Controllable Symbolic Music Generation via Stage-Aware Style Routing and Differentiable Melody Regularization

Xuanfei Zhou , Yinxuan Huang , Sining Han , Jiangyao Bai , Qianzhen Zhang *

Posted Date: 16 April 2026

doi: 10.20944/preprints202604.0984.v1

Keywords:
symbolic music generation; controllable generation; diffusion models; hierarchical generation; style routing; melody regularization

 doi:10.20944/preprints202604.0984.v1

Article

Controllable Symbolic Music Generation via Stage-Aware Style Routing and Differentiable Melody Regularization

Xuanfei Zhou [1], Yinxuan Huang [2] and Sining Han [3], Jiangyao Bai [3], Qianzhen Zhang [3]*

[1] Preschool College, Changsha Normal University, Changsha 410100, China
[2] College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China
[3] College of Systems Engineering, National University of Defense Technology, Changsha 410073, China
* Correspondence: zhangqianzhen18@nudt.edu.cn

Abstract

Controllable symbolic music generation must preserve a reference melody while remaining responsive to style prompts. Existing hierarchical diffusion systems typically reuse a shared condition vector across harmony, rhythm, and timbre stages, which can entangle stylistic factors and weaken melody preservation. We present HCDMG++, a hierarchical diffusion framework that addresses these two limitations through Stage-Aware Style Routing and Differentiable Melody Regularization. The routing module uses a residual Multi-Layer Perceptron (MLP) to project text-derived style embeddings into stage-specific subspaces, whereas the regularization branch aligns soft pitch histograms and contour trajectories with the conditioning melody during training. We evaluate the integrated system on a 384-sample benchmark covering four melodies, eight styles, four random seeds, and three denoising budgets. HCDMG++ produces valid four-track outputs in all runs and reaches a peak pitch-histogram similarity of 0.508 under a 64-step budget. A matched legacy-compatible reference further shows substantially stronger pitch-histogram alignment than Legacy-HCDMG. These results indicate that stage-specific conditioning and differentiable structural guidance improve controllability in symbolic music diffusion.

Keywords: symbolic music generation; controllable generation; diffusion models; hierarchical generation; style routing; melody regularization

1. Introduction

Conditional symbolic music generation aims to synthesize musically coherent and expressive sequences while respecting user controls such as reference melodies, style descriptors, and structural hints. A central challenge in this setting is to achieve both *melody fidelity*—that is, to keep the generated multi-track output aligned with a user-provided melodic contour—and *style controllability*—that is, to produce outputs that remain perceptibly distinct under different style prompts. Although recent advances in Transformer architectures and latent-variable models have substantially improved long-range sequence modeling [1–3], generating complex polyphonic music under multiple simultaneous constraints remains difficult.

Recently, Denoising Diffusion Probabilistic Models (DDPMs) [4] have emerged as a powerful paradigm for generative modeling. When adapted to symbolic music [5–7], diffusion models generate multi-track event sequences through iterative refinement. To manage the complexity of polyphonic music, hierarchical diffusion systems, including the Legacy-HCDMG baseline considered here, decompose generation into specialized stages: harmony skeleton diffusion (HSD), rhythmic accompaniment diffusion (RAD), and detail timbre diffusion (DTD). This hierarchy is musically intuitive, as it mirrors a common compositional workflow in which harmony is established first, rhythm is conditioned on the harmonic progression, and timbral realization is conditioned on both.

 doi:10.20944/preprints202604.0984.v1

Despite the conceptual appeal of hierarchical generation, practical systems still face substantial architectural bottlenecks. A critical limitation of the original Legacy-HCDMG framework, as well as related hierarchical models, is its reliance on a *unified conditioning mechanism*. Injecting the same style embedding into all generation stages forces the network to encode heterogeneous musical attributes such as harmonic color, rhythmic density, and timbral texture within a single latent representation. We argue that this design encourages feature entanglement, which in turn weakens style separability and obscures fine-grained stylistic nuance [2,8]. In addition, long-form iterative denoising can accumulate errors over time, causing generated tracks to drift away from the conditioning melody [9–11].

To address these limitations, we build upon a reproducible implementation of the Legacy-HCDMG baseline and propose HCDMG++, an enhanced hierarchical diffusion framework designed to decouple style representation and reinforce structural fidelity.

More specifically, we introduce Stage-Aware Style Routing, which replaces rigid unified conditioning with a lightweight residual Multi-Layer Perceptron (MLP) router. This module projects text-derived style embeddings into stage-specific subspaces, allowing HSD to focus on harmonic attributes, RAD on rhythmic patterns, and DTD on multi-track timbral rendering. To mitigate melodic drift, we further introduce Differentiable Melody Regularization. By leveraging expected token indices and soft distributions derived from diffusion logits, this module applies histogram- and contour-alignment losses during training, thereby acting as an auxiliary structural guide without breaking the differentiable computation graph.

We evaluate HCDMG++ through a reproducible long-run benchmark spanning four melodies, eight styles, four random seeds, and three denoising budgets, thereby enabling a system-level view of controllability, fidelity, and efficiency tradeoffs. In summary, this study makes three contributions to controllable symbolic music generation. First, it introduces Stage-Aware Style Routing, which replaces unified conditioning with stage-specific residual routing so that the harmony, rhythm, and timbre stages can express distinct stylistic attributes. Second, it introduces Differentiable Melody Regularization, which injects soft histogram and contour constraints into diffusion training to mitigate melodic drift. Third, it establishes a multi-melody benchmark that reveals step-dependent alignment, style-response structure, and cross-melody variability under a unified evaluation protocol.

2. Related Work

2.1. Deep Learning for Symbolic Music Generation

Symbolic music generation has witnessed a paradigm shift from early statistical heuristics to advanced deep learning architectures. Transformer-based models, leveraging self-attention mechanisms [12], have demonstrated an exceptional ability to capture long-range dependencies and recurring motifs in musical sequences [13]. To handle the complex polyphony of music, advanced tokenization strategies have emerged. For instance, the REMI (Revamped MIDI) representation [14] integrates bar and beat tokens to enforce metrical strictness, while the Compound Word Transformer [15] groups concurrent musical attributes to significantly reduce sequence length and generation latency. Concurrently, latent-variable approaches like MusicVAE [16] employ hierarchical recurrent autoencoders to compress multi-track music into continuous latent spaces, enabling semantic operations such as interpolation. For multi-track orchestration, models like SymphonyNet [17] have further pushed the boundaries by introducing multi-track coordinate representations. These developments underscore the importance of robust event-sequence representations and hierarchical modeling, which form the basis of the present framework.

2.2. Diffusion Models for Sequence and Music Generation

Denoising Diffusion Probabilistic Models (DDPMs) [4] have achieved state-of-the-art performance in generative tasks. While Latent Diffusion Models (LDMs) [18] and advanced acoustic models like MusicLM [19] have revolutionized continuous audio waveform synthesis, the research community is actively adapting diffusion principles to discrete domains. Unlike autoregressive models that generate

 doi:10.20944/preprints202604.0984.v1

tokens strictly from left to right, diffusion models offer non-autoregressive, globally aware generation through iterative refinement. In natural language processing, Diffusion-LM [20] successfully bridged the continuous-to-discrete gap by mapping discrete words to continuous embedding spaces. In the symbolic music domain, Mittal et al. [5] demonstrated that analogous continuous diffusion and rounding techniques could synthesize coherent MIDI sequences. More recent works have explored hierarchical whole-song diffusion [6] and controllable rule-guided diffusion [11]. Because music is intrinsically structural, flat sequence generation often struggles to maintain macroscopic coherence. This motivates hierarchical diffusion architectures that factorize polyphonic generation into harmony, rhythm, and detail stages. However, the condition-fusion mechanisms in existing hierarchical diffusion frameworks remain rudimentary, a bottleneck our work seeks to resolve.

2.3. Style Conditioning and Feature Disentanglement

Controllable generation relies heavily on effectively injecting auxiliary conditions (e.g., text prompts, genre tags, or latent vectors) into the generative process. Textual conditions are increasingly encoded using pre-trained language models like Sentence-BERT [21], mapping natural language descriptors into dense semantic vectors. Recent frameworks such as FIGARO [22] have successfully demonstrated text-driven controllable symbolic music generation. Meanwhile, early generative models like MuseGAN [23] established the importance of track-level conditioning for coherent multi-instrument accompaniment.

A critical challenge in conditional modeling is *feature entanglement*. When a single, unified style embedding is injected uniformly across all stages, the model tends to average out stylistic nuances [24], leading to weak style separability or posterior collapse. In computer vision, architectures like StyleGAN [25] address this by injecting style vectors at different resolutions. Mathematically, this dynamic adaptation is often achieved via Feature-wise Linear Modulation (FiLM) [26], which scales and shifts features based on conditioning inputs. Inspired by these mechanisms, our proposed Stage-Aware Style Routing utilizes a residual MLP router to dynamically decouple the unified style embedding into harmonic, rhythmic, and timbral subspaces, allowing specific musical traits to be expressed at their corresponding hierarchical stages.

2.4. Structural Fidelity and Differentiable Regularization

Maintaining long-term structural fidelity—ensuring that generated accompaniments strictly adhere to a user-provided melody—is a persistent challenge. Dedicated models like PopMAG [27] attempt to mitigate this by jointly modeling melody and accompaniment through carefully designed attention masks. However, in diffusion models, the iterative denoising steps often cause the generated tracks to drift away from the conditioning melody, especially during lengthy polyphonic sections.

To enforce structural constraints, traditional approaches rely on non-differentiable post-hoc heuristic rules during inference, which cannot be optimized during training. Implementing constraints via differentiable proxies for discrete categorical variables is highly non-trivial; classical solutions include the Gumbel-Softmax trick [28] or straight-through estimators. In this work, rather than enforcing hard discrete token matching, we introduce a soft differentiable melody-regularization objective. By aligning pitch histograms and contours directly via expected token distributions [20], we provide continuous gradient feedback to the network. This anchors the generative process to the conditioning melody without interrupting the backpropagation pipeline.

2.5. Benchmarking and Reproducibility in Creative AI

Algorithmic reproducibility has become a central concern in creative AI. Generative music models depend heavily on large MIDI corpora, which are often noisy and heterogeneous [29]. At the same time, many music-generation papers still emphasize qualitative examples more heavily than systematic benchmark analysis. This motivates evaluating controllable symbolic generators under transparent multi-factor protocols that expose fidelity, style response, runtime behavior, and cross-seed variability. In this work, reproducibility is therefore treated not as a standalone contribution, but as an experimental

 doi:10.20944/preprints202604.0984.v1

prerequisite for assessing whether the proposed architectural changes produce measurable behavioral differences.

3. Motivation: Entangled Conditioning and Melody Drift

This section focuses on the scientific limitations of the Legacy-HCDMG baseline that motivate our method: weak style disentanglement under unified conditioning and structural drift under iterative denoising.

3.1. Weak Style Separability Under Unified Conditioning

Legacy-HCDMG fuses melody, style, and latent variation into a single condition vector and reuses that vector across all hierarchical generation stages. While computationally simple, this design implicitly assumes that harmonic color, rhythmic density, and timbral texture can be expressed through the same undifferentiated style representation. In practice, these attributes operate at different musical resolutions. When they are compressed into one shared condition, the model is encouraged to average over stage-specific cues rather than specialize them, leading to feature entanglement.

This limitation is reflected in the baseline behavior observed after end-to-end generation: coarse symbolic descriptors such as note density, rhythm complexity, and pitch range exhibit only weak separation across multiple style prompts. In other words, the baseline can produce valid multi-track outputs while still failing to translate stylistic conditions into sufficiently distinct symbolic structure. This weak separability motivates a stage-aware routing mechanism that allows harmony-, rhythm-, and timbre-relevant style components to be emphasized where they are most musically meaningful.

3.2. Melody Drift in Iterative Hierarchical Denoising

The second limitation concerns structural fidelity. In melody-conditioned symbolic diffusion, the accompaniment should remain aligned with the reference melody over long denoising trajectories. However, iterative refinement does not guarantee that the generated symbolic sequence will preserve either the global pitch distribution or the local contour implied by the conditioning melody. Small deviations introduced early in the reverse process can accumulate across stages, especially when the final multi-track realization is much denser than the input melody.

Our benchmark observations are consistent with this concern: melody-alignment metrics vary substantially across denoising budgets and melodies, and the relationship between computational budget and fidelity is non-monotonic rather than guaranteed. This motivates a differentiable regularization strategy that directly constrains soft pitch distributions and contour trajectories during training, rather than relying on generation-time heuristics alone.

4. Proposed Framework

This section details the architecture and mathematical formulation of HCDMG++, a hierarchical diffusion framework designed to decouple stylistic attributes and preserve conditioning melodies. As illustrated in Figure 1, the overall pipeline consists of three main phases. First, in the Condition Encoding phase, musical inputs and text prompts are mapped into continuous latent spaces, yielding an encoded melody $\mathbf{e}_m$, a continuous latent vector $\mathbf{z}$, and a global style embedding $\mathbf{e}_s$. Second, the Stage-Aware Style Routing phase utilizes parallel residual Multi-Layer Perceptrons (MLPs) to project the global style into stage-specific subspaces ($\tilde{\mathbf{e}}_s^{(h)}$, $\tilde{\mathbf{e}}_s^{(r)}$, and $\tilde{\mathbf{e}}_s^{(d)}$). Finally, in the Hierarchical Diffusion Backbone, these tailored conditions guide the cascaded generation of the harmony skeleton, rhythmic accompaniment, and detail timbre, while a Differentiable Melody Regularization objective continuously anchors the process to the reference melody. We first revisit the foundational hierarchical generation pipeline, followed by the detailed formulation of our two core innovations.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2026 doi:10.20944/preprints202604.0984.v1

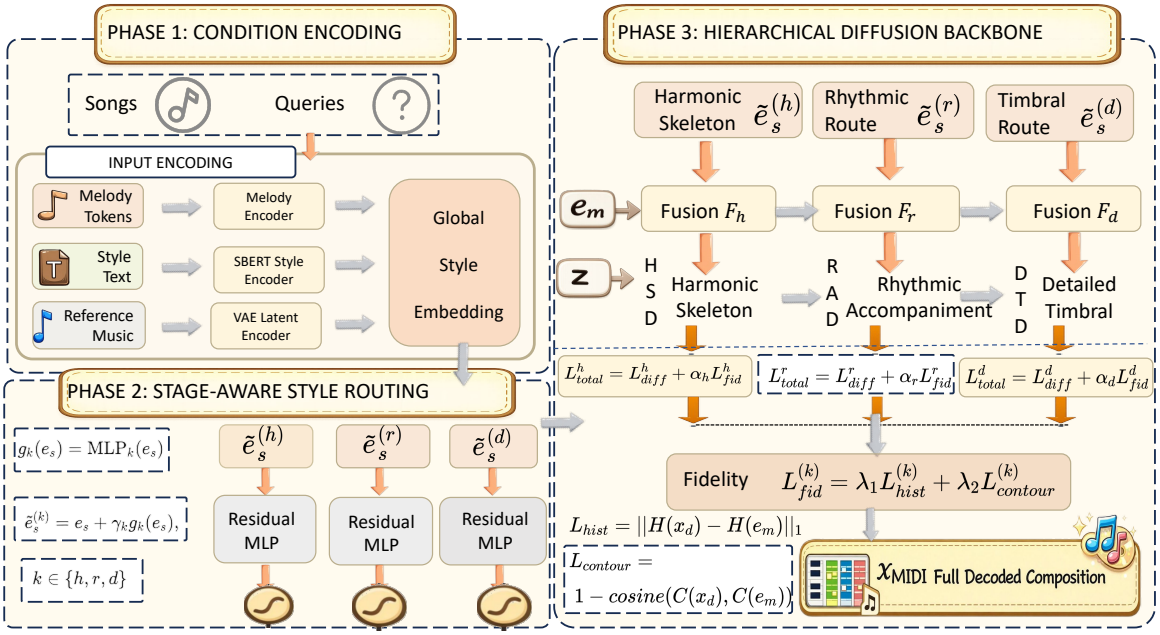


Figure 1. System overview of HCDMG++, comprising Condition Encoding, Stage-Aware Style Routing, and the Hierarchical Diffusion Backbone. Global style embeddings ($\mathbf{e}_s$) are routed into stage-specific condition vectors ($\tilde{\mathbf{e}}_s^{(h)}$, $\tilde{\mathbf{e}}_s^{(r)}$, $\tilde{\mathbf{e}}_s^{(d)}$) via parallel residual MLPs, avoiding feature entanglement across the Harmony Skeleton (HSD), Rhythmic Accompaniment (RAD), and Detail Timbre (DTD) diffusion stages. During training, a Differentiable Melody Regularization branch computes histogram ($\mathcal{L}_{\text{hist}}^{(k)}$) and contour ($\mathcal{L}_{\text{contour}}^{(k)}$) losses to preserve melodic fidelity.

4.1. Formulation of Hierarchical Symbolic Diffusion

Polyphonic symbolic music generation requires modeling complex joint distributions over time, pitch, velocity, and duration. HCDMG++ uses an event-sequence representation comprising 454 distinct MIDI-derived tokens. The generative process is factorized into three cascaded stages to mimic the human composition process: Harmony Skeleton Diffusion (HSD), Rhythmic Accompaniment Diffusion (RAD), and Detail Timbre Diffusion (DTD).

Let $\mathbf{e}_m \in \mathbb{R}^d$, $\mathbf{e}_s \in \mathbb{R}^d$, and $\mathbf{z} \in \mathbb{R}^d$ denote the encoded embeddings of the reference melody, the textual style descriptor (extracted via Sentence-BERT), and the continuous latent variation, respectively. The Legacy-HCDMG baseline unifies these conditions via a learned fusion module F to produce a global condition vector $\mathbf{c} = F(\mathbf{e}_m, \mathbf{e}_s, \mathbf{z})$.

Following the standard DDPM paradigm [4], each generation stage learns a reverse Markov transition to denoise a sequence of discrete tokens mapped to a continuous embedding space. The multi-stage generation proceeds as follows:

$$\mathbf{x}_h = \mathcal{D}_h(\mathbf{c}, \mathbf{x}_h^{(T)}), \tag{1}$$

$$\mathbf{x}_r = \mathcal{D}_r(\mathbf{c}, \mathbf{x}_r^{(T)} \mid \mathbf{x}_h), \tag{2}$$

$$\mathbf{x}_f = \mathcal{D}_d(\mathbf{c}, \mathbf{x}_f^{(T)} \mid \mathbf{x}_h, \mathbf{x}_r), \tag{3}$$

where $\mathcal{D}_h$, $\mathcal{D}_r$, and $\mathcal{D}_d$ represent the denoising networks (typically Transformer-based) for the harmony, rhythm, and final detail stages, respectively, starting from pure Gaussian noise $\mathbf{x}^{(T)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Although this hierarchical cascade partitions structural complexity, injecting the exact same unified condition $\mathbf{c}$ into all stages severely limits the model's ability to express fine-grained stylistic features, inevitably leading to feature entanglement.

 doi:10.20944/preprints202604.0984.v1

4.2. Stage-Aware Style Routing via Residual Multi-Layer Perceptron

To disentangle the semantic attributes of a musical style, we propose Stage-Aware Style Routing. In music theory, a genre is rarely defined by a single global parameter. For instance, "Jazz" implies extended chords in the harmony skeleton ($\mathcal{D}_h$), swing or syncopated patterns in the rhythmic accompaniment ($\mathcal{D}_r$), and specific instrumentations (e.g., upright bass, brass) in the timbre realization ($\mathcal{D}_d$).

To disentangle these attributes efficiently without drastically inflating the model's parameter count, we replace the naive linear concatenation with a dynamic, stage-specific Residual MLP router. For each hierarchical stage $k \in \{h, r, d\}$, we derive a customized style subspace embedding $\tilde{\mathbf{e}}_s^{(k)}$:

$$g_k(\mathbf{e}_s) = \mathbf{W}_{k,2} \cdot \sigma\left(\mathbf{W}_{k,1}\mathbf{e}_s + \mathbf{b}_{k,1}\right) + \mathbf{b}_{k,2}, \tag{4}$$

$$\tilde{\mathbf{e}}_s^{(k)} = \mathbf{e}_s + \gamma_k \cdot g_k(\mathbf{e}_s), \tag{5}$$

where $g_k(\cdot)$ is the stage-specific routing network, $\mathbf{W}$ and $\mathbf{b}$ are learnable projection weights and biases, and $\sigma(\cdot)$ denotes the Gaussian Error Linear Unit (GELU) activation function.

Crucially, γ_k is a learnable scalar gate with near-zero initialization (e.g., 10^{-4}). This zero-initialization strategy is inspired by ControlNet and Fixup architectures; it ensures strict backward compatibility with pre-trained legacy checkpoints. During the initial phases of fine-tuning, the router acts as an identity mapping ($\tilde{\mathbf{e}}_s^{(k)} \approx \mathbf{e}_s$), preventing catastrophic forgetting. As training progresses, γ_k learns to scale the residual stylistic features, dynamically allocating harmonic colors, groove templates, and timbral selections to their respective generation stages. The stage-aware condition vectors thus become $\mathbf{c}_h = F_h(\mathbf{e}_m, \tilde{\mathbf{e}}_s^{(h)}, \mathbf{z})$, $\mathbf{c}_r = F_r(\mathbf{e}_m, \tilde{\mathbf{e}}_s^{(r)}, \mathbf{z}, \mathbf{x}_h)$, and $\mathbf{c}_d = F_d(\mathbf{e}_m, \tilde{\mathbf{e}}_s^{(d)}, \mathbf{z}, \mathbf{x}_h, \mathbf{x}_r)$.

4.3. Differentiable Melody Regularization

A persistent challenge in sequence-to-sequence diffusion models is the degradation of structural constraints over prolonged denoising steps. Long-form generated sequences often drift from the initial conditioning melody, generating perceptually disconnected accompaniments.

To enforce melody adherence without relying on non-differentiable post-processing or heuristic masking, we propose a Differentiable Melody Regularization objective. As standard arg max operations used to decode continuous diffusion outputs back to discrete MIDI tokens disrupt the computational graph, we instantiate melody fidelity through continuous soft proxies based on the logits of the final diffusion projection layer.

Let $\boldsymbol{\ell}_t \in \mathbb{R}^V$ be the predicted logit vector over the vocabulary size V at sequence step t. The differentiable token distribution $\mathbf{p}_t$ is obtained via a temperature-scaled softmax: $\mathbf{p}_t = \text{softmax}(\boldsymbol{\ell}_t / \tau)$, where τ controls the distribution sharpness. Smaller τ values make the distribution closer to an arg max-like discrete selection, whereas larger values produce smoother probabilities; thus, temperature scaling provides a differentiable compromise between symbolic discreteness and stable gradient flow. To account for variable sequence lengths, we define $m_t \in \{0, 1\}$ as a binary padding mask.

We construct a Soft Pitch Histogram proxy for the generated sequence (H_{gen}) and compare it against the hard one-hot encoded histogram of the reference melody (H_{ref}):

$$H_{\text{gen}} = \frac{1}{N}\sum_{t=1}^{T_{seq}} m_t\, \mathbf{p}_t, \qquad H_{\text{ref}} = \frac{1}{N}\sum_{t=1}^{T_{seq}} m_t\, \text{onehot}(y_t), \tag{6}$$

where y_t is the ground-truth reference token, T_{seq} is the maximum sequence length, and $N = \sum m_t$ is the effective length. The global histogram alignment loss is calculated using the $\mathcal{L}_1$ norm: $\mathcal{L}_{\text{hist}} = \|H_{\text{gen}} - H_{\text{ref}}\|_1$.

Although the histogram captures global pitch distributions, it neglects temporal sequencing. To enforce temporal melody adherence, we define a Soft Pitch Contour proxy. We compute the expected token index at each time step $\hat{u}_t = \sum_{v=1}^{V} v \cdot \mathbf{p}_t(v)$. The local contour direction is represented by

 doi:10.20944/preprints202604.0984.v1

the first-order difference $\Delta\hat{u}_t = \hat{u}_t - \hat{u}_{t-1}$. The contour alignment loss penalizes angular deviations between the generated soft contour $\Delta\hat{\mathbf{u}}$ and the reference contour $\Delta\mathbf{u}$ using cosine distance:

$$\mathcal{L}_{\text{contour}} = 1 - \frac{\Delta\hat{\mathbf{u}} \cdot \Delta\mathbf{u}}{\|\Delta\hat{\mathbf{u}}\|\|\Delta\mathbf{u}\|}. \tag{7}$$

The finite difference is defined for $t \in [2, T_{seq}]$, and the initial term $\Delta\hat{u}_1$ (and its reference counterpart) is set to zero so that the contour proxy remains well-defined at the sequence boundary.

4.4. Overall Training Objective

The total composite loss function for optimizing HCDMG++ is defined stage-wise so that each diffusion stage receives its own fidelity feedback:

$$\mathcal{L}^{(k)}_{\text{total}} = \mathcal{L}^{(k)}_{\text{diff}} + \alpha_k\Big(\lambda_1\mathcal{L}^{(k)}_{\text{hist}} + \lambda_2\mathcal{L}^{(k)}_{\text{contour}}\Big), \tag{8}$$

where $k \in \{h, r, d\}$ indexes the HSD, RAD, and DTD stages, respectively. Here $\mathcal{L}^{(k)}_{\text{diff}}$ is the standard mean-squared error (MSE) between the predicted noise and the added Gaussian noise at diffusion timestep t for stage k. The coefficient α_k controls the overall intensity of the fidelity feedback at that stage, while λ_1 and λ_2 balance the global histogram alignment and the local contour consistency. This stage-specific regularization ensures that melody adherence is propagated down to the deepest hierarchical representations.

5. Experimental Setup

To validate the effectiveness of HCDMG++, we report a completed long-run multi-melody benchmark together with focused sensitivity analysis over denoising budgets.

5.1. Dataset and Inference Configurations

Benchmark evaluations span eight distinct stylistic presets: `free`, `classical`, `jazz`, `rock`, `electronic`, `pop`, `ambient`, and `cinematic`. This selection assesses the model's capacity to span sparse acoustic textures (e.g., classical, ambient) to denser rhythmic settings (e.g., rock and electronic).

For objective analysis, we standardize the generation window to sequences of 128 events and evaluate four curated reference melodies, denoted throughout the paper as Melody A, Melody B, Melody C, and Melody D. These correspond to the source MIDI files `I'm Good.mid`, `demo_melody.mid`, `example_melody.mid`, and `0-melody.mid`, respectively. The reverse diffusion process is tested under 16, 32, and 64 Denoising Diffusion Implicit Models (DDIM) steps, each combined with 8 style presets and 4 unique random seeds, yielding 384 generated multi-track samples in total. Unless otherwise stated, the figures in this section report the full 384-run sweep; the 32-step slice alone contains 128 directly comparable generations.

5.2. Objective Evaluation Metrics

To assess generative quality and controllability, we employ a set of standardized symbolic metrics. First, we evaluate *Style Separability* to measure the distinctiveness of the generated outputs under different conditioning prompts. This involves analyzing the variance of key musical descriptors across styles, specifically *Pitch-Range* (the interval between the highest and lowest active notes), *Polyphony Density* (the average number of simultaneous active notes per beat), and *Rhythm Complexity* (derived from syncopation rates and groove pattern entropy). A higher inter-style variance in these descriptors indicates that the model is effectively responding to the style conditions rather than reverting to a generic mean.

Second, we measure *Melody Fidelity* to quantify the structural adherence of the generated accompaniment to the conditioning melody. We utilize *Pitch-Histogram Intersection* to calculate the overlapping area between the probability mass functions of the reference and generated pitch classes, providing

 doi:10.20944/preprints202604.0984.v1

a global measure of harmonic compatibility. Additionally, *Contour Cosine Similarity* is employed to empirically track the alignment of melodic contours, ensuring that the generated sequences follow the intended directional trends of the input.

Finally, we implement *Diversity and Stability Diagnostics* to detect potential mode collapse and generation failures. We compute the intra-style *Note Density Variance* across different random seeds to verify that the model produces diverse variations for a fixed prompt. Concurrently, a *Degenerate Rate* metric strictly monitors the proportion of invalid outputs—defined as those containing fewer than 16 notes or lasting less than 4 seconds—to identify catastrophic failures in the HSD, RAD, or DTD diffusion stages.

5.3. Implementation Details

All models are implemented in PyTorch. Training is conducted on NVIDIA GPUs (e.g., RTX A6000) using the AdamW optimizer. To promote stable convergence, we apply linear learning-rate warmup followed by cosine annealing. For inference analysis and metric computation, generated artifacts are parsed with `pretty_midi` and aggregated with `pandas` and `seaborn`.

6. Results

This section presents the empirical evidence for HCDMG++. Rather than relying on isolated qualitative examples, we analyze the model through a unified 384-sample benchmark that exposes controllability, melody fidelity, runtime behavior, and cross-melody variability.

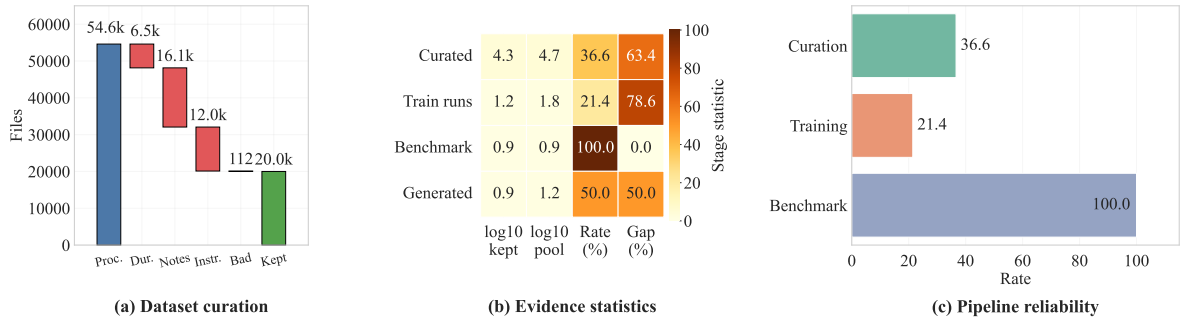
6.1. Benchmark Scope and Corpus Quality

Training hierarchical diffusion models for polyphonic music requires carefully curated symbolic data. As summarized in Table 1, the latest retraining pipeline screened 54,609 MIDI files. We applied strict filtering criteria to reject anomalies, most prominently note-count mismatches (29.44%), instrument-count deviations (21.89%), and duration inconsistencies (11.84%). Consequently, 20,000 samples were retained for training, corresponding to a retention ratio of 36.62%. These statistics underscore a point often overlooked in generative music research: raw symbolic corpora are substantially noisier than their nominal file counts suggest, and rigorous quality filtering is a prerequisite for musically meaningful modeling. Figure 2 extends this observation by linking dataset curation with benchmark readiness in a single evidence overview.

Table 1. Dataset filtering statistics ensuring the high-quality curation of the symbolic training corpus.

Statistic	**Count**	**Ratio (%)**
Processed MIDI files	54,609	100.00
Retained after filtering	20,000	36.62
Filtered out	34,609	63.38
Primary Rejection Causes		
Note-count mismatches	16,078	29.44
Instrument track mismatches	11,952	21.89
Duration anomalies	6,467	11.84
Data corruption	112	0.21

 doi:10.20944/preprints202604.0984.v1



(a) Dataset curation **(b) Evidence statistics** **(c) Pipeline reliability**

Figure 2. Dataset-to-evidence overview for the current HCDMG++ study. Panel (a) reports the dataset-filtering waterfall from 54,609 screened MIDI files to the retained 20,000-file subset, explicitly separating duration, note-count, instrument, and corruption rejections. Panel (b) summarizes the curated-dataset and benchmark-coverage stages as a statistical matrix over retained count, reference pool size, completion rate, and residual gap. Panel (c) reports the corresponding validity and benchmark-completion rates.

Table 2 shows that the integrated HCDMG++ pipeline operates stably under the long-run protocol: all 384 generations are valid four-track outputs with consistent duration and note counts. More importantly, the benchmark exposes measurable variation in alignment, runtime, and style response, enabling analysis beyond binary success or failure.

Table 2. Completed long-run HCDMG++ benchmark summary measured directly from the 384 generated MIDI artifacts.

Statistic	Value
Generated samples	384
Valid outputs	384 / 384 (100%)
Melody inputs	4
Style presets	8
Random seeds	4
Step budgets	16 / 32 / 64
Mean duration	31.50 s
Mean total notes	499.68
Mean pitch-histogram similarity	0.4251
Mean melody-track pitch similarity	0.4164
Mean interval-histogram similarity	0.4485
Mean melody-track interval similarity	0.4023

6.2. Legacy-Compatible Baseline Reference

Because the central claim of this work is that HCDMG++ addresses limitations of Legacy-HCDMG, at least one explicit baseline comparison is necessary even without a complete four-way ablation matrix. Table 3 therefore reports the closest legacy-compatible reference available in the current experimental record: the existing single-melody, eight-style Legacy-HCDMG benchmark versus a matched HCDMG++ slice generated with the same eight style prompts, four random seeds, and a 32-step budget on Melody D. This comparison is intentionally conservative. It does not replace the broader 384-run benchmark, but it provides a concrete reference point for judging whether the upgraded pipeline improves over the legacy system under a directly inspectable setup.

 doi:10.20944/preprints202604.0984.v1

Table 3. Legacy-compatible reference comparison between Legacy-HCDMG and HCDMG++ under the closest matched setting available in the current experimental record. Legacy-HCDMG statistics come from the existing single-melody eight-style benchmark, whereas HCDMG++ statistics are computed from the 32-step Melody D slice with four random seeds and the same eight styles. The two settings are aligned in prompt space but not identical in evaluation scale; accordingly, this table should be interpreted as a minimum baseline reference rather than a controlled efficiency comparison. `N/A` indicates quantities that cannot be estimated reliably from the legacy benchmark because it contains only one sample per style and does not export track-wise alignment metrics.

Metric	Legacy-HCDMG	HCDMG++
Generation success rate (%)	100.00	100.00
Mean pitch-histogram similarity	0.0043	0.3797
Mean interval-histogram similarity	0.7391	0.7310
Mean melody-track pitch similarity	N/A	0.3720
Style separability score	N/A	1.0581

As shown in Table 3, the legacy baseline exhibits near-zero full-output pitch-histogram overlap under its available benchmark, whereas HCDMG++ reaches 0.3797 on the matched 32-step slice and 0.3720 for melody-track pitch similarity. The near-zero legacy score (0.0043) is consistent with severe melodic drift during prolonged iterative denoising under unified conditioning and without explicit fidelity regularization, which can cause the generated accompaniment to lose pitch-level correspondence with the reference melody. The interval-histogram metric remains comparatively close across the two systems, suggesting that coarse interval statistics alone are insufficient to characterize controllability gains. Importantly, the legacy benchmark does not contain replicated samples per style, so a stable separability score cannot be computed for that system; however, the HCDMG++ slice already yields a positive style-separability estimate of 1.0581, indicating measurable style differentiation under the upgraded conditioning scheme. We therefore treat this table as a minimum legacy-compatible baseline reference rather than a substitute for a full ablation matrix, and we interpret the comparison in that limited but informative sense throughout the remainder of the paper.

6.3. Long-Run Multi-Melody Evaluation

Using the retrained HCDMG++ checkpoint, we executed the long-run protocol across 4 melodies, 8 styles, 4 random seeds, and 3 denoising budgets, yielding 384 valid generations. This evaluation exposes runtime scaling, melody-conditioned alignment differences, and non-trivial cross-seed/style variability under a common protocol.

Figure 3 presents step-budget diagnostics of the full 384-run sweep. Mean latency increases monotonically from 0.664 ± 0.327 s at 16 steps to 1.258 ± 0.612 s at 32 steps and 3.104 ± 0.280 s at 64 steps, but the four-panel presentation shows more than runtime growth alone. In addition to the runtime distribution, the figure overlays alignment curves with variability bands, places all samples in runtime–fidelity space, and normalizes runtime, note count, and melody fidelity on a common scale. This presentation clarifies the computational tradeoff as a joint efficiency-quality surface rather than a single latency curve.

 doi:10.20944/preprints202604.0984.v1

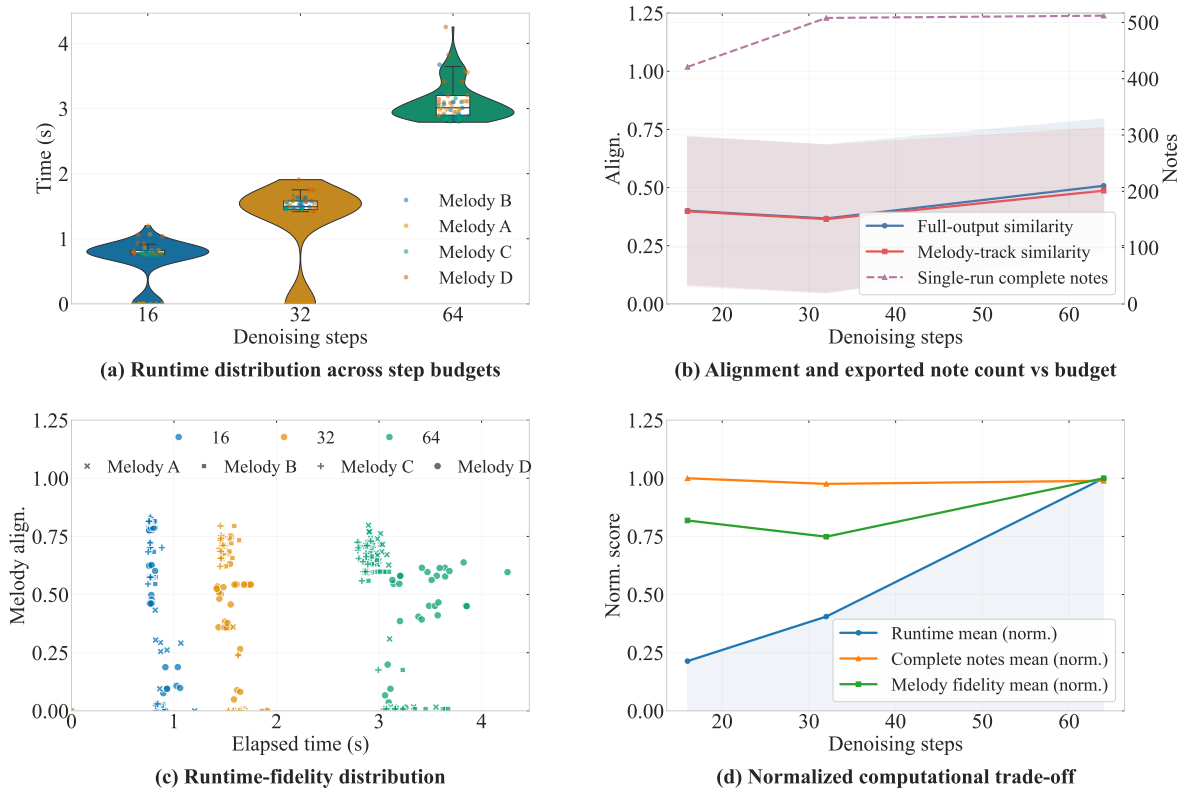
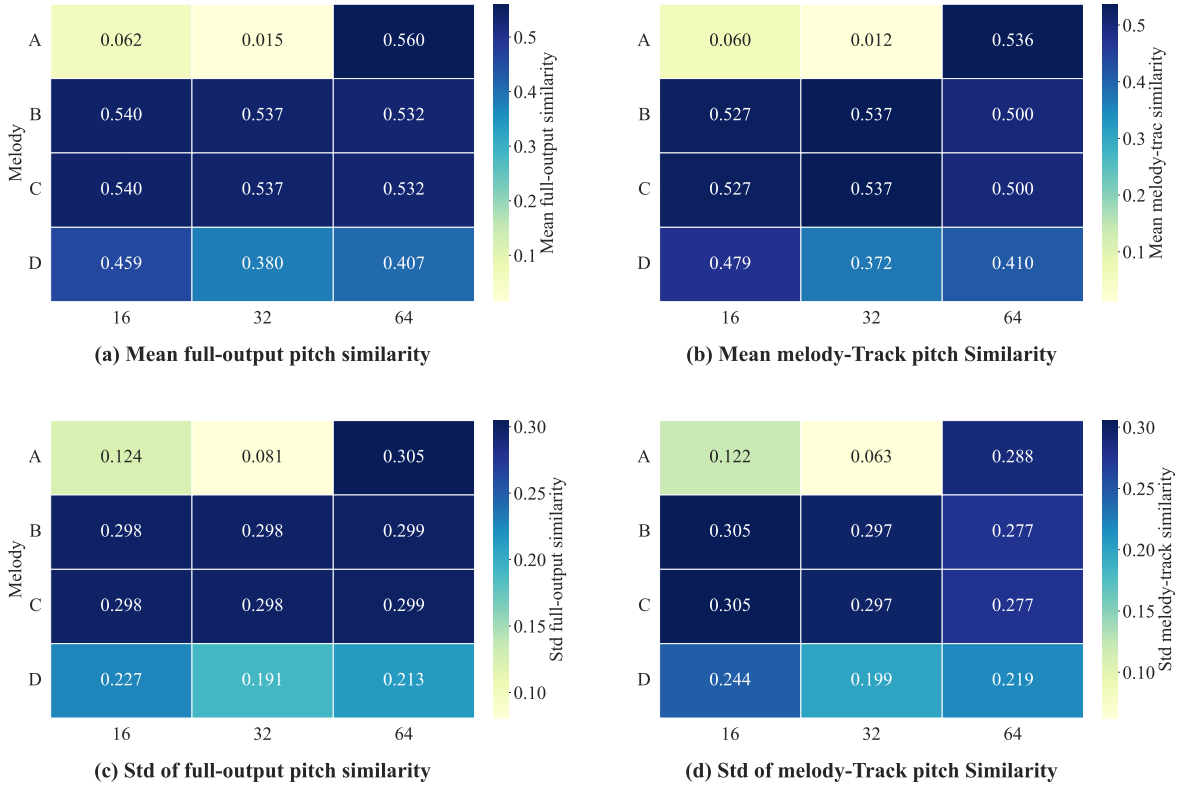


Figure 3. Step-budget diagnostics across the 384-run long-run benchmark. Panel (a) reports runtime distributions for 16, 32, and 64 denoising steps using violin, box, and sample overlays. Panel (b) combines full-output and melody-track pitch similarity with step-wise variability bands and note-count references. Panel (c) places all samples in runtime–melody-fidelity space to expose budget-dependent operating regions. Panel (d) normalizes runtime, note count, and melody fidelity on a common scale to make the efficiency-quality trade-off directly comparable. The benchmark pools four input melodies (Melody A–D: `I'm Good.mid`, `demo_melody.mid`, `example_melody.mid`, and `0-melody.mid`).

Figure 4 summarizes melody-alignment behavior with a multi-melody heatmap suite rather than a single averaged view. Mean full-output pitch-histogram similarity is 0.4004 at 16 steps, dips to 0.3671 at 32 steps, and rises to 0.5076 at 64 steps; the melody-track similarity follows the same pattern (0.3983, 0.3643, and 0.4866, respectively). Combining mean and standard-deviation heatmaps for both metrics makes clear that the alignment trend is not driven by a single outlier melody and that variability itself is melody-dependent.

Figure 5 shifts focus from melody-level dispersion to style-level response structure. The left and middle panels summarize how pitch range and melody fidelity vary jointly with style prompt and denoising budget, while the right panel decomposes style separability into between-style spread, within-style dispersion, and the resulting separability score. This view is critical, as the central question for HCDMG++ is not merely whether samples vary, but whether the variation aligns with intended style control rather than uncontrolled noise. The figure suggests that style response is real but still incomplete: separability improves only in specific budgets, and within-style dispersion remains large enough to blur stylistic boundaries for some prompt families.

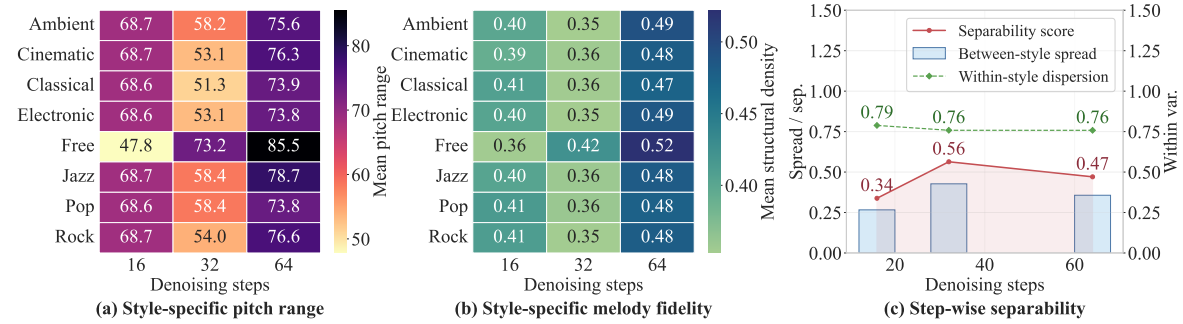
 doi:10.20944/preprints202604.0984.v1



(a) Mean full-output pitch similarity (b) Mean melody-Track pitch Similarity

(c) Std of full-output pitch similarity (d) Std of melody-Track pitch Similarity

Figure 4. Multi-melody heatmap suite for the 384-run benchmark. Panel (a) reports mean full-output pitch-histogram similarity, and Panel (b) reports mean melody-track pitch-histogram similarity. Panels (c) and (d) report the corresponding standard deviations across seed/style combinations. The combined view makes both average alignment quality and cross-run variability visible for each melody-step pair. Melody A–D denote the four curated input melodies `I'm Good.mid`, `demo_melody.mid`, `example_melody.mid`, and `0-melody.mid`, respectively.



(a) Style-specific pitch range (b) Style-specific melody fidelity (c) Step-wise separability

Figure 5. Style-response analysis across step budgets. Panel (a) reports the style-wise mean pitch range, and Panel (b) reports the style-wise mean melody-track pitch similarity. Panel (c) decomposes the step-wise style response into between-style spread, within-style dispersion, and the resulting separability score derived from multiple symbolic descriptors. All panels summarize responses aggregated over the four input melodies (Melody A–D). Together, the panels indicate that style effects remain budget-sensitive and are still only partially disentangled in the current checkpoint.

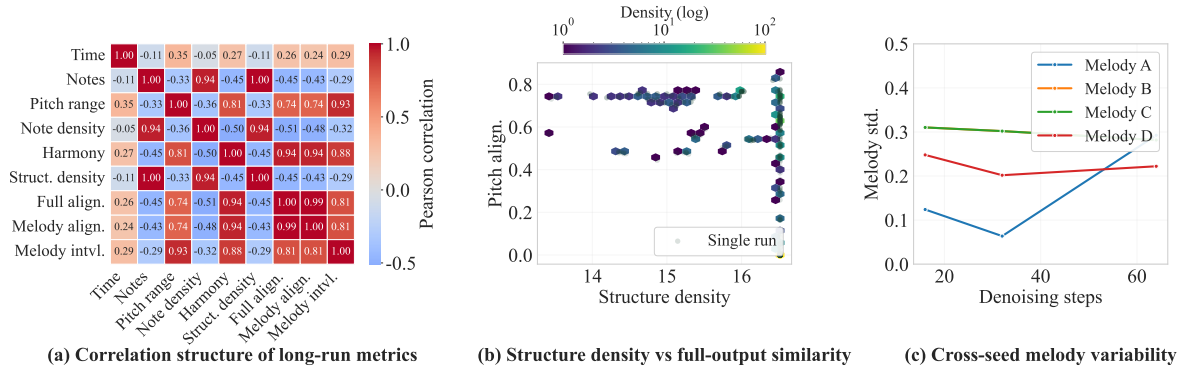
6.4. Metric Coupling and Efficiency Tradeoffs

Finally, we analyze the covariance of long-run symbolic descriptors beyond simple step-budget averages. Figure 6 summarizes the metric manifold of the benchmark by jointly showing the full correlation structure, the density of structure-versus-alignment samples, and the cross-seed melody-variability trajectories. This view is more informative than a single controlled sweep because it reveals which symbolic quantities move together across the full 384-run dataset.

The manifold confirms several non-trivial relationships. Runtime and alignment metrics are not isolated: pitch-range, structure-density, and melody-fidelity variables form partially coupled clusters,

 doi:10.20944/preprints202604.0984.v1

while some coarse descriptors remain only weakly correlated with alignment. The hexbin panel further shows that higher structural density does not guarantee stronger pitch-histogram similarity, indicating that denser symbolic realization is not equivalent to better melody preservation. Meanwhile, the melody-wise variability trajectories indicate that the 32-step regime often remains less stable than the 16- and 64-step settings for several inputs. These observations reinforce the conclusion that computational budget, controllability, and output structure interact in a genuinely multidimensional way rather than along a single monotonic axis.



(a) Correlation structure of long-run metrics (b) Structure density vs full-output similarity (c) Cross-seed melody variability

Figure 6. Metric-manifold analysis of the long-run benchmark. Panel (a) reports the Pearson correlation matrix over runtime, note-count, structure, and alignment descriptors. Panel (b) shows the density of samples in structure-density versus full-output pitch-similarity space. Panel (c) reports cross-seed variability in melody-track pitch similarity for each melody as the denoising budget changes. Together, the panels show that efficiency, structure, and controllability remain coupled but not reducible to a single scalar trend.

Given this tradeoff, 32 denoising steps represent a reasonable operating point for the current system: they provide moderate runtime cost and acceptable structural statistics, even though the 64-step budget yields the strongest average alignment and the 16-step budget can be more efficient and occasionally more stable than the 32-step midpoint. This recommendation should therefore be interpreted as an empirical operating choice for the present integrated HCDMG++ pipeline rather than a universal optimum for controllable symbolic diffusion.

7. Discussion

The results from the 384-sample benchmark indicate that HCDMG++ provides a more informative controllability profile than the legacy unified-conditioning setup. The legacy-compatible baseline reference in Table 3 is particularly useful in this regard: even under this limited comparison, HCDMG++ markedly improves full-output pitch-histogram similarity over Legacy-HCDMG and exhibits measurable style differentiation under the upgraded protocol, whereas the legacy benchmark is too sparse to support a stable separability estimate. This observation is consistent with the motivation behind Stage-Aware Style Routing, namely that harmonic, rhythmic, and timbral cues should not compete within a single undifferentiated style vector. At the same time, Differentiable Melody Regularization appears to provide a structural anchor, with pitch-histogram similarity peaking at 0.508 under the 64-step setting. The benchmark nevertheless shows that controllability is not resolved by a single architectural modification: style separability remains partial, and stronger fidelity often comes at higher computational cost.

More broadly, controllable symbolic music diffusion should be studied as a multi-objective problem rather than as a single-score optimization task. Runtime, density, style response, and melody alignment interact in ways that are clearly coupled but not reducible to a single scalar measure. For this reason, the dense statistical views in Figures 3–6 are not merely descriptive add-ons; they expose operating regimes, failure tendencies, and tradeoff surfaces that simpler aggregate summaries would obscure. Future work should therefore pair architectural advances with equally systematic evaluation protocols.

 doi:10.20944/preprints202604.0984.v1

8. Conclusions

This paper introduced HCDMG++, a hierarchical diffusion framework for controllable symbolic music generation that combines Stage-Aware Style Routing with Differentiable Melody Regularization. Across a 384-sample benchmark, the integrated system exhibits stable four-track generation together with measurable structure in melody fidelity, style response, and runtime tradeoffs. These findings support the view that hierarchical controllability benefits from stage-specific conditioning and differentiable structural guidance.

This study nevertheless has clear limitations. It evaluates HCDMG++ as an integrated pipeline and therefore does not yet isolate the marginal contribution of each proposed module through a full ablation matrix. In addition, the present evidence is objective and symbolic rather than perceptual; large-scale human listening studies are still needed to connect symbolic controllability metrics with musical preference and perceived style adherence. Future work will therefore extend the present study along two directions: more granular ablation experiments and broader subjective evaluation protocols that relate symbolic metrics to human musical judgment.

Funding: This research received no external funding.

Author Contributions: Conceptualization, Xuanfei Zhou; methodology, Xuanfei Zhou and Yinxuan Huang; software, Yinxuan Huang and Sining Han; validation, Xuanfei Zhou, Yinxuan Huang and Sining Han; formal analysis, Xuanfei Zhou and Sining Han; investigation, Xuanfei Zhou, Yinxuan Huang and Sining Han; resources, Xuanfei Zhou; data curation, Yinxuan Huang; writing—original draft preparation, Xuanfei Zhou and Yinxuan Huang; writing—review and editing, Xuanfei Zhou, Yinxuan Huang and Sining Han; visualization, Yinxuan Huang and Sining Han; supervision, Xuanfei Zhou; project administration, Xuanfei Zhou. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Min, L.; Jiang, J.; Xia, G.; Zhao, J. Polyffusion: A Diffusion Model for Polyphonic Score Generation with Internal and External Controls. In *Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR)*, Milan, Italy, 5–9 November 2023.
2. Lv, A.; Tan, X.; Lu, P.; Ye, W.; Zhang, S.; Bian, J.; Yan, R. GETMusic: Generating Any Music Tracks with a Unified Representation and Diffusion Framework. *arXiv* **2023**, arXiv:2305.10841.
3. Lu, P.; Xu, X.; Kang, C.; Yu, B.; Xing, C.; Tan, X.; Bian, J. MuseCoco: Generating Symbolic Music from Text. *arXiv* **2023**, arXiv:2306.00110.
4. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Virtual, 6–12 December 2020; pp. 6840–6851.
5. Mittal, G.; Engel, J.; Hawthorne, C.; Simon, I. Symbolic Music Generation with Diffusion Models. *arXiv* **2021**, arXiv:2103.16091.
6. Wang, Z.; Min, L.; Xia, G. Whole-Song Hierarchical Generation of Symbolic Music Using Cascaded Diffusion Models. *arXiv* **2024**, arXiv:2405.09901.
7. Yuan, R.; Lin, H.; Wang, Y.; Tian, Z.; Wu, S.; Shen, T.; et al. ChatMusician: Understanding and Generating Music Intrinsically with LLM. *arXiv* **2024**, arXiv:2402.16153.
8. Cífka, O.; Şimşekli, U.; Richard, G. Groove2Groove: One-Shot Music Style Transfer With Supervision From Synthetic Data. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2638–2650.
9. Dhariwal, P.; Jun, H.; Payne, C.; Kim, J.W.; Radford, A.; Sutskever, I. Jukebox: A Generative Model for Music. *arXiv* **2020**, arXiv:2005.00341.
10. Huang, Q.; Park, D.S.; Wang, T.; Denk, T.I.; Ly, A.; Chen, N.; Zhang, Z.; Zhang, Z.; Yu, J.; Frank, C.; Engel, J.; Le, Q.V.; Chan, W.; Chen, Z.; Han, W. Noise2Music: Text-conditioned Music Generation with Diffusion Models. *arXiv* **2023**, arXiv:2302.03917.
11. Huang, Y.; Ghatare, A.; Liu, Y.; Hu, Z.; Zhang, Q.; Shama Sastry, C.; Gururani, S.; Oore, S.; Yue, Y. Symbolic Music Generation with Non-Differentiable Rule Guided Diffusion. In *Proceedings of the International Conference on Machine Learning (ICML)*, Vienna, Austria, 21–27 July 2024; pp. 19772–19797.

 doi:10.20944/preprints202604.0984.v1

12. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
13. Huang, C.-Z.A.; Vaswani, A.; Uszkoreit, J.; Simon, I.; Hawthorne, C.; Dai, A.; Hoffman, M.D.; Dinculescu, M.; Eck, D. Music Transformer: Generating Music with Long-Term Structure. In *Proceedings of the International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, 6–9 May 2019.
14. Huang, Y.-S.; Yang, Y.-H. Pop Music Transformer: Beat-Based Modeling and Generation of Expressive Pop Piano Compositions. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, Seattle, WA, USA, 12–16 October 2020; pp. 1180–1188.
15. Hsiao, W.-Y.; Liu, J.-Y.; Yeh, Y.-C.; Yang, Y.-H. Compound Word Transformer: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Virtual, 2–9 February 2021; pp. 178–186.
16. Roberts, A.; Engel, J.; Raffel, C.; Hawthorne, C.; Eck, D. A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. In *Proceedings of the International Conference on Machine Learning (ICML)*, Stockholm, Sweden, 10–15 July 2018; pp. 4364–4373.
17. Liu, J.; Dong, Y.; Cheng, Z.; Zhang, X.; Li, X.; Yu, F.; Sun, M. Symphony Generation with Permutation Invariant Language Model. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*, Bengaluru, India, 4–8 December 2022.
18. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
19. Agostinelli, A.; Denk, T.I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; Sharifi, M.; Zeghidour, N.; Frank, C. MusicLM: Generating Music From Text. *arXiv* **2023**, arXiv:2301.11325.
20. Li, X.; Thickstun, J.; Gulrajani, I.; Liang, P.S.; Hashimoto, T.B. Diffusion-LM Improves Controllable Text Generation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA, 28 November–9 December 2022; pp. 4328–4343.
21. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong, China, 3–7 November 2019; pp. 3982–3992.
22. von Rütte, D.; Biggio, L.; Kilcher, Y.; Hofmann, T. FIGARO: Generating Symbolic Music with Fine-Grained Artistic Control. *arXiv* **2022**, arXiv:2201.10936.
23. Dong, H.-W.; Hsiao, W.-Y.; Yang, L.-C.; Yang, Y.-H. MuseGAN: Multi-Track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, 2–7 February 2018; pp. 34–41.
24. Yang, R.; Wang, D.; Wang, Z.; Chen, T.; Jiang, J.; Xia, G. Deep Music Analogy Via Latent Representation Disentanglement. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 4–8 November 2019; pp. 596–603.
25. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
26. Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; Courville, A. FiLM: Visual Reasoning with a General Conditioning Layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, 2–7 February 2018; pp. 3942–3951.
27. Ren, Y.; He, J.; Tan, X.; Qin, T.; Zhao, Z.; Liu, T.-Y. PopMAG: Pop Music Accompaniment Generation. *arXiv* **2020**, arXiv:2008.07703.
28. Jang, E.; Gu, S.; Poole, D. Categorical Reparameterization with Gumbel-Softmax. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Toulon, France, 24–26 April 2017.
29. Ens, J.; Pasquier, P. Building the MetaMIDI Dataset: Linking Symbolic and Audio Musical Data. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, Online, 7–12 November 2021; pp. 182–188.

 doi:10.20944/preprints202604.0984.v1

disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.