MUSEREASON: FROM MUSICAL EVIDENCE TO COUNTERFACTUAL EDITING VIA BIDIRECTIONAL AFFECTIVE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Music emotion understanding and emotion conditioned music generation have made significant progress in recent years, but they are usually modelled as relatively independent tasks. Existing music emotion understanding methods mainly focus on emotion state prediction and lack an explicit characterisation of the musical evidence supporting emotional judgements. Emotion conditioned music generation methods typically take the target emotion directly as a condition, with less reasoning about which musical attributes should be adjusted based on the source musical state and which constraints should be maintained. This paper proposes ***MuseReason***, a bidirectional emotion reasoning framework for music emotion understanding and counterfactual editing. MuseReason uses temporal music representations to extract verifiable musical evidence. It then combines the musical evidence with the target affect to predict structured music adjustments and preservation constraints for counterfactual editing. The experimental results show that MuseReason demonstrates capabilities in emotional understanding, structured output, and temporal emotional modelling. Structured music adjustment can support counterfactual editing by considering emotional changes while maintaining the source music. MuseReason extends music emotion modelling from a single emotional prediction to a unified framework that connects musical evidence, emotional reasoning, structured music adjustment, and generated result verification.

1 INTRODUCTION

Music emotion research concerns both understanding the affect expressed by existing music and editing music towards a desired affect. Musical affect arises from factors such as tempo, mode, rhythm, energy, timbre, and structural evolution (Armitage et al., 2025), and it also changes over time (Gu et al., 2026). Music affect modelling therefore requires more than mapping music to an emotion label. Understanding should identify the musical evidence behind an affective judgement, while editing should determine what to change and what to preserve. This motivates a framework that connects musical evidence, affect understanding, structured musical adjustment, and counterfactual editing.

Existing research has made substantial progress in both music emotion understanding and emotionally controllable generation. Music emotion models can predict discrete categories, continuous valence and arousal, and affective trajectories that vary over time (Fong et al., 2025; Lee et al., 2025; Peintner et al., 2025). Music language models further support musical description and emotion related question answering (Gardner et al., 2024; Mao et al., 2025; Liu et al., 2024; Deng et al., 2024). Yet, the musical evidence underlying these judgements often remains implicit, making it difficult to relate predicted affect to concrete musical cues. Controllable generation and editing methods can condition on emotion descriptions, musical attributes, or reference music (Mei et al., 2025; Tal et al., 2024; Tsai et al., 2025; 2024; Zhang et al., 2025), but they typically specify the target affect directly without reasoning from the current state of the source music. They also rarely identify what should change and what should be preserved. Consequently, affect understanding, musical evidence, and concrete editing decisions remain weakly connected, while generated outcomes are seldom used to assess whether the intended change was realised. Figure 1 illustrates this missing reasoning bridge.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

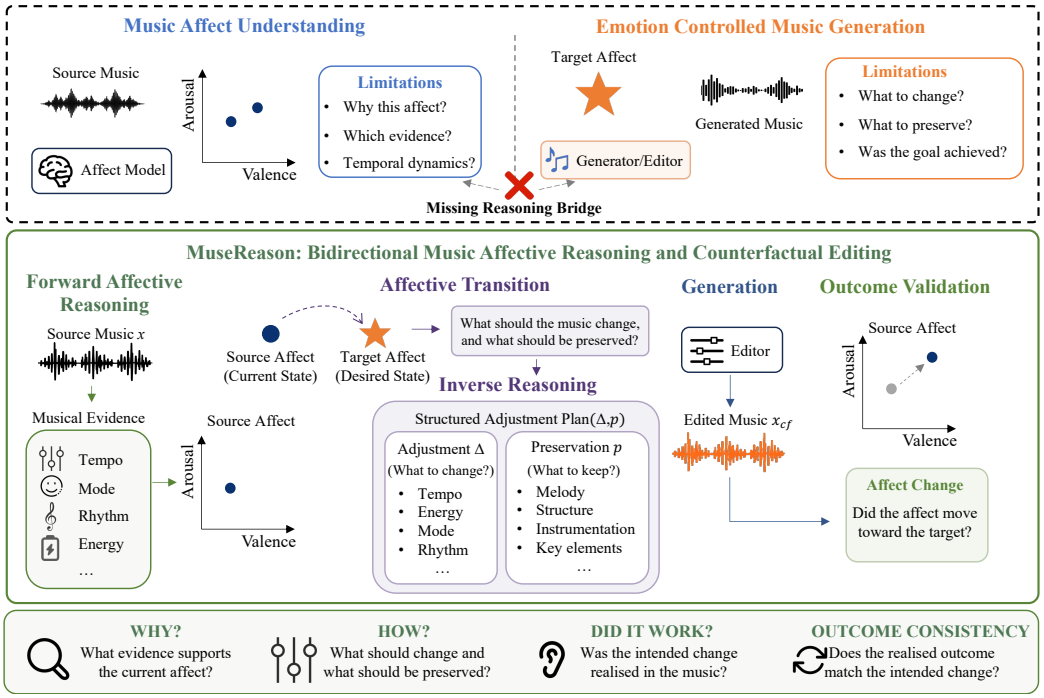


Figure 1: Motivation for MuseReason. The framework connects musical evidence, affect understanding, structured adjustment, counterfactual editing, and outcome consistency reasoning.

To address these issues, we propose ***MuseReason***, a bidirectional emotional reasoning framework for music emotion understanding and counterfactual editing. Rather than directly establishing a mapping between music and emotion, MuseReason first extracts temporal, verifiable musical evidence from the music during forward reasoning and then forms an emotional judgement on this basis. During inverse reasoning, MuseReason combines the emotional state of the original music, the musical evidence, and the target emotion to predict a structured musical adjustment plan. This design no longer treats emotional understanding and music generation as two independent tasks. Building on this, MuseReason connects intended musical adjustments with their realised outcomes through outcome consistency reasoning. This allows the model to recognise and interpret the emotions expressed by music, infer how the source should be edited towards a target affect, and assess the realised result. Our contributions are summarised as follows:

- We propose MuseReason, a bidirectional emotional reasoning framework for music that is based on evidence. This framework unifies music emotion understanding and emotional control within a single reasoning framework, rather than treating them as two independent tasks.
- We propose a structured music adjustment method for counterfactual editing. This method specifies what should change and what should be preserved before the source music is edited towards the target affect.
- We introduce outcome consistency reasoning between intended musical adjustments and their realised results. Independent affect, preservation, and musical measurements are used to assess the generated music after editing.

2 Related Work

Music Affect Understanding. Music emotion recognition has progressed from discrete emotion classification to continuous valence and arousal modelling, together with temporal affect estimation (Aljanaki et al., 2017; Lee et al., 2025; Peintner et al., 2025; Gu et al., 2026). Pretrained music representations such as MERT and MuQ further provide expressive features for music understanding (Li et al., 2024; Zhu et al., 2025). Recent music language models extend these capabilities to

natural language interaction and broader musical reasoning (Gardner et al., 2024; Deng et al., 2024; Mao et al., 2025; Ghosh et al., 2026). However, these approaches typically focus on predicting or describing musical content. They rarely connect affective judgements to explicit musical evidence and use that evidence to reason about subsequent musical changes.

Controllable Music Generation and Editing. Controllable music generation supports affect control through emotion labels, continuous affect representations, and musical attributes (Ferreira et al., 2022; Huang et al., 2024; Mei et al., 2025; Tsai et al., 2025). Methods conditioned on references or instructions further enable the modification of existing music while preserving selected source characteristics (Tsai et al., 2024; Zhang et al., 2025). Most approaches nevertheless treat the desired emotion as a direct generation condition. They do not explicitly reason from the source affect and musical evidence to determine what should change and what should remain stable. MuseReason instead predicts structured musical adjustments and preservation constraints before counterfactual editing.

Grounded and Counterfactual Reasoning. Grounded reasoning links model decisions to observable evidence, while counterfactual reasoning studies how outcomes change under hypothetical modifications (Wachter et al., 2018; Khorram & Li, 2022; Wu et al., 2026). Related consistency constraints have also been used to preserve relevant structure under transformations (Zhu et al., 2017; Khorram & Li, 2022). MuseReason brings these ideas into music affective reasoning. It connects musical evidence, source affect, target affect, structured adjustment, and generated outcome assessment within a bidirectional framework. This allows the model to reason not only about what affect the music expresses, but also about how the music should change and whether the intended change is realised.

3 Theories and Methods

MuseReason connects music affect understanding and counterfactual editing through bidirectional affective reasoning. Given source music x, forward reasoning identifies musical evidence and infers its affective state. Given a target affect $a^\star$, inverse reasoning predicts musical adjustments Δ and preservation constraints p. A frozen music editor realises the predicted adjustment. The generated music is then measured and represented again to assess whether the intended affective change and source preservation are realised. Figure 2 gives an overview of MuseReason.

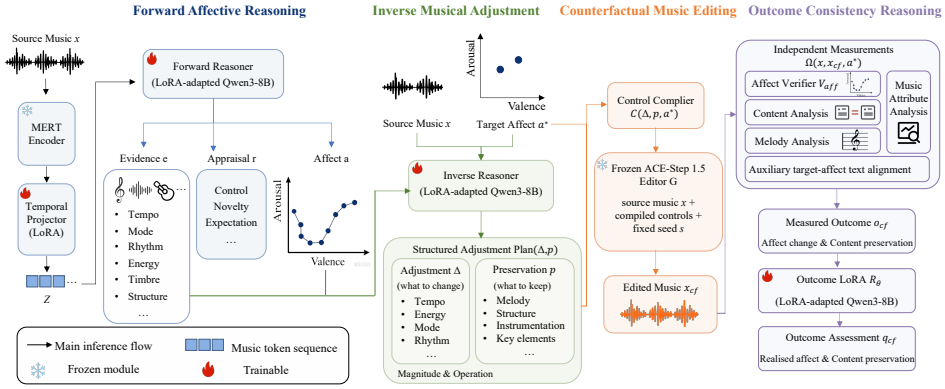


Figure 2: Overview of MuseReason. Forward reasoning infers musical evidence and source affect. Inverse reasoning predicts structured adjustments from the source and target affect. The generated music is then independently evaluated for outcome consistency.

3.1 Problem Formulation

Given source music $x \in \mathcal{X}$, we represent its observable musical evidence as

$$e = \left\{ e^{\text{tempo}}, e^{\text{mode}}, e^{\text{energy}}, e^{\text{rhythm}}, e^{\text{timbre}}, e^{\text{structure}} \right\}. \tag{1}$$

Its affective state is defined as

$$a = (v, q, d, \tau), \tag{2}$$

where v and q denote valence and arousal, d is an open vocabulary affective description, and $\tau = \{(v_t, q_t)\}_{t=1}^{T_a}$ represents temporal affect evolution. We use r to describe the appraisal that relates musical evidence to the affective judgement. For counterfactual editing, the target affect is represented as $a^\star = (v^\star, q^\star, d^\star)$, specifying the desired global valence, arousal, and affective description.

Given a target affect $a^\star$, MuseReason predicts structured musical adjustments Δ and preservation constraints p,

$$(x, e, a, a^\star) \to (\Delta, p). \tag{3}$$

The adjustment set is defined as $\Delta = \{\delta_j\}_{j=1}^{J}$, where

$$\delta_j = (\alpha_j, o_j, m_j, \mathcal{T}_j). \tag{4}$$

Here, α_j denotes the musical attribute, o_j the adjustment operation, m_j the magnitude or target value, and $\mathcal{T}_j$ an optional temporal region. The preservation constraints p specify source characteristics that should remain stable, such as melody, structure, or instrumentation.

The structured prediction is converted into controls supported by a frozen music editor,

$$c = C(\Delta, p, a^\star), \tag{5}$$

$$x_{\mathrm{cf}} = G(x, c, s), \tag{6}$$

where C is the control compiler, G is the frozen editor, and s is a fixed generation seed.

Multiple musical changes may lead towards the same target affect. MuseReason therefore does not assume a unique adjustment. The desired result approaches the target affect while preserving source characteristics that are not intended to change. We express this objective at the task level as

$$\min_{\Delta, p} D_{\mathrm{aff}}\left(V_{\mathrm{aff}}(x_{\mathrm{cf}}), a^\star\right) + \lambda_{\mathrm{pre}} D_{\mathrm{pre}}\left(x, x_{\mathrm{cf}}\right), \tag{7}$$

where V_{aff} is an independent affect verifier. D_{aff} measures target affect discrepancy and D_{pre} measures source preservation using the content and melody proxies defined in the evaluation protocol. The preservation set p at the plan level remains part of the editor controls. This objective describes the desired editing behaviour.

3.2 Bidirectional Affective Reasoning

MuseReason first extracts temporal music representations using a frozen music encoder E_{mus}. A Temporal Music Projector P_ϕ compresses the sequence of frame representations into temporal music tokens:

$$H = E_{\mathrm{mus}}(x). \tag{8}$$

The Temporal Music Projector P_ϕ divides the sequence of frame representations into temporal segments and obtains segment representations $\bar{H} = [\bar{h}_1, \ldots, \bar{h}_K]$. Learnable queries then aggregate these representations as

$$Z = \mathrm{Attn}(Q, \bar{H}, \bar{H}) W_p^\top = [z_1, \ldots, z_M]. \tag{9}$$

Compared with global pooling, Z retains information from different temporal regions and therefore supports both global affective understanding and affect analysis over time. The detailed projector architecture is provided in Appendix A.2.

Based on Z, MuseReason performs two reasoning processes conditioned on the task within a shared language model. In the forward direction,

$$F_{\theta_f}(Z) \to (\hat{e}, \hat{r}, \hat{a}), \tag{10}$$

where $\hat{e}$ denotes predicted musical evidence, $\hat{r}$ denotes the appraisal, and $\hat{a}$ denotes the predicted affective state and trajectory.

The inverse direction builds on the source representation and the inferred source state:

$$I_{\theta_i}(Z, \hat{e}, \hat{a}, a^\star) \to (\hat{\Delta}, \hat{p}). \tag{11}$$

Here, $\hat{\Delta}$ specifies the intended changes to musical attributes, whereas $\hat{p}$ identifies the source characteristics that should be preserved. F_{θ_f} and I_{θ_i} denote two complementary reasoning functions over the same temporal music representation, corresponding to forward affective reasoning and inverse musical adjustment, respectively. During training, reference evidence and affect annotations can be used as structured targets. At inference time, the inverse direction operates on the source state inferred by the forward direction. The structured reasoning interfaces are described in Appendix A.1, with training objectives in Appendix A.3.

3.3 Counterfactual Editing and Outcome Consistency

A structured adjustment is meaningful when it produces the intended change in the resulting audio. We therefore compile the inverse prediction into controls compatible with the editor and generate a counterfactual version of the source:

$$\hat{c} = C(\hat{\Delta}, \hat{p}, a^\star), \tag{12}$$

$$x_{\mathrm{cf}} = G(x, \hat{c}, s). \tag{13}$$

The resulting music is then independently evaluated. Its affective state is estimated as

$$\tilde{a}_{\mathrm{cf}} = V_{\mathrm{aff}}(x_{\mathrm{cf}}). \tag{14}$$

Content and melody similarities measure source preservation. Deterministic music analysis also records tempo, energy, rhythmic density, spectral brightness, and characteristics related to chroma after editing. These measurements separate the predicted adjustment from the change realised in audio.

Outcome consistency evaluates the realised musical result against the intended affective transition and the measured preservation criteria. Let

$$Z_{\mathrm{cf}} = P_\phi\left(E_{\mathrm{mus}}(x_{\mathrm{cf}})\right) \tag{15}$$

denote the temporal representation of the generated music. We summarise the independently measured outcome as follows:

$$o_{\mathrm{cf}} = \Omega(x, x_{\mathrm{cf}}, a^\star), \tag{16}$$

where o_{cf} records the realised affective change, content and melody preservation, and independent musical measurements.

We then define an outcome consistency reasoning function as:

$$R_{\theta_o}\left(Z_{\mathrm{cf}}, o_{\mathrm{cf}}, a^\star\right) \rightarrow q_{\mathrm{cf}}, \tag{17}$$

where q_{cf} summarises whether the intended affective transition and measured preservation criteria are supported by the generated music. This distinguishes the adjustment predicted by inverse reasoning from the change that is actually realised after editing.

The forward, inverse, and outcome consistency reasoning tasks are supervised through their corresponding structured targets, with detailed objectives provided in Appendix A.3.

4 Experiments and Results

4.1 Experimental Setup

Data. DEAM (Aljanaki et al., 2017) and PMEmo2019 (Zhang et al., 2018) provide static and temporal affect supervision and form the unified affect understanding evaluation. MERGE (Louro et al., 2026) provides additional static affect supervision. MusicQA (Liu et al., 2024) and MusicBench (Melechovsky et al., 2024) support music understanding and semantic alignment. MusicBench is also used to construct supervision for structured musical adjustment. Song Describer (Manco et al., 2023) is used only for music understanding evaluation. Dataset construction and preprocessing are described in Appendix B.

Implementation. MuseReason uses MERT-v1-330M (Li et al., 2024) as the frozen music encoder and Qwen3-8B (Yang et al., 2025) as the shared reasoning backbone. Task specific LoRA adapters are used for forward, inverse, and outcome reasoning. Counterfactual music is generated with the frozen ACE-Step 1.5 editor (Gong et al., 2026). Training settings and implementation details are given in Appendix C.

4.2 Affect Understanding Evaluation

Affect Understanding and Structured Reasoning. The experiment compares MuseReason with existing audio and music language models in terms of emotional understanding and structured output. We report Valence MAE (V-MAE) and Arousal MAE (A-MAE) for continuous emotion output. Quadrant Accuracy (Quad.), Descriptor F1 (Desc. F1), and Valence and Arousal Trend Macro F1 over three classes (V-Trend and A-Trend) are used to measure static and temporal emotion modelling ability. Table 1 shows the results. MuseReason performs strongly in continuous VA estimation and Arousal Trend. It also has a more comprehensive ability to produce structured output and supports continuous VA, numerical musical evidence, and time positioning. Compared with some baselines, MuseReason can unify emotional states, temporal changes, and musical evidence into a structured source state. More details of the baselines are procided in Appendix D.1

Table 1: Affect understanding and structured reasoning capability comparison. The classification and trend scores are reported in %. Values in parentheses denote coverage when below 100%.

	Affect Understanding						Structured Output Capability				
Method	**V-MAE↓**	**A-MAE↓**	**Quad.↑**	**Desc. F1↑**	**V-Trend↑**	**A-Trend↑**	**JSON↑**	**Schema↑**	**Cont. VA**	**Num. Evid.**	**Peak Time**
MuseReason (Ours)	**0.263**	**0.311**	46.40	52.00	29.92	**46.73**	**100.0**	**100.0**	✓	✓	✓
Qwen2.5-Omni-7B (Xu et al., 2025)	0.458 (98.8)	0.459 (98.8)	52.23 (98.8)	**54.33**	**42.81 (98.8)**	30.17 (98.8)	98.8	98.8	✓	✗	✗
Qwen2-Audio-7B-Instruct (Chu et al., 2024)	0.625 (88.4)	0.662 (88.4)	45.70 (88.4)	2.55	8.70 (22.8)	20.33 (17.6)	99.2	0.4	✓	✗	✗
Kimi-Audio-7B-Instruct (Ding et al., 2025)	0.508 (95.6)	0.503 (95.6)	**57.32 (95.6)**	50.92	22.54 (95.6)	23.06 (95.6)	95.6	90.4	✓	✗	✗
MusiLingo-MusicQA (Deng et al., 2024)	✗	✗	18.64 (47.2)	39.51	17.50 (41.6)	9.09 (14.8)	✗	✗	✗	✗	✗
Audio Flamingo 2 (Ghosh et al., 2025)	✗	✗	47.20	45.20	34.08	36.28	✗	✗	✗	✗	✗

Music Dependence of Affective Reasoning. The experiment aims to verify whether MuseReason's emotion reasoning depends on the corresponding musical content rather than being driven by linguistic priors or data distribution patterns. We use matched audio as the normal input. Mismatched audio retains the real music signal but disrupts the correspondence between the audio and emotional annotations, while zero audio further removes valid auditory information. The evaluation metrics cover emotional understanding abilities at different levels. Valence/Arousal MAE measures the continuous emotional state, Quadrant Accuracy and Descriptor F1 measure static emotional recognition, and Valence/Arousal Trend measures the modelling of emotional changes. Figure 3 shows that both audio disturbances lead to a decrease in performance. The impact of mismatched audio on static indicators is relatively limited, but there is a considerable decrease in Arousal Trend (31.6%). Under zero audio, Valence Trend and Arousal Trend performance dropped by 69% and 79.9%, respectively. The result indicates that MuseReason's emotional reasoning depends on musical content rather than being driven by linguistic priors or data distribution patterns. Further details are provided in Appendix D.2.

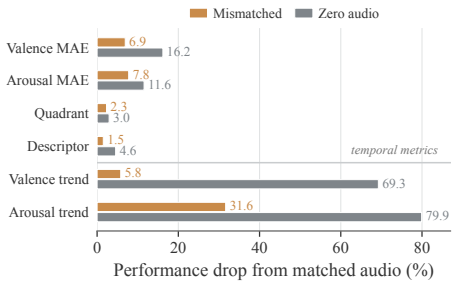


Figure 3: Music dependence of affective reasoning. Relative performance degradation under mismatched and zero audio conditions with respect to matched audio. Temporal affect metrics exhibit substantially larger degradation, indicating stronger dependence on the corresponding musical content.

4.3 Structured Musical Adjustment

The experiment evaluates the inverse reasoning module's structured musical adjustments and their dependence on the corresponding source audio. We use matched, mismatched, and zero audio evaluation conditions. The experiment is evaluated in terms of attribute selection, adjustment operation, adjustment magnitude, and preservation constraints. As shown in Table 2, matched audio performs better overall on all metrics, with Operation Accuracy reaching 68.75% and Amount MAE reduced to 46.48. These results indicate that MuseReason can utilise source music information to determine the attributes that need adjustment and how they should be adjusted. A complete comparison across DEAM, PMEmo2019, and MERGE Audio is reported in Appendix D.3.

Table 2: Inverse musical adjustment under matched, mismatched, and zero audio conditions. Attribute Precision (Attr. P), Recall (Attr. R), F1, Operation Accuracy, and Preservation F1 are reported in percentage (%). Amount MAE measures the error of the predicted adjustment magnitude and is reported in the original normalised scale.

Condition	Attr. P ↑	Attr. R ↑	Attr. F1 ↑	Operation ↑	Amount MAE ↓	Preserve F1 ↑
Matched audio	**92.67**	**40.56**	**50.12**	**68.75**	**46.48**	**9.04**
Mismatched audio	89.44	36.08	46.22	49.82	60.19	7.98
Zero audio	89.81	36.68	46.97	55.58	56.09	8.43

4.4 Counterfactual Music Editing

Method Comparison. This experiment compares MuseReason with editing methods based on reasoning and music generation systems on counterfactual music editing. The comparison focuses on affective change towards the target and preservation of the source music. Joint combines affective improvement with source preservation. Improve reports movement towards the target affect. Δ_{aff} measures affective change. Content and Melody measure source preservation. As shown in Table 3, MuseReason achieves a Joint Success of 68.77%. Its Content and Melody similarities are 0.8811 and 0.9924. M^{2}UGen obtains a larger Δ_{aff}, but preserves substantially less source content and melody. These results show that MuseReason maintains a favourable balance between affective change and source preservation. Candidate reranking based on the verifier over fixed candidates is reported in Appendix D.4.

Table 3: Counterfactual music editing comparison. Bold marks the best result. Underline marks the second best result.

Method	Joint ↑	Improve ↑	Δ_{aff} ↑	Content ↑	Melody ↑
MuseReason (ours)	**68.77%**	<u>68.77%</u>	<u>0.1468</u>	**0.8811**	**0.9924**
Direct Music LLM	20.45%	48.33%	0.0768	0.2650	0.9323
Qwen2.5-Omni-7B	20.45%	47.58%	0.0760	0.2664	0.9322
Instruct-MusicGen (Zhang et al., 2025)	39.41%	60.97%	0.0877	0.3406	0.9528
MusicGen-Melody (Copet et al., 2023)	4.46%	60.22%	0.1021	0.2527	0.8245
M^{2}UGen (Hussain et al., 2023)	36.06%	**80.67%**	**0.2675**	0.2791	0.8920

Controlled Method Comparison. This experiment compares different conditioning strategies. MuseReason derives explicit musical adjustments and preservation constraints from the source music and target affect. Reference Plan provides a structured diagnostic condition. Target Text and Target VA directly specify the desired affect. The results are shown in Table 4. MuseReason achieves higher Joint Success on all datasets while maintaining strong content and melody preservation. MuseReason also maintains positive affective change towards the target across all sets. These results show that structured adjustment derived from the source provides a more reliable conditioning strategy. MuseReason also achieves a better balance between affective change and preservation of the source music.

Table 4: Controlled counterfactual editing across different datasets. Joint and Improve are reported in %. Bold indicates the best result.

	DEAM					PMEmo2019					MERGE Audio				
Method	Joint↑	Imp.↑	Δ_{aff} ↑	Cont.↑	Mel.↑	Joint↑	Imp.↑	Δ_{aff} ↑	Cont.↑	Mel.↑	Joint↑	Imp.↑	Δ_{aff} ↑	Cont.↑	Mel.↑
MuseReason	**68.77**	**68.77**	**0.1468**	**0.8811**	**0.9924**	**72.54**	**72.54**	**0.2978**	**0.7539**	**0.9921**	**52.11**	**52.11**	**0.0211**	**0.7912**	**0.9896**
Reference Plan	20.82	48.33	0.0776	0.2664	0.9320	2.11	59.15	0.2861	0.0694	0.9537	6.34	47.18	-0.0102	0.1818	0.9551
Target Text	21.19	48.33	0.0765	0.2662	0.9322	2.11	58.45	0.2845	0.0689	0.9534	7.04	47.18	-0.0122	0.1811	0.9551
Target VA	20.45	47.96	0.0760	0.2665	0.9323	2.11	58.45	0.2856	0.0694	0.9534	7.04	47.89	-0.0133	0.1814	0.9553

Music Quality Evaluation. This experiment further compares the automatic quality of music generated by MuseReason and other controlled editing methods. We report Fréchet Audio Distance (FAD) based on MERT, Audiobox content enjoyment (CE), and production quality (PQ) (Tjandra et al., 2025), and the absolute loudness deviation $|\Delta\text{LUFS}|$. As shown in Table 5, MuseReason consistently obtains lower FAD across all sets. It also achieves higher CE and PQ scores and smaller loudness deviations. These results complement the content and melody preservation results. They

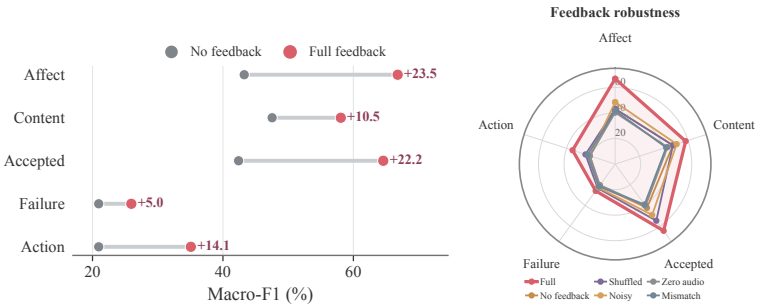
indicate that structured musical adjustment also reduces distributional distortion and quality degradation during editing.

Table 5: Music quality evaluation on different datasets. Bold indicates the best result.

Method	DEAM				PMEmo2019				MERGE Audio			
	FAD↓	CE↑	PQ↑	|ΔLUFS| ↓	FAD↓	CE↑	PQ↑	|ΔLUFS| ↓	FAD↓	CE↑	PQ↑	|ΔLUFS| ↓
MuseReason	**455.03**	**5.738**	**6.750**	**1.149**	**577.58**	**5.849**	**7.014**	**2.001**	**813.74**	**5.642**	**6.750**	**1.586**
Reference Plan	5836.62	2.823	5.278	2.964	4557.59	2.812	5.270	7.015	5254.57	2.845	5.353	4.180
Target Text	5864.23	2.830	5.264	2.973	4554.86	2.816	5.273	6.888	5288.39	2.852	5.362	4.193
Target VA	5864.54	2.830	5.261	2.994	4559.85	2.824	5.278	6.892	5280.94	2.856	5.362	4.163

4.5 Outcome Consistency Reasoning

Outcome Assessment Conditioned on Feedback. This experiment evaluates whether independent measurements of the generated music provide useful information for outcome assessment. It also examines whether the outcome reasoner depends on the measured feedback and the generated music. We compare full feedback with conditions in which the verification feedback is removed, shuffled, or perturbed. We also evaluate mismatched audio and zero audio conditions to separate the contributions of verification signals and generated music. Performance is measured using Macro-F1 for Affect, Preservation, Accepted, Failure, and Action. These metrics evaluate realised affect, satisfaction of the content and melody preservation criteria, acceptance decisions, failure modes, and subsequent actions, respectively. As shown in Figure 4 (a), full feedback improves all five outcome dimensions compared with no feedback, with particularly clear gains for Affect and Accepted. The robustness profile in Figure 4 (b) further shows that perturbing the feedback or generated audio consistently degrades outcome assessment. Zero audio and mismatched audio also lead to substantial degradation. The complete data and comparison across datasets are provided in Appendix D.5.



(a) Feedback contribution **(b) Feedback robustness**

Figure 4: Outcome assessment conditioned on feedback. (a) Macro-F1 scores (%) under full and no-feedback conditions; labels indicate the improvement from complete feedback. (b) Outcome profiles under full, perturbed, and audio ablated conditions.

Outcome Assessment over Multiple Rounds. This experiment examines whether repeated outcome assessment can further refine the judgement of a generated result when the generated music and independent verification remain fixed. Round 1 uses the original inputs after generation. Rounds 2-5 additionally condition on the structured outcome prediction from the preceding round. As shown in Figure 5, several metrics improve at intermediate rounds. Affect F1 and accepted rule consistency reach 70.48% and 82.78% at Round 3. Preservation, Accepted, and Mapping perform better at Round 4. Further increasing the number of rounds does not provide consistent gains. These results indicate that moderate assessment over multiple rounds can refine outcome judgements.

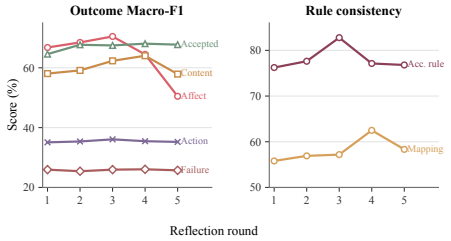


Figure 5: Outcome assessment over multiple rounds. Outcome Macro-F1 and rule consistency are reported. Several metrics improve at intermediate rounds, while additional rounds cannot provide consistent gains.

5 Discussion and Limitations

Bidirectional Affective Reasoning. The central idea of MuseReason is to connect affect understanding with music control oriented towards affect through an explicit reasoning process. Forward reasoning does not stop at an affect prediction. It represents the source music through musical evidence, affective appraisal, and temporal affect. Inverse reasoning then uses this source state together with the target affect to infer how the music should change. This formulation turns affect understanding into information that can directly support subsequent music adjustment. The affect understanding comparison shows that MuseReason maintains effective affect modelling while providing a more complete structured source state. The audio dependence and inverse adjustment experiments further show that both reasoning directions use information from the corresponding music. These findings support the role of musical evidence as the connection between understanding the current affect and reasoning towards a desired affective state.

Structured Musical Adjustment and Counterfactual Editing. MuseReason separates affective intent from audio execution through a structured adjustment plan. The adjustment specifies what should change, while the preservation specifies what should remain stable. This distinction is important for counterfactual editing. Reaching a target affect is not sufficient if the resulting music loses the identity of the source. The editing experiments reflect this trade off. MuseReason achieves substantially stronger joint editing success and source preservation. The controlled comparison further supports the value of deriving adjustments from the source state rather than directly providing a target condition. The music quality results also show a consistent pattern. These results suggest that MuseReason provides a meaningful interface between affective reasoning at a high level and music editing.

Consistency Between Reasoning and Realised Outcomes. MuseReason distinguishes an intended adjustment from its realised outcome. A structured plan describes the intended musical change. It does not confirm that the generated music reaches the target affect. MuseReason therefore evaluates the generated music before making the final outcome judgement. Independent measurements quantify affective improvement and source preservation. The generated music representation captures the resulting musical content. The consistency reasoner uses this evidence to compare the adjustment plan with the realised audio. Feedback removal, perturbation, and audio mismatch all reduce assessment performance. This result confirms the contribution of both measured outcomes and the corresponding audio. Assessment over multiple rounds provides further gains at intermediate rounds. The preceding structured judgement supplies additional context for the next assessment. Outcome consistency reasoning thus connects intended adjustment, realised audio, and verification after editing.

Limitations. The current evaluation mainly relies on independent measurements of emotion, content, melody, and musical attributes. Future work can further investigate, through listener studies, the consistency between the model-inferred musical evidence, the adjustment plan, and subjective emotional perception. The musical genres, emotion representations, and evaluation sets covered by the current experiments are still limited. Broader cultural contexts and open-ended emotion concepts remain to be further validated.

6 Conclusion

This paper proposed MuseReason, a bidirectional emotional reasoning framework for music emotion understanding and counterfactual editing. MuseReason performed forward emotional reasoning from verifiable musical evidence, and combined the source music state with the target emotion to predict structured musical adjustments and preservation constraints, thereby linking emotional understanding with goal directed music editing. Experiments showed that real musical evidence played an important role in dynamic emotional reasoning and music adjustment, that structured adjustment achieved a more reasonable balance between target emotional change and source music preservation, and that feedback from generated results further strengthened consistency assessment.

AI USE STATEMENT

In this work, we used generative AI tools to support literature discovery, error diagnosis, and data processing checks. We did not use generative AI tools as sources of scientific evidence or to produce the reported experimental measurements. Tasks involving theoretical proof construction and human participant analysis were not applicable to this work. Generative AI tools also supported reference formatting, grammar checking, and presentation refinement. All cited literature was checked against the original publications. The authors manually examined the experimental configurations, data processing procedures, model outputs, and reported numerical results. We reviewed all AI assisted work and take full responsibility for the final content, claims, and research artefacts presented in this paper.

REFERENCES

Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. Developing a benchmark for emotional analysis of music. *PLOS ONE*, 12(3):e0173392, 2017. doi: 10.1371/journal.pone.0173392. URL https://doi.org/10.1371/journal.pone.0173392.

James Armitage, Tuomas Eerola, and Andrea R. Halpern. Play it again, but more sadly: Influence of timbre, mode, and musical experience in melody processing. *Memory & Cognition*, 53(3): 869–880, 2025. doi: 10.3758/s13421-024-01614-8.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47704–47720, 2023. doi: 10.52202/075280-2066.

Zihao Deng, Yinghao Ma, Yudong Liu, Rongchen Guo, Ge Zhang, Wenhu Chen, Wenhao Huang, and Emmanouil Benetos. 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*, pp. 3643–3655, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.231. URL https://aclanthology.org/2024.findings-naacl.231/.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Lucas N. Ferreira, Lili Mou, Jim Whitehead, and Levi H. S. Lelis. Controlling perceived emotion in symbolic music generation with monte carlo tree search. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pp. 163–170, 2022. doi: 10.1609/aiide.v18i1.21960.

Hortense Fong, Vineet Kumar, and K. Sudhir. A theory-based explainable deep learning architecture for music emotion. *Marketing Science*, 44(1):196–219, 2025. doi: 10.1287/mksc.2022.0323.

Joshua P. Gardner, Simon Durand, Daniel Stoller, and Rachel M. Bittner. LLark: A multimodal instruction-following language model for music. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15037–15082. PMLR, 2024. URL https://proceedings.mlr.press/v235/gardner24a.html.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19358–19405. PMLR, 2025. URL https://proceedings.mlr.press/v267/ghosh25b.html.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sang-gil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. Music flamingo: Scaling music understanding in audio language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10009489.

Junmin Gong, Yulin Song, Wenxiao Zhao, Sen Wang, Shengyuan Xu, and Jing Guo. ACE-Step 1.5: Pushing the boundaries of open-source music generation. *arXiv preprint arXiv:2602.00744*, 2026. URL https://arxiv.org/abs/2602.00744.

Yiming Gu, Chen Shao, Jingze Li, and Yinghan Fan. Interpretable deep learning reveals distinct spectral and temporal drivers of perceived musical emotion. *Scientific Reports*, 16:4205, 2026. doi: 10.1038/s41598-025-34238-2.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jingyue Huang, Ke Chen, and Yi-Hsuan Yang. Emotion-driven piano music generation via two-stage disentanglement and functional representation. In *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 335–342, 2024.

Atin Sakkeer Hussain, Shansong Liu, Chenshuo Sun, and Ying Shan. M^2UGen: Multi-modal music understanding and generation with the power of large language models. *arXiv preprint arXiv:2311.11255*, 2023. URL https://arxiv.org/abs/2311.11255.

Saeed Khorram and Fuxin Li. Cycle-consistent counterfactuals by latent transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10193–10202, 2022.

Harin Lee, Elif Çelen, Peter M. C. Harrison, Manuel Anglada-Tort, Pol van Rijn, Minsu Park, Marc Schönwiesner, and Nori Jacoby. GlobalMood: A cross-cultural benchmark for music emotion recognition. In *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 11–19, 2025.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger B. Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. MERT: Acoustic music understanding model with large-scale self-supervised training. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=w3YZ9MSlBu.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. Music understanding LLaMA: Advancing text-to-music generation with question answering and captioning. In *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 286–290. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10447027. URL https://doi.org/10.1109/ICASSP48485.2024.10447027.

Pedro L. Louro, Hugo Redinho, Tiago F. R. Ribeiro, Ricardo Santos, Ricardo Malheiro, Renato Panda, and Rui Pedro Paiva. MERGE: A bimodal audio-lyrics dataset for static music emotion recognition. *IEEE Transactions on Affective Computing*, pp. 1–18, 2026. doi: 10.1109/TAFFC.2026.3668037. URL https://doi.org/10.1109/TAFFC.2026.3668037. Early Access.

Ilaria Manco, Benno Weck, SeungHeon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, Elio Quinton, György Fazekas, and Juhan Nam. The song describer dataset: A corpus of audio captions for music-and-language evaluation. In *Machine Learning for Audio Workshop at NeurIPS 2023*, 2023. URL https://neurips.cc/virtual/2023/79724.

Zhuoyuan Mao, Mengjie Zhao, Qiyu Wu, Hiromi Wakaki, and Yuki Mitsufuji. DeepResonance: Enhancing multimodal music understanding via music-centric multi-way instruction tuning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12926–12948, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.653. URL https://aclanthology.org/2025.emnlp-main.653/.

Jiahao Mei, Xuenan Xu, Zeyu Xie, Zihao Zheng, Ye Tao, Yue Ding, and Mengyue Wu. LARA-Gen: Enabling continuous emotion control for music generation models via latent affective representation alignment. *arXiv preprint arXiv:2510.05875*, 2025. URL https://arxiv.org/abs/2510.05875.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. Mustango: Toward controllable text-to-music generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8293–8316, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.459. URL https://aclanthology.org/2024.naacl-long.459/.

Andreas Peintner, Marta Moscati, Yu Kinoshita, Richard Vogl, Peter Knees, Markus Schedl, Hannah Strauss, Marcel Zentner, and Eva Zangerle. Nuanced music emotion recognition via a semi-supervised multi-relational graph neural network. *Transactions of the International Society for Music Information Retrieval*, 8(1):140–153, 2025. doi: 10.5334/tismir.235.

Or Tal, Alon Ziv, Itai Gat, Felix Kreuk, and Yossi Adi. Joint audio and symbolic conditioning for temporally controlled text-to-music generation. In *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 264–271, 2024.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025. URL https://arxiv.org/abs/2502.05139.

Fang-Duo Tsai, Shih-Lun Wu, Haven Kim, Bo-Yu Chen, Hao-Chung Cheng, and Yi-Hsuan Yang. Audio prompt adapter: Unleashing music editing abilities for text-to-music with lightweight finetuning. In *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 634–641, 2024.

Fang-Duo Tsai, Shih-Lun Wu, Weijaw Lee, Sheng-Ping Yang, Bo-Rui Chen, Hao-Chung Cheng, and Yi-Hsuan Yang. MuseControlLite: Multifunctional music generation with lightweight conditioners. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60266–60279. PMLR, 2025. URL https://proceedings.mlr.press/v267/tsai25a.html.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2): 841–887, 2018.

Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33577–33587, 2026.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. URL https://arxiv.org/abs/2503.20215.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang,

Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Kejun Zhang, Hui Zhang, Simeng Li, Changyuan Yang, and Lingyun Sun. The PMEmo dataset for music emotion recognition. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, ICMR '18, pp. 135–142, New York, NY, USA, 2018. Association for Computing Machinery. doi: 10.1145/3206025.3206037. URL https://doi.org/10.1145/3206025.3206037.

Yixiao Zhang, Yukara Ikemiya, Woosung Choi, Naoki Murata, Marco A. Martínez-Ramírez, Liwei Lin, Gus Xia, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. Instruct-MusicGen: Unlocking text-to-music editing for music language models via instruction tuning. In *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 328–336, 2025. doi: 10.5281/zenodo.17811375.

Haina Zhu, Yizhi Zhou, Hangting Chen, Jianwei Yu, Ziyang Ma, Rongzhi Gu, Yi Luo, Wei Tan, and Xie Chen. MuQ: Self-supervised music representation learning with mel residual vector quantization. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3653–3664, 2025. doi: 10.1109/TASLPRO.2025.3602320. URL https://doi.org/10.1109/TASLPRO.2025.3602320.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017. doi: 10.1109/ICCV.2017.244.

A Additional Method Details

A.1 Structured Reasoning Interfaces

MuseReason exposes structured variables at the transitions from music perception to affective interpretation, from source state to musical adjustment, and from generated audio to outcome assessment. The interface keeps observations, appraisals, intended actions, editor controls, and measured outcomes in distinct fields. Table 6 summarises their roles and provenance.

Table 6: Roles and provenance of the intermediate objects used by MuseReason.

Object	Role	Source	Interpretation
$e, \hat{e}$	Musical evidence	Audio-derived reference or forward prediction	Reference source properties and their predicted counterparts
$r, \hat{r}$	Affective appraisal	Reference target or joint forward prediction	Interpretation based on evidence that relates source cues to perceived affect
$a, \hat{a}$	Source affect	Affect annotation or forward prediction	Reference source affect and its predicted counterpart
$a^\star$	Target affect	Editing request	Desired global VA state and affect descriptor
$\Delta, \hat{\Delta}, p, \hat{p}$	Adjustment and preservation	Reference supervision or inverse prediction	Reference and predicted musical adjustments and preservation constraints
$c, \hat{c}$	Editor control	Deterministic compiler	Controls compiled from reference or predicted adjustment plans
o_{cf}	Measured outcome	Independent measurements after editing	Realised affect change, content and melody preservation, and musical attributes
q_{cf}	Outcome assessment	Outcome reasoner	Structured judgement of the requested transition and measured result

Figure 6 summarises the inputs and predicted records of the three reasoning tasks. Independent outcome measurements are supplied to the outcome reasoner as inputs. Assessment decisions are its outputs. The detailed field specifications follow below.

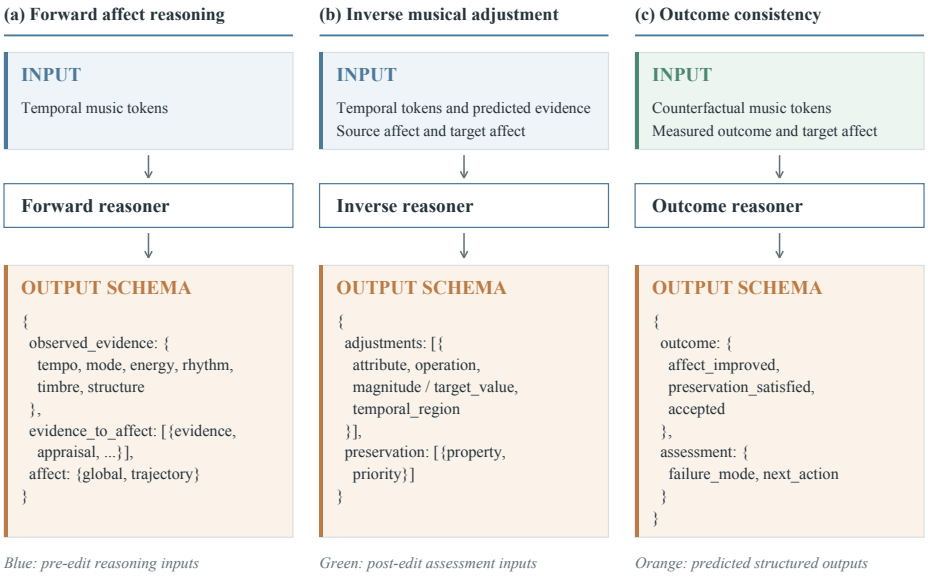


Figure 6: Structured reasoning interfaces. Each panel shows task inputs, the corresponding reasoner, and a compact output schema sketch. Source inputs, generated-audio evidence, and predicted structured outputs are distinguished by colour.

Forward Affect Reasoning. The forward interface instantiates $F_{\theta_f}(Z) \rightarrow (\hat{e}, \hat{r}, \hat{a})$. Its schema reports musical observations, appraisal based on evidence, global affect, and temporal affect in separate fields. The schema sketches in this section specify semantic fields. Braced field lists are notation for record structure.

```
Analyze the supplied music. Report audibly verifiable musical evidence.
Provide a concise appraisal that relates the evidence to perceived affect.
Estimate global valence and arousal. When temporal supervision is available,
summarize valence and arousal evolution. Return one JSON object following
FORWARD_SCHEMA.

FORWARD_SCHEMA = {
  observed_evidence: {tempo, mode, energy, rhythm, timbre, structure},
  evidence_to_affect: [{evidence, appraisal, supports, qualification}],
  affect: {global: {valence, arousal, description},
           trajectory: {available, time_points, valence, arousal,
                        valence_trend, arousal_trend,
                        start, end, peak_arousal}},
  evidence_boundary: {audio_derived, annotation_derived, model_inferred}
}
```

The `evidence_boundary` field records the provenance of each quantity. Static affect datasets set `trajectory.available=false`. Temporal supervision is drawn from datasets with annotations that vary over time. Aligned time points and VA values represent τ in the main text. Trend, endpoint, and peak fields summarise that sequence. Provenance metadata distinguishes reference annotations from model-inferred evidence and appraisal.

Inverse Musical Adjustment. The inverse interface instantiates $I_{\theta_i}(Z, \hat{e}, \hat{a}, a^\star) \rightarrow (\hat{\Delta}, \hat{p})$. The target follows the static representation $a^\star = (v^\star, q^\star, d^\star)$ defined in the main text. The interface represents the transition from source to target through structured adjustments and a separate preservation set. The main text defines each adjustment as $(\alpha_j, o_j, m_j, \mathcal{T}_j)$. The serialised record also stores the unit when a numeric magnitude is used.

```
Given SOURCE_MUSIC, SOURCE_STATE, and TARGET_AFFECT, infer a compact musical
adjustment plan. Each adjustment specifies an attribute, operation, magnitude
or target value, and an optional temporal region. State preservation
constraints separately. Report every adjustment as an intended action.
Return one JSON object following INVERSE_SCHEMA.

INVERSE_SCHEMA = {
  source_state: {affect, evidence_reference},
  target_affect: {valence, arousal, description},
  adjustments: [{attribute, operation, magnitude, unit, target_value,
                 temporal_region, rationale}],
  preservation: [{property, priority}],
  plan_status: {executable}
}
```

The rationale for an adjustment links the selected operation to source evidence and its intended contribution to the target affect. Multiple plans can satisfy the same target and preservation requirements.

Outcome Consistency Reasoning. The outcome interface instantiates $R_{\theta_o}(Z_{\mathrm{cf}}, o_{\mathrm{cf}}, a^\star) \to q_{\mathrm{cf}}$. It receives the generated music representation together with measurements taken after editing and returns a structured assessment.

```
Assess the generated counterfactual music using TARGET_AFFECT and
MEASURED_OUTCOME. Judge affect improvement, preservation satisfaction
based on content and melody, acceptance, failure mode, and next action.
Return one JSON object following OUTCOME_SCHEMA.

MEASURED_OUTCOME = {
  observed_post_edit_affect: {valence, arousal},
  target_distance_improvement,
  content_similarity, melody_similarity,
  post_edit_musical_attributes: {
    tempo, energy, rhythm_density, spectral_brightness, chroma_summary
  }
}
OUTCOME_SCHEMA = {
  outcome: {affect_improved, preservation_satisfied, accepted},
  assessment: {failure_mode, next_action}
}
```

The measurement record is produced by Ω and supplied to the outcome reasoner, as in the main text. The musical attribute fields are independent measurements taken after editing that provide additional evidence for the assessment. Affect, Preservation, Accepted, Failure, and Action evaluate the five corresponding decision fields. Here q_{cf} denotes the assessment record, while q in $a = (v, q, d, \tau)$ denotes arousal. The `next_action` field is an outcome label. The reported assessment over multiple rounds keeps the generated music and independent measurements fixed while conditioning later rounds on the preceding structured assessment.

A.2 Temporal Music Projector

Given the representation at the frame level

$$H = [h_1, \ldots, h_T], \qquad h_t \in \mathbb{R}^{d_m}, \tag{18}$$

we divide the sequence into temporal segments $\{\mathcal{S}_k\}_{k=1}^{K}$ and obtain

$$\bar{h}_k = \frac{1}{|\mathcal{S}_k|} \sum_{t \in \mathcal{S}_k} h_t + p_k, \tag{19}$$

where p_k encodes temporal position. Learnable queries aggregate the segmented sequence through cross attention,

$$Q^{(\ell+1)} = Q^{(\ell)} + \mathrm{MHA}\left(Q^{(\ell)}, \bar{H}, \bar{H}\right), \tag{20}$$

where $\bar{H} = [\bar{h}_1, \ldots, \bar{h}_K]$. The resampled queries are projected to the language model embedding space,

$$Z = Q^{(L)} W_p^\top \in \mathbb{R}^{M \times d_{\mathrm{LLM}}}. \tag{21}$$

Here queries are stacked as rows, and $W_p \in \mathbb{R}^{d_{\mathrm{LLM}} \times d_m}$ applies the tokenwise projection. This is the expanded form of the projector operation used in the main text. The implementation uses $M = 32$ temporal music tokens. During music language alignment, the music encoder and language model remain fixed while the Temporal Music Projector is optimised. The frame level MERT representation has dimension $d_m = 1024$. The projector contains two attention layers with eight heads and produces $M = 32$ music tokens. These dimensions are shared by the alignment and downstream reasoning stages so that the same temporal interface is retained throughout the curriculum.

A.3 Learning Objectives and Curriculum

Each reasoning task uses autoregressive supervision over its structured target. Let $u^{(k)}$ denote the task conditioning variables and $y^{(k)}$ the serialised target for task k. The shared objective form is

$$\mathcal{L}_k = -\mathbb{E}\left[\sum_t \log p_{\theta_k}\left(y_t^{(k)} \mid u^{(k)}, y_{<t}^{(k)}\right)\right]. \tag{22}$$

For music language alignment, $u^{(\mathrm{align})} = Z$ and the optimisation updates the Temporal Music Projector. Forward reasoning uses $y^{(\mathrm{fwd})} = \mathrm{Serialize}(e, r, a)$. Inverse reasoning uses $y^{(\mathrm{inv})} = \mathrm{Serialize}(\Delta, p)$ conditioned on $(Z, e, a, a^\star)$. Outcome consistency uses $y^{(\mathrm{out})} = \mathrm{Serialize}(q_{\mathrm{cf}})$ conditioned on $(Z_{\mathrm{cf}}, o_{\mathrm{cf}}, a^\star)$.

Training proceeds from music language alignment to forward affective reasoning and inverse musical adjustment. Counterfactual audio produced by the frozen editor is then measured to construct outcome supervision. The shared Qwen3 backbone uses separate LoRA adapters for forward, inverse, and outcome reasoning.

A.4 Counterfactual Execution and Outcome Measurement

Control Compilation and Editing. The compiler translates $(\Delta, p, a^\star)$ into controls supported by the frozen editor. Global adjustments use editing conditioned on the source. Optional temporal regions specify the scope of an adjustment through the controls supported by the editor. The compiled record can include text instructions, tempo or key conditions, temporal scope, and source preservation strength. The fixed generation seed s is passed separately to $G(x, c, s)$, as in the main text.

Independent Outcome Measurements. Generated music is analysed after editing. The outcome record o_{cf} combines independently measured affective change, content and melody preservation, and musical attributes measured after editing. These measurements support both counterfactual evaluation and outcome consistency reasoning. Concrete measurement modules are described in Appendix C, and metric definitions are provided in Appendix D.

A.5 Procedural Algorithm

Algorithm 1 expands the inference functions defined in the main text. Source evidence and affect are predicted by the forward reasoner before inverse adjustment. All model parameters are fixed during inference. The task level editing objective and supervised training objectives are distinct from this execution procedure.

Algorithm 1: Evidence-conditioned counterfactual music editing

Input: Source music x, target affect $a^\star$, fixed generation seed s; trained reasoning modules and fixed encoder, editor, and measurement modules.

1. Encode source music: $Z \leftarrow P_\phi(E_{\mathrm{mus}}(x))$.
2. Predict evidence, appraisal, and source affect: $(\hat{e}, \hat{r}, \hat{a}) \leftarrow F_{\theta_f}(Z)$.
3. Infer adjustment and preservation: $(\hat{\Delta}, \hat{p}) \leftarrow I_{\theta_i}(Z, \hat{e}, \hat{a}, a^\star)$.
4. Compile editor controls: $\hat{c} \leftarrow C(\hat{\Delta}, \hat{p}, a^\star)$.
5. Generate edited music: $x_{\mathrm{cf}} \leftarrow G(x, \hat{c}, s)$.
6. Measure the realised result: $o_{\mathrm{cf}} \leftarrow \Omega(x, x_{\mathrm{cf}}, a^\star)$.
7. Encode generated music: $Z_{\mathrm{cf}} \leftarrow P_\phi(E_{\mathrm{mus}}(x_{\mathrm{cf}}))$.
8. Assess outcome consistency: $q_{\mathrm{cf}} \leftarrow R_{\theta_o}(Z_{\mathrm{cf}}, o_{\mathrm{cf}}, a^\star)$.

Output: x_{cf}, $(\hat{\Delta}, \hat{p})$, o_{cf}, and q_{cf}.

A.6 Reasoning Trace and Structured Output Validation

Representative Reasoning Trace. Table 7 illustrates how the structured objects connect across one transition from source to target. The values are rounded for presentation.

Table 7: Representative reasoning trace from source evidence to measured outcome.

Source evidence and state	Target affect	Structured adjustment	Preservation	Measured outcome
Tempo 140.6 BPM; moderate rhythmic activation; low brightness; source $(-0.04, -0.01)$ VA	Calm and warm; target $(0.70, -0.60)$ VA	Decrease tempo by 18 BPM; reduce rhythmic activation; shift towards a warmer, brighter, major mode character	Main melody and global structure with high priority; instrumentation with medium priority	$\Delta_{\mathrm{aff}} = 0.085$; content $= 0.580$; melody $= 0.919$; accepted

The compiler converts the structured adjustment into editor controls. Independent measurements after generation then determine the realised affect change and source preservation used by the outcome reasoner.

Structured Output Validation. Structured outputs are checked at four levels. Syntactic validity requires one parseable JSON object. Structural validity checks the expected fields and value types. Semantic validity constrains VA values, operations, magnitudes, and temporal intervals to their defined domains. Consistency across fields checks relations among measured values and decision fields, including the acceptance rule. These checks support the JSON and Schema metrics reported in the main paper.

Each target is serialised as a single JSON object under the schema for the task. The model prompt requests JSON-only output and disables auxiliary reasoning text. JSON validity is computed over the complete evaluation set. Task metrics at the field level are computed only when the corresponding value can be parsed and passes its type and domain checks. Output coverage is therefore reported alongside baseline results whenever parsing is incomplete.

B Data and Supervision Details

B.1 Affective Reasoning Data

DEAM (Aljanaki et al., 2017), PMEmo2019 (Zhang et al., 2018), and MERGE (Louro et al., 2026) provide affect supervision with different temporal coverage.

DEAM. contains 1,802 valid tracks with VA annotations at the track level and over time. Scores at the track level are mapped from the original 1 to 9 scale into $[-1, 1]$. Dynamic annotations are

retained at the original 0.5 second resolution beginning at 15 seconds. Four exact duplicate audio files are removed after hashing. Observable evidence is extracted from the audio, including tempo, RMS energy, onset density, brightness, and characteristics related to chroma.

PMEmo2019. is combined with DEAM for complete affect analysis. Together they contain 2,569 unique music sources in the initial reasoning corpus. The complete analysis stage uses 2,079 training instructions, 240 validation instructions, and an isolated 250 sample test set.

MERGE. provides 3,554 tracks with VA annotations at the song level in $[0, 1]$, mapped linearly into $[-1, 1]$. After data checks, 3,545 tracks remain and yield 7,090 evidence extraction and static appraisal instances. MERGE contributes static affect supervision. Temporal affect supervision comes from DEAM and PMEmo2019.

B.2 Music Language Alignment and Evaluation Data

MusicQA (Liu et al., 2024) and MusicBench (Melechovsky et al., 2024) establish the correspondence between temporal music representations and language before affective reasoning.

MusicQA. contains 13,102 audio files, with 42,867 pretraining question answer pairs over 4,763 tracks, 70,011 finetuning pairs over 7,779 tracks, and 5,040 evaluation pairs over 560 tracks. The official evaluation partition is reserved for evaluation.

MusicBench. provides music descriptions and attributes such as BPM, key, chord, and beat. It also contains augmented audio produced by tempo, pitch, and volume modifications. Variants derived from a common source are grouped when constructing data partitions. The standardised alignment corpus contains 333,075 instructions, with 308,275 training examples, 16,736 validation examples, and 8,064 evaluation examples.

Song Describer. (Manco et al., 2023) is reserved for natural language music understanding evaluation. We use its manually validated subset with 746 descriptions over 547 tracks.

B.3 Counterfactual Editing Supervision

Counterfactual supervision combines 3,412 known edit pairs from MusicBench with target affect plans constructed from DEAM. The MusicBench pairs supervise adjustment type and direction over the available transformation set. The DEAM source set yields 3,246 target affect plans over 1,623 unique tracks. Each plan is executed with four fixed random seeds, producing 12,984 candidate counterfactual samples using the frozen ACE-Step 1.5 editor (Gong et al., 2026).

B.4 Partition Integrity and Supervision Construction

Partition Integrity. Dataset partitions are assigned at the level of the source music before prompts, task variants, or counterfactual candidates are expanded. All records derived from the same source identifier remain in one partition, and manifest validation rejects any source identifier associated with multiple partitions. DEAM uses an $80/10/10$ split stratified by quadrant. PMEmo2019 uses an $80/10/10$ split that is disjoint by artist and stratified by quadrant. MERGE follows its official $70/15/15$ training, validation, and test partition. MusicQA retains the official evaluation partition, while MusicBench groups all transformations derived from a common source before partitioning. Song Describer is used only for evaluation.

Supervision Construction. Forward targets combine listener affect annotations with deterministic music information retrieval measurements. Static examples supervise global VA and affective descriptors. Temporal DEAM and PMEmo2019 annotations additionally supervise trajectory summaries, including start and end states, valence and arousal trends in three classes, and peak arousal localisation. MERGE contributes static supervision only. Observable evidence includes tempo, energy, rhythmic density, spectral brightness, and dominant pitch or chroma characteristics. The evidence-to-affect field records an appraisal linking these cues to the annotated affect while explicitly identifying measured, annotated, and model inferred quantities.

Inverse targets are represented as adjustment and preservation pairs. MusicBench edit pairs provide supervision for available transformation types and directions. DEAM target affect pairs provide plans from source to target. Candidate audio and its independent measurements are constructed only from training and validation sources when forming outcome consistency supervision. Test music held out for evaluation and its derived candidates are excluded from all trainable stages.

C Implementation Details

C.1 Model Configuration

MuseReason uses MERT-v1-330M (Li et al., 2024) as the frozen music encoder. The Temporal Music Projector produces 32 temporal music tokens for Qwen3-8B (Yang et al., 2025). Forward, inverse, and outcome reasoning share the Qwen3 backbone and use LoRA (Hu et al., 2022) adapters specific to each task.

C.2 Optimisation Configuration

All trainable stages use AdamW with weight decay 0.01, gradient clipping at 0.5, four data loading workers per process, and random seed 42. Training is performed in full precision. The LoRA adapters have rank 16, scaling factor 32, and dropout 0.05, and are applied to the query, key, value, and output projections of the language-model attention layers.

For inverse adjustment, the five epoch task training is followed by a two epoch schema and audio conditioning refinement. Its language model learning rate is reduced from 5×10^{-6} to 3×10^{-6}, while the projector learning rate remains 1×10^{-5}. In outcome consistency training, the projector uses 2×10^{-6} while the LoRA parameters use 3×10^{-6}.

C.3 Trainable and Fixed Components

The trainable parameters are

$$\Theta_{\text{MuseReason}} = \{\phi, \theta_f, \theta_i, \theta_o\}, \tag{23}$$

where ϕ denotes the Temporal Music Projector and θ_f, θ_i, and θ_o denote the adaptation parameters specific to each task. The pretrained music encoder and music editor remain frozen,

$$\nabla_{\theta_{E_{\text{mus}}}} \mathcal{L} = 0, \qquad \nabla_{\theta_G} \mathcal{L} = 0. \tag{24}$$

Independent verification models are trained separately and remain fixed during counterfactual editing and outcome consistency training.

C.4 Execution and Measurement Modules

Counterfactual audio is generated with the frozen ACE-Step 1.5 editor (Gong et al., 2026). Independent measurements use the affect verifier based on MuQ, content and melody similarity, and deterministic music analysis of tempo, energy, rhythmic density, spectral brightness, and characteristics related to chroma. MuQ-MuLan (Zhu et al., 2025) is used as an auxiliary measure of text alignment with the target affect. The music encoder, editor, and verification modules remain fixed during the corresponding reasoning stages.

Affect Verifier Validation. The affect verifier V_{aff} is trained independently on frozen MuQ (Zhu et al., 2025) representations and evaluated on DEAM annotations reserved for evaluation. It obtains a trajectory CCC of 0.7572 and a trajectory MAE of 0.1383. The verifier is then fixed for counterfactual evaluation and outcome supervision.

D Experimental Protocols and Extended Results

D.1 Baselines and Control Conditions

Affect Understanding Baselines. For affect understanding, we compare MuseReason with Qwen2.5-Omni-7B, Qwen2-Audio-7B-Instruct, Kimi-Audio-7B-Instruct, MusiLingo-MusicQA,

and Audio Flamingo 2. Continuous VA metrics are reported when a model interface produces native continuous values. Categorical outputs from MusiLingo and Audio Flamingo 2 are evaluated on the supported discrete affect tasks. Output coverage is reported when a model does not return a valid prediction for every sample.

Counterfactual Editing Baselines. For counterfactual editing, Direct Music LLM produces an instruction in free form from the source music and target affect. Target Text and Target VA provide direct target conditions to the same frozen editor used by MuseReason. Reference Plan supplies a structured diagnostic condition. The main method comparison also includes Instruct MusicGen, MusicGen-Melody, and M^{2}UGen with their native generation interfaces.

Source Dependence Controls. Matched, mismatched, and zero audio conditions are used for music dependence analysis. Mismatched audio preserves a real music signal while breaking the source sample correspondence. Zero audio removes effective auditory content. The same conditions are applied to inverse adjustment when testing source dependence.

D.2 Affect Understanding and Music Dependence

Metrics and Output Coverage. Continuous affect is evaluated with Valence MAE and Arousal MAE. Quadrant Accuracy and Descriptor F1 measure static affect recognition. Valence Trend and Arousal Trend use Macro F1 over three classes: falling, stable, and rising trajectories. Structured output evaluation reports JSON validity, Schema validity, support for continuous VA, numerical musical evidence, and peak time localisation. Coverage is shown when a model produces valid outputs for fewer than all evaluation samples.

Music Dependence Protocol. Music dependence is quantified by comparing matched audio with mismatched and zero audio controls on the same examples. Table 1 reports affect-understanding performance, and Figure 3 summarises the relative performance degradation under audio interventions.

Results for Individual Datasets. Table 8 reports the unified affect evaluation separately on DEAM and PMEmo2019. Coverage markers identify models with incomplete outputs for at least one metric.

Table 8: Dataset specific affect understanding results. All scores except MAE are reported in %. Bold indicates the best result within each dataset and metric. † denotes models with incomplete output coverage for at least one reported metric.

	DEAM						PMEmo2019					
Method	**V-MAE↓**	**A-MAE↓**	**Quad.↑**	**Desc. F1↑**	**V-Trend↑**	**A-Trend↑**	**V-MAE↓**	**A-MAE↓**	**Quad.↑**	**Desc. F1↑**	**V-Trend↑**	**A-Trend↑**
MuseReason (Ours)	**0.256**	**0.279**	36.31	42.74	30.29	**45.86**	**0.282**	**0.392**	71.83	**75.35**	**28.80**	**44.61**
Qwen2.5-Omni-7B	0.458	0.482	45.45	**49.15**	**49.16**	34.85	0.456	0.404	69.01	67.18	25.39	18.12
Qwen2-Audio-7B-Instruct†	0.682	0.741	35.71	3.01	8.77	21.04	0.493	0.481	68.66	1.41	8.33	18.18
Kimi-Audio-7B-Instruct	0.521	0.543	**49.40**	44.86	28.38	26.26	0.476	0.408	**76.06**	65.26	6.97	16.13
MusiLingo-MusicQA†	✗	✗	10.61	32.65	16.87	0.00	✗	✗	28.85	56.52	21.89	11.76
Audio Flamingo 2	✗	✗	37.43	39.94	36.22	35.62	✗	✗	71.83	58.45	26.93	31.29

D.3 Structured Musical Adjustment Evaluation

Metrics and Source Dependence. Inverse adjustment is evaluated with attribute precision, recall, and F1, operation accuracy, adjustment magnitude MAE, and preservation F1. These metrics evaluate attribute selection, operation direction, adjustment amount, and preservation constraints. Matched, mismatched, and zero audio conditions use the same source dependence protocol described above.

Results Across Datasets. We evaluate structured musical adjustment across DEAM, PMEmo2019, and MERGE Audio. As shown in Table 9, MuseReason maintains 100% schema validity and high attribute precision across the three datasets. The results further characterise attribute selection, operation prediction, adjustment magnitude, and preservation constraints under different affect corpora.

Table 9: Complete structured musical adjustment results across affect datasets. All scores except Amount MAE are reported in %. The DEAM row uses the matched audio result reported in the main text.

Dataset	Schema	Attr. P	Attr. R	Attr. F1	Operation	Amount MAE	Preserve F1
DEAM	100.00	92.67	40.56	50.12	68.75	46.48	9.04
PMEmo2019	100.00	92.02	44.60	54.68	54.50	43.31	7.08
MERGE Audio	100.00	94.98	38.62	49.39	57.10	60.87	5.80

D.4 Counterfactual Music Editing Evaluation

Editing Metrics. Given source music x, counterfactual music x_{cf}, target affect $a^\star$, and the independent affect verifier, target affect improvement is

$$\Delta_{\text{aff}} = D_{\text{aff}}\left(V_{\text{aff}}(x), a^\star\right) - D_{\text{aff}}\left(V_{\text{aff}}(x_{\text{cf}}), a^\star\right). \tag{25}$$

A positive value indicates movement towards the target. Improve reports the proportion of samples with positive target movement. Content preservation uses MuQ representation similarity (Zhu et al., 2025). Melody preservation uses chroma similarity that is insensitive to transposition. Joint Success requires affect improvement together with the predefined content and melody criteria. The reported thresholds are $\rho_{\text{content}} = 0.25$ and $\rho_{\text{melody}} = 0.85$.

Music Quality Metrics. Music quality is evaluated with Fréchet Audio Distance based on MERT, Audiobox Content Enjoyment and Production Quality (Tjandra et al., 2025), and absolute loudness deviation $|\Delta\text{LUFS}|$. MuQ-MuLan provides an auxiliary measure of alignment with the target affect description. Figure 7 provides a supplementary view of the main editing comparison through Joint Success, Content similarity, and Melody similarity.

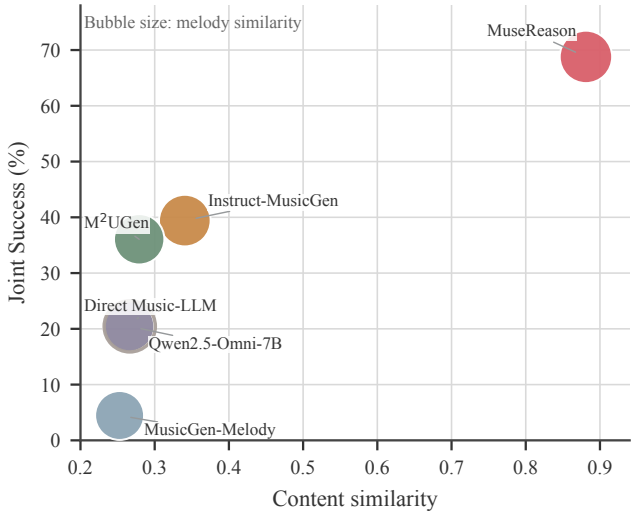


Figure 7: Supplementary visualisation of the main counterfactual editing comparison. Joint Success is shown against Content similarity, while bubble size denotes Melody similarity.

Candidate Reranking Based on the Verifier. Four fixed candidates are generated for each target and the primary verifier selects one candidate for evaluation. Because the same verifier participates in selection and scoring, reranking is reported as an analysis assisted by selection and is separate from the fixed sample comparison. The same affect change and preservation metrics are used. Table 10 reports the full reranked results, while Figure 8 compares the first sample with verifier reranking for Joint Success and Δ_{aff}.

D.5 Outcome Consistency Evaluation

Outcome Labels and Consistency Rules. Let

$$A = \mathbf{1}[\Delta_{\text{aff}} > 0], \qquad P = \mathbf{1}[S_{\text{content}} \geq 0.25 \ \wedge \ S_{\text{melody}} \geq 0.85], \qquad Y_{\text{acc}} = A \wedge P. \tag{26}$$

Table 10: Verifier-reranked results over four fixed candidates per target. The primary verifier is used for both candidate selection and scoring.

Method	Joint	Improve	Δ_{aff}	Content	Melody
MuseReason (ours)	**71.75%**	71.75%	0.2589	**0.8823**	**0.9923**
Reference Plan	24.91%	53.16%	0.1075	0.2628	0.9306
Target Text	25.28%	53.16%	0.1078	0.2626	0.9308
Target VA	25.28%	54.28%	0.1077	0.2624	0.9309
Direct Music LLM	24.16%	53.16%	0.1096	0.2615	0.9306
Qwen2.5-Omni-7B	24.91%	53.90%	0.1080	0.2623	0.9310
Instruct-MusicGen†	49.81%	72.49%	0.1549	0.3397	0.9536
MusicGen-Melody†	14.13%	71.00%	0.2228	0.2405	0.7433
M^{2}UGen†	55.02%	**91.45%**	**0.3439**	0.2971	0.8992

† Native five-second output; not duration-matched to ACE-Step.

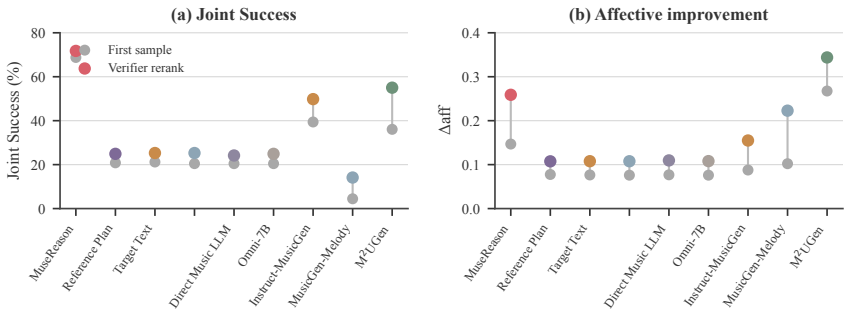


Figure 8: Candidate reranking based on the verifier over four fixed candidates. The two panels compare the first sample with verifier reranking for Joint Success and Δ_{aff}.

Here A, P, and Y_{acc} are the reference labels for Affect, Preservation, and Accepted. The preservation set p at the plan level guides control compilation. The outcome preservation label P is defined by content and melody similarity. Failure mode and next action are deterministic functions of (A, P):

Affect, Preservation, and Accepted are scored with binary Macro F1. Failure and Action use multiclass Macro F1 over the labels in Table 11. Accepted Rule measures whether the predicted acceptance decision equals the logical conjunction of the model's predicted Affect and Preservation decisions. Mapping measures whether the predicted failure mode and next action equal the deterministic mapping implied by those two predicted decisions. The two rule metrics therefore test internal consistency rather than agreement with a single class in isolation.

Table 11: Reference mapping from measured outcome decisions to failure mode and next action.

A	P	Failure mode	Next action
1	1	none	accept
0	1	affect miss	revise affect
1	0	preservation failure	restore preservation
0	0	affect miss and preservation failure	regenerate

Feedback Conditions. We evaluate the contribution of generated audio and independent measurements under six conditions. Full feedback uses the generated music representation together with the complete measurement record. No feedback removes the measurement record. Shuffled feedback breaks the correspondence between generated music and its measurements. Noisy feedback perturbs the measurement values. Zero audio removes effective audio information, while mismatched audio replaces the corresponding generated music with another sample.

Complete Feedback Results. Table 12 reports the complete results for all feedback and audio conditions. Figure 9 visualises the Macro F1 change of each perturbed condition relative to full feedback.

Table 12: Outcome assessment conditioned on feedback under full feedback, no feedback, shuffled feedback, noisy feedback, zero audio, and mismatched audio. Affect, Preservation, Accepted, Failure, and Action are reported as Macro F1 scores (%). Higher values are better.

Condition	Affect ↑	Preservation ↑	Accepted ↑	Failure ↑	Action ↑
Full feedback	**66.79**	**58.07**	**64.58**	**25.95**	**35.08**
No feedback	43.26	47.54	42.40	20.95	20.95
Shuffled feedback	43.17	47.57	54.81	23.30	24.43
Noisy feedback	48.51	50.62	49.28	22.59	20.88
Zero audio	40.40	43.08	40.47	20.33	20.83
Mismatched audio	41.87	42.00	39.13	21.66	22.78

Results Across Datasets. We evaluate outcome assessment conditioned on feedback across DEAM, PMEmo2019, and MERGE Audio. Table 13 reports schema validity and the complete set of outcome decision metrics. MuseReason maintains nearly complete schema validity across all three datasets and produces structured decisions for affect improvement, preservation satisfaction, acceptance, failure mode, and subsequent action.

Table 13: Complete results for outcome assessment conditioned on feedback across affect datasets. Schema validity and task Macro F1 scores are reported in %. The DEAM row uses the full feedback result reported in the main text.

Dataset	Schema	Affect	Preservation	Accepted	Failure	Action
DEAM	99.45	66.79	58.07	64.58	25.95	35.08
PMEmo2019	100.00	42.04	61.41	44.33	22.03	35.43
MERGE Audio	100.00	37.09	49.65	39.08	24.81	24.81

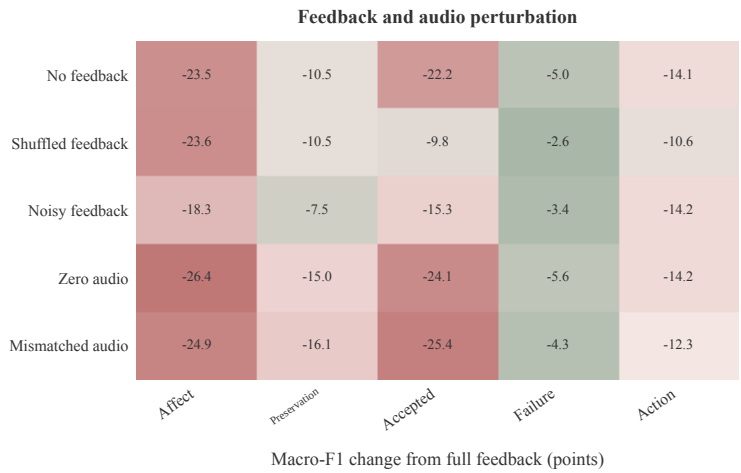


Figure 9: Changes in outcome Macro F1 relative to full feedback under feedback and audio perturbations. Negative values indicate lower performance than full feedback.

Outcome Assessment over Multiple Rounds. Outcome assessment over multiple rounds keeps the generated music and independent measurements fixed. Round 1 uses the original inputs after generation. Rounds 2 through 5 additionally condition on the structured assessment from the preceding round. Table 14 reports the complete results used in the main figure.

Algorithm 2 corresponds to the outcome assessment experiment over multiple rounds. The generated music and independent measurements remain fixed. The semicolon indicates additional prompt context containing the preceding structured assessment. It does not denote a separate model. The experiment reports each of five rounds against the same outcome references.

Algorithm 2: Repeated assessment of a fixed generated result

Input: Cached generated music tokens Z_{cf}, measurement record o_{cf}, target affect $a^{\star}$, fixed reasoner R_{θ_o}, and depth $K \geq 1$.

1. Obtain the initial assessment: $q_{\mathrm{cf}}^{(1)} \leftarrow R_{\theta_o}(Z_{\mathrm{cf}}, o_{\mathrm{cf}}, a^{\star})$.
2. For each round $k = 2, \ldots, K$, serialise the preceding assessment as $h^{(k)} \leftarrow \mathrm{Serialize}(q_{\mathrm{cf}}^{(k-1)})$ and compute
$$q_{\mathrm{cf}}^{(k)} \leftarrow R_{\theta_o}(Z_{\mathrm{cf}}, o_{\mathrm{cf}}, a^{\star}; h^{(k)}).$$
3. Retain the assessment sequence for round-wise evaluation.

Output: $(q_{\mathrm{cf}}^{(1)}, \ldots, q_{\mathrm{cf}}^{(K)})$. The reported setting is $K = 5$; no additional editor call is made within the recurrence.

Table 14: Outcome assessment over multiple rounds. Macro F1 and consistency scores are reported in %.

Round	Affect	Preservation	Accepted	Failure	Action	Accepted Rule	Mapping
1	66.79	58.07	64.58	25.95	35.08	76.24	55.80
2	68.48	59.13	67.70	25.42	35.37	77.62	56.91
3	**70.48**	62.33	67.48	25.95	**36.08**	**82.78**	57.18
4	64.48	**64.03**	**68.07**	**26.06**	35.46	77.13	**62.50**
5	50.48	57.90	67.74	25.72	35.22	76.80	58.33