MUSEADAPT: MELODY-PRESERVING MUSIC ADAPTATION WITH REPRESENTATION AND REWARD ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Melody-conditioned music generation provides fine-grained control over musical content, but adapting a reference recording to new musical conditions while faithfully preserving its melody remains challenging. Existing approaches often derive melody conditions from the same audio used as the generation target, causing the conditioning representation to retain source-specific acoustic characteristics and reducing its robustness to cross-style adaptation. To address this problem, we introduce **MAMA-20k**, a large-scale real-world dataset containing melody-aligned music pairs with shared melodic content and diverse acoustic realizations, and propose **MuseAdapt**, a framework for melody-preserving music adaptation through representation and reward alignment. For representation alignment, we train a convolutional–Transformer encoder to learn temporally structured melody representations from aligned music pairs and use them to control a pretrained flow-based music generator via ControlNet. For reward alignment, we construct preference pairs from complementary melody and text-alignment rewards and further optimize the generator with direct preference optimization (DPO). Experiments show that MuseAdapt achieves substantially stronger target-condition adherence than reconstruction-oriented baselines while maintaining comparable melodic consistency, yielding a better balance between melody preservation and musical adaptation without sacrificing generation quality. Audio examples can be found at https://museadapt.github.io/.

1 INTRODUCTION

Recent advances in generative modeling have enabled text-to-music systems to synthesize increasingly high-quality audio from natural-language descriptions Agostinelli et al. (2023); Copet et al. (2023); Evans et al. (2024a;b); Schneider et al. (2024); Chen et al. (2024). Beyond text-based generation, practical music creation often requires finer control over time-varying musical content, such as melody, rhythm, and harmony Wu et al. (2024); Tsai et al. (2025); Lin et al. (2023); Lan et al. (2024). Among these controls, melody is particularly important because it provides a recognizable musical structure that users may wish to retain while changing how the music is realized. We focus on melody-preserving music adaptation, where a reference recording provides the melody and temporal structure to be retained, while target conditions specify how the music should be realized in terms of style, instrumentation, and key.

A common strategy for learning melody control is to extract a melodic condition from the same recording that serves as the generation target Copet et al. (2023); Wu et al. (2024); Hou et al. (2025); Tsai et al. (2025). We refer to this commonly used training configuration as *self-reconstruction*. It is attractive because it requires no paired recordings: a structural representation such as a chromagram or CQT can be extracted automatically and used to reconstruct the original audio. However, this setup creates an ambiguity for music adaptation. Melody representations derived from audio generally preserve not only melodic information, but also source-dependent acoustic characteristics associated with instrumentation, articulation, note dynamics, and local spectro-temporal structure. Prior approaches have attempted to suppress such source-specific information through compact or discretized melodic representations and other information bottlenecks Copet et al. (2023); Hou et al. (2025), yet residual cues about the source realization may still remain in the conditioning signal.

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.

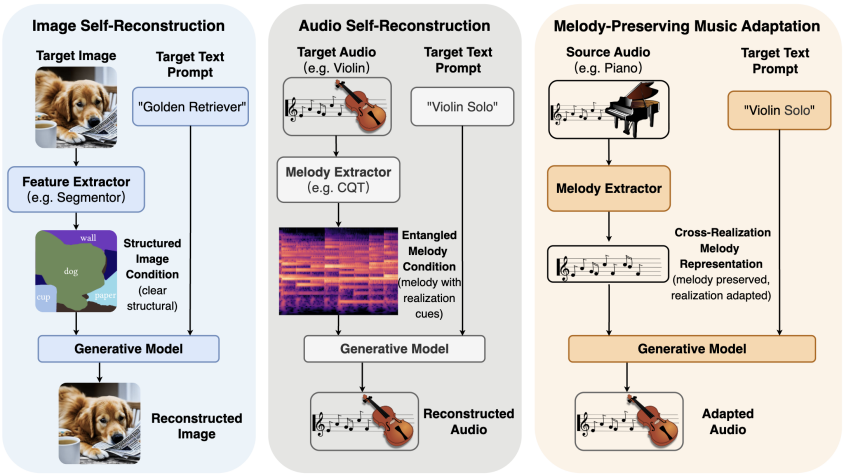


Figure 1: **Comparison between image self-reconstruction, audio self-reconstruction, and melody-preserving music adaptation.** Unlike image structural conditions that provide relatively clean structural information, audio melody conditions may contain source-specific acoustic cues that encourage reconstruction of the original realization.

During self-reconstruction, these residual source-specific cues are perfectly correlated with the target recording and can therefore become an effective shortcut for generation.

This issue is more pronounced for audio control than for many structural controls used in visual generation. Visual representations such as edges, human poses, or depth maps can often suppress appearance information to a large extent while retaining the geometry needed for conditional generation Zhang et al. (2023); Mou et al. (2024); Peng et al. (2024); Lin et al. (2025). In contrast, melody is physically realized through the same spectro-temporal patterns that also encode timbre and articulation, making it difficult to obtain an acoustic melody condition that is entirely independent of its source realization. Importantly, the problem is not simply that such representations contain source-specific information; rather, self-reconstruction explicitly rewards the model for exploiting it. This shortcut may remain hidden when the source and target conditions agree, but becomes problematic when they conflict. For example, a melody condition extracted from a piano recording may continue to favor piano-like characteristics even when the text prompt requests a violin rendition. This difference between conventional reconstruction-oriented melody control and adaptation-oriented melody modeling is illustrated in Fig. 1.

This observation also highlights an important distinction between *melody preservation* and *successful adaptation*. A reconstruction-oriented model can achieve very high melodic similarity simply by retaining more information from the source recording, yet still respond weakly to a new target style or instrumentation. Consequently, melody similarity alone is insufficient to characterize performance in realistic adaptation scenarios. What matters is the balance between preserving the melodic and temporal content of the reference and adapting its musical realization to the requested target conditions. To provide direct supervision for this setting, we move beyond same-instance self-reconstruction and adopt *cross-realization supervision*, where the conditioning and target recordings share aligned melodic content but differ in their acoustic realization. To this end, we construct **MAMA-20k**, a large-scale collection of real-world melody-aligned music pairs spanning diverse musical realizations. Such pairs provide supervision for learning which information should remain stable across different realizations of the same melody, rather than encouraging reconstruction of a particular source recording.

Building on this data, we introduce **MuseAdapt**, a framework for melody-preserving music adaptation through *representation alignment* and *reward alignment*. For representation alignment, we train a convolutional–Transformer melody encoder on melody-aligned recordings to obtain temporally structured representations that remain informative across different musical realizations. These representations directly condition a ControlNet branch attached to a pretrained flow-based music generator, while text and musical key provide complementary target conditions. Representation

alignment reduces reliance on realization-specific cues at the conditioning level, but conditional generation must still reconcile competing objectives: preserving the reference melody while following the requested style and key. We therefore further introduce reward-guided preference alignment. Complementary rewards measuring melody consistency, text–music alignment, and key adherence are used to construct preference pairs, and the generator is subsequently optimized with direct preference optimization (DPO) Rafailov et al. (2023). In this way, representation alignment determines what information should be preserved across realizations, while reward alignment encourages the generator to better balance preservation with target adaptation.

Our experiments show that reconstruction-oriented baselines can indeed achieve strong melodic consistency, but often at the expense of adaptation when source and target conditions differ. MuseAdapt maintains comparable melody preservation while achieving substantially stronger target-style and key adherence, resulting in a more favorable preservation–adaptation trade-off without sacrificing generation quality.

Our contributions are summarized as follows:

- We identify a source-specific shortcut in self-reconstruction-based melody conditioning: acoustic information leaked through the conditioning representation is rewarded during reconstruction and can hinder adaptation when source and target conditions conflict. We formulate melody-preserving music adaptation around the joint objectives of melody preservation and target adaptation, and introduce cross-realization supervision as an alternative training setting.
- We construct **MAMA-20k**, a large-scale real-world dataset of melody-aligned music pairs with shared melodic content and diverse acoustic realizations, and develop a melody encoder that learns temporal representations from these cross-realization pairs for controllable music generation.
- We propose **MuseAdapt**, which combines representation-aligned melody control with reward-guided DPO based on melody, text, and key preferences. Extensive evaluations demonstrate that MuseAdapt preserves the reference melody while providing substantially stronger adaptation to target musical conditions than reconstruction-oriented baselines.

2 Melody-Preserving Music Adaptation

2.1 Problem Definition

Existing melody-conditioned music generation methods typically aim to reconstruct or synthesize audio under a given melodic condition. However, our goal differs from reconstruction-oriented generation: we consider the setting where the melodic identity of a reference recording should be preserved while its acoustic realization is adapted according to target conditions.

Specifically, given a reference recording containing the original melodic content and a set of target conditions (e.g., textual descriptions or musical key), the objective of melody-preserving music adaptation is to generate a new musical realization that maintains the perceptually recognizable melody and temporal structure of the reference, while allowing changes in aspects such as instrumentation, style, and acoustic characteristics. Unlike waveform-level reconstruction, the generated audio is not required to reproduce the exact source realization, since variations in articulation, arrangement, and timbre are essential for meaningful adaptation.

2.2 From Self-Reconstruction to Cross-Realization Supervision

A prevalent training strategy for melody-controlled generation is self-reconstruction, where the melodic condition is extracted from the same recording used as the generation target. While effective for learning controllable generation without paired data, this setting inherently couples the desired melodic information with source-specific acoustic characteristics. As a result, models may exploit residual information related to the original realization rather than learning a representation that remains stable across different realizations of the same melody.

To address this limitation, MuseAdapt adopts a cross-realization supervision paradigm. Instead of reconstructing a recording from its own melodic condition, we leverage pairs of recordings that share

the same melodic identity but exhibit different acoustic realizations. Such supervision encourages the model to capture information that is consistent across musical realizations while reducing reliance on source-specific characteristics. Based on this paradigm, we investigate how learned melody representations and preference-based optimization can facilitate music adaptation while balancing melody preservation and target-condition alignment.

3 MAMA-20K: REAL-WORLD MELODY-ALIGNED MUSIC PAIRS

3.1 DATASET CONSTRUCTION

To enable cross-realization supervision for melody-preserving music adaptation, we construct MAMA-20k (**M**elody-**A**ligned **M**usic **A**daptation), a collection of real-world instrumental music pairs sharing the same underlying composition but exhibiting diverse acoustic realizations.

We collect real-world recordings corresponding to the same musical compositions but exhibiting diverse instrumental realizations from publicly available online sources, covering diverse genres, instruments, and performance styles. Since our focus is melody-preserving adaptation rather than vocal cover generation, we specifically consider recordings without dominant vocal components. We apply PANNs Kong et al. (2020) to filter recordings with prominent vocals and Meta Audiobox Aesthetics Tjandra et al. (2025) to remove low-quality samples, resulting in a collection of predominantly instrumental and quality-filtered musical renditions.

A key challenge in constructing such a dataset is that different realizations of the same composition often exhibit substantial temporal variations, including tempo changes, structural repetitions, and local rearrangements. Direct timestamp matching is therefore insufficient. We develop a two-stage melody alignment pipeline to identify temporally corresponding segments across different realizations.

3.2 MELODY ALIGNMENT

Our alignment pipeline first performs coarse alignment using chroma-based representations. After compensating for global key differences through optimal transposition, we construct frame-level chroma similarity matrices and apply Smith–Waterman local sequence alignment Smith et al. (1981) to identify high-confidence matching regions. To recover multiple correspondences within a pair of recordings, we further employ Waterman–Eggert-style iterative alignment Waterman & Eggert (1987). This stage prioritizes recall and produces candidate melody-aligned segments despite tempo and structural variations.

To improve alignment precision, we further refine the candidate segments using a sparse top-k chroma representation that emphasizes dominant pitch components while suppressing interference from accompaniment and timbral variation. We then perform localized re-alignment within each candidate region and retain high-confidence melody-consistent segment pairs.

The detailed alignment algorithm, including similarity computation, dynamic programming optimization, and filtering criteria, is provided in Appendix B.

3.3 DATASET CHARACTERISTICS

MAMA-20k is designed to provide cross-realization supervision rather than conventional self-reconstruction training. As summarized in Table 1, the dataset contains 13,079 compositions and 21,343 real-world recordings, forming 82,138 aligned recording pairs and 283,461 aligned segment pairs with a total duration of 4,448 hours. Each composition is associated with multiple musical realizations (3.88 renditions per composition on average), enabling models to learn melody representations that remain consistent across different acoustic realizations.

Table 1: Statistics of MAMA-20k.

Statistic	Value
Compositions	13,079
Recordings	21,343
Recording pairs	82,138
Segment pairs	283,461
Duration	4,448 h
Median length	42.5 s
Cross-instrument	79.9%
Alignment precision	93.7%

Beyond dataset scale, MAMA-20k provides substantial variation in musical realization. Among recording pairs with available instrument annotations, 79.9% correspond to cross-instrument trans-

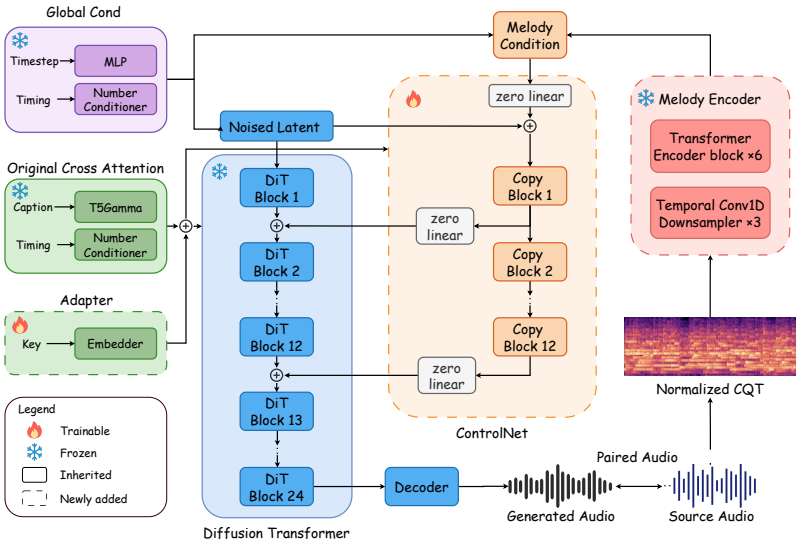


Figure 2: **Overview of MuseAdapt.** Given a source recording, MuseAdapt first extracts a learned melody representation through a cross-realization melody encoder. The melody condition is injected into a Stable Audio Open backbone through a newly introduced ControlNet branch, while textual descriptions and key information are provided through the original conditioning pathway and a lightweight key adapter.

formations, covering diverse timbral changes such as piano, guitar, saxophone, strings, and other instrumental settings. This diversity allows the model to learn melody-preserving adaptation beyond simple reconstruction of the original instrumentation. Manual inspection of the held-out test set shows that 93.7% of automatically mined pairs exhibit consistent melodic correspondence, validating the reliability of the proposed alignment pipeline.

4 MuseAdapt

4.1 Framework Overview

MuseAdapt aims to generate a new musical realization that preserves the melodic identity of a reference recording while adapting its acoustic characteristics according to target conditions. As illustrated in Fig. 2, the proposed framework consists of two main components: a cross-realization melody representation module and a controllable music generation module.

First, we learn a melody representation from aligned music pairs with shared compositions but diverse acoustic realizations. During inference, the resulting melody representation is extracted from the source recording and used as a condition for music generation. Building upon a pretrained Stable Audio Open backboneEvans et al. (2024c), MuseAdapt introduces a ControlNet-based conditioning branch to incorporate the learned melody representation, while key information is provided through a lightweight adapter to enable additional control over musical realization. After generation, we further apply reward alignment with preference optimization to improve the balance between melody preservation and target condition adherence.

4.2 Cross-Realization Melody Representation Learning

A central challenge in melody-preserving music adaptation is obtaining a melody representation that captures shared musical structures while avoiding excessive dependence on source-specific acoustic characteristics. To this end, MuseAdapt learns melody representations from cross-realization supervision. Specifically, we construct training pairs (x_a, x_b), where two recordings share the same underlying composition but differ in acoustic realization, such as instrumentation, timbre,

and performance style. Such paired recordings provide natural supervision for learning realization-consistent melodic representations.

Given an input recording, we first extract its normalized CQT representation and feed it into a temporal melody encoder. The encoder consists of temporal convolutional downsampling layers followed by a Transformer encoder, producing a time-varying latent representation z that encodes melodic evolution over time. Different from conventional melody conditions based on manually designed chroma or CQT features, the proposed representation is learned directly from real-world cross-realization pairs and is optimized for melody-preserving adaptation.

To encourage the learned representation to capture composition-level melodic information, we jointly optimize the encoder with reconstruction and cross-realization objectives. Specifically, a decoder reconstructs a melody-enhanced acoustic representation from the learned latent representation, where the target is a reweighted CQT representation designed to emphasize salient melodic components. Following the idea that realization-specific information should not be required to reside in the melody representation, we introduce cross-reconstruction, where the melody representation from one realization is combined with the auxiliary realization information of another recording. In our implementation, the decoder additionally receives global acoustic features extracted by a pretrained MERT model, allowing the decoder to model realization-related characteristics while encouraging the latent representation to focus on shared melodic structures.

The overall training objective is defined as:

$$\mathcal{L}_{mel} = \mathcal{L}_{rec} + \lambda_{cross}\mathcal{L}_{cross} + \lambda_{align}\mathcal{L}_{align} + \lambda_{tri}\mathcal{L}_{tri}, \tag{1}$$

where $\mathcal{L}_{rec}$ is the reconstruction loss that preserves sufficient musical information in the latent representation, $\mathcal{L}_{cross}$ is the cross-reconstruction loss that promotes realization consistency, $\mathcal{L}_{align}$ is a Soft-DTW based temporal alignment loss that accounts for local timing variations between different performances, and $\mathcal{L}_{tri}$ is a triplet loss that pulls together representations from the same composition while separating those from different compositions.

After training, the melody encoder is frozen and used to extract melody representations from source recordings during inference. The resulting representation is then injected into the generation backbone as the melody condition for melody-preserving music adaptation. Detailed encoder architecture, decoder design, and optimization settings are provided in Appendix C.

4.3 Melody-Preserving Music Generation

Given the learned melody representation, MuseAdapt performs controllable music generation by adapting a pretrained text-to-music model while preserving the melodic identity of the source recording. We build upon Stable Audio Open as the generation backbone and introduce melody and key conditioning pathways to guide the adaptation process.

Following recent melody-controlled generation methods such as Stable Audio Control Hou et al. (2025), we adopt a ControlNet-style conditioning branch to inject the learned melody representation into the diffusion backbone. Specifically, the melody representation extracted by the encoder is projected into the hidden space of the transformer blocks and added to the corresponding backbone features through the additional control branch. This design provides explicit melody guidance while maintaining the strong generation prior of the pretrained model, allowing the model to modify instrumentation and style according to target prompts rather than simply reconstructing the source realization.

In addition to melody conditioning, we introduce a lightweight key adapter to provide explicit control over the global musical key, inspired by prior controllable music generation frameworks such as MuseControlLite Tsai et al. (2025). Specifically, the key condition is represented by a learnable embedding:

$$e_k = \mathrm{Embedding}(k), \quad k \in \{0, \ldots, 11\}, \tag{2}$$

and injected into the generation network through the conditioning pathway. Unlike the melody representation, which captures time-varying melodic structures, the key embedding provides global pitch-level guidance and helps the model generate musically consistent adaptations under different target conditions.

By jointly conditioning on learned melody representations, text descriptions, and musical key information, MuseAdapt enables melody-preserving adaptation with flexible changes in acoustic realization.

4.4 Reward Alignment with Preference Optimization

To further align generated music with the adaptation objective, we perform preference optimization using automatically constructed rewards. Specifically, we define a composite reward that jointly considers melody preservation and text-condition alignment:

$$R(y) = \lambda_m R_{\text{mel}}(y, a) + \lambda_t R_{\text{text}}(y, c), \tag{3}$$

where a denotes the source recording and c denotes the target text condition. R_{mel} measures melody consistency between the generated audio and the source recording, while R_{text} measures semantic alignment with the target prompt using CLAP Wu et al. (2023).

Based on the composite rewards, we construct preference pairs between generated samples and optimize MuseAdapt with Direct Preference Optimization (DPO) Rafailov et al. (2023). This alignment encourages the model to preserve the source melody while generating acoustic realizations that better satisfy the target conditions.

5 Experiments

5.1 Experimental Setup

Dataset and Tasks. MAMA-20k contains 13,079 compositions, 21,343 real-world instrumental recordings, and 283,461 aligned cross-realization segment pairs. We train MuseAdapt and all baselines on the training split of MAMA-20k and evaluate them on a held-out test set constructed from cross-realization pairs. The test candidates are manually inspected, achieving a human-verified alignment precision of 93.7%, and inconsistent pairs are removed before evaluation.

The first setting is *self-reconstruction*, where the melody condition is extracted from the same recording used as the generation target. This setting follows the commonly adopted training and evaluation paradigm of existing melody-conditioned generation methods and is used to evaluate the tendency of models to reproduce the original acoustic realization.

The second setting is *melody-preserving adaptation*, where the model receives a source recording and a target condition describing the desired musical realization. The goal is to preserve the melodic identity of the source while generating a new realization consistent with the target condition.

Baselines. We compare MuseAdapt with representative melody-conditioned music generation methods covering different melody control paradigms. Specifically, we evaluate five configurations from three open-source models. MusicGen Copet et al. (2023) is a Transformer-based autoregressive text-to-music model, where we evaluate MusicGen-stereo-melody and MusicGen-stereo-melody-large, both using one-hot chroma features as melodic control signals. MuseControlLite Tsai et al. (2025) is a lightweight controllable music generation framework that enables time-varying control by injecting rotary positional embeddings into decoupled cross-attention layers and adopts Top-k CQT features as melody conditions. SongEcho Li et al. (2026), originally proposed for cover song generation, has been reimplemented within the Stable Audio Open framework by Tsai Fang-Duo et al.[1], enabling melody-controlled generation in the same latent diffusion framework. We evaluate both SongEcho-base and SongEcho-large configurations.

For fair comparison, MuseControlLite, SongEcho, and MuseAdapt are built upon the Stable Audio Open backbone, while MusicGen provides an autoregressive comparison. All controllable generation models are fine-tuned on MAMA-20k using the same training split and optimization budget, while following their corresponding training configurations and sampling strategies.

Implementation Details. MuseAdapt is fine-tuned from the Stable Audio Open backbone using the v-objectiveHo et al. (2022). We optimize the model with AdamW on four NVIDIA H200 GPUs

[1] https://huggingface.co/fundwotsai2001/Text-to-Music_control_family

Table 2: Comparison of audio quality and text–audio alignment under self-reconstruction and melody-preserving adaptation settings. $\Delta\mathrm{CLAP}_{score}$ denotes the CLAP degradation from self-reconstruction to adaptation. The best and second-best results are highlighted in **bold** and underlined, respectively.

Model	Self-Reconstruction			Adaptation			$\Delta\mathrm{CLAP}_{score}\downarrow$
	$\mathrm{FD}_{\mathrm{openl3}}\downarrow$	$\mathrm{KL}_{\mathrm{passt}}\downarrow$	CLAP $\uparrow$	$\mathrm{FD}_{\mathrm{openl3}}\downarrow$	$\mathrm{KL}_{\mathrm{passt}}\downarrow$	CLAP $\uparrow$	
MusicGen-Stereo-Melody	190.12	0.51	0.44	195.27	0.56	0.42	0.02
MusicGen-Stereo-Melody-Large	197.67	0.48	0.44	201.29	0.52	0.43	**0.01**
MuseControlLite	79.21	0.37	**0.45**	81.30	0.58	0.39	0.06
SongEcho-base	79.31	0.32	0.44	84.73	0.58	0.37	0.07
SongEcho-large	**78.92**	0.33	0.44	81.24	0.60	0.39	0.05
MuseAdapt	79.12	**0.28**	**0.45**	**80.19**	**0.30**	**0.44**	**0.01**

with a global batch size of 16. The generation model is trained for 16,000 steps with a learning rate of 5×10^{-5}. For reward alignment, we further optimize the model for 5,000 steps with a learning rate of 10^{-6}. The melody preservation and text alignment rewards are equally weighted with $\lambda_m = \lambda_t = 0.5$. For MuseAdapt inference, we employ 250-step DPM-Solver++Lu et al. (2022) sampling with classifier-free guidance (CFG) using a guidance scale of 7.

5.2 Evaluation Metrics

We evaluate model performance from three complementary perspectives: audio quality, text–audio alignment, and melody preservation.

Audio Quality and Text Alignment. Following the evaluation protocol of Stable Audio Open Evans et al. (2024c), we evaluate the quality of generated audio using distributional metrics, including $\mathrm{FD}_{\mathrm{openl3}}$ and $\mathrm{KL}_{\mathrm{passt}}$. In addition, we employ $\mathrm{CLAP}_{\mathrm{score}}$ to measure the semantic alignment between generated audio and the target text prompt.

Melody Preservation. Melody-preserving adaptation aims to maintain the melodic identity of the source recording while allowing flexible changes in acoustic realization. Therefore, we evaluate melody consistency using both frame-level pitch accuracy and alignment-aware metrics.

- **One-hot Chroma Accuracy.** Following previous melody-conditioned music generation methods Copet et al. (2023); Wu et al. (2024), we extract one-hot chroma sequences from the source and generated audio and compute the frame-level accuracy of pitch-class correspondence. This metric directly measures melodic consistency but is sensitive to temporal deviations.
- **Melody-Aligned Score (MAS).** To account for temporal variations between different musical realizations, we further introduce an alignment-aware metric. Given the similarity matrix S between source and generated melodies, we obtain the optimal alignment path $\mathcal{P}$ and calculate:

$$\mathrm{MAS} = \left(\frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} S_{i,j} \right) \cdot \min\left(\frac{L_{\mathcal{P}}}{T_{\mathrm{ref}}}, 1 \right), \tag{4}$$

  where the first term measures the average melody similarity along the alignment path and $L_{\mathcal{P}}$ denotes the duration coverage of the aligned region.
- **Key Accuracy.** We evaluate whether the generated audio preserves the desired musical key by comparing the predicted key of the generated audio with the target key condition.

6 Results

6.1 Audio Quality and Text–Audio Alignment

The quantitative results on audio quality and text–audio alignment are summarized in Table 2. In the self-reconstruction setting, most melody-conditioned baselines achieve competitive audio

Table 3: Comparison on melody preservation under self-reconstruction and melody-preserving adaptation settings. Best and second-best results are shown in **bold** and underline, respectively.

Model	Self-Reconstruction		Melody-Preserving Adaptation	
	One-hot Chroma Acc↑	MAS↑	One-hot Chroma Acc↑	MAS↑
MusicGen-Stereo-Melody	0.49	0.55	0.30	0.48
MusicGen-Stereo-Melody-Large	0.53	0.62	0.27	0.44
MuseControlLite	**0.57**	0.59	0.25	0.46
SongEcho-base	0.55	0.59	0.23	0.46
SongEcho-large	0.56	0.61	0.23	0.47
MuseAdapt	**0.57**	**0.66**	**0.31**	**0.54**

quality and text–audio alignment, reflecting the advantage of reconstructing the original acoustic realization. However, when moving from self-reconstruction to melody-preserving adaptation, several baselines exhibit noticeable degradation in text–audio consistency. Specifically, MuseControlLite and SongEcho show larger $\Delta\mathrm{CLAP}_{score}$ values, indicating that their melody conditions tend to interfere with target-condition following when the generated realization differs from the source recording.

In contrast, MuseAdapt maintains strong text–audio alignment under the adaptation setting, achieving a CLAP score of 0.44, comparable to its self-reconstruction performance (0.45). Meanwhile, MuseAdapt achieves the lowest $\mathrm{KL}_{\mathrm{passt}}$ under both settings and competitive $\mathrm{FD}_{\mathrm{openl3}}$, demonstrating that the proposed cross-realization representation learning enables flexible adaptation without sacrificing audio quality. These results suggest that learning melody conditions from cross-realization pairs reduces the reliance on source-specific acoustic cues and improves robustness in realistic adaptation scenarios.

6.2 Melody Preservation

The melody preservation results are presented in Table 3. Under the self-reconstruction setting, existing methods achieve relatively high melody consistency, as their melody conditions are extracted from the same recording used for generation. However, this advantage does not fully transfer to cross-realization adaptation.

In the melody-preserving adaptation setting, MuseAdapt achieves the best performance on both One-hot Chroma Accuracy and MAS, obtaining 0.31 and 0.54, respectively. Compared with existing methods, MuseAdapt provides more robust melodic preservation under acoustic realization changes, demonstrating that the learned melody representation captures composition-level melodic structures rather than source-specific recording characteristics.

Interestingly, while some baselines achieve comparable or even higher melody consistency in self-reconstruction, their performance decreases more substantially during adaptation. This gap highlights the limitation of self-reconstruction-based melody conditioning, where richer acoustic conditions may provide shortcuts for reproducing the original realization. By learning from cross-realization supervision, MuseAdapt better preserves melody while allowing flexible changes in instrumentation and acoustic characteristics.

7 Conclusion

In this work, we studied melody-preserving music adaptation and identified the limitation of self-reconstruction-based melody control, where source-specific acoustic cues may be inadvertently exploited for generation. To address this challenge, we proposed MuseAdapt, a cross-realization adaptation framework that learns robust melody representations from real-world instrumental pairs and enables flexible generation through controllable conditioning and reward alignment. We further introduced MAMA-20k, a large-scale dataset of melody-aligned cross-realization music pairs to facilitate this research. Experiments demonstrate that MuseAdapt better preserves melodic identity under adaptation scenarios while maintaining audio quality and target condition alignment, highlighting the importance of cross-realization supervision for controllable music generation.

AI USE STATEMENT

Generative AI tools were used to help organize the paper's narrative, edit language, draft LaTeX, and discuss figure layouts. The authors reviewed the generated text against the implementation and primary literature, and will verify all experimental claims, citations, code, and artifacts before submission. The tools were not treated as an authority for scientific results. The authors take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Persistent conversational memory introduces privacy and consent risks beyond those of a session-only assistant. Any release or deployment should make memory collection visible, restrict records to the correct user and application, support inspection and deletion, and avoid interpreting a voice embedding as real-world identity. Dataset licenses, speaker consent, demographic coverage, and access controls will be documented with the final experiments.

REPRODUCIBILITY STATEMENT

The final paper will specify the streaming schedule, layer indices, parameter accounting, data construction, speaker splits, hard-negative mining, trigger labels, all evaluation prompts, and the serving hardware used for latency and memory measurements. Code and configuration files will be included in anonymous supplementary material where policy permits.

REFERENCES

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Rachel M. Bittner, Juan J. Bosch, David Rubinstein, Gabriel Meseguer-Brocal, and Sebastian Ewert. A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation. In *Proc. ICASSP*, Singapore, 2022.

Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. MusicLDM: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *Proc. ICASSP*, Seoul, 2024.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Proc. NeurIPS*, New Orleans, 2023.

Junyu Dai, Xinyue Fan, Weiqin Li, Xiangang Li, Yunjia Li, Bin Ma, Yukun Ma, Chongjia Ni, Yufei Shi, Haoxu Wang, Menglin Wu, Jianwei Yu, Huaicheng Zhang, Han Zhao, Shengkui Zhao, and Haina Zhu. Pushing the frontier of full-song generation: Hierarchical autoregressive planning meets flow-matching rendering. *arXiv preprint arXiv:2607.20253*, 2026.

Zach Evans, Cj Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. In *Proc. ICML*, Vienna, 2024a.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Long-form music generation with latent diffusion. *arXiv preprint arXiv:2404.10301*, 2024b.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. *arXiv preprint arXiv:2407.14358*, 2024c.

Jason Ho, Ajay Jain, and Pieter Abbeel. Progressive distillation for fast sampling of diffusion models. In *Proc. ICLR*, 2022.

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 *Proc. ICASSP*, Hyderabad, 2025.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

Yun-Han Lan, Wen-Yi Hsiao, Hao-Chung Cheng, and Yi-Hsuan Yang. Musicongen: Rhythm and chord control for transformer-based text-to-music generation. 2024.

Wei-Jaw Lee, Hsuan-Yu Yeh, Ting-Yi Hu, Chih-Pin Tan, Fang-Duo Tsai, and Yi-Hsuan Yang. MPE-cho: A melody and phoneme-aware generative framework for controllable cover song generation. In *Proc. ISMIR*, Abu Dhabi, 2026.

Peike Patrick Li, Boyu Chen, Yao Yao, Yikai Wang, Allen Wang, and Alex Wang. JEN-1: Text-guided universal music generation with omnidirectional diffusion models. In *Proc. CAI*, Singapore, 2024.

Sifei Li, Yang Li, Zizhou Wang, Yuxin Zhang, Fuzhang Wu, Oliver Deussen, Tong-Yee Lee, and Wei Dong. Songecho: Towards cover song generation via instance-adaptive element-wise linear modulation. In *Proc. ICLR*, Rio de Janeiro, 2026.

Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. In *Proc. ICLR*, Singapore, 2025.

Liwei Lin, Gus Xia, Junyan Jiang, and Yixiao Zhang. Content-based controls for music large language modeling. *arXiv preprint arXiv:2310.17162*, 2023.

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, 2024.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proc. AAAI*, 2024.

Javier Nistal, Marco Pasini, Cyran Aouameur, Maarten Grachten, and Stefan Lattner. Diff-a-riff: Musical accompaniment co-creation via latent diffusion models. In *Proc. ISMIR*, San Francisco, 2024a.

Javier Nistal, Marco Pasini, and Stefan Lattner. Improving musical accompaniment co-creation via diffusion transformers. In *Proc. NeurIPS*, Vancouver, 2024b.

Julian D Parker, Janne Spijkervet, Katerina Kosta, Furkan Yesiler, Boris Kuznetsov, Ju-Chiang Wang, Matt Avent, Jitong Chen, and Duc Le. Stemgen: A music generation model that listens. In *Proc. ICASSP*, Seoul, 2024.

Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: your language model is secretly a reward model. In *Proc. Neurips*, New Orleans, 2023.

Simon Rouard, Yossi Adi, Jade Copet, Axel Roebel, and Alexandre Défossez. Audio conditioning for music generation via discrete bottleneck features. In *Proc. ISMIR*, San Francisco, 2024.

Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. Moûsai: Efficient text-to-music diffusion models. In *Proc. ACL*, Bangkok, 2024.

Temple F Smith, Michael S Waterman, et al. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

Or Tal, Alon Ziv, Itai Gat, Felix Kreuk, and Yossi Adi. Joint audio and symbolic conditioning for temporally controlled text-to-music generation. *arXiv preprint arXiv:2406.10970*, 2024.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

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 *Proc. ISMIR*, San Francisco, 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 *Proc. ICML*, Vancouver, 2025.

Michael S Waterman and Mark Eggert. A new algorithm for best subsequence alignments with application to trna-rrna comparisons. *Journal of molecular biology*, 197(4):723–728, 1987.

Haojie Wei, Xueke Cao, Tangpeng Dan, and Yueguo Chen. Rmvpe: A robust model for vocal pitch estimation in polyphonic music. In *Proc. Interspeech*, Dublin, 2023.

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.

Yusong Wu, K. Chen, Tian-Yu Zhang, Yu-Chen Hui, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proc. ICASSP*, Rhodes Island, 2023.

Jin Xu, Kangdi Wang, Ruibin Yuan, Shun Lei, Xiong Wang, Xize Cheng, Xueyao Zhang, Yang Zhang, Yiheng Chen, Yongqi Wang, Yue Wang, Zhifang Guo, Zihan Liu, Zijian Lin, Dake Guo, Hangrui Hu, Lei Xie, Linhan Ma, Wei Xue, Wenxiang Guo, Xinfa Zhu, Xipin Wei, Yangze Li, Yuanjun Lv, Yuxuan Wang, Yunfei Chu, and Zhiyong Wu. Qwen-Music technical report. *arXiv preprint arXiv:2607.11699*, 2026.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. ICCV*, Paris, 2023.

Jialing Zou, Jiahao Mei, XuDong Nan, Jinghua Li, Daoguo Dong, and Liang He. Teadapter: Supply vivid guidance for controllable text-to-music generation. In *Proc. ICME*, Niagara Falls, 2024.

A Related Work

A.1 Controllable Music Generation

Recent text-to-music models have achieved substantial progress in generating high-quality music from natural-language descriptions, spanning autoregressive approaches such as MusicLM Agostinelli et al. (2023) and MusicGen Copet et al. (2023), as well as diffusion- and flow-based models including AudioLDM 2 Liu et al. (2024), Moûsai Schneider et al. (2024), JEN-1 Li et al. (2024), and Stable Audio Evans et al. (2024a;b;c). Beyond global text conditioning, an increasing body of work has explored finer temporal control over musical structure. Coco-Mulla Lin et al. (2023), Music ControlNet Wu et al. (2024), JASCO Tal et al. (2024), TEAdapter Zou et al. (2024), and MuseControlLite Tsai et al. (2025) incorporate conditions such as melody, chords, rhythm, or other time-varying musical features into generative models. Related approaches have also investigated reference-audio conditioning for style, timbre, and accompaniment control, including MusicGen-Style Rouard et al. (2024), Audio Prompt Adapter Tsai et al. (2024), Diff-A-Riff Nistal et al. (2024a;b), and StemGen Parker et al. (2024). More recently, large-scale song generation systems such as SongEcho Li et al. (2026), MPEcho Lee et al. (2026), Qwen-Music Xu et al. (2026), and FullSong Dai et al. (2026) have further demonstrated the potential of reference-conditioned generation for cover-song creation.

A.2 Melody-Conditioned Music Generation

A central challenge in melody-conditioned music generation is how to represent melody while suppressing source-specific acoustic information. Early approaches often rely on symbolic or hand-crafted conditions. Coco-Mulla Lin et al. (2023) uses symbolic representations such as piano rolls, while MusicGen Copet et al. (2023) and Music ControlNet Wu et al. (2024) adopt one-hot chromagrams after observing that richer chromagram features can encourage reconstruction of the reference audio. More recent systems such as Stable Audio Control Hou et al. (2025) and MuseControlLite Tsai et al. (2025) adopt richer CQT-based conditions for finer pitch and temporal control, together with regularization strategies such as feature masking to mitigate over-reliance on the reference signal.

Recent cover-song generation methods, including SongEcho Li et al. (2026), MPEcho Lee et al. (2026), and Qwen-Music Xu et al. (2026), commonly exploit vocal pitch trajectories extracted with dedicated vocal F_0 estimators such as RMVPE Wei et al. (2023). This provides a relatively clean melody signal when the singing voice is the dominant melody carrier. Our setting instead focuses on instrumental, multi-instrument recordings, where the melodic content is distributed across different instruments and embedded within dense acoustic mixtures, making vocal-pitch-based extraction inapplicable. General-purpose transcription systems such as Basic Pitch Bittner et al. (2022) can estimate polyphonic notes from instrumental recordings, but are primarily designed for single-instrument transcription and do not directly provide a robust lead-melody representation for dense multi-instrument mixtures. Motivated by this gap, MuseAdapt learns a temporal melody representation directly from real-world cross-realization pairs, rather than relying solely on hand-crafted acoustic conditions or predefined vocal-pitch extractors.

B Details of the Melody Alignment Pipeline

We employ a two-stage alignment pipeline to mine temporally corresponding segments from different real-world realizations of the same composition. The first stage performs high-recall candidate mining over full recordings, while the second stage locally refines and verifies the retrieved candidates using a sparser pitch representation.

B.1 Stage I: Coarse Candidate Mining

Given two recordings, we first resample the audio to 22.05 kHz and extract Chroma Energy Normalized Statistics (CENS) features with a hop size of 512 samples. Since different realizations may be performed in different keys, we apply an Optimal Transposition Index (OTI) procedure before temporal alignment. Specifically, we circularly shift one chroma representation over the 12 pitch classes and select the transposition that maximizes the similarity between the two recordings.

We then construct a frame-level cosine-similarity matrix

$$S_{ij} = \frac{\mathbf{c}_i^{(r)\top} \mathbf{c}_j^{(t)}}{\|\mathbf{c}_i^{(r)}\|_2 \|\mathbf{c}_j^{(t)}\|_2}, \tag{5}$$

where $\mathbf{c}_i^{(r)}$ and $\mathbf{c}_j^{(t)}$ denote chroma vectors from the two realizations. To reduce recording-dependent variation in similarity magnitude, the similarities are normalized along each reference frame and mapped through a sigmoid function. We additionally apply multi-scale diagonal smoothing to favor temporally coherent correspondences over isolated high-similarity frames.

Candidate paths are extracted using Smith–Waterman local sequence alignment Smith et al. (1981). For a similarity baseline b and gap penalty $g < 0$, the dynamic-programming score is computed as

$$\begin{aligned} H(i,j) = \max \Big\{ 0,\ & H(i-1, j-1) + 2(S_{ij} - b), \\ & H(i-1, j) + g,\ H(i, j-1) + g \Big\}. \end{aligned} \tag{6}$$

We use Waterman–Eggert-style iterative extraction Waterman & Eggert (1987) to recover multiple local correspondences from the same recording pair. After extracting each path, its neighborhood is suppressed before searching for subsequent alignments. Low-confidence portions of each path are pruned according to smoothed frame-level similarity, after which temporally compatible segments are merged and redundant overlaps are removed.

To improve recall for long recordings and accommodate repeated musical sections, alignment is performed with overlapping 40-s windows with a 10-s overlap. The resulting paths are subsequently merged in global time. We use adaptive similarity thresholds derived from the similarity distribution of each recording pair and retain coarse correspondences with a final aligned duration of at least 20 s.

B.2 Stage II: Melody-Focused Local Refinement

The coarse stage is intentionally recall-oriented and may retain correspondences dominated by common harmonic or accompaniment patterns. We therefore perform a second-stage local refinement for each retrieved candidate.

For each candidate region, we extend both boundaries by 5 s and compute CQT-based chroma features. At every frame, only the three strongest pitch-class activations are retained:

$$\widetilde{\mathbf{c}}_t = \mathrm{TopK}(\mathbf{c}_t, k = 3), \tag{7}$$

with all remaining entries set to zero, followed by ℓ_2 normalization. This sparse representation emphasizes dominant pitch content and reduces the influence of weaker accompaniment components.

We again compensate for global transposition, construct a cosine-similarity matrix, and perform Smith–Waterman/Waterman–Eggert local alignment within the padded candidate region. Adaptive thresholds are recomputed for each candidate, after which low-confidence paths are pruned, nearby paths are merged, and redundant overlaps are removed. Refined segments shorter than 20 s are discarded.

Finally, each coarse candidate is assigned a verification score that jointly accounts for alignment quality and temporal coverage:

$$R_{\mathrm{align}} = \overline{S}_{\mathrm{path}} \cdot \frac{L_{\mathrm{aligned}}}{L_{\mathrm{candidate}}}, \tag{8}$$

where $\overline{S}_{\mathrm{path}}$ is the mean cosine similarity along the best refined alignment path and $L_{\mathrm{aligned}}/L_{\mathrm{candidate}}$ measures the fraction of the coarse candidate covered by the refined correspondence. Candidates with insufficient similarity or coverage are removed; in our implementation, we retain candidates with $R_{\mathrm{align}} \geq 0.35$. The remaining segments constitute the cross-realization pairs used to train MuseAdapt.

C Melody Encoder Details

C.1 Melody Encoder Architecture

The melody encoder takes normalized CQT representations as input and produces a temporal melody representation. Specifically, the input CQT is processed by three temporal convolutional downsampling blocks with kernel size 4 and stride 2, reducing the temporal resolution by a factor of 8 while increasing the feature dimension from 128 to 1024. The resulting features are further processed by a 6-layer pre-normalized Transformer encoder with 1024 hidden dimensions and 8 attention heads.

The encoder outputs a temporal latent representation:

$$z \in \mathbb{R}^{T/8 \times 1024},$$

which preserves the temporal evolution of melodic patterns and is used as the melody representation for generation.

C.2 MERT-Conditioned Reconstruction

To prevent the melody representation from encoding realization-specific acoustic information, we explicitly provide acoustic realization features to the reconstruction decoder using a pretrained MERT model. Specifically, we extract hidden representations from layers 6, 12, 18, and 24 of MERT. For each layer, we apply temporal mean and standard deviation pooling to obtain a global representation:

$$r_i = [\mu(h_i^6), \sigma(h_i^6), \mu(h_i^{12}), \sigma(h_i^{12}), \mu(h_i^{18}), \sigma(h_i^{18}), \mu(h_i^{24}), \sigma(h_i^{24})],$$

resulting in a realization feature with dimension 4×2048.

The extracted MERT feature is then mapped to FiLM parameters:

$$(\gamma, \beta) = F_\phi(r_i),$$

which modulate the melody latent representation before decoding:

$$\tilde{z} = z \odot (1 + \gamma) + \beta.$$

The modulated representation is processed by a 6-layer Transformer decoder to reconstruct the target reweighted CQT representation. During cross-realization training, the melody representation from one recording is combined with the MERT realization feature from another recording:

$$\hat{Y}_b = D(z_a, r_b),$$

forcing z_a to preserve composition-level melodic information while delegating realization-related information to the MERT branch.

C.3 Reweighted CQT Reconstruction Target

Instead of reconstructing the original CQT representation directly, we construct a reweighted CQT target that emphasizes melody-salient spectral components. Given a normalized CQT representation X_{01}, we first compute a local contrast map:

$$C = \mathrm{ReLU}(X_{01} - B),$$

where B denotes a frequency-temporal baseline estimated from local spectral statistics. We then compute temporal and frequency context terms:

$$Q = \mathrm{AvgPool}_{time}(C, 7), \qquad D = \mathrm{AvgPool}_{freq}(C, 17),$$

and obtain the saliency map:

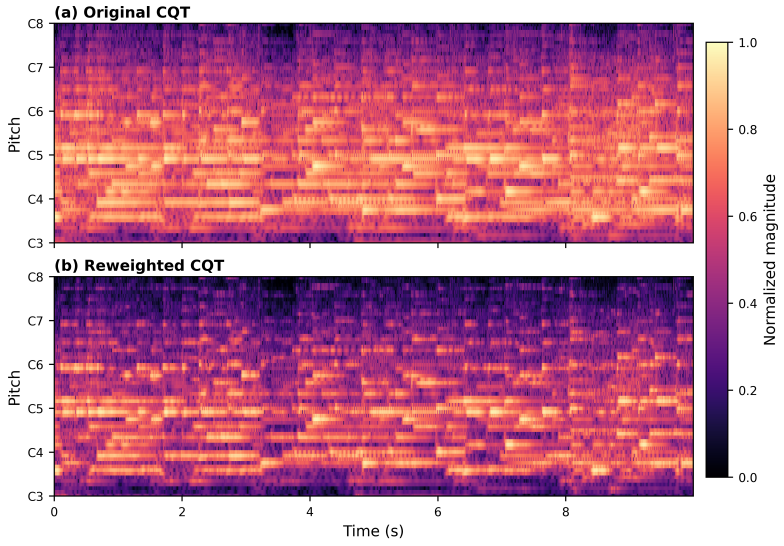


Figure 3: Visualization of the reweighted CQT reconstruction target. Compared with the original CQT, the reweighted target suppresses less melody-relevant spectral energy and emphasizes salient pitch trajectories, providing a more melody-focused supervision signal for representation learning.

$$R = \frac{Q}{D + \epsilon}.$$

After normalization and nonlinear scaling, we derive a weighting map S that highlights melody-dominant regions. The final reweighting operation is:

$$X_{focus} = \max(X_{01}, \epsilon)^{\gamma},$$

where the adaptive exponent is determined by the saliency map:

$$\gamma = 1.8 - 0.95S.$$

The reconstruction target is then generated as:

$$Y = 0.25X_{01} + 0.75X_{focus}.$$

This reweighted target preserves temporal pitch evolution while reducing the relative contribution of less melody-relevant spectral variations, providing a more suitable supervision signal for cross-realization melody representation learning. A visual comparison between the original CQT and the reweighted target is shown in Fig. 3.

C.4 Cross-Realization Objectives

For a paired sample (x_a, x_b), the encoder produces two melody representations z_a and z_b. Besides conventional reconstruction, we introduce cross-reconstruction:

$$\hat{y}_b = D(z_a, r_b),$$

where the melody representation from one realization is combined with the realization feature of another. This objective encourages the latent representation to capture shared melodic structures instead of encoding source-specific acoustic characteristics.

To account for temporal variations between different performances, we employ a Soft-DTW based alignment loss between paired melody representations:

$$\mathcal{L}_{align} = \mathrm{SoftDTW}(z_a, z_b).$$

Soft-DTW provides a differentiable temporal alignment objective and allows the model to handle local tempo deviations between different realizations.

Finally, we apply a triplet objective to improve representation discriminability. Given an anchor representation z_a, its positive sample z_p is from the same composition with a different realization, while the negative sample z_n corresponds to a different composition:

$$\mathcal{L}_{tri} = \max(d(z_a, z_p) - d(z_a, z_n) + m, 0).$$

The complete objective follows:

$$\mathcal{L}_{mel} = \mathcal{L}_{rec} + \lambda_{cross}\mathcal{L}_{cross} + \lambda_{align}\mathcal{L}_{align} + \lambda_{tri}\mathcal{L}_{tri}.$$