Transformer-Based Unpaired Piano Accompaniment Style Transfer

Hsin Ai and Yi-Hsuan Yang
National Taiwan University, Taiwan
E-mail: iivvyy0728777@gmail.com, yhyangtw@ntu.edu.tw

***Abstract*—Arranger-specific style transfer for pop piano covers requires effective content-style disentanglement. To address this, we propose a framework that uses a lead sheet (namely, melody and chords) as a style-agnostic content anchor, enabling precise style manipulation without requiring paired data. We then systematically compare several Transformer-based architectures to evaluate the efficacy of a direct token-based conditioning strategy versus more complicated embedding-based methods. While all approaches effectively capture the target style, our evaluation shows that the simpler token-based model achieves superior performance in both objective and subjective assessments of content preservation and style matching. This finding provides empirical evidences that a robust, explicit content representation (i.e., the lead sheet) is highly effective for this task, offering a practical benchmark for controllable music generation.**

I. Introduction

Symbolic music generation, which leverages MIDI's precise encoding of musical events, has made great strides with deep learning [1]. Modern architectures like the Transformer excel at generating coherent long-form music from stratch [2]–[6]. However, modeling the nuanced, arranger-specific styles found in pop piano covers for style transfer applications remains less studied. In pop piano covers [7]–[9], the "content" is typically the core melody, often played by the right hand, along with the underlying harmonic progression. The "style," in contrast, is expressed through the arranger's unique accompaniment patterns, such as characteristic left-hand rhythms, textural density, harmonic context, and right-hand embellishments that weave around the main melody. This phenomenon, where a single piece of content is naturally paired with diverse stylistic renderings, makes arranger-specific style transfer for pop piano covers an interesting music style transfer problem that entails effective content-style disentanglement.

To address this, our work explores the use of *quantized lead sheets* as a style-agnostic content anchor. The core principle of our approach is inspired by two-stage frameworks like Compose & Embellish [10], simplifies the disentanglement problem and enables style transfer without requiring paired data. Building on this foundation, we conduct a systematic investigation into the most effective strategies for representing and conditioning on content and style. We implement and comparatively analyze several Transformer-based architectures, primarily contrasting a direct, token-based approach with more complicated embedding-based methods.

The contributions of this paper are threefold:

- We propose an unpaired piano accompaniment style transfer framework that leverages a quantized lead sheet as a style-free content anchor.
- We provide a systematic comparison of token- and embedding-based models, demonstrating that a straightforward token-based method can effectively achieve both melody preservation and style matching.
- A novel symbolic approach for the less-explored task of arranger-specific style transfer, focused on capturing distinct arranger styles in single-track piano MIDI.

A demo page[1] is provided to listen to piano cover samples, along with source code that supports future research.

II. Related work

Recent years have seen great progress in symbolic music generation. Transformer models, using self-attention, have improved sequence coherence and expressiveness compared to early recurrent neural network (RNN) and convolutional neural network (CNN) methods [11], [12], as demonstrated by models such as Music Transformer [2] and Pop Music Transformer [3]. However, directly generating full performances can neglect music's hierarchical nature—particularly the interplay of melody and harmony—making it difficult to control fine-grained stylistic elements like accompaniment patterns.

To better model this hierarchy, two-stage approaches have emerged. Frameworks like Compose & Embellish [10] first generate a lead sheet (encoding melody and chords) before adding stylistic performance details, a principle that has been successfully applied across various other music generation tasks [8], [13], [14]. This mimics how human arrangers employs a clear melodic and harmonic foundation, making the lead sheet a robust, style-neutral content anchor. This principle is crucial for our primary task of music style transfer.

The definition of style in this domain is diverse, with prior works focusing on transferring broad categories like emotion [14], genre [15]–[17], composition [18], [19], expressive performance [20], [21], or timbre [22]. In contrast, our work defines style specifically as the nuanced arranger-specific accompaniment patterns found in pop piano covers, a relatively underexplored niche topic.

A central challenge across all these tasks is achieving effective content-style disentanglement. How to best encode

[1]Demo: https://ivy7couch.github.io/Piano-Accompaniment-Style-Transfer/

979-8-3315-7206-8/25/$31.00

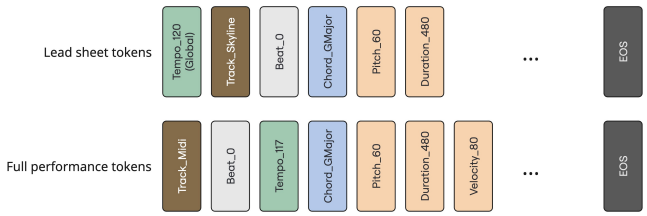

Fig. 1. Data representation for the lead sheet (M_{lead}) tokens (for representing the "content" of pop piano covers) and full piano performance (X_{perf}) tokens (for the target output).

content and style from the foundation provided by a lead sheet is therefore a critical research question. Prior works have explored two primary paradigms for representing both content and style: direct *token*-based approaches and *embedding*-based methods that learn a continuous latent space. For content, some frameworks use tokenized lead sheets to preserve melodic structure [14], while others encode musical information into content embeddings [23]. Similarly, for style, conditioning can be achieved with explicit class tokens (as in Pop2Piano [7]) for high interpretability, or with learned style embeddings from reference performances for continuous interpolation [23].

While these prior works individually demonstrate powerful strategies, a systematic comparison of these different representation trade-offs within a unified framework for arranger-specific style transfer is still lacking. Therefore, this paper provides a direct, empirical evaluation of different combinations of token- and embedding-based strategies for both content and style, grounded on a lead sheet foundation, to determine which best achieves our core task.

III. Methodology

A. Data Representation

We adopt the REMI (i.e., REvamped MIDI-derived Events) framework [3] to represent all musical data. As illustrated in Figure 1, a full piano performance, containing rich stylistic information, is encoded as a sequence of symbolic events denoted as X_{perf}. This representation captures polyphonic details where `[BAR]` and `[BEAT]` tokens provide a temporal grid with a 16th-note resolution, and other events specify note properties (`[PITCH]`, `[DURATION]`, `[VELOCITY]`), `[TEMPO_CHANGE]`, and `[CHORD]`.

A simplified content representation is central to our approach for content-style disentanglement. We define content as a lead sheet (M_{lead}), which is designed to be a style-agnostic anchor. The lead sheet is composed of two elements: 1) a monophonic melody line, extracted by the *Skyline* algorithm [24] which selects the highest-pitched note at each quantized time step, and 2) the bar-level chords, extracted using the *Chorder* rule-based chord detection.[2] The melody's timing is quantized to a coarser 8th-note resolution (compared to the 16-th note resolution adopted in modeling the piano performance). This simplified representation omits performance-specific attributes like note velocity, tempo changes, and uses a lower temporal resolution to filter out from the subtle, arranger-specific rhythmic variations present in the full performance.

For the models that utilize token-based style conditioning, style is defined by discrete class tokens (`[ARRANGER_A]`, `[ARRANGER_B]`), with each token corresponding to a specific arranger. The total vocabulary, encompassing all event types for both representations, comprises 363 tokens.

B. Model Architecture

To investigate how best to represent and integrate musical content and arranger style in accompaniment generation, we design and compare three Transformer-based architectures that differ in their encoding strategies for content and style, as illustrated in Figure 2.

- **Model 1: Decoder-only architecture, with token-based content and style representation**—Inspired by Compose & Embellish [10], this model autoregressively generates performance tokens conditioned on an input sequence interleaving one-bar segments of lead sheet tokens M_{lead} and full performance tokens X_{perf}. The interleaving ensures the accompaniment generation stays faithful to the closest lead sheet melody within the same bar. An arranger style token prepended to the sequence conditions the model toward the target arranger's stylistic patterns. In ablation studies, we also experimented with alternative style token placements (e.g., per-bar tokens) and with lead sheets omitting chord information, to examine their influence on style consistency and melody preservation.
- **Model 2: Encoder-decoder architecture, with embedding-based content representation and token-based style representation**—The bidirectional Transformer encoder transforms 8-bar lead sheet tokens into contextualized content embeddings. The decoder is initialized with a style token and autoregressively generates performance tokens, conditioning on the style token and previously generated tokens, while integrating the content embeddings via cross-attention to maintain alignment with the lead sheet.
- **Model 3: Encoder-decoder architecture, with token-based content representation and embedding-based style representation**—This model employs a β-VAE [25] framework for enhanced content–style disentanglement, using a bidirectional Transformer encoder to extract a global style embedding from an 8-bar reference performance segment. The decoder autoregressively generates interleaved lead sheet and performance tokens, incorporating the style embedding via an in-attention mechanism [21]. Ablation studies investigate the effect of different reference segments on disentanglement by comparing cases where the reference segment is either the same as or adjacent to the decoder's target 8-bar segment.

Models 1 and 2 optimize negative log-likelihood loss over autoregressive token predictions. Model 3 additionally includes

[2]https://github.com/joshuachang2311/chorder

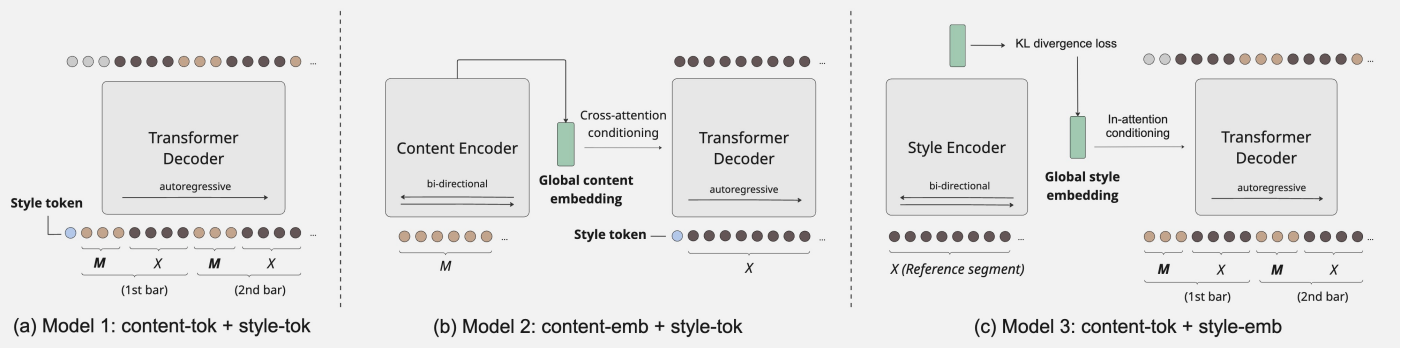

Fig. 2. Model variants with different conditioning schemes and representations of content and style. (a) Token-based content & style representations, (b) Embedding-based content with token-based style representations, (c) Token-based content with embedding-based style representations.

a Kullback-Leibler (KL) divergence regularization term to enforce a structured latent space for style.

C. Implementation and Training Details

All three models are built upon a Transformer architecture with a consistent set of shared hyperparameters (e.g., 12 layers, 8 attention heads). Model 1 is a decoder-only architecture that processes sequences up to 1,024 tokens. In contrast, Models 2 and 3 utilize an encoder-decoder framework, where their bidirectional encoders are designed to extract global embeddings from 8-bar musical segments for content or style conditioning. Model 3 further incorporates a β-VAE framework with a 128-dimensional latent style embedding, applying cyclical KL annealing [26] and a free-bits [27] strategy to boost content-style disentanglement. All three models are trained using the Adam optimizer with a cosine annealing learning rate schedule that includes a linear warmup phase. For data augmentation, we apply on-the-fly random pitch transposition (± 6 semitones) to both melody and chords. We set the batch size to four for Model 1, and to eight for the other two models.

IV. Experimental Results

A. Dataset

We adopt a symbolic dataset of piano cover performances adapted from the Pop2Piano corpus [7], containing high-quality transcriptions of popular songs performed by two distinct arrangers, referred to as Arranger A and Arranger B. These arrangers were selected for their clearly distinguishable accompaniment styles, which enables systematic arranger-specific style transfer evaluation.

The dataset comprises 811 pieces from Arranger A (on average 87 bars per piece) and 581 pieces from Arranger B (on average 92 bars per piece), totaling approximately 85 hours of music. We split the data for each arranger using an 8:1:1 ratio into training, validation, and test sets, ensuring balanced representation of both arranger styles.

To establish an evaluative baseline, we analyzed key distinguishing features such as rhythmic intensity, polyphony, and pitch range. The resulting statistics shown in Table I were calculated as detailed in the Objective Evaluation section.

TABLE I
Statistical properties of the dataset, reported as Mean ± Standard Deviation.

Arranger	**Rhythmic Intensity**	**Polyphony**	**Pitch Range**
A	0.44 ± 0.20	3.92 ± 1.73	49.52 ± 10.53
B	0.53 ± 0.21	5.74 ± 1.66	56.14 ± 5.87

B. Experimental Setup

For evaluation, we selected 100 distinct lead sheets from the test set. Each lead sheet served as the basis for generating 8-bar accompaniment segments targeting both Arranger A and Arranger B styles.

Models 1 and 2 conditioned generation on style tokens. For each lead sheet, two 8-bar segments were generated per target style, yielding 200 samples per style for each model. In contrast, Model 3 conditioned generation on style embeddings, using randomly selected 8-bar reference segments from the target arranger's training data for each generation, also producing 200 samples per style.

At the inference time, nucleus sampling (top-$p = 0.9$) with temperature $\tau = 1.2$ was used for decoding. For Model 3, the style embedding was set to the mean of the learned posterior distribution ($z_k = \mu_k$), excluding variance for improved generation stability.

C. Objective Evaluation

Evaluating arranger-specific style transfer demands assessment from two complementary perspectives: 1) whether the generated accompaniment captures the target arranger's stylistic patterns, and 2) whether it faithfully preserves the melody content specified in the lead sheet. To this end, we design a set of objective metrics that reflect both style matching and melodic fidelity, evaluated on 8-bar musical segments.

*1) **Style Matching**:* We use **Overlapping Area (OA)** [28], [29] to quantify the similarity between the feature distributions of generated samples and those from the target arranger's corpus. OA measures the area of overlap between two estimated probability density functions; a higher OA value indicates greater similarity. Ideally, generated samples conditioned on a specific arranger's style should yield a higher OA score

when compared with that arranger's reference distribution (e.g., $generated_{styleA}$ vs. $real_{styleA}$) than with a non-target arranger (e.g., $generated_{styleA}$ vs. $real_{styleB}$).

To evaluate whether the generated outputs successfully capture the target style, we first analyzed the original performances of the two arrangers to identify their key distinguishing characteristics. Our analysis revealed that their primary stylistic differences manifest in their choices of rhythmic intensity, polyphonic texture, and pitch range. Therefore, we assess style matching by analyzing the generated samples along these three key musical features:

- ***Rhythmic Intensity***: A bar-level feature measuring the rhythmic density of each bar, calculated by the number of note onsets normalized by bar duration.

$$s^{\text{rhym}} = \frac{1}{B}\sum_{b=1}^{B}\mathbf{1}(n_{onset,b}, \geq 1)\,, \tag{1}$$

  where B is the number of sub-beats in a bar and $\mathbf{1}(\cdot)$ is the indicator function.
- ***Polyphony***: Also evaluated at the bar level, this feature reflects the average number of notes (denoted as n) being hit or held simultaneously in a sub-beat, representing harmonic richness.

$$s^{\text{poly}} = \frac{1}{B}\sum_{b=1}^{B}(n_{onset,b} + n_{hold,b})\,. \tag{2}$$

- ***Pitch Range***: A segment-level feature that captures the span between the highest and lowest pitches across the entire 8-bar segment.

Through these measures, we gain a principled view of how convincingly each model captures and reproduces a target arranger's stylistic pattern.

*2) **Melodic Fidelity**:* Beyond style, our framework is designed to uphold the core melodic line of the lead sheet (i.e., the musical anchor) while letting the accompaniment vary stylistically. Since arrangers often introduce subtle timing or durational adjustments, strict note-to-note alignment is too rigid. Instead, we focus on preserving the correct sequential order of melody pitches: whether the generated performance retains the essential pitch sequence of the lead sheet in each bar, in the correct order, though not necessarily contiguously.

To measure this, we adopt the *Longest Common Subsequence (LCS)* algorithm, comparing the ordered sequence of melody pitches from the lead sheet with all pitches appearing in the generated accompaniment on the bar level. We denote the overall melodic fidelity score as:

$$\textit{Melodic Fidelity} = \frac{1}{B}\sum_{b=1}^{B}\frac{\#\text{LCS}(P_b^{\text{lead}}, P_b^{\text{gen}})}{\#P_b^{\text{lead}}}\,, \tag{3}$$

where $P_{\text{bar}}^{\text{lead}}$ and $P_{\text{bar}}^{\text{gen}}$ denote the ordered pitch sequences in the lead sheet and generated full accompaniment for that bar, respectively. The *LCS* function returns the longest subsequence present in both sequences in the correct order.

This ratio (ranging from 0 to 1) indicates how completely the lead sheet melody is respected within each performance bar. We calculate the average over 8-bar segments to summarize its melodic fidelity.

D. User Study

While objective metrics provide quantitative insights, they may not fully capture the perceptual nuances of musical style. Therefore, to assess how our generated music is perceived by human listeners, we conducted a subjective listening study.

We recruit 38 participants with diverse musical backgrounds for a 20-minute listening test. Two questionnaire versions are used, each containing four 8-bar songs excluded from training, with two songs for style transfer A→B and two for B→A.

For each song, the evaluation procedure is as follows: participants first listen to the original piano cover performance by the source arranger to familiarize themselves with the melody and style. Then, they listen to three anonymized and randomly-ordered target-style performances generated from the source lead sheet using three methods: Model 1, Model 2, and the ground-truth piano cover by the target arranger. Participants are asked to rate each performance on a 5-point Likert scale (from 1 to 5; the higher the better) for:

- **Melodic Fidelity (M)**: To what extent does this performance preserve the melody of the source version?
- **Style Matching (S)**: To what extent does this performance align with the target arranger's style?
- **Overall Quality (Q)**: How would you rate the overall musical quality of this performance?

E. Results Analysis

Our objective and subjective evaluations revealed several consistent trends. For style matching, objective metrics, as summarized in Figure 3, show that all models successfully generate outputs that are statistically closer to their intended target style than to the non-target style. Polyphony and pitch range emerged as a strong stylistic marker that was well-captured by all models. While the differences in rhythmic intensity were more subtle, this feature was also consistently reflected in the generated samples. Among the tested models, Model 1 generally exhibited a strong alignment with the target style, achieving higher OA across the evaluated features.

Ablation studies provided further insights into the models' robustness. For Model 1, variations such as applying style tokens on a per-bar basis or removing chord information from the lead sheet resulted in only negligible changes to its performance. Similarly, for Model 3, experiments with different reference segments for the style encoder showed a limited influence on the outcomes.

Regarding melodic fidelity, all models maintained a high degree of content preservation, with aggregated results shown in Table II. The token-based Model 1 achieved the highest average fidelity scores, with its LCS ratio ranged between 91% and 97%. This was followed by Model 3 (88–94%) and Model 2 (89–91%), suggesting that all architectures effectively preserved the core melodic information from the lead sheets.

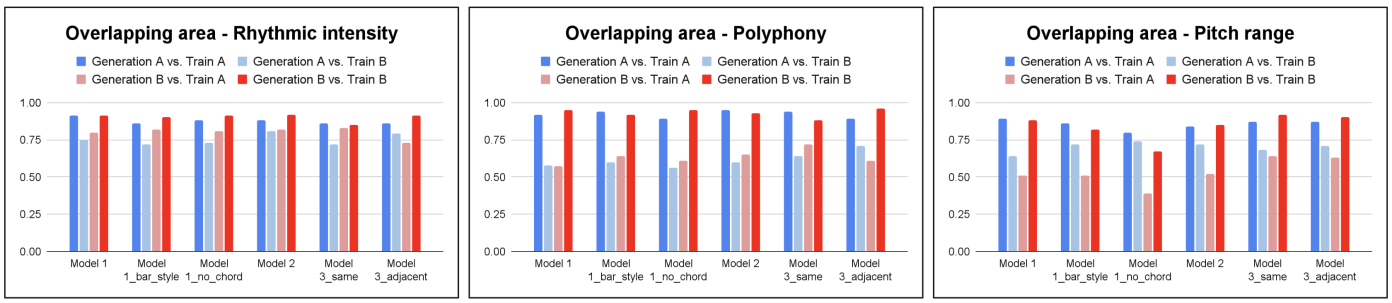

Fig. 3. Style matching performance measured by Overlapping Area (OA) on three key features. A higher OA score signifies greater distributional similarity to a reference style. The figure illustrates that for all models, the generated outputs are statistically much closer to their intended target style's distribution than to the non-target style's, demonstrating effective style discrimination.

TABLE II
OBJECTIVE EVALUATION RESULTS: AVERAGE MELODIC FIDELITY PER TARGET ARRANGER

	Arranger A	**Arranger B**
Model 1	0.97 ± 0.10	0.91 ± 0.17
– bar-level style	0.97 ± 0.10	0.96 ± 0.10
– w/o chord	0.95 ± 0.13	0.94 ± 0.13
Model 2	0.91 ± 0.23	0.89 ± 0.23
Model 3	0.92 ± 0.16	0.88 ± 0.20
– adjacent	0.93 ± 0.15	0.94 ± 0.15

TABLE III
USER STUDY MOS RESULTS (AGGREGATED OVERALL)

	Melodic Fidelity	**Style Matching**	**Overall Quality**
Model 1	3.35 ± 0.95	**3.24** ± 1.01	**3.28** ± 0.97
Model 2	2.75 ± 1.26	2.63 ± 1.23	2.91 ± 1.14
Real data	**3.53** ± 1.04	3.20 ± 1.16	3.27 ± 1.07

In the user study, the perceptual differences between the models became more pronounced, as detailed in Table III. Model 1 was rated notably higher than Model 2 across all three criteria: style matching (3.24 vs. 2.63), melodic fidelity (3.35 vs. 2.75), and overall quality (3.28 vs. 2.91). Its scores for style matching and overall quality also approached the ratings for ground-truth performances.

V. DISCUSSIONS AND FUTURE DIRECTIONS

Our experimental results present several key insights. A notable phenomenon was observed in the objective melodic fidelity scores (Table II): all models consistently scored slightly lower when targeting Arranger B's style. This suggests that the models successfully captured not only the accompaniment patterns but also Arranger B's tendency to perform with more melodic variations. Consequently, the generated outputs, while stylistically accurate, deviated marginally more from the strict lead sheet, highlighting the models' ability to learn even subtle performance traits.

A more central finding is the discrepancy between objective and subjective results. While all three models achieved comparable performance on objective style-matching metrics, a clear preference for the simplest, token-based architecture (Model 1) emerged in human evaluation. This suggests that while current metrics can capture broad statistical features, they may fail to reflect the subtle, perceptual nuances that define a convincing musical style for a human listener.

The trade-off between token-based and embedding-based representations is also central to our findings. The success of the token-based model highlights the power of a strong content anchor; when content is explicitly defined via a lead sheet, a simple style signal can be sufficient for high-quality transfer between a known set of styles. However, this approach struggles with zero-shot transfer to unseen arrangers, where embedding-based methods—capable of learning continuous style representations—show greater potential.

This points to several avenues for future work. One promising direction is to develop richer and more perceptually aligned representations for style, moving beyond broad statistical features to model more specific concepts such as characteristic rhythmic motifs or harmonic voicings. This would also necessitate creating more nuanced objective metrics that better correlate with human perception.

Another important focus is to enhance the disentanglement capabilities of Transformer-VAEs. Improving the training objectives of the embedding-based framework could lead to more effective zero-shot style transfer, significantly increasing the flexibility and generalization ability of the model.

VI. CONCLUSIONS

In this paper, we have addressed the nuanced challenge of arranger-specific style transfer for single-track piano accompaniments. We introduced a framework grounded on the use of lead sheets as a robust, style-agnostic content anchor and systematically compared token-based versus embedding-based Transformer architectures to evaluate their efficacy. Our comprehensive evaluations revealed that a straightforward token-based model consistently outperforms more complex embedding-based variants, achieving superior results in objective and subjective assessments of content preservation and style matching.

This key finding provides strong empirical evidence that for this task, a well-defined, explicit content representation can be more critical than the complexity of the style conditioning model itself, offering a practical benchmark for

music generation. Although effective, the limitation of this token-based approach in zero-shot transfer to unseen styles highlights a promising direction for future research: enhancing the disentanglement capabilities of embedding-based models, such as Transformer-VAEs, to develop a more generalized and flexible style transfer model.

References

[1] D.-V.-T. Le, L. Bigo, D. Herremans, and M. Keller, "Natural language processing methods for symbolic music generation and information retrieval: A survey," *ACM Comput. Surv.*, vol. 57, no. 7, 2025.

[2] C.-Z. A. Huang *et al.*, "Music Transformer: Generating music with long-term structure," in *Proc. Int. Conf. Learning Representations*, 2018.

[3] Y.-S. Huang and Y.-H. Yang, "Pop Music Transformer: Beat-based modeling and generation of expressive pop piano compositions," in *Proc. ACM Multimedia*, 2020.

[4] P. Pasquier, J. Ens, N. Fradet, *et al.*, "MIDI-GPT: A controllable generative model for computer-assisted multitrack music composition," in *Proc. AAAI*, 2025.

[5] Y. Wang, S. Wu, J. Hu, *et al.*, "NotaGen: Advancing musicality in symbolic music generation with large language model training paradigms," *arXiv preprint arXiv:2502.18008*, 2025.

[6] Z. Guo and S. Dixon, "Moonbeam: A MIDI foundation model using both absolute and relative music attributes," *arXiv preprint arXiv:2505.15559*, 2025.

[7] J. Choi and K. Lee, "Pop2Piano : Pop audio-based piano cover generation," in *Proc. IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 2023.

[8] C.-P. Tan, S.-H. Guan, and Y.-H. Yang, "PiCoGen: Generate piano covers with a two-stage approach," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2024.

[9] C.-P. Tan, H. Ai, Y.-H. Chang, S.-H. Guan, and Y.-H. Yang, "PiCoGen2: Piano cover generation with transfer learning approach and weakly aligned data," in *Proc. Int. Soc. Music Information Retrieval Conf.*, 2024.

[10] S.-L. Wu and Y.-H. Yang, "Compose & Embellish: Well-structured piano performance generation via a two-stage approach," in *Proc. IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 2023.

[11] G. Hadjeres, F. Pachet, and F. Nielsen, "DeepBach: A steerable model for bach chorales generation," in *Proc. Int. Conf. Machine Learning*, 2017, pp. 1362–1371.

[12] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang, "MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," in *Proc. AAAI*, 2018.

[13] J. Zhao and G. Xia, "AccoMontage: Accompaniment arrangement via phrase selection and style transfer," in *Proc. Int. Soc. Music Information Retrieval Conf.*, 2021.

[14] J. Huang, K. Chen, and Y.-H. Yang, "Emotion-driven piano music generation via two-stage disentanglement and functional representation," in *Proc. Int. Soc. Music Information Retrieval Conf.*, 2024.

[15] G. Brunner, Y. Wang, R. Wattenhofer, and S. Zhao, "Symbolic music genre transfer with cycleGAN," in *Proc. Int. Conf. Tools with Artificial Intelligence*, 2018.

[16] W. T. Lu, L. Su, *et al.*, "Transferring the style of homophonic music using recurrent neural networks and autoregressive model," in *Proc. Int. Soc. Music Information Retrieval Conf.*, 2018, pp. 740–746.

[17] O. Cífka, U. Şimşekli, and G. Richard, "Groove2groove: One-shot music style transfer with supervision from synthetic data," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 28, pp. 2638–2650, 2020.

[18] Y.-N. Hung, I. Chiang, Y.-A. Chen, Y.-H. Yang, *et al.*, "Musical composition style transfer via disentangled timbre representations," in *Proc. Int. Joint Conf. Artificial Intelligence*, 2019.

[19] D.-V.-T. Le and Y.-H. Yang, "METEOR: Melody-aware texture-controllable symbolic orchestral music generation via Transformer VAE," in *Proc. Int. Joint Conf. Artificial Intelligence*, 2025.

[20] H. Zhang and S. Dixon, "Disentangling the Horowitz factor: Learning content and style from expressive piano performance," in *Proc. IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 2023.

[21] S.-L. Wu and Y.-H. Yang, "MuseMorphose: Full-song and fine-grained piano music style transfer with one Transformer VAE," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 31, 2023.

[22] O. Cífka, A. Ozerov, U. Şimşekli, and G. Richard, "Self-supervised VQ-VAE for one-shot music style transfer," in *Proc. IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 2021.

[23] K. Choi, C. Hawthorne, I. Simon, M. Dinculescu, and J. Engel, "Encoding musical style with Transformer autoencoders," in *Proc. Int. Conf. Machine Learning*, 2020, pp. 1899–1908.

[24] A. L. Uitdenbogerd and J. Zobel, "Manipulation of music for melody matching," in *Proc. ACM Multimedia*, 1998.

[25] I. Higgins *et al.*, "Beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learning Representations*, 2017.

[26] H. Fu *et al.*, "Cyclical annealing schedule: A simple approach to mitigating KL vanishing," *arXiv preprint arXiv:1903.10145*, 2019.

[27] D. P. Kingma *et al.*, "Improved variational inference with inverse autoregressive flow," *Proc. Advances in Neural Information Processing Systems*, 2016.

[28] L.-C. Yang and A. Lerch, "On the evaluation of generative models in music," *Neural Computing and Applications*, vol. 32, no. 9, pp. 4773–4784, 2020.

[29] H.-T. Hung, C.-Y. Wang, Y.-H. Yang, and H.-M. Wang, "Improving automatic Jazz melody generation by transfer learning techniques," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2019.