecture-matched magnitude-CQT branch. These results support DS as an interpretable, complementary representation, while listener-specific perception and broader task coverage remain open problems.

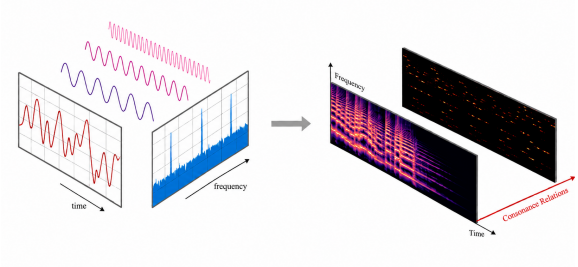

Figure 1: DS converts spectral energy into a time–frequency map of modeled consonance–dissonance relations.

Introduction

Music understanding extends beyond detecting acoustic events: listeners also respond to relations among simultaneous frequency components. Beating, critical-band interactions, periodicity, and harmonic organization contribute to consonance, dissonance, stability, tension, and emotion (von Helmholtz 1954; Plomp and Levelt 1965; Langner 1992; Stolzenburg 2015; Harrison and Pearce 2020). Neural systems learn rich musical information from waveforms, time–frequency representations, and audio–text supervision (Lu et al. 2021; Li et al. 2024; Won, Hung, and Le 2024; Huang et al. 2022; Elizalde et al. 2023; Liu et al. 2024; Deng et al. 2024). Yet spectral magnitude mainly marks active components, while learned embeddings entangle multiple attributes; neither explicitly localizes simultaneous relations under a specified consonance–dissonance model.

Music-specific inductive biases can complement end-to-end learning. Chroma and tonal features expose pitch-class organization, perceptual mid-level attributes connect audio to emotion, MERT uses a CQT teacher, and consonance-aware supervision improves chord estimation (Harte, Sandler, and Gasser 2006; Korzeniowski and Widmer 2016; Aljanaki and Soleymani 2018; Chowdhury et al. 2019; Li et al. 2024; Poltronieri, Serra, and Rocamora 2025). Lightweight mode injection likewise benefits symbolic emotion recognition (Xia et al. 2026). These results motivate a relational representation that makes information already present in the signal easier to inspect and learn.

We introduce the **Dissonance Spectrum** (Figure 1). DS combines tolerance-based rational approximation (Stolzenburg 2015) with logarithmic harmonic distance (Tenney 1984), applies the resulting kernel across a magnitude CQT, and attributes weighted pair relations back to their time–frequency locations. It is a deterministic reorganization of the signal, not an additional observed modality, and is computed efficiently by frequency-axis correlation.

We validate the operator on intervals, chord qualities, functional chord connections, and scales, then add a lightweight DS encoder to MU-LLaMA for open-ended music question answering and to Music2Emo for categorical and dimensional emotion recognition. Parameter-matched Gaussian and architecture-matched magnitude-CQT branches control for added capacity and a second pitch-resolved pathway. Six paired seeds and exploratory mechanism variants test whether gains are consistent with ordered relational structure rather than input resolution or model size alone.

Our contributions are threefold:

- We formulate a nonnegative time–frequency DS that localizes amplitude-weighted rational pitch relations and derive an efficient correlation implementation.

- We establish controlled evidence that the kernel and audio representation reproduce predefined ordinal trends across intervals, chord qualities, functional connections, and modes.
- We design lightweight plug-and-play adapters and show that DS achieves higher six-seed mean performance than the baseline, a parameter-matched Gaussian branch, and an architecture-matched CQT branch in two model families.

Related Work

Music representations and perceptual priors. The CQT provides a logarithmic frequency axis aligned with musical pitch, chroma folds energy into pitch classes, and tonal-space descriptors encode harmonic proximity (Brown 1991; Harte, Sandler, and Gasser 2006; Müller and Ewert 2011). Neural models learn broader musical attributes: SpecTNT separates spectral and temporal attention, MERT and MusicFM provide transferable music embeddings, and MuLan, CLAP, MU-LLaMA, and MusiLingo connect audio representations to language (Lu et al. 2021; Li et al. 2024; Won, Hung, and Le 2024; Huang et al. 2022; Elizalde et al. 2023; Liu et al. 2024; Deng et al. 2024). These representations are effective for downstream tasks but do not provide a stable, directly inspectable account of local consonance or dissonance. Recent human-written music-QA evaluation further emphasizes robustness to unimodal shortcuts, motivating cautious interpretation of automatically generated references and lexical-overlap metrics (Weck et al. 2026).

Music-specific priors remain complementary to learned representations. Chroma-based networks improve chord recognition; listener-rated mid-level attributes bridge audio and emotion; MERT uses a CQT teacher; and consonance-aware distances and label smoothing improve chord estimation (Korzeniowski and Widmer 2016; Aljanaki and Soleymani 2018; Chowdhury et al. 2019; Li et al. 2024; Poltronieri, Serra, and Rocamora 2025). DS follows this knowledge-guided direction but is computed continuously from audio and preserves time–frequency attribution.

Consonance perception and computational models. Consonance and dissonance denote related but nonidentical acoustic, perceptual, and music-theoretical concepts (Cazden 1980; Wand 2012). Interference accounts relate sensory dissonance to beating among nearby partials and auditory critical bands (von Helmholtz 1954; Plomp and Levelt 1965; Hutchinson and Knopoff 1978). Periodicity and harmonicity accounts associate consonance with compact common periods, virtual fundamentals, or harmonic templates (Langner 1992; Parncutt 1989; Stolzenburg 2015). Many implementations reduce these relations to a scalar, which supports ranking but does not identify the frequency regions responsible for the value.

Timbre can reshape consonance curves by changing partial locations and amplitudes (Sethares 2005; Marjieh et al. 2024). Comparative modeling and listener studies further indicate that roughness, periodicity, spectral structure, experience, register, and listener characteristics all contribute to consonance judgments (Cousineau, McDermott, and Peretz 2012; Harrison and Pearce 2020; Eerola and Lahdelma 2021; McDermott et al. 2016; Kaklamani and Simserides 2026). We therefore treat DS as a perceptually motivated low-level cue rather than a complete model of musical preference.

Recorded music and consonance-aware music AI. Applying dissonance models to recordings is difficult because real audio contains overlapping sources, transients, noise, tuning variation, and time-varying balance. Schwär, Balke, and Müller (2025) estimate time-varying sensory dissonance in multitrack recordings and attribute mixture-level dissonance to tracks. Scalar roughness descriptors and learned mid-level dissonance ratings have also supported music emotion recognition (Aljanaki and Soleymani 2018; Panda, Malheiro, and Paiva 2023). DS instead attributes aggregate relations to time–frequency bins, requires no stems or chord labels, and can be encoded as an additional input to pretrained music systems.

Method

Overview

Our method has two parts. First, the *Dissonance Spectrum* (DS) assigns each active CQT bin its amplitude-weighted relation to the other active components. A continuous kernel combines tolerance-based rational approximation (Stolzenburg 2015) with Tenney's harmonic distance (Tenney 1984); sampling it on the CQT grid permits efficient one-dimensional frequency correlation instead of a $K \times K \times T$ pair tensor. Second, a lightweight encoder and shape-preserving gated residual adapter inject cached DS maps into a host representation without changing the host outputs, heads, or losses. Zero initialization preserves the pretrained baseline function at the start of fine-tuning. We next define intrinsic and cross-reference DS, derive the correlation form, and describe both host integrations; full index derivations and optional temporal variants are in the supplement.

Pitch-Relation Kernel

Rational candidates and complexity. For a maximum numerator and denominator Q, the reduced rational candidate set is

$$\mathcal{R}_Q = \left\{ \frac{p}{q} \;\middle|\; 1 \le p \le q \le Q,\ p, q \in \mathbb{Z}^+,\ \gcd(p, q) = 1 \right\}. \tag{1}$$

For a reduced ratio, we use the complexity function

$$C\left(\frac{p}{q}\right) = \log_2(pq), \tag{2}$$

which is the logarithmic product form of Tenney's harmonic distance (Tenney 1984). Here Q is a finite search-resolution hyperparameter rather than a direct estimate of a listener attribute. For visualization only, simplicity is the reversed complexity scale,

$$S\left(\frac{p}{q}\right) = \max_{r \in \mathcal{R}_Q} C(r) - C\left(\frac{p}{q}\right). \tag{3}$$

Given two frequencies, we order their ratio as

$$r = \frac{\min(f_1, f_2)}{\max(f_1, f_2)}, \tag{4}$$

and define the relative-error neighborhood

$$\mathcal{N}_\alpha(r) = \left\{ x \in \mathbb{R}^+ \;\middle|\; |x - r| \le \alpha r \right\}. \tag{5}$$

We use $\alpha = .01$, following the 1% relative tolerance used in Stolzenburg's smoothed periodicity formulation (Stolzenburg 2015). Let

$$\frac{p^*}{q^*} = \arg\min_{\frac{p}{q} \in \mathcal{R}_Q} \left| r - \frac{p}{q} \right|. \tag{6}$$

The pairwise value is then

$$D(f_1, f_2) = \begin{cases} \min\limits_{\frac{p}{q} \in \mathcal{R}_Q \cap \mathcal{N}_\alpha(r)} C\left(\frac{p}{q}\right), & \text{if } \mathcal{R}_Q \cap \mathcal{N}_\alpha(r) \neq \emptyset, \\ C\left(\frac{p^*}{q^*}\right), & \text{otherwise.} \end{cases} \tag{7}$$

Thus ratios admitting a simpler approximation inside the tolerance receive a lower value; otherwise the closest candidate is used. This is a periodicity/harmonic-distance cue, not a critical-band roughness model.

Continuous pitch intervals and octave folding. We map pitch to frequency by

$$\text{freq}(\text{pitch}) = 440 \cdot 2^{\frac{\text{pitch}-69}{12}} \tag{8}$$

and evaluate

$$D_{\text{interval}}(\Delta p) = D\big(\text{freq}(\text{pitch}_{\text{ref}} + \Delta p),\, \text{freq}(\text{pitch}_{\text{ref}})\big). \tag{9}$$

With

$$D_{\text{int,max}} = \max_{\Delta p \in [0,12)} D_{\text{interval}}(\Delta p), \tag{10}$$

for $\Delta p \in [0, 12)$, the normalized function is

$$D_{\text{norm}}(\Delta p) = \frac{D_{\text{interval}}(\Delta p)}{D_{\text{int,max}}} \in [0, 1]. \tag{11}$$

Let $\delta_{12}(\Delta p) = |\Delta p| \bmod 12 \in [0, 12)$ denote the octave-folded interval magnitude. The resulting even function is

$$\bar{D}(\Delta p) = D_{\text{norm}}(\delta_{12}(\Delta p)), \qquad \bar{D}(0) = 0. \tag{12}$$

The continuous pitch argument allows evaluation between equal-tempered bins, while octave folding implements a pitch-chroma assumption. Because pitch height and pitch chroma can both affect perception (Wagner et al. 2022), folding is an explicit modeling choice rather than a claim that register never matters. The resulting curve is shown in the supplement.

Intrinsic Dissonance Spectrum

CQT representation and preprocessing. Let $\mathbf{X} \in \mathbb{R}^{K \times T}_{\ge 0}$ be a magnitude CQT and $\vec{x}^{\,t} = \mathbf{X}(:, t)$. With B bins per octave, its pitch and frequency grids are

$$\vec{p} = \left[\text{pitch}_{\min} + \frac{12k}{B} \right]_{k=0}^{K-1}, \qquad \vec{f} = [\text{freq}(\text{pitch}_k)]_{k=0}^{K-1}. \tag{13}$$

Values below a fixed floor are set to zero, and a local spectral-peak mask may suppress nonpeak bins. We use excerpt-level global-maximum normalization in the downstream experiments and the context-dependent normalization specified for the controlled validation. DS and its magnitude-CQT control always share the same normalization within a comparison.

Amplitude-weighted attribution. Motivated by the pairwise aggregation of spectral partials in dissonance-curve models (Sethares 1994, 2005), we use the simple amplitude-product coefficient and pair value

$$A(x_1, x_2) = x_1 x_2, \qquad \bar{D}(\Delta p;\, x_1, x_2) = x_1 x_2 \bar{D}(\Delta p). \tag{14}$$

We define the intrinsic DS element at bin k and frame t as

$$d^t_k = \frac{1}{K} \sum_{l=0}^{K-1} x^t_k \, x^t_l \, \bar{D}(p_l - p_k), \tag{15}$$

and collect the frame vectors as

$$\mathbf{D} = \begin{bmatrix} \vec{d}^{\,0} & \cdots & \vec{d}^{\,T-1} \end{bmatrix} \in \mathbb{R}^{K \times T}. \tag{16}$$

The factor $1/K$ averages the K reference-bin contributions. Equation 15 differs from a frame-level scalar: every pair contributes to both participating locations through their own target factors, so the map preserves where the modeled relations occur. All terms are nonnegative; inactive bins have zero attribution.

Cross-Reference Dissonance Spectrum

The same operator can attribute the relation between a target spectrum $\vec{x}$ and a separate reference spectrum $\vec{x}^{(\text{ref})}$. In vector form,

$$\vec{d}^{\,\vec{x} \to \vec{x}^{(\text{ref})}} = \frac{1}{K} \left(\vec{\mathcal{D}}^{\pm} \star_{\text{valid},f} \vec{x}^{(\text{ref})} \right) \odot \vec{x}. \tag{17}$$

Here $\star_{\text{valid},f}$ denotes the target-aligned valid frequency-axis correlation defined below. Intrinsic DS is the special case $\vec{x}^{(\text{ref})} = \vec{x}$. A fixed tonic, chord, or spectral template can instead provide an explicit harmonic reference while the final factor $\vec{x}$ preserves attribution to the target bins. Figure 2 contrasts the two constructions.

Efficient Correlation Form

Because $\bar{D}$ depends only on pitch difference, the K^2 pair table need not be stored. Define the length-$(2K-1)$ sampled kernel

$$\mathcal{D}^{\pm}_m = \bar{D}\left(\frac{12m}{B} \right), \qquad m = -(K-1), \ldots, K-1, \qquad \vec{\mathcal{D}}^{\pm} \in \mathbb{R}^{2K-1}. \tag{18}$$

stored in increasing m order, with center $\bar{D}(0) = 0$. The target-aligned index convention is given in the supplement. Since

$$\bar{D}(p_l - p_k) = \vec{\mathcal{D}}^{\pm}{}_{K+l-k-1}, \tag{19}$$

Equation 15 becomes

$$\vec{d}^{\,t} = \frac{1}{K} \left(\vec{\mathcal{D}}^{\pm} \star_{\text{valid},f} \vec{x}^{\,t} \right) \odot \vec{x}^{\,t}. \tag{20}$$

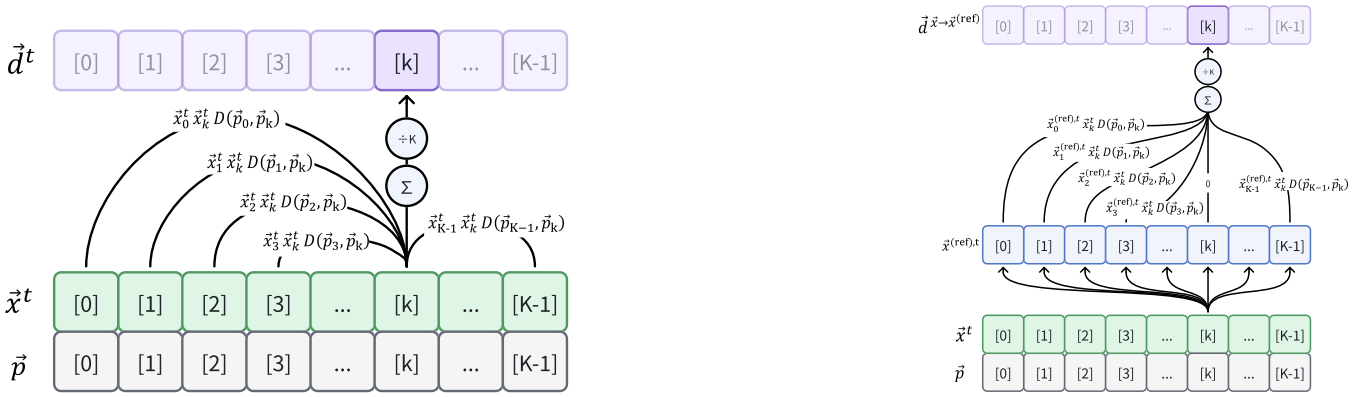

Figure 2: Conceptual DS attribution. Left: intrinsic DS aggregates pairwise relations within one spectrum. Right: cross-reference DS evaluates the target spectrum against a separate reference while retaining target-bin attribution.

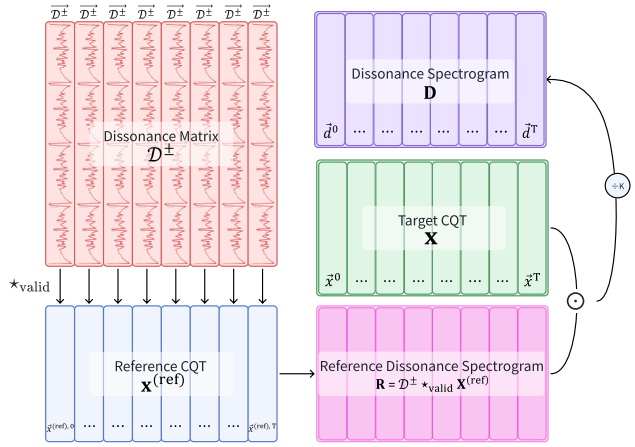

Figure 3: Frequency-axis matrix implementation of cross-reference DS. Intrinsic DS follows by setting the reference and target CQT matrices equal.

For all frames, repeat the same kernel across time as

$$\mathcal{D}^{\pm} = \begin{bmatrix} \vec{\mathcal{D}}^{\pm} & \vec{\mathcal{D}}^{\pm} & \cdots & \vec{\mathcal{D}}^{\pm} \end{bmatrix} \in \mathbb{R}^{(2K-1)\times T}, \tag{21}$$

and compute

$$\mathbf{D} = \frac{1}{K} \left(\mathcal{D}^{\pm} \star_{\mathrm{valid},f} \mathbf{X} \right) \odot \mathbf{X}. \tag{22}$$

The correlation is one-dimensional along frequency and returns $K \times T$ values. It avoids the $O(K^2T)$ intermediate storage of direct pairwise attribution and can be batched with standard correlation operators. Figure 3 shows the corresponding matrix computation; the full index derivation and optional temporal variants are included in the supplement.

Representation Properties and Extraction

Local attribution rather than a scalar score. A conventional spectrum reports energy at (k, t); DS reports how strongly that component participates in the modeled relations of its frame. Unlike a scalar score that collapses register and can only modulate a representation globally, the $K \times T$ map preserves register and instrumentation cues. A downstream encoder can therefore distinguish a narrow high-valued interaction among upper partials from a broad interaction spanning the bass and midrange. Each high value coincides with nonzero target magnitude and can be traced to the reference components selected by the shifted kernel.

Nonnegativity, symmetry, and translation structure. Nonnegative magnitudes and kernel values make DS nonnegative. The ordered ratio and octave-folded difference make the pair relation symmetric, while the final target factor restores bin-specific attribution. Because coefficients depend on pitch difference rather than absolute bin index, equal intervals share coefficients across register. This translation structure permits cross-correlation and differs from an unconstrained learned two-dimensional filter: before training, the same specified relation is treated consistently throughout the CQT range, subject to octave folding.

Continuous pitch grid. The observed frequency ratio is not quantized to the rational candidate set. Equation 7 selects a candidate within tolerance or the nearest candidate, and Equation 12 is sampled at $12/B$-semitone CQT increments. DS therefore responds to detuning, expressive variation, and inharmonic partials rather than only twelve integer pitch classes. The supplement gives qualitative microtonal and timbral examples without assigning a perceptual ranking.

Offline extraction procedure. For each excerpt, we compute magnitude CQT with the shared floor, peak mask, and normalization; sample $\vec{\mathcal{D}}^{\pm}$ from Q, α, and the CQT grid; and evaluate Equation 22. Finite outputs are transformed by $\log(1 + \mathbf{D})$ and cached with the audio hash, sample rate, hop, pitch range, B, Q, α, and preprocessing flags. Training retrieves the DS segment at the baseline branch's temporal indices, with deterministic validation and test crops. This avoids repeated spectral computation and ensures that the baseline and DS branches observe the same recording portion.

Plug-and-Play Augmentation of Music Understanding Models

DS supplements rather than replaces the host waveform encoder or task representation. Let $\mathbf{H}^{(l)} \in \mathbb{R}^{N\times d}$ be the hidden sequence, or a pooled token when $N = 1$, at insertion layer l. A lightweight encoder maps the cached feature to temporal tokens,

$$\mathbf{Z} = E_\phi(\log(1 + \mathbf{D})) \in \mathbb{R}^{M\times d_s}, \tag{23}$$

using convolutional frequency compression followed by temporal convolution or attention. DS tokens share the baseline branch's excerpt indices and valid-frame mask.

A shape-preserving adapter lets the module attach to host representations of different dimensions. The baseline tokens query the DS tokens and receive a gated residual update:

$$\mathbf{C} = \mathrm{softmax}\left(\frac{(\mathbf{H}^{(l)}\mathbf{W}_Q)(\mathbf{Z}\mathbf{W}_K)^\top}{\sqrt{d_a}}\right)(\mathbf{Z}\mathbf{W}_V), \tag{24}$$

$$\mathbf{G} = \sigma\Big(\mathbf{W}_G[\mathbf{H}^{(l)}; \mathbf{C}] + \mathbf{b}_G\Big), \tag{25}$$

$$\widetilde{\mathbf{H}}^{(l)} = \mathbf{H}^{(l)} + \mathbf{G} \odot (\mathbf{C}\mathbf{W}_O). \tag{26}$$

When $N = 1$, the same equations give a pooled adapter. The unchanged output shape preserves the decoder, heads, losses, and evaluation. We initialize $\mathbf{W}_O$ to zero and $\mathbf{b}_G$ negatively, so $\widetilde{\mathbf{H}}^{(l)} = \mathbf{H}^{(l)}$ initially. Disabling the branch recovers the baseline checkpoint without conversion. Independent extraction, shape-preserving insertion, and exact fallback define the plug-and-play property.

For MU-LLaMA, the 576-bin map is pooled to 144 channels and processed by three convolutional blocks, a 128-dimensional temporal encoder, depthwise temporal convolution, and single-head attention. The pooled adapter follows the original audio projection while MERT and LLaMA remain frozen. For Music2Emo, temporal DS tokens are queried after the unchanged 512-dimensional projection of MERT, chord, and key-mode features, before the original emotion heads. Figure 4 shows both insertion paths and frozen versus trainable modules.

To control for adding capacity or a second pitch-resolved pathway, the CQT and Gaussian conditions reuse the same encoder, fusion point, output dimension, and trainable-parameter budget. They replace $\mathbf{D}$ with the aligned normalized magnitude CQT or a fixed random input, respectively.

Interpretive scope. DS operationalizes one periodicity/harmonic-distance-inspired relation and its amplitude-weighted localization. It does not explicitly model auditory-filter bandwidths, masking, learned tonal syntax, or cultural preference. The task experiments therefore test whether this representation is useful as an inductive bias, not whether it is a complete perceptual theory of consonance.

Experiments

Experimental Setup

Implementation, compute, and reproducibility. Controlled calculations are deterministic and parameter-free. PyTorch downstream systems run on a Slurm-managed Linux cluster with one NVIDIA A100 per job and no multi-GPU parallelism. CQT and DS features are extracted once per exact excerpt, cached with metadata, and reused across conditions; the submitted environment lock records software and pretrained-model revisions.

We use paired seeds $\{17, 42, 101, 2025, 2026, 3407\}$. Within each dataset and seed, all four conditions share splits, excerpts, masks, minibatch order, optimization, early stopping, checkpoint selection, and evaluation; random generators use the listed seed. Except for PMEmo's released chorus clips, inputs are 24-kHz, 45-second excerpts with zero padding or one deterministic crop shared by all branches. CQT uses a 1,024-sample hop, eight octaves, 72 bins per octave, $f_{\min} = \mathrm{C1}$, and 576 bins; DS applies the proposed transform and $\log(1 + \mathbf{D})$. Primary endpoints are BERTScore-R for MusicQA and $\overline{R^2}_{\mathrm{VA}}$, the mean of six valence/arousal R^2 values, for Music2Emo. Tests use unrounded seed-level values.

Table 1: Controlled validation. Parentheses give two-sided p-values; the four-class row is an ordering check.

Test	n	Spearman ρ	Kendall τ_b
Intervals	13	.951 (6.36×10^{-7})	.846 (5.20×10^{-6})
Chord quality (4 exemplars)	4	1.000 (–)	1.000 (–)
Chord quality (13 voicings)	13	.626 (.022)	.462 (.030)
Functional connections	7	.794 (.033)	.655 (.054)
Church modes	7	.893 (.0068)	.810 (.0107)

Controlled Music-Theory Validation

We first test whether the relation operator recovers controlled musical structure from rendered piano audio: 13 dyads from unison through the octave, four representative and 13 extended chord voicings, seven diatonic connections to C major, and seven church modes. CQT uses 22.05 kHz audio, four frames per second, eight octaves, 72 bins per octave, $f_{\min} = \mathrm{C1}$, and $K = 576$; the rational search uses $Q = 60$ and $\alpha = .01$. Intervals and scales use a C4 reference, chord qualities equally average tonic-reference and intrinsic DS, and connections use C major. These settings test the same operator under phenomenon-specific reference contexts; downstream models use intrinsic DS.

Interval, chord, and connection maps are reduced to $D_{\max} = \max_t \sum_k \mathbf{D}(k, t)$; sequential scales use $D_{\mathrm{sum}} = \sum_t \sum_k \mathbf{D}(k, t)$. Spearman ρ and Kendall τ_b compare these summaries with predefined orders. The interval order broadly agrees with classic dyad studies (Malmberg 1918; Schwartz, Howe, and Purves 2003); chord, function, and mode orders are theory-derived hypotheses, so they test internal music-theoretical consistency rather than population-level perception.

Intervals show the strongest agreement: C–C♯ is maximal, the tritone is high, and the perfect fifth and octave are among the lowest (Figure 5). Audio DS maxima track both the sampled curve and predefined ranks. Four chord exemplars follow major $<$ minor $<$ suspended $<$ diminished; across 13 voicings the positive but nonmonotonic associa-

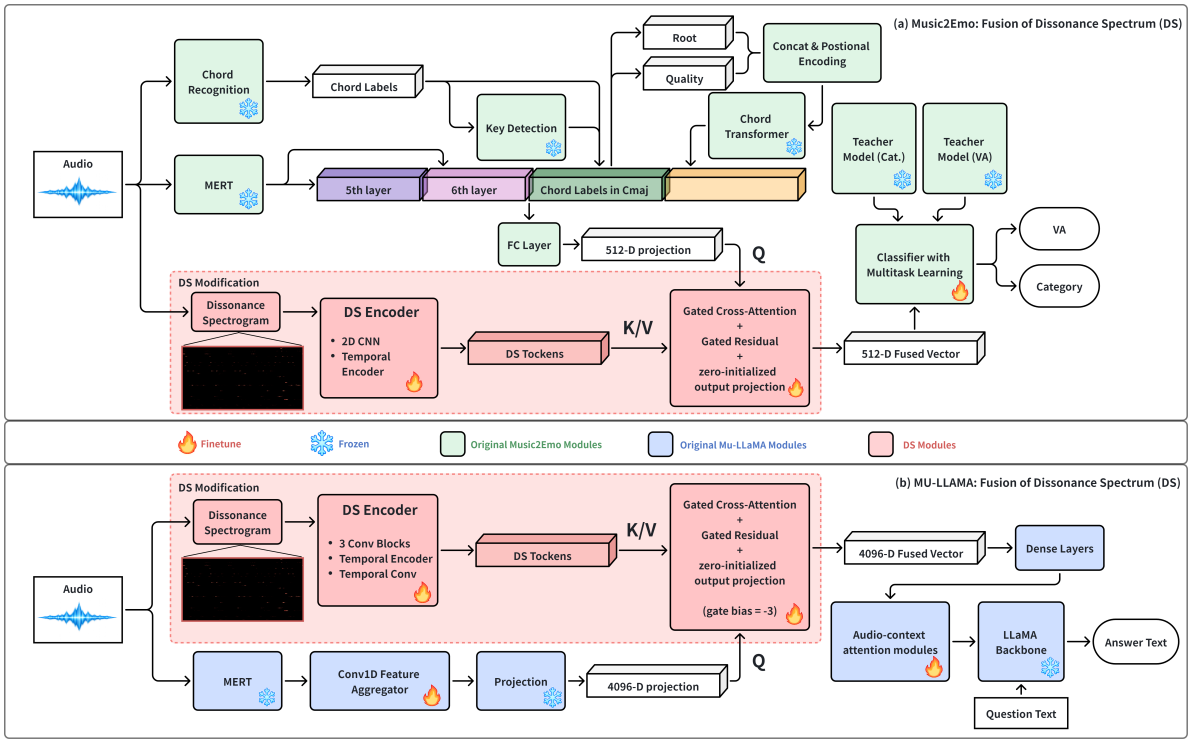

Figure 4: Plug-and-play DS integration in Music2Emo (top) and MU-LLaMA (bottom). The DS branch provides key/value tokens to a gated residual adapter while the original projection provides the query; snowflakes and flames mark frozen and fine-tuned modules.

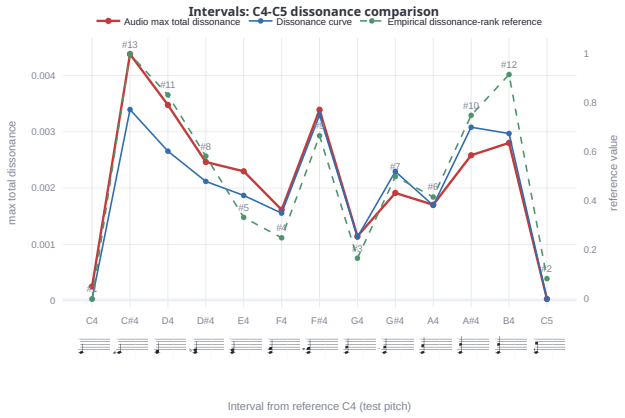

Figure 5: Interval validation: audio DS maxima align with the sampled dissonance curve and the predefined rank reference.

tion reflects spacing and inversion rather than chord labels alone. Tonic-function connections are generally below predominant and dominant connections, while Ionian is near the low end and Locrian highest among modes, with local reversals. Item values, spectral maps, loudness sensitivity, timbre, and microtonal analyses are in the supplement.

Table 2: MusicQA results over six paired training seeds (mean±SD). Δ is DS minus Baseline.

Metric	Baseline	Gaussian	CQT	DS	Δ
BLEU ↑	.2987±.0019	.2985±.0019	.3056±.0015	**.3074±.0014**	+.0087
METEOR ↑	.3761±.0017	.3759±.0018	.3838±.0014	**.3857±.0013**	+.0096
ROUGE-L ↑	.4556±.0018	.4554±.0019	.4643±.0016	**.4671±.0015**	+.0115
BERTScore-R ↑	.8952±.0007	.8950±.0006	.8996±.0006	**.9024±.0015**	+.0072
Test loss ↓	.625±.004	.626±.004	.607±.003	**.600±.002**	−.025
Perplexity ↓	1.868±.008	1.870±.008	1.836±.005	**1.822±.004**	−.046

Music Question Answering

Protocol. MU-LLaMA is fine-tuned on 70,011 question–answer pairs from 7,779 tracks and evaluated on 5,040 pairs from 560 audio-disjoint MTG-Jamendo tracks. All conditions use frozen LLaMA-2 7B and MERT-v1-330M with the released initialization (Touvron et al. 2023; Li et al. 2024; Liu et al. 2024). The Baseline has 4.21M trainable parameters; each matched branch has 5.57M. BF16 AdamW uses gradient accumulation to an effective batch of 32, at most four epochs, and validation-loss early stopping; the best checkpoint is evaluated once on the held-out set. One fixed evaluator reports BLEU, METEOR, ROUGE-L, BERTScore-R, loss, and perplexity (Papineni et al. 2002; Banerjee and Lavie 2005; Lin 2004; Zhang et al. 2020); exact tokenization, generation, and package settings are in the supplement.

DS has the highest six-seed mean on every MusicQA met-

Table 3: Music2Emo results over six paired seeds. J-PR/J-ROC are macro tag averages; V/A denote valence/arousal R^2. Only $\overline{R^2}_{\mathrm{VA}}$ reports mean±SD.

Metric	Baseline	Gaussian	CQT	DS
J-PR	.1539	.1537	.1564	**.1580**
J-ROC	.7806	.7801	.7828	**.7841**
DEAM-V	.5169	.5164	.5272	**.5355**
DEAM-A	.6209	.6202	.6260	**.6291**
Emo-V	.6487	.6479	.6575	**.6642**
Emo-A	.7598	.7590	.7642	**.7668**
PM-V	.5451	.5445	.5532	**.5587**
PM-A	.7926	.7920	.7970	**.7992**
$\overline{R^2}_{\mathrm{VA}}$	.6473±.0014	.6467±.0014	.6542±.0016	**.6589±.0018**

ric. BERTScore-R increases by .0072 over Baseline, .0074 over Gaussian, and .0028 over CQT, with positive paired differences for all six seeds. The largest Holm-adjusted paired-t p-value is .0017, but the exact two-sided sign test gives .03125 per comparison and .09375 after Holm correction across the three DS-versus-control comparisons. We therefore emphasize repeated-seed direction and effect size, not familywise distribution-free significance. Text metrics measure reference similarity rather than factual musical understanding.

Music Emotion Recognition

Model and data. Music2Emo projects MERT layers five and six, chord features, and key mode to 512 dimensions before the DS adapter (Kang and Herremans 2025). The Baseline task network has 1.07M trainable parameters; each matched branch adds 0.21M while the 95M-parameter MERT remains frozen. We retain the official MTG-Jamendo split (Bogdanov et al. 2019) and fixed 70/15/15 track splits for DEAM, EmoMusic, and PMEmo (Aljanaki, Yang, and Soleymani 2017; Soleymani et al. 2013; Zhang et al. 2018; Kang and Herremans 2025). Weighted binary cross-entropy covers 56 tags, and mean squared error covers valence and arousal. All conditions share the knowledge-distillation objective, Adam at 10^{-4}, early stopping, and checkpoint rule; details are in the supplement.

DS has the highest mean on every Music2Emo endpoint. $\overline{R^2}_{\mathrm{VA}}$ increases by .0116 over Baseline, .0122 over Gaussian, and .0047 over CQT, with positive differences for all six seeds. The Holm-adjusted exact sign-test value is .09375, so we again emphasize direction and effect size. Average valence R^2 rises from .5702 to .5861, and arousal R^2 from .7244 to .7317. CQT is second; the further DS gain is consistent with useful pairwise organization, although compression and normalization differences mean the comparison does not isolate the relation transform alone.

Conclusion

This work addresses a gap between energy-based spectra and latent learned representations by making one class of simultaneous frequency relations explicit. As a spectrogram reorganizes a waveform without adding a new observation, DS reorganizes magnitude information so that modeled pair relations become directly visible to both researchers and downstream encoders. It applies a continuous, tolerance-based harmonic-distance kernel to magnitude CQT, localizes amplitude-weighted pair relations in time and frequency, and avoids a quadratic pair tensor through frequency-axis correlation. A shape-preserving, zero-initialized adapter then adds this representation to existing systems without changing their output interfaces or baseline function at initialization. Controlled tests recovered strong ordinal agreement for intervals, functional connections, and church modes, with moderate positive agreement across diverse chord voicings. Across six paired seeds, DS also achieved the highest mean on every reported MusicQA and Music2Emo endpoint relative to the baseline, a parameter-matched Gaussian branch, and an architecture-matched magnitude-CQT branch. The Gaussian control indicates that capacity alone is insufficient, while the smaller advantage over magnitude CQT suggests value beyond an added pitch-resolved pathway, within the stated compression and normalization caveat. Together, these results support explicit relational structure as a useful and inspectable complement to learned music representations.

Beyond aggregate scores, the retained time–frequency layout gives DS a diagnostic role that scalar dissonance summaries cannot provide. Researchers can inspect when and where a modeled relation is concentrated, while the cross-reference construction can condition that attribution on an explicit tonic, chord, or spectral template. Because extraction is deterministic and cached, and the adapter preserves host shapes and offers exact fallback, the representation can be evaluated alongside existing systems without replacing their audio encoders. These properties make DS a concrete basis for theory-conditioned probing and model comparison, although their practical value outside the evaluated tasks remains to be tested.

The evidence does not establish DS as a complete model of consonance or musical preference. The kernel encodes one periodicity/harmonic-distance prior with deliberate octave folding and omits auditory-filter bandwidths, masking, tonal syntax, harsh high-frequency content, rhythmic tension, and culturally or individually learned preference. The controlled rankings are partly theory-derived, no new listening study was conducted, and automatic MusicQA references measure similarity rather than factual or expert-level harmonic reasoning. Evaluation is also limited to two host families and does not fully isolate the relation transform from branch-input compression and normalization. Future work should package the extraction and visualization pipeline as a reusable library, calibrate the representation with listener data and individual perceptual variation, and test its transfer to broader audio and symbolic tasks such as generation, style analysis, and standardized symbolic dissonance encoding.

References

Aljanaki, A.; and Soleymani, M. 2018. A Data-Driven Approach to Mid-Level Perceptual Musical Feature Modeling. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*, 615–621.

Aljanaki, A.; Yang, Y.-H.; and Soleymani, M. 2017. Developing a Benchmark for Emotional Analysis of Music. *PLOS ONE*, 12(3): e0173392.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72.

Bogdanov, D.; Won, M.; Tovstogan, P.; Porter, A.; and Serra, X. 2019. The MTG-Jamendo Dataset for Automatic Music Tagging. In *Machine Learning for Music Discovery Workshop at the 36th International Conference on Machine Learning*.

Brown, J. C. 1991. Calculation of a Constant-Q Spectral Transform. *The Journal of the Acoustical Society of America*, 89(1): 425–434.

Cazden, N. 1980. The Definition of Consonance and Dissonance. *International Review of the Aesthetics and Sociology of Music*, 11(2): 123–168.

Chowdhury, S.; Vall, A.; Haunschmid, V.; and Widmer, G. 2019. Towards Explainable Music Emotion Recognition: The Route via Mid-Level Features. In *Proceedings of the 20th International Society for Music Information Retrieval Conference*, 237–243.

Cousineau, M.; McDermott, J. H.; and Peretz, I. 2012. The Basis of Musical Consonance as Revealed by Congenital Amusia. *Proceedings of the National Academy of Sciences*, 109(48): 19858–19863.

Deng, Z.; Ma, Y.; Liu, Y.; Guo, R.; Zhang, G.; Chen, W.; Huang, W.; and Benetos, E. 2024. MusiLingo: Bridging Music and Text with Pre-trained Language Models for Music Captioning and Query Response. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 3643–3655.

Eerola, T.; and Lahdelma, I. 2021. The Anatomy of Consonance/Dissonance: Evaluating Acoustic and Cultural Predictors Across Multiple Datasets with Chords. *Music & Science*, 4: 1–19.

Elizalde, B.; Deshmukh, S.; Al Ismail, M.; and Wang, H. 2023. CLAP: Learning Audio Concepts from Natural Language Supervision. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5.

Harrison, P. M. C.; and Pearce, M. T. 2020. Simultaneous Consonance in Music Perception and Composition. *Psychological Review*, 127(2): 216–244.

Harte, C.; Sandler, M.; and Gasser, M. 2006. Detecting Harmonic Change in Musical Audio. In *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*, 21–26.

Huang, Q.; Jansen, A.; Lee, J.; Ganti, R.; Li, J. Y.; and Ellis, D. P. W. 2022. MuLan: A Joint Embedding of Music Audio and Natural Language. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference*, 559–566.

Hutchinson, W.; and Knopoff, L. 1978. The Acoustic Component of Western Consonance. *Interface*, 7(1): 1–29.

Kaklamani, S.; and Simserides, C. 2026. Psychoacoustic Study of Simple-Tone Dyads: Frequency Ratio and Pitch. *Acoustics*, 8(1): 14.

Kang, J.; and Herremans, D. 2025. Towards Unified Music Emotion Recognition across Dimensional and Categorical Models. arXiv:2502.03979.

Korzeniowski, F.; and Widmer, G. 2016. Feature Learning for Chord Recognition: The Deep Chroma Extractor. In *Proceedings of the 17th International Society for Music Information Retrieval Conference*, 37–43.

Langner, G. 1992. Periodicity Coding in the Auditory System. *Hearing Research*, 60(2): 115–142.

Li, Y.; Yuan, R.; Zhang, G.; Ma, Y.; Chen, X.; Yin, H.; Xiao, C.; Lin, C.; Ragni, A.; Benetos, E.; Gyenge, N.; Dannenberg, R. B.; Liu, R.; Chen, W.; Xia, G.; Shi, Y.; Huang, W.; Wang, Z.; Guo, Y.; and Fu, J. 2024. MERT: Acoustic Music Understanding Model with Large-Scale Self-Supervised Training. In *International Conference on Learning Representations*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81.

Liu, S.; Hussain, A. S.; Sun, C.; and Shan, Y. 2024. Music Understanding LLaMA: Advancing Text-to-Music Generation with Question Answering and Captioning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 286–290.

Lu, W.-T.; Wang, J.-C.; Won, M.; Choi, K.; and Song, X. 2021. SpecTNT: A Time–Frequency Transformer for Music Audio. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference*, 396–403.

Malmberg, C. F. 1918. The Perception of Consonance and Dissonance. *Psychological Monographs*, 25(2): 93–133.

Marjieh, R.; Harrison, P. M. C.; Lee, H.; Deligiannaki, F.; and Jacoby, N. 2024. Timbral Effects on Consonance Disentangle Psychoacoustic Mechanisms and Suggest Perceptual Origins for Musical Scales. *Nature Communications*, 15: 1482.

McDermott, J. H.; Schultz, A. F.; Undurraga, E. A.; and Godoy, R. A. 2016. Indifference to Dissonance in Native Amazonians Reveals Cultural Variation in Music Perception. *Nature*, 535(7613): 547–550.

Müller, M.; and Ewert, S. 2011. Chroma Toolbox: Matlab Implementations for Extracting Variants of Chroma-Based Audio Features. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, 215–220.

Panda, R.; Malheiro, R.; and Paiva, R. P. 2023. Audio Features for Music Emotion Recognition: A Survey. *IEEE Transactions on Affective Computing*, 14(1): 68–88.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318.

Parncutt, R. 1989. *Harmony: A Psychoacoustical Approach*. Berlin: Springer-Verlag.

Plomp, R.; and Levelt, W. J. M. 1965. Tonal Consonance and Critical Bandwidth. *The Journal of the Acoustical Society of America*, 38(4): 548–560.

Poltronieri, A.; Serra, X.; and Rocamora, M. 2025. From Discord to Harmony: Decomposed Consonance-Based Training for Improved Audio Chord Estimation. In *Proceedings of the 26th International Society for Music Information Retrieval Conference*, 492–502. Daejeon, South Korea.

Schwär, S.; Balke, S.; and Müller, M. 2025. Measuring Sensory Dissonance in Multi-Track Music Recordings: A Case Study with Wind Quartets. In *Proceedings of the 26th International Society for Music Information Retrieval Conference*, 117–126. Daejeon, South Korea.

Schwartz, D. A.; Howe, C. Q.; and Purves, D. 2003. The Statistical Structure of Human Speech Sounds Predicts Musical Universals. *The Journal of Neuroscience*, 23(18): 7160–7168.

Sethares, W. A. 1994. Adaptive Tunings for Musical Scales. *The Journal of the Acoustical Society of America*, 96(1): 10–18.

Sethares, W. A. 2005. *Tuning, Timbre, Spectrum, Scale*. London: Springer, 2 edition.

Soleymani, M.; Caro, M. N.; Schmidt, E. M.; Sha, C.-Y.; and Yang, Y.-H. 2013. 1000 Songs for Emotional Analysis of Music. In *Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, 1–6.

Stolzenburg, F. 2015. Harmony Perception by Periodicity Detection. *Journal of Mathematics and Music*, 9(3): 215–238.

Tenney, J. 1984. John Cage and the Theory of Harmony. In Garland, P., ed., *Soundings 13: The Music of James Tenney*, 55–83. Santa Fe, New Mexico: Soundings Press.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

von Helmholtz, H. L. F. 1954. *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. New York: Dover Publications. Second English edition, translated by Alexander J. Ellis; original German work published in 1863.

Wagner, B.; Czoschke, S.; Tillmann, B.; Koelsch, S.; Vuust, P.; and Brattico, E. 2022. Pitch Chroma Information Is Processed in Addition to Pitch Height Information with More Than Two Pitch-Range Categories. *Attention, Perception, & Psychophysics*, 84(5): 1757–1771.

Wand, A. 2012. On the Conception and Measure of Consonance. *Leonardo Music Journal*, 22: 73–78.

Weck, B.; Puentes, P.; Poltronieri, A.; Prabhu, S.; and Bogdanov, D. 2026. HumMusQA: A Human-written Music Understanding QA Benchmark Dataset. In *Proceedings of the 4th Workshop on NLP for Music and Audio (NLP4MusA 2026)*, 58–67. Rabat, Morocco: Association for Computational Linguistics.

Won, M.; Hung, Y.-N.; and Le, D. 2024. A Foundation Model for Music Informatics. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1226–1230.

Xia, H.; Huang, Z.; Tan, Y.; and Song, S. 2026. Let the Model Learn to Feel: Mode-Guided Tonality Injection for Symbolic Music Emotion Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(3): 2182–2190.

Zhang, K.; Zhang, H.; Li, S.; Yang, C.; and Sun, L. 2018. The PMEmo Dataset for Music Emotion Recognition. In *Proceedings of the 2018 ACM International Conference on Multimedia Retrieval*, 135–142.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Supplementary Material: Beyond Frequency: Dissonance Spectrum for Perceptually Motivated Music Understanding

Tianle Wang[1,2], **Xinyi Tong**[1,2], **Liangke Zhao**[1,2], **Jishang CHEN**[1,2], **Sirui Zhang**[1,2], **Haoxin Zhang**[1,2], **Xin Jin**[1], **Duo XU**[1], **Xiaobing Li**[2], **Song-Chun Zhu**[1,3]

[1]Beijing Institute for General Artificial Intelligence, Beijing, China
[2]Central Conservatory of Music, Beijing, China
[3]Peking University, Beijing, China

Reproducibility Package

The separately submitted Code and Data Archive contains the DS implementation, extraction configuration, cached-feature metadata schema, fixed dataset manifests, audio hashes, training configurations, environment lock files, checkpoint-selection rules, evaluation scripts, seed-level predictions, and scripts that regenerate every reported table. It includes scripts and instructions for obtaining public datasets and pretrained models from their cited official sources. The archive is packaged without author-identifying repository metadata.

Full Derivation and Extensions of Dissonance Spectrum

This section records the complete algebra and optional variants underlying the concise main-paper description and clarifies implementation shapes and experimental conventions.

Preprocessing Conventions

Let $\mathbf{X}(k,t)$ denote the magnitude CQT, $\vec{x}^{\,t} = \mathbf{X}(:,t) \in \mathbb{R}^K$, and $x_k^t = \mathbf{X}(k,t)$. The denoising and peak-selection operators are

$$\mathbf{X}_{\text{denoised}}(k,t) = \mathbf{X}(k,t) \cdot \mathbf{1}\left(\mathbf{X}(k,t) \geq \theta\right), \tag{1}$$

$$\vec{x}^{\,t}_{\text{peak}} = \vec{x}^{\,t} \odot \left[\mathbf{1}\left(x_k^t \text{ is a peak point in } \vec{x}^{\,t}\right)\right]_{k=0}^{K-1}. \tag{2}$$

Two normalization choices are useful in different settings:

$$\begin{aligned} m(t) &= \max_k x_k^t, \quad \vec{x}^{\,t}_{\text{frame}} = \vec{x}^{\,t}/m(t), \\ m &= \max_{k,t} x_k^t, \quad \mathbf{X}_{\text{global}} = \mathbf{X}/m. \end{aligned} \tag{3}$$

The downstream model comparisons use excerpt-level global-maximum normalization m, matching the magnitude-CQT control and avoiding framewise loudness equalization. The controlled music-theory tests instead use the context-window normalization described below, with target and reference spectra normalized separately. Each comparison uses one fixed convention recorded in its configuration.

Direct Pairwise Form

The intrinsic element is

$$d_k^t = \frac{1}{K} \sum_{l=0}^{K-1} x_k^t \, x_l^t \, \bar{D}(p_l - p_k), \tag{4}$$

and the complete representation is

$$\mathbf{D} = \left[\vec{d}^{\,0} \quad \vec{d}^{\,1} \quad \cdots \quad \vec{d}^{\,T-1}\right] \in \mathbb{R}^{K \times T}. \tag{5}$$

The factor $1/K$ averages the reference-bin contributions and removes their direct linear count factor under otherwise matched grids; it is not a guarantee of invariance to CQT resolution. The direct pair-attribution concept is visualized in the main paper.

Index Derivation of the Correlation Form

The CQT pitch grid and the zero-indexed stored relation vector are

$$p_k = \text{pitch}_{\min} + \frac{12k}{B}, \qquad k = 0, \ldots, K-1, \tag{6}$$

$$\left[\vec{\mathcal{D}^{\pm}}\right]_j = \bar{D}\left(\frac{12(j-K+1)}{B}\right), \qquad j = 0, \ldots, 2K-2. \tag{7}$$

Because the kernel depends only on pitch difference,

$$\bar{D}(p_l - p_k) = \vec{\mathcal{D}^{\pm}}_{K+l-k-1}. \tag{8}$$

For a length-$(2K-1)$ vector $\vec{a}$ and a length-K vector $\vec{b}$, the implementation uses the target-aligned convention

$$[\vec{a} \star_{\text{valid},f} \vec{b}]_k = \sum_{l=0}^{K-1} a_{K+l-k-1} b_l, \quad k = 0, \ldots, K-1. \tag{9}$$

With increasing-lag storage, this equals a standard valid cross-correlation followed by a fixed frequency-axis reversal; an equivalent lag-reversed layout avoids the explicit reversal. Consequently,

$$d_k^t = \frac{1}{K} x_k^t \sum_{l=0}^{K-1} x_l^t \, \bar{D}(p_l - p_k) \tag{10}$$

$$= \frac{1}{K} x_k^t \sum_{l=0}^{K-1} x_l^t \, \mathcal{D}^{\pm}_{K+l-k-1}. \tag{11}$$

Sliding the complete relation vector over the CQT gives the reference-relation frame

$$\vec{r}^{\,t} = \vec{\mathcal{D}^{\pm}} \star_{\text{valid},f} \vec{x}^{\,t}, \tag{12}$$

and therefore

$$\vec{d}^{\,t} = \frac{1}{K}\vec{r}^{\,t} \odot \vec{x}^{\,t} = \frac{1}{K}\left(\vec{\mathcal{D}}^{\pm} \star_{\text{valid},f} \vec{x}^{\,t}\right) \odot \vec{x}^{\,t}. \tag{13}$$

Figure 1 visualizes the index alignment. Equation 9 is applied independently at every time frame and returns the K target locations in their original frequency order.

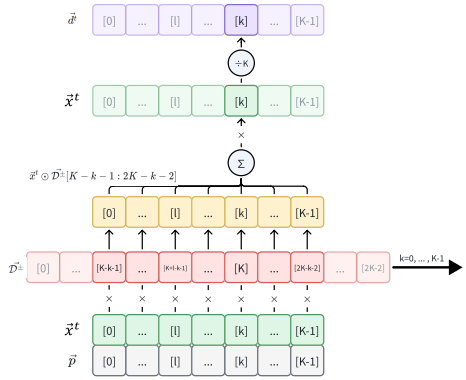

Figure 1: Index alignment for the frequency-axis correlation form of intrinsic DS.

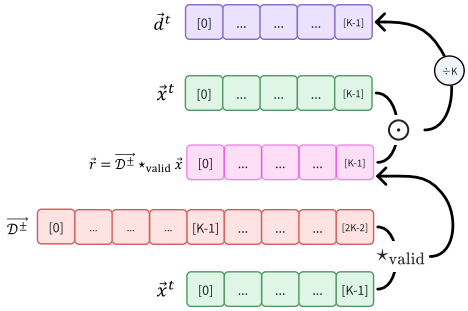

Figure 2: Vectorized intrinsic-DS computation after replacing the explicit pairwise sum by valid correlation.

Cross Dissonance Spectrum

For target $\vec{x}$ and an arbitrary reference spectrum $\vec{x}^{(\text{ref})}$, the cross-spectrum is

$$d_k^{\vec{x}\to\vec{x}^{(\text{ref})}} = \frac{1}{K}\sum_{l=0}^{K-1} x_k x_l^{(\text{ref})}\, \bar{D}(p_l - p_k) \tag{14}$$

$$= \frac{1}{K} x_k \sum_{l=0}^{K-1} x_l^{(\text{ref})}\, \mathcal{D}^{\pm}_{K+l-k-1}. \tag{15}$$

Its reusable reference map and final attribution are

$$\vec{r} = \vec{\mathcal{D}}^{\pm} \star_{\text{valid},f} \vec{x}^{(\text{ref})}, \tag{16}$$

$$\vec{d}^{\vec{x}\to\vec{x}^{(\text{ref})}} = \frac{1}{K}\left(\vec{\mathcal{D}}^{\pm} \star_{\text{valid},f} \vec{x}^{(\text{ref})}\right) \odot \vec{x}. \tag{17}$$

Intrinsic DS is the special case $\vec{x}^{(\text{ref})} = \vec{x}$. A fixed reference may instead encode a tonic template or an instrument tone.

Figure 3 shows the corresponding cross-reference computation.

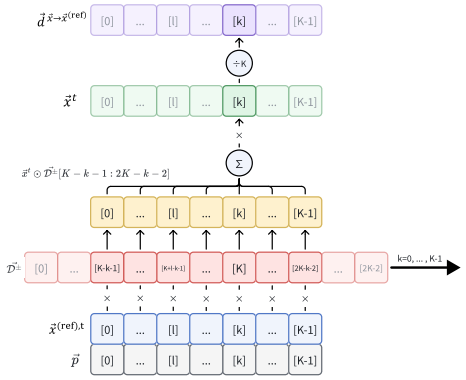

Figure 3: Index alignment for cross-DS correlation using a fixed or time-varying reference spectrum.

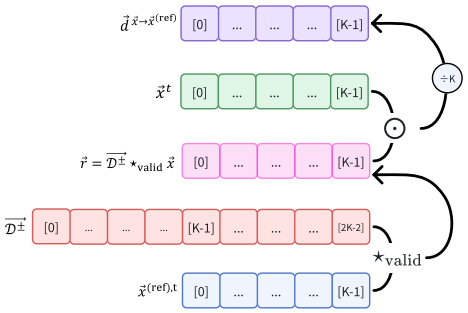

Figure 4: Vectorized cross-DS computation using a reusable reference-relation vector.

Matrix and Temporal Variants

Repeating the same relation vector across time gives

$$\mathcal{D}^{\pm} = \begin{bmatrix}\vec{\mathcal{D}}^{\pm} & \vec{\mathcal{D}}^{\pm} & \cdots & \vec{\mathcal{D}}^{\pm}\end{bmatrix} \in \mathbb{R}^{(2K-1)\times T}. \tag{18}$$

For a general reference CQT,

$$\mathbf{R} = \mathcal{D}^{\pm} \star_{\text{valid},f} \mathbf{X}^{(\text{ref})}, \tag{19}$$

$$\mathbf{D} = \frac{1}{K}\mathbf{R} \odot \mathbf{X} = \frac{1}{K}\left(\mathcal{D}^{\pm} \star_{\text{valid},f} \mathbf{X}^{(\text{ref})}\right) \odot \mathbf{X}. \tag{20}$$

Both $\mathbf{R}$ and $\mathbf{D}$ have shape $K \times T$ under the frequency-axis valid-correlation convention used here. Optional temporal DS averages references from the preceding n frames:

$$\vec{d}^{\,(\text{temporal}),\,t} = \frac{1}{n}\sum_{i=0}^{n-1} \vec{d}^{\left(\vec{x}^{\,t}\to\vec{x}^{\,t-i}\right)} = \vec{d}^{\left(\vec{x}^{\,t}\to\frac{1}{n}\sum_{i=0}^{n-1}\vec{x}^{\,t-i}\right)}. \tag{21}$$

A tonic–intrinsic combination is

$$\vec{d}^{\,(\text{combine}),\,t} = w_{\text{intrinsic}}\vec{d}^{\,t} + w_{\text{tonic}}\vec{d}^{\,(\text{tonic}),\,t}. \tag{22}$$

The downstream model comparisons use intrinsic DS. The controlled music-theory tests use tonic, chord, or tonic–intrinsic references to isolate the relation being examined.

Controlled Music-Theory Validation Details

Stimuli, References, and Aggregation

The validation suite uses MIDI-rendered stimuli with fixed note events and piano timbre unless otherwise stated. Nominal durations describe MIDI events; WAV files include onset and release tails. The controlled tests are not a listener study. They ask whether the relation operator and its audio realization reproduce predefined ordinal structures under transparent reference choices.

Table 1: Controlled-validation configuration.

Item	Setting
Audio / frame rate	22.05 kHz mono / 4 frames s^{-1}
CQT grid	8 octaves, 72 bins $octave^{-1}$, $K = 576$, $f_{\min} = \mathrm{C1}$
Kernel	$Q = 60$, relative tolerance $\alpha = .01$, octave folding
Preprocessing	linear magnitude, relative-frame soft gate at 0.1, context window 4 s
Peak selection	prominence .015; other peak filters disabled
Scaling	target and reference normalized separately; final factor $1/K$
Intervals / scales	C4 tonic reference
Chord qualities	equal-weight C4-tonic and intrinsic DS
Chord connections	C-major chord reference
Aggregation	$D_{\max}$ for intervals/chords/connections; D_{sum} for scales

For a frame, $D(t) = \sum_k \mathbf{D}(k,t)$. We use $D_{\max} = \max_t D(t)$ when a stimulus contains a sustained interval or chord event and $D_{\mathrm{sum}} = \sum_t D(t)$ when an entire sequential scale is the analysis unit. Rank hypotheses are min–max mapped only for visualization; all reported correlations use the original ranks and unscaled DS summaries.

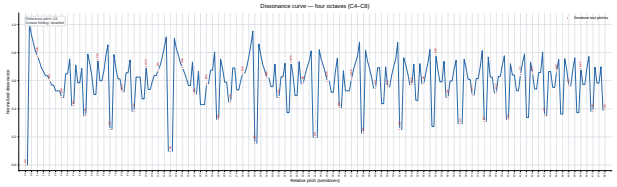

Figure 5: Four-octave view of the normalized relation kernel, illustrating its octave-folded repetition. Integer-semitone samples are marked for reference.

Intervals

Thirteen notes from C4 to C5 are compared against a C4 reference. The historical order used by the validation assets is related to classic dyad-consonance experiments (Malmberg 1918; Schwartz, Howe, and Purves 2003), but it is not presented as a direct transcription of any single published table.

Table 2: Interval results under the C4 tonic reference.

Note	Rank	$D_{\max}$	Note	Rank	$D_{\max}$
C4	1	.000253	G4	3	.001145
C♯4	13	.004379	G♯4	7	.001913
D4	11	.003476	A4	6	.001703
D♯4	8	.002463	A♯4	10	.002583
E4	5	.002298	B4	12	.002802
F4	4	.001614	C5	2	.000030
F♯4	9	.003391			

Spearman $\rho = .95055$ ($p = 6.36 \times 10^{-7}$) and Kendall $\tau_b = .84615$ ($p = 5.20 \times 10^{-6}$). The minor second is maximal, the tritone is high, the perfect fifth is low, and the octave is minimal. The small nonzero unison value reflects slight differences between independently rendered target and reference recordings.

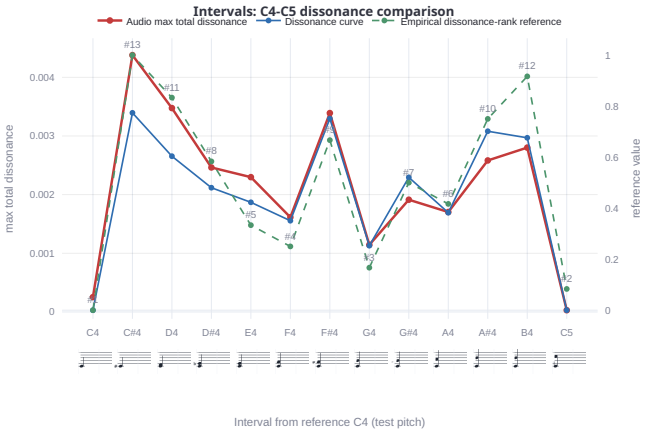

Figure 6: Interval validation details: audio DS maxima, sampled kernel values, and the predefined rank reference.

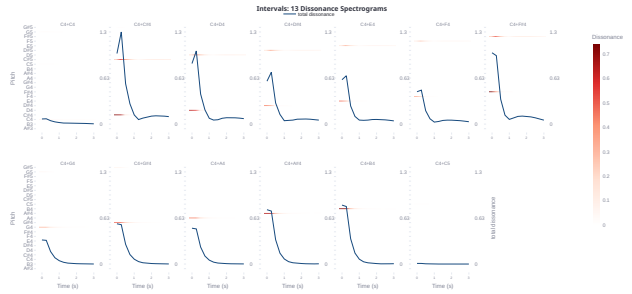

Figure 7: Representative interval DS maps for unison, minor second, perfect fifth, and octave.

Chord Qualities

The four representative classes are perfectly ordered as major $<$ minor $<$ suspended $<$ diminished. Because $n = 4$, this is treated as an ordering check rather than a meaningful significance test. The larger set varies chord quality and voicing and therefore provides a stricter test.

Table 3: Extended chord-quality validation. Semitone arrays are relative to C4.

Chord	Semitones	Rank	$D_{\max}$	Chord	Semitones	Rank	$D_{\max}$
Major-1	[0,4,7]	1	.005691	Sus-1	[0,7,17]	7	.005641
Major-2	[0,3,8]	5	.006941	Sus-2	[0,2,7]	6	.006758
Major-3	[0,9,17]	3	.006299	Sus-3	[0,10,17]	4	.007769
Minor-1	[0,3,7]	2	.006058	Dim-1	[0,3,6]	12	.015950
Minor-2	[0,4,9]	10	.006738	Dim-2	[0,9,15]	9	.008581
Minor-3	[0,8,17]	8	.006363	Dim-3	[0,9,18]	11	.008172
Aug	[0,4,8]	13	.007305				

Across 13 voicings, Spearman $\rho = .62637$ ($p = .02199$) and Kendall $\tau_b = .46154$ ($p = .03048$). The positive association coexists with local reversals, such as Sus-1 below Major-1 and the augmented triad below Dim-1. These deviations are informative rather than errors: DS analyzes realized spectra and therefore responds to spacing and inversion instead of assigning one constant to a chord label.

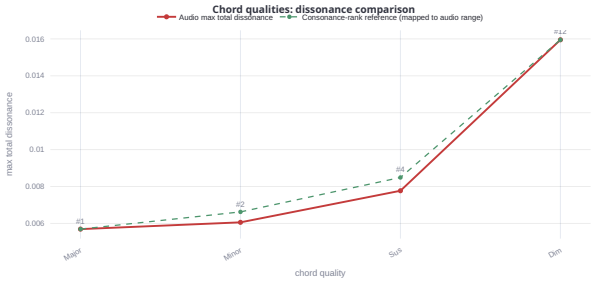

Figure 8: Representative chord-quality ordering.

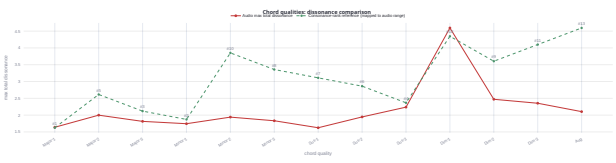

Figure 9: Extended chord-quality and voicing results.

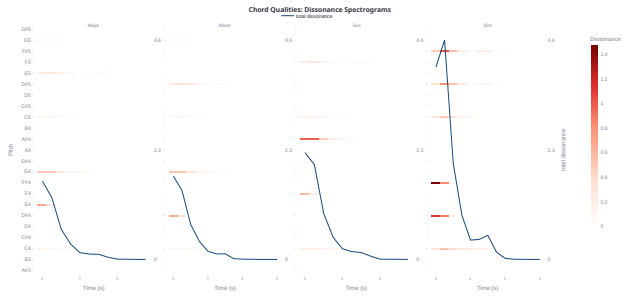

Figure 10: Representative tonic-referenced chord-quality spectra.

Functional Chord Connections

Each stimulus plays one diatonic chord followed by C major. The ordinal code assigns 1 to tonic-function chords (I, iii, vi), 2 to predominant chords (ii, IV), and 3 to dominant-function chords (V7, vii°). This reference-conditioned construction operationalizes a connection as the first chord's relation to C major; it does not model voice leading or learned temporal expectation. The code is a theory-derived trend hypothesis, not a psychophysical scale.

Table 4: Chord connections to C major.

File	First chord	Group	Rank	$D_{\max}$
CC	C–E–G	tonic	1	.003848
DC	D–F–A	predominant	2	.009701
EC	E–G–B	tonic	1	.005311
FC	F–A–C	predominant	2	.008495
G7C	G–B–D–F	dominant	3	.010312
AC	A–C–E	tonic	1	.005695
BC	B–D–F	dominant	3	.006699

Spearman $\rho = .79373$ ($p = .03310$) and Kendall $\tau_b = .65465$ ($p = .05363$). The overall group trend is recovered, although vii° and V7 differ substantially, showing that group membership does not determine the complete spectral value.

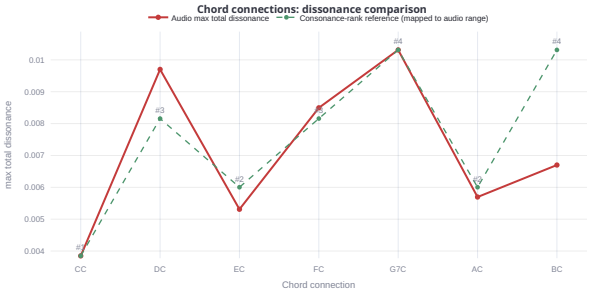

Figure 11: Functional chord-connection results.

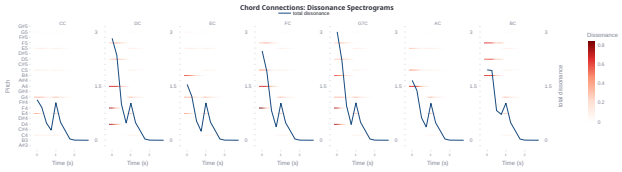

Figure 12: Reference-conditioned spectra for the seven chord connections.

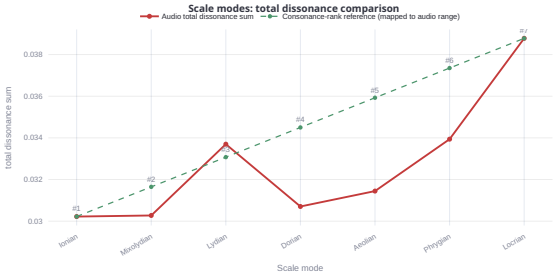

Figure 13: Church-mode total DS and the predefined ordinal reference used in the main-paper correlation.

Scales and Modes

Scale stimuli ascend from C4 with one second per note. For the seven church modes, D_{sum} correlates with the predefined order at Spearman $\rho = .89286$ ($p = .00681$) and Kendall $\tau_b = .80952$ ($p = .0107$).

Table 5: Church-mode results.

Mode	Rank	D_{sum}
Ionian	1	.030220
Mixolydian	2	.030273
Lydian	3	.033697
Dorian	4	.030704
Aeolian	5	.031444
Phrygian	6	.033937
Locrian	7	.038779

The trend is strong but not strictly monotonic because Dorian lies below Lydian. This test aggregates tonic-relative interval content and should not be read as a complete model of modal perception. Figure 14 includes all 33 available scale files. For scales without an externally grounded rank, the values should be interpreted as a quantitative "consonance palette": DS can compare and visualize their realized interval content, but the ordering is not claimed as a universal preference scale. The code defines a natural-minor file that is absent; the equivalent Aeolian pitch-class set is available.

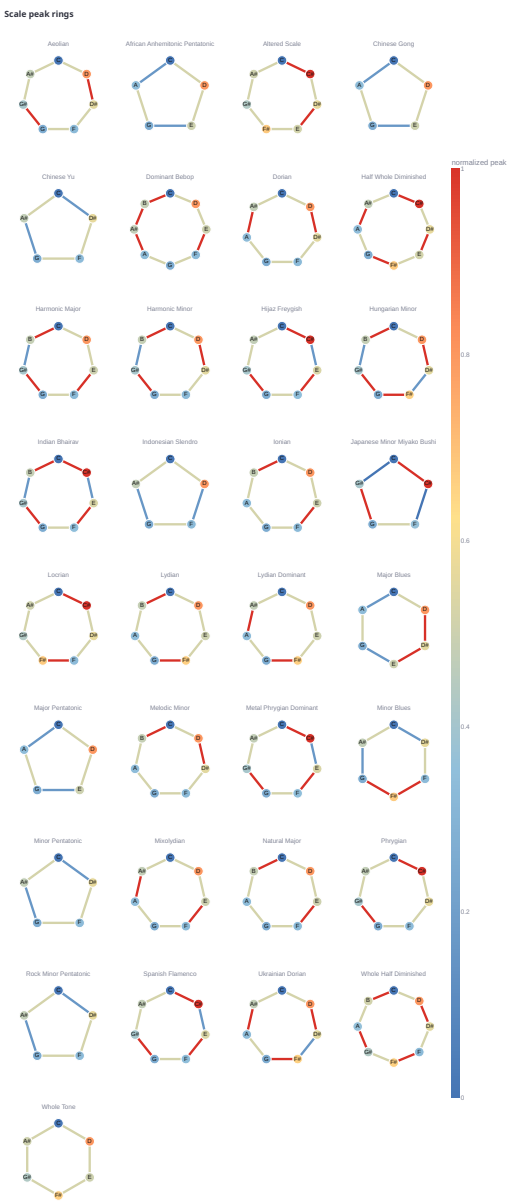

Figure 14: Circular peak patterns for all available scale recordings. Only the seven church modes are used for the ordinal correlation in the main paper.

Figure 15: DS curves for the available scale recordings, illustrating scale-dependent consonance profiles.

Timbre and Microtonal Examples

The same C4–C5 chromatic sequence yields different maxima across seven software instruments, demonstrating sensitivity to partial amplitudes, envelopes, and noise. Because loudness, spectral centroid, and envelope are not matched, these values are descriptive comparisons of complete rendered sounds rather than a causal timbre ranking. A 24-tone-equal-temperament sequence further demonstrates that the continuous kernel can analyze 50-cent steps; no experiential ranking is imposed.

Table 6: Exploratory timbre and microtonal examples.

Audio	Duration (s)	$D_{\max}$	Audio	Duration (s)	$D_{\max}$
Saxophone	14.005	.003134	Violins	15.531	.005633
Guitar	14.005	.003170	Alto	15.214	.007333
Piano	14.099	.004385	Trumpet	14.005	.011913
Flute	15.724	.029903	24-TET sequence	26.005	.003751

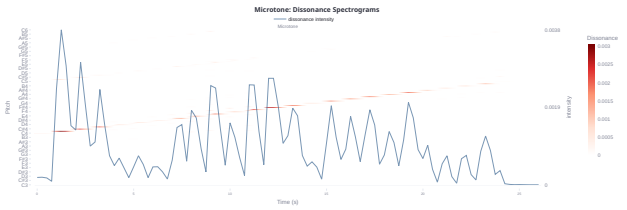

Figure 16: Exploratory 24-TET spectra at 50-cent resolution. The timbre values in Table 6 are descriptive and are not assigned a universal ranking.

Loudness Disentanglement under the Normalized Pipeline

The primary quantitative test uses seven independently rendered C4–C♯4 dyads spanning the configured level conditions ($n = 7$). Let $L_i = \sum_n |y_i[n]|$ be the absolute-amplitude level proxy used by the validation suite (not a perceptual loudness measure), and let Y_i be the DS peak. For positive L_i and Y_i, we fit the robust log–log relation

$$\ln Y_i = a + \beta \ln L_i + \epsilon_i, \tag{23}$$

$$D_{\text{elastic}} = \max(0, 1 - |\beta|), \tag{24}$$

where β is the Theil–Sen slope and a larger D_{elastic} indicates lower proportional sensitivity. The estimate is $\beta = -.000$ with a Theil–Sen 95% CI of $[-.014, .000]$, giving $D_{\text{elastic}} = 1.000$ after rounding. The DS peak varies by only .94% relative standard deviation and 2.49% relative range, defined as $s_Y/\bar{Y}$ and $(\max_i Y_i - \min_i Y_i)/\bar{Y}$, respectively. Thus even the lower confidence bound corresponds to approximately a .014% DS change per 1% loudness change. Pearson $r = -.744$ ($p = .0551$) is also reported, but it describes the ordering of small residual deviations rather than their proportional magnitude; a sizeable $|r|$ can therefore coexist with near-zero elasticity. Because the controlled target and reference contexts are normalized, this sanity check supports near gain-invariance of the normalized DS summary over the tested range rather than universal statistical independence from perceptual loudness.

The companion crescendo stimulus visualizes the time-resolved behavior within one file. It is not included in the seven-sample elasticity estimate because adjacent events share a rendering and temporal context. Future tests should expand the number of independent levels and instruments, report LUFS and clipping diagnostics, and repeat rendering under multiple gain and normalization conventions.

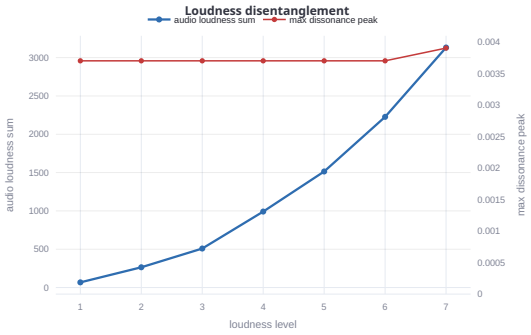

Figure 17: Primary seven-level result: the absolute-amplitude level proxy changes substantially while the maximum DS peak remains tightly concentrated.

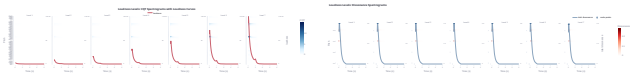

Figure 18: Level-controlled stimuli. Left: CQT magnitude. Right: intrinsic DS.

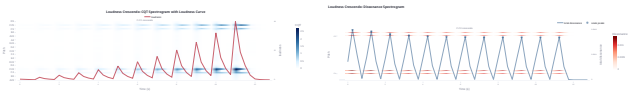

Figure 19: Within-file crescendo visualization. Left: CQT magnitude. Right: intrinsic DS.

Detailed Configuration

Compute and software. All downstream runs use one NVIDIA A100 GPU per Slurm job on Linux; no multi-GPU parallelism is used. The implementations are in PyTorch. CQT and DS extraction is cached before training, and the environment lock file in the code archive records exact framework, CUDA, evaluation-package, and pretrained-model revisions.

Shared preprocessing. Except for PMEmo's released chorus clips, all branches use 24-kHz, 45-second excerpts. Magnitude CQT uses a 1,024-sample hop, eight octaves, 72 bins per octave, $f_{\min} = \text{C1}$, and excerpt-level global-maximum normalization. DS is nonnegative and uses $\log(1 + \mathbf{D})$ compression after the relation transform. CQT and DS use the same 576-bin grid and identical 576-to-144 pooling, temporal masks, encoder, fusion location, parameter count, optimizer, schedule, and checkpoint selection. The comparison therefore matches architecture and spectral grid but does not isolate the relation transform from branch-input compression and normalization. Fixed Gaussian inputs are generated once per excerpt and training seed and reused across epochs.

MU-LLaMA parameters. The Baseline contains 4,205,568 trainable parameters. The matched parallel branch adds 1,360,193 parameters, yielding 5,565,761 trainable parameters in Gaussian, CQT, and DS. The branch uses three convolutional blocks, a 128-dimensional temporal encoder, a kernel-3 depthwise temporal convolution, single-head temporal attention, and gated residual fusion. The output projection is zero-initialized and the gate bias is -3.

Music2Emo parameters. The Baseline task network contains 1,071,617 trainable parameters. The parallel encoder and gated cross-attention residual module add 209,345 parameters, yielding 1,280,962 parameters in Gaussian, CQT, and DS. The added branch is 0.22% of the frozen 95M-parameter MERT encoder and 19.5% of the trainable task network. Following the Music2Emo 70/15/15 protocol (Kang and Herremans 2025), we generate the track-level split once and keep it fixed across all conditions: 1,261/271/270 for DEAM, 495/124/125 for EmoMusic, and 536/116/115 for the 767-track labeled PMEmo subset. These are experiment-specific splits, not official dataset splits. PMEmo contains 794 items in total and uses the released chorus clip.

Evaluation Implementation

MusicQA uses the archive's explicitly specified evaluator, which differs from the released MU-LLaMA scoring script. NLTK word-punctuation tokenization is applied before sentence BLEU with weights $(.25, .25, .25, .25)$ and

method-1 smoothing. METEOR uses exact, stem, and WordNet synonym matching. ROUGE-L uses `rouge-score` with stemming and reports F-measure. BERTScore uses `roberta-large` in English without IDF or baseline rescaling and reports recall. Loss covers answer tokens; perplexity is exponentiated per seed before averaging. For MTG-Jamendo, PR-AUC and ROC-AUC are macro-averaged over the 56 tags. The Music2Emo weighted binary cross-entropy uses $w_i = 2/(1 + p_i)$ for the positive term and $\bar{w}_i = 2p_i/(1 + p_i)$ for the negative term, where p_i is the training prevalence of tag i. The lock file records exact package and model revisions.

Additional MusicQA Comparisons

The MusicQA references and model outputs are free-form text. To avoid cherry-picking fluent examples, we summarize additional corpus-level comparisons rather than selecting answers by visual inspection. The questions follow the MU-LLaMA data-generation setting and cover attributes such as mood, instrumentation, tempo, genre, and overall tone (Liu et al. 2024). Table 7 reports the difference between the DS condition and each matched control using the six-seed means from the fixed 5,040-pair evaluation set. Higher is better for text-similarity metrics and lower is better for loss and perplexity. Seed-level predictions, questions, and references are included in the separately submitted archive for answer-level inspection.

Table 7: Additional MusicQA comparisons. Entries are DS minus the indicated control; negative values are improvements for loss and perplexity.

Metric	vs. Baseline	vs. Gaussian	vs. CQT
BLEU ↑	+.0087	+.0089	+.0018
METEOR ↑	+.0096	+.0098	+.0019
ROUGE-L ↑	+.0115	+.0117	+.0028
BERTScore-R ↑	+.0072	+.0074	+.0028
Test loss ↓	-.025	-.026	-.007
Perplexity ↓	-.046	-.048	-.014

The Gaussian comparison shows that the improvement is not explained by the added parameter budget alone. The smaller but consistently favorable mean difference over the architecture-matched CQT branch indicates that the pitch-resolved input accounts for part, but not all, of the gain. These comparisons remain reference-similarity analyses: automatically generated references can reward paraphrase overlap and do not by themselves establish factual correctness or expert-level harmonic reasoning. We therefore treat the qualitative prediction files as audit material and the paired BERTScore-R endpoint as the prespecified aggregate comparison.

Answer-level audit examples. Tables 8–10 reproduce all ten rows in the supplied audit set rather than selecting a fluent subset. Table 8 maps each case to the audio identifier and MTG-Jamendo track index recorded by the accompanying demo. The *Original Token F1* column is the token-overlap score between the original MU-LLaMA output and the reference answer; it is not a score for the DS-augmented output. Tables 9 and 10 provide the corresponding questions, reference answers, and both inference outputs. Relative to the original outputs, the DS-augmented outputs tend to replace generic genre labels, inferred lyric narratives, or vague "weird" descriptions with more specific acoustic, stylistic, and affective attributes. These examples are qualitative audit material and do not replace the aggregate paired evaluation.

Exploratory Mechanism Analysis

The mechanism analyses are secondary and are not used to select the reported architecture. MusicQA uses seeds $\{17, 42, 101\}$; Music2Emo uses all six paired seeds. Global pooling removes temporal localization, temporal shuffling preserves marginal token statistics while destroying order, pre-projection or early fusion changes the insertion point, and the randomized kernel preserves symmetry and its value distribution while permuting frequency correspondence.

Table 11: Exploratory MusicQA mechanism analysis.

Variant	BERTScore-R ↑	Δ vs. proposed
Global DS	.8985±.0007	−.0030
Randomized kernel	.8992±.0009	−.0023
Temporal DS, shuffled	.8997±.0007	−.0018
Temporal DS, pre-projection	.9004±.0006	−.0011
Temporal DS, proposed	**.9015±.0013**	–

Table 12: Exploratory Music2Emo mechanism analysis.

Variant	Emotion $\overline{R^2}_{\mathrm{VA}}$ ↑
Global DS	.6524±.0015
Temporal DS, shuffled	.6550±.0017
Global DS, early concatenation	.6558±.0016
Temporal DS, proposed	**.6589±.0018**

All perturbations reduce the mean relative to the proposed ordered temporal configuration. These results are consistent with a contribution from frequency correspondence and temporal organization, but they do not isolate a single causal mechanism and support no inferential claim.

Table 8: Provenance and score ownership for the MusicQA audit examples. Audio ID and dataset track index reproduce the `audio` and `trackId` fields in the accompanying demo. Original Token F1 applies only to the original MU-LLaMA output.

ID	Audio ID	Dataset track index	Original Token F1	ID	Audio ID	Dataset track index	Original Token F1
01	`012895.wav`	661609	.000	06	`013055.wav`	873804	.071
02	`012815.wav`	477309	.038	07	`012553.wav`	8308	.111
03	`012725.wav`	267904	.063	08	`012986.wav`	796003	.073
04	`012786.wav`	387501	.087	09	`012972.wav`	768506	.094
05	`012750.wav`	345903	.000	10	`013063.wav`	875005	.103

Table 9: MusicQA answer-level audit examples, cases 01–05. Reference answers and dataset identifiers are reproduced from the accompanying demo.

ID	Question	Reference answer	Original MU-LLaMA	DS-augmented MU-LLaMA
01	What is the term used to describe a style of music that features a strong emphasis on melody and harmony?	Acoustic rock is a term used to describe a style of music that features a strong emphasis on melody and harmony. Acoustic rock typically combines elements of rock and folk music, with acoustic guitar, harmonica, and vocals as the main instruments.	Classical.	Acoustic rock.
02	What type of orchestral music is characterized by complex harmonies and intricate compositions?	Progressive orchestral. Progressive orchestral music incorporates elements of progressive rock, avant-garde, and experimental music, and is characterized by complex harmonies and intricate compositions.	Symphonic music.	Progressive orchestral music.
03	What can be inferred from the audio?	The audio blends folk, punk rock, rock, and ska. It uses acoustic guitars, bass, drums, and occasional brass and strings, with gruff vocals, harmonies, and call-and-response sections.	The song is about a man's love for a woman.	The music blends folk, punk rock, ska, and acoustic instrumentation.
04	Describe the audio in detail.	The audio is a soothing and upbeat tune with a catchy melody and harmonious vocals. Harp, synthesizer, and piano create an elegant, layered, and enjoyable arrangement.	The music is described as slow and weird.	The music is described as soothing, upbeat, and harmonious.
05	What is the genre of music that is characterized by a slow and smooth rhythm, often played on a saxophone?	Bossanova. It originated in Brazil and is known for its relaxed and romantic sound, often featuring a slow and smooth rhythm.	Jazz.	Bossanova.

Reference Comparisons under Modified Protocols

Table 13: Contextual MU-LLaMA comparison. The original values use the released scoring script on 4,500 QA pairs from 500 tracks (Liu et al. 2024); our values use the explicitly specified evaluator and are six-seed means on the custom 5,040-pair, 560-track set.

Source	BLEU	METEOR	ROUGE-L	BERTScore-R
Original report (released script)	.3060	.3850	.4660	.9010
Our baseline (specified metrics)	.2987	.3761	.4556	.8952
Absolute difference	.0073	.0089	.0104	.0058

The rows are not treated as a direct reproduction comparison because both the evaluation set and metric implementation differ.

Table 14: Contextual Music2Emo comparison. The original report uses its 30-second segment augmentation and original training protocol (Kang and Herremans 2025); our baseline uses the modified fixed-excerpt protocol described here.

Source	J-PR	J-ROC	DEAM-V	DEAM-A	Emo-V	Emo-A	PM-V	PM-A
Original report	.1543	.7810	.5184	.6228	.6512	.7616	.5473	.7940
Our baseline	.1539	.7806	.5169	.6209	.6487	.7598	.5451	.7926
Absolute difference	.0004	.0004	.0015	.0019	.0025	.0018	.0022	.0014

Table 10: MusicQA answer-level audit examples, cases 06–10 (continued).

ID	Question	Reference answer	Original MU-LLaMA	DS-augmented MU-LLaMA
06	What can be inferred from the audio?	The audio appears to be a movie or TV soundtrack with a moody, atmospheric character. Its slow, mellow melody and layered guitars and piano evoke longing, nostalgia, and introspection.	The song is about love and heartbreak.	The music feels moody, mellow, nostalgic, and introspective.
07	Describe the audio in detail.	A relaxing instrumental composition combining electronic, lounge, and oriental music. Its slow, smooth tempo and minimal instrumentation create a calming atmosphere.	The music is described as having a weird and otherworldly quality.	The music is described as relaxing, smooth, and gently oriental.
08	What can be inferred from the audio?	A relaxing and upbeat trip-hop track for chilling out. Synthesizer, strings, bass, drums, Rhodes, and trumpet create a dreamy, smooth, and sophisticated atmosphere.	The song is about love and heartbreak.	The track is relaxing, upbeat, dreamy, and sophisticated.
09	Describe the audio in detail.	A calm orchestral composition. Acoustic bass, cello, piano, violin, bell, and oboe create a warm, elegant, magical, and serene piece for relaxing.	The music is described as having a weird and otherworldly quality.	The music is described as calm, warm, elegant, and serene.
10	Describe the audio in detail.	A fusion of classical and ambient music with flute and synthesizer, a slow and mellow pace, bass and electric guitar foundation, and soulful background vocals.	The music is described as having a weird and scary atmosphere.	The music is described as slow, mellow, soulful, and ambient.

Seed-Level Primary Endpoints

Table 15: MusicQA BERTScore-R for the six paired seeds.

Seed	Baseline	Gaussian	CQT	DS
17	.8943	.8942	.8990	.9001
42	.8950	.8948	.8996	.9017
101	.8954	.8951	.8998	.9027
2025	.8948	.8947	.8989	.9028
2026	.8962	.8960	.8997	.9024
3407	.8955	.8952	.9006	.9047
Mean	.8952	.8950	.8996	**.9024**
SD	.0007	.0006	.0006	.0015

Table 16: Music2Emo $\overline{R^2}_{\mathrm{VA}}$ for the six paired seeds.

Seed	Baseline	Gaussian	CQT	DS
17	.6453	.6449	.6519	.6560
42	.6472	.6466	.6542	.6587
101	.6481	.6476	.6552	.6596
2025	.6464	.6457	.6525	.6580
2026	.6494	.6488	.6557	.6602
3407	.6476	.6468	.6555	.6609
Mean	.6473	.6467	.6542	**.6589**
SD	.0014	.0014	.0016	.0018

Paired Comparisons

Table 17: Paired comparisons on the two designated primary endpoints using unrounded seed-level values. $p_{\mathrm{H}}^{(t)}$ is Holm-adjusted paired-t; $p_{\mathrm{H}}^{(\mathrm{sign})}$ is the Holm-adjusted exact two-sided sign test based only on paired directions. Each task has one three-comparison Holm family.

Task	Comparison	Mean Δ	$p_{\mathrm{H}}^{(t)}$	$p_{\mathrm{H}}^{(\mathrm{sign})}$
MusicQA	DS vs. Baseline	+.0072	$< .001$	.0938
MusicQA	DS vs. Gaussian	+.0074	$< .001$	.0938
MusicQA	DS vs. CQT	+.0028	.0017	.0938
Music2Emo	DS vs. Baseline	+.0116	$< .001$	.0938
Music2Emo	DS vs. Gaussian	+.0122	$< .001$	.0938
Music2Emo	DS vs. CQT	+.0047	$< .001$	.0938

References

Kang, J.; and Herremans, D. 2025. Towards Unified Music Emotion Recognition across Dimensional and Categorical Models. arXiv:2502.03979.

Liu, S.; Hussain, A. S.; Sun, C.; and Shan, Y. 2024. Music Understanding LLaMA: Advancing Text-to-Music Generation with Question Answering and Captioning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 286–290.

Malmberg, C. F. 1918. The Perception of Consonance and Dissonance. *Psychological Monographs*, 25(2): 93–133.

Schwartz, D. A.; Howe, C. Q.; and Purves, D. 2003. The Statistical Structure of Human Speech Sounds Predicts Musical Universals. *The Journal of Neuroscience*, 23(18): 7160–7168.