Is Symbolic Music a Specific Language? Exploring Inspiration-to-Structure Machine Composition via LLMs

Zhejing Hu[1], Yan Liu[1]*, Zhi Zhang[1], Aiwei Zhang[2], Sheng-hua Zhong[3], Bruce X.B. Yu[4], Gong Chen[5],

[1] Department of Computing, The Hong Kong Polytechnic University
[2] School of Interactive Computing, Georgia Institute of Technology
[3] College of Computer Science and Software Engineering, Shenzhen University
[4] Zhejiang University-University of Illinois Urbana-Champaign Institute
[5] FireTorch Partners
zhejing.hu@connect.polyu.hk, yan.liu@polyu.edu.hk, zhi271.zhang@connect.polyu.hk, azhang677@gatech.edu, csshzhong@szu.edu.cn, xinboyu@intl.zju.edu.cn, heinz@clozzz.com

Abstract

Large Language Models (LLMs) have demonstrated remarkable proficiency in diverse tasks. This success raises a fundamental question in machine composition: Can symbolic music be considered a special form of language that can be jointly modeled with natural language for composition tasks? Recent studies validate that symbolic music can be modeled as a human language, yet composing structured music from partial symbolic inputs through natural language interaction remains underexplored. Even LLMs struggle to generate structurally coherent compositions in such hybrid input-output scenarios, highlighting a fundamental gap that calls for a domain-specific learning paradigm. To this end, we propose Inspiration-to-Structure (IoS), a cognitively inspired framework that enables LLMs to generate structured musical sections from melodic ideas. IoS employs a three-phase process—semantic, structural, and collaborative cognition—and is supported by two key components: (1) a new dataset and construction protocol called Structured Triplet Data (STD), and (2) a training method, Dual-Instance Structural Contrastive Optimization (DiSCO), designed to enhance structural awareness. Experiments show that IoS improves structural coherence by 47.8% and artistic creativity by 21.8% compared to conventional language modeling paradigm, supervised fine-tuning, and even enables smaller LLMs to surpass larger LLMs. These results suggest that symbolic music, while language-like, demands specialized modeling beyond standard language modeling paradigms. IoS enables LLMs to transform music theory knowledge into structured composition, empowering users to compose music interactively via language and advancing toward general creative AI.

Introduction

Large Language Models (LLMs) have shown remarkable capabilities across tasks such as text summarization (Xu et al. 2022), mathematical reasoning (Wei et al. 2022), and code generation (Liu et al. 2023). This progress raises a fundamental question in machine composition: *Can symbolic music be considered a special form of language that can be jointly modeled with natural language for composition tasks*? If so, language modeling paradigms such as supervised fine-tuning (SFT) may be transferable to machine composition, enabling users to compose music via natural language interaction. This is supported by parallels between the two domains—both use symbolic units (e.g., notes, words) and rely on rule-based systems (e.g., music theory, grammar).

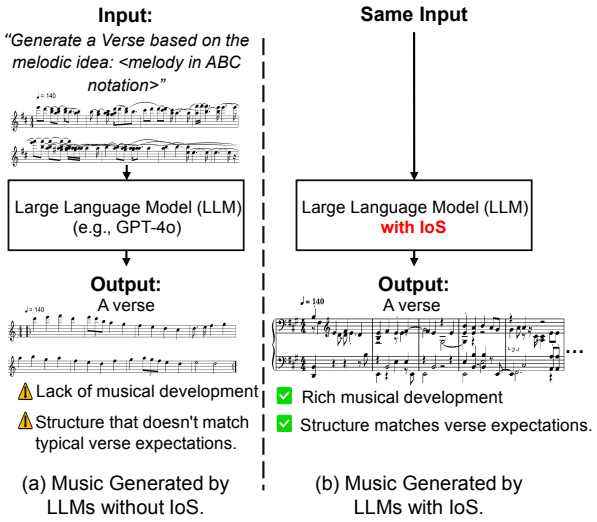

Figure 1: Demonstration of generating a verse from a melodic idea using LLMs with and without IoS. Music is shown in sheet notation for demonstration, while conversion to symbolic music sequences is performed during interaction with the LLMs.

This hypothesis has inspired pioneering explorations that apply LLMs to machine composition (Yuan et al. 2024; Zhou et al. 2024; Ma et al. 2024; Ding et al. 2024; Pasquier et al. 2025). For example, ChatMusician applies SFT to help LLMs learn music theory and composition (Yuan et al. 2024). These studies demonstrate that symbolic music can indeed be modeled as a sequence language and LLMs are ca-

*Corresponding author.

pable of learning music-theoretic rules and generating rule-compliant music.

However, modeling symbolic music as a sequence does not imply it can be fully treated as a language. As shown in Figure 1(a), even advanced LLMs (e.g., GPT-4o), while capable of understanding music theory, struggle to generate a coherent section from a given melodic idea—a fundamental skill in human composition and creative development. In contrast, human composers—such as in Beethoven's sketches for the Eroica Symphony—often used verbal annotations to guide the development of simple motifs into structured sections (Johnson, Tyson, and Winter 1985). This contrast reveals a deeper challenge: although symbolic music exhibits linguistic properties, its compositional process requires LLMs to have a cognitive leap from inspiration to structured realization—a gap not easily bridged by standard language modeling alone.

To further explore this challenge, we propose the *Inspiration-to-Structure (IoS)* framework—a cognitively informed paradigm designed to guide LLMs in learning how structured music is created (Figure 1(b)). IoS simulates how human composers transform melodic inspiration into structured sections through a three-phase process (Figure 1(b)): (1) Semantic cognition (intra-song perspective): drawing on theories of thematic coherence (Margulis 2013), the model learns to recognize how repeated sections within a piece maintain shared melodic ideas. (2) Structural cognition (inter-song perspective): drawing from hierarchical theories of musical form (Lerdahl and Jackendoff 1996), the model captures asymmetries in structural relationships—strong dependencies between sections within a song and weaker, more variable associations across songs of the same section type. This structured differentiation introduces compositional priors that are typically absent in conventional approaches. (3) Collaborative cognition: through a multi-turn process, the model progressively refines sections with human input, mirroring how composers evolve ideas into structured, context-aware music (Collins 2005).

To enable the learning of these cognitive mechanisms, we first construct *Structured Triplet Data (STD)*, a composition-oriented conversational dataset that pairs user language instructions with triplets of music sections (Figure 2): an anchor, a positive, and a negative. Next, we propose *Dual-instance Structural Contrastive Optimization (DiSCO)*, a self-supervised training method that adapts classical contrastive learning to capture semantic and structural divergence among sections. Experimental results validate that IoS significantly outperforms across multiple LLMs, achieving up to 47.8% gain in structural coherence and 21.8% improvement in artistic creativity. This improvement comes not from model architecture design, but from targeting the unique nature of symbolic music. We conclude that symbolic music can be treated as a specific language—but one whose structural logic is not fully captured by generic language modeling. The value of the IoS framework lies in showing that, with a task- and domain-adapted strategy, even standard LLMs can be guided to transform music theory knowledge into structured composition—empowering interactive language-based music creation and advancing toward general creative AI capabilities within LLMs.

Related Work

Machine Composition, as a subfield of music generation (Muhamed, others, and Smola 2021; Jiang et al. 2020; Hu et al. 2025a; Yang et al. 2025; Zuo et al. 2025; Yao et al. 2025), has long been a central focus of computational creativity research (Ferreira et al. 2022; Bodily and Ventura 2024; Ji and Yang 2024; Pasquier et al. 2025; Cosenza et al. 2023; Cancino-Chacón et al. 2023; Deng et al. 2024; Hu et al. 2025c,b). Early work, such as Hiller and Isaacson's Markov-based composition system (Hiller Jr and Isaacson 1957), laid the foundation for automated music generation. With the rise of deep learning, models like MusicVAE (Roberts et al. 2018) and MuseGAN (Dong et al. 2018) improved generation quality, while Transformer-based models (Huang et al. 2018; Huang and Yang 2020; Hsiao et al. 2021; Peng et al. 2023) enabled better modeling of long-range dependencies in symbolic sequences. More recent efforts emphasize structure-aware modeling (Dai et al. 2021; Lu et al. 2022; Naruse et al. 2022; Shih et al. 2022; Hu et al. 2023; Wang, Min, and Xia 2024; Hu et al. 2024; Bhandari, Wiggins, and Colton 2025; Wang et al. 2024; Bhandari and Colton 2024), highlighting the importance of compositional structure in generating musically coherent outputs. The sequential nature of symbolic music has led to growing interest in applying LLMs to composition tasks. Some studies focus on music lyrics, a naturally text-based task (Sheng et al. 2021; Zhang et al. 2024; Chai and Wang 2025; You et al. 2025). In terms of symbolic music generation, ChatMusician (Yuan et al. 2024) showed that LLMs can process symbolic prompts after supervised fine-tuning on music data. However, whether LLMs can achieve the leap from inspiration to structured music through natural language interaction, using conventional language training paradigms, remains underexplored. This work fills this gap by introducing a new learning paradigm tailored to the unique demands of musical composition—enabling section-level music generation from a melodic idea in an interactive language-based setting.

Contrastive Learning is a powerful method for representation learning, widely adopted in Computer Vision (Hua et al. 2023; Liu et al. 2021) and Natural Language Processing (NLP) (Zhang et al. 2023, 2022) for its ability to capture semantic similarity and structural distinctions. In the music domain, contrastive learning has been applied to tasks such as music classification and retrieval (Spijkervet and Burgoyne 2021; Yao et al. 2022), melody extraction (Yu 2024), and bridging audio and symbolic formats (Banar, Colton et al. 2022). Recent work has further demonstrated its potential in capturing motif-level patterns for symbolic composition (Wu, Dannenberg, and Xia 2023), highlighting its effectiveness in learning structured musical representations. Building on this foundation, we explore whether such a classical learning approach can be adapted to machine composition. In this work, we design a novel structure-aware contrastive training method—*DiSCO*—tailored to the characteristics of musical form. By explicitly encoding intra- and inter-song structural relations, DiSCO aligns the model's representation space with compositional logic.

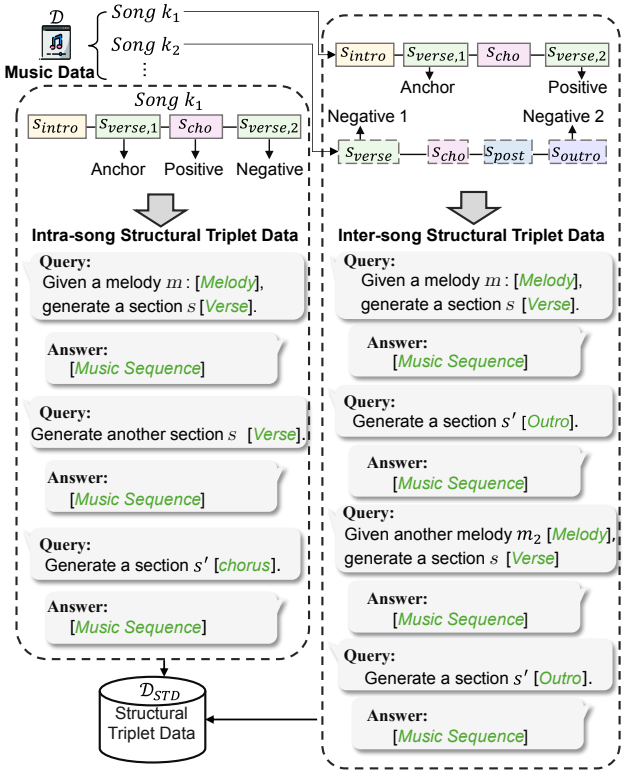

Figure 2: Structural Triplet Data Construction.

Our goal is to fine-tune the LLM to generate structured symbolic music (e.g., verse or chorus) from a melodic idea and a natural language instruction specifying the section label (Figure 3). Training uses Dual-Instance Structural Contrastive Optimization (DiSCO) and Supervised Fine-Tuning (SFT) on Structured Triplet Data (STD). At inference, the model generates coherent music reflecting learned structure and semantics.

Formally, the input $x = (m, s)$ consists of a melody $m \in \mathcal{M}$, represented as a sequence of tokens in ABC notation (i.e., $m = (m_1, m_2, \ldots, m_T)$, where $m_t \in \mathcal{V}_{\text{ABC}}$), and a section label $s \in \mathcal{S}$, such as "verse" or "chorus", which encodes the core structural intention of the composition. These components are combined into a textual prompt that is fed to the language model. The output $y \in \mathcal{Y}$ is also a sequence in ABC notation representing the target symbolic music section of type s, i.e., $y = (y_1, y_2, \ldots, y_L)$ with $y_\ell \in \mathcal{V}_{\text{ABC}}$.

Background of Supervised Fine-Tuning

SFT adapts a pre-trained LLM to specific tasks using labeled data. It minimizes the negative log-likelihood of target outputs given inputs:

$$L_{\text{SFT}} = -\mathbb{E}_{(x,y)\sim D}\left[\log \pi_\theta(y \mid x)\right], \tag{1}$$

where $\pi_\theta(y \mid x)$ is the model's predicted distribution and D is the dataset.

Structural Triplet Data Construction

Traditional SFT treats symbolic music as plain text, limiting the model's ability to distinguish musical sections and explore new patterns. Inspired by contrastive learning in preference-labeled tasks (Xu et al. 2024), we propose a Structured Triplet Data (STD) method, which introduces positive and negative samples to enhance structural discrimination.

Data Formulation Given a symbolic music dataset $\mathcal{D} = \{k_1, k_2, \ldots, k_N\}$ containing N music samples, for each music piece k_i where $i = 1, \ldots, N$, we construct either (Figure 2) an Intra-song triplet $(x^{(i)}, y_a^{(i)}, y_p^{(i)}, y_n^{(i)})$, or an Inter-song triplet $(x^{(i)}, y_a^{(i)}, y_p^{(i)}, y_{n_1}^{(i)}, y_{n_2}^{(i)})$, where:

- $x^{(i)} = (m^{(i)}, s^{(i)})$: input melody and section type.
- $y_a^{(i)}$ (Anchor Sample): ground-truth section of type $s^{(i)}$.
- $y_p^{(i)}$ (Positive Sample): same-type section (intra-song) or any section from the same song (inter-song).
- Negative Samples:
  - $y_n^{(i)}$: intra-song negative—section with a different type from the same song.
  - $y_{n_1}^{(i)}, y_{n_2}^{(i)}$: inter-song negatives—sections from different songs.

Data Construction Real-world composition is iterative, not one-shot. To reflect this collaborative cognition, we embed STD into a multi-turn conversational framework, where the model generates sections step-by-step in response to sequential queries.

Intra-Song example:

$$\begin{cases} q_1^{(i)} : \text{"Given } m^{(i)}, \text{ generate } s^{(i)} \text{ (e.g., verse)"}, & a_1^{(i)} = y_a^{(i)}, \\ q_2^{(i)} : \text{"Generate another } s^{(i)}\text{"}, & a_2^{(i)} = y_p^{(i)}, \\ q_3^{(i)} : \text{"Generate } s'^{(i)} \text{ (e.g., chorus)"}, & a_3^{(i)} = y_n^{(i)}. \end{cases} \tag{2}$$

Inter-Song example:

$$\begin{cases} q_1^{(i)} : \text{"Given } m^{(i)}, \text{ generate } s^{(i)} \text{ (e.g., verse)"}, & a_1^{(i)} = y_a^{(i)}, \\ q_2^{(i)} : \text{"Generate another } s'^{(i)} \text{ (e.g., outro)"}, & a_2^{(i)} = y_p^{(i)}, \\ q_3^{(i)} : \text{"Given another } m_2^{(i)}, \text{ generate } s^{(i)}\text{"}, & a_3^{(i)} = y_{n_1}^{(i)}, \\ q_4^{(i)} : \text{"Generate another } s'^{(i)}\text{"}, & a_4^{(i)} = y_{n_2}^{(i)}. \end{cases} \tag{3}$$

Finally, this process transforms the original dataset $\mathcal{D}$ into a structured triplet dataset $\mathcal{D}_{\text{STD}} = \{\{(q_1^{(i)}, a_1^{(i)}), \ldots, (q_{R^{(i)}}^{(i)}, a_{R^{(i)}}^{(i)})\}\}_{i=1}^{N}$, where each sample is organized as a multi-turn conversation. Typically, intra-song samples contain $R^{(i)} = 6$ turns (three query-response pairs), while inter-song samples contain $R^{(i)} = 8$ turns (four query-response pairs).

Dual-instance Structural Contrastive Optimization

Building on $\mathcal{D}_{\text{STD}}$, we propose *Dual-Instance Structural Contrastive Optimization* (DiSCO) to help the model learn both *semantic* and *structural* distinctions in symbolic music.

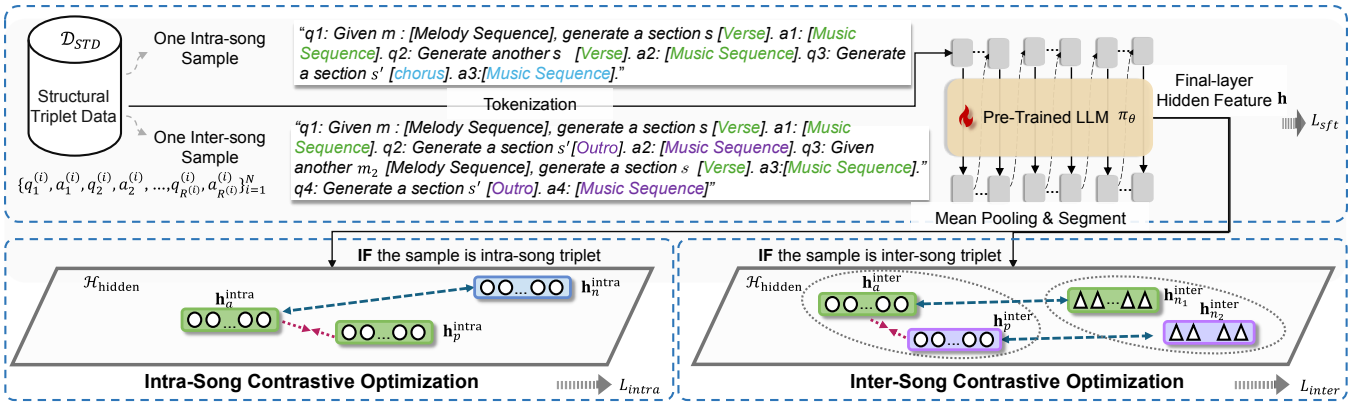

Figure 3: An illustration of DiSCO. Each STD sample is tokenized and fed into the LLM π_θ, which is trained via SFT and our proposed intra- and inter-song contrastive objectives.

DiSCO captures (*i*) intra-song contrasts for semantic cognition and (*ii*) inter-song contrasts for structural cognition, guiding composition from both content and form perspectives.

Intra-Song Contrastive Optimization Music often repeats and contrasts ideas within a piece—e.g., similar verses vs. distinct choruses. Intra-song contrastive learning captures this semantic landscape by recognizing coherence across repeated sections and distinguishing differing expressive roles.

Formally, for each data sample $\{(q_1^{(i)}, a_1^{(i)}), \ldots, (q_{R^{(i)}}^{(i)}, a_{R^{(i)}}^{(i)})\}$, the full multi-turn sequence is passed into the autoregressive transformer LLM π_θ, which produces final-layer hidden states $\mathbf{h} \in \mathbb{R}^{l \times d}$, where l denotes the total token length and d is the hidden feature dimension. We segment $\mathbf{h}$ using known token boundaries corresponding to the anchor, positive, and negative music sections, and apply mean pooling over each segment to obtain fixed-dimensional representations $\mathbf{h}_a^{\text{intra}}, \mathbf{h}_p^{\text{intra}}, \mathbf{h}_n^{\text{intra}} \in \mathbb{R}^d$. They are then used to compute a margin-based contrastive loss (Shah et al. 2022):

$$L_{\text{intra}} = \mathbb{E}_{\mathcal{D}_{STD}}\left[\log\left(1+\exp\left(\text{sim}(\mathbf{h}_a^{\text{intra}}, \mathbf{h}_p^{\text{intra}}) - \text{sim}(\mathbf{h}_a^{\text{intra}}, \mathbf{h}_n^{\text{intra}}) + \gamma\right)\right)\right], \tag{4}$$

where $\text{sim}(\mathbf{a}, \mathbf{b})$ is cosine similarity, and γ is a margin hyperparameter. This loss encourages the model to assign higher similarity to semantically consistent sections (e.g., two verses in the same song) while separating semantically distinct ones (e.g., verse vs. chorus).

Inter-Song Contrastive Optimization At the song level, sections like verse and chorus are structurally linked within a piece, while similarly labeled sections across songs may serve similar roles but remain structurally independent. Inter-song contrastive learning captures these higher-order structural distinctions.

Formally, let $\left(\mathbf{h}_a^{\text{inter}}, \mathbf{h}_p^{\text{inter}}, \mathbf{h}_{n_1}^{\text{inter}}, \mathbf{h}_{n_2}^{\text{inter}}\right)$ denote the representations of the anchor, positive, and two negative sections from the triplet. The inter-song contrastive loss is defined as:

$$L_{\text{inter}} = \mathbb{E}_{\mathcal{D}_{STD}}\Bigg[\log\Big(1+\exp\Big(\text{sim}\big(\mathbf{h}_a^{\text{inter}}, \mathbf{h}_p^{\text{inter}}\big) - \frac{1}{2}\Big(\text{sim}\big(\mathbf{h}_a^{\text{inter}}, \mathbf{h}_{n_1}^{\text{inter}}\big) + \text{sim}\big(\mathbf{h}_a^{\text{inter}}, \mathbf{h}_{n_2}^{\text{inter}}\big)\Big) + \gamma\Big)\Big)\Bigg]. \tag{5}$$

Here, $\mathbf{h}_a^{\text{inter}}$ and $\mathbf{h}_p^{\text{inter}}$ originate from the same song, while $\mathbf{h}_{n_1}^{\text{inter}}$ and $\mathbf{h}_{n_2}^{\text{inter}}$ come from different songs. This contrastive setup guides the model to capture high-level structural cohesion across sections of a composition, while distinguishing similarly labeled sections across songs that may not share structural dependence.

Overall Optimization Objective To ensure fluency and adherence to real musical syntax, we incorporate a SFT objective over the multi-turn conversational dataset $\mathcal{D}_{STD}$. Each training sample $\{(q_1^{(i)}, a_1^{(i)}), \ldots, (q_{R^{(i)}}^{(i)}, a_{R^{(i)}}^{(i)})\}$ encodes a sequence of melody-section prompts and symbolic music responses, reflecting anchor, positive, and negative sections from structured triplets.

The SFT loss is computed autoregressively across all response tokens:

$$L_{\text{SFT}} = -\mathbb{E}_{\mathcal{D}_{STD}}\left[\sum_{t=1}^{R^{(i)}} \log \pi_\theta(a_t^{(i)} \mid q_{\leq t}^{(i)}, a_{<t}^{(i)})\right], \tag{6}$$

where π_θ denotes the model's output distribution, and $a_t^{(i)}$ is decoded token-by-token. The loss is accumulated over all response tokens.

This supervised loss is combined with the DiSCO, resulting in the overall training objective:

$$\min_\theta \begin{cases} L_{\text{SFT}} + \lambda L_{\text{intra}} & \text{if intra-song contrastive input} \\ L_{\text{SFT}} + \lambda L_{\text{inter}} & \text{if inter-song contrastive input} \end{cases} \tag{7}$$

where $\lambda \in [0, 1]$ controls the relative contributions of *intra-song* and *inter-song* contrastive losses. During inference, given a user-provided melody and a target section type, the model generates a corresponding symbolic music section by sampling from $\pi_\theta(y \mid x)$ in an autoregressive manner.

Category	Model	Melody		Accompaniment		Overall		
		PD ↑	DD ↑	PD ↑	DD ↑	PD ↑	DD ↑	CD ↑
Music-Specific Models	RWKV4-Music (Peng et al. 2023)	0.62	0.47	0.53	0.40	0.58	0.44	0.69
	CRG (Hu et al. 2024)	0.61	0.41	0.51	0.32	0.56	0.36	0.68
	ChatMusician (Yuan et al. 2024)	0.61	0.79	0.50	0.68	0.55	0.73	0.61
General-Purpose Models	DeepSeek-R1 (Guo et al. 2025)	0.56	0.61	0.23	0.31	0.40	0.46	0.62
	GPT-4o (Islam and Moushi 2024)	0.61	0.59	0.46	0.31	0.54	0.45	0.69
	GPT-o3 (OpenAI 2025)	0.62	0.61	0.55	0.49	0.58	0.54	0.62
	LLama3.2-1B (Dubey et al. 2024)	0.38	0.50	0.35	0.42	0.36	0.46	0.52
	+SFT	0.69	0.72	0.61	0.65	0.68	0.63	0.63
	+IoS	**0.78**	**0.88**	**0.67**	**0.78**	**0.72**	**0.83**	**0.82**
	LLama3.2-3B (Dubey et al. 2024)	0.42	0.54	0.32	0.35	0.37	0.44	0.55
	+SFT	0.66	0.79	0.62	0.71	0.64	0.75	0.76
	+IoS	**0.80**	**0.89**	**0.67**	**0.77**	**0.74**	**0.83**	**0.87**
	Qwen2.5-0.5B (Yang et al. 2024)	0.06	0.10	0.00	0.01	0.03	0.05	0.13
	+SFT	0.58	0.77	0.47	0.61	0.53	0.69	0.67
	+IoS	**0.76**	**0.87**	**0.57**	**0.76**	**0.67**	**0.82**	**0.87**
	Qwen2.5-1.5B (Yang et al. 2024)	0.08	0.10	0.00	0.01	0.04	0.05	0.16
	+SFT	0.66	0.72	0.49	0.68	0.58	0.70	0.74
	+IoS	**0.77**	**0.87**	**0.58**	**0.77**	**0.68**	**0.82**	**0.87**
	Qwen2.5-3B (Yang et al. 2024)	0.41	0.51	0.16	0.24	0.28	0.38	0.53
	+SFT	0.68	0.79	0.51	0.69	0.60	0.74	0.78
	+IoS	**0.78**	**0.88**	**0.62**	**0.77**	**0.70**	**0.83**	**0.89**

Table 1: Comparison of objective results between IoS and baseline methods across Music-Specific and General-Purpose Models. "+SFT" indicates the model is supervised fine-tuning on our STD. Text in bold indicates the better performance within that model, while text in red indicates the highest model performance among all models.

Experiment

Dataset We construct a structural music dataset by first defining eight section types grounded in music theory: intro, verse, chorus, bridge, pre-chorus, post-chorus, interlude, and outro. Based on these definitions, we manually annotate section-level structures for each piece in the POP909 dataset (Wang et al. 2020). Three trained musicians independently annotated all pieces, followed by discussion to resolve disagreements and reach consensus, ensuring annotation quality under strict quality control. We choose POP909 for its clear section boundaries, melody-accompaniment separation, and strong internal consistency compared to other open-source datasets. Following annotation, we construct a triplet-based, multi-turn conversational dataset for supervised and self-supervised training.

Dataset Statistics. The dataset consists of 909 complete pop songs with 9,132 section-level annotations. Most pieces contain 6–14 sections, reflecting common structures in pop music. To support contrastive learning, we construct a triplet-based multi-turn conversational dataset containing 9,132 samples, where each conversation consists of three or four turns (see Figure 2). The dataset is balanced, with 4,566 intra-contrastive samples and 4,566 inter-contrastive samples. The detailed statistics and annotation process are illustrated in the Appendix.

Evaluation Metrics We evaluate performance using both objective and subjective metrics. Objective metrics assess the statistical properties of the generated music and the model's executability, while subjective metrics measure the perceived quality of the generated music based on human evaluation.

Objective Evaluation. We evaluate the model based on how well the generated sections aligns with human-composed ground truth using the average overlapped area of Pitch Distribution (PD), Duration Distribution (DD), and Chord Distribution (CD), following (Hu et al. 2025c). Higher scores indicate closer alignment.

Subjective Evaluation. To assess both musical quality and interactive composition capabilities, we designed a human evaluation protocol with optional multi-turn interaction. Participants either uploaded a user-defined melodic idea or selected one from a provided MIDI test set. They then wrote a text instruction specifying the target section type, which, along with the melody, was as input to the model. The model generated a symbolic music section, which was converted to MIDI format for playback.

Participants were allowed up to three interaction rounds, during which they could revise their instructions to refine or regenerate the output. After the final interaction, they rated the result on five criteria: *Structure* (alignment with section type), *Relevance* (fit to the input melody), *Musicality* (naturalness and aesthetic quality), *Creativity* (originality and expressiveness), and *Interactive Satisfaction* (responsiveness to user feedback). Each was scored from 1 to 5, and an *Overall* score was calculated as the average.

Model	Structural ↑	Relevance ↑	Musicality ↑	Creativity ↑	Satisfaction ↑	Overall ↑
RWKV-Music (Peng et al. 2023)	2.80±0.61	2.13±0.72	3.16±1.13	3.18±0.84	2.30±0.86	2.71
CRG (Hu et al. 2024)	2.57±0.83	2.01±0.94	3.02±0.81	2.89±0.81	2.15±0.94	2.53
ChatMusician (Yuan et al. 2024)	2.72±0.85	2.98±0.54	2.89±1.01	2.94±0.52	2.48±0.63	2.80
DeepSeek-R1 (Guo et al. 2025)	2.97±0.43	3.11±0.70	2.95±0.93	2.53±0.93	3.40±0.88	2.99
GPT-4o (Islam and Moushi 2024)	3.03±0.79	3.05±1.04	3.15±0.94	2.74±1.07	3.50±0.95	3.09
Llama3.2-3B (Dubey et al. 2024)	2.18±0.76	2.55±1.02	2.75±1.08	2.57±1.00	2.52±0.96	2.51
+SFT	2.96±0.69	3.01±0.93	3.06±0.95	2.93±0.88	3.05±0.84	3.00
+IoS	**3.93**±0.59	**3.57**±0.71	**3.64**±0.73	**3.54**±0.68	**3.60**±0.70	**3.66**
Qwen2.5-3B (Yang et al. 2024)	2.30±0.69	2.84±0.51	2.57±1.02	2.40±0.98	2.50±0.90	2.52
+SFT	3.01±0.66	3.04±0.77	2.80±0.90	2.93±0.85	2.68±0.81	2.89
+IoS	**3.71**±0.59	**3.77**±0.65	**3.24**±0.80	**3.57**±0.79	**3.55**±0.74	**3.57**

Table 2: Comparison of subjective results. IoS largely improves the music quality. Text in bold indicates better performance within that model, while text in red indicates the highest model performance among all models.

We collaborated with a music technology company to collect 115 evaluation sessions. Among the participants, 51 were male and 64 were female; 33 were under 20 years old, 60 were aged 20–30, and 22 were aged 31–40. To assess musical background, we collected self-reported years of training and composition experiences: 34 participants had 1–3 years, 49 had 4–6 years, 21 had 7–10 years, and 11 had over 10 years. We followed a standard blind-review protocol to ensure unbiased evaluation across all model outputs.

Comparable Methods We evaluate our method on both music-specific models—RWKV-Music (Peng et al. 2023), CRG (Hu et al. 2024), and ChatMusician (Yuan et al. 2024), all trained on symbolic music data—and general-purpose LLMs, including GPT-4o/o3 (Islam and Moushi 2024; OpenAI 2025), DeepSeek-R1 (Guo et al. 2025), LLaMA 3.2 (Dubey et al. 2024), and Qwen 2.5 (Yang et al. 2024). To assess the added value of our framework, we further apply SFT on the STD dataset to representative general-purpose models (LLaMA 3.2 and Qwen 2.5), and fine-tune them again with our proposed DiSCO method. Our evaluation focuses primarily on smaller-scale models suitable for practical deployment. Due to their closed-source nature, larger proprietary models like GPT-4o/o3 cannot be fine-tuned and are evaluated solely via prompting.

Implementation Details For data representation, we encode symbolic music using ABC notation for two key reasons. First, it has higher length efficiency. ABC notation is more compact than MIDI-based representations, significantly reducing computational costs for long sequences. Second, previous studies (Qu et al. 2024) have demonstrated that LLMs can effectively understand ABC notation, making it a viable choice for music modeling. For model training, we utilize Unsloth (Daniel Han and team 2023) to facilitate the training process with max length setting as 8192 tokens. We train the model on 2 Nvidia RTX 3090 GPUs with an initial learning rate of 1e-4 and a batch size of 4 for 200 epochs. For hyperparameter selection, we set $\lambda = 0.5$ and $\gamma = 0.1$. Additionally, we reserved 100 music pieces and generate a total of 600 turns of conversations for testing the performance. We also conducted a sensitivity analysis and found that performance is robust to moderate variations in hyperparameters.

Main Result

Model	STD	DiSCO		Overall		
		Intra	Inter	PD ↑	DD ↑	CD ↑
Baseline	✗	✗	✗	0.37	0.44	0.55
SFT	✗	✗	✗	0.60	0.61	0.61
Ours	✓	✗	✗	0.64	0.75	0.76
	✓	✓	✗	0.68	0.79	0.79
	✓	✗	✓	0.68	0.80	0.80
	✓	✓	✓	**0.74**	**0.83**	**0.87**

Table 3: Ablation of the IoS. We evaluate the contribution of the STD and each component of the DiSCO module.

Evaluation Results Analysis. Experimental findings confirm that the proposed IoS framework significantly enhances the compositional ability of LLMs. As shown in Table 1 and Table 2, IoS consistently outperforms both base models and those enhanced with SFT across objective and subjective metrics. In subjective evaluations, IoS yields up to a *47.8% increase in structural coherence* and a *21.8% improvement in artistic creativity*, reflecting its effectiveness in reinforcing musical form. Among all models, LLaMA 3.2-3B+IoS achieves the highest overall score (3.66), surpassing both its SFT counterpart and larger models like GPT-4o. Objective metrics show similar trends, validating the robustness of the framework.

Group-Specific Observations. Further analysis based on musical background reveals that the 32 participants with over six years of training gave higher ratings for structure and relevance, appreciating the model's ability to maintain musical logic. Those with less training focused more on melodic appeal and creativity. Across all groups, models trained with IoS received higher satisfaction scores, suggesting that both novice and expert users benefited from more structured and expressive outputs.

Why Does IoS Improve Performance? IoS improves performance by teaching models to distinguish structural roles through contrastive learning and by enhancing context retention via multi-turn training, leading to more coherent generation. These results show that IoS boosts symbolic music quality across model types, underscoring the value of structure-aware training for scalable, composition-ready musical intelligence.

Ablation Analysis

Table 3 presents an ablation study evaluating the individual contributions of the STD and the DiSCO algorithm. The Baseline refers to direct prompting, while SFT denotes supervised fine-tuning on standard music data without STD. Compared to both the baseline and the SFT-only variant, incorporating STD alone yields notable improvements across all metrics, highlighting the value of structured multi-turn data in enhancing the model's compositional reasoning. Introducing the DiSCO module further boosts performance. The combined use of intra-song and inter-song contrastive learning achieves the highest overall performance, confirming their complementary benefits.

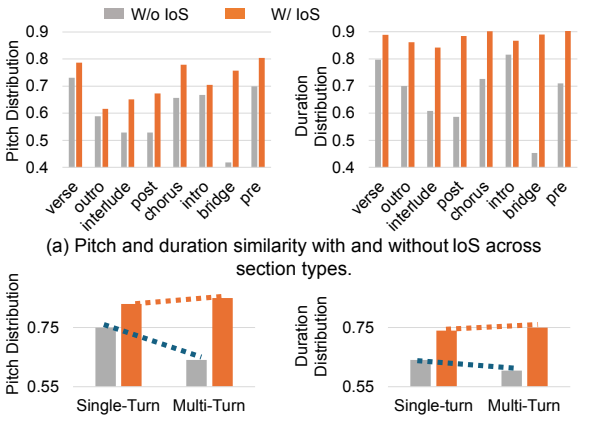

Figure 4: (a) IoS Enhances Sectional Structure. (b) IoS Stabilizes Multi-Turn Machine Composition.

Discussion

LLMs Perceive Structural Variations. One of the fundamental characteristics of human-like composition is the ability to structure music into distinct and meaningful sections. Figure 4(a) shows IoS enhances this ability via PD and DD. IoS-trained LLMs achieve higher pitch and duration similarity across sections, reflecting a finer grasp of musical form and variation. This structural awareness boosts diversity and signals progress toward human-like composition, where sections are intentionally arranged.

LLMs as Collaborative Composers. Human composers refine music iteratively, and IoS enables LLMs to do the same. Figure 4(b) shows that without IoS, multi-turn performance deteriorates, while IoS-trained models maintain stability, preserving context and structural coherence. This suggests that IoS prevents error accumulation, allowing LLMs to refine compositions interactively, making them more effective AI collaborators in music creation.

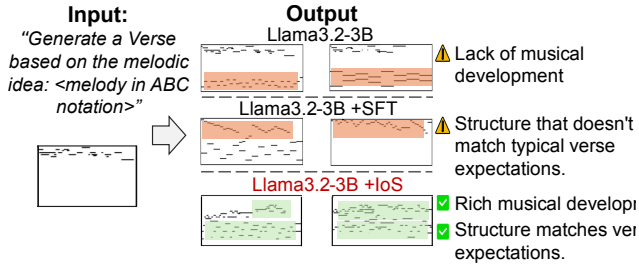

Figure 5: Comparison of verses generated by our method and baselines.

LLMs Understand Semantics Within Music. Figure 5 compares our IoS output with LLaMA3.2-3B baselines. While the baselines often fail to differentiate sections, IoS generates clearer contrasts and rich musical development, demonstrating stronger structural and semantic control.

Efficient Music Processing for Human-Like Composition. We use ABC notation instead of MIDI to reduce token length—average drops from 5,803 to 3,229, and max from 22,390 to 7,956. This improves context retention and lowers computational cost.

Limitations

This work focuses on the pop genre due to the lack of symbolic datasets with accurate structural annotations—not due to methodological limits, extending to other genres will require further annotation. Additionally, our framework generates individual sections from a melodic idea but does not model structural development across sections. While full-length composition relies on multi-turn interaction and user-defined planning.

Conclusion

In this work, we propose Inspiration-to-Structure (IoS), a cognitively inspired learning framework that enables LLMs to generate structured musical sections from simple melodic ideas through natural language interaction. Instead of scaling model size or modifying architectures, IoS incorporates a three-phase process supported by Structured Triplet Data (STD) and Dual-instance Structural Contrastive Optimization (DiSCO). Experimental results show that LLMs trained with IoS can generate eight musical section types and engage in multi-turn collaboration with human. These findings suggest that, while symbolic music shares linguistic features, its structured composition demands a distinct modeling approach. IoS provides a scalable and generalizable pathway toward equipping LLMs with deeper musical intelligence, enabling interactive, language-driven music composition. In future work, we aim to curate additional datasets across diverse genres and explore how to seamlessly connect different sections into a coherent full-length composition.

Acknowledgements

This work is supported by A Large Language Model-powered Clinical Communication Training System Under Cantonese-English Code-switching Scenario ITS/397/23 and MusicGPT: Revolutionizing the Soundscape with AI-Powered Interactive Music Creation (P0048382).

References

Banar, B.; Colton, S.; et al. 2022. Connecting audio and graphic score using self-supervised representation learning-a case study with Gyorgy Ligeti's artikulation. In *ICCC*.

Bhandari, K.; and Colton, S. 2024. Motifs, phrases, and beyond: The modelling of structure in symbolic music generation. In *EvoMUSART*, 33–51. Springer.

Bhandari, K.; Wiggins, G. A.; and Colton, S. 2025. Yin-yang: Developing motifs with long-term structure and controllability. In *EvoMUSART*, 1–17. Springer.

Bodily, P. M.; and Ventura, D. 2024. Operationalizing essential characteristics of creativity in a computational system for music composition. In *AAAI*, volume 38, 447–455.

Cancino-Chacón, C.; et al. 2023. The accompanion: combining reactivity, robustness, and musical expressivity in an automatic piano accompanist. In *IJCAI*, 5779–5787.

Chai, L.; and Wang, D. 2025. CSL-L2M: Controllable Song-Level Lyric-to-Melody Generation Based on Conditional Transformer with Fine-Grained Lyric and Musical Controls. In *AAAI*, volume 39, 23541–23549.

Collins, D. 2005. A synthesis process model of creative thinking in music composition. *Psychology of music*, 33(2): 193–216.

Cosenza, E.; et al. 2023. Graph-based polyphonic multitrack music generation. In *IJCAI*, 5797–5805.

Dai, S.; Jin, Z.; Gomes, C.; and Dannenberg, R. B. 2021. Controllable deep melody generation via hierarchical music structure representation. In *ISMIR*.

Daniel Han, M. H.; and team, U. 2023. Unsloth.

Deng, Q.; et al. 2024. ComposerX: Multi-Agent Symbolic Music Composition With LLMs. In *ISMIR*, 669–679.

Ding, S.; et al. 2024. Songcomposer: A large language model for lyric and melody composition in song generation. *arXiv preprint arXiv:2402.17645*.

Dong, H.-W.; et al. 2018. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *AAAI*, 34–41.

Dubey, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ferreira, L. N.; et al. 2022. Controlling perceived emotion in symbolic music generation with monte carlo tree search. In *AAAI*, volume 18, 163–170.

Guo, D.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hiller Jr, L. A.; and Isaacson, L. M. 1957. Musical composition with a high speed digital computer. In *Audio Engineering Society Convention 9*.

Hsiao, W.-Y.; et al. 2021. Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs. In *AAAI*, 178–186.

Hu, J.; Li, J.; Pan, Z.; Chen, C.; Li, Z.; Wang, P.; and Zhang, L. 2025a. SongSong: A Time Phonograph for Chinese SongCi Music from Thousand of Years Away. In *AAAI*, volume 39, 26229–26237.

Hu, Z.; Liu, Y.; Chen, G.; Ma, X.; Zhong, S.; and Luo, Q. 2024. Responding to the Call: Exploring Automatic Music Composition Using a Knowledge-Enhanced Model. In *AAAI*, volume 38, 521–529.

Hu, Z.; Liu, Y.; Chen, G.; and Yu, B. X. 2025b. CompLex: Music Theory Lexicon Constructed by Autonomous Agents for Automatic Music Generation. In *IJCAI*.

Hu, Z.; Liu, Y.; Chen, G.; and Yu, B. X. 2025c. Compose with Me: Collaborative Music Inpainter for Symbolic Music Infilling. In *AAAI*, volume 39, 1327–1335.

Hu, Z.; et al. 2023. The beauty of repetition: an algorithmic composition model with motif-level repetition generator and outline-to-music generator in symbolic music generation. *TMM*.

Hua, Y.; Wu, W.; Zheng, C.; Lu, A.; Liu, M.; Chen, C.; and Wu, S. 2023. Part Aware Contrastive Learning for Self-Supervised Action Recognition. *arXiv preprint arXiv:2305.00666*.

Huang, C.-Z. A.; Vaswani, A.; Uszkoreit, J.; Shazeer, N.; Simon, I.; Hawthorne, C.; Dai, A. M.; Hoffman, M. D.; Dinculescu, M.; and Eck, D. 2018. Music transformer. *arXiv preprint arXiv:1809.04281*.

Huang, Y.-S.; and Yang, Y.-H. 2020. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *ACMMM*, 1180–1188.

Islam, R.; and Moushi, O. M. 2024. Gpt-4o: The cutting-edge advancement in multimodal llm. *Authorea Preprints*.

Ji, S.; and Yang, X. 2024. Muser: Musical element-based regularization for generating symbolic music with emotion. In *AAAI*, volume 38, 12821–12829.

Jiang, N.; Jin, S.; Duan, Z.; and Zhang, C. 2020. Rl-duet: Online music accompaniment generation using deep reinforcement learning. In *AAAI*, 710–718.

Johnson, D. P.; Tyson, A.; and Winter, R. 1985. *The Beethoven sketchbooks: History, reconstruction, inventory*. 4. Univ of California Press.

Lerdahl, F.; and Jackendoff, R. S. 1996. *A Generative Theory of Tonal Music, reissue, with a new preface*. MIT press.

Liu, C.; Fu, Y.; Xu, C.; Yang, S.; Li, J.; Wang, C.; and Zhang, L. 2021. Learning a few-shot embedding model with contrastive learning. In *AAAI*, volume 35, 8635–8643.

Liu, J.; et al. 2023. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *NeurIPS*, 36: 21558–21572.

Lu, P.; Tan, X.; Yu, B.; Qin, T.; Zhao, S.; and Liu, T.-Y. 2022. MeloForm: Generating melody with musical form based on expert systems and neural networks. *arXiv preprint arXiv:2208.14345*.

Ma, W.; et al. 2024. Do music LLMs learn symbolic concepts? A pilot study using probing and intervention. In *NeurIPS 2024 Workshop*.

Margulis, E. H. 2013. *On repeat: How music plays the mind*. Oxford University Press.

Muhamed, A.; others, Z. C.; and Smola, A. J. 2021. Symbolic music generation with transformer-gans. In *AAAI*, volume 35, 408–417.

Naruse, D.; Takahata, T.; Mukuta, Y.; and Harada, T. 2022. Pop Music Generation with Controllable Phrase Lengths. In *ISMIR*, 125–131.

OpenAI. 2025. Official website. //https://openai.com/. Accessed: 2025-08-01.

Pasquier, P.; Ens, J.; Fradet, N.; Triana, P.; Rizzotti, D.; Rolland, J.-B.; and Safi, M. 2025. MIDI-GPT: A controllable generative model for computer-assisted multitrack music composition. In *AAAI*, volume 39, 1474–1482.

Peng, B.; Alcaide, E.; Anthony, Q.; Albalak, A.; Arcadinho, S.; Biderman, S.; Cao, H.; Cheng, X.; Chung, M.; Grella, M.; et al. 2023. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*.

Qu, X.; Bai, Y.; Ma, Y.; Zhou, Z.; Lo, K. M.; Liu, J.; Yuan, R.; Min, L.; Liu, X.; Zhang, T.; et al. 2024. Mupt: A generative symbolic music pretrained transformer. *arXiv preprint arXiv:2404.06393*.

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

Shah, A.; Sra, S.; Chellappa, R.; and Cherian, A. 2022. Max-margin contrastive learning. In *AAAI*, volume 36, 8220–8230.

Sheng, Z.; et al. 2021. Songmass: Automatic song writing with pre-training and alignment constraint. In *AAAI*, volume 35, 13798–13805.

Shih, Y.-J.; Wu, S.-L.; Zalkow, F.; Muller, M.; and Yang, Y.-H. 2022. Theme Transformer: Symbolic Music Generation with Theme-Conditioned Transformer. *TMM*, 1–1.

Spijkervet, J.; and Burgoyne, J. A. 2021. Contrastive learning of musical representations. *arXiv preprint arXiv:2103.09410*.

Wang, Y.; et al. 2024. MeloTrans: A Text to Symbolic Music Generation Model Following Human Composition Habit. *arXiv preprint arXiv:2410.13419*.

Wang, Z.; Chen, K.; Jiang, J.; Zhang, Y.; Xu, M.; Dai, S.; and Xia, G. 2020. POP909: A Pop-Song Dataset for Music Arrangement Generation. In *ISMIR*, 38–45.

Wang, Z.; Min, L.; and Xia, G. 2024. Whole-Song Hierarchical Generation of Symbolic Music Using Cascaded Diffusion Models. In *ICLR*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35: 24824–24837.

Wu, Y.; Dannenberg, R. B.; and Xia, G. 2023. Motif-Centric Representation Learning for Symbolic Music. *arXiv preprint arXiv:2309.10597*.

Xu, H.; et al. 2024. Contrastive preference optimization: pushing the boundaries of LLM performance in machine translation. In *ICML*, 55204–55224.

Xu, S.; Zhang, X.; Wu, Y.; and Wei, F. 2022. Sequence level contrastive learning for text summarization. In *AAAI*, volume 36, 11556–11565.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yang, H.; Su, K.; Zhang, Y.; Chen, J.; Qian, K.; Liu, G.; and Gan, C. 2025. UniMuMo: Unified Text, Music, and Motion Generation. In *AAAI*, volume 39, 25615–25623.

Yao, D.; et al. 2022. Contrastive learning with positive-negative frame mask for music representation. In *Proceedings of the ACM Web Conference 2022*, 2906–2915.

Yao, Y.; Li, P.; Chen, B.; and Wang, A. 2025. Jen-1 composer: A unified framework for high-fidelity multi-track music generation. In *AAAI*, volume 39, 14459–14467.

You, M.; Zhang, F.; Zhang, S.; and Xu, L. 2025. S^2MILE: Semantic-and-Structure-Aware Music-Driven Lyric Generation. In *AAAI*, volume 39, 22208–22217.

Yu, S. 2024. MCSSME: multi-task contrastive learning for semi-supervised singing melody extraction from polyphonic music. In *AAAI*, volume 38, 365–373.

Yuan, R.; et al. 2024. ChatMusician: Understanding and Generating Music Intrinsically with LLM. In *ACL*, 6252–6271.

Zhang, T.; et al. 2022. Frequency-aware contrastive learning for neural machine translation. In *AAAI*, volume 36, 11712–11720.

Zhang, Z.; Lasocki, K.; Yu, Y.; and Takasu, A. 2024. Syllable-level lyrics generation from melody exploiting character-level language model. In *ACL*, 1336–1346.

Zhang, Z.; et al. 2023. NerCo: a contrastive learning based two-stage chinese NER method. In *IJCAI*, 5287–5295.

Zhou, Z.; et al. 2024. Can LLMs" Reason" in Music? An Evaluation of LLMs' Capability of Music Understanding and Generation. *arXiv preprint arXiv:2407.21531*.

Zuo, H.; et al. 2025. Gvmgen: A general video-to-music generation model with hierarchical attentions. In *AAAI*, volume 39, 23099–23107.