Regular Manuscript

Style-VT: Style Conditioned Chord Generation by Variational Transformer with Chord Substitution

Submission ID 9d4a00f3-34ef-44b9-aa09-2270da1c6ff6

Submission Version Initial Submission

PDF Generation 19 Jan 2025 21:07:48 EST by Atypon ReX

Authors

Mr. Kyowon Song

Affiliations
- Division of Computer Convergence, Chungnam National University, Daejeon, 34134 South Korea

Mr. Dongyoung Seo

Affiliations
- Division of Computer Convergence, Chungnam National University, Daejeon, 34134 South Korea

Mr. Junsu Na

Affiliations
- Division of Computer Convergence, Chungnam National University, Daejeon, 34134 South Korea

Mr. Heecheol Yang
Corresponding Author
Submitting Author

ORCiD
https://orcid.org/0000-0002-2802-2143

Affiliations
- Division of Computer Convergence, Chungnam National University, Daejeon, 34134 South Korea

Additional Information

Keywords

Machine learning

Machine learning algorithms

Music

Music information retrieval

Subject Category

Computational and artificial intelligence

Signal processing

Files for peer review

All files submitted by the author for peer review are listed below. Files that could not be converted to PDF are indicated; reviewers are able to access them online.

Name	Type of File	Size	Page
IEEE_Access_Style_VT.pdf	Main Document - PDF	4.6 MB	Page 3

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Style-VT: Style Conditioned Chord Generation by Variational Transformer with Chord Substitution

Kyowon Song[1], Dongyoung Seo[1], Junsu Na[1], (Student Member, IEEE), and Heecheol Yang[1], (Member, IEEE)

[1]Division of Computer Convergence, Chungnam National University, Daejeon, 34134 South Korea (e-mail: hcyang@cnu.ac.kr)

Corresponding author: Heecheol Yang (e-mail: hcyang@cnu.ac.kr).

ABSTRACT Recent advancements in music modeling have shown the potential of Transformer in tasks such as chord progression generation. Despite these successes, controlling chord progressions to reflect genre-specific characteristics remains underexplored. In this work, we present Style-VT, a style-conditioned chord generation model that combines the Transformer architecture with a variational autoencoder. Our model leverages genre embeddings to generate chord progressions that align with the stylistic norms of specific genres. We also introduce a newly defined chord substitution into the objective function, allowing for greater flexibility in generated chord progressions. By extensive experiments, we demonstrate that Style-VT generates a more flexible chord progressions and achieves higher harmonic similarities with human-composed chords than the existing methods.

INDEX TERMS Symbolic music generation, Chord generation, Variational autoencoder, Transformer, Chord substitution

I. INTRODUCTION

TRansformer [1] has shown significant potential in music modeling tasks, such as piano score generation [2] and video game sound synthesis [3]. Building on this success, recent research has focused on integrating various techniques into the Transformer framework. For instance, the Reinforcement Learning (RL) Transformer combines music transformers with RL by fine-tuning the pre-trained model using a reward function based on music theory [4]. Another approach, proposed in [5], introduces a new training method where various musical features are extracted during data pre-processing, and training is only performed when the generated melody sequences align with the chord sequences, thus reducing dissonance. There have also been attempts to incorporate chord segmentation and recognition within a Transformer framework to effectively capture the hierarchical structure of music [6]. Additionally, latent space learning has been employed to capture and analyze further musical features. The integration of Transformer and Variational Autoencoder (VAE) architectures [7] has proven effective in enhancing both the structural coherence and interpretability of music generation models [8]–[10]. Furthermore, a latent diffusion architecture has been introduced to enable rule-based controllability [11].

In this work, we focus on the task of chord generation for a given melody using a machine learning approach. Chord progressions are fundamental elements that not only construct the harmonic structure but also reflect the genre of the music. Existing studies have explored the classification of musical genres based on chord progressions [12], [13]. However, to the best of our knowledge, no research has focused on incorporating genre-specific characteristics directly into the chord progression generation. To fill this gap, we integrates genre-specific characteristics through genre embeddings within a variational transformer framework, which combines the strengths of Transformer and VAE architectures.

The main contribution of this work is to propose a style-conditioned chord generation model, termed *Style-VT*, which is based on variational transformer architecture. Beyond incorporating genre embeddings, we introduce chord substitution into the newly defined objective function to further refine the model's capabilities. Recognizing the inherently subjective nature of music, where multiple chord progressions can suit the given melody, these variations are crucial for arranging music across different genres. According to music theory [14], chord substitution involves replacing one chord with another that serves a similar harmonic function within a progression. We incorporate this chord substitution and integrate it into the model's objective function, thereby enhancing the genre-specific chord generation process. This approach effectively addresses the limitations of previous models, which often treated the original chord progression as the only correct solution, and instead promotes a more

flexible and theoretical method for chord generation.

II. RELATED WORK

A. MELODY CONDITIONED CHORD GENERATION

Automatic melody harmonization offers a compositional guidance while maintaining the original melody, thereby ensuring that human creativity is effectively integrated into the composition. There has been rapid progress in the development of machine learning models for automatic melody harmonization. In the early stages, extensive researches have been conducted using Hidden Markov Models (HMM) [15], [16] and Genetic Algorithms (GA) [17]. Afterwards, deep learning models using Long Short-Term Memory (LSTM) have shown great performance [18]–[20]. However, it has been revealed that these approaches have a limitation in generating structured chord progressions [9], [21]. Recently, it has been shown that Transformer effectively addresses the aforementioned issues through the attention mechanism, thereby enabling more sophisticated and consistent chord generation. The most relevant work to ours is [9] that introduces a probabilistic encoder and Transformer to specify global properties of given melody and chords. In [9], the key signature is used as a condition in a probabilistic encoder to acquire a specific harmonic context. In [22], the authors have focused on conveying the emotional characteristics of music based on key and chords by proposing a novel functional representation that utilizes Roman numerals relative to the musical keys to represent melody and chords. In addition, MelodyT5 [23] combines a Transformer-based encoder-decoder architecture with linear projection and patch-level encoder-decoder components.

B. STYLE CONDITIONED MUSIC GENERATION

Music is an abstract modality that does not possess clear meanings, specific referents, or defined objectives unlike language, speech, and images [24]. The emotions and atmosphere experienced from a specific music are unique subjective characteristics that vary by individuals. These are usually defined by "style", which can account for most descriptions of the music. DeepJ [25] has utilized a variant of the LSTM model, specifically biaxial LSTM, to create polyphonic music conditioned on a composer's style by applying style conditioning. In addition, theme-conditioned music generation has been proposed using Transformer that generates music by extending a given thematic sequence [26].

III. METHODOLOGY

In this section, we introduce Style-conditioned chord generation Variational Transformer coined *Style-VT*.

A. CHORD SUBSTITUTION

To apply chord substitution to our model, we first clarify the scope of chords that are treated as substitute chords. We propose using three types of substitute chords: Diatonic Substitute, Parallel Substitute, and Tritone Substitute.

The Diatonic Substitute refers to substituting a chord with another chord within the same key that shares a similar harmonic function. In diatonic harmony, chords are grouped by function: $\mathrm{I}, \mathrm{III}^{\mathrm{m}}, \mathrm{VI}^{\mathrm{m}}$ as Tonic, $\mathrm{II}^{\mathrm{m}}, \mathrm{IV}$ as Subdominant, and $\mathrm{V}, \mathrm{VII}^{\circ}$ as Dominant. Chords within the same functional group can substitute for one another, as they fulfill a similar harmonic role. For example, in the key of C major, substituting C major (C, I) with A minor $(\mathrm{Am}, \mathrm{VI}^{\mathrm{m}})$ maintains the tonic function, while substituting F major $(\mathrm{F}, \mathrm{IV})$ with D minor $(\mathrm{Dm}, \mathrm{II}^{\mathrm{m}})$ retains the subdominant function. The Parallel Substitute involves substituting a chord with another chord that shares the same root but belongs to a different mode, such as major to minor or vice versa. For example, substituting C major (C) with C minor (Cm) is a common application. The Tritone Substitute refers to substituting a dominant chord with another dominant chord that is a tritone (3 whole tones) away, sharing key harmonic tones. This works because the tritone interval between the third and seventh of the original dominant chord is preserved in the substitute chord. For example, substituting G^7 $(\mathrm{G}^7, \mathrm{V}^7)$ with $\mathrm{D}\flat^7$ $(\mathrm{D}\flat^7)$ achieves a similar dominant function.

Fig. 1 illustrates a substitute chord table that presents chord symbols and their corresponding analysis symbols across 12 keys. An analysis symbol represents the harmonic functions that each chord typically performs when used in a specific key and are primarily represented using Roman numerals. Chords that are not typically used in the given key are represented with an "X", while a "/" in the analysis symbols indicates that the chord is used as a secondary dominant. The red, green, and yellow groups indicate that the chords within the same color group can be used as substitute chords for each other.

B. MODEL ARCHITECTURE

Style-VT is composed of two primary components: Transformer and VAE. Transformer plays a role in translating a given melody into a chord sequence, and VAE learns global latent representations based on the genre. A whole model architecture is described in Fig. 2.

Transformer in Style-VT mostly follows the conventional architecture, except for the input and output representations. Inspired by [9], we adopt a note-wise representation method for depicting melodies because the conventional event-based representations differ from how humans perceive a melody for harmonization and thus it is hard to design patterns between melodies and chord labels. We utilize a binary melody piano roll and serialized chord labels. Each frame of the melody piano roll represents the same temporal length.

Let $m_{1:T} \in \{0,1\}^{T\times|P|}$ be a one-hot vector sequence of a given melody, where T is the length of the melody, $|P|$ is the number of pitches, and t is a time index by the length of a sixteenth note. A conditional token c is formed by concatenating the embedded genre vector and the key signature vector to inform the style of given input to each part of the model.

The Transformer melody encoder (TME) receives the input

C KEY					
C	Cm	Cdim	CM7	Cm7	C7
I	Im	X	I	X	V7/IV
Db	Dbm	Dbdim	DbM7	Dbm7	Db7
V	X	VIIdim/II	V	X	V7
D	Dm	Ddim	DM7	Dm7	D7
V/V	IIm	X	X	IIm	V7/V
Eb	Ebm	Ebdim	EbM7	Ebm7	Eb7
X	X	VIIdim/III	X	X	subV7/II
E	Em	Edim	EM7	Em7	E7
V/VI	IIIm	VIIdim/IV	X	IIIm	V7/VI
F	Fm	Fdim	FM7	Fm7	F7
IV	IVm	X	IV	IVm	subV7/III
Gb	Gbm	Gbdim	GbM7	Gbm7	Gb7
bV	X	VIIdim/V	X	Rel.II-V-I	subV7/IV
G	Gm	Gdim	GM7	Gm7	G7
V	X	X	X	Rel.II-V-I	V7
Ab	Abm	Abdim	AbM7	Abm7	Ab7
I	X	VIIdim/VI	IV	X	subV7/V7
A	Am	Adim	AM7	Am7	A7
V/II	VIm	X	X	VIm	V7/II
Bb	Bbm	Bbdim	BbM7	Bbm7	Bb7
IV or V	X	X	IV or V	X	subV7/VI
B	Bm	Bdim	BM7	Bm7	B7
V/III	X	VII	X	X	V/III

Db KEY					
Db	Dbm	Dbdim	DbM7	Dbm7	Db7
I	Im	X	I	X	V7/IV
D	Dm	Ddim	DM7	Dm7	D7
V	X	VIIdim/II	V	X	V7
Eb	Ebm	Ebdim	EbM7	Ebm7	Eb7
V/V	IIm	X	X	IIm	V7/V
E	Em	Edim	EM7	Em7	E7
X	X	VIIdim/III	X	X	subV7/II
F	Fm	Fdim	FM7	Fm7	F7
V/VI	IIIm	VIIdim/IV	X	IIIm	V7/VI
Gb	Gbm	Gbdim	GbM7	Gbm7	Gb7
IV	IVm	X	IV	IVm	subV7/III
G	Gm	Gdim	GM7	Gm7	G7
X	X	VIIdim/V	X	Rel.II-V-I	subV7/IV
Ab	Abm	Abdim	AbM7	Abm7	Ab7
V	X	X	X	Rel.II-V-I	V7
A	Am	Adim	AM7	Am7	A7
I	X	VIIdim/VI	IV	X	subV7/V7
Bb	Bbm	Bbdim	BbM7	Bbm7	Bb7
V/II	VIm	X	X	VIm	V7/II
B	Bm	Bdim	BM7	Bm7	B7
IV or V	X	X	IV or V	X	subV7/VI
C	Cm	Cdim	CM7	Cm7	C7
V/III	X	VII	X	X	V/III

Bb KEY					
Bb	Bbm	Bbdim	BbM7	Bbm7	Bb7
I	Im	X	I	X	V7/IV
B	Bm	Bdim	BM7	Bm7	B7
V	X	VIIdim/II	V	X	V7
C	Cm	Cdim	CM7	Cm7	C7
V/V	IIm	X	X	IIm	V7/V
Db	Dbm	Dbdim	DbM7	Dbm7	Db7
X	X	VIIdim/III	X	X	subV7/II
D	Dm	Ddim	DM7	Dm7	D7
V/VI	IIIm	VIIdim/IV	X	IIIm	V7/VI
Eb	Ebm	Ebdim	EbM7	Ebm7	Eb7
IV	IVm	X	IV	IVm	subV7/III
E	Em	Edim	EM7	Em7	E7
bV	X	VIIdim/V	X	Rel.II-V-I	subV7/IV
F	Fm	Fdim	FM7	Fm7	F7
V	X	X	X	Rel.II-V-I	V7
Gb	Gbm	Gbdim	GbM7	Gbm7	Gb7
I	X	VIIdim/VI	IV	X	subV7/V7
G	Gm	Gdim	GM7	Gm7	G7
V/II	VIm	X	X	VIm	V7/II
Ab	Abm	Abdim	AbM7	Abm7	Ab7
IV or V	X	X	IV or V	X	subV7/VI
A	Am	Adim	AM7	Am7	A7
V/III	X	VII	X	X	V/III

B KEY					
B	Bm	Bdim	BM7	Bm7	B7
I	Im	X	I	X	V7/IV
C	Cm	Cdim	CM7	Cm7	C7
V	X	VIIdim/II	V	X	V7
C#	C#m	C#dim	C#M7	C#m7	C#7
V/V	IIm	X	X	IIm	V7/V
D	Dm	Ddim	DM7	Dm7	D7
X	X	VIIdim/III	X	X	subV7/II
D#	D#m	D#dim	D#M7	D#m7	D#7
V/VI	IIIm	VIIdim/IV	X	IIIm	V7/VI
E	Em	Edim	EM7	Em7	E7
IV	IVm	X	IV	IVm	subV7/III
F	Fm	Fdim	FM7	Fm7	F7
X	X	VIIdim/V	X	Rel.II-V-I	subV7/IV
F#	F#m	F#dim	F#M7	F#m7	F#7
V	X	X	X	Rel.II-V-I	V7
G	Gm	Gdim	GM7	Gm7	G7
I	X	VIIdim/VI	IV	X	subV7/V7
G#	G#m	G#dim	G#M7	G#m7	G#7
V/II	VIm	X	X	VIm	V7/II
A	Am	Adim	AM7	Am7	A7
IV or V	X	X	IV or V	X	subV7/VI
A#	A#m	A#dim	A#M7	A#m7	A#7
V/III	X	VII	X	X	V/III

FIGURE 1. **Table of substitute chords and chord analysis symbols for each key.**

$m_{1:T}$ to capture the note-wise melodic context as follows.

$$e_{time} = \text{Embedding}(m_{1:T}), \tag{1}$$

$$e_{note} = \text{EmbConv}(e_{time} + w_{time}, M), \tag{2}$$

$$e_{note+1} = \text{Concat}(c, e_{note}), \tag{3}$$

$$e'_{note+1} = \text{TMEs}(e_{note+1} + w_{note}). \tag{4}$$

Embedding and TMEs denote the embedding layer and L transformer melody encoder blocks. e_{time}, e_{note}, and N denote the time-level embedding vectors, note-level embedding vectors, and the number of melody notes, respectively. w denotes a sinusoidal positional embedding, and EmbConv converts the time-wise embedding to the note-wise embedding to capture the note patterns in a melody. The conditional token c is concatenated at the beginning of the note-wised melody embedding e_{note}. In EmbConv block, we add the scaled positional embedding w_{time} to e_{time}. We transfer it to the note-wise embedding e_{note} with an average pooling by an alignment matrix $M \in \{0,1\}^{T\times N}$, where M denotes the

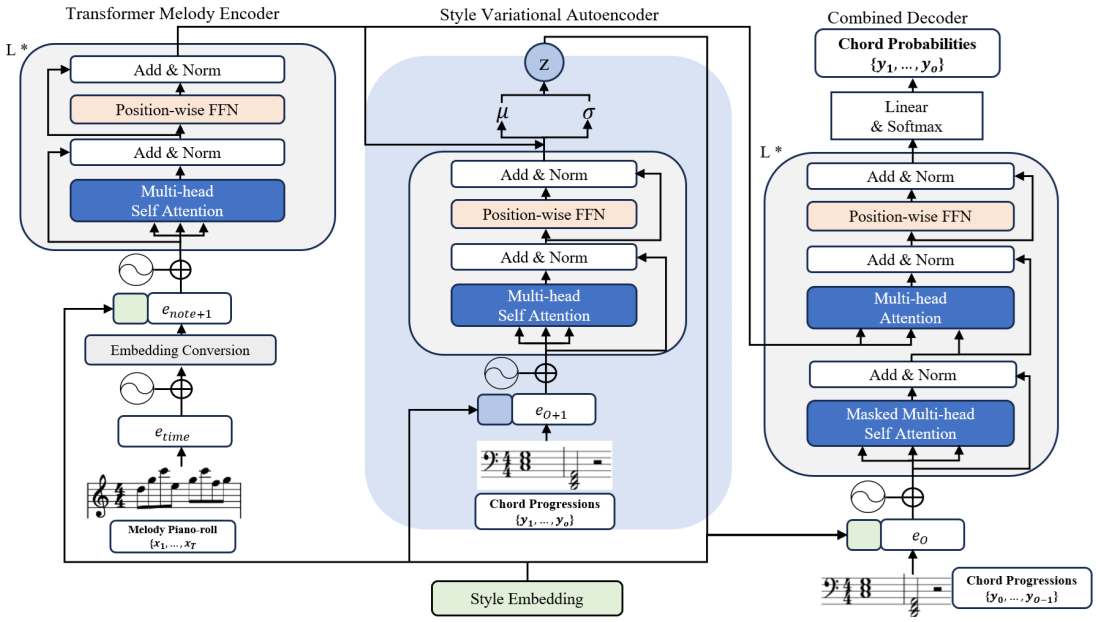


FIGURE 2. Model architecture

alignment path between a piano roll and a series of notes as

$$\text{EmbConv}(e, M) = \text{Linear}\left(\frac{M^T * e}{\sum_{t=1}^{T} M_{t,1:N}}\right). \quad (5)$$

The style variational encoder (SVE) infers the latent representation z from the output of the transformer melody encoder, chord input y, and conditional token c as follows.

$$e_{O+1} = \text{Concat}(c, \text{Embedding}(y_{1:O})), \quad (6)$$
$$e'_{O+1} = \text{SVE}(e_{O+1} + w_{O+1}), \quad (7)$$
$$r = \text{Concat}(\text{Pool}(e'_{note+1}), \text{Pool}(e'_{O+1})), \quad (8)$$
$$[\mu, \sigma] = \text{Linear}(r), \quad \text{where } z \sim \mathcal{N}(\mu, \sigma). \quad (9)$$

Pool denotes the average pooling. The outputs of SVE and TMEs are both mean-aggregated over time and concatenated, resulting in two parameters μ and σ, after passing through a fully connected layer. The latent variable z is inferred from μ and σ through the reparameterization trick, and its prior is assumed to follow a normal distribution.

The combined decoder (CD) reconstructs the right-shifted target chords and encoder output, conditioned by c and latent variable z as follows.

$$e_O = \text{Concat}(z + c, \textit{Embedding}(y_{1:O-1})), \quad (10)$$
$$e'_O = \text{CDs}(e_O + w_O, \textit{Enc}(x_{1:T})), \quad (11)$$
$$p(y_{1:O}) = \text{Softmax}(\textit{Linear}(e'_O)). \quad (12)$$

The latent variable z and the style token c are added to the beginning, which corresponds to the start-of-sequence part of the chord embedding. The following attention blocks transfer the aggregated information from z and c to all frames of the embedding. The rest of the CD reconstructs the target chords.

C. OBJECTIVE FUNCTION

In Style-VT, the primary objective is to approximate the marginal distribution of y by optimizing the negative evidence lower bound (ELBO). We define the objective function to train Style-VT as follows.

$$\mathcal{L}_{\text{Style-VT}} = \text{RE} + \lambda_S \text{SE} + \lambda_K \text{KLD}. \quad (13)$$

RE denotes the Reconstruction Error that measures the difference between the model output and the target value as

$$\text{RE} = \mathbb{E}_{q(z|x,y,c)}\left[-\log p_\theta(y \mid x, z, c)\right], \quad (14)$$

where the chord probability $p_\theta(y)$ and posterior distribution $q_\phi(z)$ are conditioned by the melody input x and style token c, whereas the prior $p_\theta(z)$ is normal distribution following the conditional VAE framework. SE denotes the Substitution reconstruction Error that reflects the impact of chord substitution on the reconstruction process, which is given by

$$\text{SE} = -\sum_{i=1}^{N}\sum_{k=1}^{C} v_{yk} \log(p_{ik}), \quad (15)$$

where N and C represent the length of the chord sequence and the chord set, respectively. v_y is a vector of size $|C|$ that represents whether it is a substitution chord for chord y, the target chord, across all chord sets. If $C_k \in \mathcal{S}(C_y)$, then $v_{yk} = 1$; otherwise, $v_{yk} = 0$. The vector p_i is a probability vector whose elements sum to 1, representing the model's predicted probabilities for the i-th chord sequence. KLD shows the Kullback-Leibler Divergence that regularizes the distribution of the latent variables, which is given by

$$\text{KLD} = \text{KL}\left(q_\phi(z|x, y, c) \parallel p(z)\right). \quad (16)$$

TABLE 1. Comparison of VAE Loss (Style-VT w/o genre embedding)

Method	Style-VT ($\lambda_S = 0$)	Style-VT ($\lambda_S = 0.05$)	Style-VT ($\lambda_S = 0.1$)	Style-VT ($\lambda_S = 0.15$)
RE+λ_KKLD	0.763	0.707	**0.681**	0.696

TABLE 2. Comparison of VAE Loss (Style-VT with genre embedding)

Method	Style-VT ($\lambda_S = 0$)	Style-VT ($\lambda_S = 0.05$)	Style-VT ($\lambda_S = 0.1$)	Style-VT ($\lambda_S = 0.15$)
RE+λ_KKLD	0.702	0.675	**0.658**	0.664

TABLE 3. Comparison on Harmonic Similarity, Chord Diversity, and Chord Coherence

Model \ Metric	Harmonic Similarity			Diversity and Coherence					
	TPSD↓	DICD↓	LD↓	CHE↑	CC↑	CTnCTR↑	PCS↑	CTD↓	MCTD↓
GT	-	-	-	1.629	6.688	1.319	0.385	0.623	1.403
VTHarm [9]	2.987	144.567	0.905	1.9571	8.0039	1.2162	0.3671	0.6862	1.4321
MelodyT5 [23]	2.695	142.047	0.907	1.538	6.238	1.210	**0.441**	0.695	**1.4054**
Style-VT (w/o style embedding, $\lambda_S = 0.1$)	2.799	135.533	0.862	**1.9660**	**8.0643**	1.2301	0.3798	0.7053	1.4297
Style-VT (style embedding, $\lambda_S = 0$)	2.903	134.795	0.867	1.9258	7.693	1.2136	0.3763	**0.6653**	1.4138
Style-VT (style embedding, $\lambda_S = 0.1$)	**2.647**	**131.493**	**0.843**	1.9426	7.8536	**1.2632**	0.3858	0.6942	1.4226

IV. EXPERIMENTS

A. DATASET

Since we could not find the previous research on a melody harmonization task for a given genre of music, we construct the dataset with three categories based on genre: jazz, rock, and the other genres. We rearrange the dataset from the Chord Melody Dataset (CMD) [27] by classifying the genres to include genre labels. We additionally collect jazz and rock genre music from the sheet music site, Musescore, to balance the genre of musics. All sheet music data are in a major key, and each piece of music is transposed into 12 keys. The sheet musics are formatted in MusicXML that can be parsed using a MusicXML parser library. We ensure that the all sheet music data follows that time signature is $4/4$ and the tempo is 120 BPM. We use triad and seventh chords for major, minor, and diminished chords, while excluding tension and added tone chords. We recognize a total of 72 chords based on 12 pitch classes. Each song is divided into 8-bar segments for training, with a 2-bar overlap with the previous training unit. The training, validation, and test sets are divided into 8:1:1. The datasets consist of 126 jazz songs with 91,845 training units, 74 rock songs with 80,369 training units, and 179 songs from other genres with 109,271 training units.

B. TRAINING

We train Style-VT for 100 epochs with a batch size of 128. A dropout layer with a rate of 0.2 is applied after each scaled positional encoding. We utilize Adam optimizer, starting with an initial learning rate of 1e-4 that decreases by 5% after every epoch. The embedding sizes for the melody and chord are 128 and 256, respectively. We use 4 attention blocks, each of which has an attention head size of 4, a hidden size of 256, and a latent variable z size of 3. We empirically set both λ_K and λ_S to 0.1. Comparisons for the effect of λ_S are provided in Table 1 (without genre embedding) and Table 2 (with genre embedding). Style-VT ($\lambda_S = 0$) without genre embedding is structurally identical to VTHarm [9].

C. METRICS

We utilize the two categories of metrics: harmonic similarity and chord diversity & coherence. We measure the harmonic similarity between the generated and human-composed chords using three metrics proposed in [9]. The tonal pitch step distance (TPSD) and directed interval class distance (DICD) measure the distance between two chord progressions. TPSD measures the geometrical dissimilarity between generated chords and ground-truth chords based on the Tonal Pitch Space (TPS) [28] chord distance rule. This metric evaluates the tonal relationship between chords by calculating their distance within the TPS framework. DICD calculates the city block distance between Directed Interval Class (DIC) representation vectors for chord transitions [29]. The DIC is a histogram vector representing directional pitch interval classes, which range from -5 to 6, computed for all pairs of chord notes between two adjacent chords. The Levenshtein edit distance (LD) is the global matching score between two chord sequences. LD measures the similarity between generated chord labels and ground-truth labels by calculating the minimum number of edits. In addition, we use the metrics proposed in [20] for chord diversity and coherence, which are widely utilized in melody harmonization studies [9], [23]. Chord histogram entropy (CHE) and chord coverage (CC) measure the diversity of chords. CHE evaluates the diversity of chord classes within a chord sequence by calculating entropy. The computation is based on a histogram, where each bin corresponds to a chord class, and the probabilities of their occurrences serve as input. CC is the total count of distinct chord labels with non-zero occurrences in the chord histogram of a given chord sequence. Chord tonal distance (CTD) measures the coherence of the chord transition. CTD is the average value of the tonal distance [30] computed between every pair of adjacent chords in a given

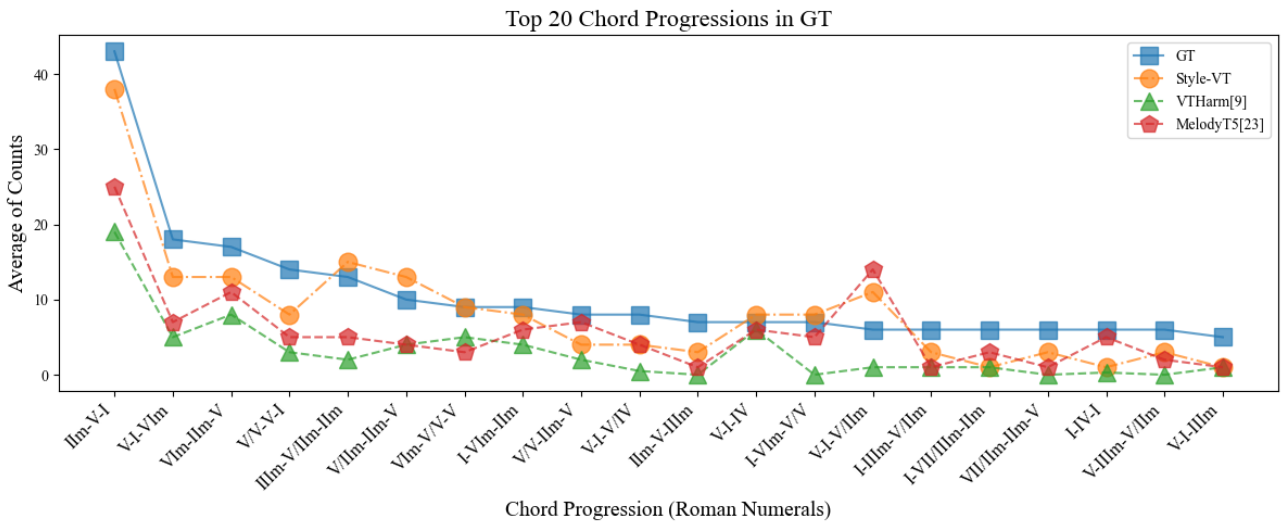


FIGURE 3. Chord Progression Frequency Analysis

chord sequence. The tonal distance is a canonical way to measure the closeness of two chords. Chord tone to non-chord tone ratio (CTnCTR), pitch consonance score (PCS), and melody-chord tonal distance (MCTD) measure the coherence between the melody and chords. CTnCTR is a metric that represents the ratio of chord tones to the sum of nonchord tones and proper nonchord tones. Proper nonchord tones are defined as those with a maximum interval of two semitones to the note immediately following. PCS evaluates consonance based on the pitch intervals between melody notes and their corresponding chord tones. It assumes that the melody notes are consistently higher in pitch than the chord tones. PCS assigns a score of 1 for intervals such as perfect unisons, perfect fifths, major/minor thirds, and major/minor sixths; 0 for perfect fourths; and -1 for all other intervals. The PCS values are averaged within each sixteenth-note window, and the overall PCS is calculated by averaging these values across all windows over time. MCTD measures the tonal distance between each melody note and its corresponding chord label. The calculation follows the same method as CTD. Each MCTD value is weighted by the duration of the corresponding melody note, and the overall MCTD is obtained by averaging the values across all melody notes and their associated chord labels.

D. EVALUATION

We conduct an ablation study to evaluate the impact of genre embeddings and chord substitution on model performance. We set VTHarm [9] and MelodyT5 [23] as the baseline and set the original progressions of the dataset as a ground truth. Evaluation results are provided in Table 3.

We evaluate the harmonic similarity between model-generated chords and human-composed chords, under the assumption that chord progressions in human-composed music are well structured. Style-VT (style embedding, $\lambda_S = 0.1$) exhibits the best performance across all three metrics: TPSD, DICD, and LD. Style-VT (with style embedding, $\lambda_S = 0$) and Style-VT (without style embedding, $\lambda_S = 0.1$) demonstrate superior performance compared to both VTHarm and MelodyT5. This implies that chord progressions generated by Style-VT are the most similar to those composed by humans, and thus Style-VT is capable of creating structured harmonies by style-embeddings and chord substitution.

From the perspective of chord diversity, Style-VT (without style embedding, $\lambda_S = 0.1$) achieves the highest values for CC and CHE than Style-VT ($\lambda_S = 0$). In this case, the use of genre embedding led to a decrease in chord diversity, while the application of substitute chords was found to enhance it. This difference can be attributed to the direct chord guidance provided by style embeddings, in contrast to the broader range of options introduced by chord substitution. Style-VT (style embedding, $\lambda_S = 0.1$) achieves the highest CTnCTR, implying that more harmonic chord tones are generated in the melody. In the other three metrics for evaluating coherence, Style-VT demonstrates similarly high performance compared to the baseline.

In addition, we conduct an additional experiment to evaluate whether the music generated by Style-VT effectively captures the intended style. Under the assumption that the distribution of chord progression patterns varies by genre, we analyze the chord progressions generated by each model in 50 test songs from the jazz genre. Fig. 3 shows the comparison of chord progression frequencies between the original songs and the generated compositions from VTHarm, MelodyT5, and Style-VT. The frequency of chord progressions is accumulated using the sliding window of size 3 with a stride of 1 in the music generated by each model, and the most frequently occuring progressions in GT are plotted in Fig. 3. The chords generated by each model are converted into the corresponding analysis symbols for each key. In Fig. 3, it

is observed that Style-VT closely follows the frequent chord progressions in GT from 1st to 8th, which means that Style-VT resembles the frequent chord progression patterns in GT compared to VTHarm and MelodyT5. Beyond the 8th most frequent progression, Style-VT continues to exhibit a distribution that closely follows GT, although slight deviations are observed. In contrast, VTHarm and MelodyT5 exhibit more significant deviations in chord progression frequencies across the overall progressions. This implies that Style-VT is more closely aligned with the chord progression patterns of the original dataset over the entire range of progressions.

V. CONCLUSION

In this paper, we introduced Style-VT, a style-conditioned chord generation model that integrates Transformer and VAE architectures. By incorporating style embeddings and a newly defined chord substitution, Style-VT effectively captures and generates chord progressions that are coherent with specific genres. Our experimental results demonstrate that Style-VT outperforms existing methods in generating more flexible and musically appropriate chord progressions. This work not only advances the field of automatic chord generation but also opens a future research direction in style-conditioned music generation, potentially expanding to other musical elements and genres.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st NeurIPS*, pp. 6000–6010, Dec. 2017.

[2] M. Suzuki, "Score Transformer: Generating Musical Score from Note-level Representation," in *Proc. 3rd ACM Int. Conf. Multimedia in Asia*, 2021.

[3] C. Donahue, H. H. Mao, Y. E. Li, G. W. Cottrell, and J. McAuley, "LakhNES: Improving multi-instrumental music generation with cross-domain pre-training," in *20th ISMIR*, Delft, Netherlands, pp. 685–692, Jul. 2019.

[4] X. Guo, H. Xu, and K. Xu, "Fine-Tuning Music Generation with Reinforcement Learning Based on Transformer," in *Proc. IEEE 16th Int. Conf. Anti-counterfeiting, Security, and Identification*, Xiamen, China, pp. 1-5, 2022.

[5] J. Tang, L. Yin, and J. Yu, "Partially Trained Music Generation based on Transformer," in *Proc. ICSCSS*, Coimbatore, India, pp. 1133–1138, 2023.

[6] T.-P. Chen, L. Su, and others, "Harmony Transformer: Incorporating chord segmentation into harmony recognition," *Neural Netw.*, vol. 12, pp. 15, 2019.

[7] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv:1312.6114*, Dec. 2013.

[8] J. Jiang, G. G. Xia, D. B. Carlton, C. N. Anderson, and R. H. Miyakawa, "Transformer VAE: A Hierarchical Model for Structure-Aware and Interpretable Music Representation Learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, pp. 516–520, 2020.

[9] S. Rhyu, H. Choi, S. Kim, and K. Lee, "Translating melody to chord: Structured and flexible harmonization of melody with Transformer," *IEEE Access*, vol. 10, pp. 28261–28273, Feb. 2022.

[10] S. Ji and X. Yang, "Emotion-Conditioned Melody Harmonization with Hierarchical Variational Autoencoder," in *Proc. IEEE Int. Conf. Sys. Man. Cybern. (ICSMC)*, pp. 228–233, 2023.

[11] Y. Huang, A. Ghatare, Y. Liu, Z. Hu, Q. Zhang, C. S. Sastry, S. Gururani, S. Oore, and Y. Yue, "Symbolic music generation with non-differentiable rule guided diffusion," *arXiv:2402.14285*, Feb. 2024.

[12] C. Pérez-Sancho, D. Rizo, and J. M. Inesta, "Genre classification using chords and stochastic language models," in *Proc. Connection Science*, vol. 21, pp. 145–159, Nov. 2009.

[13] C. Pérez-Sancho, D. Rizo, J. M. Iñesta, P. J. P. de León, S. Kersten, and R. Ramirez, "Genre classification of music by tonal harmony," in *Proc. Intelligent Data Analysis*, vol. 14, no. 5, pp. 533–545, Sep. 2010.

[14] A. Latham, *Oxford Companion to Music*, Oxford University Press, 1938.

[15] M. Kaliakatsos-Papakostas and E. Cambouropoulos, "Probabilistic harmonization with fixed intermediate chord constraints," in *Proc. ICMC*, Sep. 2014.

[16] I. Simon, D. Morris, and S. Basu, "MySong: automatic accompaniment generation for vocal melodies," in *Proc. SIGCHI Conf. Human Factors in Computing Systems*, Florence, Italy, Apr. 2008, pp. 725–734.

[17] A. Freitas and F. Guimaraes, "Melody harmonization in evolutionary music using multiobjective genetic algorithms," in *Proc. 8th Sound Music Comput. Conf. (SMC)*, Padova, Italy, pp. 1–8, Jan. 2011.

[18] H. Lim, S. Rhyu, and K. Lee, "Chord generation from symbolic melody using BLSTM networks," *arXiv:1712.01011*, pp. 621–627, Dec. 2017.

[19] C.-E. Sun, Y.-W. Chen, H.-S. Lee, Y.-H. Chen, and H.-M. Wang, "Melody Harmonization Using Orderless Nade, Chord Balancing, and Blocked Gibbs Sampling," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process (ICASSP)*, Toronto, ON, Canada, pp. 4145–4149, 2021.

[20] Y.-C. Yeh, W.-Y. Hsiao, S. Fukayama, T. Kitahara, B. Genchel, H.-M. Liu, H.-W. Dong, Y. Chen, T. Leong, and Y.-H. Yang, "Automatic melody harmonization with triad chords: A comparative study," *Journal of New Music Research*, vol. 50, no. 1, pp. 37–51, Jan. 2021.

[21] S. Li and Y. Sung, "Transformer-Based Seq2Seq Model for Chord Progression Generation," *Mathematics*, vol. 11, no. 5, Art. no. 1111, Feb. 2023.

[22] J. Huang and Y.-H. Yang, "Emotion-Driven Melody Harmonization via Melodic Variation and Functional Representation," *arXiv:2407.20176*, Jul. 2024.

[23] S. Wu, Y. Wang, X. Li, F. Yu, and M. Sun, "MelodyT5: A Unified Score-to-Score Transformer for Symbolic Music Processing," *arXiv:2407.02277*, Jul. 2024.

[24] R. B. Dannenberg, "Style in music," *The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning*, Springer, pp. 45–57, 2010.

[25] H. H. Mao, T. Shin, and G. Cottrell, "DeepJ: Style-specific music generation," in *Proc. IEEE 12th Int. Conf. Semant. Comput. (ICSC)*, pp. 377–382, 2018.

[26] Y.-J. Shih, S.-L. Wu, F. Zalkow, M. Müller, and Y.-H. Yang, "Theme transformer: Symbolic music generation with theme-conditioned transformer," *IEEE Transactions on Multimedia*, vol. 25, pp. 3495–3508, 2022.

[27] S. Hiehn, "Chord Melody Dataset," Accessed: Aug. 5, 2024. [Online]. Available: https://github.com/shiehn/chord-melody-dataset.

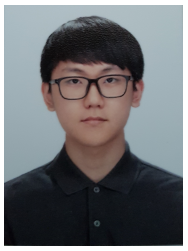
[28] W. B. de Haas, F. Wiering, and R. C. Veltkamp, "A geometrical distance measure for determining the similarity of musical harmony," *Int. J. Multimedia Inf. Retr.*, vol. 2, no. 3, pp. 189–202, Sep. 2013.

[29] E. Cambouropoulos, "A directional interval class representation of chord transitions," in *Proc. 12th Int. Conf. Music Percept. Cogn. (ICMPC)*, Thessaloniki, Greece, pp. 1–5, 2012.

[30] C. Harte, M. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," in *Proc. 1st ACM Workshop Audio Music Comput. Multimedia (AMCMM)*, pp. 21–26, 2006.

KYOWON SONG is currently pursuing the B.S. degree in computer science and engineering at Chungnam National University, South Korea. His academic interests lie in the intersection of music generation and machine learning, with a particular focus on applying artificial intelligence to analyze and create musical compositions. He is deeply motivated by the potential of machine learning models to enhance creativity and improve understanding of musical structures. His current work explores topics such as music harmonization, generative models for music, and the development of computational tools to support music creation and analysis.

DONGYOUNG SEO is currently pursuing the B.S. degree in computer science and engineering at Chungnam National University, South Korea. His academic focus is centered on the application of machine learning techniques to music generation and analysis. He is particularly interested in exploring how artificial intelligence can enhance musical creativity, automate composition processes, and uncover new insights into harmonic and melodic structures.



JUNSU NA is currently pursuing the B.S. degree in computer science and engineering at Chungnam National University, South Korea. His research interests revolve around leveraging machine learning for music generation and computational creativity. He is particularly fascinated by the role of AI in generating harmonically rich compositions and developing algorithms that mimic human-like musical intuition.

HEECHEOL YANG received the B.S. degree and the Ph.D. degree in electrical and computer engineering from Seoul National University, South Korea, in 2013 and 2018, respectively. He was a staff engineer of standard research team in Samsung Research at Samsung Electronics, Co., Ltd. from 2018 to 2019, and was an assistant professor with the School of Electronic Engineering, Kumoh National Institute of Technology, South Korea, from 2019 to 2021. He is currently an assistant professor with the division of computer convergence, Chungnam National University, South Korea. His research interests include wireless communication systems for beyond 5G and 6G, distributed storage systems and distributed computing, and data security and privacy in distributed computing and learning systems. He was awarded the Best Ph.D. Dissertation Award from Seoul National University in 2018, the Bronze Prize from the 23th Samsung HumanTech Paper Contest, and the Gold Prize from the IEEE Student Paper Contest (Seoul Section) 2017.