CSL-L2M: Controllable Song-Level Lyric-to-Melody Generation Based on Conditional Transformer with Fine-Grained Lyric and Musical Controls

Li Chai, Donglin Wang*

Westlake University
{chaili, wangdonglin}@westlake.edu.cn

Abstract

Lyric-to-melody generation is a highly challenging task in the field of AI music generation. Due to the difficulty of learning strict yet weak correlations between lyrics and melodies, previous methods have suffered from weak controllability, low-quality and poorly structured generation. To address these challenges, we propose CSL-L2M, a controllable song-level lyric-to-melody generation method based on an in-attention Transformer decoder with fine-grained lyric and musical controls, which is able to generate full-song melodies matched with the given lyrics and user-specified musical attributes. Specifically, we first introduce REMI-Aligned, a novel music representation that incorporates strict syllable- and sentence-level alignments between lyrics and melodies, facilitating precise alignment modeling. Subsequently, sentence-level semantic lyric embeddings independently extracted from a sentence-wise Transformer encoder are combined with word-level part-of-speech embeddings and syllable-level tone embeddings as fine-grained controls to enhance the controllability of lyrics over melody generation. Then we introduce human-labeled musical tags, sentence-level statistical musical attributes, and learned musical features extracted from a pre-trained VQ-VAE as coarse-grained, fine-grained and high-fidelity controls, respectively, to the generation process, thereby enabling user control over melody generation. Finally, an in-attention Transformer decoder technique is leveraged to exert fine-grained control over the full-song melody generation with the aforementioned lyric and musical conditions. Experimental results demonstrate that our proposed CSL-L2M outperforms the state-of-the-art models, generating melodies with higher quality, better controllability and enhanced structure.

Demos — https://lichaiustc.github.io/CSL-L2M/
Code — https://github.com/LiChaiUSTC/CSL-L2M

Introduction

Deep learning techniques have been increasingly applied to various music generation tasks (Duan, Yu, and Oyama 2024; Hahn et al. 2023; Yu, Srivastava, and Canales 2021; Tian et al. 2023). Lyric-to-melody generation, one of the most essential and common tasks in songwriting, has attracted growing interest from both academia and industry. A high-quality lyric-to-melody generation is required to generate melodies not only following good musical patterns but also aligning with the given lyrics. Due to the scarcity of paired lyric-melody data with alignment information and the difficulty of learning the strict but weak correlations between lyrics and melodies, this task remains under-explored.

Many deep learning methods have been explored for lyric-to-melody generation. A sequence-to-sequence based melody composition model is proposed in (Bao et al. 2019) , which is the first work to use an end-to-end network model to generate melodies from lyrics. Subsequently, Yu (Yu, Srivastava, and Canales 2021) proposes a conditional LSTM-GAN generative model for melody generation from lyrics. In (Srivastava et al. 2022), a novel architecture, three branch conditional LSTM-GAN is proposed to further improve generation quality. However, the direct mapping from lyrics to melodies is difficult to learn because they are weakly correlated (e.g., a melody can correspond to many different lyrics and vise versa.). Accordingly, these end-to-end generation methods suffer from low generation quality due to the limited available parallel lyric-melody data. To this end, an unsupervised method is proposed in (Sheng et al. 2021), which performs self-supervised masked sequence to sequence pre-training on large amount of unpaired lyric and melody data. In addition, a two-stage generation method with music template is proposed in (Ju et al. 2021), which is data efficient and addresses the issues of limited paired data to some extent. With the tremendous success of large language models (LLMs) (Touvron et al. 2023), more recently, the work in (Ding et al. 2024) attempts to leverage the capability of LLMs to model the lyric-melody relationship.

Currently, the lyric-to-melody generation methods including those mentioned above focus on generating short melodies from lyrics typically consisting of one sentence or a few sentences, where a full-song melody is usually composed by simply concatenating these sentence-level melodies resulting in incoherent musical structure without both repetition patterns and distinguishable verse-chorus structure. In addition, controllability is a crucial aspect of the lyric-to-melody generation task, which allows users to interact with the generation process to create their expected melodies. Nevertheless, only a few lyric-to-melody works have explored the controllability. In (Ju et al. 2021), the gen-

*Corresponding author.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

erated melodies can be controlled by adjusting the musical elements in music templates including tonality and chord progression. A reference style embedding technique is proposed in (Zhang, Yu, and Takasu 2023) to achieve the control over the style of generated melodies. The research of (Duan et al. 2022) enables users to interact with the generation process and recreate music by selecting from recommended musical attributes. However, these works only provide a few coarse-grained musical attribute controls. One more thing, since one syllable may correspond to one or more notes, the alignment between the given lyrics and corresponding melodies could be "one-to-one "or "one-to-many ". Most of previous methods only consider the "one-to-one "alignment, which introduces bias into the melody composition.

To address the aforementioned issues, we propose a controllable song-level lyric-to-melody generation method called CSL-L2M, which is capable of generating melodies aligning with lyrics and user-specified musical attributes at the full-song level. Specifically, we first introduce a novel music representation called REMI-Aligned. This representation incorporates strict syllable- and sentence-level lyric-melody alignments, which makes both exact and "one-to-many "alignment learning feasible. Inspired by (Wu and Yang 2023), which equips conditional Transformer with the capability to model long sequences under fine-grained time-varying conditions through in-attention, we integrate the in-attention technique into our CSL-L2M model. Multiple multi-granularity lyric controls (including sentence-level semantic embeddings, word-level part-of-speech (POS) embeddings, and syllable-level tone embeddings) and musical controls (including coarse-level human-labeled musical tags, sentence-level statistical musical attributes, and learned high-fidelity musical features (von Rütte et al. 2023) from a pre-trained Vector Quantized-Variational AutoEncoder (VQ-VAE)) are extracted and fed into the conditional Transformer decoder through in-attention to realize tight fine-grained control of lyrics and musical attributes over melody generation process. This enables the generation of high-quality melodies from lyrics, precisely tailored to the user's desired musical style. Moreover, the musical controls not only enable user-controllable generation but also provide the model with additional musical information that is beneficial for melody modeling. Experiments conducted on our 10,170 Chinese pop songs demonstrate that CSL-L2M could generate melodies that are both well-matched with the lyrics and consistent with the user-specified musical attributes. Compared to the state-of-the-art lyric-to-melody generation methods, CSL-L2M generates melodies with higher quality, better controllability and enhanced structure.

Related Work

Lyric-to-Melody Generation The development of lyric-to-melody generation has evolved from traditional rule-based (Nichols 2009; Monteith, Martinez, and Ventura 2012) and statistical methods (Long, Wong, and Sze 2013) to deep learning methods. The traditional methods usually rely on specific hand-designed musical rules and suffer from low generation quality. Currently, the end-to-end deep generative models are the mainstream methods but they suffer from several challenges: 1) weak correlations between lyrics and melodies are difficult to capture by the network models, where much paired training data with alignment information is required; 2) strict alignment between each syllable in the given lyric and note in the corresponding melody is required, which needs additional alignment modeling. As for the first challenge, limited available paired lyric-melody data affects generation quality. The end-to-end models which directly learn the mapping from lyrics to melodies with the limited paired data often lead to poor generation quality. To this end, SongMASS (Sheng et al. 2021) improves the generation performance of end-to-end models by leveraging self-supervised pre-training on much unpaired lyric and melody data. Furthermore, TeleMelody (Ju et al. 2021), a two-stage generation pipeline based on musical templates, is proposed to enhance data efficiency and further improve generation performance. In addition, ReLyMe (Zhang et al. 2022) introduces several principles of lyric-melody relationships from music theory into the decoding process, enhancing the harmony between lyrics and melodies. However, these methods fail to exploit melody-related lyric information and additional musical information for tightly controlling over the melody generation. Consequently, they are unable to adequately capture the intricate relationships between lyrics and melodies, resulting in limited generation quality. Moreover, few of them generate melodies at the full-song level, causing poor musical structure. As for the second challenge, most existing works either focus solely on the "one-to-one "lyric-melody alignment or do not ensure precise alignment, which can easily degrade generation quality.

Controllable Music Generation Controllability in music generation aims to provide user control over the process in a desired direction (Briot and Pachet 2020). According to the levels of controllability, it can be divided into global/coarse-grained control and fine-grained control. The former refers to the fact that generation process is guided by time-invariant controls. Instead, the later refers to the fact that the generation process is guided by time-varying controls, which can provide more flexible and precise control, especially in the generation of long sequences. Controllable music generation has attracted increasing research interest. In (Dong et al. 2018; Yang, Chou, and Yang 2017; Neves, Fornari, and Florindo 2022), global conditions are injected into the training procedure of generative adversarial networks. Some works (Payne 2019; Sarmento et al. 2023) achieve global control through conditional Transformer models with prompt-based control tokens. Many methods based on VAE enable users to exert global control by manipulating latent conditioning vectors (Brunner et al. 2018; Roberts et al. 2018; Tan and Herremans 2020). Transformer autoencoders are used in (Choi et al. 2020) to realize improved control by learning global performance representations. However, these global controls often become less effective during long sequence generation, as the model may forget or weaken the global conditions over time. In contrast, MuseMorphose (Wu and Yang 2023) and FIGARO (von Rütte et al. 2023) introduce fine-grained control, where the for-

mer is realized through one Transformer VAE based on an in-attention conditioning technique and the later is achieved through description-to-sequence learning. Existing research on lyric-to-melody generation rarely pays attention to controllability. Only a few works have explored this area and do not offer fine-grained and flexible control. In this paper, we delve into the controllability of lyric-to-melody generation.

Methodology

To overcome the difficulty of learning strict yet weak correlations between lyrics and melodies and enable user controls over full-song melody generation, we propose a controllable song-level lyric-to-melody generation method called CSL-L2M, as shown in Figure 2. This method is capable of generating full-song melodies that match the given lyrics and adhere to user-specified musical attributes. We achieve this by employing the in-attention technique, as proposed in (Wu and Yang 2023), to tightly control the conditional autoregressive Transformer decoder's generation process under multiple multi-granularity lyric and musical conditions.

Technical Background

The unconditional Transformer decoder's autoregressive generation process can be formulated as $p(x_t|x_{<t})$, where x_t is the element of a sequence to predict at timestep t, and $x_{<t}$ represents all previously generated elements of the sequence. If a global condition vector $\boldsymbol{c}$ is offered to the model, the modeling could be formulated as $p(x_t|x_{<t}, \boldsymbol{c})$. However, the global control tends to lose its effectiveness during long sequence generation. It is needed to incorporate fine-grained control mechanisms. Assuming that the target sequence consists of N segments and each timestep index $t \in [1, T]$ belongs to one of the N sets of indices $I_1, I_2, ..., I_N$, where $I_n \cap I_{n'} = \varnothing$ for $n \neq n'$ and $\bigcup_{n=1}^{N} I_n = [1, T]$, the fine-grained control is achieved by providing the generation model with each segment-level condition vector $\boldsymbol{c}_n$ during the corresponding time interval I_n, formulated as:

$$p(x_t|x_{<t}; \boldsymbol{c}_n), \quad \text{for } t \in I_n, \tag{1}$$

where the time-varying condition vectors $\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_N$ provide a high-level blueprint of the sequence to model. This could be helpful for long sequence generation, particularly for full-song music generation.

There are many ways to condition autoregressive Transformer decoders at fine-grained level, where the in-attention conditioning (Wu and Yang 2023) offers tight control. Specifically, the in-attention method projects the segment-level condition vector $\boldsymbol{c}_n$ to the same space as the self-attention hidden stats via

$$\boldsymbol{e}_n^\top = \boldsymbol{c}_n^\top W_{in}, \quad W_{in} \in \mathbb{R}^{d_c \times d}. \tag{2}$$

Then the hidden condition state $\boldsymbol{e}_n$ is added to each hidden state of all the self-attention layers to obtain the input to the subsequent layer, formulated as:

$$\begin{aligned} \tilde{\boldsymbol{h}}_t^l &= \boldsymbol{h}_t^l + \boldsymbol{e}_n, \quad \forall l \in \{0, ..., L-1\}; \\ \boldsymbol{h}_t^{l+1} &= \text{SelfAttention}(\tilde{\boldsymbol{h}}_t^l), \end{aligned} \tag{3}$$

which serves as a frequent reminder of the segment-level conditions for the Transformer decoder, thereby achieving tight control over the generation process.

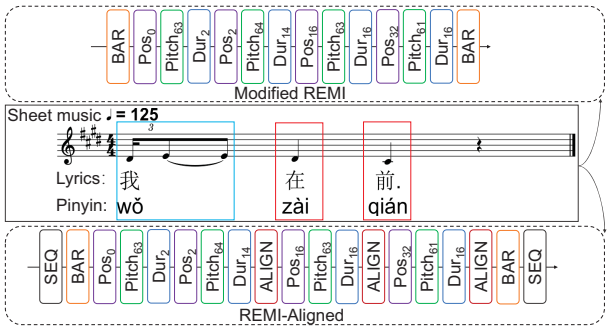


Figure 1: Two representations of the same music piece.

REMI-Aligned Representation

To apply neural sequence models to symbolic music generation tasks, it is necessary to first convert a musical piece into a time-ordered sequence of discrete tokens. There are several ways to implement the conversion, leading to different sequence representations of the same music piece. One prevalent music representation is based on revamped MIDI-derived events (REMI) (Huang and Yang 2020). In REMI, a musical piece is represented as a time-ordered sequence of event tokens including bar, position, pitch, duration, velocity, tempo and chord.

To precisely model the strict alignments between lyrics and melodies, we propose incorporating both sentence-level and syllable-level alignments into the music representation to enable explicit learning. Accordingly, we extend REMI to form a REMI-Aligned music representation by adding these alignments, while discarding tempo, chord and note velocity tokens given the fixed tempo in our dataset, irrelevance of note velocity to our task, and potential issues with chord accuracy. Furthermore, since 64th notes are the shortest notes in our dataset, we improve the temporal resolution of note position from 4 to 16 sub-beats per quarter note, enabling precise quantization of each note in our fixed 4/4 time signature dataset. Examples of a music sequence encoded in modified REMI and REMI-Aligned are shown in Figure 1.

Model Architecture

Figure 2 illustrates the architecture of our proposed CSL-L2M, consisting of a sentence-wise bidirectional Transformer encoder and an autoregressive Transformer decoder equipped with the capability to model full-song melodies under both lyric and musical multi-granularity controls.

Lyric Controls In CSL-L2M, we fully utilize the lyric information related to melodies in training, which improves the model capability to capture the correlations between lyrics and melodies. Specifically, we extract syllable-level tone embeddings, word-level POS embeddings, and sentence-level semantic embeddings serving as time-varying conditions at different granularities to exert fine-grained controls over the Transformer decoder's generation via the in-attention technique.

1) Tone: Tone [1], in tonal languages, refers to the pitch

[1] https://en.wikipedia.org/wiki/Tone

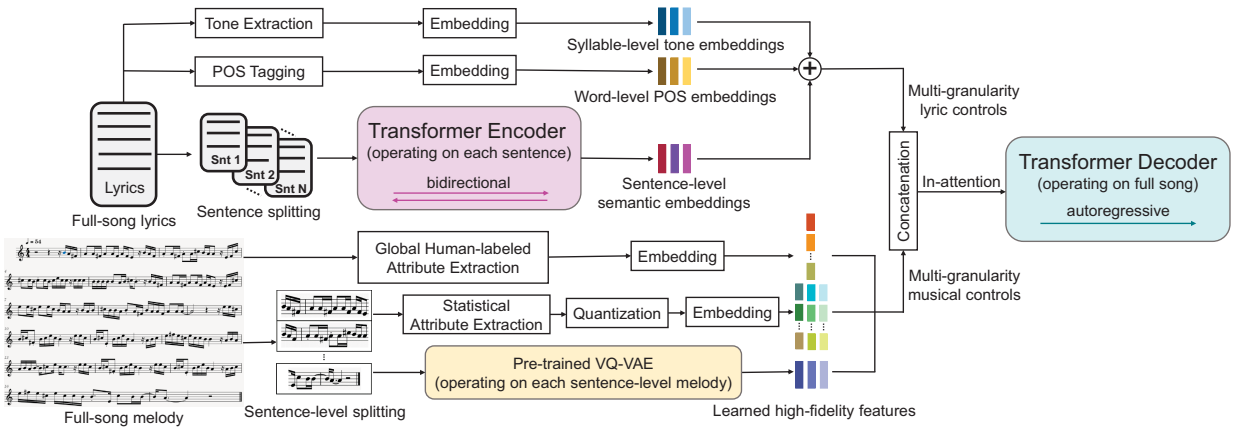


Figure 2: Architecture of CSL-L2M.

variations that help distinguish words with the same spelling but different meanings, playing a crucial role in minimizing semantic ambiguities. Around 60% of languages have tone. In contrast to English, Chinese is a tonal language containing four main tones and one light tone in its characters. The pitch flow of a melody in Chinese songs is usually closely related to the tones of the corresponding lyrics. Accordingly, we incorporate the tone information into our model training to help learn pitch flow of the generated melodies to match with the given lyrics. The s^{th} syllable-level tone attribute c_s^{tone} of each song is converted to an embedding vector $\boldsymbol{c}_s^{\text{tone}} = \mathbf{Emb}^{\text{tone}}(c_s^{\text{tone}})$, before being fed into the decoder as a syllable-varying condition.

2) POS: POS contains potential information of prosodic boundaries between words, which is helpful for enhancing rhythms and structures of generated melodies. Consequently, the POS information is utilized for our model training. Specifically, we first perform POS tagging on the given lyrics by Jieba [2], an open-source tool that supports 56 tags commonly used in Chinese. Then the p^{th} word-level POS attribute c_p^{POS} of each song is transformed into an embedding vector $\boldsymbol{c}_p^{\text{POS}} = \mathbf{Emb}^{\text{POS}}(c_p^{\text{POS}})$, before being fed into the decoder as a word-varying condition.

3) Semantic Embeddings: Since most previous works divide the input lyrics into sentences and then compose each piece of melody from the sentences one by one, we set the granularity of lyric text conditions to a sentence. We employ a bidirectional Transformer encoder (Vaswani et al. 2017) to learn to extract sentence-level latent semantic embeddings of the given lyrics, which is jointly trained with the Transformer decoder using the negative log-likelihood (NLL) training objective. More concretely, the input lyrics of each song are divided into sentences, formulated as $I = \{I_1, I_2, ..., I_N\}$, where I_n is the n^{th} sentence of the lyrics. Then the Transformer encoder encodes these sentences in parallel. We treat the encoder's attention output at the first timestep (corresponding to the SEQ token of the sentence sequence tokens), i.e., $\boldsymbol{h}_{n,1}^{L_{\text{Enc}}}$, as the contextualized representation of the sequence. Finally, it is performed an affine transformation via a learnable weight W to obtain the semantic embedding. These operations can be summarized as follows:

$$\begin{aligned} \boldsymbol{h}_{n,1}^{L_{\text{Enc}}} &= \mathbf{Enc}(I_n) \quad \text{for } 1 \le n \le N; \\ \boldsymbol{z}_n^{\text{sem}} &= \boldsymbol{h}_{n,1}^{L_{\text{Enc}}\top} W, \quad W \in \mathbb{R}^{d \times d_l}, \end{aligned} \tag{4}$$

where $\boldsymbol{z}_n^{\text{sem}}$ is the semantic embedding for the n^{th} sentence.

Musical Controls To enable user control over the melody generation, we introduce human-labeled musical tags, sentence-level statistical musical attributes, and learned latent musical representations extracted from a pre-trained VQ-VAE serving as coarse-grained, fine-grained and high-fidelity controls, respectively, to the generation process.

1) Human-Labeled Musical Tags: We offer a high-quality, precisely annotated parallel lyric-melody dataset with alignment information consisting of 10,170 Chinese pop songs with time signature 4/4. Moreover, tags of key [3], emotion, and structure [4] of each song are meticulously annotated, encompassing 24 [5], 3 [6], and 5 [7] distinct categories respectively. The three types of musical tags serving as coarse-grained conditions are introduced into the decoder to realize human-interpretable control over the melody generation. Specifically, key is highly correlated with pitch distribution of the entire melody. The key of each song c^{key} is transformed into an embedding vector by an embedding layer, i.e., $\boldsymbol{c}^{\text{key}} = \mathbf{Emb}^{\text{key}}(c^{\text{key}})$ and then fed into the decoder as a global condition. Similarly, the emotion of each song c^{emot} is transformed into an embedding vector $\boldsymbol{c}^{\text{emot}} = \mathbf{Emb}^{\text{emot}}(c^{\text{emot}})$ and then fed into the decoder as a global condition. The verse-chorus form, serving as the cornerstone of pop songs, comprises two core sections—a verse and a chorus—that typically contrast melodically, rhythmically, harmonically and dynamically. We convert the u^{th}

[2] https://github.com/fxsjy/jieba

[3] https://en.wikipedia.org/wiki/Key_(music)

[4] https://en.wikipedia.org/wiki/Song_structure

[5] 12 major keys: C, D♭, D, E♭, E, F, F♯, G, A♭, A, B♭, B; 12 minor keys: c, c♯, d, d♯, e, f, f♯, g, g♯, a, b♭, b.

[6] 3 emotions: Neutral, Positive, Negative.

[7] 5 structure sections: Verse, Chorus, Insertion, Bridge, Outro.

structure-level attribute of each song into an embedding vector $\boldsymbol{c}_u^{\text{struc}} = \mathbf{Emb}^{\text{struc}}(c_u^{\text{struc}})$, and then fed it into the decoder as a structure-varying condition.

2) Statistical Musical Attributes: To enable fine-grained user controls over melody generation and help the model better learn correlations between lyrics and melodies, we introduce sentence-level statistical musical attributes. Their granularity is set to a sentence instead of a bar to maintain consistency with the semantic lyric embedding control granularity. We utilize 12 types of statistical musical attributes: pitch mean (PM), pitch variance (PV), pitch range (PR), direction of melodic motion (DMM), amount of arpeggiation (AA), chromatic motion (CM), duration mean (DM), duration variance (DV), duration range (DR), prevalence of most common note duration (MCD), note density (ND), fraction of syllables in lyrics to notes in the corresponding melodies (Align) [8]. They are calculated for each sentence-level melody sequence. We first quantize these attributes into K classes with approximately equal sample sizes, where $K = 64$ in our work. Then the n^{th} sentence-level attributes of the 12 statistical musical attributes for each song are converted into embedding vectors $\boldsymbol{c}_n^{\text{PM}}$, $\boldsymbol{c}_n^{\text{PV}}$, $\boldsymbol{c}_n^{\text{PR}}$, $\boldsymbol{c}_n^{\text{DMM}}$, $\boldsymbol{c}_n^{\text{AA}}$, $\boldsymbol{c}_n^{\text{CM}}$, $\boldsymbol{c}_n^{\text{DM}}$, $\boldsymbol{c}_n^{\text{DV}}$, $\boldsymbol{c}_n^{\text{DR}}$, $\boldsymbol{c}_n^{\text{MCD}}$, $\boldsymbol{c}_n^{\text{ND}}$, $\boldsymbol{c}_n^{\text{Align}}$, respectively and fed into the decoder. These controls can be grouped into four categories, i.e., pitch-related controls (pitch Ctls=Concat($\boldsymbol{c}^{\text{PM}}; \boldsymbol{c}^{\text{PV}}; \boldsymbol{c}^{\text{PR}}; \boldsymbol{c}^{\text{DMM}}; \boldsymbol{c}^{\text{AA}}; \boldsymbol{c}^{\text{CM}}$)), duration-related controls (Dur Ctls=Concat($\boldsymbol{c}^{\text{DM}}; \boldsymbol{c}^{\text{DV}}; \boldsymbol{c}^{\text{DR}}; \boldsymbol{c}^{\text{MCD}}$)), rhythm-related controls ($\boldsymbol{c}^{\text{ND}}$), and note-number-related controls ($\boldsymbol{c}^{\text{Align}}$).

3) Learned Musical Features: Inspired by (von Rütte et al. 2023), we introduce learned musical features extracted from the latent space of a pre-trained VQ-VAE model to provide high-fidelity information to the decoder. This helps to alleviate non-injectivity problem in the lyric-to-melody generation task. The VQ-VAE model consists of a Transformer encoder, a Transformer decoder, and a vector quantization. For training, first, the full-song melody of each song is split into sentence-level melody sequences $X = \{X_1, X_2, ..., X_N\}$, where X_n is the n^{th} sentence-level melody sequence and tokenized by REMI-Aligned. Then the Transformer encoder maps these sequences to the latent space in parallel. The encoder's attention output at the first timestep (corresponding to the SEQ token of the sentence-level melody sequence tokens) is considered as contextualized representation of the sequence. Finally, the quantized latent representations are obtained via the vector quantization and then fed into the decoder through in-attention to reconstruct the original full-song melody. Note that the hyperparameters in the vector quantization, i.e., latent group and codebook sizes, are set to 64 and 2048 respectively in our work. Thus the sentence-level quantized latent representations $\boldsymbol{z}_n^{\text{learned}}$ extracted from the pre-trained VQ-VAE serving as learned musical features are introduce into our CSL-L2M training to provide high-fidelity information.

Feeding Multi-Granularity Controls into the Transformer Decoder Borrowing the in-attention technique from (Wu and Yang 2023), which conditions Transformer decoders with time-varying conditions during long sequence generation, we feed aforementioned multi-granularity controls into our Transformer decoder to achieve firm control over the full-song melody generation. Specifically, the word-level POS embeddings and sentence-level semantic embeddings are expanded to the syllable level by replication. Then they are added to the tone embeddings to get the syllable-level lyric-related controls $\boldsymbol{c}_s^{\text{lyric}} = \boldsymbol{c}_s^{\text{tone}} + \boldsymbol{c}_s^{\text{POS}} + \boldsymbol{z}_s^{\text{sem}}$. Next, aforementioned multi-granularity musical controls are expanded to the syllable level by replication according to the alignment information between lyrics and melodies. Finally, these syllable-level lyric and musical controls are fed into the decoder through in-attention after concatenation, i.e.,

$$\begin{aligned}\boldsymbol{c}_s = \text{concat}([&\boldsymbol{c}_s^{\text{lyric}}; \boldsymbol{c}_s^{\text{key}}; \boldsymbol{c}_s^{\text{emot}}; \boldsymbol{c}_s^{\text{struc}}; \boldsymbol{c}_s^{\text{PM}}; \boldsymbol{c}_s^{\text{PV}}; \boldsymbol{c}_s^{\text{PR}}; \boldsymbol{c}_s^{\text{DMM}}; \\ &\boldsymbol{c}_s^{\text{AA}}; \boldsymbol{c}_s^{\text{CM}}; \boldsymbol{c}_s^{\text{DM}}; \boldsymbol{c}_s^{\text{DV}}; \boldsymbol{c}_s^{\text{DR}}; \boldsymbol{c}_s^{\text{MCD}}; \boldsymbol{c}_s^{\text{ND}}; \boldsymbol{c}_s^{\text{Align}}; \boldsymbol{z}_s^{\text{learned}}]), \\ \boldsymbol{y}_t = \mathbf{Dec}&(x_{<t}; \boldsymbol{c}_s).\end{aligned} \tag{5}$$

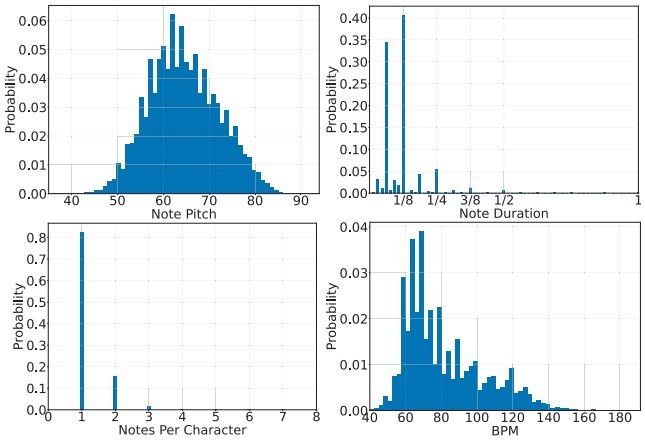


Figure 3: Distributions of music attributes in our paired lyric-melody dataset.

Experiments

Experimental Settings

Dataset The collection of paired lyric-melody data is difficult due to the need for precise synchronization between lyrics and melodies as shown in the sheet music in Figure 1, which requires detailed annotation and specific expertise. Currently, available paired lyric-melody dataset with alignment information is limited and of insufficient quality. To this end, we offer a high-quality, precisely annotated parallel lyric-melody dataset, encompassing 10,170 Chinese pop songs with time signature 4/4. Moreover, musical tags for each song including key, lyric emotion, song structure, and beats per minute (BPM) are precisely annotated. We perform some statistics on this dataset shown in Figure 3. The following observations are made: 1) the most pitch/MIDI numbers fall within the range of 50 to 80; 2) in contrast to melodies in the English dataset (Yu, Srivastava, and Canales 2021), the melodies in our Chinese dataset feature a predominance of short musical notes, specifically 8th and 16th notes; 3)

[8] Details of these musical attributes are available at https://lichaiustc.github.io/CSL-L2M/

more than 80% of characters/syllables correspond to a single musical note (i.e. “one-to-one ”alignment), and nearly 20% of characters correspond to two or more notes (i.e. “one-to-many ”alignment); 4) the BPM of most songs falls within the range of 60 to 120. The 10,170 Chinese pop songs are split into the training, validation, and test sets in an 9:0.5:0.5 ratio for our experiments.

Implementation Details Both the encoder and decoder of our CSL-L2M and VQ-VAE models consist of 12 self-attention layers with 8 self-attention heads, 512 hidden size and 2048 feed-forward dimension. The dimension of each lyric attribute embedding as well as learned musical feature is 128. And the dimension of each human-annotated and statistical musical attribute embedding is 32. The models are trained with Adam optimizer and teacher forcing. We use linear warm-up to increase the learning rate to 10^{-4} in the first 200 steps, followed by a 150k-step cosine decay down to 5×10^{-6}. The batch size is set to 4. During inference, nucleus sampling (Holtzman et al. 2020) is used to sample from the decoder output distribution at each timestep with a softmax temperature $\tau = 1.2$ and truncating the distribution at cumulative probability $p = 0.9$.

Evaluation Metrics Unlike previous works that evaluate generated melodies from lyrics at the sentence level, we conduct evaluations on full-song melodies.

1) Objective Metrics: Objective evaluations are conducted on our test set comprising around 500 songs from our 10,170 Chinese pop songs. We focus on assessing the similarity between the generated and the ground-truth melodies. The following objective metrics proposed by (Sheng et al. 2021) are adopted: 1) Pitch Distribution Similarity (PD); 2) Duration Distribution Similarity (DD); 3) Melody Distance (MD).

2) Subjective Metrics: Subjective evaluations are conducted on 10 songs randomly selected from our test set. We invite 70 participants (including 50 amateurs and 20 professionals) to score the melody properties using a scale from 1 (Poor) to 5 (Perfect). The following subjective metrics are considered: 1) Harmony: Is the melody itself harmonious as well as harmonized with the lyrics ? 2) Rhythm: Does the rhythm sound natural and match the rhythm of the lyrics? 3) Structure: How well does the melody structure match lyric structure? Specifically, whether lyrics with similar rhythm patterns have similar melodies? Does the melody feature a distinguishable verse-chorus structure? Are the transitions between contiguous phrases natural and coherent? 4) Emotion: Does the melody convey a consistent emotion with the lyrics? 5) Quality: What is the overall quality of the melody?

Experimental Results

Main Results Since learned musical features need to be extracted from existing melodies, unless otherwise specified, our reference to CSL-L2M refers to the version without learned musical controls. The evaluation of the full version will be conducted later in the context of style transfer and controllable generation. We first compare our CSL-L2M with two state-of-the-art models, i.e. TeleMelody (Ju et al. 2021) and SongComposer (Ding et al. 2024). As shown in Table 1, CSL-L2M significantly outperforms advanced

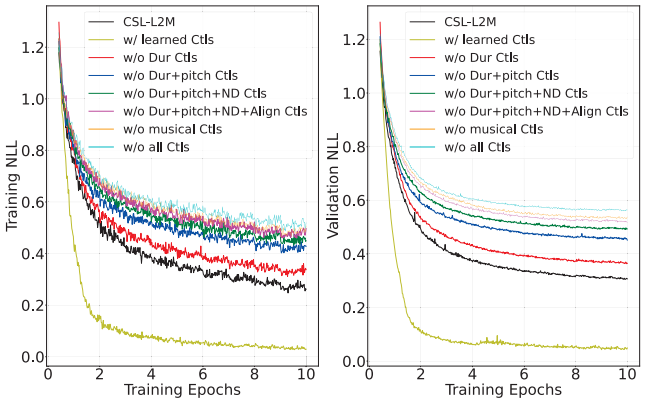


Figure 4: Training dynamics of our CSL-L2M under different controls.

models, namely SongComposer and TeleMelody, in both objective and subjective evaluations, demonstrating the effectiveness of CSL-L2M in generating high-quality song-level melodies from lyrics. We further perform ablation study to verify the effectiveness of lyric and musical controls in CSL-L2M. As illustrated in Table 2 and Figure 4, by successively removing duration-related controls, pitch-related controls, note density and alignment controls, and human-annotated musical attribute controls, we observe a continuous performance degradation. This indicates that the musical controls, in addition to enabling user-controlled generation, can help conditional Transformer in modeling melodies because they offer the model more musical information for reference. Besides, the learned musical features include high-fidelity information of melodies, which aids in reducing non-injectivity of the generation model. As a result, we find that CSL-L2M equipped with learned musical controls achieves top performance that nearly reaches the ceiling. Moreover, CSL-L2M with only lyric controls exceeds the performance of the two state-of-the-art models. This confirms the effectiveness of our designed fine-grained lyric controls and the lyric-to-melody generation framework based on conditional Transformer with the in-attention conditioning mechanism.

Controllability Study Given that our statistical musical controls are ordinal by nature, following (Kawai, Esling, and Harada 2020) and (Wu and Yang 2023), we use the Spearman’s rank correlation coefficient ρ to quantitatively assess the strength of statistical musical attribute control. To simultaneously evaluate the impact on other unrelated attributes when transferring a specific attribute, we calculate Spearman’s rank correlation coefficient matrix between the user-specified attribute classes and attribute raw scores derived from the generated melodies. Results in Figure 5 reveal the strong and independent controllability strengths of CSL-L2M in attribute control. Specifically, for example, $\rho_{\text{PM}} = 0.98$ denotes a strong and positive correlation between the user-specified attribute class c^{PM} and the attribute raw class $\hat{c}^{\text{PM}}$ computed from the generated melodies, which demonstrates strong controllability of the pitch mean attribute. In contrast, $\rho_{\text{PM|Align}} = 0.09$ is the correlation co-

Model	Objective			Subjective				
	PD(%)↑	DD(%)↑	MD↓	Harmony↑	Rhythm↑	Structure↑	Emotion↑	Quality↑
SongComposer	33.37	44.98	3.12	2.32	2.56	2.33	2.51	2.47
TeleMelody	40.02	49.82	2.93	2.71	2.90	2.45	2.42	2.72
CSL-L2M	86.35	93.50	1.27	3.74	4.03	4.20	3.86	3.94

Table 1: Objective and subjective evaluation results of our CSL-L2M and compared models.

	PD(%)↑	DD(%)↑	MD↓
CSL-L2M	86.35	93.50	1.27
+ w/ learned Ctls	97.98	98.62	0.25
- w/o Dur Ctls	85.82	86.41	1.37
- w/o Dur+pitch Ctls	66.07	85.65	1.78
- w/o Dur+pitch+ND+Align Ctls	63.18	63.59	2.03
- w/o musical Ctls	49.20	59.13	2.26

Table 2: Objective evaluation results of our CSL-L2M under different controls.

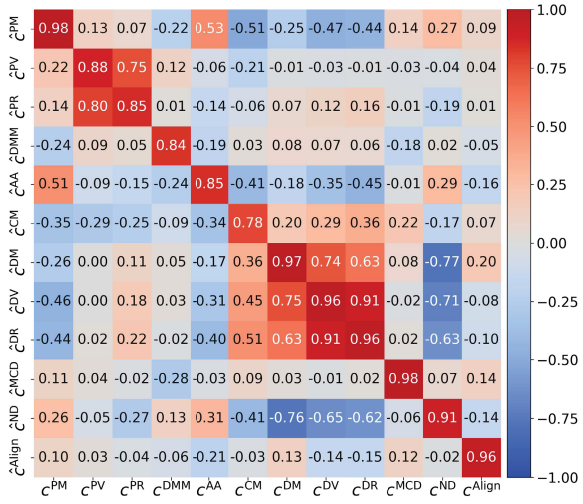
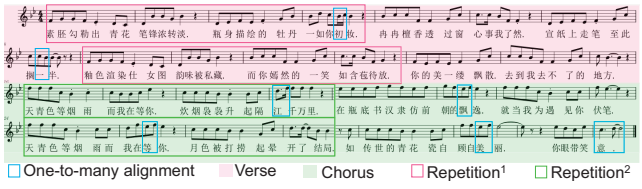


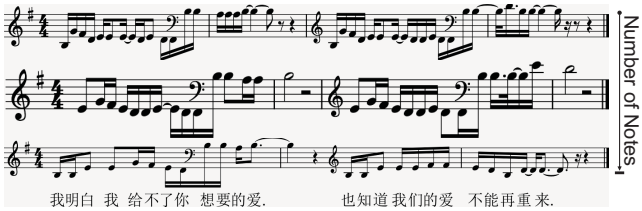
Figure 5: Spearman's rank correlation coefficients between user-specified attribute classes and the attribute raw classes computed from the generated melodies.

efficient between the user-specified alignment attribute class c^{Align} and the unrelated attribute class $\hat{c}^{\text{PM}}$ computed from the generated melodies, revealing the independent controllability of attributes in the multi-attribute scenario.

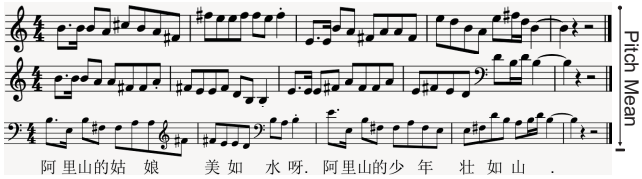
Case Study In Figure 6, we present some generated sheet music given lyrics to demonstrate the advantages of our CSL-L2M in terms of generation quality and controllability. Specifically, Figure 6a shows that the generated melodies not only harmonize with given lyrics but also exhibit a coherent and distinguishable verse-chorus structure, along with repetition patterns matching lyrics. Besides, it is observed that our model can well model the "one-to-many" alignment relationship between lyrics and melodies. In Figure 6b, the generate melodies well adhere to user-specified musical attributes. Figure 6c presents high-fidelity style transfer results of CSL-L2M equipped with the learned musical features, confirming that the learned features provide high-fidelity melody information to the Transformer decoder. In summary, our proposed CSL-L2M could generate melodies that not only match with the given lyrics but also adhere to user-specified musical attributes.



(a) High-quality and well-structured generation.



(b) Strong controllability.



(c) High-fidelity controllability.

Figure 6: Case study on generated melodies from CSL-L2M.

Conclusion

To address weak controllability, low-quality and poorly structured generation issues in the lyric-to-melody generation task, we propose CSL-L2M in this paper towards controllable song-level melody generation conditioning on lyrics and user-specified musical attributes. We first introduce a novel music representation named REMI-Aligned to facilitate precise lyric-melody alignment relationship modeling. Then multiple multi-granularity lyric and musical attribute controls are extracted and fed into the conditional Transformer decoder through in-attention to achieve firm control over the generation process. Experiments demonstrate that our proposed CSL-L2M outperforms the state-of-the-art models in terms of generation quality and controllability. We believe our contributions will further advance the under-explored field of lyric-to-melody generation.

Acknowledgments

This work was supported by the National Science and Technology Innovation 2030 – Major Project (Grant No.2022ZD0208800) and the National Natural Science Foundation of China (Grant No. 62176215).

References

Bao, H.; Huang, S.; Wei, F.; Cui, L.; Wu, Y.; Tan, C.; Piao, S.; and Zhou, M. 2019. Neural melody composition from lyrics. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part I 8*, 499–511. Springer.

Briot, J.-P.; and Pachet, F. 2020. Deep learning for music generation: challenges and directions. *Neural Computing and Applications*, 32(4): 981–993.

Brunner, G.; Konrad, A.; Wang, Y.; and Wattenhofer, R. 2018. MIDI-VAE: Modeling dynamics and instrumentation of music with applications to style transfer. In *ISMIR*, 747–754.

Choi, K.; Hawthorne, C.; Simon, I.; Dinculescu, M.; and Engel, J. 2020. Encoding musical style with transformer autoencoders. In *International conference on machine learning*, 1899–1908. PMLR.

Ding, S.; Liu, Z.; Dong, X.; Zhang, P.; Qian, R.; He, C.; Lin, D.; and Wang, J. 2024. Songcomposer: A large language model for lyric and melody composition in song generation. *arXiv preprint arXiv:2402.17645*.

Dong, H.-W.; Hsiao, W.-Y.; Yang, L.-C.; and Yang, Y.-H. 2018. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Duan, W.; Yu, Y.; and Oyama, K. 2024. Semantic dependency network for lyrics generation from melody. *Neural Computing and Applications*, 36(8): 4059–4069.

Duan, W.; Zhang, Z.; Yu, Y.; and Oyama, K. 2022. Interpretable melody generation from lyrics with discrete-valued adversarial training. In *Proceedings of the 30th ACM international conference on multimedia*, 6973–6975.

Hahn, S.; Zhu, R.; Mak, S.; Rudin, C.; and Jiang, Y. 2023. An Interpretable, Flexible, and Interactive Probabilistic Framework for Melody Generation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4089–4099.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The curious case of neural text degeneration. *The Eighth International Conference on Learning Representations*.

Huang, Y.-S.; and Yang, Y.-H. 2020. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM international conference on multimedia*, 1180–1188.

Ju, Z.; Lu, P.; Tan, X.; Wang, R.; Zhang, C.; Wu, S.; Zhang, K.; Li, X.; Qin, T.; and Liu, T.-Y. 2021. Telemelody: Lyric-to-melody generation with a template-based two-stage method. *arXiv preprint arXiv:2109.09617*.

Kawai, L.; Esling, P.; and Harada, T. 2020. Attributes-Aware Deep Music Transformation. In *ISMIR*, 670–677.

Long, C.; Wong, R. C.-W.; and Sze, R. K. W. 2013. T-music: A melody composer based on frequent pattern mining. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, 1332–1335. IEEE.

Monteith, K.; Martinez, T. R.; and Ventura, D. 2012. Automatic Generation of Melodic Accompaniments for Lyrics. In *ICCC*, 87–94.

Neves, P.; Fornari, J.; and Florindo, J. 2022. Generating music with sentiment using Transformer-GANs. In *ISMIR*, 717–725.

Nichols, E. 2009. Lyric-based rhythm suggestion. In *ICMC*.

Payne, C. 2019. MuseNet. https://openai.com/blog/musenet.

Roberts, A.; Engel, J.; Raffel, C.; Hawthorne, C.; and Eck, D. 2018. A hierarchical latent vector model for learning long-term structure in music. In *International conference on machine learning*, 4364–4373. PMLR.

Sarmento, P.; Kumar, A.; Chen, Y.-H.; Carr, C.; Zukowski, Z.; and Barthet, M. 2023. GTR-CTRL: instrument and genre conditioning for guitar-focused music generation with transformers. In *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*, 260–275. Springer.

Sheng, Z.; Song, K.; Tan, X.; Ren, Y.; Ye, W.; Zhang, S.; and Qin, T. 2021. Songmass: Automatic song writing with pre-training and alignment constraint. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 13798–13805.

Srivastava, A.; Duan, W.; Shah, R. R.; Wu, J.; Tang, S.; Li, W.; and Yu, Y. 2022. Melody generation from lyrics using three branch conditional LSTM-GAN. In *International Conference on Multimedia Modeling*, 569–581. Springer.

Tan, H. H.; and Herremans, D. 2020. Music FaderNets: Controllable Music Generation Based On High-Level Features via Low-Level Feature Modelling. In *ISMIR*, 109–116.

Tian, Y.; Narayan-Chen, A.; Oraby, S.; Cervone, A.; Sigurdsson, G.; Tao, C.; Zhao, W.; Chen, Y.; Chung, T.; Huang, J.; et al. 2023. Unsupervised melody-to-lyric generation. *arXiv preprint arXiv:2305.19228*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

von Rütte, D.; Biggio, L.; Kilcher, Y.; and Hofmann, T. 2023. FIGARO: Controllable music generation using learned and expert features. In *The Eleventh International Conference on Learning Representations*.

Wu, S.-L.; and Yang, Y.-H. 2023. MuseMorphose: Full-song and fine-grained piano music style transfer with one transformer VAE. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 1953–1967.

Yang, L.-C.; Chou, S.-Y.; and Yang, Y.-H. 2017. MidiNet: A convolutional generative adversarial network for symbolic-domain music generation. In *ISMIR*, 324–331.

Yu, Y.; Srivastava, A.; and Canales, S. 2021. Conditional LSTM-GAN for melody generation from lyrics. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(1): 1–20.

Zhang, C.; Chang, L.; Wu, S.; Tan, X.; Qin, T.; Liu, T.-Y.; and Zhang, K. 2022. Relyme: improving lyric-to-melody generation by incorporating lyric-melody relationships. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1047–1056.

Zhang, Z.; Yu, Y.; and Takasu, A. 2023. Controllable lyrics-to-melody generation. *Neural Computing and Applications*, 35(27): 19805–19819.