

Moodify: An Explainable Post-Processing and Quality Evaluation Framework for AI-Generated Music

Xuanzhen Zhu

Moodify Sound Laboratory, Rongjing Wenchuan (Shenzhen) Technology Co., Ltd., Shenzhen, China

Manuscript type: Engineering Report draft for internal review and experimental completion

Abstract—AI-generated music can be compositionally complete while still exhibiting production-level defects such as spectral imbalance, weak vocal presence, unnatural dynamics, unstable stereo imaging, and processing artifacts. This engineering report presents Moodify, an explainable post-processing and quality-evaluation framework designed for full-mix AI-generated music. The current stable implementation provides an operational pipeline from audio import to six-band analysis, rule-based diagnosis, preset-driven digital signal processing, waveform export, and treatment-record generation. A research layer extends this baseline with a five-dimensional acoustic state representation, Wave Health Score (WHS), Emotion Development Score (EDS), Moodify Reality Score (MRS), risk models, quality gates, and versioned processing history. MRS measures weighted distance from a reference feature distribution across spectrum, dynamics, transients, space, texture, temporal stability, and artifacts. The report also defines a theory-experiment-engineering-data loop intended to prevent unsupported theoretical concepts and unvalidated code from entering production. The contribution is an auditable engineering architecture and a reproducible validation program; superiority over human mastering or commercial systems is not claimed. Controlled-degradation tests are implemented, while large-scale listening evaluation and cross-genre calibration remain future work.

Keywords—audio quality assessment; automatic music production; digital signal processing; explainable systems; music mastering

1 Introduction

Text-to-music and audio-generation systems have substantially reduced the cost of producing complete musical recordings. The generated output, however, is not necessarily equivalent to a release-ready master. In practice, a generated full mix may contain excessive or insufficient low-frequency energy, recessed vocal presence, harsh high-frequency texture, narrow or unstable dynamics, exaggerated stereo width, inconsistent spatial cues, or cumulative processing artifacts. These defects are difficult to address with a single global objective because production quality, perceptual quality, and artistic intention are related but non-identical.

Automated mastering and intelligent music production have been studied through rule-based systems, machine-learning approaches, differentiable audio effects, and online services. Comparative work has shown that automated mastering can be evaluated meaningfully only when objective measurements are combined with subjective listening tests [1]. Differentiable signal-processing systems can estimate effect parameters and achieve competitive results in selected tasks [2], while data-driven automatic mixing remains constrained by data availability, generalization, and perceptual evaluation [3]. These observations motivate a system that is not merely an “enhance” button but an auditable workflow in which measurements, processing decisions, risks, and results are recorded.

This report introduces Moodify, an engineer-

ing framework for AI-music post-processing. Its present contribution is not a new generative model. It is a structured layer after generation that attempts to answer seven operational questions: (1) what is the acoustic state of the input; (2) which defects are measurable; (3) which processing chain is appropriate; (4) what risks may be introduced; (5) whether the output is technically improved; (6) whether the artistic target is better represented; and (7) whether the result and its history can be reproduced.

The primary contributions are:

1. an operational, local-first baseline pipeline for full-mix audio;
2. a transparent mapping from measured features to diagnosis and DSP parameters;
3. three complementary internal scores for technical health, target-direction movement, and reference-distribution distance;
4. a multi-stage quality-gate and rollback design; and
5. a theory-experiment-engineering-data loop for cumulative knowledge production.

2 Related Work

2.1 Automatic Production and Mastering

Automatic music production has progressed from heuristic equalization and dynamics control toward learned parameter estimation and end-to-end optimization. Ramírez et al. integrated non-differentiable black-box effects into a train-

able system and demonstrated applications including automatic mastering [2]. Later work explored automatic mixing with out-of-domain multitrack data and subjective evaluation by experienced engineers [3]. A comparative JAES study found that human engineers and automated services can produce measurably and perceptually different masters, reinforcing the need for controlled, multidimensional evaluation rather than simple loudness comparison [1].

Recent unified restoration and mastering research has begun to address multiple defects in a single controllable model, including tonal imbalance, clipping, reverberation, and stereo narrowing [4]. Moodify differs in its current stage and design goal: it uses a conservative rule-based baseline and a modular research layer, emphasizing interpretability, rollback, treatment records, and incremental calibration rather than claiming end-to-end generative restoration.

2.2 Objective and Subjective Quality Assessment

Objective audio quality metrics have well-known limitations. ITU-R BS.1770 specifies programme loudness and true-peak measurement [5]; ITU-R BS.1387 specifies perceptual evaluation of audio quality [6]; and ITU-R BS.1534 defines MUSHRA for controlled subjective assessment [7]. Fr chet Audio Distance (FAD) was proposed as a reference-free distributional metric for music enhancement and showed higher correlation with human judgments than several signal-domain distances in its validation setting [8]. These methods demonstrate that no single scalar fully represents musical quality. Accordingly, Moodify treats scalar scores as internal decision aids rather than universal truth.

2.3 Open Audio Engineering Components

The current implementation uses established open-source components. `librosa` provides music and audio analysis utilities [9]; Spotify's `pedalboard` provides programmatic audio effects and rendering [10]; and harmonic-percussive source separation by median filtering provides a lightweight decomposition route for experimental workflows [11]. The contribution of Moodify lies in how measurement, diagnosis, processing, risk, evaluation, and version records are organized, not in claiming ownership of these underlying operators.

3 Design Scope and Principles

Moodify is defined as an acoustic diagnosis, secondary processing, quality evaluation, and craft-memory system for AI-generated music. It does not generate lyrics, melody, or performance, and

it is not a complete digital audio workstation. The stable baseline is intentionally deterministic and conservative.

Five principles constrain the system:

1. **Explainability:** each processing action must correspond to a diagnosed condition or selected craft objective.
2. **Non-destructive operation:** the source file is preserved and outputs are separately versioned.
3. **Multiple criteria:** loudness, spectral balance, dynamics, space, artifacts, and artistic direction are not collapsed prematurely.
4. **Candidate before replacement:** a new algorithm may generate a candidate output but does not automatically become the accepted version.
5. **Reproducibility:** parameters, metrics, software version, and decision records should be recoverable.

4 System Architecture

4.1 Stable v0.1 Mainline

The stable mainline implements the sequence shown in Fig. 1. The public repository identifies `v01_pipeline.py` as the sole orchestrator for version 0.1.0. It calls analysis, rule-based diagnosis, a preset-selected DSP chain, export, and report generation [12].

The API exposes health, preset-listing, and processing endpoints. The canonical process path accepts an uploaded file, resolves one of three presets, runs the stable pipeline, and returns either a WAV file or JSON metadata. A 50 MB upload limit is used in the current implementation. A repository snapshot for tag `v0.1.0-alpha.1` reports 20 tests for the v0.1 path and 104 tests across baseline, legacy, and experimental modules, all passing at the time of that snapshot [12]. This is an implementation-status result, not evidence of perceptual superiority.

4.2 Acoustic Analysis

The baseline analyzer computes file duration, sample rate, channel count, peak level, crest factor, a short-time dynamic-range estimate, left-right correlation, and relative energy in six bands. Let $X[k]$ be the windowed discrete Fourier transform and $\mathcal{B}_b$ the bins of band b . The relative band-energy descriptor is

$$E_b = \frac{\sum_{k \in \mathcal{B}_b} |X[k]|^2}{\sum_k |X[k]|^2 + \epsilon}, \quad (1)$$

with the reported decibel feature proportional to $20 \log_{10} \sqrt{E_b + \epsilon}$. The six bands are listed in Table 1.

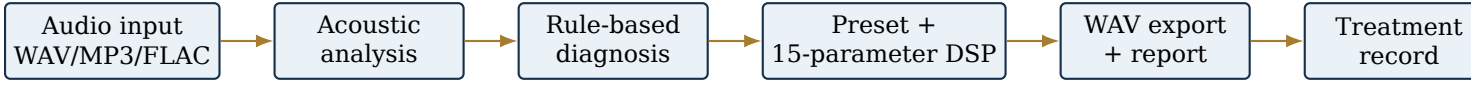


Figure 1: Stable Moodify v0.1 processing mainline. Research modules are maintained separately until validation and integration requirements are met.

Table 1: Baseline frequency bands.

Band	Range (Hz)
Sub	20-60
Bass	60-250
Low-mid	250-500
Mid	500-2000
Presence	2000-5000
Air	8000-16000

Dynamic range is estimated as the difference between the 95th and 5th percentiles of RMS level in 100 ms windows. Stereo correlation uses Pearson correlation between left and right channels. These features are inexpensive and interpretable, but they do not fully represent timbre, masking, musical structure, or perceived artistic quality.

4.3 Rule-Based Diagnosis

The stable diagnostic layer applies explicit thresholds to identify prominent or weak low-frequency energy, excessive or insufficient presence, missing air-band energy, narrow or extreme dynamics, and potentially problematic stereo correlation. The output includes issues, strengths, an overall label (good, fair, or poor), and suggested presets. This layer deliberately does not use machine learning, retrieval-augmented generation, or online calibration. Its value is traceability; its weakness is limited genre and context sensitivity.

4.4 DSP Chain and Presets

The baseline contains three presets: `warm_vocal`, `clean_master`, and `wide_space`. Each defines 15 parameters covering vocal-presence equalization, low-frequency warmth, compressor ratio/attack/release/threshold, reverberation, width, harmonic drive, and high-shelf air enhancement. The current effect order is

Peak EQ → Low Shelf → Compressor → Reverb → Distortion → High Shelf → Gate → Limiter (2)

The limiter uses a nominal -1 dBFS ceiling. The presets are starting points, not universal targets. The current baseline does not yet synthesize all 15 parameters dynamically from the diagnosis; it either uses a selected preset or a suggested preset path.

4.5 Research Layer

The repository also contains diagnosis, knowledge, risk, optimization, orchestration, physics, safety, calibration, and memory modules that are not all connected to the v0.1 public processing path. Table 2 separates stable and research capabilities.

5 Quality Representation and Scores

5.1 Five-Dimensional Acoustic State

The research layer represents a recording by

$$\mathbf{W} = \{\mathbf{S}, \mathbf{D}, \mathbf{P}, \mathbf{L}, \mathbf{E}\}, \quad (3)$$

where $\mathbf{S}$ denotes spectrum, $\mathbf{D}$ dynamics, $\mathbf{P}$ spatial properties, $\mathbf{L}$ layer-related descriptors, and $\mathbf{E}$ emotion-related descriptors. This is an internal engineering coordinate system rather than a claim that music has exactly five objectively complete dimensions.

Processing is modeled as a state transition

$$\mathbf{W}_1 = F(\mathbf{W}_0, \mathbf{p}), \quad (4)$$

where $\mathbf{p}$ is a vector of controllable DSP parameters. A target-oriented optimization may be written as

$$\mathbf{p}^* = \arg \min_{\mathbf{p}} d(F(\mathbf{W}_0, \mathbf{p}), \mathbf{W}_T) + \lambda R(\mathbf{p}), \quad (5)$$

with distance d , target state $\mathbf{W}_T$, risk R , and risk weight λ .

5.2 Wave Health Score

Wave Health Score (WHS) is an internal 0-100 technical-health indicator. For diagnostic dimensions d with health scores q_d and weights w_d ,

$$\text{WHS} = \sum_d w_d q_d, \quad \sum_d w_d = 1. \quad (6)$$

The current implementation derives dimension scores from classified defect severities. WHS is distorted to high shelf when no measurable technical defects are present; it is not a complete artistic score.

5.3 Emotion Development Score

Emotion Development Score (EDS) measures whether processing moves an internal process

Table 2: Implementation maturity separation.

Layer	Implemented stable baseline	Research or partial integration
Input/output	WAV/MP3/FLAC input; WAV export; JSON report	asynchronous catalogue processing; platform-scale storage
Analysis	six-band energy, peak, crest factor, dynamic range, L/R correlation	five-dimensional state, 18-parameter diagnosis, temporal and texture descriptors
Decision	rule thresholds; three preset suggestions	candidate search, RAG recommendation, craft cards
Processing	fixed 15-parameter DSP presets	HPSS-assisted processing, multi-candidate rendering, adaptive parameter selection
Evaluation	before metrics and diagnosis	WHS, EDS, MRS, multi-stage risk and quality gates
Memory	report files and treatment records	calibrated case retrieval, version capital, full-catalogue re-rendering

vector toward a selected target vector. Let $\mathbf{z}_0$, $\mathbf{z}_1$, and $\mathbf{z}_T$ denote the original, processed, and target vectors. The implementation uses

$$\text{EDS} = 100 \left(1 - \frac{\|\mathbf{z}_1 - \mathbf{z}_T\|_2}{\|\mathbf{z}_0 - \mathbf{z}_T\|_2 + \epsilon} \right). \quad (7)$$

Positive values indicate movement toward the target under the defined mapping. Current target vectors are provisional and require listening-based calibration; therefore EDS must not be presented as a validated emotion-perception standard.

5.4 Moodify Reality Score

MRS is a reference-distribution distance score. The implementation extracts groups of features for spectrum, dynamics, transients, space, texture, temporal stability, and artifacts. For feature j in group g , with reference mean μ_{gj} and standard deviation σ_{gj} , the group distance is

$$d_g = \sqrt{\frac{1}{n_g} \sum_{j=1}^{n_g} \left(\frac{x_{gj} - \mu_{gj}}{\sigma_{gj} + \epsilon} \right)^2}. \quad (8)$$

The total distance and score are

$$D_R = \sum_g \alpha_g d_g, \quad \text{MRS} = 100e^{-D_R}, \quad (9)$$

where α_g are nonnegative weights. The current default weights are 0.20 for spectrum, 0.15 for dynamics, 0.10 for transients, 0.15 for space, 0.20 for texture, 0.15 for temporal stability, and 0.05 for artifacts. MRS is conceptually related to reference-free distributional assessment such as FAD [8], but it is a handcrafted feature-distance measure and has not yet demonstrated comparable perceptual validation.

5.5 Composite Decision and Quality Gates

A candidate may be ranked by a composite objective

$$J = \beta_1 \Delta \text{WHS} + \beta_2 \text{EDS} - \beta_3 R, \quad (10)$$

but no single scalar is sufficient for release. The research design includes three gates: diagnosis completeness, decomposition quality, and output compliance. The output gate considers loudness, true peak, loudness range, peak-to-loudness ratio, width, low-frequency correlation, and fatigue risk. Some checks, including complete distortion and mono-compatibility logic, remain placeholders and are explicitly treated as unfinished.

6 Theory-Experiment-Engineering-Data Loop

A recurring failure mode in experimental software is divergence between conceptual documents and production code. Moodify addresses this by requiring bidirectional synchronization among theory, experiments, engineering, and data, as illustrated in Fig. 2.

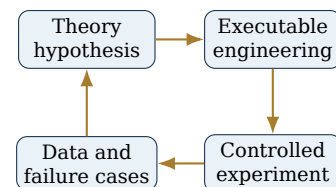


Figure 2: Theory-experiment-engineering-data loop. Evidence may confirm, bound, or reject a hypothesis.

A theoretical concept becomes operational only after the sequence

$$\text{concept} \rightarrow \text{variable} \rightarrow \text{measurement} \rightarrow \text{experiment} \rightarrow \text{imp} \quad (11)$$

Conversely, repeated disagreement between predictions and listening evidence requires revision of the theory. The experiment layer permits aggressive or failed approaches; the stable layer accepts only reproducible modules with boundaries, bypass, rollback, fallback, and regression testing.

7 Preliminary Validation Program

7.1 Controlled-Degradation Tests

The repository contains a directional MRS validation script that constructs controlled degradations from reference audio. Table 3 summarizes the current test design. These tests examine whether a metric responds in the expected direction; they do not establish correlation with human preference.

7.2 Required Listening Validation

A submission-ready perceptual study should use loudness-matched conditions and a double-blind procedure. A practical first protocol would include at least 30 tracks across multiple AI-generation systems and genres, original and processed conditions, hidden identity, and separate ratings for naturalness, clarity, fatigue, spatial coherence, intention preservation, and overall preference. MUSHRA principles should be used where appropriate [7]. Listener expertise should be recorded, and reliability, confidence intervals, effect sizes, and inter-rater consistency should be reported.

7.3 Ablation and Benchmarking

The following ablations are necessary:

- analysis plus no processing;
- each preset separately;
- full chain without reverb, compression, harmonic drive, or width change;
- fixed preset versus candidate search;
- quality gate enabled versus disabled;
- MRS components individually removed.

Comparative baselines should include the unprocessed input, a simple loudness-normalized output, at least one established automated mastering service where licensing permits evaluation, and preferably a human-engineered condition. These experiments are future work and no comparative superiority is asserted in this report.

8 Discussion

8.1 Engineering Value

The baseline demonstrates that an AI-music post-processing product can be organized as a transparent sequence rather than a black-box en-

hancement claim. The measurements are inexpensive, the rules are interpretable, and the effects chain is reproducible. The modular separation between stable and research paths is particularly important: it allows ambitious metrics and candidate-selection logic to evolve without destabilizing the current processing entry point.

8.2 Knowledge Accumulation

The long-term value of the system depends less on any individual equalizer or compressor than on paired causal records:

$$\mathcal{R}_i = \{\mathbf{W}_{i,0}, \mathbf{p}_i, \mathbf{W}_{i,1}, \Delta Q_i, R_i, L_i, F_i\}, \quad (12)$$

where L_i is listening feedback and F_i is a failure label. Such records support retrieval, calibration, boundary discovery, and regression testing. Public audio files alone do not provide this causal structure.

8.3 Limitations

The current system has important limitations. The stable diagnosis is based on coarse fixed thresholds. Preset parameters are not fully adapted to each track. Some research metrics rely on provisional mappings and weights. The MRS reference distribution has not yet been documented with a sufficiently large, diverse, and independently reviewed dataset. The output quality gate contains unfinished checks. Full-mix processing cannot replace multitrack mixing, and a technically healthier output may still violate artistic intention. Finally, generated music is heterogeneous; defects associated with one model or genre may not generalize to another.

9 Roadmap to a Research-Grade System

The immediate research program is:

1. freeze a versioned evaluation dataset with provenance and licensing records;
2. implement standards-aligned loudness, true-peak, mono compatibility, and distortion checks;
3. conduct controlled MUSHRA-style listening experiments;
4. estimate correlations between WHS, EDS, MRS, and human ratings;
5. calibrate parameter ranges by genre and defect type;
6. publish failure cases and negative results;
7. connect validated research modules to the stable pipeline only after regression testing; and
8. implement non-destructive versioned rendering from preserved source masters.

Table 3: Controlled-degradation validation design implemented for MRS.

Category	Example manipulations	Directional expectation
Spectrum	high-frequency boost; high-frequency attenuation; multi-band flattening	degraded variants should generally obtain lower MRS than the reference
Dynamics	strong compression; clipping/limiting; section-dependent gain instability	reduced naturalness or stability should increase distance
Space	excessive side-channel gain; inter-channel delay; mono collapse	spatial anomalies should be reflected in the space component
Temporal stability	late-section high-frequency boost, gain drift, and spectral spikes	late degradation should increase temporal and artifact distance
Artifacts	clipping ratio, abnormal peaks, narrow spectral spikes	higher artifact incidence should reduce MRS

10 Conclusion

This engineering report presented Moodify, an explainable framework for post-processing and evaluating AI-generated music. The current stable system implements audio import, interpretable measurement, rule-based diagnosis, three DSP presets, WAV export, and treatment records. The research layer adds multidimensional state representation, WHS, EDS, MRS, risk models, quality gates, candidate generation, and historical memory. The central architectural contribution is not a claim that one score defines good music, but a disciplined evidence chain in which theory must become measurable, processing must be reversible, and candidate outputs must be evaluated before release. The system has reached an operational engineering-prototype stage. Its next scientific requirement is not more conceptual complexity, but reproducible datasets, standards-aligned measurements, ablation studies, and controlled human listening evidence.

Code and Data Availability

The current Moodify source repository is publicly accessible under the license stated in the repository [12]. Audio samples, third-party generated music, and future listening-test data require separate rights and consent management. No proprietary or private audio was used in preparing this manuscript draft.

Competing Interests

The author is the founder of Rongjing Wenchuan (Shenzhen) Technology Co., Ltd. and the originator of the Moodify project. This relationship

may create a commercial interest in the system described.

AI-Assistance Disclosure

OpenAI ChatGPT (GPT-5.6 Thinking, accessed 23 July 2026) was used for substantive assistance in manuscript organization, initial drafting, language editing, and LaTeX preparation. The human author is responsible for verifying all technical claims, equations, references, experimental results, and interpretations before submission. No AI-generated experimental data are reported.

Author Biography

Xuanzhen Zhu is the founder and chief artistic director of Rongjing Wenchuan (Shenzhen) Technology Co., Ltd. His current work focuses on AI-generated music post-processing, explainable audio engineering workflows, cultural intellectual property, and the integration of technical standards with content platforms. A formal biography, author photograph, correspondence address, and ORCID should be added before journal submission.

References

- [1] S. H. Chon, H. Lee, and K. Lee, "A Comparative Study of Music Mastered by Human Engineers and Automated Services," *J. Audio Eng. Soc.*, vol. 70, no. 9 (2022 Sep.), doi: 10.17743/jaes.2022.0050.
- [2] M. A. Martínez Ramírez, O. Wang, P. Smaragdis, and N. J. Bryan, "Differentiable Signal Processing With Black-Box Audio Effects," in *Proc.*

- IEEE Intl. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 66–70 (2021), doi: 10.1109/ICASSP39728.2021.9415103.
- [3] M. A. Martínez-Ramírez, W.-H. Liao, G. Fabbro, S. Uhlich, C. Nagashima, and Y. Mitsufuji, “Automatic Music Mixing With Deep Learning and Out-of-Domain Data,” in *Proc. Intl. Soc. Music Information Retrieval Conf. (ISMIR)* (2022).
- [4] J. Melechovsky, A. Mehrish, and D. Herremans, “SonicMaster: Towards Controllable All-in-One Music Restoration and Mastering,” arXiv:2508.03448 (2025).
- [5] International Telecommunication Union, “Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level,” Recommendation ITU-R BS.1770-5 (2023 Nov.).
- [6] International Telecommunication Union, “Method for Objective Measurements of Perceived Audio Quality,” Recommendation ITU-R BS.1387-2 (2023 May).
- [7] International Telecommunication Union, “Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems,” Recommendation ITU-R BS.1534-3 (2015 Oct.).
- [8] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, “Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms,” in *Proc. Interspeech*, pp. 2350–2354 (2019), doi: 10.21437/Interspeech.2019-2219.
- [9] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and Music Signal Analysis in Python,” in *Proc. Python in Science Conf.*, pp. 18–25 (2015), doi: 10.25080/Majora-7b98e3ed-003.
- [10] P. Sobot, “Pedalboard,” software release, Spotify (2021–2026), doi: 10.5281/zenodo.7817838.
- [11] D. FitzGerald, “Harmonic/Percussive Separation Using Median Filtering,” in *Proc. Intl. Conf. Digital Audio Effects (DAFx)*, Graz, Austria (2010 Sep.).
- [12] X. Zhu, “Moodify: AI Music Post-Processing and Listening Engineering System,” version 0.1.0, software repository (2026), <https://github.com/huliye24/moodify-o3is>.