

TPSMG：テキスト制御可能なポリフォニックシンボリック音楽生成

朱 金龍[†], 櫻井 慶悟^{††}, 藤後 廉^{†††}, 小川 貴弘^{†††}, 長谷山 美紀^{†††}

概要 拡散モデルに基づく新しいテキスト制御可能なポリフォニックシンボリック音楽生成法を提案する。記号的音楽生成は、その柔軟性とデジタルオーディオワークステーション(DAW)とのシームレスな統合により、波形音楽と比較して変更が容易なMIDIファイル生成が可能となり、大きな注目を集めている。既存の手法では、コードやその他のメタデータによる制御が可能であるが、テキストプロンプトによる直感的な制御が可能な手法はほとんどなく、ユーザーの好みにより合致している。この限界に対処するため、テキスト条件付き記号音楽生成のために特別に設計された拡散モデルであるテキスト制御可能なポリフォニック記号音楽生成(TPSMG)を紹介する。我々のアプローチは、ノイズ除去拡散確率モデル内のU-Netバックボーンにテキスト条件モジュールを組み込む。このモジュールは、テキストプロンプトを埋め込みに変換し、ノイズ除去プロセスを制御することで、音楽生成に対するテキストベースの正確な制御を可能にする。実験結果は、本手法が意図したテキスト入力を忠実に反映した高品質のポリフォニックシンボリック音楽出力を生成することを示している。

キーワード：シンボリック音楽生成、ノイズ除去拡散確率モデル、シンボリック信号処理、自動作曲

1. まえがき

自動音楽生成は1950年代から広範な研究の対象となっており[1]、インスピレーションがない場合でも、より効率的に音楽を生成したいという願望が動機となっている。近年、動画、映画、ゲームなどのマルチメディアアプリケーションの急激な成長により、オリジナル音楽に対する需要がさらに高まっている[2]。このような需要の急増は、数多くの自動音楽生成モデルの開発に拍車をかけている[3]。音楽家が創造的なプロセスを効果的に支援するためには、音楽生成モデルが音楽家自身によって容易に修正可能な出力を生成することが不可欠である。しかし、ほとんどの波形音楽生成モデルは、音声フォーマットで音楽を出力するため、直接的に音を表し、音楽家が生成されるコンテンツを変更する柔軟性が制限される[3]。これに対して、記号的音楽生成モデルは、音楽をピアノロールや記号的イベントのシー

ケンスとして表現し、通常、音楽楽器デジタルインターフェース(MIDI)フォーマットで音楽を出力する[4]。このフォーマットにより、音楽家はデジタルオーディオワークステーション(DAW)を通じて音楽を簡単に編集・カスタマイズすることができ、大きな柔軟性と創造的なコントロールを提供することができる。その結果、記号的な音楽生成は、音楽編集に実用的な有用性を提供するため、近年大きな研究関心を集めている。従来の記号的音楽生成手法の進歩にもかかわらず、プロンプトで生成されたサンプルを正確に制御することは依然として課題である。特にテキストは、ユーザーが自分の音楽的意図を説明するための非常に効果的な手段を提供する。画像生成のような他の生成領域では、テキストプロンプトはその単純さと表現力から標準的なアプローチとなっている[5, 6]。テキストベースのプロンプトは、ジャンル、ムード、和音進行などの本質的な音楽要素を捉えることができ、音楽理論に関する広範な知識を持たないユーザーでも、音楽生成にアクセスしやすくなる。本研究では、テキスト入力は主に自由形式の音楽説明(例:「D minorのリラックスできるポップソング」)、感情的な手がかり(例:「メランコリックで夢想的強い」)、およびシステムの入力空間の範囲を定義するジャンルやテンポなどのカテゴリラベルから構成される。テキスト制御可能な記号的音楽生成は、熟練した音楽家とアマチュアの両方を鼓舞する可能性があるにもかかわらず、現在この領域を扱ったモデルはほとんどない。効果的なテキスト

control relies on selecting an appropriate architecture.

Received May 28, 2025; Revised September 4, 2025; Accepted October 8, 2025

[†]Graduate School of Information Science and Technology, Hokkaido University

(Kita-14 Nishi-9 Kita-ku Sapporo, 060-0814, Japan)

^{††}Data-Driven Interdisciplinary Research Emergence Department (D-RED), Institute for Integrated Innovations, Hokkaido University

(Kita-12 Nishi-7 Kita-ku Sapporo, 060-0812, Japan)

^{†††}Faculty of Information Science and Technology, Hokkaido University

(Kita-14 Nishi-9 Kita-ku Sapporo, 060-0814, Japan)

[機械翻訳コンテンツの著作権について]

当サイトに掲載されている原著論文の著作権は映像情報メディア学会に帰属します。原著論文および翻訳論文の無断使用は禁止します。
引用の際には、必ず原著論文の書誌情報をご記載ください。詳細は映像情報メディア学会著作権規定をご覧ください。

TPSMG: Text-Controllable Polyphonic Symbolic Music Generation

Jinlong Zhu[†], Keigo Sakurai^{††}, Ren Togo^{†††},
Takahiro Ogawa (member)^{†††}, Miki Haseyama (fellow)^{†††}

Abstract We propose a novel text-controllable polyphonic symbolic music generation method based on diffusion models. Symbolic music generation has garnered significant attention due to its flexibility and seamless integration with Digital Audio Workstations (DAWs), as it enables the generation of MIDI files, facilitating easier modification compared to waveform music. Although existing techniques enable control through chords or other metadata, few methods allow intuitive control via text prompts, which better align with user preferences. To address this limitation, we introduce *Text-Controllable Polyphonic Symbolic Music Generation (TPSMG)*, a diffusion model specifically designed for text-conditioned symbolic music generation. Our approach incorporates a text condition module into a U-Net backbone within a Denoising Diffusion Probabilistic Model. This module translates text prompts into embeddings that steer the denoising process, thereby enabling precise, text-based control over music generation. Experimental results demonstrate that our method generates high-quality polyphonic symbolic music outputs that closely reflect the intended textual input.

Key words: Symbolic music generation, denoising diffusion probabilistic model, symbolic signal processing, automatic composition

1. Introduction

Automatic music generation has been a subject of extensive research since the 1950s [1], motivated by the aspiration to produce music more efficiently, even in the absence of inspiration. In recent years, the exponential growth of multimedia applications, including videos, films, and games, has further intensified the demand for original music [2]. This surge in demand has spurred the development of numerous automatic music generation models [3]. To effectively assist musicians in the creative process, it is essential for music generation models to produce outputs that are easily modifiable by the musicians themselves. However, most waveform music generation models output music in audio format, which represents sound directly and restricts the flexibility for musicians to alter the generated content [3].

In contrast, symbolic music generation models represent music as piano rolls or sequences of symbolic events

and typically output music in the Musical Instrument Digital Interface (MIDI) format [4]. This format allows musicians to easily edit and customize music through Digital Audio Workstations (DAWs), offering significant flexibility and creative control. Consequently, symbolic music generation has attracted considerable research interest in recent years, as it provides practical utility for music editing. Despite advances in conventional symbolic music generation methods, achieving precise control over generated samples with prompts remains a challenge. Text, in particular, offers a highly effective means for users to describe their musical intentions. In other generative domains, such as image generation, text prompts have become a standard approach due to their simplicity and expressiveness [5, 6]. Text-based prompts can capture essential musical elements, including genre, mood, and chord progressions, making music generation more accessible even to users without extensive knowledge of music theory. In this study, the textual inputs mainly consist of free-form musical descriptions (e.g., “a relaxing pop song in D minor”), emotional cues (e.g., “melancholic and dreamy”), and categorical labels such as genre or tempo, which define the scope of our system’s input space. Despite the potential of text-controllable symbolic music generation to inspire both expert musicians and amateurs, few models currently address this domain. Effective text control relies on selecting an appropriate architecture.

Received May 28, 2025; Revised September 4, 2025; Accepted October 8, 2025

[†]Graduate School of Information Science and Technology, Hokkaido University
(Kita-14 Nishi-9 Kita-ku Sapporo, 060-0814, Japan)

^{††}Data-Driven Interdisciplinary Research Emergence Department (D-RED), Institute for Integrated Innovations, Hokkaido University
(Kita-12 Nishi-7 Kita-ku Sapporo, 060-0812, Japan)

^{†††}Faculty of Information Science and Technology, Hokkaido University
(Kita-14 Nishi-9 Kita-ku Sapporo, 060-0814, Japan)

畳み込みニューラルネットワーク、リカレントニューラルネットワーク、トランスフォーマーなどの従来のアーキテクチャは、特に複数の時間スケールでモデリングする場合、制御性と一貫性の限界に直面する可能性がある[7]。これらの課題に対処するため、ノイズ除去拡散確率モデル(DDPM)は、その効果的な制御可能性から注目されている[8]。DDPMは、自然言語入力を条件とした場合、テキストプロンプトに密接に沿った出力を生成する能力を実証している[9]。この利点は、ノイズの多い入力を徐々にきれいなサンプルに洗練していく、その反復的なサンプリングプロセスに起因する。このステップバイステップの洗練は、正確な制御を可能にし、画像生成だけでなく、記号的な音楽生成においても、高品質で制御可能な出力を生成するのに有効であることが証明されている[10-14]。

本研究では、音楽要素のテキスト記述を条件としたポリフォニックな記号音楽を生成するための、新しいDDPMベースのアプローチであるテキスト制御可能なポリフォニック記号音楽生成(TPSMG)を提案する。我々のモデルにおけるテキスト制御の最も効果的な戦略を特定するために、様々なテキストエンコーダを評価し、記号的音楽生成への影響を分析する。我々の研究は、テキスト制御可能な記号的音楽生成の初期段階の探求であり、この分野における将来の研究に有望な方向性を提供する。実験結果は、我々の方法が、提供されたテキストプロンプトに密接に沿った記号的な音楽出力を生成できることを示している。本論文の貢献は以下のようにまとめられる：

- テキストプロンプトからポリフォニックな記号音楽を生成する方法を提示し、テキスト制御可能な記号音楽生成の比較的未開拓の分野に貢献する。
- DDPMベースのアプローチを採用することで、テキストプロンプトによる音楽生成プロセスの正確な制御を実現し、生成された音楽が多様な音楽特徴にわたって入力テキストと密接に整合することを保証するのに有効であることを実証する。

2. 関連研究

記号的音楽生成は長年の研究分野であり、その主な理由は DIDI 出力の使用によるもので、DAW で容易に編集でき、広範なカスタマイズが可能である。テキスト制御可能な記号的音楽生成の実現には、音楽表現、外部制御メカニズム、モデルアーキテクチャを注意深く考慮する必要がある。そこで、(1)記号的音楽表現、(2)テキスト条件付き記号的音楽生成、

(3)音楽生成のための拡散ベースのモデル、の3つの主要分野にまたがる関連文献をレビューする。

2.1 シンボリック音楽表現

記号的音楽生成の初期の研究は、主に曲の音楽的特徴を効果的に捉えるための適切な表現を特定することに焦点を当てていた。音楽は本質的に時系列データの一種態であることから、これらの研究の多くは音楽を逐次構造としてモデル化している。

Huangら[15]は、音符の持続時間とバーイベントを組み込んだシングルトラック表現であるREvamped MIDI-derived events (REMI)を提案した。REMIは微妙なリズムの繰り返しを持つ音楽の生成を可能にするが、過度に長いトークン列を生成する傾向があり、マルチトラック音楽生成のサポートに欠ける。これらの限界に対処するために、Dongら[16]は、Transformer-XLベースのデコーダのみのモデルであるMultitrack Music Transformer (MMT)と共に、六重奏トークンのシーケンスに基づくマルチトラック表現を導入した。このアーキテクチャは、多次元の入力と出力を効果的に処理し、従来のアプローチと比較して、より長く複雑なマルチトラック構成の生成を可能にする。しかし、これらの方法は、ジャンル、気分、テキスト記述の条件付けのような外部制御のメカニズムを欠いている。ユーザーは通常、最初のシードを操作することに限定され、出力に対する限定的で不都合な制御しか提供しない。記号的音楽生成における制御性を向上させるために、Minら[12]は2チャンネルの2値テンソルとしてフォーマットされたピアノロール表現を提案し、外部制御への従順性で知られる拡散モデルの使用を容易にした。同様に、拡散モデルを活用し、ユーザー入力に沿ったより効果的な外部制御を可能にするために、時系列シーケンスではなくピアノロール表現を採用する。

2.2 テキスト条件付き音楽生成

テキスト制御可能な音楽生成は、外部条件付き音楽生成のより広い領域におけるサブタスクである。先行研究では、和音[17, 18]、歌詞[19]、音符の密度や声調[20]などの他の音楽的特徴を条件とした音楽生成が研究されている。テキスト制御可能な記号的音楽生成に関する研究は、最近いくつかの方法が登場しているが、まだ初期段階にある。Luら[21]は、音楽属性の中間表現を介して、自由形式のテキストを編集可能なMIDIに変換する2段階のフレームワークであるMuseCocoを紹介した。

Traditional architectures such as convolutional neural networks, recurrent neural networks, and transformers may face limitations in controllability and consistency, particularly when modeling across multiple temporal scales [7]. To address these challenges, Denoising Diffusion Probabilistic Models (DDPMs) have gained prominence due to their effective controllability [8]. DDPMs have demonstrated an ability to generate outputs that closely align with textual prompts when conditioned on natural language input [9]. This advantage stems from their iterative sampling process, which gradually refines a noisy input into a clean sample. This step-by-step refinement allows for precise control and has proven effective in generating high-quality, controllable outputs not only in image generation but also in symbolic music generation [10–14].

In this study, we propose *Text-Controllable Polyphonic Symbolic Music Generation (TPSMG)*, a novel DDPM-based approach for generating polyphonic symbolic music conditioned on textual descriptions of musical elements. To identify the most effective strategy for text control in our model, we evaluate a range of text encoders and analyze their influence on symbolic music generation. Our work represents an early-stage exploration of text-controllable symbolic music generation, offering a promising direction for future research in this field. Experimental results demonstrate that our method can generate symbolic music outputs that closely align with the provided text prompts. The contributions of this paper are summarized as follows:

- We present a method for generating polyphonic symbolic music from text prompts, contributing to the relatively underexplored field of text-controllable symbolic music generation.
- We demonstrate that employing a DDPM-based approach is effective for achieving precise control over the music generation process via text prompts, ensuring that the generated music closely aligns with the input text across diverse musical features.

2. Related Work

Symbolic music generation has been a longstanding research area, largely due to the use of MIDI output, which is easily editable in DAWs and enables extensive customization. Achieving text-controllable symbolic music generation requires careful consideration of music representation, external control mechanisms, and model architecture. Accordingly, we review the relevant literature across three key areas: (1) symbolic mu-

sic representation, (2) text-conditioned symbolic music generation, and (3) diffusion-based models for music generation.

2.1 Symbolic Music Representation

Early research in symbolic music generation primarily focused on identifying suitable representations to effectively capture the musical characteristics of a song. Given that music is inherently a form of time-series data, many of these studies have modeled music as a sequential structure.

Huang *et al.* [15] proposed REvamped MIDI-derived events (REMI), a single-track representation that incorporates note durations and bar events. Although REMI enables the generation of music with nuanced rhythmic repetition, it tends to produce overly long token sequences and lacks support for multitrack music generation. To address these limitations, Dong *et al.* [16] introduced a multitrack representation based on sequences of sextuple tokens, along with the Multitrack Music Transformer (MMT), a Transformer-XL-based decoder-only model. This architecture effectively handles multi-dimensional inputs and outputs, enabling the generation of longer and more complex multitrack compositions compared to previous approaches. However, these methods lack mechanisms for external control, such as conditioning on genre, mood, or textual descriptions; users are typically limited to manipulating the initial seed, which provides only limited and inconvenient control over the output. To improve controllability in symbolic music generation, Min *et al.* [12] proposed a piano-roll representation formatted as a two-channel binary tensor, facilitating the use of diffusion models known for their amenability to external control. Similarly, we adopt a piano-roll representation rather than a time-series sequence to leverage diffusion models and enable more effective external control aligned with user input.

2.2 Text-Conditioned Music Generation

Text-controllable music generation is a subtask within the broader domain of externally conditioned music generation. Prior research has explored music generation conditioned on chords [17, 18], lyrics [19], and other musical features such as note density and voicing [20]. Research on text-controllable symbolic music generation remains at an early stage, although several methods have recently emerged. Lu *et al.* [21] introduced MuseCoco, a two-stage framework that converts free-form text into editable MIDI via an inter-

MuseCocoは、テキストから属性への理解モジュールと、それに続く属性から音楽への生成モジュールから構成される。このモジュール設計は実用的で制御可能な生成を容易にするが、個別に学習されたコンポーネントが必要とされるため、パイプラインの複雑が増し、エンドツーエンドの最適化が制限される。これに対して、Bhandariら[22]は、自由形式のテキストプロンプトからMIDI形式の記号音楽を直接生成する、最先端のエンドツーエンドモデルであるText2midiを提案した。Text2midiは、入力キャプションを埋め込むために凍結されたFLAN-T5エンコーダと、MIDIトークンを生成するために自己回帰的に訓練されたTransformerデコーダを採用し、MuseCocoと比較して優れた性能を達成した。このモデルは、SymphonyNet [23]の擬似キャプションを用いて半教師付きで事前学習され、MidiCapsデータセット [24]で微調整される。Text2midiは直接的なテキストガイド付き音楽生成を可能にするが、生成された音楽と入力テキストとの間の整合性はまだ限定的であり、キー、モード、ジャンルにしばしば矛盾が見られる。これに対し、我々の提案するTPSMGはDDPMベースのアーキテクチャを採用しており、以下のサブセクションで詳述するように、外部制御信号を取り込む優れた能力を実証している。

2.3 音楽生成のための拡散に基づくモデル

近年、複雑な構造化離散データのモデリングにおける有望性を反映し、記号音楽生成のための拡散モデルへの注目が高まっている。元来、画像や音声のような連続的な領域のために開発された拡散モデルは、適切な表現や条件付けのメカニズムと組み合わせることで、記号的な音楽イベントの複雑な分布を捉える強い可能性を示している。

潜在変数モデリングと拡散フレームワークの反復的なノイズ除去の精密化を効果的に統合する。このアプローチは、潜在空間表現と拡散ベースの生成技術を組み合わせる可能性を示している。Cheukら[25]は、DiffWaveアーキテクチャ[26]内のピアノロール形式を適応させることで、拡散モデルを音楽トランスクリプションに拡張した。彼らの研究は、生成よりも転写を主な対象としているが、条件付け信号としてグランドトゥルースのスペクトログラムを採用している。この設計は、記号音楽タスクにおける構造化条件付けの実現可能性を効果的に示し、きめ細かな外部制御をサポートするピアノロール表現を持つ拡散モデルを用いた記号音楽生成のための有望なアプローチを示唆する。

Polyfusion[12]は、条件付き記号音楽生成に拡散モデルを明示的に適用した最初のもので、生成モデリングと記号音楽表現の組み合わせにおいて大きな進歩を示した。これはLatent Diffusion [27]にインスパイアされたクロスアテンションメカニズムを組み込んでおり、従来のU-Netバックボーンをクロスアテンションレイヤーで補強し、ノイズ除去プロセス中に条件付け信号を中間潜在表現に投影する。この設計により、生成される出力を柔軟かつきめ細かく制御することができる。このアーキテクチャにより、Polyfusionはメロディ生成、和音条件付け、インペインティングを含む様々な記号音楽タスクにおいて高い制御性を実現している。このような制御は、構造的な一貫性と意味的な整合性を保つことが困難な複雑な生成シナリオにおいて、特に価値がある。

Polyfusionは条件生成において強力な性能を示すが、テキスト制御可能なタスクには拡張されていないため、自然言語プロンプトと記号的音楽作成を直接結びつけることはできない。その結果、自由形式テキストを条件とした記号音楽生成におけるDDPMのような拡散ベースのモデルの可能性は、ほとんど未解明のままである。Polyfusionのアプローチに触発され、我々のTPSMG手法は、拡散メカニズムを通して言語埋め込みを記号的な音楽表現に整合させることを目的とし、より表現力豊かで直感的、かつユーザーフレンドリーなテキストガイド付き音楽生成システムの作成に努めている。

3. 提案モデル

我々のTPSMGフレームワークは、音楽表現、モデルアーキテクチャ、学習戦略の3つの中核的な側面を扱う。第3節1では、拡散モデリングに合わせた音楽データの表現方法について述べる。第3.2節では、テキストコンディショニングモジュールで補強されたU-Netバックボーンからなるモデルアーキテクチャと、学習手順の包括的な説明を詳述する。TPSMGフレームワークの概要を図1に示す。

3.1 シンボリック音楽表現

音楽作品を2チャンネルの2値テンソル $v \in \mathbb{R}^{2 \times T} \times P$ として表現し、ピアノロールとして解釈できる。2つのチャンネルは音符のオンセットとサステイン音符を区別し、音符の開始と継続の両方を正確にモデル化することを可能にする。各時間ステップは16分音符(1/4拍)に対応し、細かい時間分解能を提供する。ピッチ次元Pは、C-1からG9までの128のMIDIピッチにまたがり、音楽で一般的に使用される標準的なMIDIレ

mediate representation of musical attributes. MuseCoco comprises a text-to-attribute understanding module followed by an attribute-to-music generation module. While this modular design facilitates practical and controllable generation, the requirement for separately trained components increases pipeline complexity and restricts end-to-end optimization. In contrast, Bhandari *et al.* [22] proposed Text2midi, a state-of-the-art, end-to-end model that directly generates MIDI-format symbolic music from free-form text prompts. Text2midi employs a frozen FLAN-T5 encoder to embed input captions and an autoregressively trained Transformer decoder to produce MIDI tokens, achieving superior performance compared to MuseCoco. The model is pre-trained in a semi-supervised manner using pseudo-captions from SymphonyNet [23] and fine-tuned on the MidiCaps dataset [24]. While Text2midi enables direct text-guided music generation, it still exhibits limited alignment between the generated music and input text, with inconsistencies frequently observed in key, mood, and genre. In contrast, our proposed TPSMG employs a DDPM-based architecture, which has demonstrated superior capability in incorporating external control signals, as detailed in the following subsection.

2.3 Diffusion-Based Models for Music Generation

Recently, diffusion models have attracted increasing attention for symbolic music generation, reflecting their promise in modeling complex, structured discrete data. Originally developed for continuous domains such as images and audio, diffusion models demonstrate strong potential in capturing the intricate distributions of symbolic musical events when combined with suitable representations and conditioning mechanisms.

, effectively integrating latent-variable modeling with the iterative denoising refinement of diffusion frameworks. This approach illustrates the potential of combining latent space representations with diffusion-based generation techniques. Cheuk *et al.* [25] extended diffusion models to music transcription by adapting the piano-roll format within the DiffWave architecture [26]. Although their work primarily targets transcription rather than generation, it employs ground-truth spectrograms as conditioning signals. This design effectively demonstrates the feasibility of structured conditioning in symbolic music tasks, suggesting a promising approach for symbolic music generation using diffusion models with piano-roll representations that support fine-grained external control. Polyffusion [12] was the

first to explicitly apply diffusion models to conditional symbolic music generation, marking a significant advance in combining generative modeling with symbolic music representation. It incorporates a cross-attention mechanism inspired by Latent Diffusion [27], augmenting the traditional U-Net backbone with cross-attention layers that project conditioning signals into intermediate latent representations during the denoising process. This design enables flexible and fine-grained control over the generated outputs. Owing to this architecture, Polyffusion achieves high controllability across various symbolic music tasks, including melody generation, chord conditioning, and inpainting. Such control is particularly valuable in complex generation scenarios where preserving structural coherence and semantic alignment is challenging.

While Polyffusion demonstrates strong performance in conditional generation, it has not been extended to text-controllable tasks and thus cannot directly link natural language prompts with symbolic music creation. As a result, the potential of diffusion-based models like DDPMs for symbolic music generation conditioned on free-form text remains largely unexplored. Inspired by Polyffusion’s approach, our TPSMG method aims to align language embeddings with symbolic music representations through diffusion mechanisms, striving to create a more expressive, intuitive, and user-friendly text-guided music generation system.

3. Proposed Model

Our *TPSMG* framework addresses three core aspects: music representation, model architecture, and training strategy. In Subsection 3.1, we describe our approach to representing music data tailored for diffusion modeling. Subsection 3.2 details the model architecture, comprising a U-Net backbone augmented with a text-conditioning module, along with a comprehensive explanation of the training procedure. An overview of the TPSMG framework is illustrated in Fig. 1.

3.1 Symbolic Music Representation

We represent a music piece as a 2-channel binary tensor $v \in \mathbb{R}^{2 \times T \times P}$, interpretable as a piano roll, where T denotes the temporal dimension and P the pitch dimension. The two channels distinguish between note onsets and sustained notes, enabling precise modeling of both the initiation and continuation of notes. Each time step corresponds to a sixteenth note (1/4 beat), providing fine temporal resolution. The pitch dimension P spans 128 MIDI pitches, from C-1 to G9, covering the full

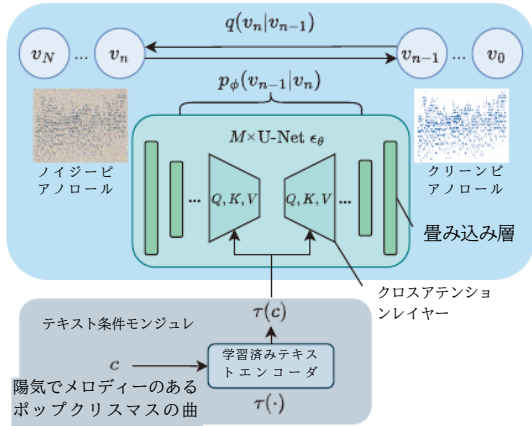


図1:TPSMGの概要。テキストキャプション c は、まずテキストエンコーダー $\tau(\cdot)$ によって処理され、埋め込み $\tau(c)$ を生成する。次に、埋め込みはクロスアテンションレイヤーに統合され、その後、学習プロセスを導くために中間レイヤーにマッピングされる。

ンジの全範囲をカバーする。

3.2 拡散公式

拡散過程をモデル化するために、2次元のU-Netを用いて記号音楽を生成する。我々のTPSMGは、順方向ノイズ処理 $q(v_n|v_{n-1})$ から始まり、複数のステップにわたって元のクリーンなピアノロール v_0 に徐々にノイズを加える：

$$q(v_n|v_{n-1}) = \mathcal{N}(\sqrt{1-\beta_n}v_{n-1}, \beta_n I), \quad (1)$$

$$q(v_n|v_0) = \prod_{n=0}^{N-1} q(v_n|v_{n-1}), \quad (2)$$

ここで、 N は拡散ステップの最大数を表し、 v_n と β_n はそれぞれ n 番目のステップにおけるピアノロールと分散スケジューリングパラメータの状態を表す。各ステップで追加されるノイズの独立性を維持するために、恒等行列 I が採用される。

この順方向ノイズ処理に続いて、モデル ϕ は逆方向演算 $p_\phi(v_{n-1}|v_n)$ を実行するように学習され、これは最終的なノイズバージョン $v_N \equiv N(0, I)$ から元のピアノロール v_0 を再構成することを目的とする：

$$p_\phi(v_{n-1}|v_n) = \mathcal{N}(\mu_\phi(v_n), \sigma_\phi(v_n)), \quad (3)$$

$$p_\phi(v_0) = p(v_N) \prod_{n=0}^{N-1} p_\phi(v_{n-1}|v_n), \quad (4)$$

ここで、 $\mu_\phi(v_n)$ と $\sigma_\phi(v_n)$ はそれぞれ v_n の平均と分散を表す。

ノイズ除去処理では、生成された出力を洗練させる際に、モデルを導くためにテキスト条件付け信号 c が導入される。

埋め込み $\tau(c)$ を、事前に学習されたテキストエンコーダを用いて行う。埋め込み $\tau(c)$ は、U-Netのデコーダ層内に埋め込まれたクロスアテンション層を介してモデルに組み込まれる。これらのクロスアテンションレイヤーは、テキストベースの情報をU-Netの中間特徴表現に投影し、テキスト条件がノイズ除去の軌跡に影響を与えることを可能にする。このガイダンスは、最適化プロセスを適宜ステアリングすることで、生成された音楽が入力テキストプロンプトと密接に整合することを保証する。学習手順は、以下のように定義される損失関数 $L(\phi)$ によって監督される：

$$L(\phi) = \mathbb{E}_{v_0, c, \delta, n} \left[\|\delta - \delta_\phi(\sqrt{\alpha_n}v_0 + \sqrt{1-\alpha_n}\delta, n, \tau(c))\|^2 \right], \quad (5)$$

この信号はまず、 δ が標準正規分布 $N(0, I)$ からサンプリングされたランダムに生成されたノイズを表し、 δ_ϕ はこのノイズを予測するために使用される関数である。パラメータ $\alpha_t = 1 - \beta_t$ と $\Delta_t = \sum_{k=1}^t \alpha_k$ を用いてノイズスケジュールを変調し、時間ステップ t は $[1, N]$ の範囲から一様にサンプリングされる。モデル δ_ϕ は学習中の損失 $L(\phi)$ を最小化するように最適化される。この損失は、生成プロセスを導くテキスト条件付け信号 $\tau(c)$ に大きく影響される。学習終了後、モデルはノイズの多い入力からきれいなピアノロールを効果的に再構成し、元の構図に酷似した記号的な音楽を生成し、提供されたテキストプロンプトとよく一致する。

4. 実験結果

4.1 実験セットアップ

MidiCapsデータセット[24]を用いて実験を行った。このデータセットには、168,000のMIDIファイルと、様々な音楽属性を記述したテキストキャプションが対になっている。データセットはトレーニングセット(80%)、検証セット(10%)、テストセット(10%)に分割された。時間的な粒度を細かくするため、時間分解能を四半期ノートごとに4つの時間ステップに設定した。データ補強は、 $s \sim U(-5, 6)$ ($s \in \mathbb{Z}$) の半音でランダムにピッチを転置し、 U は一様分布を表し、さらにロバスト性を高めるためにランダムな開始ビートを割り当てることによって適用された。拡散プロセスは1,000ステップで構成され、学習率 $5e-5$ のAdamオプティマイザ[28]を用いて100エポック学習された。MIDI変換のために、生成された2チャンネルピアノロールテンソルは、音楽の一貫性を確保するために、オンセットなしで音符を破棄し、0または1に丸めることによって2値化された。

TPSMGモデルの性能を評価するために、テストセットの全テキストキャプションの出力を生成した。

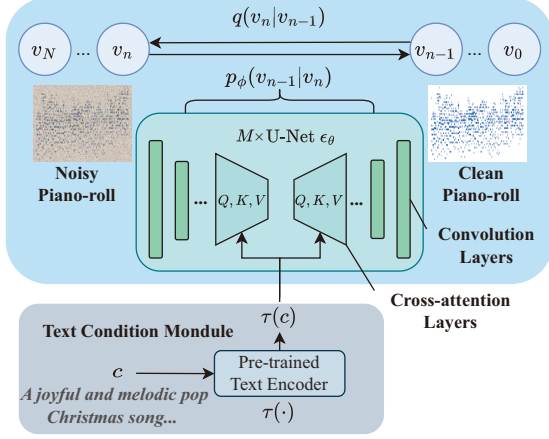


Fig. 1: Overview of TPSMG. The text caption c is first processed by the text encoder $\tau(\cdot)$, generating the embedding $\tau(c)$. The embeddings are then integrated into cross-attention layers and subsequently mapped into intermediate layers to guide the training process.

standard MIDI range commonly used in music.

3.2 Diffusion Formula

We generate symbolic music using a two-dimensional U-Net to model the diffusion process. Our TPSMG begins with a forward noising process $q(v_n|v_{n-1})$, which progressively adds noise to the original clean piano roll v_0 over multiple steps:

$$q(v_n|v_{n-1}) = \mathcal{N}(\sqrt{1 - \beta_n}v_{n-1}, \beta_n I), \quad (1)$$

$$q(v_n|v_0) = \prod_{n=0}^{N-1} q(v_n|v_{n-1}), \quad (2)$$

where N represents the maximum number of diffusion steps, and v_n and β_n denote the state of the piano roll and the variance scheduling parameter at the n -th step, respectively. The identity matrix I is employed to maintain the independence of the noise added at each step.

Following this forward noising process, the model ϕ is trained to execute the reverse operation $p_\phi(v_{n-1}|v_n)$, which aims to reconstruct the original piano roll v_0 from the final noisy version $v_N \sim \mathcal{N}(0, I)$:

$$p_\phi(v_{n-1}|v_n) = \mathcal{N}(\mu_\phi(v_n), \sigma_\phi(v_n)), \quad (3)$$

$$p_\phi(v_0) = p(v_N) \prod_{n=0}^{N-1} p_\phi(v_{n-1}|v_n), \quad (4)$$

where $\mu_\phi(v_n)$ and $\sigma_\phi(v_n)$ represent the mean and variance of v_n , respectively.

During the denoising process, a text-conditioning signal c is introduced to guide the model in refining the generated output. This signal is first encoded into an

embedding $\tau(c)$ using a pre-trained text encoder. The embedding $\tau(c)$ is then incorporated into the model via cross-attention layers embedded within the decoder layers of the U-Net. These cross-attention layers project the text-based information into the intermediate feature representations of the U-Net, allowing the text condition to influence the denoising trajectory. This guidance ensures that the generated music aligns closely with the input textual prompt by steering the optimization process accordingly. The training procedure is supervised by a loss function $L(\phi)$, defined as follows:

$$L(\phi) = \mathbb{E}_{v_0, c, \delta, n} \left[\|\delta - \delta_\phi(\sqrt{\alpha_n}v_0 + \sqrt{1 - \alpha_n}\delta, n, \tau(c))\|^2 \right], \quad (5)$$

where δ denotes randomly generated noise sampled from a standard normal distribution $\mathcal{N}(0, I)$, and δ_ϕ is a function used to predict this noise. The parameters $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{k=1}^t \alpha_k$ are used to modulate the noise schedule, and the time step t is uniformly sampled from the range $[1, N]$. The model δ_ϕ is optimized to minimize the loss $L(\phi)$ during training. This loss is heavily influenced by the text-conditioning signal $\tau(c)$, which guides the generation process. Upon completion of training, the model effectively reconstructs a clean piano roll from a noisy input, producing symbolic music that closely resembles the original composition and aligns well with the provided text prompt.

4. Experiments

4.1 Experimental Setup

We conducted experiments using the MidiCaps dataset [24], which contains 168,000 MIDI files paired with text captions describing various musical attributes. The dataset was split into training (80%), validation (10%), and testing (10%) sets. To ensure fine temporal granularity, we set the temporal resolution to four time steps per quarter note. Data augmentation was applied by randomly transposing pitches by $s \sim U(-5, 6)$ ($s \in \mathbb{Z}$) semitones, where U denotes a uniform distribution, and by assigning randomized starting beats to further enhance robustness. The diffusion process was configured with 1,000 steps, and the model was trained for 100 epochs using the Adam optimizer [28] with a learning rate of $5e-5$. For MIDI conversion, the generated 2-channel piano roll tensors were binarized by rounding values to 0 or 1, discarding any notes without onsets to ensure musical coherence.

To evaluate the performance of our TPSMG model, we generated outputs for all text captions in the test

TPSMGを代表的な7つの音楽生成手法と比較した。音響領域でのテキストから音楽への生成には、2つの最先端手法を含めた: Mou[^]sai [29]とMustango [30]である。記号領域では、Text2midi [22]をベースラインとして選択し、MuseCocoやテキスト制御可能な記号音楽生成における以前の手法よりも優れた性能を持つことが認識された。TPSMGによって生成された記号音楽の品質をさらに評価するために、4つの最先端の記号音楽生成モデルも含めた: マルチトラック音楽機械(MMM) [31]、Final Generation of ARranged Orchestration (FIGARO) [32]、MMT [16]、MMT-Bidirectional Encoder Representations from Transformers (BERT) [33]である。さらに、異なるテキストエンコーダ-BERT [34]、T5 [35]、ADA-002 [36]がTPSMGの性能に与える影響を調べるアブレーション研究を行った。提案手法の美的性能をさらに評価するために、Text2midi [22]を比較手法として、18人の参加者を対象に主観的リスニングテストを実施した。

4.2 評価基準

我々は、3つの定量的基準に基づいてモデルの性能を評価した: 精度、キャプションの類似性、および音楽の質。精度。精度の指標は、生成された音楽が、キー、ムード、ジャンルの観点から、テキストプロンプトとどの程度一致しているかを評価する。精度が高いほど、キャプション情報をより忠実に捉えることができる。グランドトゥールスキーはスカラー値として提供され、ムードとジャンルはMidiCapsデータセットからの確率分布として表現される。生成された音楽について、キーはMusic21 [37]とMido *を用いて推定され、気分とジャンルはESSENTIA [38]を用いて予測される。キー精度は、推定キーがグランドトゥールスキーと完全に一致する生成サンプルの割合として計算される。気分とジャンルは確率分布であり、意味的に関連する項が重なることがあるため、それらの精度は埋め込みg aとg bの間の余弦類似度を平均化することによって測定される。ここで、g aは真実の気分またはジャンルの埋め込み、g bは潜在空間における最も近いマッチング生成気分またはジャンルの埋め込みである。なお、精度評価はMidiCapsで学習したテキストから音楽へのモデルにのみ適用できるため、異なるテキストエンコーダとText2midiを用いた3つのTPSMGバリエーションに限定される。

キャプションの類似度。キャプションの類似度は、入力されたキャプションの埋め込みと、生成された音楽作品に対して生成されたキャプションとの間の余弦類似度として計算された。生成された記号音楽のキャプションを得るために、MidiCaps [24]で紹介されたMIDI-to-textモデルを利用した。この指標は、生成された音楽が提供されたテキストプロンプトとどの程度一致するかを評価するもので、本研究の重要な目的である。音声ベースの手法Mou[^]sai [29]とMustango [30]については、音声テキストモデルLP-MusicCaps [39]を用いてキャプションを生成した。音声ベースの手法とMIDIベースの手法では、異なるキャプションモデルを使用しているが、完全性を期すために、両者でこの実験を行った。

品質。生成された音楽の品質を評価するために、JazzTransformer [40]によって導入された、記号的音楽生成分野から広く使われている3つのメトリクスを採用した: ピッチクラスエントロピー(PCE)、スケール一貫性(SC)、グルーヴ一貫性(GC)。これらのメトリクスの絶対値は音楽の質を直接示すものではないので、ベンチマークとして人間の作曲から計算された対応する値と比較する。具体的には、MidiCapsの人間の構成と生成された出力の両方について、平均PCE、SC、GCを計算し、これらの値間の類似度を測定する。この結果、ピッチクラスエントロピー類似度(PCES)、スケール一貫性類似度(SCS)、グルーヴ一貫性類似度(GCS)が得られ、類似度が高いほど、生成された音楽がこれらの側面において人間の構図により近いことを示す。PCE、SC、GCの計算には記号情報が必要であり、音声では利用できないため、Mou[^]saiとMustangoが生成した出力に対してこれらのメトリクスを計算することはできなかった。主観的な測定基準。主観的な実験では、参加者に以下の5つの質問をした。

requested that they rated each on a five-point scale.

- リッチネス(R): 音楽は和声やリズムに多様性を示すか?
- 忠実度(F): 与えられたプロンプトはどの程度音楽に反映されているか?
- 構造的ー貫性(S): 音楽は、繰り返されるテーマや音楽的アイデアの展開など、明確な構造的パターンを示すか?
- 全体(0): 以上の点から、楽曲の総合的な評価はどのようなものですか?

4.3 評価結果

における精度の定量的評価結果

キー、ムード、ジャンルを表1に示す。として

*/github.com/bzamecnik/midi2audio

set. We compared TPSMG against seven representative music generation methods. For text-to-music generation in the acoustic domain, we included two state-of-the-art methods: Moûsai [29] and Mustango [30]. In the symbolic domain, we selected Text2midi [22] as a baseline, recognized for its superior performance over MuseCoco and earlier methods in text-controllable symbolic music generation. To further assess the quality of symbolic music generated by TPSMG, we also included four state-of-the-art symbolic music generation models: Multitrack Music Machine (MMM) [31], FInal Generation of ARranged Orchestration (FIGARO) [32], MMT [16], and MMT-Bidirectional Encoder Representations from Transformers (BERT) [33]. Additionally, we performed an ablation study examining the influence of different text encoders—BERT [34], T5 [35], and ADA-002 [36]—on TPSMG’s performance. To further assess the aesthetic performance of the proposed method, we conducted a subjective listening test with 18 participants, using Text2midi [22] as the comparative method.

4.2 Evaluation Criteria

We evaluated model performance based on three quantitative criteria: accuracy, caption similarity, and music quality.

Accuracy. The accuracy metric assesses how well the generated music aligns with the text prompts in terms of key, mood, and genre. Higher accuracy indicates better fidelity in capturing the caption information. The ground-truth key is provided as a scalar value, while mood and genre are represented as probability distributions from the MidiCaps dataset. For the generated music, the key is estimated using Music21 [37] and Mido*, whereas mood and genre are predicted using ESSENTIA [38]. Key accuracy is computed as the proportion of generated samples whose estimated key exactly matches the ground-truth key. Since mood and genre are probability distributions and semantically related terms may overlap, their accuracy is measured by averaging the cosine similarity between embeddings $\mathbf{g}^a$ and $\mathbf{g}^b$, where $\mathbf{g}^a$ is the embedding of the ground-truth mood or genre, and $\mathbf{g}^b$ is the embedding of the closest matching generated mood or genre in the latent space. Note that accuracy evaluation is only applicable to text-to-music models trained on MidiCaps and is thus limited to the three TPSMG variants using different text encoders and Text2midi.

Caption similarity. Caption similarity was calculated as the cosine similarity between the embeddings of the input caption and the caption generated for the produced music piece. To obtain captions for the generated symbolic music, we utilized the MIDI-to-text model introduced with MidiCaps [24]. This metric evaluates how closely the generated music matches the provided text prompts, which is a key objective of this study. For the audio-based methods Moûsai [29] and Mustango [30], captions were generated using the audio-to-text model LP-MusicCaps [39]. Although audio-based methods and MIDI-based methods use different captioning models, we conducted this experiment with both for the sake of completeness.

Quality. To assess the quality of the generated music, we adopted three widely used metrics from the symbolic music generation field, originally introduced by JazzTransformer [40]: pitch class entropy (PCE), scale consistency (SC), and groove consistency (GC). Because the absolute values of these metrics do not directly indicate music quality, we compare them to the corresponding values calculated from human compositions as a benchmark. Specifically, we compute the average PCE, SC, and GC for both the human compositions in MidiCaps and the generated outputs, then measure the similarity between these values. This results in pitch class entropy similarity (PCES), scale consistency similarity (SCS), and groove consistency similarity (GCS), where higher similarity indicates that the generated music more closely resembles human compositions in these aspects. Since the calculation of PCE, SC, and GC requires symbolic information, which is not available for audio, these metrics could not be computed for the outputs generated by Moûsai and Mustango.

Subjective metrics. In subjective experiment, we asked the participants the following five questions and requested that they rated each on a five-point scale.

- **Richness (R):** Does the music exhibit diversity in harmony and rhythm?
- **Fidelity (F):** To what extent is the given prompt reflected in the music?
- **Structural Coherence (S):** Does the music display clear structural patterns such as recurring themes or the development of musical ideas?
- **Overall (O):** Based on the above aspects, what is your overall assessment of the music piece?

4.3 Evaluation Results

The quantitative evaluation results for accuracy in key, mood, and genre are presented in Table 1. As

* <https://github.com/bzamecnik/midi2audio>

表1: キー、ジャンル、ムードに対する精度結果。

	Key Accuracy $\uparrow$ (%)	Mood Accuracy $\uparrow$ (%)	Genre Accuracy $\uparrow$ (%)
TPSMG (BERT)	52.43 $\pm$ 33.63	91.99 $\pm$ 8.28	92.69$\pm$8.74
TPSMG (T5)	60.53$\pm$40.37	92.29$\pm$8.09	91.93 $\pm$ 8.92
TPSMG (ADA-002)	54.36 $\pm$ 33.07	91.74 $\pm$ 8.20	91.92 $\pm$ 8.79
Text2midi [22]	35.88 $\pm$ 29.58	73.45 $\pm$ 11.58	64.98 $\pm$ 15.53

表2: 主観的リスニングテストの結果各指標は5段階で評価され、平均点が算出される。

	R	F	S	O
Text2midi [22]	2.56 $\pm$ 1.04	2.48 $\pm$ 0.88	2.18 $\pm$ 0.68	2.58 $\pm$ 0.94
TPSMG	3.48$\pm$0.95	3.31$\pm$0.97	3.49$\pm$0.82	3.67$\pm$0.69

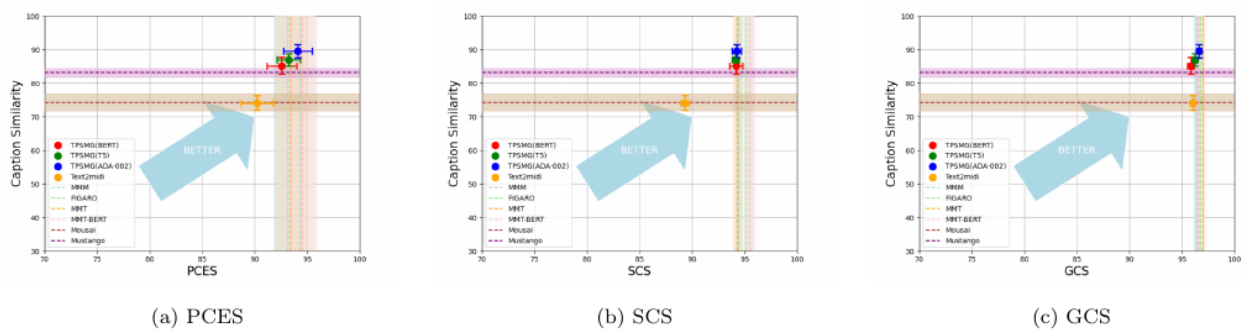


図2: キャプションの類似度と音楽の質に関する定量的な結果。(a)、(b)、(c)はそれぞれ異なる横軸-PCES、SCS、GCSを持つ3つの異なる空間における結果の分布を示す。右上に近いデータポイントは、より良いとみなされる。破線はCMを表し、薄い色の部分は標準偏差の範囲を示す。

TPSMGの3つのバージョンはすべて、90%を超えるムードとジャンルの正確さを達成しており、これらの音楽属性を捉えるモデルの強力な能力を示している。TPSMG(T5)は、どのバージョンも気分とジャンルの精度において同等の結果を出す、キー精度において比較的高い性能を示す。注目すべきは、3つの評価指標すべてにおいて、3つのTPSMGバージョンすべてが最先端のText2midi手法を上回り、TPSMGの有効性を実証していることである。キー精度が低いのは、キー情報がテキストプロンプトのごく一部に過ぎないことが多いことに起因している可能性がある。このため、モデルはこの側面を十分に捉えず、符号化することが難しく、観測されたキー精度の低下を説明できる可能性がある。さらに、キー精度の標準偏差が比較的高いのは、各予測が正しいか正しくないかに分類されるというメトリックの二値性に起因する可能性があり、それによって集約された結果にばらつきが生じる。音楽の質とキャプションの類似性を比較するために、さらにこれらの関係を検証する。

つの重要な次元がある。

具体的には、平均キャプション類似度と3つの異なる品質メトリクスを組み合わせ、3つの2D可視化空間を作成し、それぞれ図2(a)、(b)、(c)に示す。MMM、FIGARO、MMT、MMTBERTではキャプションの類似性データが利用できず、Mou'saiとMustangoでは品質メトリクスが欠落しているため、公平で有益な視覚的比較を保証するために、水平線と垂直線を用いてこれらのベースラインを表現する。音楽の品質の観点から、図2は、TPSMGによって生成された音楽が、Text2midiのそれを上回るだけでなく、ピッチ、スケール、グルーブなどの主要な音楽次元において、テキスト制御性を組み込んでいない他の最先端の記号的音楽生成手法に匹敵する結果を達成していることを明確に示している。このことは、音楽的に首尾一貫した美的感覚を持つコンポジションを生み出すTPSMGの実用性と頑健性の可能性を強調している。キャプションの類似性については、TPSMGの3つのバージョンはすべて、音響領域と記号領域の両方で既存のモデルを一貫して上回り、

Table 1: Accuracy results for key, genre, and mood.

	Key Accuracy $\uparrow$ (%)	Mood Accuracy $\uparrow$ (%)	Genre Accuracy $\uparrow$ (%)
TPSMG (BERT)	52.43 $\pm$ 33.63	91.99 $\pm$ 8.28	92.69$\pm$8.74
TPSMG (T5)	60.53$\pm$40.37	92.29$\pm$8.09	91.93 $\pm$ 8.92
TPSMG (ADA-002)	54.36 $\pm$ 33.07	91.74 $\pm$ 8.20	91.92 $\pm$ 8.79
Text2midi [22]	35.88 $\pm$ 29.58	73.45 $\pm$ 11.58	64.98 $\pm$ 15.53

Table 2: Subjective listening test results. Each metric is rated on a five-point scale, with average score calculated.

	R	F	S	O
Text2midi [22]	2.56 $\pm$ 1.04	2.48 $\pm$ 0.88	2.18 $\pm$ 0.68	2.58 $\pm$ 0.94
TPSMG	3.48$\pm$0.95	3.31$\pm$0.97	3.49$\pm$0.82	3.67$\pm$0.69

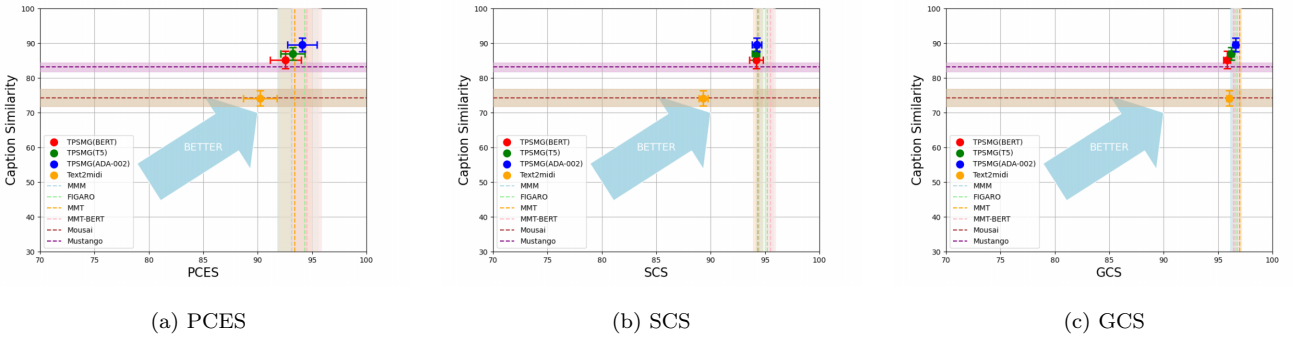


Fig. 2: Quantitative results of caption similarity and music quality. (a), (b), and (c) show the distribution of results in three different spaces, each with a different horizontal axis—PCES, SCS, and GCS, respectively. Data points closer to the upper right are viewed as better. Dashed lines represent CMs, and light-colored areas indicate the range of standard deviation.

shown, all three versions of TPSMG achieve mood and genre accuracy rates exceeding 90%, indicating the model’s strong capability in capturing these musical attributes. Among the variants, TPSMG (T5) shows relatively higher performance in key accuracy, while all versions produce comparable results in mood and genre accuracy. Notably, all three TPSMG versions outperform the state-of-the-art Text2midi method across all three evaluation metrics, demonstrating TPSMG’s effectiveness. The lower performance in key accuracy may stem from the fact that key information often represents only a small portion of the textual prompts. This likely causes the model to inadequately capture and encode this aspect, which could explain the observed decrease in key accuracy. Additionally, the relatively high standard deviation in key accuracy can be attributed to the binary nature of the metric, where each prediction is classified as either correct or incorrect, thereby introducing variability in the aggregated results. To compare music quality and caption similarity, we further examine the relationship between these

two important dimensions.

Specifically, we combine the average caption similarity with three distinct quality metrics to create three 2D visualization spaces, which are shown in Fig. 2 (a), (b), and (c), respectively. Since caption similarity data is unavailable for MMM, FIGARO, MMT, and MMT-BERT, and quality metrics are missing for Moûsai and Mustango, we represent these baselines using horizontal and vertical lines to ensure a fair and informative visual comparison. From the perspective of music quality, Fig. 2 clearly demonstrates that the music generated by TPSMG not only surpasses that of Text2midi but also achieves results comparable to other state-of-the-art symbolic music generation methods that do not incorporate text controllability, across key musical dimensions such as pitch, scale, and groove. This underscores the potential practicality and robustness of TPSMG in producing musically coherent and aesthetically pleasing compositions. Regarding caption similarity, all three versions of TPSMG consistently outperform existing models in both acoustic and symbolic do-

Ex. 1

D minorのリラックスしたポップソングで、Viva ceテンポと4/4タイムシグネチャーを特徴とする。メロディの構成は、愛や感情を呼び起こすので、映画での使用に適している。Fdim、F6、C6、F6、Fmaj7のコード進行は、作品に深さを加える。



ポップ、エレクトロニクス、ロック、サウンドトラックのタッチのハメミックスで、この曲は幸せてエネルギッシュなムードを愛のタッチで包含し、映画に完全に適している。120bpmの適度なテンポで構成され、4/4の時間シグネチャーで動作し、D minorの鍵にセットされる。具体的な楽器や和音進行がないにもかかわらず、そのジャンルやムード・タグは、さまざまな感情を伝える魅力的で活気ある作品を約束する。

Ex. 2

メロディックな電子・ポップクリスマス曲ピアノ、弦楽アンサンブル、オーボエ、スチールドラム、オーバードライブギターを備え、Cメジャーの鍵盤に4/4時間シグネチャーと93拍/分のモデレートテンポでセットされ、その時間を通して幸せでリラックスした雰囲気を作り出している。



このメロディーとエネルギッシュな曲の中で、電子ポップ、クラシックエレメント、楽しいサウンドトラックの味が完璧に融合していることを発見する。93BPMの適度なテンポに設定された幸せでありながらリラックスした気分を体現し、この4/4時間の特徴的なピースはC Majorの鍵で構成されている。その魅惑的なリズムは、最初から最後まであなたを魅了することだろう。

Ex. 3

フルート、アコーディオン、シンセベース、弦楽アンサンブル、シンセリードの活気あるブレンドを特徴とする、エプメジャーのエネルギッシュな電子ダンストラック。レトロな魅力をほめめかすメロディーと夢のような雰囲気を作り出し、速いテンポと4/4時間の特徴によって推進され、Eb、Bb、Abの和音進行がその魅力的な魅力に加わる。



エネルギッシュなメロディー。ロックツイストを搭載した、幸せてエネルギッシュな電子ポップサウンドトラック。Eb Majorで構成されるこのcaptivatingメロディーは、All egoテンポで4/4時間シグネチャーに従う。その高揚感にもかかわらず、微妙に暗いテーマを示唆している。魅力的で映画的な気分を設定することを完璧にする。

図3: TPSMG(ADA-002)から生成されたサンプルの例。生成された3つのサンプルは、すべてプロンプトで指定されたキーとテンポに正確に沿ったものである。また、生成されたキャプションは、サンプルが期待されるジャンルや気分を正確に反映していることを示している。ハイライトされた部分は、テキストプロンプトと生成されたキャプションの間の類似した部分を示す。

提供されたテキスト入力と意味的に一致する音楽を生成するTPSMGの能力を裏付けている。主観的リスニングテストの結果を表2に示す。表2より、TPSMGはText2midiと比較して、全てのメトリクスで一貫して高いスコアを達成していることがわかる。このことは、反復的なノイズ除去プロセスが、繰り返されるテーマや和声の発達といった首尾一貫した音楽構造を保持するのに対し、Text2midiはしばしば不均衡な声や首尾一貫性の喪失に悩まされることを示唆している。さらに、より多様なハーモニーやリズムの生成も可能である。一方、忠実度スコアの高さも結果に反映されており、本手法が気分とジャンルの分類において90%以上の精度を達成していることがわかる。Text2midiのスコアが低いのに比べ、TPSMGは意図したテキスト記述をより忠実に捉えていることがわかる。

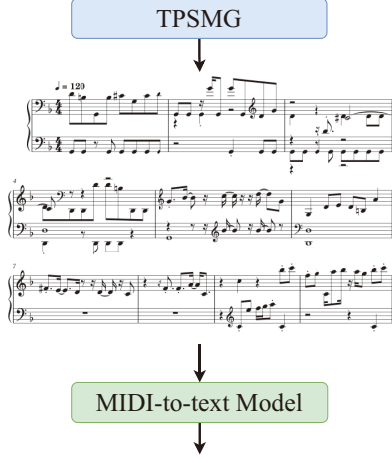
さらに、生成された音楽ス

コアの定性的な例と、それに対応するテキスト記述を図3に示す。これらの例は、TPSMG(ADA002)によって生成された出力が、ジャンル、ムード、その他の文体要素など、意図された音楽属性と明確な整合性を示すことを示している。さらに、TPSMGの3つのバージョンはすべてText2midiを一貫して上回り、ユーザーの入力プロンプトを忠実に反映した記号音楽を生成するモデルの強力な能力を強調している。これらの視覚的な例は、TPSMGがテキストプロンプトを効果的に記号表現に変換し、望ましい音楽的特徴を保持し伝えるというさらなる証拠を提供する。

これらの結果を総合すると、TPSMGは高品質な記号音楽を生成することに長けているだけでなく、自然言語プロンプトに埋め込まれたニュアンスの特徴を捉え、反映することにも優れていることが示唆される。このことは、TPSMGがテキストから音楽

Ex. 1

A relaxing **pop song in D minor**, characterized by a **Vivace tempo** and a **4/4 time signature**. The melodic composition **evokes feelings of love and emotion**, making it **suitable for use in films**. The chord progression of Fdim, F6, C6, F6, and Fmaj7 adds depth to the piece.



A melodious mix of **pop**, electronic, rock, and a touch of **soundtrack**, this song encapsulates happy and energetic moods with a **touch of love** and is **perfectly suited for film**. Composed in a moderate tempo of 120 bpm, it operates in a **4/4 time signature** and is set in the key of **D minor**. Despite the lack of specific instruments and chord progressions, its genre and mood tags promise a charming, vibrant piece that conveys a range of emotions.

Ex. 2

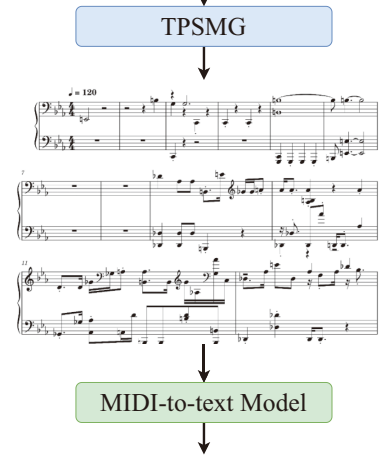
A melodic electronic and pop Christmas song featuring piano, string ensemble, oboe, steel drums, and overdriven guitar, **set in the key of C major** with a **4/4 time signature** and a **moderate tempo of 93 beats per minute**, creating a **happy and relaxing atmosphere** throughout its duration.



Discover the perfect blend of **electronic pop**, classical elements, and a delightful **soundtrack** flavor in this melodic and energetic song. Embodying a **happy yet relaxing mood** set to a moderate tempo of 93 BPM, this **4/4 time signature** piece is composed in the key of **C Major**. Its enchanting rhythm is sure to captivate you from beginning to end.

Ex. 3

An **energetic electronic dance track in Eb major**, featuring a vibrant blend of flute, accordion, synth bass, string ensemble, and synth lead, creating a melodic and dreamy atmosphere with a hint of retro charm, **propelled by a fast tempo and a 4/4 time signature**, with a chord progression of Eb, Bb, and Ab adding to its captivating appeal.



Energetically Melodic: A **Happy, Energetic Electronic Pop Soundtrack** with a **Rock Twist**. This captivating melody, composed in **Eb Major**, follows a **4/4 time signature** at an **Allegro** tempo. Despite its uplifting tone, it subtly hints at darker themes. Perfect for setting an engaging, cinematic mood.

Fig. 3: Examples of the generated samples from TPSMG (ADA-002). All three generated samples precisely align with the key and tempo specified in their prompts. The generated captions also indicate that the samples accurately reflect the expected genre and mood. The highlighted part denotes similar parts between text prompts and generated captions.

mains, confirming TPSMG's ability to generate music that semantically aligns with the provided text inputs.

The result of subjective listening test is shown in Table 2. Table 2 indicates that TPSMG achieves consistently higher scores across all metrics compared to Text2midi. This suggests that the iterative denoising process preserves coherent musical structures, such as recurring themes and harmonic development, whereas Text2midi often suffers from imbalanced voices and a loss of coherence. Moreover, it enables the generation of more diverse harmonies and rhythms. On the other hand, the high fidelity score is also reflected in the results, showing that our method achieves over 90% accuracy in mood and genre classification. Compared to the lower scores of Text2midi, this demonstrates that TPSMG more faithfully captures the intended textual descriptions.

Furthermore, qualitative examples of generated mu-

sic scores alongside their corresponding textual descriptions are presented in Fig. 3. These examples illustrate that the outputs produced by TPSMG (ADA-002) exhibit clear alignment with the intended musical attributes, such as genre, mood, and other stylistic elements. Additionally, all three versions of TPSMG consistently outperform Text2midi, highlighting the model's strong ability to generate symbolic music that closely reflects the user's input prompts. These visual examples provide further evidence that TPSMG effectively translates text prompts into symbolic representations that preserve and convey the desired musical characteristics.

Taken together, these results suggest that TPSMG is not only proficient in generating high-quality symbolic music but also excels at capturing and reflecting the nuanced features embedded within natural language prompts. This demonstrates that TPSMG holds sig-

への生成システムとして大きな可能性を秘めていることを示し、音楽の自動生成における将来の研究開発のための強固な基盤を確立するものである。

5. むすび

本研究では、TPSMGと名付けた、テキスト制御可能な新しい記号的音楽生成法を提案した。我々のアプローチは、テキストキャプションを潜在空間に埋め込み、これらの埋め込みをU-Netアーキテクチャのクロスアテンションレイヤーに統合する。この情報を中間層にマッピングすることで、学習中にモデルはテキストプロンプトに沿った記号的な音楽を生成するように導かれる。実験結果は、TPSMGがテキスト入力に優れたアライメントを持つ出力を効果的に生成し、気分やジャンル情報を正確に捉え、全体的に高い音楽品質を維持することを示している。今後の課題として、DDPMに基づく手法をマルチトラック音楽生成に拡張すること、現在のピアノロール表現には明示的な楽器情報が欠けていること、が残されている。現在の焦点は制御性を優先しており、マルチトラック機能を犠牲にしている。今後、マルチトラック生成のためのDDPMベースのモデルを開発するのか、それとも変換器ベースのモデルの制御性を高めるのかが未解決の問題であり、最終的には、完全にテキスト制御可能なマルチトラック記号音楽生成を実現することを目指している。

謝辞

本研究の一部は、日本学術振興会科研費 JP23K21676、JP23K11141、JP23KJ0044 の助成を受けた。英文校閲を担当したエナゴ(www.enago.jp)に感謝する。

References

- Frederick P Brooks, AL Hopkins, Peter G Neumann, and William V Wright. An experiment in musical composition. *IRE Transactions on Electronic Computers*, (3):175–182, 1957.
- Weiming Liu. Literature survey of multi-track music generation model based on generative confrontation network in intelligent composition. *The Journal of Supercomputing*, 79(6):6560–6582, 2023.
- Jean-Pierre Briot and François Pachet. Deep learning for music generation: challenges and directions. *Neural Computing and Applications*, 32(4):981–993, 2020.
- Shulei Ji, Jing Luo, and Xinyu Yang. A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions. *arXiv preprint arXiv:2011.06801*, 2020.
- James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023.
- Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pages 4055–4075, 2023.
- Jean-Pierre Briot and François Pachet. Music generation by deep learning-challenges and directions. *arXiv preprint arXiv:1712.04371*, 2017.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6007–6017, 2023.
- Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18381–18391, 2023.
- Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 16784–16804, 2022.
- Lejun Min, Junyan Jiang, Gus Xia, and Jingwei Zhao. Polyfusion: A diffusion model for polyphonic score generation with internal and external controls. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 231–238, 2023.
- Ziyu Wang, Lejun Min, and Gus Xia. Whole-song hierarchical generation of symbolic music using cascaded diffusion models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.
- Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. Symbolic music generation with diffusion models. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 468–475, 2021.
- Yu-Siang Huang and Yi-Hsuan Yang. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pages 1180–1188, 2020.
- Hao-Wen Dong, Ke Chen, Shlomo Dubnov, Julian McAuley, and Taylor Berg-Kirkpatrick. Multitrack music transformer. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.
- Ian Simon, Dan Morris, and Sumit Basu. Mysong: automatic accompaniment generation for vocal melodies. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 725–734, 2008.
- Chris Donahue, Huanru Henry Mao, Yiting Ethan Li, Garrison W Cottrell, and Julian McAuley. Lakhnes: Improving multi-instrumental music generation with cross-domain pre-training. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 685–692, 2019.
- Zeqian Ju, Peiling Lu, Xu Tan, Rui Wang, Chen Zhang, Songruoyao Wu, Kejun Zhang, Xiangyang Li, Tao Qin, and Tie-Yan Liu. Telemelody: Lyric-to-melody generation with a template-based two-stage method. *arXiv preprint arXiv:2109.09617*, 2021.
- Jingwei Zhao and Gus Xia. Accomontage: Accompaniment arrangement via phrase selection and style transfer. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 833–840, 2021.
- 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.
- Keshav Bhandari, Abhinaba Roy, Kyra Wang, Geeta Puri, Simon Colton, and Dorian Herremans. Text2midi: Generating symbolic music from captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23478–23486, 2025.
- Jiafeng Liu, Yuanliang Dong, Zehua Cheng, Xinran Zhang, Xiaobing Li, Feng Yu, and Maosong Sun. Symphony generation with permutation invariant language model. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 551–558, 2022.
- Jan Melechovsky, Abhinaba Roy, and Dorian Herremans.

nificant potential as a text-to-music generation system and establishes a solid foundation for future research and development in automatic music creation.

5. Conclusion

In this study, we proposed a novel text-controllable symbolic music generation method, named TPSMG. Our approach embeds text captions into a latent space and integrates these embeddings into the cross-attention layers of a U-Net architecture. By mapping this information into intermediate layers, the model is guided during training to generate symbolic music aligned with textual prompts. Experimental results demonstrate that TPSMG effectively produces outputs with superior alignment to text inputs, accurately captures mood and genre information, and maintains high overall musical quality. For future work, a significant challenge remains in extending DDPM-based methods to multitrack music generation, as the current piano-roll representation lacks explicit instrument information. Our current focus prioritizes controllability, which comes at the expense of multitrack capabilities. Going forward, it is an open question whether to develop DDPM-based models for multitrack generation or to enhance transformer-based models' controllability, with the ultimate goal of achieving fully text-controllable multitrack symbolic music generation.

Acknowledgements

This work was partly supported by JSPS KAKENHI Grant Numbers JP23K21676, JP23K11141, and JP23KJ0044. The authors would like to thank Enago (www.enago.jp) for the English language review.

References

- Frederick P Brooks, AL Hopkins, Peter G Neumann, and William V Wright. An experiment in musical composition. *IRE Transactions on Electronic Computers*, (3):175–182, 1957.
- Weiming Liu. Literature survey of multi-track music generation model based on generative confrontation network in intelligent composition. *The Journal of Supercomputing*, 79(6):6560–6582, 2023.
- Jean-Pierre Briot and François Pachet. Deep learning for music generation: challenges and directions. *Neural Computing and Applications*, 32(4):981–993, 2020.
- Shulei Ji, Jing Luo, and Xinyu Yang. A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions. *arXiv preprint arXiv:2011.06801*, 2020.
- James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023.
- Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pages 4055–4075, 2023.
- Jean-Pierre Briot and François Pachet. Music generation by deep learning-challenges and directions. *arXiv preprint arXiv:1712.04371*, 2017.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- Bahjat Kavar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6007–6017, 2023.
- Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18381–18391, 2023.
- Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 16784–16804, 2022.
- Lejun Min, Junyan Jiang, Gus Xia, and Jingwei Zhao. Polyfusion: A diffusion model for polyphonic score generation with internal and external controls. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 231–238, 2023.
- Ziyu Wang, Lejun Min, and Gus Xia. Whole-song hierarchical generation of symbolic music using cascaded diffusion models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.
- Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. Symbolic music generation with diffusion models. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 468–475, 2021.
- Yu-Siang Huang and Yi-Hsuan Yang. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pages 1180–1188, 2020.
- Hao-Wen Dong, Ke Chen, Shlomo Dubnov, Julian McAuley, and Taylor Berg-Kirkpatrick. Multitrack music transformer. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.
- Ian Simon, Dan Morris, and Sumit Basu. Mysong: automatic accompaniment generation for vocal melodies. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 725–734, 2008.
- Chris Donahue, Huanru Henry Mao, Yiting Ethan Li, Garrison W Cottrell, and Julian McAuley. Lakhnes: Improving multi-instrumental music generation with cross-domain pre-training. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 685–692, 2019.
- Zeqian Ju, Peiling Lu, Xu Tan, Rui Wang, Chen Zhang, Songruoyao Wu, Kejun Zhang, Xiangyang Li, Tao Qin, and Tie-Yan Liu. Telemelody: Lyric-to-melody generation with a template-based two-stage method. *arXiv preprint arXiv:2109.09617*, 2021.
- Jingwei Zhao and Gus Xia. Accomontage: Accompaniment arrangement via phrase selection and style transfer. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 833–840, 2021.
- 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.
- Keshav Bhandari, Abhinaba Roy, Kyra Wang, Geeta Puri, Simon Colton, and Dorian Herremans. Text2midi: Generating symbolic music from captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23478–23486, 2025.
- Jiafeng Liu, Yuanliang Dong, Zehua Cheng, Xinran Zhang, Xiaobing Li, Feng Yu, and Maosong Sun. Symphony generation with permutation invariant language model. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 551–558, 2022.
- Jan Melechovsky, Abhinaba Roy, and Dorian Herremans.

- MidiCaps—a large-scale MIDI dataset with text captions. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 858–865, 2024.
- 25) Kin Wai Cheuk, Ryosuke Sawata, Toshimitsu Uesaka, Naoki Murata, Naoya Takahashi, Shusuke Takahashi, Dorien Herremans, and Yuki Mitsufuji. Diffroll: Diffusion-based generative music transcription with unsupervised pretraining capability. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.
 - 26) Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.
 - 27) Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.
 - 28) P Kingma Diederik. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.
 - 29) Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. Moúsa: Efficient text-to-music diffusion models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 8050–8068, 2024.
 - 30) Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. Mustango: Toward controllable text-to-music generation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 8286–8309, 2024.
 - 31) Jeff Ens and Philippe Pasquier. MMM: Exploring conditional multi-track music generation with the transformer. *arXiv preprint arXiv:2008.06048*, 2020.
 - 32) Dimitri von Rütte, Luca Biggio, Yannic Kilcher, and Thomas Hofmann. FIGARO: Generating symbolic music with fine-grained artistic control. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
 - 33) Jinlong Zhu, Keigo Sakurai, Ren Togo, Takahiro Ogawa, and Miki Haseyama. MMT-BERT: Chord-aware symbolic music generation based on multitrack music transformer and musicbert. pages 470–477, 2024.
 - 34) Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, page 2, 2019.
 - 35) Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
 - 36) Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.
 - 37) Michael Scott Cuthbert and Christopher Ariza. Music21: A toolkit for computer-aided musicology and symbolic music data. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 637–642, 2010.
 - 38) Dmitry Bogdanov, Nicolas Wack, Emilia Gómez Gutiérrez, Sankalp Gulati, Herrera Boyer, Oscar Mayor, Gerard Roma Trepas, Justin Salamon, José Ricardo Zapata González, Xavier Serra, et al. ESSENTIA: An audio analysis library for music information retrieval. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 493–498, 2010.
 - 39) SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. Lp-musicaps: Llm-based pseudo music captioning. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 409–416, 2023.

- 40) Shih-Lun Wu and Yi-Hsuan Yang. The jazz transformer on the front line: Exploring the shortcomings of ai-composed music through quantitative measures. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 142–149, 2020.



Jinlong Zhu received the B.S. degree from the School of Water Resources and Hydropower Engineering at Wuhan University, China, in 2023. He is currently pursuing an M.S. degree at the Graduate School of Information Science and Technology, Hokkaido University. His research interests include deep learning methods for symbolic music generation and style transfer in music mixing. He is a student member of IEEE and ISMIR.



Keigo Sakurai received the B.S. degree in Electronics and Information Engineering from Hokkaido University, Japan in 2021, and the M.S. and Ph.D. degrees from the Graduate School of Information Science and Technology, Hokkaido University, in 2023 and 2025, respectively. He is currently a specially appointed assistant professor in the Data-Driven Interdisciplinary Research Emergence Department (D-RED), Institute for Integrated Innovations, Hokkaido University. His research interests lie in the areas of multi-domain data mining and recommender systems. He is a graduate student member of IEEE and a student member of IPSJ, JSAI and ITE.



Ren Togo received the B.S. degree in Health Sciences from Hokkaido University, Japan, in 2015, and the M.S. and Ph.D. degrees from the Graduate School of Information Science and Technology, Hokkaido University, in 2017 and 2019, respectively. He is also a Radiological Technologist. He is currently an associate professor in the Faculty of Information Science and Technology, Hokkaido University. His research interests include machine learning and its applications. He is a member of IEEE.



Takahiro Ogawa received his B.S., M.S., and Ph.D. degrees in Electronics and Information Engineering from Hokkaido University, Japan, in 2003, 2005, and 2007, respectively. He joined the Graduate School of Information Science and Technology, Hokkaido University, in 2008. He is currently a professor in the Faculty of Information Science and Technology at Hokkaido University. His research interests include AI, IoT, and big data analysis for multimedia signal processing and its applications. He has served in various leadership roles, including Special Session Chair of IEEE ISCE 2009, Doctoral Symposium Chair of ACM ICMR 2018, Organized Session Chair of IEEE GCCE 2017–2020, TPC Vice Chair of IEEE GCCE 2018, and Conference Chair of IEEE GCCE 2019. He is also an Associate Editor of ITE Transactions on Media Technology and Applications. He is a senior member of IEEE and a member of ACM, IEICE, and ITE.



Miki Haseyama received her B.S., M.S., and Ph.D. degrees in Electronics from Hokkaido University, Japan, in 1986, 1988, and 1993, respectively. She joined the Graduate School of Information Science and Technology at Hokkaido University as an associate professor in 1994. From 1995 to 1996, she was a visiting associate professor at Washington University, USA. She is currently a professor in the Faculty of Information Science and Technology at Hokkaido University. Her research interests include image and video processing and their development into semantic analysis. She has served as Vice-President of the Institute of Image Information and Television Engineers, Japan (ITE), Editor-in-Chief of ITE Transactions on Media Technology and Applications, and Director of International Coordination and Publicity for The Institute of Electronics, Information and Communication Engineers (IEICE). She is a Fellow of ITE, a senior member of IEEE, and a member of IEICE and ASJ.

- MidiCaps—a large-scale MIDI dataset with text captions. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 858–865, 2024.
- 25) Kin Wai Cheuk, Ryosuke Sawata, Toshimitsu Uesaka, Naoki Murata, Naoya Takahashi, Shusuke Takahashi, Dorien Herremans, and Yuki Mitsufuji. Diffroll: Diffusion-based generative music transcription with unsupervised pretraining capability. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.
 - 26) Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.
 - 27) Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.
 - 28) P Kingma Diederik. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.
 - 29) Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. Mousai: Efficient text-to-music diffusion models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 8050–8068, 2024.
 - 30) Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. Mustango: Toward controllable text-to-music generation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 8286–8309, 2024.
 - 31) Jeff Ens and Philippe Pasquier. MMM: Exploring conditional multi-track music generation with the transformer. *arXiv preprint arXiv:2008.06048*, 2020.
 - 32) Dimitri von Rütte, Luca Biggio, Yannic Kilcher, and Thomas Hofmann. FIGARO: Generating symbolic music with fine-grained artistic control. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
 - 33) Jinlong Zhu, Keigo Sakurai, Ren Togo, Takahiro Ogawa, and Miki Haseyama. MMT-BERT: Chord-aware symbolic music generation based on multitrack music transformer and musicbert. pages 470–477, 2024.
 - 34) Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, page 2, 2019.
 - 35) Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
 - 36) Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.
 - 37) Michael Scott Cuthbert and Christopher Ariza. Music21: A toolkit for computer-aided musicology and symbolic music data. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 637–642, 2010.
 - 38) Dmitry Bogdanov, Nicolas Wack, Emilia Gómez Gutiérrez, Sankalp Gulati, Herrera Boyer, Oscar Mayor, Gerard Roma Trepas, Justin Salamon, José Ricardo Zapata González, Xavier Serra, et al. ESSENTIA: An audio analysis library for music information retrieval. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 493–498, 2010.
 - 39) SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. Lp-musicaps: Llm-based pseudo music captioning. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 409–416, 2023.

- 40) Shih-Lun Wu and Yi-Hsuan Yang. The jazz transformer on the front line: Exploring the shortcomings of ai-composed music through quantitative measures. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 142–149, 2020.



Jinlong Zhu received the B.S. degree from the School of Water Resources and Hydropower Engineering at Wuhan University, China, in 2023. He is currently pursuing an M.S. degree at the Graduate School of Information Science and Technology, Hokkaido University. His research interests include deep learning methods for symbolic music generation and style transfer in music mixing. He is a student member of IEEE and ISMIR.



Keigo Sakurai received the B.S. degree in Electronics and Information Engineering from Hokkaido University, Japan in 2021, and the M.S. and Ph.D. degrees from the Graduate School of Information Science and Technology, Hokkaido University, in 2023 and 2025, respectively. He is currently a specially appointed assistant professor in the Data-Driven Interdisciplinary Research Emergence Department (D-RED), Institute for Integrated Innovations, Hokkaido University. His research interests lie in the areas of multi-domain data mining and recommender systems. He is a graduate student member of IEEE and a student member of IPSJ, JSAI and ITE.



Ren Togo received the B.S. degree in Health Sciences from Hokkaido University, Japan, in 2015, and the M.S. and Ph.D. degrees from the Graduate School of Information Science and Technology, Hokkaido University, in 2017 and 2019, respectively. He is also a Radiological Technologist. He is currently an associate professor in the Faculty of Information Science and Technology, Hokkaido University. His research interests include machine learning and its applications. He is a member of IEEE.



Takahiro Ogawa received his B.S., M.S., and Ph.D. degrees in Electronics and Information Engineering from Hokkaido University, Japan, in 2003, 2005, and 2007, respectively. He joined the Graduate School of Information Science and Technology, Hokkaido University, in 2008. He is currently a professor in the Faculty of Information Science and Technology at Hokkaido University. His research interests include AI, IoT, and big data analysis for multimedia signal processing and its applications. He has served in various leadership roles, including Special Session Chair of IEEE ISCE 2009, Doctoral Symposium Chair of ACM ICMR 2018, Organized Session Chair of IEEE GCCE 2017–2020, TPC Vice Chair of IEEE GCCE 2018, and Conference Chair of IEEE GCCE 2019. He is also an Associate Editor of ITE Transactions on Media Technology and Applications. He is a senior member of IEEE and a member of ACM, IEICE, and ITE.



Miki Haseyama received her B.S., M.S., and Ph.D. degrees in Electronics from Hokkaido University, Japan, in 1986, 1988, and 1993, respectively. She joined the Graduate School of Information Science and Technology at Hokkaido University as an associate professor in 1994. From 1995 to 1996, she was a visiting associate professor at Washington University, USA. She is currently a professor in the Faculty of Information Science and Technology at Hokkaido University. Her research interests include image and video processing and their development into semantic analysis. She has served as Vice-President of the Institute of Image Information and Television Engineers, Japan (ITE), Editor-in-Chief of ITE Transactions on Media Technology and Applications, and Director of International Coordination and Publicity for The Institute of Electronics, Information and Communication Engineers (IEICE). She is a Fellow of ITE, a senior member of IEEE, and a member of IEICE and ASJ.