# Unifying Multitrack Music Arrangement via Reconstruction Fine-Tuning and Efficient Tokenization

**Anonymous Authors**

## Abstract

Automatic music arrangement streamlines the creation of musical variants for composers and arrangers, reducing reliance on extensive music expertise. However, existing methods suffer from inefficient tokenization, underutilization of pre-trained music language models (LMs), and suboptimal fidelity and coherence in generated arrangements. This paper introduces an efficient multitrack music tokenizer for unconditional and conditional symbolic music generation, along with a unified sequence-to-sequence reconstruction fine-tuning objective for pre-trained music LMs that balances task-specific needs with coherence constraints. Our approach achieves state-of-the-art results on band arrangement, piano reduction, and drum arrangement, surpassing task-specific models in both objective metrics and perceptual quality. Additionally, we demonstrate that generative pretraining significantly contributes to the performance across these arrangement tasks, especially when handling long segments with complex alignment.[1]

## 1 Introduction

Music arrangement, the art of adapting musical compositions for different performance contexts, plays a fundamental role in music creation by enabling artistic reinterpretation and expanding the accessibility of musical works. It is ubiquitous across various musical domains, from professional music production and education to live performance and digital content creation. However, manual arrangement requires extensive expert knowledge of harmony, instrumentation, and style-specific conventions, making it a highly specialized and demanding process. Automating this task could make arrangement more accessible to individuals with limited musical training and allow composers to focus more on the creative aspects by relieving them of technical groundwork.

Despite the potential benefits, existing methods for automatic arrangement face significant challenges, leading to substantial quality gaps that hinder practical applications. One critical issue lies in tokenization, the foundation of any unconditional or conditional generative music modeling. The commonly adopted REMI tokenizer [Huang and Yang, 2020] often results in excessively long sequence lengths, a known bottleneck for symbolic music generation [Hsiao *et al.*, 2021]. Moreover, as an REMI's extension to multitrack music, REMI+ [von Rütte *et al.*, 2023] contains redundant information and suboptimal design choices that lead to the *content fragmentation* issue in the tokenized sequences, burdening both generative modeling and arrangement performance.

Another limitation of prior arrangement methods is their reliance on task-specific model structures. While generative pre-training has demonstrated great potential in related fields such as NLP, integrating similar approaches into these highly specialized structures is not straightforward. Consequently, these methods fail to fully leverage the generative power of pre-trained music language models (LMs). Given the scarcity of annotated music data, this limitation further constrains their effectiveness.

Finally, existing models often struggle with fidelity and coherence. Specialized arrangement models [Zhao *et al.*, 2023; Zhao *et al.*, 2024] frequently fail to preserve melody and texture, retaining only harmony, and thus are perceived by users as lacking resemblance to the original piece. Also, our experiments show that even the state-of-the-art piano reduction model [Terao *et al.*, 2023] exhibit suboptimal arrangement coherence since different musical segments are handled independently.

To address these limitations, we propose a novel framework for automatic symbolic music arrangement. First, we design a more efficient tokenizer to solve the content fragmentation issue and reduce sequence length and redundancy. Second, we explore the potential of sequence-to-sequence fine-tuning with our pre-trained music LM, enabling task specific models to harness the power of generative pre-training. This approach allows flexible adaptation to various arrangement tasks with limited annotated data. Finally, we validate our unified methodology across multiple tasks, including band arrangement, piano reduction, and drum arrangement.

Our key contributions are summarized as follows:

- We propose an **efficient tokenizer for multitrack music** that reduces sequence length and redundant information, improves effectiveness for unconditional pre-

[1] Demos available at ae4rtjsyr.github.io. All code will be released upon acceptance.

training, enhances arrangement performance, and potentially extensible to general music generation tasks.

- We introduce a **self-supervised reconstruction objective** to effectively adapt pre-trained music LMs to various music arrangement tasks, ensuring both task-specific requirements and coherence.
- We achieve **state-of-the-art performance** on band arrangement, piano reduction, and drum arrangement tasks, surpassing task-specific models in both objective and subjective evaluations.
- Through ablation studies, we demonstrate how **generative pre-training significantly improves** bar-to-bar transformations and plays a crucial role in processing longer segments, where models without pre-training fail.

# 2 Related Work

## 2.1 Automatic Arrangement in Multitrack Symbolic Music

Music arrangement encompasses various tasks like chord progression creation [Yi *et al.*, 2022], lead sheet orchestration [Wang *et al.*, 2024], and composition adaptation [Zhao *et al.*, 2024]. This paper focuses on adapting existing music to new instruments or ensembles, which is crucial for multi-style music generation and style transfer.

Earlier supervised approaches trained end-to-end models using parallel data (e.g., piano-to-orchestra [Crestel and Esling, 2016] or band-to-piano [Terao *et al.*, 2022; Terao *et al.*, 2023]), which are extremely challenging to prepare and inherently restrict the direction of arrangement. Classification-based methods such as [Dong *et al.*, 2021] attempt to predict instrument labels for each note, but they lack creativity by generating no new content and suffer from inflexibility, being incapable of handling unseen instrument combinations. Recent self-supervised methods like Q&A [Zhao *et al.*, 2023; Zhao *et al.*, 2024] offer more flexible solutions without requiring parallel datasets or fixed instrument set assumptions. However, [Zhao *et al.*, 2023] assumes prior knowledge of track-wise output distributions (i.e., note density and pitch histograms), which is impractical for arrangement tasks since such knowledge is unavailable before generation. While [Zhao *et al.*, 2024] attempts to solve this by introducing a separate *prior model* to estimate the track-wise distributions, this approach comes at the cost of compromised arrangement fidelity. Additionally, infilling-based models such as Composer's Assistant [Malandro, 2023; Malandro, 2024], although capable of handling additive arrangement scenarios such as drum arrangement, do not preserve the contents of the original composition in the generation output and are thus fundamentally unsuited for band arrangement, where the primary aim is to reinterpret existing musical content.

## 2.2 Generative Symbolic Music Models

Transformer networks have advanced music generation tasks like chord-to-melody generation [Madaghiele *et al.*, 2021], accompaniment arrangement [Ren *et al.*, 2020], and music inpainting [Malandro, 2023; Malandro, 2024]. However, these models remain task-specific, with limited exploration of unconditional pre-training and task-specific fine-tuning. Regarding controlled generation, existing works primarily focus on using simple high-level attributes (style [Choi *et al.*, 2020; Lu *et al.*, 2023], structure [Zhang *et al.*, 2022], and sentiment [Makris *et al.*, 2021]), with minimal investigation of music-based content conditioning. As large symbolic music LMs emerge [Dong *et al.*, 2023; Qu *et al.*, 2024; Wu *et al.*, 2024], it becomes crucial to explore more complex control signals, such as music itself, and expand their application to broader scenarios.

## 2.3 Symbolic Music Tokenizers for Transformers

Transformer-based music modeling requires converting data into token sequences, which is challenging due to music's multi-stream nature. Many tokenizers treat music as a single stream using MIDI-like note-event encoding, representing each musical note with multiple tokens (onset time, pitch, duration, velocity, instrument). Some use absolute timing [Huang *et al.*, 2018; Gardner *et al.*, 2021; Zeng *et al.*, 2021; Ens and Pasquier, 2020], while others use metrical durations [Ren *et al.*, 2020; Huang and Yang, 2020; von Rütte *et al.*, 2023]. Among them, the REMI [Huang and Yang, 2020] tokenizer is widely adopted for single-track music and has been extended to multitrack in [von Rütte *et al.*, 2023] by adding an instrument token as an attribute of each note.

However, such extension, along with all approaches that strictly order notes temporally for multitrack music, has an inherent limitation—**content fragmentation**, i.e., notes from one instrument being interleaved with notes from other instruments playing simultaneously. As shown in Figure 1a, the content of instrument `i-29` (distorted electric guitar, highlighted in orange) spreads across the entire sequence, interleaved by notes from instrument `i-80` (synth lead, highlighted in red). Consequently, tokenization results can be highly inconsistent for same instrument-wise contents once its context (interleaving contents) differs, bringing extra difficulty for generative modeling.

## 2.4 Prompt-based Fine-Tuning for Conditional Generation

Prompt-based fine-tuning in NLP [Liu *et al.*, 2023] uses control tokens to guide pre-trained LM outputs, enabling attribute control over style [Sennrich *et al.*, 2015], length [Kikuchi *et al.*, 2016], and pronunciation [Ou *et al.*, 2023]. Combining such strategy with LMs pre-trained with standard left-to-right language modeling objective [Radford *et al.*, 2019; Brown *et al.*, 2020] can effectively generates fluent, condition-aligned outputs, providing insights for the music arrangement problems, where similar pre-training and control mechanisms can potentially generate music that is musically appealing and meets task-specific requirements.

# 3 Method

To bridge the gaps in previous research, we adopt a pre-train and fine-tune paradigm for the arrangement task, design an efficient multitrack music representation for Transformer networks that makes both pre-training and fine-tuning more effective, and develop a unified objective for music-to-music

**o-0** i-26 p-60 d-26 **o-0** i-33 p-36 d-23 **o-0** i-29 p-36 d-10 **o-12** i-29 p-36 d-12 **o-18** i-80 p-74 d-14 **o-18** i-29 p-48 d-12 **o-24** i-29 p-36 d-8 **o-30** i-29 p-52 d-11 **o-36** i-80 p-76 d-11 **o-36** i-29 p-36 d-10 **o-42** i-29 p-52 d-7 b-1

(a) REMI+ tokenization, demonstrated with REMI-z vocabulary.

**i-80** o-18 p-74 d-14 o-36 p-76 d-11 **i-26** o-0 p-60 d-26 **i-29** o-0 p-36 d-10 o-12 p-36 d-12 o-18 p-48 d-12 o-24 p-36 d-8 o-30 p-52 d-11 o-36 p-36 d-10 o-42 p-52 d-7 **i-33** o-0 p-36 d-23 b-1

(b) A REMI-z bar sequence containing four track sequences.

Figure 1: REMI-z and REMI+ tokenization for a same bar. Content of the same instruments are highlighted with same color.

| Meaning | Token | X's range |
|---|---|---|
| Instrument type | i-X | 0∼128 |
| Note's within-bar position | o-X | 0∼127 |
| Note's pitch (non-drum) | p-X | 0∼127 |
| Note's pitch (drum) | p-X | 128∼255 |
| Note's duration | d-X | 0∼127 |
| End of a bar | b-1 | - |
| Time signature | s-X | 0∼253 |
| Tempo | t-X | 0∼48 |

Table 1: REMI-z vocabulary with pitch token distinctions

transformation that addresses both task-specific requirements and coherence. We test the effectiveness of this approach on multiple typical music arrangement tasks.

### 3.1 REMI-z Representation

Instruments have inherent syntactic rules governed by their physical properties and conventional playing patterns. A good arrangement, or in broader terms, any musical composition, must ensure each instrumental part is both executable and idiomatic. However, as mentioned earlier, time-ordered tokenizers struggle with instrument-wise content fragmentation, making it difficult to model these syntactic rules. To address this issue, we propose *REMI-z*, a tokenization scheme that does not sort notes strictly by time but instead prioritizes the sequence continuity of content for each instrument.

The vocabulary of REMI-z is shown in Table 1, similar to that of REMI+ [von Rütte *et al.*, 2023] (detailed in Appendix A). The key innovation lies in how notes are organized. REMI-z tokenize arbitrary-length MIDI files into a list of *bar sequences*, each ending with a special end-of-bar token (`b-1`). Within each bar, notes are further grouped into *track sequences*, where each track corresponds to a specific instrument. Each track sequence begins with an instrument token specifying the instrument type, followed by all notes played by that instrument within the bar. Track sequences are ordered by their average pitch (from high to low), and within each track sequence, notes are represented as triplets of (position, pitch, duration). These notes are sorted first by position (left to right) and then by pitch (high to low) before being flattened into a 1-D sequence. If two adjacent notes share the same position, the second position token is omitted.

The primary difference between REMI-z and REMI+ is that in REMI-z, notes belonging to the same instrument are grouped together within each bar (as in Figure 1b), significantly alleviating the content fragmentation issue and facilitating instrument-specific syntactic modeling. Another advantage is that the sequence length is reduced, as instrument tokens appear only once per track sequence in REMI-z, whereas REMI+ needs an instrument token for each note.

We adopt REMI-z tokenization to pre-train a Transformer decoder using a standard next-token-prediction objective on a large unlabeled corpus and then proceed to convert the pre-trained model to task-specific models with the objective in the following section.

### 3.2 Arrangement Fine-Tuning

Music arrangement involves adapting given musical content to a new set of instrumental constraints. With a well-trained content- and instrument-controllable generation model, novel arrangements can be generated from the original composition by assigning new instrumentation constraints at inference time. To train this capability without relying on parallel data, such as multiple versions of the same song arranged for different instrumentations, we propose a context-aware self-supervised segment-level reconstruction objective. Let $y^{(t)}$ denote the $t$-th *segment* of a music piece. The fine-tuning objective is defined as:

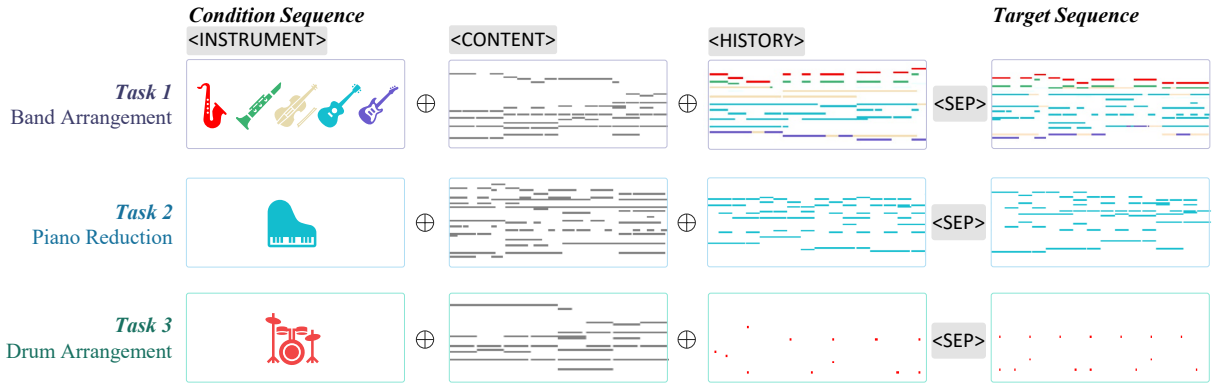
$$\mathcal{L}(\theta) = -\log p_\theta\big(\mathcal{T}_{\text{task}}(y^{(t)}) \,\big|\, \mathrm{I}(\mathcal{T}_{\text{task}}(y^{(t)})), \mathrm{C}(\mathcal{S}_{\text{task}}(y^{(t)})),\ \mathcal{T}_{\text{task}}(y^{(t-1)})\big), \quad (1)$$

where $\theta$ represents the model parameters, $\mathrm{I}(\cdot)$ and $\mathrm{C}(\cdot)$ extract *instrument* and *content* information, $\mathcal{S}_{\text{task}}$ and $\mathcal{T}_{\text{task}}$ represent selecting task-specific source and target tracks, and $y^{(t-1)}$ provides target-side *history*. Equation 1 is implemented with a standard LM objective to the sequence `[condition]<sep>[target]`, as illustrated in Figure 2b, with cross-entropy loss computed only on the target subsequence.

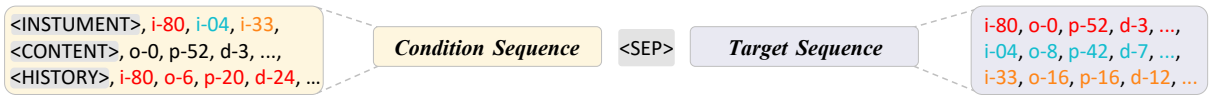
During training, the instrument, content, and history conditions are derived from the same music being reconstructed. The model is trained to reconstruct the original music, learning to interpret the given musical content with a specific instrument set in a certain context. During inference, users can flexibly define instrument constraints, while the content comes from the music to be arranged, and history from the model's previously generated outputs.

The instrument condition, derived from $\mathrm{I}(\cdot)$, specifies the desired instruments to be used in the segment. During training, it includes all instrument tokens present in the target-side sequence. Additionally, the relative positions of tokens within this sequence encode important information about the voice relationships between instruments: tokens earlier in the sequence correspond to instruments with higher average pitches, and vice versa. At inference, users have the flexibility to specify both the instruments and their voice relationships for each segment.

The content condition from $\mathrm{C}(\cdot)$ represents the musical notes from the original composition, encoded as a *content sequence*. This sequence provides precise information about what notes are played and their timing, but without specifying instruments. To construct the content sequence, the following steps are applied: (1) remove all instrument tokens from REMI-z, (2) sort notes strictly by their temporal positions, (3) sort notes within the same position by pitch in descending




(a) Detailed settings for each arrangement task. The symbol $\bigoplus$ denotes concatenation of component sequences.




(b) An example of the tokenized sequence for a band arrangement task. Special tokens <SEP>, <INSTRUMENT>, <CONTENT>, and <HISTORY> are used to separate different components. Notes of each instrument are highlighted with different colors.

Figure 2: Music segments are decomposed into three subsequences: *instruments*, *content*, and target-side *history*. These components form the *condition sequence*, with the original music as the *target sequence*. The model is trained to reconstruct the music from these components.

order, and (4) merge duplicate notes with identical positions and pitches. This representation ensures a clean, time-ordered view of the musical content of a multitrack segment without any instrument-related information.

Lastly, we introduce a history condition $\mathcal{T}_{\text{task}}(y^{(t-1)})$ that encourages contextual coherence between segments during the arrangement process to tackle the style inconsistency issue brought by segment-level processing, inspired by the success of historical context in machine translation [Tiedemann and Scherrer, 2017]. During training, the history is provided through teacher-forcing using the ground truth REMI-z sequence from the previous segment's target sequence, while at inference time, it is generated autoregressively based on the model's output from the preceding segment. This mechanism ensures that each segment is generated with awareness of its musical context, leading to coherent arrangements.

## 3.3 Tasks

We evaluate our method on the following tasks, which are representative of typical arrangement scenarios in music composition studies. Each task assesses different aspects of model capabilities.

### Band Arrangement

This task involves arranging an existing piece of music for arbitrary combinations of instruments. The model must understand the properties and typical playing styles of each instrument to allocate or generate new notes appropriately. In this setup, both $\mathcal{S}_{\text{task}}$ and $\mathcal{T}_{\text{task}}$ are defined as identity transformations, directly returning the original REMI-z sequence, i.e., $\mathcal{S}_{\text{task}}(y) = \mathcal{T}_{\text{task}}(y) = y$. Note that we focuses exclusively on pitched instruments and pre-remove drum tracks from $y$ due to REMI-z's drum-specific pitch tokens. Drum arrangement is addressed as a separate task.

To foster creativity, we randomly delete a small proportion of tracks in $\mathrm{C}(\mathcal{S}_{\text{task}}(y^{(t)}))$ during training, encouraging the model to generate new notes that are compatible with the musical context but absent from the input music during inference. Special care is taken to preserve the melody during random deletions by ensuring that the track with the highest average pitch remains untouched. Additionally, duration tokens are removed from the content sequence to allow the model to generate durations that best suit the playing styles of the desired instruments, rather than merely copying from inputs. The segment length is set to 1 bar.

### Piano Reduction

This task focuses on simplifying multi-instrumental musical pieces into solo piano arrangements that ensure pianistic playability while preserving the original musical essence, i.e., harmonies and textures. In the training process, $\mathcal{T}_{\text{task}}$ is defined to extract the original piano tracks, while $\mathcal{S}_{\text{task}}$ is an identity transformation. To ensure the piano part is sufficiently prominent in the training data, we filter the dataset to include only segments where the piano pitch range constitutes more than 40% of the total pitch range, reducing the dataset to approximately 40% of its original size. The segment length is set to 1 bar. Additionally, drum tracks are also pre-removed from $y$.

### Drum Arrangement

This task involves creating a drum track for songs that lack one. The model needs to recognize the groove of the music and enhance it using the drum set, and further, handle transitions between musical phrases to drive the music forward and make it more engaging, which consequently requires a better understanding of musical structure. Here, $\mathcal{S}_{\text{task}}$ extracts tracks belonging to pitched instruments, while $\mathcal{T}_{\text{task}}$ extracts the drum track. A longer segment length of 4 bars is used

for this task, as drum patterns generally span across multiple bars.

# 4 Experiments

## 4.1 Implementation Details

Our model, an 80M-parameter decoder-only Transformer, has a hidden dimension of 768, 12 layers, 16-head attention, and a context length of 2048 tokens (around 8× the longest bar in our dataset). The model first undergoes a standard LM pre-training, and then was fine-tuned with the proposed objective. Pre-training used four RTX A5000 GPUs (batch size 12, 1 epoch), while fine-tuning used a single A40 GPU (variable batch size, 3 epochs). Training was done with two publicly accessible datasets. Pre-training adopted the Los-Angeles-Project dataset [Lev, 2024] (405K MIDI files, 4.3B tokens after REMI-z tokenization, 2% validation split) and fine-tuning was done with Slakh2100 [Manilow *et al.*, 2019] (1,289 training, 270 validation, 151 test MIDI files), featuring 34 pitched instruments and drums, with ≥5 tracks per piece. Detailed hyperparameter settings are in Appendix B.

## 4.2 Tokenizer Efficiency Comparison

We compare REMI+ and REMI-z through theoretical and empirical analyses. First, we tokenize the Slakh2100 dataset using both methods and evaluate compactness via the **average number of tokens per bar** ($\bar{T}_{\text{bar}}$) and **per note** ($\bar{T}_{\text{note}}$). Lower values indicate reduced computational costs during training and autoregressive decoding.

Next, we quantify token sequence complexity using **bar-level Shannon entropy**. For a discrete random variable $X$ representing token distributions, entropy of a token sequence $H(X)$ is defined as:

$$H(X) = -\sum_{i=1}^{N} P(x_i) \log_2 P(x_i) \quad \text{(bits per token)}, \quad (2)$$

where $\{x_1, \ldots, x_N\}$ are all tokens appeared in the sequence, and $\{P(x_1), \ldots, P(x_N)\}$ are their probabilities. We calculate the entropy for each bar sequence to get $H_{\text{bar}}$, and then average across all bars in the dataset to obtain $\bar{H}_{\text{bar}}$. Lower $\bar{H}_{\text{bar}}$ value indicates less inherent complexity of the bar sequence, and by extension, the multitrack music, reflecting efficient encoding with less redundant information.

Then, we train unconditional LMs on the Slakh2100 dataset and observe the **note-level perplexity**:

$$\text{PPL}_{\text{note}} = \exp\left(-\frac{1}{M}\sum_{j=1}^{M} \log P(n_j \mid n_{1:j-1})\right), \quad (3)$$

where $M$ is the number of notes in a bar, and $P(n_j \mid n_{1:j-1})$ aggregates probabilities for all tokens corresponding to note $n_j$ (instrument, position, pitch, duration). Unlike token-level perplexity ($\text{PPL}_{\text{token}}$), this metric normalizes by the number of notes, ensuring fair comparison across tokenization schemes. Lower note-level perplexity indicates better modeling of musical notes, as it reflects reduced uncertainty in predicting all associated attributes.

Finally, we assess task-specific performance with different tokenizers, discussed in the next section.

## 4.3 Task-Specific Evaluation

This section introduces baseline models and metrics used for evaluating the music arrangement models.

**Baseline Models**

For each task, we compare our model against a state-of-the-art (SOTA) task-specific baseline. For band arrangement, we adopt **Transformer-VAE** from [Zhao *et al.*, 2024], the strongest previously reported model for multitrack arrangement without assumptions on track type or number. It combines Transformer-based long-term and inter-track modeling with a VQ-VAE generation module. For piano reduction, we compare with [Terao *et al.*, 2023] (**UNet**), the most recent work in this area. For drum arrangement, existing drum-specific models (e.g., [Barnabò *et al.*, 2023], [Dahale, 2022]) are unsuitable as they generate drums from a single melody or instrumental track. Instead, we adopt Composer's Assistant 2 (**CA v2**) [Malandro, 2024], a strong track infilling model capable of handling multitrack inputs and generating drum outputs. Baseline's implementation details are in Appendix B.

To demonstrate the impact of generative pre-training, an ablation variant of our model without pre-training (**w/o PT**) was used as a baseline. Additionally, we include rule-based baselines to showcase the difficulty of the tasks. For band arrangement, **Rule-Based** distributes notes evenly by pitch across instruments. For piano reduction, we use **Rule-F** (flattened multitrack where the piano plays all notes) and **Rule-O** (original piano track). Note that Rule-O is not considered ground truth since the piano track is likely to not fulfil the $\geq$ 40% pitch range requirements. For drum arrangement, the original drum track (**Ground Truth**) is included anonymously for human evaluation as a golden standard.

**Segment-level Objective Evaluation**

Objective metrics measure similarity between model outputs and target sequences, assuming closer resemblance indicates higher naturalness and musicality. Given the extreme sparsity of note events in the track-wise piano roll (e.g., only 0.04% non-zero elements in the Slakh2100 dataset under 48th-note quantization), F1-based metrics are more suitable than accuracy-based metrics for similarity evaluation. Following [Malandro, 2024; Terao *et al.*, 2023], we use *note-level F1* to measure similarity between model outputs and target sequences. Specifically, we compute **Note F1** (correct onset and pitch) and **Note**$_i$ **F1** (additional correct instrument prediction), both under 16th-note quantization for fair comparison.

For piano reduction and drum arrangement, the same models are used in objective and subjective evaluations. For band arrangement, models are separately trained without random track deletion to ensure deterministic outputs. Baseline models are also modified by excluding the prior model to remove long-term context hints, ensuring evaluation fairness.

In addition to the modifications described above, we introduce three additional metrics specific to the band arrangement task. First, **Instrument Intersection over Union (I-IoU)** evaluates the accuracy of instrument control. Second, **Voice Word Error Rate (V-WER)** measures the similarity in voice features between the generated output and the reference, indicating the effectiveness of the voice control. Third, **Melody**

**F1 (Mel F1)** calculates Note F1 on the tracks with the highest average pitch in the output and target music, assessing the preservation of melody, which is critical for arrangement fidelity. The detailed computation of these metrics is provided in Appendix C.

Beyond comparing with the SOTA model, the objective evaluation for band arrangement serves two additional purposes: (1) to assess the task-specific performance of different tokenizers and (2) to validate the design of the fine-tuning objective. The band arrangement task is particularly challenging because it imposes no assumptions on the types or numbers of target instruments, making it more representative of the model's overall capabilities compared to the other tasks.

**Song-Level Subjective Evaluation**

To complement similarity metrics and account for perceptual quality and creativity, we conducted human evaluations. Full-piece arrangements were generated by all models and evaluated phrase-by-phrase on a 5-point scale (1: very low, 5: very high). For band arrangement, models were tested across three instrument combinations with different complexity: string trio (3 tracks), rock band (4 tracks), and jazz band (7 tracks). Three metrics were used across band, piano, and drum arrangement tasks: **Coherence**, which evaluates the natural flow of the arrangement and the consistency of each instrument's performance and style throughout the piece; **Creativity**, which assesses the degree of innovation in the arrangement under the constraints of the music's content and style; and **Musicality**, which measures the overall musical appeal and aesthetic quality of the arrangement. Further details on the metrics, ensemble settings, questionnaire, and evaluation process are in Appendix D.

Task-specific metrics were also introduced. For band arrangement, **Faithfulness** measures resemblance to the original in melody and overall feel, while **Instrumentation** assesses the appropriateness of each instrument's role within the band and their harmony. For piano reduction, **Faithfulness** is also adopted but without melody preservation requirements, and **Playability** assesses the feasibility of the generated contents played by human pianists. For drum arrangement, **Compatibility** measures how well the drum track blends with other instruments, and **Phrase Transition** assesses the smoothness of transitions between phrases.

Additionally, significance tests were conducted between our model and the SOTA baselines using within-subject (repeated-measures) ANOVA [Scheffe, 1999]. The significance level are showed on the result tables: $*$ for p-values less than 0.05, † for p-values less than 0.01, and ‡ for p-values less than 0.001.

# 5 Results

## 5.1 Effectiveness of REMI-z

As shown in Table 2, our proposed tokenization method significantly reduces the average sequence length from 225.91 tokens (REMI+) to 151.68 tokens per bar, achieving a 32.9% reduction. The number of tokens per note also decreases from 4.03 (REMI+) to 2.77 with REMI-z. These results demonstrate that our tokenizer generates a more compact representation for multitrack music, which not only reduces the computational complexity for generative modeling but also accelerates autoregressive decoding and potentially facilitates the modeling of long-term dependencies.

| Tokenizer | $\bar{T}_{\text{bar}}$ | $\bar{T}_{\text{note}}$ | $\bar{H}_{\text{bar}}$ | $\text{PPL}_{\text{note}}$ | $\text{PPL}_{\text{token}}$ |
|---|---|---|---|---|---|
| REMI+ | 225.91 | 4.03 | 41.68 | 116.20 | **3.00** |
| REMI-z (Ours) | **151.68** | **2.77** | **29.43** | **84.11** | 4.50 |

Table 2: Tokenizer efficiency comparison.

| Model | I-IOU | V-WER | Note F1 | $\text{Note}_i$ F1 | Mel F1 |
|---|---|---|---|---|---|
| Transformer-VAE | 97.5 | 35.0 | 49.5 | 40.0 | 24.7 |
| Transformer w/ REMI+ | 95.0 | 18.2 | 94.4 | 76.0 | 68.8 |
| Transformer w/ REMI-z | 99.5 | 9.9 | **97.8** | 77.5 | 77.8 |
| + Pre-training (Ours) | *99.8* | **7.6** | *97.5* | **87.0** | **84.5** |
| − voice | 99.6 | 17.6 | 97.2 | *84.3* | *81.5* |
| − history | **100.0** | *9.0* | 97.6 | 77.4 | 79.4 |

Table 3: Objective evaluation results of the band arrangement task.

Moreover, this increased compactness does not introduce additional complexity. In terms of bar-level Shannon entropy, REMI-z achieves 29.43 bits/token compared to REMI+'s 41.68 bits/token, representing a 29.4% relative reduction. This indicates that sequences tokenized with REMI-z contain less redundant information and can be further encoded with a shorter expected code length, requiring simpler encoding rules for lossless compression and potentially enabling more effective generative modeling.

We confirm this advantage with empirical analysis. The LM trained with REMI-z achieves lower note-level perplexity (84.11 vs 116.20), demonstrating that modeling the basic unit of music—the note—is more effective with REMI-z. This improvement in note-level modeling naturally extends to better performance in generating bars and complete compositions. Conversely, while individual REMI+ tokens are easire to model (-1.5 token-level perplexity), this does not translate to better note-level and song-level performance.

## 5.2 Band Arrangement

**Objective Evaluation**

We then evaluate REMI-z's task-specific performance. As shown in Table 3, REMI-z consistently outperforms REMI+ on all metrics. It achieves better instrument control (-4.5% I-IOU), reducing undesired or missing instruments, and more accurate voice control with lower V-WER (-8.3%). This indicates REMI-z enable the model to better follows user-specified instrument constraints. The higher Note F1 and $\text{Note}_i$ F1 scores demonstrate improved reconstruction quality, highlighting REMI-z's effectiveness in the arrangement training. The significantly higher Mel F1 (+9.0%) also indicates better melody preservation and arrangement fidelity.

Pre-training further enhances our model's performance in three major aspects: (1) improved voice control (-2.3% V-WER), indicating better understanding of voice requirement in the instrumental prompts, (2) more accurate per-instrument reconstruction (+9.5% $\text{Note}_i$ F1), suggesting enhanced comprehension of instrumental roles in ensembles, and (3) better

| Model | Fa. | Co. | In. | Cr. | Mu. |
|---|---|---|---|---|---|
| Rule-Based | *3.46* | *3.05* | *2.89* | *3.00* | *3.07* |
| Transformer-VAE | 2.65 | 2.70 | 2.72 | *3.00* | 2.72 |
| Ours | $^{\ddagger}$**3.77** | $^{\ddagger}$**3.47** | $^{\ddagger}$**3.49** | $^{*}$**3.40** | $^{\ddagger}$**3.47** |
| w/o PT | 3.19 | 2.82 | 2.86 | 2.93 | 2.75 |

Table 4: Band arrangement subjective evaluation results. Fa., Co., In., Cr., and Mu. represent Faithfulness, Coherence, Instrumentation, Creativity, and Musicality, respectively.

| Model | F1 | Fa. | Co. | Pl. | Cr. | Mu. |
|---|---|---|---|---|---|---|
| Rule-F | - | **3.93** | *3.59* | 3.14 | *2.96* | *3.34* |
| Rule-O | - | 2.75 | 3.49 | **4.07** | 2.62 | 2.96 |
| UNet | 58.3 | 2.97 | 2.90 | 3.47 | 2.82 | 2.78 |
| Ours | **85.5** | $^{\ddagger}$*3.63* | $^{\ddagger}$**3.64** | $^{*}$*3.86* | $^{*}$**3.14** | $^{\ddagger}$**3.48** |
| w/o PT | *78.4* | 2.25 | 2.58 | 3.29 | 2.67 | 2.26 |

Table 5: Piano reduction results. The Pl. represents Playability score.

melody preservation (+6.7% Mel F1), leading to improved arrangement fidelity. For detailed knowledge probing analysis of how pre-training aligns learned representations with musical properties, please refer to Appendix E.

Our model consistently outperforms Transformer-VAE across all metrics. The gap is particularly significant in Note F1 (97.5% vs 49.5%) and Note$_i$ F1 (87.0% vs 40.0%), indicating Transformer-VAE's limitation in inferring instrument-specific content within musical contexts. Our model also achieves substantially better melody preservation (+59.8% Mel F1), ensuring higher-fidelity song-level arrangements.

Ablation studies validate our design choices for voice and history conditioning. Excluding relative voice information from the instrument condition sequence leads to degraded voice relationships (-10.0% V-WER) and consequently lower Note$_i$ F1 (-2.7%) and Mel F1 (-3.0%). This demonstrates that voice conditioning enables effective voice control and facilitates instrument role inference. Similarly, excluding history from the condition sequence noticeably reduces Note$_i$ F1 (-9.6%) and Mel F1 (-5.1%), laying the foundation for generating coherent outputs in song-level inference.

**Subjective Evaluation**

As shown in Table 4, our model consistently achieves the highest scores across all metrics, demonstrating its capability to generate high-quality arrangements that adhere to the original musical essence while achieving good Coherence, Creativity, and Musicality. Transformer-VAE scores lower than the Rule-Based model, particularly in Faithfulness and Musicality, indicating its limitations in preserving musical essence and expression. This also suggests that the rule-based method can serve as competitive baselines for band arrangement research. Notably, removing pre-training degrades all metrics, highlighting its importance for generation quality.

## 5.3 Piano Reduction

Table 5 presents the piano reduction results. Our model outperforms all baselines in F1 score, Coherence, and Musicality. Unlike Rule-O and Rule-F which excel only at Playability or Faithfulness respectively, our model achieves a balance between these two aspects. Pre-training also proves essential: without it, the model has lower F1 and underperforms UNet from [Terao *et al.*, 2023] across all subjective metrics. In contrast, the model with pre-training significantly surpasses UNet in subjective evaluations. Interestingly, F1 scores do not strictly correlate with perceptual quality, as evidenced by the w/o PT model's higher F1 but lower subjective evaluation scores compared to UNet.

| Model | F1 | Comp. | Co. | Tr. | Cr. | Mu. |
|---|---|---|---|---|---|---|
| Ground Truth | **100.0** | **4.31** | **4.18** | *3.36* | *3.16* | **3.78** |
| CA v2 | 20.3 | 3.82 | *4.05* | 2.86 | 2.58 | 3.19 |
| Ours | *79.3* | *3.91* | 4.03 | $^{\ddagger}$**3.77** | $^{\ddagger}$**3.27** | $^{\dagger}$*3.57* |
| w/o PT | 1.2 | 2.49 | 2.19 | 2.21 | 2.82 | 2.05 |

Table 6: Drum arrangement results. Comp. and Tr. represent Compatibility and Phrase Transition score respectively.

## 5.4 Drum Arrangement

As in Table 6, our model significantly outperforms baseline CA v2 in F1 score. On subjective metrics, our model performs comparably with CA v2 on Coherence, and better at all remaining metrics. Particularly, we notice that CA v2 tends to generate repetitive content throughout the song. Conversely, our model excels in handling phrase transitions and demonstrating creativity, achieving a Musicality score comparable to the ground truth. This indicates that our model effectively captures high-level musical structures during generation, introducing variations between conventional drum patterns and producing more engaging arrangements. The model without pre-training fails to perform meaningfully (1.2% F1), with a notably larger performance gap compared to other arrangement tasks. This demonstrates pre-training's crucial role in sequence-to-sequence tasks when handling longer musical structures, where complex alignments between source and target sequences require enhanced structural understanding.

# 6 Conclusion

In this paper, we proposed a unified framework for automatic music arrangement that addresses challenges in tokenization, pre-trained model utilization, and arrangement quality. Our tokenizer demonstrates superior efficiency and enhances modeling effectiveness across various generation tasks, while our training objective shows compatibility with diverse music arrangement tasks. Further, LM pre-training significantly enhances downstream performance, particularly for tasks requiring complex sequence alignment. Taken together, our method achieves state-of-the-art results on band arrangement, piano reduction, and drum arrangement tasks, generating natural, faithful, coherent, creative, and musically appealing outputs. The performance advantage of our framework suggests broader applications in music-to-music transformation tasks like infilling and call-and-response generation. Furthermore, our tokenizer's compact and efficient design makes it a standalone contribution applicable to general multitrack symbolic music generation tasks beyond the scope of arrangement.

# References


[Barnabò *et al.*, 2023] Giorgio Barnabò, Giovanni Trappolini, Lorenzo Lastilla, Cesare Campagnano, Angela Fan, Fabio Petroni, and Fabrizio Silvestri. Cycledrums: automatic drum arrangement for bass lines using cyclegan. *Discover Artificial Intelligence*, 3(1):4, 2023.

[Brown *et al.*, 2020] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[Choi *et al.*, 2020] Kristy Choi, Curtis Hawthorne, Ian Simon, Monica Dinculescu, and Jesse Engel. Encoding musical style with transformer autoencoders. In *International conference on machine learning*, pages 1899–1908. PMLR, 2020.

[Crestel and Esling, 2016] Léopold Crestel and Philippe Esling. Live orchestral piano, a system for real-time orchestral music generation. *arXiv preprint arXiv:1609.01203*, 2016.

[Dahale, 2022] Rishabh Dahale. *Automatic Drum Accompaniment Generation from Melody*. PhD thesis, Indian Institute of Technology Bombay, 2022.

[Dong *et al.*, 2021] Hao-Wen Dong, Chris Donahue, Taylor Berg-Kirkpatrick, and Julian McAuley. Towards automatic instrumentation by learning to separate parts in symbolic multitrack music. *arXiv preprint arXiv:2107.05916*, 2021.

[Dong *et al.*, 2023] Hao-Wen Dong, Ke Chen, Shlomo Dubnov, Julian McAuley, and Taylor Berg-Kirkpatrick. Multitrack music transformer. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Ens and Pasquier, 2020] Jeffrey Ens and Philippe Pasquier. Mmm : Exploring conditional multi-track music generation with the transformer. *ArXiv*, abs/2008.06048, 2020.

[Gardner *et al.*, 2021] Josh Gardner, Ian Simon, Ethan Manilow, Curtis Hawthorne, and Jesse Engel. Mt3: Multi-task multitrack music transcription. *ArXiv*, abs/2111.03017, 2021.

[Hsiao *et al.*, 2021] Wen-Yi Hsiao, Jen-Yu Liu, Yin-Cheng Yeh, and Yi-Hsuan Yang. Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs. In *AAAI Conference on Artificial Intelligence*, 2021.

[Huang and Yang, 2020] Yu-Siang Huang and Yi-Hsuan Yang. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1180–1188, 2020.

[Huang *et al.*, 2018] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam M. Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer: Generating music with long-term structure. In *International Conference on Learning Representations*, 2018.

[Kikuchi *et al.*, 2016] Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Controlling output length in neural encoder-decoders. *arXiv preprint arXiv:1609.09552*, 2016.

[Lev, 2024] Aleksandr Lev. Los angeles midi dataset: Sota kilo-scale midi dataset for mir and music ai purposes. In *GitHub*, 2024.

[Liu *et al.*, 2023] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[Lu *et al.*, 2023] Peiling Lu, Xin Xu, Chenfei Kang, Botao Yu, Chengyi Xing, Xu Tan, and Jiang Bian. Musecoco: Generating symbolic music from text. *arXiv preprint arXiv:2306.00110*, 2023.

[Madaghiele *et al.*, 2021] Vincenzo Madaghiele, Pasquale Lisena, and Raphaël Troncy. Mingus: Melodic improvisation neural generator using seq2seq. In *ISMIR*, pages 412–419, 2021.

[Makris *et al.*, 2021] Dimos Makris, Kat R Agres, and Dorien Herremans. Generating lead sheets with affect: A novel conditional seq2seq framework. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[Malandro, 2023] Martin E Malandro. Composer's assistant: An interactive transformer for multi-track midi infilling. *arXiv preprint arXiv:2301.12525*, 2023.

[Malandro, 2024] Martin Malandro. Composer's Assistant 2: Interactive Multi-Track MIDI Infilling with Fine-Grained User Control. In *Proc. 25th Int. Society for Music Information Retrieval Conf.*, San Francisco, CA, USA, 2024.

[Manilow *et al.*, 2019] Ethan Manilow, Gordon Wichern, Prem Seetharaman, and Jonathan Le Roux. Cutting music source separation some slakh: A dataset to study the impact of training data quality and quantity. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 45–49. IEEE, 2019.

[Ou *et al.*, 2023] Longshen Ou, Xichu Ma, Min-Yen Kan, and Ye Wang. Songs across borders: Singable and controllable neural lyric translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–467, Toronto, Canada, July 2023. Association for Computational Linguistics.

[Qu *et al.*, 2024] Xingwei Qu, Yuelin Bai, Yinghao Ma, Ziya Zhou, Ka Man Lo, Jiaheng Liu, Ruibin Yuan, Lejun Min, Xueling Liu, Tianyu Zhang, et al. Mupt: A generative symbolic music pretrained transformer. *arXiv preprint arXiv:2404.06393*, 2024.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Ren *et al.*, 2020] Yi Ren, Jinzheng He, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu. Popmag: Pop music accompaniment generation. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1198–1206, 2020.

[Scheffe, 1999] Henry Scheffe. *The analysis of variance*, volume 72. John Wiley & Sons, 1999.

[Sennrich *et al.*, 2015] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*, 2015.

[Terao *et al.*, 2022] Moyu Terao, Yuki Hiramatsu, Ryoto Ishizuka, Yiming Wu, and Kazuyoshi Yoshii. Difficulty-aware neural band-to-piano score arrangement based on note-and statistic-level criteria. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 196–200. IEEE, 2022.

[Terao *et al.*, 2023] Moyu Terao, Eita Nakamura, and Kazuyoshi Yoshii. Neural band-to-piano score arrangement with stepless difficulty control. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Tiedemann and Scherrer, 2017] Jörg Tiedemann and Yves Scherrer. Neural machine translation with extended context. *arXiv preprint arXiv:1708.05943*, 2017.

[von Rütte *et al.*, 2023] Dimitri von Rütte, Luca Biggio, Yannic Kilcher, and Thomas Hofmann. Figaro: Controllable music generation using learned and expert features. In *The Eleventh International Conference on Learning Representations*, 2023.

[Wang *et al.*, 2024] Ziyu Wang, Lejun Min, and Gus Xia. Whole-song hierarchical generation of symbolic music using cascaded diffusion models. *arXiv preprint arXiv:2405.09901*, 2024.

[Wu *et al.*, 2024] Shangda Wu, Yashan Wang, Xiaobing Li, Feng Yu, and Maosong Sun. Melodyt5: A unified score-to-score transformer for symbolic music processing. *arXiv preprint arXiv:2407.02277*, 2024.

[Yi *et al.*, 2022] Li Yi, Haochen Hu, Jingwei Zhao, and Gus Xia. Accomontage2: A complete harmonization and accompaniment arrangement system. *arXiv preprint arXiv:2209.00353*, 2022.

[Zeng *et al.*, 2021] Mingliang Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, and Tie-Yan Liu. Musicbert: Symbolic music understanding with large-scale pre-training. *ArXiv*, abs/2106.05630, 2021.

[Zhang *et al.*, 2022] Xueyao Zhang, Jinchao Zhang, Yao Qiu, Li Wang, and Jie Zhou. Structure-enhanced pop music generation via harmony-aware learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1204–1213, 2022.

[Zhao *et al.*, 2023] Jingwei Zhao, Gus Xia, and Ye Wang. Q&a: query-based representation learning for multi-track symbolic music re-arrangement. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5878–5886, 2023.

[Zhao *et al.*, 2024] Jingwei Zhao, Gus Xia, Ziyu Wang, and Ye Wang. Structured Multi-Track Accompaniment Arrangement via Style Prior Modelling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.