CREATING MUSIC AS A GO PLAYER: A PROCESS REWARD MODEL OF MUSIC GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating and aligning music generation systems remains challenged by the gap between static, outcome-level labels and the intrinsically dynamic, process-dependent nature of musical quality. Existing post-training methods rely primarily on coarse global rewards, restricting fine-grained policy optimization and leaving models incapable of localized error correction. Inspired by value function temporal difference (TD) learning in computer Go (e.g., AlphaGo)—where dense intermediate values are derived from sparse outcomes—we introduce MusicGo, a process reward model that assigns dense temporal rewards to partial musical sequences. By framing process reward estimation as predicting the likelihood that a prefix leads to a high-quality completion within a TD framework, MusicGo converts sparse track-level preferences into fine-grained process supervision with minimal human process annotations. Correspondingly, we release MusPR, the first dataset for training and validating process reward models in music generation. Furthermore, we evaluate MusicGo in offline and online reinforcement learning settings and present PassGRPO (Passage-level Group Relative Policy Optimization), an enhanced GRPO variant that executes policy updates at passage granularity to produce structurally coherent generations. By establishing frame-level reward diagnostics and passage-wise gradient updating, our framework lays a precedent for targeted music revision, localized inpainting, and fine-grained interactive editing. We make our audio samples available here.

1 INTRODUCTION

Recent generative models have made it possible for end-to-end synthesis of high-fidelity, musically coherent compositions (Agostinelli et al., 2023; Evans et al., 2024; Yuan et al., 2025; Lei et al., 2025). Evaluating these systems, however, requires more than an assessment of overall audio quality. Musical judgments depend on events distributed across time, including motif development, rhythmic continuity, harmonic transitions, and the relationship between local passages and larger-scale structure. A single track-level score does not explicitly reveal where a generation becomes compelling, loses coherence, or recovers from a local defect. Consequently, a static sequence-level score does not explicitly indicate where quality improves or deteriorates, or how individual passages contribute to the overall assessment.

This limitation affects both evaluation and post-training alignment algorithms, such as Reinforcement Learning from Human Feedback (RLHF) and Group-Relative Policy Optimization (GRPO). Such approaches rely on global outcome rewards assigned at the full-song level, unable to provide sufficient dense supervision for token- or segment-level optimization. Consequently, attributing an overall judgment to particular passages remains difficult, especially when acceptable and problematic regions coexist within the same track. Temporally resolved feedback offers a complementary source of information for reward-guided decoding and policy optimization.

Our approach is motivated by intermediate-state evaluation in computer Go. AlphaGo learns a value network from self-play outcomes to assess positions before a game terminates via temporal difference (TD) learning. (Silver et al., 2016). Music generation presents a structural parallel: by predicting the probability that a partial audio prefix will evolve into a high-quality piece, we can formulate a continuous process reward that respects musical temporality without relying on dense, expensive human step-by-step annotations. We therefore use this analogy to motivate prefix evaluation, without assuming that subjective musical judgments constitute calibrated winning probabilities.

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.

We introduce MusicGo, a causal music **Process Reward Model (PRM)**. Unlike conventional **Outcome Reward Models (ORMs)** that score an audio track holistically, a music PRM provides continuous, causal feedback $r(c, x_{\leq t})$ for a partial audio sequence $x_{\leq t}$ given prompt c. It measures the margin by which a musical prefix can be successfully extended into a coherent full composition. By capturing transient quality fluctuations—such as acoustic degradation, rhythmic drift, or harmonic collapse—music process reward converts coarse sequence-level evaluation into fine-grained temporal supervision. Our contributions are threefold:

- **MusicGo**: We introduce a process reward framework that leverages temporal difference learning on both coarse outcome labels and fine-grained annotations, deriving dense temporal reward signals without requiring dense step-by-step supervision.
- **MusPR**: We construct a benchmark dataset comprising $36{,}098$ pieces with binary quality labels and $1{,}274$ tracks with fine-grained frame-level annotations (highlight/degradation spans, overall quality, and individual preferences) for training and validating music process reward models.
- **PassGRPO**: We propose **PassGRPO**, a passage-granularity GRPO variant that fosters structurally coherent generation, and evaluate its use of MusicGo in offline and online RL settings. Furthermore, we showcase MusicGo's potential in segment-level rewriting, offering a paradigm for interactive human-AI co-creation while maintaining global contextual continuity.

2 Related Work

Music Evaluation and Music Reward Model Music evaluation has progressed from handcrafted, reference-free audio descriptors toward reference-based metrics, multimodal embedding representations, and learned song-level evaluators. Conventional metrics assess specific dimensions of musicality—such as pitch and rhythmic regularity, tonal consistency, and structural self-similarity—as well as audio fidelity and text-audio alignment (Duan et al., 2026; Wang et al., 2023). Modern approaches further leverage learned audio-language representations and distributional comparisons to evaluate generated music against reference corpora (Kilgour et al., 2019). While providing reproducible diagnostics, these methods assess complementary aspects of musical quality: reference-based metrics are highly sensitive to the composition of reference distributions (Kilgour et al., 2019), whereas reference-free statistical metrics often capture surface regularity while may not fully reflect long-range structure, expressive development, and perceptual preference.

To provide a unified evaluation across multiple holistic quality dimensions, learned song-level evaluators and benchmarks have been introduced, including SongEval (Yao et al., 2025a), SongBench (Wu et al., 2026), MusicEval (Liu et al., 2025), and AudioBox (Vyas et al., 2023; Duan et al., 2026; Kim et al., 2026; Wang et al., 2023). Nevertheless, these systems primarily output aggregate, track-level quality scores. Due to the lack of explicit temporal resolution, their aggregate outputs do not directly supply temporally localized rewards for autoregressive generation.

Generative Process Reward Process Reward Models (PRMs) shift supervision from final outcomes to intermediate generative steps, facilitating error localization, search guidance, and iterative correction (Zheng et al., 2026). This paradigm has proven effective in mathematical reasoning and code generation, where intermediate steps align with verifiable objective ground truths (e.g., derivational constraints or execution and unit-test feedback) (Zhang et al., 2026; Lin et al., 2025b; Li et al., 2025). Beyond symbolic domain reasoning, recent multimodal studies extend process supervision to visual generation by verifying intermediate reasoning trajectories, employing visual programs, or assigning feedback across diffusion timesteps (Guo et al., 2025; Gao et al., 2025; Lin et al., 2025a). Furthermore, to enhance long-horizon consistency and scalability, recent architectures reformulate reward prediction into generative QA tasks to mitigate reward hacking (Wu et al., 2025), or optimize context selection via policy optimization for long video generation (Fang et al., 2025).

Despite these advances, process supervision beyond language reasoning remains underexplored, as existing methods rely heavily on task-specific representations, domain mechanics (e.g., diffusion timesteps), or external visual verifiers. Applying PRMs to music generation poses distinct challenges: open-ended compositions lack canonical trajectories, local validity does not guarantee global quality, and opposing sequences share minimal token overlap. Consequently, exact-answer rollouts and preference-tree constructions from symbolic reasoning cannot be directly applied, highlighting the need for temporally localized, preference-grounded process supervision in audio generation.

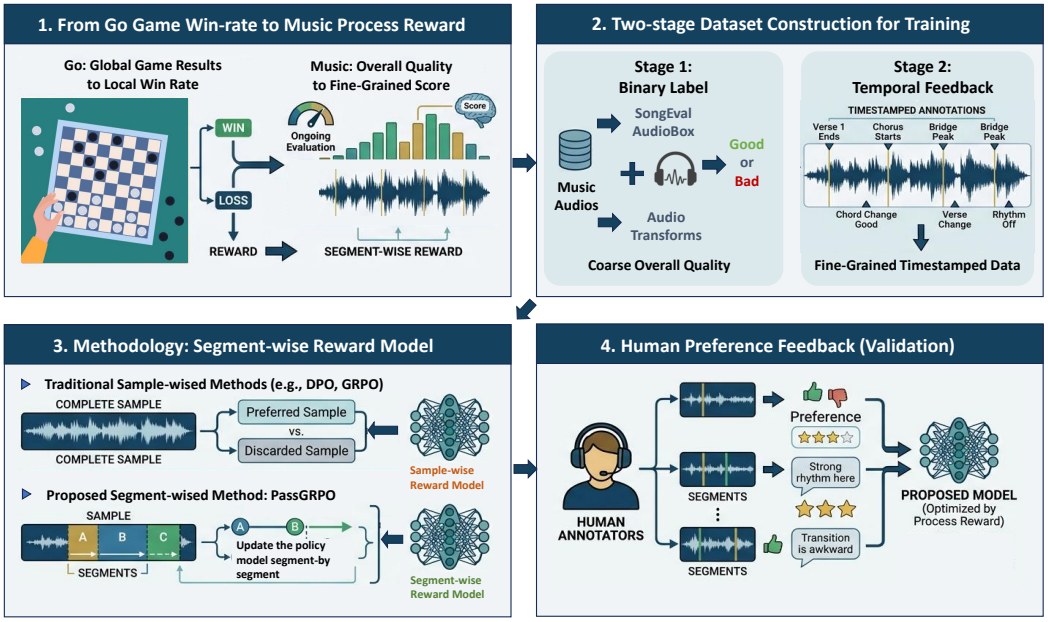

Figure 1: Overview of the whole process of formulating the problem of music process reward.

3 Dataset Construction

We construct **MusPR**, a dataset of 36,098 tracks built through a two-stage pipeline: **Stage I** provides coarse-grained, song-level quality labels across diverse sources, while **Stage II** equips a 1,274-track subset with fine-grained, timestamped human feedback. All audio tracks are encoded via the shared tokenizer (Section 4), with conditioning information, label provenance, and temporal annotations stored in a unified manifest.

Stage I: Coarse-grained quality supervision. To capture broad quality extremes, we collect real-world recordings and outputs from Suno (Muse; Jiang et al., 2026), ACE-Step 1.5 (Gong et al., 2026), and HeartMuLa (Yang et al., 2026). High- and low-confidence tracks are isolated using aggregated assessments from SongEval (Yao et al., 2025b) and Audiobox Aesthetics (Tjandra et al., 2025), while intermediate tracks are pseudo-labeled via a human-rated seed set. Crucially, quality labels are assigned based on track characteristics rather than data sources. To enrich negative coverage, we apply time-/frequency-domain perturbations and continuation-based structural disruptions, retaining verified failures as targeted negative samples with timestamped transformation boundaries.

Stage II: Fine-grained human feedback. To collect dense temporal signals, human annotators log timestamped highlights and degradations during live playback on a dedicated platform. Nearby events are clustered and assigned musicological tags covering composition, performance, arrangement, or audio quality, complemented by global ratings for personal preference and production quality. Unmarked regions remain unannotated. Complete platform architecture and annotation calibration details are provided in Appendix C.2.

Process Data Transformation Temporal annotations are converted into win-rate targets aligned with the audio codec sequence. In Stage I and Stage II, degradation intervals and clustered highlight/degradation events are mapped from seconds onto codec sequence x as temporal envelopes $s, h \in [0,1]^T$, where s_t lowers and h_t raises the win-rate targets. The polarity of the target at the Track-level is provided by binary labels $y \in \{0,1\}$ in Stage I and calibrated slider targets V_T in Stage II.

To normalize heterogeneous raw slider ratings $q \in [-1,1]$ while preserving original polarities ($q \geq 0$ for highlights, $q < 0$ for degradations), we avoid standard z-score normalization and instead apply

a sign preservation scale per annotator a:

$$\sigma_a = \max\big(\mathrm{P}_{95}(|q|;\, a),\, 0.25\big), \qquad \tilde{q} = \mathrm{clip}(q/\sigma_a, -1, 1), \qquad V_T = \tfrac{1}{2}(1+\tilde{q}). \tag{1}$$

This ensures that $q = 0$ aligns with $V_T = 1/2$, matching both the classification limit y and the evaluation threshold $w_T \geq 1/2$. Finally, a monotonic baseline interpolates from $1/2$ to V_T with local highlight/degradation residuals, pinning the terminal token target to V_T, constructing the dataset $\mathcal{D}_{\mathrm{I}}$ and $\mathcal{D}_{\mathrm{II}}$.

4 MusicGo: Trajectory Win-Rate Reward Model

Generation and reward-model backbones. Our policy model is built upon a 1B-parameter music generation backbone. It comprises three primary components: a trainable Llama-based Touvron et al. (2023); Grattafiori et al. (2024) language model, along with the frozen tokenizer and detokenizer inherited from LeVo Lei et al. (2025). Furthermore, MusicGo adopts the same architectural design as the policy model. This decision stems from the rationale that pre-trained generative priors facilitate stable, frame-level contribution learning across the entire audio sequence. Details of the music generation model are shown in the Appendix A.

Trajectory win-rate prediction. Let π_θ denote an autoregressive audio language model generating discrete codec tokens $x = (x_1, \ldots, x_T) \in \mathcal{V}^T$ conditioned on prompt c. To provide granular feedback, a process reward model (MusicGo) must be *causal*—where $r(c, x_{\leq t})$ depends strictly on observed prefixes—and *temporally localized*, capturing localized quality drops and recoveries rather than collapsing credit into a single terminal scalar.

We propose MusicGo, a trajectory win-rate reward model r_ϕ outputting local win rates $w_t = \sigma(u_t) \in (0,1)$ at frame t, with neutral prior $w_{\mathrm{BOS}} = 1/2$. The model is trained in two progressive stages: Stage I acquires coarse global quality with limited temporal sensitivity from synthetic noise, while Stage II aligns r_ϕ with human preference annotations.

4.1 Stage I: Coarse Supervision

We partition sequence x into g-frame segments with endpoints $e_s = \min(sg, T)$ ($s \in \{1, \ldots, S\}$) and $e_0 \triangleq 0$. Segment increments $\Delta_s = w_{e_s} - w_{e_{s-1}}$ track local degradation ($\Delta_s < 0$) or recovery ($\Delta_s > 0$).

Stage I constructs a paired dataset $\mathcal{D}_{\mathrm{I}}$ of clean sequences x and synthetically degraded variants x', paired with global quality labels $y \in \{0,1\}$ and frame-aligned degradation envelopes $s_t \in [0,1]$. We optimize a joint objective to obtain parameter set $\phi_{\mathrm{I}}^\star$:

$$\mathcal{L}_{\mathrm{I}} = \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}} + \lambda_{\mathrm{soft}}\mathcal{L}_{\mathrm{soft}} + \lambda_{\mathrm{pair}}\mathcal{L}_{\mathrm{pair}} + \lambda_{\mathrm{CE}}\mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{reg}}. \tag{2}$$

Loss Formulation. (1) *Alignment Loss* $\mathcal{L}_{\mathrm{align}} = \mathbb{E}[\mathbb{I}_{\{y=1\}}[\frac{1}{2} - w_T]_+^2 + \mathbb{I}_{\{y=0\}}[w_T - \frac{1}{2}]_+^2 + (w_T - w_{T-g})^2]$ enforces terminal polarity and boundary continuity. (2) *Classification Loss* $\mathcal{L}_{\mathrm{CE}} = \mathbb{E}[-\log p_\phi(y \mid c, x)]$ trains a global head on pooled audio hidden states. (3) *Pairwise Loss* $\mathcal{L}_{\mathrm{pair}} = [\bar{w}_T^- - \bar{w}_T^+ + m]_+$ maintains a margin m between terminal win rates of clean and degraded sequences. (4) *Soft Ceiling Loss* $\mathcal{L}_{\mathrm{soft}}$ penalizes scores exceeding an envelope-bounded threshold:

$$\mathcal{L}_{\mathrm{soft}} = \mathbb{E}\left[\frac{\sum_{t=1}^T s_t\,[w_t - (\bar{w}_{\mathrm{ref}} - \alpha s_t)]_+^2}{\sum_{t=1}^T s_t + \epsilon}\right], \tag{3}$$

where reference $\bar{w}_{\mathrm{ref}} = \mathrm{sg}\left(\frac{1}{|\mathcal{S}_{\mathrm{pre}}|}\sum_{s\in\mathcal{S}_{\mathrm{pre}}} w_{e_s}\right)$ averages boundary scores across pre-degradation segments $\mathcal{S}_{\mathrm{pre}}$.

Regularization. $\mathcal{L}_{\mathrm{reg}} = \lambda_{\mathrm{shape}}\mathcal{L}_{\mathrm{shape}} + \lambda_{\mathrm{TD}}\mathcal{L}_{\mathrm{TD}} + \lambda_{\mathrm{sv}}\mathcal{L}_{\mathrm{sv}}$. Here, $\mathcal{L}_{\mathrm{shape}}$ smooths clean frames by penalizing token-level total variation, intra-segment oscillation, and frequent sign changes in Δ_s. $\mathcal{L}_{\mathrm{TD}} = \mathbb{E}[\frac{1}{S-1}\sum_{s=2}^S (w_{e_{s-1}} - \gamma\ \mathrm{sg}(w_{e_s}))^2]$ enforces discounted temporal consistency across adjacent segment boundaries. Finally, $\mathcal{L}_{\mathrm{sv}} = \mathbb{E}[\frac{1}{S}\sum_{s=1}^S |\nu_s - \mathrm{sg}(w_{e_s})|]$ fits a auxiliary segment-value head $\nu_s = f_\phi^{\mathrm{sv}}(\bar{h}_s)$. Hyperparameter details are provided in Appendix C.1.

4.2 Stage II: Human Process Alignment

While synthetic transformations grant acoustic sensitivity, they fail to model complex generational artifacts. Stage II fine-tunes $\phi_{\mathrm{I}}^{\star}$ on human annotations comprising localized degradation (s_t) and highlight (h_t) envelopes. We optimize $\mathcal{L}_{\mathrm{II}} = \mathcal{L}_{\mathrm{I}} + \lambda_{\mathrm{H}}\mathcal{L}_{\mathrm{H}}$, where $\mathcal{L}_{\mathrm{H}}$ imposes a symmetric lower floor on highlight passages:

$$\mathcal{L}_{\mathrm{H}} = \mathbb{E}\left[\frac{\sum_{t=1}^{T} h_t \left[(\bar{w}_{\mathrm{ref}} + \alpha' h_t) - w_t\right]_+^2}{\sum_{t=1}^{T} h_t + \epsilon}\right]. \tag{4}$$

The optimal parameters $\phi_{\mathrm{II}}^{\star}$ are frozen to serve as the MusicGo $r_{\phi_{\mathrm{II}}^{\star}}$ in policy optimization.

5 Policy Optimization Methods

5.1 Sequence-Level Preference Baseline: TDPO

We adopt Token-level Direct Preference Optimization (TDPO; Zeng et al., 2024) as a preference-based alignment baseline for music generation. TDPO constructs a token-aware objective from sequence-level preference pairs by incorporating sequential forward-KL corrections. This enables direct policy optimization without an explicit reward or value model, while its trajectory-wise correction regularizes policy deviations along both preferred and dispreferred sequences.

Given a prompt c and a preferred-dispreferred pair of music token sequences (x^+, x^-), we compute their policy-to-reference log-probability ratios and the forward KL divergences accumulated along each sequence.

$$\ell_\theta(c, x) = \sum_{t=1}^{|x|} \log \frac{\pi_\theta(x_t \mid c, x_{<t})}{\pi_{\mathrm{ref}}(x_t \mid c, x_{<t})}, \tag{5}$$

$$K_\theta(c, x) = \sum_{t=1}^{|x|} D_{\mathrm{KL}}(\pi_{\mathrm{ref}}(\cdot \mid c, x_{<t}) \,\|\, \pi_\theta(\cdot \mid c, x_{<t}))\,, \tag{6}$$

where π_{ref} is the frozen reference policy. The resulting objective is

$$\begin{aligned} \mathcal{L}_{\mathrm{TDPO}} &= -\mathbb{E}_{(c,x^+,x^-)\sim\mathcal{D}_{\mathrm{pref}}}\left[\log \sigma\big(\beta\left[\ell_\theta(c, x^+) - \ell_\theta(c, x^-) - \alpha \Delta K_\theta\right]\big)\right], \\ \Delta K_\theta &= K_\theta(c, x^-) - \mathrm{sg}\big[K_\theta(c, x^+)\big]\,, \end{aligned} \tag{7}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator, β controls the scale of the preference logits, and α determines the strength of the sequential KL correction.

TDPO therefore provides a sequence-preference baseline that directly optimizes the policy without training an explicit reward or value model. For music generation, it encourages preferred sequence patterns while constraining deviations from the reference generator. However, its supervision remains defined over complete preferred–dispreferred track pairs: it neither estimates the evolving quality of an audio prefix nor explicitly localizes quality changes to individual musical passages.

5.2 Process-Reward Baseline: Process GRPO

Unlike TDPO, Process GRPO uses the frozen MusicGo model to derive temporally localized rewards from the win-rate trajectory. However, it still postpones policy optimization until complete trajectories have been sampled.

We optimize the policy π_θ against a frozen Stage II MusicGo $r_{\phi_{\mathrm{II}}^{\star}}$, using a frozen copy of the supervised music LM as the reference policy π_{ref}. Below, we outline Process GRPO as our primary baseline and present **PassGRPO**, which redefines the temporal unit of policy optimization at the level of contiguous musical passages.

Process GRPO adapts the process-supervision framework of DeepSeekMath (Shao et al., 2024) to long-form audio generation. Given a group of G complete sequence completions $\{x^{(i)}\}_{i=1}^{G}$ sampled from prompt c, step rewards are defined as win-rate increments across a coarse grid of step size

L. Starting from a neutral onset prior $w_{e_0} \triangleq 1/2$, the step reward $r_k^{(i)}$ at step boundary $e_k = \min(kL, T)$ is given by:

$$r_k^{(i)} = w_{e_k}^{(i)} - w_{e_{k-1}}^{(i)}. \tag{8}$$

These step rewards are standardized across all steps and completions within the group to obtain $\tilde{r}_k^{(i)}$. The token-level advantage $\hat{A}_{i,t}$ is then computed as the return of future steps:

$$\hat{A}_{i,t} = \sum_{k:e_k \geq t} \tilde{r}_k^{(i)}. \tag{9}$$

Finally, a single post-hoc GRPO gradient step is applied across the full sequence with a per-token KL divergence penalty ($\beta = 0.02$).

5.3 Proposed Passage-Level Optimization: PassGRPO

PassGRPO further departs from this post-hoc optimization paradigm by interleaving passage generation, reward evaluation, and policy updates.

Instead of delaying optimization until full trajectories are completed, **PassGRPO** performs online policy updates over local passages of length L (Algorithm 1). This enables the reward model to use passage-level feedback before the full trajectory is completed, dynamically updating θ at each passage boundary.

We define block increments on a fine-grained grid stride g as $\delta_s^{(g)} = w_{e_s} - w_{e_{s-1}}$, anchored by the initial audio win rate $w_{e_0} \triangleq w_1$. Given an existing prefix $p^{(k)}$, iteration k samples G candidate continuations of length L, denoting the newly generated token set as $\mathcal{U}_i$. The passage-level quality scalar R_i is computed as the average win rate over $\mathcal{U}_i$:

$$R_i = \frac{1}{|\mathcal{U}_i|} \sum_{t \in \mathcal{U}_i} w_t^{(i)}. \tag{10}$$

To construct token advantages, the passage scalars R_i are standardized across the candidate group to yield $\tilde{R}_i$. Concurrently, local increments $\delta_t^{(g,i)}$ evaluated on a fine grid ($g = 25$, corresponding to 1 s) are standardized within each sequence to yield $\tilde{\delta}_t^{(i)}$. For any token $t \in \mathcal{U}_i$, the advantage is formulated as an equal combination of both metrics:

$$\hat{A}_{i,t} = \frac{1}{2}\tilde{R}_i + \frac{1}{2}\tilde{\delta}_t^{(i)}. \tag{11}$$

An immediate policy gradient update ($\beta_{\text{in}} = 0.08$) is applied exclusively to $\mathcal{U}_i$. To retain exploration during continuation selection, we mix score-based sampling with uniform sampling, the prefix for the next iteration $p^{(k+1)} = p^{(k)} \| u^{(i^\star)}$ is drawn using an ε-greedy soft-max distribution over R_i:

$$\pi_{\text{cont}}(i) = (1-\varepsilon)\frac{\exp((R_i - R_{\max})/\tau)}{\sum_{j=1}^{G} \exp((R_j - R_{\max})/\tau)} + \frac{\varepsilon}{G}, \tag{12}$$

where $R_{\max} = \max_j R_j$, and hyperparameters are set to $(\tau, \varepsilon) = (0.4, 0.15)$.

Once the trajectory terminates, the final candidate group is reused without additional sampling for a holistic song-level GRPO step ($\beta = 0.02$). The song-level scalar combines the terminal score $w_T^{(i)}$ with mean L-grid increments $\Delta_k^{(L,i)}$:

$$R_i^{\text{song}} = w_T^{(i)} + \frac{1}{2}\text{mean}_k \Delta_k^{(L,i)} + \frac{1}{2}\text{mean}_k \left[\Delta_k^{(L,i)}\right]_+. \tag{13}$$

Song-level token advantages are formed by evaluating Eq. equation 11 with $R_i \leftarrow R_i^{\text{song}}$ and L-grid increments standardized across the entire track.

6 Experiments

All prompts are drawn from MusPR's validation pool, comprising three subsets: **MusPR-S2-Eval** evaluates process-label prediction accuracy, **MusPR-GenEval** benchmarks full-song generation,

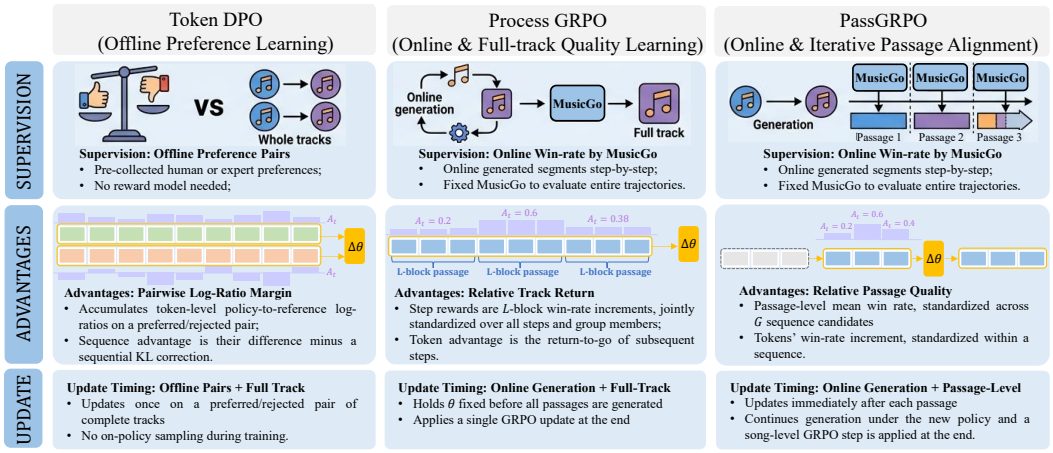

Figure 2: Comparison of different token-level policy updating methodologies.

Table 1: Offline reward-guided sampling on MusPR-GenEval with frozen B0. Δ and SE wins are reported relative to B0. AB win/lose/tie percentages (%) denotes comparison against B0 on the MusPR-GenEval evaluation subset.

Method	SE	Δ vs. B0	SE Win vs. B0	AB win/lose/tie
B0	3.882	—	—	—
Best-of-4 (w_T)	**3.903**	**+0.021**	**56.0%**	24.14/54.74/21.12
SBR-mean (Stage-I MusicGo)	3.827	−0.056	46.0%	**66.67**/25.00/8.33
SBR-mean (Stage-II MusicGo)	3.872	−0.011	48.0%	**55.74**/24.68/19.57

and **MusPR-RewriteEval** provides a hard evaluation set for local audio editing. Unless specified, all methods are compared against the baseline generator B0 (Section 4), using offline inference-time steering, online RL, or continue-SFT. Evaluation relies on both automatic metrics (e.g., SongEval, Table 10) and blinded subjective A/B listening tests—our primary metric to faithfully capture human auditory preference.

6.1 Offline Reward-Guided Sampling

We first assess whether test-time reward steering can enhance quality without updating model weights, comparing sequence-level selection (Best-of-4) against fine-grained prefix rescoring (SBR-mean).

As shown in Table 1, inference-time steering yields minimal shifts in automatic SongEval scores, yet uncovers a key post-training insight: sequence-level reward steering degrades human auditory experience, whereas fine-grained, segment-level rescoring effectively drives subjective alignment. While trajectory-level Best-of-4 slightly boosts SE, it loses under human AB judgment. Conversely, segment-level prefix rescoring (SBR-mean) substantially improves human win rates over B0, demonstrating that reward-guided inference must operate at local temporal granularities rather than full sequences. Pairwise comparisons against online policies are compiled in Table 3.

6.2 Online Policy Optimization

We evaluate parameter-updating online policies across token-level preference learning (Token DPO), sequence DPO, scalar outcome RL (Sample GRPO), and process supervision models (PassGRPO and Process GRPO).

Results in Tables 2 and 3 reveal three fundamental implications for music RL post-training:

- **TDPO Underperforms in the Tested Configuration:** Token DPO severely collapses generation quality, confirming that unguided token-level preference loss disrupts dense acoustic dependencies.
- **Process Supervision Outperforms Outcome Reward:** Process-supervised objectives (PassGRPO and Process GRPO) systematically outperform scalar outcome RL (Sample GRPO) and

Table 2: Online policy optimization on MusPR-GenEval under unsteered decoding. Δ denotes SE gain and AB win/lose/tie (%) denotes the comparison against B0.

Method	SE	Δ	SE Win (%)	AB win/lose/tie (%)
Token DPO	3.621	-0.261	36.0	–
Sample DPO	3.848	-0.035	54.0	**51.89**/41.08/7.03
Sample GRPO (w_T, Stage-I)	3.868	-0.014	38.0	**59.78**/21.79/18.44
Process GRPO (L=200, Stage-I)	3.902	$+0.020$	48.0	–
Process GRPO (L=400, Stage-I)	**3.910**	$\mathbf{+0.028}$	50.0	**63.21**/32.12/4.66
Process GRPO (L=600, Stage-I)	3.786	-0.096	32.0	–
PassGRPO (L=200, Stage-I)	3.906	$+0.024$	50.0	–
PassGRPO (L=400, Stage-I)	3.910	$+0.027$	48.0	**59.66**/16.31/24.03
PassGRPO (L=600, Stage-I)	3.867	-0.015	58.0	–
Process GRPO (L=200, Stage-II)	3.873	-0.009	42.0	–
Process GRPO (L=400, Stage-II)	3.884	$+0.002$	54.0	**46.85**/39.76/13.39
Process GRPO (L=600, Stage-II)	**3.927**	$\mathbf{+0.045}$	52.0	–
PassGRPO (L=200, Stage-II)	3.911	$+0.028$	54.0	–
PassGRPO (L=400, Stage-II)	3.871	-0.012	46.0	**53.85**/36.69/9.47
PassGRPO (L=600, Stage-II)	3.861	-0.021	46.0	–

Table 3: Other pairwise AB listening tests' results on the MusPR-GenEval subset. B win/lose/tie (%) are reported. Direct comparisons against B0 are listed in Tables 1–2.

A	B	B win/lose/tie (%)
Process GRPO (L=400, Stage-II)	PassGRPO (L=400, Stage-II)	**66.48**/26.70/6.82
Sample GRPO (w_T, Stage-I)	PassGRPO (L=400, Stage-II)	**58.33**/25.00/16.67
Sample GRPO (w_T, Stage-I)	Process GRPO (L=400, Stage-II)	41.67/41.67/16.67
PassGRPO (Stage-I)	PassGRPO (Stage-II)	**52.74**/39.04/8.22
Process GRPO (Stage-I)	Process GRPO (Stage-II)	**47.62**/42.86/9.52
SBR-mean (Stage-II)	PassGRPO (Stage-II)	**43.53**/32.94/23.53
SBR-mean (Stage-II)	Process GRPO (Stage-II)	**67.26**/30.95/1.79
Best-of-4	PassGRPO (Stage-II)	**81.60**/9.60/8.80
Best-of-4	Process GRPO (Stage-II)	**74.38**/17.36/8.26
Token DPO	Sample DPO	**58.33**/33.33/8.33
Token DPO	Process GRPO (Stage-II)	**50.00**/33.33/16.67

offline DPO in human preference. Segment-based credit assignment is essential for guiding autoregressive music decoding.

- **Reward Quality Bounds Policy Performance:** Policy alignment is strictly bounded by reward model capacity. Higher-capacity reward signals (Stage-II) yield policies that decisively outperform Stage-I variants, and parameter-updated online policies consistently dominate offline test-time steering.

6.3 Individual Preference Alignment

We examine personalized preference optimization by training annotator-specific reward models and conducting personalized SBR-mean decoding. On the Stage II holdout set, the pooled MusicGo model achieves 65% accuracy (n=64) in matching binary human labels (highlight vs. degraded) via terminal win rate $w_T \geq 0.5$. Crucially, personalized checkpoints elevate this accuracy to 80% on individual holdouts (Table 4), confirming both their capacity to capture idiosyncratic musical taste and that the pooled gap stems primarily from subjective taste variance rather than modeling limits.

However, global automated metrics (SE) remain insensitive to personalized alignment, showing negligible variance across individuals. In subjective listening tests, personalized reward steering consistently achieves superior individual alignment over both B0 and pooled models, demonstrating that music post-training must account for individual user preference rather than relying exclusively on consensus-based reward models.

Table 4: Performance of personalized Stage-II MusicGo paired with SBR-mean on Gen-50. Acc. denotes last-WR accuracy on individual MusicGo-S2-Eval holdouts. AB metrics reflect that specific annotator's win / lose / tie (%) against B0 and against the pooled Stage-II baseline.

Method	Acc.	SE	Δ	SE-win	AB vs. B0	AB vs. Pooled
Pooled	65%	3.872	-0.011	48.0	55.74/24.68/19.57	—
Ann. A	80%	3.891	$+0.008$	48.0	33.33/41.67/25.00	50.00/50.00/0.00
Ann. B	80%	**3.994**	$\mathbf{+0.111}$	**64.0**	50.00/41.67/8.33	41.67/41.67/16.67
Ann. C	80%	3.865	-0.017	52.0	75.00/8.33/16.67	50.00/25.00/25.00

6.4 Segment-Level Reward-Guided Rewriting

To fix local audio flaws without resampling entire tracks, we edit targeted spans of frozen B0 songs. On MusPR-RewriteEval, frozen MusicGo identifies the lowest-reward window, where B0-initialized policies draw $K=4$ candidates (Table 5):

- **Continue-SFT** completes left-to-right from the window start without re-attaching the suffix; candidates are ranked by Last WR (w_T).
- **Cloze-SFT** performs Fill-In-the-Middle (FIM) restoration over a central hole (25–100 tokens) given frozen context; candidates are ranked by Join WR (w_{t_1}) at the post-edit boundary.

Results in Table 5 highlight three primary findings across rewriting protocols:

- **Unconstrained Continuation:** Continue-SFT substantially boosts Last WR and achieving a better MusicGo score for a majority of samples, though regenerating downstream audio induces a slight decline in overall SongEval.
- **Context-Constrained Infilling:** Cloze-SFT significantly improves boundary transition quality (Join WR) while preserving global SongEval, confirming that bidirectional context constraints effectively mend local splices without disturbing overall song structure.
- **Infilling Variants:** Under token-splice evaluation, Prefix infill achieves superior boundary alignment and prompt win rates compared to standard FIM, with negligible impact on global audio quality.

Table 5: Segment-level rewriting on MusPR-RewriteEval ($K=4$). Last WR (w_T) and Join WR (w_{t_1}) measure terminal and edit-junction win rates; Win (%) uses Last WR for Continue and Join WR for Cloze/infill tasks. Top block: pooled Stage-II lastmos MusicGo. Bottom block: token-splice protocol under Stage-II MusicGo (Original baselines are non-comparable across blocks).

Method	Last WR	Δ_{last}	Join WR	Δ_{join}	Win	SE (Δ)
Original (lastmos)	0.596	—	0.568	—	—	3.642
Continue-SFT	**0.680**	$\mathbf{+0.083}$	0.586	$+0.005$	72.0%	3.606 (-0.036)
Cloze-SFT	0.593	-0.003	**0.630**	$\mathbf{+0.062}$	80.0%	3.641 (-0.000)
Original (Stage-II MusicGo)	**0.536**	—	0.484	—	—	3.668
Prefix infill (Tier-1)	0.544	$+0.008$	**0.585**	$\mathbf{+0.101}$	**86.0**%	3.660 (-0.008)
FIM + join-WR select	0.497	-0.039	0.556	$+0.071$	72.0%	3.648 (-0.001)

7 Conclusion

In this paper, we pioneer process reward modeling for LLM-based music generation. To overcome process data scarcity, we propose MusicGo, a trajectory win-rate reward model trained via two-stage temporal difference learning that significantly outperforms strong baselines. Leveraging this temporal signal, we introduce PassGRPO, a passage-level policy optimization algorithm that achieves marked gains over token-DPO and Process GRPO, particularly on low-scoring baseline outputs and sparse preference data. Finally, we demonstrate that process supervision effectively guides targeted passage revision and continuation, laying a foundation for fine-grained interactive co-creation.

AI use statement

In this work, we used generative AI tools for writing assistance and code debugging. Specifically, we used generative AI tools (specifically LLMs) for grammar polishing and rephrasing selected text

sections to enhance readability. We have reviewed all AI-assisted work. All LLM-suggested edits and code snippets were manually verified and tested for correctness and technical precision by the authors. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics. Our research focuses on algorithm design and model architecture evaluation in music generation. The training dataset of the music generation model used in our experiments is custom and will not be released.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our findings, we provide complete implementation details throughout the paper and supplementary material. The detailed network architecture, hyperparameters, and training objectives are specified in methodology-related sections and Appendix C. All models and evaluation benchmarks used in this paper will be publicly accessible, with preprocessing and filtering protocols fully documented in Section 6.

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.

Wei Duan, Yi Yu, and Jean-Pierre Briot. Recent advances in music generation: methods, evaluation, and challenges. *Artificial Intelligence Review*, 59, 06 2026. doi: 10.1007/s10462-026-11582-x.

Zach Evans, CJ Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. *arXiv preprint arXiv:2402.04825*, 2024.

Xueji Fang, Liyuan Ma, Zhiyang Chen, Mingyuan Zhou, and Guo jun Qi. Inflvg: Reinforce inference-time consistent long video generation with grpo, 2025. URL https://arxiv.org/abs/2505.17574.

Minghe Gao, Xuqi Liu, Zhongqi Yue, Yang Wu, Shuang Chen, Juncheng Li, Siliang Tang, Fei Wu, Tat-Seng Chua, and Yueting Zhuang. Benchmarking multimodal cot reward model stepwise by visual program, 2025. URL https://arxiv.org/abs/2504.06606.

Junmin Gong, Yulin Song, Wenxiao Zhao, Sen Wang, Shengyuan Xu, Jing Guo, and Xuerui Yang. Ace-step 1.5: Pushing the boundaries of open-source music generation. *arXiv preprint arXiv:2602.00744*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid

El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy,

Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Rui Huang, Haoquan Zhang, Manyuan Zhang, Jiaming Liu, Shanghang Zhang, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step, 2025. URL https://arxiv.org/abs/2501.13926.

Changhao Jiang, Jiahao Chen, Zhenghao Xiang, Zhixiong Yang, Hanchen Wang, Jiabao Zhuang, Xinmeng Che, Jiajun Sun, Hui Li, Yifei Cao, et al. Muse: Towards reproducible long-form song generation with fine-grained style control. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 22492–22512, 2026.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms, 2019. URL https://arxiv.org/abs/1812.08466.

Yonghyun Kim, Junwon Lee, Haiwen Xia, Yinghao Ma, Junghyun Koo, Koichi Saito, Yuki Mitsufuji, and Chris Donahue. Tunejury: An open metric for improving music generation preference alignment, 2026. URL https://arxiv.org/abs/2606.17006.

Shun Lei, Yaoxun Xu, Zhiwei Lin, Huaicheng Zhang, Wei Tan, Hangting Chen, Jianwei Yu, Yixuan Zhang, Chenyu Yang, Haina Zhu, Shuai Wang, Zhiyong Wu, and Dong Yu. Levo: High-quality song generation with multi-preference alignment, 2025. URL https://arxiv.org/abs/2506.07520.

Qingyao Li, Xinyi Dai, Xiangyang Li, Weinan Zhang, Yasheng Wang, Ruiming Tang, and Yong Yu. CodePRM: Execution feedback-enhanced process reward model for code generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8169–8182, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.428. URL https://aclanthology.org/2025.findings-acl.428/.

Wang Lin, Liyu Jia, Wentao Hu, Kaihang Pan, Zhongqi Yue, Wei Zhao, Jingyuan Chen, Fei Wu, and Hanwang Zhang. Reasoning physical video generation with diffusion timestep tokens via reinforcement learning, 2025a. URL https://arxiv.org/abs/2504.15932.

Yuanyuan Lin, Xiangyu Ouyang, Teng Zhang, and Kaixin Sui. Rpm-mcts: Knowledge-retrieval as process reward model with monte carlo tree search for code generation, 2025b. URL https://arxiv.org/abs/2511.19895.

Cheng Liu, Hui Wang, Jinghua Zhao, Shiwan Zhao, Hui Bu, Xin Xu, Jiaming Zhou, Haoqin Sun, and Yong Qin. Musiceval: A generative music dataset with expert ratings for automatic text-to-music evaluation, 2025. URL https://arxiv.org/abs/2501.10811.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, January 2016. ISSN 1476-4687. doi: 10.1038/nature16961. URL https://www.nature.com/articles/nature16961.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, Jeff Wang, Ivan Cruz, Bapi Akula, Akinniyi Akinyemi, Brian Ellis, Rashel Moritz, Yael Yungster, Alice Rakotoarison, Liang Tan, Chris Summers, Carleigh Wood, Joshua Lane, Mary Williamson, and Wei-Ning Hsu. Audiobox: Unified audio generation with natural language prompts, 2023. URL https://arxiv.org/abs/2312.15821.

Lei Wang, Ziyi Zhao, Hanwei Liu, Junwei Pang, Yi Qin, and Qidi Wu. A review of intelligent music generation systems, 2023. URL https://arxiv.org/abs/2211.09124.

Dapeng Wu, Shun Lei, Wei Tan, Guangzheng Li, Yunzhe Wang, Huaicheng Zhang, Lishi Zuo, and Zhiyong Wu. Songbench: A fine-grained multi-aspect benchmark for song quality assessment, 2026. URL https://arxiv.org/abs/2604.25937.

Jie Wu, Yu Gao, Zilyu Ye, Ming Li, Liang Li, Hanzhong Guo, Jie Liu, Zeyue Xue, Xiaoxia Hou, Wei Liu, et al. Rewarddance: Reward scaling in visual generation. *arXiv preprint arXiv:2509.08826*, 2025.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Dongchao Yang, Yuxin Xie, Yuguo Yin, Zheyu Wang, Xiaoyu Yi, Gongxi Zhu, Xiaolong Weng, Zihan Xiong, Yingzhe Ma, Dading Cong, et al. Heartmula: A family of open sourced music foundation models. *arXiv preprint arXiv:2601.10547*, 2026.

Jixun Yao, Guobin Ma, Huixin Xue, Huakang Chen, Chunbo Hao, Yuepeng Jiang, Haohe Liu, Ruibin Yuan, Jin Xu, Wei Xue, Hao Liu, and Lei Xie. Songeval: A benchmark dataset for song aesthetics evaluation, 2025a. URL https://arxiv.org/abs/2505.10793.

Jixun Yao, Guobin Ma, Huixin Xue, Huakang Chen, Chunbo Hao, Yuepeng Jiang, Haohe Liu, Ruibin Yuan, Jin Xu, Wei Xue, et al. Songeval: A benchmark dataset for song aesthetics evaluation. *arXiv preprint arXiv:2505.10793*, 2025b.

Ruibin Yuan, Hanfeng Lin, Shuyue Guo, Ge Zhang, Jiahao Pan, Yongyi Zang, Haohe Liu, Yiming Liang, Wenye Ma, Xingjian Du, Xinrun Du, Zhen Ye, Tianyu Zheng, Zhengxuan Jiang, Yinghao Ma, Minghao Liu, Zeyue Tian, Ziya Zhou, Liumeng Xue, Xingwei Qu, Yizhi Li, Shangda Wu, Tianhao Shen, Ziyang Ma, Jun Zhan, Chunhui Wang, Yatian Wang, Xiaowei Chi, Xinyue Zhang, Zhenzhu Yang, Xiangzhou Wang, Shansong Liu, Lingrui Mei, Peng Li, Junjie Wang, Jianwei Yu, Guojian Pang, Xu Li, Zihao Wang, Xiaohuan Zhou, Lijun Yu, Emmanouil Benetos, Yong Chen, Chenghua Lin, Xie Chen, Gus Xia, Zhaoxiang Zhang, Chao Zhang, Wenhu Chen, Xinyu Zhou, Xipeng Qiu, Roger Dannenberg, Jiaheng Liu, Jian Yang, Wenhao Huang, Wei Xue, Xu Tan, and Yike Guo. Yue: Scaling open foundation models for long-form music generation, 2025. URL https://arxiv.org/abs/2503.08638.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. *arXiv preprint arXiv:2404.11999*, 2024.

Ruiyi Zhang, Peijia Qin, Qi Cao, Eric Xue, and Pengtao Xie. Funprm: Function-as-step process reward model with meta reward correction for code generation, 2026. URL https://arxiv.org/abs/2601.22249.

Congmin Zheng, Jiachen Zhu, Zhuoying Ou, Yuxiang Chen, Kangning Zhang, Rong Shan, Zeyu Zheng, Mengyue Yang, Jianghao Lin, Yong Yu, and Weinan Zhang. A survey of process reward models: From outcome signals to process supervisions for large language models, 2026. URL https://arxiv.org/abs/2510.08049.

APPENDIX

A MUSIC GENERATION MODEL

This section details the architecture and pretraining recipe for B0, the foundational music generation checkpoint serving as the policy model throughout this work (Section 4). The proposed MusicGo framework shares the same backbone architecture and is initialized directly from B0.

A.1 ARCHITECTURE AND TOKENIZATION

B0 is a 1B-parameter decoder-only Transformer belonging to the Llama family Touvron et al. (2023); Grattafiori et al. (2024). It adopts a frozen tokenizer and detokenizer inherited from LeVo Lei et al. (2025). Table 6 summarizes its detailed structural configuration. The input embedding layer handles a 200,000-token text vocabulary from Qwen2 (Yang et al., 2024). In contrast, the language modeling head predicts exclusively over audio tokens (32,775 logits); consequently, lyric tokens are never treated as reconstruction targets during loss calculation.

Audio signals are represented via a single discrete codebook of size 32,768 operating at a sampling rate of 25 Hz ($g = 25$ tokens per second). Special audio identifiers occupy the remaining output logit slots, including $\langle\text{BOS}\rangle$, $\langle\text{EOS}\rangle$, a text-mask token used for classifier-free guidance (CFG), and two fill-in-the-middle (FIM) boundary tokens reserved for the Cloze-SFT setting (Appendix B). Text token IDs are shifted by a fixed scalar offset to avoid index collisions with the audio codebook. The unified sequence layout during training is formatted as:

$$z = \big(c_{\text{tag}},\, c_{\text{lyr}},\, \langle\text{BOS}\rangle,\, x,\, \langle\text{EOS}\rangle\big), \tag{14}$$

where c_{tag} denotes a comma-separated style control string (`[tags:...]`), and c_{lyr} represents lyric sentences interspersed with structural tags (e.g., intro, verse, chorus) reformatted into a semicolon-delimited sequence. Rotary Position Embeddings (RoPE) are configured with a base frequency $\theta = 10{,}000$ over a maximum context length of 10,000 tokens.

Table 6: Architectural configuration of the pretrained music generation base model, B0.

Component	Value
Parameters	1B
Hidden size / FFN intermediate size	1536 / 6144
Number of layers	18
Attention / KV heads	16 / 8 (GQA)
Activation / Normalization	SiLU / RMSNorm ($\epsilon = 10^{-6}$)
Maximum context length	10,000
Text vocabulary size	200,000
Audio codebook size / LM head dimension	32,768 / 32,775
Codec frame rate	25 Hz
Precision / Attention implementation	`bf16` / FlashAttention-2

A.2 Pretraining Objective

B0 is pretrained using standard teacher-forced next-token cross-entropy loss applied solely to audio sequence positions. Specifically, the loss mask m_t is set to 1 for codec tokens and the $\langle\mathrm{EOS}\rangle$ token, and 0 for the text prompt prefix and $\langle\mathrm{BOS}\rangle$. Positions corresponding to non-audio tokens are assigned the ignore index in the cross-entropy computation. Optimization is carried out using `bf16` mixed precision, gradient checkpointing, the AdamW optimizer, cosine learning rate decay, and DeepSpeed ZeRO-2.

To enable lyric-conditioned classifier-free guidance at inference time, conditioning inputs are randomly dropped during pretraining: with a probability of 0.1, c_{tag} is replaced by an empty tag string (`[tags:]`); independently, lyric token IDs are replaced with the text-mask token with 0.1 probability; with another 0.1 probability, the entire lyric span is fully masked out. When training on tracks with synchronized timestamps, a contiguous lyric segment is sampled with high probability, and the corresponding audio slice (computed via start/end timestamps $\times$ 25 Hz) is extracted as the target sequence. This strategy ensures that the model learns both full-song synthesis and section-level continuation.

A.3 Decoding Protocol

Unless stated otherwise, decoding of music generation model is conducted with a temperature of 0.8 and a lyric CFG scale of 1.3. During unconditional forward passes for CFG, the text-mask ID is substituted for all lyric tokens while preserving $\langle\mathrm{BOS}\rangle$ and the audio prompt prefix. A repetition penalty of 1.1 is applied specifically during rewriting experiments. All online policy search algorithms described in the main text are initialized from this frozen B0 checkpoint; the reference policy π_{ref} is kept continuously frozen as an identical copy.

B Music Rewriting Task

Because full-song generation under online alignment exhibits minimal margin for improvement under standard evaluation metrics (Section 6.4), we design additional fine-tuning policies dedicated to local segment editing. Rather than regenerating the entire audio track, these policies edit lower-quality segments extracted from songs synthesized by the frozen B0 model. MusPR-RewriteEval identifies target spans of 400 tokens (≈ 16 s) buffered by 200-token boundary margins, prioritized by the lowest mean score w_t under pooled Stage-II MusicGo. We fine-tune two supervised policies initialized from B0: Continue-SFT and Cloze-SFT, both trained on the curated clean human-recorded subset of MusPR.

B.1 Training Data and Protocol

We filter the dataset to retain Stage-I tracks whose identifiers originate exclusively from custom human-composed and synthetic data. The resulting dataset is partitioned into a $98/2$ train/validation split using fixed random seed 3407, yielding $14{,}700$ training and 300 validation songs balanced across the three data sources. The maximum sequence context is set to $16{,}384$. Both rewriting policies are trained across four GPUs using Distributed Data Parallel (DDP) with a per-device batch size of 1, optimized via AdamW at a learning rate of 2×10^{-5} and gradient norm clipping of 1.0. The checkpoint achieving the lowest validation negative log-likelihood (NLL) is selected for evaluation.

B.2 CONTINUE-SFT

Continue-SFT models left-to-right audio completion starting from a random truncation point. For an audio codec sequence x of total length T, a split index s_0 is drawn uniformly at random from $\{50, \ldots, T-100\}$, guaranteeing at least 2 s of prefix context and 4 s of target continuation. The input format matches that of B0 pretraining, but the cross-entropy loss mask is active only for sequence indices $x_{s_0:T}$ and $\langle\text{EOS}\rangle$. A new truncation point s_0 is dynamically re-sampled for each track at every epoch. Training runs for 5 epochs (18,375 optimization steps), yielding a final validation NLL of 3.80.

At test time, the model is conditioned on the prefix generated by frozen B0 up to the cut location s_0 identified in MusPR-RewriteEval, and proceeds to generate through the remainder of the original duration; the pre-existing suffix is *not* concatenated back. $K{=}4$ candidate completions are sampled at temperature 0.8 with lyric CFG 1.3, ranked according to the terminal win rate w_T provided by the pooled Stage-II lastmos MusicGo evaluator.

B.3 CLOZE-SFT

Cloze-SFT is designed as an auxiliary Fill-In-the-Middle (FIM) infilling model. During training, up to eight local spans (holes) are randomly sampled per track. The length of each hole is drawn uniformly from $\{25, \ldots, 100\}$ tokens ($\approx$ 1–4 s), constrained by a 50-token margin at both sequence boundaries and a minimum 25-token separation between adjacent holes. Representing the sequence as $x = \text{prefix} \parallel \text{middle} \parallel \text{suffix}$, the causal autoregressive ordering is structured as:

$$\big(c,\ \langle\text{BOS}\rangle,\ \text{prefix},\ \langle\text{FIM-S}\rangle,\ \text{suffix},\ \langle\text{FIM-M}\rangle,\ \text{middle},\ \langle\text{EOS}\rangle\big). \tag{15}$$

Cross-entropy loss is computed exclusively on the `middle` target tokens. Since eight holes are packaged within each sequence, one training epoch effectively yields eight times as many optimization updates as Continue-SFT; model training is conducted for 3 epochs (88,200 total steps), achieving a validation NLL of 4.42.

During evaluation, a single hole spanning 25–100 tokens is introduced at the center of the MusPR-RewriteEval target window. Both the prefix and suffix retain the original audio tokens generated by B0. The lyric CFG scale is set to 1.0 (guidance is omitted as the lyric context is implicitly constrained by the observed suffix). $K{=}4$ infill candidates are generated and ranked by their join win rate w_{t_1}, defined as the MusicGo score evaluated at the immediate boundary token following the restored hole. Note that join win rate is uniquely applicable to the cloze formulation where the trailing suffix remains fixed. At a decoding temperature of 0.8, the exact token-level match rate between the newly generated middle segment and the ground-truth original span is effectively 0.

C MUSICGO AND PASSGRPO TRAINING DETAILS

C.1 STAGE I HYPERPARAMTERS

The default hyperparameter configurations used in Stage I training are summarized in Table 7.

C.2 STAGE II FINE-GRAINED DATA CONSTRUCTION

Stage II collects timestamped human process labels on full songs via a browser interface, converting them into continuous envelopes (s, h) and terminal targets (V_T) to train MusicGo.

C.2.1 INTERFACE AND SESSION FLOW

Annotators complete a 5-step pipeline: (i) latency calibration, (ii) instructions with audio examples, (iii) a practice track (discarded), (iv) a background questionnaire, and (v) formal annotation in 50-song batches. The multilingual desktop/touch UI displays lyrics and style tags throughout listening. Users trigger `H` for highlights, `F` for degradations, and Space for latency probes.

Latency calibration. Listeners undergo a synthetic probe test followed by two songs with embedded probes to establish their median reaction offset ($\text{latency_offset_sec}$). Raw logs store

Table 7: Hyperparameter settings for Stage I supervision.

Category	Hyperparameter	Value
Segment	Frame step size g	25 (1 s)
Loss Weights	Alignment weight λ_{align}	7.0
	Soft ceiling weight λ_{soft}	0.55
	Pairwise weight λ_{pair}	4.0
	Classification weight λ_{CE}	0.15
	Shape regularization weight λ_{shape}	3.2
	TD residual weight λ_{TD}	0.05
	Value head weight λ_{sv}	0.3
Loss Parameters	Margin m ($\mathcal{L}_{\text{pair}}$)	0.12
	Ceiling drop rate α ($\mathcal{L}_{\text{soft}}$)	0.22
	Numerical stability ϵ	10^{-8}
	Discount factor γ ($\mathcal{L}_{\text{TD}}$)	0.95
	Oscillation threshold δ ($\mathcal{L}_{\text{shape}}$)	0.02

unshifted player time, while cluster and token-aligned envelopes use latency-corrected times $t - \text{latency_offset_sec}$.

C.2.2 Highlights, degradations, and clustering

Listeners mark salient musical moments without being required to annotate every bar:

- **Highlight (`H`):** Auditorily pleasing or memorable passages (e.g., catchy hooks, strong vocals).
- **Degradation (`F`):** Technical flaws or unnatural audio artifacts (e.g., vocal distortion, drum drift, clipping, pitch instability). Stylistic distortions are explicitly excluded.

Repeated keypresses encode intensity, while unmarked regions remain unannotated. Same-polarity events are clustered using a 3 s gap threshold and assigned at least one aspect tag (Tables 8 and 9).

Table 8: Highlight cluster tags.

Id	Meaning
`melody_hook`	Catchy / memorable melody
`harmony_rich`	Strong harmony
`rhythm_groove`	Groove / rhythm
`vocal_expressive`	Nuanced or skilled vocal
`timbre_contrast`	Instrument timbre that clicks
`arrangement_clever`	Clever arrangement / layers
`emotion_touching`	Emotionally spot-on
`lyrics_strong`	Strong lyrics
`other_highlight`	Other (use sparingly)

Table 9: Degradation cluster tags.

Id	Meaning
`vocal_artifact`	Distorted, muddy, cracked, or synthetic vocal
`timing_drift`	Drum drift / dropped beat / groove collapse
`clip_mud`	Harsh, clipped, or muddy instruments
`timbre_inconsistent`	Unintended vocal-timbre jump
`pitch_unstable`	Pitch off or drifting
`rhythm_unstable`	Unstable rhythm
`arrangement_abrupt`	Abrupt arrangement cut
`mix_distortion`	Mix distortion / poor audio quality
`other_degrade`	Other (use sparingly)

C.2.3 Track-level ratings and validity

Post-listening, annotators move two independent bipolar sliders in $[-1, 1]$ accompanied by mandatory song-level reason tags:

- **Personal preference (q^{pref}):** Measures general like/dislike.
- **Work quality (q):** Measures production integrity versus degradation; scaled directly to V_T.

Validity & Token Painting: Records are valid if both sliders/tags are complete, all clusters are tagged, and audio coverage reaches $\geq 80\%$. The finalized pool yields 1,270 complete annotations (1,281 unique tracks) from 16 listeners (mean coverage $\approx 90\%$). Corrected cluster times are painted

onto the 25 Hz codec sequence using a Gaussian kernel ($\sigma = 0.4\,\mathrm{s}$) to form s_t and h_t. A per-annotator holdout split ($n = 60$, seed 42) is reserved for RM-S2 evaluation.

C.3 PassGRPO Algorithm Explanation

Algorithm 1 PassGRPO Algorithm

Require: Prompt c; Policy π_θ; MusicGo $r_{\phi^\star_{\mathrm{II}}}$; Passage length L; Group size G

$p \leftarrow c$ ▷ Initialize prefix with text prompt and audio BOS
while p has not terminated **do**
  Sample completions $u^{(1)}, \ldots, u^{(G)} \sim \pi_\theta(\cdot \mid p)$ of length L
  Evaluate token win rates $\{w_t\}$ on $x^{(i)} = p \| u^{(i)}$ via $r_{\phi^\star_{\mathrm{II}}}$
  Calculate passage score R_i via Eq. equation 10
  Form passage advantage $\hat{A}_{i,t} \leftarrow \frac{1}{2}\tilde{R}_i + \frac{1}{2}\tilde{\delta}_t^{(i)}$ on $\mathcal{U}_i$ using fine grid $g = 25$
  Update policy $\theta \leftarrow \theta + \eta \nabla_\theta \mathcal{J}_{\mathrm{GRPO}}(\theta)$ restricted to $\mathcal{U}_i$ ($\beta_{\mathrm{in}} = 0.08$)
  Sample continuation index $i^\star \sim \pi_{\mathrm{cont}}$ via Eq. equation 12
  Append chosen continuation: $p \leftarrow p \| u^{(i^\star)}$
end while
Reuse final sequence group $\{p \| u^{(i)}\}$ for global alignment
Compute track score R_i^{song} via Eq. equation 13
Form global advantage $\hat{A}_{i,t}$ across all audio tokens using coarse grid $g = L$
Update policy $\theta \leftarrow \theta + \eta \nabla_\theta \mathcal{J}_{\mathrm{GRPO}}(\theta)$ ($\beta = 0.02$)

D Objective Evaluation Metrics

Table 10: Objective Metrics

Metric	Definition
Last-WR accuracy	Fraction of RM-S2 tracks where $w_T \geq 1/2$ iff the annotator marked the track intact.
MOS Pearson/Spearman/MAE	Agreement between the model last win-rate output and the annotator's global MOS (mapped from the $[-1, 1]$ slider).
Q-tag P/R/F1	Multi-label match of predicted song-level quality-reason tags to the annotator's multi-hot, at a fixed probability threshold.
SongEval (SE)	Average score of five-dim outputs on the waveform.
Join WR $w_{\mathrm{join}} = w_{t_1}$	Causal MusicGo win rate at the first token after the rewritten span.
Last WR w_T	Terminal win rate of the candidate selected by w_{join} (K=4).
$\mathrm{Win}_{\mathrm{join}}$	Fraction of tracks with $w_{\mathrm{join}}(m) > w_{\mathrm{join}}(\mathrm{original})$.