

EVALUATING MUSIC CONTEXT PRESERVATION: A MULTI-FACET FRAMEWORK FOR MUSIC EDITING SYSTEMS

Yash Vishe Eric Xue Xunyi Jiang
Zachary Novack Junda Wu Julian McAuley Xin Xu

University of California, San Diego

{yvishe, exue, xuj003, znovack, juw069, jmcauley, xinxucs}@ucsd.edu

ABSTRACT

Music editing plays a vital role in modern music production, with applications in film, broadcasting, and game development. Recent advances in music editing systems have enabled diverse editing tasks such as timbre transfer, instrument substitution, and genre transformation. However, many existing works overlook evaluating their ability to preserve musical facets that should remain unchanged during editing, which we define as Music Context Preservation (MuseCP). While some studies do consider MuseCP, their evaluation protocols and metrics are not comprehensive. To address this, we introduce the first MuseCP evaluation framework, **MuseCPEval**, that covers four categories of music facets with fine-grained and well-tailored metrics to capture nuanced changes in music attributes. Objective validation and a human study demonstrate the effectiveness of these metrics. Moreover, the case studies on diverse music editing systems illustrate the practical utility of these metrics as a testbed and diagnostic tool, providing insights into the strengths and limitations of existing systems. We hope our metrics and findings can offer practical guidance for developing more effective and reliable music editing strategies with strong MuseCP capability¹.

1. INTRODUCTION

Music editing has emerged as an important direction in recent research and real-world music applications [1–6]. Although recent approaches have achieved promising progress in these areas, their evaluation for **Music Context Preservation** (MuseCP) capability remains limited in scope or even missing, which we define as the ability of a music editing system to retain the fundamental musical attributes of the source material that are not intended to be edited. These attributes constitute essential facets of music theory, such as harmony, rhythm, and structure, and

¹ Code: <https://github.com/Yashvishe13/MuseCPEval>, Demo: <https://musecpeval-demo.netlify.app/>

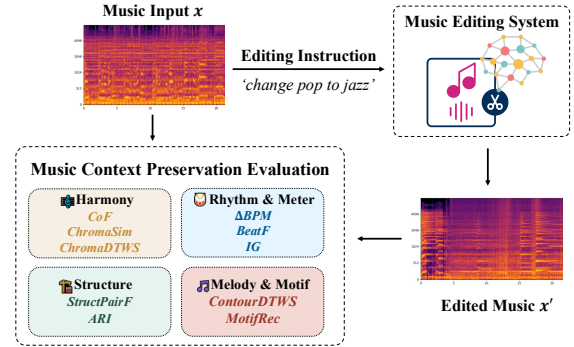


Figure 1: Music context preservation evaluation overview.

their persistence during editing is therefore highly significant and warrants thorough evaluation. However, as shown in Table 1, the MuseCP evaluation protocols of existing music editing works are not comprehensive. For instance, ZETA [7], Audio Prompt Adapter [8], and SteerMusic [9] only evaluated structure, harmony, and melody, respectively. Even some work didn’t consider MuseCP evaluation. This lack of well-rounded MuseCP evaluation limits our understanding of the strengths and weaknesses of current music editing systems. Meanwhile, to our best knowledge, there is no framework or tool for systematically measuring the MuseCP capability of music editing systems.

To address this issue, we construct **MuseCPEval** (Figure 1), the first comprehensive MuseCP evaluation framework for music editing systems, which covers four music facets with carefully tailored fine-grained metrics. Overall, here are the main contributions of our work:

1. **Introducing MuseCPEval (§2):** We propose the first MuseCP evaluation framework MuseCPEval. It reorganizes the music facets that are commonly considered to be preserved into four categories: harmony, rhythm & meter, structure, and melody & motif. For each facet, fine-grained metrics are tailored and collected.
2. **Validating MuseCP evaluation metrics (§3):** The metrics are validated through both objective experiments and human listening studies, demonstrating their effectiveness in capturing nuanced music attribute changes and strong alignment between metric values and human perception of musical differences.
3. **Conducting case studies on diverse music editing systems (§4):** We apply MuseCPEval to four rep-



Method	Backbone	Task	MuseCP Evaluation?	Evaluated Facet
AUDIT [10]	Diffusion	Add, Drop, Replace, Inpaint, Super-resolution	✗	—
InstructME [11]	Diffusion	Add, Remove, Extract, Replace, Remix	✓	Harmony, Rhythm
MusicMagus [12]	Diffusion	Music Attribute Modification	✓	Harmony, Structure
ZETA [7]	Diffusion	Music Attribute Modification	✓	Structure
Audio Prompt Adapter [8]	Diffusion	Timbre and Genre Transfer	✓	Harmony
Instruct-MusicGen [13]	Transformer	Add, Separate, and Remove Stems	✗	—
Melodia [14]	Diffusion	Music Attribute Modification	✓	Harmony, Melody, Structure
SteerMusic [9]	Diffusion	Music Attribute Modification	✓	Melody

Table 1: MuseCP evaluation of recent music editing systems (sorted by publication year, 2023-2026).

representative music editing systems spanning diverse methodologies, model architectures, and editing tasks. Through detailed analyses of their strengths and limitations in preserving different music attributes, we show that MuseCPEval serves as a practical testbed and diagnostic tool, offering insights for developing more reliable music editing systems.

2. MUSECPEVAL

2.1 Music Context Preservation Evaluation

For most existing music systems, given an original audio music input x and a user’s editing instruction, they will produce an edited music x' , where certain music attributes are expected to be edited while all other contexts should remain preserved. In this work, we introduce a framework for evaluating the extent to which non-target music attributes are preserved under editing. We reorganize widely considered music attributes into four music facets: **Harmony, Rhythm & Meter, Structure, and Melody & Motifs**, covering fundamental dimensions of music [15]. For each facet, we define fine-grained metrics that measure different aspects of similarity between the reference x and the edited output x' within the facet. The overall evaluation thus provides a multi-facet view of MuseCP. In this section, we introduce these metrics in detail.

2.2 Harmony

Harmony is an important concept in music, representing the basic color of the music. We measure this across high-level keys and scales to low-level chords and chroma using three metrics.

Circle-of-fifths (CoF) Distance. As for key comparison, we use CoF Distance [16] to measure tonal relatedness between x and x' . The tonic $k \in \{0, 1, \dots, 11\}$ (12 pitch classes: C, C $^\sharp$, ..., B) and mode (major or minor) of x and x' (k') are estimated via the Krumhansl–Schmuckler algorithm [17]. To account for the perceptual equivalence of a minor key and its relative major, each minor tonic is normalized to its relative major via $\tilde{k} = (k+3) \bmod 12$, while major tonics are left unchanged. Each $\tilde{k}$ is then mapped to a position $\phi(\tilde{k}) \in \{0, 1, \dots, 11\}$ on the circle of fifths, where adjacent positions differ by a perfect fifth. The distance is then calculated and normalized to $[0, 1]$ as:

$$\text{CoF}(x, x') = \frac{1}{6} \min \left(\left| \phi(\tilde{k}) - \phi(\tilde{k}') \right|, 12 - \left| \phi(\tilde{k}) - \phi(\tilde{k}') \right| \right) \quad (1)$$

A score of 0 means the x and x' share the same key signature; A score of 1 indicates maximally distant keys (e.g., C major vs. F $^\sharp$ major).

Chroma Similarity. Chroma Similarity [8, 12] is used to capture the chroma distribution similarity between x and x' . Concretely, x and x' are each converted into a chroma matrix $\mathbf{H} \in \mathbb{R}^{T \times 12}$ via Constant-Q Transform [18], where T is the number of frames and the 12 dimensions correspond to 12 chromas. The mean chroma vector $\bar{\mathbf{h}} \in \mathbb{R}^{12}$ is obtained by averaging $\mathbf{H}$ over the time axis, representing the global chroma energy distribution of the music. Chroma Similarity is then the cosine similarity between $\bar{\mathbf{h}}$ and $\bar{\mathbf{h}}'$, the mean chroma vectors of x and x' respectively:

$$\text{ChromaSim}(x, x') = \frac{\bar{\mathbf{h}} \cdot \bar{\mathbf{h}}'}{\|\bar{\mathbf{h}}\| \|\bar{\mathbf{h}}'\|} \quad (2)$$

It ranges in $[0, 1]$, where 1 indicates identical chroma distributions and 0 indicates completely dissimilar.

Chroma DTW Similarity. To further measure the temporal dynamics of pitch-class content, we additionally compute Chroma Dynamic Time Warping (DTW) Similarity [19]. We compute a pairwise cosine distance matrix $\mathbf{D} \in \mathbb{R}^{T \times T'}$, where $\mathbf{h}_i \in \mathbb{R}^{12}$ and $\mathbf{h}'_j \in \mathbb{R}^{12}$ denote the i -th $T_{i=1}^T$ and j -th $T'_{j=1}^{T'}$ frame of $\mathbf{H}$ and $\mathbf{H}'$ respectively, and $D_{ij} = 1 - \frac{\mathbf{h}_i \cdot \mathbf{h}'_j}{\|\mathbf{h}_i\| \|\mathbf{h}'_j\|}$. DTW finds the optimal path $P^* = \arg \min_P \sum_{(i,j) \in P} D_{ij}$, where P is the set of all valid monotonic and continuous index-pair sequences from $(1, 1)$ to (T, T') . The Chroma DTW Similarity is defined as the mean cosine similarity along the optimal path:

$$\text{ChromaDTWS}(x, x') = \frac{1}{|P^*|} \sum_{(i,j) \in P^*} (1 - D_{ij}) \quad (3)$$

It’s in $[0, 1]$ where 1 means the two chroma sequences are identical and 0 means completely dissimilar.

2.3 Rhythm & Meter

Rhythm and meter describe the temporal organization of music. We evaluate their preservation between x and x' using three metrics at different levels of granularity.

Folded Beat Difference. This measures the global tempo difference. The tempos of x and x' are estimated in beats-per-minute (BPM) via *librosa* [20] beat tracker, as BPM and BPM' respectively. The absolute BPM difference is then computed, but naively comparing BPM values is susceptible to octave errors, where the beat tracker may

estimate the tempo of the same music as BPM, $2 \times \text{BPM}$, or $\frac{1}{2} \times \text{BPM}$ depending on the metric level it locks onto. To handle this, we fold the estimate by taking the minimum difference folded over the metric octave:

$$\Delta_{\text{BPM}}(x, x') = \min \left(|\text{BPM} - \text{BPM}'|, |\text{BPM} - 2 \cdot \text{BPM}'|, \left| \text{BPM} - \frac{\text{BPM}'}{2} \right| \right) \quad (4)$$

A value of 0 indicates identical tempo; larger values indicate greater tempo deviation.

Beat F-measure. It assesses whether beats fall in the right places. Beat positions are first extracted from x and x' as a sequence of timestamps in seconds. A beat in x' is counted as a true positive if it falls within a tolerance window ($\pm 70\text{ms}$) around a reference beat in x . The F-measure is then:

$$\text{BeatF}(x, x') = \frac{2 \cdot P \cdot R}{P + R} \quad (5)$$

where P and R are the precision and recall of matched beats under the *mir_eval* framework [21]. It ranges in $[0, 1]$, where 1 means every beat in x' perfectly aligns with a beat in x , and 0 means no correct hits.

Information Gain (IG). We adopt the beat information gain from *mir_eval* to measure how consistently the beat phase of x' aligns with that of x . For each reference beat, we compute its signed timing error relative to the nearest estimated beat, normalized by the local beat period. The procedure is repeated in the reverse direction to penalize both missed and spurious beats. All normalized errors are binned into a circular histogram of $K = 41$ bins, yielding an empirical distribution p . The IG is then defined as the KL divergence from p to the uniform distribution:

$$\text{IG}(x, x') = (\log_2 K - H(p)) / \log_2 K \quad (6)$$

where $H(p) = -\sum_{k=1}^K p_k \log_2 p_k$ is the entropy of p . It ranges in $[0, 1]$. A high score indicates that timing errors concentrate at a consistent phase offset, reflecting a stable, regular pulse, while a low score indicates that errors are spread uniformly, implying no coherent metrical relationship between x and x' .

2.4 Structure

Structure describes the high-level organization of music, such as how it is divided into sections and how those sections relate to one another. We evaluate its preservation by using two complementary metrics.

Structural Pairwise F-measure. Music segments are extracted from x and x' via *msaf* [22]. Each segment spans a set of time points, where each time point inherits the label of the segment it belongs to. Given the reference labeling from x and the estimated labeling from x' , we examine all pairs of time points and ask whether each pair shares a label in both labelings. A true positive occurs when both labelings agree that a pair belongs to the same segment; a false positive when only x' groups them together; a false negative when only x does. The precision P_{pair} and recall R_{pair} are the true-positive rate with respect to x' and x ,

respectively. The Structural Pairwise F-measure is:

$$\text{StructPairF}(x, x') = \frac{2 \cdot P_{\text{pair}} \cdot R_{\text{pair}}}{P_{\text{pair}} + R_{\text{pair}}} \quad (7)$$

It ranges in $[0, 1]$, where 1 means the two labelings induce identical pairwise groupings and 0 means no agreement.

Adjusted Rand Index (ARI). The Rand Index additionally accounts for pairs that both labeling assign to different segments, i.e., agreement on both togetherness and separation [23]. ARI further corrects for chance by subtracting the expected agreement under random labeling:

$$\text{ARI}(x, x') = \frac{\text{RI} - \mathbb{E}[\text{RI}]}{\max(\text{RI}) - \mathbb{E}[\text{RI}]} \quad (8)$$

where RI is the fraction of all time-point pairs on which the two labelings agree. A score of 1 indicates perfect agreement, 0 indicates chance-level agreement, and negative values (rare in practice) indicate worse-than-chance clustering. By additionally rewarding agreement on separation and discounting accidental matches, ARI provides a more comprehensive view of structural preservation.

2.5 Melody & Motif

Melody and its recurring short motifs are among the most salient cues by which listeners recognize a piece. Since this facet focuses solely on melodic content, we first isolate it from key-level variation and then use the following two metrics to measure its preservation.

Contour DTW Similarity. This metric captures whether the pitch-class content of the piece is preserved [24]. We characterize frame-level pitch content with 12-dimensional CQT chromagrams [25], yielding per-frame pitch-class sequences $\mathbf{d} \in \mathbb{R}^{T \times 12}$ and $\mathbf{d}' \in \mathbb{R}^{T' \times 12}$, whose rows $\mathbf{d}_i$ and $\mathbf{d}'_j$ encode the pitch-class distribution of a single frame. We then compute a pairwise cosine distance matrix $\mathbf{D} \in \mathbb{R}^{T \times T'}$ with $D_{ij} = 1 - \frac{\mathbf{d}_i \cdot \mathbf{d}'_j}{\|\mathbf{d}_i\| \|\mathbf{d}'_j\|}$. DTW finds the optimal warping path $P^* = \arg \min_P \sum_{(i,j) \in P} D_{ij}$, where P is the set of all valid monotonic and continuous index-pair sequences from $(1, 1)$ to (T, T') . The Contour DTW Similarity is defined as the mean cosine similarity along the optimal path:

$$\text{ContourDTWS}(x, x') = \frac{1}{|P^*|} \sum_{(i,j) \in P^*} (1 - D_{ij}) \quad (7)$$

It clips to $[0, 1]$, where 1 indicates identical pitch-class trajectories up to tempo-elastic alignment. Lower values indicate greater divergence in pitch-class content.

Motif 3-gram Recall. To assess whether a motif survives, we tokenize each track into a sequence of pitch-class intervals between successive notes and measure how many short interval patterns from x reappear in x' [26]. Let $\mathcal{G}_n(x)$ denote the set of all contiguous length- n subsequences of this interval sequence. Each one is a short interval pattern that represents a local melodic figure. Using $n = 3$ to capture short but distinctive figures, the Motif

Music Facets	Harmony						Structure			
Editing	CoF ↓	Expected	ChromaSim↑	Expected	ChromaDTWS↑	Expected	StructPairF↑	Expected	ARI↑	Expected
+7 semitones (+Perfect Fifth)	0.18 ± 0.09	≈ 0.167	0.94 ± 0.05	< 1	0.81 ± 0.12	< 1	0.67 ± 0.14	≈ 1	0.30 ± 0.29	≈ 1
-3 semitones (- Minor Third)	0.50 ± 0.11	0.5-0.67	0.82 ± 0.11	< 1	0.71 ± 0.14	< 1	0.69 ± 0.16	≈ 1	0.37 ± 0.28	≈ 1
ABC→AAA	0.06 ± 0.13	≈ 0	0.99 ± 0.01	≈ 1	0.89 ± 0.12	≈ 1	0.58 ± 0.17	< 1	0.19 ± 0.24	< 1
ABC→ABA	0.05 ± 0.12	≈ 0	0.99 ± 0.00	≈ 1	0.92 ± 0.12	≈ 1	0.65 ± 0.16	< 1	0.31 ± 0.27	< 1
Tempo shifted +50%	0.03 ± 0.12	≈ 0	0.99 ± 0.00	≈ 1	0.84 ± 0.14	≈ 1	0.66 ± 0.16	≈ 1	0.25 ± 0.29	≈ 1
All beats shifted later by a constant -150 ms	0.02 ± 0.10	≈ 0	0.99 ± 0.00	≈ 1	0.96 ± 0.13	≈ 1	0.74 ± 0.16	≈ 1	0.48 ± 0.30	≈ 1
Interval direction flipped for the main melody	0.06 ± 0.16	0-0.167	0.99 ± 0.02	< 1	0.92 ± 0.14	< 1	0.76 ± 0.19	≈ 1	0.52 ± 0.34	≈ 1
The climax phrase of the main melody is inverted	0.02 ± 0.11	≈ 0	0.99 ± 0.00	≈ 1	0.96 ± 0.13	≈ 1	0.78 ± 0.17	≈ 1	0.56 ± 0.32	≈ 1
Music Facets	Rhythm & Meter						Melodic Content & Motif			
Editing	Δ BPM↓	Expected	BeatF↑	Expected	IG↑	Expected	ContourDTWS↑	Expected	MotifRec↑	Expected
+7 semitones (+Perfect Fifth)	1.26 ± 8.79	≈ 0	0.90 ± 0.19	≈ 1	0.76 ± 0.19	≈ 1	0.78 ± 0.05	< 1	0.61 ± 0.10	< 1
-3 semitones (- Minor Third)	0.00 ± 0.00	≈ 0	0.92 ± 0.16	≈ 1	0.79 ± 0.18	≈ 1	0.80 ± 0.06	< 1	0.58 ± 0.09	< 1
ABC→AAA	0.87 ± 2.72	≈ 0	0.54 ± 0.22	< 1	0.39 ± 0.17	< 1	0.76 ± 0.04	< 1	0.65 ± 0.09	< 1
ABC→ABA	0.29 ± 1.15	≈ 0	0.74 ± 0.14	< 1	0.57 ± 0.14	< 1	0.85 ± 0.05	< 1	0.74 ± 0.09	< 1
Tempo shifted +50%	26.10 ± 12.96	> 0	0.24 ± 0.08	< 1	0.19 ± 0.11	≈ 0	0.69 ± 0.04	< 1	0.60 ± 0.07	< 1
All beats shifted later by a constant -150 ms	1.35 ± 8.80	> 0	0.03 ± 0.07	≈ 0	0.63 ± 0.13	< 1	0.82 ± 0.05	< 1	0.66 ± 0.09	< 1
Interval direction flipped for the main melody	0.13 ± 0.89	≈ 0	0.94 ± 0.15	≈ 1	0.83 ± 0.20	≈ 1	0.78 ± 0.20	< 1	0.61 ± 0.30	< 1
The climax phrase of the main melody is inverted	1.26 ± 8.79	≈ 0	0.98 ± 0.08	≈ 1	0.93 ± 0.12	≈ 1	0.95 ± 0.10	< 1	0.91 ± 0.17	< 1

Table 2: Objective validation for MuseCPEval metrics.

n -gram Recall is defined as the fraction of reference motifs recovered in the edited output:

$$\text{MotifRecall}(x, x') = \frac{|\mathcal{G}_n(x) \cap \mathcal{G}_n(x')|}{|\mathcal{G}_n(x)|} \quad (8)$$

It lies in $[0, 1]$, where 1 means every motif is reproduced in x' , and 0 means no motif survives.

3. METRIC VALIDATION

The metrics proposed in Section 2 for MuseCP evaluation are implemented by the algorithms based on probability, making it essential to verify their validity. In this section, we conduct validation from two complementary perspectives: *i*) objective validation using edits with known effects on specific musical facets; *ii*) human listening studies to assess alignment with perception.

3.1 Objective Validation

We evaluate whether the proposed metrics respond correctly to edits with known effects on specific musical facets. 50 MIDI samples of high musicality with clear melodic lines and structure are selected from the Lakh dataset² for correct editing. We apply eight tailored editing operations covering four music facets to each sample as shown in Table 2. After editing, the samples are transferred to audio. We compute all metrics and report the mean and standard deviation across all samples in Table 2, where the *Expected* columns specify the anticipated metric values under each editing operation. Details are in Appendix B.

According to the results, most metrics behave consistently with expectations, demonstrating the effectiveness of our evaluation framework. The metrics are illustrated to be sensitive to changes in musical attributes. For example, pitch-shifting leads to clear changes in harmonic metrics such as CoF and ChromaSim, while leaving rhythm metrics (e.g., ΔBPM, BeatF) unaffected. Conversely, tempo scaling (+50%) results in a large ΔBPM and small BeatF and IG, demonstrating strong sensitivity to rhythmic changes. For the structural metrics, although Struct-PairF and ARI show values below 1.0 for editing without

effects on musical structure, such as harmony transposition and tempo shift, due to the segmentation pipeline being sensitive to spectral changes beyond structural boundaries, both metrics successfully differentiate structural edits from structure-preserving edits, confirming their utility as relative indicators. Overall, these results show that the MuseCPEval metrics can effectively measure MuseCP capability by capturing nuanced changes in attributes.

3.2 Human Listening Study

Music Facet	Human-Gold	Metric-Human	Metric-Gold
Harmony	72.7	65.9	93.2
Rhythm & Meter	100.0	100.0	100.0
Structure	72.7	65.9	72.7
Melody & Motif	72.7	45.5	72.7

Table 3: Agreement rates (%) between human perception, metric values, and gold labels across different music facets.

We also conduct a human listening study, as objective validation cannot guarantee that metrics align with human perception and may be biased by the limited edits we consider. For each music facet, we design 4 pairwise comparison questions. Each question consists of one original audio clip and two edited versions with different editing strengths, which serve as gold labels indicating the relative degree of deviation from the original. Participants are asked to determine which edited sample is farther from the original with respect to the specified facet, while ignoring other aspects. Their choices are human labels. The original music data used in the questionnaire are from the MIDI samples used in §3.1. Corresponding metric values are also computed for all origin-edited sample pairs, enabling direct comparison between metric values, gold labels, and human judgments. We obtained 11 valid responses from human participants that passed the sanity checks. The details are shown in the Appendix C. Finally, we report the agreement rates among human labels, gold labels, and metric-based comparisons in Table 3.

Overall, the results demonstrate that the proposed metrics are effective and largely consistent with both gold labels and human perception. In particular, the Metric-Gold

² <https://colinraffel.com/projects/lmd/>

agreement is high across all facets, indicating that the metrics correctly capture the intended nuanced differences in music attributes. The Metric–Human agreement is also reasonably aligned for most facets, showing that the metrics reflect perceptually meaningful differences. The relatively lower agreement on melody & motif suggests this facet is more subjective and harder to quantify. Overall, these results validate the reliability and perceptual relevance of MuseCPEval.

4. CASE STUDIES: MUSIC EDITING SYSTEM ANALYSIS THROUGH MUSECP

While the evaluations in existing music editing works typically report aggregate quality scores, they provide limited insight into how their methodologies affect MuseCP capability and their pros and cons. In this section, we demonstrate that our MuseCPEval metrics go beyond evaluation and serve as a diagnostic tool for music editing systems. We apply our framework to four representative and diverse music editing systems spanning different architectures, editing paradigms, and tasks: MusicMagus [12], ZETA [7], Audio Prompt Adapter [8], and Instruct-MusicGen [13]. Our analysis reveals the strengths and limitations of each system through MuseCPEval, providing fine-grained insights that conventional evaluations miss.

4.1 Setups

MusicMagus. MusicMagus is a zero-shot diffusion-based music editing method that performs text-guided latent embedding steering with cross-attention consistency constraints. We follow its official protocol by generating 60 samples with AudioLDM2 [27] and then constructing paired editing instructions through its captioning process, covering instrument and style transfer. The 60 paired editing instructions and audio are fed into MusicMagus³ to obtain edited outputs.

ZETA. ZETA is a zero-shot audio editing method that inverts the input music into the latent trajectory of a pre-trained diffusion model, then performs reverse denoising conditioned on a target text prompt to generate the edited output while preserving the original coarse structure. Following its supervised text-guided editing evaluation, we use 34 musical excerpts from the MedleyDB [28] with 324 editing instructions. And then we directly use ZETA⁴ to generate 324 edited outputs for MuseCP evaluation.

Audio Prompt Adapter. AP-Adapter is a lightweight adapter-based extension of AudioLDM2, which injects AudioMAE-extracted features from the input audio through decoupled cross-attention layers, enabling text-guided music editing. Following its official evaluation⁵, we use 40 in-domain samples with 3 instrument editing instructions for each sample to conduct timbre transfer.

Instruct-MusicGen. Instruct-MusicGen is a text-based music editing system that extends MusicGen [29] to

follow natural language instructions for modifying specific musical attributes, such as instrumentation. Following the official setups⁶, we adopt instrument-level editing tasks and data from Slakh [30], including adding, removing, and extracting instruments. For each edit, 150 input audio and editing instruction pairs are evaluated.

4.2 Case Analysis

MusicMagus. MusicMagus demonstrates strong MuseCP ability in harmony and structural form, achieving near-perfect chroma-based similarity, such as ChromaSim and ChromaDTWS, and high structural consistency (StructPairF). These results suggest that the model effectively retains global harmonic content and large-scale arrangement while modifying target attributes. It also shows relatively good preservation of melodic contour and moderate motif retention according to ContourDTWS and MotifRec, respectively. They suggest that enforcing cross-attention consistency during diffusion sampling in MusicMagus significantly contributes to MuseCP. The low Δ BPM shows MusicMagus performs well on rhythm preservation, while its BeatF and IG are near-zero, which we think may reflect musically meaningful rhythmic transformations induced by style-transfer edits rather than unintended context loss. Meanwhile, the strong overall MuseCP capability suggests that MusicMagus’ inference-time intervention, without additional fine-tuning, could avoid distorting the original musical content.

Music Editing System	Δ BPM
MusicMagus	5.573 ± 10.315
ZETA	4.253 ± 12.234
Audio Prompt Adapter	20.856 ± 20.584
Instruct-MusicGen	20.012 ± 14.310

Table 4: Δ BPM for each music editing systems.

ZETA. ZETA adopts a zero-shot DDPM inversion framework, where the edited audio is generated by reusing the inverted latent trajectory of the source signal while replacing the conditioning text prompt. This design naturally encourages preservation of low- and mid-level musical context inherited from the original audio. Consistent with this mechanism, ZETA demonstrates particularly strong harmony and beat preservation in our MuseCP evaluation. High chroma-based similarity scores suggest that the tonal center and harmonic content are largely retained after editing, while strong beat-related metrics indicate that temporal periodicity, onset patterns, and rhythmic pulse remain stable throughout the generation process. These results imply that the inversion process effectively anchors the edited output to the source music, allowing semantic modifications without substantially disrupting core harmonic or rhythmic foundations. At the same time, ZETA’s prompt-driven resampling may still alter higher-level musical organization or melodic details when stronger edits are applied, highlighting a common trade-off in text-guided diffusion editing between editability and preservation of fine-

³ <https://github.com/ldzhangyx/MusicMagus>

⁴ <https://github.com/HilaManor/AudioEditingCode/>

⁵ <https://github.com/fundwotsai2001/AP-adapter>

⁶ <https://github.com/ldzhangyx/instruct-MusicGen>

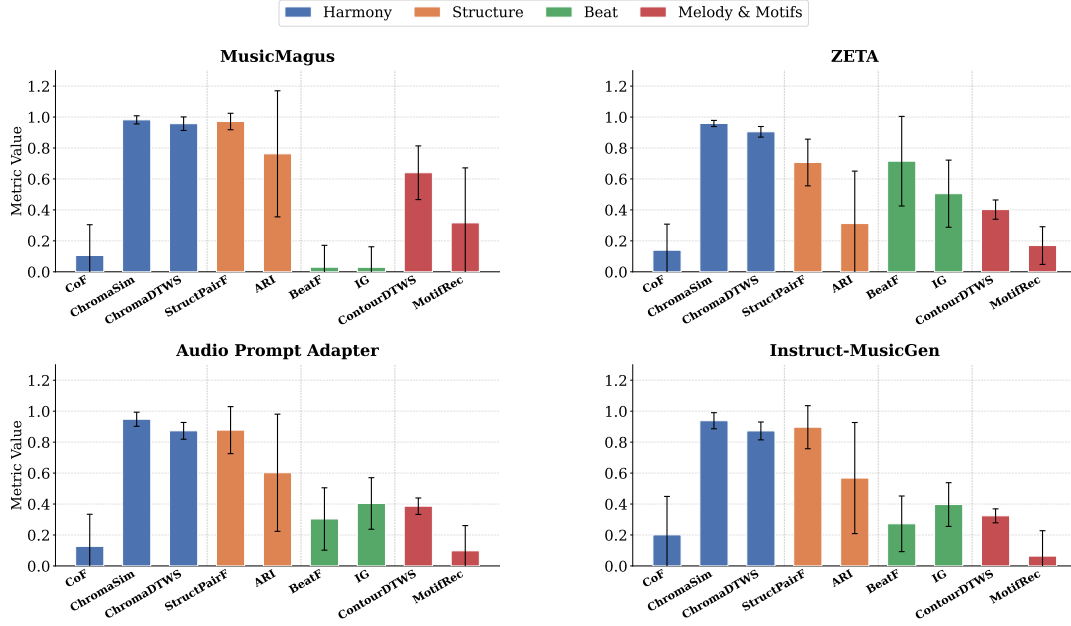


Figure 2: MuseCPEval evaluation from four different music editing systems. Note that these systems cannot be compared because the evaluation setups of the systems are different from each other.

grained long-range context. Overall, the MuseCP results suggest that ZETA is especially effective when the desired edit should preserve the original harmonic progression and rhythmic groove.

Audio Prompt Adapter. This system demonstrates a preservation profile that strongly aligns with its adapter-based conditioning design. By injecting source-audio representations into a pretrained text-to-music diffusion backbone through decoupled cross-attention, it explicitly encourages reuse of high-level contextual information from the input audio. This is reflected in its strong harmony and structure preservation, achieving competitive scores on 0.95 ChromaSim, 0.88 StructPairF, and the relatively high ARI, suggesting that harmonic content and large-scale form are largely retained during editing. However, its beat and melody preservation are notably weaker, with very large tempo deviations, high Δ BPM, and only moderate ContourDTW. We suggest that global semantic audio embeddings provide weaker constraints on fine-grained temporal alignment and exact melodic trajectories during diffusion sampling. Overall, Audio Prompt Adapter represents a strong solution for high-level context preservation, while highlighting the need for future editing systems to better model local temporal and melodic consistency.

Instruct-MusicGen. Instruct-MusicGen shows a facet-dependent preservation pattern under instrument-level editing. It preserves harmony and structure relatively well, but shows clear degradation in rhythm & meter and melodic content, which is consistent with its parameter-efficient architecture. Instruct-MusicGen extends MusicGen with an audio fusion module and a text fusion module, while keeping most pretrained components frozen and only training lightweight fusion parameters and LoRA modules in the text cross-attention. Such a design is effective for injecting the source audio and instruction

into a strong pretrained MusicGen backbone, which helps retain coarse musical context such as global tonal distribution and segment-level organization. However, because the model is not explicitly constrained to align beat positions, tempo, or melodic trajectories with the input, the autoregressive generation process can still introduce timing drift and local melodic changes, leading to poor Δ BPM, BeatF, and IG, and lower ContourDTW/MotifRec scores. The melody degradation should also be interpreted together with the task setting. In adding, removing, and extracting instrument editing, the target stem may carry part of the main melody, so a semantically correct edit can naturally alter melodic contours and motif statistics. Overall, these results suggest that Instruct-MusicGen is effective for instruction-following stem manipulation and coarse context retention, but its lightweight fusion strategy is less sufficient for fine-grained MuseCP, especially for time-locked rhythmic structure and melody preservation.

5. CONCLUSION

In this work, we introduced MuseCPEval, the 1st music context preservation evaluation benchmark to measure the capability of music editing systems to preserve music attributes that are not intended to be changed during editing. 10 metrics covering four music facets are collected and tailored for comprehensive evaluation. Through objective validation and a human listening study, it is shown that these metrics are effective at capturing fine-grained changes in music attributes for MuseCP evaluation. The case studies on four representative music editing systems with MuseCPEval offer practical insights. We hope our framework can serve as a testbed to understand the cons and pros of music editing systems and provide insights for constructing more reliable editing systems in the future.

6. AI USAGE AND ETHICS STATEMENT

ChatGPT is used only for limited language polishing. All technical content, experimental design, implementation, result organization, and conclusions were developed by the human authors. This work presents evaluation metrics for music editing systems and does not involve high-risk applications. For the human listening study, participation was voluntary and anonymous, and no sensitive personal data was collected. Participants were fairly compensated. All data and code used are publicly available.

7. REFERENCES

- [1] G. L. Lan, B. Shi, Z. Ni, S. Srinivasan, A. Kumar, B. Ellis, D. Kant, V. Nagaraja, E. Chang, W. Hsu, Y. Shi, and V. Chandra, “High fidelity text-guided music editing via single-stage flow matching,” *CoRR*, vol. abs/2407.03648, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2407.03648>
- [2] S. Hou, S. Liu, R. Yuan, W. Xue, Y. Shan, M. Zhao, and C. Zhang, “Editing music with melody and text: Using controlnet for diffusion transformer,” in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*. IEEE, 2025, pp. 1–5. [Online]. Available: <https://doi.org/10.1109/ICASSP49660.2025.10890309>
- [3] P. Ramoneda, M. Rocamora, and T. Akama, “Music proofreading with refinpaint: Where and how to modify compositions given context,” in *Proceedings of the 25th International Society for Music Information Retrieval Conference, ISMIR 2024, San Francisco, California, USA and Online, November 10-14, 2024*, B. Kaneshiro, G. J. Mysore, O. Nieto, C. Donahue, C. A. Huang, J. H. Lee, B. McFee, and M. C. McCallum, Eds., 2024, pp. 240–247. [Online]. Available: <https://doi.org/10.5281/zenodo.14877319>
- [4] L. Zhang, “Not that groove: Zero-shot symbolic music editing,” *CoRR*, vol. abs/2505.08203, 2025. [Online]. Available: <https://doi.org/10.48550/arXiv.2505.08203>
- [5] G. Mundada, Y. Vishe, A. Namburi, X. Xu, Z. Novack, J. McAuley, and J. Wu, “WildScore: Benchmarking MLLMs in-the-wild symbolic music reasoning,” in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 16 847–16 863. [Online]. Available: <https://aclanthology.org/2025.emnlp-main.853/>
- [6] B. Wang, Y. Vishe, X. Xu, Z. Novack, X. Jiang, J. McAuley, and J. Wu, “Csymr: Benchmarking compositional music information retrieval in symbolic music reasoning,” *arXiv preprint arXiv:2601.11556*, 2025.
- [7] H. Manor and T. Michaeli, “Zero-shot unsupervised and text-based audio editing using DDPM inversion,” in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, ser. Proceedings of Machine Learning Research, R. Salakhutdinov, Z. Kolter, K. A. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Eds. PMLR / OpenReview.net, 2024, pp. 34 603–34 629. [Online]. Available: <https://proceedings.mlr.press/v235/manor24a.html>
- [8] F. Tsai, S. Wu, H. Kim, B. Chen, H. Cheng, and Y. 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 2024, San Francisco, California, USA and Online, November 10-14, 2024*, B. Kaneshiro, G. J. Mysore, O. Nieto, C. Donahue, C. A. Huang, J. H. Lee, B. McFee, and M. C. McCallum, Eds., 2024, pp. 634–641. [Online]. Available: <https://doi.org/10.5281/zenodo.14877417>
- [9] X. Niu, K. W. Cheuk, J. Zhang, N. Murata, C. Lai, M. Mancusi, W. Choi, G. Fabbro, W. Liao, C. P. Martin, and Y. Mitsufuji, “Steermusic: Enhanced musical consistency for zero-shot text-guided and personalized music editing,” in *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, S. Koenig, C. Jenkins, and M. E. Taylor, Eds. AAAI Press, 2026, pp. 2000–2010. [Online]. Available: <https://doi.org/10.1609/aaai.v40i3.37181>
- [10] Y. Wang, Z. Ju, X. Tan, L. He, Z. Wu, J. Bian, and sheng zhao, “AUDIT: Audio editing by following instructions with latent diffusion models,” in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: <https://openreview.net/forum?id=EO1KuHoR0V>
- [11] B. Han, J. Dai, W. Hao, X. He, D. Guo, J. Chen, Y. Wang, Y. Qian, and X. Song, “Instructme: An instruction guided music edit framework with latent diffusion models,” in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*. ijcai.org, 2024, pp. 5835–5843. [Online]. Available: <https://www.ijcai.org/proceedings/2024/645>
- [12] Y. Zhang, Y. Ikemiya, G. Xia, N. Murata, M. A. M. Ramírez, W. Liao, Y. Mitsufuji, and S. Dixon, “Music-magus: Zero-shot text-to-music editing via diffusion models,” in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*. ijcai.org, 2024, pp. 7805–7813. [Online]. Available: <https://www.ijcai.org/proceedings/2024/864>

- [13] Y. Zhang, Y. Ikemiya, W. Choi, N. Murata, M. A. M. Ramírez, L. Lin, G. Xia, W. Liao, Y. Mitsufuji, and S. 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 2025, Daejeon, South Korea, September 21-25, 2025*, J. Nam, D. Jeong, K. Choi, L. Su, M. Fuentes, T. Nakano, X. Hu, and H. H. Dong, Eds., 2025, pp. 328–336. [Online]. Available: <https://doi.org/10.5281/zenodo.17811375>
- [14] Y. Yang, H. Li, T. Li, B. Cao, X. Zhang, L. Chen, and Q. Liu, “Melodia: Training-free music editing guided by attention probing in diffusion models,” in *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, S. Koenig, C. Jenkins, and M. E. Taylor, Eds. AAAI Press, 2026, pp. 2209–2217. [Online]. Available: <https://doi.org/10.1609/aaai.v40i3.37204>
- [15] F. B. Kader and S. Karmaker, “A survey on evaluation metrics for music generation,” 2025. [Online]. Available: <https://doi.org/10.48550/arXiv.2509.00051>
- [16] Y. Zhu and M. S. Kankanhalli, “Precise pitch profile feature extraction from musical audio for key detection,” *IEEE Trans. Multim.*, vol. 8, no. 3, pp. 575–584, 2006. [Online]. Available: <https://doi.org/10.1109/TMM.2006.870727>
- [17] C. L. Krumhansl, *Cognitive Foundations of Musical Pitch*. Oxford University Press, 11 2001. [Online]. Available: <https://doi.org/10.1093/acprof:oso/9780195148367.001.0001>
- [18] C. Schoerhuber and A. Klapuri, “Constant-q transform toolbox for music processing,” 7th Sound and Music Computing Conference, 2010. [Online]. Available: <https://api.semanticscholar.org/CorpusID:12358579>
- [19] H. Sakoe and S. Chiba, *Dynamic programming algorithm optimization for spoken word recognition*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990, p. 159–165.
- [20] 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 *Proceedings of the 14th Python in Science Conference, SciPy 2015, Austin, Texas, USA, July 6-12, 2015*, K. D. Huff and J. Bergstra, Eds. scipy.org, 2015, pp. 18–24. [Online]. Available: <https://doi.org/10.25080/Majora-7b98e3ed-003>
- [21] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, “Mir_eval: A transparent implementation of common MIR metrics,” in *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, H. Wang, Y. Yang, and J. H. Lee, Eds., 2014, pp. 367–372. [Online]. Available: http://www.terasoft.com.tw/conf/ismir2014/proceedings/T066_320_Paper.pdf
- [22] O. Nieto and J. P. Bello, “Systematic exploration of computational music structure research,” in *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York City, United States, August 7-11, 2016*, M. I. Mandel, J. Devaney, D. Turnbull, and G. Tzanetakis, Eds., 2016, pp. 547–553. [Online]. Available: https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/043_Paper.pdf
- [23] W. M. Rand, “Objective criteria for the evaluation of clustering methods,” *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971. [Online]. Available: <http://www.jstor.org/stable/2284239>
- [24] R. M. Bittner, J. Salamon, S. Essid, and J. P. Bello, “Melody extraction by contour classification,” in *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26-30, 2015*, M. Müller and F. Wiering, Eds., 2015, pp. 500–506. [Online]. Available: http://ismir2015.uma.es/articles/247_Paper.pdf
- [25] J. C. Brown, “Calculation of a constant q spectral transform,” *Journal of the Acoustical Society of America*, vol. 89, pp. 425–434, 1991. [Online]. Available: <https://api.semanticscholar.org/CorpusID:44239681>
- [26] K. Frieler, D. Basaran, F. Höger, H. Crayencour, G. Peeters, and S. Dixon, “Don’t hide in the frames: Note- and pattern-based evaluation of automated melody extraction algorithms,” in *6th International Conference on Digital Libraries for Musicology, DLfM 2019, The Hague, The Netherlands, November 19, 2019*, D. Rizo, Ed. ACM, 2019, pp. 25–32. [Online]. Available: <https://doi.org/10.1145/3358664.3358672>
- [27] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, “Audioldm 2: Learning holistic audio generation with self-supervised pretraining,” *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 2871–2883, 2024. [Online]. Available: <https://doi.org/10.1109/TASLP.2024.3399607>
- [28] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, “Medleydb: A multitrack dataset for annotation-intensive MIR research,” in *ISMIR*, 2014, pp. 155–160.

- [29] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, “Simple and controllable music generation,” in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: <https://openreview.net/forum?id=jtiQ26sCJi>
- [30] E. Manilow, G. Wichern, P. Seetharaman, and J. L. Roux, “Cutting music source separation some slakh: A dataset to study the impact of training data quality and quantity,” in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2019, New Paltz, NY, USA, October 20-23, 2019*. IEEE, 2019, pp. 45–49. [Online]. Available: <https://doi.org/10.1109/WASPAA.2019.8937170>
- [31] M. I. Mandel and D. P. W. Ellis, “Song-level features and support vector machines for music classification,” in *International Society for Music Information Retrieval Conference*, 2005. [Online]. Available: <https://archives.ismir.net/ismir2005/paper/000106.pdf>
- [32] B. Logan, “Mel frequency cepstral coefficients for music modeling,” in *ISMIR 2000, 1st International Symposium on Music Information Retrieval, Plymouth, Massachusetts, USA, October 23-25, 2000, Proceedings*, 2000. [Online]. Available: http://ismir2000.ismir.net/papers/logan_paper.pdf

A. TIMBRE PRESERVATION EVALUATION

Thanks to ISMIR reviewers’ suggestions, we also explore timbre context preservation. Timbre is the attribute by which listeners distinguish two sounds of the same pitch, loudness, and duration, and it is the primary cue for identifying which instrument is playing. We evaluate timbre context preservation, i.e., whether x' still sounds as though it were played on the same instruments as x .

SKL Similarity. This metric compares the global timbral content of x and x' [31]. Concretely, x and x' are each converted into an MFCC [32] feature sequence $\mathbf{M} \in \mathbb{R}^{T \times F}$ and $\mathbf{M}' \in \mathbb{R}^{T' \times F}$, where the 0-th cepstral coefficient is discarded because it encodes overall energy rather than spectral shape, and the remaining 12 coefficients are stacked with their first and second temporal derivatives, giving $F = 36$. Each sequence is summarized by a single multivariate Gaussian fitted to its frames, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathcal{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$, where $\boldsymbol{\mu} \in \mathbb{R}^F$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{F \times F}$ are the sample mean and covariance of the rows of $\mathbf{M}$, and likewise for $\mathbf{M}'$. The timbral distance is the symmetrized KL divergence between them, which admits the closed form $\text{SKL}(x, x') = \frac{1}{2} \left[\text{tr}(\boldsymbol{\Sigma}'^{-1} \boldsymbol{\Sigma}) + \text{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}') + \delta^\top (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}'^{-1}) \delta - 2F \right]$ where $\delta = \boldsymbol{\mu}' - \boldsymbol{\mu}$ and the log-determinant terms cancel in the symmetric sum. Since SKL is unbounded and heavy-tailed, we map it to a bounded similarity:

$$\text{SKLSim}(x, x') = \frac{1}{1 + \text{SKL}(x, x')/50} \quad (9)$$

It lies in $[0, 1]$, where 1 indicates identical timbral distributions and values near 0 indicate that the two recordings occupy disjoint regions of timbre space.

Mean MFCC Cosine. To grade the magnitude of editing, we additionally compare the average spectral envelopes of x and x' . Let $\bar{\mathbf{m}} = \frac{1}{T} \sum_{t=1}^T \mathbf{m}_t$ and $\bar{\mathbf{m}}'$ denote the time-averaged MFCC frame vectors of x and x' . The Mean MFCC Cosine is their cosine similarity:

$$\text{MMCos}(x, x') = \frac{\bar{\mathbf{m}} \cdot \bar{\mathbf{m}}'}{\|\bar{\mathbf{m}}\| \|\bar{\mathbf{m}}'\|} \quad (10)$$

It lies in $[-1, 1]$, where 1 means the two recordings share the same average spectral envelope.

In addition to the editing curated in Section 3.1, we also tailor two editing for timbre context preservation evaluation: changing the instrument, the grand piano, into other instruments, the electric piano and acoustic guitar. Table 5 shows the results.

B. OBJECTIVE VALIDATION SETUPS

This section provides the details of the objective validation referenced in Section 3.1.

B.1 Definition of the Editing Operations

Table 6 gives the transformation applied by each of the eight operations abbreviated in Table 2 and Table 5, together with the facet it targets.

B.2 Evaluation Set and Audio Rendering

The 50 MIDI files are drawn from the Lakh MIDI Dataset⁷ and selected to exhibit a clearly identifiable main melodic line, a stable metrical structure, and a recognizable sectional form, so that a facet-targeted edit yields a well-defined change in the corresponding facet. Editing in the symbolic domain is deliberate. It lets each operation alter exactly one intended facet while leaving the others untouched by construction, which is the controlled condition required to test whether a metric responds only to the facet it is designed to measure. The main melody used by the two melodic operations is taken to be the monophonic track with the highest mean pitch among the tracks carrying continuous note activity, and the three sections used by the two structural operations are taken from the sectional annotation of each source, so that both are defined on the symbolic score rather than recovered from audio. Each source yields ten edited variants, giving $50 \times 10 = 500$ origin-edited pairs. Both the original and the edited MIDI are then rendered to audio with FluidSynth⁸ using the FluidR3_GM general MIDI soundfont, so that the reference and the edited signal share an identical instrument mapping and any measured difference is attributable to the edit rather than to synthesis.

⁷ <https://colinraffel.com/projects/lmd/>

⁸ <https://github.com/fluidsynth/fluidsynth>

Editing	Timbre			
	SKLSim↑	Expected	MMCos↑	Expected
Grand Piano → Electric Piano	0.16 ± 0.15	< 1	0.78 ± 0.07	< 1
Grand Piano → Acoustic Guitar	0.35 ± 0.12	< 1	0.90 ± 0.04	< 1
+7 semitones (+Perfect Fifth)	0.72 ± 0.13	≈ 1	0.97 ± 0.05	≈ 1
-3 semitones (- Minor Third)	0.84 ± 0.11	≈ 1	0.99 ± 0.02	≈ 1
ABC→AAA	0.94 ± 0.13	≈ 1	1.00 ± 0.01	≈ 1
ABC→ABA	0.97 ± 0.11	≈ 1	1.00 ± 0.01	≈ 1
Tempo shifted +50%	0.87 ± 0.13	≈ 1	0.99 ± 0.02	≈ 1
All beats shifted later by a constant -150ms	0.97 ± 0.11	≈ 1	1.00 ± 0.01	≈ 1
Interval direction flipped for the main melody	0.87 ± 0.16	≈ 1	0.99 ± 0.02	≈ 1
The climax phrase of the main melody is inverted	0.98 ± 0.11	≈ 1	1.00 ± 0.01	≈ 1

Table 5: Objective validation for the timbre metrics, where the two instrument-replacement editings are referenced against the Grand Piano render of each sample.

Editing Operation	Target Facet	Description
+7 semitones (Perfect Fifth)	Harmony	Transpose every pitch up by a perfect fifth. The tonic moves one step clockwise on the circle of fifths, while note timing, sectional layout, and melodic shape are unchanged.
-3 semitones (Minor Third)	Harmony	Transpose every pitch down by a minor third, moving the tonic three steps along the circle of fifths while preserving all non-harmonic facets.
ABC→AAA	Structure	Replace the second and third sections with repetitions of the first, collapsing a three-part form into a single repeated block.
ABC→ABA	Structure	Reorder the sections into a ternary reprise form, changing the sectional layout while reusing the original material.
Tempo shifted +50%	Rhythm & Meter	Scale the global tempo by a factor of 1.5, compressing all inter-onset intervals while leaving pitch content, harmony, and sectional form intact.
All beats displaced by a constant 150 ms	Rhythm & Meter	Apply the same timing offset to every beat, introducing a fixed phase shift without changing tempo, note content, or note durations.
Interval direction flipped for the main melody	Melody & Motif	Invert the sign of every melodic interval in the main melody, reversing the contour direction while keeping interval magnitudes and rhythm.
Climax phrase of the main melody inverted	Melody & Motif	Melodically invert only the climax phrase, a localized contour change confined to a single salient region.
Grand Piano → Electric Piano	Timbre	Set the general MIDI program number of every pitched track to Grand Piano and then change it into Electric Piano, leaving the percussion track and all note content, timing, and sectional layout unchanged.
Grand Piano → Acoustic Guitar	Timbre	Set the general MIDI program number of every pitched track to Grand Piano and then change it into Acoustic Guitar, leaving the percussion track and all note content, timing, and sectional layout unchanged.

Table 6: Definition of the ten editing operations used for objective validation. Each operation targets one single music facet and leaves the remaining facets unchanged.

B.3 Derivation of the Expected Values

The *Expected* entries in Table 2 are derived from the definition of each metric under a known edit rather than chosen post hoc. We give the derivation per facet.

Harmonic Transpositions. A perfect fifth advances the tonic by one position on the circle of fifths, so Equation 1 gives $\frac{1}{6} \min(1, 11) \approx 0.167$. A downward minor third advances it by three positions, giving $\frac{3}{6} = 0.5$ whenever the key estimate is correct for both signals; the stated upper end of 0.67 accommodates the cases in which the Krumhansl–Schmuckler estimate lands on an adjacent key, which costs one further step. Both transpositions rotate the pitch-class distribution without changing its shape, so ChromaSim and ChromaDTWS should fall below 1 while remaining well above 0. Neither operation touches onset times or sectional layout, so the rhythmic and structural

metrics should stay near their preservation ceiling.

Structural Reorderings. Reordering or duplicating sections reuses the original material, so the global key, the pitch-class distribution, and the global tempo are unchanged, and the harmonic metrics together with ΔBPM should remain at their preservation ceiling. Beat positions, however, are displaced past the first reordered boundary, so BeatF and IG should fall below 1 even though no local rhythmic content is altered.

Rhythmic Perturbations. A +50% tempo scaling multiplies the underlying pulse rate by 1.5, which is not a metric-octave relation and therefore survives the folding in Equation 4, so ΔBPM should be large, and both BeatF and IG should fall. A constant offset leaves the pulse rate untouched, so ΔBPM will be near 0. Because 150ms exceeds the $\pm 70\text{ms}$ matching tolerance, BeatF will collapse

toward 0, while IG will remain moderate, since a fixed offset concentrates the timing error at a single phase rather than spreading it uniformly.

Melodic Transformations. Contour inversion is applied to the main melody track only, which accounts for a median of 9% of the notes in the mix. It is a chromatic inversion and therefore does alter the melody’s pitch-class content, but the harmonic, rhythmic, and structural metrics read the full polyphonic mix and are dominated by the unchanged accompaniment, so they should remain high. ContourDTWS will fall, since inversion changes the pitch-class trajectory of the melody. MotifRec will fall, since inverted interval 3-grams are distinct tokens from the originals. The localized climax inversion perturbs only a small fraction of the interval sequence, so it will move both melodic metrics less than the global interval flip.

C. HUMAN LISTENING QUESTIONNAIRE

We design a questionnaire to collect human auditory perception of the tailored editing settings⁹. The questionnaire comprises paired edited music clips, organized as 4 questions per music facet. Each question presents an original music clip together with two edited versions of that clip. To validate the metric’s ability to capture musical nuances that align with human auditory perception, the two edited clips are produced by applying the same facet-targeted operation at two different strengths. We ask the participants to compare the two edited clips and decide which one is farther from the original one with respect to one specific facet. We also construct samples for a *sanity check* to filter out low-quality responses. Participants are anonymous and have diverse music backgrounds. We received 33 questionnaire responses and retained the 11 that passed our sanity checks for the final analysis in Section 3.2.

⁹ <https://github.com/Yashvishe13/MuseCPEval/blob/main/questionnaire.pdf>