CLEF: LEARNING CLOSED-LOOP CONTROL FOR TEXT-GUIDED MUSIC EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we propose CLEF, a framework that learns closed-loop control for text-guided music editing. Our key idea is to use the outcome of each editing round to adjust the guidance for the next, preserving the source audios structure while faithfully realizing the requested edits. Specifically, we construct two complementary feedback signals to assess unintended structural changes and insufficient adherence to the editing instruction. Temporal structural feedback uses our proposed TACons metric to evaluate how well the source audios temporal structure is preserved. Target semantic feedback measures alignment between the edited audio and the target text using CLAP, assessing whether the requested attributes have been faithfully realized. To translate these evaluations into control over subsequent edits, we learn a feedback-aware Compensation Controller that combines both feedback signals with the current editing state to predict the direction and magnitude of updates to the target textual conditioning representations. At inference, the updated conditioning guides the next editing round, whose output is evaluated again to inform subsequent compensation. This closes the feedback loop and allows the same trained controller to refine the conditioning over multiple rounds, adapting the editing process to the observed structural and semantic discrepancies. Moreover, CLEF can flexibly incorporate an additional reference feedback signal to assess attribute consistency with reference audio, further enhancing personalized editing. Extensive experiments demonstrate that CLEF consistently outperforms state-of-the-art baselines, achieving a 23.78%improvement in CQT1-PCC and a 4.18% gain in CLAP, along with superior subjective listening quality.

1 INTRODUCTION

Recent advances in text-to-music generation, exemplified by MusicGen (Copet et al., 2023), AudioLDM 2 (Liu et al., 2024a), and MusicLM (Agostinelli et al., 2023), have enabled the synthesis of high-quality music from natural language descriptions. Beyond generating new music, practical creative workflows often involve revising an existing piece to meet specific artistic requirements. Text-guided music editing addresses this need by allowing users to modify selected musical attributes through natural language instructions (Zhang et al., 2024). For example, a user may ask to replace a piano with an acoustic guitar or change the style of a passage from jazz to classical. Such capabilities support iterative refinement in music production and the adaptation of existing tracks to the stylistic and emotional requirements of film scenes and video games (Niu et al., 2026).

Some studies train or adapt generative models on collected or synthesized editing datasets to follow textual instructions (Han et al., 2023; Zhang et al., 2025). Other methods exploit the capabilities of pretrained diffusion-based audio generation models (Liu et al., 2024a) without additional training. Methods such as ZETA (Manor & Michaeli, 2024), MusicMagus (Zhang et al., 2024), MEDIC (Liu et al., 2024b), and SteerMusic (Niu et al., 2026) use textual conditions to specify the desired semantic change and guide the editing of source music. However, achieving sufficient target-attribute modification while preserving the temporal structure of the source music remains challenging: stronger editing may disturb non-target melody and rhythm, whereas overly conservative updates may leave the requested modification insufficiently realized. In the example shown in Figure 1, ZETA achieves a pronounced attribute change with structural distortion, whereas SteerMusic better preserves the

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.

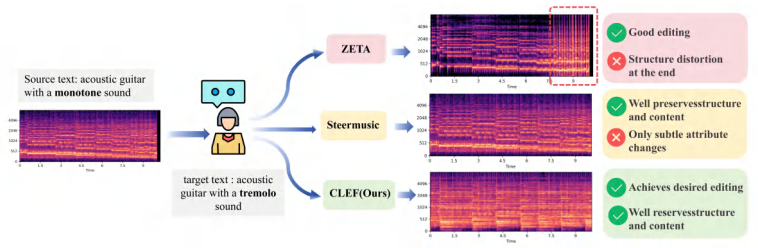

Figure 1: Spectrograms for editing an acoustic guitar track from monotone to tremolo. Source: left; edited results: right. ZETA introduces late timefrequency artifacts (red dashed box); SteerMusic preserves structure but yields weak tremolo; CLEF produces distinct tremolo-like modulation while preserving harmonic trajectories and temporal structure.

source structure but yields a weaker edit. These observations motivate adaptive editing guidance that responds to both target realization and source-structure preservation.

To better coordinate target-attribute modification and source structure preservation, we seek to adapt editing guidance according to the actual editing outcome. Although textual conditions specify the intended semantic change, they do not directly indicate how sufficiently the requested attributes have been realized or how well the source temporal structure has been preserved. This motivates incorporating semantic and structural evaluations of the edited output as feedback to inform adjustments to the target textual condition. Since a condition adjustment can affect both attribute realization and structural preservation, its effect needs to be reassessed through the resulting audio. We therefore formulate music editing as a closed-loop process of editing, evaluation, and condition compensation, in which feedback from each editing outcome guides the conditioning update for the subsequent round.

Based on this formulation, we propose CLEF, a closed-loop framework built on a Delta Denoising Score (DDS)-based music editor (Hertz et al., 2023; Niu et al., 2026). During training, we evaluate an initial editing result using two complementary feedback signals: temporal structural preservation measured by our proposed DTW-based Temporal Alignment Consistency (TACons), and target semantic alignment measured by CLAP (Elizalde et al., 2023). TACons temporally aligns source and edited music representations and emphasizes low-consistency regions to reveal local structural deviations. For personalized editing, we additionally incorporate reference feedback to assess consistency with the reference audio. We train a lightweight Feedback-aware Compensation Controller (FACC) that uses these feedback signals and the current editing context to predict the direction and magnitude of target-condition compensation. The pretrained diffusion model remains frozen throughout training. At inference, the same trained controller can be reused over multiple edit–evaluate–compensate rounds. Each editing outcome provides updated feedback for the next conditioning adjustment, and each subsequent round re-edits the original source music under the updated condition, enabling the editing guidance to adapt to the observed semantic and structural discrepancies.

We evaluate CLEF on ZoME-Bench (Niu et al., 2026) and MusicBench (Melechovsky et al., 2024) for text-guided and personalized music editing. Objective and subjective evaluations show improved target-attribute realization and source-structure preservation over SteerMusic and SteerMusic+.

Our main contributions are summarized as follows:

- **A closed-loop editing framework.** We propose CLEF, which uses feedback from editing outcomes to iteratively refine target textual conditions and coordinate attribute modification with source-structure preservation.
- **Feedback design and controller learning.** We introduce TACons for temporal structural feedback and combine it with semantic and optional reference feedback to train a lightweight condition-compensation controller.
- **Improved editing performance.** Experiments on two benchmarks show gains in target-attribute realization and source-structure preservation, supported by objective metrics and subjective listening evaluations.

2 Related Work

Learning-Based Music Editing Learning-based methods link textual instructions to source music through trained editing models or conditioning modules. InstructME (Han et al., 2023) learns from paired data, using multi-scale source features and chord conditioning to preserve harmony. Instruct-MusicGen (Zhang et al., 2025) adapts pretrained MusicGen with audio–text fusion modules and synthesized instructions for stem addition, removal, and extraction. Audio Prompt Adapter (Tsai et al., 2024) injects source-audio features into frozen AudioLDM2 via lightweight cross-attention adapters. SmartDJ (Lan et al., 2026) trains an audio-language planner and diffusion editor to decompose instructions into sequential atomic operations. Related controllable music-generation methods learn explicit musical-feature controls. Music ControlNet (Wu et al., 2024) provides time-varying melody, dynamics, and rhythm controls, while *Editing Music with Melody and Text* (Hou et al., 2025) trains a melody-conditioned ControlNet on a frozen diffusion transformer for text-guided editing and style transfer. MuseControlLite (Tsai et al., 2025) learns lightweight conditioners for temporal features and reference audio, supporting generation, inpainting, and outpainting. Together, these methods provide learned interfaces for specifying musical content and editing requirements.

Training-Free Music Editing Training-free methods adapt pretrained generative models for editing without further training. Inversion-based methods recover source-specific noise or latent trajectories before target-text guidance. ZETA (Manor & Michaeli, 2024) extracts noise via DDPM inversion and reuses them for sampling under a target prompt. MusicMagus (Zhang et al., 2024) constructs a source-to-target text-embedding direction and constrains cross-attention consistency to preserve non-target content; for real-audio editing, it uses DDIM inversion. MEDIC (Liu et al., 2024b) combines source-branch trajectory correction with coordinated self- and cross-attention control to improve editing fidelity and content preservation. For flow matching, MelodyFlow (Le Lan et al., 2024) extends inversion-based editing with regularized latent inversion to improve reversibility and source consistency. Inversion-free methods edit without reconstructing the source inversion trajectory. SteerMusic (Niu et al., 2026) uses Delta Denoising Score (DDS) (Hertz et al., 2023) guidance the difference between source- and target-conditioned denoising predictions to directly optimize the edited audio representation. In broader audio editing, DirectAudioEdit (Ge et al., 2026) contrasts source- and target-conditioned reverse diffusion updates under shared-noise re-noising. EchoEdit (Fu et al., 2026) constructs a direct editing field from source- and target-conditioned drift differences, with an optimal-transport extension to stabilize the trajectory. CLEF builds on a frozen DDS-based editor and learns a lightweight controller mapping semantic and temporal structural feedback from editing outcomes to target-condition compensation. The trained controller is reused across editing rounds, allowing subsequent guidance to respond to the observed output.

3 Preliminaries

CLEF builds upon the score-distillation editing framework of SteerMusic (Niu et al., 2026). Given source music X_{src}, source description y^{src}, and target instruction y^{tgt}, the editable music is represented as $x = g(\theta)$, where $g(\theta) = \theta$ and θ is initialized with X_{src}. SteerMusic optimizes the editable representation using Delta Denoising Score (DDS) (Hertz et al., 2023), whose gradient is

$$\nabla_\theta \mathcal{L}_{\mathrm{DDS}} = \mathbb{E}_{t,\epsilon}\left[w(t) \left(\epsilon_\phi(x_t, y^{\mathrm{tgt}}, t) - \epsilon_\phi(x_t^{\mathrm{src}}, y^{\mathrm{src}}, t) \right) \frac{\partial x}{\partial \theta} \right],$$

where x_t and x_t^{src} share the same timestep t and noise ϵ. The target- and source-conditioned predictions guide the desired modification and source preservation, respectively.

For personalized editing, SteerMusic+ extends DDS to Personalized Delta Score (PDS) via a personalized diffusion model $\epsilon_{\phi'}$, fine-tuned on reference audio with concept token [S]. PDS replaces the target-conditioned prediction with the personalized model while retaining the original model for source guidance, yielding:

$$\nabla_\theta \mathcal{L}_{\mathrm{PDS}} = \mathbb{E}_{t,\epsilon}\left[w(t) \left(\epsilon_{\phi'}(x_t, y^{\mathrm{tgt}}, t) - \epsilon_\phi(x_t^{\mathrm{src}}, y^{\mathrm{src}}, t) \right) \frac{\partial x}{\partial \theta} \right].$$

We adopt DDS and PDS as the underlying editing objectives for zero-shot and personalized music editing, respectively, while freezing diffusion backbones.

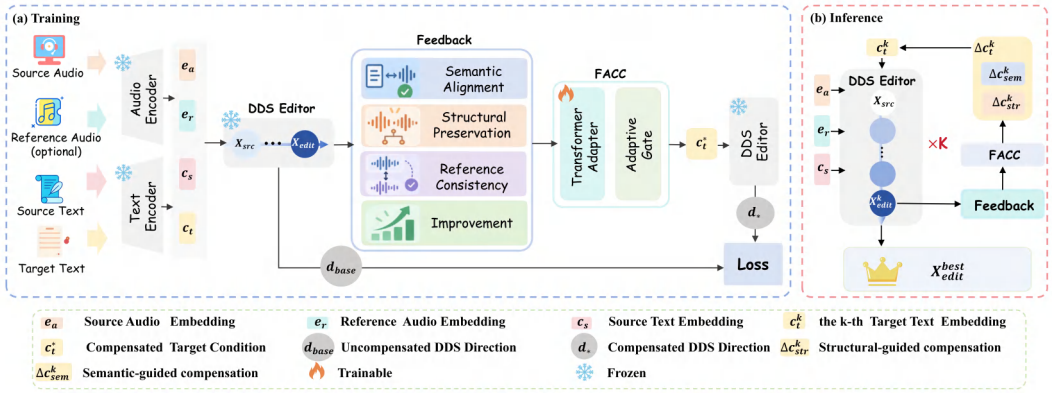

Figure 2: Overview of the proposed CLEF framework. (a) Training: multi-dimensional editing feedback guides FACC to learn compensatory updates to the target textual condition while keeping the pretrained DDS editor frozen. (b) Inference: iterative feedback evaluation and condition compensation loop for K rounds to progressively refine the edit.

4 METHOD

4.1 OVERVIEW

Built upon SteerMusic (Niu et al., 2026), CLEF adapts the target textual condition using editing feedback while keeping the pretrained editor frozen, as illustrated in Fig. 2.Given the source music, source and target descriptions, and optional reference audio, the editor first generates an initial result under the target condition. Its multi-dimensional feedback module evaluates each edit, and the Feedback-aware Compensation Controller (FACC) predicts a gated residual correction. FACC learns from a local objective and compensation regularization during training(Fig. 2(a)), then runs an iterative editevaluatecompensate loop at inference (Fig. 2(b)). Each round re-edits the source music under the updated condition; DDS and PDS support text-guided and personalized editing, respectively.

4.2 MULTI-DIMENSIONAL EDITING FEEDBACK

The feedback representation F^k combines temporal structural, target semantic, and optional reference feedback to assess the result of the k-th editing round.

For temporal structural feedback, we introduce DTW-based Temporal Alignment Consistency (TACons). Given the source music X_{src} and the k-th edited result X^k_{edit}, we extract frame-level representations using MERT (Li et al., 2024). We then apply Dynamic Time Warping (DTW) (Sakoe & Chiba, 1978) within a constrained temporal neighborhood to accommodate local alignment variations while restricting excessive warping. Let $\pi^k = \{(i_l, j_l)\}_{l=1}^{L_k}$ denote the resulting alignment path. Averaging similarity over the entire path can obscure deviations confined to short regions. We therefore also evaluate the least-consistent aligned positions. Specifically, let Q^k_q denote the indices of the lowest $q\%$ cosine similarities along π^k. We define

$$r^k_{\text{TACons}_q} = \frac{1}{|Q^k_q|} \sum_{l \in Q^k_q} \cos\big(f_{i_l}(X_{\text{src}}), f_{j_l}(X^k_{\text{edit}})\big)\,, \tag{1}$$

where $f_i(\cdot)$ denotes the i-th MERT representation. Setting $q = 100\%$ measures average consistency over the full alignment path, whereas $q = 10\%$ emphasizes localized deviations. We denote these two measurements as r^k_{TACons} and $r^k_{\text{TACons}_{0.1}}$, and combine them as

$$r^k_{\text{str}} = \alpha r^k_{\text{TACons}} + (1-\alpha) r^k_{\text{TACons}_{0.1}}. \tag{2}$$

where α balances global consistency and sensitivity to low-consistency regions. We use $\alpha = 0.8$ throughout the experiments, with its sensitivity analyzed in Appendix C.2.

Target semantic feedback measures alignment between the edited audio and the target text using CLAP (Elizalde et al., 2023), computed as $r^k_{\text{sem}} = \cos(E_{\text{CLAP}}(X^k_{\text{edit}}), E_{\text{CLAP}}(y_{\text{tgt}}))$. For personalized editing, we additionally compute reference feedback as $r^k_{\text{ref}} = \cos(E_r(X^k_{\text{edit}}), E_r(X_{\text{ref}}))$, where E_r extracts the audio representations used for reference comparison.

The feedback vector for the k-th round is $\mathbf{r}^k = [r^k_{\text{sem}},\, r^k_{\text{str}},\, m_r r^k_{\text{ref}}]$, where $m_r \in \{0, 1\}$ indicates the availability of reference feedback. We also include the inter-round feedback difference $\Delta r^k = r^k - r^{k-1}$, with $\Delta r^1 = 0$. The final feedback representation is defined as $F^k = [\mathbf{r}^k,\, \Delta r^k,\, m_r] \in \mathbb{R}^7$.

4.3 Feedback-aware Compensation Controller

FACC performs feedback control in the target conditioning space by predicting a compensation residual and an adaptive gating coefficient.

The controller receives the current target condition c^k_t, the source textual condition c_s, the source-audio representation e_a, and local DDS state representation h^k_n extracted from the frozen U-Net. Modality-specific pooling and projection map these inputs into a common-dimensional token space. For personalized editing, the reference representation e_r is processed by cross-attention, with the projected target condition as the query and reference features as keys and values. A lightweight Transformer Adapter P_ϕ jointly processes the resulting tokens to predict

$$u^k = P_\phi(c^k_t, c_s, e_a, m_r e_r, F^k, h^k_n), \tag{3}$$

where $m_r \in \{0, 1\}$ indicates the availability of reference audio, , c^k_t represents the target textual conditioning, u^k denotes the residual update applied to the target condition. Furthermore, An adaptive feedback gate G_ψ uses the feedback and local DDS state to determine the compensation coefficient:

$$\lambda^k = \lambda_{\max}\sigma\big(G_\psi(F^k, h^k_n)\big)\,, \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function and $\lambda_{\max}$ bounds the gating coefficient. The resulting compensation is $\Delta c^k_t = \lambda^k u^k$, and the target condition is updated as $c^{k+1}_t = c^k_t + \Delta c^k_t$. The residual update preserves the original target condition while incorporating feedback-based correction.

During training, FACC learns feedback-to-conditioning compensation without ground-truth condition corrections. For each sample, the base edit X_{base} provides initial feedback F_{base}, with the inter-round difference initialized to zero. We construct matched uncompensated and compensated DDS branches from the same source-target pair using identical timestep and noise realization. At timestep n, the DDS directions are defined as $d_{\text{base}} = \text{sg}[\epsilon_\theta(x_n, c_t, n) - \epsilon_\theta(x^{\text{src}}_n, c_s, n)]$ and $d_* = \epsilon_\theta(x_n, c^*_t, n) - \text{sg}[\epsilon_\theta(x^{\text{src}}_n, c_s, n)]$, where c^*_t is the compensated target condition, ϵ_θ is the frozen backbone and $\text{sg}[\cdot]$ denotes stop-gradient. Matching the state, timestep, and noise isolates the effect of condition compensation on the local DDS direction. We weight a local control objective using feedback deficiencies: for $j \in \{\text{sem}, \text{str}\}$, the adaptive weight is defined as $w_j = [\tau_j - r^{\text{base}}_j]_+ / \max(|\tau_j|, \epsilon)$. where $[z]_+ = \max(z, 0)$,τ_j is the corresponding feedback target, and $\epsilon > 0$ ensures numerical stability. We construct two local control utilities from the compensated DDS norm: $\tilde{s}_{\text{edit}} = \tanh(\|d_*\|_2/\tau_e)$ and $\tilde{s}_{\text{pres}} = \exp[-\tau_p(\|d_*\|_2/\max(\|h_{\text{src}}\|_2, \epsilon))^2]$, where τ_e and τ_p control their sensitivity and the source-dependent term supplies the normalization scale. The feedback loss is defined as $\mathcal{L}_{\text{fb}} = -\left(w_{\text{sem}}\tilde{s}_{\text{edit}} + w_{\text{str}}\tilde{s}_{\text{pres}}\right)$. Minimizing the semantic term favors a larger local DDS norm when target alignment is insufficient, whereas minimizing the structural term favors a smaller norm when source preservation is insufficient. These utilities regulate local update strength; they are not estimates of CLAP or TACons. The original target-conditioned DDS direction provides a local editing prior for this adjustment. We encourage compensation to remain consistent with that direction and penalize excessive relative magnitude: $\mathcal{L}_{\text{dir}} = 1 - \cos(d_*, d_{\text{base}})$ and $\mathcal{L}_{\text{mag}} = [\frac{\|d_*\|_2}{\|d_{\text{base}}\|_2+\epsilon} - \gamma]^2_+$. The rationale is that, within a neighborhood of a useful DDS editing direction, increasing update strength may alleviate under-editing, while reducing it may limit unintended changes. The regularizers encourage this regime without assuming a monotonic relationship between DDS magnitude and audio-level feedback scores.The remaining terms $\mathcal{L}_\Delta$ and $\mathcal{L}_\lambda$ regularize the compensation residual and gate activation, respectively.The overall objective is:

$$\mathcal{L} = \mathcal{L}_{\text{dir}} + \beta_{\text{mag}}\mathcal{L}_{\text{mag}} + \beta_\Delta\mathcal{L}_\Delta + \beta_\lambda\mathcal{L}_\lambda + \beta_{\text{fb}}\mathcal{L}_{\text{fb}}. \tag{5}$$

Gradients propagate through d_* and c^*_t into FACC, while F_{base}, d_{base}, and the backbone parameters remain fixed. Training therefore learns a one-step feedback-to-conditioning operator without unrolling multiple editing rounds or differentiating through the final audio-level feedback metrics.

4.4 DDS-BASED FEEDBACK CLOSED-LOOP EDITING

At inference, CLEF reuses the trained controller across editing rounds. The result X^k_{edit} provides the feedback F^k , which guides condition compensation for round $k+1$. Each round starts from the original source music under the updated conditioning. Within a round, we use a two-stage conditioning strategy to emphasize structural and semantic feedback at different stages of DDS optimization. For N optimization iterations and a transition point M, the target condition at iteration m of round $k+1$ is defined as

$$c_{\text{t}}^{k+1,m} = \begin{cases} c_t^k + \Delta c^k_{\text{str}}, & m \le M, \\ c_t^k + \Delta c^k_{\text{sem}}, & m > M, \end{cases} \tag{6}$$

where both compensations are produced by the same jointly trained FACC using different masked configurations of F^k.Structural compensation retains structural feedback while masking semantic feedback; semantic compensation retains semantic feedback while masking structural feedback. For personalized editing, reference feedback is retained in both configurations. We use the structure-to-semantics ordering with $M = N/2$.

The compensated condition enters the target-conditioned denoising prediction of the DDS update. At iteration m of editing round $k+1$, the denoising direction is

$$d^{k+1,m}_{\text{DDS}} = \epsilon_\theta\Big(x^{k+1,m}_{n_m}, c_{\text{t}}^{k+1,m}, n_m\Big) - \epsilon_\theta\big(x^{\text{src}}_{n_m}, c_s, n_m\big)\,. \tag{7}$$

where n_m is the diffusion timestep sampled at iteration m. For personalized editing, the target-conditioned branch is replaced by its PDS counterpart described in Section 3. CLEF modifies the editing trajectory through target conditioning while keeping the backbone parameters fixed.

After each round, the generated audio is evaluated again using the feedback measures 4.2. This output-level evaluation allows subsequent compensation to respond to the observed effects of previous condition updates, rather than relying on the local training utilities to predict their perceptual outcome. During multi-round inference, CLEF retains the edited result from each feedback round as a candidate and selects the final output using a fixed inference-time rule. The final candidate is selected by maximizing the following score:

$$S^k_{\text{sel}} = r^k_{\text{sem}} + r^k_{\text{str}} + m_r r^k_{\text{ref}} - \mathcal{P}^k_{\text{trade}}, \tag{8}$$

where r^k_{sem}, r^k_{str}, and r^k_{ref} are the corresponding feedback scores. The trade-off penalty is defined as $\mathcal{P}^k_{\text{trade}} = [r^{k-1}_{\text{str}} - r^k_{\text{str}}]_+ [\epsilon - (r^k_{\text{sem}} - r^{k-1}_{\text{sem}})]_+$, where ϵ is the semantic improvement margin,and We set $\mathcal{P}^1_{\text{trade}} = 0$ for the first round. The penalty is active when structural consistency decreases and the semantic gain falls below the specified margin. After all feedback rounds, CLEF returns the candidate with the highest selection score, using feedback measured from the generated audio across rounds.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets. We evaluate CLEF on ZoME-Bench (Niu et al., 2026) and MusicBench (Melechovsky et al., 2024). ZoME-Bench contains 1,000 ten-second audio clips from MusicCaps (Agostinelli et al., 2023), split into 425 training, 115 validation, and 460 test samples. The test set covers instrument, genre, mood and background editing, with 131, 100, 134 and 95 editing pairs, respectively. These categories evaluate modifications to different musical attributes under the common requirement of preserving instruction-irrelevant source content. We additionally evaluate on 400 one-minute clips from MusicBench. For personalized editing, we use 20 ten-second reference clips collected from YouTube and Freesound, covering Guitar, Sitar, Reggae and Hiphop. Reference audio specifies the desired target characteristics, while the textual instruction defines the editing scope.

Evaluation Metrics. We use FAD_{CLAP} (Kilgour et al., 2019) and $\text{FAD}_{\text{VGGish}}$ (Hershey et al., 2017) for distributional audio quality, and CLAP (Elizalde et al., 2023) for target-text alignment. Source preservation is measured by LPAPS (Hershey et al., 2017) (perceptual deviation), CQT1-PCC (Niu et al., 2026) (melodic correspondence), and TACons (consistency of temporally aligned

Table 1: Zero-shot music editing results on ZoME-Bench and MusicBench.

Method	$\text{FAD}_{\text{CLAP}}\downarrow$	$\text{FAD}_{\text{VGGish}}\downarrow$	CQT1-PCC↑	LPAPS↓	CLAP↑	TACons↑	$\text{TACons}_{0.1}\uparrow$	MOS-P↑	MOS-T↑
ZoME-Bench									
DDIM Inversion (2021)	1.1358	26.7389	0.0315	5.9443	0.0361	0.2831	0.1684	1.26	0.92
MusicGen (2023)	0.3221	4.7524	0.0911	6.6938	0.3091	0.2961	0.1715	1.70	1.45
AudioLDM2 (2024)	0.2074	5.2318	0.0008	5.0438	0.2739	0.2944	0.1687	2.17	1.20
MusicMagus (2024)	0.4382	6.3348	0.2703	3.4418	0.2804	0.5574	0.4214	2.30	1.37
ZETA (2024)	0.5755	4.0161	0.2756	4.3873	0.2661	0.5171	0.3631	1.82	1.74
MelodyFlow (2025)	0.2289	4.4365	0.4329	4.9878	**0.3107**	0.4712	0.3250	2.73	2.20
SmartDJ (2026)	0.2704	2.345	0.5346	2.9973	0.2130	0.6691	0.5407	1.97	1.67
SteerMusic (2026)	0.1401	2.8100	0.5500	2.0800	0.2800	0.6500	0.6038	2.54	2.23
CLEF	**0.0401**	**0.9202**	**0.6808**	**1.6700**	0.2917	**0.7921**	**0.6861**	**2.80**	**2.68**
MusicBench									
DDIM Inversion (2021)	1.1871	3.9506	0.0410	6.0462	0.0181	0.2864	0.1745	1.33	1.01
MusicGen (2023)	0.6600	10.0583	0.0101	6.8158	0.1759	0.2994	0.1731	1.30	1.83
AudioLDM2 (2024)	0.5205	6.3544	-0.0114	5.4846	0.0705	0.2842	0.1607	1.45	1.29
MusicMagus (2024)	0.1308	3.8667	0.1854	4.2288	0.3058	0.4969	0.3623	1.99	1.36
ZETA (2024)	0.6502	4.7226	0.3290	4.3811	0.2308	0.5206	0.3656	1.78	1.30
MelodyFlow (2025)	0.1109	5.5953	0.3398	5.7493	**0.4009**	0.6395	0.4553	**2.23**	2.60
SmartDJ (2026)	0.2208	3.4640	0.6046	2.9367	0.1646	0.6665	0.5108	1.99	1.76
SteerMusic (2026)	0.0500	0.7900	0.6500	2.8300	0.3500	0.7000	0.5621	2.01	2.39
CLEF	**0.0410**	**0.7063**	**0.7145**	**1.6924**	0.3627	**0.7982**	**0.6867**	2.21	**2.65**

Table 2: Personalized music editing results on ZoME-Bench and MusicBench.

Method	$\text{FAD}_{\text{CLAP}}\downarrow$	$\text{FAD}_{\text{VGGish}}\downarrow$	CQT1-PCC↑	LPAPS↓	CDPAM↓	TACons↑	$\text{TACons}_{0.1}\uparrow$	MOS-P↑	MOS-T↑
ZoME-Bench									
Textual Inv. (2022)	0.6767	4.9703	0.1590	5.3523	0.4543	0.2968	0.1653	0.78	0.56
DreamSound (2024)	0.7268	6.0530	0.0918	5.0300	0.3187	0.3270	0.1968	1.34	1.87
SteerMusic+ (2026)	0.3612	3.1765	0.3288	4.6869	0.3131	0.4160	0.2792	2.87	2.30
CLEF	**0.0448**	**1.4320**	**0.6046**	**1.9569**	**0.2297**	**0.7616**	**0.6551**	**3.09**	**2.70**
MusicBench									
Textual Inv. (2022)	1.1807	14.3436	0.3814	5.6865	0.6204	0.2905	0.1610	0.66	0.33
DreamSound (2024)	1.0982	12.7945	0.1229	5.1159	0.3759	0.3296	0.2041	2.01	1.89
SteerMusic+ (2026)	0.6070	1.8271	0.4368	2.5246	0.2508	0.6821	0.5574	2.66	2.59
CLEF	**0.0481**	**1.0549**	**0.6380**	**1.9569**	**0.2382**	**0.7613**	**0.6568**	**2.99**	**2.80**

representations). Full-path TACons and bottom-10% TACons ($\text{TACons}_{0.1}$) capture average consistency and consistency in the least-matching aligned regions, respectively. For personalized editing, CDPAM (Manocha et al., 2021) measures perceptual distance to reference audio. MOS-P and MOS-T assess source preservation and target-edit fidelity, respectively. Since CLAP and TACons also guide condition compensation and candidate selection, we interpret them alongside other objective metrics and listening evaluations.

Baselines. We compare CLEF with six text-guided music editing baselines: DDIM Inversion (Hershey et al., 2017), MusicMagus (Zhang et al., 2024), ZETA (Manor & Michaeli, 2024), MelodyFlow (Le Lan et al., 2024), SmartDJ (Lan et al., 2026), and SteerMusic (Niu et al., 2026). AudioLDM2 (Liu et al., 2024a) and MusicGen (Hershey et al., 2017) serve as text-to-music baselines without source-audio conditioning. Personalized editing baselines include Textual Inversion, DreamSound, and SteerMusic+. All methods use the same test samples and target instructions within each setting, with public implementations and checkpoints where available. Configurations are provided in Appendix A.1.

Implementation details. Experiments use an NVIDIA RTX 4090 GPU. FACC is trained for 3 epochs with a learning rate of 10^{-4}, while the pretrained DDS editor remains frozen. Training uses the single-step feedback compensation in Section 4.3, with gradients propagated through the compensated target-conditioned branch. At inference,each feedback round uses 500 DDS optimization steps and a CFG scale of 30. Each round is initialized from the original source representation. We use two feedback rounds, a structure-to-semantics transition at the midpoint of DDS optimization, and $\lambda_{\max} = 0.03$.The final output is selected using the rule in Section 4.4. Additional implementation details are provided in Appendix A.2.

Table 3: Ablation studies of CLEF. We evaluate the contributions of feedback signals, the compensation controller, and training objectives.

(a) Feedback and controller ablation (zero-shot).

Variant	$\text{FAD}_{\text{CLAP}}\downarrow$	$\text{FAD}_{\text{VGGish}}\downarrow$	CQT1-PCC↑	LPAPS↓	CLAP↑	TACons↑	$\text{TACons}_{0.1}\uparrow$
w/o FACC	0.04	1.45	0.57	1.92	0.28	0.21	0.15
w/o R_{str}	0.03	0.79	0.67	1.68	0.28	0.80	0.70
w/o R_{sem}	0.18	4.12	0.44	2.55	0.30	0.70	0.55
w/o H_n^k	0.04	1.49	0.56	1.98	0.27	0.76	0.68
Full CLEF	0.04	0.92	0.68	1.67	0.29	0.79	0.69

(b) Reference feedback ablation (personalized).

Variant	$\text{FAD}_{\text{CLAP}}\downarrow$	$\text{FAD}_{\text{VGGish}}\downarrow$	CQT1-PCC↑	LPAPS↓	CDPAM↓	TACons↑	$\text{TACons}_{0.1}\uparrow$
w/o R_{ref}	0.0678	1.6234	0.5567	2.0052	0.2511	0.7067	0.6023
Full CLEF	0.0448	1.4320	0.6046	1.9569	0.2297	0.7616	0.6551

(c) Training-objective ablation.

Removed objective	Δ CLAP (%)	Δ CQT1-PCC (%)	Δ CDPAM (%)
w/o $\mathcal{L}_{\text{dir}}$	-0.10	-1.61	+0.4
w/o $\mathcal{L}_{\text{mag}}$	≈0.00	-0.51	≈0.00
w/o $\mathcal{L}_{\Delta}$	-0.03	-0.17	+0.8
w/o $\mathcal{L}_{\text{fb}}$	-3.01	-4.22	+10.91

5.2 Results

Zero-shot Text-guided Music Editing. Table 1 shows that CLEF achieves the best FAD, LPAPS, CQT1-PCC, and both TACons scores on both benchmarks. Compared with SteerMusic, it improves source preservation and target alignment; on ZoME-Bench, CQT1-PCC increases by 25.6% and CLAP by 4.2%.Although MelodyFlow obtains higher CLAP scores, its lower CQT1-PCC and higher LPAPS indicate weaker source preservation. These results highlight the trade-off between target alignment and source retention. Additional task-wise results for instrument,genre, mood, and background editing are provided in Appendix B.1.

Personalized Music Editing. CLEF achieves the best aggregate objective scores on both benchmarks(Table 2). Compared with SteerMusic+ On ZoME-Bench, it reduces CDPAM from 0.3131 to 0.2297 and increases CQT1-PCC from 0.3288 to 0.6046, improving references similarity and melodic preservation. On MusicBench, the results show a smaller reference-similarity gain but consistent preservation improvements. Further concept-wise results for Reggae,Hiphop, Guitar, and Sitar are reported in Appendix B.1.

5.3 Ablation Studies and Analysis

Component Ablation. Table 3(a)–(b) summarizes component ablations. Removing FACC lowers CQT1-PCC from 0.68 to 0.57 and TACons from 0.79 to 0.21, supporting its contribution to source preservation. Removing R_{sem} slightly improves CLAP (0.29 to 0.30) but worsens CQT1-PCC (0.44) and LPAPS (2.55). Removing R_{str} lowers CQT1-PCC and CLAP while slightly improving FAD and TACons, indicating metric-dependent benefits. In personalized editing, removing R_{ref} increases CDPAM from 0.2297 to 0.2511 and reduces CQT1-PCC and both TACons measures, supporting its contribution to reference similarity and source retention. Removing H_n^k reduces CQT1-PCC from 0.68 to 0.56 and increases LPAPS from 1.67 to 1.98, supporting compensation conditioned on local editing state.

Training-objective Ablation. Table 3 (c) examines the contribution of the training objectives and compensation regularizers. Removing $\mathcal{L}_{\text{fb}}$ causes the largest degradation, with decreases of 3.01% in CLAP and 4.22% in CQT1-PCC, and an increase of 10.91% in CDPAM. In comparison, removing $\mathcal{L}_{\text{dir}}$, $\mathcal{L}_{\text{mag}}$, or $\mathcal{L}_{\Delta}$ produces relatively smaller changes, demonstrating the dominant role of $\mathcal{L}_{\text{fb}}$ in balancing instruction alignment and source preservation. These ablations evaluate the contribution of each training component through generated outputs rather than establishing a direct relationship between DDS magnitude and perceptual improvement.



Figure 3: User preference comparison in zero-shot and personalized music editing.

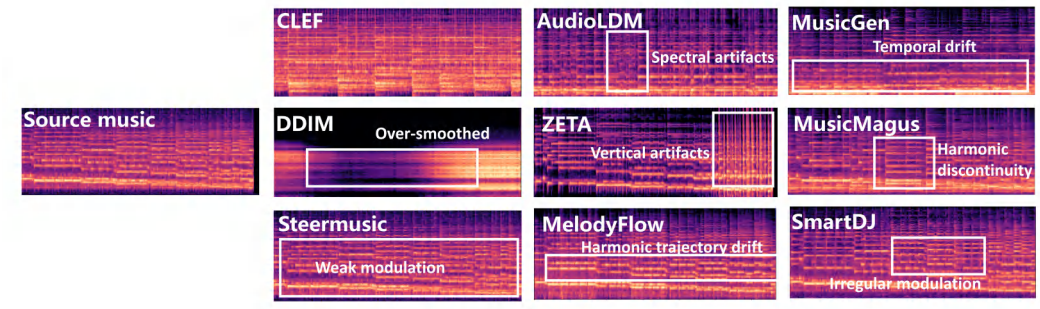

Figure 4: Qualitative comparison of personalized music editing. While transforming a monotone acoustic guitar into one with a tremolo effect, CLEF preserves source temporal and harmonic structures. Conversely, baselines exhibit structural degradation, including over-smoothing, modulation inconsistency, harmonic drift, or spectral artifacts as highlighted in white boxes.

Effect of Stage-wise Conditioning and Parameter Sensitivity. We compare three compensation schedules in DDS optimization: structure-to-semantics (Str→Sem), the reverse order (Sem→Str), and alternation every 25 DDS steps. Structure-to-semantics conditioning (Str→Sem) achieves higher CLAP and CQT1-PCC than both alternatives, indicating that its preservation advantage is accompanied by stronger target alignment. This supports the selected ordering within each editing round. We further analyze CLEF's sensitivity to the number of feedback rounds k, the maximum compensation coefficient $\lambda_{\max}$, and the bottom-q ratio in TACons, as well as its robustness to temporal shifts and its computational overhead relative to SteerMusic. Detailed analyses are provided in Appendices B and C.

5.4 Qualitative Analysis and Subjective Evaluation

We evaluate perceptual quality through a study with 25 music-literate volunteers, evaluating source preservation (MOS-P), target-edit fidelity (MOS-T), and anonymized pairwise preference. Tables 1–2show higher MOS-P and MOS-T scores for CLEF than for SteerMusic/SteerMusic+ on both benchmarks. On ZoME-Bench, CLEF achieves MOS-P/MOS-T of 2.80/2.68 for zero-shot editing and 3.09/2.70 for personalized editing; corresponding MOS-T scores on MusicBench are 2.65 and 2.80. Listeners also prefer CLEF in the majority of comparisons in both settings (Fig. 3). The tremolo example (Fig. 4) shows periodic energy modulation consistent with the requested effect while retaining the source's harmonic trajectories and temporal organization. Study details and additional examples are provided in Appendices B.5 and B.6.

6 Conclusion

We presented CLEF, a closed-loop framework balancing instruction adherence and source structure preservation in text-guided music editing. Its observe–compensate–edit loop uses a Feedback-aware Compensation Controller to adapt target textual conditioning from editing outcomes while keeping the pretrained diffusion editor frozen. We introduced TACons for temporal structural evaluation and complementary semantic and structural feedback for conditioning refinement. Experiments demonstrate improved editing quality and controllability over existing approaches in zero-shot and personalized editing, with effectiveness maintained on longer audio. However, feedback evaluation and iterative compensation add inference cost, especially for long-form and personalized editing. Future work will explore efficient feedback mechanisms and richer perceptual signals to improve scalability and generalization.

AI USE STATEMENT

We used generative AI tools for manuscript drafting and translation, literature review, feedback on experimental design, implementation and debugging, result analysis, and visualization. AI-assisted work was reviewed and revised under author supervision. The authors take responsibility for the accuracy, originality, and integrity of the final submission.

ETHICS STATEMENT

CLEF supports text-guided music editing and carries risks of unauthorized adaptation, imitation of artists' styles, and misleading presentation of edited recordings. Biases inherited from pretrained models and their training data may also affect editing outcomes. Source and reference music, pretrained models, and edited outputs should be used in ways that respect applicable licenses and creators' rights, with permission obtained where required. Synthetically generated or edited audio should be clearly identified as such. Our subjective listening study involved 25 music-literate volunteers and evaluated source preservation, target-edit fidelity, and anonymized A/B preferences. Appendix B.5 describes the protocol, including randomized presentation of anonymized system outputs.

REPRODUCIBILITY STATEMENT

Section 4 describes the feedback signals, compensation controller, training objective, and iterative editing procedure. Section 5.1 documents the benchmarks, dataset splits, evaluation metrics, hardware, and main training and inference settings. Appendices A.1 and A.2 provide baseline configurations and details of CLEF's architecture, optimization, and hyperparameters. Section 5.3 and its associated appendix analyses document component ablations and the effects of conditioning schedules, feedback rounds, and compensation magnitude. Appendix B.5 describes the listening-study protocol, rating scales, and randomized anonymized comparisons. Together, these descriptions specify the implemented method and evaluation procedures used for the reported zero-shot and personalized music editing experiments.

REFERENCES

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. MusicLM: Generating Music From Text, 2023.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Defossez. Simple and Controllable Music Generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP Learning Audio Concepts from Natural Language Supervision. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, Rhodes Island, Greece, 2023. IEEE. doi: 10.1109/ICASSP49357.2023.10095889.

Zhongyuan Fu, Yuhang Jia, Hui Wang, Pengjun Chen, Jian Gao, Cun Liu, Wenjia Zeng, Yong Chen, and Yong Qin. Echoedit: Stabilizing inversion-free audio editing via optimal transport geometry, 2026.

Zhengkun Ge, Xiaoqian Liu, Haoran Zhang, Yuan Ge, Junxiang Zhang, Zhengtao Yu, Jingbo Zhu, and Tong Xiao. Directaudioedit: Inversion-free text-guided audio editing via diffusion prediction contrast, 2026.

Bing Han, Junyu Dai, Weituo Hao, Xinyan He, Dong Guo, Jitong Chen, Yuxuan Wang, Yanmin Qian, and Xuchen Song. InstructME: An Instruction Guided Music Edit And Remix Framework with Latent Diffusion Models, 2023.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss,

and Kevin Wilson. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135. IEEE, 2017. doi: 10.1109/ICASSP.2017.7952132.

Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta Denoising Score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2328–2337, 2023. doi: 10.1109/ICCV51070.2023.00221.

Siyuan Hou, Shansong Liu, Ruibin Yuan, Wei Xue, Ying Shan, Mangsuo Zhao, and Chao Zhang. Editing Music with Melody and Text: Using ControlNet for Diffusion Transformer. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10890309.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech 2019*, pp. 2350–2354, 2019. doi: 10.21437/Interspeech.2019-2219.

Zitong Lan, Yiduo Hao, and Mingmin Zhao. Smartdj: Declarative audio editing with audio language model. In *International Conference on Learning Representations (ICLR)*, 2026.

Gael Le Lan, Bowen Shi, Zhaoheng Ni, Sidd Srinivasan, Anurag Kumar, Brian Ellis, David Kant, Varun Nagaraja, Ernie Chang, Wei-Ning Hsu, Yangyang Shi, and Vikas Chandra. High Fidelity Text-Guided Music Editing via Single-Stage Flow Matching, 2024.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger 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 *International Conference on Learning Representations (ICLR 2024)*, 2024. URL `https://openreview.net/forum?id=w3YZ9MSlBu`.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024a. doi: 10.1109/TASLP.2024.3399607.

Huadai Liu, Jialei Wang, Xiangtai Li, Wen Wang, Qian Chen, Rongjie Huang, Yang Liu, Jiayang Xu, and Zhou Zhao. Medic: Zero-shot music editing with disentangled inversion control. *arXiv preprint arXiv:2407.13220*, 2024b. doi: 10.48550/arXiv.2407.13220.

Pranay Manocha, Zeyu Jin, Richard Zhang, and Adam Finkelstein. CDPAM: Contrastive Learning for Perceptual Audio Similarity. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 196–200. IEEE, 2021. doi: 10.1109/ICASSP39728.2021.9413711.

Hila Manor and Tomer Michaeli. Zero-Shot Unsupervised and Text-Based Audio Editing Using DDPM Inversion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34603–34629. PMLR, 2024.

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.

Xinlei Niu, Kin Wai Cheuk, Jing Zhang, Naoki Murata, Chieh-Hsin Lai, Michele Mancusi, Woosung Choi, Giorgio Fabbro, Wei-Hsiang Liao, Charles Patrick Martin, and Yuki Mitsufuji. SteerMusic: Enhanced Musical Consistency for Zero-Shot Text-Guided and Personalized Music Editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2000–2010, 2026. doi: 10.1609/aaai.v40i3.37181.

Hiroaki Sakoe and Seibi Chiba. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, February 1978. ISSN 0096-3518. doi: 10.1109/TASSP.1978.1163055.

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.

Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J. Bryan. Music ControlNet: Multiple Time-Varying Controls for Music Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2692–2703, 2024. doi: 10.1109/TASLP.2024.3399026.

Yixiao Zhang, Yukara Ikemiya, Gus Xia, Naoki Murata, Marco A. Martínez-Ramírez, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. MusicMagus: Zero-Shot Text-to-Music Editing via Diffusion Models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 7805–7813, 2024. doi: 10.24963/ijcai.2024/864.

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.

A Implementation Details

A.1 Baseline Implementations

For a consistent comparison, we reproduce all baselines using their publicly released code, pretrained checkpoints, and recommended configurations. We do not tune baseline-specific hyperparameters on ZoME-Bench. When a parameter is not explicitly specified by the authors, we retain the default value in the released implementation. All methods are evaluated on the same test samples and editing instructions using the evaluation protocol described in Section 5. The implementation details are summarized in Table 4.

Table 4: Implementation details of the baseline methods.

Baseline	**Key hyperparameters**	**Implementation details**
DDIM Inversion	200 diffusion steps; $t_{\text{start}} = 200$; CFG $= 7.5$	Source and target descriptions are used as the source and target conditions, respectively; other settings follow the official implementation.
AudioLDM2	FP16; 200 inference steps; CFG $= 3.5$; seed $= 1234 + i$	The target description is used as the generation prompt without source-audio conditioning. Outputs are saved as 16-kHz WAV files.
MusicGen	10 s; batch size $= 1$; top-$k = 250$; top-$p = 0$; temperature $= 1.0$; CFG $= 3.0$; seed $= 42$	The target description is used as the generation prompt, with generation parameters fixed through AudioCraft.
MusicMagus	100 inference steps; 10 s; inversion guidance $= 1.0$; CFG $= 3.5$; cross-attention guidance $= 0.04$; seed $= 42$	AudioLDM2 is used with DDIM inversion. Source and target descriptions specify the before and after concepts; the negative prompt is "low quality."
ZETA	$t_{\text{start}} = 100$; source guidance $= 1.0$; target CFG $= 7.5$	We follow the recommended Stable Audio Open configuration and retain the official defaults for unspecified parameters.
SmartDJ	$N_{\text{diff}} = 100$; CFG $= 4$; guidance rescale $= 0.8$	We use the released SmartDJ-Editor checkpoint at 24 kHz without modifying its pretrained DiT architecture.
SteerMusic / SteerMusic+	SteerMusic: guidance scale $= 30$, $\text{weight}_{\text{aug}} = 3$; SteerMusic+: guidance scale $= 15$	Both use AudioLDM2 as the backbone. SteerMusic+ additionally loads the corresponding DreamSound checkpoint and personalized concept token.
Textual Inversion	Batch size $= 1$; gradient accumulation $= 4$; 3000 steps; learning rate $= 5 \times 10^{-4}$	Only the placeholder-token embedding is optimized for each personalized concept and loaded at inference.
DreamSound	Batch size $= 1$; gradient accumulation $= 4$; 300 steps; learning rate $= 1 \times 10^{-5}$; 50 inference steps	We follow the AudioLDM2-based DreamBooth configuration and adapt a separate model to each personalized target concept.

A.2 CLEF Implementation and Hyperparameters

Component	Key hyperparameters	Implementation details
Trainable modules	Prompt Compensation Network; Adaptive Feedback Gate	Only the newly introduced Transformer Adapter and Adaptive Feedback Gate are optimized. AudioLDM2, including its VAE, U-Net, and text encoder, as well as the DDS backbone of SteerMusic, remains frozen throughout training.
Audio projection	$32000 \to 512 \to 512$	Source and reference audio are represented as 32,000-dimensional features and projected into a 512-dimensional conditioning space through a two-layer MLP.
Reference audio	Hidden dim $= 512$; 8 attention heads; FFN $= 512 \to 2048 \to 512$	Reference-audio features are processed by an 8-head cross-attention module with a hidden dimension of 512, followed by a feed-forward network with dimensions $512 \to 2048 \to 512$.
Prompt Compensation Network	2 Transformer Adapter layers; hidden dim $= 512$; 8 attention heads; FFN dim $= 2048$	The compensation network jointly incorporates target/source text conditions, source/reference audio features, a 7-dimensional feedback state, and a 128-dimensional timestep representation.
Adaptive Feedback Gate	$135 \to 512 \to 256 \to 1$; $\lambda_{\max} = 0.03$	The gate is implemented as a three-layer MLP. Its output is passed through a sigmoid function and scaled by $\lambda_{\max}$ to bound the magnitude of target-condition compensation.
Training objective	$\lambda_{\text{dir}} = 1.0$; $\lambda_{\text{mag}} = 0.20$; $\lambda_{\Delta} = 0.05$; $\lambda_{\lambda} = 0.01$; $\lambda_{\text{over}} = 0.20$	For each sample, the base feedback F_{base} is first computed from the uncompensated edit. The controller then produces the compensated target condition, from which the compensated feedback CLEF is obtained. The objective combines editing-direction consistency, editing-magnitude constraint, condition-compensation regularization, gate-magnitude regularization, and feedback-improvement loss. The comparison between F_{base} and CLEF further discourages ineffective or excessive compensation.
Optimization	AdamW; 3 epochs; learning rate $= 1 \times 10^{-4}$; weight decay $= 1 \times 10^{-4}$; max gradient norm $= 1.0$	The Transformer Adapter and Adaptive Feedback Gate are the only modules updated during optimization; all pretrained backbone parameters remain frozen.
Inference	500 DDS optimization steps per feedback round; CFG $= 30$; $\lambda_{\max} = 0.03$	Feedback is computed online from the current editing result and used to iteratively update the target condition. The same $\lambda_{\max} = 0.03$ is used for both zero-shot and personalized editing.
Hardware	$1\times$ NVIDIA RTX 4090	All training and evaluation experiments are conducted on a single NVIDIA RTX 4090 GPU.

B Additional Experimental Results

B.1 Full Task-wise Results

We further break down the aggregate results in the main paper by editing category. This provides a closer view of how CLEF behaves across different types of transformations, rather than averaging their effects into a single score. We report the zero-shot results for instrument, genre, mood, and background editing in Table 5.

To further analyze CLEF across different editing types, we conduct a task-wise comparison with SteerMusic under the same setting. As shown in Table 8, CLEF performs favorably across all four tasks, with particularly clear gains for instrument and genre editing. For these two tasks, CQT_1-PCC increases from 0.5569 to 0.7135 and from 0.4829 to 0.6637, respectively. CLAP improves from

Table 5: Task-wise zero-shot editing results on ZoME-Bench. Best results within each editing category are shown in bold. ↓ indicates lower is better and ↑ indicates higher is better.

Task	Method	FAD $_{\text{CLAP}}$↓	FAD $_{\text{VGGish}}$↓	CQT1-PCC↑	LPAPS↓	CLAP↑	TACons↑	TACons$_{0.1}$ ↑
Instrument	CLEF	**0.0617**	**1.1468**	**0.7135**	**1.6527**	**0.2932**	**0.7936**	**0.6861**
	SteerMusic	0.0937	2.9868	0.5569	2.0179	0.2700	0.7456	0.6233
Genre	CLEF	**0.0689**	**1.2993**	**0.6637**	**1.6826**	**0.2635**	**0.8007**	**0.6876**
	SteerMusic	0.0989	2.5651	0.4829	2.2430	0.2500	0.7260	0.6098
Mood	CLEF	**0.0743**	**1.3185**	**0.6325**	**1.6894**	**0.3159**	**0.7979**	**0.6806**
	SteerMusic	0.0832	1.5723	0.6023	1.8198	0.3000	0.7667	0.6584
Background	CLEF	**0.0838**	**1.1223**	**0.7217**	**1.6824**	**0.2833**	**0.7885**	**0.6891**
	SteerMusic	0.0939	1.8786	0.6028	1.8040	0.2700	0.7690	0.6640

Figure 1: Time-Memory Efficiency for Zero-shot Music Editing

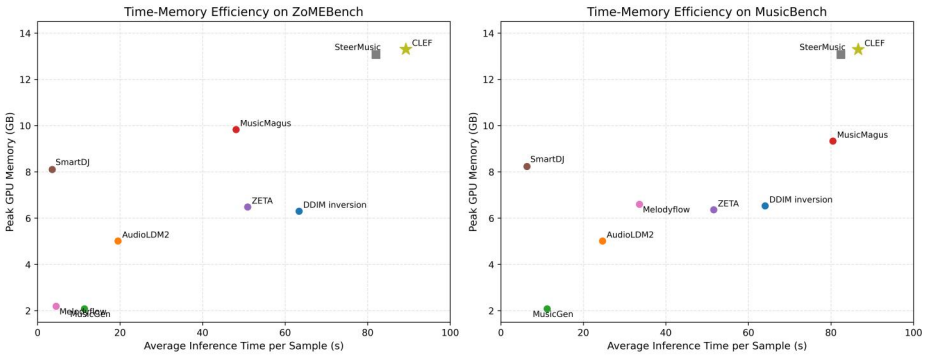

Figure 5: Time–memory efficiency for zero-shot music editing. CLEF incurs additional inference time while maintaining peak GPU memory close to SteerMusic.

Figure 2: Time-Memory Efficiency for Personalized Music Editing

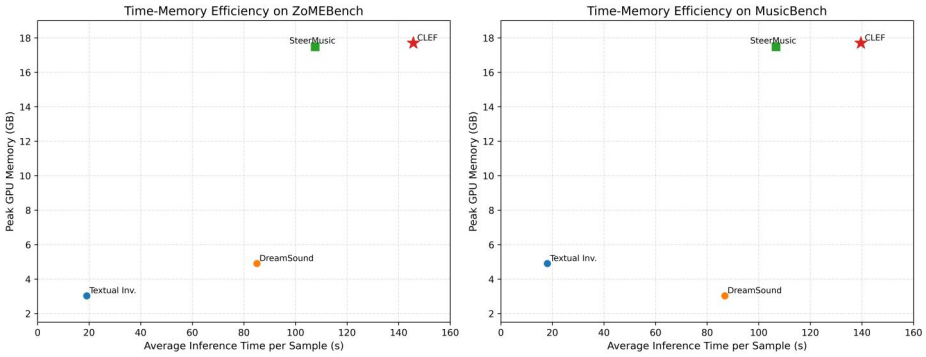

Figure 6: Runtime-memory efficiency for zero-shot music editing. Runtime denotes the average wall-clock time for one complete editing run per test sample, including the initial editing pass and all feedback rounds.

0.2700 to 0.2932 for instrument editing and remains comparable for genre editing. Background editing also shows clear gains, while improvements for mood editing are more moderate. Overall, these results suggest that CLEF better balances target alignment and source preservation across editing types.

B.2 Efficiency Analysis

To evaluate the computational overhead introduced by the feedback mechanism, we compare the average per-sample runtime and peak GPU memory across methods. The runtime is measured as the wall-clock time of one complete editing run on a single test sample, including the initial editing

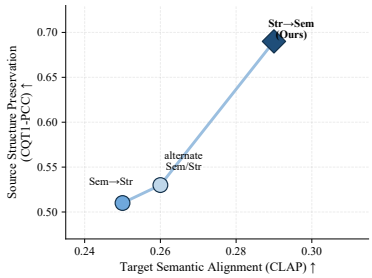

Figure 7: Feedback ordering analysis based on the trade-off between target semantic alignment and source structure preservation.

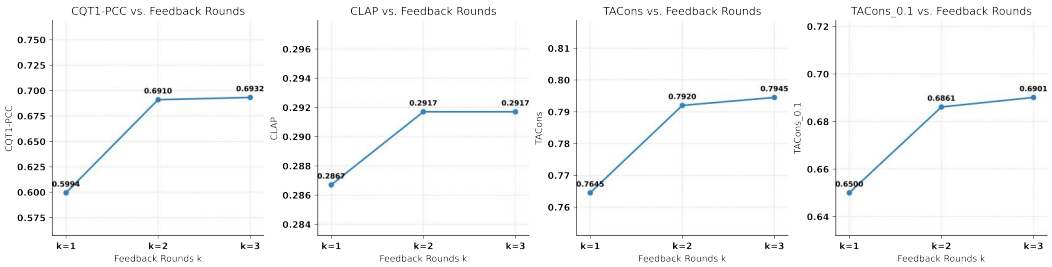

Figure 8: Effect of the number of feedback rounds. The subplots report the changes in CQT1-PCC, CLAP, TACons, and $\text{TACons}_{0.1}$ as the number of feedback rounds increases from $k = 1$ to $k = 3$.

process and all configured feedback refinement rounds. We report the average runtime across test samples, with all methods evaluated under the same hardware and inference settings. In **zero-shot editing**, compared with SteerMusic, CLEF increases runtime and peak memory by only **8.8% and 1.8%** on ZoME-Bench and **5.1% and 1.7%** on MusicBench, respectively. The overhead is more pronounced in **personalized editing**: the average per-sample runtime increases from approximately **108 s to 146 s** on ZoME-Bench and from **107 s to 140 s** on MusicBench, while peak memory remains nearly unchanged (**17.4 GB to 17.7 GB**). These results indicate that the computational cost of CLEF is mainly reflected in additional inference time rather than memory consumption. The increased runtime accompanies improved editing quality and structural preservation, while further reducing feedback-related computation remains an important direction for future work.

B.3 Effect of Stage-wise Conditioning.

(Fig. 7) compares three compensation schedules in DDS optimization: structure-to-semantics conditioning(Str→Sem), the reverse ordering(Sem→Str), and alternation every 25 DDS steps. Structure-to-semantics conditioning(Sem→Str) achieves higher CLAP and CQT1-PCC than both alternatives, indicating that its preservation advantage is accompanied by stronger target alignment. This supports the selected ordering within each editing round.

B.4 Number of Feedback Rounds.

Finally, we study the effect of the number of feedback rounds k. As shown in Figure 8, increasing k from 1 to 2 improves CQT1-PCC, CLAP, TACons, and $\text{TACons}_{0.1}$ by 15.3%, 1.7%, 3.6%, and 5.6%, respectively. The substantial improvement from the first to the second round indicates that iterative feedback can exploit the previous editing state to further refine the target condition.

Increasing the number of rounds to $k = 3$ yields only marginal additional gains, while introducing further computational cost. We therefore use $k = 2$ as the default setting, providing a practical balance between editing performance and inference efficiency.

B.5 HUMAN SUBJECTIVE STUDIES

We recruited 25 participants for the subjective listening study, which consisted of a MOS evaluation and a pairwise preference evaluation. In the MOS evaluation, participants rated anonymized outputs from nine systems for zero-shot editing and four systems for personalized editing. Each system was evaluated using six listening questions. The zero-shot questions covered instrument, genre, mood, and background editing; the personalized questions covered the Reggae, Hiphop, Guitar, and Sitar concepts. For each output, participants rated source content preservation (MOS-P) and target editing fidelity (MOS-T) on a 1–5 scale. We subtract 1 from each raw rating when reporting MOS; thus, the reported scores range from 0 to 4. MOS-P measures how well the output preserves source content unrelated to the instruction, while MOS-T measures how well it achieves the requested edit. For the pairwise preference evaluation, participants compared anonymized CLEF and baseline outputs for the same source and editing instruction. The baseline was SteerMusic for zero-shot editing and SteerMusic+ for personalized editing. The preference rate is the percentage of comparisons in which CLEF was selected. The study interface and evaluation procedure are illustrated in Figures **??**–10.

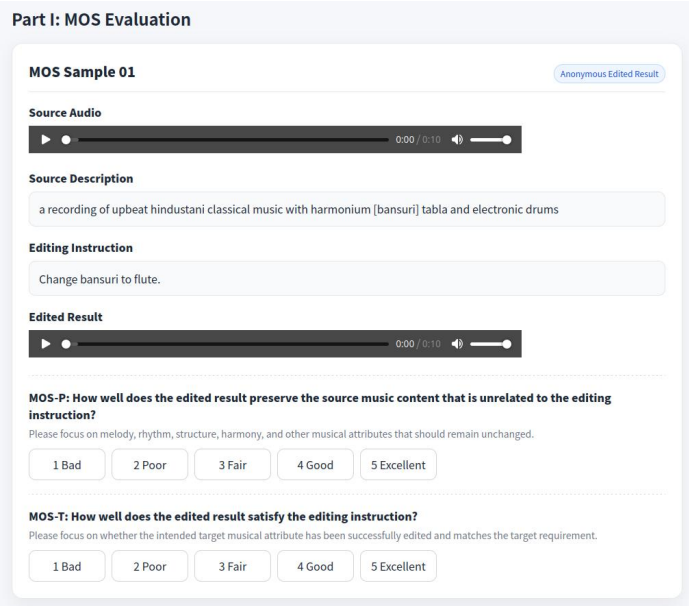

Figure 9: MOS evaluation interface used in the listening study. The two questions separately measure source content preservation (MOS-P) and target editing fidelity (MOS-T). MOS-P and MOS-T are reported on a 0–4 scale.

For the MOS evaluation, each edited sample is rated on two dimensions: source content preservation (MOS-P) and target editing fidelity (MOS-T). Participants use a five-point scale from 1 (Bad) to 5 (Excellent). For reporting, we subtract 1 from each rating before computing the mean; therefore, the MOS-P and MOS-T values in the tables are on a 0–4 scale.

Specifically, participants answer the following two questions:

MOS-P (Source Content Preservation). To what extent does the edited result preserve source content unrelated to the editing instruction Participants focus on attributes that should remain unchanged, including melody, rhythm, structure, and harmony.

MOS-T (Target Editing Fidelity). To what extent does the edited result achieve the target musical attribute specified by the editing instruction Participants assess whether the target attribute is successfully modified and consistent with the intended edit.

Together, MOS-P and MOS-T separately measure **source preservation** and **target editing effectiveness**, avoiding the trade-off between them being obscured by a single overall-quality score.

User Preference. We further conduct an anonymized A/B comparison. For zero-shot editing, CLEF is compared with SteerMusic; for personalized editing, CLEF is compared with SteerMusic+. For

each comparison, participants evaluate six matched samples generated by the two anonymized systems from the same source music and editing instruction. The presentation order is randomized, and participants select the result they prefer by jointly considering **target editing effectiveness and source content preservation**. Unlike MOS-P and MOS-T, which assess these dimensions separately, this comparison measures listeners' **overall preference**.

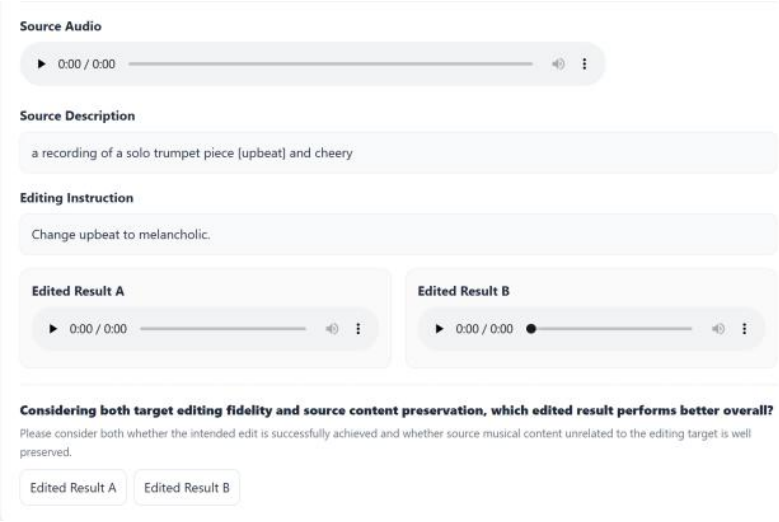

Figure 10: User preference evaluation interface for anonymized A/B comparison.

Table 6: User preference results comparing CLEF with SteerMusic. Preference denotes the percentage of trials in which CLEF was selected as the better overall result.

Task	Baseline	CLEF Pref.	95% CI / p-value
Zero-shot Music Editing	SteerMusic	**76.00%**	[54.87%, 90.64%] / 0.015
Personalized Music Editing	SteerMusic	**88.00%**	[68.78%, 97.45%] / < 0.001

User Preference Results. CLEF is preferred over SteerMusic in both settings, achieving preference rates of **76.00%** for **zero-shot editing** and **88.00%** for **personalized editing**, both significantly above the 50% chance level ($p = 0.015$ and $p < 0.001$, respectively). The stronger preference in **personalized editing** suggests that feedback-guided compensation is particularly beneficial when introducing a specific target concept while preserving instruction-irrelevant source content. Together with the MOS results, these findings provide complementary perceptual evidence that CLEF achieves a **better balance between target editing fidelity and source content preservation**.

B.6 More Cases

We provide additional qualitative comparisons to examine how different methods balance **target-attribute modification and source-content preservation**. Figures 11 show zero-shot editing examples with different target attributes. Some baselines produce noticeable target-related changes but substantially alter the source time–frequency structure, while others better preserve the source at the cost of weaker target expression. In contrast, **CLEF achieves clear target modifications while better preserving the dominant harmonic trajectories, transient locations, and temporal structure of the source**. This suggests that feedback-guided adjustment helps localize modifications to target-related attributes while reducing unnecessary structural drift.

Figure 12 further examines the more challenging setting of **long-form editing**, where local deviations may accumulate into global structural drift. MusicGen, AudioLDM2, and MusicMagus noticeably alter the original time–frequency distribution, DDIM exhibits large regions of missing spectral energy, and ZETA preserves some transient patterns but still deviates from the source in its overall spectral texture. In comparison, **CLEF maintains stronger correspondence with the source in the principal harmonic trajectories, transient locations, and energy distribution over the full duration**. Notably, SteerMusic and SmartDJ also preserve substantial source structure in this example; the main distinction is that **CLEF maintains a more stable balance between the requested modification and global structural consistency over long sequences**.

Figure 13 presents two **personalized editing** examples. For **guitar-to-keyboard**, Textual Inversion and DreamSound substantially reconstruct the original harmonic and energy patterns, while the periodic time–frequency texture produced by SteerMusic also differs noticeably from the source. CLEF instead **preserves the major temporal organization and harmonic contour while introducing the target instrument characteristics**. The **monotone-to-tremolo** transformation provides a more fine-grained test, as it modifies the performance attribute rather than the underlying musical content, making melody preservation particularly important. CLEF retains a harmonic trajectory closely aligned with the source while primarily modifying local time–frequency textures associated with tremolo, whereas several baselines introduce broader structural changes. Overall, these cases show that CLEF can **localize modifications to target-related attributes while minimizing changes to instruction-irrelevant source content**, effectively preserving ***what is played*** while modifying ***how it is played***.

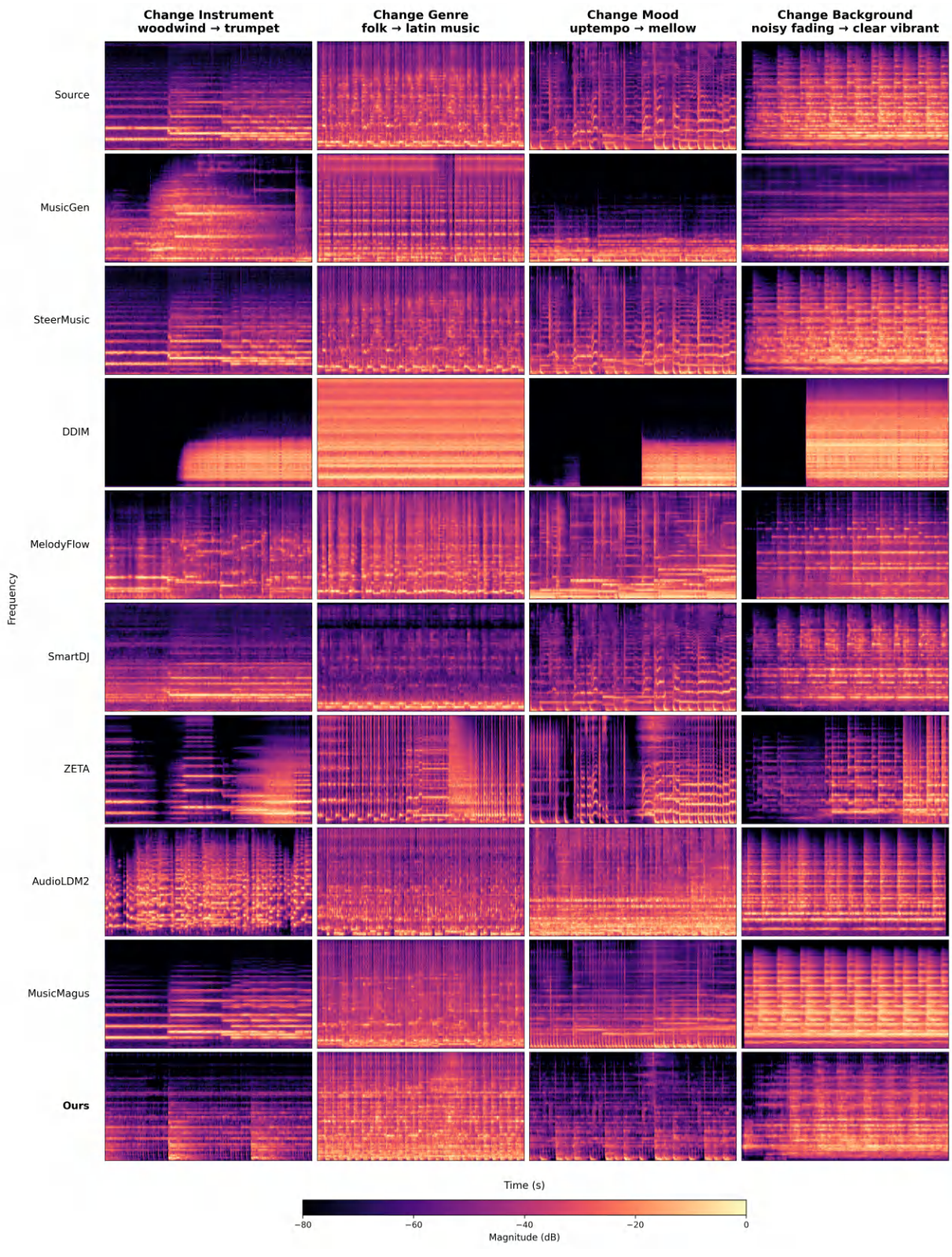

Figure 11: Additional qualitative comparison for zero-shot music editing. CLEF introduces the requested target attribute while preserving the dominant time–frequency structure of the source music.

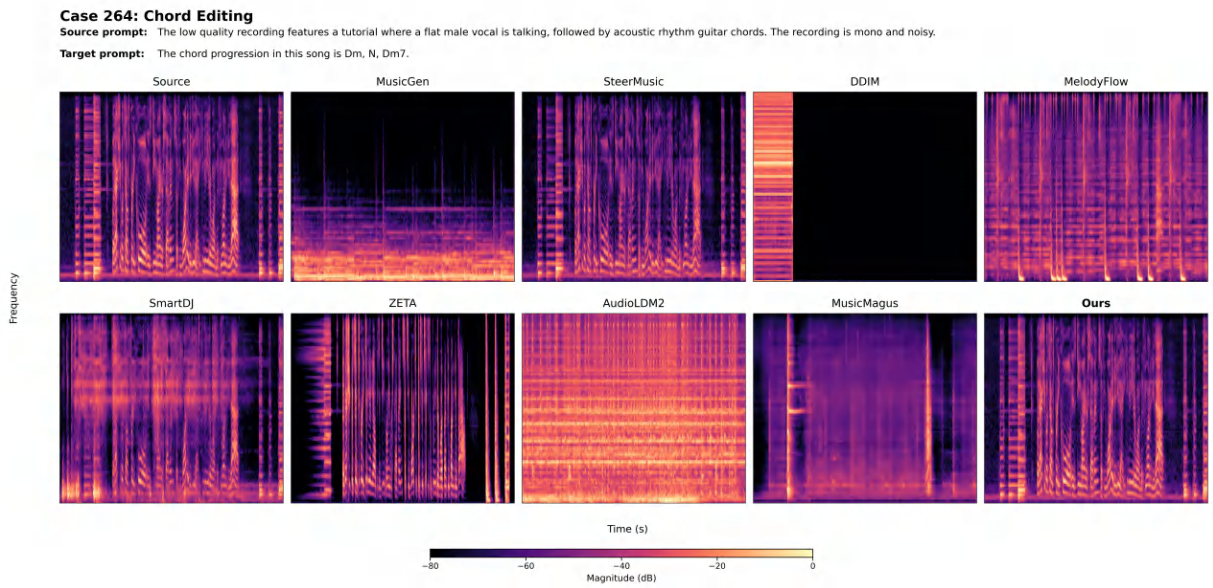

Figure 12: Qualitative comparison on a long-form editing example. CLEF preserves the global temporal and harmonic organization of the source while introducing the requested target attribute.

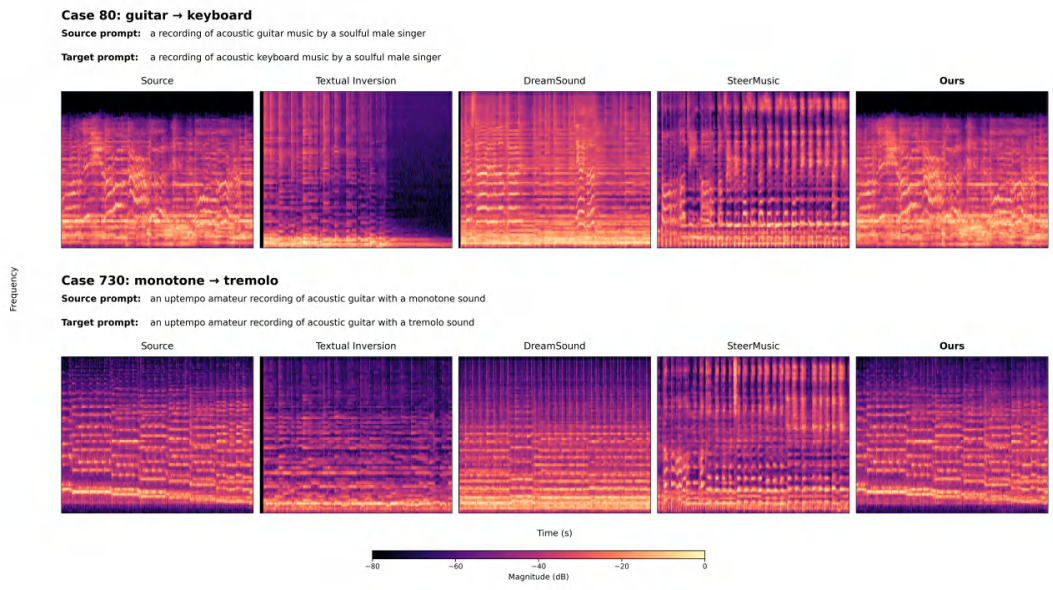

Figure 13: Personalized editing examples for guitar-to-keyboard and monotone-to-tremolo transformations. CLEF preserves the dominant temporal and harmonic structure of the source while incorporating the target reference attribute.

B.7 Failure Cases

Figure 14 presents two representative failure modes of CLEF. In the **rap-to-jazz** example, the edited result remains highly similar to the source, indicating that the model overemphasizes content preservation and fails to sufficiently introduce the target attribute, resulting in **under-editing**. In contrast, the **eighties-pop-to-modern-powerful** example exhibits segment-wise time–frequency reconstruction and discontinuous boundaries that disrupt the original harmonic structure. This suggests that for more **abstract target attributes**, feedback adjustment may induce overly strong or unstable local modifications.

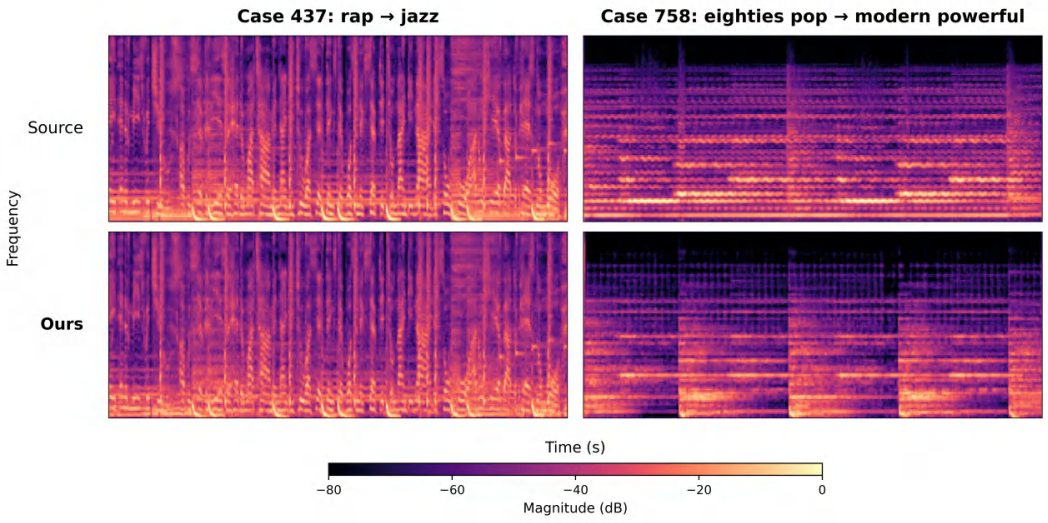

Figure 14: Representative failure cases of CLEF: **under-editing** due to overly conservative feedback (top) and **unstable local modification** under a more abstract editing target (bottom).

Together, these cases reveal two complementary limitations: **overly conservative feedback may lead to insufficient target expression, whereas overly strong or unstable feedback may damage source structure**. Improving feedback robustness for abstract editing concepts and adaptively adjusting the compensation strength according to editing difficulty therefore remain important directions for future work.

C Sensitivity Analysis

In this section, we further examine the sensitivity of CLEF to key hyperparameters and analyze several design choices in TACons. Unless otherwise specified, all training, inference, and evaluation settings are kept unchanged, with only the factor under investigation varied.

C.1 Sensitivity to the Maximum Feedback-Gate Magnitude

We examine the effect of the maximum compensation coefficient $\lambda_{\max}$ on both zero-shot and personalized editing. As shown in Figure 15, zero-shot editing exhibits only minor metric variations across the tested range, indicating low sensitivity to $\lambda_{\max}$. In contrast, personalized editing shows a clear improve-then-degrade trend. Increasing $\lambda_{\max}$ from 0.01 to 0.03 reduces LPAPS from 2.7005 to 1.9685 (**-27.1%**) and increases TACons from 0.6605 to 0.7681 (**+16.3%**), suggesting that an overly restrictive gate limits effective feedback adjustment. Further increasing $\lambda_{\max}$ to 0.07 reverses this trend, with CQT_1-PCC decreasing from 0.6402 to 0.5648 and bottom-10% TACons decreasing from 0.6938 to 0.6213, indicating that excessive compensation gradually disrupts source content and local temporal structure. Overall, $\lambda_{\max} = 0.03$ lies within the stable region for zero-shot editing while corresponding to a clear performance turning point for personalized editing, and is therefore used as the default setting.

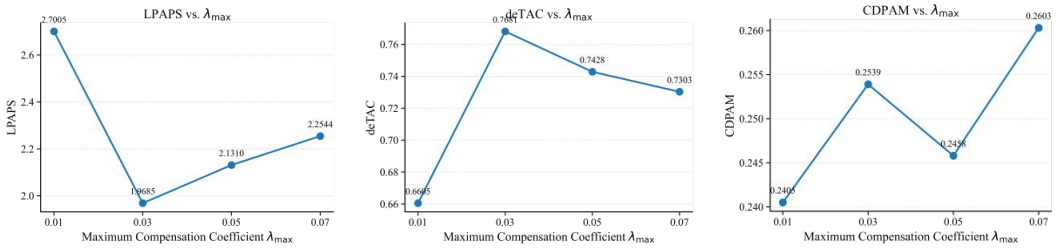

Figure 15: Sensitivity to $\lambda_{\max}$. Zero-shot editing remains relatively stable, whereas personalized editing exhibits an **improve-then-degrade** trend, with $\lambda_{\max} = 0.03$ providing a favorable operating point.

C.2 Additional Analysis of TACons

Effect of DTW alignment. To validate the use of DTW in TACons, we introduce a controlled **3-frame temporal shift** to the validation-set editing results and compare DTW alignment with direct frame-wise cosine similarity. As shown in Table 7, under the original alignment, frame-wise and DTW-based mean similarities are nearly identical (**0.8025 vs. 0.8029**), indicating that DTW **does not artificially inflate structural similarity** when temporal correspondence is already accurate. After the 3-frame shift, however, the frame-wise mean drops from **0.8025 to 0.4979** (-38.0%), with its bottom-10% similarity decreasing from **0.6713 to 0.2548**. In contrast, the DTW-based mean remains nearly unchanged (**0.8029 to 0.8021**), as does the bottom-10% score (**0.6722 to 0.6688**). These results show that **DTW compensates for minor temporal misalignment without artificially increasing similarity**, allowing TACons to better reflect genuine temporal structural changes rather than absolute positional shifts.

Table 7: Effect of DTW alignment on temporal consistency evaluation.

Setting	**Frame-wise**		**DTW**	
	Mean↑	Bottom-10%↑	Mean↑	Bottom-10%↑
Original alignment	0.8025	0.6713	0.8029	0.6722
Shifted by 3 frames	0.4979	0.2548	0.8021	0.6688

Sensitivity to the bottom-q ratio. S_{global} measures the average similarity over all DTW-aligned MERT feature pairs, whereas $S_{\text{bottom-}q}$ focuses on the lowest-similarity $q\%$ to capture **localized structural degradation**. As shown in Table 8, increasing q from **5% to 20%** raises $S_{\text{bottom-}q}$ from **0.6454 to 0.7027**, gradually approaching $S_{\text{global}} = \mathbf{0.8029}$. A smaller q is more sensitive to severe local deviations but also more susceptible to extreme frames, whereas a larger q incorporates more normally aligned regions and weakens the focus on local degradation. We therefore use $q = 10\%$ as a balance between **local structural sensitivity and statistical stability**.

Table 8: Sensitivity of local temporal consistency to the bottom-q ratio.

Bottom-q	Global↑	Bottom-q↑	Global–Bottom Gap↓	Bottom-q Std.↓
5%	0.8029	0.6454	0.1574	0.0723
10%	0.8029	0.6722	0.1307	0.0700
20%	0.8029	0.7027	0.1001	0.0657