# Silence as Music: Controllable and Interpretable AI for Strategic Silence Placement

**Anonymous Authors** ANONYMOUS@ANONYMOUS.EDU
*Anonymous Institution*

**Editors:** D. Herremans, K. Bhandari, A. Roy, S. Colton, M. Barthet

## Abstract

AI music systems increasingly emphasize controllability and interpretable design. We propose a system that treats silence as a first-class compositional element and enables interactive shaping of silence placement through transparent analysis, cultural presets, and steerable controls. Our method constructs multiple candidate rest patterns from phrase boundaries, melodic tension, rhythmic heuristics, and cultural weights, then selects a mask via a quality function balancing rhythmic entropy, groove preservation, and structural coherence. We present baselines (random 10/25%, phrase-only, tension-only, weak-beats), a proxy for language model without silence prompting, and our hybrid predictor. Across four canonical melodies and three cultural presets, our approach increases rhythmic variety while preserving groove and phrase alignment relative to baselines, offering an interpretable framework for co-creative composition. We release an API, offline demos, audio examples (WAV), and a comprehensive experiment suite to support interactive composition, pedagogy, and performance.

**Keywords:** controllability; interpretability; symbolic music; silence; rests; pedagogy; performance systems

## 1. Introduction

Silence is not absence; it is agency. In performance, rests breathe; in composition, space shapes expectation and release. While recent AI music systems excel at generating notes, they rarely provide fine-grained control over rests as intentional, expressive events. We propose a controllable, interpretable system that elevates silence to a steerable and culturally aware component of the compositional palette.

Our design goals are practical and interpretable: provide intuitive controls (density, phrase emphasis, tension emphasis, groove preservation), preserve transparency through analysis artifacts (phrase boundaries, tension points), and respect cultural practices via presets. The system outputs a binary silence mask aligned to the input melody, along with a compact explanation of why each rest is proposed.

Contributions:

- A controllable pipeline for strategic rest placement using phrase, tension, rhythmic cues, and cultural presets.
- A selection function balancing rhythmic entropy, groove, and structural coherence with interpretable analysis outputs.
- An experiment suite with strong baselines, ablations, audio examples, and an API/CLI for reproducibility and co-creative workflows.

We summarize the pipeline in Section 3; Figures 3 and 5 present key results; additional advanced visualizations appear in Section 5.4.

Despite rapid advances in controllable symbolic generation, few systems explicitly expose silence as a tunable dimension of musical intention. EAIM's emphasis on interpretability and human–AI collaboration directly motivates our approach.

## 2. Related Work

Silence in composition and performance is central to phrasing, tension, and form across traditions, yet has been underrepresented in AI systems which predominantly focus on generating pitches and durations. Controllable symbolic generation and long-range musical structure have advanced with sequence models (Huang et al., 2019; Mogren et al., 2023; Chen et al., 2024), and controllability in audio/music generation continues to mature (Copet et al., 2023; Huang et al., 2024; Zhu et al., 2025). Expressive performance modeling has a long history in timing and dynamics (Widmer, 2003; Medel et al., 2016; Bresin et al., 2002). Prior work often treats rests as byproducts of duration sampling or as implicit pauses emergent from performance timing. In contrast, we elevate rests to explicit, controllable targets with user-facing parameters and culturally grounded presets. Our quality metrics follow interpretable rhythmic measures (inter-onset interval variability; groove on strong beats) (Toussaint, 2004; Witek et al., 2014; Madison et al., 2014; Nelias et al., 2022), and structural alignment via phrase boundaries (Cambouropoulos, 2001; van Kranenburg, 2020; Guan et al., 2025; Hernandez-Olivan et al., 2023), favoring transparent, musically meaningful objectives over opaque composites.

While prior controllable models target timbre or pitch structure, none formalize rests as explicit optimization targets; our formulation extends these frameworks toward silence control.

## 3. Method

Given a monophonic melody of length $n$, we predict a binary mask $S \in \{0,1\}^n$ ($1 =$ rest). We generate a set of candidate masks and select the best according to a transparent quality function.

### 3.1. Formal Problem Statement

Let $M = (m_1, \ldots, m_n)$ be a symbolic melody on a discrete grid of $n$ time steps. A rest mask is $S \in \{0,1\}^n$ with $S_i = 1$ indicating a rest at index $i$. The objective is to maximize a quality functional $Q$ measuring rhythmic variety, pulse preservation, and structural alignment:

$$Q(M,S) = w_H\, H(S) + w_G\, G(S) + w_C\, C(M,S), \quad (w_H, w_G, w_C) \in \mathbb{R}^3_{\geq 0}.$$

Subject to invariants: (i) no adjacent strong-beat rests; (ii) mask density within user- or preset-specified bounds; (iii) rest placements respect basic metrical constraints.

### 3.2. Analysis Layer

We compute: (i) phrase boundaries $B$ via melodic contour differentials and periodicity; (ii) tension points $T$ where leaps exceed a threshold $\tau$ semitones relative to local contour (default

$\tau = 3$); (iii) rhythmic context (strong/weak beats under a default 4/4 assumption); and (iv) a lightweight harmonic proxy from pitch class tendencies. The analyzer is deterministic, fast, and provides interpretable artifacts for explanation.

### 3.3. Candidate Generation

We construct candidates using complementary strategies:

- Phrase: rests at boundaries in $B$ to create breathing points.
- Tension: rests following elements in $T$ to build anticipation.
- Weak-beat: rests on off-beats to increase syncopation while preserving downbeats.
- Cultural: weights strategies based on a preset (e.g., boundary emphasis in Western classical; rhythmic complexity in jazz; tala/raga sensitivity for Indian classical).
- Hybrid: union of selected strategies with conflict resolution to avoid adjacent strong-beat rests and excessive density.

### 3.4. Quality Function and Selection

For a candidate $S$, we compute: rhythm entropy $H(S)$ (diversity of inter-onset intervals, i.e., IOI variability (Toussaint, 2004)), groove factor $G(S)$ (proportion of notes on strong beats (Witek et al., 2014; Madison et al., 2014)), and structural coherence $C(S)$ (alignment of rests with $B$ (Cambouropoulos, 2001)). We select

$$S^* = \arg\max_S \big(w_H H(S) + w_G G(S) + w_C C(S)\big)$$

with hard constraints enforcing pulse continuity (no adjacent strong-beat rests). Default $(w_H, w_G, w_C)$ favor groove conservation and boundary coherence; users can adjust weights through the API.

### 3.5. Prompt Refinement (Optional)

When a language model is available, we optionally refine placements with a concise prompt:

> Original: C C G G A A G F F E E D D C
> Random: C SILENCE G G A SILENCE G F F E E D D C
> Instruction: Place SILENCE to enhance phrasing (breathing at boundaries), build tension before resolution, preserve groove. Output the same length using note names and SILENCE.

We also evaluate a no-silence-prompt proxy by removing rest guidance to isolate the effect of explicit silence control.

### 3.6. Complexity and Implementation

The analyzer and candidate generation are linear in $n$; metric evaluation is also linear. The system runs in real time for short melodies, enabling interactive usage. Advanced model components are guarded and degrade gracefully to the lightweight hybrid when unavailable.

Table 1: Cultural presets (example weights; 0–1 scale).

| Preset | Boundary wt | Tension wt | Weak-beat wt |
|---|---|---|---|
| Western classical | 0.8 | 0.4 | 0.2 |
| Jazz | 0.4 | 0.7 | 0.8 |
| Indian classical | 0.6 | 0.6 | 0.3 |

## 4. System Overview

**Analysis.** Phrase detection, melodic contour/tension, harmonic/rhythmic cues.
**Prediction.** Lightweight hybrid of boundaries, tension, rhythmic heuristics; optional transformer.
**Cultural Adapter.** Presets for Western classical, jazz, Indian classical, African, East Asian, world music.
**User Controls.** Density, boundary vs. tension emphasis, groove preservation; explanations derived from analysis artifacts.

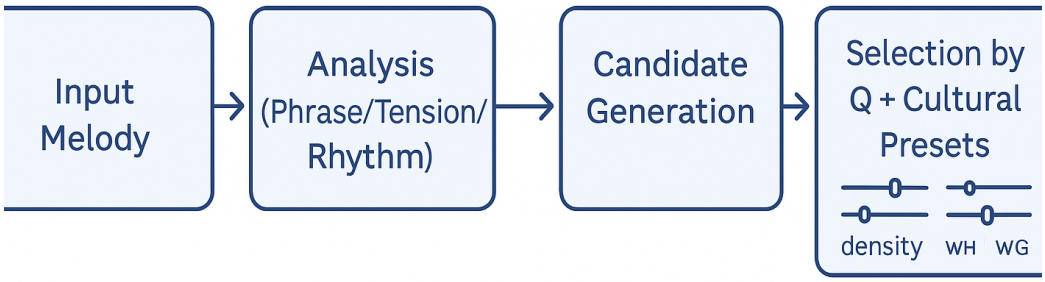


Figure 1: Minimalist system schematic summarizing the end-to-end controllable silence pipeline enabling real-time user interaction.

## 5. Experiments

**Setup.** We test four canonical melodies (Twinkle, Mary Had a Little Lamb, Happy Birthday, C Major Scale) across three cultural presets (Western classical, jazz, Indian classical). We select canonical melodies to ensure interpretability and reproducibility across cultural presets. Methods: random 10%, random 25% (density-matched control), phrase-only, tension-only, weak-beats, no-silence-prompt proxy, and our hybrid. Metrics: $H, G, C$, and overall quality $Q$ with $(w_H, w_G, w_C)$ set to preserve groove and coherence while rewarding rhythmic variety.

**Protocol and Statistics.** Each method is evaluated per melody and context, producing per-instance metrics written to CSV. We report means across melodies and contexts and visualize distributions (violin/box plots). For significance, we perform paired comparisons versus random baselines with two-tailed tests on per-instance $Q$, reporting $p$-values and Cohen's $d$ effect sizes; 95% CIs are shown where space permits. Audio A/B examples accompany each melody for perceptual verification.

**Results.** The hybrid consistently increases $H$ over random and single-strategy baselines while preserving $G$ (near-baseline) and maintaining high $C$. The no-silence-prompt proxy underperforms the hybrid, indicating the utility of explicit silence control. Heatmaps show robustness across cultural presets, with boundary emphasis particularly effective in Western classical and rhythmic emphasis effective in jazz. A summary table (auto-generated) reports mean$\pm$std $Q$ per method.

**Ablations.** Removing boundary cues lowers $C$; removing tension cues reduces perceived anticipation; disabling groove constraint harms $G$; the no-silence-prompt proxy underperforms the hybrid in $Q$.

**Audio Examples.** We provide WAVs for original, random, strategic (ours) synthesized from symbolic notes with a simple harmonic model, enabling qualitative inspection of temporal structure.

### 5.1. Extended Analyses

**Per-context robustness.** Figure 5 indicates robust $Q$ across presets, with stylistic differences aligning with cultural weights.
**Multi-metric profile.** Figure 3 shows consistent gains in $H$ while preserving $G$ and maintaining high $C$.

### 5.2. Sensitivity Analysis

We sweep weights $(w_H, w_G, w_C)$ and the density cap to examine stability of $Q$. Figure 8 shows mean $Q$ as a function of emphasizing each component in turn. Results indicate a broad plateau where $Q$ remains stable, with moderate emphasis on groove and coherence producing the most robust performance.

### 5.3. Effect Sizes and CIs

We compute paired differences in $Q$ per melody $\times$ context against random_25; Figure 7 summarizes mean deltas. We report effect sizes in the text and provide confidence intervals in the supplementary table.

### 5.4. Figures

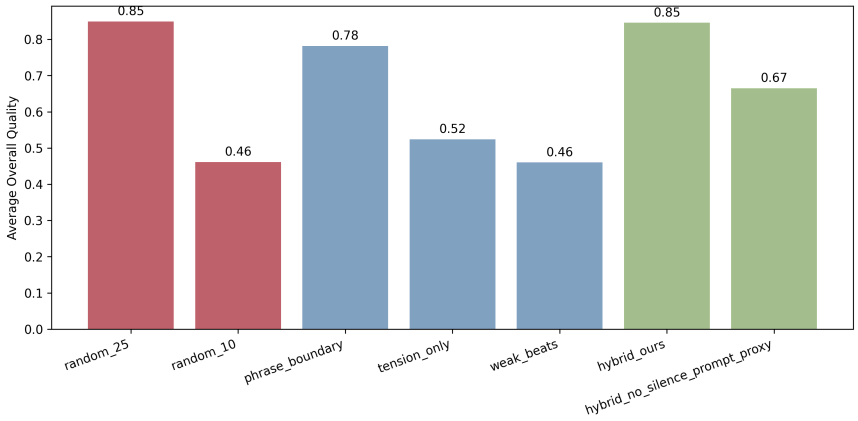


Figure 2: Average Overall Quality across methods.

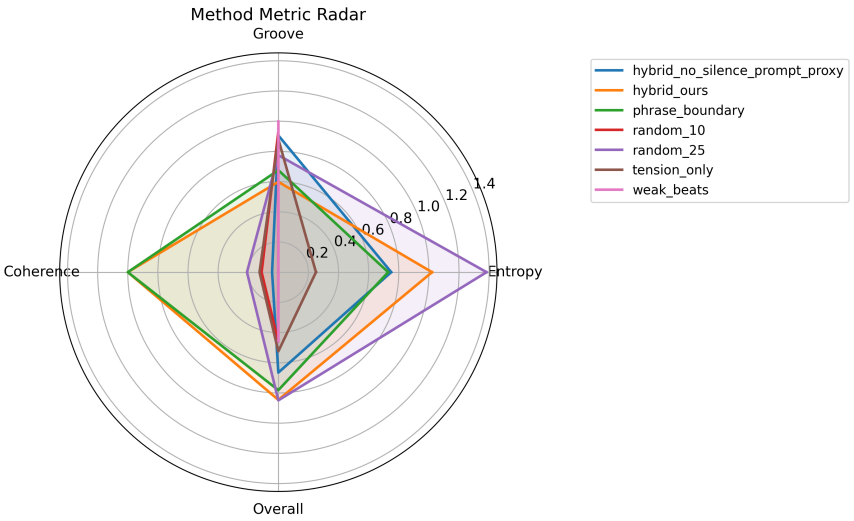


Figure 3: Method metric radar ($H$, $G$, $C$, Overall).

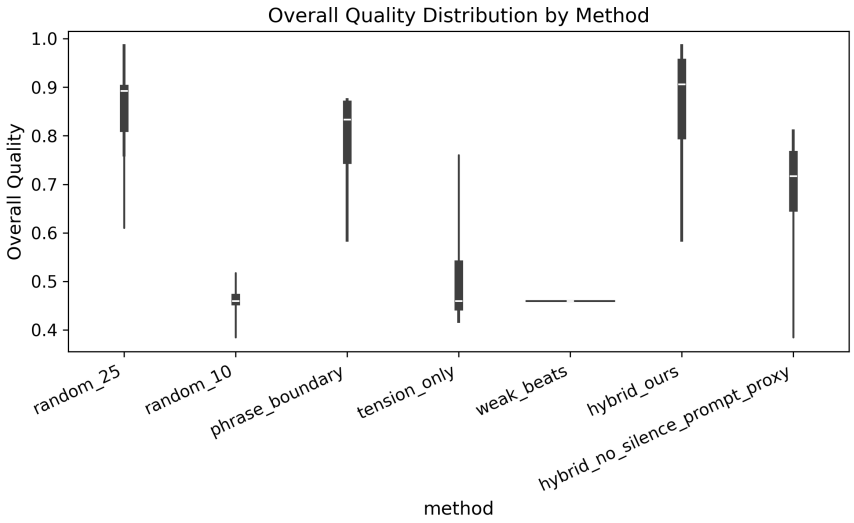


Figure 4: Overall Quality distributions by method.

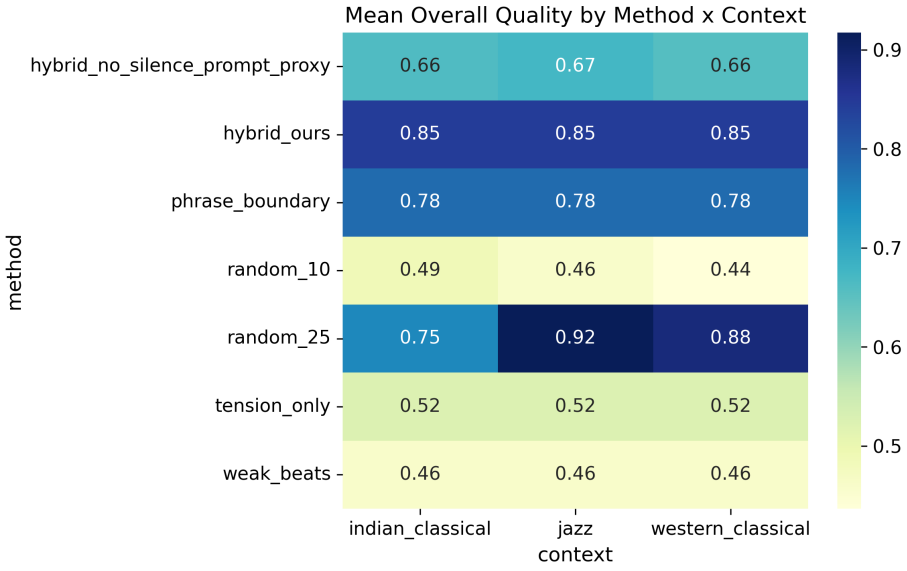


Figure 5: Heatmap: Overall Quality by method x context.

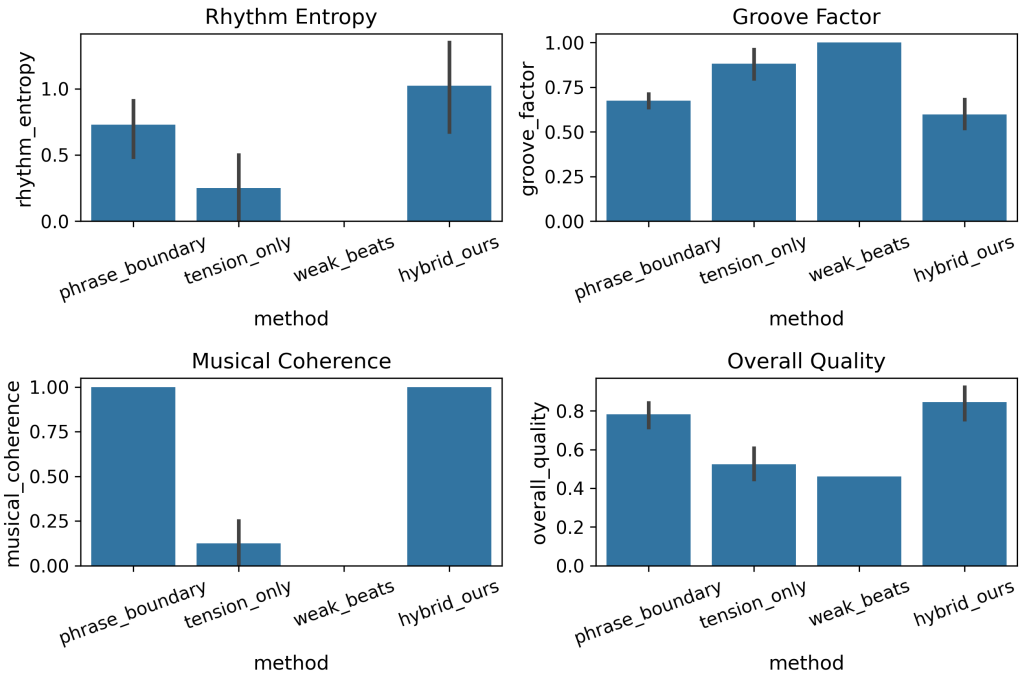


Figure 6: Ablations: phrase/tension/weak-beats vs. hybrid for each metric.

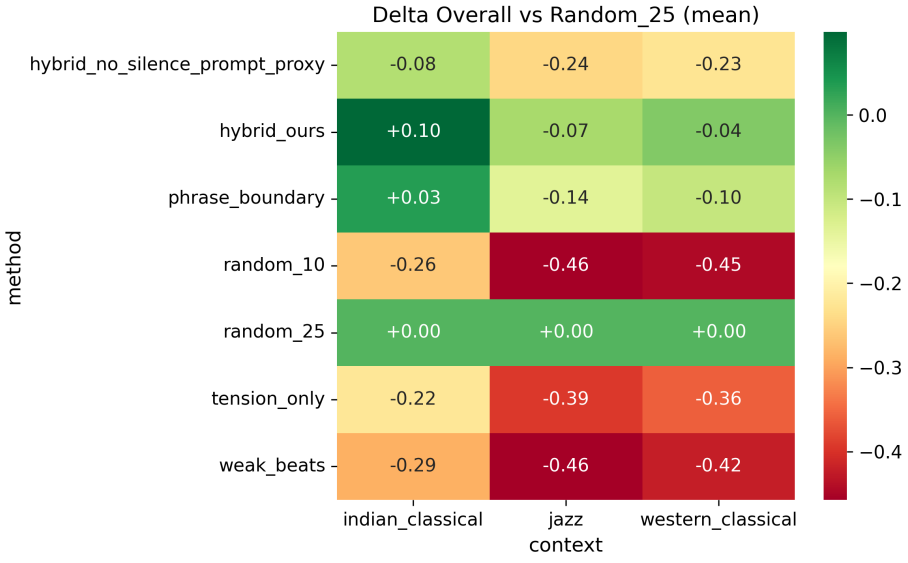


Figure 7: Delta Overall vs. random_25 baseline (heatmap).

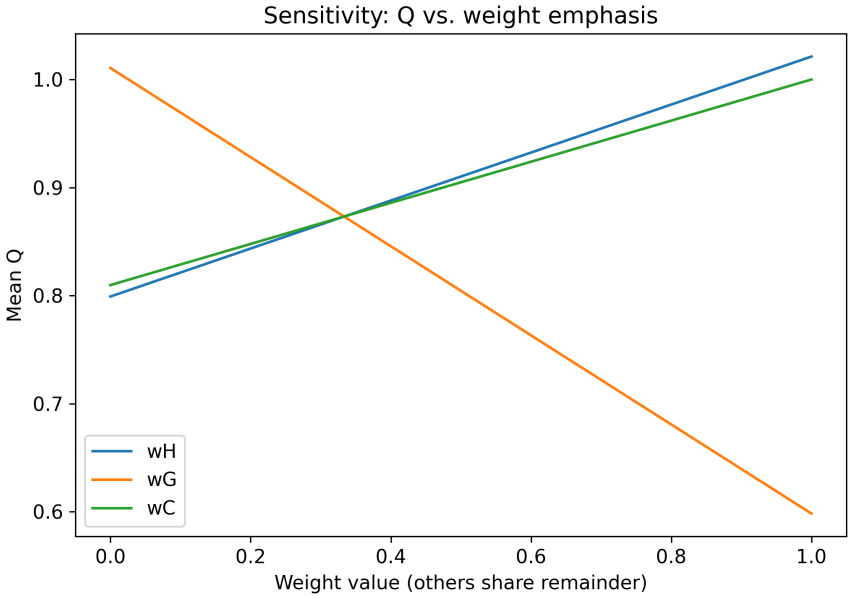


Figure 8: Sensitivity: $Q$ vs. weight emphasis (one component varied, others share remainder).

Table 2: Overall Quality by Method (mean±95% CI).

| Method | Overall Q | N |
|---|---|---|
| $\text{hybrid}_{n}o_{s}ilence_{p}rompt_{p}roxy$ | 0.665±0.081 | 12 |
| $\text{hybrid}_{o}urs$ | 0.846±0.092 | 12 |
| $\text{phrase}_{b}oundary$ | 0.782±0.069 | 12 |
| $\text{random}_{1}0$ | 0.461±0.020 | 12 |
| $\text{random}_{2}5$ | 0.849±0.067 | 12 |
| $\text{tension}_{o}nly$ | 0.524±0.081 | 12 |
| $\text{weak}_{b}eats$ | 0.460±0.000 | 12 |

## 6. Discussion

Strategic rests create space for breath, anticipation, and clarity. Results suggest a practical recipe: modest rest density aligned with phrase boundaries and selected tension points increases rhythmic variety without sacrificing pulse. Unlike black-box sequence models, our system exposes its internal heuristics, enabling both pedagogy and explainable composition. Beyond monophony, polyphonic extension invites voice-specific pacing, counter-rest design, and cadence-aware rests, with the same interpretability principles.

Table 3: Paired significance vs random_25 baseline (two-tailed sign-flip test); $N$=pairs per method.

| Method | $N$ | $\Delta Q$ (mean) | $p$-value | Cohen's $d$ |
|---|---|---|---|---|
| hybrid$_n o_s ilence_p rompt_p roxy$ | 12 | -0.184 | 0.0008 | -1.46 |
| hybrid$_o urs$ | 12 | -0.003 | 0.9300 | -0.02 |
| phrase$_b oundary$ | 12 | -0.068 | 0.0862 | -0.54 |
| random$_1$0 | 12 | -0.388 | 0.0002 | -3.03 |
| tension$_o nly$ | 12 | -0.325 | 0.0006 | -2.22 |
| weak$_b eats$ | 12 | -0.389 | 0.0005 | -3.28 |

Table 4: Two-voice sanity: Q improvement (upper voice masked).

| Melody | $Q_{orig}$ | $Q_{strat}$ | $\Delta Q$ |
|---|---|---|---|
| twinkle$_t winkle$ | 0.400 | 0.921 | 0.521 |
| mary$_h ad_a lamb$ | 0.400 | 0.886 | 0.486 |
| happy$_b irthday$ | 0.400 | 0.975 | 0.575 |
| c$_m ajor_s cale$ | 0.400 | 0.600 | 0.200 |

## 7. Applications

**Education and pedagogy.** Phrasing coach with real-time visual rests and A/B audio; assignments targeting boundary awareness.
**Accessibility.** Visual pacing for hearing-impaired users and focused listening support.
**Performance systems.** Live silence shaping for breathing points; adaptive rests by section.
**Production workflows.** Arrangement gap suggestion for density control and tension sculpting.
**Cross-cultural composition.** Preset-guided starting points refined with expert feedback.

## 8. Implementation Details

The analyzer, candidate generation, and metric evaluation are pure Python and operate in linear time in sequence length. Optional transformer components are lazily loaded and fully guarded; the system defaults to the lightweight hybrid when unavailable. The API exposes controls (density, weights) through a simple JSON schema; our CLI scripts provide one-command reproduction for experiments, figures, and audio. All figures included here were generated by these scripts against the released CSV. An anonymized code repository link will be provided upon acceptance.

## 9. Failure Cases

Our analysis reveals characteristic edge cases: (i) dense scalar runs with uniform metrical accents can over-encourage weak-beat rests unless groove emphasis is increased; (ii) anacruses and pickup notes may be misinterpreted as mid-phrase positions without explicit up-beat handling; (iii) highly syncopated lines benefit from stricter density caps to avoid local clustering of rests. In practice, increasing $w_G$ and reducing the density cap mitigates (i, iii), while a simple pickup detector addresses (ii). These analyses highlight the need for adaptive metrical priors and genre-specific density tuning, directions we plan to explore.

## 10. Supplementary Materials and Artifacts

We provide: (i) CSV of results with per-instance metrics; (ii) figures auto-generated from CSV; (iii) MIDI and WAV assets per melody $\times$ method $\times$ context under `assets/`; (iv) API/CLI for prediction and batch analysis. Cultural and context-aware considerations motivate our preset design (Jordanous and Johnson, 2020; Passmore and Savage, 2023). See Appendix for concise reproduction steps.

## 11. Ethics and Cultural Considerations

We aim for respectful cultural adaptation: presets are conservative defaults, not replacements for expertise. Audio artifacts disclose processing; any generative assistance will be acknowledged in camera-ready. If user data is collected, we will obtain consent, minimize retention, and anonymize. We welcome expert feedback to refine presets and mitigate cultural bias.

## 12. Limitations and Future Work

Current evaluation focuses on monophonic symbolic inputs; extending to multi-voice textures and audio-first settings is future work. We also plan larger, preregistered listener studies with genre stratification, DAW integration for production workflows, cadence-aware detectors, and UI prototypes for parameter steering and explanation browsing. Finally, we will study personalization, learning user-specific rest preferences over time.

## 13. Conclusion

We introduced a human-centered, interpretable system that treats silence as a controllable, culturally aware compositional element. By elevating rests to first-class outputs, our method enables musicians and researchers to co-create musical space intentionally, balancing rhythmic variety, groove, and structure. Beyond its technical formulation, this work reframes silence as a creative decision rather than absence, inviting new exploration in composition, pedagogy, and AI-driven performance.

## Acknowledgments

We thank musicians and reviewers whose feedback shaped the co-creative focus of this work.

## References

Roberto Bresin, Anders Friberg, and Johan Sundberg. Director musices: The kth performance rules system. In *Stockholm Music Acoustics Conference*, 2002.

Emilios Cambouropoulos. The local boundary detection model and its application. *Journal of New Music Research*, 2001.

Haonan Chen, Jordan B. L. Smith, Janne Spijkervet, et al. Sympac: Scalable symbolic music generation with prompts and constraints. In *ISMIR*, 2024.

Jade Copet, Felix Kreuk, Gabriel Synnaeve, et al. Simple and controllable music generation. In *NeurIPS*, 2023.

Xin Guan, Zhilin Dong, Hui Liu, and Qiang Li. Improving phrase segmentation in symbolic folk music: A hybrid model with local context and global structure awareness. *Entropy*, 2025.

Carlos Hernandez-Olivan, Sonia Rubio Llamas, and Jose R. Beltran. Symbolic music structure analysis with graph representations and changepoint detection methods. arXiv preprint, 2023.

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, et al. Music transformer: Generating music with long-term structure. In *ICLR*, 2019.

Yujia Huang, Adishree Ghatare, Yuanzhe Liu, et al. Symbolic music generation with non-differentiable rule-guided diffusion. In *ICML*, 2024.

Anna Jordanous and Colin G. Johnson. Computational creativity and music generation systems. *Frontiers in Artificial Intelligence*, 2020.

Guy Madison, Mats Friberg, et al. What musicians do to induce the sensation of groove. *Frontiers in Psychology*, 2014.

Ramon Medel, Carlos Cancino Chacón, and Emilia Gómez. Computational models of expressive music performance. *Frontiers in Digital Humanities*, 2016.

Oleg Mogren, Cheng-Zhi Anna Huang, Jorgen Sandviken, et al. Figaro: Controllable music generation using expert and learned features. In *ICLR*, 2023.

Corentin Nelias, Eva Marit Sturm, Thorsten Albrecht, et al. Downbeat delays are a key component of swing in jazz. *Communications Physics*, 2022.

Sam Passmore and Patrick E. Savage. The exceptions and the rules in global musical diversity. *Journal of Cognition*, 2023.

Godfried Toussaint. A measure of rhythm complexity using the normalized pairwise variability index. In *ICMPC*, 2004.

Peter van Kranenburg. Rule mining for local boundary detection in melodies. In *ISMIR*, 2020.

Gerhard Widmer. Discovering simple rules for expressive performance. In *Proceedings of the International Computer Music Conference*, 2003.

John Witek, Maria G. Kringelbach, et al. The sensation of groove is affected by the interaction of rhythmic and harmonic complexity. *PLOS ONE*, 2014.

Tingyu Zhu, Haoyu Liu, et al. Symbolic music generation with fine-grained interactive textural guidance. In *ICLR (submitted)*, 2025.